Manuscript: Intercomparison of remote sensing retrievals: an examination of prior-induced biases in averaging kernel corrections

H. Nguyen and J. Hobbs

August 2020

1 Responses to Reviewer 1

In this paper, the authors perform a linear analysis and accompanying numerical simulation to highlight the error induced by comparing multiple retrievals that is due to the differences in their averaging kernels alone. Such differences arise from different observational error specifications, prior state uncertainties, and prior mean states. The results are of some theoretical interest, but it is hard to place them in the context of other sources of errors in remote sensing retrievals

We would like to thank the reviewer for the kind comments, and for providing constructive inputs which helped improve our manuscript. Please see our detailed responses below.

1. **The results are of some theoretical interest, but it is hard to place them in the context of other sources of errors in remote sensing retrievals, and the assumptions that are made in the text simplify the retrieval problem to the point where it would likely be hard to do so by simple comparison with previous literature. To name a few examples: different retrievals use different forward models and different state vectors; different instruments will have different spectral resolutions, inter-calibration errors, etc.**

   This paper examines the averaging kernel correction process that is proposed by Rodgers and Connor (2003), and therefore we have made the same assumptions about the basic requirements. For convenience, we reproduce the relevant passage from Rodgers and Connor (2003) below:

   > For simplicity, the observing systems should both retrieve the same target quantity. The definition of the state vector $\mathbf{x}$ includes the parameter retrieved, such as trace gas mixing ratio or

> concentration, but there may also be instrument-specific parameters in the state vectors that are not targets for comparison but need to be retrieved, such as surface reflectivity or instrumental offsets. The definition of $\mathbf{x}$ also includes the grid on which the targets for comparison are retrieved, and any interpolation rule implied. For the purpose of this paper we take all profiles to be represented on the same vertical grid, fine enough that representation error can be ignored. If a retrieval is carried out on a coarse grid, then an interpolation must be used before the methods described here are applied. It is possible to compare observing systems that retrieve different representations of the state, but this leads to extra complications in the comparison that we prefer to leave to a future publication. (Rodgers and Connor, 2003, Section 1.2, Basic Requirements)

The fundamental assumption in the passage above is that the two retrievals share a common state vector $\mathbf{x}$, which consists of the same geophysical quantities and at the same vertical grid. We have made the same assumption in our paper in Section 2.1.

Regarding the referee's comment about different forward models and spectral resolutions, we have allowed for it in the paper in the paragraph immediately below Eqn. (1), where we noted that the retrievals may have different forward models $\mathbf{F}_i(\cdot)$, where the subscript i "allows for the possibility of the state being observed on different sets of spectral channels across multiple instruments." Throughout the paper, we also used the subscript i on $\mathbf{K}_i$ to indicate that the two retrievals may have different Jacobians.

Regarding the inter-calibration errors, we assume that the referee is thinking of the measurement-error $\mathbf{S}_{\epsilon,i}$ or the misspecification of the forward model parameters. We note that Rodgers and Connor (2003) assumed that the forward model is correct, instrument calibration is done correctly, and $\mathbf{S}_{\epsilon,i}$ are correct for both instruments, and we have made the same assumptions in this paper. We also elaborate a bit more on this issue in the response to point 4) below.

To reiterate, we have duplicated the basic requirements of Rodgers and Connor (2003) in our paper, and therefore our conclusions should apply wherever it is appropriate to apply the averaging kernel correction process in Rodgers and Connor (2003).

2. **The one implication that the authors focus on, namely attempted algorithm improvements, is important but not the only one. Comparisons between satellite retrievals and validation data are often used to correct the satellite data. What are the implications of this work for the "corrected" satellite observations?**

To the best of our knowledge, the current practice in validation is to assess biases between remote sensing data and validation data (this is where we would apply the averaging kernel correction process). If any bias is found, then an attempt is made to trace it to factors such as calibration, spectroscopy, radiative transfer models, geolocation, non-linear optimizer or other potential causes. Our contribution to the literature is to note that there is a component of bias that is due to the choice of comparision prior $\{\mathbf{x}_C, \mathbf{S}_C\}$ alone. As we noted in the paper, if there is a prior-induced bias, then no amount of tinkering with calibration, spectroscopy, radiative transfer models, geolocation, non-linear optimizer or other potential causes will remove this bias. Our paper should be useful for validation team in identifying an extra component of contributory causes of bias.

