CO2 Human Emissions

Progress report on service elements for data assimilation methodology

Wouter Peters

Maarten Krol

che-project.eu

Co-ordinated by ECMWF

D5.5 Progress report on service elements for data assimilation methodology

Dissemination Level: Public

Author(s): Wouter Peters and Maarten Krol
(Wageningen University)

Date: 19/12/2019

Version: 1.0

Contractual Delivery Date: 31/12/2019

Work Package/ Task: WP5/ T5.5

Document Owner: Wageningen University

Contributors: Nicolas Bousserez and Anna Agusti-Panareda (ECMWF), plus CHE WP leaders.

Status: Final

CHE: CO2 Human Emissions Project

Coordination and Support Action (CSA)
H2020-EO-3-2017 Preparation for a European capacity to monitor CO2 anthropogenic emissions

Project Coordinator: Dr Gianpaolo Balsamo (ECMWF)
Project Start Date: 01/10/2017
Project Duration: 39 months

Published by the CHE Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX,
gianpaolo.balsamo@ecmwf.int

The CHE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776186.

Table of Contents

Tables

1 Executive Summary

This report defines the necessary components for a multi-scale and multi-species data assimilation (DA) system that targets anthropogenic CO_2 emissions. This DA system will use multiple streams of observations, including satellite observations. Out of the many viable options to serve as the basis for such a system on the global scale, we see a hybrid 4d-VAR-ensemble approach, implemented in an online transport model, and operated within a Numerical Weather Prediction environment, as a fundamental building block. On top of this a DA system should use multiple tracers, be adoptable to long- and short windows, and optimize both the atmospheric state as well as surface fluxes.

Such a system does currently not yet exist, and we recommend a number of concrete research and development needs, including to:

- Allow mass-conserving transport in the operational Integrated Forecast System (IFS) of ECMWF
- Improve the treatment of background covariances and building of long-window information transfer in the IFS DA system
- Develop Fossil Fuel DA models and capacity (FFDAS) for global and regional scales
- Expand Biospheric Carbon Cycle DA models and capacity (CCDAS) for global and regional scales
- Invest in multi-tracer transport+source modelling on all scales
- Improve the seamless coupling of regional DA systems to the global IFS
- Investigate plume-based methods for fast DA, also in a plume-in-grid approach

Many of these developments are ongoing in the community and to facilitate their uptake in an MVS for anthropogenic CO_2, we see an important role for:

(a) a prototype MVS built around the IFS and focusing on available high-resolution CO and NO_2 satellite data
(b) a multi-scale integration tool that allows local- and regional scale DA systems to feed into the global analyses.

2 Introduction

2.1 Background

The CHE prototype aims at building a system to monitor the exchange of CO_2 and potentially other important man-made greenhouse gases like CH_4 between the Earth surface and the atmosphere with the use of observations (mostly in the atmosphere), models and prior information, as well as their uncertainties to leverage the different sources of information. The system is designed to support the Paris Agreement and follows the directive of the European Commission CO_2 Task Force (Pinty et al., 2017). The general strategy and rationale for the CHE prototype is provided in CHE D5.9, stemming from the discussions in the first WP5 workshop (Reading, 25-26 September 2019). The main challenges in the approach are:

- **Multi-scale** approach to monitor emission from point sources (power stations or industrial facilities), cities and countries using different model domains from global, regional to local and model resolutions (e.g. from 25km to 100m).

- **Multi-species** approach to detect and attribute the observed atmospheric signal to specific sources/sinks (e.g. natural and anthropogenic emissions with sectorial distribution).
- **Multi-stream** approach to support different applications and users with a near-real time stream focusing on shorter synoptic timescales designed to provide early warnings and giving feedback to data producers, and a re-analysis stream that uses consolidated quality-controlled data, products and models with their individually associated uncertainties to estimate trends.

This report focuses data assimilation methodology linking with on the modelling and prior components of the prototype (D5.3) and the Earth observations (D5.1). Data assimilation methods can also support uncertainty estimation as posterior error estimation, although a complete consideration of uncertainties is covered by a dedicated report deliverable (D5.7).

The use of atmospheric measurements to constrain CO_2 exchange processes with the Earth surface is called the "source inversion problem". More specifically, within CHE we are interested in constraining anthropogenic CO_2 emissions. Anthropogenic emissions are generally confounded by CO_2 exchange with the biosphere and oceans on scales larger than individual point sources. Separation of the different exchange processes requires (1) the use of additional constraints from other trace gases (e.g. NO_2, CO, $^{14}CO_2$, and further observations like that of sun-induced fluorescence (SIF) and other human activity proxies) (2) a multi-scale approach to separate the anthropogenic hotspots from regional and global exchange with the biosphere and oceans.

Several data-assimilation (DA) methods are being employed to solve the source inversion problem considering different options for the optimal assimilation time-window, to trade-off computational pragmatism with model errors and with observational constraints. Two of the most widely employed techniques are 4-dimensional variational data-assimilation (4D-Var) and the Ensemble Kalman Filter approach (EnKF). Both approaches start with a statistical description of a state-vector. This state-vector $\boldsymbol{x}$ commonly describes the CO_2 exchange fluxes and the associated error structures. Propagation of this state-vector using an atmospheric transport model produces simulated observations (CO_2 mixing ratios, satellite columns, etc.) which are compared to atmospheric observations. Again, using proper error statistics (measurement errors, model errors) a cost function $\boldsymbol{J(x)}$ is defined which quantifies the goodness-of-fit with the observations (Tarantola, 2005).

$$\boldsymbol{J}(\boldsymbol{x}) = \tfrac{1}{2}\,(\boldsymbol{x} - \boldsymbol{x_b})^{\mathrm{T}}\,\mathbf{B}^{-1}\,(\boldsymbol{x}-\boldsymbol{x_b}) + \tfrac{1}{2}\,(H(\boldsymbol{x})-\boldsymbol{y})^{\mathrm{T}}\,\mathbf{R}^{-1}\,(H(\boldsymbol{x})-\boldsymbol{y}) \qquad (1)$$

Here, $\boldsymbol{x}$ represents the state to be optimized and $\boldsymbol{x_b}$ represents the prior information. $H(\boldsymbol{x})$ represents the simulated observations for state $\boldsymbol{x}$, and $\boldsymbol{y}$ are the observations. **B** and **R** are the matrices that represent the error statistics of the state $\boldsymbol{x}$ and of the model-data comparison.

In subsequent steps, which are different for 4D-Var and EnKF, the cost function is minimized. The 4D-Var method uses an adjoint model to calculate the gradient of $\boldsymbol{J}$ w.r.t. $\boldsymbol{x}$, and to iteratively run the model forward and backward in time, while the EnKF approach spans the uncertainty in $\boldsymbol{x}$ by an ensemble that is run forward in time in a sequential time-stepping mode. Table 1.1 summarizes the most important pros and cons for 4D-Var and EnKF.

Table 1: Accuracy, Computational and Maintenance pros and cons of 4D-Var and EnKF data assimilation methods.

