CoCO2

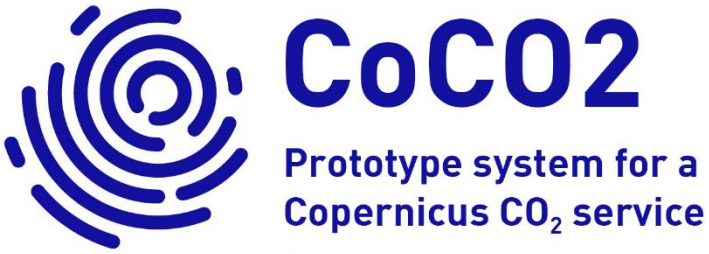
Prototype system for a Copernicus CO_2 service

D2.6 PED uncertainty 2018

Ingrid Super, Arjan Droste (TNO)

Margarita Choulga (ECMWF)

Co-ordinated by ECMWF



D2.6 PED uncertainty 2018 and uncertainties based on Monte Carlo simulation using the emission model from D2.5 v1

Dissemination Level:	Public/ Confidential
Author(s):	Ingrid Super (TNO), Arjan Droste (TNO), Margarita Choulga (ECMWF)
Date:	03/08/2022
Version:	1.1
Contractual Delivery Date:	30/06/2022
Work Package/ Task:	WP2/ T2.5
Document Owner:	TNO
Contributors:	TNO, ECMWF
Status:	Reviewed

CoCO2: Prototype system for a Copernicus CO_2 service

Coordination and Support Action (CSA)
H2020-IBA-SPACE-CHE2-2019 Copernicus evolution – Research activities in support of a European operational monitoring support capacity for fossil CO2 emissions

Project Coordinator: Dr Richard Engelen (ECMWF)
Project Start Date: 01/01/2021
Project Duration: 36 months

Published by the CoCO2 Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX,
richard.engelen@ecmwf.int

The CoCO2 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958927.

Table of Contents

Figures

Tables

Executive Summary and introduction

This deliverable report describes the development of prior uncertainty dataset related to the European and global prior emission datasets (PED) for the year 2018. We provide an overview of the developed methodology to estimate prior uncertainties in the aggregated PED starting from a detailed set of uncertainties. This also includes the estimation of gridded uncertainties and spatial error correlation lengths. For the global emissions only CO_2 is considered, whereas for the European emissions CO and NO_x are also included. Since the data underlying the global and European datasets is different, a slightly different approach is used for the datasets, but where possible we kept our methods consistent.

Two methods are developed that are discussed and compared in this report. The first method is an extensive Monte Carlo simulation, which can take into account non-Gaussian error distributions and error correlations. However, since this method is computationally expensive, especially for a high-resolution dataset, we also used a relatively simple error propagation approach. We find that this simpler method gives comparable results, especially considering the uncertainties in the prior uncertainty data.

At the end of the document a guideline is provided for use of the data. The data are available through an FTP site (coco2@ftp.ecmwf.int), following the directory structure data-exchange/WP2/D2-6-prior_emission_uncertainties. Please contact the CoCO2-Coord to obtain the FTP password access.

Background

Prior emission data is important input data for the modelling efforts done in CoCO2. For inverse modelling, and also to understand discrepancies between models and observations, the uncertainties in the prior emissions also need to be quantified. The challenge is to get a consistent set of uncertainties for the emission datasets, making use of uncertainty estimates at a very detailed level and considering different error distributions and correlations.

First efforts have been made in the CHE project to quantify these uncertainties, which form the basis for the work in T2.5 (Choulga et al. (2021a), Super et al. (2020)). Whereas in CHE work was done mostly on country-level uncertainties for CO_2, we want to extend this work by including other sources of uncertainties and co-emitted species. Main focus points are:

- Create a consistent set of country-level uncertainties, both globally and for Europe.
- Include CO_2 and co-emitted species CO and NO_x.
- Assess error correlations between sectors and species.
- Assess uncertainties in spatial/temporal proxies, including spatial/temporal correlation lengths.

Scope of this deliverable

Objectives of this deliverable

The aim is to provide a first set of uncertainties for the PED products from WP2, including spatial uncertainties and error correlations. Other WPs can use these data in inverse modelling efforts and provide feedback on the current product. This feedback will be taken into account for the next deliverable.

Work performed in this deliverable

There are two deliverables for T2.5 (D2.6 and D2.7), of which this report (and accompanying dataset and user documentation) is the first one. We made a list of priorities and started from there. This report describes the work that has been done so far. The current dataset will be extended and updated next year with those aspects that are not yet considered in this deliverable.

Deviations and counter measures

Originally, the focus year was 2016, but throughout the project this has been updated to 2018. Since the PED is for 2018 also the uncertainties are provided for the 2018 data. The PED data for 2021 is not yet finished and is therefore not included in this deliverable.

The uncertainties related to the emission model from D2.5 is not yet included, because the emission model is not yet finished. These uncertainties will be part of the next deliverable (D2.7). This has no consequences for the project, as the users of this product are also responsible for its preparation and the usage is not planned until later.

Originally, ECMWF had the lead for this deliverable. However, for practical reasons this was changed to TNO. Nevertheless, the work has been done by both parties, as planned.

1 Input data and pre-processing

1.1 Prior emissions

1.1.1 Global emissions

The global prior emissions are based on EDGARv6.0 for the year 2018 (JRC (2022), Crippa et al. (2021a), Crippa et al., (2021b)) with a spatial resolution of 0.1x0.1 degrees. Emissions are calculated as the product of activity data (AD) and emission factors (EF). For EDGAR the AD primarily come from international statistics (International Energy Agency). For most countries and sectors default EFs are used from the IPCC Guidelines for National Greenhouse Gas Inventories (IPCC, 2006). A total of 21 sectors is differentiated, which are grouped into 6 aggregated sectors for this deliverable (Table 4, for link with IPCC sectors see Table 4).

Table 1: Overview of emission categories in the emission data (EDGAR sector) and the uncertainty data (grouped).

Grouped category name	EDGAR sector	Note
PublicPower	ENE	Power industry
	SWD_INC	Solid waste incineration
Industry	CHE	Chemical processes
	IND	Combustion for manufacturing
	IRO	Iron and steel production
	NEU	Non energy use of fuels
	NFE	Non-ferrous metals production
	NMM	Non-metallic minerals production
OtherStatComb	RCO	Energy for buildings
RoadTransport	TNR_Other	Railways, pipelines, off-road transport
	TRO	Road transportation no resuspension
Shipping	TNR_Ship	Shipping
Other	AGS	Agricultural soils
	PRO	Fuel exploitation
	PRU_SOL	Solvents and products use
	REF_TRF	Oil refineries and Transformation industry
	TNR_avi_CDS	Aviation climbing, descent
	TNR_avi_CRS	Aviation cruise
	TNR_avi_LTO	Aviation landing, take-off

In addition to the EDGAR emissions database we use the Large Scale International Boundary (LSIB) dataset (U.S. Office of the Geographer (2017)). This dataset contains country borders of 279 countries, which are separated into areas with well- and less well-developed statistical infrastructures. The LSIB shapefiles were used to create country fraction maps at the resolution of the EDGAR v6.0 emission data. First, country shapefiles were rasterised globally

to a regular grid at 100m resolution and then averaged per country to the EDGAR resolution. The emissions over oceans (i.e. all emissions left after removing all countries) were inspected and 56 grid cells with emissions from the energy sector (ENE), manufacturing industry (IND), fuel exploitation (PRO) or oil refineries/transformation industry (REF_TRF) were assigned to a country. Finally, Japan, Maldives and Philippines were additionally buffered with one extra grid cell over the perimeter of these countries. This results in a global country fraction map.

The emissions from the energy sector (ENE) were separated into ENE_sup and ENE_ave, fluxes from super emitting power plants and fluxes from average emitting power plants (grid cell threshold of 7.9 x10^{-6} kg substance/m^2/s). Power plants are separated into CO_2 emitting (fossil fuel use) and non-emitting (renewable energy use) ones using the Global Power Plant Database (Global Energy Observatory et al. (2018)). This database contains information on geographical locations, owning country and fuels used for around 28,500 power plants. The CO_2 emitting power plants are rasterised globally to a regular grid at 100m resolution, where each grid cell contains a country ID. Finally, we checked whether each power plant was matched with an energy sector grid cell at the native resolution from one country only. If so, the emissions from that power plant are added to the country budget. Otherwise, we assume that the emissions are shared between countries according to the country fraction map.

Then all flux maps were translated into emission mass maps in kiloton (kt) taking into account the grid cell areas. Emission budgets per country per sector were computed by summing grid cell emission values. Then emission budgets per country were refined by taking into account additional yearly emission information from EDGAR v6.0 "Timeseries" EXCEL file (Appendix A, Table 4) and a list of CO_2-emitting activities which must not be neglected according to the IPCC 2006 guidelines, resulting in 51 sub-sectors. Following the same proportion of country's yearly budget yearly emission grid-cell values were disaggregated into 51 sub-sector maps.

1.1.2 European emissions

The European prior emissions used as a basis for this work is the TNO GHGco v4 (delivered end of 2021 by T2.1) for 2018. The dataset is described in D2.1 and by Kuenen et al. (2022). The dataset is compiled from emissions reports delivered to EMEP/CEIP and the UNFCCC by individual countries. The reports are very detailed and contain a long list of sectors (NFR categorisation) and fuel splits. Because the level of detail in reporting is a bit different for greenhouse gases and air pollutants, the emissions are (dis)aggregated to a total of 246 sector-fuel combinations. In the final dataset these emissions are aggregated into 12 sectors (GNFR categorisation, see Table 2), with an extra split for road transport, but for this task we work with the detailed sector split. Not all countries report their emissions and in that case other emission datasets are used for gap filling. For these countries emissions are only available at the GNFR level. Note that the uncertainty estimates are different for those countries, as we lack detailed information.

The detailed country-level emissions are spatially downscaled using proxy maps, as explained by Kuenen et al. (2022). The proxy maps describe the fraction of the country-level emissions for a particular sector that is assigned to one grid cell. Therefore, the fractions sum up to 1 for each country-sector combination. There are many different proxy maps and some of them are shared between multiple sectors. All pollutants are spatially disaggregated with the same proxy maps for a particular sector. Also for the spatial downscaling some exceptions apply. For some countries the spatial proxies are not available or replaced with other datasets.

To avoid a lot of work on sectors which are not that relevant we first made a list of sector-fuel combinations which make up more than 95% of the emissions for CO_2, CO and NO_x for the whole domain. When we combine those lists for all pollutants we end up with 90 sector-fuel combinations that describe 96% of CO_2 emissions, 98% of CO emissions, and 97% of NO_x emissions. For these sector-fuel combinations we look for uncertainty data and the remainder gets an uncertainty of zero.

