

Benchmarking data-driven inversion methods for the estimation of local CO₂ emissions from XCO₂ and NO₂ synthetic satellite images

Diego Santaren¹, Janne Hakkarainen², Gerrit Kuhlmann³, Erik Koene³, Frédéric Chevallier¹, Iolanda Ialongo², Hannakaisa Lindqvist², Janne Nurme², Johanna Tamminen², Laia Amorós², Dominik Brunner³ and Grégoire Broquet¹

¹Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France

²Finnish Meteorological Institute, Helsinki, Finland

³Swiss Federal Laboratories for Materials Science and Technology (EMPA), Dübendorf, Switzerland

Correspondence to: diego.santaren@lsce.ipsl.fr

Abstract.

The largest anthropogenic emissions of carbon dioxide (CO₂) come from local sources such as cities and power plants. The upcoming Copernicus CO₂ Monitoring Mission (CO2M) will provide satellite images of the CO₂ and NO₂ plumes associated with these sources at a resolution of 2 km × 2 km and with a swath of 250 km. These images could be exploited with atmospheric plume inversion methods to estimate local CO₂ emissions at the time of the satellite overpass and the corresponding uncertainties. To support the development of the operational processing of satellite column-averaged CO₂ dry air mole fraction (XCO₂) and tropospheric column NO₂ imagery, this study evaluates “data-driven inversion methods”, i.e., computationally light inversion methods that directly process information from satellite images, local winds and meteorological data, without resorting to computationally expensive dynamical atmospheric transport models. We have designed an objective benchmarking exercise to analyse and compare the performance of five different data-driven inversion methods: two implementations with different complexity for the cross-sectional flux approach (CSF and LCSF) and one implementation for the Integrated Mass Enhancement (IME), the Divergence (Div) and the Gaussian Plume model inversion (GP) approaches. This exercise is based on pseudo-data experiments with simulations of synthetic “true” emissions, meteorological and concentration fields, and CO2M observations in a domain of 750 km × 650 km centred on Eastern Germany over 1-year. The performance of the methods is quantified in terms of accuracy in the single-image (from individual images) or annual average (from the full series of images) emission estimates and in terms of number of instant estimates for the city of Berlin and 15 power plants in this domain. Several ensembles of estimations are conducted, using different scenarios for the available synthetic datasets. These ensembles are used to analyse the sensitivity of the performance to the loss of data due to cloud cover, to the uncertainty in the wind or to the added value of simultaneous NO₂ images. The GP and the LCSF methods generate the most accurate estimates from individual images. The deviations between the emission estimates and the true emissions from these two methods have similar Interquartile Ranges (IQR):

32 between ~20 % and ~60 % depending on the scenario. When taking the cloud cover into account, these methods produce
33 respectively 274 and 318 instant estimates from the ~500 daily images that cover significant portions of the plumes from the
34 sources. Filtering the results based on the associated uncertainty estimates can improve the statistics of the IME and CSF
35 methods, but at the cost of a large decrease in the number of estimates. Due to a reliable estimation of uncertainty and thus a
36 suitable selection of estimates, the CSF method achieves similar if not better statistics of accuracy for instant estimates
37 compared to the GP and LCSF methods after filtering. In general, the performances for retrieving single-image estimates are
38 improved when, in addition to XCO₂ data, collocated NO₂ data are used to characterise the structure of plumes. With respect
39 to the estimates of annual emissions, the root mean square errors (RMSE) are for the most realistic benchmarking scenario
40 20 % (GP), 27 % (CSF), 31 % (LCSF), 55 % (IME) and 79 % (Div). This study suggests that the Gaussian plume and/or the
41 cross-sectional approaches are currently the most efficient tools to provide estimates of CO₂ emissions from satellite images
42 and their relatively light computational cost will enable analysis of the massive amount of data provided by future missions
43 of satellite XCO₂ imagery.

44 **1 Introduction**

45 The satellite imagery of column-averaged CO₂ dry air mole fractions (XCO₂) has been identified as an essential
46 component of a future atmospheric observing system to monitor anthropogenic CO₂ emissions, and in particular to detect
47 and monitor hotspot atmospheric plumes and thus emissions, in order to verify emission reductions or assess national
48 budgets (Ciais et al., 2015; Pinty et al., 2017). The Copernicus CO₂ Monitoring (CO2M mission was designed to meet these
49 objectives with a constellation of two to three Low Earth Orbit (LEO) satellites flying in a sun-synchronous low-earth orbit
50 crossing the Equator around 11:30 local time. Each satellite will carry an imaging spectrometer providing images of XCO₂
51 and of NO₂ tropospheric column densities (referred to as NO₂ hereinafter) along a 250 km wide swath with a resolution of 2
52 km × 2 km (Sierk et al., 2019). Current satellite missions, like Sentinel-5 Precursor (Sentinel-5P) and the third Orbiting
53 Carbon Observatory (OCO-3, when targeting specific sources in its Snapshot Area Map -SAM- mode), already deliver NO₂
54 column-density and XCO₂ images, albeit, for the former, at a resolution coarser than CO2M, and for the latter, over areas
55 and at a frequency much smaller than with CO2M. Upcoming missions, such as Global Observing SATellite for Greenhouse
56 gases and Water cycle (GOSAT-GW, Kasahara et al., 2020), MicroCarb (in its “city-mode”, Pascal et al., 2017) and Twin
57 ANthropogenic Greenhouse gas Observers (TANGO, Landgraf et al., 2020), are expected to increase the amount of CO₂ and
58 NO₂ images of the plumes from emission hotspots.

59 Operational services are being developed such as the Copernicus capacity for anthropogenic CO₂ emissions monitoring
60 and verification support (CO2MVS, Pinty et al., 2017; Janssens-Maenhout et al., 2020), to process these XCO₂ and NO₂
61 images for the monitoring of emissions in a systematic and global way at spatial and time scales that are relevant for
62 policymakers and to support emission mitigation actions. Plume inversion systems are used to derive estimates of the CO₂
63 emissions from local sources using satellite images of the corresponding atmospheric plumes. One of the key elements of

64 operational services will thus be standard plume inversion methods providing precise and reliable data in an automated and
65 fast manner. Various plume inversion approaches and implementations are now regularly used to process the existing
66 spaceborne atmospheric plumes images (Varon et al., 2018; Zheng et al. 2020; Kuhlmann et al., 2021; Nassar et al., 2021;
67 Jacob et al., 2022; Hakkarainen et al., 2023a). Therefore, there is a need to benchmark in a quantitative way the plume
68 inversion methods for the estimation of local emissions of CO₂, and more generally of greenhouse gases and pollutants.

69 Monitoring anthropogenic CO₂ emissions of point sources or cities from satellite XCO₂ images is challenging as
70 corresponding column-average enhancements are often small compared with the local fluctuations of the “background” CO₂
71 field due to biogenic CO₂ fluxes and to neighbour anthropogenic sources, and with the typical level of errors in the XCO₂
72 retrievals (Buchwitz et al., 2013). Despite this challenge, the potential of CO₂ imagers to estimate anthropogenic emissions
73 has been demonstrated with observing system simulation experiments (OSSEs) using synthetic data, for power plants
74 (Bovensmann et al., 2010), cities (Pillai et al., 2016; Broquet et al., 2018; Wang et al., 2020) and in a more general way, at
75 local to national scales (Santaren et al., 2021). Furthermore, several studies have shown that the joint analysis of co-located
76 NO₂ satellite observations strongly enhances the skill to detect the XCO₂ enhancement plumes from sources in XCO₂
77 images, and consequently to estimates the corresponding CO₂ emissions (Reuter et al., 2019; Kuhlmann et al., 2021). NO₂
78 observations are indeed characterised by a better signal-to-noise ratio and a generally small and low-amplitude background
79 field, due to the relatively short lifetime of nitrogen oxides (NO_x).

80 CO₂ emissions of large point sources and cities can be estimated from satellite images by plume inversion systems
81 integrating the observations with dynamical transport model simulations of atmospheric CO₂ concentrations (e.g., Broquet et
82 al., 2018; Ye et al., 2020; Santaren et al., 2021). In principle, the use of such dynamical models could support the analysis of
83 the 3D dynamical patterns of the observed plume and thus the accuracy of the inversion. They could also support the
84 derivation of the spatial distribution of the emissions within cities, and of the temporal variation of the emissions
85 corresponding to a plume in the hours preceding each satellite overpass. However they can be strongly impacted by
86 modelling errors which become critical at local scale, when trying to model plumes from emission hotspots over a few tens
87 to a few hundreds of kilometres (Brunner et al., 2023). Furthermore, their computational burden hampers their use for a
88 global and routine coverage of the sources in an operational context. *Data-driven plume inversion methods* appear to be
89 currently more suitable for such wide-scale applications (Ehret et al., 2022). These are computationally light inversion
90 methods that directly process information from satellite images and local winds and meteorological data (typically from
91 operational weather analyses), without resorting to dynamical atmospheric transport models.

92 The main data-driven approaches for estimating local emissions based on satellite images of plumes that have been tested
93 and analysed in a significant number of studies are:

94 1) the Integrated Mass Enhancement (IME) approach, which relates the total mass of plumes to the corresponding
95 emissions; it has been used for retrieving CH₄ emissions from airborne observations (Frankenberg et al., 2016) or from fine-
96 scale satellite data (Varon et al., 2018)

97 2) the Gaussian plume approach which extracts emissions from the fit of plume shapes by Gaussian functions and was
98 applied for instance to estimate power plant CO₂ emissions from OCO-2 satellite data (Nassar et al. 2017; 2021)

99 3) the cross-sectional flux approach which infers emissions from the fluxes passing through cross-sections of the plumes
100 and whose potential to estimate CO₂ emissions of power plants with CO₂ and NO₂ satellite imagery data was assessed, for
101 instance, by Kuhlmann et al. (2021)

102 4) the divergence (Div) approach, which derives emissions from the application of the divergence operator to fields of
103 fluxes and which was originally designed to estimate nitrogen oxide (NO_x) emissions from NO₂ data provided by the
104 TROPOMI satellite imagery (e.g. Beirle et al., 2019; 2021, 2023) and was more recently adapted to the quantification of CO₂
105 emissions (Hakkaraïnen et al., 2022). Contrarily to the other methods of this study, the Div method is generally used to
106 generate annual estimates from average fields extracted from multiple images.

107 Against this background, the aim of this study is to benchmark these four data driven plume inversion approaches for the
108 monitoring of CO₂ emission hotspots with CO2M images. We present a benchmarking framework to objectively evaluate
109 and compare the performance of different implementations of the four data-driven approaches (Sect. 2.1) to estimate CO₂
110 local emissions from such satellite data. For this purpose, we use one year of synthetic satellite observations closely
111 mimicking those expected from the upcoming CO2M mission (Sect. 2.2) that were generated in the European Space Agency
112 (ESA) funded SMARTCARB project from high-resolution atmospheric transport simulations (e.g. Brunner et al., 2019;
113 Kuhlmann et al., 2020). The emissions of the city of Berlin and 15 large power plants are estimated from these synthetic
114 satellite data and the ability of the different inversion methods is assessed by comparing their estimates to the corresponding
115 *true* values used by the atmospheric transport model. Performances of the different inversion approaches are evaluated for 1)
116 single-image estimates that are retrieved from daily images (Sect. 3) and, 2) annual estimates that are computed from the
117 inversion of one year of data (Sect. 4). Furthermore, performances are analysed for different scenarios regarding the data
118 used by the inversions, where the impacts of considering the cloud cover in the data, the uncertainties in the wind and the use
119 of collocated NO₂ data are assessed. Finally, results are discussed by analysing 1) the potential of ensemble approaches that
120 would gather different inversion methods and, 2) the trade-off between overall accuracy and number of estimates when the
121 cases are filtered based on the uncertainties in the estimates computed by the plume inversion methods (Sect. 5).

122 **2 Data and methods**

123 **2.1 Data-driven inversion methods**

124 Five different emission quantification methods are evaluated in this study: (1) the integrated mass enhancement method
125 (IME), (2) the cross-sectional flux (CSF) method, (3) the light cross-sectional flux (LCSF) method, (4) the Gaussian plume
126 (GP) method and (5) the divergence (Div) method. More precisely, what is studied here are specific configurations of certain
127 methods as is the case for the CSF and LCSF “methods” which are derived from the same general approach. But, hereinafter
128 we will refer to these configurations as methods to avoid weighing down the text. The general approaches have been widely

129 used and described in previous papers such as Varon et al. (2018) and Beirle et al. (2019, 2021). The specific
130 implementations of the CSF and Div methods tested here have been used extensively by the authors in previous studies
131 (Kuhlmann et al., 2019, 2020, 2021 and Hakkarainen et al., 2022). They have been slightly upgraded in the course of this
132 benchmarking exercise to improve their stability, accuracy, and capability of running in a fully automated way. Details of the
133 methods are presented in an accompanying study by Kuhlmann et al. (2023). Further details about the theory of the Div
134 method and its application are given in Koene et al. (2023) and Hakkarainen et al. (2022, 2023b). All algorithms and tools
135 used in this work have been integrated into a Python library for *data-driven emission quantification* (ddeg), which has been
136 made publicly available and is described in Kuhlmann et al. (2024). We provide below a short description of these methods
137 with an emphasis on their relative advantages and limitations and on the way they estimate uncertainty. The main features of
138 the methods are summarised in Table 1 and illustrated in Figure 1 and Figure A1. Table 1 also lists the computation times of
139 the methods calculated for the same inversion example using the same hardware. As the methods have all been implemented
140 in the same Python package, the timings are directly comparable.

141 All methods except the Div method can provide estimates derived from individual satellite images. The Div approach as
142 implemented here is based on the averaging of information contained within multiple images and hence typically delivers
143 annual estimates. We will hereinafter refer to the IME, CSF, LCSF and GP methods as single-image methods. These
144 methods share a common algorithmic sequence that starts with identifying clusters of enhancements above a background in
145 satellite images. Subsequently, these clusters are assigned to plumes from specific known sources, and finally, the emissions
146 of the corresponding sources are estimated. The plume detection combines the first two stages and can be used to discern
147 plumes from unreported sources; however the ability of the different approaches to detect unknown point sources has not
148 been studied here, as the primary focus is to analyse their potential to detect and process plumes of known sources from
149 CO₂M-like satellite images (see Sect. 2.2). It is worth mentioning that the divergence, cross-sectional flux and machine-
150 learning approaches are particularly well-suited for automatic detection of plumes from unknown sources (Zheng et al.,
151 2020; Beirle et al., 2021; Schuit et al., 2023). Moreover, as previously mentioned, a benefit of the CO₂M mission is the
152 availability of co-registered XCO₂ and NO₂ columns, which can further benefit the plume detection and emission
153 quantification steps.

154 Obtaining the column enhancements over the background can be achieved with different thresholding techniques as
155 detailed below. When it comes to NO₂, the global background field is insignificant but in the case of CO₂, its amplitude is
156 important and can vary significantly in space and time due to biogenic and other anthropogenic fluxes surrounding the
157 sources of interest and due to gradients in the background. Another common feature is the need for defining an effective
158 wind speed, which describes the average mass transport of CO₂ within the plumes. This a major challenge as wind speed
159 varies with altitude whereas satellite images contain integrated column measurements with no vertical resolution.
160 Additionally, the horizontal resolutions of wind products are generally different from those of satellite images. To address
161 these limitations, the methods determine effective winds in a more or less sophisticated manner.

162 Finally, all methods have implemented some quality control on their estimates. These checks are more or less restrictive
163 depending on the methods and may filter out, for example, cases with overlapping plumes originating from neighbouring
164 sources. Further details are provided in Kuhlmann et al. (2023). It is worth emphasizing the fact that our implementation of
165 the GP method discards values that are below 1/4 or beyond 4 times the “true” values averaged one hour before the satellite
166 overpass (10:00 to 11:00 UTC); this filtering stabilises the otherwise underdetermined inversion. Unlike the other methods,
167 the GP method thus uses a priori information about the source strength, which artificially improves its performance.

168 **2.1.1 Cross-sectional flux (CSF) inversion method**

169 The cross-sectional flux inversion method has been used in many studies such as for example the determination of CH₄
170 emissions of point sources from high-resolved satellite data for which its superiority over other methods has been
171 demonstrated within the framework of the study of Varon et al. (2018). In brief, this method calculates the fluxes through
172 single or multiple cross-sections of the plumes as the product of effective winds and integrals of column mass enhancements
173 along plume transects (line densities). Under the assumption of steady-state conditions, these fluxes are equivalent to the
174 emissions. The CSF method used in this study has been used by Kuhlmann et al. (2020, 2021) for the estimation of CO₂
175 emissions from CO₂ and NO₂ images. These studies have demonstrated that the inclusion of NO₂ observations significantly
176 increases the number and precision of the estimates.

177 The plume detection module of the CSF approach determines in a first stage the CO₂ or NO₂ pixels that are significantly
178 enhanced above the background with a statistical z-test (Kuhlmann et al., 2021). To perform this, a Gaussian kernel to
179 average local observations values is applied and the background field is at this stage computed by applying a median filter.
180 The parameters defining the z-test were carefully assessed in order to get enough valid pixels to describe a plume while
181 avoiding false detections (Kuhlmann et al. 2019). The detected pixels are then grouped by a labelling algorithm and assigned
182 to a source. Finally, a curve representing the centerlines of the plume is fitted to the detected pixels.

183 For the quantification of CO₂ emissions, the CSF method groups the detected plume pixels into sub-polygons along the
184 curved plume, whose width equals ~5 km (2-3 pixels of CO₂M data). All detected pixels within a sub-polygon are used to
185 construct a single estimate of the line density. Following Reuter et al. (2019), the CSF method assumes that the plume
186 transect follows a Gaussian behaviour, after removing the background signal with a normalised convolution. To obtain the
187 line densities, the integration of the fitted Gaussian functions does not require any additional computation as the line
188 integrals are simply equal to the amplitude parameters of the fitted Gaussian functions. Then, in order to be converted into
189 fluxes, line densities are multiplied by effective winds which are the horizontal winds at the corresponding source locations
190 and times of the satellite overpasses, vertically weighted by the GNFR-A/SNAP-1 emission profile (Brunner et al., 2019).

191 Finally, the CO₂ emission of a given source retrieved from a given satellite image is computed by averaging the CO₂
192 estimated fluxes of all the sub-polygons describing the plume downstream of the source. The uncertainty in the emission
193 estimate is then computed by propagation of the uncertainties in the line densities computation and in the wind; the

194 uncertainties in the line densities are extracted from the standard deviation of the sub-polygon estimates and capture mostly
195 satellite data noise through uncertainty in the Gaussian fitting.

196 When NO₂ data are used in conjunction with CO₂, detections of plumes are first performed for NO₂, while the CO₂ and
197 NO₂ enhancements are fitted simultaneously by Gaussian functions that share the same mean (or central location) and the
198 same standard deviation. Thus, the fit of CO₂ enhancements takes advantage of the better signal-to-noise ratio of NO₂ data
199 by better constraining the parameters of the Gaussian functions, which provides more accurate estimates of CO₂ line
200 densities and hence CO₂ emissions.

201 **2.1.2 Light cross-sectional flux (LCSF) inversion method**

202 The light cross-sectional flux method shares the same theoretical foundations as the CSF method, but its implementation
203 is largely different. It is derived from the method originally developed by Zheng et al. (2020) to estimate the CO₂ emissions
204 of cities and industrial areas in China that produce atmospheric plumes clearly detectable in transects of OCO-2 data which
205 are characterised by a resolution of few km² and by a swath about 10 km wide, which is almost 25 times narrower than the
206 ~250 km wide swath of the CO2M instruments. This method has been applied to the routine and automatic estimation of
207 isolated clusters of CO₂ emissions worldwide (Chevallier et al., 2020) and to study the temporal variability of the emissions
208 based on several years of OCO-2 and OCO-3 data (Chevallier et al., 2022). The method has undergone significant
209 modifications for this comparative study, where the location of the emission sources is known, in order to fully harness the
210 potential of high-resolution satellite imagery.

211 For a given source and satellite overpass, the LCSF method performs a simple detection of the plume by extracting from
212 the satellite image an area which is 100 km wide in across-wind (perpendicular) direction and which extends downwind the
213 source over a distance equal to the distance travelled by the wind in one hour. The method then selects the pixels of the
214 extracted area where XCO₂ or NO₂ enhancements – simply defined as the difference between data values and the average
215 data of the area – are greater than the spatial variability, i.e. the standard deviation of the data contained within the area.

216 The quantification of the source emission is then performed on each selected enhancement by extracting again a 100 km
217 wide across-wind area centred at the enhancements and extending 10 km (~5 CO2M pixels) downwind from the
218 enhancements. The sums of a linear term accounting for large scale variations in the background fields and a Gaussian
219 function describing the plume cross-section perpendicular to the wind direction are then fitted to the data contained within
220 these areas. The plume detection and fitting of the enhancements can be carried out in the same way when NO₂ data are
221 available. And, standard deviations and means of the Gaussian functions fitted with NO₂ data are then used for fitting CO₂
222 enhancements; CO₂ data constrain in this case only the amplitudes of the CO₂ Gaussian functions. This allows transferring
223 information derived from NO₂ data when estimating CO₂ emissions from CO₂ data.

224 CO₂ line densities are, as for the CSF method, derived from the Gaussian functions fitted with CO₂ data and converted
225 into emission estimates by the multiplication of an effective wind. For the LCSF method, this effective wind is extracted at
226 the location of the enhancements and at an altitude above ground of 100 m, as preliminary tests have shown that extracting

227 winds at the altitude of 100 m yields, for the LCSF approach, better inversion results compared to other altitudes or
228 alternative methods of computing the effective winds. This result may be reflecting a trade-off between the need to account
229 for emission injection heights higher than 100 m when considering isolated power plants, and lower than 100 m when
230 considering the mix of sources within cities, whose emissions are not dominated by large power plants (Brunner et al.,
231 2024). The automatic process of sources limits the ability to derive a case by case selection of the height for the wind
232 extraction, but a finer option for future analysis might be to discriminate this selection as a function of the type of target
233 (considering at least isolated power plants vs. urban areas).

234 Finally, under steady-state atmospheric conditions, the cross-sectional CO₂ flux derived at each selected enhancement is
235 equivalent to the upwind source emissions. Therefore, as several enhancements belonging to a same atmospheric signature of
236 a source are generally processed, the algorithm produces multiple individual estimates of the source emission; the estimate
237 computed by the method for a given source and from a given image is then computed as the median value of these individual
238 estimates; the use of the median helping to reduce the impact of outliers. Moreover, uncertainties in the individual estimates
239 provided by the LCSF method are computed by propagation of the errors derived by the fitting algorithm when generating
240 the line densities; uncertainties in the final estimates are finally the median of these uncertainties.

241 **2.1.3 Gaussian plume (GP) inversion method**

242 The Gaussian plume inversion approach assumes that observed plumes can be described with Gaussian plume models. This
243 approach has been widely used such as for example in the determination of CH₄ point source emissions (Varon et al., 2018),
244 the use of OCO-2 data to quantify CO₂ emissions from power plants (Nassar et al., 2017), or in a framework to estimate at
245 the global scale CO₂ emissions from large cities and point sources (Wang et al., 2020). Compared to previous Gaussian
246 plume inversions, the GP inversion method used in this work allows the Gaussian plume model (like the CSF method) to
247 handle curved plumes (see Sect 3.2.1 in Hakkarainen et al., 2023b).

248 The detection of plumes, i.e. of the CO₂ or NO₂ enhancements from the background, is carried out using the same
249 algorithm as for the CSF method. Then, the inversion uses a Levenberg-Marquardt least-squares optimization to find the
250 optimal parameters of the Gaussian functions fitting the enhancements and, of the Bézier curves describing the centre lines
251 of the plumes (Hakkarainen et al., 2023b). If NO₂ data and CO₂ data are simultaneously available, then the Gaussian plume
252 model is first fitted to the NO₂ observations and the optimised parameters regarding the plume shape are subsequently used
253 as first guesses for the fitting to CO₂ observations. These derived parameters are constrained to remain close to the optimised
254 parameters obtained from the fitting of NO₂ data. Finally, the uncertainties in the Gaussian plume estimates are obtained by
255 propagation of the uncertainties in the fitted parameters for the wind speed and for the source strength.

256 To ensure the convergence of the minimization algorithm, first-guessed values of the fitted parameters need to be
257 carefully prescribed: parameters of the centre-line curves, for example, are initialised from the curves retrieved by the plume
258 detection algorithm, and the initial wind speed is calculated as in the CSF method (see Sect. 2.1.1). Most importantly, the
259 prior values of emission parameters are set to the *true* summertime source emission strength. Thus, unlike any of the other

260 methods studied in this work, the GP method integrates an important constraint on the emissions which implies that the
261 estimated values, hence the method's performance, are not entirely determined by the information contained within the
262 synthetic satellite observations alone. This limitation should be taken into account when applying this method to invert from
263 real satellite data emissions of sources whose amplitudes are barely known.