Data Assimilation methods	4D-Var	EnKF
Accuracy	(+) accurate solution possible	(-) noise-generation for restricted ensemble size
Computational Cost	(-) poorly scalable on multiple core computers	(+) scalable on multiple core computers
Maintenance Cost	(-) adjoint code needed	(+) no adjoint code needed

The main weakness of EnKF algorithms resides in the low-rank nature of the error covariance matrix represented by the ensemble. Two techniques are commonly employed to mitigate the resulting sampling noise: inflation of the ensemble error variance, and localization of the impact of observations on the analysis (Houtekamer and Mitchell, 2005; Anderson, 2009). Recently, there has been a growing interest in hybrid approaches to DA, leveraging advantages of both ensemble and variational methods. In this report we describe in detail the possible DA configurations, based on variational, EnKF and hybrid ensemble-variational approaches, that can be used to build a prototype CO_2 source inversion system.

Past efforts in CO_2 DA have mostly focused on the uncertain biosphere and ocean fluxes. Constraining anthropogenic emissions therefore requires a DA system that considers these natural fluxes, but also needs to start from a proper quantification of the anthropogenic fluxes: the so-called bottom up inventories. These inventories are currently derived from proxies like fossil fuel use and activity data, which are not available instantly as required for an operational system. Multiple solutions are investigated (e.g. within CHE) that entail an anthropogenic emission model that is driven by proxies such as night-lights, temperature, etc. This is similar to approaches that use biosphere models to calculate the expected exchange of CO_2 with the biosphere. In combination with DA methods, these approaches are generally called FFDAS (fossil-fuel data-assimilation) and CCDAS (carbon cycle data-assimilation), and here the traditional gridded flux-state $\boldsymbol{x}$ is replaced by parameters that drive the biosphere and/or anthropogenic emission models.

Another major uncertainty in traditional flux inversions is the quantification of the errors that are associated with atmospheric transport. Ideally, flux inversions are performed with a transport model that accounts for transport errors (e.g. by propagating an ensemble that accounts for uncertainties in fluxes ***and*** transport), but traditional approaches account for transport errors on the right-hand side of the cost function (in the $\boldsymbol{R}$ matrix).

Within CHE, multiple challenges for a DA system targeting anthropogenic CO_2 emissions are addressed. Such a system should quantify anthropogenic emissions at the scale of individual hot-spots and countries, but also need to account for CO_2 exchange with the biosphere and ocean at global scales. Like mentioned above, separation of anthropogenic and natural CO_2 likely requires a multi-scale and multi-species approach, both in modelling and in observations. Candidate tracers that are linked to anthropogenic CO_2 are NO_2, atmospheric potential oxygen (APO), CO and $^{14}CO_2$ (CO_2 produced by burning fossil fuels is void of ^{14}C). Some of these tracers are chemically reactive and require consideration of atmospheric chemistry and surface processes like fractionation.

The multi-scale aspect refers to the need to quantify emissions from hotspots like large power-plants. Large CO_2 emitting facilities make up a substantial fraction of the global fossil CO_2 emissions, but are poorly represented in global models. Therefore, a multi-scale and multi-model DA approach is necessary to provide CO_2 emission products that are relevant for environmental decision-making.

In this report we propose solutions to build a CO_2 Monitoring and Verification Support (MVS) capacity that integrates information on CO_2 emissions from a wide range of spatial and temporal scales.

2.2 Scope of this deliverable

2.2.1 Objective of this deliverable

This deliverable defines the necessary components for a multi-scale and multi-species DA system that targets anthropogenic CO_2 emissions using multiple streams of observations, including satellite observations.

2.2.2 Work performed in this deliverable

The work benefitted from the CHE Work-package 1 developments.

2.2.3 Deviations and counter measures

Not applicable.

3 Data Assimilation components

3.1 Assimilation Methods

The following table presents a number of state-of-the art global/regional DA systems. Specifics of the systems are given below. The table presents the observation operator ***H***, which can be offline (using stored meteorological fields) or online (atmospheric dynamics is solved along with the flux inversion); the control vector ***x***, which can contain fluxes, the initial state, and meteorology (for online ***H***); characteristics of the error matrices ***B*** and ***R***; and the most important benefits and downsides.

Table 2: Assimilation methods and their components: Observation operators, Control vectors, Background and Observation error covariance matrices, and pros and cons. This table refers to atmospheric transport enabled systems, while a more comprehensive list is provided in Section 3.

Method (examples)	**Observation operator** *H*	**Control vector** *x*	B	R	**Pros (+)**	**Cons (-)**
Offline 4DVAR (CAMS) Chevallier et al 2019 & Rödenbeck et al., (2005)	Offline transport + Tangent Linear/Adjoint (TL/AD)	Fluxes, initial conditions	Static model	Transport model error + measurement error	Long window (years-decades) facilitates mass conservation, implicit full rank propagation	Uncoupled data assimilation, transport, TL/AD required, error characterisa

					of error statistics	tion is challenging
Online 4DVAR (IFS**)** (Agusti-Panareda et al., 2016)	Online transport TL/AD	Fluxes, initial conditions and meteo	Static or hybrid model	Measurement error	Coupled data assimilation (meteo,flux), transport error implicitly accounted for.	Short window (strong non-linearities), TL/AD required
Online EnKF **(**Environment Canada GEMS)	Online transport model	Fluxes, initial conditions and meteo	4D-ensemble	Measurement error	No TL/AD required, transport error implicitly accounted for	Short window, sampling noise
Offline EnKF **(**CarbonTracker**)**	Offline transport model	Fluxes, initial conditions	4D-ensemble	Transport error + measurement error	No TL/AD required	Limited window, sampling noise
Offline analytical **(**CHE regional**)**	Offline transport model + Full Jacobians	Fluxes, initial conditions	Static model or matrix	Transport error + measurement error	Exact solution	Limited size of x and/or observations
Online Hybrid **Ensemble variational** **(**IFS, Bousserez, tech memo**)**	Online transport model + TL/AD	Fluxes, initial conditions and meteo	4D-ensemble + TL/AD propagation	Measurement error	Coupled DA+ long window +transport error implicitly accounted for +potential to include processes missing from TL/AD	TL/AD

At this stage, it is useful to introduce a distinction between systems that, in analogy to weather prediction, predict the greenhouse gas emissions and distribution: Numerical Weather and Greenhouse gas Prediction (**NWGP**) systems, and systems that re-analyse the emissions using full knowledge of the entire system (e.g. measurements made in a ground network, fossil-fuel use statistics, ..): Numerical Weather and Greenhouse gas reanalysis systems (**NWGR**). Systems like the CAMS system and Carbontracker fall in this latter category, while the online systems are NWGP systems.

3.1.1 Offline 4DVAR

Finding the minimum of the cost function J (equation 1) involves iteratively progressing towards the solution with some optimization software and preconditioning strategy. At each iteration, the following gradient indicates the descent direction:

$$\nabla_{\mathbf{x}} J(\mathbf{x})=\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^{b})+\mathbf{H}^{T}\mathbf{R}^{-1}(\mathbf{H}\,\mathbf{x}-\mathbf{y}) \qquad (2)$$

The advantage of such a formulation lies in the fact that all heavy computations (i.e. those that involve square matrices) can be prepared beforehand in a generic way:

- **H** is the forward model and normally exists in a computationally tractable form. Note that **H** may be the tangent-linear code of a non-linear model $H(\mathbf{x})$ where all lines have been analytically derived once and for all, either automatically or by hand.
- the right-multiplication of $\mathbf{H}^T$ with a column vector can be made with the adjoint code of **H** where all lines have been analytically derived and transposed once and for all, either automatically or by hand.
- **B** can be designed in such a way that its inverse is numerically convenient, for instance through singular value decompositions and Kronecker products of correlation matrices in space and time.
- **R** is usually considered to be diagonal. Its potential non-diagonality is the main difficulty of the variational approach, but some solutions exist on a case-by-case basis (e.g., Chevallier 2007).