Table 2: Overview of aggregated emission categories in the European emission data (GNFR) and the uncertainty data (grouped). Also relative uncertainties (95% confidence interval) in AD and EF per GNFR sector are given.

GNFR category	GNFR category name	Grouped category name	AD	EF
A	A_PublicPower	PublicPower	0.02	0.049
B	B_Industry	Industry	0.03	0.049
C	C_OtherStationaryComb	OtherStatComb	0.15	0.049
D	D_Fugitives	Other	0.05	0.75
E	E_Solvents	Other		
F	F_RoadTransport	RoadTransport	0.05	0.05
G	G_Shipping	Shipping	0.05	0.015
H	H_Aviation	Other	0.5	0.05
I	I_OffRoad	Other	0.5	0.02
J	J_Waste	Other	0.13535	0.0707
K	K_AgriLivestock	Other		
L	L_AgriOther	Other	0.2	0.2

It was decided to group together some sectors which are of minor importance for CO_2 (<10%) in the final product, so we describe uncertainties for six sectors (see Table 2). For two of these sectors we include spatial errors and correlation lengths, namely other stationary combustion (mostly residential/commercial heating) and road transport. The proxy maps used for these sectors are given in Table 3. We made this decision based on knowledge on the underlying proxies and the possibility to estimate errors for them with relative ease for purpose of developing the methodology. For all other sectors we assume no spatial errors for now and this also applies for countries for which emissions are not downscaled using the selected proxies. Although the purpose of this deliverable is to start developing the methodology and not providing a complete product at once, users of this preliminary product should be aware that spatial errors are missing for some sectors. This will mainly affect the correlation between CO_2 and co-emitted species, which is strongly related to the spatial patterns. For the next deliverable we aim to include the other sectors as well.

1.2 Country-level uncertainties

1.2.1 Global uncertainties

Uncertainties for country budgets are based on the IPCC 2006 guidelines (IPCC, 2006) and its 2019 refinements (IPCC, 2019). From these guidelines a table with major CO_2 emitting activities (92 in total) and their upper and lower uncertainty bounds (i.e. 95% confidence interval) for countries with well- and less well-developed statistical infrastructures was compiled (full information can be accessed through Choulga et al. (2021b)). Note that uncertainty bounds are rarely symmetrical and can differ substantially between countries with well- and less well-developed statistical infrastructures. This list is filtered again to match the 51 sub-sectors for which country budgets and flux maps were calculated.

1.2.2 European uncertainties

1.2.2.1 Greenhouse gases

Most countries in the European domain are Annex-I countries, which report their greenhouse gas emissions to the UNFCCC on an annual basis following standardized reporting guidelines. In most cases these countries also include an uncertainty estimate in their NIR reports (National Inventory Report), with separate uncertainty estimates for AD and EF. From these country reports we extract all available uncertainty information. There are different methods for uncertainty calculations and we take the most detailed one (Tier 2 (Monte Carlo) if available, else Tier 1 (error propagation)). For many countries Tier 2 estimates are available, but often the differences with the Tier 1 estimates are small.

The level of detail of the reported uncertainties differs per country and sector, but in all cases gap filling is needed to cover all relevant sectors. For example, most countries report uncertainty per fuel type for road transport, whereas emissions are available per vehicle type as well. We assume that the uncertainty given for road transport is then valid for all vehicle types. In the case that no uncertainty data is available for a sector we use the median uncertainty from all countries that do report something for this sector. The variability between countries in reported uncertainties for AD and the EF of CO_2 is limited and therefore this approach is suitable. We also use this median uncertainty for countries with emission reporting, but without an uncertainty estimate.

For countries that do not report their emissions we estimate the uncertainties at the GNFR level from the IPCC guidelines (IPCC, 2006). Since the emission factor for CO_2 depends only on the fuel type and not the combustion technology the uncertainty ranges are rather generic and consistent across sub-sectors. Where needed we have averaged uncertainty ranges, selected the largest uncertainties from a given range, or made an estimate based on the reported uncertainties from Annex-I countries. This results in the uncertainties given in Table 2. Note that GNFR sectors E and K are missing, because they are irrelevant for CO_2.

1.2.2.2 Air pollutants

For air pollutants (NO_x and CO) we assume that the same AD is used to calculate the emissions as for greenhouse gases. Therefore, the same uncertainties are applied as well. We use global EF uncertainty data per NFR sector-fuel combination from the most recent EMEP guidebook (European Environment Agency (2019)). These uncertainties are generic and applied to all countries, irrespective of whether emissions are reported or taken from another emission dataset.

The data are available for varying degrees of aggregation, starting at a tier-1 level (generic for the whole sector, often fuel-based) to tier-2 and tier-3 (technology-based and very specific). We typically use the tier-1 values that represent NFR sectors, since we do not consider a split per combustion technology. Tier-1 EF uncertainties are typically given for a number of fuel types, i.e. solid, gaseous, liquid and biomass. For power production these fuel types are split out further (e.g. solid into hard & brown coal) and we use the EF uncertainty of the dominant fuel type in Europe, making use of Eurostat data. As with the greenhouse gases, the tier-1 uncertainties for air pollutants are sometimes given at a more aggregated level than the emissions (e.g. industrial combustion as a whole) and we assume the same uncertainty applies to all sub-categories belonging to that sector.

In the cases where there is no uncertainty data for a specific sector we make use of the generic error ranges (95% confidence interval) also provided in Chapter 5 of the guidebook, on uncertainty estimation. This generic error range is different per sector, but for most sectors with missing data this amounts to an interval of -50% and +100% of the EF. For road transport we also consider biofuel emissions, but no uncertainty is given, so we apply the generic standard uncertainty range. Mopeds are a separate category, for which the uncertainties are copied from motorcycles. For NO_x, the chemical industry is dominated by nitric acid production, hence we use the tier-2 value for nitric acid there.

1.2.3 Estimating uncertainties in emissions

We start with pre-processing the uncertainties gathered from the NIR reports and the IPCC/EMEP guidebook. In all cases the uncertainties represent a 95% confidence interval, which we assume is 2 times the standard deviation (σ). In the NIR reports the AD and EF uncertainties are sometimes reported as one value. When the lower and upper value are similar (less than 5% difference) we assume a Gaussian error distribution, otherwise we assume a lognormal error distribution. For the air pollutants the error distribution is often lognormal. When the reported standard deviation exceeds 25% we also assume a lognormal error distribution to avoid getting negative values.

Since we do not always have the expected values of AD and EF we will work with emissions (E). Therefore, the uncertainties in AD and EF need to be combined to get the uncertainty in the emissions (E). We use a simple error propagation to estimate the uncertainty in the emissions from the standard deviations of AD and EF:

$$\sigma_{E,n} = \sqrt{\left(\frac{\sigma_{AD}}{AD}\right)^2 + \left(\frac{\sigma_{EF}}{EF}\right)^2} = \sqrt{{\sigma_{AD,n}}^2 + {\sigma_{EF,n}}^2} \quad (1)$$

The subscript ‘n’ means that we work with normalized (or relative) standard deviations, since we do not know the exact values of AD and EF. Note that we assume there is no error correlation between AD and EF, which is probably realistic.

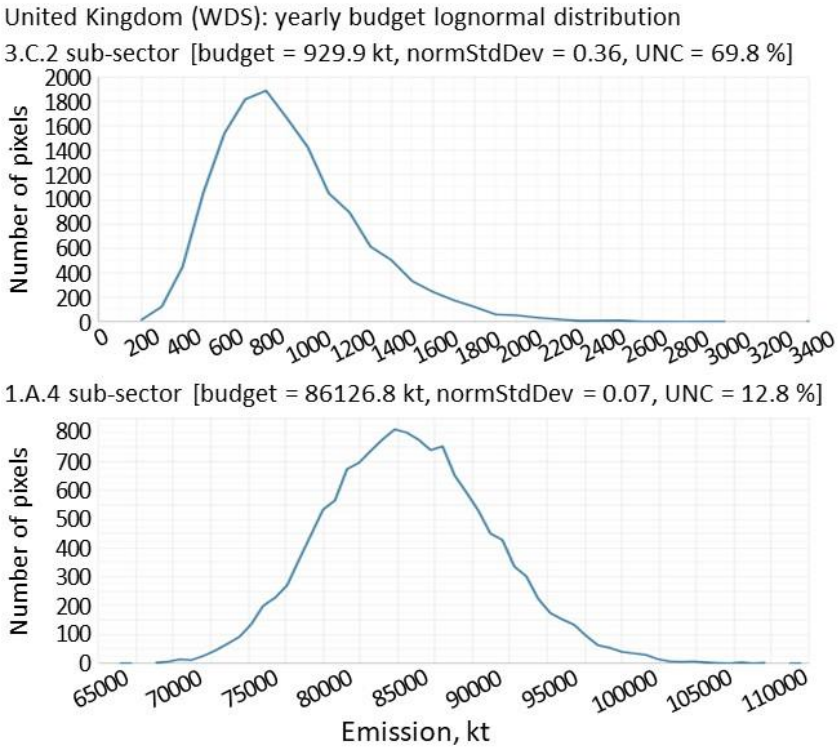


Figure 1: Two lognormal error distributions for emissions from agricultural sub-sector 3.C.2 (upper) and residential sub-sector 1.A.4 (lower) for the UK. Distributions are computed by taking a random sample (N = 14110) using the standard deviation and mean of the Gaussian distribution, based on EDGAR.

Normal error propagation works well with Gaussian error distributions, but we also have to deal with lognormal error distributions. A lognormal distribution is simply an exponential function of a Gaussian distribution. Therefore, we can calculate the standard deviation of the Gaussian function belonging to a lognormal function as:

$$\sigma_n = \frac{(\ln(lim_{upper}) - \mu}{\sqrt{2} \cdot erf^{-1}(2 \cdot 0.975 - 1)} \quad (2)$$

Or use a simple approximation:

$$\sigma_n = \frac{(\ln(lim_{upper}) - \ln(lim_{lower}))}{4} \quad (3)$$

Where lim_{upper} is the 97.5 percentile and lim_{lower} the 2.5 percentile of the lognormal distribution (i.e. the numerator in Eq. 3 is the 95% confidence interval) and μ in Eq. 2 is the mean of the Gaussian distribution. In Eq. 3 we divide by 4 to get the σ value. It is also possible to use one of the percentiles in combination with the expected value and then divide by two.

With Eq. 3 we can translate the lognormal function into a Gaussian one, extract the relevant parameters for the covariance matrix and still apply Eq. 1. An example is shown in Figure 1. Note that officially the combination of a normal and lognormal function doesn't result in a lognormal function, because the result can be negative. However, here we make the assumption that the combined distribution is lognormal, because the uncertainty of the Gaussian distribution is often relatively small.

