Quantification of uncertainty ranges from European multi-model inversions and ways to benchmark inversion systems

M. Scholze, S. Houweling, U.
Karstens, A.G.C.A. Meesters
& G. Monteil

CoCO2
Prototype system for a
Copernicus CO$_2$ service

Co-ordinated by
ECMWF

D5.6 Quantification of uncertainty ranges from European multi-model inversions and ways to benchmark inversion systems

Dissemination Level:	Public
Author(s):	M. Scholze (ULUND), S. Houweling (VUA), U. Karstens (ICOS-CP), A.G.C.A. Meesters (VUA) & G. Monteil (ULUND)
Date:	08/11/2023
Version:	2.0
Contractual Delivery Date:	30/09/2023
Work Package/ Task:	WP5/ T5.6
Document Owner:	ULUND
Contributors:	WP5.6 Partners
Status:	Version for submission after internal review

CoCO2: Prototype system for a Copernicus CO₂ 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

Table of Contents

Figures

Tables

1 Executive Summary

This document reports on the work performed in Task 5.6 within the work package WP5 'Connecting scales and uncertainties' of the CoCO2: Prototype system for a Copernicus CO$_2$ service project. The aim of Task 5.6 'Assessment of uncertainties in European inversion of CO$_2$ and CH$_4$' is essentially to perform multi-model intercomparisons to quantify the uncertainty range in European CO$_2$ and CH$_4$ land-atmosphere fluxes from atmospheric inversions. The work in this task also includes preparing a roadmap on benchmark atmospheric transport model inversions.

For CO$_2$ an analysis of the impact of changing various components (foreground and background atmospheric transport model, prior uncertainty specification) in an atmospheric tracer transport inversion on posterior CO$_2$ fluxes was performed. For this, we used two such inversion systems (CSR and LUMIA) over Europe for the year 2018. The objective is to identify the dominant driver of uncertainty in the posterior CO$_2$ estimates. Two Lagrangian transport models (STILT and FLEXPART) were used to assess the impact of foreground (regional) transport on posterior CO$_2$ fluxes. Two Eulerian transport models (TM3 and TM5) were used to quantify the impact of the background (lateral boundary conditions) on posterior CO$_2$ fluxes, and finally two different schemes to set spatio-temporal prior uncertainties as employed by LUMIA and CSR were used to quantify the impact on posterior CO$_2$ fluxes. These variations lead to an ensemble of eight inversions. The results from this ensemble show a large spread in the annual terrestrial posterior fluxes over the whole domain of 0.92 PgC yr^{-1} ranging between −0.72 and 0.20 PgC yr^{-1}, which is almost twice as large as the assumed prior uncertainty of 0.47 PgC yr^{-1}. The largest part of the spread in the results could be accounted for by the regional transport model component. The global transport models used for providing background contribution were responsible for a smaller part of the spread but with a quasi-constant offset, hence acting like a bias. The differences arising from using different inversion systems (i.e. prior uncertainty specifications) were the smallest.

In the CH$_4$ inversion intercomparison eight different atmospheric transport inversion systems have been used. The range in posterior fluxes obtained from the inversions is currently too large to provide a strong constraint on national emissions. The only countries for which the inversions deviate systematically from the prior are The Netherlands and Italy, where the inventory reports respectively lower and higher emissions than the inversions. In the case of the Netherlands this may include the region of intensive agriculture (extending into north-western Germany). Across most measurement sites the models perform quite well in capturing the timing of mixing ratio anomalies, an important requirement for emission estimation. However, the signal of emissions in the mixing ratio time series is most evident in the amplitude of the observed variability. These amplitudes also vary between models, making the inversion-derived emission adjustments sensitive to transport model uncertainty. A logical next step would be to further investigate such uncertainties using simulations of ^{222}Rn (which is also proposed in the roadmap as an essential part of the benchmarking system). So far, however, the number of participants who provided information on radon is unfortunately too low for such an assessment.

The results obtained here from the intercomparisons emphasize the need for a robust evaluation of atmospheric transport inversion systems. Hence, a roadmap for setting up a community benchmarking system is presented. Since the posterior emissions estimates from atmospheric transport inversions cannot be directly validated, such a benchmarking system needs to include a) a range of different observations and b) the possibility of comparing results from different inversion systems to each other (i.e. a cross-comparison of inversion results).

2 Introduction

2.1 Background

To support EU countries in assessing their progress for reaching their targets agreed in the Paris Agreement, the European Commission is supporting the development of an anthropogenic CO$_2$ emissions monitoring and verification capacity. Such a capacity would deliver consistent and reliable information to support policy- and decision-making processes.

The CoCO2 Coordination and Support Action builds the prototype system for such a European monitoring and verification capacity. Is the continuation of an initiative to explore the development of a European system to monitor human activity related carbon dioxide (CO$_2$) emissions across the world – the CO$_2$ Human Emissions (CHE) project

The main objective of CoCO2 is to perform R&D activities identified as a need in the CHE project and strongly recommended by the European Commission's CO$_2$ monitoring Task Force. The activities shall sustain the development of a European capacity for monitoring anthropogenic CO$_2$ emissions. The activities will address all components of the system with the aim to have prototype systems at the required spatial scales ready by the end of the project as input for the foreseen Copernicus CO$_2$ service element.

The overall objective of WP5 is to improve the representation of inversion uncertainties, which are important not only for the uncertainty of the generated flux estimates, but also to determine the weight that different elements of information that are used should receive. Also, recent atmospheric tracer transport inverse modelling intercomparison activities, such as the EUROCOM project, focusing on the regional European CO$_2$ inversions, have shown that there is a broad range in the posterior estimates of the net European terrestrial CO$_2$ fluxes (Monteil et al., 2020).

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

The objective of Task 5.6 is to investigate ways to evaluate posterior estimates and their uncertainties by means of model inter-comparisons both for CO$_2$ and CH$_4$ as well as the preparation of a roadmap for a dedicated benchmarking system for atmospheric inversions.

2.2.2 Work performed in this deliverable

The WP5 team has performed transport inversion experiments for CO$_2$ and CH$_4$ and coordinated the intercomparison of these inversion experiments. In addition, the task has prepared a roadmap for setting up a benchmarking system for atmospheric inversions and illustrating the benchmarking concept with some prototype results.

The inversions systems participating in the intercomparison include:

- LUMIA (ULUND, both CO$_2$ and CH$_4$)
- CarboScope Regional (MPG, both CO$_2$ and CH$_4$)
- WRF-STILT (NIM, CH4)
- CIF-FLEXPART (NILU, CH4)
- CIF-CHIMERE (LSCE, CH4)
- NIES-FLEXPART (NIES, CH4)
- TM5 CT-Europe (FMI, CH4)
- ICONDA (EMPA, CH4)

The work on the CO_2 transport inversion intercomparison has already been published in Munassar et al. (2023). Therefore, only a brief summary is given here in Section 3 of this deliverable report and the full paper is included as an appendix (Section 7).

2.2.3 Deviations and counter measures

No deviations were encountered.

3 Assessment of uncertainties in European inversion of CO_2

3.1 Set-up of the intercomparison

The work performed here is based on the outcome of the regional atmospheric CO_2 transport inversion intercomparison project EUROCOM (Monteil et al., 2020). Within the EUROCOM project six state-of-the-art inversion systems provided posterior estimates of the net European terrestrial CO_2 fluxes. The ensemble of regional inversions suggest that European ecosystems are a relatively small sink of −0.09 PgC yr⁻¹ (ensemble mean posterior estimate of the 2006-2015 average net flux), however with values from individual inversion systems ranging from a net source of 0.28 PgC yr⁻¹ to a net sink of −0.41 PgC yr⁻¹. Overall, the range of posterior estimates from the inversions (0.7 Pg C yr−1) is narrower than that of the priors (1.06 Pg C yr⁻¹ compared to 0.7 PgC yr⁻¹) but the convergence of the regional inversions at this scale is not better than that obtained in state-of-the-art global inversions.

As part of the CoCO2 project we now calculated posterior fluxes from an ensemble based on two of the inversion systems that participated in EUROCOM to better quantify the sources of uncertainties in the posterior estimates. While in the EUROCOM project the intercomparison protocol was, on purpose, rather loose to explore the full range of reasonable posterior flux estimates, we defined a very strict protocol for the ensemble used here. In detail, we tested the impact of various components in the setup of regional atmospheric CO_2 inversion systems through a set of inversions that differ only by the component in question in the atmospheric transport inversion system. The components we looked at are the atmospheric transport models used for calculating both the foreground and the background response, as well as the inversion system, i.e., the specification of prior uncertainties. For this, we performed an ensemble of eight inversions using the two regional atmospheric transport inversion systems LUMIA (Monteil and Scholze, 2021) and CarboScope-Regional (CSR; Kountouris et al., 2018; Munassar et al., 2022). The ensemble of inversions consists of the following configurations (also shown in Table 1):

- In its default configuration, LUMIA relies on the FLEXPART Lagrangian particle dispersion model to compute CO_2 transport within Europe (foreground), with lateral boundary conditions taken from a TM5-4DVAR simulation. These boundary conditions are provided in the form of time series of "background" concentrations (i.e., far-field contributions), computed directly at the observation sites by the TM5 model, using the 2-step inversion scheme of Rödenbeck et al. (2009). The temporal shape of the prior terrestrial flux uncertainty is determined as a weekly uncertainty from the standard deviation of NEE. For the spatial domain, a Gaussian function of the spatial correlation decay is applied to the prior uncertainty structure with a spatial length scale of 500 km. These values are then scaled such that the prior flux uncertainty over the full domain of Europe is 0.47 PgC yr⁻¹. That corresponds to experiment LF5 in Table 1.

- CSR, by default uses the STILT Lagrangian particle dispersion model to compute the foreground with the background provided by TM3 from the global CarboScope following also the 2-step scheme. In CSR the same total European prior flux uncertainty of 0.47 PgC yr⁻¹ is used, however, this uncertainty is uniformly distributed

spatially and temporally over the domain using a hyperbolic function with correlation length scales chosen to be 66.4 km spatially and 30 d temporally, respectively. That corresponds to experiment CS3 in Table 1.

- In addition, LUMIA inversions were performed using the Lagrangian transport model STILT, as an alternative to FLEXPART, and using background concentrations from global (TM3-based) CarboScope inversions, as an alternative to TM5-4DVAR. And vice versa, CSR inversions were performed using FLEXPART to compute the foreground and using background concentrations from a TM5-4DVAR simulation.

- One additional experiment (not listed in Table 1) was performed with LUMIA using entirely the same specification as in CSR (STILT foreground, TM3 background and flat hyperbolic prior uncertainties) and compared against the default CSR set-up (CS3) to ensure that there were no other differences affecting the inversion results.

Table 1: List of the inversion setups for the CO₂ intercomparison with the two inversion systems LUMIA and CSR

Experiment	Inversion System	Transport model		Prior uncertainty	
		Foreground	Background	Shape	Decay
LF5	LUMIA	FLEXPART	TM5	Variable	Gaussian
LF3	LUMIA	FLEXPART	TM3	Variable	Gaussian
LS5	LUMIA	STILT	TM5	Variable	Gaussian
LS3	LUMIA	STILT	TM3	Variable	Gaussian
CS3	CSR	STILT	TM3	Flat	Hyperbolic
CS5	CSR	STILT	TM5	Flat	Hyperbolic
CF3	CSR	FLEXPART	TM3	Flat	Hyperbolic
CF5	CSR	FLEXPART	TM5	Flat	Hyperbolic

All simulations were done for the year 2018 using the same VPRM (Mahadevan et al., 2008) NEE prior terrestrial fluxes, optimized at a weekly resolution on a 0.25° grid, and using continuous observations from 45 tall-tower sites in Europe. Anthropogenic emissions are taken from the EDGAR v4.3 inventory (for the purpose of this comparison studies this version of EDGAR is fully sufficient) and are updated to recent years according to statistics from the energy company BP of fossil fuel consumption, and they are distributed spatially and temporally based on fuel type, category, and country-specific emissions, using the COFFEE approach (Steinbach et al., 2011). The emissions are remapped to a 0.25° spatial grid and to an hourly temporal resolution. Ocean fluxes are taken from Fletcher et al. (2007), who provide climatological fluxes at a spatial resolution of 5°×4°, remapped to 0.25° to be compatible with the biosphere model fluxes.

3.2 Results CO₂ fluxes

We first analyzed the impact of the various component changes in the inversion systems on the annual estimates of the posterior terrestrial biosphere flux over the whole domain. The spread among posterior estimates is relatively large, ranging between −0.72 and 0.20 PgC yr^{-1} (Figure 1b) with an average of −0.29 PgC yr^{-1} among the ensemble for the annual estimates. This range is larger than the assumed prior uncertainty of the terrestrial biosphere flux of 0.47 PgC yr^{-1}. This value for the prior uncertainty on the annual European terrestrial

flux has been derived from the global uncertainty (2.80 PgC) assumed in the CarboScope global inversion for the annual biogenic fluxes (Rödenbeck et al., 2003) On a monthly basis, the mean standard deviation of monthly posterior estimates over the ensemble of inversions is 0.72 PgC yr^{-1}. The largest deviations occur between inversions that differ by the foreground, i.e., regional transport models (e.g., CS3 versus CF3 or LS5 versus LF5). Also, the seasonal amplitude was found to be different between the STILT and FLEXPART inversions. The STILT-based inversions led to a larger amplitude of posterior NEE than the FLEXPART-based inversions (Figure 1a).
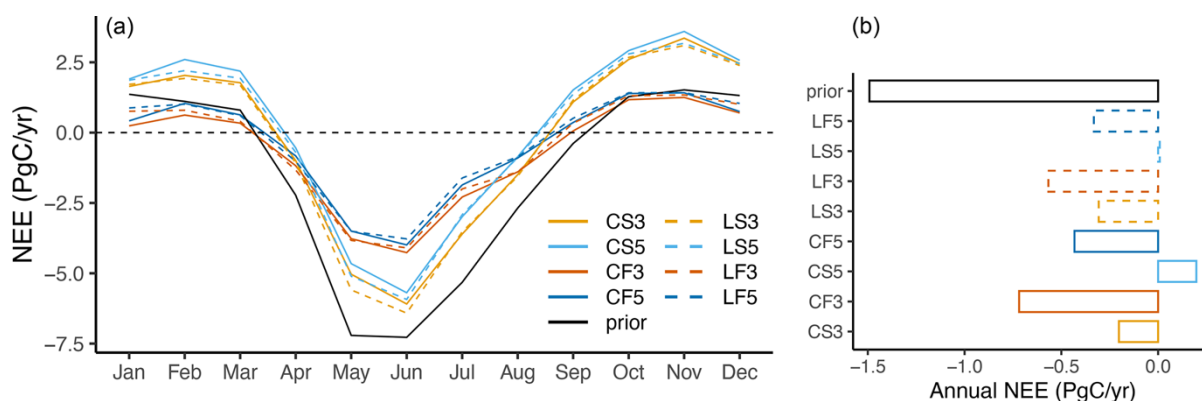
Figure 1: Panel (a) refers to posterior monthly NEE estimated using eight inversions, including prior NEE shown in black, with CSR (solid lines) and LUMIA (dashed lines), and panels (b) denotes the corresponding annually aggregated fluxes. Orange and red colours correspond to TM3, and dark or light blue correspond to TM5. Orange and light blue colours refer to STILT, and red and dark blue refer to FLEXPART.

When looking at domain wide averages the largest differences in posterior flux estimates result from the models used for calculating the foreground response (the regional transport models STILT and FLEXPART). The differences in monthly estimates of NEE calculated with CS3 and CF3 inversion setups that vary in regional transport models are shown in Figure 2 ("transport"). The differences caused by transport have a clear seasonal pattern: differences between CS3 and CF3 peak in November and June, reaching 2.11 and −1.82 PgC yr^{-1}, respectively. The best agreement between both inversions is obtained during the transitional months (August and April) with differences of −0.10 and −0.18 PgC yr^{-1}, respectively.
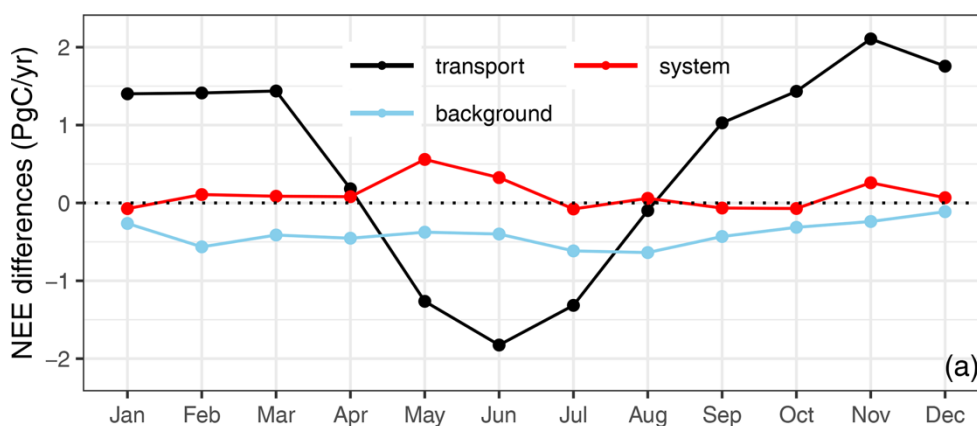
Figure 2: Differences in optimized fluxes calculated with the regional transport models STILT and FLEXPART ("transport"; CS3-CF3) and background provided through TM3 and TM5 ("background"; CS3-CS5). "system" refers to the differences between CSR and LUMIA inversion for optimized fluxes (CS5-LS5).

The second largest differences in posterior flux estimates result from the background, i.e., which model was used for calculating the lateral boundary conditions. The impact of using different background fields was analysed by assessing the differences in the posterior NEE estimated with CS3 and CS5 inversions, which use boundary conditions from the global inversions CarboScope and TM5-4DVAR, respectively. Figure 2 ("background") shows that this impact has rather the form of a bias (almost consistent differences over time) when aggregated over the entire domain of Europe: differences in the monthly posterior fluxes between CS3 and CS5 inversions amount to a range of 0.11 to 0.64 PgC yr^{-1} with the smallest differences occurring in the winter months. Larger flux corrections are modelled when using background fields from TM5 (CS5) than from TM3 (CS3).

The smallest differences in posterior flux estimates result from using different systems, here we compare CS5 and LS5 (Figure 2; "system"). The monthly differences between CS5 and LS5 range between 0.06 and 0.56 PgC yr^{-1}, the differences peak during May, June, and November, while the differences remained rather small during the rest of the year. Generally, LF5 predicts larger CO_2 releases compared to CS5.
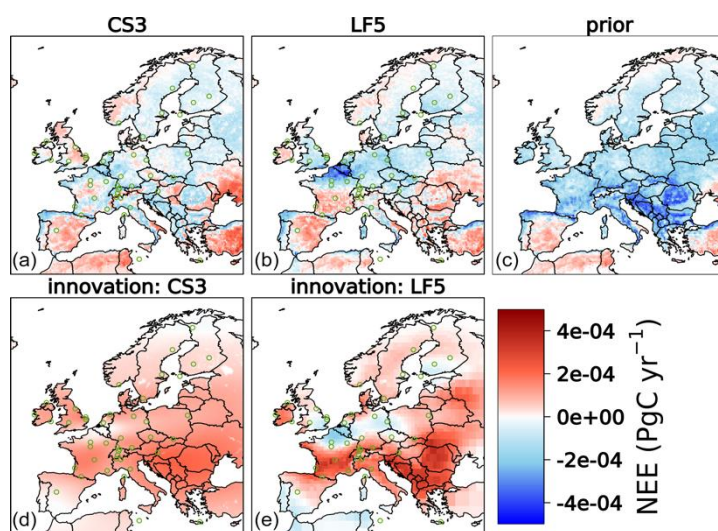
Figure 3: Panels (a)–(c) show the spatial distributions of annual NEE estimated with the base inversions CS3 and LF5, as well as their prior. Panels (d) and (e) depict the innovations of fluxes calculated for the inversions CS3 and LF5. Green circles denote the locations of observational sites.

In terms of spatial distributions, the base cases of CSR and LUMIA inversions, i.e., CS3 and LF5 (default configurations of both systems), exhibit a reasonable agreement in the annual terrestrial source/sink distribution over Europe (Figure 3). Major corrections compared to the prior fluxes are obtained over western and southern Europe suggesting an overestimation of the CO_2 uptake by the prior biogenic fluxes. The exceptionally dry summer in 2018 in Europe (Bastos et al., 2020) turned some areas in central, northern, and western Europe into a net source of CO_2. The discrepancies between CS3 and LF3 noticed in the innovations (that is the difference between posterior and prior), e.g., in northern France, the Netherlands, and south-eastern UK, are attributable to the combination of differences in regional transport models, lateral boundaries, and system configurations.
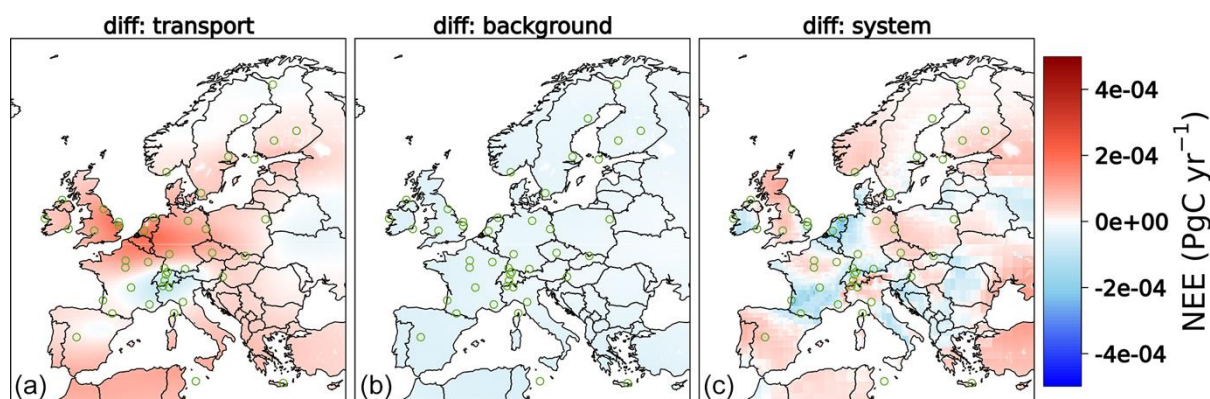
Figure 4: Panels (a)–(c) indicate differences in annual posterior NEE estimated with STILT and FLEXPART models, referred to as "transport" (CS3-CF3); TM3 and TM5 are referred to as "background" (CS3-CS5); and CSR and LUMIA are referred to as "system" (CF3-LF3).

In terms of the spatial differences in annual flux estimates, using STILT generally leads to predicting larger sources of CO_2 in the regional inversions, in particular over central Europe and the UK compared to using FLEXPART (Figure 4, "diff: transport"). In turn, inversions using FLEXPART suggest less uptake over northern Italy, Switzerland, and south-eastern France. However, this impact refers to a spatial pattern of transport differences that might be caused either by meteorological data or by problematic sites that transport models have difficulty representing.

The distributions of spatial differences in posterior fluxes caused by using different background fields indicate a homogeneous impact across the full domain of Europe (Figure 4, "diff: background"). These findings confirm the results obtained in Figure 2 ("background") for the temporal domain. This impact is consistent in space and time, with coherent deviation over all months, and is therefore not expected to affect the seasonal and interannual variability.

The spatial differences shown in Figure 4 "diff: system" alternate between positive and negative differences over the domain (but these tend to compensate when aggregating the flux estimates over the full domain). It should be noted that the inversion systems mainly differ in the definition of the shape and structure of the prior uncertainty. Therefore, applying different structure and magnitude of prior flux uncertainty in the inversions may inflate the error in CO_2 flux estimates over the underlying regions in the domain, in particular if the spatial differences do not cancel out. The spatial results indicate that the impact of inversion systems should not be neglected, especially at national and subnational scales.