This approach has been designed to get the mode of the posterior distribution of **x**. It also gives access to the first eigenvectors of the inverse of the covariance of this posterior distribution (Fisher and Courtier, 1995), but if the eigenvalue spectrum converges slowly, this information is of little use (Chevallier et al., 2005). Alternatively, ensembles of variational inversions can be designed to reconstruct the posterior distribution of **x** (Chevallier et al., 2007).

For atmospheric inversion (where **H** is mainly a transport model, or a linearized version of it), the target surface fluxes include the time dimension directly in the control vector **x**, while for the 4D-Var systems designed for Numerical Weather Prediction (next section), the control vector is mainly the state of a model at the initial time step, the later states being obtained through the transport model. We therefore refrain from calling 4D-Var the former, in order to account for the different nature of the estimation problem. However, with the future Copernicus CO_2 support service, the two approaches may merge together in a single system, with equal importance given to the initial state of the atmosphere and to the surface boundary conditions. Examples of current variational global atmospheric inversion systems are CAMS (Section 3.4.2) and Jena-Carboscope (Rödenbeck, 2005).

3.1.2 Online 4DVAR

An online 4D-Var system consists in allowing the variational system described in 3.1.1 to optimize jointly emissions with meteorological variables. In that context the control vector and its associated **B** matrix includes both the CO_2 emissions and prognostics meteorological variables of the numerical weather prediction system, and the forward model ***H(x)*** corresponds to the integration of the equations of motion of the atmosphere together with the transport of atmospheric tracers after emissions (e.g., CO_2). Such a system presents several advantages. For instance, it implicitly accounts for model transport errors associated with uncertainties in initial meteorological conditions. It also enables transport adjustment based on observed CO_2 concentrations in a statistically and dynamically consistent manner. One disadvantage of such technique is the cost associated with the non-quadratic minimization of the variational cost function, as well as the need to define a short assimilation window to mitigate the effect on non-linearities on the convergence performance of the algorithm (i.e., the presence of multiple minima can severely hamper the efficiency of the variational optimization).

3.1.3 Online EnKF

In the papers of Kang et al. (2012) and Liu et al. (2016) a Local Ensemble Transform Kalman Filter (LETKF) DA system for the combined atmospheric state (weather and CO_2 mole fractions) and surface flux (CO_2) was demonstrated. Using a DA window of only 6-hours and an observation network representing a GOSAT + AIRS unbiased satellite view, they were able to retrieve detailed surface fluxes successfully over time-scales of a few days, for a full year.

An important role was played by the background error covariance matrix: it evolved dynamically because of the full atmospheric state, and contained covariances between the weather variables (specifically winds), and transported CO_2 mole fractions, allowing for updates to the CO_2 mole fractions based on extensive weather observations. In turn, error covariance between CO_2 mole fractions and CO_2 fluxes allowed the update of fluxes across all spatial scales contained in the background covariance. This effectively was the synoptic scale (high- and low-pressure areas and associated fronts), thus projecting local CO_2 observation information out to scales much larger than covered in the DA window. Tests presented in 2012 were done at coarse atmospheric transport and flux resolution (T32). In recent years, Environment and Climate Change Canada has worked to develop a similar capacity (Polavarapu et al., 2016), based on the GEM-MACH-GHG DA system at 0.9 degrees global resolution.

3.1.4 Offline EnKF

The offline ENKF for carbon flux estimation was introduced in Peters et al. (2005), where the offline component refers to both (a) transport in the observation operator, and (b) propagation of the state vector. In this type of application, the observation operator is a full atmospheric transport model to turn CO_2 fluxes into simulated mole fractions, driven by offline archived mass-fluxes from a parent model. This approach is efficient because offline transport modelling is a convenient way to reproduce atmospheric tracer transport, without the need to solve the full Navier-Stokes equations nor for full data assimilation of observed weather. This allows mass-conserving advection schemes, including two-way nesting, to be used within a large ensemble (N>100). Moreover, it facilitates large DA windows of several weeks that are impossible in an online DA system. Large windows in turn can include explicit transport between fluxes and mole fractions over large scales, reducing the dependence on a statistical representation of the covariance matrix that is typical for the online systems. A downside of this approach is that propagation of transport errors, as well as of background covariance is not done with a physical state model, but rather with a simple now-casting. This limits the power of the ENKF to improve the state estimate over many consecutive cycles. Like the online ENKF, this approach offers an alternative to the explicit representation of the relation between local CO_2 mole fractions and surface fluxes (i.e. transport Jacobians), which in the community had been leveraged in atmospheric inversions with windows ranging from weeks to years.

3.1.5 Offline analytical

In the presence of Gaussian probability density functions and of a linear forward model, it can also be shown that the most likely values of the control variables can be expressed as:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\,\mathbf{H}^T(\mathbf{H}\,\mathbf{B}\,\mathbf{H}^T + \mathbf{R})^{-1}\,(\mathbf{y} - \mathbf{H}\,\mathbf{x}^b) \qquad (3)$$

or equivalently as:

$$\mathbf{x}^a = \mathbf{x}^b + (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\,\mathbf{H}^T\mathbf{R}^{-1}\,(\mathbf{y} - \mathbf{H}\,\mathbf{x}^b) \qquad (4)$$

Similar expressions exist for the error covariance matrix of $\mathbf{x}^a$ (Tarantola, 2005).

Depending on the respective dimension of $\mathbf{x}$ and $\mathbf{y}$, one may prefer one of the two equivalent formulations. However, in both cases, square matrices have to be multiplied together and one intermediate square matrix needs to be inverted. The way that $\mathbf{H}$ is used also suggests that it has to be stored as a (potentially large) Jacobian matrix before the computation.

Atmospheric inversions have used this approach a lot in the past (e.g., Gurney et al., 2002) until the need for higher resolution inversions (large $\mathbf{x}$) assimilating time series of in situ or satellite observations with little averaging (large $\mathbf{y}$) made the computation of $\mathbf{H}$ problematic. Its algorithmic simplicity and its efficiency, after $\mathbf{H}$ has been obtained, make it still interesting for some applications, like Observing System Simulation Experiments (as is done in WP3 and

WP4 of CHE, or, e.g., in Bousquet et al., 2018). Note that **H** can be obtained in a parallel way by repeated runs of the tangent-linear or adjoint codes of *H*, or by finite-differences perturbations of *H*, or by finite-differences perturbations of *H,* or using Lagrangian transport model (e.g., Pisso et al., 2019).

3.1.6 Online Hybrid Ensemble Variational

An online hybrid ensemble-variational system has been proposed to implement an efficient joint state/source DA system in the IFS, that would allow to both extend the current operational 12-hour assimilation window and to include chemical mechanisms in the variational minimization algorithm (currently only the transport and physical processes are included in the TL/AD for CAMS). The method is based on: 1) hybridization of ensemble information with full-rank statistical modelling by combining an ensemble-based increment with an adjoint-based increment propagation, allowing one to increase current spatial resolution and/or include forward model processes missing from the adjoint integration (for instance, chemical reactions and processes); 2) combination of tangent-linear and adjoint solvers with ensemble-based least-square approximations of transport Jacobians to construct a long-window 4D-Var with timescales relevant to greenhouse gas source inversion. The proposed methodology is non-intrusive in the sense that the main structure of the current IFS incremental 4D-Var algorithm remains unchanged, while the additional computational cost associated with the source inversion component is minimized. This methodology is described in detail in Bousserez (2019) [Tech Memo].

