Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-407-RC1, 2021
© Author(s) 2021. This work is distributed under
the Creative Commons Attribution 4.0 License.

Geoscientific Model Development Discussions

Interactive comment on "The Community Inversion Framework v1.0: a unified system for atmospheric inversion studies" *by* Antoine Berchet et al.

Anonymous Referee #1

Received and published: 17 February 2021

The submitted paper presents the Community Inversion Framework (CIF) to help rationalise development efforts and leverage the strengths of individual inversion systems into a comprehensive framework. The CIF is primarily a programming protocol to allow various inversion bricks to be exchanged among researchers. The system is supposed to allow running different atmospheric transport models, different observation streams and different data assimilation approaches. The paper describes a system that will bring a major software advances for future inversion frameworks. The presentation is clear and well written. I have few main major concerns detailed in the main comments below.

Main comments:

The paper advertises and the multi-CTM or model capability, but this is not shown in the paper only mentioning a future paper. This capability needs to be showcased in the paper or such claims should be removed or rephrased.

It is a bit limiting to narrow the scope only on GHG as inversion of reactive species and primary aerosols is quite relevant as well. Especially if CIF plans to, or already includes the CHIMERE CTM which is originally designed for air quality applications.

The ensemble and analytical results point out possible problems in the implementation of the EnKF and in the analytical solution using correlation length scales. It is not entirely clear on how the system uses the ensemble information to perform the inversion. This ensemble-based inversion should not be called an EnKF as it is not sequential. This needs to be revised or diagnosed or explained.

Finally, the testing of system using the Gaussian plumes on my local machine (Mac) was impossible for the time I gave in. I regularly use python and successfully setup conda on my local machine. But it seems that the CIF is dependent on many specific packages that bring conflict issues and are not straightforward to install (specifically GDAL seems problematic). It required some research to find different commands to install the packages from what is indicated on the online documentation. After trying for hours, I had to give up as my time to revise this paper is limited.

Specific comments:

P3, L8: Define better what a unified system would be here.

P3, L12: OOPS main goals are on NWP. The CAMS system will run on OOPS (which is based on IFS) which will run inversions in the future. Maybe it is worth mentioning the CoCO2 projects (formerly CHE).

The authors should also mention the Joint Effort for Data Integration led by UCAR/JCSDA. https://www.jcsda.org/jcsda-project-jedi.

P3, L13-14: DART is an EnKF (EAKF to be more precise) that allows to run on the CESM earth system model. There is also a WRF mode. I would not call this unified or modular system compared to OOPS, CIF or JEDI.

Also, DART can run atmospheric composition and recently chemical species inversions as well (see Gaubert et al., 2020).

P5, L10: Some of the models mentioned here use semi-lagrangian advection scheme. Maybe remove Eulerian and use a different word.

P5, L11: Mizzi et al., 2016 is a WRF CHEM paper and also do not perform any source inversion but only concentration assimilation. Please consider removing this reference.

Equation 2: Does N indicates a normal distribution? Please clarify. Also, in the second brace element, is this zero in the first moment? If yes, I guess this assumes unbiased observation and prior. I know this is a requirement for data assimilation formulation. But often not the case in practice in atmospheric composition. Maybe you can detail and develop this in the text.

P6, L22-25: This is an important point, maybe the authors could expand more on this for the non-data assimilation expert?

P7, L8: Are control space and target space the same thing or different? Please be consistent (I would recommend control space as it is more generally used in atmospheric DA) or explain the differences.

P7, L12: This part of the sentence is not very clear. Replace "... problems such as the variational and the ensemble Kalman filter methods which are described below."

P7, L18: “limited non-linearity”: This is not necessarily true, or it is misleading. EnKFs are frequently applied successfully on chemical transport models which can be significantly non-linear systems in the polluted boundary layer for example.

P7, L25: “running computation window”: I believe the authors refer to what is commonly

called assimilation window? (Note that inversion is part of data assimilation techniques, so it is fine to call it data assimilation window in the inversion framework). Please be consistent in the terminology in this paragraph. “Smaller”, smaller than what? I understand that for GHG inversions especially CO2 you need long window given the observation network and the CO2 atmospheric lifetime, but this is not general to all atmospheric composition inversion.

P7, L29-31: “for very dense . . . may not be sufficient”: This is not necessarily true, or the statement is misleading. This also needs to be proven. If the reference exists please add it. One can foresee that with a better coverage and higher satellite sensitivity towards the surface the window length could be reduced hence the number of observations to go through sequentially is reduced. Please correct or remove this statement or justify this more clearly.

P8, L11: “under-estimating”: This is not necessarily true. Small ensemble size can over and underestimate the uncertainty. More importantly they introduce spuriousness. Please clarify the statement.