264 **2.1.4 Integrated mass enhancement (IME) method**

265 The IME method integrates the total mass enhancements of CO₂ or NO₂ above the background that can be associated with
266 detectable plumes. Then, following Frankenberg et al. (2016), the relationship between IMEs and emissions (Q) can be
267 approximated by a linear relationship defined by the residence times (τ) of the species within the plumes (Eq. 1):

$$Q = \frac{1}{\tau} IME \quad (1)$$

$$\tau = \frac{U_{eff}}{L} \quad (2)$$

268 The residence time can in turn be expressed as a characteristic plume length L divided an effective wind speed U_{eff} (Eq.
269 2) . For example, Varon et al. (2018), who applied the IME method with CH₄ observations, derived U_{eff} from 10 m wind
270 speeds using large eddy simulations (LES). Here, the plume detection algorithm which identifies either CO₂ or NO₂
271 enhancements from the background is the same as the one used by the CSF and GP methods, but the detected area of the
272 plume over which the integration is performed is dilated using a circular kernel in order to increase the number of integrated
273 pixels (Hakkarainen et al., 2023b). Missing values are filled using a normalised convolution and estimates are rejected when
274 less than 75 % of valid pixels are available for the detected plume. The characteristic length L is computed from the centre-
275 line of the plume as the arc length to the most distant detected pixel minus 10 km, but at least 10 km. Moreover, the effective
276 wind speed U_{eff} is extracted by using the same vertically weighted average as the CSF method. If NO₂ observations are used
277 in conjunction with CO₂ observations, the integration area is established by the application of the plume detection algorithm
278 with NO₂ data. Then, to estimate CO₂ emissions, the IME is calculated over this area with CO₂ observations. Finally, the
279 uncertainty in the IME estimates is computed by propagation of uncertainty from the single sounding precision of satellite
280 data and an estimate of the uncertainty in the wind speed.

281 **2.1.5 Divergence method**

282 The divergence method, initially introduced by Beirle et al. (2019, 2021), was used to estimate NO_x emissions based on
283 TROPOMI NO₂ observations. For this study, the method has been modified in order to estimate CO₂ emissions, as outlined
284 in Hakkarainen et al. (2022) where a detailed theoretical analysis of this approach can be found in the supplementary
285 material. The divergence method is based on the continuity equation at steady state (Jacob, 1999), where the divergence of a
286 vector field F (flux) is defined as the difference between emissions E and sinks S (Eq. 3):

$$\nabla \cdot F = E - S \quad (3)$$

288
$$F = (F_x, F_y) = (\Delta I \cdot U_{eff}, \Delta I \cdot V_{eff}) \quad (4)$$

289 Since CO₂ lifetime is extremely long, the sink term can be neglected. However, before applying the divergence operator to
290 XCO₂ images, the atmospheric background needs to be removed in order to extract purely the XCO₂ enhancements. For this
291 purpose, a median filter is applied to the data and the resulting field is subtracted from the original data. Moreover, in order
292 to improve the accuracy of the estimates when CO₂ noise levels are high, data first undergo a denoising process using a 5×5
293 pixel mean filter. The flux field F is then defined at each pixel by the Eq. 4 where ΔI is the vertical column density
294 enhancement above background, and U_{eff} and V_{eff} are the eastward and northward winds, respectively, interpolated at the
295 location of the pixel and at the time of the satellite observations, and vertically averaged using the GNFR-A/SNAP-1
296 emission profile (Brunner et al., 2019).

297 Divergence maps are computed from the mass flux field using a finite difference approximation. The divergence map is
298 then averaged over a long period to enhance the emission signal, while reducing the impact of noise and the spatio-temporal
299 variations of the CO₂ background. Here, divergence maps are averaged over one year. In theory, the divergence method can
300 also be used to estimate emissions from single-overpass images such as the cross-sectional flux method (as the two methods
301 are in theory similar, see Koene et al. 2024). However, we choose in this study to focus on the standard application of this
302 method (e.g., Beirle et al. 2019, 2021, 2023; Hakkarainen et al., 2022, Sun et al., 2022), which provides temporally averaged
303 estimates. Appendix A provides a brief overview of the performance when estimating emissions from individual images with
304 different versions of the divergence approach.

305 For a specific source, the annual estimate of the emissions is then computed from the enhancement in the averaged
306 divergence field by using a peak fitting approach which fits the divergence map by a function including a Gaussian and a
307 linear term centred at the source (Beirle et al, 2021). Emissions, and more generally the parameters, of the peak function are
308 determined by an adaptive Markov chain Monte Carlo (MCMC) that also provides the uncertainties in the estimates from the
309 standard deviations of the sampled posterior distributions of the parameters.

310 **2.2. Synthetic satellite observations of CO₂ and NO₂**

311 In this study, synthetic satellite observations of CO₂ and NO₂ were generated from atmospheric simulations in order to
312 evaluate and compare the ability of the methods described in Sect. 2.1 for retrieving CO₂ or NO₂ emissions from point
313 sources or urban areas using satellite imagery akin to that provided by the upcoming CO2M mission. These simulated
314 satellite data are readable by the ddeq Python library and were produced as part of the SMARTCARB project and have been
315 extensively described and used in previous works (e.g. Brunner et al., 2019; Kuhlmann et al., 2019; 2020; 2021). They are
316 openly accessible from <https://doi.org/10.5281/zenodo.4048227> (Kuhlmann et al., 2020b).

317 Atmospheric concentrations of CO₂ and NO₂ were simulated by the COSMO-GHG atmospheric transport model (Jähn et
318 al., 2020) with a vertical resolution of 60 levels up to an altitude of 24 km and with a horizontal resolution of about 1 km × 1
319 km for a domain centred over the city of Berlin. The domain extends about 750 km in the east-west and 650 km in the south-

320 north direction. Simulations provided hourly outputs for nearly the entire year 2015. In order to generate realistic
321 simulations, initial and lateral boundary conditions for meteorological variables and tracers were extracted from products of
322 the European Centre for Medium-Range Weather Forecasts (ECMWF) and MeteoSwiss (Kuhlmann et al., 2019).
323 Furthermore, CO₂ emissions included both the anthropogenic and biospheric components which were interpolated onto the
324 COSMO grid at a temporal resolution of one hour: anthropogenic emissions were largely derived from the TNO/MACC-3
325 inventory (Kuenen et al., 2014) and biospheric fluxes were simulated with the Vegetation Photosynthesis and Respiration
326 Model (VPRM, Mahadevan et al., 2008). NO_x emissions were also derived from the TNO-MACC-3 inventory and
327 atmospheric simulations used a simplified NO_x chemistry with a fixed NO_x decay time of 4 hours. NO_x concentrations were
328 converted to NO₂ concentrations using an empirical equation for the evolution of NO₂ : NO_x ratios downwind of emission
329 sources (Düring et al., 2011).

330 To generate synthetic satellite observations similar to CO2M observations, the XCO₂ and NO₂ column densities derived
331 from the COSMO-GHG simulations were sampled at the resolution of 2 km × 2 km along 250 km wide satellite tracks
332 (Kuhlmann et al., 2019); these tracks were computed using an orbit simulator and correspond to a hypothetical constellation
333 of six CO2M satellites. In addition to XCO₂ and NO₂ column-average data, a cloud mask was generated from the total cloud
334 fraction computed by the COSMO-GHG model. For CO₂ data, all pixels with cloud fraction larger than 1 % were removed
335 as CO₂ retrievals are strongly impacted by clouds (Taylor et al., 2016). For NO₂ data, less sensitive to clouds, a threshold of
336 30 % on the cloud fraction was used to select valid pixels (e.g. Boersma et al., 2011). Figure 2 illustrates a COSMO-GHG
337 simulation of XCO₂ over the SMARTCARB domain, on which are represented synthetic XCO₂ data corresponding to a
338 CO2M satellite overpass.

339 For the purposes of this benchmarking study, we use the configuration of the SMARTCARB dataset where the CO2M
340 constellation consists of three satellites. By choosing this, we follow the recommendation of Kuhlmann et al. (2021) that a
341 constellation of at least three CO2M satellites is necessary for a proper estimation of the annual emissions from weak
342 sources and in regions such as central Europe where cloud cover dramatically reduces the number of estimates. When
343 ignoring clouds, this constellation of three satellites leads to observing each local source within the SMARTCARB domain
344 once every other day; if we consider that a satellite image is usable if there are at least 50 data pixels next and downwind to
345 the source, then we can use about 3000 images to determine the emissions of the 16 local sources considered in this study.
346 But, if we consider the cloud cover, only 500 images remain usable.

347 The characteristics of the uncertainties in the synthetic CO2M observations were computed using three different
348 uncertainty scenarios (low, medium, high). Simulated XCO₂ column densities were thus assigned random errors by
349 employing various levels of instrumental noise in the error parameterization formula. This formula, used for generating the
350 errors, takes into account the Solar Zenith Angle (SZA) and surface albedos (Buchwitz et al., 2013). The NO₂ column
351 densities were assumed to be characterised by random uncertainties of different constant values depending on the chosen
352 uncertainty scenario. These values are defined for clear sky conditions and increase in the presence of clouds; nearly
353 doubling for a cloud fraction of 30 %. No systematic errors were prescribed for either XCO₂ or NO₂ column averaged data.

354 In this study, the characteristics of the random uncertainties prescribed to the synthetic data are chosen according to the
355 requirements of the CO2M mission (Meijer et al., 2019). For XCO₂ retrievals, random errors are generated using the error
356 parameterization formula with a single sounding precision of 0.7 ppm for vegetation albedos and a SZA of 50°. For NO₂
357 retrievals, a single sounding precision in cloud-free conditions of 2×10^{15} molecules cm⁻² is prescribed.

358 **2.3. Benchmarking scenarios**

359 The relative performance of the different inversion methods to estimate CO₂ emissions are evaluated for the 15 strongest
360 point sources of the SMARTCARB domain and for the city of Berlin (Fig. 2 and Table 1 in Kuhlmann et al., 2021). These
361 16 sources cover a large emission range that extends from 3.7 MtCO₂.yr⁻¹ for the power plant located in Chvaletice (CZ) to
362 40.3 MtCO₂.yr⁻¹ for the power plant located in Jänschwalde (DE); these values being the annual mean emissions at the time
363 of the satellite overpass (10:30 UTC) used in the COSMO-GHG simulations. It is worth mentioning that the distribution of
364 the source emissions is skewed towards the lowest value as the median emission rate in the collection is around 9.6
365 MtCO₂.yr⁻¹ and 75 % of the sources emit less than 14 MtCO₂.yr⁻¹.

366 In order to thoroughly evaluate the relative performance of the different methods and the sensitivity of these
367 performances to different factors, the benchmarking study is carried out according to several scenarios that share the same
368 features for the simulated data and for the source collection that have been described above. The most optimistic or ideal
369 scenario corresponds to the application of inversions to CO₂ and NO₂ images without the removal of pixels associated to
370 cloud-cover (ignoring the clouds modelled with the COSMO-GHG model; we label such inversions “cloud-free” hereafter)
371 and with a perfect knowledge of the wind field (i.e. using directly the winds from the COSMO-GHG model, denoted
372 SMARTCARB winds). It is the ideal case because 1) the joint analysis of NO₂ and CO₂ images strengthen the estimates
373 compared to the analysis of CO₂ images only; 2) ignoring the potential loss of data due to cloud cover in the CO₂ and NO₂
374 images yield full images, whose analysis is more robust than that of partial images, and thus provides a higher number and
375 precision of estimates. . The results derived from this benchmarking scenario should be seen as an upper limit of what the
376 inversion methods could achieve in terms of accuracy and number of estimates. The most realistic scenarios take cloud cover
377 into account and use winds extracted from the ERA5 wind product (Hersbach et al., 2020) that is independent from the
378 inverted data and whose resolution (~0.25°) is much coarser than that of the SMARTCARB winds (~0.01°). The results
379 derived from this benchmarking scenario should be seen as a lower limit for the method's performance.

380 The differences between the ERA5 and SMARTCARB wind products are significant at the 16 sources considered in this
381 study: the annual mean biases between these two wind products in 2015 range from 0.1 ms⁻¹ to 1.5 ms⁻¹ depending on the
382 source with an average value across the sources of 0.6 ms⁻¹ while RMSEs range from 1.1 ms⁻¹ to 2.1 ms⁻¹ depending on the
383 source with an average value across the sources of 1.5 ms⁻¹ (Fig. A2). The biases per source are systematically positive since
384 SMARTCARB tends to provide larger winds than ERA5. With such differences, comparing scenarios with the same
385 characteristics but using different wind products allows us to gain insight into the method's sensitivity to wind uncertainties.
386 Additional benchmarking scenarios were designed to test the sensitivity of the methods with respect to other factors,

387 including the consideration of cloud cover in satellite data and the use of NO₂ for plume detection and characterization. All
388 benchmarking scenarios are listed in Table 2.

389 **2.4. Benchmarking metrics**

390 For a given benchmarking scenario, the performances of the different inversion methods can be evaluated through the
391 number of single-image estimates that can be retrieved regarding the number of available satellite images: ~500 or ~3000
392 considering or ignoring the cloud cover in the data. Performances can be assessed as well through the quality of the
393 estimates; the accuracies of the methods are then assessed by comparing the estimates retrieved from single satellite
394 overpasses to the corresponding *true* values that were used to generate the synthetic satellite data. More precisely, inversion
395 results are analysed in terms of distributions of the differences between the estimated and the true emissions of all the
396 sources considered in this study. We will refer to these differences in the following as *deviations*. More precisely, our
397 analysis will mostly focus on examining the distributions of the *relative* deviations, i.e. the differences between estimated
398 and true emissions divided by the true emissions, in order to fairly compare results across sources with significantly different
399 magnitudes (Sect. 2.3). Furthermore, to properly describe distributions that may be very different from Gaussian
400 distributions, box plots are used, in which the median values, the interquartile ranges (IQRs), the 10th and the 90th percentiles
401 of the distributions are represented.

402 The ability of the different inversion methods to estimate source emissions can also be analysed from the study of the
403 annual or monthly averages of the single-image estimates. Benchmarking results are then evaluated for each source in terms
404 of relative deviations of the annual/monthly estimates from the annual/monthly true emissions and, in terms of Root Mean
405 Square Errors (RMSE) in order to provide a global indicator for the accuracy of the annual/monthly estimates across all
406 sources.

407 In this study, the annual/monthly averages of the single-image estimates for a given source are computed using three
408 different methods which are 1) the arithmetic means of all the single-image estimates of the source emission that have been
409 generated from inverting one year/month of data, 2) the means of these estimates weighted by the inverse of their computed
410 variances (Sect. 2.1) and 3) the medians of these estimates. The annual/monthly inverse variance weighted means
411 incorporate the information provided by the methods on the quality of the estimates when averaging, whereas the
412 annual/monthly medians are statistical indicators that are more robust to outliers than the means. Moreover, since the Div
413 method is applied by temporally averaging satellite observations over the year, it produces only a single annual estimate for
414 each source; we will thus consider that the three types of annual/monthly estimates are all equal to this single estimate.

415 It is important to note that the annual and monthly estimates are affected by temporal sampling biases when inversion
416 methods use data filtered by cloud cover. Specifically, the presence of denser cloud cover during winter generally results in
417 over-representation of emission estimates during summer and hence could lead to an underestimation of annual estimates as
418 emissions are higher during winter due to increased fossil fuel consumption associated with electricity and heat production.
419 Although more advanced methods, such as fitting periodic curves to capture seasonal cycles as demonstrated by Kuhlmann

420 et al. (2021) could potentially enhance the accuracy of estimates, they are not included in this study. However, these
421 temporal sampling biases are integrated in the results as the annual/monthly estimates are compared to the true
422 annual/monthly emissions which are computed by considering all the days of the year/months.

423 **3 Results on emission estimates based on individual images**

424 The following subsections present a comparative study of the CSF, GP, IME, and LCSF methods for estimating emissions
425 from single images. In the following, we will refer to these kinds of estimates as *single-image* estimates. It is worth
426 mentioning that, as the methods use different algorithms for plume detection and emission quantification, which include
427 different rejection criteria (Sect. 2.1), they produce different sets of estimates.

428 **3.1 Sensitivity to the emission strengths of the sources**

429 In the optimal scenario (cloud-free, SMARTCARB winds, CO₂ and NO₂ data), all methods tend to provide more accurate
430 estimates for strong sources than for weak sources, and this trend is particularly noticeable for the IME and CSF methods
431 (Fig. 3). The median values of the absolute relative deviations for weak sources (emissions ranging from 0 to 6.9 MtCO₂/yr
432 in the 1st row of Fig. 3) are 207 % (IME method) and 54 % (CSF method), respectively. In contrast, for strong sources
433 (emissions ranging from 15.6 to 53.2 MtCO₂/yr in the 4th row of Fig. 3), they are approximately 47 % (IME) and 28 %
434 (CSF), respectively. The inversion methods are also more prone to produce unrealistic values for weak sources as the
435 distributions are strongly skewed for this type of sources: the 95th percentile accuracy indicator is indeed 1128 %, 584 %,
436 172 % and 178 % for the IME, CSF, GP and LCSF inversion models respectively (1st row in Fig. 3). For strong sources, this
437 indicator is significantly lower, decreasing to 200 %, 108 %, 90 % and 76 %, respectively (4th row in Fig. 3). Atmospheric
438 signals generated by strong sources are more distinct from the background than those from weak sources and as a result, the
439 signal-to-noise ratio in the XCO₂ and NO₂ images is better which helps to reduce uncertainties in the determination of their
440 emissions. For low-emitting sources, the performance of the inversion methods can be degraded by the limited number of
441 enhanced pixels that are detected in images with noise; this limitation makes the identification of plume centre-lines by the
442 CSF, IME and GP methods challenging (Sect. 2.1). This problem could have impacted the GP method, but its current
443 implementation incorporates prior knowledge filtering out estimates that fall outside the 25 % to 400 % range from the prior.
444 This filtering process is expected to improve the accuracy of the GP method, especially for weak sources.

445 Biases in the emission estimates may also depend on the strength of the source, as observed in the IME and CSF methods
446 which strongly overestimate the emissions of weak sources compared to strong sources. For weak sources, the median of the
447 deviation distributions for the IME and CSF models (blue bars, 1st row of Fig. 3) are +116 % and +50 %, respectively,
448 compared to +16 % and +11 % for strong sources (blue bars, 4th row of Fig. 3). This discrepancy is probably due to the
449 plume detection algorithm, which, for weak sources, may wrongly attribute enhancements from other sources in the vicinity
450 of the source of interest and thus artificially increase the amplitude of the detected emissions. Conversely, the LCSF

451 approach tends to underestimate the emissions of strong sources while slightly overestimating those of weak sources, with
452 the median of the deviation distribution being -26% (blue bar, 4th row of Fig. 3) and $+12\%$ (blue bar, 1st row of Fig. 3)
453 respectively. The underestimation of source emissions could be attributed to a tendency of the method to overestimate the
454 amplitudes of the background for non-isolated sources: contrary to the other methods, the LCSF method does not remove the
455 influence of neighbouring plumes when computing the background around a given source. Another explanation could lie in
456 the fact that this method uses 100-m winds as effective winds while, especially for strong emitting sources, these winds are
457 lower than the GNFR-A average winds used by the other methods.

458 **3.2 Impact of the use of NO₂ images for the detection of plumes**

459 The use of NO₂ data to identify and characterise plumes increases the number of estimates for all inversion methods
460 compared to CO₂-only inversions, as shown in Figure 4 (blue vs orange bars). The increase is significant for the IME and GP
461 methods ($\sim 93\%$ and $\sim 70\%$), moderate for the CSF method ($\sim 34\%$), and slight for the LCSF method ($\sim 4\%$). The IME, GP,
462 and CSF methods rely on a plume detection algorithm that is less reliable when using only CO₂ observations (Kuhlmann et
463 al. 2019). Of these three, the CSF method requires fewer pixels to detect and quantify plumes, resulting in a larger proportion
464 of still quantified plume cases than the IME and GP methods when having CO₂ data only. The detection of plumes by the
465 LCSF method is performed on data slices whose pixels are relatively close to sources and where XCO₂ enhancement signals
466 due to emissions are thus relatively strong; this may explain the only small benefit for this method of using joint CO₂ and
467 NO₂ images to better determine the shape of the plumes.

468 When using CO₂ and NO₂ data, the maximum number of estimates obtained from each inversion method varies
469 significantly: the IME method produces the smallest number of estimates, with 1661, while the LCSF method produces the
470 largest, with 2722. The GP and CSF methods, based on the same algorithm of plume detection as the IME method, produce
471 up to 1776 and 2012 estimates, respectively. These differences can be attributed to the differences in the number of detected
472 pixels below which the algorithm rejects plumes and, in the emission quantification algorithms used by the different
473 methods. In addition, the overall complexity of the IME, CSF and GP methods, which use a relatively large number of
474 rejection criteria likely explains why these three methods deliver much fewer estimates than the LCSF method. The relative
475 efficiency and robustness of the plume detection algorithm of the LCSF method is evidenced when using CO₂ data only to
476 determine emissions: the number and accuracy of estimates is hardly changed compared to the inversions performed with
477 CO₂ and NO₂ data; contrarily to the other methods whose algorithms are more sensitive to uncertainties in XCO₂ data and
478 which need NO₂ data to accurately fit a plume coordinate system to the data.

479 The inclusion of NO₂ data does not appear to significantly improve the overall performance of the GP and LCSF methods
480 in terms of accuracy of the CO₂ emission estimates (lower panel in Fig. 4). However, for the LCSF method, there is a notable
481 reduction in the 95th percentile of the relative absolute deviations from 175% without NO₂ to 115% with NO₂. For the CSF
482 method, the use of NO₂ data strongly improves its overall performance as the 3rd quartile and the median of the absolute
483 residuals are for example significantly decreased, from $\sim 127\%$ down to $\sim 74\%$ and from $\sim 54\%$ to $\sim 36\%$, respectively. As

484 the CSF method rejects fewer estimates when using CO₂ data only than the GP method, its accuracy decreases because with
485 a more permissive filtering, it may include complex cases for which emissions are difficult to estimate. This may also
486 explain why the CSF estimates are less biased, with a significantly lower median relative deviation, in cases where
487 inversions also use NO₂ data (upper panel in Fig. 4).

488 In contrast, the precision of the IME method decreases when using NO₂ data, but this fact could be related to a numerical
489 artefact: the IME method performs much better for high-emitting sources than for low-emitting sources (see Sect. 3.1) and
490 the use of NO₂ data likely allows constraining small sources more efficiently than with CO₂ data only. Therefore, when
491 adding NO₂ data, the number of low-emitting sources which are estimated increases more than for the high-emitting sources
492 and then the overall performance degrades. This bias associated to the relative bad estimation of low-emitting sources is
493 confirmed when deviations are used to assess performance instead of relative deviations: the absolute deviations associated
494 to the IME estimates globally decrease with the use of NO₂ data with for example the median error decreasing from ~15 to
495 ~11.5 MtCO₂/yr.

496 **3.3 Impact of the cloud cover**

497 The impact of clouds is studied by comparing inversions with cloud-free images to inversions with cloud-filtered images
498 (Sect. 2.3). When disregarding cloudy pixels in the XCO₂ and column-averaged NO₂ data, the number of estimates from all
499 the methods is considerably reduced, with a decrease of 94 %, 85 %, 85 % and 88 % for the IME, CSF, GP and LCSF
500 methods respectively (Table 3). The number of estimates that can be provided for the cloud-filtered configuration with
501 SMARTCARB winds is at the maximum equal to 313 (LCSF) and decreases to 96 for the IME method which can provide
502 robust estimates for images free of clouds only as this method requires integrating enhancements over the full extent of
503 plumes. As sources are characterized by different cloud covers, the number of estimates per year and per source ranges from
504 1 to 12 (IME), from 6 to 28 (CSF), from 8 to 23 (GP) and from 15 to 26 (LCSF).

505 Furthermore, the filtering of data pixels removing those with a significant cloud cover not only affects the number of
506 estimates but also impacts the performance of the methods, although to a much lesser extent. When comparing results
507 obtained from the same images, cloud-free inversions produce slightly better results than cloud-filtered inversions (Fig A3).
508 This is because, in images partially masked by cloud cover, some pixels containing useful information are likely removed,
509 which can lead to less accurate determination of emissions. Consistently, if the threshold of cloud cover above which XCO₂
510 images are discarded for the analysis is increased from 1 % to 2 % or 5 %, the performance of the methods does not
511 significantly increase, unlike the number of estimates, which can increase, e.g. by 12 % and 29 % respectively when using
512 the LCSF method (Fig. A4).

513 **3.4 Impact of uncertainty in the wind**

514 As mentioned above, in order to assess the impact of potential uncertainties in the wind, a series of inversions is carried out
515 with a different wind product than the one used to generate the synthetic XCO₂ and NO₂ data. For this purpose, the

516 SMARTCARB winds are replaced by ERA5 winds and the differences between these two wind products are characterised at
517 the sites of this study by random and systematic components (Sect 2.3 and Fig. A3). Notably, ERA5 winds show
518 systematically lower values.

519 For all inversion methods, the global accuracies of the estimates, evaluated in terms of relative absolute deviations, are
520 only slightly reduced when using ERA5 winds instead of SMARTCARB winds (lower panel in Fig. 4, green vs red bars).
521 There are a few possible explanations for this: the temporal or spatial uncertainties in wind components are only a minor
522 source of uncertainty compared to other factors impacting the determination of the estimates by the different inversion
523 methods such as, for example, uncertainties in the XCO₂ and NO₂ columns densities (Sect. 2.2) or over-simplified
524 assumptions in plume detection or quantification algorithms. Kuhlmann et al. (2020, 2021) showed, for instance, that the
525 determination of the CO₂ background field could introduce significant uncertainties in the estimates. Furthermore, as
526 indicated by Reuter et al. (2019), one of the important benefits of satellite imagery is that uncertainties related to
527 meteorological variables likely average out when emission estimates are sampled along significant areas of plumes.

528 However, the fact that ERA5 wind values are systematically lower than those of SMARTCARB winds has an impact on
529 the median values of the relative deviations, i.e. on the biases in the estimates. While the accuracies in terms of relative
530 absolute deviations are slightly affected by using either wind product (bottom panel in Fig. 4, green vs red bars), biases can
531 be significantly increased, as in the cases of the GP and LCSF methods whose estimates are on average underestimated if
532 inversions use ERA5 winds instead of SMARTCARB winds. The lower amplitudes of the ERA5 winds explains also that the
533 results for the IME and CSF methods improve, especially for the 95th percentiles of the absolute deviation distributions
534 which respectively decrease from around 504 % and 411 % to 370 % and 286 % respectively. The systematic overestimation
535 of the estimates evidenced above for the CSF and the IME methods is therefore mitigated when using ERA5 winds (top
536 panel in Fig. 4).

