FLEXINVERT+

Bayesian Inversion Framework

User's Guide

Table of Contents

1. Introduction

FLEXINVERT+ is a framework for the optimization of surface-atmosphere fluxes of trace species, such as greenhouse gases or aerosols. It is based on Bayesian statistics and optimizes a prior estimate of the fluxes to best fit the atmospheric observations within prescribed bounds of uncertainty. The fluxes are related to the observations through a model of atmospheric transport: in FLEXINVERT+ the Lagrangian particle dispersion model, FLEXPART is used. The user should be familiar with the principles of inverse modelling and data assimilation as well as with running FLEXPART before attempting to run FLEXINVERT+.

FLEXINVERT+ has its foundation in the code FLEXINVERT, but is a complete rewrite of the code to make it more modular and uses an object orientated programming style. The code has also been modified to include special treatment of CO_2 fluxes (described in Section 6.6).

2. System requirements

FLEXINVERT+ is written in Fortran90 and has been tested using the gfortran compiler on the Linux Ubuntu operating system. To compile and run FLEXINVERT+ the following libraries are required:

- NetCDF (tested with package libnetcdff6)
- LAPACK (tested with package liblapack3)

Output from the transport model FLEXPART will need to be prepared prior to running FLEXINVERT+. For details about installing and running FLEXPART, the user is referred to the FLEXPART website: https://www.flexpart.eu.

3. Input data and pre-processing

As with any atmospheric inversion framework, FLEXINVERT+ requires four basic types of input: 1) atmospheric observations, 2) a model of atmospheric transport, 3) prior estimates of the fluxes, and 4) initial concentrations. In addition, FLEXINVERT+ has the possibility to optimize the fluxes on a spatially variable grid or by regions, in which case a grid (or regions) definition file is required. Each of these input components is discussed in detail below.

3.1. Atmospheric observations

FLEXINVERT+ can assimilate any type of atmospheric observation, from stations, aircrafts or ships. In principle satellite data can also be used, however, currently the pre-processing steps for setting-up the FLEXPART runs and formatting the observations does not include satellite data. An important revision compared to the previous FLEXINVERT version, is that the observations need to be pre-processed and a pre-processor is provided for this.

The pre-processor, *prep_flexpart*, needs to be run before running the inversion. The pre-processor reads all observations in their raw format, performs any averaging or data selection (based on the SETTINGS file), prepares the input files needed for running FLEXPART, and writes the (averaged and/or selected) observations to the file format required by the inversion. Currently, *prep_flexpart* can process three different raw observation formats: 1) NOAA surface flasks, 2) ObsPack (versions 2.1 and 3.2) and 3) WDCGG hourly data. (Note that *prep_flexpart* currently only handles one format type at a time). To read another data format

will require a small modification: the best practice for this is to introduce the new format as an option in "main.f90" and to make a copy of one of the subroutines "read_noaa.f90", "read_obspack.f90" or "read_wdcgg.f90" (which ever is closest to the new format) and modify this.

The processed observations are saved in the directory "path_obsout" (see SETTINGS file) with one file per station or campaign. Note that all data selection and averaging must be done in the pre-processing step, as this is not done in the inversion.

3.2. Atmospheric transport

Atmospheric transport is modelled using output from FLEXPART. The pre-processor, *prep_flexpart*, prepares the following input files required to run FLEXPART:

- COMMAND
- AGECLASS
- RELEASES
- OUTGRID

These files are prepared for each station (or aircraft/ship campaign) and month corresponding to one FLEXPART backwards mode run. The RELEASES file contains all the *releases* for the given month with one *release* per observation. This release will determine the so-called source-receptor-relationship (SRR), i.e. the relationship of the fluxes to an observation. The SRRs for each observation are saved in separate files (grid_time) with the timestamp of the observation in the file name. In addition, one file containing the sensitivity of the observation to the initial concentrations (grid_initial) is saved for each observation.

Note that for FLEXINVERT+ these files have a different format to what was required by the previous version of FLEXINVERT. To get the new formatted files, FLEXPART needs to be run with the COMMAND file options: LINVERSIONOUT = 1 and SURF_ONLY = 1 (these are the default settings given by *prep_flexpart*). The first setting will write grid_time and grid_initial files for each release with a time dimension for each footprint (this is in contrast to the standard output format in which the grid_time and grid_initial files are written for each footprint with a time dimension for each release). The second setting will write only the surface SRRs to the grid_time file, as only the surface values are required by the inversion, while the grid_initial files will contain all vertical layers (without this option all vertical layers will be written in both files, which will require more memory to store and take longer to read).

