Variational regional inverse modeling of reactive species emissions with PYVAR-CHIMERE

Audrey Fortems-Cheiney, Isabelle Pison, Gaelle Dufour, Grégoire Broquet, Antoine Berchet, Elise Potier, Adriana Coman, Guillaume Siour, and Lorenzo Costantino

1Laboratoire des Sciences du Climat et de l’Environnement, LSCE-IPSL (CEA-CNRS-UVSQ), Université Paris-Saclay, 91191 Gif-sur-Yvette, France.
2Laboratoire Interuniversitaire des Systèmes Atmosphériques, UMR CNRS 7583, Université Paris Est Créteil et Université Paris Diderot, Institut Pierre Simon Laplace, Créteil, France.

Abstract

Up-to-date and accurate emission inventories for air pollutants are essential for understanding their role in the formation of tropospheric ozone and particulate matter at various temporal scales, for anticipating pollution peaks and for identifying the key drivers that could help mitigate their emissions. This paper describes the Bayesian variational inverse system PYVAR-CHIMERE, which is adapted to the inversion of reactive species. Complementarily with bottom-up inventories, this system aims at updating and improving the knowledge on the high spatio-temporal variability of emissions of air pollutants and their precursors. The system is designed to use any type of observations, such as satellite observations or surface stations. The potential of PYVAR-CHIMERE is illustrated with one-day inversions of CO and NO2 emissions in Europe, using the MOPITT and OMI satellite observations (for CO and for NO2, respectively).

1. Introduction

The degradation of air quality is a worldwide environmental problem: 91% of the world’s population have breathed polluted air in 2016 according to the World Health Organization (WHO), resulting in 4.2 millions of premature deaths every year [WHO, 2016]. The recent study of Lelieveld et al. [2019] even suggests that the health impacts attributable to outdoor air pollution are substantially higher than previously assumed (with 790,000 premature deaths in the 28 countries of the European Union against the previously estimated 500,000 [EEA, 2018]). The main regulated primary (i.e. directly emitted in the atmosphere) anthropogenic air pollutants are carbon monoxide (CO), nitrogen oxides (NOx =NO+NO2), sulfur dioxide (SO2), ammonia (NH3), volatile organic compounds (VOCs ), and primary particles. These primary gaseous air pollutants are precursors of secondary (i.e. produced in the atmosphere through chemical reactions) pollutants such as ozone (O3) and Particulate Matter (PM), which are also threatening to both human health and ecosystems. Monitoring concentrations and quantifying emissions are still challenging and limit our capability to forecast air quality to warn population and to assess i) the exposure of population to air pollution and ii) the efficiency of mitigation policies.
Bottom-up (BU) inventories are built in the framework of air quality policies such as The Convention on Long-Range Transboundary Air Pollution (LRTAP) for air pollutants. Based on national annual inventories, research institutes compile gridded global or regional, monthly inventories (mainly for the US, Europe and China) with a high spatial resolution (currently regional or city scale inventories are typically finer than 0.1°x0.1°). These inventories are constructed by combining available (economic) statistics data, from different detailed activity sectors, with the most appropriate emission factors (defined as the average emission rate of a given species for a given source or process, relative to the unit of activity). It is important to note that the activity data (often statistical data) has an inherent uncertainty and its reliability may vary between countries or regions. In addition, the emission factors bear large uncertainties in their quantification [Kuenen et al., 2014; EMEP/EEA, 2016; Kurokawa et al., 2013]. Moreover, these inventories are often provided at the annual or monthly scale with typical temporal profiles to build the weekly, daily and hourly variability of the emissions. The combination of uncertain activity data, emission factors and emission timing can be a large source of uncertainties, if not errors, for forecasting or analyzing air quality [Menut et al., 2012]. Finally, since updating the inventories and gathering the required data for a given year is costly in time, manpower and money, only a few institutes have offered estimates of the gaseous pollutants for each year since 2011 (i.e., EMEP updated until the year 2017, MEIC updated until the year 2017 to our knowledge). Nevertheless, using knowledge from inventories and air quality modelling, emissions could have been mitigated. For example, from 2010 to nowadays, emissions in various countries have been modified and/or regional trends have been reversed (e.g., the decrease of NOx emissions over China since 2011 [de Foy et al., 2016]), leading to significant changes in the atmospheric composition. Consequently, the knowledge of precise and updated budgets, together with seasonal, monthly, weekly and daily variations of gaseous pollutants driven, amongst other processes, by the emissions are essential for understanding their role in the formation of tropospheric ozone and PMs at various temporal scales, for anticipating pollution peaks and for identifying the key drivers that could help mitigate these emissions.