3.3 Concluding remarks

In this section, we have presented an analysis of the impact of changing various components (foreground and background atmospheric transport model, and prior uncertainty specification as referred to as inversion system) in atmospheric tracer transport inversion systems on estimating net terrestrial CO_2 fluxes using two such systems (CSR and LUMIA) over Europe in 2018. The main focus here is to quantify the dominant drivers of uncertainties in the posterior CO_2 estimates derived from atmospheric tracer inversions. Two Lagrangian transport models (STILT and FLEXPART) were used to assess the impact of foreground (regional) transport on posterior CO_2 fluxes. Two Eulerian transport models (TM3 and TM5) were used to quantify the impact of the background (lateral boundary conditions) on posterior CO_2 fluxes, and finally two different schemes to set spatio-temporal prior uncertainties as employed by LUMIA and CSR were used to quantify the impact on posterior CO_2 fluxes. These variations lead to an ensemble of eight inversions. The results from this ensemble show a large spread in the annual terrestrial posterior fluxes over the whole domain of 0.92 PgC yr^{-1} ranging between

−0.72 and 0.20 PgC yr^{-1}, which is almost twice as large as the assumed prior uncertainty of 0.47 PgC yr^{-1}.

The choice of the regional transport model accounts for the largest part of the differences in the posterior fluxes. Temporal and spatial differences in posterior fluxes demonstrate a spatial pattern over certain areas during June and December, suggesting rather systematic differences between STILT and FLEXPART. The differences in the regional transport are mainly caused by the transport schemes (meteorological forcing data only partially contribute to these differences, not shown here). In addition, the global transport models used in the global inversions that provide the far-field contributions to the regional domain are responsible for small but non-negligible differences in the inversion estimates. These differences appeared to be homogeneous spatially and temporally, and hence can be considered as a bias. The differences arising from using different inversion systems integrated over the entire domain of Europe were in contrast rather small in comparison. However, such an impact is a result of applying different structure and shape in the prior flux uncertainty reflecting the importance of the way the uncertainty is prescribed in the tracer inversion systems.

The results obtained here emphasize the need for further evaluation of atmospheric transport models in order to improve the performance of the models and hence the inversion systems. A first aspect on this is to perform an intercomparison of regional atmospheric transport inversions for CH$_4$, which is believed to be more robust since there are mainly emission sources and no large sink terms for CH$_4$ as opposed to the net CO$_2$ flux which is the difference between two large gross fluxes. This topic is further discussed in the next section (Section 4) of this report. Ultimately, what is needed to reduce uncertainties and quantify the fidelity of inversion systems is an objective benchmarking system. A roadmap for setting up such a system is presented in Section 5 of this report.

4 Assessment of uncertainties in European inversion of CH$_4$

4.1 Introduction and background

The CH$_4$ inversion intercomparison builds on initial efforts in the H2020 project VERIFY to collect prior fluxes and surface measurements over Europe for use in the regional inversion systems developed in the project. We made use of WMO-IG^{3}IS/Transcom meetings to discuss an international intercomparison with the inverse modelling community based on the protocol developed in VERIFY. This gave other groups within and outside of Europe the opportunity to join the intercomparison, increasing the number of inverse models to 8. Support from COCO2 and WMO-IG^{3}IS was used to coordinate the experiment, update the protocol, make the input data package available to potential participants, collect submissions on a server, analyze the results, organize meetings with the participants to discuss the outcomes, and present the status of the experiment at international meetings.

The aim of the experiment is to investigate the use of the inverse modelling technique to support the national emission reporting to the UNFCCC by improving the consistency between the national emission inventories and atmospheric measurements. The case of CH$_4$ emissions over Europe was chosen, because of the good availability of measurements from the ICOS regional network and the relatively large uncertainty of anthropogenic CH$_4$ emissions (compared with CO$_2$), which atmospheric measurements may help reduce. The experiment focused on emissions trends starting in the year 2008, when the INGOS project started delivering data to ICOS, until 2018.

A second, equally important, aim of the intercomparison experiment is to develop benchmarking methods to evaluate and compare the performance of inverse modelling

systems and monitor their development in time. For this purpose, the protocol made an evaluation dataset available of the coordinates and sampling times of independent measurements that were not used in the inversions, for example because of incomplete measurement time series (making them less suitable for use in the inversion).

This section describes the experimental protocol (Section 4.2) and presents and discusses the results that were collected (Sections 4.3 and 4.4). The results section includes comparisons of inverse models and surface measurements that are used Section 5 to discuss a future benchmarking system.

4.2   Protocol and submissions

Participants are requested to optimize surface fluxes of methane using their inverse modelling systems for the European domain, covering at least the area of 15°W- 35°E and 35°N – 70°N. Inversion results should cover as many years as possible in the time range of 2005-2018. If continuous coverage of all those years is too computationally expensive then at least the years 2008, 2013, and 2018 are requested to be able to compare fluxes from as many inverse models as possible, including their trends. The requested output is in the form of gridded surface fluxes and country integrals for at least EU27 + UK.

The inverse modelers are requested to make use of a common set of prior fluxes, measurements, and regional domain boundary conditions (for regional models) in their inversion setups. However, it was decided not to prescribe prior or data uncertainties. For the surface fluxes, this is because the different resolutions of the models requires a regridding of emissions that would make the prior uncertainties at the model grid ambiguous without detailed information about spatiotemporal correlation of uncertainties, which inventories do not provide. For the measurements, estimates of measurement uncertainties were provided and requested to be used, however, the model / data representation error components of the data uncertainty are model dependent and were therefore not prescribed. Because of these rules, the intercomparison focuses primarily on impacts of transport model differences and differences in the optimization method and setup.

Table 2: Prior methane emissions used in the CH4 inversion intercomparison.

Category	Data source	Original Resolution		Time period
Peatlands, Mineral soils (emissions & uptake), inundated	JSBACH-HIMMELI[1]	0.1°x 0.1°	daily	2005-2020
Inland water[2]	ULB	0.1°x 0.1°	monthly	Climatology
Termites	Saunois, 2020	-	annually	Climatology
Ocean	Weber, 2021	0.25°x 0.25°	monthly	Climatology
Geological	Etiope, 2015	-	annually	Climatology
Fossil fuels	EDGAR v6.0	0.1°x 0.1°	monthly	2005-2018
Agriculture and waste	EDGAR v6.0	0.1°x 0.1°	monthly	2005-2018
Biofuels & biomass burning	GFED-4.1s	0.25°x 0.25°	monthly	2005-2020

[1]Covers Europe from 10.5°W to 33°E and 34.5°N to 73.5°N

[2]Covers Eurasia from 26°W to 55°E and 34°N to 78°N

Prior fluxes are made available as monthly gridded flux fields at 0.25ox0.25o resolution. Alternatively, fluxes are also provided at their native resolution (see Table 1), for inversions that can benefit from the higher resolution information offered by those datasets. Modelers are

free to represent the OH sink of methane in their European inversion, for which an OH field has been made available.

The VERIFY project delivered a harmonized set of CH4 measurements. This set contains observations from 66 surface sites from the InGOS project (2005- 2016), NOAA flask sampling sites in Europe (2005-2018), AGAGE, and additional data from European sites (ICOS, WDCGG and personal communications). The station list classifies sites as 'core' (39), 'other' (22), and 'validation' (5) (see Figure 5). Participants are asked to perform separate inversion using measurements from 'core' (39) sites and 'core + other' (61) sites. 'core' sites are those that deliver the most complete datasets throughout the whole inversion time window. The 'core + other' are used in an inversion that will be referred to as 'Exp 1'.

Participants who use regional models are requested to use a common set of initial and lateral boundary conditions from the CAMS v19r1 reanalysis, based on a global inversion using surface measurements at background sites. For groups that make use of the Rödenbeck method (Rödenbeck et al, 2009) consistent baseline concentrations have been made available.

All groups are asked to perform inversions using the 'core' and 'core+other' measurement datasets. To evaluate the transport model performance participants are asked to perform a forward run of ^{222}Rn, for which a climatological flux field has been made available based on Karstens et al (2015). After a first evaluation of results, discussed with the participating groups, it was decided to allow for one round of updated inversion for those who want to improve methodological issues that came up in the analysis and discussions.
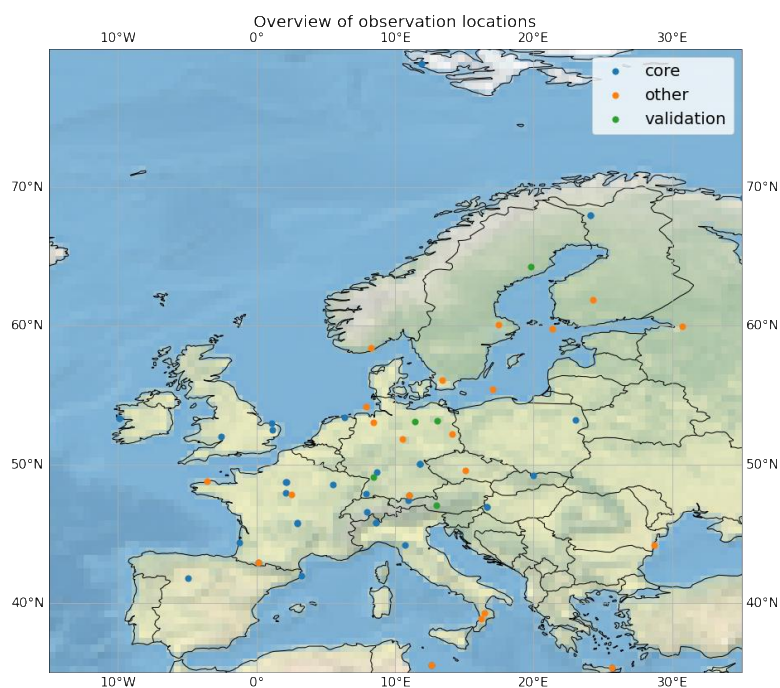
Figure 5: Map of CH$_4$ sites used in inversions and for validation

Since submissions are on a voluntary basis, some groups were only able to submit results that were incomplete on, for example, the validation dataset, posterior CH4 mixing ratios. Or ^{222}Rn abundances. Such submissions could still be used to assess methane fluxes, but not for performance evaluation and benchmarking. Table 3 provides an overview of participants, models, and the information that has been submitted. CIF-Chimere and CIF-Flexpart were performed within VERIFY, which explains why the validation dataset that was requested in the COCO2/IG[3]IS protocol and country integrals (using the country masks we provided) are not available. For these submissions country totals were reconstructed from the gridded fluxes.

For CarboScope an updated submission, which should make the missing mixing ratio output available, did not come in time for this report.

4.3 Results

First, we examine spatial patterns of surface flux adjustments from the gridded surface flux fields that were provided by all participants. The prior fluxes show a good agreement between the models, confirming a successful implementation of the protocol regarding these fluxes. Figure 6 compares annual mean differences between posterior and prior fluxes for the year 2013. The geometry of the observation network (Figure 5) suggests that the surface fluxes of Germany and the BeNeLux are resolved relatively well by the inversions. According to Figure 6, the flux adjustments for these countries are reasonably consistent between the inversion systems with emission increases over The Netherland and Northern Germany. In the remainder of Germany the flux adjustments are smaller and more ambiguous between the inversions at the southern border. This pattern is the least clear in the global NIES inversion, which shows smaller emission adjustments than other models and systematically lower across the European domain. Nevertheless, The Netherlands is among the few regions where upward flux adjustments are made in NIES.

Table 3: Participants and available datasets.

Group	Model	Gridded Fluxes	Country totals	CH$_4$ mixing ratios	Valid data	Exp.
Uni Lund	Lumia	🟩	🟩	🟩	🟩	Base/Exp1
MPI-Jena	CarboScope	🟩	🟩			Base/Exp1
FMI	CTE-CH$_4$	🟩	🟩	🟩	🟩	Base
NIES	NTFVAR2.0	🟩	🟩	🟩	🟩	Base
LSCE	CIF-Chimere	🟩		🟩		Base
NILU	CIF-Flexpart	🟩		🟩		Base
EMPA	ICONDA	🟩	🟩	🟩	🟩	Base/Exp 1
NIM	NTLB	🟩	🟩	🟩	🟩	Base/Exp 1

Italy shows emission adjustments that are downward on average, but the size of this emission reduction seems less well constrained by the data. Some inversions, such as CIF-Flexpart ('NILU') and Lumia show strong local emission adjustments in the north of Italy, which could point to difficulties representing the measurements at Mt. Cimone in these models. The most notable disagreement between the inversion optimized fluxes is for Ireland, where emissions are either increased (CSR), decreased (CIF-Flexpart, Iconda), or not adjusted much (the other inversions). This points most likely to the influence of the western lateral domain boundary condition that is used, with Ireland being the first country (for the predominantly westerly winds) where emission can be adjusted to improve the agreement with the most western measurement sites in the network. Similar disagreements are found for Spain, which is not well detected by the network. Emission adjustments in Eastern Europe are less prominent than in the west, which could be explained by the western European centre of weight of the ICOS network in combination with a predominantly westerly flow.

Despite some minor differences between different years (see Figure 8) the general pattern in Figure 6 is remarkably robust in time. Even different months of the year show a similar pattern as the annual mean.
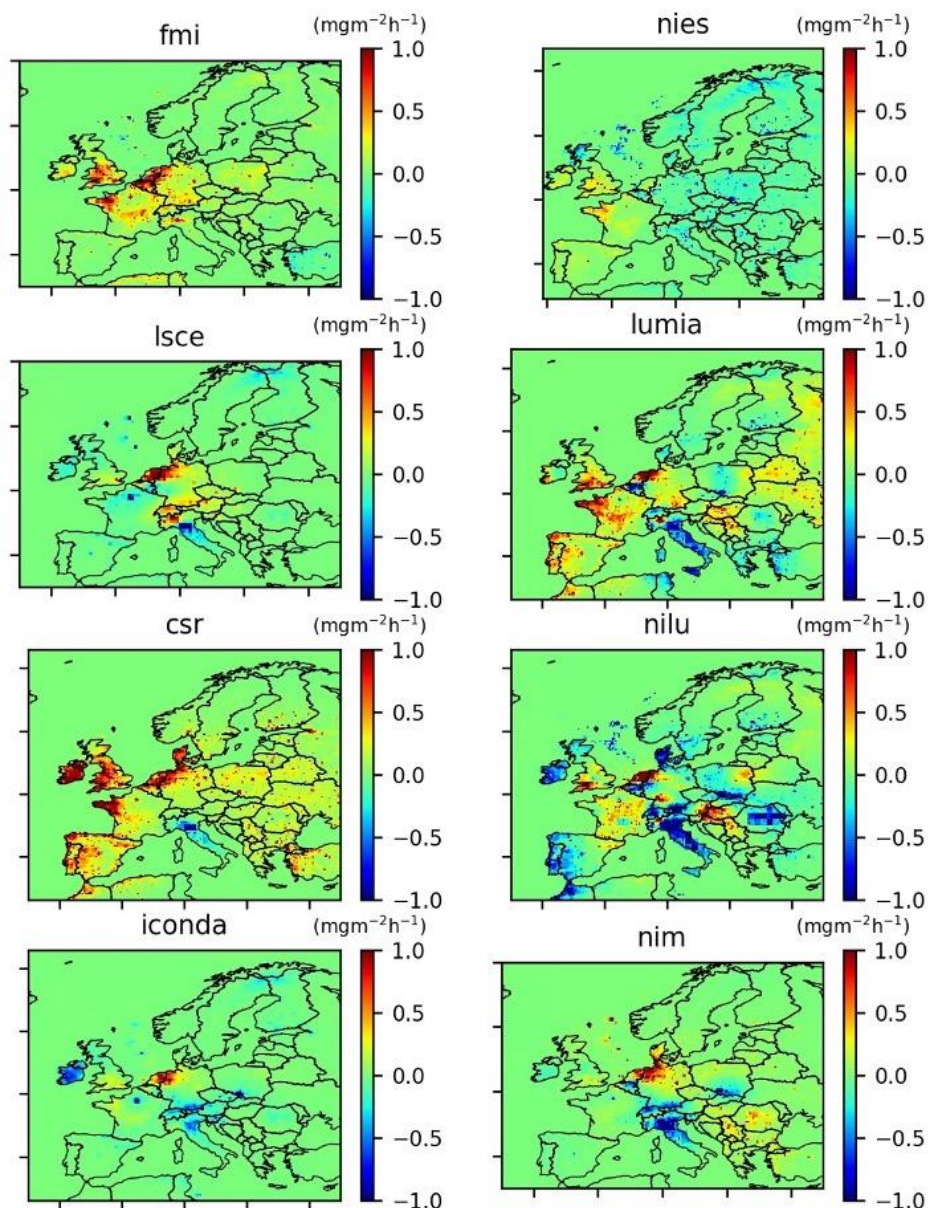
Figure 6: Annual mean surface flux adjustments (posterior – prior) for the year 2013 in kg/m²/h

The mean emission over 2006 – 2018 from the prior inventory is for most countries within the range of posterior flux estimates, suggesting that the inventory is consistent with atmospheric data within the uncertainty of the flux inversions. Exceptions are the aforementioned countries Italy and the Netherlands, and to a lesser extend also Hungary. As can be seen in Figure 7, estimates for The Netherlands and Hungary are significantly and systematically higher than the inventory, with the exception in both cases of the NIES model. Conversely, for Italy the inversion estimates are systematically lower than the inventory, again except for the NIES model. As discussed earlier the flux adjustments of the NIES inversion are smaller and smoother than for the other inversions. However, for some countries NIES does deviate significantly from the prior in a way that is not very different from the other models. Therefore, the NIES inversion cannot simply be disregarded as an outlier.
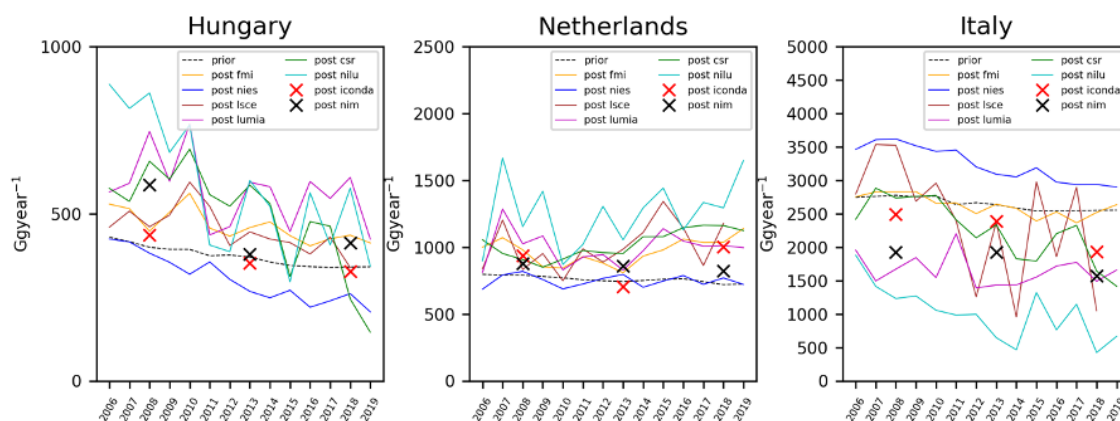
Figure 7: Comparison of posterior and prior CH$_4$ emission for selected countries.

To assess emission trends, it is useful to subtract the mean as shown in Figure 8. For this figure, countries have been chosen with relatively large methane emissions in the region that is expected to be best sampled by the ICOS network considering the network geometry and the importance of the west to east component of the wind. For the UK, Poland, and France the trend of the emission inventory is within the range of trends from the inversions. In the case of Italy (not shown), this range in trend and year-to-year variability is quite large among the different inversions and can therefore not be estimated robustly. For The Netherlands and Germany the inversions show a relatively consistent emission pattern with a decrease between 2006 and 2011, turning into an increase since then. The inventory estimates for these countries shows a steady decline in emissions. For Hungary the inversions show a trend that is on average somewhat larger than estimated by the inventory. However, the range is again large, with NILU's CIF-Flexpart showing the largest deviation from inventory, contrasted by FMI's CT-Europe being in reasonably agreement with the prior.
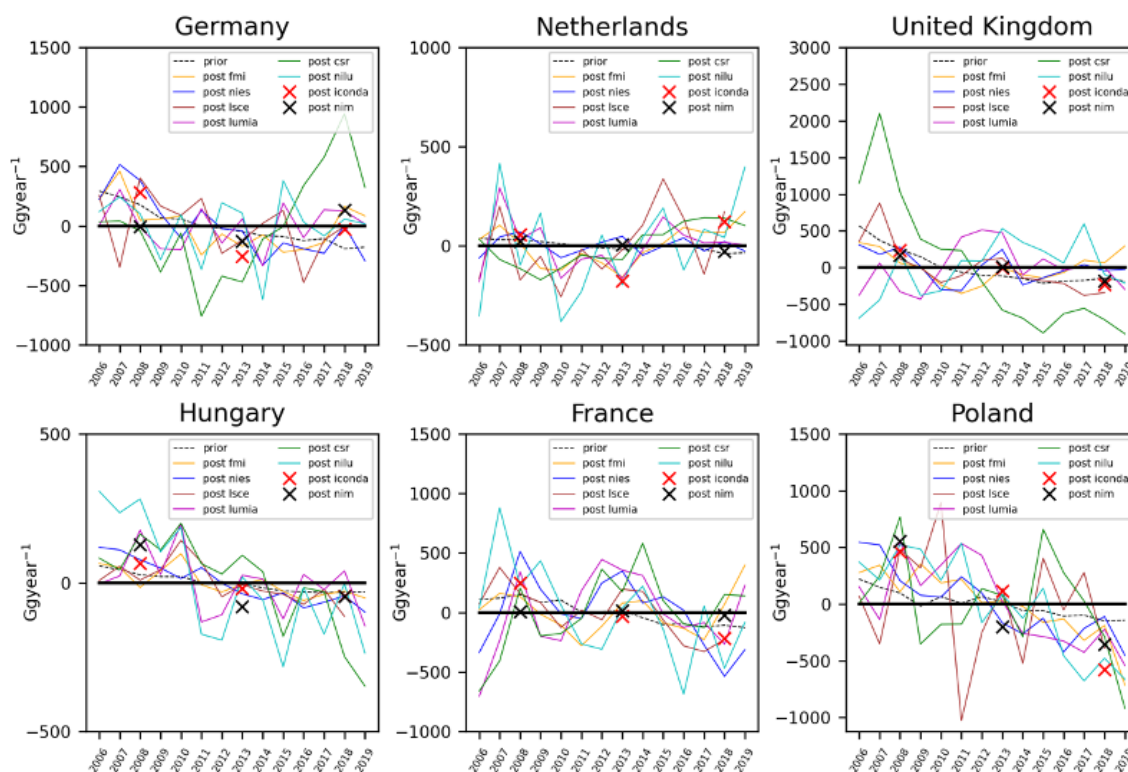
Figure 8: Comparison of posterior and prior CH$_4$ emission anomalies for selected countries.

Finally, to assess the performance of the inversions and the atmospheric transport models that are used, we have compared the prior and posterior fits to the observations. Figure 9

shows a few examples of such comparisons along a west to east transect from Biscarosse (France), Ochsenkopf (Germany), and Hegyhatsal (Hungary) for the year 2013. At Biscarosse, near the Atlantic coast of France, all inversions are quite well capable of capturing the observed variability already in the prior model setup. Notice that the night time measurements, which are not used in the inversions except for mountain sites, are left out of this comparison. Biscarosse is a core station in the protocol, meaning that these data have been used in the core inversion and are therefore not an independent evaluation of the inversions.