537 As mentioned previously (Sect. 2.3), the benchmarking scenario for which inversions are performed with ERA5 winds
538 and data filtered for cloud cover, is the closest to real conditions of monitoring emissions from data images delivered by
539 satellites. For this scenario with CO₂ and NO₂ data, the GP and LCSF methods show the best performances in terms of
540 global accuracies with respectively IQRs of 25–62 % and 17–55 % for the distributions of the absolute relative deviations
541 (red boxes in Fig. 4). It is interesting to note that the overall accuracies of these methods are similar for this realistic scenario
542 and the ideal scenario where inversions are performed with cloud-free data and SMARTCARB winds. Contrarily, the
543 number of estimates strongly decreases when inversions are performed with cloud-filtered data such as, for example, from
544 2722 to 318 estimates for the LCSF method (see Table 3).

545 **4 Results on annual and monthly averages of the emissions**

546 **4.1 Annual estimates**

547 To evaluate how well an inversion method performs on an annual basis, we include all image estimates generated by the
548 method, regardless of their uncertainty. We calculate annual estimates for a given source using three methods, as described
549 in Sect. 2.4: 1) by taking the average of all available image estimates for the source over the entire year, 2) by taking the
550 weighted average of these image estimates based on their uncertainty, and 3) by taking the median value of these image
551 estimates. Because the Div method only provides one estimate per year, its annual estimates are the same, irrespective of the
552 calculation method used. In order to compare for a given source the three estimated annual values to the true emission, we
553 define this latter as the arithmetic mean of the true emissions values for the source over all 365 days of the year.

554 As noted earlier (Sect. 2.1.5), the Div method computes the annual emission estimate for a given source by averaging the
555 divergence map from all available overpasses in 2015. However, the other methods select overpasses for which they succeed
556 to detect plumes, likely increasing the reliability of their estimates. These selections generally correspond to conditions — in
557 terms of wind, of background variability or of emission strength — that should be favorable to all methods, including the
558 Div method. The lack of selection and thus the use of unfavourable overpasses when applying the Div method may therefore
559 hamper the comparison between the annual estimates of the Div method and that from the other methods.

560 When annual estimates are calculated as arithmetic means or medians of individual image estimates, the GP and LCSF
561 methods generally outperform the other methods. Indeed, for cloud-free inversions with CO₂ and NO₂ data, the median
562 deviations for the annual arithmetic means (solid lines, 2nd column of Fig. 5) are 8 % (GP), 14 % (LCSF), 73 % (IME), 35
563 % (CSF), and 64 % (Div), and the median deviations for the annual medians (dotted lines, 2nd column of Fig. 5) are 14 %
564 (GP), 21 % (LCSF), 54 % (IME), 13 % (CSF), and 64 % (Div). However, if annual estimates are calculated as the means of
565 image estimates weighted by their uncertainty, the relative performance of the methods changes. In this case, the median
566 deviations for annual weighted means (dashed lines, 2nd column of Fig. 5) are 28 % (GP), 48 % (LCSF), 46 % (IME), and 12
567 % (CSF). Thus, using weighted means to calculate annual estimates significantly improves, especially for low-emitting
568 sources, the performance of the IME and CSF methods while having a negative impact on the GP and LCSF methods. This
569 finding indicates the reliability of the uncertainties in the estimates produced by the IME and CSF methods compared to the
570 other methods and, if we use weighted means to compute annual estimates, the accuracies of the IME and CSF methods
571 increase significantly.

572 Figure 6 displays the inversion results for the annual estimates in a different but complementary way compared to Fig. 5:
573 the estimated annual emissions are represented with respect to the true ones which in particular allows illustrating whether
574 annual estimates are over- or under-estimated for a certain type of source and by a given inversion method. In order to
575 consider the best performance for each method according to what has been shown above, annual estimates represented in the
576 figure, and used for the analysis of the results made below, are arithmetic means of single-image estimates for the LCSF and
577 the GP methods, while they are weighted means for the IME and CSF methods. Furthermore, Fig. 6 illustrates more clearly

578 than Fig. 5 the fact that, when weighted averages are used as annual estimates, the latter methods produce annual estimates
579 whose precision is comparable for weak *and* strong sources while the global precision of estimates derived from single
580 images by these methods is significantly lower for weak sources (Fig. 3); averaging single-image estimates weighted by their
581 uncertainty thus strongly increases the performance of the IME and CSF methods at the annual scale for low-emitting
582 sources. However, even though the amplitudes of the relative deviations are similar between strong and weak sources, they
583 have opposite signs: annual estimates for strong sources are generally underestimated while annual estimates for weak
584 sources are generally overestimated.

585 Contrary to the results for the estimates retrieved from single images (Fig. 4), the CSF, GP and LCSF approaches show
586 similar performance, with a slight advantage for the GP method, when estimating annual emissions if we consider the
587 ensemble of the benchmarking scenarios. For example, in the case of inversions from cloud-filtered CO₂ and NO₂ data and,
588 with SMARTCARB/ERA5 winds, the relative RMSEs are 18/27 % (CSF), 20/20 % (GP) and 17/31 % (LCSF). The analysis
589 of Fig. 3 shows that the LCSF method produces single-image estimates that are slightly more accurate but more biased than
590 that of the GP method. Thus, the compensation of errors when averaging single-image estimates over a year may be less
591 effective for the LCSF method than for the GP method leading to similar global accuracies for both methods. For instance,
592 the LCSF method has a greater tendency to underestimate high emissions (4th row of Fig. 3) which likely explain why,
593 contrarily to the GP method, it systematically underestimates the emissions of the strong emitting power plant located in
594 Jämschwalde, regardless of the inversion scenario (Fig. 6). With respect to its results for single-image estimates, the CSF
595 method has significantly better results at the annual scale when annual estimates are computed as weighted averages of
596 single-image estimates.

597 Even when annual estimates are computed for the IME method as weighted averages of the single-image estimates, this
598 method still show smaller accuracies compared to the CSF, GP and LCSF methods: the median values of the deviations for
599 the annual estimates are for example 39 % (IME), 20 % (CSF), 11 % (GP) and 21 % (LCSF) when considering the best
600 scores for the inversions performed with ERA5 winds and cloud-filtered data (4th column of Fig. 5). The relative
601 performance of the IME method is even worse when analysing the performance in terms of RMSE because, despite a
602 weighting of estimates according to their quality or uncertainty in the annual averages, this method produces for some
603 sources annual estimates that strongly deviate from the actual values, as in the cases of Boxberg or Schwarze Pumpe power
604 plants (Fig. 6). Moreover, the deviations of the Div method compared to that of the CSF, GP and LCSF methods are higher
605 for most of sources except for strong sources (true annual emissions > 15 MtCO₂/yr) when inversions are performed using
606 cloud-filtered data and ERA5 winds (4th column of Fig. 5).

607 It is noteworthy that annual estimates for most inversion methods are comparable between inversions using data with or
608 without clouds (comparison between the 2nd and 3rd columns, Fig. 5), and surprisingly the deviations of the IME and Div
609 approaches are even smaller for inversions with cloud-filtered data. Despite significant differences in the number of image
610 estimates between those two (i.e., cloud-filtered and cloud-free) inversion configurations, annual estimates are *on average*
611 slightly affected when cloud cover is considered in the data, at least for the year and sources examined in this study.

612 However, even though the relatively small number of image estimates in the inversion configuration with clouds does not
613 hinder most methods from determining annual emissions of most sources, discrepancies can be high for some sources when
614 estimates do not sample correctly the entire year and thus introduce an important temporal bias. For example, the GP method
615 mostly estimates emissions during summer for the Jämschwalde power plant when it uses the cloud-filtered inversion setup,
616 explaining the strong underestimation of the annual emission of this source compared to the cloud-free case (top-left vs
617 bottom-left panel of Fig. 6); this explains additionally why the RMSE increases significantly for the GP method (from 13 %
618 to 20 % when inversions use SMARTCARB winds) when the cloud cover limits the number of single-image estimates. The
619 IME method is also impacted by this temporal bias when the number of estimates is too small to properly capture the
620 seasonal cycle of the emissions, as in the case of the Boxberg power plant. Moreover, whatever the benchmarking scenario,
621 most inversion methods produce annual estimates for all the sources studied in this work, with the notable exception of the
622 Div approach, which estimates annual emissions for only 10 out of 16 sources. This limitation, also present for cloud-free
623 data configurations, is related to the fact that some sources don't produce strong enough divergence peaks from which
624 annual estimates can be made by this method.

625 As for the results concerning single-image estimates, the use of ERA5 winds instead of SMARTCARB winds has on
626 average a very low impact on annual estimates delivered by the IME, CSF, GP and LCSF methods. For emissions estimated
627 from cloud-free CO₂ and NO₂ data, the median deviations when inversions use SMARTCARB winds are indeed 46 %
628 (IME), 12 % (CSF), 8 % (GP) and 14 % (LCSF), and when inversions use ERA5 winds, they are equal to 46 % (IME), 12 %
629 (CSF), 9 % (GP) and 12 % (LCSF) as shown in the comparison between the 2nd and 4th columns of Fig. 5. On the other hand,
630 the overall accuracy of the Div method improves when inversions use ERA5 winds rather than SMARTCARB winds to
631 estimate emissions. In this case, annual estimates are less prone to overestimation due to the generally lower amplitude of
632 ERA5 winds compared to SMARTCARB winds (Fig. A2). This also explains a stronger underestimation of the emissions of
633 strong sources by the LCSF method, resulting in a decrease in the accuracy of the annual estimates for this kind of sources
634 when this method uses ERA5 instead of SMARTCARB winds (left-bottom vs right-bottom panel of Fig. 6).

635 The overall precision of the annual estimates computed by the IME, CSF, GP and LCSF methods are, for all the
636 benchmarking scenarios, significantly higher than the overall precision of their single-image estimates. For example, when
637 inversions are performed with ERA5 winds and cloud-filtered data, which is the benchmarking scenario with the poorest
638 results, the median deviations of the annual estimates are 39 %, 20 %, 11 % and 21 % whereas the median deviations of the
639 single-image estimates are 73 %, 35 %, 46 % and 37 % for the IME, CSF, GP and LCSF methods. Despite the biases that
640 can hamper the image estimates, the compensation for errors when averaging across a year allow to generate annual
641 estimates that are more precise and this positive effect is amplified when error-weighted averages are used, as in the case of
642 the IME and CSF methods.

643 **4.2 Monthly estimates and seasonal cycle**

644 Monthly estimates can be computed using the same three methods as the annual estimates but, according to the results
645 analysed in the former section, we choose to estimate monthly emissions with the method leading to the best performance at
646 the annual scale: monthly estimates are thus calculated as the arithmetic means for the GP and LCSF methods and, as
647 weighted means for the CSF and IME methods. Then, considering the distributions of image estimates month by month
648 allows us to study how well inversion approaches capture the seasonal cycle of the true emissions. The analysis of Fig. 7
649 shows however that none of them are able to do this when the cloudy pixels are masked: the seasonal cycle of the actual
650 monthly emissions, i.e. maximal/minimal emissions for winter/summer months, is not reproduced by the inversion methods
651 whose estimates are characterised by an erratic monthly evolution leading to inconsistent seasonal cycles. Even though a
652 method correctly estimates annual emissions, some of its monthly estimates can be in important disagreement with the *true*
653 monthly emissions as it is the case for the CSF method on the Heyden source or for the LCSF method on the Dolna Odra
654 source (Fig. 7). Moreover, the methods generally fail to produce estimates for the winter months of the year due to the
655 temporal sparsity of data when the impact of the cloud cover is taken into account.

656 If the number of estimates is higher, i.e. when clouds are not considered in the data, seasonal cycles derived from
657 monthly estimates are in better agreement with that of the observations for most of inversion methods: the amplitude of the
658 seasonal cycle of the data can be well reproduced as it is the case for the Jänschwalde and Dolna Odra sources for example
659 (Fig. A5). But, the averaged values of the seasonal cycles of the monthly estimates, i.e. the annual estimates, can still be in
660 strong disagreement with that of the data even though the number of estimates is higher; this fact supports the presence of
661 systematic biases in the estimates that was evidenced for most of the methods in the analysis of the results for single-image
662 image estimates (Sect. 3.1).

663 **5 Discussion**

664 **5.1 Accuracy vs number of estimates**

665 For a given benchmarking scenario, the analysis conducted in Section 3 has evaluated the performance of the different
666 methods in inferring estimates from individual images by considering all the estimates provided by each method for this
667 scenario. In other terms, the analysis did not integrate any diagnostic regarding the quality of the estimates from these
668 methods. However, we demonstrated in Sect. 4.1 that computing annual means of estimates weighted by their uncertainties
669 can significantly improve the accuracy of the annual estimates when uncertainties are effectively characterised as in the case
670 of the IME and CSF methods. Therefore, a study of the performance of inversion methods for estimating single-image
671 estimates from synthetic XCO₂ images should as well integrate a characterization of the quality of its estimates. More
672 precisely, different performance indicators or error estimates can be derived from the application of the inversion methods
673 and such indicators can be used to identify and select the most reliable estimates. Nevertheless, there are no objective criteria

674 to impose a threshold on the quality of the estimates; higher quality thresholds come with smaller sets of estimates, and
675 optimal values depend on the inversion method. Indeed, not only do the different inversion methods calculate the
676 uncertainties in the estimates in different ways but also the computed uncertainties only reflect part of the total/actual
677 uncertainties, focusing on subsets of sources of uncertainties which differ across the different methods.

678 For a given inversion method, we attempt an effective quality indicator (QI) which would allow selecting estimates in a
679 manner that the global accuracy of the method increases when the QI increases, and which would provide indications on the
680 actual/total errors. We assume that the uncertainties in the estimates derived by the methods provide the best basis we can
681 get from the algorithms described in Sect. 2.1 for the derivation of such an indicator. In principle, since dealing with sources
682 of quantitatively different amplitudes (see Sect. 2.3) we should derive the QI in terms of *relative* uncertainties. And, if we
683 define the QI as a threshold selecting the estimates whose relative uncertainties are below it, we should select the most
684 reliable estimates regardless of the strength of the source they are associated with. However, this would be true if the
685 methods perform independently with respect to the amplitudes of the emissions and this is not the case for most methods as
686 illustrated in Sect 3.1. The CSF and IME methods for example strongly overestimate low-emitting sources compared to
687 high-emitting sources which implies that the relative uncertainties of weak sources are underestimated by these methods
688 (Fig. 3). Therefore, if the threshold value of relative uncertainty was decreased, we would tend to select more bad than good
689 estimates and the overall performance would decrease. Therefore, for these methods, we prefer to select estimates with
690 respect to their uncertainties, and not to their *relative* uncertainties, which will mitigate the impact of the bias in the
691 estimation of low-emitting sources.

692 In any case, determining whether a QI should be based on absolute or relative uncertainties depends on whether the
693 overall performance of the method improves when estimates with decreasing absolute or relative uncertainties are chosen.
694 Preliminary tests (not shown here) have established that the overall accuracy of the IME and CSF methods increases when
695 the *absolute* uncertainty below which estimates are selected is decreased. For the GP and LCSF methods, this behaviour is
696 obtained when *relative* uncertainties are used to discriminate estimates. Consistently, for all methods, the increase of
697 performance is then associated with a reduction in the number of estimates and, in order to get a significant number of high-
698 quality estimates, the value of uncertainty corresponding to the maximal accuracy of the method is arbitrarily set to the 10th
699 percentile of the distribution of the absolute/relative uncertainties. Then, by varying its QI between this value and the
700 maximal uncertainty of its estimates, each method can be thus associated to a range of accuracies with their respective
701 number of estimates for a specific benchmarking scenario (e.g. cloud-filtered or cloud-free). In other words, inversion results
702 can be represented by curves of accuracy *vs* number of estimates, which gives for each inversion method a complete
703 overview of its performance in terms of accuracy and number of estimates.

704 To assess the inherent performance of the methods without considering the impact of the cloud cover or of the
705 uncertainty in the winds, inversion results are analysed for the inversion configuration using XCO₂ and NO₂ cloud-free data
706 and SMARTCARB winds, *i.e.* the same winds used to generate the synthetic XCO₂ and NO₂ observations. Figure 8
707 illustrates that the overall accuracies of the CSF and IME methods are highly dependent on the selection of their estimates,

708 and are therefore strongly correlated with their number of estimates. For instance, the IME and CSF methods exhibit large
709 increases in the 3rd quartiles of their deviation distribution when the QIs of their estimates decrease: from 81 % to 231 %
710 (IME) and from 43 % to 75 % (CSF) respectively. For these methods, the selection of estimates based on their quality
711 indicators appears to be effective, as the 3rd quartiles and 95th percentiles, which indicate the proportion of poor estimates,
712 significantly decrease with increasing quality index, *i.e.* with decreasing number of estimates. Therefore, the IME and CSF
713 methods are very likely to produce reliable uncertainty estimates in the individual emission estimates and the definition and
714 derivation of their QI reflect the level of accuracy of their estimates.

715 The LCSF and GP methods display a slight correlation between most of their accuracy indicators and the number of
716 estimates. For instance, the 3rd quartiles of the distributions of relative absolute deviations remain relatively stable, varying
717 only from 46 % to 56 % and from 51 % to 59 % for the LCSF and GP methods respectively, over their entire range of
718 number of estimates. For these methods, the tradeoff between precision and number of estimates is not a critical issue and
719 retrieving an important number of estimates does not imply a significant deterioration in accuracy. On the other hand, this
720 also indicates that the current quality indicators for the GP and LCSF methods do not reflect the total/actual uncertainties in
721 their estimates.

722 As the methods present different sensitivities of the accuracy to the number of estimates, the relative performances of the
723 methods in terms of accuracy change according to the number of estimates. In other terms, as is the case for the LCSF and
724 CSF methods in Fig. 8, one method may outperform another method depending on the number of estimates we consider.
725 Indeed, below 1000 estimates, the CSF method is characterised by a better precision than the LCSF method for all the
726 statistical indicators and in particular for the 95th percentile of the deviation distribution. The best performance of the CSF
727 methods in terms of precision is then reached for ~400 estimates where the median of the deviations is ~25 % compared to
728 ~29 % for the LCSF method. But, if the number of estimates increases beyond 1000, the LCSF method starts outperforming
729 the CSF method with respect to the 95th percentile and when estimates are not filtered by their QI (right ends of the curves of
730 Fig. 8), it totally outperforms the CSF method not only in terms of precision but also in terms of number of estimates: if all
731 estimates are considered, the LCSF/CSF method generates 2722/2028 estimates whose deviations from the truth are
732 characterised by an IQR of 17 %–56 %/17 %–75 %. Furthermore, the LCSF method discards outliers much more efficiently
733 than the CSF method insofar as the 95th percentile of the deviation distribution is much lower for the former (118 %) than for
734 the latter method (341 %).

735 Selecting one method over another involves making a trade-off between precision and the number of estimates obtained.
736 Taking the example from Fig. 8, if the primary objective of an application is to obtain as many estimates as possible, the
737 LCSF method would be the preferred choice, as it can provide 2722 estimates with an IQR of the deviations ranging from 17
738 % to 56 %. On the contrary, if the main priority is to obtain estimates with the highest precision, the CSF method would be
739 more suitable, providing approximately 400 estimates with an IQR of the deviations ranging from 11 % to 45 %. The trade-
740 off between accuracy and number of estimates in the choice of method is even more accentuated in the case where inversions
741 are made with ERA5, as the use of this wind product increases the accuracy of the CSF method through bias compensation

742 (Sect. 3.4): in this case, using the CSF method, a maximum precision can be obtained, with an IQR equal to 11 %–42 %, for
743 650 estimates. If, on the other hand, the LCSF method is used, a maximum number of estimates, 2670, can be obtained with
744 an IQR of 18 %–55 % (Fig. A6).

745 The difficulty in achieving the best possible precision for a given method lies in determining an appropriate QI for their
746 estimates. Here, we adopted a relatively simple approach by defining high-quality estimates as those with relative or absolute
747 errors below the 10th percentile of the distribution relative to all the uncertainties of the estimates. However, as seen in the
748 curves of Fig. 8, highest precision may not be achieved at this value but at a higher one as in the examples of the IME and
749 CSF method. This is because misleading estimates, such as those resulting from the overlap of plumes from two sources, can
750 be characterised by very small uncertainties but at the same time by important deviations from the truth, and their impact on
751 the results becomes significant when the number of estimates gets relatively small. More generally, the QIs defined in this
752 study reflect the actual uncertainties in the estimates more or less well and the definition of a more reliable QI that ensures
753 increased accuracy with higher values of the indexes and deliver the maximum achievable precisions for all of the methods
754 is beyond the scope of this study, as it likely requires extensive studies in order to provide a common and an accurate
755 characterization of the total uncertainties in the estimates for all the inversion methods. Finally, we will note that all the
756 qualitative insights stated above about the relationships between accuracy and number of estimates are also valid when
757 considering inversions using cloud-filtered data and ERA5 winds (Fig. A7).

758 **5.3 Single methods vs ensemble approaches**

759 In this study, we create ensemble approaches by averaging the single-image estimates – for the same source and from the
760 same individual image – produced by different inversion methods. The aim is to obtain more robust and reliable predictions
761 if individual biases and errors associated with each approach compensate each other. We want thus to analyse whether an
762 ensemble method, although more expensive from a computational point of view, would perform quantitatively better than a
763 single method among CSF, GP and LCSF; these methods clearly outperforming the IME method in terms of accuracy and
764 number of estimates.

765 Four sets of ensemble approaches are considered: the first one integrates the CSF, GP and LCSF inversion methods, and
766 the remaining three ensemble approaches integrate pairs of methods (CSF & GP, CSF & LCSF and GP & LCSF). Moreover,
767 in order to assess the impact of the QIs of the different inversion methods on the performance of the ensemble methods,
768 results are analysed by considering 1) all the estimates and 2) only the best estimates produced by each method. As results
769 are assessed for the inversions using ERA5 winds and cloud-filtered data which provide a relatively small number of
770 estimates, we consider the best estimates as the estimates whose relative/absolute errors are below the 25th percentile of their
771 respective error distribution.

772 The ensemble approaches do not provide clear improvements in terms of estimate accuracy over the individual methods
773 from which they are derived (Fig. 9), with the exception of the important number of outliers produced by the CSF method
774 when estimates are not filtered: the 95th percentile of the deviation distribution is equal to 286 % for the CSF method only,

775 while it decreases to 160 % for the ensemble approach gathering the CSF, GP and LCSF methods. On the other hand, the
776 skewness of the CSF distribution of deviations lead to an increase of the 95th percentile of the deviations of the ensemble
777 approaches compared to the 95th percentiles of the LCSF and GP methods. Otherwise, the IQR of the deviations are similar
778 for all the ensemble and individual approaches and roughly ranges from 15 % to 65 % when estimates are not selected based
779 on their uncertainty and from 15 % to 60 % when the best estimates are selected. Therefore, errors and biases in the
780 estimates produced by a given method are generally not compensated by the estimates of other inversion methods which
781 suggest that in general, for the same images and sources, the estimates produced by other inversion methods may also
782 present larger errors or similar biases.

783 The great benefit of using ensemble approaches lies in the significant increase in the number of estimates, which is a
784 crucial issue in the real world when the amount of satellite data is strongly limited by the cloud cover. The ensemble
785 approach gathering the CSF, GP and LCSF methods can supply a maximum of 412 estimates over the year analysed in this
786 study, representing a 30 % increase compared to the LCSF method which is the individual method that supplies the most
787 estimates (318). This result indicates that the CSF, GP and LCSF methods can provide estimates from different images, i.e. if
788 one method does not provide an estimate from a given image, another method from the ensemble may, conversely, provide
789 one (Fig. A8). This allows the ensemble method to produce a maximum number of estimates (412) that is close to the
790 number of usable satellite images (~500). When only best estimates are considered, the ensemble approach generates more
791 than twice as many values compared to the LCSF method (195 vs 80) whereas the other ensemble approaches (CSF & GP,
792 CSF & LCSF and GP & LCSF) only provide about 140 estimates.

793 While combining the estimates generated by the CSF, GP and LCSF methods seems to be the optimal choice for an
794 ensemble approach providing the largest number of predictions, the computational cost of using these methods together may
795 not outweigh the benefits in terms of number of estimates compared to using a single method. For example, in the most
796 realistic scenario of inversions conducted with cloud-filtered data and ERA5 winds, the computational time required for the
797 CSF-GP-LCSF ensemble method is more than three times that of the LCSF method alone (see Sect. 2.1) whereas the overall
798 precision of the LCSF method is better and the increase in the number of estimates is only 30 % when using the ensemble
799 approach. Therefore, if the performance of computer systems remains an important factor to take into account, one would
800 prefer to use the LCSF method, which is the fastest method of this study, instead of using an ensemble approach.