In this context, complementary methods have been developed for estimating emissions using atmospheric observations. They operate in synergy between a chemistry-transport model (CTM) - which links the emissions to the atmospheric concentrations-, atmospheric observations of the species of interest, and statistical inversion techniques. A number of studies using inverse modeling were first carried out for long-lived species such as greenhouse gases (GHGs) (e.g., carbon dioxide CO2 or methane CH4) at the global or continental scales [Hein et al., 1997; Bousquet et al. 1999], using surface measurements. Later, following the development of monitoring station networks, the progress...
of computing power, and the use of inversion techniques more appropriate to non-linear problems, these methods were applied to shorter-lived molecules such as CO. For these various applications (e.g., for CO₂, CH₄, CO), the quantification of sources was solved at the resolution of large regions [Pétron et al., 2002]. Finally, the growing availability and reliability of observations since the early 2000s (in-situ surface data, remote sensing data such as satellite data), the improvement of the global CTMs, of the computational capacities and of the inversion techniques have increased the achievable resolution of global inversions, up to the global transport model grid cells, i.e. typically with a spatial resolution of several hundreds of square kilometers [Stavrakou and Muller, 2006; Pison et al., 2009; Fortems-Chouzy et al., 2011; Hooghiemstra et al., 2012; Yin et al., 2015; Miyazaki et al., 2017, Zheng et al., 2019].

Today, the scientific and societal issues require an up-to-date quantification of pollutant emissions at a higher spatial resolution than the global one and imply to widely use regional inverse systems. However, although they are suited to reactive species such as CO, NOₓ, and their very large spatial and temporal variability, they have hardly been used to quantify pollutant emissions. Some studies inferred NOₓ [Pison et al., 2007; Tang et al., 2013] and VOC emissions [Koochkan et al., 2013] from surface measurements. Konovalov et al. [2006, 2008, 2010], Mijling et al. [2012, 2013], Van der A. [2008], Lin et al. [2012] and Ding et al. [2017] have also shown that satellite observations are a suitable source of information to constrain the emissions of NOₓ. These regional inversions using satellite observations were often based on Kalman Filter (KF) schemes [Mijling et al., 2012, 2013; Van der A., 2008; Lin et al., 2012; Ding et al., 2017]. However, these inversion problems may be impacted by the non-linearity of atmospheric chemistry, for which variational methods are more suitable than KFs by design.

Here, we present the Bayesian variational inverse system PYVAR-CHIMERE, that has reached a good level of maturity (robustness of calculations, clarity, portability, modularity of the code) and that is adapted to reactive species such as CO and NO₂. This adaptation for reactive species takes advantage of the previous developments for the quantification of fluxes of long-lived GHG species such as CO₂ [Broquet et al., 2011] and CH₄ [Pison et al., 2018] at the regional to the local scales, paving the way for monitoring anthropogenic emissions at high spatial resolutions. The PYVAR-CHIMERE inverse modeling system is based on the Bayesian variational assimilation code PYVAR [Chevallier et al. 2005] and on the regional state-of-the-art CTM CHIMERE, dedicated to the study of regional atmospheric pollution events [Menut et al., 2013, Mailler et al., 2017]. Variational techniques require the adjoint of the model to compute the sensitivity of simulated atmospheric concentrations to corrections of the fluxes. CHIMERE is one of the few CTMs all over the world...
possessing its adjoint code (e.g., for global models: IMAGES [Stavrakou and Muller, 2006], TM5 [Krol et al., 2008], GELKA [Belikov et al., 2016] and LMDz [Chevallier et al., 2005; Pison et al., 2009]; for limited-area models: CMAQ [Hakami et al., 2007], EURAD-IM [Elbern et al., 2007], RAMS/CTM-4DVAR [Yumimoto et Uno, 2006]). Originally, the sequential adjoint code was developed at LMD [Menut et al., 2000; Menut et al., 2003]. Then, it has been further developed and parallelized at LISA [Pison et al., 2007] and LSCE (see Section 3).