The emission adjustments in the run with posterior fluxes change the amplitudes rather than the phasing of the simulated variability. This may be expected as the short-term variability is largely dictated by the synoptic weather conditions, rather than the emissions. However, the amplitude of variability is also sensitive to the transport model that is used. For example, mixing ratios in July are underestimated by Lumia, but well simulated by CIF-Chimere (LSCE) and Iconda with the same prior fluxes. In December, Iconda overestimates the observations, while CIF-Chimere and Lumia are at the observed level. The peak in the beginning of March is either underestimated (CIF-Chimere) or overestimated (Iconda). These transport model differences are known to be sensitive to vertical mixing and the simulated height of the planetary boundary layer, which is critical for the emissions that are estimated by the inversions. As expected the posterior fits to the measurements are closer than the prior, although mismatches remain visible in these plots.

The picture at Biscarosse is not very different further inland as shown for the comparison at Ochsenkopf. Here the observed time series show more measurement gaps complicating the comparison. Nevertheless, a systematic overestimation using the prior fluxes is visible towards the end of the year in CIF-Flexpart (NILU), Lumia, and NIM. This is not reproduced by Iconda and CIF-Chimere, however, raising again the question of the role of transport model uncertainty. At Hegyhatsal, strong enhancements in methane are observed in the winter, again with strong difference in their representation by the Iconda and CIF-Chimere in this case. The inconsistency in the model – observation mismatches might well explain the large range of posterior emission adjustments discussed earlier.
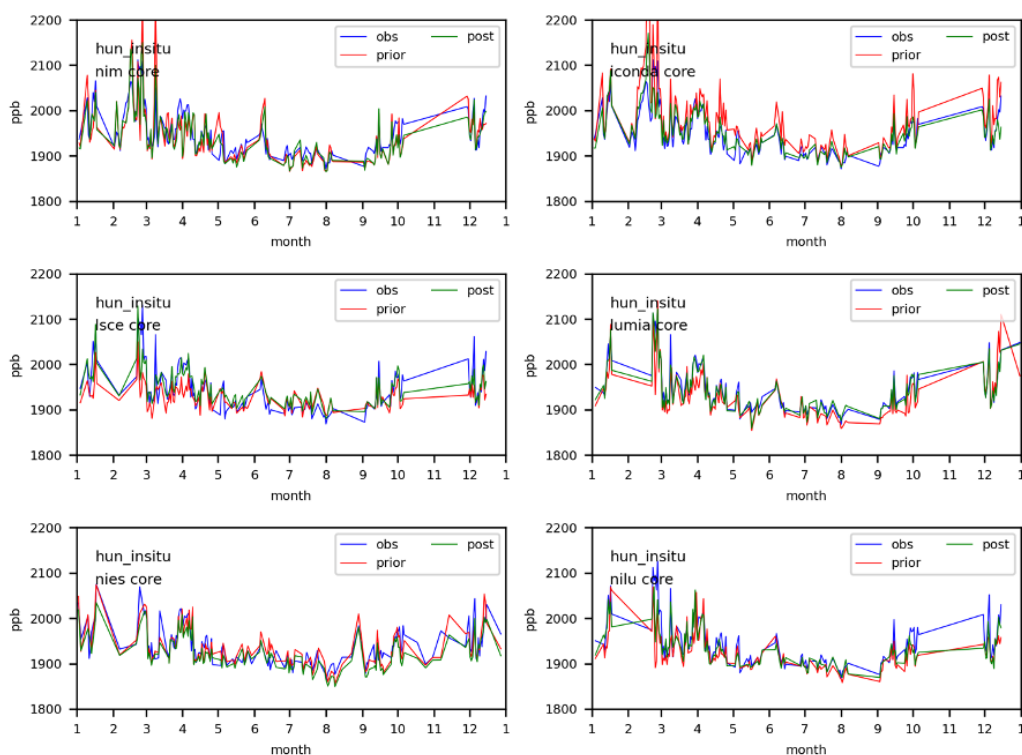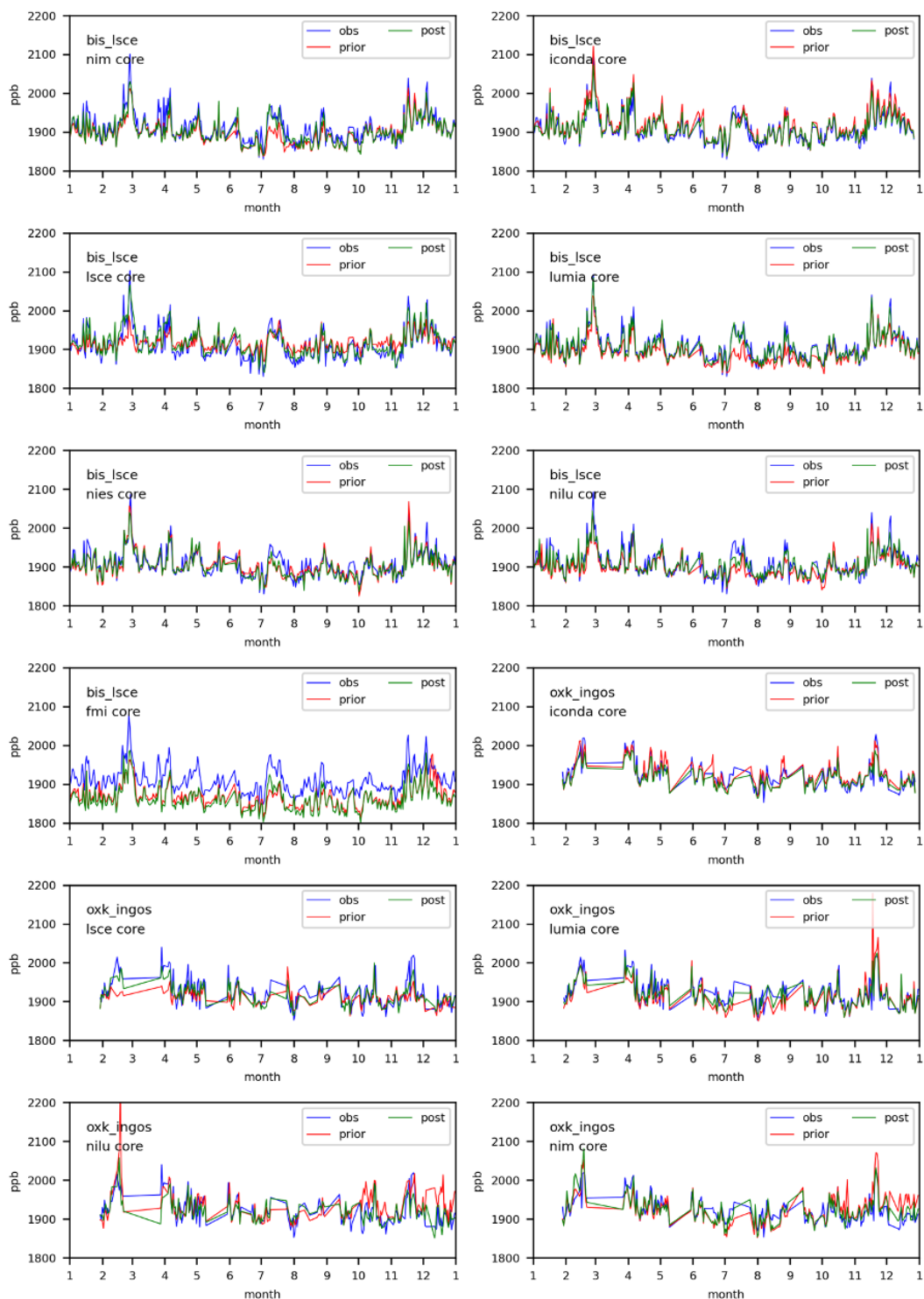
Figure 9: Modelled and measured CH$_4$ time series at selected sites comparing a priori and a posteriori model fits to the measurements.

Figure 10 summarizes the prior and posterior fits of the inversions to the measurements that are optimized in the inversion, showing the RMS error at each station for each model. The results for CT-Europe show RMS errors that are almost a factor two higher than the other models. The measured and modeled time series show systematic offsets (not shown) pointing to a remaining problem with the inversion that the authors have been contacted about. The other models show overlapping circles and crosses indicating similar fits of 'core' end 'exp1' inversions to the measurements. The inversions have difficulty fitting the site in Ispra (which

has a complicated flow pattern that is difficult to represent in the models), the mountain site Mt Cimone, and the high latitude site Zeppelin, which show very high RMS errors compared to other sites. The Lumia inversion system shows the lowest RMS of all models at many sites, starting from prior fits to the stations that are not much different from the other inversions.
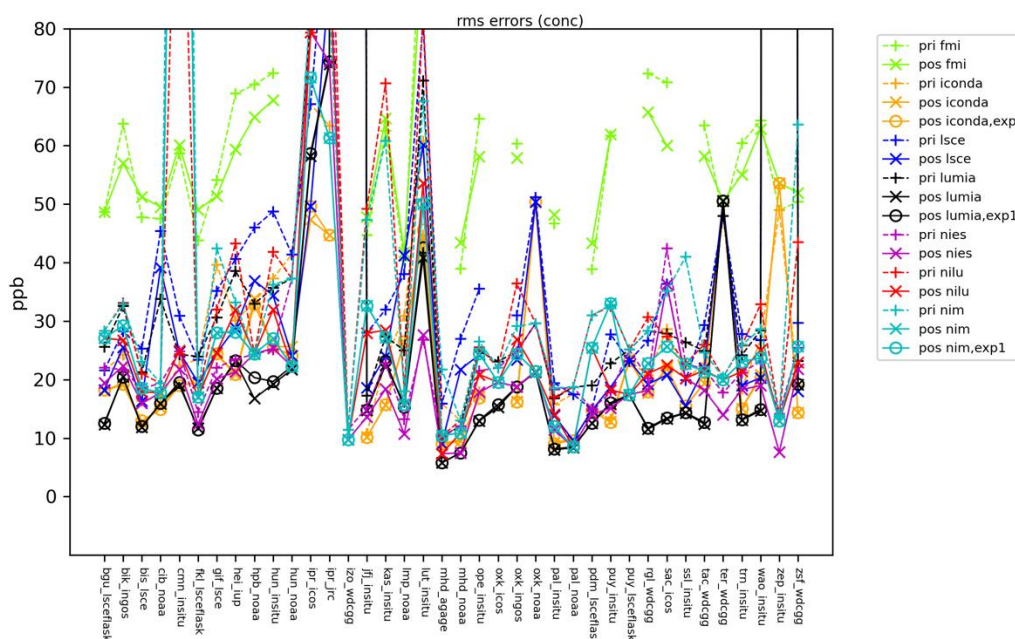
Figure 10: RMS error of the fit between CH4 inversions and measurements at each optimized measurement site.

To facilitate the comparison of prior and posterior RMS errors between models we averaged them over all sites (see Figure 11) except Izana (IZO) and Zeppelin (ZEP) (both sites are not inside the model domain of most of the regional models). The latter two were disregarded because of large residual errors (see Figure 10), most likely explained by their large distances from the European emissions that are optimized. Lumia has the lowest posterior RMS errors, followed by NIES and Iconda, which have the lowest prior RMS errors. CIF-Flexpart (NILU) and Lumia show the strongest error reductions between prior and posterior.
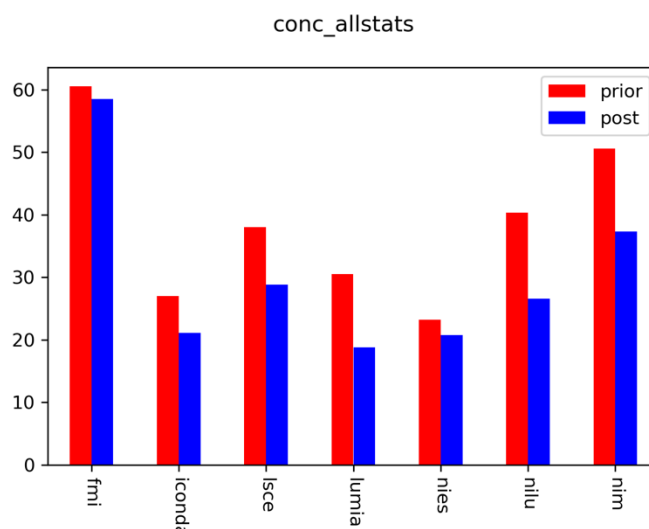
Figure 11: As Figure 9 averaged over all measurement sites.

4.4   Concluding Remarks

The inversion results for methane presented in this section highlight the challenge of estimating national greenhouse gases emission using a regional monitoring network, which is the case not only for carbon dioxide but also for methane. The models used in the intercomparison are quite well capable of capturing the timing of mixing ratio anomalies, an important requirement for emission estimation which is satisfied. However, the signal of emissions that is most evident in the mixing ratio time series is the amplitude of the observed variability, which is sensitive to transport model uncertainty. A logical next step would be to further investigate such uncertainties using simulations of ^{222}Rn. So far, however, the number of participants who provided information on radon is unfortunately too low for such an assessment.

The range in posterior fluxes obtained from the inversions is currently too large to provide a strong constraint on national emissions. The inventory estimates of mean emissions and emission trends are mostly within the range of the inversion results. So, it can be concluded that they are consistent with the atmospheric information. However, the limited constraint of the inversions on the inventory fluxes is more due to the quality of the inversions than intrinsic limitations in the information on fluxes provided by the measurements. Therefore, there is scope left to strengthen the observational constraint on the fluxes by improving the quality of the inversions. The only countries for which the CH$_4$ inversions deviate systematically from the prior are The Netherlands and Italy, where the inventory under- and overestimates the emission according to the inversions. In the case of The Netherlands, as well as for north-western Germany, this may include regions of intensive agriculture. Further research is needed to investigate if these differences can really be attributed to inventory uncertainties.

5 Roadmap for benchmarking atmospheric inversion systems

5.1   Conceptual framework

Following the results presented in the previous two sections showing large ranges in posterior fluxes among the various inverse modelling systems it is rather obvious that an objective and robust evaluation/benchmarking system will be instrumental for reducing the large spread in posterior estimates of either CO$_2$ or CH$_4$ emissions from regional atmospheric inversions. This also follows the fact that the inverse problem is usually underdetermined, because observational networks are sparse in coverage and hence resolving fluxes (spatially and temporally) at scales relevant for scientific or policy interest is not possible without prior information. In addition, the problem is ill-conditioned because diffusion in atmospheric transport can render rather small variations or errors in observed or modelled atmospheric concentrations to relatively large changes or errors in the posterior flux estimates. Hence, it is paramount to objectively evaluate the results of such inverse problems. Robust evaluation/benchmarking systems have already been used in the past for other modelling systems such as numerical weather prediction models (e.g. Rodwell et al., 2010, or also here 'ECMWF: Quality of Our Forecasts, http://www.ecmwf.int/en/forecasts/quality-our-forecasts (last access: 17 September 2023). More recently, such developments have also taken place for e.g., terrestrial biosphere models such as the International Land Model Benchmarking system (ILAMB; Collier et al., 2018) or also the CoCO2 modelevaluation.org instance (as part of the CoCO2 project and described in deliverable report D5.2 'Report on error assessment data base and toolbox for simulated terrestrial CO$_2$ fluxes') that provides a toolbox to assess errors of simulated terrestrial CO$_2$ fluxes (Nelson and Walther, 2022). For the evaluation of atmospheric inversions it is useful to distinguish between the quality and the skill of an inversion system. The former refers to whether an inversion system provides correct results ('it got it right') whereas the latter refers to the reasons for providing correct results ('why did it

get it right'). Here we describe a way forward for establishing a community benchmarking system for atmospheric greenhouse gas inverse modelling systems with a focus on assessing the quality of the system.

Main components of a community benchmarking system

Because direct observation of the major long-lived greenhouse gas (GhG, here CO_2 and CH_4) fluxes at the relevant spatiotemporal scales for evaluating posterior emission estimates from inversions are rarely available, any evaluation/benchmarking system must rely on indirect measures. The roadmap for a community benchmarking system laid out here consists of two main components: First, a diagnostic performance check of an individual inverse modelling system against a range of independent information as well as statistical diagnostics of inversion results. And second, a platform for comparison of inversion results from an individual inverse modelling system against results from ideally an ensemble of different inverse modelling systems. This should be seen as collaborative work among atmospheric CO_2 modelers to assess atmospheric GhG inversion models and help individual groups to study the performance of their system against an ensemble of systems. This ensemble should be based on a set of well-defined 'base inversions', such that any potential user can upload their (recent version of) inversion results and compare these against the ensemble.

A comprehensive assessment of available methods for diagnosing the quality of inverse modelling results for long-lived greenhouse gases for individual inverse modelling systems has been published by Michalak et al. (2017). They mainly distinguish between assessment against independent information, statistical diagnostics of inversion results, sensitivity tests and analysis of robustness, and synthetic data experiments as diagnostic methods. The first three are highly relevant here: the assessment against independent information and the statistical diagnostics are most important for individual systems, whereas the analysis of robustness can also be done as part of an ensemble described below.

As mentioned before the direct evaluation against independent information from surface fluxes is usually not possible because of the differences in scales. But posterior flux estimates can be evaluated against unused atmospheric observations, which are withheld from the inversion. These observations are additional data from in situ station measurements, aircraft profiles, ground-based remote sensing (such as the Total Carbon Column Observing Network, TCCON), or satellite remote sensing. A more detailed description of the various datasets is given in section 5.2. The evaluation is performed by an additional 'forward' transport simulation of the posterior fluxes translating the flux estimates into corresponding atmospheric concentrations (either at the ground-based measurement station, an atmospheric profile or the total atmospheric column). Such comparisons against various independent atmospheric observations provide information on whether the posterior flux estimates are consistent with the main constraints provided by the atmosphere, i.e. seasonal cycle, latitudinal gradients, or regional patterns of concentrations (e.g., Jiang et al., 2014; Díaz Isaac et al., 2014; Pandey et al., 2016; Liu and Bowman, 2016; Kountouris et al., 2018; Bergamaschi et al., 2018; Crowell et al., 2019). Typical metrics to compare against these time series data are bias and standard deviation.

Another way of benchmarking posterior flux estimates is the evaluation at larger scales based on independent information. This type of evaluation typically involves comparisons of the inversion-derived estimates against flux magnitudes at larger (e.g. continental to larger countries) scales assessing the direction and magnitude of the net flux in such a region (i.e., source versus sink, e.g. Berchet et al., 2013), flux seasonality (e.g. Houweling et al., 2014), or spatial and temporal gradients (e.g. Cressort et al., 2014). For assessing the direction and magnitude of the net flux, independent estimates based on either remotely sensed vegetation data (such as NDVI or FAPAR) or upscaling of direct measurements (such as the FLUXCOM product) can be used, however, these products carry their own uncertainties such that a quantitative evaluation is difficult. On continental scales and in regions with dense direct (eddy

covariance) flux measurement networks (such as Europe) the posterior flux estimates from an inversion can be compared directly against average quantities of the direct measurements. For example, Broquet et al. (2013) compared the temporal (monthly) evolution of European-wide flux estimates against the average flux (flux aggregates) calculated from continuous, quality controlled eddy covariance measurements from a large set of sites that are spread over the main regions and ecosystems of Europe. The data availability is discussed in Section 5.2. Relevant metrics for the direct comparisons are the amplitude of the seasonal cycle and the root mean square errors in the monthly/daily fluxes.

As outlined in Section 3 above model transport errors are the main contributors to the uncertainty in the posterior flux estimates. The best way to obtain an estimate of these errors, and benchmark atmospheric tracer transport models, is by assessing the transport model quality in a controlled tracer release experiment. However, such experiments are limited in their scale and therefore the estimated errors are not specific for certain locations and meteorological conditions. An alternative is a forward transport model simulation for a tracer, for which the fluxes are widespread and both fluxes and concentrations are relatively well known. Examples for this are the simulations of ^{222}Radon (^{222}Rn; Karstens et al., 2015) or SF$_6$ (Krol et al., 2018), however, for both tracers the emissions fields are not known well enough to provide robust estimates of the large scale transport errors. Another way to benchmark transport models is to analyse the quality of the vertical mixing. This can be done by comparing the simulated depth of the mixing layer against derived mixing heights from radiosonde measurements (Gerbig et al., 2008). Typical metrics for the evaluation are here again bias and standard deviation against the observations at measurement sites.

Another strategy for evaluating inversion results is based on assessing whether the prior and posterior flux estimates, their uncertainties and covariances are consistent with the assumptions that went into the setup inverse modelling system. This strategy then rather focuses on evaluating the statistical self-consistency of the inversion framework. This helps to identify errors and misconceptions in the setup that eventually can lead to unreliable results. A standard statistical diagnostic for inversions (assuming Gaussian, unbiased errors) is to test whether the sum of squared posterior errors follows a chi-squared distribution with a known number of degrees of freedom (e.g. Desroziers et al., 2005; Cressot et al., 2014). Such a chi-square test should be included in the benchmarking system to verify the self-consistency of a given inversion system.

The second component of the proposed community benchmarking system is a platform for comparing posterior emissions from an individual inverse modelling system against results from an ensemble of inverse modelling systems (including a group of different transport models and inversion methods) to understand the impact of transport model error and inversion method on estimated fluxes. An example for such a platform is the OCO-2 v10 MIP available at https://gml.noaa.gov/ccgg/OCO2_v10mip/ (last accessed 17 September 2023). In order to facilitate the comparison of posterior results and to be able to track the reasons for differences in the posterior flux estimates from the various inverse modelling system a strict pre-formulated experiment protocol for performing a 'base inversion' experiment including a well-defined common format for the submitted results is required. The details of a proposed protocol are given below. In short, the protocol needs to provide a list of mandatory measurement stations to be used in the inversions as well as prescribe prior flux fields to be used and any other flux components not to be optimised in the inversions (such as e.g., emissions from fossil fuel burning in the case of CO$_2$). Both CO$_2$ and CH$_4$ are included as inverted tracers in the community benchmarking system. Additional tracers, such as ^{222}Radon should also be included for evaluating the quality of the transport model component of the inverse modelling systems. In addition, the community benchmarking system needs to include advanced and easy-to-use collaborative visualization tools to facilitate the uptake of the comparison/benchmarking platform by the community.

The community benchmarking system should be hosted centrally at one place that can provide both storage space to store the data as well as collaborative tools to analyse and visualise the

data. The ICOS Carbon Portal facilities can provide both these aspects by making available storage space through their data servers, and even more importantly, by running a jupyter hub (https://jupyter.icos-cp.eu/) that can be accessed by the community and can host jupyter notebooks allowing collaborative analysis and visualisation of the submitted data.

Any submissions from individual inverse modelling groups to the benchmarking platform should include both posterior emission estimates and their uncertainties. Ideally, the uncertainties are provided on the same spatio-temporal resolution as the emission estimates. However, if an individual inverse modelling group cannot include posterior uncertainty estimates at the same resolution as the posterior emission estimates then, at least, aggregated uncertainty estimates on national and annual scale should be submitted.

The spatial focus of the community benchmarking system should be on a data rich region such as Europe as defined in the EUROCOM project (15° W–35° E by 33° N–73° N). But submission from global inverse modelling systems that include the European region in a sufficiently high spatial resolution (~1° by 1° in a regular longitude/latitude grid) should be allowed as well.

A key requirement for the benchmarking system is transparency. That means any group wanting to participate and submit results to the community benchmarking platform needs also to provide all the necessary meta data. This includes a full description of the components of their inversion system (i.e. transport model, inversion methodology as well as set-up of the system beyond any specifications given in the comparison/benchmarking protocol). In particular, the model setup needs to be described providing details on how the background contribution is treated in regional systems and how the prior uncertainty is specified.

To increase the numbers of participants anonymous initial submissions should be allowed and new users should be encouraged to upload their results themselves and test if their files comply with the requested format and fall within the results of the ensemble. Also, the public visualisation of the results from the comparison can be done such that individual inversion systems are anonymised.