FLEXINVERT+ has the new feature of allowing for the use of nested FLEXPART domains (in SETTINGS file: LNESTED = 1). If the nested option is used, then two sets of grid_time files are written for each observation, one for the global domain (at lower resolution) and one for the nested domain (at higher resolution). In this case, the global domain files are used to calculate the so-called background mixing ratios or concentrations while the nested domain files are used in the inversion step (see also Sections 3.3, 3.5 and 6.1). If using this option, the nested domain for the FLEXPART runs must be the same size or larger than the domain of the inversion.

After running the pre-processor (to set-up the FLEXPART input files and format the observations), the FLEXPART jobs can be run by simply editing and executing the file "run_flexpart.sh".

3.3. Prior estimates of the fluxes

FLEXINVERT+ requires a prior estimate of the fluxes. The fluxes are read from NetCDF files (one file for each year) and must be at the same spatial resolution as the FLEXPART

grid_time files. A pre-processor, *prep_fluxes*, is provided to reformat the fluxes to the resolution required, this can be either by averaging or interpolating. If using a nested domain (see Section 3.2), then the fluxes must be provided for both the global and the nested domain with the corresponding spatial resolutions (both can be prepared by separate runs of *prep_fluxes*). Otherwise only the global fluxes are required.

The output flux file contains the latitude and longitude dimensions (given for the mid-point of each grid-cell in degrees) and the time dimension (given in days since 1-Jan-1900).

3.4. Initial concentrations

For long-lived atmospheric species (with an atmospheric lifetime of more than a few weeks) the atmospheric background must be accounted for. This is because the FLEXPART runs only account for the influence on the species for as long as the length of the backwards trajectory (set by AGECLASS). The background mixing ratio (or concentration) is modelled for each observation by coupling the grid_initial files to global fields of initial mixing ratios (or concentrations). The initial mixing ratio fields can be provided from prior runs of a global Eulerian model or by a bivariate (latitude and longitude) interpolation of observations representing the *well-mixed troposphere* from e.g. the NOAA flask sampling network. The initial mixing ratio fields need to be provided as NetCDF files.

In the case that initial mixing ratio fields cannot be provided (because there are no existing model runs or insufficient observations to interpolate) the user can modify the code to read his/her own background mixing ratio estimates. This would involve:

1) modifying "init_ghg.f90" (or "init_co2.f90") to comment-out the call to "init_cini.f90" and replacing it with a call to the user's own sub-routine providing estimates of the background mixing ratios for each observation. This sub-routine must write the background estimate for each observation to the variable "cini" in the data structure "obs" (a template for such a routine is provided in Appendix A).
2) modifying "simulate.f90" to comment-out the lines assigning the values of "obs%bkg(i)" and "bkgerr" immediately below the comment "*background contribution from fluxes outside inversion domain*" and setting the value of these variables to zero.

3.5. Aggregated grid or regions definition

FLEXINVERT+ has the option of optimizing the fluxes for each grid-cell (at the resolution of the grid_time files) or on an aggregated grid (or regions). These regions may be based on an aggregation of grid-cells using the information from the SRRs or on the user's own definition. Using an aggregated grid has the advantage of reducing the dimension of the inversion problem, thus reducing the computation time and memory required. On the other hand, depending on the aggregation, it may increase the aggregation error.

Using an aggregated grid based on the SRRs means that there is a minimal increase in the aggregation error, as grid-cells are aggregated only where there is little information provided by the observations about the fluxes. A pre-processor, *prep_regions*, is provided to calculate the aggregated grid. This step requires the grid_time files (from *prep_flexpart*) and optionally the prior fluxes. The *prep_regions* pre-processor uses the same SETTINGS_files and SETTINGS_config files as used for running the inversion (namely the paths to the FLEXPART output, land-sea mask, and prior fluxes, as well as the inversion domain settings). In addition, there are user options in the SETTINGS_regions file in the same directory as *prep_regions*.

