CHEEREIO: a generalized, open-source ensemble-based chemical data assimilation and emissions inversion platform for the GEOS-Chem chemical transport model

Drew C. Pendergrass[1], Daniel J. Jacob[1], Hannah O. Nesser[1], Daniel J. Varon[1], Melissa Sulprizio[1], Kazuyuki Miyazaki[2], and Kevin W. Bowman[2]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]NASA Jet Propulsion Laboratory, Pasadena, CA, USA.

Abstract. We present a general, open-access, user-friendly chemical data assimilation toolkit for simultaneously optimizing emissions and concentrations of chemical species based on atmospheric observations from satellites or suborbital platforms. The CHemistry and Emissions REanalysis Interface with Observations (CHEEREIO) exploits the GEOS-Chem chemical transport model and a localized ensemble transform Kalman filter algorithm (LETKF) to determine the Bayesian optimal (posterior) emissions and/or concentrations of a set of species based on observations and prior information, using an easy-to-modify configuration file with no change to the GEOS-Chem or LETKF code base. The LETKF algorithm readily allows for non-linear chemistry and includes posterior error covariances from the ensemble spread. The object-oriented Python-based design of CHEEREIO allows users to easily add new observation operators such as for satellites. CHEEREIO takes advantage of the HEMCO modular structure of input data management in GEOS-Chem to update emissions and concentrations from the assimilation process independently from the GEOS-Chem code, and thus can seamlessly support GEOS-Chem version updates or other chemical transport models with similar modular input data structure. A postprocessing suite combines ensemble output into consolidated NetCDF files and supports a wide variety of diagnostic plots and animations. We demonstrate CHEEREIO's capabilities with an out-of-the-box application, assimilating global methane emissions at weekly temporal resolution and 2°x2.5° spatial resolution for 2019 using TROPOMI satellite observations.

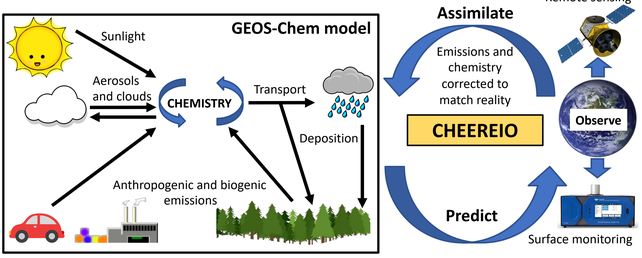
What is data assimilation?

Data assimilation is a field of applied mathematics that studies the most probable combination of a physical model and observational data to define the state of a system. In atmospheric chemistry, we run a chemical transport model (CTM; in our case, GEOS-Chem) and compare its output with data from the real world. We then update model parameters, like emissions, to match reality.

Most data assimilation algorithms involve the optimization of a Bayesian scalar cost function $J(\boldsymbol{x})$:

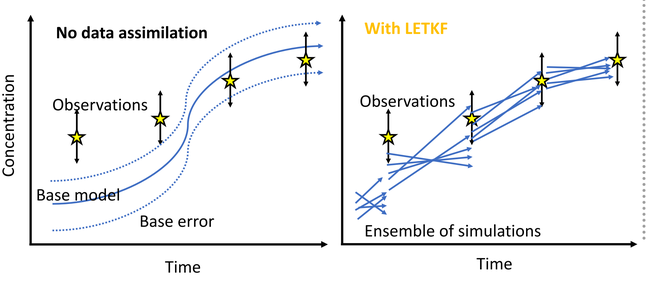
$$J(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x^b})^T (\boldsymbol{P^b})^{-1} (\boldsymbol{x} - \boldsymbol{x^b}) + (\boldsymbol{y} - H(\boldsymbol{x}))^T \boldsymbol{R}^{-1} (\boldsymbol{y} - H(\boldsymbol{x}))$$

Here $\boldsymbol{x}$ is the state vector to be optimized, $\boldsymbol{x^b}$ is the physical model prediction of the state vector, $\boldsymbol{P^b}$ is the background error covariance matrix of the model prediction, $\boldsymbol{y}$ is the suite of observed atmospheric concentrations arranged as a vector, $H(\cdot)$ is an observation operator that transforms the state vector $\boldsymbol{x}$ from the state space to the observation space, and $\boldsymbol{R}$ is the observational error covariance matrix. Solving for the minimum of the cost function ($\nabla J(\boldsymbol{x}) = \boldsymbol{0}$) defines the optimized analysis (also called posterior) estimate $\boldsymbol{x^a}$ for the state vector.