The principle of variational atmospheric inversion and the configuration of PYVAR-CHIMERE are described in Section 2 and in Section 3, respectively. Details about the forward and adjoint codes are also given. Then, the potential of PYVAR-CHIMERE is illustrated in Section 4 with the optimization of European CO and NO$_2$ emissions, constrained by observations from the Measurement of Pollution in the Troposphere (MOPITT) instrument and from the Ozone Monitoring Instrument (OMI).

2. Principle of Bayesian variational atmospheric inversion

The Bayesian variational atmospheric inversion method adjusts a set of control parameters in input of the CTM, including parameters related to the emissions whose estimate is the primary target of the inversion. The control parameters may also include initial and boundary conditions for the atmospheric species of interest in the CTM. The adjustments are applied to prior values, usually taken, for the emissions, from pre-existing BU inventories. The principle is to minimize, on the one hand, the departures from the prior estimates of the control parameters (term $J_b$ in Equation 1), which are weighted by the uncertainties in these estimates (called hereafter “prior uncertainties”), and, on the other hand, the differences between simulated and observed concentrations (term $J_o$ in Equation 1), which are weighted by all other sources of uncertainties explaining these differences (called hereafter all together “observation errors”). In statistical terms, the inversion searches for the most probable estimate of the control parameters given their prior estimates, the observations, the CTM and the associated uncertainties. The solution, which will be called posterior estimate in the following, is found by the iterative minimization of a cost function $J$ [Talagrand et al., 1997], defined as:

$$J(x) = \frac{(x - x_p)^T B^{-1} (x - x_p) + (H(x) - y)^T R^{-1} (H(x) - y)}{J_b} + \frac{J_o}{J_o}$$

(Eq. 1)

The control vector $x$ contains the variables to be optimized during the inversion process (surface fluxes but also initial or boundary conditions for example, see Section 3.3). $H$ is the non-linear observation operator that projects the state vector $x$ onto the observation space. In most of the variational atmospheric inversion cases (such as those described in Section 4), the observation operator includes the CTM and an interpolation or an extraction and averaging of the simulated fields.
(see Section 3.4). The observations in $y$ could be surface measurements and/or remote sensing data such as satellite data. The prior uncertainties and the observation errors are assumed to be centered and to have a Gaussian distribution. Consequently, the prior uncertainties are characterized by their covariance matrix $B$ and the observation errors are characterized by their covariance matrix $R$. By definition, the observation errors combine errors in both the data and the observation operator, in particular the combination of measurement errors and errors in the conversion of satellite measurement into concentration data, errors from the CTM, representativity errors due to the comparison between point measurements and gridded models or due to the representation of the fluxes as gridded maps at a given spatial resolution, and aggregation errors associated with the control of emissions at a given spatial and/or temporal resolution that is different from (usually coarser than) that of the CTM.

For non-linear problems, as it is the case with reactive species, the minimum of $J$ may be reached iteratively with a descent algorithm. In this case, the iterative minimization of $J$ is based on a gradient method. $J$ is calculated with the forward observation operator (including the CTM) and its gradient relative to the control parameters $x$ is provided by the adjoint operator (including the adjoint of the CTM):

$$\text{Grad J} = B^{-1}(x - x_b) + H^TR^{-1}(Hx - y) (\text{Eq. 2})$$

As shown in Figure 1, the minimization algorithm repeats the forward-adjoint cycle to seek an optimal solution for the control parameters.

In PYVAR-CHIMERE, we use the M1QN3 limited memory quasi-Newton minimization algorithm [Gilbert and Lemaréchal, 1989]. As most quasi-Newtonian methods, it requires an initial regularization of $x$, the vector to be optimized, for better efficiency. We adopt the most generally used regularization, made by optimizing $\chi = B^{1/2}(x - x_b)$ instead of $x$. Although more optimal regularizations can be chosen, the minimization with $\chi$ is preferred for its simplifying the equation to solve. This transformation translates in Equation 2 as follows: $\text{Grad J} = \chi + B^{1/2}H^*(R^{-1}(H(x) - y))$. The criterion for stopping the algorithm is based on a threshold set on the ratio between the final and initial gradient norms or on the maximum number of iterations to perform. Due to the non-linearity of the problem, the minimization may reach only a local minimum.