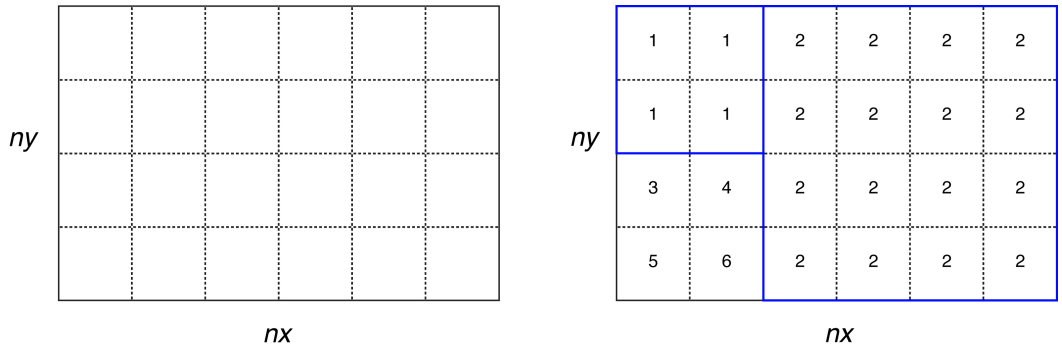

Figure 1. Schematic of the grid aggregation. Left is the grid at the original resolution (*nx* × *ny*) and right is the aggregated grid with *N*=6 regions with grid cell aggregate 1 at two times the original resolution, grid cell aggregate 2 at four times the original resolution and the remaining grid cells at the original resolution.

The grid definition is saved in NetCDF format with the filename as specified by "file_regions" defined in SETTINGS_files. This file contains a 2D matrix (lon × lat) covering the inversion domain (as defined in SETTINGS_config) at the same resolution as the grid_time files. The elements of this matrix contain a number from 1 to *N* where *N* is the total number of grid-cell aggregates (or regions). If the option USELANDUSE = true (see SETTINGS_regions) then land regions will be numbered 1 to N_{land} and ocean regions will be numbered -1 to $-N_{ocean}$.

Note if USELANDUSE = FALSE, then no land/ocean distinction will be made in FLEXINVERT+.

The user can also provide his/her own regions definition file (instead of using *prep_regions*) as long as it follows the same format as the one described above. This file must cover at least the inversion domain.

4. Running FLEXINVERT+

The user settings for FLEXINVERT+ are specified in two files, SETTINGS_config and SETTINGS_files in the sub-directory "settings". The settings files must be provided as input arguments when running FLEXINVERT+ (see "job_flexinvert.sh").

4.1. Configuration options

This section describes the configuration options specified in SETTINGS_config.

1. **run_mode**: FLEXINVERT+ has three modes: i) forward run in which only the prior mixing ratios are modelled, ii) optimization run in which the fluxes are optimized and iii) random perturbation run in which random perturbations are added to the prior fluxes and observations following the error characteristics of the prior and observation error covariance matrices (this mode is needed to run a Monte Carlo ensemble, see Section 4.2).
2. **seed**: this is only used for run_mode = 2 and sets the seed for the random number generation.
3. **datei/datef**: the start and end dates for the inversion interval. These may cover multiple years but must start at the beginning a month and end at the end of a month.
4. **method**: three options for solving the inverse problem are provided: i) the analytical method, which requires the full transport matrix to be stored in memory, ii) the conjugate gradient method, and iii) the M1QN3 Quasi-Newton method. Methods (ii) and (iii) require only the gradients of the cost function to be stored (for details see Section 6).

5. **offsets**: this option determines whether the inversion optimizes the offsets from the prior fluxes (true) or the fluxes themselves (false). (Note this is only applicable to GHG species since for CO_2 it is always the offsets that are optimized.)
6. **inc_ocean**: if this option is set to true fluxes over the ocean will be optimized in addition to land boxes (note if spatial aggregation is used, then the regions must contain the distinction between land and ocean regions, see Section 3.5).
7. **spa_corr**: if this option is set to true then the error covariance matrix will contain spatial correlations between fluxes (note if large regions are used in the inversion then no spatial correlation between them should be used).
8. **prior_bg**: this option only applies to the conjugate gradient method and allows the "analysis.nc" file from a previous inversion to be used as a starting point for continuing the optimization (note that this is different from the prior flux estimate, which must be the same as that used in the previous inversion).
9. **restart**: this option can be used with the conjugate gradient method (which is iterative) to continue further iterations or to pick-up a run that has crashed.
10. **verbose**: if this option is set to true, then additional output is saved that can be used for debugging (note that using this option for large runs is not recommended as it will increase the computation time and produce large numbers of output files).
11. **spec**: currently two species types are defined: i) a generic greenhouse gas species and ii) CO_2. Note for CO_2 a special treatment is applied (see Section 6.5).
12. **molarmass**: the molar mass of the species optimized. This is used to convert the SRRs from mass mixing ratios to volume mixing ratios (if mixing ratios is used and not concentrations). Note the molar mass must be consistent with the mass used in the prior fluxes.
13. **coeff**: this is used to convert the SRRs mass mixing ratios (in units of ppt) to the unit of the observations (e.g. ppm or ppb).
14. **nested**: if this option is set to true, the nested FLEXPART output will be used to describe the relationship between the fluxes to be optimized (in the inversion domain) and the observations (see also Sections 3.2 and 6.1)
15. Inversion domain: the domain is specified by the left lower corner longitude and latitude (w_edge_lon and s_edge_lat) and the upper right corner (e_edge_lon and n_edge_lat) and the resolution (xres and yres). Note although two variables are used to specify the resolution, currently the same resolution must be used in the longitude and latitude directions (this is the same resolution as the FLEXPART output).
16. **regions**: if this option is set to true, then a spatially aggregated grid is used as defined by the file "file_regions" in SETTINGS_files.
17. **stateres**: the temporal resolution of the state vector in days (must be an integer)
18. **stateres_hr**: (CO_2 only) the sub-daily resolution of the state vector in hours (must be an integer and there must exist an integer by which this can be multiplied to give 24 hours, see also Section 6.5).
19. **nt_flx**: the number of intervals per year in the input prior flux file(s) (e.g. for monthly fluxes nt_flx = 12).
20. **num_nee_day**: (CO_2 only) the number of intervals per 24 hours in the prior NEE flux files (e.g. for 3-hourly NEE fluxes num_nee_day = 8).
21. **measerr**: the minimum measurement uncertainty (only used if measurement uncertainty is not specified in the observation files or if it is larger than that specified)
22. **cinierr**: the estimated uncertainty in the initial concentration fields.
23. **flxerr**: the prior flux uncertainty specified as a fraction.
24. **ffferr**: (CO_2 only) the uncertainty estimated for the prior fossil fuel fluxes specified as a fraction.