1.2.4 Error correlations

The errors in AD and EF are not always independent. For example, at the country level the total gasoline consumption by road transport is relatively well-known. However, the AD we use for road transport is split into different vehicle types and road types. The error that is made in assigning the fuel consumption to each of these sub-categories needs to be compensated for by another sub-category. Hence, the error in the AD is in this case negatively correlated. Additionally, the AD is shared between pollutants, meaning a positive error correlation exists between the AD per sector and fuel for all pollutants. Also for EFs error correlations may occur, for example for sectors that receive the same EF estimate because they use similar technologies. It is likely that if the EF of one of those sectors is overestimated that this is also true for the other related sector(s). In this case, the error correlation is positive.

Unfortunately, the error correlations are unknown and difficult to estimate. Additionally, it is unclear how to estimate the error correlation between emissions from two sectors when both the AD and EF are correlated. The error correlation tends to go towards the error correlation for the most uncertain parameter (either AD or EF). To examine the importance of error correlations we do a sensitivity analysis on the European emissions. When two sectors have a correlation in either AD or EF we use the correlation for that variable. Otherwise we take the error correlation of the variable with the largest uncertainty in either of the two sectors. We use different correlation coefficients for a range of sectors:

- Correlations in AD occur for:
  - All road transport (different vehicle and road types) per fuel type
  - Biomass consumption for sub-sectors under OtherStatComb
  - Liquid fuel consumption for sub-sectors under OtherStatComb
  - All off-road transport and machinery per fuel type
- Correlations in EF occur for:
  - Road transport (different road types) per vehicle type and fuel type
  - Off-road transport (different fuel types) per sub-sector

The results and implications are discussed in section 3.1.

1.3 Spatial errors and correlations

1.3.1 Errors in global proxy maps

For the global emission dataset we make a first estimate of the grid-cell uncertainties using the country's normalised standard deviation as a constraint. Hence, the input for the global

grid cell uncertainties are the country-level sector-specific emission uncertainties. The global gridded uncertainties are only calculated using an error propagation approach.

1.3.2 Errors in European proxy maps

For the European dataset we take a different approach. The spatial errors in the emissions are caused partly by the resolution of the grid, but more importantly by uncertainties in the proxy maps used for downscaling the national emissions. There are different sources of uncertainty in the proxy maps:

- The value of each pixel, e.g. the population density
- The quality of the proxy, e.g. whether there are cells missing that in reality contain activity or vice versa
- The representativeness of the proxy for the activity causing the emissions

Table 3: Overview of proxy maps included in the uncertainty analyses for the European emissions, their 95% confidence interval and correlation length.

Proxy map	Uncertainty (95% CI)	Correlation length (km)
RoadTransport_Urban_PC	0.1	18
RoadTransport_Urban_Mopeds	0.1	18
RoadTransport_Urban_Motorcycles	0.1	18
RoadTransport_Highway_HDV	0.1	30
RoadTransport_Highway_LDV	0.1	30
RoadTransport_Highway_Buses	0.1	30
RoadTransport_Highway_PC	0.1	30
RoadTransport_Highway_Motorcycles	0.1	30
RoadTransport_Highway_Mopeds	0.1	30
RoadTransport_Rural_Buses	0.1	24
RoadTransport_Rural_LDV	0.1	24
RoadTransport_Rural_HDV	0.1	24
RoadTransport_Rural_Motorcycles	0.1	24
RoadTransport_Rural_Mopeds	0.1	24
RoadTransport_Rural_PC	0.1	24
RoadTransport_Urban_HDV	0.1	18
RoadTransport_Urban_LDV	0.1	18
RoadTransport_Urban_Buses	0.1	18
Population_total_2015	0.1	24
Population_rural_2015	0.1	24
Population_urban_2015	0.1	24
Wood_use_2014	0.1	24
CORINE_Arable_land_2018	0.1	24

The proxy quality is a difficult error to work with. If a grid cell falsely lacks activity there is no way to correct for that, as scaling a value of zero always gives zero. Similarly, if the location of a point source is wrong (for example the location of the head office is given instead of the actual emission stack) it is difficult to estimate where it should be instead. Since we cannot reliably compensate for this error, we choose to not account for it, while acknowledging it as a local source of uncertainty.

The errors in the pixel values are estimated from metadata for the proxy data. The population density map is based on Landscan, which estimates ambient population. This represents not only residential population but also working and travelling population. Archila Bustos et al. (2020) report an accuracy rating of around 93% for densely populated areas in Sweden, whereas for China the accuracy is around 88% (Ma et al., 2021). In general, sparsely populated areas seem to be more prone to errors, but in an absolute sense we expect larger uncertainties for densely populated errors. The proxies for road transport are based on OpenTransportMap (Jedlička et al., 2016), which combines different datasets which each will have their uncertainty. Unfortunately, little is known about the overall uncertainty in this dataset.

The representativeness error behaves different from the error in the pixel values. Besides that it adds an uncertainty to each pixel, it also causes errors to be correlated between pixels that have similar characteristics. For example, the heating demand for residential buildings depends on the population density. People that live closer together generally need less heating, e.g. in high-rise buildings. This means that the heating emissions are not linearly related to population density. If we make an error in describing this relationship it will affect pixels with similar characteristics in a similar fashion. Hence errors are spatially correlated. By combining the error in the pixel values and representativeness we estimate a CI of 10% for all proxy maps in Table 3. Note that for those countries for which emissions are not downscaled using these proxies no spatial error is included. The spatial error correlation will be explored in the next section.

1.3.3 Spatial error correlation lengths

We define the error correlation length as the maximum distance at which two grid cells are still correlated to each other. The correlation length is based on proxy values (not uncertainties), as the representativeness error causes correlations between pixels with similar values, and is estimated by constructing semi-variograms for each proxy map. A semi-variogram is a measure of the variance between two points related to their distance, where points nearer to each other show less variance and more correlation than points which are further apart (spatial autocorrelation). A semi-variogram model can be fitted to this data representation to obtain a curve showing the semi-variance as a function of distance. The semi-variogram range parameter is the distance of maximum semi-variance, i.e. the distance where values are no longer related to each other in space. We take this range parameter as an approximation of the correlation length for our proxy maps, rounded towards the nearest multiple of 6/10 km to represent the grid spacing (Europe/global).

The main proxy map for EDGARv6.0 are population density and settlement type based on Global Human Settlement Layer (GHSL) dataset (Melchiorri et al., 2019). The GHSL dataset is multitemporal, here last available year 2015 maps were used. Population density and settlement type maps were first re-gridded to a regular grid at 1 km resolution, and then sum and mode values at EDGARv6.0 grid (i.e. regular grid at ~10 km horizontal resolution) were computed respectively. Finally, non-populated grid-cells (i.e. population density equals zero) were masked over both maps and resulting point list was formed. Distances between all populated grid-cells within 150 km radius were computed. Distances were binned with 10 km intervals to compute Pearson's correlation r and significance p values between population grid-cells pairs. Correlation length is the greatest distance where correlation values constantly decrease by 1 % or more and p values are continuously significant at $p < 0.05$ (i.e. p not greater 0.05) for both population density and settlement type values.

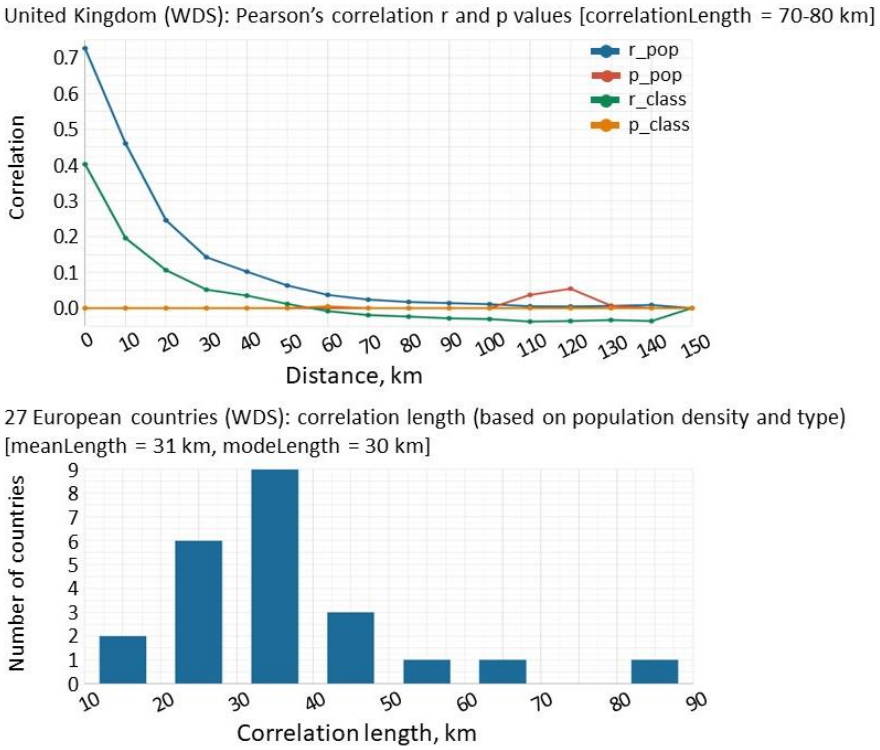


Figure 2: Pearson's correlation *r* for population density and settlement type class as a function of distance for the UK (upper); derived correlation lengths for all countries

For the United Kingdom the correlation length is 70-80 km. The most common correlation length between 27 European countries is 30-40 km (Figure 2). This value can be also used as the most appropriate value globally. However, it was observed that the error correlation length as it is defined here depends on country's urbanisation level. Countries with a high urbanisation level tend to have huge heavily populated urban clusters, e.g. Japan and the UK. In contrast, countries with lower urbanisation levels tend to have more smaller urban clusters that might not correlate by type with each other, like Finland and Sweden.

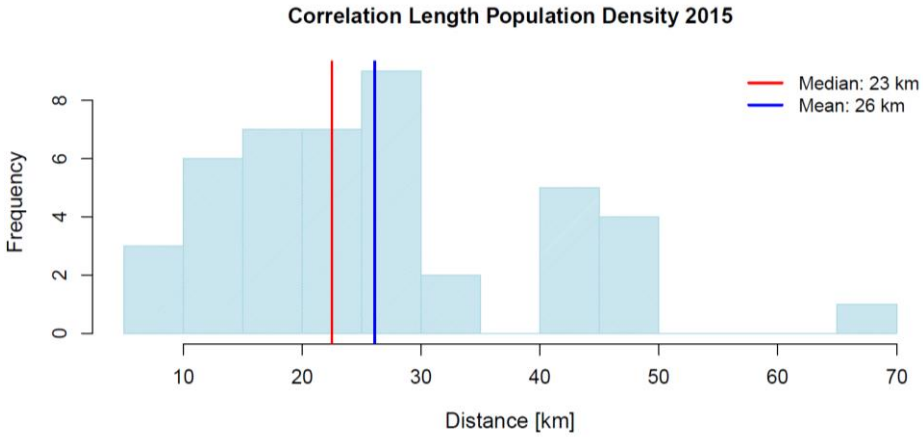


Figure 3: Derived correlation lengths for the European proxy map of total population density (2015), binned per 5 km.