801 In order to investigate the benefit of using ensemble approaches for the estimation of annual emissions, we use the same
802 three individual methods that produce much better results than the IME and Div methods (see Sect. 4.1), but we consider
803 different definitions of the annual estimates depending on the inversion method: annual estimates are arithmetic means of
804 image estimates for the LCSF and the GP methods whereas they are weighted means for the CSF method. This choice
805 corresponds to the best performance at the annual scale that has been found in this study for each method (Sect. 4.1.)
806 Besides, no selection of the estimates was performed to compute the annual estimates although the quality of the estimates is
807 integrated within the annual estimates of the CSF method which are averages weighted by the errors in the estimates. Among
808 the ensemble methods considered here, only the approach gathering the CSF and GP methods yields better results than the

809 best individual method composing it for most of benchmarking scenarios (Fig. A9). For example, when inversions are
810 performed with cloud-filtered data and SMARTCARB winds, the CSF, GP and their ensemble approach are characterised by
811 relative RMSE equal to 18 %, 20 % and 16 %, respectively. The benefit of using ensemble methods for estimating annual
812 estimates is thus questionable, especially considering that the gain in accuracy, if any, is very small compared to the
813 individual methods which, depending on the inversion scenario, produce the more accurate annual estimates. This is due to
814 the fact that the inversion methods generate annual estimates that are generally biased in the same way: emissions of strong
815 sources are generally underestimated while emissions of weak sources are generally overestimated (see median values in
816 Fig. 6).

817 **6 Conclusions**

818 In this paper, we tested and benchmarked several lightweight data-driven inversion methods for estimating local (city and
819 power plant) emissions from XCO₂ and NO₂ satellite images. The five methods that have been studied are the Integrated
820 Mass Enhancement (IME), the Cross-Sectional Flux (CSF), the Gaussian Plume (GP), the Light Cross-Sectional Flux
821 (LSCF) and the Divergence (Div); this last method generating only annual estimates. In a domain centred over the city of
822 Berlin, which extends about 750 km in the east-west and 650 km in the south-north direction, inversions were performed
823 with almost one year of synthetic SMARTCARB XCO₂ and tropospheric column NO₂ satellite observations with similar
824 characteristics as the upcoming CO2M mission. The ability of the inversion methods to estimate emissions has been assessed
825 by comparing the deviations of estimates from the corresponding “true” values used in the simulations, for 16 sources
826 including the city of Berlin and 15 power plants. To get a complete overview of performance, several benchmarking
827 scenarios were considered in order to analyse the benefit of using auxiliary NO₂ data or the impacts of the cloud cover in the
828 data or of uncertainties in the wind data.

829 In terms of quantifying emissions from single satellite images, the implementations of the CSF, GP and LCSF methods
830 used in this study outperform that of the IME method. Furthermore, we have demonstrated that the performance in terms of
831 accuracy and number of estimates varies, to a greater or a lesser extent depending on the method, with the selection of the
832 estimates based on their relative or absolute uncertainty. The overall accuracies of the IME and CSF methods are
833 significantly enhanced when a strict screening for high quality estimates is applied but at the cost of an important decrease in
834 the number of estimates. The GP and LCSF methods, on the other hand, perform more robustly showing only a variation in
835 their global precisions with increasing quality screening. This behaviour points out the need for these methods of a better
836 characterization of the uncertainties in the estimates. When estimates are filtered, the CSF method yields the best results in
837 terms of accuracy while, when estimates are not filtered, the LCSF method provides the highest number of estimations with
838 a slight decrease in accuracy. Overall, the CSF, GP and LCSF methods show similar accuracies for all the benchmarking
839 scenarios and when the less reliable estimates of the CSF method are removed: most of IQRs of the absolute deviations
840 range from 15 % to 60 % with an average median around 35 %. Moreover, for the most realistic benchmarking scenario, i.e.

841 for the inversions using cloud-filtered NO₂ & CO₂ data and ERA5 winds, the IME, CSF, GP and LCSF methods generate on
842 average 6 (IME), 18 (CSF), 17 (GP) and 20 (LCSF) estimates per source and per year with great differences between sources
843 (See Sect. 3.3), which is equivalent to a maximum number of estimates equal to 96 (IME), 295 (CSF), 274 (GP) and 318
844 (LCSF) for all 16 sources. These figures are significantly lower than the number of usable images (~500) that can provide a
845 hypothetical constellation of 3 satellites as analysed here; this suggests that methodological improvements could increase the
846 number of estimates.

847 The accuracy of the CSF and IME methods was found to depend on the strength of the sources with important errors
848 when determining low emissions; the GP and LCSF methods, in contrast, show similar performances across different ranges
849 of emissions. Moreover, the advantage of using co-located NO₂ signal for plume detection and quantification appeared to be
850 clear for the CSF, IME and GP methods, for which the number of single-image estimates significantly increased, while it
851 was rather weak for the LCSF method. When a cloud cover mask was taken into account in the data, the number of estimates
852 significantly decreased for all the inversion methods with an average reduction of 85 %; the global precision however hardly
853 decreased and even improved for the IME method. For all the inversion methods, the sensitivities of the results to wind
854 uncertainties were surprisingly found to be insignificant when replacing the SMARTCARB winds (used in the simulation)
855 by ERA5 reanalysis winds. Finally, if we do not take computational cost into account, the interest in using ensemble
856 approaches instead of a single method lies mainly in an increased number of single-image estimates as the availability of
857 estimates from the different methods complements each other.

858 Part of the effectiveness of the implementations of the cross-sectional flux method may come from the generation of
859 multiple estimates of cross-sectional fluxes along plumes and the subsequent averaging in order to get an unique emission
860 estimate for a given source and satellite overpass. Probably, errors in the satellite data or in the simplifying assumptions of
861 the cross-sectional approaches partly cancel out when averaging. The CSF implementation uses a complex algorithm of
862 plume detection which makes it possible to use the total detectable plume, probably leading to more accurate estimates than
863 for the LCSF implementation, which only uses observations near the source. However, the plume detection and the
864 computation of the curved centreline can fail for weak sources (i.e. short plumes) at the cost of having a large number of
865 outliers. On the contrary, the LCSF implementation uses a simpler but more robust algorithm that uses the wind vector to
866 estimate the location of the plume, which likely explains why this method generates more estimates, and without the need of
867 NO₂ data, compared to the CSF implementation. However, efforts should be made to correct the systematic underestimation
868 of strong emissions by the LCSF implementation. A way forward can be merging the CSF and LCSF method into a single
869 algorithm that takes the advantages of both approaches.

870 When compared to other methods, the relative ability of the GP method in estimating emissions probably relies on the
871 use of a Gaussian function whose optimization determines the emissions while taking into account the entire structure of the
872 plumes, and calculating effective winds that are consistent with that of the plumes. However, this optimization and thus the
873 performance of the GP method highly depend on the first-guessed values to be assigned to its parameters (not shown). And,
874 in this study, the first-guessed values of the emissions are the summer average emissions for each source; this could be a

875 strong constraint on the estimated values and could lead to an overestimation of the GP performance in this benchmarking
876 study. Finally, the GP method is computationally expensive due to the heavy plume detection algorithm and to the multi-
877 parameter optimization required for the Gaussian fitting of the plumes (Table 1).

878 The IME method also integrates information retrieved from the entire structure of the plumes but, contrarily to the GP
879 method, it does not use this information when computing effective winds. Therefore, these winds may be inconsistent with
880 the characteristic lengths of plumes used by the IME method to estimate CO₂ emissions (Sect. 2.1.4) and this could explain
881 the relatively poor performance of the IME method in this study. Varon et al. (2018) probably found that the IME method
882 was adapted to estimate CH₄ emissions from high-resolution plumes because they inferred a relationship between the
883 effective winds and the characteristic lengths through LES simulations. Another drawback of the IME method is that it is
884 very sensitive to missing data as it needs an entire coverage of the plume area by data to efficiently integrate the total mass
885 enhancement. Other single-image methods (GP, CSF and LCSF) are less sensitive to missing data as they fit functions to the
886 data and can handle data gaps; this explains why these methods provide a much larger number of estimates when the impact
887 of cloud cover on the data is considered (see Sect. 3.3).

888 In this study, we chose not to analyze the potential of the divergence method for estimating instant emissions from single
889 satellite overpasses because of the lack of studies on such an application of this method. As highlighted in the introduction
890 section, our aim is to compare proven approaches for the local scale estimation of strong sources (such as the application of
891 the divergence method to time-averages of satellite images). Moreover, the strong spatial variability of the divergence fields
892 derived from single images suggest that only averaged fields could be processed properly with the version of the divergence
893 approach which is used here for annual estimates and which relies on the peak-fitting of temporally averaged divergence
894 fields. However, we have conducted some preliminary analysis on a version of the divergence method which instead
895 integrates the divergence signal spatially (over disks centered on the sources). The results, documented in appendix A,
896 demonstrate that with a range of integration radii close to that of the spatial resolution of image, this approach can yield
897 estimates that would be comparable in terms of accuracy and quantity to that of the best inversion methods of our benchmark
898 evaluation for single-image based estimates. A better understanding of the behavior of this approach as a function of the
899 integration radius, and an assessment of the estimation errors are needed to conduct a proper comparison to the other
900 methods. This deserves further investigations. However, these preliminary results raise optimistic perspectives regarding the
901 potential of using the divergence method for estimating instant emissions from single-overpass images.

902 For estimating annual emissions, the CSF, GP and LCSF methods outperform the Div and IME methods when annual
903 estimates are computed as error-weighted means of single-image estimates for the CSF method and as arithmetic means of
904 these estimates for the GP and LCSF methods. Across the different benchmarking scenarios, the GP method shows better
905 precisions in its annual estimates because its single-image estimates have similar absolute deviations from the truth but are
906 less affected by biases compared to the CSF and LCSF methods (see Fig. 3). However, despite biases, errors in the single-
907 image estimates provided by the CSF, GP and LCSF methods likely compensate when averaging and these methods also
908 generate annual estimates with a better precision than for their single-image estimates. In the most realistic benchmarking

909 scenario – where inversions use cloud-filtered XCO₂ & NO₂ data and ERA5 winds and where performances are the lowest
910 compared to other scenarios – the relative RMSE for the annual emissions of the 16 sources is 20 % (GP), 27 % (CSF), 31 %
911 (LCSF), 55 % (IME) and 79 % (Div). The relatively weak performance of the Div method could be explained by the fact that
912 this method was originally developed for the estimation of NO_x emissions and the fields of this chemical species are
913 generally characterised by stronger divergence peaks than for CO₂ fields. Its performance may also be hindered by the fact
914 that our implementation of this method does not select the overpasses from which the annual divergence maps are derived
915 (see Sect. 4.1). Further investigation is needed to determine whether the filtering of overpasses which could be favourable to
916 the method could strongly increase the accuracy of its annual estimates. The performances of ensemble approaches gathering
917 several inversion methods in terms of annual estimations is not better, and in some cases even worse, than the individual
918 methods. Finally, none of the methods were able to correctly reproduce the monthly seasonal cycle of the emissions when
919 data underwent a cloud-filtering, i.e. when data were not available for some months, which points out the need for an
920 extensive temporal coverage of the observations when aiming to capture the monthly variability in emissions.

921 In addition to the technical improvements that could be made on the algorithms of the methods, further developments
922 could extend this study such as the integration of new data streams for estimating CO₂ emissions such as satellite data of
923 other co-emitted gases than NO₂, e.g. CO data provided by the TROPOMI instrument. A companion paper (Hakkarainen et
924 al., 2024) analyses the ability of the inversion methods in determining NO_x emissions, from synthetic and TROPOMI NO₂
925 satellite data for the Matimba and Medupi power plants in South-Africa. The NO₂ synthetic data are extracted from the high-
926 resolution MicroHH Large Eddy Simulations (LES) (Van Heerwaarden et al., 2017) and used in particular to study the
927 nitrogen dioxide to nitrogen oxide scaling factors that are required for satellite-based estimations of NO_x emissions.
928 Moreover, the capacity of the inversion methods to estimate city emissions has been analysed in this study on the single
929 example of the city of Berlin and, as most of the methods have provided correct estimates for its emissions, it would be
930 interesting to expand this study to other cities and other local sources. Finally, this benchmarking study has not integrated the
931 new and promising type of inversion methods that are the methods derived from deep learning techniques (e.g. Lary et al.,
932 2016). After a potentially complex training phase, deep-learning methods could quickly process large amounts of data and
933 provide estimations with similar or better accuracy than the methods studied here (Dumont le Brazidec et al., 2023). They
934 could also complement these methods by allowing a fine differentiation of the plumes compared to the background with
935 advanced image segmentation techniques.

936 The aim of this study is to contribute to the development of the CO₂ Monitoring and Verification Support system that
937 will use the upcoming CO2M satellite data. And, although this benchmarking study has been performed with synthetic
938 observations, the methods studied here can be easily adapted to the analysis of real satellite observations and to deal with
939 sources of unknown location as demonstrated in Hakkarainen et al. (2024).

940
941
942

944 **Appendix A: Potential of the divergence approach to estimate local CO₂ emissions from single-overpass satellite**
 945 **images of XCO₂ and NO₂**

946 In this study, the performance of the divergence approach to estimate local CO₂ emissions from XCO₂ and NO₂ synthetic
 947 satellite images is assessed with a standard version of this approach (e.g., Beirle et al., 2021; Hakkarainen et al., 2022),
 948 which provides temporally averaged estimates. Results concerning the divergence approach are thus analyzed in the main
 949 part of this paper in terms of annual means. However, following the suggestions of a reviewer (S. Beirle), we also tested the
 950 potential of this method to estimate instant emissions using single-overpass images. For this purpose, we have used two
 951 versions of the divergence approach that have been modified for single image geometry as in Beirle et al. (2023).

952 For both versions, the computation of the divergence fields is performed by only considering the “advective” term
 953 ($10^6 * M_{air} * U * \nabla(VCD)$) of the full expression of the horizontal flux divergence ($\nabla(10^6 M_{air} * U * VCD)$) where M_{air} is
 954 the dry air mass, U is the wind vector and VCD is the vertical column density in parts per million. Such reformulation of the
 955 divergence method that does not compute the divergence of the wind term was also used by Beirle et al. (2023) for NO₂. The
 956 advantage of this reformulation for CO₂ is that the background (e.g., a constant offset of 400 ppm) is implicitly removed.

957 These versions of the divergence approach differ from each other in their way of computing emissions from the
 958 divergence maps associated with single-overpass images: the first version integrates the divergence fields on disks centered
 959 on the sources (Figure A10). And, to mitigate the impact of the uncertainties in the observations, the emission estimate for a
 960 given satellite overpass and source can be computed as the average of the estimates when integrating the divergence signal
 961 on disks of different radii. This version of the divergence approach will be referred to hereinafter as the *integral* divergence
 962 method. The second version proceeds in a similar way to the one used in the main part of the article and fits a 2-D Gaussian
 963 function to the divergence maps in order to retrieve source emissions (e.g. Beirle et al. 2020). The modified peak fitting
 964 model is similar to the original but with a reduced number of estimated parameters. Namely, the parameters related to the
 965 background and to the location correction are removed from the model parameters. This version of the divergence approach
 966 will be referred to hereinafter as the *peak-fitting* divergence method.

967 For both versions, potential peaks are detected by using NO₂ fields which are integrated over disks of 6 km radius
 968 centered on the sources. If the integral of the divergence map on the disk is larger than the integral on the area outside the
 969 disk, then the enhancement, related to a given source and for a given satellite overpass, is considered strong enough and the
 970 emission estimation can be carried out. Many sources in the SMARTCARB dataset are weak and enhancements may be
 971 barely visible which causes challenges for both versions.

972 To evaluate the potential of these two versions of the divergence approach, we use the SMARTCARB dataset described
 973 in section 2.2. which provides about 3000 images to determine the emissions of the 16 local sources that are considered in
 974 this study (if we take into account the cloud cover, only 500 images remain usable). Furthermore, we consider two

975 benchmark scenarios (see table 2 and section 2.3) where inversions are performed using CO₂ and NO₂ data with
976 SMARTCARB winds. In one case, we use cloud-free data, while in the other, cloud-filtered data.

977 The analysis of the deviations from the truth of the instant estimates shows that the integral divergence approach is
978 strongly sensitive to the radius of the integration disks (Fig. A11). No clear trend appears except that errors increase sharply
979 for a radius greater than 10 km, with a significant presence of outliers. Below this value, the absolute relative deviations
980 (bottom panel of Fig. A11) can increase or decrease depending on the value of the radius. Furthermore, the integral
981 divergence approach can underestimate or overestimate emissions depending if the radius is lower or greater than ~4 km. A
982 possible explanation for this behavior could be that the impacts of the two main sources of errors in the divergence method
983 — namely, the uncertainties in the observations and the influence of additional but unwanted sources on the background of
984 the divergence fields — evolve in opposite directions as the integration radius increases. The impact of the uncertainties is
985 mitigated when the area of the integration disk increases because errors have more probability to cancel out. Conversely, the
986 impact of neighboring sources on the background of the divergence field intensifies as the integration radius increases,
987 because the likelihood of capturing features in the divergence maps that are not directly related to the emissions of the
988 targeted sources grows. This impact consistently introduces a positive bias in the estimates (as we capture more sources) and
989 is likely more important than the one related to the uncertainties as performance overall degrades when the integration radius
990 increases.

991 The peak-fitting divergence method is characterized by a poor performance compared to the integral divergence method
992 for the ensemble of integration radii that we have considered here (Fig. A11). The estimation of small emitting sources may
993 be more difficult for the peak-fitting version as the fit of the 2-D Gaussian function to the data associated to these sources
994 often fails and does not provide optimal and reliable parameter combinations, yielding poor and often overestimated
995 emission estimations. Therefore, even though the peak-fitting divergence method is generally more efficient at the annual
996 scale, these results suggest that it is not the case when estimating instant emissions from single overpass images.

997 The configuration of the integral divergence method which averages estimates across the integration radii of 2, 3 and 4
998 km shows the best performance amongst the configurations that we have tested. Probably, the impacts of the data
999 uncertainties and the background are well balanced for this range of radii and the fact of averaging estimates across three
1000 different radii further reduces the influence of the data uncertainties on the results. When compared to other inversion
1001 methods analyzed in this study, the performance of this configuration of the integral divergence method is similar to that of
1002 the best inversion methods (Fig. A12). For the benchmarking scenario considering cloud-free data, its relative absolute
1003 deviations are for example characterized by a median value of ~38 % and Interquartile Range (IQR) of [~19 % – ~64 %]
1004 which are comparable to deviations associated to the Light Cross-Sectional Flux (LCSF) method which have a median value
1005 of ~32 % and an IQR of [~15 % – ~56 %]. Notably, the integral divergence method generates fewer estimates (2174)
1006 compared to the LCSF method (2722), but more than the Gaussian Plume (GP) method (1776).

1007 These preliminary results regarding the potential of the integral divergence method for estimating local CO₂ emissions
1008 from single-overpass images of XCO₂ and NO₂ appear promising, especially since this method allows for the detection of

1009 plumes from unknown sources (Beirle et al., 2021). However, further investigation is required to properly assess factors such
1010 as the integration radius based on data resolution, and to generalize this method to various types of satellite data.
1011 Additionally, a thorough quantitative error assessment is essential to evaluate the accuracy of the estimates, enabling the
1012 classification and selection of estimates, which would enhance the method's overall performance.

1013
1014 *Code and data availability.* The code repository of the python package *ddeq* is available on Gitlab.com:
1015 <https://gitlab.com/empa503/remote-sensing/ddeq>. The SMARTCARB dataset is available on Zenodo:
1016 <https://doi.org/10.5281/zenodo.4048227>.

1017
1018 *Author contributions.* DS made the diagnostics and led the analysis for the intercomparison of the results from the different
1019 inversion methods. All co-authors contributed to the decisions for the configuration, diagnostics and analysis of the
1020 intercomparison. DS wrote the manuscript with inputs from all co-authors. DS, GB and FC carried out the analysis specific
1021 to the LCSF method. JH, II, HL, JN and LA carried out the analysis specific to the Div method. GK developed the original
1022 *ddeq* library that has been used as a basis for the application of the different methods. GK provided the SMARTCARB
1023 dataset used to test the different methods. GK carried out the analysis specific to the IME method. EK carried out the
1024 analysis specific to the CSF and GP inversion methods. The project was coordinated by JT, DB and GB.

1025
1026 *Competing Interests.* Some authors are members of the editorial board of Atmospheric Measurement Techniques. The
1027 authors have no other competing interests to declare.

1028
1029 *Acknowledgements.* Most of the work performed in this paper was done in the framework of EU H2020 project CoCO2
1030 (grant No. 958927). The FMI team would like to thank the Research Council of Finland project 353082. All authors would
1031 like to thank the ICOS Carbon Portal for providing access to their JupyterLab servers, which were used for code development
1032 and data sharing. Finally, the authors would like to thank the two reviewers for their insightful comments, and especially S.
1033 Beirle for his suggestions on the application of the divergence approach for estimating instant emissions.

1034 **References**

1035 Beirle, S., Borger, C., Dörner, S., Li, A., Hu, Z., Liu, F., et al.: Pinpointing nitrogen oxide emissions from space. *Science*
1036 *Advances* 5. doi:10.1126/sciadv.aax9800, 2019.

1037 Beirle, S., Borger, C., Dörner, S., Eskes, H., Kumar, V., de Laat, A., et al.: Catalog of NO_x emissions from point sources as
1038 derived from the divergence of the NO₂ flux for TROPOMI. *Earth System Science Data* 13, 2995–3012. doi:10.5194/essd-
1039 13-2995-2021, 2021.

1040 Beirle, S., Borger, C., Jost, A., and Wagner, T.: Improved catalog of NO_x point source emissions (version 2), *Earth Syst. Sci.*
1041 *Data*, 15, 3051–3073, <https://doi.org/10.5194/essd-15-3051-2023>, 2023.

1042 Boersma, K. F., Eskes, H. J., Dirksen, R. J., van der A, R. J., Veefkind, J. P., Stammes, P., Huijnen, V., Kleipool, Q. L.,
1043 Sneep, M., Claas, J., Leitão, J., Richter, A., Zhou, Y., and Brunner, D.: An improved tropospheric NO₂ column retrieval
1044 algorithm for the Ozone Monitoring Instrument, *Atmos. Meas. Tech.*, 4, 1905–1928, [https://doi.org/10.5194/amt-4-1905-](https://doi.org/10.5194/amt-4-1905-2011)
1045 2011, 2011.

1046 Bovensmann, H., Buchwitz, M., Burrows, J. P., Reuter, M., Krings, T., Gerilowski, K., et al.: A Remote Sensing Technique
1047 for Global Monitoring of Power Plant CO₂ Emissions from Space and Related Applications. *Atmos. Meas. Tech.*
1048 3, 781–811. doi:10.5194/amt-3-781-2010, 2010.

1049 Broquet, G., Bréon, F.-M., Renault, E., Buchwitz, M., Reuter, M., Bovensmann, H., et al.: The Potential of Satellite Spectro-
1050 Imagery for Monitoring CO₂ Emissions from Large Cities. *Atmos. Meas. Tech.* 11, 681–708. doi:10.5194/amt-11-681-2018,
1051 2018.

1052 Brunner, D., Kuhlmann, G., Marshall, J., Clément, V., Fuhrer, O., Broquet, G., Löscher, A., and Meijer, Y.: Accounting for
1053 the vertical distribution of emissions in atmospheric CO₂ simulations, *Atmos. Chem. Phys.*, 19, 4541–4559,
1054 <https://doi.org/10.5194/acp-19-4541-2019>, 2019.

1055 Brunner, D., Kuhlmann, G., Henne, S., Koene, E., Kern, B., Wolff, S., ...Fix, A.: Evaluation of simulated CO₂power plant
1056 plumes from six high-resolution atmospheric transport models. *Atmospheric Chemistry and Physics*, 23(4), 2699-2728, 2023

1057 Buchwitz, M., Reuter, M., Bovensmann, H., Pillai, D., Heymann, J., Schneising, O., et al.: Carbon Monitoring Satellite
1058 (CarbonSat): Assessment of Atmospheric CO₂ and CH₄ Retrieval Errors by Error Parameterization. *Atmos. Meas. Tech.* 6,
1059 3477–3500. doi:10.5194/amt-6-3477-2013, 2013.

1060 Chevallier, F., Feng, L., Bösch, H., Palmer, P. I., and Rayner, P. J.: On the impact of transport model errors for the
1061 estimation of CO₂ surface fluxes from GOSAT observations, *Geophys. Res. Lett.*, 37,
1062 21, <https://doi.org/10.1029/2010GL044652>, 2010.

1063 Chevallier, F., Zheng, B., Broquet, G., Ciais, P., Liu, Z., Davis, S. J., et al.: Local anomalies in the column-averaged dry air
1064 mole fractions of carbon dioxide across the globe during the first months of the coronavirus recession. *Geophysical Research*
1065 *Letters*, 47, e2020GL090244. <https://doi.org/10.1029/2020gl090244>, 2020.

1066 Chevallier, F., Broquet, G., Zheng, B., Ciais, P., & Eldering, A.: Large CO₂ emitters as seen from satellite: Comparison to a
1067 gridded global emission inventory. *Geophysical Research Letters*, 49, e2021GL097540.
1068 <https://doi.org/10.1029/2021GL097540>, 2022.

1069 Ciais, P., Crisp, D., v. d. Gon, H., Engelen, R., Heimann, M., Janssens-Maenhout, G., Rayner, P., and Scholze, M.: Towards
1070 a European Operational Observing System to Monitor Fossil CO₂emissions – Final Report from the expert group,
1071 Copernicus climate Change Service, Report, European Commission, Brussels, 2015.

1072 Crisp, D., Pollock, H. R., Rosenberg, R., Chapsky, L., Lee, R. A. M., Oyafuso, F. A., et al.: The on-orbit performance of the
1073 Orbiting Carbon Observatory-2 (OCO-2) instrument and its radiometrically calibrated products. *Atmos. Meas. Tech.* 10, 59–
1074 81. doi:10.5194/amt-10-59-2017, 2017.

1075 Dumont Le Brazidec, J., Vanderbecken, P., Farchi, A., Broquet, G., Kuhlmann, G., & Bocquet, M.: Deep learning applied to
1076 CO₂ power plant emissions quantification using simulated satellite images. *Geoscientific Model Development*, 17(5), 1995–
1077 2014. 2024.