25. **flxerr_ll**: (GHG only) the lower limit for the prior flux uncertainty (in the same units as the input fluxes). This is used to avoid having zero uncertainty where the prior flux estimate is zero.
26. **sigma_land**: the spatial correlation scale length over land in km (only used if spa_corr is true).
27. **sigma_ocean**: the spatial correlation scale length over ocean in km (only used if spa_corr is true and ocean fluxes are optimized).
28. **sigmatime**: the temporal correlation scale length in days.
29. **globerr**: an estimate of the domain total error in Tg/y. This is used to scale the prior error covariance matrix.

4.2. Monte Carlo ensembles

If the conjugate gradient method is used, the estimate of the posterior uncertainty found by the Lanczos algorithm gives only a poor approximation of the actual uncertainty (see Section 6.3), and for the M1QN3 method, the posterior uncertainty is not estimated at all. In these cases, FLEXINVERT+ can be used to generate a Monte Carlo ensemble of inversions, which can be used to approximate the posterior uncertainty more reliably. This requires multiple inversion runs, as one inversion equates to one member of the ensemble, with each run using "run_mode" of 2 and a different value for "seed". The standard deviation of the posterior fluxes from the ensemble provides an estimate of the posterior uncertainty.

5. Output data

All output from running FLEXINVERT+ are saved to the directory specified by "path_output" in SETTINGS_files.

1. analysis.nc: the prior and posterior fluxes and uncertainties, as well as the flux increments (posterior minus prior) as 3D arrays (lon × lat × time) (for CO_2, the prior fossil fuel, biomass burning and ocean fluxes are also included). Note if using the conjugate gradient method, the posterior uncertainties in this file should not be used but instead calculated using a Monte Carlo ensemble (see also Table 2).
2. analysis_nee.nc: (CO_2 only) the prior and posterior NEE fluxes and flux increments (at the same temporal resolution as the prior) as 4D arrays (lon × lat × day × hour).
3. area_box.txt: the areas of the cell aggregates (or regions) in square meters.
4. cort.txt: the temporal covariance matrix as a 2D array (lon × lat).
5. evals.txt: the eigenvalues of the prior error covariance matrix for 1 time-step.
6. evecs.txt: the eigenvectors of the prior error covariance matrix for 1 time-step.
7. flexinvert.log: the logfile of the inversion run
8. grad_xx.txt: (congrad method only) the gradient of the cost function at iteration xx.
9. hessian_evals.txt: (congrad method only) the eigenvalues of the Hessian matrix.
10. hessian_evecs.txt: (congrad method only) the eigenvectors of the Hessian matrix.
11. hloc_box.txt: the difference from UTC in hours for each grid cell aggregate (or region).
12. lsm_box.txt: the land-sea mask for the grid cell aggregates (or regions).
13. monitor.txt: the modelled and observed atmospheric mixing ratios (or concentrations).
14. nbox_xy.txt: the definition of the aggregated grid as a 2D array (lon × lat).
15. obsfiles.txt: the number of observation files read (first entry) and a list of the files.
16. obsread.txt: the observed mixing ratios (or concentrations) and their uncertainty for all sites and times.
17. sqcort.txt: the square root of the temporal covariance matrix as a 2D array (lon × lat).
18. cost_obs.txt: the cost at each iteration (needed for picking-up an M1QN3 inverion)

