Responses to Anonymous Referee #1:

1) Major issues:

- *My question about retrieval error as a function of satellite altitude: a same instrument receives less photons on average when placed at a higher altitude, and the uncertainty of the column retrieval consequently (and significantly) increases. This effect does not seem to be accounted for here.*

Response:

The effect of sensor altitude on the errors is generally mitigated in radiance space by integrating the measurement over a longer time sample to increase the signal to noise (SNR) of the final radiance. It has been shown with other sensors (e.g. TEMPO, Chance et al., (2013)) that GEO SNR's are similar to LEO SNR's. Moreover, in a geostationary configuration, the possibility to sample a scene at high time frequency can further reduce the effect of altitude on the measurement error.

- *My question about the absence of prior error correlations: the authors justify their diagonal matrix by the impossibility to compute "accurate" or "rigorous" prior error correlations (l. 188, l. 439). However, in the absence of perfect knowledge, the authors do not justify why zero is better than, e.g., an e-folding length of 300 km. Taking the extreme case may not be a fair choice, which the authors implicitly acknowledge l. 308 by speculating on a resulting systematic bias in their results (actually more details on this speculation would be needed because, if the existence of a bias seems obvious, its "sign" is not to me).*

Response:

After having given more thoughts on the sign of the bias in the DOF when correlations are neglected, we concluded that the bias can be either positive or negative. We rigorously demonstrate that statement below.

The matrix **B** can be written:

$$\mathbf{B} = \Sigma\mathbf{C}\Sigma, \qquad (1)$$

where Σ is the diagonal matrix of variances and $\mathbf{C}$ the matrix of correlations. For the sake of simplicity, let us assume that all variances are unity ($\Sigma = \mathbf{Id}$). If there is no correlation, one has $\mathbf{C} = \mathbf{Id}$ and $\mathbf{B} = \mathbf{Id}$.

Adding correlations results in:

$$\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^T (\neq \mathbf{Id}), \qquad (2)$$

where $\mathbf{V}$ is the matrix column of singular vectors $\mathbf{v}_i$ of $\mathbf{B}$ and $\mathbf{D}$ is a diagonal matrix with at least one element $d_p > 1$ and at least one element $d_l < 1$. This follows from the invariance of the trace of a matrix by similarity (i.e., change of basis) which requires $\sum_i d_i = n$ for unit variances.

In the general case, it can be shown that the averaging kernel (or model resolution) matrix can be written:

$$\mathbf{A} = \mathbf{B}^{1/2}\mathbf{W}\mathbf{\Omega}\mathbf{W}^T\mathbf{B}^{-1/2}, \quad (3)$$

where $\mathbf{W}$ is the matrix column of singular vectors $\mathbf{w}_i$ of $\mathbf{B}^{1/2}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2} = \mathbf{W}\Lambda\mathbf{W}^T$ (where **H** and $\mathbf{H}^T$ are the tangent linear and adjoint models, respectively, and **R** the covariance matrix of observational errors), and the diagonal elements of $\mathbf{\Omega}$ and Λ are related by $\omega_i = \lambda_i(1+\lambda_i)^{-1}$. Therefore the DOF is given by:

$$\mathrm{DOF} = \sum_i \frac{\lambda_i}{1+\lambda_i}, \quad (4)$$

It is a monotonic increasing function of the λ_i.

Now let us consider the particular case where $\mathbf{B} = \mathbf{W}\mathbf{D}\mathbf{W}^T$ and $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} = \mathbf{W}\Delta\mathbf{W}^T$, i.e. a prior error covariance matrix that is diagonalizable in the same basis as $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ (the observational errors projected onto the control space). In practice, identity of the singular modes for the prior and projected observational errors would correspond to error correlation structures that are identical (e.g., same correlation lengths in control space). In this case one has:

$$\mathrm{DOF} = \sum_i \frac{d_i\delta_i}{1+d_i\delta_i}, \quad (5)$$

where the d_i are the diagonal elements of **D** and the δ_i are the diagonal elements of Δ. Note that $d_i = 1, \forall i$, correspond to the case with unit variances and no correlations. Adding correlations results in the existence of a subset of *p* indices $i_1,..,i_p$ with $d_i > 1, i \in i_1,..,i_p$, $d_i < 1, i \notin i_1,..,i_p$. Therefore, in this particular case it appears clearly that the DOF can either increase or decrease when prior error correlations are added, since the only conditions are $\sum_i d_i = n$ (**B** with unit variance).

In dimension *n*=2, consider for instance $(d_1 = 0.5,\ d_2 = 1.5)$ and $(\delta_1 = 1,\ \delta_2 = 10)$, which results in a DOF decrease of ~+0.14 compared to a case with no correlations, and $(d_1 = 1.5,\ d_2 = 0.5)$ and $(\delta_1 = 1,\ \delta_2 = 10)$, which corresponds to a DOF decrease of ~-0.02.

Based on this analysis, the text L.310-325 has been modified, and it is now stated that the fact that prior error correlations are neglected **may** lead to an overestimation of the DOF and partially explain the discrepancy found between our results and those from a previous study by Turner et al. (2015). It is now stated that: "the DOFs we derived should therefore be interpreted with caution, but can provide useful insights into the relative magnitude of the constraints afforded by different instruments and orbit configurations. These results also correspond to the limit to which the observational constraints would tend as the effective spatial resolutions of the bottom-up CH4 inventories are increased."

- L. 251: am I missing something or have the authors forgotten about model temporal error correlations? There are likely very large from one hour to the next (see, e.g., Lauvaux et al., doi:10.5194/bg-6-1089-2009).

Response:

In our setup temporal variability of the emission is assumed to be a hard constraint at scales smaller than the assimilation window. This is now explicitly stated at the end of section 2.2 (see text in red).

2) Minor issues:

- What is the link between l. 89 "their computational cost can be prohibitive, since many perturbed inversions (typically about 50) are needed" and l. 129 "Here an ensemble of 500 random gradients of the cost function are used". The word "prohibitive" seems to be author-centric.

Response:

The difference between these examples is that one entails 50 inversions (each with ~50 iterations costly 2500 gradient calculations) vs 500 gradient calculations. However, the term prohibitive has been removed. Also, the text has been replace by: "[...] their computational cost can be extremely high. Indeed, many perturbed inversions (typically about 50) are needed, each of them requiring numerous forward and adjoint model integrations (iterations) in case the problem is not well-conditioned (about 50 iterations for our methane inversion)."

- At several places (at least l. 223 and 255), the authors use the generic word "error" in place of "error standard deviation". The authors should use the exact expression.

Response:

This has been corrected in the revised manuscript.

- L. 228: some details are needed about the vertically resolved covariance matrix.

Response:

We have updated the text in this section to include more information about the multi-spectral error covariance matrix (see text in red, L.230-237).

- L. 429: "nation" likely means the US, but it should be explicited.

Response:

"Nation" has been replaced by "US".