The LETKF algorithm

There are many algorithms for data assimilation, but we choose the Localized Ensemble Transform Kalman Filter because of its flexibility. LETKF can be readily applied to nonlinear problems; however, unlike the other methods, LETKF avoids the need for the adjoint of the CTM because it is powered by an ensemble of CTM simulations which capture the nonlinearity of the system. Each ensemble member is initialized with random perturbations applied to emissions of interest, and the ensemble is evolved for the assimilation time window using the CTM. At assimilation time, the ensemble spread is used to approximate the background error covariance matrix $\boldsymbol{P^b}$ and from there solve for the minimum of the cost function. The ensemble is then updated to reflect the optimized state, including emissions and concentrations, and the cycle repeats. The LETKF continuously improves with each iteration.



CHEEREIO allows you to easily calculate emissions updates from observations of atmospheric composition, and lots more!

Download and use* the code

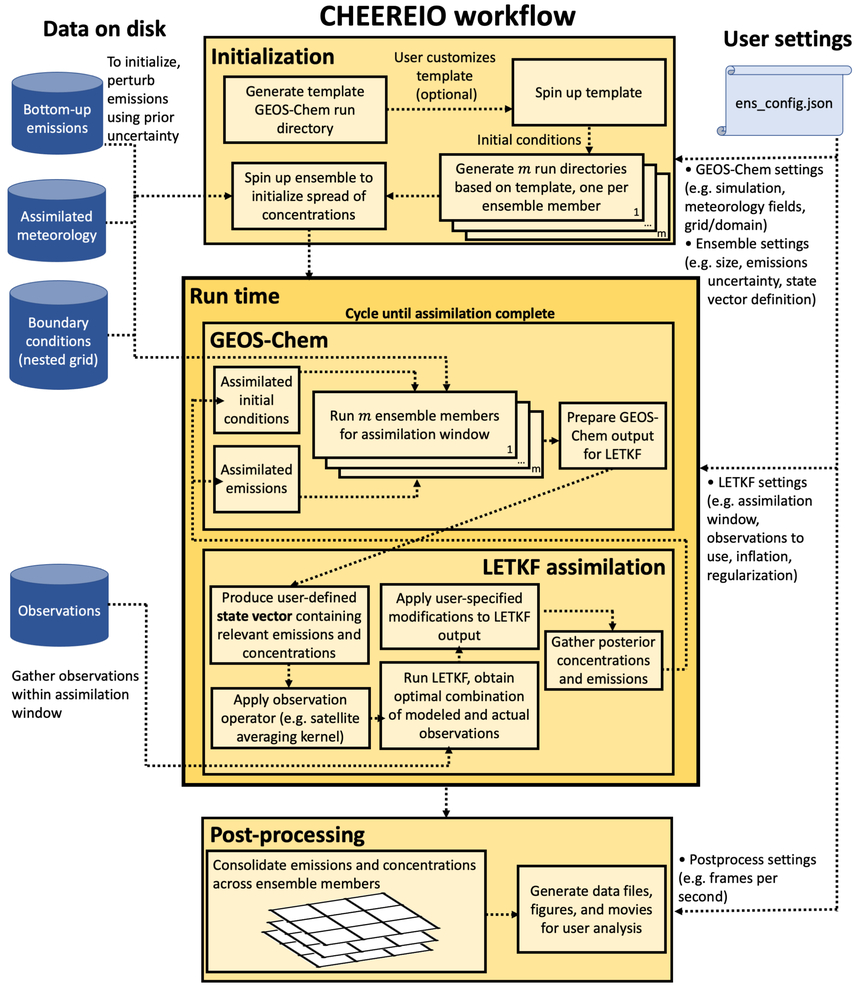
bit.ly/cheereio

Read the documentation

bit.ly/CheereioDocs

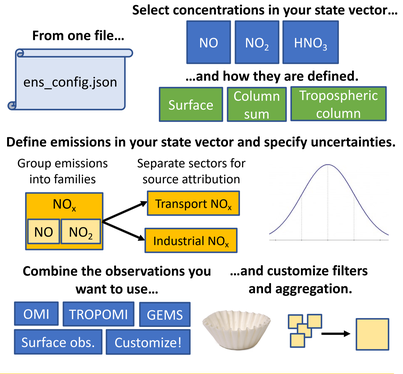
**Official release and model description paper coming soon!*

CHEEREIO workflow



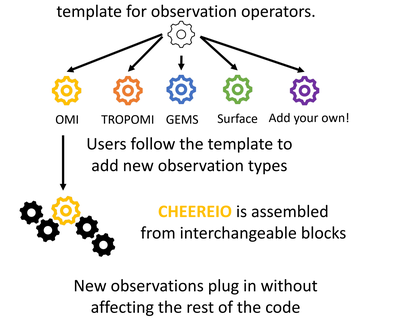
Configuring CHEEREIO

LETKF assimilation is implemented in CHEEREIO using a structure of nested and interchangeable Python objects. CHEEREIO gathers building blocks according to the settings users specify in their configuration file, and then automatically combines them into a unique simulation that can be submitted in one line.