Table 1. Definition of the mixing ratio variables in monitor.txt

Variable	Description
conc	observed mixing ratio (or concentration)
cini	initial mixing ratio (contribution from particle termination points)
bkg	background mixing ratio (contribution from fluxes outside domain)
ghg	total contribution from inside domain fluxes (only if offsets = true)
nee	contribution from NEE fluxes (CO_2 only)
fff	contribution from fossil fuel fluxes (CO_2 only)
ocn	contribution from ocean fluxes (for GHG only if offsets = false)
bbg	contribution from biomass burning fluxes (CO_2 only)
prior	contribution from the prior state vector (zero for CO_2 or if offsets = true)
post	contribution from the posterior state vector
diff	the model-observation differences
error	the observation error

The total prior and posterior modelled mixing ratios can be calculated from the monitor.txt file as follows:

for CO_2:

$$y^{pri} = y^{cini} + y^{bkg} + y^{nee} + y^{fff} + y^{ocn} + y^{bbg} + y^{prior}$$
$$y^{pos} = y^{cini} + y^{bkg} + y^{nee} + y^{fff} + y^{ocn} + y^{bbg} + y^{post}$$

for GHG:

offsets = true:

$$y^{pri} = y^{cini} + y^{bkg} + y^{ghg} + y^{prior}$$
$$y^{pos} = y^{cini} + y^{bkg} + y^{ghg} + y^{post}$$

offsets = false:

$$y^{pri} = y^{cini} + y^{bkg} + y^{ocn} + y^{prior}$$
$$y^{pos} = y^{cini} + y^{bkg} + y^{ocn} + y^{post}$$

Table 2. Definition of the flux variables in analysis.nc and analysis_nee.nc

File	Species	Variable	Description
analysis.nc	GHG, CO_2	fpri	prior fluxes (lon × lat × time)
		fpos	posterior fluxes (lon × lat × time)
		epri	prior flux uncertainty (lon × lat × time)
		epos*	posterior flux uncertainty (lon × lat × time)
		fincrement	flux increment (lon × lat × time)
	CO_2	focn	ocean fluxes (lon × lat × time)
		fff	fossil fuel emissions (lon × lat × time)
		fbbg	biomass burning emissions (lon × lat × time)
analysis_nee.nc	CO_2	nee_pri	prior NEE (lon × lat × day × hour)
		nee_pos	posterior NEE (lon × lat × day × hour)
		nee_increment	NEE increment (lon × lat × day × hour)

*Note epos is only valid for the analytical method (for conjugate gradient and M1QN3 methods the posterior uncertainty must be calculated via a Monte Carlo ensemble).

6. Theoretical aspects

6.1. Forward model

An atmospheric mixing ratio, y_i for a given time and location can be modelled as follows:

$$y_i = H(\mathbf{x}) + y_i^{ini} \tag{1}$$

where H is the atmospheric chemistry and transport function (continuity equation), $\mathbf{x}$ is a vector of the fluxes and y_i^{ini} is the initial mixing ratio. Running FLEXPART in backwards time mode calculates the relationship H as a matrix operator, $\mathbf{H}$ (provided that any chemistry can be approximated to be linear). If the inversion domain is not global, then FLEXINVERT+ makes a distinction between the contribution from fluxes inside and outside the *nested* domain. Fluxes outside the nested domain are attributed to the *background* and those inside are optimized. In this case Eq.1 becomes:

$$y_i = \mathbf{H}_i^{nest}\mathbf{x}^{nest} + \mathbf{H}_i^{bkg}\mathbf{x}^{bkg} + y_i^{ini} \tag{2}$$

where $\mathbf{H}^{nest}$ and $\mathbf{H}^{bkg}$ are the transport operators for the nested domain and for the background (i.e. everywhere outside the domain), respectively. If LNESTED = true (in SETTINGS_config) then nested FLEXPART output will be used to calculate $\mathbf{H}^{nest}$ and global output will be used to calculate $\mathbf{H}^{bkg}$, otherwise the global output will be used to calculate both. The inversion can optionally optimize offsets from the prior fluxes rather than the fluxes themselves (the option "offsets"). In this case, Eq.2 becomes:

$$y_i = \mathbf{H}_i^{nest}\mathbf{x}^{offset} + \mathbf{H}_i^{nest}\mathbf{x}^{nest} + \mathbf{H}_i^{bkg}\mathbf{x}^{bkg} + y_i^{ini} \tag{3}$$

where $\mathbf{x}^{\text{offset}}$ contains the offsets from the prior fluxes.