A set of basic diagnostics needs to be performed with the ensemble of submissions. Among them are the ensemble mean (weighted evenly by model submissions, but at a later stage this can be complemented by a weighting according to specific model performance from individual benchmarks against independent data as described above), ensemble spread and standard deviation across the ensemble on top of the individual results.

Protocol

The starting point for a common protocol for the community comparison/benchmarking system is the protocol for the CH4 inverse modelling intercomparison described in Section 4. This protocol is provided in Appendix 8 as an example, which can also be used and revised accordingly for CO$_2$.

Potential submissions should focus on a few selected years to also allow the analysis of emission trends. The base year can be 2018 because this year has already been intensively studied (e.g. Monteil et al., 2020; Thompson et al., 2020; Munassar et al., 2023) and is also included in the CH4 intercomparison from Section 4. As mentioned before, any submission needs to also include uncertainty estimates on the posterior emissions. An important aspect that simplifies the collaborative analysis (by e.g. jupyter notebooks) of submitted results tremendously is to comply to a given output format as specified in the protocol.

The exact specifications in the protocol depend also on the initial participants and what they can provide and what they would like to see, but the protocol needs to include information on (see Appendix 8 for details):

- Spatial and temporal domain definition
- Prior fluxes to be used

- Emission fields for any flux components not to be optimised by the inversions (e.g. anthropogenic emissions for CO$_2$)
- Atmospheric observations to be assimilated for which concentrations need to be submitted
- Atmospheric observations withheld for validation but for which concentrations need to be submitted
- Boundary conditions to be used by regional inversion systems (either as lateral fluxes or as background fields)
- Uncertainties on both observations and prior fluxes
- ^{222}Rn emissions fields to be transported
- Specification of the output data
  - Gridded fluxes and uncertainties
  - Mixing rations (also for ^{222}Rn) and uncertainties
- Required documentation and meta data
  - Contact person
  - Version of the inversion system
  - Details about the inversion framework:
    - Definition of the state vector
    - Correlation length
    - Optimization method
    - Treatment of background
  - Details about the transport model
    - Resolution
    - Meteorological input data, and other relevant settings
    - Estimate of the transport model error as used in the inversion
  - Prior emissions and their uncertainties
    - Prior flux fields to compare priors as used in the model to priors specified in the protocol
    - Source (if other sources are used than given in the protocol and resolution)
    - Assumed uncertainty and covariance
  - Assimilated observations
    - Deviations from list in the protocol need to be explained
  - Chi square statistics of the inversion
  - Overview of the performed experiments and simulated years.

5.2 Required datasets

Table 4 provides an overview on the required datasets, their intended use and their availability. More information on observational datasets is provided in the deliverable reports from Work Package 7, in particular the gap analysis report D7.6 (Parampil et al., 2023).

Table 4: Datasets for the benchmarking system

Dataset	Usage	Source
Atmospheric in-situ observations (CO$_2$ and CH$_4$)	Assimilation and withheld observations for validation	ICOS Carbon Portal: https://www.icos-cp.eu/data-products/atmosphere-release European ObsPack: https://www.icos-cp.eu/data-products/PEKQ-M4T1

		https://www.icos-cp.eu/data-products/9CQ4-W69K
Satellite total column observations (CO$_2$ and CH$_4$)	Assimilation and withheld observations for validation	OCO2 CO$_2$: https://disc.gsfc.nasa.gov/datasets?keywords=oco-2&page=1 Sentinel 5P CH$_4$: https://sentinels.copernicus.eu/web/sentinel/data-products/-/asset_publisher/fp37fc19FN8F/content/tropomi-level-2-methane
Aircraft Profiles	Validation	North America: https://gml.noaa.gov/ccgg/aircraft/ CONTRAIL: https://cger.nies.go.jp/contrail/download/index.html IAGOS: https://iagos.aeris-data.fr
Aircore Profiles (CO$_2$ and CH$_4$)	Validation	Europe: https://aircore.aeris-data.fr/
Ground-based column observations	Validation	TCCON: https://tccondata.org COCCON: https://www.imk-asf.kit.edu/english/3884.php
^{222}Rn emission map atmospheric ^{222}Rn observations	Validation	ICOS Carbon Portal: https://doi.org/10.18160/2ST9-3NAD atm. ^{222}Rn available soon at selected ICOS stations
Radiosonde mixing heights Climatological mixing heights	Validation	Integrated Global Radiosonde Archive: https://www.ncei.noaa.gov/products/weather-balloon/integrated-global-radiosonde-archive Climatological mixing heights: Provided as a supplement to Seidel et al., 2012, https://doi.org/10.1029/2012JD018143
Direct eddy covariance (EC) flux observations	Validation	ICOS Carbon Portal: https://www.icos-cp.eu/data-products/ecosystem-release
Flux map from upscaled EC observations (FLUXCOM)	Validation	https://www.fluxcom.org/CF-Download/

5.3 Demonstration from test cases

An evaluation of CH$_4$ inversions at validation sites is shown in Figure 12. This is taken from the CH4 inversion intercomparison described in Section 4 where the participating modelling system are known. Except for some site-to-site variations, the results are fairly systematic across the sites of the validation dataset. The ranking of models by RMS-error is consistent with the prior and posterior fits to inverted measurement sites (as discussed earlier in Section 4.3), except NIES performing less well than Lumia and Iconda at validation sites compared to inversion optimized sites. Because of the low number of validation sites it is difficult to judge if this is caused by the independence of the validation data or by the poorer spatial representation of the validation dataset.
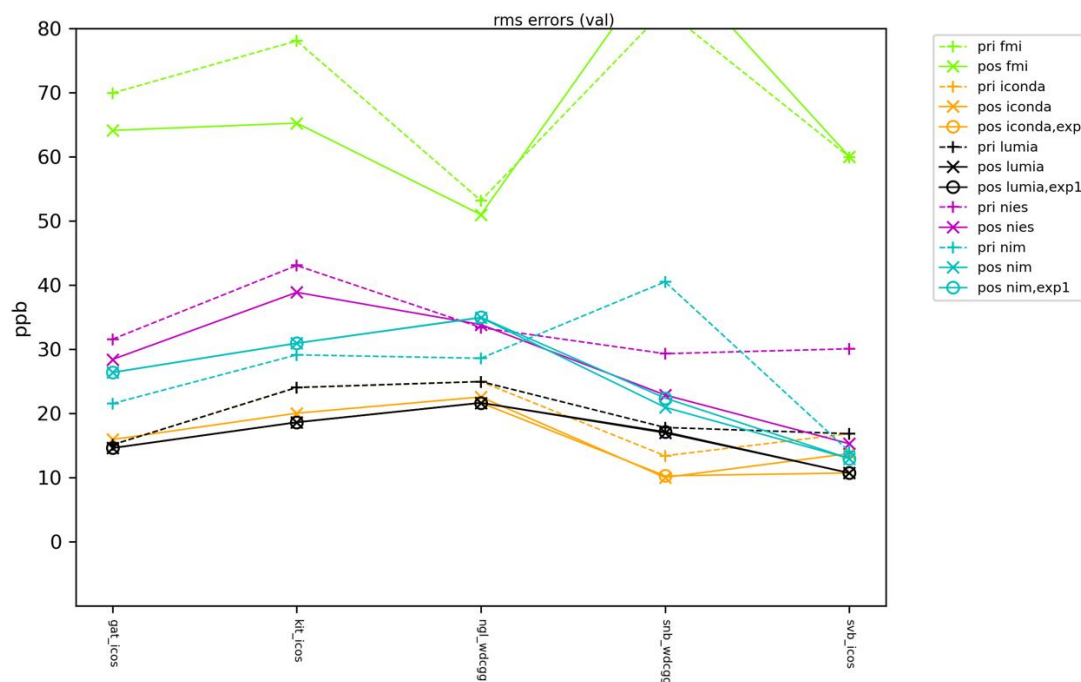
Figure 12: Comparison of RMS errors of CH$_4$ inversions at validation sites.

Averages of RMS errors across all validation sites (see Figure 13) highlight the differences between the inversions discussed above. As expected, the difference between prior and posterior models is less for the validation dataset than for the inverted dataset. For NIES the performance gain at validation sites as fraction of the prior RMS error is somewhat larger than for the other models. Because of the limited number of validation sites, however, it is difficult to judge how robust these averaged differences are. The optimization process improves the correlation between model and measurements at all validation sites, most notably for models that do not score so well using prior fluxes. The model bias improves for all models also. As seen in the timeseries already, the high RMS of FMI is largely explained by a bias, that does not improve much from prior to posterior. We are in contact with the data providers to check this outcome.

The validation would benefit from more sites, however, at the cost of the measurements available for emission estimation. A solution may be to perform more inversions, each reserving different sites for validation. This would of course make the inversions more computationally expensive, which has a trade-off with the resolution or the number of iterations or ensemble members that can be used in a single inversion.
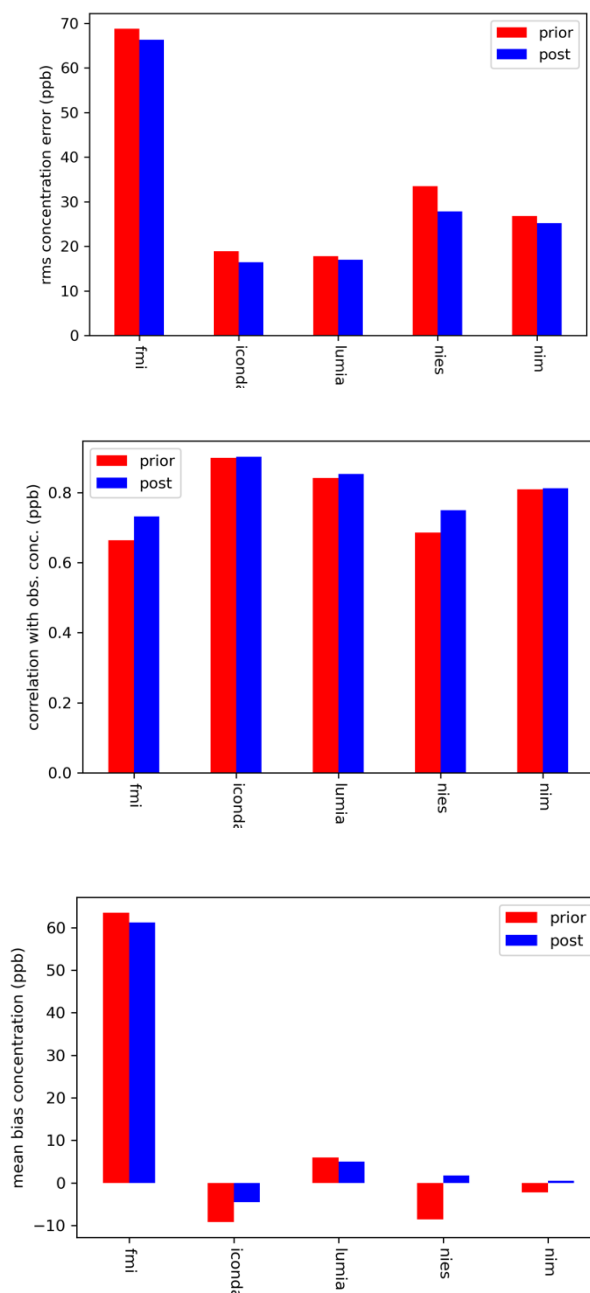
Figure 13: Comparison of the performance of CH$_4$ inversions at validation sites, comparing RMS (top), correlation (middle) and bias (bottom).

5.4 Concluding Remarks

The roadmap for benchmarking atmospheric inversion systems describes a community based benchmarking system based on two pillars: a) benchmarking individual inversion results against a range of different observations and b) comparing results from different inversion systems to each other (i.e. a cross-comparison of inversion results).

The first pillar constitutes a 'traditional' benchmarking system in which the results (posterior GhG fluxes) of an inversion system should ideally be compared against independent observations of GhG fluxes. This direct benchmarking against flux observations based on

eddy-covariance measurements is usually not possible because of the scale differences between the gridcell-based inversion estimates and point scale measurements. But the posterior flux estimates can be assessed against unused atmospheric observations, which are withheld from the inversion, however, such an evaluation against concentration data is not conclusive for evaluating the quality of an inversion system as the example test case in Section 5.3 shows.

The second pillar is based on a platform for comparison of inversion results from an individual inverse modelling system against results from an ensemble of different inverse modelling systems. Here, the idea is not to establish an intercomparison project but rather to provide a platform where the results of an individual inversion system can be assessed against the ensemble results and help individual groups to study the performance of their system against an ensemble of systems. This ensemble should be based on a set of well-defined 'base inversions', such that any potential user can upload their (recent version of) inversion results and compare these against the ensemble. For both pillars various performance metrics (such as bias, standard deviations, root mean square errors, amplitude of the seasonal cycle) are suggested.

6   References

Bastos, A., Ciais, P., Friedlingstein, P., Sitch, S., Pongratz, J., Fan, L., Wigneron, J. P., Weber, U., Reichstein, M., Fu, Z., Anthoni, P., Arneth, A., Haverd, V., Jain, A. K., Joetzjer, E., Knauer, J., Lienert, S., Loughran, T., McGuire, P. C., Tian, H., Viovy, N., and Zaehle, S.: Direct and seasonal legacy effects of the 2018 heat wave and drought on European ecosystem productivity, Sci. Adv., 6, eaba2724, https://doi.org/10.1126/sciadv.aba2724, 2020.

Bergamaschi, P., Karstens, U., Manning, A. J., Saunois, M., Tsuruta, A., Berchet, A., et al. (2018). Inverse modelling of European CH4 emissions during 2006-2012 using different inverse models and reassessed atmospheric observations. Atmospheric Chemistry and Physics, 18(2), 901– 920. https://doi.org/10.5194/acp-18-901-2018.

Berchet, A., Pison, I., Chevallier, F., Bousquet, P., Conil, S., Geever, M., Laurila, T., Lavric, J., Lopez, M., Moncrieff, J., Necki, J., Ramonet, M., Schmidt, M., Steinbacher, M., and Tarniewicz, J.: Towards better error statistics for atmospheric inversions of methane surface fluxes, Atmos. Chem. Phys., 13, 7115–7132, https://doi.org/10.5194/acp-13-7115-2013, 2013.

Broquet, G., Chevallier, F., Bréon, F.-M., Kadygrov, N., Alemanno, M., Apadula, F., Hammer, S., Haszpra, L., Meinhardt, F., Morguí, J. A., Necki, J., Piacentino, S., Ramonet, M., Schmidt, M., Thompson, R. L., Vermeulen, A. T., Yver, C., and Ciais, P.: Regional inversion of CO2 ecosystem fluxes from atmospheric measurements: reliability of the uncertainty estimates, Atmos. Chem. Phys., 13, 9039–9056, https://doi.org/10.5194/acp-13-9039-2013, 2013.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. Journal of Advances in Modeling Earth Systems, 10, 2731–2754. https://doi.org/10.1029/2018MS001354

Cressot, C., Chevallier, F., Bousquet, P., Crevoisier, C., Dlugo- kencky, E. J., Fortems-Cheiney, A., Frankenberg, C., Parker, R., Pison, I., Scheepmaker, R. A., Montzka, S. A.,

Krummel, P. B., Steele, L. P., and Langenfelds, R. L.: On the consistency between global and regional methane emissions inferred from SCIAMACHY, TANSO-FTS, IASI and surface measurements, Atmos. Chem. Phys., 14, 577–592, https://doi.org/10.5194/acp- 14-577- 2014, 2014.

Crowell, S., Baker, D., Schuh, A., Basu, S., Jacobson, A. R., Chevallier, F., Liu, J., Deng, F., Feng, L., McKain, K., Chatterjee, A., Miller, J. B., Stephens, B. B., Eldering, A., Crisp, D., Schimel, D., Nassar, R., O'Dell, C. W., Oda, T., Sweeney, C., Palmer, P. I., and Jones, D. B. A.: The 2015–2016 carbon cycle as seen from OCO-2 and the global in situ network, Atmos. Chem. Phys., 19, 9797–9831, https://doi.org/10.5194/acp-19-9797-2019, 2019.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, Q. J. Roy. Meteor. Soc., 131, 3385–3396, https://doi.org/10.1256/qj.05.108, 2005.

Díaz Isaac, L. I., Lauvaux, T., Davis, K. J., Miles, N. L., Richardson, S. J., Jacobson, A. R., and Andrews, A. E.: Model-data comparison of MCI field campaign atmospheric CO2 mole fractions, J. Geophys. Res.-Atmos., 119, 10536–10551, https://doi.org/10.1002/2014JD021593, 2014.

Fletcher, S. E. M., Gruber, N., Jacobson, A. R., Gloor, M., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Muller, S. A., and Sarmiento, J. L.: Inverse estimates of the oceanic sources and sinks of natural CO2 and the implied oceanic carbon transport, Global Biogeochem. Cy., 21, Gb1010, https://doi.org/10.1029/2006gb002751, 2007.

Gerbig, C., Körner, S., & Lin, J. C. (2008). Vertical mixing in atmospheric tracer transport models: Error characterization and propagation. Atmospheric Chemistry and Physics, 8(3), 591– 602. https://doi.org/10.5194/acp-8-591-2008.

Houweling, S., Krol, M., Bergamaschi, P., Frankenberg, C., Dlugokencky, E. J., Morino, I., Notholt, J., Sherlock, V., Wunch, D., Beck, V., Gerbig, C., Chen, H., Kort, E. A., Röckmann, T., and Aben, I.: A multi-year methane inversion using SCIAMACHY, accounting for systematic errors using TCCON measurements, Atmos. Chem. Phys., 14, 3991–4012, https://doi.org/10.5194/acp-14-3991-2014, 2014.

Jiang, F., Wang, H. M., Chen, J. M., Machida, T., Zhou, L. X., Ju, W. M., Matsueda, H., and Sawa, Y.: Carbon balance of China constrained by CONTRAIL aircraft CO2 measurements, Atmos. Chem. Phys., 14, 10133–10144, https://doi.org/10.5194/acp-14-10133-2014, 2014.

Karstens, U., Schwingshackl, C., Schmithüsen, D., & Levin, I. (2015). A process-based [222]radon flux map for Europe and its comparison to long-term observations. Atmospheric Chemistry and Physics, 15(22), 12845–12865. https://doi.org/10.5194/acp-15-12845-2015.

Kountouris, P., Gerbig, C., Rödenbeck, C., Karstens, U., Koch, T. F., and Heimann, M.: Technical Note: Atmospheric CO2 inversions on the mesoscale using data-driven prior uncertainties: methodology and system evaluation, Atmos. Chem. Phys., 18, 3027– 3045, https://doi.org/10.5194/acp-18-3027-2018, 2018.

Krol, M., de Bruine, M., Killaars, L., Ouwersloot, H., Pozzer, A., Yin, Y., Chevallier, F., Bousquet, P., Patra, P., Belikov, D., Maksyutov, S., Dhomse, S., Feng, W., and Chipperfield, M. P.: Age of air as a diagnostic for transport timescales in global models, Geosci. Model Dev., 11, 3109–3130, https://doi.org/10.5194/gmd-11-3109-2018, 2018.

Liu, J. and Bowman, K.: A method for independent vali- dation of surface fluxes from atmospheric inversion: Application to CO2, Geophys. Res. Lett., 43, 3502–3508, https://doi.org/10.1002/2016GL067828, 2016.

Mahadevan, P., Wofsy, S. C., Matross, D. M., Xiao, X. M., Dunn, A. L., Lin, J. C., Gerbig, C., Munger, J. W., Chow, V. Y., and Got- tlieb, E. W.: A satellite-based biosphere parameterization for net ecosystem CO2 exchange: Vegetation Photosynthesis and Res- piration Model (VPRM), Global Biogeochem. Cy., 22, Gb2005, https://doi.org/10.1029/2006gb002735, 2008.

Monteil, G., Broquet, G., Scholze, M., Lang, M., Karstens, U., Gerbig, C., Koch, F.-T., Smith, N. E., Thompson, R. L., Luijkx, I. T., White, E., Meesters, A., Ciais, P., Ganesan, A. L., Manning, A., Mischurow, M., Peters, W., Peylin, P., Tarniewicz, J., Rigby, M., Rödenbeck, C., Vermeulen, A., and Walton, E. M.: The regional European atmospheric transport inversion comparison, EUROCOM: first results on European-wide terrestrial carbon fluxes for the period 2006–2015, Atmos. Chem. Phys., 20, 12063–12091, https://doi.org/10.5194/acp-20-12063-2020, 2020.

Monteil, G. and Scholze, M.: Regional CO$_2$ inversions with LUMIA, the Lund University Modular Inversion Algorithm, v1.0, Geosci. Model Dev., 14, 3383–3406, https://doi.org/10.5194/gmd-14-3383-2021, 2021.

Munassar, S., Rödenbeck, C., Koch, F. T., Totsche, K. U., Gałkowski, M., Walther, S., and Gerbig, C.: Net ecosystem ex- change (NEE) estimates 2006–2019 over Europe from a pre-operational ensemble-inversion system, Atmos. Chem. Phys., 22, 7875–7892, https://doi.org/10.5194/acp-22-7875-2022, 2022.

Munassar, S., Monteil, G., Scholze, M., Karstens, U., Rödenbeck, C., Koch, F.-T., Totsche, K. U., and Gerbig, C.: Why do inverse models disagree? A case study with two European CO2 inversions, Atmos. Chem. Phys., 23, 2813–2828, https://doi.org/10.5194/acp-23-2813-2023, 2023.

Nelson, J. and Walther, S.: Toolbox to derive customized model forcing data and for assessing errors of simulated terrestrial CO$_2$ fluxes from data base of biogenic CO$_2$ flux measurements. Deliverable report D5.2, CoCO2 H2020 project, 16/12/2022.

Pandey, S., Houweling, S., Krol, M., Aben, I., Chevallier, F., Dlugokencky, E. J., Gatti, L. V., Gloor, E., Miller, J. B., Detmers, R., Machida, T., and Röckmann, T.: Inverse modeling of GOSAT-retrieved ratios of total column CH4 and CO2 for 2009 and 2010, Atmos. Chem. Phys., 16, 5043–5062, https://doi.org/10.5194/acp-16-5043-2016, 2016.

Parampil, S. R., Kasurinen, V. and E. Saltikoff, CoCO2 project Deliverable Report D7.6: Gap analysis report of the current in situ measurement capacity, 2023.

Rödenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: CO2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919–1964, https://doi.org/10.5194/acp-3-1919-2003, 2003.

Rödenbeck, C., Gerbig, C., Trusilova, K., and Heimann, M.: A two- step scheme for high-resolution regional atmospheric trace gas inversions based on independent models, Atmos. Chem. Phys., 9, 5331–5342, https://doi.org/10.5194/acp-9-5331-2009, 2009.

Rodwell, M.J., Richardson, D.S., Hewson, T.D. and Haiden, T. (2010), A new equitable score suitable for verifying precipitation in numerical weather prediction. Q.J.R. Meteorol. Soc., 136: 1344-1363. https://doi.org/10.1002/qj.656

Seidel, D. J., Zhang, Y., Beljaars, A., Golaz, J.-C., Jacobson, A. R., and Medeiros, B. (2012), Climatology of the planetary boundary layer over the continental United States and Europe, J. Geophys. Res., 117, D17106, doi:10.1029/2012JD018143.

Steinbach, J., Gerbig, C., Rödenbeck, C., Karstens, U., Minejima, C., and Mukai, H.: The CO2 release and Oxygen uptake from Fossil Fuel Emission Estimate (COFFEE) dataset: effects from varying oxidative ratios, Atmos. Chem. Phys., 11, 6855–6870, https://doi.org/10.5194/acp-11-6855-2011, 2011.