1078 Düring, I., Bächlin, W., Ketzler, M., Baum, A., Friedrich, U., and Wurzler, S.: A New Simplified NO/NO₂ Conversion Model
1079 under Consideration of Direct NO₂-Emissions. *metz* 20, 67–73. doi:10.1127/0941-2948/2011/0491, 2011.

1080 Ehret, T., De Truchis, A., Mazzolini, M., Morel, J. M., D’aspremont, A., Lauvaux, T., ... & Facciolo, G.: Global tracking and
1081 quantification of oil and gas methane emissions from recurrent sentinel-2 imagery. *Environmental science & technology*,
1082 56(14), 10517-10529, 2022.

1083 Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski,
1084 K., Sweeney, C., and Conley, S.: Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners
1085 region, *P. Natl. Acad. Sci. USA*, 113, 9734–9739, <https://doi.org/10.1073/pnas.1605617113>, 2016.

1086 Hakkarainen, J., Ialongo, I., and Tamminen, J.: Direct space-based observations of anthropogenic CO₂ emission areas from
1087 OCO-2. *Geophysical Research Letters* 43, 11,400–11,406. doi:10.1002/2016GL070885, 2016.

1088 Hakkarainen, J., Ialongo, I., Koene, E., Szelağ, M., Tamminen, J., Kuhlmann, G., and Brunner, D.: Analyzing local carbon
1089 dioxide and nitrogen oxide emissions from space using the divergence method: An application to the synthetic
1090 SMARTCARB dataset. *Frontiers in Remote Sensing* 3. doi:10.3389/frsen.2022.878731, 2022.

1091 Hakkarainen, J., Ialongo, I., Oda, T., Szelağ, M. E., O’Dell, C. W., Eldering, A., and Crisp, D.: Building a bridge:
1092 Characterizing major anthropogenic point sources in the South African Highveld region using OCO-3 carbon dioxide
1093 Snapshot Area Maps and Sentinel-5P/TROPOMI nitrogen dioxide columns. *Environmental Research Letters*, 18(3),
1094 doi:10.1088/1748-9326/acb837, 2023a.

1095 Hakkarainen, J., Tamminen, J., Nurmela, J., Lindqvist, H., Santaren, D., Broquet, G., Chevallier, F., Koene, E., Kuhlmann,
1096 G. and Brunner, D.: Benchmarking of plume detection and quantification methods. Technical Report. FMI. URL:
1097 <https://www.coco2-project.eu/node/366>. CoCO₂: Prototype system for a Copernicus CO₂ service, 2023b. Hakkarainen, J.,
1098 Kuhlmann, G., Koene, E., Santaren, D., Meier, S., Krol, M.C., van Stratum, B.J.H, Ialongo, I., Chevallier, F., Tamminen, J.,
1099 Brunner, D., Broquet, G.: Analyzing nitrogen dioxide to nitrogen oxide scaling factors for data-driven satellite-based
1100 emission estimation methods: a case study of Matimba/Medupi power stations in South Africa, *Atmospheric Pollution*
1101 *Research*, Volume 15, Issue 7, 2024, 102171, ISSN 1309-1042, <https://doi.org/10.1016/j.apr.2024.102171>, 2024.

1102

1103 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al.: The ERA5 global reanalysis.
1104 *Quarterly Journal of the Royal Meteorological Society*, 1, 51. <https://doi.org/10.1002/qj.3803>, 2020.

1105 Houweling, S., Aben, I., Breon, F.-M., Chevallier, F., Deutscher, N., Engelen, R., Gerbig, C., Griffith, D., Hungershofer,
1106 K., Macatangay, R., Marshall, J., Notholt, J., Peters, W., and Serrar, S.: The importance of transport model uncertainties for
1107 the estimation of CO₂ sources and sinks using satellite measurements, *Atmos. Chem. Phys.*, 10, 9981–
1108 9992, <https://doi.org/10.5194/acp-10-9981-2010>, 2010.

1109 Jacob, D. J.: Introduction to Atmospheric Chemistry, (Princeton University Press), 1999.

1110 Jacob, D. J., Varon, D. J., Cusworth, D. H., Dennison, P. E., Frankenberg, C., Gautam, R., ... & Duren, R. M.: Quantifying
1111 methane emissions from the global scale down to point sources using satellite observations of atmospheric methane.
1112 *Atmospheric Chemistry and Physics*, 22(14), 9617-9646, 2022.

1113 Jähn, M., Kuhlmann, G., Mu, Q., Haussaire, J. M., Ochsner, D., Osterried, K., ... & Brunner, D.: An online emission module
1114 for atmospheric chemistry transport models: implementation in COSMO-GHG v5. 6a and COSMO-ART v5. 1-
1115 3.1. *Geoscientific Model Development*, 13(5), 2379-2392, 2020.

1116 Janssens-Maenhout, G., Pinty, B., Dowell, M., Zunker, H., Andersson, E., Balsamo, G., et al.: Toward an Operational
1117 Anthropogenic CO₂ Emissions Monitoring and Verification Support Capacity. *Bull. Am. Meteorol. Soc.* 101, E1439–E1451.
1118 doi:10.1175/BAMS-D-19-0017.1, 2020.

1119 Kasahara, M., Kachi, M., Inaoka, K., Fujii, H., Kubota, T., Shimada, R., & Kojima, Y.: Overview and current status of
1120 GOSAT-GW mission and AMSR3 instrument. In *Sensors, Systems, and Next-Generation Satellites XXIV* (Vol. 11530, p.
1121 1153007). SPIE. 2020.

1122 Koene, E., Brunner, D. and Kuhlmann, G.: Documentation of plume detection and quantification methods. Tech. rep., Empa.
1123 CoCO₂: Prototype system for a Copernicus CO₂ service. <https://coco2-project.eu/node/329>, 2021.

1124 Koene, E. and Brunner, D.: Assessment of plume model performance. Technical Report. Empa. URL: [https://www.coco2-](https://www.coco2-project.eu/node/357)
1125 [project.eu/node/357](https://www.coco2-project.eu/node/357). CoCO₂: Prototype system for a Copernicus CO₂ service, 2023.

1126 Koene, E. F. M., Brunner, D., & Kuhlmann, G. On the theory of the divergence method for quantifying source emissions
1127 from satellite observations. *Journal of Geophysical Research: Atmospheres*, 129, e2023JD039904.
1128 <https://doi.org/10.1029/2023JD039904>, 2024.

1129 Kort, E. A., Frankenberg, C., Miller, C. E., and Oda, T.: Space-based observations of megacity carbon dioxide, *Geophys.*
1130 *Res. Lett.*, 39, L17806, <https://doi.org/10.1029/2012gl052738>, 2012.

1131 Kuenen, J. J. P., Visschedijk, A. J. H., Jozwicka, M., and Denier van der Gon, H. A. C.: TNO-MACC_II emission inventory;
1132 a multi-year (2003–2009) consistent high-resolution European emission inventory for air quality modelling, *Atmos. Chem.*
1133 *Phys.*, 14, 10963–10976, <https://doi.org/10.5194/acp-14-10963-2014>, 2014.

1134 Kuhlmann, G., Broquet, G., Marshall, J., Clément, V., Löscher, A., Meijer, Y., et al.: Detectability of CO₂ emission plumes
1135 of cities and power plants with the Copernicus Anthropogenic CO₂ Monitoring (CO₂M) mission. *Atmospheric Measurement*
1136 *Techniques* 12, 6695–6719. doi:10.5194/amt-12-6695-2019, 2019.

1137 Kuhlmann, G., Brunner, D., Broquet, G., and Meijer, Y.: Quantifying CO₂ emissions of a city with the Copernicus
1138 Anthropogenic CO₂ Monitoring satellite mission. *Atmospheric Measurement Techniques* 13, 6733–6754. doi:10.5194/amt-
1139 13-6733-2020, 2020.

1140 Kuhlmann, G., Clément, V., Marshall, J., Fuhrer, O., Broquet, G., Schnadt-Poberaj, C., et al.: Synthetic XCO₂, CO and NO₂
1141 Observations for the CO₂M and Sentinel-5 Satellites. doi:10.5281/zenodo.4048228, 2020b.

1142 Kuhlmann, G., Henne, S., Meijer, Y., and Brunner, D.: Quantifying CO₂ Emissions of Power Plants With CO₂ and NO₂
1143 Imaging Satellites. *Frontiers in Remote Sensing* 2, 14. doi:10.3389/frsen.2021.689838. 2021.

1144 Kuhlmann, G., Koene, E. F. M., Meier, S., Santaren, D., Broquet, G., Chevallier, F., Hakkarainen, J., Nurmela, J., Amorós,
1145 L., Tamminen, J., and Brunner, D.: The ddeq Python library for point source quantification from remote sensing images
1146 (Version 1.0). *Geoscientific Model Development*, 17(12), 4773–4789, <https://doi.org/10.5194/gmd-17-4773-2024>, 2024.

1147 Landgraf, J., Rusli, S., Cooney, R., Veeffkind, P., Vemmix, T., de Groot, Z., Bell, A., Day, J., Leemhuis, A., and Sierk, B.:
1148 The TANGO mission: A satellite tandem to measure major sources of anthropogenic greenhouse gas emissions, EGU
1149 General Assembly 2020, Online, 4–8 May 2020, EGU2020-19643, <https://doi.org/10.5194/egusphere-egu2020-19643>, 2020.

1150 Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L.: Machine learning in geosciences and remote
1151 sensing. *Geoscience Frontiers*, 7(1), 3-10, 2016.

1152 Mahadevan, P., Wofsy, S. C., Matross, D. M., Xiao, X., Dunn, A. L., Lin, J. C., ... & Gottlieb, E. W.: A satellite-based
1153 biosphere parameterization for net ecosystem CO₂ exchange: Vegetation Photosynthesis and Respiration Model
1154 (VPRM). *Global Biogeochemical Cycles*, 22(2), 2008.

1155 Meijer, Y., Boesch, H., Bombelli, A., Brunner, D., Buchwitz, M., Ciais, P., et al.: Copernicus CO₂ monitoring mission
1156 Requirements document (MRD). Netherlands, Europe: European Space Agency, Earth and Mission Science Division. 2019.

1157 Nassar, R., Hill, T. G., McLinden, C. A., Wunch, D., Jones, D. B. A., and Crisp, D.: Quantifying CO₂ emissions from
1158 individual power plants from space. *Geophys. Res. Lett.* 44, 10045-10053. doi:10.1002/2017GL074702, 2017.

1159 Nassar R, Moeini O, Mastrogiacomo J-P, O'Dell CW, Nelson RR, Kiel M, Chatterjee A, Eldering A and Crisp D: Tracking
1160 CO₂ emission reductions from space: A case study at Europe's largest fossil fuel power plant. *Front. Remote Sens.*
1161 3:1028240. doi: 10.3389/frsen.2022.1028240, 2022.

1162 Pascal, V., Buil, C., Loesel, J., Tauziede, L., Jouglet, D., & Buisson, F.: An improved microcarb dispersive instrumental
1163 concept for the measurement of greenhouse gases concentration in the atmosphere. In *International Conference on Space*
1164 *Optics—ICSO 2014* (Vol. 10563, pp. 1028-1036). SPIE. 2017.

1165 Pillai, D., Buchwitz, M., Gerbig, C., Koch, T., Reuter, M., Bovensmann, H., et al.: Tracking City CO₂ Emissions from Space
1166 Using a High-Resolution Inverse Modelling Approach: a Case Study for Berlin, Germany. *Atmos. Chem. Phys.* 16, 9591–
1167 9610. doi:10.5194/acp-16-9591-2016, 2016.

1168 Pinty, B., Janssens-Maenhout, G., Dowell, M., Zunker, H., Brunhes, T., Ciais, P., Dee, D., Denier van der Gon, H. A. C.,
1169 Dolman, H., Drinkwater, M., Engelen, R., Heimann, M., Holmlund, K., Husband, R., Kentarchos, A., Meyer, A., Palmer, P.,

1170 and Scholze, M.: An operational anthropogenic CO₂ emissions monitoring and verification support capacity. Baseline
1171 requirements, model components and functional architecture, EUR28736 EN, European Commission Joint Research Centre,
1172 Ispra, Italy, <https://doi.org/10.2760/08644>, 2017.

1173 Reuter, M., Buchwitz, M., Schneising, O., Krautwurst, S., O'Dell, C. W., Richter, A., et al.: Towards Monitoring Localized
1174 CO₂ Emissions from Space: collocated Regional CO₂ and NO₂ Enhancements Observed by the OCO-2 and S5P Satellites.
1175 *Atmos. Chem. Phys.* 19, 9371–9383. doi:10.5194/acp-19-9371-2019, 2019.

1176 Santaren, D., Broquet, G., Bréon, F.-M., Chevallier, F., Siméoni, D., Zheng, B., and Ciais, P.: A local- to national-scale
1177 inverse modeling system to assess the potential of spaceborne CO₂ measurements for the monitoring of anthropogenic
1178 emissions, *Atmos. Meas. Tech.*, 14, 403–433, <https://doi.org/10.5194/amt-14-403-2021>, 2021.

1179 Schuit, B. J., Maasackers, J. D., Bijl, P., Mahapatra, G., van den Berg, A.-W., Pandey, S., Lorente, A., Borsdorff, T.,
1180 Houweling, S., Varon, D. J., McKeever, J., Jervis, D., Girard, M., Irakulis-Loitxate, I., Gorroño, J., Guanter, L., Cusworth,
1181 D. H., and Aben, I.: Automated detection and monitoring of methane super-emitters using satellite data, *Atmos. Chem.*
1182 *Phys.*, 23, 9071–9098, <https://doi.org/10.5194/acp-23-9071-2023>, 2023.

1183 Sierk, B., Bézy, J.-L., Löscher, A., and Meijer, Y.: The European CO₂ Monitoring Mission: Observing Anthropogenic
1184 Greenhouse Gas Emissions from Space 11180. Proceedings, International Conference on Space Optics—ICSO 2018. 12 July
1185 2019. Chania, Greece. 111800M. doi:10.1117/12.2535941. 2019

1186 Singer, A.M., Branham, M., Hutchins, M.G., Welker, J., Woodard, D. L., Badurek, C. A., et al.: The role of CO₂ emissions
1187 from large point sources in emissions totals, responsibility and policy. *Environ. Sci. Policy* 44, 190–200.
1188 doi:10.1016/j.envsci.2014.08.001, 2014.

1189 Sun, K.. Derivation of emissions from satellite-observed column amounts and its application to TROPOMI NO₂ and CO
1190 observations. *Geophysical Research Letters*, 49(23), e2022GL101102. <https://doi.org/10.1029/2022gl101102>, 2022.

1191 Taylor, T. E., O'Dell, C. W., Frankenberg, C., Partain, P. T., Cronk, H. Q., Savtchenko, A., Nelson, R. R., Rosenthal, E. J.,
1192 Chang, A. Y., Fisher, B., Osterman, G. B., Pollock, R. H., Crisp, D., Eldering, A., and Gunson, M. R.: Orbiting Carbon
1193 Observatory-2 (OCO-2) cloud screening algorithms: validation against collocated MODIS and CALIOP data, *Atmos. Meas.*
1194 *Tech.*, 9, 973–989, <https://doi.org/10.5194/amt-9-973-2016>, 2016.

1195 Van Heerwaarden, C. C., Van Stratum, B. J., Heus, T., Gibbs, J. A., Fedorovich, E., & Mellado, J. P.: MicroHH 1.0: A
1196 computational fluid dynamics code for direct numerical simulation and large-eddy simulation of atmospheric boundary layer
1197 flows. *Geoscientific Model Development*, 10(8), 3145–3165, 2017.

1198 Varon, D. J., Jacob, D. J., McKeever, J., Jervis, D., Durak, B. O. A., Xia, Y., et al.: Quantifying methane point sources from
1199 fine-scale satellite observations of atmospheric methane plumes. *Atmospheric Measurement Techniques* 11, 5673–5686.
1200 doi:10.5194/amt-11-5673-2018, 2018.

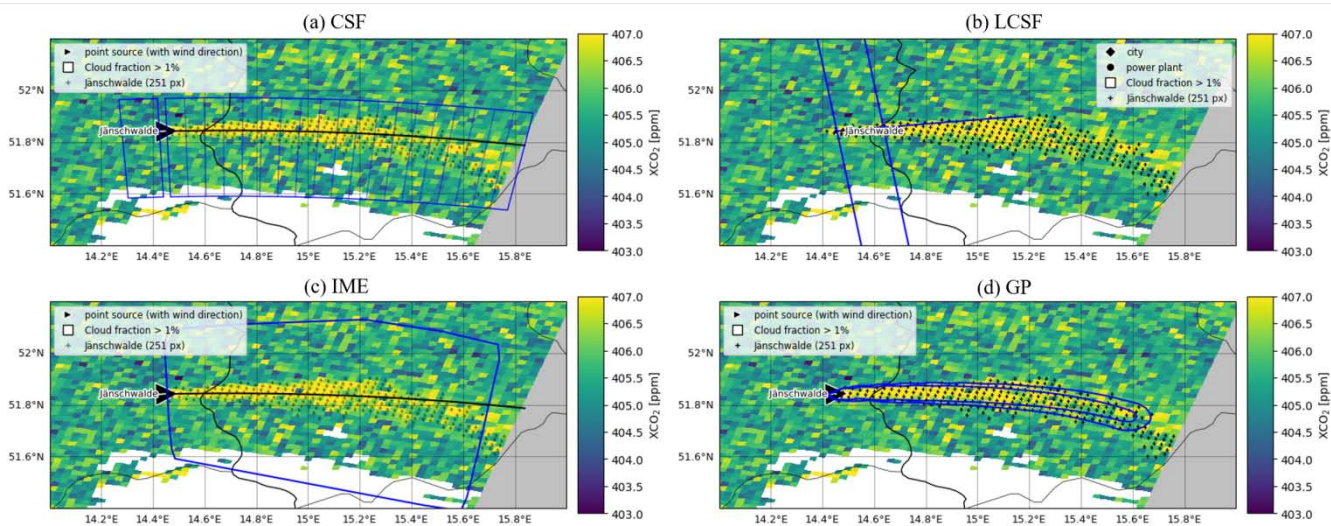
1201 Wang, Y., Broquet, G., Bréon, F.-M., Lepsinas, F., Buchwitz, M., Reuter, M., et al.: PMIF v1.0: Assessing the Potential of
1202 Satellite Observations to Constrain CO₂ Emissions from Large Cities and point Sources over the globe Using Synthetic Data.
1203 *Geosci. Model. Dev.* 13, 5813–5831. doi:10.5194/gmd-13-5813-2020. 2020.

1204 Worden, J. R., Doran, G., Kulawik, S., Eldering, A., Crisp, D., Frankenberg, C., O'Dell, C., and Bowman, K.: Evaluation
 1205 and attribution of OCO-2 XCO₂ uncertainties, *Atmos. Meas. Tech.*, 10, 2759–2771, [https://doi.org/10.5194/amt-10-2759-](https://doi.org/10.5194/amt-10-2759-2017)
 1206 2017, 2017.

1207 Ye, X., Lauvaux, T., Kort, E., Oda, T., Feng, S., Lin, J., Yang, E., & Wu, D.: Constraining Fossil Fuel CO₂ Emissions From
 1208 Urban Area Using OCO-2 Observations of Total Column CO₂. *Journal of Geophysical Research: Atmospheres*, 1-29, 2020.

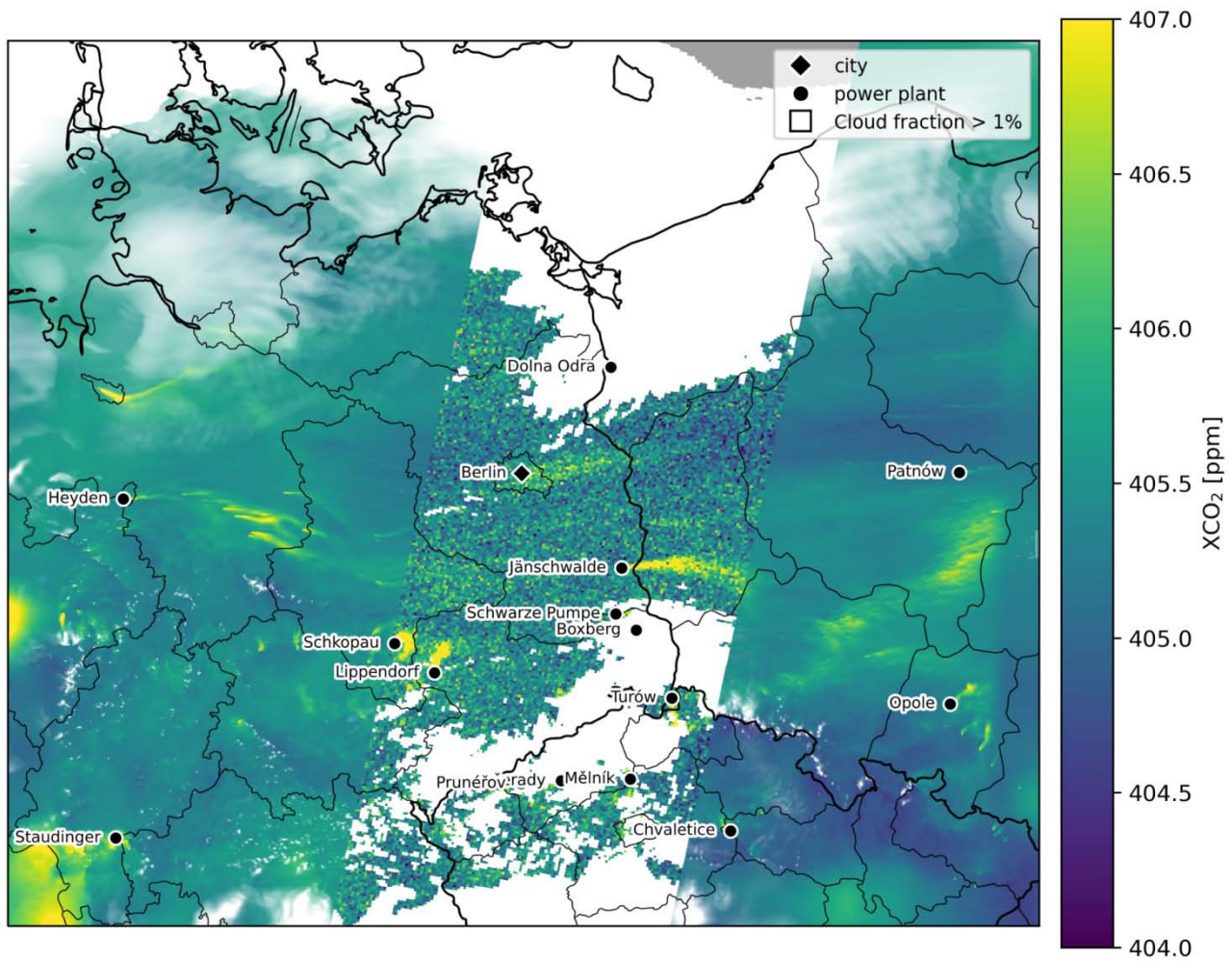
1209 Zheng, T., Nassar, R., and Baxter, M.: Estimating power plant CO₂ emission using OCO-2 XCO₂ and high resolution WRF-
 1210 Chem simulations. *Environ. Res. Lett.* 14, 085001. doi:10.1088/1748-9326/ab25ae. 2019.

1211 Zheng, B., Chevallier, F., Ciais, P., Broquet, G., Wang, Y., Lian, J., et al.: Observing Carbon Dioxide Emissions over
 1212 China's Cities and Industrial Areas with the Orbiting Carbon Observatory-2. *Atmos. Chem. Phys.* 20, 8501–8510.
 1213 doi:10.5194/acp-20-8501-2020. 2020.