Finally, the calculation of the posterior uncertainty is challenging in a variational inverse system. Even though the analysis error covariance matrix can be explicitly written in various analytical forms, it requires the inversion of matrices that are too large to invert given the current computational resources in our variational approach. As a trade-off between computing resources and
comprehensiveness, the analysis error may be evaluated by an approach based on the spread of sensitivity tests (e.g., as in [Fortems-Cheiney et al., 2012]). It can also be estimated through a Monte Carlo Ensemble [Chevallier et al., 2007], implemented in PYVAR.

3. The PYVAR-CHIMERE configuration

3.1. PYVAR adapted to CHIMERE

The PYVAR-CHIMERE inverse modeling system is based on the Bayesian variational assimilation code PYVAR [Chevallier et al. 2005] and on the previous inversion system based on CHIMERE [Pison et al., 2007]. PYVAR is an ensemble of Python scripts, which deals with preparing the vectors and the matrices for the inversion, driving the required Fortran codes of the transport model and computes the minimization of the cost function to solve the inversion. Previously used for global inversions with the LMDz model (e.g., Pison et al., 2009; Chevallier et al., 2010; Fortems-Cheiney et al., 2011; Yin et al., 2015; Locatelli et al., 2015; Zheng et al., 2019), PYVAR has been adapted to the state-of-the-art CHIMERE regional CTM. Including the elements of the previous inversion system built around CHIMERE (coded in Fortran90) [Pison et al., 2007] in PYVAR (coded in Python, see Section 3.5) lead to an up-to-date, more flexible and more robust system.

3.2. Development of the adjoint of CHIMERE

To compute the sensitivity of simulated atmospheric concentrations to corrections to the fluxes, the adjoint of CHIMERE has been developed. Originally, the sequential adjoint was coded at LISA
Menut et al., 2000; Menut et al., 2003; Pison et al., 2007]. Then, it has been parallelized and further
developed at LSCE and LISA, together with its tangent-linear (TL) code. The adjoint has been coded
by hand line by line, following the principles formulated by Talagrand [1997]. It contains exactly the
same processes as the CHIMERE forward model. PYVAR-CHIMERE is currently operational for
the full module of gaseous chemistry. As a compromise between the robustness of the method for
reactive species, the time required coding the adjoint and the computational cost with a full chemical
scheme, the aerosols modules of CHIMERE have not been included in the adjoint of CHIMERE yet
and are therefore not available in PYVAR-CHIMERE.

It should also be noted that other minor changes against the CHIMERE 2013 version [Menut et al.,
2013] have been implemented, including:

• For the geometry, the possibility of polar domains and the use of the coordinates of the corners
of the cells instead of only the centers
• For the transport, the non-uniform Van Leer transport scheme on the horizontal,
• For chemistry, various switches have been added to avoid going into the chemistry, deposition
and wet deposition routines when no species requires them (e.g. no chemistry for methane at a
regional scale).

It should be noted that the development and maintenance of the adjoint means that the version used
is necessarily one or two versions behind the distributed CHIMERE version
(http://www.lmd.polytechnique.fr/chimere/).

As an example, Figure 2 presents a simplified scheme of how PYVAR scripts are used to drive this
version of CHIMERE for forward simulations and inversions using satellite observations. A mode is
also available to test the adjoint: it runs the TL code.
3.3. Definition of the control vector

The control vector is specified by the user in a text file. This file is formatted following Figure 3. The inputs to constrain could be fluxes, initial concentrations, boundary conditions at the top or boundary conditions on the four lateral sides. The spatial resolution of the inversion could be the grid-cell resolution or one whole encompassing region. A choice of simple but efficient ways of building the covariance matrix $B$ are implemented in PYVAR-CHIMERE. The user has only to choose:

- the $B$ variance coefficient, to indicate how to get the variances for these components,
- Correlation lengths $L$ through time and space to define the correlations. For example, the following formula is used to construct the distance correlation $r$ between two fluxes $x_i$ and $x_j$: $r(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)}{L}\right)$

where $r$ is the correlation between the fluxes $x_i$ and $x_j$, $d$ is the distance between these two fluxes in kilometers.

3.4 Equivalents of the observations

The individual data given as constraints in the system are first formatted into a text file described in Figure 4. During forward simulations, the equivalents of the components of $y$ (i.e., the equivalents of the individual data) are calculated by PYVAR-CHIMERE. It includes the CTM and an interpolation (see below the vertical interpolation from the model’s grid to the satellite levels) or an extraction and averaging (e.g., extracting the grid cell matching the geographical coordinates of a surface station and averaging over one hour). As a compromise between technical issues such as the time required for reading/writing files, the observation operator $H$ that generates the equivalent of the observations by
the model (i.e. $H(x)$) has been so far partly embedded in the code of CHIMERE. It makes it easier to use finer time intervals than available in the usual hourly outputs of CHIMERE to compute the required information (e.g., within the finer CTM physical time steps).