New observation operators

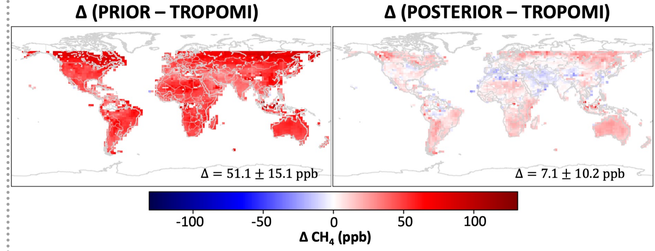
A major benefit of CHEEREIO's modular design is that new observation operators can immediately plug into CHEEREIO and work automatically. Users do not need deep knowledge of the CHEEREIO code structure to add new observations.



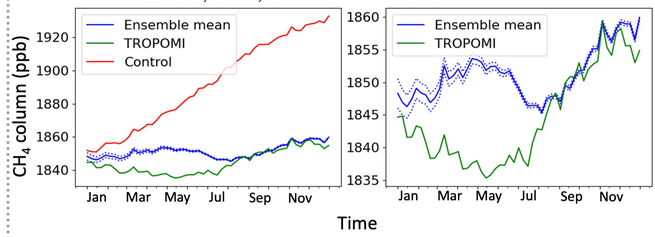
Demonstration with 2019 TROPOMI CH_4

Here we demonstrate an end-to-end example application of CHEEREIO to the problem of optimizing global emissions of methane at 2.0° x 2.5° spatial and weekly temporal resolution by assimilation of dense TROPOMI satellite observations of CH_4 for 2019 using a 24-member ensemble. The simulation here was specified purely using CHEEREIO configuration files, and all figures and statistics in this section are automatically produced by CHEEREIO — no programming was required to obtain these results and corresponding figures, and the simulation could be reproduced with minimal effort by any CHEEREIO user who has the same configuration files.

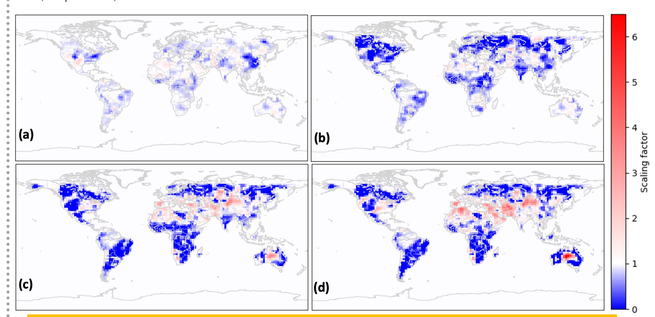
The figure below compares simulated methane columns from GEOS-Chem with TROPOMI observations for all of 2019. We find that model bias is considerably reduced by the LETKF assimilation procedure, with a mean bias of 7.1±10.2 ppb in the ensemble (assimilated) mean against 51.1±15.1 ppb in the prior (no assimilation) simulation.



The figure below shows that GEOS-Chem without assimilation (control line at left) leads to a high modelled CH_4 column relative to TROPOMI, likely due to emissions inventories. For the first few months of the simulation, CHEEREIO falls between the control simulation and TROPOMI values before following TROPOMI closely for the second half of 2019. Despite a high methane overestimate in the prior, CHEEREIO is in time able to adjust the system to better match observations.



The below figure shows assimilated methane emissions expressed as a multiplicative scaling factor adjustment from the prior inventory (both anthropogenic and natural sources) for the start of March, June, September, and December 2019.



Acknowledgements

This work was funded by the NASA Carbon Monitoring System. DCP is funded by an NSF Graduate Research Fellowship Program (GRFP) grant.

Contact information and links

Contact Drew Pendergrass at **pendergrass@g.harvard.edu**

Download and use the CHEEREIO code: **bit.ly/cheereio**

Read more in the documentation: **bit.ly/CheereioDocs**

Ask me why you should use CHEEREIO in your research!