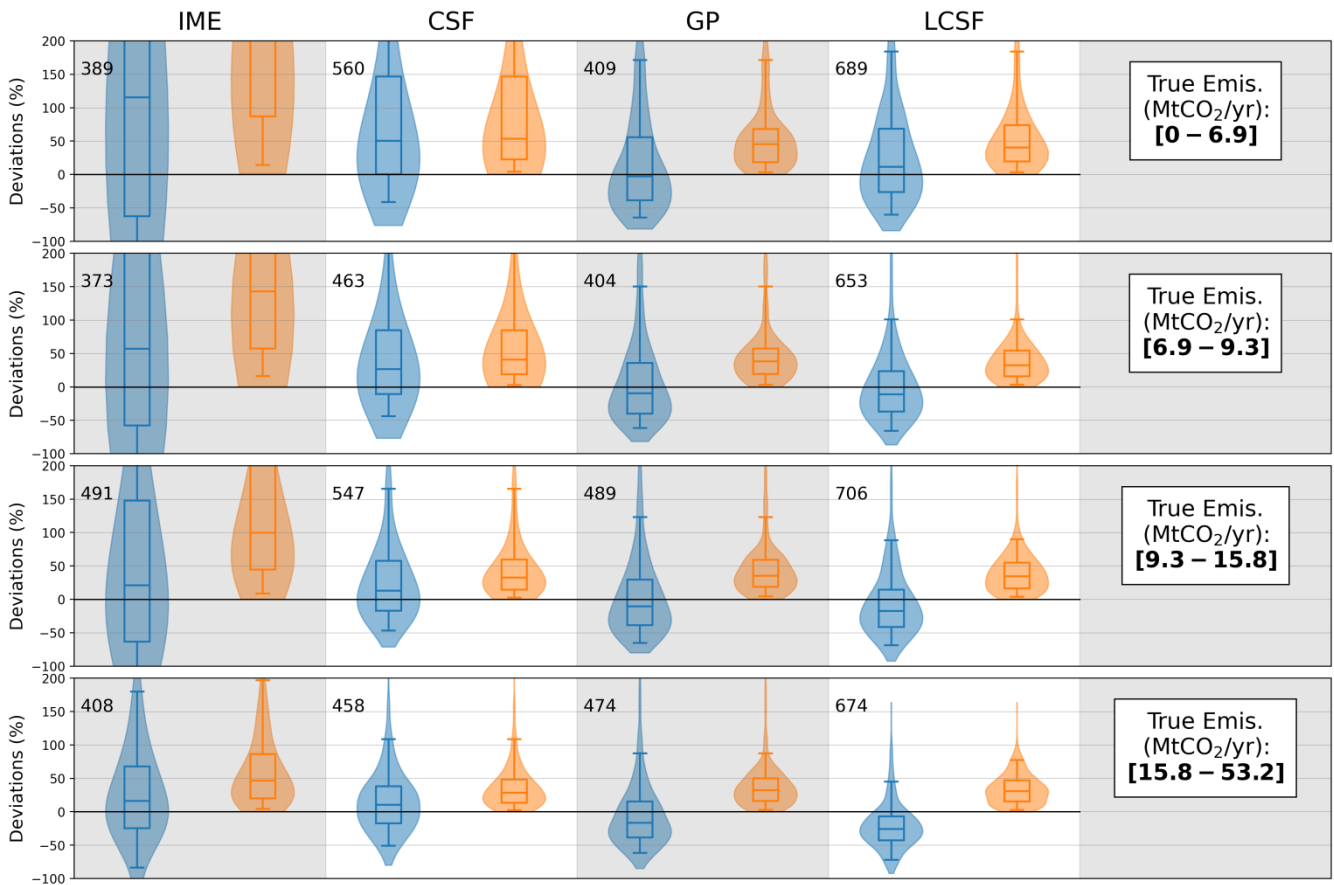
1214
 1215 **Figure 1. Illustration of different inversion methods for a plume produced by the Janschwalde power plant on April 23rd, 2015.**
 1216 **For all figures, pixels with dots are the selected enhancements representing the plume a) CSF method: the blue boxes depict the**
 1217 **areas where the Gaussian fits of the plume cross-sections are made and the black line the centre-line of the plume. b) LCSF**
 1218 **method: the blue lines represent the domain where the Gaussian fits of the plume cross-sections are made and the black line the**
 1219 **along-wind direction at the source. c) IME method: the blue curve represents the domain on which mass enhancements are**
 1220 **integrated. d) GP method: Blue curves depict contour lines of the 2-dimensional Gaussian curve that fits the plume.**

1221



1222
 1223
 1224
 1225
 1226

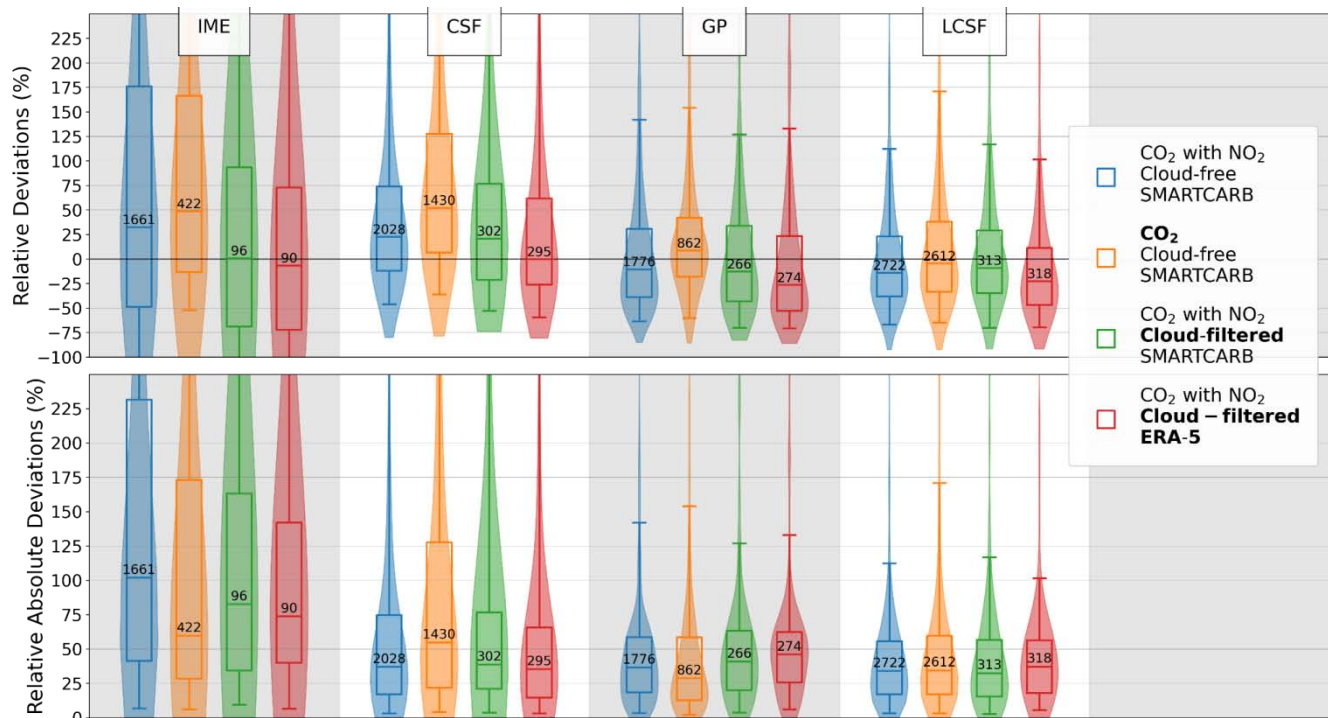
Figure 2. Simulations of XCO₂ on 23 April 2015 over the SMARTCARB domain. Synthetic XCO₂ observations over a 250 km wide swath are represented in the centre of the figure for a low noise scenario. Missing XCO₂ observations due to a cloud fraction larger than 1 % are shown in white. The 16 emission sources considered in this study are highlighted along with their names



1227
 1228 **Figure 3. Performance when estimating CO₂ emissions from individual images of the different single-image inversion methods**
 1229 **(columns) across different ranges of true emissions (rows) using SMARTCARB winds and cloud-free CO₂ and NO₂ data. The**
 1230 **distributions of relative deviations (in blue) and relative absolute deviations (in orange) are illustrated using violin plots. The inter-**
 1231 **quartiles are represented by the boxes, while the whiskers indicate the 5th and 95th percentiles, and medians are the lines inside the**
 1232 **boxes. The numbers alongside boxes show the numbers of estimates corresponding to true emissions ranges and inversion**
 1233 **methods.**

1234

1235



1236

1237

1238

1239

1240

1241

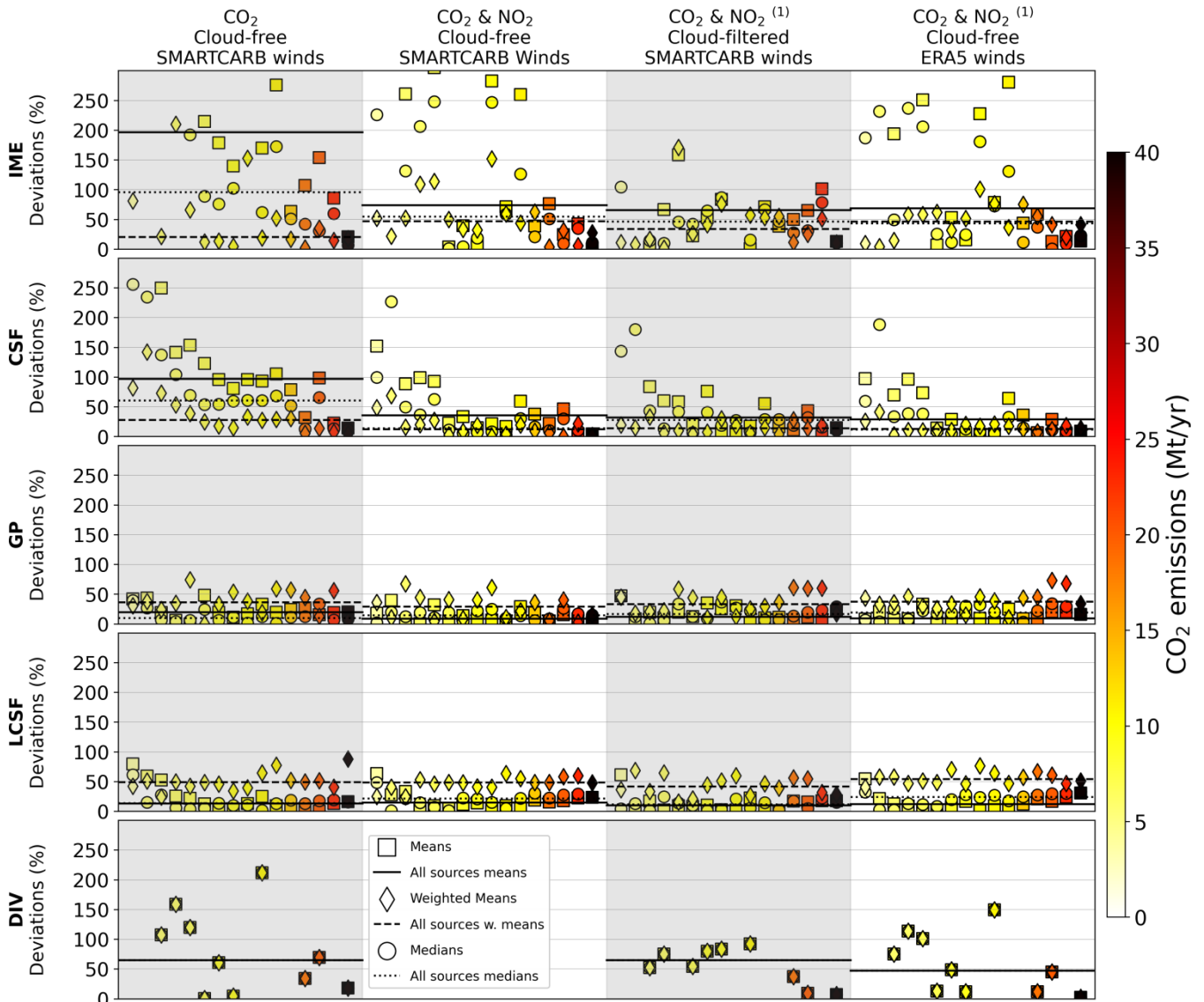
1242

1243

1244

1245

Figure 4. Performances of the inversion methods when estimating emissions from single images for different benchmarking scenarios: cloud-free CO₂ and NO₂ data with SMARTCARB winds (in blue), cloud-free CO₂ data only with SMARTCARB winds (in orange), cloud-filtered CO₂ and NO₂ data with SMARTCARB winds (in green), cloud-filtered CO₂ and NO₂ data with ERA5 winds (in red). **Bold texts in the legend indicate the elements of benchmarking scenarios that differ from those in the ideal benchmarking scenario.** Distributions of the relative deviations (top panel) and relative absolute deviations (bottom panel) are illustrated using violin plots. Boxes are the inter-quartiles of the distributions, the whiskers are the 5th and 95th percentiles, and the lines within boxes are the medians. Numbers in the inter-quartile boxes are the number of estimates for each benchmarking scenario and inversion method.



1246

1247

1248

1249

1250

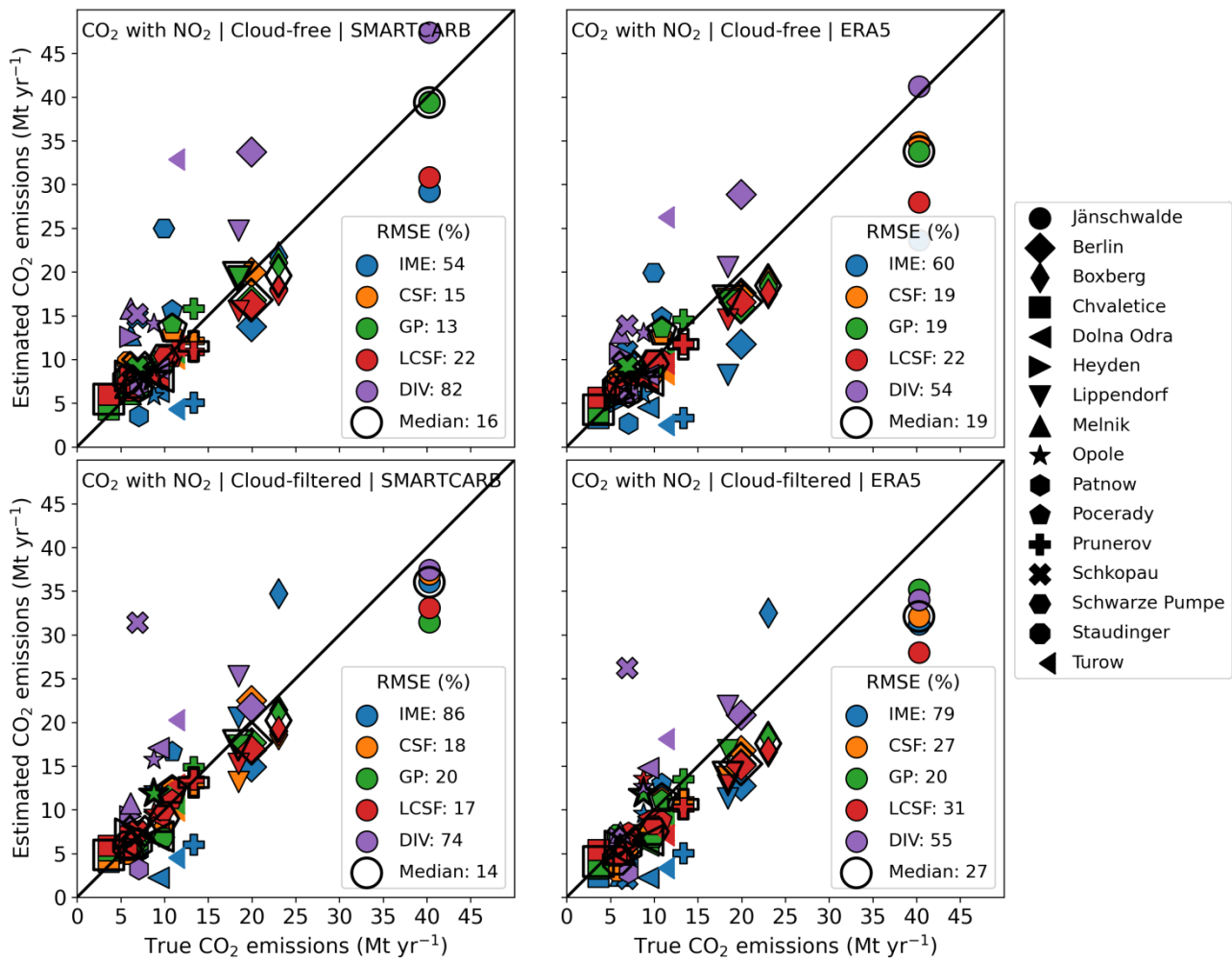
1251

1252

1253

1254

Figure 5. Performance of the inversion methods for annual estimates of CO₂ emissions. The markers represent for a given source the relative absolute deviations from the true annual emissions of the arithmetic means (squares), the weighted means (diamonds) and the medians (circles) of the estimates over a year. The lines represent the median values of the annual estimates over the entire set of sources. The inversions are performed using CO₂ cloud-free data and SMARTCARB winds (1st column), using CO₂ and NO₂ cloud-free data and with SMARTCARB winds (2nd column), using CO₂ and NO₂ cloud-filtered data and SMARTCARB winds (3rd column), and using CO₂ and NO₂ cloud-free data and with ERA5 winds (4th column). (1) For the Divergence methods, the inversions of the 3rd and 4th columns are performed using CO₂ data only. Markers color indicates the true CO₂ annual emissions of the corresponding source.



1255

1256

1257

1258

1259

1260

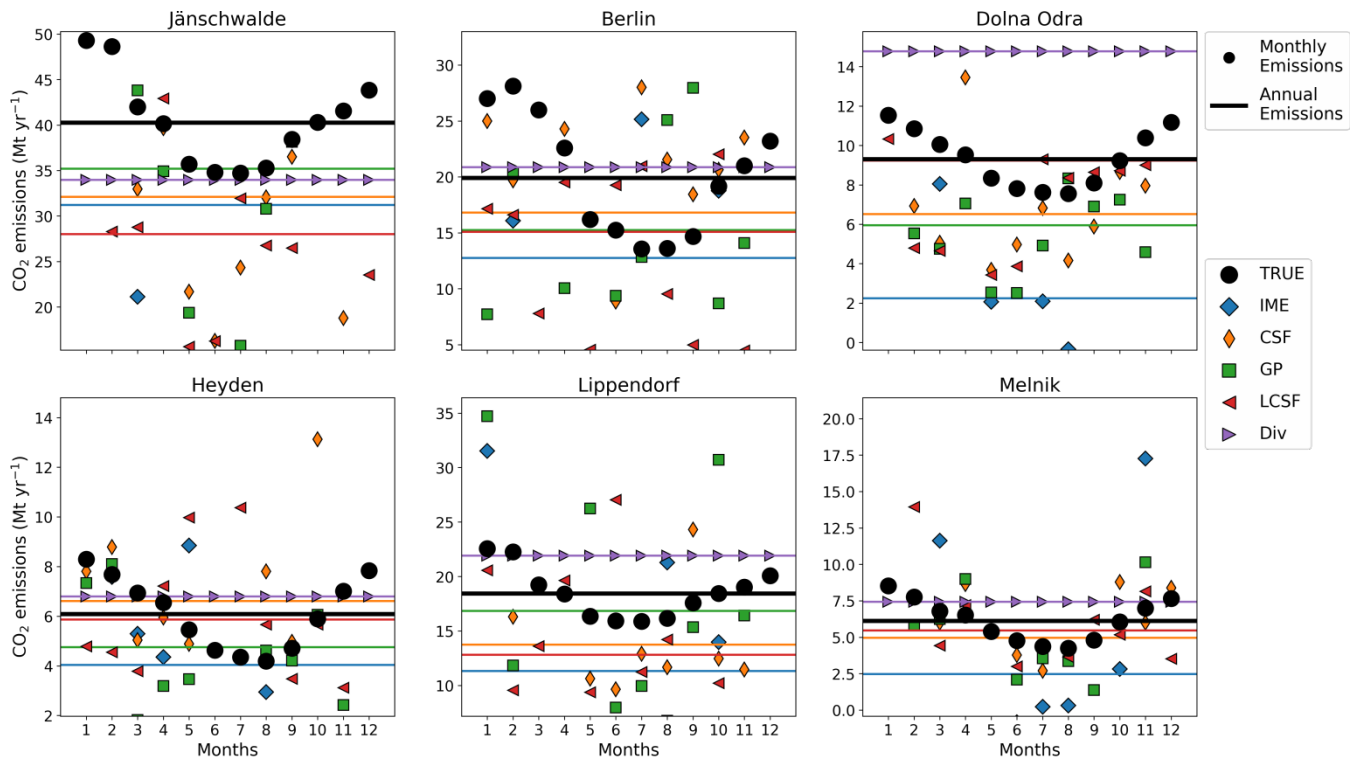
1261

1262

1263

1264

Figure 6. Estimated vs true annual emissions for 4 inversion scenarios (titles of the panels). For the IME and CSF methods, annual estimates are weighted means of the single-image estimates while they are arithmetic means for the GP, LCSF and Divs methods. Each marker represents a given emission source and each color a given inversion method. The unfilled markers represent the median values of all the estimates for each source. The divergence inversion method uses CO₂ data for all the inversion scenarios. The plain line represents the 1:1 line. The bottom-right legends display for each inversion method the relative RMSE which is the RMSE between estimated and true annual emissions divided by the median of true annual CO₂ emissions of all sources (~9.6 Mt yr⁻¹).



1265

1266

1267

1268

1269

1270

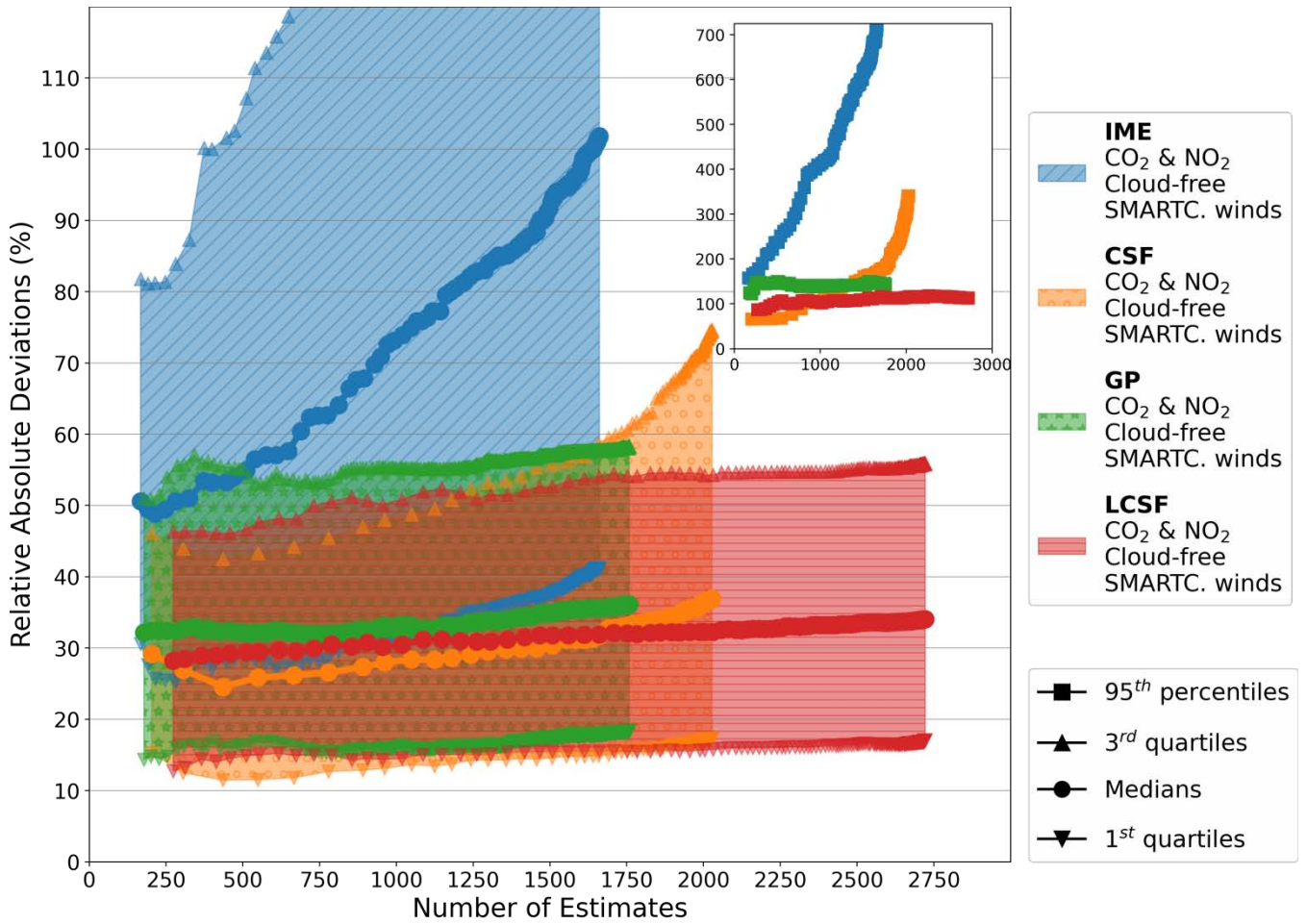
1271

1272

1273

1274

Figure 7. Annual and monthly estimates of the true and estimated emissions for different sources and for different inversion methods. Each panel is associated with a given source. Plain lines and markers represent annual averages and monthly averages respectively. Colors and markers are associated with different inversion methods (true emissions are represented by black circles). Annual and monthly estimates for the IME and CSF methods are weighted means of image estimates. Annual and monthly estimates for the GP and LCSF are means of image estimates while for the divergence method, we use the annual estimate also for monthly estimates. All inversion methods use CO₂ and NO₂ cloud-filtered data (CO₂ data only for the Div method) with ERA5 winds.



