Reply to Review by Filipe Aires

First, thank you for your thorough comments. See itemized responses below.

* The experiments are conducted on a synthetic dataset built on purpose for this experiment. I understand that some techniques can be much better illustrated in synthetic cases, but it is frustrating to not see the impact of such a technique on real data considering that the authors has the possibility to easily make test on real data.

Our reason for using synthetic data was to conduct a controlled experiment that isolates the aspect of the problem that our method is intended to address without the complications that afflict real data, including non-linearity, non-Gaussianness, and complex geophysical noise. By creating a database that was in most respects well-behaved and easily visualizable in a low-dimensional space, we are able to demonstrate most clearly and unambiguously the potential failings of a traditional Bayesian retrieval method and to demonstrate a clear improvement following our dimensional reduction technique. The application to real data (in this case 9-channel brightness temperature data from the TRMM Microwave Imagers) is covered in two papers that have now been accepted for *J. Atmos. Ocean. Tech.*

* What is the content of the cited paper "Petty and Li" (in review)? Isn't it applying this technique to real data? Is there a true interest in this case in having a new paper on data that are so synthetic?

The purpose of the present paper using synthetic data is to describe the conceptual basis for our method and to illustrate its value in a way that isolates the problem our method is designed to address. In the *JTech* papers, we refer to the methods and findings in this paper to justify the specific implementation for TMI.

* It is clear that reducing the size of the input data before the retrieval can have a strong impact: it reduces the number of parameters in the retrieval scheme which regularizes the inverse problem, it can extract important features and suppress part of the noise. However, the paper doesn't do a good job in citing the other approaches and even compare to them. There is a lot of litterature on this subject: - PCA-regression (PCA before a regression), - PCA as a pre-processing of a NN network, - projected- PCA to perform a PCA to facilitate the retrieval of a particular quantity, - "intelligent" distances (such as the Mahalanobis distance) before the Bayesian search step in the retrieval - ... Since you are really focusing in this paper on the technical aspects of PCA, I believe that you really need to spend more time comparing with the existing approaches otherwise the reader cannot judge on the potential of your method. For example, work has been done using a PCA to compress and extract the cloudy signal on IASI data. I believe that the PCA was used to cloud-clear the data (dimension of data is 8461 in this case, the number of channels in IASI). I believe that this type of experiments is very close to what you are doing so a discussion, even comparison would be interesting.

PCA as a general technique is very well known. Our method utilizes a specific two-stage application of PCA, one which we believe to be novel in the way we explicitly separate a

potentially large geophysical noise source from a potentially small signal. The most important point to keep in mind is that this paper isn't intended to be *about* PCA but rather about the potential shortcomings of a traditional Bayesian database approach and about how (1) a reduction in dimensionality and (2) a decorrelation and normalization of background noise can sharply improve the performance of the Bayesian method. That we used applications of PCA to achieve (1) and (2) is in some ways incidental. We therefore see it as well beyond the scope of this narrowly focused paper to embark on a review of PCA in remote sensing.

* I find the synthetic database a disturbing factor. First, you should make a 3D figure with the 3D ellipsoid of the inputs space, and then represent the direction of the rain signature. Maybe a 3D plot of the raining and non-raining data.

We have added a 3D plot, as requested.

Second, the approach you describe here is based on the fact that you can extract linearly a raining signature from the input data. This is true in your synthetic case, and therefore the technique as to work.

It isn't necessary that the raining signature be linear, only that it be separable from the background noise via linear operators, which is a much less restrictive condition. We're basically saying that there's separation between the signal and the noise in 3D space, but that one needs to be extremely careful about how to preserve that separation when utilizing a "Bayesian" matching algorithm.

I have two questions:
(1) Is it true that the raining signature is just a linear signature in the input data in real cases?

No. But it will usually have a projection onto a linear operator that, if chosen carefully, will help isolate the raining signature from the spectrally distinct background noise. This is the essence of the Spencer et al. (1989) "polarization corrected temperature" (PCT), to give just one well-known example.

(2) Would your approach work if this is not true?

The striking validation results in Petty and Li (2013, Part 2, now accepted for *J. Atmos. Ocean. Tech.*) suggest that it does.

(3) Since your raining signature is linear in the input data, would a linear regression work as fine a your approach? A test on (3) would be very easy and is necessary here I believe.