The initial mixing ratio, y_i^{ini} is the contribution from mixing ratios at the time and locations where the FLEXPART back-trajectories terminate, in other words, it accounts for the *history* of the atmosphere up to this moment. The initial mixing ratio for a given time and location can be modelled as:

$$y_i^{ini} = \mathbf{H}_i^{ini}\mathbf{c}^{ini} \tag{4}$$

where $\mathbf{c}^{ini}$ is a global field of mixing ratios at the time when the trajectories terminate and the elements of $\mathbf{H}^{ini}$ are defined as:

$$h_{ij}^{ini} = \frac{n_{ijk}}{J_k} \tag{5}$$

where n is the number of virtual particles terminating in grid cell j from trajectory k and J is the total number of particles released for the trajectory.

6.2. Transformations to the aggregated grid

If an aggregated grid is used for the inversions, then the SRRs (from the grid_time files) and the fluxes are transformed to this grid. The aggregated grid is described by a 2D array (lon × lat) at the same resolution as the SRRs (and prior fluxes) containing values of 1 to N (or if USELANDUSE = true, then 1 to N_{land} for land grid cells and -1 to $-N_{ocean}$ for ocean grid cells). This array is saved in nbox_xy.txt. The transformation from the input resolution to the aggregated grid is simply a mapping operation

$$\mathbf{H}_{i,n}^{box} = \mathbf{M}\mathbf{H}_{i,n}^{nest} \tag{6}$$

where $\mathbf{H}_{i,n}{}^{box}$ is the transport operator for the grid cell aggregates (or *boxes*), **M** is the mapping operator, and *n* is the footprint time-step. The fluxes are similarly transformed from the input resolution to the aggregated grid to give the state vector, **p**. The expression for the mixing ratios the becomes:

$$y_i = \mathbf{H}_i^{box}\mathbf{p} + \mathbf{H}_i^{bkg}\mathbf{x}^{bkg} + y_i^{ini} \tag{7}$$

or in the case "offsets" is true:

$$y_i = \mathbf{H}_i^{box}\mathbf{p} + \mathbf{H}_i^{nest}\mathbf{x}^{nest} + \mathbf{H}_i^{bkg}\mathbf{x}^{bkg} + y_i^{ini} \tag{8}$$

in which case **p** contains the offsets from the prior fluxes and the contribution from the prior fluxes ($\mathbf{H}_i^{nest}\mathbf{x}^{nest}$) is modelled at the full resolution of the SRRs.

6.3. Baye's Theorem

FLEXINVERT+ uses Bayesian statistics to optimize the fluxes. Baye's theorem states that the probability of the fluxes, *x* given a set of observations, *y* can be expressed as:

$$\rho(x|y) = \frac{\rho(y|x)\rho(x)}{\rho(y)} \tag{9}$$

where $\rho(y|x)$ is the conditional probability of observing *y* given a set of fluxes. Assuming that the probability of the observations is one, and that the probability distribution is Gaussian, the following cost function can be derived:

$$J(\mathbf{p}) = \frac{1}{2}(\mathbf{p}-\mathbf{p}_0)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{p}-\mathbf{p}_0) + \frac{1}{2}(H(\mathbf{p})-\mathbf{y})^{\mathrm{T}}\mathbf{R}^{-1}(H(\mathbf{p})-\mathbf{y}) \tag{10}$$

where **p** is the state vector and $\mathbf{p}_0$ is its prior estimate, **y** is the observed mixing ratios minus the fixed model contributions ($\mathbf{H}^{nest}\mathbf{x}^{nest}$ + $\mathbf{H}^{bkg}\mathbf{x}^{bkg}$ + $\mathbf{y}^{ini}$), **B** is the prior error covariance matrix and **R** is the observation error covariance matrix. Since (8) is quadratic, the first derivative equals zero at the **p** which minimizes this equation. The first derivative is:

$$J'(\mathbf{p}) = \mathbf{B}^{-1}(\mathbf{p}-\mathbf{p}_0) + (H'(\mathbf{p}))^{\mathrm{T}}\mathbf{R}^{-1}(H(\mathbf{p})-\mathbf{y}) \tag{11}$$

6.4. Analytical solution

The analytical solution solves (11) directly to find the state vector that minimizes the cost function. Since *H* can be defined as a matrix operator **H**, then $(H'(\mathbf{p}))^{\mathrm{T}}$ simply becomes $\mathbf{H}^{\mathrm{T}}$. The optimal fluxes are then found according to:

$$\mathbf{p} = \mathbf{p}_0 + (\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{p}_0) \tag{12}$$

for which the matrix ($\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ + $\mathbf{B}^{-1}$) of dimension $nq \times nq$ (where *n* is the number of flux time-steps and *q* is the number of state variables per time-step) needs to be inverted. An alternative equivalent formulation is possible:

$$\mathbf{p} = \mathbf{p}_0 + \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}(\mathbf{y}-\mathbf{H}\mathbf{p}_0) \tag{13}$$

for which the matrix ($\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ + **R**) of dimension $m \times m$ (where *m* is the number of observations) needs to be inverted. FLEXINVERT+ uses the formulation for which the smaller of the two matrices is inverted. In cases where the transport matrix **H** is too big to store in memory, FLEXINVERT+ has the option of using the conjugate gradient method (see Section 6.5).