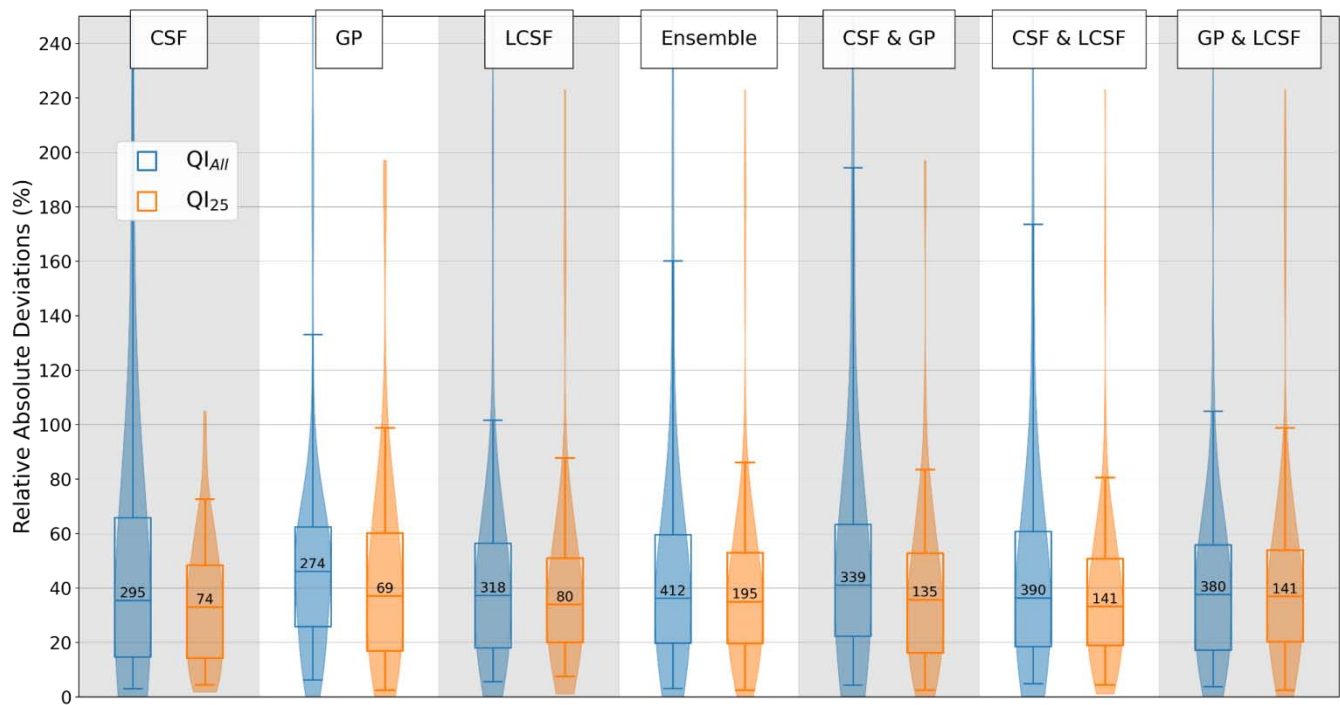
1275

1276

Figure 8. Accuracy of inversions vs number of single-image estimates. The inversion methods shown here use CO₂ and NO₂ cloud-free data and SMARTCARB winds. The filled areas represent the inter-quartiles of the distributions of the relative absolute deviations depending on the number of estimates. The 95th percentiles of the distributions are represented in the inset. Points belonging to a same curve are associated to different QIs and from left to right along curves, points are associated with a decreasing QI; the points at the left and right ends of the curves are associated with the maximal and minimal QIs respectively.

1281

1282



1283

1284

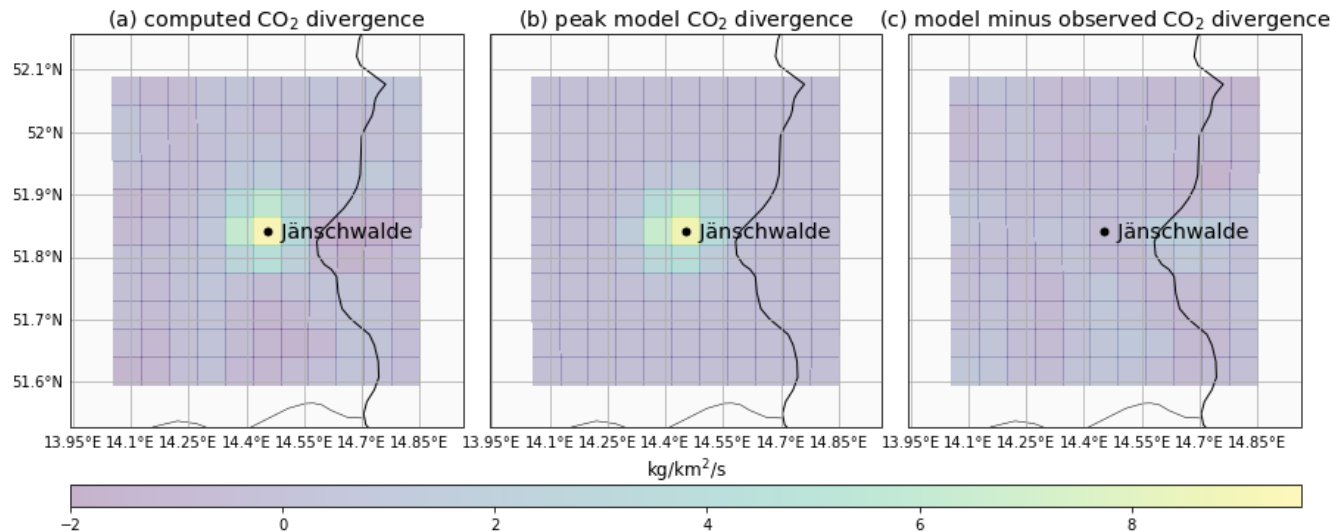
1285

1286

1287

1288

Figure 9: Performance of the inversion methods and ensemble approaches for estimating the emissions with cloud-filtered CO₂ & NO₂ data and with ERA5 winds. The distributions of the relative absolute deviations for all the inversion results (in blue) and for the best estimates (in orange) provided by each method (see text) are illustrated using violin plots. Boxes represent the inter-quartiles of the distributions, the whiskers the 5th and 95th percentiles, and the lines within boxes the medians. Numbers in the inter-quartile boxes are the number of estimates for each benchmarking scenario and inversion method.



1289

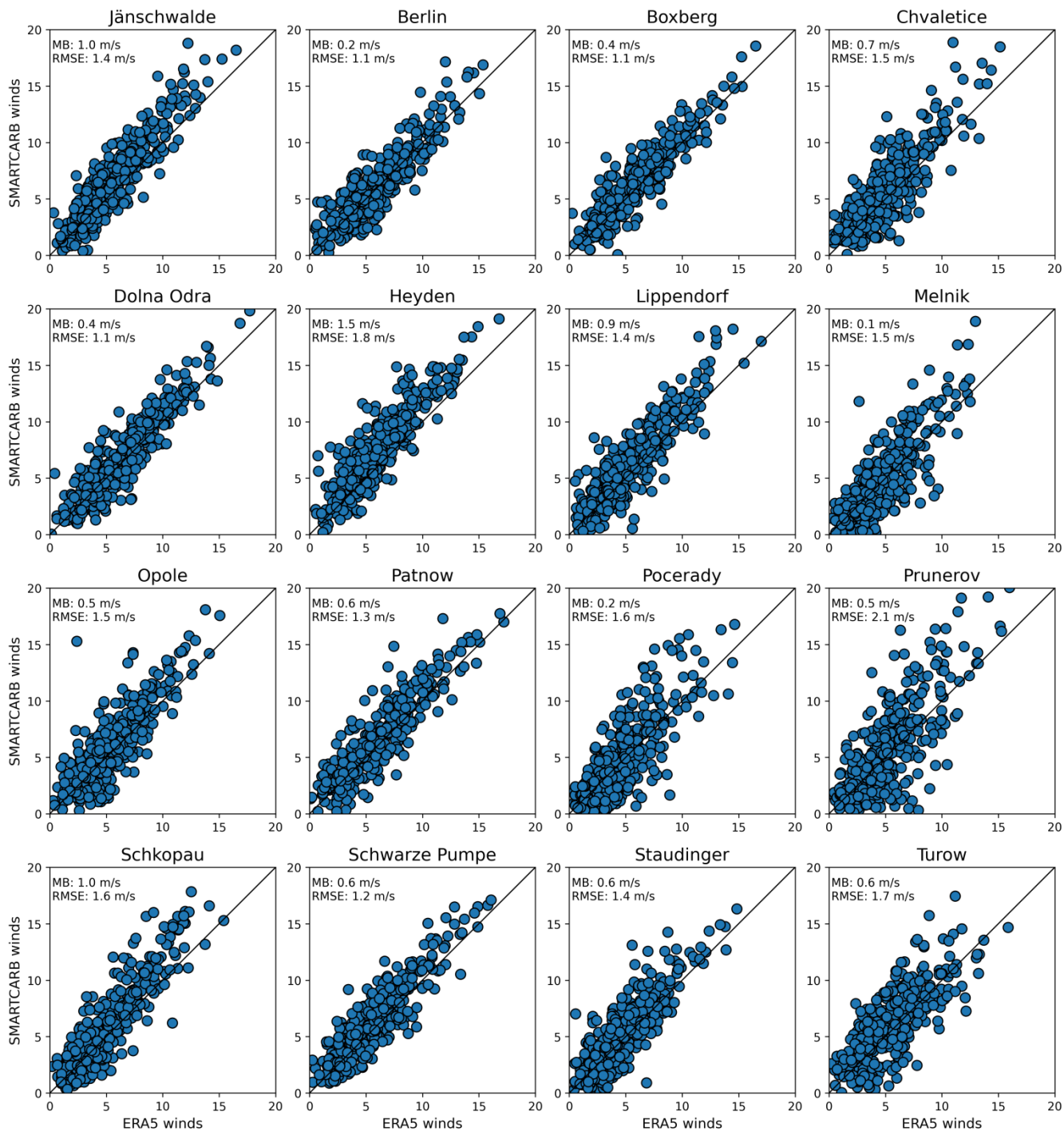
1290

1291

1292

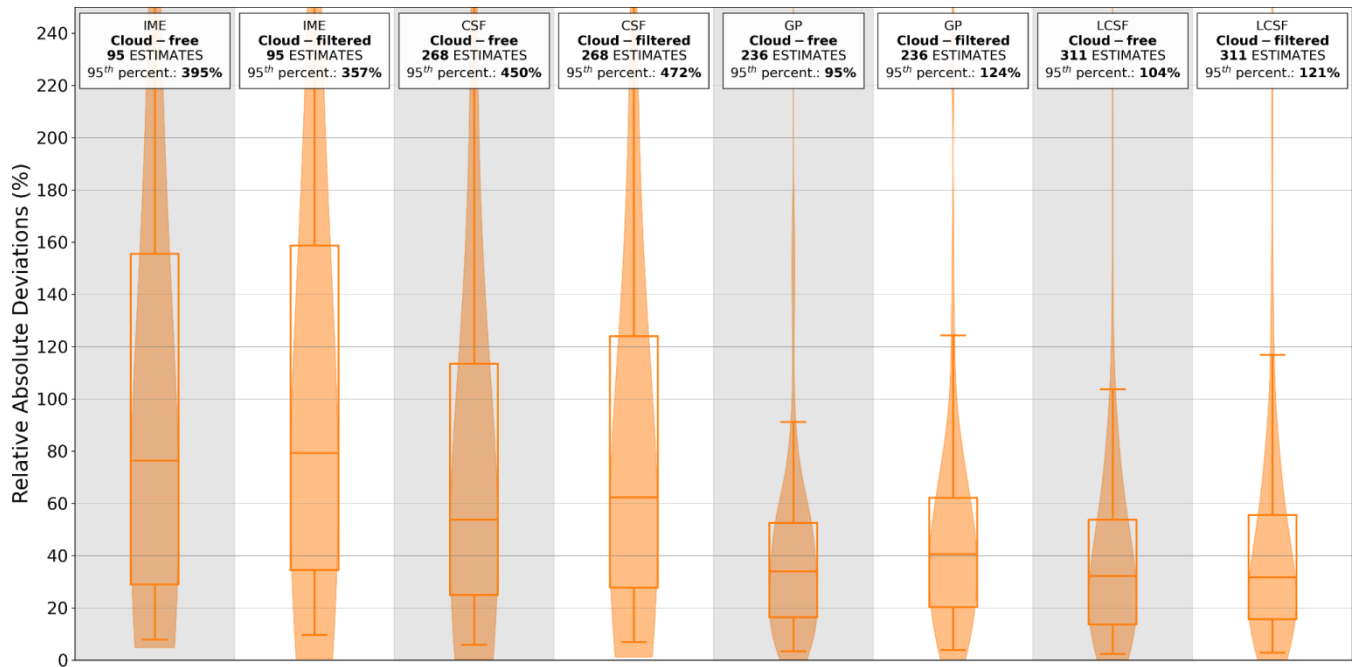
1293

Figure A1: Illustration of the divergence method for the Jämschwalde power station in 2015 based on the synthetic SMARTCARB dataset (see text). The figures represent the annual fields of the computed CO₂ divergence (a), the modelled CO₂ divergence (b) and the difference of both quantities (c). Sink terms are considered negligible for CO₂, divergence fields are considered equal to the emission fields for CO₂.



1294
 1295 **Figure A2: Norms of the ERA5 winds vs norms of the SMARTCARB winds at the sources considered in this study and for all the**
 1296 **days of 2015. Black lines represent the 1:1 agreement line. Mean biases of the SMARTCARB norms minus the ERA5 norms and**
 1297 **RMSEs are noted at the top left of the figures.**

1298



1299

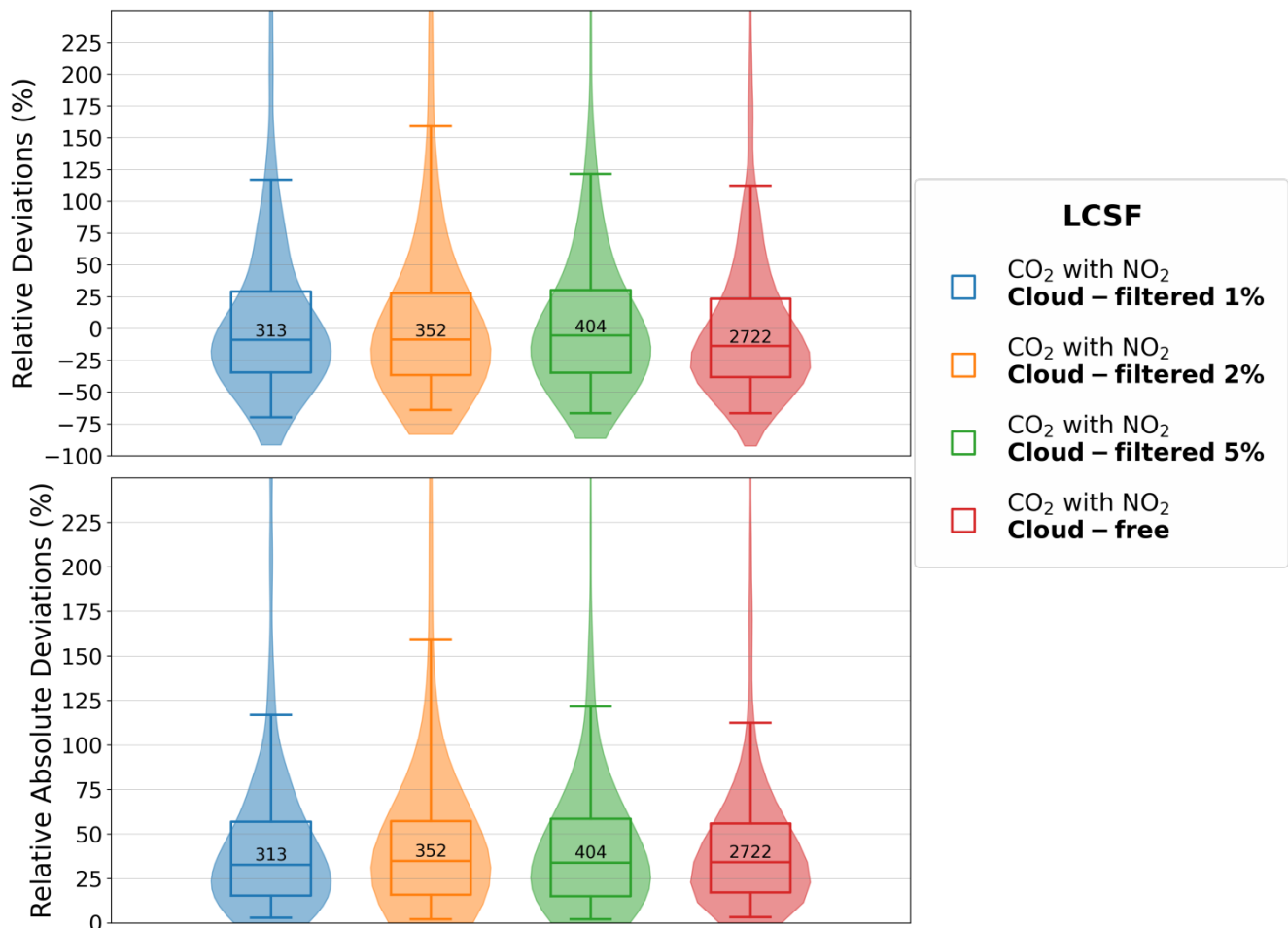
1300

1301

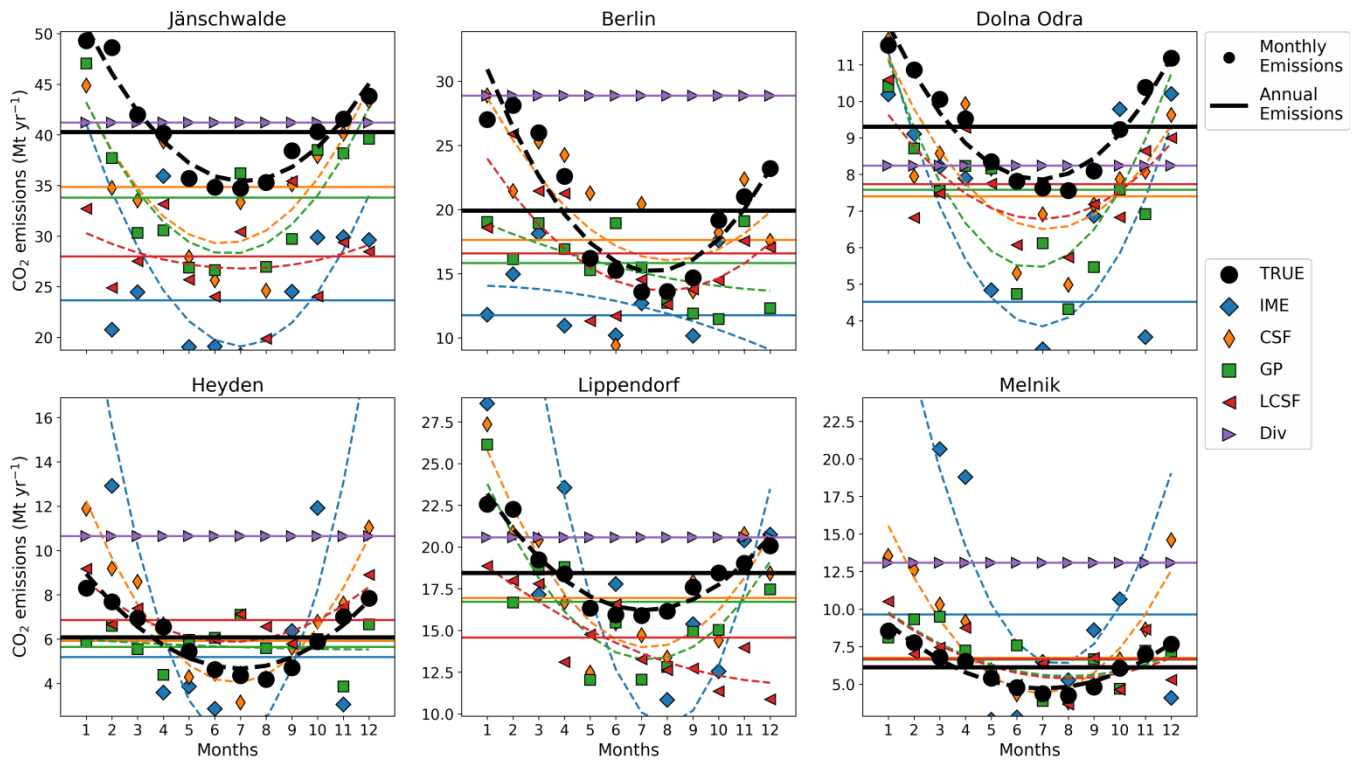
1302

1303

Figure A3: Performance of the inversion methods when using data with or without clouds for the emissions estimated from the same images. The inversion methods use CO₂ and NO₂ data and SMARTCARB winds. The boxes represent the inter-quartiles of the distributions of the absolute relative deviations, the whiskers the 5th and 95th percentiles, and the lines within boxes the medians.



1304
 1305 **Figure A4:** Performance of the LCSF method when estimating emissions from single images of CO₂ and NO₂ without considering
 1306 clouds (in red) and for different cloudiness thresholds: 1 % (in blue), 2 % (in orange) and 5 % (in green). Distributions of the
 1307 relative deviations (top panel) and relative absolute deviations (bottom panel) are illustrated using violin plots. Boxes are the inter-
 1308 quartiles of the distributions, the whiskers are the 5th and 95th percentiles, and the lines within boxes are the medians. Numbers in
 1309 the inter-quartile boxes are the number of estimates for each benchmarking scenario.
 1310



1311

1312

1313

1314

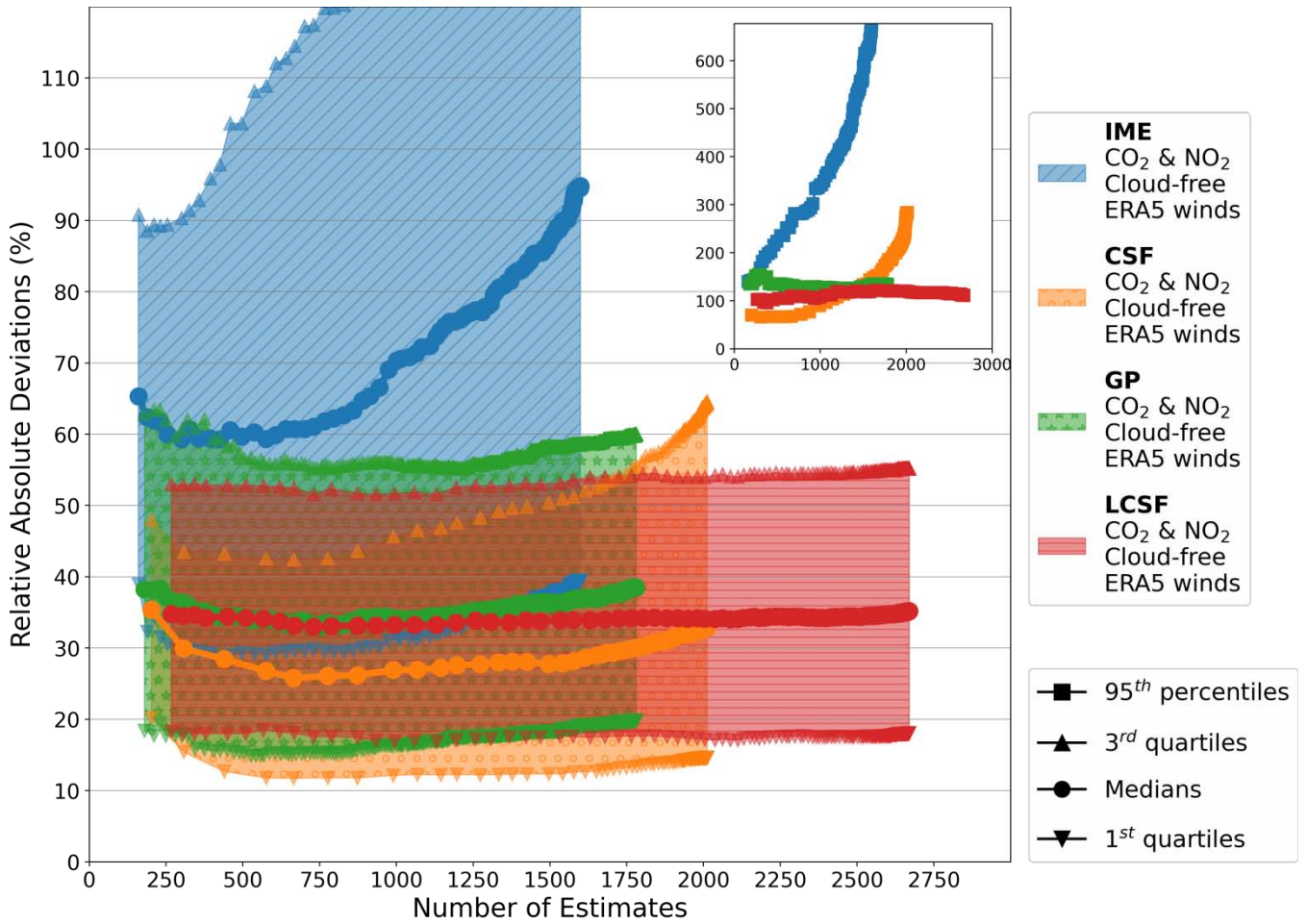
1315

1316

1317

1318

Figure A5: Annual and monthly estimates of the true and estimated emissions for different sources and for different inversion methods. Each panel is associated with a given source. Plain lines and markers represent annual averages and monthly averages respectively. Dashed lines represent the fits by a 2nd order polynomial of the monthly estimates. Colours are associated with different inversion methods (true emissions are in black). Annual and monthly estimates for the IME and CSF methods are weighted means of image estimates. Annual and monthly estimates for the GP and LCSF are means of image estimates while for the divergence method, we use the annual estimate also for monthly estimates. All inversion methods use CO₂ and NO₂ cloud-free data (CO₂ data only for the Divs methods) with ERA5 winds.