For Europe the semi-variogram model construction is done using the fit.variogram function from the *gstat* geostatistical package in the R software. For each proxy map listed in Table 3 and each country for which emissions are downscaled using that proxy (excl. Russia) we construct a semi-variogram and fit a spherical semi-variogram model, setting the limits of the considered distance between 6 km (grid spacing) and 120 km. The fitting procedure then tries to optimize the model parameters to provide the best fit to the data. We do this twice: once

without setting an initial sill (the semi-variance at distance 0), and once by setting it to 0. This is to ensure that the resulting range values are not just the cause of the initial values set in the model. This results in two range values per country per proxy map and we pick the value that is within our set boundary (between 6 and 120 km) or the average of the two if both values are within this range.

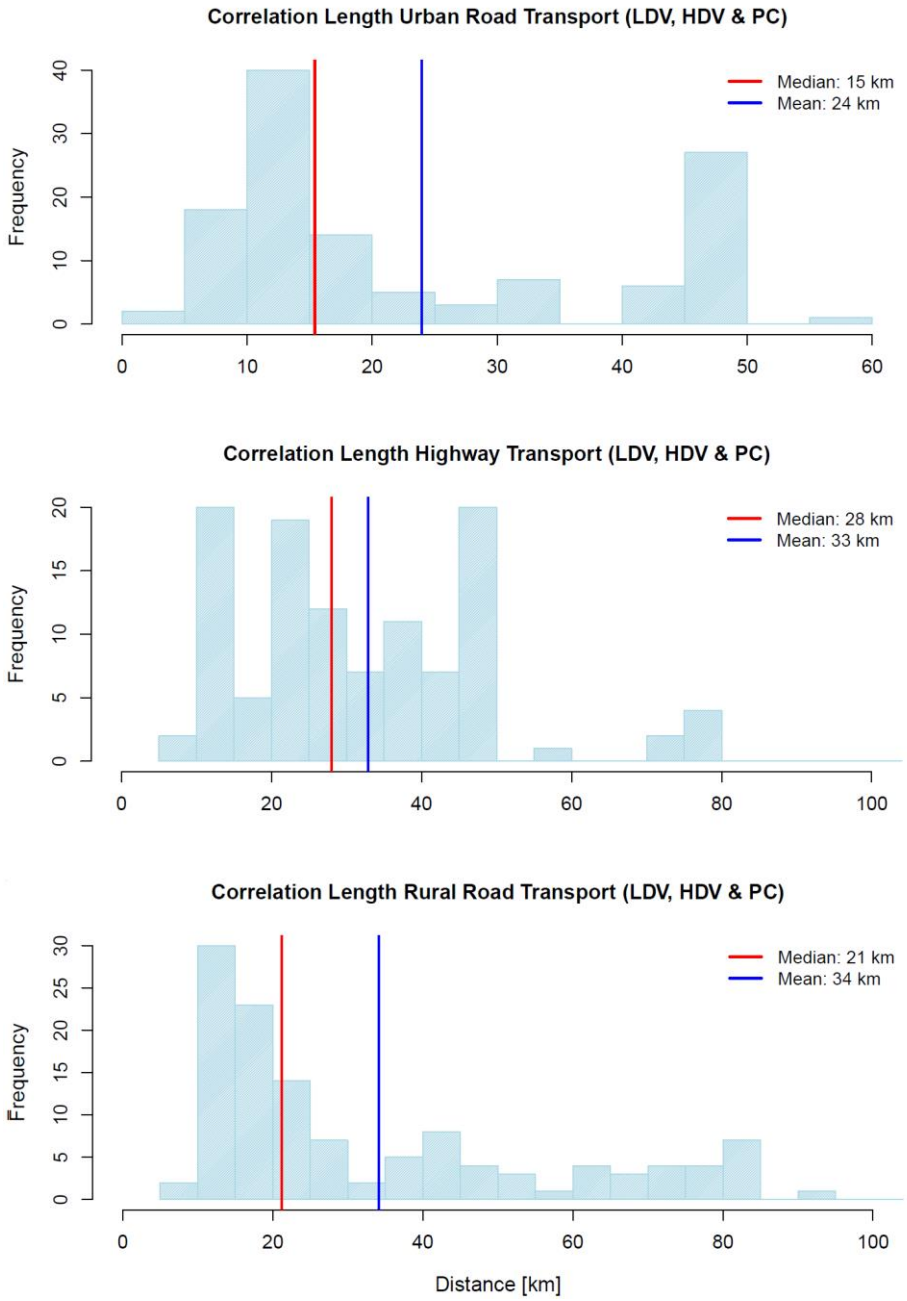


Figure 4: Derived correlation lengths for the European road transport proxy maps per road type for personal cars (PC), Heavy Duty Vehicles (HDV) and Light Duty Vehicles (LDV) combined, binned per 5 km.

This results in a histogram of correlation lengths (Figure 3 and Figure 4). Similar to what was found for the global dataset, we see large differences between countries with similar characteristics. For example, there seems to be a systematic difference in the values between countries in North/South Europe and East/West Europe. However, using country-specific correlation lengths would create irregularities near country borders. Therefore, we take the median value, rounded up towards the nearest 6 km to match the grid spacing, to be a fairly

representative value of the correlation length for that proxy map for all countries considered (Table 3).

For the road transport proxy maps the various vehicle types over a given road type do not show much variability in correlation lengths. However, different road types for a given vehicle type provide distinctly different correlation lengths. Hence we have combined the major vehicle types (passenger cars, light and heavy duty vehicles) for each road type (urban, highway, rural) to obtain road transport correlation lengths per road type (Figure 4).

1.3.4 Geographical uncertainty

For the global emission dataset we calculate the geographical uncertainty based on the country fraction map. From the definition given by Andres et al. (2016) the geographical uncertainty depends on number of neighbouring pixels from the same country: if a grid cell has no neighbours from the same country, that grid cell is assigned an uncertainty of 100%. In case there is 1 neighbour from the same country the uncertainty is 87.5%, in case of 2 – 75%, 3 – 62.5%, 4 – 50%, 5 – 37.5%, 6 – 25%, 7 – 12.5%, and in case when all neighbours are from the same country a grid cell is assigned an uncertainty of 0%. First, a country binary map was computed (if country fraction is greater than 0 grid-cell is assigned 1, otherwise – grid-cell is assigned 0). Geographical uncertainty values were assigned to each grid cell and all country maps were finally merged into one global map by summation.

2 Methods

Uncertainty estimates are mostly available at a detailed level (sectors, fuel-specific) and these need to be propagated to get an uncertainty estimate at the aggregated level. Simple error propagation techniques work well under specific circumstances, but when errors follow a non-Gaussian distribution or they are correlated things become more complicated. In such cases a Monte Carlo simulation can provide a more reliable estimate of the final uncertainty.

A Monte Carlo simulation produces a range of possible outcome values for any variable based on its probability distribution. It relies on random sampling from the probability distribution, which allows for different distribution shapes (e.g. Gaussian, lognormal, uniform, etc.). Another major advantage of the Monte Carlo simulation is that it can deal with interdependent input variables. A major downside is that it is a computationally expensive method, especially when dealing with a high spatiotemporal resolution. Therefore, we compare the results from the Monte Carlo simulation with the error propagation method to see if the latter gives decent results and could be used as well.

2.1 The Monte Carlo approach

2.1.1 Building the covariance matrix

To take into account error correlations a covariance matrix is needed. A covariance matrix is a square matrix that describes the covariance between each pair of elements in a vector. The diagonal values describe the variances of each element. A covariance matrix is always symmetric and positive semi-definite.

For a variable *A* with a Gaussian error distribution the variance can be calculated from the standard deviation:

$$var_{A,n} = \sigma_{A,n} \cdot \sigma_{A,n} \quad (4)$$

The covariance between variables *A* and *B* can be described as:

$$cov_{AB,n} = r \cdot \sigma_{A,n} \cdot \sigma_{B,n} \quad (5)$$

Where r is the correlation between the errors of both variables. The correlation takes a value between -1 (completely negatively correlated) and 1 (completely positively correlated). A

correlation of zero means the variables are uncorrelated. Note that there may be a need to scale error correlations to comply with the positive semi-definite requirement. The maximum negative correlation is:

$$r_{min} = \frac{1}{N-1} \quad (6)$$

Where N is the number of sectors that are correlated between themselves. To implement negative error correlations in AD we calculate the largest negative correlation and multiply that with the overall error correlations assumed for the whole group of sectors.

When all these values are calculated a covariance matrix ***P*** can be made, which takes this shape:

$$\boldsymbol{P} = \begin{bmatrix} var_{A,n} & cov_{AB,n} & cov_{AC,n} \\ cov_{AB,n} & var_{B,n} & cov_{BC,n} \\ cov_{AC,n} & cov_{BC,n} & var_{C,n} \end{bmatrix} \quad (7)$$

We make the assumption that each of the Gaussian functions has a mean value (μ) of one and therefore the standard deviations in Eq. 4 and 5 are normalized. This makes it easier to do calculations without an emission budget at hand.

2.1.2 Creating an ensemble

After preparing the covariance matrix we can start the actual Monte Carlo simulation. We start with a Cholesky decomposition of the covariance matrix:

$$\boldsymbol{P} = \boldsymbol{L}\boldsymbol{L}^* \quad (8)$$

In the case of only uncorrelated errors this results in a matrix ***L*** with standard deviations (i.e. the square root of the covariance matrix). However, if there are error correlations then this creates a matrix which also contains off-diagonal values. Combining this matrix with a vector of uncorrelated random samples (***u***) from a normal distribution with μ = 0 and σ = 1 through a dot product gives us a sample vector with the covariance properties of the whole system.

$$\boldsymbol{p} = \boldsymbol{L} \cdot \boldsymbol{u} \quad (9)$$

This vector ***p*** contains one sample from the whole system of variables, i.e. it represents one ensemble member. Now, we need to translate this normalized vector into one that represents the actual values of the variables. For each variable with a Gaussian distribution we can calculate the actual sample value as:

$$X_m = \bar{X}(\boldsymbol{p}[x] + 1) \quad (10)$$

Where ***p****[x]* represents one element of the ***p*** vector corresponding to variable *X*, X_m is the estimated value of variable *X* for ensemble member *m*, and $\bar{X}$ is the expected value of variable *X*. Note that we add one to the vector element to ensure we sample from a Gaussian function with μ = 1 instead of μ = 0. For each variable with a lognormal distribution we can calculate the actual sample value as:

$$X_m = \bar{X}e^{\boldsymbol{p}[x]} \quad (11)$$

Now we have one vector that contains an estimated value for each variable, taking into account random errors and correlations. We can use this vector in a model, for example to calculate the emissions on a more aggregated sector level. However, we want to estimate the error distribution of the model output and for that we need to create an ensemble of vectors by repeating the previous steps to get more vectors ***p*** with incorporated random errors.