Figure 4. Simplified scheme of how PYVAR scripts prepares the $y$ observations, using satellite data. PYVAR, and text sources are respectively displayed in blue and in grey.

To make comparisons between simulations and satellite observations, the simulated vertical profiles are first interpolated on the satellite’s levels (with a vertical interpolation on pressure levels). Then, the averaging kernels (AKs), when available, are applied to represent the vertical sensitivity of the satellite retrieval. Two types of formula, depending on the satellite observations used, have been detailed in PYVAR-CHIMERE for the use of AKs: $C_m = AK.C_{m(o)}$ (e.g., used by the satellite instrument OMI, see Section 4.3) or $C_m = x_a + AK(C_{m(o)} - x_a)$ where $C_m$ is the modeled column, AK contains the averaging kernels, $x_a$ is the prior profile (provided together with the AKs when relevant) and $C_{m(o)}$ is the vertical distribution of the original model partial columns interpolated to the pressure grid of the averaging kernels.

3.5. Numerical language

The PYVAR code is in Python 2.7, the CHIMERE CTM is coded in Fortran90. The CTM requires several numerical tools, compilers and libraries. The CTM was developed and tested using the software versions as described in Table 1.
### Table 1. URL addresses for the development and the use of the PYVAR-CHIMERE system and its modules.

<table>
<thead>
<tr>
<th>Software</th>
<th>URL</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td><a href="https://www.python.org/downloads/">https://www.python.org/downloads/</a></td>
<td>2.7</td>
</tr>
<tr>
<td>Libraries or</td>
<td></td>
<td></td>
</tr>
<tr>
<td>packages</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UnidataNetCDF</td>
<td><a href="https://www.unidata.ucar.edu/">https://www.unidata.ucar.edu/</a></td>
<td>3</td>
</tr>
<tr>
<td>Open MPI</td>
<td><a href="https://www.open-mpi.org/">https://www.open-mpi.org/</a></td>
<td>1.10.5</td>
</tr>
<tr>
<td>GRIB_API</td>
<td><a href="https://confluence.ecmwf.int/display/GRIB/Releases">https://confluence.ecmwf.int/display/GRIB/Releases</a></td>
<td>1.14</td>
</tr>
<tr>
<td>nco</td>
<td><a href="http://nco.sourceforge.net/#Source">http://nco.sourceforge.net/#Source</a></td>
<td>4.6.3</td>
</tr>
</tbody>
</table>

PYVAR-CHIMERE’s computation time for one node of 10 CPUs is about 3h for 1 day of inversion (with ~10 iterations) for the European domain size of 101 (longitude) x 85 (latitude) x 17 (vertical levels) used in Section 4. It is also important to note that an integration of CHIMERE’s adjoint is about three times longer than a CHIMERE forward simulation. The model parallelism results from a Cartesian division of the main geographical domain into several sub-domains, each one being processed by a worker process. To configure the parallel sub-domains, the user has to specify two parameters in the model parameter file: the number of sub-domains for the zonal and meridional directions. The total number of CPUs used is therefore the product of these two numbers plus one for the master process.

4. Potential of PYVAR-CHIMERE for the inversion of CO and NO$_2$ emissions

The potential of the PYVAR-CHIMERE system to invert emissions of reactive species is illustrated with the inversion of the emissions of CO and NO$_2$ over Europe. We present results for 1-day inversions. We select two different days, respectively for CO and NO$_2$: the 7th March 2015 and the 19th February 2015. These particular days have been chosen as they present a typical number of supers­observations (with respectively, 1587 and 3330) during winter, when the prior emissions are high.