6.5. Conjugate gradient solution

The conjugate gradient method is a numerical method to find the state vector that minimizes the cost function. FLEXINVERT+ uses a method based on the Lanczos algorithm, which finds the eigenvalues and eigenvectors of the Hessian matrix ($\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}$). The algorithm is iterative requiring the gradient of the cost function to be re-calculated for each iteration. The conjugate gradient method solves equations of the form:

$$f(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T\mathbf{A}\mathbf{z} - \mathbf{z}^T\mathbf{h} + \mathbf{c} \tag{14}$$

where $\mathbf{A}$ is unknown but the gradient $f'(\mathbf{z})$ is known:

$$f'(\mathbf{z}) = \mathbf{A}\mathbf{z} - \mathbf{h} \tag{15}$$

For the problem at hand, $\mathbf{A} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})$ and $\mathbf{z} = \mathbf{p} - \mathbf{p}_0$, thus $\mathbf{A}$ is the Hessian matrix.

A pre-conditioning is applied to transform the state vector from *physical* space to the optimization or *chi* space. The transformation is:

$$\chi = \mathbf{B}^{-1/2}(\mathbf{p} - \mathbf{p}_0) \tag{16}$$

so that $\mathbf{A} = (\mathbf{I} + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})$ where $\mathbf{I}$ is the identity matrix. With this transformation, the eigenvalues of the Hessian have a minimum value of 1 and the algorithm should converge faster. Although $\mathbf{A}$ is the reciprocal of the posterior error covariance matrix, the approximation of $\mathbf{A}$ from the Lanczos algorithm gives only a poor estimate of posterior error, since the largest eigenvalues of $\mathbf{A}$ are the smallest eigenvalues of the posterior error covariance matrix. Therefore, to find the posterior uncertainty, the Monte Carlo method is recommended.

6.6 Quasi-Newton solution

An alternative numerical method to the conjugate gradient is provided, that is, the M1QN3 Quasi-Newton algorithm. This algorithm has been developed for very large numerical problems and, unlike the conjugate gradient algorithm, may be used in cases where the first derivative of the cost function is non-linear (e.g. if a log-normal probability distribution function is used). Note, however, that this method does not approximate $\mathbf{A}$, so the posterior uncertainty must be estimated using the Monte Carlo method.

6.7. Optimization of CO_2 fluxes

FLEXINVERT+ currently only optimizes the net ecosystem exchange (NEE) fluxes of CO_2. However, the other fluxes of CO_2, namely, biomass burning, fossil and bio-fuel emissions as well as ocean fluxes need to be accounted for. For this, the user has to provide flux estimates of these components in addition to the prior estimate of NEE (these additional fluxes are specified in SETTINGS_files). Atmospheric mixing ratios of CO_2 are then modelled as:

$$y_i = \mathbf{H}_i^{box}\mathbf{p} + \mathbf{H}_i^{nest}\mathbf{x}_{fix}^{nest} + \mathbf{H}_i^{nest}\mathbf{x}_{nee}^{nest} + \mathbf{H}_i^{bkg}\mathbf{x}_{total}^{bkg} + y_i^{ini} \tag{17}$$

with the extra terms $\mathbf{H}_i^{nest}\mathbf{x}_{fix}^{nest}$ and $\mathbf{H}_i^{nest}\mathbf{x}_{nee}^{nest}$ compared to (6). The first of these is the contribution to the mixing ratio from fluxes inside the inversion domain that are not optimized, i.e., $\mathbf{x}_{fix} = \mathbf{x}_{ff} + \mathbf{x}_{bbg} + \mathbf{x}_{ocn}$ (where *ff*, *bbg*, and *ocn* are the fossil (and bio) fuel, biomass burning and ocean emissions). The second is the prior diurnal cycle of NEE. Note that in the case of CO_2, the state vector $\mathbf{p}$ contains *increments* of NEE (rather than NEE itself), which can be optimized at sub-daily time intervals (e.g. 6 or 12 hours) allowing the diurnal cycle also to be adjusted. The sub-daily time intervals may also be averages over more than one day. It is also important to note that the prior fluxes for NEE and fossil fuel are

averaged/interpolated to the time resolution of the SRRs used in $\mathbf{H}^{nest}$ while the biomass burning and ocean fluxes are used at monthly resolution.