Another simple corollary of our results, which we did not mention in the paper so as not to encourage bad research practice, is that the averaging kernel correction process could be tinkered with to produce whatever result a researcher wished. Since the bias is proportional to the difference $(\mathbf{x}_T - \mathbf{x}_C)$, a researcher can choose $\mathbf{x}_C$ so as to produce a positive bias, negative bias, or even zero bias depending on the desired narrative. In practice, it is common to pick $\{\mathbf{x}_C, \mathbf{S}_C\}$ from one of the two retrieval priors given in the datasets, but we have shown that in general $\mathrm{E}(\mathbf{x}_{12} - \mathbf{x}_1) \neq \mathrm{E}(\mathbf{x}_{21} - \mathbf{x}_2)$, and hence an unscrupulous researcher could cherry-pick the comparision ensemble as the one that produces a better 'story'.

As a mean to mitigate this situation, we have discussed some guidelines for ameliorating or eliminating this prior-induced bias in Section 4 (specifically the last 3 bullet points).

3. **Could large ensembles of retrievals and averaging kernels somehow reduce the impacts of this error? The proportionality of the bias to x_C-x_T suggests that this could be the case if the retrievals contain the true states in their envelope, as would be the case with an unbiased posterior state with scatter driven by noise.**

   This is a great comment. Our understanding of the comment is that since the bias is proportional to $(\mathbf{x}_C - \mathbf{x}_T)$, then if the misspecification of the true profile is random (that is, $\mathbf{x}_C$ is equally likely to 'undershoot' as it is to 'overshoot' $\mathbf{x}_T$), then over large ensembles of retrievals, the prior-induced bias will be reduced or eliminated all-together.

   There are two problems with the above. The first is that if the misspecification $(\mathbf{x}_C - \mathbf{x}_T)$ 'averages out' over many ensembles (and note that this is a *big* 'if'), then the biases that we discussed in Section 2 would be pushed

into the variability component. Recall that the posterior uncertainty in OE is computed as

$$\mathbf{\Sigma} \equiv \mathrm{var}(\hat{\mathbf{x}}_i - \mathbf{x}) = (\mathbf{S}_{a,i}^{-1} + \mathbf{K}_i' \mathbf{S}_\epsilon^{-1} \mathbf{K}_i)^{-1}, \tag{1}$$

Since the expression above does not take into account the extra variability introduced by the prior-misspecification, then this posterior uncertainty estimate would *underestimate* the true uncertainty. And hence we have traded one undesirable property for another.

A second issue with this, and this is more relevant to the problem, is that we *do not know* the distribution of $(\mathbf{x}_C - \mathbf{x}_T)$ over many ensembles, and hence we can not say how much, if at all, it would reduce the prior-induced bias. It is tempting to make conjectures that $(\mathbf{x}_C - \mathbf{x}_T)$ would 'average out' over many ensembles, but that is an extraordinary claim and it would require extraordinary evidence. In the absence of evidence or knowledge about the distribution of $(\mathbf{x}_C - \mathbf{x}_T)$ across multiple retrievals, we prefer to stay conservative and discuss only the fixed-point case.

4. **To make the results in this paper most useful to the community, I suggest the authors consider:**

   - **misspecified observation errors (i.e. the true observation error is unknown or the radiances have a non-random error) - this could make a large difference in how important this error source is- [and] a simple forward model error**

     We agree that it would be useful to have some demonstration of the relative magnitude of this prior-induced bias compared to those arising from instrument calibration, observation error). One problem, however, is that the relative magnitude of these bias are dependent upon the choice of Jacobian $\mathbf{K}_i$, measurement errors $\mathbf{S}_{\epsilon,i}$, working priors $\{\mathbf{x}_{a,i}, \mathbf{S}_{a,i}\}$, and the comparison ensemble $\{\mathbf{x}_C, \mathbf{S}_C\}$ in additional to whatever parameterization we choose to characterize the misspecification in instrument inter-calibration/observation error.

     For instance, we have shown that the prior-induced bias is proportional to the difference $(\mathbf{x}_T - \mathbf{x}_C)$, so the larger this difference, the larger the bias would be. In this paper we have chosen $\mathbf{x}_C$ to be misspecified by 1 or 2 % of the true value (which is probably overly optimistic), and we found that it already returned fairly significant errors (.6-1 ppm) with our chosen set of $\mathbf{K}_i$, $\mathbf{S}_{\epsilon,i}$, $\{\mathbf{x}_{a,i}, \mathbf{S}_{a,i}\}$, $\{\mathbf{x}_C, \mathbf{S}_C\}$. If we had chosen the misspecification to be 4-5 %, then the error would be proportionally larger.