4.1. Set-up

For both examples, CHIMERE is run over a 0.5°x0.5° regular grid (about 50x50km$^2$) and 17 vertical layers, from the surface to 200hPa (about 12km), with 8 layers within the first two kilometers. The domain includes 101 (longitude) x 85 (latitude) grid-cells (15.5°W-35°E; 31.5°N-74°N, see Figure 5). Meteorological fields are provided by ECMWF meteorological forecast [Owens and Hewson, 2018]. The chemical scheme used in PYVAR-CHIMERE is MELCHIOR-2, with more than 100 reactions [Derognat et al., 2003], including 22 for inorganic chemistry.
4.1.1. Control vector $x$

All the information required by the inverse system to run the inversion for CO or NO\textsubscript{2} emissions are listed in Table 1. The control vector $x$ has 17542 components:

- the CO or NO\textsubscript{2} emissions at a 1-day and at a 0.5° ×0.5° (longitude, latitude) resolutions i.e. 1 day times 101x85 grid cells = 8585 components in $x$ for each one-day-long inversion
- the CO or NO\textsubscript{2} 3D initial conditions at a 0.5° ×0.5° (longitude, latitude) resolution i.e. 8585 components in $x$
- the CO or NO\textsubscript{2} boundary conditions at a 1-day and at a 0.5° ×0.5° (longitude, latitude) resolutions i.e. 2x101 + 2x85=372 components in $x$

The prior anthropogenic emissions for CO and NO\textsubscript{x} emissions come from the TNO inventory [Dellaert et al., 2018]. The prior anthropogenic emissions for VOCs come from the EMEP inventory. Climatological values from the LMDZ-INCA global model [Szopa et al., 2008] are used to prescribe concentrations at the lateral and top boundaries and the initial atmospheric composition in the domain.

4.1.2. Observations $y$

The observational constraints for CO emissions come from the MOPITT instrument. It has been flown onboard the NASA EOS-Terra satellite, on a low sun-synchronous orbit that crosses the equator at 10:30 and 22:30 LST. The spatial resolution of its observations is about 22x22 km\textsuperscript{2} at nadir. It has been operated nearly continuously since March 2000. MOPITT CO products are available in three variants: thermal-infrared TIR only, near-infrared NIR only and the multispectral TIR-NIR product, all containing total columns and retrieved profiles (expressed on a ten-level grid from the surface to 100 hPa). We choose to constrain CO emissions with the MOPITT surface product for our illustration. Among the different MOPITTv8 products, we choose to work with the multispectral MOPITTv8-NIR-TIR one, as it provides the highest number of observations, with a good evaluation against in situ data from NOAA stations [Deeter et al., 2019]. The MOPITTv8-NIR-TIR surface concentrations are sub-sampled into “super-observations” in order to reduce the effect of correlated errors between neighboring observations in the inversion system: we selected the median of each subset of OMI data within each 0.5°×0.5° grid-cell and each physical time step (about 5-10 minutes). After this screening, 1587 “super-observations” remain in the 1-day inversion (from 10667 raw observations). These super-observations are provided to PYVAR-CHIMERE as constraints $y$, and treated as described in Section 3.4. Any other pre-treatment of the data (e.g., no screening, or different subsampling, etc) could have been chosen. It is important to note that the potential of MOPITT to provide information at a high temporal resolution, up to the daily scale, is hampered by the cloud coverage (see the blanks in Figure 5b).
The observational constraint on NO$_2$ emissions comes from the OMI QA4ECV tropospheric columns [Muller et al., 2016; Boersma et al., 2016, Boersma et al., 2017]. The Ozone Monitoring Instrument (OMI), a near-UV/Visible nadir solar backscatter spectrometer, was launched onboard EOS Aura in July 2004. It has been flying on a 705 km sun-synchronous orbit that crosses the Equator at 13:30 LT. Our data selection follows the criteria of the OMI QA4ECV data quality statement. As the spatial resolution of the OMI data is finer than that of the chosen CHIMERE model grid (13x24 km$^2$ against 0.5°×0.5°, respectively), the OMI tropospheric columns are sub-sampled into “super-observations” (median of the OMI data within the 0.5°×0.5° grid-cell and each physical time step and its corresponding AK).

### 4.1.3. Covariance matrices $B$ and $R$

The covariance matrix $B$ of the prior errors is defined as diagonal (i.e. only variances are taken into account). Even though total emissions are well known, large uncertainties still affect such emission inventories [Kuenen et al. 2014] at the pixel scale. Consequently, the error standard deviations assigned to the CO and NO$_2$ prior emissions in the covariance matrix $B$ are set at 100%. The variance of the individual observation errors in $R$ is defined as the quadratic sum of the measurement error reported in the MOPITT and the OMI data sets, and of the CTM errors (including transport errors and representativity errors) set to 20% of the retrieval values. The representativity errors could have been reduced with the choice of a finer CTM resolution (e.g., with a resolution closer to the size of the satellite pixel). Error correlations between the super-observations are neglected, so that the covariance matrix $R$ of the observation errors is diagonal.