3.2 Control Vector Configuration

The control vector ***x*** of a DA system plays a very important role. It needs to be balanced between the desired detail of information needed, and the observational density available to constrain the problem, while avoiding aggregation errors. To estimate anthropogenic (and biogenic) CO_2 fluxes a number of choices have demonstrated to be successful, being (i) direct flux estimation on a grid, (ii) estimation of model parameters that control the fluxes, and (iii) joint estimation of atmospheric CO_2 state and fluxes.

Whereas (i) offers a reanalyses of the variable targeted for interpretation, (ii) offers a better compromise between observational capacity and the number (and nature) of the unknowns, while (iii) allows atmospheric transport patterns to be included in the background covariance which allows efficient spatial extrapolation of local results. Each of these approaches has been used in published applications, however, they have not, or very scarcely, been applied to estimating anthropogenic CO_2 emissions. Option (iii) is also least developed scientifically because of the large task involved in DA of full atmospheric states, and thus the need to have a NWP centre involved. Some of these properties are summarized in the table below, followed by a more explicit description of the type of flux representations possible.

Table 3: Choice of control vectors for surface emissions and their pros and cons.

Control Vector Choice	Description of state	Pros (+)	Cons (-)
Direct flux estimation	Each emission represents an unknown in a gridded structure	+ Most direct path to target variable, least assumptions	- No process information, requires dense observational network

Parameter optimization	Underlying unknowns in a dynamic model that generates CO_2 fluxes	+ Efficient interpolation of observations to smaller scales + extra observable information can be added (e.g., energy, traffic, and consumer statistics) + Dependency between weather and CO_2 exchange can be leveraged	- Model structural errors need to be handled - Propagation of model parameters often non-physical
Joint atmospheric state and flux/parameter optimization	The 4D structure of the atmospheric CO_2 field is optimized along with the underlying emissions of CO_2	+ Representation of large-scale atmospheric structures in background covariance possible + Allows short-windows through statistical relation between fluxes and mole fractions	- Requires full NWP setting, only a few groups can deliver - Mass-conservation can be difficult to maintain - Approximation of large background covariance structure with small ensemble number, how well does it work?

Generally, there is consensus that to estimate anthropogenic emissions from satellite observations, the state vector needs to keep track of the full atmospheric 4D mole fractions of CO_2, but also contain information about auxiliary tracers and their relation to the emissions. Because anthropogenic emissions will occur at scales below those resolved by the global transport models we employ (at least for the next decade), it is also likely that we need them to be represented not as individual flux elements, but as the result of a dynamical model that generates emission estimates based on activity data (also observable), emission factors (measurable), and variations in driver data (temperature, wind, solar radiation). In that case it is also important to include a bias correction term to account for structural model errors. Currently, the optimal stratification in space/time of model parameters is not yet known, but being investigated in projects such as VERIFY.

3.2.1 Direct flux estimation (NWGR)

When fluxes are placed in the state vector, each element either needs to be observed directly, or its value must be inferred from nearby flux elements that have a co-varying error, as carried in the background error covariance matrix. Typically flux error covariances only span a small spatiotemporal scale though (Chevallier et al., 2007) leaving many degrees of freedom. Because fluxes often change rapidly over time (diurnal cycles of for example traffic emissions, heat generation, and biosphere fluxes) this challenge becomes even larger. Moreover, the representation of anthropogenic emissions, often small point sources, is difficult in a global model with 5-10 km resolution, rendering direct flux estimation for anthropogenic emissions impractical. Since this approach has a very long history in the inverse modelling community that came from continental scale and annual flux estimates, much experience was built around this concept in the last two decades.

3.2.2 Dynamic biosphere flux description model for CCDAS (NWGP, NWGR)

An example of this building block is VPRM or another simple model that feeds off NWP weather variables and/or easily available remote sensing data (LAI, faPAR, SIF, NIRv, NDVI,…). It provides the biosphere exchange at the surface and focuses on the shorter time scales where weather variations cause much of the variability that manifests itself nearly as "noise" for the data assimilation problem. Since this noise is not observable everywhere all the time, the CCDAS allows this to be captured by tuning simple model parameters across space and time. Examples are the light-use efficiency of plants, or the water-use efficiency, or the stomatal VPD-response, or the Q10 factor for respiration. Since these can vary stochastically and respond to the NWP variables, both noise and uncertainty can be captured in CCDAS, and moreover the forecast-analysis cycle of NWP allows for updates to the model over time. For NWGR, this biosphere exchange model could include longer relevant time scales of weeks to perhaps even decades. An important challenge is to propagate information for time-varying parameters through consecutive states, as needed for the DA system: the lack of a dynamical model with realistic error (growth and propagation) puts limits on our ability to keep meaningful structures in space/time, which are needed for the envisioned statistical methods in the hybrid ensemble-variational system to work. For more sophisticated surface flux models, initial values for example for carbon pool sizes can furthermore be included in the state. Examples of recent CCDAS efforts include: Koffi et al., 2012, Santaren et al., 2014, Kaminski et al., 2017 (within CHE).

3.2.3 Dynamic anthropogenic flux description model for FFDAS (NWGP, NWGR)

Similar to the CCDAS, the FFDAS model provides anthropogenic surface fluxes as a function of weather variables from the NWP, as well as other recorded variables such as activity of traffic, power plants, ships, or industry over space and time. The relation between activity and emissions can then be optimized, and is a relevant variable that relates directly to the uncertainties also in national emission reports. Since the underlying parameters can vary stochastically and respond to the NWP variables, both noise and uncertainty can be captured in FFDAS, and moreover the forecast-analysis cycle of NWP allows for updates to the model over time. For NWGR, this anthropogenic exchange model could include longer relevant time scales of weeks to perhaps even decades. Auxiliary remote sensing data such as NO_2 and CO columns or nightlight data can furthermore be incorporated readily in this system. Examples of recent FFDAS efforts include: Brophy et al, (2018) , Super et al. (2019), Asefi-Najafa et al. (2014).

3.3 Error Covariance Statistics

Uncertainty representation needs to be carefully considered in the operational prototype. Two critical components of the inversion systems are the prior error covariance and transport (or forward model) error covariance matrices.

Prior flux uncertainties will be first based on the available knowledge from state-of-the-art bottom-up inventories. However, given the high level of uncertainty in those estimates, further adjustments for the prescribed prior error covariance matrix will be needed. In particular, flux error correlations are poorly known in current bottom-up inventories. Within CHE, we will leverage available wavelet-based modelling tools in the IFS to construct an efficient model for the prior error covariance matrix that can account for spatially heterogeneous error correlation structure. The parameters of this **B** matrix will be optimized based on observed CO_2 atmospheric concentrations, using standard high-dimensional adjoint-based optimization techniques.

In the CC-FF-DAS approach uncertainty on the model parameters must be based as much as possible on the error structure of the underlying data. This consists of (a) Activity data (traffic counts, energy demand, population and livestock density, productivity of factories, etc), and (b) the transfer parameters from Activity to emissions (emission factors, emission ratios, temporal profiles, etc). The latter are often based on literature and laboratory measurements (for vegetation) or from laboratory and in-situ measurements (e.g., emission factors of coal burning). An important aspect mentioned in Table 3 is that both (a) and (b) can be weather-dependent: total energy consumption depends on outside temperature, the need for fossil energy depends on the availability of wind and sunlight for renewable, but also emission factors of traffic are temperature-dependent (cold starts). This means that the **B** matrix can have covariances between such NWP variables and anthropogenic & natural fluxes.