2.1.3 A spatial ensemble

The approach described here can also be used to make a spatial ensemble, but then the covariances in the matrix ***P*** are a function of distance. That means for each element on the matrix the distance to other elements needs to be known and the correlation coefficient between elements *i* and *j* can be calculated as:

$$r_{i,j} = e^{-x_{i,j}/l} \quad (12)$$

Where $x_{i,j}$ is the distance between elements *i* and *j* and *l* is the correlation length scale. The correlation thus follows an exponential decay function, which is divided by the sum of all correlations within distance *l*. After distance *l* we assume the correlation is zero, following the definition of Kunik et al. (2019). When combined with the gridded uncertainty, which takes the same value per proxy map for each pixel and is assumed to have a Gaussian error distribution, we can build the covariance matrix (Eq. 7).

We create one set of perturbations for each proxy map and apply this equally to all pollutants using that proxy map. The proxy maps basically describe the spatial pattern in activity, which is the same for all pollutants. Therefore, the strongest correlation between pollutants is in their spatial distribution and we want to maintain this by giving each pollutant the same perturbation. For the selected sectors, which are related to fossil fuel combustion, this seems valid. However, in some other cases this may not be true, especially when pollutants result from different processes. Also, for now we have no sectors using the same proxy maps, but if this is the case this poses an additional challenge. Although the proxy value uncertainty is always the same, the representativeness error may be different. We will take this into consideration for the next deliverable.

2.2 Error propagation

For comparison we also use error propagation to calculate the aggregated uncertainties. If the results look similar to the Monte Carlo-based uncertainties this may provide a good alternative for large datasets. We are not looking for a perfect match, as the uncertainty in the prior uncertainty data is already quite large. But we do want a method that correctly captures the order of magnitude and variability between countries and pollutants.

There are several equations to calculate the uncertainty from the combination of multiple real variables. We have already shown how to calculate the error of the product of two uncorrelated variables (Eq. 1). This is used to calculate the total uncertainty in the emissions from the uncertainty in AD and EF. However, this is still very detailed and we want to know the uncertainty for an aggregated sector. In that case we need to sum emissions and the uncertainty in the aggregated emissions:

$$\sigma_{agg} = \sqrt{\sum_{i=m}^{n} \sigma_{sub,m}{}^2} \quad (13)$$

Where the subscript ‘agg’ refers to the aggregated emissions and uncertainty and the subscript ‘sub’ refers to the sub-sectors part of that aggregated sector. To use Eq. 13 we do need the emission budgets, because it uses actual standard deviations instead of normalized ones. The outcome of Eq. 13 can be compared to the relative standard deviation in the ensemble of country-level emissions per sector resulting from the Monte Carlo simulation.

For the European proxies a similar approach is used as for the country-level emissions, but to calculate the standard deviations a weighted average proxy map per aggregated sector is calculated. This means that for each combination of pollutant and country we determine the relative contribution of each sector to the aggregated sector, assign a weight to the corresponding proxy map and multiply that value with the fraction in each grid cell. This results in a new proxy map per sector with a sum of 1 per country. Next, we calculate the uncertainty in the averaged proxy map using Eq. 13, where the standard deviation is now related to the fraction in each grid cell. We only do this for CO_2 to create one set of perturbations, as was discussed before. Since all proxy maps have the same uncertainty (for now) this is a valid assumption. The actual values of the averaged proxy maps are pollutant-specific.

To generate global gridded uncertainties (i.e. disaggregating country yearly uncertainties to each country’s grid cell) as a first attempt it was assumed that grid cell uncertainties within a country are uncorrelated, and are equal to country’s normalised standard deviation N multiplied by a boosting parameter α – specific for each country and sub-sector, following Eq. 14:

$$(U_cB_c)^2 = (U_1B_1)^2 + (U_2B_2)^2 + \cdots + \left(U_pB_p\right)^2, where\ N = \frac{U_c}{B_c}\ and\ U_p = \frac{U_c \cdot \alpha \cdot B_p}{B_c} = N \cdot \alpha \cdot B_p \Rightarrow$$

$$\alpha = \sqrt{\frac{B_c^4}{\left[B_1^4+B_2^4+\cdots+B_p^4\right]}}, \quad (14)$$

where U and B represent standard deviation of distribution and emission budget, $_c$ and $_{1,2,\ldots,p}$ represent countries and pixel values respectively. To merge country's gridded uncertainties per sub-sector into more general groups Eq. 13 was used. Similar as to the European proxies also for the global emissions a weighted average proxy map per aggregated sector was calculated. Figure 5 shows the results for the industry sector. The majority of all values globally (i.e. 95 %) are between 0.6 and 50.3, with median value 13.4.

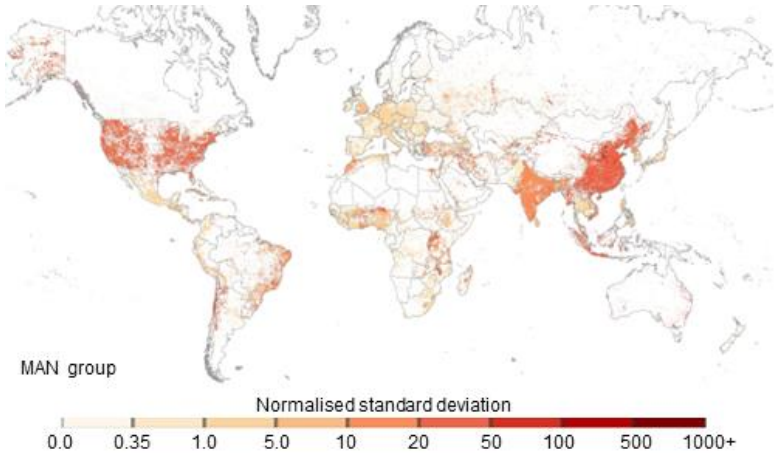


Figure 5: Map (upper) and value distribution (bottom) of normalised standard deviations per grid cell for the Industry sector.

In cases when emissions per grid cell are quite small, but country has lots of these grid-cells (i.e. emissions were not allocated but distributed over certain areas, e.g. according to population density), the boosting parameter α can be unrealistically huge. This leads to extremely high normalised standard deviation values per grid cell, e.g. for Residential/ commercial combustion group for Brazil and China, and for RoadTransport group – for United States, Canada and Russia. This issue is investigated further in more detail and a more suitable solution is being tested. Preliminary results for the industry sector show that majority of all values globally (i.e. 95 %) are between 0.2 and 2.4, with median value being 0.8 and maximum value 7.6.

Finally, we need to determine the error correlation length for the aggregated sectors. For this we also calculate a weighted average correlation length, using the same approach as for the proxy maps and only for CO_2. However, because the combined correlation length is also slightly sensitive to the uncertainty in each proxy map we calculate the weight based on both the emissions and the uncertainty (i.e. relative emission share multiplied by relative uncertainty share).

3 Results

For clarity we will discuss the results at different levels of detail. First, we start with the importance of error correlations at the country-level. Then we show results for country-level annual emissions from the Monte Carlo vs. the error propagation method. Finally, we will have a look at the spatial errors.

3.1 The importance of error correlations

To better understand the impact of assuming a particular correlation strength in AD and in EF we run the Monte Carlo simulation for different correlations and compare the emissions (only

for the European data). Interestingly, we find that the industry sector shows quite some variability due to the correlations, while there are no correlations for this sector. It seems that this is caused almost entirely by the high uncertainty for biomass from the paper and pulp industry in Sweden. Sweden reports an uncertainty of 100% in their CO_2 emissions for this sector, which makes up almost 20% of the total CO_2 emissions from the industry. It seems that with such large uncertainties for a sector that is relatively important, the sample that we used here is not large enough to get a stable result. This means that the distribution looks different if we create a new ensemble using the same settings, so differences between the runs may not just be caused by choosing different error correlations. Therefore, we removed Sweden from the sample and worked with the other countries only to illustrate the importance of error correlations.

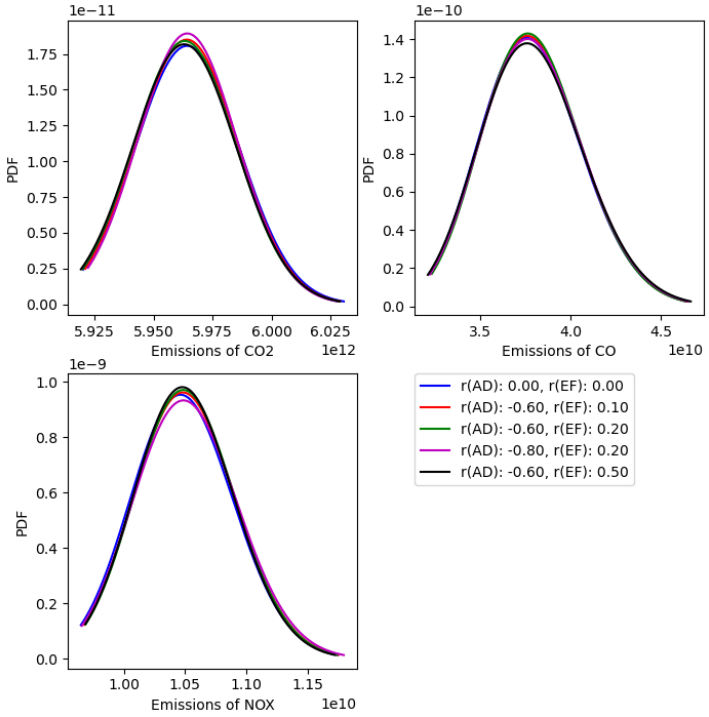


Figure 6: PDF of emissions of CO_2, CO and NOx (kg/yr) resulting from a Monte Carlo simulation (N=2000).

Figure 6 shows the probability density function (PDF) of the emissions for each pollutant. For CO and NO_x the presence of correlations has no significant impact. This was to be expected; the uncertainties in the emissions are relatively large for these pollutants and therefore the correlations have little effect. For CO_2 we also barely see an impact of the correlations, despite the smaller uncertainties. Perhaps this is due to the absence of error correlations in the most dominant source sectors for CO_2.

As discussed before, we assume error correlations exist between sub-sectors from road transport, other stationary combustion and from the other sector (off-road transport). Figure 7 shows almost no impact from the correlations on other stationary combustion, which is due to the fact that we only assumed correlation in AD for the less relevant fuels. For road transport and the other sector the impact is visible, although very small. From this we conclude that the error correlations in AD and EF are of limited importance, irrespective of their exact values.