To reiterate, while a simulation can illustrate the relative magnitude of the different sources of biases, those relative biases are highly dependent upon how we choose to parameterize the various sources of errors, and the relative magnitudes are likely to change for different applications with different forward models and Jacobians. We feel that one such simulation would be of limited use to the readers (especially those who are not familiar with OCO-2 and its XCO2 retrievals, since the results are unlikely to generalize to their instrument of interest).

Therefore, a more useful and informative approach would be to analytically derive effect of bias arising from different sources (intercalibration, observation error) under the effect of averaging kernel correction, similar to what we have done in Section 2. However, this is a massive undertaking, and certainly not one that we can do within the 2 week revision period given by the editor. This topic, we feel, would be better served as a subject for a different paper.

We note that our derivations can be used to give some indication of how large the prior-induced bias would be. We have shown that the prior-induced bias is given by the following equation,

$$\mathrm{E}(\hat{\mathbf{x}}_{12} - \hat{\mathbf{x}}_1) = (\mathbf{A}_1\mathbf{S}_{a,1}\mathbf{A}_2' \left(\mathbf{A}_2\mathbf{S}_{a,1}\mathbf{A}_2' + \mathbf{G}_2\mathbf{S}_{\epsilon,2}\mathbf{G}_2'\right)^{-1} \mathbf{A}_2 - \mathbf{A}_1)(\mathbf{x}_T - \mathbf{x}_{a,1})$$

where the only unknown is the true prior mean $\mathbf{x}_T$ (highlighted in red in the Equation above). Therefore, in validation exercises, the researcher can estimate the magnitude of the prior induced bias by choosing some reasonable choices of $\mathbf{x}_T$ and computing the resulting bias, which they can then compare to known sources of biases or validation errors.

We have added a paragraph to the manuscript summarizing the point above, which can be found on line 248, which we reproduce below

> In some validation exercises, it might be important to assess the relative magnitude of the prior-induced bias relative to known validation errors, and Equation (10) offers a means to do so. Note that Equation (10) only has one unknown, which is the true prior means $\mathbf{x}_T$. Here, the researcher can make reasonable guesses at some choices of $\mathbf{x}_T$, and then compute what would be there resulting prior-induced biases. For instance, the researcher may assume that $\mathbf{x}_C$ is misspecified by 1% or 2% (i.e., $\mathbf{x}_C = .99\,\mathbf{x}_T$ or $\mathbf{x}_C = .98\,\mathbf{x}_T$), which can then be used in Equation (10) to produce a numeric result that the researcher can compare to other sources of error.

We have also paraphrased the above paragraph in the Conclusion section as the last bullet point.

- **- a bias corrected posterior state - note that this is especially applicable to different satellites with different bias characteristics**

  We assume this comment is talking about cases where a post-hoc bias correction is applied to an OE retrieval. We are most familiar with the bias-correction for XCO2 as formulated in O'Dell et al. (2018), which adds a (varying) offset to retrieved XCO2 in order to correct the bias, so we will respond to this comment in context of OCO-2, although the same concerns will apply to other post-hoc corrections.

  At the present moment, there is a theoretical gap between the O'Dell bias correction formulation and the OCO-2 operational uncertainties. Recall from Rodgers (2000) that the OE uncertainties are computed by first finding the global minimum $\hat{\mathbf{x}}$ of the cost function:

  $$L(\mathbf{x}) = (\mathbf{y} - \mathbf{F}(\mathbf{x}))' \, \mathbf{S}_\epsilon^{-1} \, (\mathbf{y} - \mathbf{F}(\mathbf{x})) + (\mathbf{x} - \mathbf{x}_T)' \, \mathbf{S}_T^{-1} \, (\mathbf{x} - \mathbf{x}_T),$$

  where $\mathbf{F}(\mathbf{x})$ is the forward model. Then a first-order Taylor expansion is computed from the forward model centered on $\hat{\mathbf{x}}$, and then linear error analysis is applied (Rodgers, 2000, Chapter 5.5).