The transport error covariance matrix will be estimated through the use of ensemble information within a Monte-Carlo framework (McNorton et al., submitted). However, such estimations are associated with the forward model parameters uncertainty, and several, non-parametric, additional sources of transport errors remain. Therefore, additional benchmarking based on high-resolution model simulations and in situ comparisons will be necessary to capture model errors not accounted for by the standard ensemble-based methods. More information on the methods to estimate transport error statistics are provided in the CHE WP5 report on uncertainty quantification (CHE D5.7).

Finally, due to the large dimension of the inverse problem, an efficient computational approach is required to approximate the posterior error covariance matrix and related information content diagnostics for the estimated CO_2 fluxes. To this aim, low-rank approximations of the posterior error covariance and model resolution matrices will be used as described in the CHE WP5 report on uncertainty quantification (CHE D5.7). These can be compared to full-rank estimates like provided in the system of Rödenbeck et al. 2005.

3.4 Examples of Existing Inversion Systems

As referred to several times above, relevant experience with DA systems for CO_2 fluxes is available in the scientific community, albeit not at the scale targeted with the prototype MVS. Below, several of the world-wide acknowledged systems are mentioned in the table, and expanded in the text to provide context to the current state-of-the-art.

Table 4: State-of-the-art data-assimilation systems for Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). Important pros and cons are listen

Name or acronym	*NWGP*	*NWGR*	*Details*	+	-
IFS	☑	☑	12-24 hour window with propagation of information from outside the window, demonstrated for CO_2 atmospheric state	Versatile, shared with NWP	Short assimilations window & Strong dependence on B-matrix and small ensemble for large scales
Carbon-Tracker		☑	Long-window Ensemble Kalman	Explicit representations of intermediate	Max resolution offline and CPU- scalability limited. No transport errors

			Smoother system, demonstrated for decadal CO_2 flux reanalyses including satellite XCO_2 and sun-induced fluorescence (**CHE outcome**)	scales, mass conserving	
CAMS		✓	Full-window 4D-VAR system, demonstrated for CO_2 flux reanalyses including satellite XCO_2 and sun-induced fluorescence (**CHE outcome**)	Explicit representations of intermediate +long scales, mass conserving	Max resolution offline and CPU- scalability limited. No transport errors
Satellite Mass-Balance Approach	✓	✓	Direct estimation of point source strengths from satellite images using simple (mass-balance, **CHE outcome)**	Direct approach using limited additional information	Emission estimates depend critically on wind-speed. Complicated for multiple sources.
CC-FF-DAS	✓	✓	Parameters from flux models are optimized	Allows easy coupling to NWP models, enables the use of further observational constraints (SIF, FAPAR, nightlights, activity/census data …)	Error propagation difficult, non-linearity of underlying system hard to include in Gaussian minimizations

3.4.1 CarbonTracker

The CarbonTracker data assimilation system for CO_2 (Peters et al., 2005) was the first of its kind to make use of ensembles of fluxes, and the square-root sequential filter described in Whitaker and Hamill (2001). Directly from its first launch in 2006, it moreover used a large number of global regions (more than 200, later replaced by a gridded state vector) and a weekly time step, providing many more degrees of freedom than systems used before. Its success in assimilating data from the global CO2 monitoring networks, and especially from semi-continuous tower observations contributed to the wide use of its flux products and atmospheric mole fractions in the community. CarbonTracker has since forked into two development branches, one at NOAA ESRL in the USA, and one in Europe (Peters et al., 2010) which was later incorporated in the python-based CTDAS shell (Laan-Luijkx et al., 2017). CT Europe uses 150 ensemble members to represents the spread of a set of ~9000 linear flux scaling factors across the globe, which are updated every week. Their error covariances scale across distance and across terrestrial ecosystem types, reducing the weekly degrees of freedom to ~1200. It moreover uses a five-week lagged smoother window to allow observations to change scaling factors even at longer travel times of the observed air masses. The system is typically run with a highest resolution of 1x1 degree, and covers two decades in its reanalysis. The TM5 model, driven by offline mass-conserving and positive definite mass-fluxes from the IFS reanalyses (ERA5, and ERA-I) provides transport, and is parallelized to scale well up to hundreds of CPUs, making IO of offline fluxes one of its computational bottlenecks.

3.4.2 CAMS inversion system

For 10 years, the operational Copernicus Atmosphere Monitoring Service (https://atmosphere.copernicus.eu/) and its precursor projects Monitoring Atmospheric Composition and Climate have been analysing CO_2, methane and nitrous oxide surface fluxes over the recent decades and over the globe by assimilating near-surface or column mole fraction observations in global atmospheric chemistry-transport models. In the case of CO_2, the system minimizes a Bayesian cost function to optimize the $3.75^o \times 1.9^o$ grid-cell eight-day surface fluxes over the globe (with a distinction between local night-time fluxes and daytime fluxes, but without fossil fuel emissions, that are prescribed) and the initial state of CO_2. To do so, it assimilates a series of CO_2 atmospheric observations over a given time window within an off-line version of the global atmospheric general circulation model of the Laboratoire de Météorologie Dynamique (LMDz, Hourdin et al., 2013) run at global resolution $3.75^o \times 1.9^o$ and nudged towards ECMWF re-analyses The minimization approach is called 'variational' because it explicitly computes the gradient of the cost function using the adjoint code of LMDz (Chevallier et al. 2005). Thanks to a double parallelisation of the transport model (Chevallier 2013), it allows the inversion window to extend over several decades (currently four) seamlessly, while still producing and delivering the CAMS CO_2 inversion twice per year. Corresponding Bayesian uncertainty statistics are available on request, based on Monte Carlo simulations (Chevallier et al., 2007). Prior information about the surface fluxes is provided to the Bayesian system by a combination of climatologies and other types of measurement-driven flux estimates. Assigned prior error variances vary in space and time, and are associated to temporal and spatial error correlations. Details can be found in Chevallier (2018a). The main CAMS product assimilates near-surface measurements, but a satellite-driven product is now also available. Details about both products can be found in Chevallier (2019).

3.4.3 IFS inversion system

As part of CHE and preceding efforts a global NWGP system was built, based around the IFS model at ~9km spatial resolution. It currently uses a 4D-Var window length of 12-hours, but efforts are well underway to adopt a hybrid ensemble-variational method. This will allow ensemble information to extend the current operational 12-hour window to past and future assimilation windows [Tech Memo N. Bousserez (2019)]. Currently the system has been demonstrated to improve the atmospheric state analysis and 10-day forecast of CO_2 concentrations, but fluxes have not yet been estimated. CHE ensemble nature runs performed with IFS confirm that the signal from the biogenic fluxes is dominant over anthropogenic signals in many areas, and at the resolutions being considered. Therefore, it is deemed critical to: 1) include the biospheric fluxes or CCDAS parameters in the control vector; 2) consider the potential of co-emitters (e.g., NO_2, CO) in disentangling the anthropogenic and biospheric flux signal. Recommendations for next steps in its development are provided in the next section.