Thompson RL et al., 2020, Changes in net ecosystem exchange over Europe during the 2018 drought based on atmospheric observations. Phil. Trans. R. Soc. B 375: 20190512. http://dx.doi.org/10.1098/rstb.2019.0512

Document History

Version	Author(s)	Date	Changes
0.1	M. Scholze (ULUND)	25/08/2023	Initial draft
1.0	M. Scholze (ULUND), S. Houweling (VUA), U. Karstens (ICOS-CP), A.G.C.A. Meesters (VUA) & G. Monteil (ULUND)	08/10/2023	Complete draft for internal review
2.0	M. Scholze (ULUND), S. Houweling (VUA), U. Karstens (ICOS-CP), A.G.C.A. Meesters (VUA) & G. Monteil (ULUND)	08/11/2023	Revised version for submission

Internal Review History

Internal Reviewers	Date	Comments
Cicero (R. Andrew)	12/10/2023	
ENPC (J. Dumont Le Brazidec)	31/10/2023	

7 Appendix: Munassar et al., 2023, publications

Atmospheric
Chemistry
and Physics

Research article

Why do inverse models disagree? A case study with two European CO_2 inversions

Saqr Munassar[1,2], Guillaume Monteil[3], Marko Scholze[3], Ute Karstens[4], Christian Rödenbeck[1], Frank-Thomas Koch[1,5], Kai U. Totsche[6], and Christoph Gerbig[1]

[1]Department of Biogeochemical Signals, Max Planck Institute for Biogeochemistry, Jena, Germany
[2]Department of Physics, Faculty of Sciences, Ibb University, Ibb, Yemen
[3]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[4]ICOS Carbon Portal at Lund University, department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[5]Meteorological Observatory Hohenpeissenberg, Deutscher Wetterdienst, Hohenpeißenberg, Germany
[6]Institute of Geoscience, Friedrich Schiller University, Jena, Germany

Correspondence: Saqr Munassar (smunas@bgc-jena.mpg.de)

Abstract. We present an analysis of atmospheric transport impact on estimating CO_2 fluxes using two atmospheric inversion systems (CarboScope-Regional (CSR) and Lund University Modular Inversion Algorithm (LUMIA)) over Europe in 2018. The main focus of this study is to quantify the dominant drivers of spread amid CO_2 estimates derived from atmospheric tracer inversions. The Lagrangian transport models STILT (Stochastic Time-Inverted Lagrangian Transport) and FLEXPART (FLEXible PARTicle) were used to assess the impact of mesoscale transport. The impact of lateral boundary conditions for CO_2 was assessed by using two different estimates from the global inversion systems CarboScope (TM3) and TM5-4DVAR. CO_2 estimates calculated with an ensemble of eight inversions differing in the regional and global transport models, as well as the inversion systems, show a relatively large spread for the annual fluxes, ranging between -0.72 and $0.20\,\mathrm{PgC\,yr^{-1}}$, which is larger than the a priori uncertainty of $0.47\,\mathrm{PgC\,yr^{-1}}$. The discrepancies in annual budget are primarily caused by differences in the mesoscale transport model ($0.51\,\mathrm{PgC\,yr^{-1}}$), in comparison with 0.23 and $0.10\,\mathrm{PgC\,yr^{-1}}$ that resulted from the far-field contributions and the inversion systems, respectively. Additionally, varying the mesoscale transport caused large discrepancies in spatial and temporal patterns, while changing the lateral boundary conditions led to more homogeneous spatial and temporal impact. We further investigated the origin of the discrepancies between transport models. The meteorological forcing parameters (forecasts versus reanalysis obtained from ECMWF data products) used to drive the transport models are responsible for a small part of the differences in CO_2 estimates, but the largest impact seems to come from the transport model schemes. Although a good convergence in the differences between the inversion systems was achieved by applying a strict protocol of using identical prior fluxes and atmospheric datasets, there was a non-negligible impact arising from applying a different inversion system. Specifically, the choice of prior error structure accounted for a large part of system-to-system differences.

1 Introduction

Inverse modelling has been increasingly used to infer surface–atmosphere fluxes of carbon dioxide (CO$_2$) from observations of dry mole fractions made at spatio-temporal points across an observational network (Enting and Newsam, 1990; Bousquet et al., 1999). Reducing uncertainty in the flux estimates is, therefore, essential to reliably quantify the carbon budget (Friedlingstein et al., 2022; Le Quéré et al., 2018) as well as to improve our understanding about the variability and trends of the carbon cycle over times at finer regional scales, in particular in response to the climate perturbation caused by the increase in anthropogenic emissions (Shi et al., 2021). The estimates obtained from atmospheric tracer inversions still demonstrate large deviations due to manifold sources of uncertainty such as using different data, inversion schemes, and atmospheric transport models (Baker et al., 2006; Gurney et al., 2016), either at global scales or, to a larger extent, at regional scales. Although the global inversions can provide convergent estimations of the global carbon budgets, they are limited by the coarse resolution of atmospheric transport that may not allow for a realistic representation of the observations at complex mesoscale terrains. In turn, performing regional inversions with mesoscale transport models has offered a better opportunity to represent and make use of the dense measurements available at all the sites across regional domains (Broquet et al., 2013; Kountouris et al., 2018a; Lauvaux et al., 2016), specifically after the expanding coverage of data over large areas in recent years as has been established, for example, over Europe by the Integrated Carbon Observation System (ICOS). Although CO$_2$ fluxes constrained by atmospheric data in the Bayesian inversion framework inherit a dominant spatial and temporal pattern from the atmospheric signal, the a posteriori fluxes still suffer from a large spread when using different global and mesoscale transport models (Rivier et al., 2010).

As a first intercomparison between six regional inversions covering a wide range of system characteristics (e.g. prior fluxes, inversion approaches, and transport models), the EUROCOM experiment (Monteil et al., 2020) suggested large spreads in posterior estimates over Europe, particularly over regions that are poorly constrained by atmospheric data. This, on the one hand, partly indicates the sensitivity of the a posteriori estimates to the observations and to the a priori models as explained in Munassar et al. (2022). On the other hand, inaccuracies in atmospheric transport (Schuh et al., 2019), far-field contributions, and the configurations of inversions are responsible for part of that spread. A further study suggests that uncertainties in both transport and CO$_2$ fluxes contribute equally to the uncertainties in CO$_2$ dry mole fraction simulations, displaying similar temporal and spatial patterns (Chen et al., 2019).

The atmospheric transport relates the measured tracer concentration to its possible sources and sinks, which are adjusted in order to fit the modelled concentrations to observed data. However, inaccuracies in representing the real atmospheric dynamics by transport models lead to uncertainties in CO$_2$ flux estimates. This kind of error can emerge from both simplified parameterizations of real physics and model parameters themselves (Engelen et al., 2002). The atmospheric transport models rely on a mesoscale representation of air mass movements, which cannot completely reproduce the observed fine-scale variability of tracer concentration, leading to the so-called representation error. As a result, inversions cannot solve for fluxes at lower spatial and temporal resolutions than that of their transport model, resulting in aggregation errors (Kaminski et al., 2001). Additionally, atmospheric transport models are typically driven by meteorological data available from operational weather forecast models or reanalysis data optimized against observations and dynamical model forecasts. However, such meteorological fields have uncertainties owing to errors and gaps in the observations and errors in the weather forecast models (Deng et al., 2017; Liu et al., 2011; Tolk et al., 2008).

As the lateral boundaries are provided from a global model run at lower resolution than the regional model (Davies, 2014), this leads to biases in CO$_2$ lateral concentrations and thus affects the inversion estimates (Chen et al., 2019). The information of providing boundary conditions to regional inversions is necessary to isolate the influence of far-field contributions before performing the regional inversion. In Bayesian inversion setups, a proper information on prior error structures is also essential to determine the spatial pattern of the flux corrections based on the assumed error, especially at high spatial resolution inversions (Chevallier et al., 2012; Kountouris et al., 2015; Lauvaux et al., 2016). Therefore, the spatial pattern of flux corrections is dependent on the way the error covariance matrices are constructed, which can lead to large spatial discrepancies between the estimates from different inversion systems.

This study is dedicated to quantify the relative contributions of the differences in optimized fluxes resulting from varying as follows: (1) atmospheric transport models, (2) lateral boundary conditions, and (3) inversion configurations on flux estimates, as the error contributions from each component to the inversion's spread remain unclear in regional inversions, specifically at finer spatial scales over a continental domain such as Europe (Monteil et al., 2020; Petrescu et al., 2021; Thompson et al., 2020). We analysed results of a posteriori net ecosystem exchange (NEE) estimated from the two inversion systems CarboScope-Regional (CSR; Kountouris et al., 2018b; Munassar et al., 2022) and LUMIA (Monteil and Scholze, 2021). Both inversions employ pre-computed sensitivities of atmospheric mole fractions to surface fluxes, so-called source-weight functions or "footprints", via two Lagrangian transport models at regional scales, and they make use of the two-step inversion approach established by Rödenbeck et al. (2009) to provide the lateral boundary conditions. The regional atmospheric transport models were used at a horizontal resolution of 0.25°. The im-

pacts of both global and regional models were compared by analysing the differences in space and time.

Section 2 presents detailed descriptions of the inversion setups, the transport models, and the a priori fluxes used. The observational stations that provide CO$_2$ dry mole fraction data are described within the Methods section as well. We introduce the results obtained from eight inversions in Sect. 3. The results are discussed and interpreted through a spatial and temporal analysis of the differences between the elements of inversions in Sect. 4. Finally, Sect. 5 highlights a few concluding remarks on the impacts of regional transport, boundary conditions, and inversion setups on CO$_2$ estimates in the inverse modelling.

2   Methods

An atmospheric tracer inversion framework is mainly made up of transport model, data source for boundary conditions (in case of regional inversions), datasets of atmospheric mole fractions, and surface flux fields. In this study, several inversion runs differing in atmospheric transport models are conducted using two tracer inversion systems, CSR and LUMIA (see Table 2). The default CSR inversion system utilizes pre-calculated footprints from the Stochastic Time-Inverted Lagrangian Transport (STILT) model (Lin et al., 2003) at the regional domain and the TM3 model at the global scale, applying the two-step scheme inversion approach (Rödenbeck et al., 2009), to provide the far-field contributions to the regional domain. In the default setup of the inversion system LUMIA, the footprints are pre-calculated using the Lagrangian particle dispersion model FLEXPART (Pisso et al., 2019), and the far-field contributions are calculated using the global transport model TM5 in a separate global inversion run, applying the two-step scheme inversion as well. These default configurations in both systems constitute the base cases. We strive to restrict the differences in the inversion runs to the targeted components, i.e. regional transport, boundary conditions, and the inversion systems, so as to outline the impact of each suite. That is, input data such as measurements of CO$_2$ dry mole fraction and the a priori fluxes, used as constraints based on Bayesian inference, are identical for all runs. We exchangeably make use of the four combinations of transport model components, the regional and global models, in the two inversion systems. The impacts were evaluated using forward model runs to quantify the differences in CO$_2$ concentrations (simulated with prior fluxes) and inversion runs to quantify the magnitude of differences in the flux space. The inversion setups and implementation are explained in the comparison protocol (Sect. 2.6).

2.1   Inversion framework

In the following description, we remind the reader about the basic principles of the inversion schemes. For detailed information about the mathematical schemes, the reader

is referred to Rödenbeck (2005) for CSR and to Monteil and Scholze (2021) for LUMIA. Both systems rely on the Bayesian inference that accounts for observations and a priori knowledge to regularize the solution of the ill-posed inverse problem where a unique solution does not exist due to the spatial scarcity of observations. Therefore, the optimal state vector (x) is searched for in the Bayesian formalism by minimizing the cost function $J(x)$ that is typically composed of the observational constraint term $J_c(x)$ and the a priori flux constraint term $J_b(x)$:

$$J(x) = J_c(x) + J_b(x), \qquad (1)$$

where

$$J_b(x) = \frac{1}{2}(x - x_b)^T \mathbf{B}^{-1}(x - x_b), \qquad (2)$$

$$J_c(x) = \frac{1}{2}(H(x) - y)^T \mathbf{Q}^{-1}(H(x) - y). \qquad (3)$$

The a priori flux uncertainty defined in the covariance matrix **B** limits the departure of the control vector (x) to the a priori flux vector (x_b). Similarly, the observational constraint is weighted by the observational covariance matrix **Q** that contains the so-called model–data mismatch error, including uncertainty of measurement, representativeness, and transport. This uncertainty is assigned to the diagonal of the matrix **Q** for the respective sites based on the ability of the transport model to represent the atmospheric circulation at such locations. $H(x)$ represents the atmospheric transport operator (i.e. calculated by STILT and FLEXPART in our inversions) that determines the relation between fluxes and the modelled tracer concentration, which corresponds spatially and temporally to a given vector of measurements y. Following the gradient descent method, a variational algorithm is applied iteratively to reach the best convergence (global minimum) of the cost function that satisfies the optimal solution of the control vector. The default configurations for constructing the covariance matrices of a priori uncertainty are slightly different in CSR and LUMIA. A priori flux uncertainty is assumed to be around 0.47 PgC yr^{-1} over the full domain of Europe derived from the global uncertainty (2.80 PgC) assumed in the CarboScope global inversion for the annual biogenic fluxes (Rödenbeck et al., 2003). In CSR, this uncertainty is uniformly distributed spatially and temporally in a way that the annual uncertainty aggregated over the entire domain should arrive at the same value. The uncertainty structure is fit to a hyperbolic decay function in space (Eq. 4) and to an exponential function (Eq. 5) for the temporal decay as explained in Kountouris et al. (2015).

$$r(s) = \frac{1}{1 + \frac{s}{ds}}, \qquad (4)$$

$$r(t) = e^{\frac{-t}{dt}}. \qquad (5)$$

The correlation length scales ds and dt applied to flux uncertainties are chosen to be 66.4 km spatially and 30 d temporally, respectively, following Kountouris et al. (2018a) and

Munassar et al. (2022). The spatial length in the zonal direction is set to be longer than that in the meridional direction by a factor of 2 (anisotropic), owing to larger spatial climate variability in the meridional as compared to zonal direction.

The spatio-temporal shape of the a priori uncertainty in LUMIA is computed in a way that each control vector comprises weekly uncertainty calculated as the standard deviation of NEE based on weekly flux variance; however, LUMIA agrees on the overall annually aggregated flux uncertainty over the entire domain with CSR. A Gaussian function of the spatial correlation decay (Eq. 6) is applied to the a priori uncertainty structure with a spatial length scale of 500 km,

$$r(s) = e^{-\left(\frac{s}{L}\right)^2}, \tag{6}$$

whereas the effective temporal decay was set to 30 d (same as in CSR). Given the difference in the spatial correlation decay of the a priori uncertainty, LUMIA is set to draw larger flux corrections in a broader radial area where stations exist following the Gaussian decay with a longer length scale compared to the hyperbolic decay in CSR. In turn, the hyperbolic function has a larger impact in the further radial distances than the Gaussian function does, regardless of the longer spatial scale assumed with the Gaussian decay in a factor of around 7.5 in comparison with the hyperbolic decaying function.

2.2 Atmospheric transport models

Surface sensitivities are calculated using the STILT (Lin et al., 2003) and FLEXPART (Pisso et al., 2019) models at a horizontal resolution of 0.25° and hourly temporal resolution. Both models simulate the transport of air masses via releasing an ensemble of virtual particles at the locations of stations. The virtual particles are transported backward in time and driven by meteorological fields obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). STILT particles are transported 10 d backward in time and forced by forecasting data obtained from the high-resolution implementation of the Integrated Forecasting System (IFS HRES). For the FLEXPART model in standard operation, particles are followed for 15 d backward in time driven by ERA-5 reanalysis data. To keep the consistency with STILT footprints, the backward time of FLEXPART footprints was limited to 10 d in the inversions. After this backward time integration, the particles are assumed to leave the domain, even though a large number of particles are expected to escape after a few days. To better represent air sampling in the mixed layer, day-time observations are considered, except for mountain stations where night-time observations are used instead (Geels et al., 2007). To ensure best mixing conditions, temporal windows were considered for simulating CO₂ dry mole fractions over stations as explained in Sect. 2.4 (Table 1). In addition, release heights of particles are taken as the highest sampling level above ground

at each measurement site. For high-altitude receptors, such as mountains, a correction height is used in STILT in a way that the actual elevation of the station can be represented in the corresponding vertical model level (Munassar et al., 2022). In FLEXPART, the elevation above sea level is taken as the model sampling height.

2.3 A priori and prescribed fluxes

Three components of prior and prescribed surface-to-atmosphere fluxes of CO₂ are obtained from (1) biogenic terrestrial fluxes, (2) ocean fluxes, and (3) anthropogenic emissions and kept identical in both systems. Prior net terrestrial CO₂ exchange fluxes, net ecosystem exchange (NEE), are calculated using the diagnostic biogenic model Vegetation Photosynthesis and Respiration Model (VPRM) (Mahadevan et al., 2008). VPRM calculates NEE at hourly temporal and 0.25° spatial resolutions, and it provides a partitioning of the net flux into gross ecosystem exchange (GEE) and ecosystem respiration. Data obtained from remote sensing provided through the MODIS instrument and meteorological parameters from ECMWF drive both quantities of the light-dependent GEE and the light-independent ecosystem respiration. The model parameters were also optimized against eddy covariance data selected within the global FLUXNET site network across Europe in 2007 (Kountouris et al., 2015). For more details on the VPRM model, the reader is referred to Mahadevan et al. (2008).

Ocean fluxes are taken from Fletcher et al. (2007), who provide climatological fluxes at a spatial resolution of 5° × 4°, remapped to 0.25° to be compatible with the biosphere model fluxes. In addition, anthropogenic emissions are taken from the EDGAR_v4.3 inventory and are updated to recent years according to British Petroleum (BP) statistics of fossil fuel consumption, and they are distributed spatially and temporally based on fuel type, category, and country-specific emissions, using the COFFEE approach (Steinbach et al., 2011). The emissions are remapped to a 0.25° spatial grid and to an hourly temporal resolution.

Biogenic terrestrial fluxes are optimized in the inversions, while the ocean fluxes and anthropogenic emissions are prescribed, given the better knowledge about their spatial and temporal distribution in comparison with the heterogeneity, variability, and uncertainty of the biogenic fluxes. Moreover, in the absence of observational constraints that help discriminate the contributions from the three categories, we chose to prescribe the ocean fluxes and anthropogenic CO₂ emissions. This is also justified by the fact that the observation sites are located in areas where the biospheric flux influence is expected to dominate the variability of CO₂ concentration, but it means that errors in the fossil fuel or ocean fluxes might be compensated by the inversions, resulting in changes in the posterior NEE.

Table 1. Atmospheric sites used in the inversions.

Site code	Site name	Coordinates (lat, long)°	STILT release height (m a.g.l.)	FLEXPART release height (m a.g.l.)	Time window (UTC)	Uncertainty (ppm)
SM3	Hyytiälä	61.85, 24.29	125	125	10:00–14:00	1.5
BI5	Białystok	53.23, 23.03	300	300	10:00–14:00	1.5
FKL	Finokalia	35.34, 25.67	15	15	10:00–14:00	1.5
PAL	Pallas	67.97, 24.12	12	12	10:00–14:00	2.5
PUI	Puijo	62.91, 27.65	84	84	10:00–14:00	1.5
UTO	Utö, Baltic Sea	59.78, 21.37	57	57	10:00–14:00	1.5
BIR	Birkenes Observatory	58.389, 8.25	3	3	11:00–15:00	2.5
BR5	Beromünster	47.19, 8.17	212	212	11:00–15:00	1.5
DEC	Delta de l'Ebre	40.74, 0.79	10	10	11:00–15:00	1.5
EEC	El Estrecho	36.0586, −5.664	20	20	11:00–15:00	1.5
GIC	Sierra de Gredos	40.3457, −5.1755	20	20	11:00–15:00	2.5
HEI	Heidelberg	49.417, 8.674	30	30	11:00–15:00	4
HP4	Hohenpeissenberg	47.8011, 11.0246	300	131	11:00–15:00	1.5
ER2	Ersa	42.9692, 9.3801	40	40	11:00–15:00	1.5
HT3	Hyltemossa	56.0969, 13.4189	150	150	11:00–15:00	1.5
HU4	Hegyhátsál	46.95, 16.65	115	115	11:00–15:00	1.5
IP3	Ispra	45.8147, 8.636	100	100	11:00–15:00	1.5
KR3	Křešín u Pacova	49.572, 15.08	250	250	11:00–15:00	1.5
LMU	La Muela	41.5941, −1.1003	80	79	11:00–15:00	1.5
LMP	Lampedusa	35.53, 12.62	10	10	11:00–15:00	1.5
LUT	Lutjewad	53.4036, 6.3528	60	60	11:00–15:00	2.5
NO3	Norunda	60.0864, 17.4794	100	100	11:00–15:00	1.5
SV3	Svartberget	64.256, 19.775	150	150	11:00–15:00	1.5
TR4	Trainou	47.9647, 2.1125	180	180	11:00–15:00	1.5
OHP	Observatoire de Haute Provence	43.931, 5.712	100	100	11:00–15:00	1.5
SA3	Saclay	48.7227, 2.142	100	100	11:00–15:00	1.5
LHW	Laegern-Hochwacht	47.4822, 8.3973	400	32	11:00–15:00	2.5
BS3	Bilsdale	54.359, −1.15	248	248	12:00–16:00	1.5
RG2	Ridge Hill	51.9976, −2.54	90	90	12:00–16:00	1.5
TA3	Tacolneston	52.5177, 1.1386	185	185	12:00–16:00	1.5
WAO	Weybourne, Norfolk	52.9502, 1.1219	10	10	12:00–16:00	1.5
OP3	Observatoire pérenne de l'environnement	48.5619, 5.5036	120	120	14:00–17:00	1.5
GA5	Gartow	53.0657, 11.4429	341	341	14:00–18:00	1.5
LIN	Lindenberg	52.1663, 14.1226	98	98	14:00–18:00	1.5
BIS	Biscarrosse	44.3781, −1.2311	47	47	14:00–18:00	2.5
CRP	Carnsore Point	52.18, −6.37	14	14	14:00–18:00	1.5
MHD	Mace Head	53.3261, −9.9036	24	24	14:00–18:00	1.5
MLH	Malin Head	55.355, −7.333	47	47	14:00–18:00	1.5
JFJ	Jungfraujoch	46.5475, 7.9851	720	3570 (m a.s.l.)	23:00–03:00	1.5
KAS	Kasprowy Wierch	49.2325, 19.9818	480	1989 (m a.s.l.)	23:00–03:00	1.5
PUY	Puy de Dôme	45.7719, 2.9658	400	1465 (m a.s.l.)	23:00–03:00	1.5
SI2	Schauinsland	47.91, 7.91	450	1205 (m a.s.l.)	23:00–03:00	1.5
PTR	Plateau Rosa	45.94, 7.71	500	3480 (m a.s.l.)	23:00–03:00	1.5
PD2	Pic du Midi	42.9372, 0.1411	1458	2877 (m a.s.l.)	23:00–03:00	1.5
CMN	Monte Cimone	44.1963, 10.6999	670	2165 (m a.s.l.)	23:00–03:00	1.5

2.4 Observations

Measurements of CO$_2$ dry mole fractions are collected through ICOS, NOAA, and pre-ICOS stations across the domain of Europe provided by Drought 2018 Team and ICOS Atmosphere Thematic Centre (2020). In total, datasets from 44 stations are used covering the domain of Europe in 2018, in which a maximum number of stations is present compared to the other years. Regarding model–data mismatch errors, in LUMIA a weekly value of 1.5 ppm is assumed for all sites, except for the Heidelberg site where 4 ppm was assumed due

to the anthropogenic influence from the neighbourhood. Table 1 denotes the weekly values of uncertainty used in CSR for the corresponding sites. The uncertainty for the surface sites is inflated to 2.5 ppm as a slight difference to LUMIA. The inflation of uncertainty from weekly to hourly values is basically calculated by multiplying weekly errors by $\sqrt{7 \times n}$ (where n refers to the number of hours in the daily measurements used in the inversion). The observations are mostly assimilated as hourly continuous measurements and are taken from the highest level, avoiding large vertical gradients near the surface that are hard to represent in the transport mod-

Table 2. List of the inversion setups.