### 4.2. Inversion of CO emissions

Large discrepancies are found between the MOPITT CO observations and the prior simulation of their equivalents by CHIMERE over Europe (Figure 5). These discrepancies might be explained by an underestimation in the BU inventory due to a general trend in emissions (if the underestimation persists throughout the year) or to an underestimation regarding particular activity sectors or the time profiles at given scales (daily, monthly). This can also be explained by uncertainties from the satellite data or from the CTM (e.g., atmospheric production, chemistry with OH).
Figure 5. CO collocated surface concentrations a) simulated by CHIMERE using the prior TNO emissions and b) observed by MOPITT\textsuperscript{v8-NIR-TIR} in ppbv, c) relative differences between the MOPITT observations and the CHIMERE simulated concentrations using the prior TNO emissions, and d) relative differences between the CHIMERE simulated concentrations using the posterior emissions and the CHIMERE simulated concentrations using the prior TNO emissions, in %, at the 0.5°\times0.5° grid-cell resolution, over Europe for the 7\textsuperscript{th}, March 2015.

About 10 iterations are needed to reduce the norm of the gradient of $J$ by 90\% with the minimization algorithm M1QN3 and obtain the increments i.e. the corrections provided by the inversion. The prior CO emissions over Europe on the 7\textsuperscript{th}, March 2015 and their increments are shown in Figure 6. As expected from the underestimation of the prior surface concentrations in Figure 5, local increments may be significant, reaching more than +50\%. The posterior emissions and their uncertainties will have to be evaluated and may bring hints to the cause of the discrepancies. The analyzed
concentrations in Figure 5d are the concentrations simulated by CHIMERE with the posterior fluxes: as expected, the optimization of the fluxes improves the fit of the simulated concentrations to the observations.

a) b)

Figure 6. a) TNO CO anthropogenic prior emissions, in ktCO/grid-cell and b) increments provided by the inversion with constraints from MOPITT v8-NIR-TIR for the 7th, March 2015, in %. Note that part of Figure 6b presents blanks, as there is no observation to constrain the fluxes (see Figure 5).

4.3. Inversion of NO2 emissions

Large discrepancies are found between the OMI NO2 super-observations and the prior simulation of their equivalents by PYVAR-CHIMERE (Figure 7). Over Europe, the prior simulation strongly underestimates the tropospheric columns, and particularly over Po Valley.
Figure 7. NO$_2$ collocated tropospheric columns a) simulated by CHIMERE using the prior TNO emissions, b) observed by OMI, in 1e16 molec.cm$^{-2}$, c) relative differences between the OMI observations and the CHIMERE simulated tropospheric columns using the prior TNO emissions, and d) relative differences between the CHIMERE simulated tropospheric columns using the posterior emissions and the CHIMERE simulated tropospheric columns using the prior TNO emissions, in %, at the 0.5°x0.5° grid-cell resolution, for the 19th February 2015, over Europe.

The prior NO$_2$ emissions and the increments obtained after inversion over Europe for the 19th, February 2015 are shown in Figure 8. As expected from the underestimation of the prior tropospheric columns in Figure 7, local increments may be large, for example over industrial areas (e.g., over the Po Valley) and over large cities (e.g., Madrid), with increments of more than +50% (Figure 8). The analyzed NO$_2$ tropospheric columns in Figure 5c are the columns simulated by CHIMERE with the NO$_2$ posterior fluxes: as expected, the optimization of the NO$_2$ fluxes has only a small impact on the differences between the simulated and observed NO$_2$ columns. NO$_x$ emissions are speciated as
9.2 % of NO$_2$, 0.8 % of HONO, and 90 % of NO [Menut et al., 2013; Liu et al., 2018], following the Generation of European Emission Data for Episodes (GENEMIS) recommendations [Friedrich, 2000; Kurtenbach et al., 2001; Aumont et al., 2003]. Consequently, the NO$_2$ fluxes contribute only to about 10% to the NO$_2$ tropospheric column. Further work will be done to simultaneously optimize NO and NO$_2$ fluxes, which together contribute to more than 99% to the NO$_2$ tropospheric column.