  The derivations of OE uncertainties and averaging kernels are clearly only applicable to the non-bias-corrected state $\hat{\mathbf{x}}$, or $\hat{x} = \mathbf{h}'\hat{\mathbf{x}}$ in XCO2-space, where $\mathbf{h}$ is the pressure weighting vector. Once we compute a bias-corrected value by adding an offset (i.e., computing $\hat{x}_{bc} = \hat{x} + \delta$), the OE uncertainties are no longer valid since they were computed based on a Taylor-series expansion around the estimate $\hat{\mathbf{x}}$. That is, the OCO-2 uncertainties for XCO2 are *no longer accurate* after the bias-correction procedure has been applied. Similarly, recall that the averaging kernel $\mathbf{A}_i$ has the following form:

  $$\mathbf{A}_i = (\mathbf{S}_{a,i}^{-1} + \mathbf{K}_i' \mathbf{S}_{\epsilon,i}^{-1} \mathbf{K}_i)^{-1} \mathbf{K}_i' \mathbf{S}_{\epsilon,i}^{-1} \mathbf{K}_i \tag{2}$$

  where $\mathbf{K}_i$ is the Jacobian evaluated at the non-bias-corrected state $\hat{\mathbf{x}}$. After the bias correction procedure, the associated averaging kernel is no longer correct due to the changing Jacobian.

  We are not aware of any effort to modify the operational OCO-2 posterior uncertainties and averaging kernel to account for the effect of the O'Dell bias correction procedure. If it were easy to characterize this bias-correction-induced error, then we would incorporate it into our paper in addressing this comment. Unfortunately, a preliminary examination of the bias-correction procedure indicates that it is a deep topic that requires treatment in a separate paper.

To reiterate, bias-correction procedures tend to add an offset to the retrieved value, without making a corresponding update to the Jacobian, posterior error, or averaging kernels. Due to this gap in the theoretical framework in the existing literature, we are not in a position to assess the impact of the bias-correction formulation on prior misspecification or averaging kernel correction.

5. **171: typo in specification of prior covariance**

   Good catch. The typo $\{\mathbf{x}_{a,i}, \mathbf{x}_{a,i}\}$ has been changed to $\{\mathbf{x}_{a,i}, \mathbf{S}_{a,i}\}$. Thank you!

6. **294: biases in the forward model and data also are mitigated by the prior assumption, so this suggestion of loosening the prior uncertainty will have other impacts.**

   Thank you for the comment. We note that the effects of the priors are applied twice in intercomparison of remote sensing retrievals: the first time in the retrieval process where $\hat{\mathbf{x}}_i$ is estimated from the radiance vector $\mathbf{y}_i$, and a second time when we apply averaging kernel correction on the retrieved state $\hat{\mathbf{x}}_i$. It is true that 'biases in the forward model and data also are mitigated by the prior assumption' in both steps, although we note that the theoretical development in this manuscript applies *entirely* to the second step, and that we are not 're-retrieving' $\hat{\mathbf{x}}_i$ from $\mathbf{y}_i$. That is, the retrieval team's choice of priors and its impact on 'biases in the forward model and data' in the first step are still preserved.

   The referee is correct in that 'loosening the prior uncertainty will have other impacts'. In a previous paper, we have demonstrated that loosening the prior will increase the 'validity' of an estimate (i.e., unbiasedness) at the expense of 'efficiency' (Nguyen et al., 2019). In the context of averaging kernel corrections, loosening the prior in the comparison ensemble implies that the variability of the difference, in general, will be increased. We have included a caveat on this following the referenced sentence in the updated draft at line 297: "In the case of averaging kernel corrections, letting $\mathbf{S}_C \to \boldsymbol{\infty}$ will ensure that $\mathrm{E}(\hat{\mathbf{x}}_{12} - \hat{\mathbf{x}}_1) \to \mathbf{0}$, but this choice may lead to larger variability in the difference (i.e., $\mathrm{var}(\hat{\mathbf{x}}_{12} - \hat{\mathbf{x}}_1)$ may be 'larger')."

References

Nguyen, H., Cressie, N., and Hobbs, J. (2019). Sensitivity of optimal estimation satellite retrievals to misspecification of the prior mean and covariance, with application to OCO-2 retrievals. *Remote Sensing*, 11(23):2770.

O'Dell, C., Eldering, A., Wennberg, P. O., Crisp, D., Gunson, M., Fisher, B., Frankenberg, C., Kiel, M., Lindqvist, H., Mandrake, L., et al. (2018).

Improved retrievals of carbon dioxide from Orbiting Carbon Observatory-2 with the version 8 ACOS algorithm. *Atmospheric Measurement Techniques*, 11(12):6539–6576.

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Press, Singapore.

Rodgers, C. D. and Connor, B. J. (2003). Intercomparison of remote sounding instruments. *Journal of Geophysical Research: Atmospheres*, 108(D3).