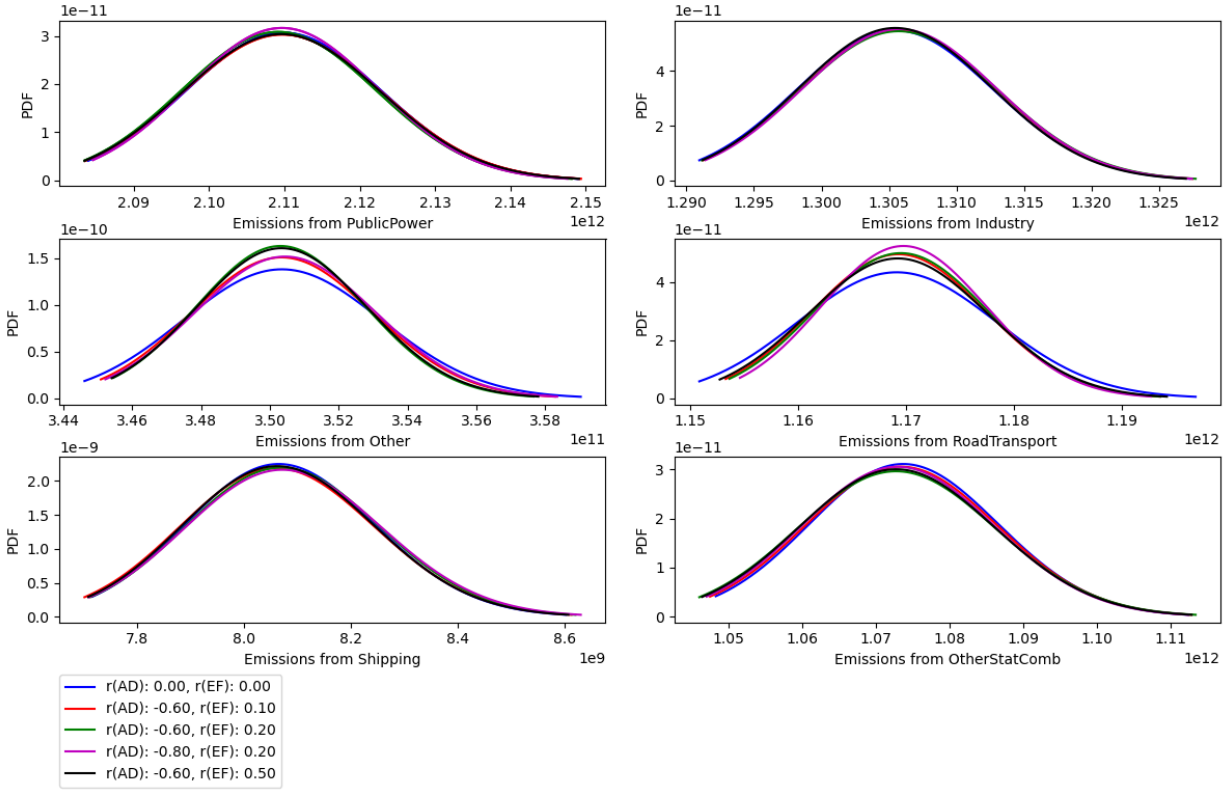


Figure 7: PDF of emissions of CO_2 (kg/yr) for different sectors resulting from a Monte Carlo simulation (N=2000).

3.2 Country-level uncertainties

The results from the Monte Carlo for the country-level annual emissions (UK only) in EDGAR are shown in Figure 8. The normalized standard deviation for the OtherStatComb sectors (upper panel) is estimated at 0.07 (representing 12.7 % uncertainty). For RoadTransport (lower panel) the normalized standard deviation resulting from the Monte Carlo simulation is about 0.03 (representing 5.5 % uncertainty). For countries with less well-developed statistical infrastructure the uncertainties are larger – for OtherStatComb the relative standard deviation is around 0.15. For RoadTransport there is a strong difference between countries, with often values around 0.04-0.06, but with quite some outliers up to 0.15-0.18. Some values can be slightly different from the reported values as global approach starts only from two country types – with well- and less well-developed statistical infrastructure, and also there might be small inconsistencies in emission budgets – for the global approach global emission flux map is used and then flux values are translated into emission mass and summed according country masks which also can have minor inaccuracies.

Since the UK is also part of the European emissions dataset we can compare the results from both datasets. With the European data we find normalized standard deviations of 1.9% for the OtherStatComb sector and of 0.3% for the RoadTransport sector. Although this seems quite low the UK reports uncertainties of 2% in each sub-sector of the RoadTransport group and in the Monte Carlo simulation this results in an aggregated uncertainty which is a bit lower then this by partly compensating sub-sectors. So the results are in line with their own reported uncertainties.

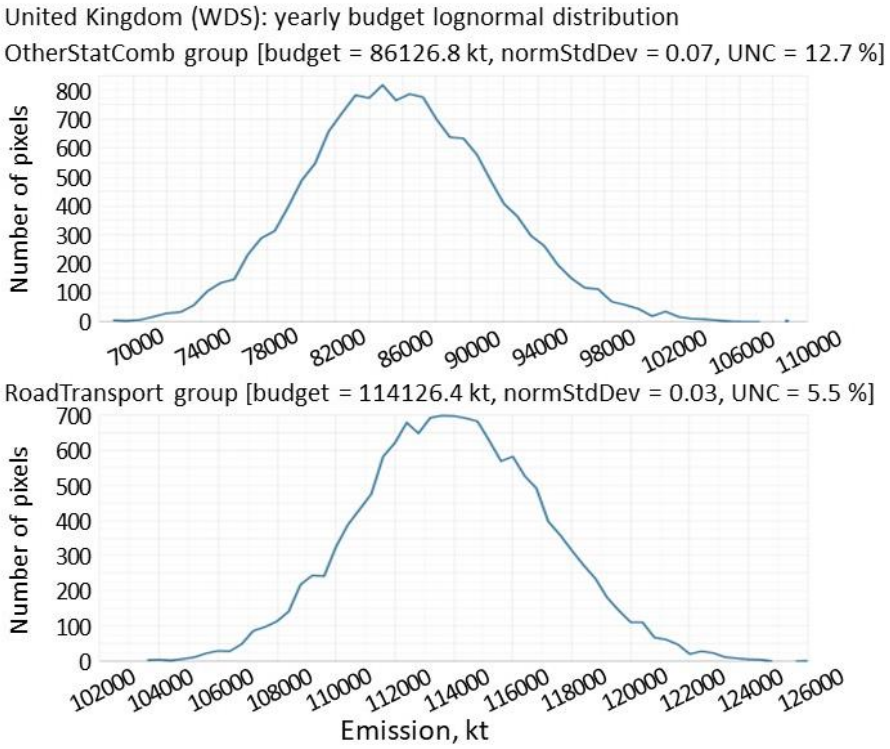


Figure 8: Distribution of CO_2 emissions resulting from the Monte Carlo simulation (N = 14110, no error correlations) for the OtherStatComb (upper) and RoadTransport (lower) grouped sectors for the UK.

Figure 9 shows the normalized spread in total country-level emissions for the European emissions (including small sectors not part of the Monte Carlo). For CO_2 we generally see a very small range of just a few percent. For CO and NO_x the spread is larger and more skewed with more high values. This is the result of the EF uncertainties with are often strongly lognormal. We assumed no error correlations between AD and EF here.

The results from the error propagation method are also illustrated with red dots showing the 95% confidence interval. The ranges and order of magnitude look similar and clearly follow the variability between countries that is also visible from the Monte Carlo simulation. Especially the lower limits show some stronger deviations. From this we can conclude that the Monte Carlo has added value over normal error propagation, which in absence of error correlations should be attributed to a better representation of different error distributions. Nevertheless, the simple approach gives reasonable results.

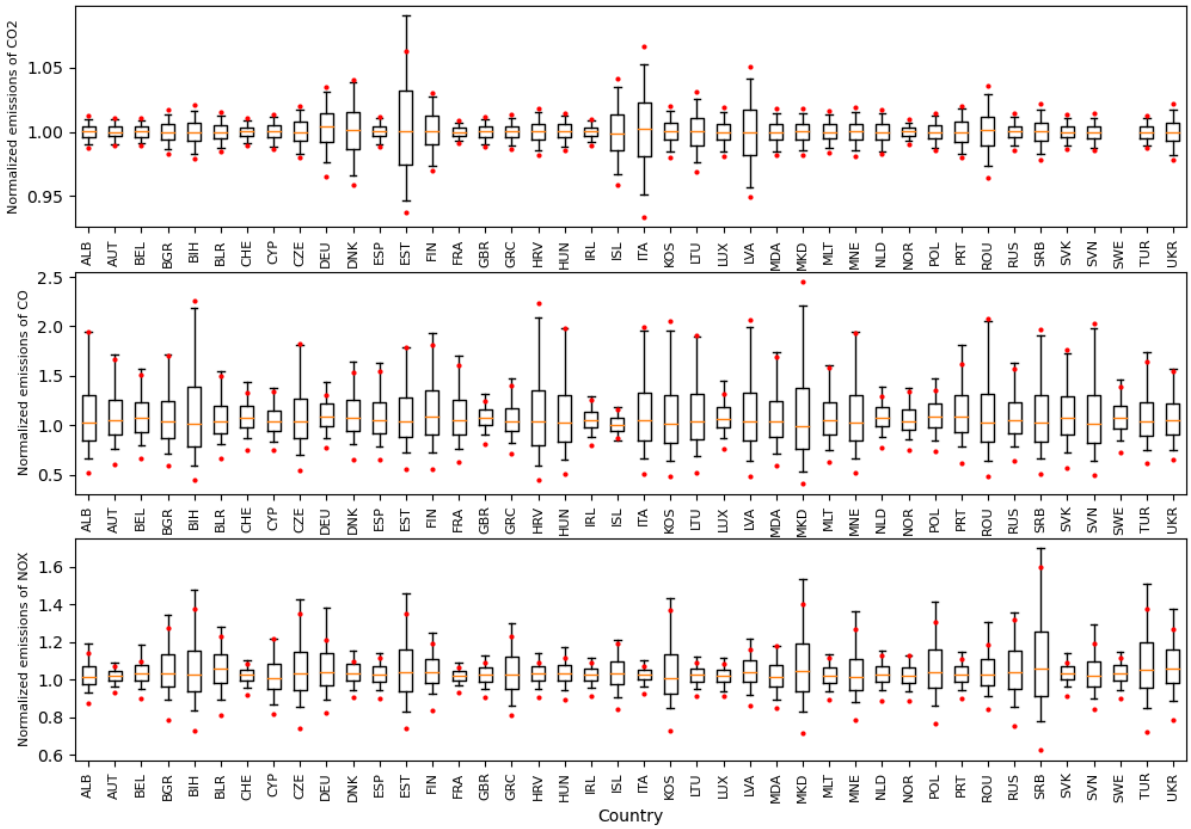


Figure 9: Normalized spread in emissions of CO_2, CO and NOx for European countries resulting from a Monte Carlo simulation (N=2000, no error correlations) and the 95% confidence interval as calculated from the error propagation method (red dots).