1319

1320

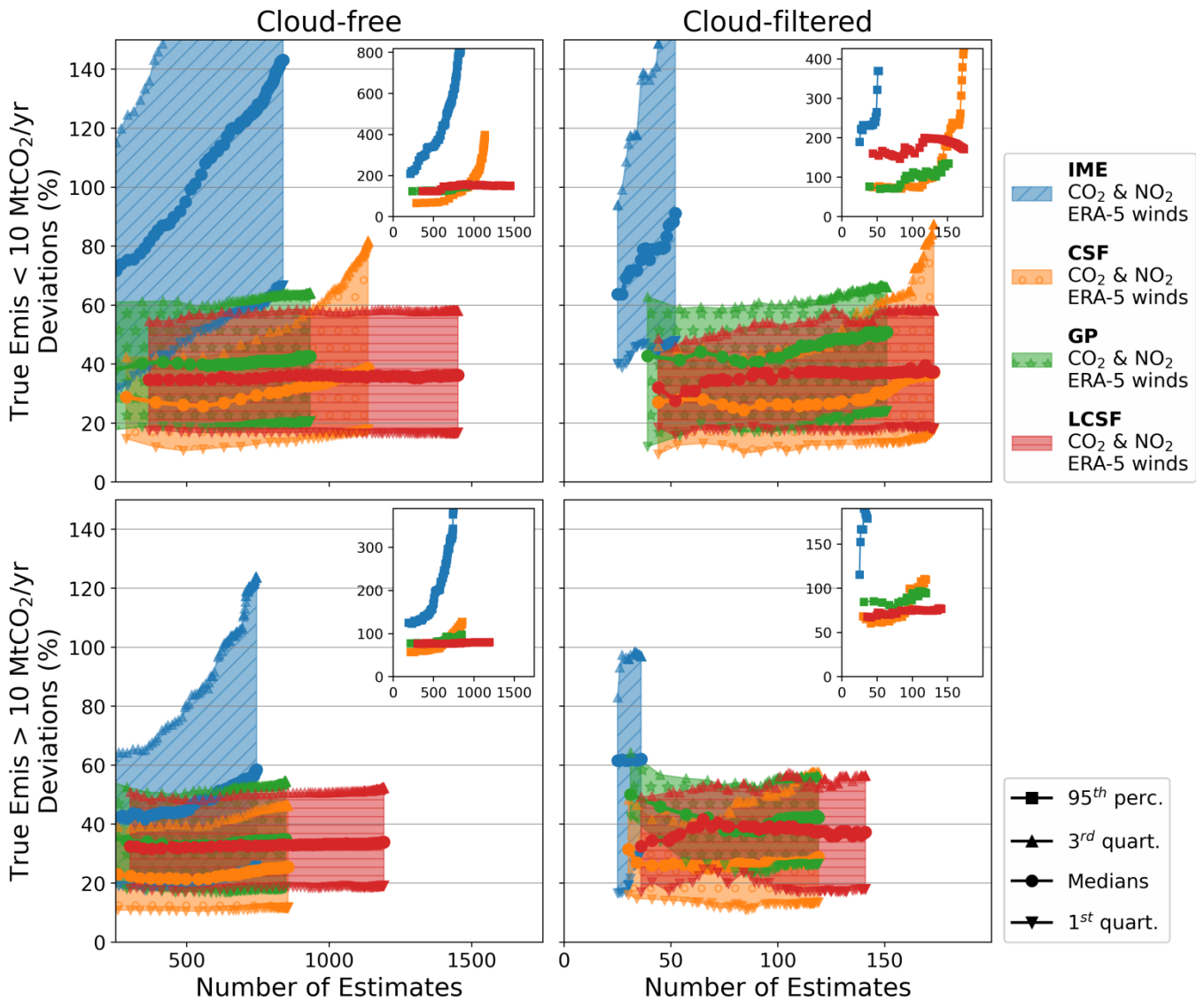
1321

1322

1323

1324

Figure A6. Accuracy of inversions vs number of instant estimates. The inversion methods shown here use CO₂ and NO₂ cloud-free data and ERA5 winds. The filled areas represent the inter-quartiles of the distributions of the relative absolute deviations depending on the number of estimates. The 95th percentiles of the distributions are represented in the inset. Points belonging to a same curve are associated to different QIs and from left to right along curves, points are associated with a decreasing QI; the points at the left and right ends of the curves are associated with the maximal and minimal QIs respectively.



1325

1326

1327

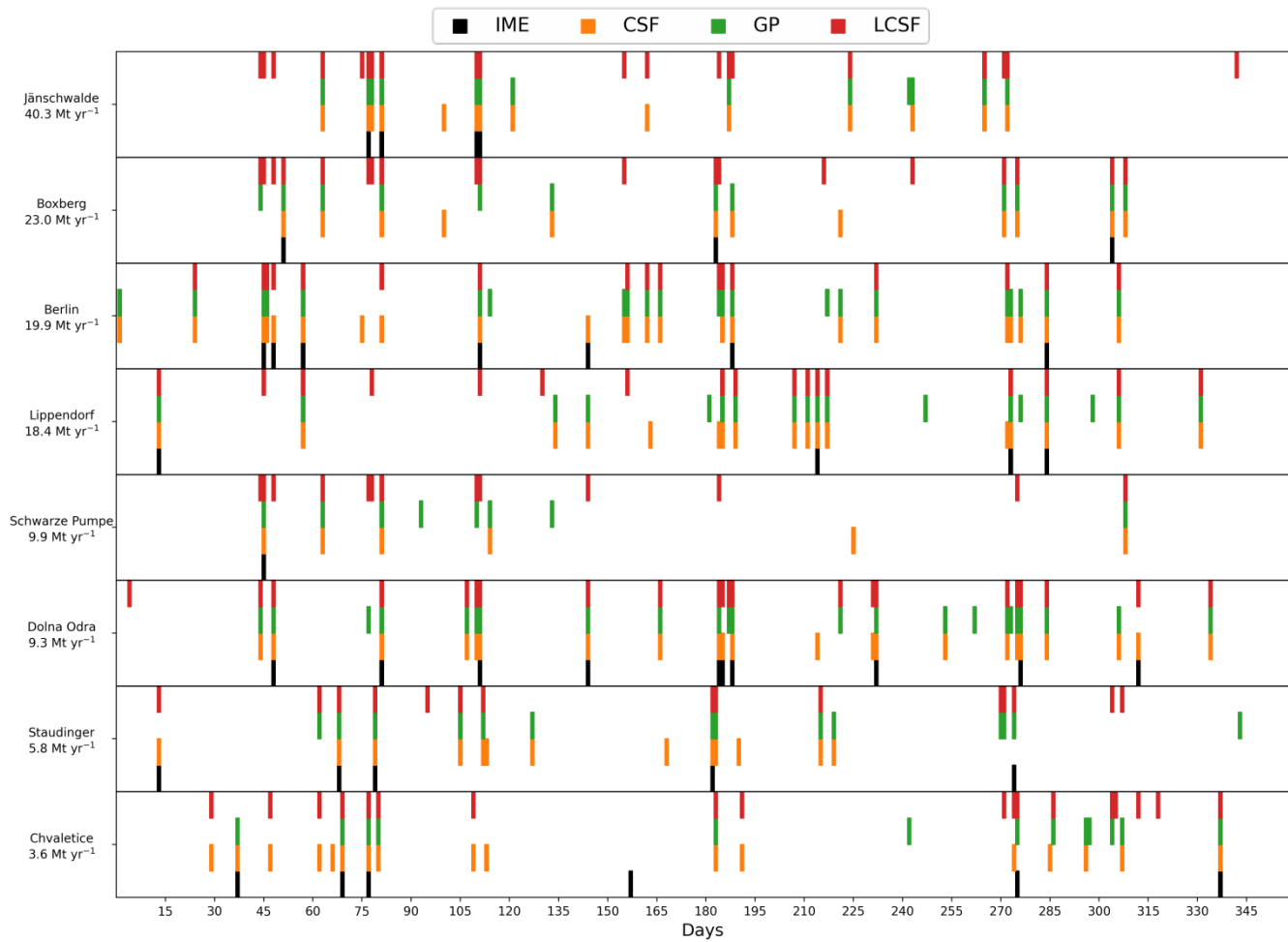
1328

1329

1330

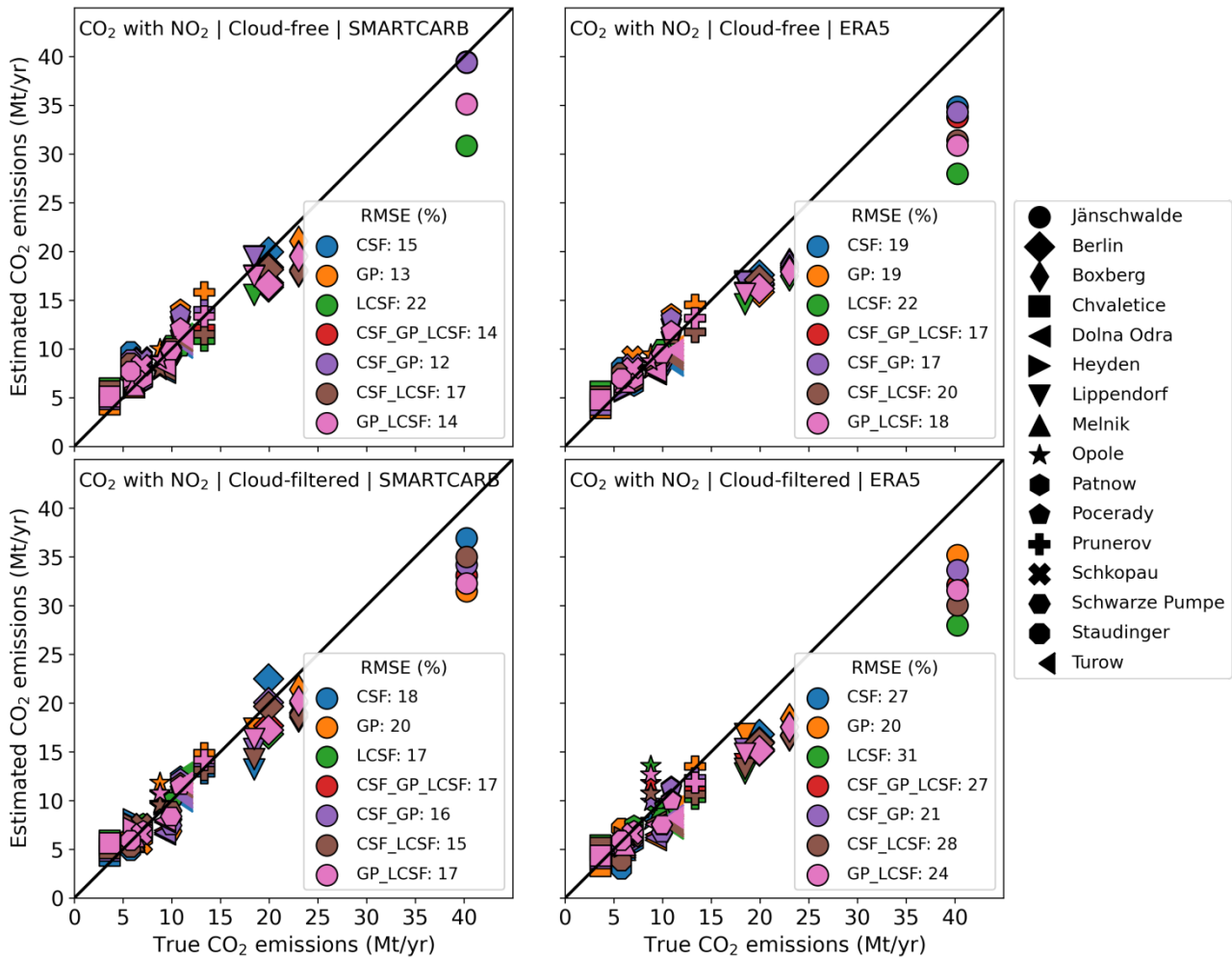
1331

Figure A7: Accuracy of inversions vs number of instant estimates. The inversion methods shown here use CO₂ and NO₂ data, ERA5 winds and for cloud-free (1st column) and cloud-filtered data (2nd column). Results are shown for the cases where true CO₂ emissions of sources are below (1st row) and above (2nd row) 10 Mt yr⁻¹. The filled areas represent the inter-quartiles of the distributions of the relative absolute deviations depending on the number of estimates. The 95th percentiles of the distributions are represented in the insets. Each point belonging to a same curve is associated with a different QI and from left to right along a same curve; points are associated with a decreasing QI.



1332
 1333
 1334
 1335
 1336
 1337

Figure A8: Days of 2015 (x-axis) for which the IME, CSF, GP and LCSF methods produce estimates for the CO₂ emissions of eight sources (y-axis). For a given day, the availability of an estimate from a given inversion method is illustrated by a color bar (for color explanation, see legend of the figure). Inversions use CO₂ and NO₂ cloud-filtered data and ERA5 winds.



1338

1339

1340

1341

1342

1343

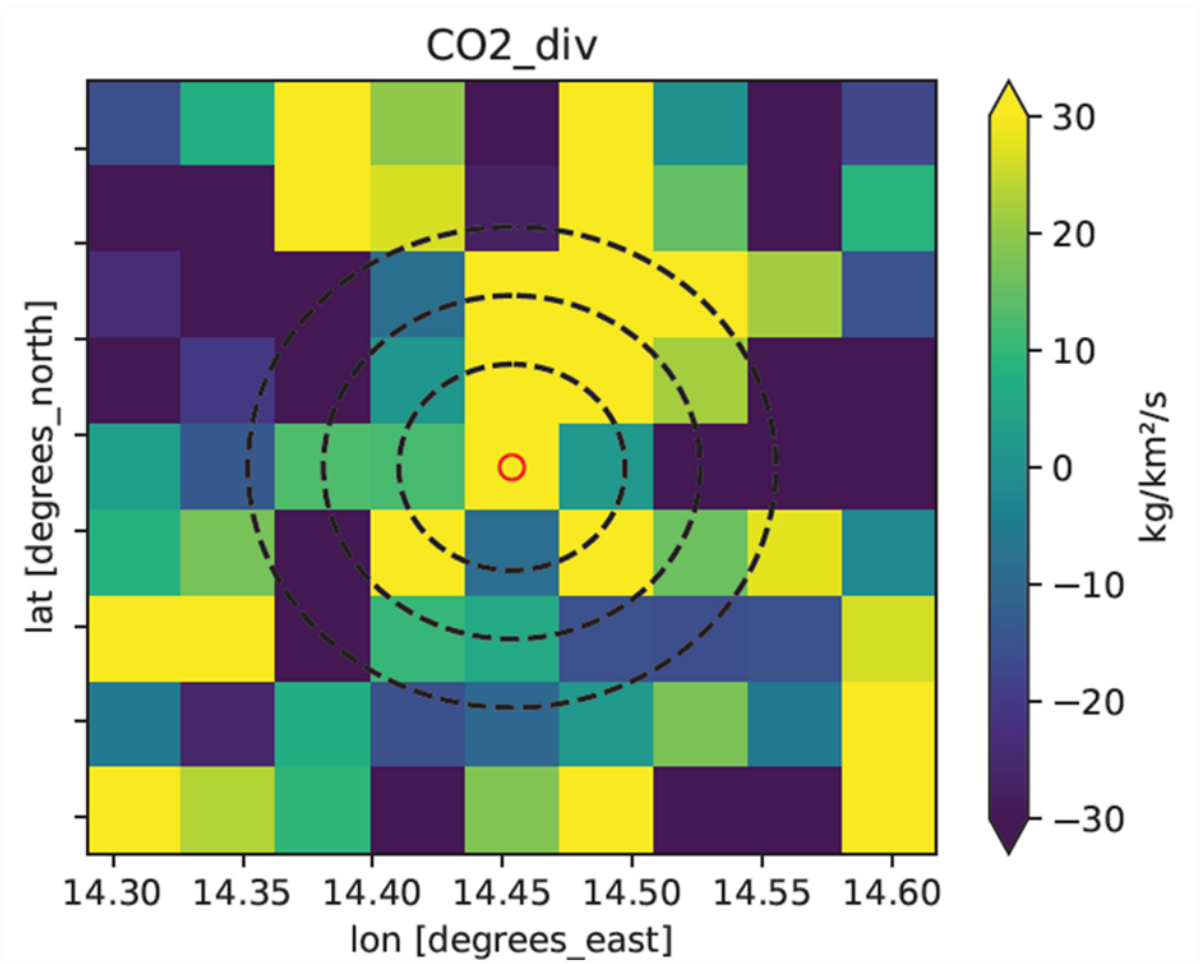
1344

1345

Figure A9: Estimated vs true annual emissions for 4 inversion scenarios (titles of the panels). Results are displayed for the CSF, GP, LCSF and ensemble methods that gather 2 or 3 of these individual methods. For the CSF method, annual estimates are weighted means of the instant estimates while they are arithmetic means for the GP and LCSF methods. Each marker represents a given emission source and each color a given inversion method. The divergence inversion method uses CO₂ data only for all the inversion scenarios. The plain line represents the 1:1 line. The bottom-right legends display for each inversion method the relative RMSE which is the RMSE between estimated and true annual emissions divided by the median of true annual emissions of all sources (~9.6 MtCO₂/yr).

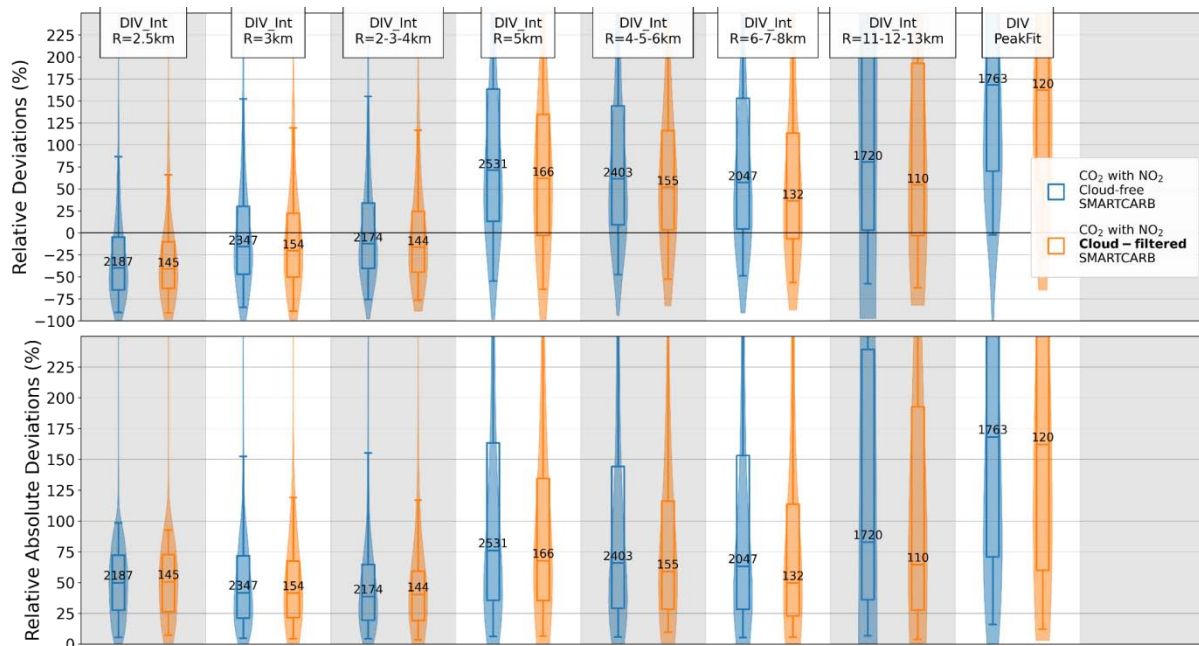
1346

1347



1348
 1349
 1350

Figure A10: Divergence map estimated around the Jämschwalde power station on January 2015 the 12th. Dotted circles show different radii (3 km, 5 km and 7 km) which define integration disks that could be used by the integral divergence method.

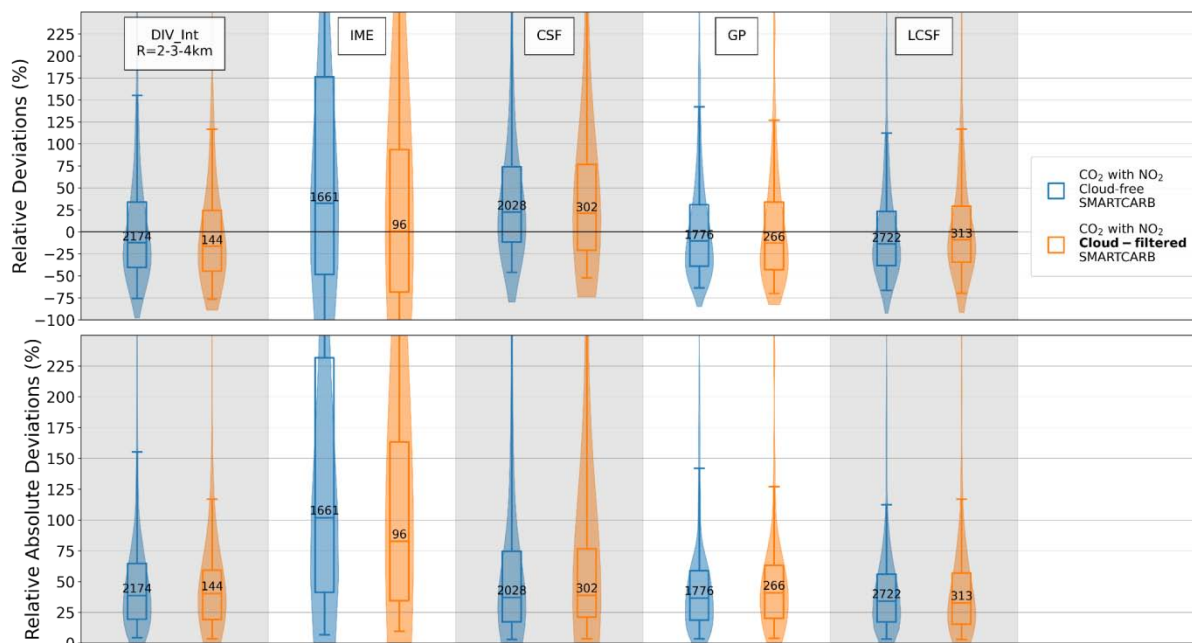


1351

1352

1353 **Figure A11: Performances of the different versions of the divergence inversion method when estimating emissions from one year**
 1354 **of single images for different benchmarking scenarios: cloud-free CO₂ and NO₂ data with SMARTCARB winds (in blue) and**
 1355 **cloud-filtered CO₂ and NO₂ data with SMARTCARB winds (in orange). Distributions of the relative deviations (top panel) and**
 1356 **relative absolute deviations (bottom panel) are illustrated using violin plots. Boxes are the inter-quartiles of the distributions, the**
 1357 **whiskers are the 5th and 95th percentiles, and the lines within boxes are the medians. Numbers in the inter-quartile boxes are the**
 1358 **number of estimates for each benchmarking scenario and inversion method. Methods DIV_int_R=xkm and DIV_PeakFit are the**
 1359 **integral (for an integration radius of x km) and peak-fitting versions of the divergence approach respectively. For a given overpass**
 1360 **and source, the emission estimate of the method DIV_int_R=x-y-zkm is the average of the estimates when integrating over circles**
 1361 **of x, y and z km radius around the source.**

1361



1362

1363

1364

1365

1366

1367

1368

1369

1370

Figure A12: Performances of the inversion methods when estimating emissions from one year of single images for different benchmarking scenarios: cloud-free CO₂ and NO₂ data with SMARTCARB winds (in blue) and cloud-filtered CO₂ and NO₂ data with SMARTCARB winds (in orange). Distributions of the relative deviations (top panel) and relative absolute deviations (bottom panel) are illustrated using violin plots. Boxes are the inter-quartiles of the distributions, the whiskers are the 5th and 95th percentiles, and the lines within boxes are the medians. Numbers in the inter-quartile boxes are the number of estimates for each benchmarking scenario and inversion method. Methods DIV_int_R=2-3-4km and DIV_PeakFit are the integral and peak-fitting versions of the divergence approach respectively. For a given overpass and source, the emission estimate of the method DIV_int_R=2-3-4km is the average of the estimates when integrating over circles of 2,3 and 4 km radius around the source.

1371

1372

Method	Time frame	Computational cost (1)
Integrated Mass Enhancement (IME)	Single-Image estimates	Medium: ~20 min
Cross-Sectional Flux (CSF)	Single-Image estimates	Medium: ~25 min

Gaussian Plume (GP)	Single-Image estimates	High: ~110 min
Light Cross-Sectional Flux (LCSF)	Single-Image estimates	Low: ~10 min
Divergence (Div)	Averaged estimates from ensemble of images	Medium: ~23 min

1373 **Table 1: Summary of characteristics of the benchmarked methods. (1) Computation time was estimated by inverting one month of**
1374 **CO₂ and NO₂ cloud-free SMARTCARB data on the same server using the ddeq package (Kuhlmann et al., 2023)**

1375

Benchmark Scenario	Wind dataset	Cloud fraction thresholds	Joint use of NO ₂ and CO ₂
Scenario 1	SMARTCARB	100 % (no clouds)	Yes
Scenario 2	SMARTCARB	1 % for CO ₂ , 30 % for NO ₂	No
Scenario 3	SMARTCARB	100 % (no clouds)	No
Scenario 4	SMARTCARB	1 % for CO ₂ , 30 % for NO ₂	Yes
Scenario 5	ERA5	100 % (no clouds)	Yes
Scenario 6	ERA5	1 % for CO ₂ , 30 % for NO ₂	No
Scenario 7	ERA5	100 % (no clouds)	No
Scenario 8	ERA5	1 % for CO ₂ , 30 % for NO ₂	Yes

1376 **Table 2: List of the different benchmarking scenarios: from the most optimistic (scenario 1) which considers inversions with cloud-**
1377 **free data and SMARTCARB winds to the most realistic (Scenario 8) with cloud-filtered data and with ERA5 winds. Note that a**
1378 **cloud fraction threshold of x % corresponds to the rejection of data pixels if their cloud cover exceeds x %, so that a cloud fraction**
1379 **of 100 % yields full images without a loss of data pixels.**

1380

1381

1382

Inversion method	Cloud-free data	Cloud-filtered data
IME	1661	96
CSF	2028	302
GP	1776	266
LCSF	2722	313

1383 **Table 3. Number of estimates for each inversion method when data with or without clouds are used. Inversions are**
1384 **performed with CO₂ and NO₂ data and, with SMARTCARB winds.**

1385