Figure 8. a) TNO NO$_2$ anthropogenic prior emissions, in ktNO$_2$/grid-cell and b) increments provided by the inversion with constraints from OMI for the 19th, February 2015, in %. The configuration of the inversion is detailed in Table 1.

5. Conclusion/Discussion

This paper presents the Bayesian variational inverse system PYVAR-CHIMERE, which has been adapted to the inversion of reactive species such as CO and NO$_2$, taking advantage of the previous developments for long-lived species such as CO$_2$ [Broquet et al., 2011] and CH$_4$ [Pison et al., 2018]. We show the potential of PYVAR-CHIMERE, with inversions for CO and NO$_2$ illustrated over Europe for an example day. PYVAR-CHIMERE will be used to infer CO and NO$_x$ emissions over long periods, e.g. first for a whole season or year and then for the recent decade 2005-2015 in the framework of the H2020 VERIFY project over Europe, and in the framework of the ANR PolEASIA over China, to quantify their trend and their spatio-temporal variability.

The PYVAR-CHIMERE system can handle any large number of both control parameters and observations. It will be able to cope with the dramatic increase in the number of data in the near future.
with, for example, the high-resolution imaging (pixel of 7x3.5 km²) of the new Sentinel-5P/TROPOMI program, launched in October 2017. These new space missions with high-resolution imaging have indeed the ambition to monitor atmospheric chemical composition for the quantification of anthropogenic emissions. Moreover, a step forward in the joint assimilation of co-emitted pollutants will soon be possible with the PYVAR-CHIMERE system and the availability of TROPOMI co-localized images of NO₂ and CO for example. This should improve the consistency of the inversion results and can be used to inform inventory compilers, and subsequently improve emission inventories. Moreover, this development will help in further understanding of air quality problems and help address air quality related emissions at the national to subnational scales.

Author Contribution
All authors have contributed to the manuscript writing (main authors: AFC, GB, IP and GD) and to the development of the present version of the PYVAR-CHIMERE system (main developer: IP). IP and GD have parallelized the adjoint version from Menut et al., [2000], Menut et al., [2003] and Pison et al., [2007]. IP has complemented the adjoint of new parameterizations since the CHIMERE release in 2011 and the tangent-linear model.

Code and Data Availability
OMI QA4ECV NO₂ product can be found here: http://temis.nl/qa4ecv/no2.html.
MOPITTv8-NIR-TIR CO product can be found here: ftp://l5ftl01.larc.nasa.gov/MOPITT/
The CHIMERE code is available here: www.lmd.polytechnique.fr/chimere/.
The associated documentation of PYVAR-CHIMERE is available on the website https://pyvar.lsce.ipsl.fr/doku.php/3chimere:headpage. The documentation includes a whole description of PYVAR-CHIMERE and several tutorials on how to run a first PYVAR-CHIMERE simulation or how to run an inversion.

Competing interests
The authors declare that they have no conflict of interest.

Acknowledgements
We acknowledge L. Menut and C. Schmechtig for their contributions to the development work on the adjoint code of CHIMERE and its parallelization. We acknowledge the TNO team (H.A. Denier van der Gon, J. Kuenen, S. Dellaert, S. Jonkers, A. Visschedijk, et al.) for providing NOₓ and CO emissions over Europe. We also acknowledge the free use of tropospheric NO₂ column data from the OMI sensor from http://temis.nl/qa4ecv/no2.html and the free use of CO surface concentrations from the MOPITT sensor from ftp://l5ftl01.larc.nasa.gov/MOPITT/. For this study, A. Fortems-Cheiney was funded by the FrenchSpace Agency-Centre National d’Etudes Spatiales CNES and by the H2020 VERIFY project, funded by the European Commission Horizon 2020 research and innovation programme, under agreement number 776810. L. Costantino was funded by the PoIEASIA ANR project under the allocation ANR-15-CE04-0005. This work was granted access to the HPC resources of TGCC under the allocation A0050107232 made by GENCI. Finally, we wish to thank F. Marabelle (LSCE) and his team for computer support.

References


Ding, J., Miyazaki, K., van der A, R. J., Mijling, B., Kurokawa, J.-I., Cho, S., Janssens-Maenhout, G., Zhang, Q., Liu, F., and Levelt, P. F.: Intercomparison of NOx emission inventories...


EMEP/EEA air pollutant emission inventory guidebook, 2016.