Inversion system	Transport model	Global boundary condition	Identifier code	Flux uncertainty	
				Shape	Decay
LUMIA	FLEXPART	TM5	LF5	Variable	Gaussian
LUMIA	FLEXPART	TM3	LF3	Variable	Gaussian
LUMIA	STILT	TM5	LS5	Variable	Gaussian
LUMIA	STILT	TM3	LS3	Variable	Gaussian
CSR	STILT	TM3	CS3	Flat	Hyperbolic
CSR	STILT	TM5	CS5	Flat	Hyperbolic
CSR	FLEXPART	TM3	CF3	Flat	Hyperbolic
CSR	FLEXPART	TM5	CF5	Flat	Hyperbolic

els. Model error in representing observations in the planetary boundary layer (PBL) is expected to be largest when the PBL is shallow. Therefore, for most sites, we considered data only when the PBL was expected to be well developed, i.e. during the afternoon, local time (LT). The exception is at high-altitude sites, which tend to sample the free troposphere during night (Kountouris et al., 2018b). The assimilated windows are reported in Table 1.

2.5 Boundary conditions

Far-field contributions of CO$_2$ concentrations (originating from sources outside of the regional domain) are taken from global inversions. As default setups of the global runs, the Eulerian transport model TM3 is used in the CarboScope global inversion at 5° (long) × 4° (lat), while TM5-4DVAR (Transport Model 5 – Four Dimensional Variational model) is used to provide boundary conditions to LUMIA using the global transport model TM5 at 6° (long) × 4° (lat) (Babenhauserheide et al., 2015; Monteil and Scholze, 2021). Both inversion systems apply the two-step scheme inversion, explained in Rödenbeck et al. (2009), in which a global inversion is first used to estimate CO$_2$ fluxes globally (based on observations inside and outside Europe). In a second step, the global transport model is used to estimate the influence of European CO$_2$ fluxes on European CO$_2$ observations. That regional influence is then subtracted from the total concentration to obtain a time series of the far-field influence directly at the locations of the observation sites. This prevents introducing biases by passing concentration fields from one model to another. For detailed information about the approach methodology, the reader is referred to Rödenbeck et al. (2009).

2.6 Comparison protocol

The results of the study are based on eight variants of inversions differing in global and regional transport models, as well as in inversion systems, as explained in Table 2. This implies that the two inversion systems (CSR and LUMIA) make use of two regional transport models (STILT and FLEX-

PART) and two global transport models (TM3 and TM5), which represent the boundary conditions (background) calculated from two global inversions. Hereafter, the identifier codes (see corresponding column in Table 2) will be used to refer to the individual runs within the inversion ensemble. For instance, to highlight the impact of regional transport models, we compare the inversions that only differ in regional transport models, regardless of the inversion system or boundary conditions used, such as CS3 and CF3 or LS5 and LF5. Similarly, we use the same specifications of transport models (indicated through the identifier codes) for the forward runs to outline the differences in CO$_2$ concentrations simulated using prior fluxes with different transport models. In this case, using a different system should not result in discrepancies as long as prior fluxes remain identical. In terms of system-to-system comparison, the impact of flux uncertainty should be taken into account as the prior error structure is specific for each inversion system. With that said, this has been investigated by conducting additional tests in CSR and LUMIA using identical uncertainties with flat shape and Gaussian correlation decay.

3 Results

Estimates of the regional biosphere–atmosphere fluxes over the domain of Europe are calculated using CSR and LUMIA for 2018 from an ensemble of eight inversions as listed in Table 2. Generally, all the inversions showed that the estimates of NEE are constrained by the atmospheric data as can be seen from the positive flux corrections made by the inversions in comparison with the a priori fluxes calculated from the biosphere flux model VPRM, which obviously overestimates CO$_2$ uptake, specifically during the growing season (Fig. 1a). This is also obvious in the ensemble-averaged annual estimates of posterior fluxes −0.29 PgC versus −1.49 PgC in the a priori fluxes (Fig. 1b). However, the spread among posterior estimates is still relatively large, ranging between −0.72 and 0.20 PgC yr^{-1} for the annual estimates, which is larger than the a priori uncertainty of 0.47 PgC yr^{-1}. Likewise, the mean standard deviations of

the monthly estimates over the ensemble of inversions is 0.72 PgC yr^{-1}. The largest deviations occur between inversions that differ by the regional transport models (e.g. CS3 versus CF3 or LS5 versus LF5). In addition, the seasonal amplitude was found to be different between the STILT and FLEXPART inversions. The STILT-based inversions led to a larger amplitude of posterior NEE than the FLEXPART-based inversions.

In terms of spatial distributions, the base cases of CSR and LUMIA inversions, i.e. CS3 and LF5 (default configurations of both systems), exhibit good agreement in predicting smaller uptake of CO$_2$ compared to the a priori fluxes (Fig. 2a–c). The magnitude of flux corrections suggests additional sources inferred from the atmospheric signal, as shown in the innovations of fluxes (Fig. 2d, e). Major corrections are obtained over western and southern Europe where the inversions point to an overestimation of the CO$_2$ uptake by the prior biogenic fluxes. The weak annual uptake of CO$_2$ in 2018 was exceptional and caused by the drought episode in Europe (Bastos et al., 2020; Rödenbeck et al., 2020; Thompson et al., 2020), which even turned some areas in central, northern, and western Europe into a net source of CO$_2$. The discrepancies between CS3 and LF3 noticed in the innovations, e.g. in northern France, the Netherlands, and south-eastern UK, are attributable to the combination of differences in regional transport models, lateral boundaries, and system configurations.

In the following, we will focus on separating and quantifying the contributions of such differences caused by each driver.

3.1 Impact of mesoscale transport

Inversions that differ in the regional transport models (STILT and FLEXPART) demonstrate the largest differences in posterior fluxes, resulting in a relative contribution of about 61 % of the total differences compared to the boundary conditions and inversion systems. The differences in monthly estimates of NEE calculated with CS3 and CF3 inversion setups that vary in regional transport models are shown in Fig. 3a ("transport"). Additionally, the discrepancies caused by transport have an obvious seasonal pattern. The differences between CS3 and CF3 peak in November and June, reaching 2.11 and −1.82 PgC yr^{-1}, respectively. The best agreement between both inversions is obtained during the transitional months (August and April) with differences of −0.10 and −0.18 PgC yr^{-1}, respectively. This might be attributed to the decline of the net flux magnitude during these months.

Furthermore, we assessed the impact of atmospheric transport in the simulations of CO$_2$ concentrations, because this directly translates into differences in the optimized fluxes. These simulations were calculated using the total components of prior fluxes (biosphere, ocean, and fossil fuel emissions) with STILT and FLEXPART in forward model runs to sample the atmospheric concentrations at hourly time steps at the station locations across the site network. Note that since all runs use identical prior fluxes, it does not matter for the differences whether the prior fluxes were precise enough to reproduce the true concentration or not. Figure 3b ("transport") illustrates the monthly differences in the forward simulations between STILT and FLEXPART averaged over all observational stations. Similarly to the discrepancies in the optimized fluxes, the differences in the forward simulations demonstrate a dominant impact of the regional transport model, preserving the same temporal pattern as seen in the flux differences but with opposite signs. The absolute difference ranges from 0.39 to 4.37 ppm when computed for the monthly means throughout all the sites. Geels et al. (2007) even found a larger spread up to 10 ppm when calculated with five transport models over 10 stations distributed across Europe. The notably large difference reported in that study is likely attributed to the large discrepancies in the model configurations, especially regarding the horizontal resolution and vertical levels used. The harmonized configurations used in STILT and FLEXPART lead to a reasonably consistent representation of the atmospheric variability at synoptic and diurnal timescales. The largest differences are observed during November and May with −4.37 and 3.60 ppm, respectively. On the other hand, the smallest differences were found to be −0.39, −0.42, and 0.56 ppm during September, April, and August, respectively. These results suggest a maximum impact of the mesoscale transport during the growing season and winter, while the impact converges to the minimum during transitional months such as May and September. Overall, the differences in posterior fluxes are consistent in the timing with the differences in the simulated concentrations computed using the prior fluxes.

Further diagnostics of model–data mismatches are provided in the Supplement, indicating the performances of STILT and FLEXPART with respect to the observations using prior and posterior fluxes across the site network at hourly, weekly, and yearly time steps (see Fig. 1S and Table 1S).

In terms of the spatial discrepancies in annual flux estimates, using STILT generally leads to predicting a larger sources of CO$_2$ in the regional inversions, in particular over central Europe and the UK compared to using FLEXPART (Fig. 4, "diff: transport"). In turn, inversions using FLEXPART suggest less uptake over northern Italy, Switzerland, and south-eastern France. However, this impact refers to a spatial pattern of transport differences that might be caused either by meteorological data or by problematic sites that are hard to represent by transport models. Some areas such as north-western Italy exhibit a persistent impact over time as shown in Fig. 4 ("SD: transport"), which shows the standard deviation of monthly differences calculated for the CS3 and CF3 inversions. In terms of temporal variations, the inversions performed with different regional transport models indicate larger monthly flux variations in comparison with
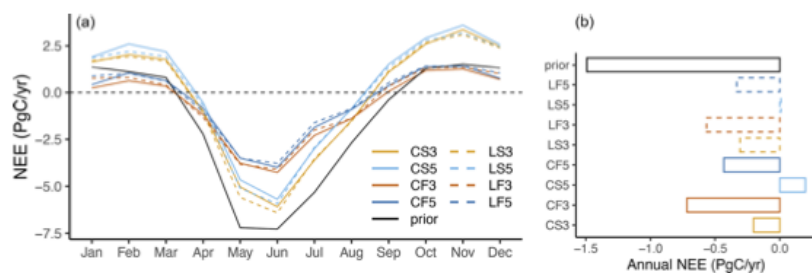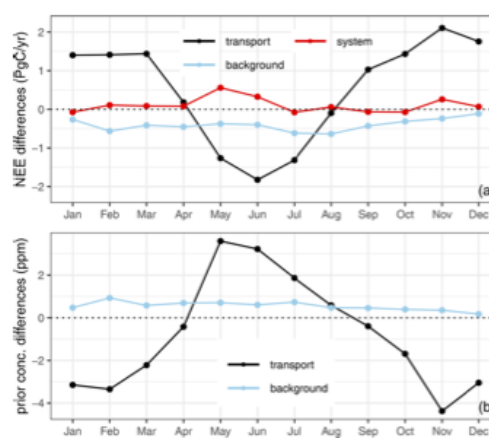
Figure 1. Panel **(a)** refers to a posteriori monthly NEE estimated using eight inversions, including a priori NEE shown in black, with CSR (solid lines) and LUMIA (dashed lines), and panels **(b)** denotes the corresponding annually aggregated fluxes. Orange and red colours correspond to TM3, and dark or light blue correspond to TM5. Orange and light blue colours refer to STILT, and red and dark blue refer to FLEXPART.
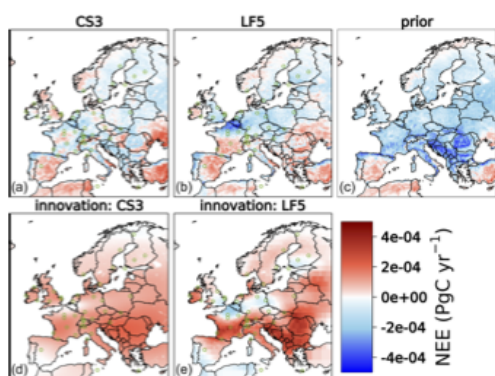
Figure 2. Panels **(a)**–**(c)** show the spatial distributions of annual NEE estimated with the base inversions CS3 and LF5, as well as their prior. Panels **(d)** and **(e)** depict the innovations of fluxes calculated for the inversions CS3 and LF5. Green circles denote the locations of observational sites.

Figure 3. Differences in optimized fluxes **(a)** and prior concentrations **(b)** calculated with the regional transport models STILT and FLEXPART (CS3-CF3) and background provided through TM3 and TM5 (CS3-CS5). "system" refers to the differences between CSR and LUMIA inversion for optimized fluxes (CS5-LS5).

those differing in global models and inversion systems (see Fig. 4, "SD: background" and "SD: system").

Figure 5 shows the spatial flux differences together with differences in prior concentrations simulated using STILT and FLEXPART during June and December. Note that the differences in NEE, to a large extent, agree in their spatial patterns with the differences in prior concentrations calculated over the station network. In addition, there are notably particular areas that exhibit opposite signs of the spatial impact in the differences in posterior fluxes and prior concentrations such as western Europe during June and northern Europe during December. One important difference between STILT and FLEXPART is that the STILT model has higher sensitivities during summer than FLEXPART, while the opposite holds true during winter. However, there are excep-

tions at individual sites such as Weybourne (WAO) in the UK and Ispra (IPR) in Italy, indicating either difficult terrains that cannot be well represented by the models or real synoptic features that are resolved by one model but not by the other. The differences in forward simulations are inversely manifested in the posterior flux differences as large surface sensitivities result in smaller posterior flux corrections and vice versa. In this case, STILT computes higher surface sensitivities than FLEXPART in June; therefore, the CS3 inversion needs to adjust the prior fluxes less to fit the observations. On the contrary, a weaker uptake is suggested by the STILT inversion during December over Europe, except for
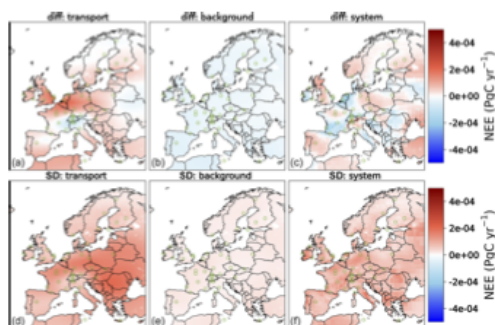
Figure 4. Panels **(a)–(c)** indicate differences in annual posterior NEE estimated with STILT and FLEXPART models, referred to as "transport" (CS3-CF3); TM3 and TM5 are referred to as "background" (CS3-CS5); and CSR and LUMIA are referred to as "system" (CF3-LF3). Panels **(d)–(f)** demonstrate the standard deviations of the corresponding monthly differences.
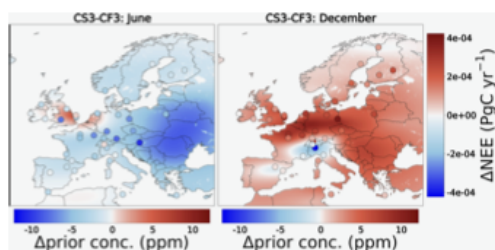
Figure 5. Spatial differences of posterior NEE estimated from the inversions CS3 and CF3 with STILT and FLEXPART transport models during June and December; filled circles indicate the differences in prior concentrations at the locations of sites (horizontal legend explains the magnitude of differences).

the abovementioned areas around northern Italy and southeastern France. The differences appeared to be larger during the months of growing season and winter, following the seasonal amplitude of CO_2.

3.2 Impact of lateral boundary conditions

The differences in lateral boundary conditions were found to account for about 27 % of the total differences resulting from the regional transport, lateral boundaries, and systems. This is a non-negligible contribution, albeit smaller than the regional transport contribution. The impact of using different far-field contributions was analysed by assessing the differences in the posterior NEE estimated with CS3 and CS5 inversions, which use boundary conditions from the global inversions CarboScope and TM5-4DVAR, respectively. Figure 3 ("background") shows consistent differences over time

between these inversion estimates aggregated over the entire domain of Europe. Larger flux corrections are suggested by CS5 than by CS3. This is because the global TM3-based inversion predicts higher influence at the lateral boundaries than the global TM5-based inversion does. Discrepancies in the monthly posterior fluxes between CS3 and CS5 inversions amount to a range of 0.11 to 0.64 $PgC\,yr^{-1}$ and absolute differences with a mean of 0.40 $PgC\,yr^{-1}$. Monthly-mean differences in CO_2 concentrations throughout all sites simulated using TM3 and TM5 boundary conditions were found to range from 0.17 to 0.93 ppm with a mean of 0.55 ppm.

The distributions of spatial differences of posterior fluxes indicate a homogeneous impact across the full domain of Europe (Fig. 4, "diff: background"). Likewise, the standard deviations of the monthly posterior fluxes obtained from CS3-CS5 ("SD: background") denote flat temporal variations throughout all the grid cells. These findings confirm the results obtained in Fig. 3 ("background"). This impact is consistent in space and time, with coherent deviation over all months, and is therefore expected to not affect the seasonal and interannual variability.

3.3 Impact of inversion systems

CS3 and LF5 differ by more than their regional transport and boundary conditions. In particular, the uncertainties are, by default, set up differently in CSR and LUMIA. The two systems optimize a different set of variables (weekly NEE offsets in LUMIA and 3-hourly NEE in CSR). Here we compare CS5 and LS5, which differ by their inversion systems but not by their transport model and boundary conditions. The differences in flux estimates between CS5 and LS5 inversions amount to 12 % relative to the total differences, including that caused by the mesoscale transport and lateral boundaries. This impact is, however, dependent upon system configurations, in particular the way the prior flux uncertainty is prescribed. The absolute monthly differences between CS5 and LS5 range between 0.06 and 0.56 $PgC\,yr^{-1}$ with a mean of 0.15 $PgC\,yr^{-1}$ (Fig. 3, "system"). This demonstrates the smallest differences amid inversions in comparison with the transport and lateral boundary differences, which yielded absolute monthly means of 1.27 and 0.40 $PgC\,yr^{-1}$, respectively. The differences peaked during May, June, and November, while the differences remained rather small during the rest of the year. LS5 infers −6.42 and 2.39 $PgC\,yr^{-1}$ during June and December, respectively, which is higher than CS5 estimates by 0.33 and 0.07 $PgC\,yr^{-1}$. Generally, LS5 predicts slightly larger CO_2 releases compared to CS5, which is partially due to differences in how uncertainties are assumed in both systems.

The impact of uncertainty definition is quantitatively assessed by using identical uncertainties for model–data mismatch as well as for prior fluxes in both CSR and LUMIA. The spatial flux corrections (innovation of fluxes) shown in

Fig. 8 denote quite good agreement between CSR and LU-MIA estimates. In this experiment, the differences in June and December decreased to 0.23 and 0.04 PgC yr^{-1}, respectively, in comparison with the corresponding differences obtained from the default configurations of both systems. That is to say, the impact of uncertainty definition alone amounts to 0.09 and 0.03 PgC yr^{-1} in June and December, respectively, leading to approximately 30 % and 50 % of the overall system-to-system differences. The rest of the differences may be attributed to differences in the convergence of the cost function to reach the minimum values.

The spatial differences shown in Fig. 4 "diff: system" alternate between positive and negative differences over the domain (but these tend to compensate when aggregating the flux estimates over the full domain). It should be noted that the inversion systems mainly differ in the definition of the shape and structure of the prior uncertainty. Therefore, applying different structure and magnitude of prior flux uncertainty in the inversions may inflate the error in CO$_2$ flux estimates over the underlying regions in the domain, in particular if the spatial differences do not cancel out. In addition, the corresponding standard deviations of monthly estimates ("SD: system") show large temporal variations, specifically over areas that have large spatial differences. The spatial results indicate that the impact of inversion systems should not be neglected, especially at national and subnational scales.

4 Discussion

The regional inversions computed over Europe showed that posterior NEE is largely derived from the atmospheric signal. The seasonality of posterior NEE, inferred from the atmospheric signal, is strongly impacted by differences in the representation of atmospheric transport. Given the identical priors and observational datasets used in the inversions, using different mesoscale transport models leads to 61 % of the differences in posterior fluxes in comparison with 27 % and 12 % of the differences caused by the use of different boundary conditions and different inversion systems, respectively. In agreement with these results, Schuh et al. (2019) also found a large impact of mesoscale transport on estimating CO$_2$ fluxes. Hence, any error in the atmospheric transport is translated into posterior fluxes as flux corrections. For instance, CS3 and LS3 suggest annual CO$_2$ flux budgets of -0.20 and -0.72 PgC, respectively, indicating a difference of 0.51 PgC in the annual flux budget. This difference is even larger than the prior flux uncertainty (0.47 PgC). The transport also showed a large impact on flux seasonality, leading to a difference of 49 % relative to the mean seasonal cycle. However, Schuh et al. (2019) found smaller differences, amounting to about 10 %–15 % of the mean seasonal cycle. Unlike the regional transport model error, the impact of boundary conditions does not show any striking seasonality and thus can be thought of as a bias in dry mole fractions.

The consistency of the lateral boundary impact over time and space is in agreement with results of lateral boundary uncertainties assessed by Chen et al. (2019) using four different global transport models, albeit over a different domain. Therefore, such an impact may be dealt with as a constant correction in mixing ratios before performing the regional inversions, which are potentially site-specific corrections. But there should be a reference for these corrections, e.g. taking the most robust model that has been validated against observations or simply a factor of the relative mean of the relevant models/approaches. Although the inversion systems showed the smallest differences in CO$_2$ flux estimates, the specification of the control vector (regarding the construction of covariance matrices) that devises the flux correction can result in larger differences, specifically in the spatial flux patterns.

The large number of stations within central and western Europe leads to a strong observational constraint that is reflected in the spatially optimized fluxes over that area. Therefore, large spatial differences between the inversions are pronounced around areas where stations exist, precisely for grid cells that have non-zero footprints. The large temporal variations indicate a systematic error that possibly arises from the transport models themselves as well as from meteorological forcing data. Additionally, systematic differences between transport models occur due to discrepancies in representing vertical mixing and horizontal and vertical resolution of the models (Peylin et al., 2002). Gerbig et al. (2008) found large discrepancies in derived mixing heights between meteorological analysis from ECMWF and radiosonde data, which reached about 40 % for the day-time and about 100 % for the nocturnal boundary layer. The vertical mixing in tracer dispersion models was found to result in a significant variability in methane emission estimations (up to a factor of 3) given the same meteorology as investigated by Karion et al. (2019).

Drivers of STILT–FLEXPART differences

Although STILT and FLEXPART are run at the same spatio-temporal resolution, employing similar schemes to parametrize the atmospheric motion unresolved by meteorological forcing data such as turbulence, and similar diagnostics to determine mixing heights, they still exhibit large spatial and temporal differences. A first assumption was that the differences between STILT and FLEXPART could be caused by differences in the calculation of mixing height. However, we did not find a correlation between the differences in mixing heights, calculated with the two models, and the differences in prior concentrations (Fig. 6). This finding concludes that the discrepancies in representing mixed layer heights do not explain the major differences in simulated CO$_2$ concentrations nor the differences in footprints.