3.3 Spatial errors

As mentioned before we are unable to make a large ensemble for the European gridded emissions at the level of detail required to test the method. Therefore, we do a Monte Carlo simulation for the Dutch road transport sector with an ensemble size of N = 200. We compare the results against the error propagation method, from which we also create an ensemble using the Monte Carlo simulation with aggregated uncertainties and the averaged proxy map. The reason is that we have a relatively small ensemble size (N=200) to reduce computational expenses and this may not be sufficient to compare to the error propagation method directly. Moreover, the inverse modellers may want to create an ensemble of possible emission input maps for their models from the aggregated errors and this should be similar to the ensemble from the detailed Monte Carlo.

When perturbing the spatial proxy maps the sum of all fractions may no longer add up to 1 per country. We have tested the importance of correcting for this, but we found that given the relatively small correlation lengths these deviations have a negligible impact on the country-level emissions, except maybe for very small countries.

A first test is to compare the standard deviation in the emissions of each pixel in the two ensembles. The result is shown in Figure 10. We see a strong correlation in the standard deviation per grid cell. This suggests that the error propagation method gives a good estimate of the overall grid cell uncertainty and variations between grid cells. Moreover, when we sum emissions from all grid cells we get a standard deviation that is similar to the country-level uncertainty for this sector.

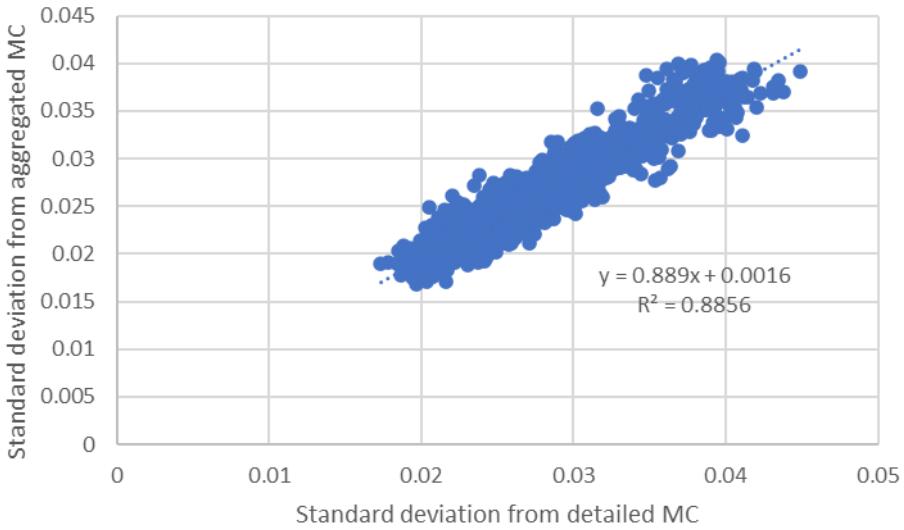


Figure 10: Scatter plot of standard deviation per pixel with road transport emissions (N = 1099) in the ensemble resulting from the detailed Monte Carlo simulation (x-axis) and the Monte Carlo simulation with aggregated uncertainties from the error propagation method (y-axis).

Next, we look at the error correlation length in the aggregated sector emissions. We calculate the correlation coefficient between two grid cells from the full Monte Carlo ensemble (i.e. the larger the correlation coefficient the stronger the perturbations in those grid cells are correlated). The average correlation coefficient per grid cell distance (6 km) is shown in Figure 11. From the error propagation method we estimate an average correlation length of approximately 24 km for road transport (dashed vertical line in Figure 11). This matches well with the decay in the correlation coefficient.

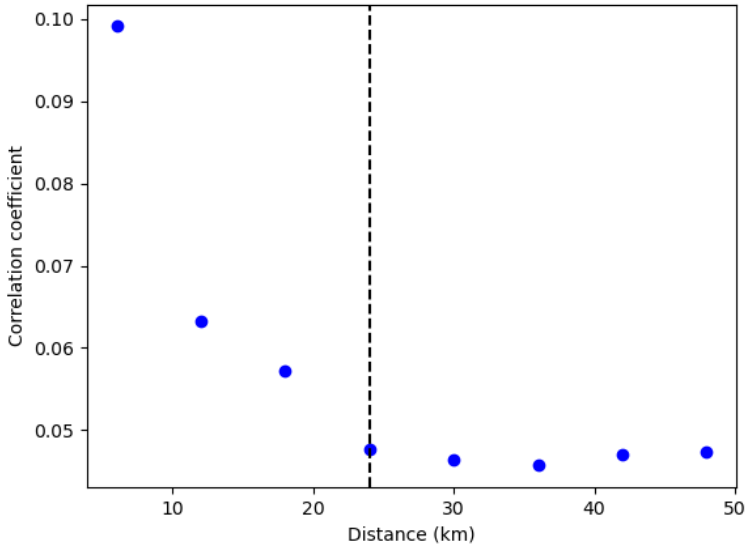


Figure 11: The average correlation coefficient between pixels at certain distance from each other, as calculated from the detailed Monte Carlo ensemble.

4 Conclusion and discussion

In this report we describe different approaches to calculate the prior emission uncertainty. Although the Monte Carlo approach is more flexible when error correlations and non-Gaussian errors are present, the more simple error propagation method mimics the aggregated errors and variability between sectors, countries and pixels well.

The uncertainties in country-level emissions are not necessarily accurate. Often IPCC guidelines are followed to calculate these uncertainties, containing several assumptions and generalisations. Also for the air pollutants the uncertainty ranges for emission factors are often based on expert judgement. Although for specific technologies measurements may be available, it is difficult to get a better foundation for these numbers. Nevertheless, the uncertainties are well-documented, which makes this work easier to repeat for other datasets and that is a major advantage of using these data. TNO aims to improve the uncertainty estimate for some major sectors, especially for CO and NO_x, but this is outside the scope of this project.

The uncertainties in proxy maps are even more difficult to estimate. Unfortunately, data on the underlying datasets is often lacking. Some datasets are processed to serve as a better proxy for a particular sector and it is very difficult to validate these spatial patterns. We believe it is important to develop a standard to estimate these uncertainties, but we realize that with the large amount of data this may be too challenging. Nevertheless, we will try to work on this further in the coming years.

This work is surrounded by uncertainties, both in the input data and in the methodology (e.g. assumptions). Unfortunately, validation of our results is extremely difficult. We believe that the uncertainties in the country-level emissions of CO_2 are rather accurate for most developed countries, as the uncertainties in itself are small. However, for co-emitted species we use one uncertainty range for the emission factor of all countries. Since these ranges are often lognormal, high EF values can occur in an ensemble, which are probably unrealistic in developed countries with state-of-the-art technologies. How much this affects the final result is difficult to estimate, since we do not have a better estimate for the uncertainty range in developed countries. The uncertainty in the spatial distribution is likely to be underestimated, since we ignore the proxy quality error and assume the representativeness error only to affect the correlation length and not the total error per pixel. I addition, for the global approach it is important to remember that a methodology suitable for some countries can be less suitable for others, e.g. on different continents.

Next year further work will be done to extend the number of sectors for which spatial uncertainty data is provided. Moreover, we will include temporal errors and error correlations as well to allow for temporal downscaling of the emissions. We will also consider what to do with the aggregated proxy map uncertainty when different pollutants have very different shares (e.g. for CO residential wood combustion is a dominant process with a much larger spatial uncertainty than natural gas usage).

The uncertainty product is a good step towards a full quantification of prior emission uncertainties, which is already a huge improvement compared to the back-of-the-envelope calculations that are now often used to estimate uncertainties for inverse modelling studies. It will provide a better basis for the MVS capacity.

Appendix A: Overview EDGAR sectors

Table 4: Overview of sectors part of the EDGAR emissions, including sub-sectors and related IPCC activities (to match uncertainties).

Nr	EDGAR sector	EDGAR sub-sector [nr. non-zero unc/ total sub-sectors]	IPCC2006 activity	Note
1	AGS	[2/4] – 2 budgets: 3.C.2, 3.C.3	3.C.2, 3.C.3, 3.C.4, 3.C.7	
2	CHE	[9/14] – 1 budget: 2.B	2.B.1, 2.B.2, 2.B.3, 2.B.4_CPR, 2.B.4_GLA, 2.B.4_GLL, 2.B.5, 2.B.6, 2.B.8.a, 2.B.8.b, 2.B.8.c, 2.B.8.d, 2.B.8.e, 2.B.8.f	2.B (sum of 9 sub-sectors)
3	ENE (ENEa, ENEs)	[1/1] – 1 budget: 1.A.1.a	1.A.1.a	Divide ENE to ENEa and ENEs based on pixel flux threshold.
4	IND	[1/1] – 1 budget: 1.A.2	1.A.2	Compare difference between map and EXCEL budgets.
5	IRO	[2/2] – combined budget: IRO+NFE	2.C.1, 2.C.2	
6	NEU	[2/3] – combined budget: NEU+PRU_SOL	2.D.1, 2.D.2, 2.D.4	
7	NFE	[4/5] – combined budget: IRO+NFE	2.C.3, 2.C.4, 2.C.5, 2.C.6, 2.C.7	
8	NMM	[4/4] – 4 budgets: 2.A.1, 2.A.2, 2.A.3, 2.A.4	2.A.1, 2.A.2, 2.A.3, 2.A.4	
9	PRO	[11/17] – 3 budgets: 1.B.1, 1.B.2_PRO, 1.C	1.B.1.a.i.1, 1.B.1.a.i.2, 1.B.1.a.i.3, 1.B.1.a.ii.1, 1.B.1.a.ii.2, 1.B.2.a.ii, 1.B.2.a.iii.2, 1.B.2.a.iii.3, 1.B.2.b.ii, 1.B.2.b.iii.2, 1.B.2.b.iii.4_STR, 1.B.2.b.iii.4_TRN, 1.B.2.b.iii.5, 1.C.1.a_PIP, 1.C.1.a_SHP, 1.C.2.a_INJ, 1.C.2.a_STR	1.B.1 (sum of 3 [5] sub-sectors); 1.B.2_PRO (sum of 7 [8] sub-sectors) = 1.B.2 – 1.B.2_REF; 1.C (sum of 1 [4] sub-sectors) = PRO – 1.B.1 – 1.B.2_PRO.
10	PRU_SOL	[1/5] – combined budget: NEU+PRU_SOL	2.D.3, 2.E.1, 2.E.5, 2.F, 2.G	
11	RCO	[1/4] – 1 budget: 1.A.4	1.A.4, 1.A.5.b.ii, 1.A.5.a, 1.A.5.b.i	! RCO has 1.A.5.b.ii (and 1.A.5.a, 1.A.5.b.i) with zero unc – might be part of