Yes, multiple linear regression would clearly work very well in this case, and I don't believe a test is even necessary to demonstrate this. But the purpose of this paper isn't to identify an *alternative* to Bayesian retrievals in this idealized experiment but rather to point out the profound limitations of a Bayesian method when applied in a manner similar to that used by GPROF, for example. **If the Bayesian method can fail so badly when applied to idealized, quasi-Gaussian, linear data, it is will fail at least as badly in a non-linear, non-Gaussian case.** Again, our purpose (in keeping with the title) is specifically to improve the robustness of the Bayesian approach, the expectation being that the improvements will carry over at least partly to more realistic data sets.

* (1) The study on the minimal distance for the MC integration doesn't seem to include any uncertainty in the input parameters. When you perform these test, I think that adding a random error in the inputs of the data should bring a more realistic behavior. With this input noise, you optimal distance should behave slightly differently, especially when tested in the training dataset.

There is already noise in the data set that has a component in the same direction as the "rain" signature (this is why the separation isn't perfect in the new Figs. 5b and 6. Whether one regards that noise as instrument noise or geophysical noise is irrelevant for the experiment at hand. The signal to noise ratio is obviously adjustable, but changing it won't affect our qualitative conclusions.

(2) Using a small distance gives your a un-biased estimator but with high variance, using a higher distance provides you a biased estimator but with lower variance (this is the bias-variance dilemma). Such a discussion can be find for example in k-Nearest Neighboors discussion, varying the k like you vary the minimal distance.

The issue we address in this paper is not so much bias vs. variance but rather bias vs. a failure to find any valid matches whatsoever. I have never worked with K-Nearest Neighbors, but I know it's a clustering technique, not a retrieval technique. I'm reluctant to invest the time (and space) that would be needed to include a competent discussion of KNN in a paper that isn't focused on clustering or classification.

* I find the discussion about the stage 2 not detailed enough. I don't understand exactly what is been done here. I believe that you could illustrate and explain what this stage 2 does geometrically.

I have added a figure that illustrates geometrically the results of the two-stage process.

* page 2330, Eq. (1): I find the description of the "empirical" Bayesian scheme not good enough for readers not familiarized with this technique. Eq. (1) is the traditional Bayes formula. If we had information on the involved PDF, the expression could be used analytically, but this is not the case in a lot of cases. As a consequence, the integration of the Eq. (1) is performed empirically with the search and weighting procedure the authors describes, in a Monte-Carlo way. I believe you could improve this part and that this would be beneficial for new-comers in this field.

For the newcomers to the technique, we now recommend several papers by Evans, Kummerow, Marzano, and L'Ecuyer that describe the conceptual basis, implementation, and error analysis of Bayesian retrieval schemes for precipitation. I hope that this will be sufficient, as it could take considerable space to rehash the theoretical issues associated with the weighting of solutions.

* page 2330, line 20: supress "only"

Fixed .. thank you.

* page 2330, line 25: a way to accelerate this search is to order in some way the dataset so that the search is more efficient.

While this is true, it does not invalidate our statement.

* Page 2331, line 4: Could you explain why RTE simulations cannot preserve the physical correlation structure among the TBs?

We didn't say that they "cannot", only that it is very difficult to reproduce the microwave spectral dependence and distribution of real observations so accurately that the cloud of real data points in a high-dimensional channel space are likely to fully coincide with the corresponding cloud of simulated data points -- see the Panegrossi paper for more discussion of this issue.

* Page 2331, line 10-15: I don't understand the numbers you provide (10^4 and 10^12). From where they come from?

Imagine a 3D volume of, say, unit dimensions populated by a uniformly distributed set of 10^4 points. This implies an average separation of order 0.05 units between neighboring points. To achieve the same average Euclidean separation in a 9D volume with the same overal dimensions, one needs O(10^12) points.

* Page 2334, line 4: "as as".

Fixed... thank you.

* Page 2335, Eq. (4): please, descrive better the MC integration of Eq. (1).

Our analysis is constructed to be consistent with the way GPROF performs the "Bayesian" retrieval. Rather than re-justify the implementation details of GPROF, I'd rather let the reader refer to the relevant papers by Kummerow et al. for those details.

* Fig 1: It seems that there is no difference on distribution between rainy and non rainy data. This cannot be true otherwise the retrieval couldn't work. Could you try to find a representation (maybe 3D) that could illustrate the distribution of non-rainy points and the raining signature?

We have added the requested 3D figures.