The second assumption was that differences in the forcing data of meteorological products might lead to the discrepancies in both models, given that STILT uses meteorological parameters from IFS HRES, while FLEXPART uses the
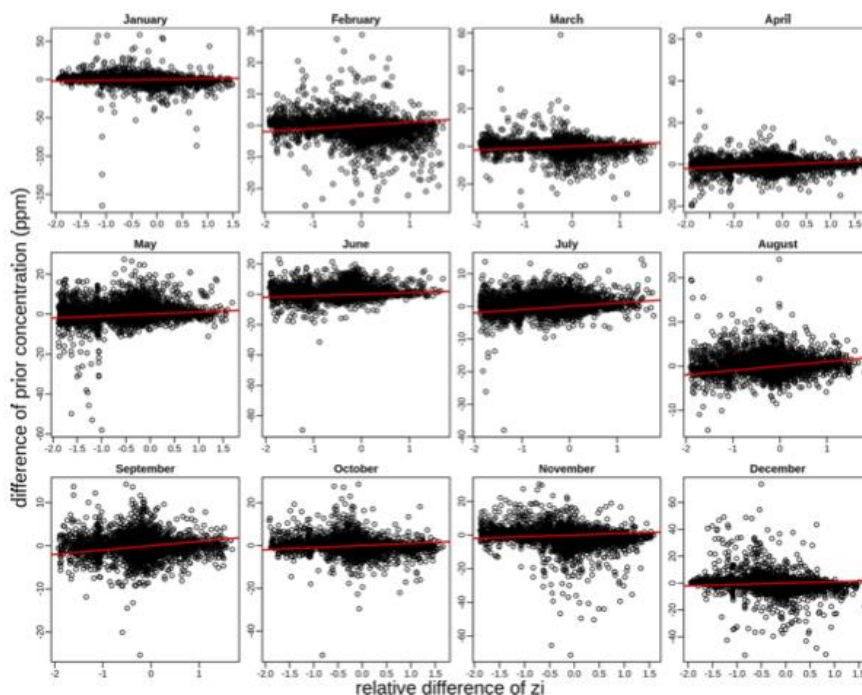
Figure 6. Scatter plot of differences of prior concentrations and mixing heights calculated with STILT and FLEXPART models (i.e. STILT-FLEXPART on the x and y axes). Red lines indicate the slopes.

ERA-5 reanalysis. Results in Fig. 7, "meteo", indicate that using different meteorological data results in pronounced differences when the FLEXPART model was forced by operational forecast data instead of the ERA-5 reanalysis. These differences notably occur during the time of net CO$_2$ release, corresponding to quite small differences during the time of growing season. This, however, only explains a small part of the overall differences (shown in Fig. 7, "base") that dominate all the months except August and September. In a previous study, Liu et al. (2011) concluded that uncertainties in meteorological fields lead to a significant contribution to the total transport error, as well as to an underestimation of the vertical turbulent mixing even when the same circulation model and mixing parameterizations were used to reconstruct vertical mixing from a single meteorological analysis. Tolk et al. (2008) also found meteorology to be a key driver of representation error, which varies spatially and temporally. They indicated that a large contribution to representation error is caused by unresolved model topography at coarse spatial resolution during night, while convective structures, mesoscale circulations, and the variability of CO$_2$ fluxes dominate during day-time. Deng et al. (2017) found

that assimilating meteorological observations such as wind speed and wind direction in transport models significantly improved the model performances, achieving an uncertainty reduction of about 50 % in wind speed and direction, especially when measurements in the mixed layer were assimilated. Nonetheless, they concluded that the differences in CO$_2$ emissions reached up to 15 % at local-scale corrections after inversion and were limited to 5 % for the total emissions integrated across the regional domain of interest. These results refer to the limited impact of meteorological data. Note however that the main aim of this experiment was to test whether differences in driving meteorological data could explain the differences between STILT and FLEXPART, but we are not assessing the overall impact of meteorological uncertainties. Doing so would in particular require testing non-ECMWF meteorological products.

Furthermore, we tested the possible impact of surface layer heights (the height up to which particles are sensitive to the fluxes) that may affect the particle dispersion, provided that STILT relies on the assumption of defining the surface layer as a half of the mixed layer height, while in FLEXPART it is defined as a fixed height of 100 m (these
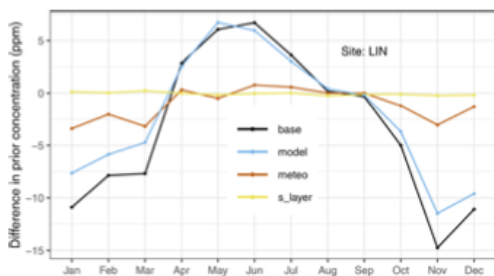
Figure 7. Differences in prior concentration simulated at LIN with STILT and FLEXPART using different configurations. "s_layer", yellow line, refers to the difference calculated with STILT using two assumptions of defining the surface layer height, once with the default as 0.5 of the mixed layer and once with 100 m as used in FLEXPART; "meteo", red line, indicates the differences calculated with FLEXPART using two different types of meteorological data, IFS (the STILT default) and ERA-5; "model", blue line, denotes the differences calculated with STILT and FLEXPART, given identical meteorological data (IFS) and surface layer height (100 m); "base", black line, refers to the base configurations of STILT and FLEXPART encompassing all possible differences between models – i.e. (1) STILT with IFS forecasting data and a surface layer height as 0.5 times that of the mixed layer height and (2) FLEXPART with ERA-5 reanalysis and the surface layer height of 100 m.

are default configurations of the models). In this experiment, STILT was run with a surface layer height of 100 m, so the impact of the surface layer on CO$_2$ simulations is outlined by the comparison with another run using the default configurations of STILT. The differences in simulated CO$_2$ concentrations due to differences in the surface layer were found to be quite small (Fig. 7, "s_layer") and, therefore, can be negligible in both magnitude and temporal pattern compared to the overall differences. However, varying the models STILT and FLEXPART with identical meteorological data and identical surface layer leads to the largest differences, in particular during the growing season months and winter months (Fig. 7, "model"). As a result, model-to-model differences largely affect the simulations of CO$_2$ concentrations and are likely originating from the transport model schemes. It is clearly noticeable that the overall differences combine the underlying differences of "model", "meteo", and "s_layer" and are yielded as the arithmetic summation of this partitioning.

## How do our results explain the range of uncertainties reported in scientific literature?

To shed more light on the drivers of differences in optimized CO$_2$ fluxes, we analyse the spread in our inversions in line with the spreads in other inversion estimates that were reported in two previous studies over the same domain of Europe. Figure 9 shows the spreads amid the three studies:
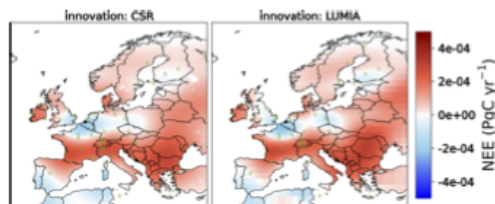
Figure 8. Innovation of fluxes calculated from CSR and LUMIA using identical uncertainties of prior flux and measurements. The uncertainty flux shape was flat and the decaying spatial correlation was fit to a Gaussian function with 500 km scale. FLEXPART and TM5 models were used in this experiment.

(1) eight inversions conducted in our results denoted as "Ensemble", (2) six inversions of the EUROCOM experiment (denoted as "EUROCOM") done by Monteil et al. (2020), and (3) five inversions of the drought study by Thompson et al. (2020), focusing on analysing the 2018 drought impact on NEE, denoted as "Drought". Note that in "EUROCOM" and "Drought", the tracer inversions differed in the atmospheric regional transport models, the definition of boundary conditions, the definition of control vector, the selection of atmospheric datasets, and the a priori fluxes. These differences are expected to span a large range of uncertainty sources in the posterior NEE. The climatological monthly estimates of NEE were averaged over "EUROCOM" inversion members for the respective years 2006–2015, except for one inversion (NAME), which was limited to 2011–2015. "Ensemble" and "Drought" were confined to the analysis year of 2018. The monthly NEE estimates were calculated for all ensembles as the average over their respective inversion members. The annual mean of NEE estimated with "EUROCOM", "Ensemble", and "Drought" amounts to -0.19, -0.29, and -0.05 PgC with standard deviations of 0.34, 0.29, and 0.46 PgC, respectively.

The spreads amid each ensemble of inversions are illustrated by the min and max values bounded around the mean on the error bars (Fig. 9). The monthly mean of NEE estimates shows a good consistency in all the ensembles. The spreads are also relatively comparable, albeit variable over months. For instance, "EUROCOM" and "Drought" exhibit larger spreads during the growing season (April–August), while "Ensemble" has a larger spread in the rest of the months – i.e. during winter. Note that all ensembles experience large spreads during June and May. Although the participating inversions to "EUROCOM" and "Drought" had different configurations, the spreads were not largely different from our inversion spreads. This implies that the use of different atmospheric transport models could account for a large fraction of differences in posterior fluxes, although differences in the definition of uncertainty covariance matrices and lateral boundary conditions likely contribute as well. More-
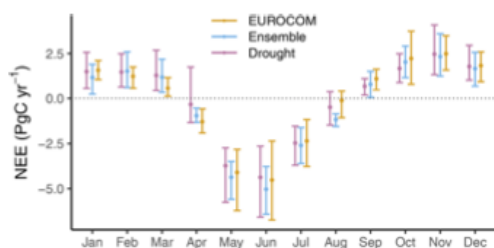
Figure 9. Comparison of monthly NEE estimates calculated as the mean of six inversions taken from Monteil et al. (2020), denoted as "EUROCOM"; eight inversion members conducted in our study (setups listed in Table 2), denoted as "Ensemble"; and five inversions used in Thompson et al. (2020) for the 2018 drought study, denoted as "Drought". The error bars refer to the spreads (min/max) over the respective members amid each ensemble of inversions.

over, the discrepancies in "EUROCOM" and "Drought" estimates are expected to be partially caused by using different atmospheric datasets in the inversion systems. Munassar et al. (2022) found that posterior fluxes can be more sensitive to changing the number of stations than changing the prior flux models.

5 Conclusions

Estimating atmospheric tracer fluxes through inverse modelling systems has been widely used, in particular for targeting the major greenhouse gases (GHGs) to improve the quantification of natural (both terrestrial and oceanic) sources and sinks. Here, an analysis of differences in posterior fluxes of CO_2 was carried out using inversion systems deploying different regional transport models. The difference between minimum and maximum spreads for annually integrated fluxes was found to be $0.92 \, \mathrm{PgC \, yr^{-1}}$ for the ensemble range of 0.20 and $-0.72 \, \mathrm{PgC \, yr^{-1}}$, with a mean estimate of $-0.29 \, \mathrm{PgC \, yr^{-1}}$ calculated over the full domain of Europe in 2018. We tested the regional transport, the boundary conditions, and the inversion systems. The regional transport accounts for the largest part of the discrepancies in the optimized fluxes as well as in the estimation of CO_2 concentration. Temporal and spatial differences in posterior fluxes are consistent with the differences in simulated CO_2 concentration sampled with STILT and FLEXPART over the station network. They demonstrate a spatial pattern over certain areas during June and December, suggesting rather systematic differences between STILT and FLEXPART. The differences in the regional transport are mainly caused by the transport schemes, while meteorological forcing data partially contribute to these differences, especially during the months in which net release of CO_2 occurs. However, the differences in CO_2 simulations did not show large sensitivities to other pa-

rameters such as the way the surface layer height (maximum altitude considered sensitive to the fluxes in Lagrangian models) and the mixing height are defined. In addition, the global transport models used in the global inversions that provide the far-field contributions to the regional domain are responsible for small but non-negligible differences in the inversion estimates. These differences appeared to be homogeneous spatially and temporally, which can be considered as bias-like. The differences arising from using different inversion systems integrated over the entire domain of Europe were on the contrary rather small, once differences such as the transport model and the uncertainties are controlled for. However, such an impact is partially a result of applying different structure and shape in the prior flux uncertainty, while the rest may be attributed to differences in the cost function convergence to reach the minimum. This reflects the importance of the way the uncertainty is prescribed in the tracer inversion systems.

The divergence in CO_2 flux estimates resulting from swapping the regional transport model emphasizes the need for further evaluation of atmospheric transport models in order to improve the performance of the models. At the same time, it is important to realistically account for the transport errors in the tracer inversions. Errors in meteorology parameters assimilated in transport models as forcing data should also be accounted for explicitly, potentially through making use of an ensemble of meteorology data to estimate such errors. Despite the non-negligible difference between inversion systems, this study indicates the importance of following a common inversion protocol when reporting flux estimates from different inversion frameworks.

Author contributions. SM and GM designed the study. CG, MS, UK, and KUT gave valuable suggestions in frequent discussions that helped improve the design and the structure of the study. SM wrote the paper and performed the simulations of CSR. GM performed the simulations of LUMIA. CG prepared and provided the fluxes of VPRM, and FTK processed the anthropogenic emission

References

Babenhauserheide, A., Basu, S., Houweling, S., Peters, W., and Butz, A.: Comparing the CarbonTracker and TM5-4DVar data assimilation systems for CO_2 surface flux inversions, Atmos. Chem. Phys., 15, 9747–9763, 10.5194/acp-15-9747-2015, 2015.

Baker, D. F., Law, R. M., Gurney, K. R., Rayner, P., Peylin, P., Denning, A. S., Bousquet, P., Bruhwiler, L., Chen, Y. H., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S., and Zhu, Z.: TransCom 3 inversion intercomparison: Impact of transport model errors on the interannual variability of regional CO_2 fluxes, 1988–2003, Global Biogeochem. Cy., 20, Gb1002, https://doi.org/10.1029/2004gb002439, 2006.

Bastos, A., Ciais, P., Friedlingstein, P., Sitch, S., Pongratz, J., Fan, L., Wigneron, J. P., Weber, U., Reichstein, M., Fu, Z., Anthoni, P., Arneth, A., Haverd, V., Jain, A. K., Joetzjer, E., Knauer, J., Lienert, S., Loughran, T., McGuire, P. C., Tian, H., Viovy, N., and Zaehle, S.: Direct and seasonal legacy effects of the 2018 heat wave and drought on European ecosystem productivity, Sci. Adv., 6, eaba2724, https://doi.org/10.1126/sciadv.aba2724, 2020.

Bousquet, P., Ciais, P., Peylin, P., Ramonet, M., and Monfray, P.: Inverse modeling of annual atmospheric CO_2 sources and sinks: 1. Method and control inversion, J. Geophys. Res.-Atmos., 104, 26161–26178, https://doi.org/10.1029/1999JD900342, 1999.

Broquet, G., Chevallier, F., Breon, F. M., Kadygrov, N., Alemanno, M., Apadula, F., Hammer, S., Haszpra, L., Meinhardt, F., Morgui, J. A., Necki, J., Piacentino, S., Ramonet, M., Schmidt, M., Thompson, R. L., Vermeulen, A. T., Yver, C., and Ciais, P.: Regional inversion of CO_2 ecosystem fluxes from atmospheric measurements: reliability of the uncertainty estimates, Atmos. Chem. Phys., 13, 9039–9056, https://doi.org/10.5194/acp-13-9039-2013, 2013.

Chen, H. W., Zhang, F. Q., Lauvaux, T., Davis, K. J., Feng, S., Butler, M. P., and Alley, R. B.: Characterization of Regional-Scale CO_2 Transport Uncertainties in an Ensemble with Flow-Dependent Transport Errors, Geophys. Res. Lett., 46, 4049–4058, https://doi.org/10.1029/2018gl081341, 2019.

Chevallier, F., Wang, T., Ciais, P., Maignan, F., Bocquet, M., Arain, M. A., Cescatti, A., Chen, J. Q., Dolman, A. J., Law, B. E., Margolis, H. A., Montagnani, L., and Moors, E. J.: What eddy-covariance measurements tell us about prior land flux errors in CO_2-flux inversion schemes, Global Biogeochem. Cy., 26, Gb1021, https://doi.org/10.1029/2010gb003974, 2012.

Davies, T.: Lateral boundary conditions for limited area models, Q. J. Roy. Meteorol. Soc., 140, 185–196, https://doi.org/10.1002/qj.2127, 2014.

Deng, A. J., Lauvaux, T., Davis, K. J., Gaudet, B. J., Miles, N., Richardson, S. J., Wu, K., Sarmiento, D. P., Hardesty, R. M., Bonin, T. A., Brewer, W. A., and Gurney, K. R.: Toward reduced transport errors in a high resolution urban CO_2 inversion system, Element.-Sci. Anthrop., 5, 20 pp., https://doi.org/10.1525/elementa.133, 2017.

Drought 2018 Team and ICOS Atmosphere Thematic Centre: Drought-2018 atmospheric CO_2 Mole Fraction product for 48 stations (96 sample heights), ICOS [data set], https://doi.org/10.18160/ERE9-9D85, 2020.

Engelen, R. J., Denning, A. S., Gurney, K. R., and Modelers, T.: On error estimation in atmospheric CO_2 inversions, J. Geophys. Res.-Atmos., 107, 4635, https://doi.org/10.1029/2002jd002195, 2002.

Enting, I. G. and Newsam, G. N.: Inverse Problems in Atmospheric Constituent Studies, 2. Sources in the Free Atmosphere, Inverse Probl., 6, 349–362, https://doi.org/10.1088/0266-5611/6/3/005, 1990.

Fletcher, S. E. M., Gruber, N., Jacobson, A. R., Gloor, M., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Muller, S. A., and Sarmiento, J. L.: Inverse estimates of the oceanic sources and sinks of natural CO_2 and the implied oceanic

carbon transport, Global Biogeochem. Cy., 21, Gb1010, https://doi.org/10.1029/2006gb002751, 2007.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Bates, N. R., Becker, M., Bellouin, N., Bopp, L., Chau, T. T. T., Chevallier, F., Chini, L. P., Cronin, M., Currie, K. I., Decharme, B., Djeutchouang, L. M., Dou, X., Evans, W., Feely, R. A., Feng, L., Gasser, T., Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Luijkx, I. T., Jain, A., Jones, S. D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lienert, S., Liu, J., Marland, G., McGuire, P. C., Melton, J. R., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S. I., Niwa, Y., Ono, T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A. J., Sweeney, C., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F., van der Werf, G. R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A. J., Willis, D., Wiltshire, A. J., Yuan, W., Yue, C., Yue, X., Zaehle, S., and Zeng, J.: Global Carbon Budget 2021, Earth Syst. Sci. Data, 14, 1917–2005, https://doi.org/10.5194/essd-14-1917-2022, 2022.

Geels, C., Gloor, M., Ciais, P., Bousquet, P., Peylin, P., Vermeulen, A. T., Dargaville, R., Aalto, T., Brandt, J., Christensen, J. H., Frohn, L. M., Haszpra, L., Karstens, U., Rodenbeck, C., Ramonet, M., Carboni, G., and Santaguida, R.: Comparing atmospheric transport models for future regional inversions over Europe – Part 1: mapping the atmospheric CO$_2$ signals, Atmos. Chem. Phys., 7, 3461–3479, https://doi.org/10.5194/acp-7-3461-2007, 2007.

Gerbig, C., Korner, S., and Lin, J. C.: Vertical mixing in atmospheric tracer transport models: error characterization and propagation, Atmos. Chem. Phys., 8, 591–602, https://doi.org/10.5194/acp-8-591-2008, 2008.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B. C., Sarmiento, J., Taguchi, S., Takahashi, T., and Yuen, C.-W.: TransCom 3 CO$_2$ inversion intercomparison: 1. Annual mean control results and sensitivity to transport and prior flux information, Tellus B, 55, 555–579, https://doi.org/10.3402/tellusb.v55i2.16728, 2016.

Kaminski, T., Rayner, P. J., Heimann, M., and Enting, I. G.: On aggregation errors in atmospheric transport inversions, J. Geophys. Res.-Atmos., 106, 4703–4715, 2001.

Karion, A., Lauvaux, T., Coto, I. L., Sweeney, C., Mueller, K., Gourdji, S., Angevine, W., Barkley, Z., Deng, A. J., Andrews, A., Stein, A., and Whetstone, J.: Intercomparison of atmospheric trace gas dispersion models: Barnett Shale case study, Atmos. Chem. Phys., 19, 2561–2576, https://doi.org/10.5194/acp-19-2561-2019, 2019.

Kountouris, P., Gerbig, C., Totsche, K. U., Dolman, A. J., Meesters, A. G. C. A., Broquet, G., Maignan, F., Gioli, B., Montagnani, L., and Helfter, C.: An objective prior error quantification for regional atmospheric inverse applications, Biogeosciences, 12, 7403–7421, https://doi.org/10.5194/bg-12-7403-2015, 2015.

Kountouris, P., Gerbig, C., Rödenbeck, C., Karstens, U., Koch, T. F., and Heimann, M.: Atmospheric CO$_2$ inversions on the mesoscale using data-driven prior uncertainties: quantification of the European terrestrial CO$_2$ fluxes, Atmos. Chem. Phys., 18, 3047–3064, https://doi.org/10.5194/acp-18-3047-2018, 2018a.

Kountouris, P., Gerbig, C., Rödenbeck, C., Karstens, U., Koch, T. F., and Heimann, M.: Technical Note: Atmospheric CO$_2$ inversions on the mesoscale using data-driven prior uncertainties: methodology and system evaluation, Atmos. Chem. Phys., 18, 3027–3045, https://doi.org/10.5194/acp-18-3027-2018, 2018b.

Lauvaux, T., Miles, N. L., Deng, A. J., Richardson, S. J., Cambaliza, M. O., Davis, K. J., Gaudet, B., Gurney, K. R., Huang, J. H., O'Keefe, D., Song, Y., Karion, A., Oda, T., Patarasuk, R., Razlivanov, I., Sarmiento, D., Shepson, P., Sweeney, C., Turnbull, J., and Wu, K.: High-resolution atmospheric inversion of urban CO$_2$ emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX), J. Geophys. Res.-Atmos., 121, 5213–5236, https://doi.org/10.1002/2015jd024473, 2016.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P. A., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Arneth, A., Arora, V. K., Barbero, L., Bastos, A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Doney, S. C., Gkritzalis, T., Goll, D. S., Harris, I., Haverd, V., Hoffman, F. M., Hoppema, M., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Johannessen, T., Jones, C. D., Kato, E., Keeling, R. F., Goldewijk, K. K., Landschützer, P., Lefèvre, N., Lienert, S., Liu, Z., Lombardozzi, D., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S., Neill, C., Olsen, A., Ono, T., Patra, P., Peregon, A., Peters, W., Peylin, P., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rocher, M., Rödenbeck, C., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Steinhoff, T., Sutton, A., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F. N., van der Laan-Luijkx, I. T., van der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J., Wright, R., Zaehle, S., and Zheng, B.: Global Carbon Budget 2018, Earth Syst. Sci. Data, 10, 2141–2194, https://doi.org/10.5194/essd-10-2141-2018, 2018.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model, J. Geophys. Res.-Atmos., 108, 4493, https://doi.org/10.1029/2002jd003161, 2003.

Liu, J. J., Fung, I., Kalnay, E., and Kang, J. S.: CO$_2$ transport uncertainties from the uncertainties in meteorological fields, Geophys. Res. Lett., 38, L12808, https://doi.org/10.1029/2011gl047213, 2011.

Mahadevan, P., Wofsy, S. C., Matross, D. M., Xiao, X. M., Dunn, A. L., Lin, J. C., Gerbig, C., Munger, J. W., Chow, V. Y., and Gottlieb, E. W.: A satellite-based biosphere parameterization for net ecosystem CO$_2$ exchange: Vegetation Photosynthesis and Respiration Model (VPRM), Global Biogeochem. Cy., 22, Gb2005, https://doi.org/10.1029/2006gb002735, 2008.

Monteil, G. and Scholze, M.: Regional CO$_2$ inversions with LUMIA, the Lund University Modular Inversion Algorithm, v1.0, Geosci. Model Dev., 14, 3383–3406, https://doi.org/10.5194/gmd-14-3383-2021, 2021.

Monteil, G., Broquet, G., Scholze, M., Lang, M., Karstens, U., Gerbig, C., Koch, F. T., Smith, N. E., Thompson, R. L., Luijkx, I. T., White, E., Meesters, A., Ciais, P., Ganesan, A. L., Manning, A.,

Mischurow, M., Peters, W., Peylin, P., Tarniewicz, J., Rigby, M., Rödenbeck, C., Vermeulen, A., and Walton, E. M.: The regional European atmospheric transport inversion comparison, EURO-COM: first results on European-wide terrestrial carbon fluxes for the period 2006–2015, Atmos. Chem. Phys., 20, 12063–12091, https://doi.org/10.5194/acp-20-12063-2020, 2020.