P8, L13-14: Could you possibly provide few examples of different problems having different approximations?

P8 L15-28: I recommend revising this introduction of the variational method. Add a sentence to present the cost function. Put the equation right after. Explain in few sentences that the aim is to minimize this cost function and that to achieve this the gradient is calculated. Then put the gradient formula.

P8, L25: I am not sure this is true. If the state vector is large, i.e. increasing resolution or augmenting the control variables, or increasing the number of observations, the minimization is expected to be slower.

P12, Equation 12: Please, be explicit about what the index represents. "... of elementary interchangeable N transformations ..."

P13, L10: It is a bit of pity that air quality application is only mentioned here. I think a system such as CIF could greatly benefit the air quality community as much as the GHG community, which are getting more connected lately.

P13, L13: I am not sure what the authors meant by spatial gradient as observations? Please clarify or remove.

P13, L14-15: Maybe the authors could mention here that the main idea behind “super-obing” is to lower the discrepancies of the representativeness between model and observations.

P13, L15-16: “This is the case for . . . time and locations,” This sentence is very unclear. Please rephrase or explain better.

P15, L21-31: What about satellite observations? I think the authors should add statement about this.

P17, L11: Those studies use Gaussian plume technique but do not provide a continuous inversion constrain globally. Please correct the statement as it makes the reader believes that the Gaussian plume technique would allow a global inversion similar to a CTM based inversion.

P19, L28: 10000m is 4 times the size of the domain here. No data assimilation system would use such long length scale. Is this on purpose? If yes, please justify.

P20, L16, Fig 5 and similar other figures. With all the fluxes and fluxes differences plots it would be better to use python colour maps that are diverging and centred around 0. This will improve the readability of the figures. P20, L20: I am unsure if I understand correctly but I believe the authors are speaking about iterations in the variational sense and members in the EnKF sense. If yes, please have few sentences to explain what the word ’simulation’ means. Alternatively, I would replace the word ’simulations’ by ’iterations or members.’

P20, L32: 10000m is 4 times the size of the domain here. See the related com-

ment above. Such long length scales would provide very uniform structure on the increments. Structure similar to the 500m increments is seen on the result using the 10000m length scale. I expect there is typo through the text a 0 should be removed to 1000m?

P20,L33-P21,L2: This look to me more like a bug rather noise as vertical lines or “chess board” like pattern appear in the increments results. Such pattern also occurs in the analytical solution. This not looking like sampling noise to me. I think this points out possible problems in the implementation of the EnKF and of the systems other than the variational methods. It is not entirely clear if the system uses the ensemble information to derive jacobians for the calculation of H and perform a minimization? In this case this inversion cannot be called EnKF. This needs to be revised or diagnosed or explained.

Technical comments:

P2, L27-29: an order in all those references, chronological or alphabetical.

P7, L16: replace "... linear, simple cases." by "... linear and simple cases."

P7, L19: change “characterized” to “characterize”.

P9, L13: Chevallier et al., 2005: put this reference to the previous parenthesis for clarity.

P10, L5: There is a colon here but nothing after. Should we expect equations?

P15, L18: I understand what the authors means but this need to be reformulated. “Meteos” is not really correct English.

P19, L3-10: Maybe add the equations in the text where they are mentioned.

P22, L6: This is not a fair statement considering on how the “EnKF” has been implemented here. It doesn’t seem to be strictly an EnKF as the sequential assimilation is not performed. Consider changing or removing this in the conclusion. Also, this is not the scope of GMD paper here to compare methods but to describe and showcase

model/software developments for geoscience.

P22, L10-12: The authors could add some sentences about the challenges in implementing actual CTMs in the CIF framework. For example, among many other challenges, I/O is a limiting factor for data assimilation especially with CTMs.

References:

Gaubert, B., Emmons, L. K., Raeder, K., Tilmes, S., Miyazaki, K., Arellano Jr., A. F., Elguindi, N., Granier, C., Tang, W., Barré, J., Worden, H. M., Buchholz, R. R., Edwards, D. P., Franke, P., Anderson, J. L., Saunois, M., Schroeder, J., Woo, J.-H., Simpson, I. J., Blake, D. R., Meinardi, S., Wennberg, P. O., Crounse, J., Teng, A., Kim, M., Dickerson, R. R., He, H., Ren, X., Pusede, S. E., and Diskin, G. S.: Correcting model biases of CO in East Asia: impact on oxidant distributions during KORUS-AQ, Atmos. Chem. Phys., 20, 14617–14647, https://doi.org/10.5194/acp-20-14617-2020, 2020.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-407, 2020.