				available budget 1.A.5
12	REF_TRF	[4/7] – 2 budgets: 1.A.1.bc, 1.B.2_REF	1.A.1.b, 1.B.2.a.iii.4, 1.A.1.c, 1.A.5.b.iii, 1.B.1.c, 1.B.2.a.iii.6, 1.B.2.b.iii.3	1.A.1.bc (sum of 2[2] sub-sectors); 1.B.2_REF (sum of 2 [3] sub-sectors) = REF_TRF – 1.A.1.bc; ! REF_TRF has 1.A.5.b.iii with zero unc – might be part of available budget 1.A.5
13	SWD_INC	[1/1] – 1 budget: 4.C	4.C	
14	TNR_avi_CDS	[1/1] – combined budget: CDS+CRS+LTO	1.A.3.a_CDS	
15	TNR_avi_CRS	[1/1] – combined budget: CDS+CRS+LTO	1.A.3.a_CRS	
16	TNR_avi_LTO	[1/1] – combined budget: CDS+CRS+LTO	1.A.3.a_LTO	
17	TNR_Other	[2/3] – 2 budgets: 1.A.3.c, 1.A.3.e	1.A.3.c, 1.A.3.e_OFF, 1.A.3.e_PIP	
18	TNR_Ship	[1/1] – 1 budget: 1.A.3.d	1.A.3.d	
19	TRO	[1/1] – 1 budget: 1.A.3.b_noRES	1.A.3.b	

Appendix B: User guidelines

The European uncertainty data is delivered as a NetCDF file following a similar format as the emission data. Information is provided on the domain (latitude, longitude), countries, pollutants, sectors and source type (area or point). All data is aggregated into the six source sectors listed in Table 2.

The variable '*emissions'* contains the emissions per country, pollutant and sector. This variable contains all countries and sea regions part of the domain covered by the emission dataset. The uncertainty is given by the variable '*unc_emis*', which follows the same structure as the emissions. The uncertainty is a relative standard deviation. This value is zero when we lack information, which is mostly true for sea regions and some country-sector combinations which have no emissions at all. When working with these data it means that those particular country-pollutant-sector combinations receive no uncertainty and thus the emissions are constant.

Spatial data are presented as a list of sources, similar to the emission data. Each source in this list has a specific location, country, sector and source type. The variable '*distribution'* describes the spatial distribution of emissions for each pollutant and consists of fractions which for each country-pollutant-sector combination should add up to one. When multiplying these values with the corresponding emissions it gives the original emission data. Uncertainties

related to the spatial proxies are represented by the variable '*unc_dist*' and again the uncertainty is a relative standard deviation. There is only one list of uncertainties, which is based on CO_2 as was discussed before. Finally, the variable '*corr_length*' gives the spatial correlation length for each aggregated sector. The unit of the correlation length is the number of grid cells in the native grid resolution of 1/10 x 1/20 degrees.

Important to note is that we only include the two sectors for which we have determined spatial errors (road transport and other stationary combustion). So the spatial patterns of the other sectors need to be taken from the original prior emission dataset. The correlation length is zero for sectors which are not included yet for the spatial uncertainties.

For modellers who might want to create an ensemble of possible emission maps there is another important consideration. Since error propagation makes use of Gaussian error definitions the uncertainties provided are all relative standard deviations. However, emissions may follow lognormal distributions as well, especially for CO and NO_x. Therefore, we may consider all uncertainties in the emissions larger than 0.05 (5%) to be lognormal and Eq. 11 can be applied. For the spatial distribution we assume all uncertainties are Gaussian and Eq. 10 can be applied (also for other Gaussian error distributions).

The data are available through an FTP site (coco2@ftp.ecmwf.int), following the directory structure data-exchange/WP2/D2-6-prior_emission_uncertainties. Please contact the CoCO2-Coord to obtain the FTP password access.

References

Andres, R. J., Boden, T. A., and Higdon, D. M.: Gridded uncertainty in fossil fuel carbon dioxide emission maps, a CDIAC example, Atmos. Chem. Phys., 16, 14979–14995, https://doi.org/10.5194/acp-16-14979-2016, 2016.

Archila Bustos, M.F., Hall, O., Niedomysl, T, and Ernstson, U.: A pixel level evaluation of five multitemporal global gridded population datasets: a case study in Sweden, 1990–2015. Popul Environ 42, 255–277, https://doi.org/10.1007/s11111-020-00360-8, 2020.

Choulga, M., Janssens-Maenhout, G., Super, I., Solazzo, E., Agusti-Panareda, A., Balsamo, G., Bousserez, N., Crippa, M., Denier van der Gon, H., Engelen, R., Guizzardi, D., Kuenen, J., McNorton, J., Oreggioni, G., and Visschedijk, A.: Global anthropogenic CO_2 emissions and uncertainties as a prior for Earth system modelling and data assimilation, Earth Syst. Sci. Data, 13, 5311–5335, https://doi.org/10.5194/essd-13-5311-2021, 2021a.

Choulga, M., Janssens-Maenhout, G., and McNorton, J.: Anthropogenic CO2 emission uncertainty calculation tool CHE_UNC_APP. Zenodo. https://doi.org/10.5281/zenodo.5196190, 2021b.

Crippa, M., Guizzardi, D., Schaaf, E., Solazzo, E., Muntean, M., Monforti-Ferrario, F., Olivier, J.G.J., and Vignati, E.: Fossil CO2 and GHG emissions of all world countries - 2021 Report, in prep., 2021a.

Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Lo Vullo, E., Solazzo, E., Monforti-Ferrario, F., Olivier, J., and Vignati, E.: EDGAR v6.0 Greenhouse Gas Emissions. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/97a67d67-c62e-4826-b873-9d972c4f670b, 2021b.

European Environment Agency, EMEP/EEA air pollutant emission inventory guidebook 2019: technical guidance to prepare national emission inventories, Publications Office, https://data.europa.eu/doi/10.2800/293657, 2019.

Global Energy Observatory, Google, KTH Royal Institute of Technology in Stockholm, University of Groningen, World Resources Institute: Global Power Plant Database, https://developers.google.com/earth-engine/datasets/catalog/WRI_GPPD_power_plants#description, last access: June 2022, 2018.

IPCC: 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme, Eggleston, H.S., Buendia, L., Miwa, K., Ngara, T. and Tanabe, K. (eds). Published: IGES, Japan, 2006.

IPCC: 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, Buendia, E., Guendehou, S., Limmeechokchai, B., Pipatti, R., Rojas, Y., Sturgiss, R., Tanabe, K., Wirth, T., Romano, D., Witi, J., Garg, A., Weitz, M.M., Bofeng, C., Ottinger, D.A., Dong, H., MacDonald, J.D., Ogle, S.M., Theoto Rocha, M., Sanz Sanchez, M.J., Bartram, D.M., and Towprayoon, S. (aut.); Gomez, D. and Irving, W. (eds.), Vol1. Ch.8, 2019.

Jedlička, K., Hájek, P., Cada, V., Martolos, J., Šťastný, J., Beran, D., Kolovský, F., and Kozhukh, D.: Open transport map — Routable OpenStreetMap, *2016 IST-Africa Week Conference*,1-11, https://doi.org/10.1109/ISTAFRICA.2016.7530657, 2016.

JRC, EDGAR – Emissions Database for Global Atmospheric Research, Global Greenhouse Gas Emissions, EDGAR v6.0, https://edgar.jrc.ec.europa.eu/index.php/dataset_ghg60, last access: June 2022, 2022.

Kuenen, J., Dellaert, S., Visschedijk, A., Jalkanen, J.-P., Super, I., and Denier van der Gon, H.: CAMS-REG-v4: a state-of-the-art high-resolution European emission inventory for air quality modelling, Earth Syst. Sci. Data, 14, 491–515, https://doi.org/10.5194/essd-14-491-2022, 2022.

Kunik, L., Mallia, D.V., Gurney, K.R., Mendoza, D.L., Oda, T., and Lin, J.C.; Bayesian inverse estimation of urban CO_2 emissions: Results from a synthetic data simulation over Salt Lake City, UT. Elementa: Science of the Anthropocene, 7, https://doi.org/10.1525/elementa.375, 2019.

Ma, J., Sun, Y., Meng, D., Huang, S., Li, N., and Zhu, H.: Accuracy Assessment of Two Global Gridded Population Dataset: A Case Study in China. In *The 4th International Conference on Information Science and Systems (ICISS 2021)*. Association for Computing Machinery, New York, NY, USA, 120–125, https://doi.org/10.1145/3459955.3460610, 2021.

Melchiorri, M.P., Florczyk, M., Corbane, A.J., and Kemper, C.: Principles and applications of the global human settlement layer as baseline for the land use efficiency indicator—SDG 11.3.1 ISPRS Int. J. Geo-Inf. 8 96, 2019.

Super, I., Dellaert, S. N. C., Visschedijk, A. J. H., and Denier van der Gon, H. A. C.: Uncertainty analysis of a European high-resolution emission inventory of CO_2 and CO to support inverse modelling and network design, Atmos. Chem. Phys., 20, 1795–1816, https://doi.org/10.5194/acp-20-1795-2020, 2020.

United States Office of the Geographer: Large Scale International Boundary Polygons, Detailed, https://developers.google.com/earth-engine/datasets/catalog/USDOS_LSIB_2017#description, last access: June 2022, 2017.

Document History

Version	Author(s)	Date	Changes
	Name (Organisation)	dd/mm/yyyy	
V0.1	Ingrid Super (TNO), Arjan Droste (TNO)	8/6/2022	First version containing input TNO
V0.2	Ingrid Super (TNO), Arjan Droste (TNO), Margarita Choulga (ECMWF)	21/6/2022	Added input from ECMWF
V1.1	Ingrid Super (TNO), Arjan Droste (TNO), Margarita Choulga (ECMWF)	3/8/2022	Processed review comments

Internal Review History

Internal Reviewers	Date	Comments
Name (Organisation)	dd/mm/yyyy	
Martin Jung (MPG)	1/7/2022	
Hannakaisa Lindqvist (FMI)	04/07/2022	A detailed report with a clear structure and well-presented content. Minor comments and edits have been suggested.

Estimated Effort Contribution per Partner

Partner	Effort
TNO	5 PM
ECMWF	7 PM
Other partners (for task meetings)	0.5 PM
Total	**12.5**