6.8. Prior error covariance matrix

The prior error covariance matrix, **B** can be broken down into spatial and temporal covariances:

$$\mathbf{B} = \mathbf{C}_T \otimes \mathbf{B}_S \tag{18}$$

where $\mathbf{C}_T$ is the temporal error correlation matrix, $\mathbf{B}_S$ is the spatial error covariance matrix for a single time step, and $\otimes$ is the Kronecker product. Each element of $\mathbf{B}_S$ is calculated as:

$$b_{ij} = \sigma_i \sigma_j \exp\left(-\frac{d_{ij}}{D}\right) \tag{19}$$

where σ_i is the prior uncertainty for grid cell i, d_{ij} is the distance between grid cells i and j, and D is the temporal correlation scale length. The temporal error correlation matrix $\mathbf{C}_T$ is calculated as:

$$c_{ij} = \exp\left(-\frac{t_{ij}}{T}\right) \tag{20}$$

where t_{ij} is the time difference between two flux time-steps and T is the temporal correlation scale length. Note that the matrix **B** is not stored in memory, but only the eigen-decomposition of $\mathbf{B}_S$ and the matrix $\mathbf{C}_T$, which are used in all calculations involving **B**.

6.9. Observation error covariance matrix

The observation error covariance matrix **R** is currently defined as a diagonal matrix with the diagonal elements equal to the observation variance:

$$\sigma_i^2 = \sigma_{meas}^2 + \sigma_{ini}^2 + \sigma_{bkg}^2 + (\sigma_{ff}^2) \tag{21}$$

where σ_{meas} is the measurement uncertainty, σ_{ini} is the uncertainty in the initial mixing ratio y^{ini}, σ_{bkg} is the uncertainty in the background mixing ratio $\mathbf{H}^{bkg}\mathbf{x}^{bkg}$ and, for CO_2, σ_{ff} is the uncertainty in the modelled mixing ratio contribution from fossil fuel fluxes, $\mathbf{H}^{nest}\mathbf{x}_{ff}^{nest}$.

7. Quick reference guide

This section provides a quick reference guide summarizing all the steps needed to prepare and run FLEXINVERT+.

1. **Prepare observations and FLEXPART runs**
   Using the pre-processor *prep_flexpart*:
   1) Compile using "make"
   2) Edit the SETTINGS file
   3) Edit the bash script: job_prep_flexpart.sh (or alternatively for slurm: sbatch_prep_flexpart.sh)
   4) Run the bash script: ./job_prep_flexpart.sh or ./sbatch_prep_flexpart.sh)

2. **Prepare fluxes on the FLEXPART grid**
   Using the pre-processor *prep_fluxes*:
   1) Compile using "make"
   2) Edit the file FLUXES
   3) Edit the bash script: job_prep_flux.sh (or alternatively for slurm sbatch_prep_flux.sh)
   4) Run the bash script: ./job_prep_flux.sh (or ./sbatch_prep_flux.sh)
   5) If using a nested domain, repeat steps (2) to (4) for the nested resolution

3. **Prepare regions definition file (optional)**
   Using the pre-processor *prep_regions*:
   1) Compile using "make"
   2) Edit settings/SETTINGS_config and settings/SETTINGS_files (i.e. same settings files as used for the inversion)
   3) Edit SETTINGS_regions
   4) Edit the bash script: job_prep_regions.sh (or alternatively for slurm: sbatch_prep_regions.sh)
   5) Run the bash script: ./job_prep_regions.sh (or ./sbatch_prep_regions.sh)

4. **Run FLEXINVERT+**
   1) Compile using "make"
   2) Edit settings/SETTINGS_config and settings/SETTINGS_files
   3) Edit the bash script: job_flexinvert.sh (or alternatively for slurm: sbatch_flexinvert.sh)
   4) Run the bash script ./job_flexinvert.sh (or ./sbatch_flexinvert.sh)

Appendix A

Pseudo-code to replace "init_cini.f90" for the user's own method of calculating the background contribution. Note the data structure "obs" must be passed to the subroutine as in this example code.

```
subroutine myroutine(obs)

  use mod_var
  use mod_obs

  implicit none

  type (obs_t), intent (in out) :: obs
  integer                       :: i
  real, dimension(nobs)         :: mybkg

  ! my calculation of the background for each observation
  ! and assignment to the variable mybkg

  do i = 1, nobs
    obs%cini(i) = mybkg(i)
  end do

end subroutine myroutine
```