Munassar, S. and Monteil, G.: JenaCarboScopeRegional and LU-MIA inversion results for 2018, ICOS ERIC – Carbon Portal [data set], https://doi.org/10.18160/QE4G-TP7T, 2023.

Munassar, S., Rödenbeck, C., Koch, F. T., Totsche, K. U., Gałkowski, M., Walther, S., and Gerbig, C.: Net ecosystem exchange (NEE) estimates 2006–2019 over Europe from a pre-operational ensemble-inversion system, Atmos. Chem. Phys., 22, 7875–7892, https://doi.org/10.5194/acp-22-7875-2022, 2022.

Petrescu, A. M. R., McGrath, M. J., Andrew, R. M., Peylin, P., Peters, G. P., Ciais, P., Broquet, G., Tubiello, F. N., Gerbig, C., Pongratz, J., Janssens-Maenhout, G., Grassi, G., Nabuurs, G. J., Regnier, P., Lauerwald, R., Kuhnert, M., Balkovič, J., Schelhaas, M. J., Denier van der Gon, H. A. C., Solazzo, E., Qiu, C., Pilli, R., Konovalov, I. B., Houghton, R. A., Günther, D., Perugini, L., Crippa, M., Ganzenmüller, R., Luijkx, I. T., Smith, P., Munassar, S., Thompson, R. L., Conchedda, G., Monteil, G., Scholze, M., Karstens, U., Brockmann, P., and Dolman, A. J.: The consolidated European synthesis of CO₂ emissions and removals for the European Union and United Kingdom: 1990–2018, Earth Syst. Sci. Data, 13, 2363–2406, https://doi.org/10.5194/essd-13-2363-2021, 2021.

Peylin, P., Baker, D., Sarmiento, J., Ciais, P., and Bousquet, P.: Influence of transport uncertainty on annual mean and seasonal inversions of atmospheric CO₂ data, J. Geophys. Res.-Atmos., 107, 4385, https://doi.org/10.1029/2001jd000857, 2002.

Pisso, I., Sollum, E., Grythe, H., Kristiansen, N. I., Cassiani, M., Eckhardt, S., Arnold, D., Morton, D., Thompson, R. L., Groot Zwaaftink, C. D., Evangeliou, N., Sodemann, H., Haimberger, L., Henne, S., Brunner, D., Burkhart, J. F., Fouilloux, A., Brioude, J., Philipp, A., Seibert, P., and Stohl, A.: The Lagrangian particle dispersion model FLEX-PART version 10.4, Geosci. Model Dev., 12, 4955–4997, https://doi.org/10.5194/gmd-12-4955-2019, 2019.

Rivier, L., Peylin, P., Ciais, P., Gloor, M., Rödenbeck, C., Geels, C., Karstens, U., Bousquet, P., Brandt, J., and Heimann, M.: European CO₂ fluxes from atmospheric inversions using regional and global transport models, Climatic Change, 103, 93–115, https://doi.org/10.1007/s10584-010-9908-4, 2010.

Rödenbeck, C.: Estimating CO₂ sources and sinks from atmospheric mixing ratio measurements using a global inversion of atmospheric transport, Technical Report, Max-Planck Institute for Biogeochemistry, 2005.

Rödenbeck, C., Houweling, S., Gloor, M., and Heimann, M.: CO₂ flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport, Atmos. Chem. Phys., 3, 1919–1964, https://doi.org/10.5194/acp-3-1919-2003, 2003.

Rödenbeck, C., Gerbig, C., Trusilova, K., and Heimann, M.: A two-step scheme for high-resolution regional atmospheric trace gas inversions based on independent models, Atmos. Chem. Phys., 9, 5331–5342, https://doi.org/10.5194/acp-9-5331-2009, 2009.

Rödenbeck, C., Zaehle, S., Keeling, R., and Heimann, M.: The European carbon cycle response to heat and drought as seen from atmospheric CO₂ data for 1999–2018, Philos. T. R. Soc. Lond. B, 375, 20190506, https://doi.org/10.1098/rstb.2019.0506, 2020.

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., Chevallier, F., Crowell, S., Davis, K. J., Deng, F., Denning, S., Feng, L., Jones, D., Liu, J. J., and Palmer, P. I.: Quantifying the Impact of Atmospheric Transport Uncertainty on CO₂ Surface Flux Estimates, Global Biogeochem. Cy., 33, 484–500, https://doi.org/10.1029/2018gb006086, 2019.

Shi, H., Tian, H. Q., Pan, N. Q., Reyer, C. P. O., Ciais, P., Chang, J. F., Forrest, M., Frieler, K., Fu, B. J., Gadeke, A., Hickler, T., Ito, A., Ostberg, S., Pan, S. F., Stevanovic, M., and Yang, J.: Saturation of Global Terrestrial Carbon Sink Under a High Warming Scenario, Global Biogeochem. Cy., 35, e2020GB006800, https://doi.org/10.1029/2020GB006800, 2021.

Steinbach, J., Gerbig, C., Rödenbeck, C., Karstens, U., Minejima, C., and Mukai, H.: The CO₂ release and Oxygen uptake from Fossil Fuel Emission Estimate (COFFEE) dataset: effects from varying oxidative ratios, Atmos. Chem. Phys., 11, 6855–6870, https://doi.org/10.5194/acp-11-6855-2011, 2011.

Thompson, R. L., Broquet, G., Gerbig, C., Koch, T., Lang, M., Monteil, G., Munassar, S., Nickless, A., Scholze, M., Ramonet, M., Karstens, U., van Schaik, E., Wu, Z., and Rodenbeck, C.: Changes in net ecosystem exchange over Europe during the 2018 drought based on atmospheric observations, Philos. T. R. Soc. Lond. B, 375, 20190512, https://doi.org/10.1098/rstb.2019.0512, 2020.

Tolk, L. F., Meesters, A. G. C. A., Dolman, A. J., and Peters, W.: Modelling representation errors of atmospheric CO₂ mixing ratios at a regional scale, Atmos. Chem. Phys., 8, 6587–6596, https://doi.org/10.5194/acp-8-6587-2008, 2008.

8 Appendix: CH$_4$ inverse modelling intercomparison protocol

Protocol for the intercomparison of national CH$_4$ emissions estimated by inverse modelling systems for Western Europe – Phase 1

- version November 2021 -

L. Florentie, S. Houweling

1. Goal

This CH$_4$ inversion intercomparison effort aims at assessing the constraints of atmospheric measurements on national methane fluxes, in particular the long-term trend in those emissions. The main focus of this intercomparison is on national CH$_4$ emission estimates in (Western) Europe (including separation into anthropogenic and natural emissions) and their uncertainties, as well as the added value of satellite-based XCH$_4$ observations in this respect. In addition, we aim to formulate benchmarking procedures and best practices for inversion-based emission estimates.

We will perform sensitivity experiments to search for answers to the following questions:
-   What is the uncertainty range of inversion-optimized national CH$_4$ emissions? Is this as expected given the uncertainties of input datasets, observations, transport models and the findings in other research efforts (e.g. previous intercomparison efforts and the work done in other CoCO2 work packages)?
-   How does the uncertainty in the absolute emissions compare to that of year-to-year changes and the long-term trend? Are inversions capable of robustly resolving changes in national emissions on the time-scale relevant for the global stock take?
-   What validation measurements and benchmarking metrics are needed to quantify the performance of different inversion systems?
-   Can we formulate guidelines and best practices for atmospheric inversion setups that are general enough to be used by non-scientific parties, and that will yield consistent results w.r.t. national CH$_4$ emission estimates?
-   What is the readiness level of inversion systems for use of the CO2M (which includes XCH$_4$ observations) in refining CH$_4$ emission reporting?

The desired outcome of this intercomparison consists of a top-down assessment of trends in national methane emissions and their consistency with bottom-up reporting. This includes the formulation of a set of benchmarking methodologies, performance metrics, and best-practices for regional emission inversions.

2. Relation to the VERIFY CH$_4$ inversion intercomparison

The current effort is closely linked to the experiment performed in work package 4 of the VERIFY project. Both efforts investigate the ability of regional inversion frameworks to

estimate CH$_4$ fluxes in Europe. However, this experiment extends the effort of the VERIFY project in the sense that

(i)   we are particularly interested in national emission estimates, and the role of atmospheric inversions for verification of those estimates,

(ii)   we will conduct a 2-stage assessment: the first stage focusses on the use of ICOS surface data, whereas the 2nd stage extends this with the use of satellite observations. Moreover, phase 2 will be used to test findings/guidelines from the first round of inversions.

(iii)   this assessment continues beyond the end of the VERIFY project, allowing us to study more recent years in the 2nd phase. This facilitates a better assessment of satellite-based inversions, but also lets us prepare for and eventually include the first global stock take.

(iv)   we explicitly welcome participation from groups not involved in the VERIFY project to increase the size of the ensemble of models and inversion methods.

To minimize the effort for research groups, this experiment starts from the same basis as the VERIFY intercomparison. This implies that inversions performed to participate in the VERIFY intercomparison will automatically satisfy the requirements for participation in the first phase of this experiment as well, although some additional output is requested to allow for the specific analyses that we intend to perform for the current study.

The specifications for this first phase can be found in the current protocol document, which is based largely on the VERIFY protocol. We therefore gratefully acknowledge the large effort done by the VERIFY team to set up this protocol, to gather and produce input datasets, and to harmonize the observation datasets.

3. Overview of the experiment

An important aim of this study consists of the definition of a benchmarking strategy to assess the performance of different emission estimates, and the formulation of best practices to define country emission totals. This study will therefore consist of 2 modelling & analysis phases:

- **Modelling phase 1**: National emission estimates based on atmospheric CH$_4$ inversions using surface measurements, according to the protocol described in this document.
- **Analysis phase 1**: Assessment of the submitted results, with focus on uncertainty reduction due to the observational constraints, the detection of long-term trends, and the impact of the performance of the atmospheric transport model. This phase will result in an updated benchmarking strategy and formulation of a first version of best practices (in consultation with the modelling groups).
- **Modelling phase 2**: New round of inversions according to an updated protocol (based on the outcome of phase 1). In this round the inversions will at least be extended (we aim for 2021). It includes the possibility to rerun the full period for those groups that wish to make changes to their setup, according to the new protocol. In this round we will also strongly encourage (new) submissions based on satellite-based XCH$_4$ observations.

- **Analysis phase 2**: This analysis round will strongly focus on the added observational constraint due to the availability of satellite observations. Moreover, the preliminary findings and best practices as resulting from phase 1 are tested. The desired outcome of this phase consists of the formulation of a set of benchmarking methodologies, performance metrics and best-practices for national CH$_4$ inversion-based emission estimates. Our findings will be written down in a technical report and a scientific publication.

We strongly encourage groups that participate in modelling phase 1 to also participate in modelling phase 2, for consistency of the analysis phases. New submissions for modelling phase 2 only are welcomed as well.

4. Timeline

When?	What?
Nov., 2021	Distribution of intercomparison protocol and input datasets for phase 1 to the modelling community
March 1, 2022	**Deadline for submission of inversion results for phase 1 (<= 2018)**
April, 2022	Discussion with the modelling groups about the results of phase 1
Oct., 2022	Distribution of intercomparison protocol and input datasets for phase 2 to the modelling community
April 1, 2023	**Deadline for submission of inversion results for phase 2 (<= 2021)**
June, 2023	Discussion with the modelling groups about the results of phase 2

5. Domain definition (spatial & temporal)

The regional inversions should cover at least the area from 15W to 35E and 35N to 70N. Modelers are free to choose the spatial resolution they deem most appropriate. National emission totals are to be provided for at least the EU27 countries and the UK.

The inversions should cover as many years as possible in the range 2005-2018. If it is not possible to provide results for the full period it is preferred if groups submit results for a selection of (separate) years, chosen such as to cover the full period as good as possible, and including at least the years 2008, 2013 and 2018. This way we can still study the trend in emissions.

6. Input Data

6.1   Prior CH$_4$ fluxes

An overview of the prior total CH$_4$ fluxes that we expect modelling groups to use for this experiment is given below in Table 1. This is the same dataset as used for the VERIFY experiment. The fluxes area available on our Research Drive, both on monthly 0.25°x 0.25°

Table 1: Prior CH$_4$ fluxes

Category	Data source	Original Resolution		Time period
Peatlands, Mineral soils (emissions & uptake), inundated	JSBACH-HIMMELI[1]	0.1°x 0.1°	daily	2005-2020
Inland water[2]	ULB	0.1°x 0.1°	monthly	Climatology
Termites	Saunois, 2020	-	annually	Climatology
Ocean	Weber, 2021	0.25°x 0.25°	monthly	Climatology
Geological	Etiope, 2015	-	annually	Climatology
Fossil fuels	EDGAR v6.0	0.1°x 0.1°	monthly	2005-2018
Agriculture and waste	EDGAR v6.0	0.1°x 0.1°	monthly	2005-2018
Biofuels & biomass burning	GFED-4.1s	0.25°x 0.25°	monthly	2005-2020

[1]Covers Europe from 10.5°W to 33°E and 34.5°N to 73.5°N
[2]Covers Eurasia from 26°W to 55°E and 34°N to 78°N

resolution and in their original resolution, such that you can choose to use the resolution most suitable for your system. If you want to make use of the original flux dataset, please make sure that the total flux is conserved upon regridding. Note that fluxes for peatlands, mineral soils and inland water (lakes) are only given for a limited domain. If groups want to include fluxes for these categories *outside* the provided domain as well, they are free to use the fluxes there that they deem appropriate.

Modelling groups are free to decide whether or not to explicitly model the atmospheric OH sink (as its influence is expected to be very small over Europe). For groups that do wish to include it, please use the OH field provided on our Research Drive. Modelling groups that decide to submit an extra inversion (apart from the proposed experiments, see section 6.5), which makes use of isotope data, we strongly advise to include the atmospheric OH sink.

6.2   Atmospheric observations

As part of the VERIFY experiment a harmonized set of atmospheric observations of CH$_4$ mixing ratios was created. This set contains observations provided from the InGOS project (2005-2016), from NOAA flask sampling sites in Europe (2005-2018), from AGAGE, and additional data from European sites (ICOS, WDCGG and personal communications). The full site list is available with the input data. The list contains a label for each site to indicate whether it may be assimilated (thereby also distinguishing between a set of 'core' and 'other' observation sites, see section 6.5) or must be kept for validation purpose.

We request groups to submit at least an inversion using daytime (12:00 to 16:00 local time) surface observations only. As an additional submission, groups are free to use (satellite-based) observations in addition to or instead of the provided observations dataset. In that case, please clearly document the assimilated observations.

For mountain site observations (indicated in the observation dataset overview), we ask to only assimilate nighttime observations (0:00 to 6:00 local time), sampled at the real altitude. If groups wish to submit an additional inversion that also assimilates daytime observations, please use a sampling height in the middle between the model topography and the real altitude.

6.3 Boundary conditions

For consistency in the regional contribution, we require modelling groups that perform a regional inversion to use the CAMS v19r1 reanalysis data as background/boundary conditions. This data is based on assimilated surface observations only and can be downloaded from the Copernicus atmospheric data store:
https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-greenhouse-gas-inversion?tab=overview
For groups that have a system which makes use of the Rödenbeck 2-step inversion approach [3], consistent baseline fluxes are available on our Research Drive.

6.4 Uncertainties

All assumed uncertainties and covariance structures must be documented well. To facilitate a meaningful comparison, please adhere to the following:
- Use the observational uncertainties as provided with the observed value for the surface observations. Note that some observations are reported without uncertainties. Uncertainties to be used in those cases can be found in the header of the particular file.
- On top of the observational uncertainty a model uncertainty should be included as deemed appropriate by the modelling groups. Please document the choice made.
- Uncertainties for the prior anthropogenic and natural fluxes are free to be chosen by the modelers, but should be in reasonable agreement with the current scientific consensus. So please support your choice by appropriate references. Useful sources may include Solazzo et al. 2021, Saunois et al. 2020, and Petrescu et al. 2020

6.5 Inversion experiments

We request modelling groups to submit inversion results for at least the baseline configuration, and we highly encourage submission of additional results for the experiments described below:
- **Baseline** (mandatory): assimilating the 'core' observations only, using each model's best set-up, but adhering to the requirements outlined in sections 6.1 to 6.4 of this protocol. Please submit results for as many years as possible within the range 2005-2020, but including at least 2008, 2013 and 2018.
- **Experiment 1**: Similar to the baseline inversion, but assimilating the 'other' observations in addition to the 'core' observations.

- **Experiment 2**: Similar to the baseline inversion, but doubling the relative uncertainty of the anthropogenic prior fluxes compared to the natural fluxes,
- **Extra**: Any additional inversion results a group would like to submit. Please try to change only one aspect at a time compared to the baseline set-up, for ease of analysis.

When submitting the results, please use the following designation to distinguish between the different experiments: base, exp1, exp2, extra.

7. Output Data

To accommodate the analysis of the inversion results, all groups are asked to provide prior and posterior values and uncertainties for different types of output: gridded fluxes, national totals and simulated mole fractions. Please see the sections below for specifications of the desired output.

7.1 Gridded fluxes and uncertainties

Please provide yearly NetCDF4 files with gridded methane fluxes on a 0.25x0.25 degrees spatial grid and monthly resolution, containing:

- Dimensions:
  - latitude [degrees North, center of gridbox, at least 35.125 to 69.875]
  - longitude [degrees East, center of gridbox, at least -14.875 to 34.875]
  - time [day of year]
- Variables [lon x lat x time, kg CH_4 m^{-2} h^{-1}, double precision]:
  - prior [prior fluxes]
  - error_prior [prior flux uncertainty]
  - post [optimized fluxes]
  - error_post [uncertainty of optimized fluxes]

Filename format: grid_[yyyy]_[exp]_[model].nc

7.2 National total emissions and uncertainties

In addition to the gridded fluxes we ask you to also provide estimates for the national land-only (so excluding territorial waters) CH_4 emissions and their uncertainties. A country mask (both a high resolution 0.01°x 0.01° mask and a corresponding fractional 0.25°x 0.25° mask are provided) to be used for this aggregation can be downloaded from our Research Drive (see section 10). Please distinguish between the following categories: (fossil fuel) energy use, agriculture and waste, biofuel and biomass burning, wetlands, and other natural emissions. If your set-up does not facilitate separate optimization of these categories you can aggregate the optimized total fluxes according to the prior distribution.

Please provide yearly NetCDF4 files that contain monthly totals for:

- Dimensions:
  - country [name of country, as specified in the country mask]

- month [month number]
- Variables [country x month, Gg CH$_4$, double precision]:
  - prior_[category]
  - error_prior_[category]
  - post_[category]
  - error_post_[category]
  with *category* denoting: *total, energy, agriwaste, bioburning, wetlands, other*

Filename format: country_[yyyy]_[exp]_[model].nc

7.3   Mixing ratios and uncertainties

Please provide both prior and optimized mixing ratios and the imposed uncertainty for all assimilated observations. Include these results in the provided .csv files, or as a NetCDF4 file with variables:

site, yyyy, mm, dd, hh, mi, ss, yobs, ybkg, ypri, ypos, yerr, (ybkgpos)

all having the number of observations as dimension. In the above, 'site' represents the 3-character site identifier, 'yobs' the observed value, 'ybkg' the background mixing ratio (use '-999' if not applicable, i.e. for global models), 'ypri' and 'ypos' the prior and posterior mixing ratios respectively, 'yerr' the model-data mismatch used for assimilation (i.e. sum of observation and model uncertainty), and (if applicable) 'ybkgpos' the optimized background mixing ratio. Report all mixing ratios in ppb.

Filename format: conc_[yyyy]_[exp]_[model].csv/nc

8. Validation

8.1   CH$_4$ mixing ratios

Some observations are denoted as 'validation' in the site list. Please provide simulated mixing ratio's corresponding to these observations (and matching the timestamps) for both the prior and posterior fluxes, using the same format as specified above (use '-999' for 'yerr').

Filename format: val_[yyyy]_[exp]_[model].csv/nc

8.2   Radon mixing ratios

To assess the impact of the performance of the different transport models on national emission estimates we ask you to include radon (^{222}Rn) as a tracer for atmospheric transport in your baseline set-up. A climatological flux map of ^{222}Rn is provided on our research drive, as constructed by Karstens et al. 2015. Note that we regridded the fluxmap from its original resolution of 0.087°x 0.087° to 0.25°x 0.25°, thereby imposing average flux values per ecosystem type for the cells within our domain that are outside the bounds of the original

map. Please provide hourly simulated concentrations for the year 2018 for the sites indicated in the observations site list, in NetCDF4 files:

- Variables:
  - o Time [samples x 6, as [year,month,day,hour,minute,second] for each sample]
  - o rn_model [samples, in Bq/m^3]

Filename format: rn_[site]_[exp]_[model].nc

9. Documentation

All submissions should be accompanied by a document outlining the inversion set-up(s). This document should include at least:

- Name and email of one contact person
- Name and identifier (used in the filenames) of your inversion system
- Details about the inversion framework:
  - o Definition of the state vector
  - o Correlation length
  - o Optimization method
  - o Treatment of background mole fractions (if applicable)
- Details about the transport model:
  - o Resolution
  - o Name, driving meteorological data, and other relevant settings
  - o Estimate for the transport model error as used in the inversion
- Overview of prior emissions and their uncertainties
  - o Source (if additional sources are used apart from those listed in section 6.1) and resolution
  - o Assumed uncertainty and covariance
- Assimilated observations
  - o If you deviated from the provided list, please indicate how and why
- Chi square statistics of the inversion
- Overview of the performed experiments and simulated years.

Filename format: info_[model].txt

10. Logistics

The **deadline for submission of results for phase 1 is March 1, 2022**. We kindly ask potential participants to declare their intention to participate in advance, by sending an email to l.florentie@vu.nl. We encourage modeling groups to participate in both phases of the study, however, this is not a strict requirement.

Input data can be downloaded from our Research Drive. You will be given access after notifying us about your intention to participate.

Your results can be submitted by uploading them to the same Research Drive. Please make sure to keep the filename formats of submitted files as specified above, where [model] represents and identifier for your inversion framework.

For questions and further information, please contact Liesbeth Florentie (l.florentie@vu.nl) or Sander Houweling (s.houweling@vu.nl).

11.   Acknowledgements

12.   References

[1] Karstens, U., Schwingshackl, C., Schmithüsen, D. and Levin, I., *A process-based [222]radon flux map for Europe and its comparison to long-term observations*, Atmos. Chem. Phys., 15, 12845–12865, doi: 10.5194/acp-15-12845-2015, 2015

[2] Petrescu, A.M., Qiu, C., Ciais, P., et al., *The consolidated European synthesis of CH4 and N2O emissions for EU27 and UK: 1990-2018*, Earth Syst. Sci. Data, 13, 2307-2362, doi: 10.5194/essd-13-2307-2021, 2021

[3] Rödenbeck, C., Gerbig, C., Trusilova, K. and Heimann, M., *A two-step scheme for high-resolutionregional atmospheric trace gas inversions based on independent models*, Atmos. Chem. Phys. Discuss., 9, 1727–1756, doi: 10.5194/acpd-9-1727-2009, 2009

[4] Saunois, M., Stavert, A.R., Poulter, B., et al., *The global methane budget 2000-2017*, Earth Syst. Sci. Data, 12, 1561-623, doi 10.5194/essd-12-1561-2020, 2020

[5] Solazzo, E., Crippa, M., Guizzardi, D., Muntean, M., Choulga, M., Janssens-Maenhout, G., *Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) emission inventory of greenhouse gases*, Atmos. Chem. Phys., 21, 5655–5683, doi: 10.5194/acp-21-5655-2021, 2021