3.4.4 Satellite Mass Balance methods

Local point-source emissions, or even city-domes, show strongest signals at scales below what can be resolved by the global or regional systems. Tailor-made approaches to estimate their source strength are needed. Available methodologies include various mass-balance approaches, Gaussian plume modeling, Lagrangian dispersion modeling, and increasingly also machine-learning techniques. Within a DA framework, such approaches can be used as a separate building block to monitor local sources separately, or they can be integrated in the parent global IFS system, e.g. as a plume-in-grid-approach. In the latter, each system would profit from the two-way interactions of meteorology and concentrations. The point-source estimation approach should be scalable from doing a limited (top-N, with N=50-100) number of large emitters, to a system where the majority of global (observed) point-sources monitored.

University of Bremen is developing an inversion system aiming at the quantification of CO_2 emissions from localized CO_2 emission sources such as power plants and cities using satellite retrievals of column-average dry-air mole fractions of CO_2 (XCO_2) and NO_2 tropospheric columns (Reuter et al., 2019). For satellite XCO_2 the Level 2 XCO_2 product form NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite is used. However, regional column-average enhancements of individual point sources are usually small, compared to the background concentration and its natural variability, and often not much larger than the satellite's measurement noise. This makes the unambiguous identification and quantification of anthropogenic emission plume signals challenging. NO_2 is co-emitted with CO_2 when fossil fuels are combusted at high temperatures. NO_2 has a short lifetime on the order of hours so that NO_2 columns often greatly exceed background and noise levels of modern satellite sensors near sources, which makes it a suitable tracer of recently emitted CO_2. Based on six case studies (Moscow, Russia; Lipetsk, Russia; Baghdad, Iraq; Medupi and Matimba power plants, South Africa; Australian wildfires; and Nanjing, China), Univ. Bremen demonstrated the usefulness of simultaneous satellite observations of NO_2 and XCO_2. For this purpose, they analyzed co-located regional enhancements of XCO_2 observed by OCO-2 and NO_2 from the Sentinel-5 Precursor (S5P) satellite and estimate the CO_2 plume's cross-sectional fluxes taking advantage of the nearly simultaneous NO_2 measurements with S5P's wide swath and small measurement noise by identifying the source of the observed XCO_2 enhancements, excluding interference with remote upwind sources. This allows to adjust the wind direction and constrains the shape of the CO_2 plumes. They compare the inferred cross-sectional fluxes with the Emissions Database for Global Atmospheric Research (EDGAR), the Open-Data Inventory for Anthropogenic Carbon dioxide (ODIAC), and, in the case of the Australian wildfires, with the Global Fire Emissions Database (GFED). The inferred cross-sectional fluxes range from 31 $MtCO_2$/year to 153 $MtCO_2$/year with uncertainties (1σ) between 23 % and 72 %. For the majority of analyzed emission sources, the estimated cross-sectional fluxes agree, within their uncertainty, with either EDGAR or ODIAC or lie somewhere between them. They assessed the contribution of multiple sources of uncertainty and found that the dominating contributions are related to the computation of the effective wind speed normal to the plume's cross section. The flux uncertainties are expected to be significantly reduced by the planned European Copernicus anthropogenic CO_2 monitoring mission (CO2M), which will provide not only precise measurements with high spatial resolution but also imaging capabilities with a wider swath of simultaneous XCO_2 and NO_2 observations.

An example of a fast mass-balance method, aided by fine-scale transport modeling, was given by Wu et al., (2018), using the column X-STILT Lagrangian dispersion model. In this system, the difference between up- and downwind XCO_2 observations ($DXCO_2$) from OCO-2 was projected onto the city in between, using X-STILT footprints. This footprint determined the sensitivity of each height represented in the column to the detailed city emission landscape below (resolved at high resolution), and thus allowed the $DXCO_2$ to correctly scale underlying emissions.

3.4.5 CCFFDAS

The Carbon Cycle Fossil Fuel Data Assimilation approach tackles the inverse problem of inferring surface fluxes of trace gases through assimilation of atmospheric and other observations into process models simulating the surface fluxes. The atmospheric transport acts as (part of the) observation operator for the atmospheric observations (Kaminski and Mathieu, 2017). The control space can be any combination of process parameters, initial- and boundary conditions of these process models and the observation operators. In the weak-constraint (Zupanski, 1997) setup, deviations from the simulated fluxes are added to the control vector (see, e.g., Lewis et al., 2012). Fluxes are then simulated from the posterior control vector. Background on the approach is provided, e.g. by Rayner et al. (2010), Kaminski et al. (2013), and Asefi-Najafabady et al. (2014). We highlight the following aspects of the approach:

- An appropriate choice of the process models implicitly addresses the attribution to flux categories (e.g. sectorial fossil fuel emissions or natural fluxes).
- The process models can be regarded as a way to bring extra information into the assimilation system, in particular if (with appropriate observation operators) they can enable the use of further observations (e.g. nightlights, SIF, or FAPAR) or variables that are provided by the operational system (e.g. to simulate fossil fuel emissions from heating and cooling).
- The (linearised) process models implicitly provide the uncertainty structure in flux space. In this respect the role of the process models is analogous to the role of observation operators for lower level EO products, the uncertainty structure of which is usually much easier. Residual errors in the process models can be addressed by the weak-constraint version. More accurate process model will serve better to remove uncertainty from the inverse problem. This is because they provide lower forecast residual errors, and thus allow lower prior uncertainties. Owing to the non-linearity of the process models, the uncertainty structure in the flux space changes in the course of the minimisation (in contrast to the B matrix in an atmospheric inversion that directly solves for fluxes).

A “light” prototype of a global-scale CCFFDAS was developed in the ESA CCFFDAS study (http://ccffdas.inversion-lab.com/) based on existing components, i.e. the Fossil Fuel Data Assimilation System by Asefi-Najafabady et al. (2014), the Carbon Cycle Data Assimilation System by Kaminski et al. (2017), and the Atmospheric Transport model TM3 (Heimann and Koerner, 2003). The prototype is being applied in CHE by Lund and iLab to explore design choices for the operational MVS. A particular asset is the availability of a full Jacobian representation of the modelling chain. The full Jacobian allows, for example, to assess approximations of posterior uncertainties in low-dimensional subspaces of the control space. A limitation is the coarse resolution of the atmospheric transport (4 by 5 degrees) that is currently used, even though an update of this global system to higher resolution is feasible. A regional CCFFDAS prototype is currently being developed within the same ESA study.

4 Recommendations for operational CHE prototype

- ★ *Details of different configurations/streams can be tested within the next three years to address different temporal/spatial scales and user needs*
- ★ *Incremental step to implementation (emphasizing the added value of each step).*

Each scale on which we need to develop a prototype MVS system is associated with different configurations, and different needs. These require different, but closely coordinated, development strategies.

4.1 Global inversions

Table 5: Immediate development needs linked to the domain (global) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort (Person Months)
IFS	Global	NWGP	Facilitate use of IFS model to other partners, including capacity to “replay” the forecast ensemble from EPS (online, not through static-fields and massive I/O)	12 months
IFS hybrid ensemble-4dvar	Global	NWGP, NWGR	Demonstrate joint state-flux estimation including propagation of information from outside the 12-hour window. Assimilate high-frequency near-real time observations from satellites and surface network. Use OSSEs, as well as real-world test with CO as tracer.	9 months
IFS hybrid ensemble-4dvar	Global	NWGP, NWGR	Develop long-window inversions using a hybrid ensemble-variational method, estimating joint biospheric/anthropogenic fluxes in a CC-FF-DAS approach.	6 months
IFS hybrid ensemble-4dvar	Global	NWGP, NWGR	Through OSSEs, assess the requirements, and statistical performance of the ensemble in the DA framework: is the	6 months

			linearization good enough to allow propagation of the state across time scales from hours-to-weeks? How much localization is needed and how to apply it in space/time? Is the low-dimensional (or wavelet) representation of the B-matrix sufficient to assess posterior errors? Do these methods work for long-window re-analyses too?	
IFS multi-scale assimilation	Global	NWGR	Show the feasibility of the multi-scale integration of a reanalysis with 2-3 systems (CTE, CAMS, IFS)	6 months
CarbonTracker	Global	NWGR	Build a hybrid long-window-short-window assimilation system, with consistent propagation of covariances.	6 months, planned under CHE
CarbonTracker	Global	NWGR	Improve numerical satellite-assimilation capacity to handle larger volumes of data.	2 months
CarbonTracker	Global	NWGR	Replace offline meteo with IFS online alternative (+analysis replay-mode)	12 months, partly started for EC-Earth
CAMS	Global	NWGR	Identify and implement innovations to estimate anthropogenic fluxes in re-analysis mode	12 months
CCDAS	Global to regional	NWGP, NWGR	Identify suitable model, and optimizable parameters, decide how to approach slow changes and hysteresis in carbon fluxes. Design a dynamical model for the propagation of the mean state and covariance. Build TL/AD codes.	12 months
FFDAS	Global to regional	NWGP, NWGR	Identify suitable model, and optimizable parameters, decide how	12 months

			to approach. Decide how to approach point-sources and integration into IFS. Design a dynamical model for the propagation of the mean state and covariance. Build TL/AD codes.	
Satellite Data	Global	NWGP, NWGR	Assimilate satellite data to estimate emissions. Real-case scenarios using CO system & TROPOMI. Compare NWGP and NWGR mode and document requirements + path forward for CO_2.	8 months

4.2 Regional inversion

Table 6: Immediate development needs linked to the domain (regional) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort (Person Months)
COSMO	Regional	NWGR	Investigate use of boundary conditions within regional inversions capabilities	6 months
Testbed for components	Regional	NWGR	Identify testbed regions, invest in interface to integrate with NWGP and NWGR within Copernicus, incorporate ICOS resources to facilitate continuous exchange of information	8 months

4.3 Local inversions

Table 7: Immediate development needs linked to the domain (local) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort

				(Person Months)
Bremen approach	Local	NWGR	Investigate application for CO2M mission	6 months
Plume-in-grid	Local	NWGP, NWGR	Investigate split in state-vector: large point-sources are estimated separately and merged with the NWGP & NWGA systems.	9 months
Plume Chemistry	Local	NWGR	Investigate effects of non-linear atmospheric chemistry on NO_2 lifetime	6 months

5 Research priorities

5.1 Multi-scale Integration System

A proposed multi-model system could be used to integrate spatiotemporally heterogeneous posterior emission products. This system, outlined by Bousserez (2019,Tech Memo), would treat the local and regional posterior flux products as observations in a global IFS-driven CO_2 inversion. In practice, each regional and local inversion outputs to be assimilated in the global multi-model product would be required to provide an ensemble of prior and posterior samples of 4D CO_2 emissions and CO_2 concentrations fields. In order to avoid any detrimental effects from the integration of poorly estimated posterior emissions and/or inaccurately prescribed posterior errors on the multi-model product, a strict quality control mechanism will need to be implemented. The complexity of assimilating inversion products across different spatiotemporal scales in consistent manner may require an efficient integration tool similar to CIF (VERIFY), in particular to standardize model inputs/outputs. Within the IFS global model, the multi-model assimilation algorithm will be implemented using the modular OOPS DA system.

Table 8: Research priorities linked to the domain (global, regional, local) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort (Person Months)
Transport	Global Regional	NWGP, NWGR	Test new transport schemes developed in NWP, e.g. MPDATA advection in FVM IFS (Kühnline et al, 2019)	8 months

5.2 Multi-species constraints (NO_2, CO, CO_2)

The multi-species constraints are a necessity driven by the CO_2 underdetermined data assimilation problem, a number of recommendations for future research especially on turbulent scales that is largely unchartered territory.

One important recommendation for all scales, and all systems, is to start investing in multi-species data assimilation. This is based on the recognition that anthropogenic CO_2 emissions can never completely be constrained with CO_2 observations alone, and the signal-to-noise of co-emitted species is often much better than that of CO_2 (and especially XCO_2). Global, local, and regional scale DA systems so far have only focused on one or two species simultaneously. A large leap is needed.

Table 9: Research priorities linked to the domain (global, regional, local) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort (Person Months)
Chemistry for Multi-species approach	Global Regional	NRT RA	NO_2, CO, VOCs via reduced complexity models, AI, multi-tracer approach	6 months
Use LES and DNS for error covariances characterisation	Local/Regional	NRT RA	Urban CO_2 distribution for error characterisation and representativity	4 months

5.3 Assimilation of satellite data to estimate surface fluxes

Operational weather centers usually focus on the atmospheric state rather than on surface fluxes. To gain experience in estimating surface fluxes in the proposed hybrid DA system, CO surface flux estimation using S5P TROPOMI data should receive high priority

Table 9: Research priorities linked to the domain (global, regional, local) and stream for application in the prototype: Numerical Weather and Greenhouse Gas prediction (NWGP) and Reanalysis (NWGR). An estimate of the effort required is given in person months.

Component	Domain	Stream	Recommendation	Estimated effort

				(Person Months)
Assimilation of TROPOMI data	Global	NWGP, NWGA	Use CO as a tracer to test the hybrid DA system	8 months

6 Conclusions

The specification of the configuration recommended for the data assimilation components at different scale is provided following the work done in WP1.

7 References

Agustí-Panareda, *et al*., 2016. "A Biogenic CO2 Flux Adjustment Scheme for the Mitigation of Large-Scale Biases in Global Atmospheric CO2 Analyses and Forecasts." *ACP* 16 (16). Copernicus Publications: 10399–418. doi:10.5194/acp-16-10399-2016.

Anderson, J. L. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A* **61,** 72–83 (2009).

Asefi-Najafabady, S., *et al* : A multiyear, global gridded fossil fuel CO2 emission data product: Evaluation and analysis of results, Journal of Geophysical Research: Atmospheres, 119, 10,213–10,231, https://doi.org/10.1002/2013JD021296, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD021296, 2014.

Bousserez, Nicolas. 2019. "Towards a Prototype Global CO2 Emissions Monitoring System for Copernicus." https://arxiv.org/abs/1910.11727

Bousquet, P., C. *et al*., (2018): Error budget of the MEthane Remote LIdar missioN (MERLIN) and its impact on the uncertainties of the global methane budget. J. Geophys. Res. Atmos., doi:10.1029/2018JD028907.

Brophy, K. *et al.* Characterizing Uncertainties in Atmospheric Inversions of Fossil Fuel CO2; Emissions in California. *Atmos Chem Phys Discuss* 1–44 (2018). doi:10.5194/acp-2018-473

CHE D5.1: Progress report on service elements for CO_2 Earth observation integration

CHE D5.3: Progress report on service elements for CO2 emission and transport models integration

CHE D5.5: Progress report on service elements for data assimilation methodology

CHE D5.7: Progress report on service element requirements for uncertainty representation.

Chevallier, F.: On the parallelization of atmospheric inversions of CO2 surface fluxes within a variational framework, Geosci. Model Dev., 6, 783–790, https://doi.org/10.5194/gmd-6-783-2013, 2013.

Chevallier, F., *et al*., 2005: Inferring CO_2 sources and sinks from satellite observations: method and application to TOVS data. J. Geophys. Res., 110, D24309, doi:10.1029/2005JD006390.

Chevallier, F., 2007: Impact of correlated observation errors on inverted CO_2 surface fluxes from OCO measurements, Geophys. Res. Lett., 34, L24804, doi:10.1029/2007GL030463

Chevallier, F., *et al*, 2007: The contribution of the Orbiting Carbon Observatory to the estimation of CO_2 sources and sinks: Theoretical study in a variational data assimilation framework. J. Geophys. Res., 112, D09307, doi:10.1029/2006JD007375

Chevallier, F., *et al*., 2019: Objective evaluation of surface- and satellite-driven CO2 atmospheric inversions, Atmos. Chem. Phys., accepted.

Gurney, K. R., *et al*.: Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, Nature, 415, 626–630, https://doi.org/10.1038/415626a, 2002.

Heimann, M. and Körner, S.: The global atmospheric tracer model TM3, Tech. Rep. 5,Max-Planck-Institut für Biogeochemie, Jena, Germany, 2003.

Hourdin, F *et al*., Impact of the LMDZ atmospheric grid configuration on the climate and sensitivity of the IPSL-CM5A coupled model, Clim. Dynam., 40, 2167–2192, https://doi.org/10.1007/s00382-012-1411-3, 2013.

Houtekamer, P. L. & Mitchell, H. L. Ensemble Kalman filtering. *Q J Roy Meteor Soc* **131,** 3269–3289 (2005).

Kaminski, T.T., *et al*., 2013. The BETHY/JSBACH carbon cycle data assimilation system: experiences and challenges. J. Geophys. Res. Biogeo. 118 (4), 1414–1426. http://dx.doi.org/10.1002/jgrg.20118.

Kaminski, T. and Mathieu, P.-P.: Reviews and syntheses: Flying the satellite into your model: on the role of observation operators in constraining models of the Earth system and the carbon cycle, Biogeosciences, 14, 2343–2357, https://doi.org/10.5194/bg-14-2343-2017, URL http://www.biogeosciences.net/14/2343/2017/, 2017.

Kaminski, T., *et al*.: Constraining a terrestrial biosphere model with remotely sensed atmospheric carbon dioxide, Remote Sensing of Environment, pp. 109 – 124, https://doi.org/http://dx.doi.org/10.1016/j.rse.2017.08.017, URL http://www.sciencedirect.com/science/article/pii/S0034425717303838, 2017.

Kang, J.-S., Kalnay, E., Miyoshi, T., Liu, J. & Fung, I. Estimation of surface carbon fluxes with an advanced data assimilation methodology. *J Geophys Res* **117,** n/a–n/a (2012).

Koffi, E. N., *et al*.,. Atmospheric constraints on gross primary productivity and net ecosystem productivity: Results from a carbon-cycle data assimilation system. *Global Biogeochemical Cycles* **26,** (2012).

Laan Luijkx, I. T. *et al.* The CarbonTracker Data Assimilation Shell (CTDAS) v1.0: implementation and global carbon balance 2001–2015. *Geosci Model Dev* **10,** 2785–2800 (2017).

Lewis, P. E., J. *et al*., (2012), An Earth Observation Land Data Assimilation System (EO-LDAS), Remote Sens. Environ., 120, 219–235.

Liu, J., *et al*. Comparison between the Local Ensemble Transform Kalman Filter (LETKF) and 4D-Var in atmospheric CO 2flux inversion with the Goddard Earth Observing System-Chem model and the observation impact diagnostics from the LETKF. *J Geophys Res Atmos* **121,** 13,066–13,087 (2016).

Peters, W. *et al.* An ensemble data assimilation system to estimate CO2 surface fluxes from atmospheric trace gas observations. *J Geophys Res* **110,** D24304 (2005).

Peters, W. *et al.* Seven years of recent European net terrestrial carbon dioxide exchange constrained by atmospheric observations. *Global Change Biology* **16,** 1317–1337 (2010).

Pisso, I., Patra, P., Takigawa, M. *et al*. Assessing Lagrangian inverse modelling of urban anthropogenic CO2 fluxes using in situ aircraft and ground-based measurements in the Tokyo area. Carbon Balance Manage 14, 6 (2019) doi:10.1186/s13021-019-0118-8

Polavarapu, S. M. *et al.* Greenhouse gas simulations with a coupled meteorological and transport model: the predictability of CO 2. *Atmos. Chem. Phys.* **16,** 12005–12038 (2016).

Rayner, P.J.P.J., *et al*., 2010. A new global gridded data set of CO2 emissions from fossil fuel combustion: methodology and evaluation. J. Geophys. Res. Atmos. 115 (D19). http://dx.doi.org/10.1029/2009JD013439. D19306 n/a–n/a.

Reuter, M. *et al.* Towards monitoring localized CO2 emissions from space: co-located regional CO2 and NO2 enhancements observed by the OCO-2 and S5P satellites. *Atmos. Chem. Phys.* **19,** 9371–9383 (2019).

Rödenbeck, C.: Estimating CO2 sources and sinks from atmospheric mixing ratio measurements using a global inversion of atmospheric transport, no. 06, Max Planck Institute for biogeochemistry Technical Reports, Max Planck Society, 2005

Santaren, D., *et al*., Ecosystem model optimization using in situ flux observations: benefit of Monte Carlo versus variational schemes and analyses of the year-to-year model performances. *Biogeosciences* **11,** 7137–7158 (2014).

Super, I., *et al*. Optimizing a dynamic fossil fuel CO2 emission model with CTDAS (v1.0) for an urban area using atmospheric observations of CO2, CO, NOx, and SO2. *Geosci Model Dev* 1–38 (2019). doi:10.5194/gmd-2019-283

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, Society for Industrial and Applied Mathematics, 89, 339 pp., 2005.

Whitaker, J. & Hamill, T. Ensemble data assimilation without perturbed observations. **130,** 1913—1924 (2001).

Wu, D. *et al.* A Lagrangian approach towards extracting signals of urban CO2 emissions from satellite observations of atmospheric column CO2 (XCO2): X-Stochastic Time-Inverted Lagrangian Transport model (‘X-STILT v1’). *Geosci Model Dev* **11,** 4843–4871 (2018).

Zupanski, D.: A general weak constraint applicable to operational 4DVAR data assimilation systems, Mon. Weather Rev., 125, 2274–2292, 1997.

Document History

Version	Author(s)	Date	Changes
V0.1	Peters & Krol	04/12/2019	First draft for review
V1.0	Peters & Krol	18/12/2019	Comments from 2 reviewers incorporated, all textual

Internal Review History

Internal Reviewers	Date	Comments
I Pisso (NILU)	15/12/2019	Provided review and assessment
T Kaminski (iLab)	15/12/2019	Provided review and assessment

Estimated Effort Contribution per Partner

Partner	Effort
WU	8 PM
ECMWF	2 PM
Total	**10**