Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-308-AC1, 2019

Geoscientific Model Development Discussions

Interactive comment on "Scalable Diagnostics and Data Compression for Global Atmospheric Chemistry using Ristretto Library (version 1.0)" *by* Meghana Velegar et al.

Meghana Velegar et al.

kutz@uw.edu

Received and published: 27 March 2019

Referee 1 comments:

The research article entitled "Scalable Diagnostics and Data Compression for Global Atmospheric Chemistry using Ristretto Library" for global environment monitoring deals with huge volume of data. It is necessary to develop an efficient method to reduce the data obtained from the sensor for reasonable analysis. Here are some of the clarifications/minor revision required from the authors.

Thank you for the comments about the paper and for the clarifications/minor revisions

suggested.

1. Authors have applied the already existing methods to the Global Atmospheric Chemistry Data. What is the novelty in this work compared to the existing methods?

This work is aimed at bringing emerging data methods, such as the proposed randomized algorithms posited, to the atmospheric chemistry community. Much like the application of neural networks (NNs), for instance, for forecasting and prediction, the goal is to use methods developed in the computational/statistical/computer science community to large-scale scientific problems. In emerging NNs, there is typically very little innovation on the NNs themselves, but rather on what the NNs can reveal for scientific discovery. Indeed, we have demonstrated that critical diagnostics, such as the production of EOF/PCA patterns of global chemical variability, can be extracted from global data sets on laptop/desktop architectures. Unless one had access to large scale computing and HPC coding, this simply is not tractable. Our innovation is in developing a suite of methods whereby practitioners from the atmospheric sciences community can now do practical computations from desktop/laptop architectures using simple open source python code. We know of no other existing computational methods that can yield such diagnostic features on such large scale data on laptop/desktop computing.

2. What is the reason for choosing Ristretto Library and it would be better to mention the advantages of using Ristretto Library package.

The Ristretto library was developed to provide a robust and stable code base for producing a variety of dimensionality reduction algorithms from large scale data that has traditionally been intractable. Not only can it produce PCA/EOF/POD/SVD modes, but it can also produce (as demonstrated) nonnegative matrix factorization and sparse PCA versions as well. There is no other package that integrates all these features in one package. Scikit-learn does have a randomized PCA routine, but not the other physically important variants demonstrated here. Moreover, the sparse PCA routine in Scikit-learn does not perform well on the data considered here as it was developed

with an eye towards other applications. As such, we instead recommend the integrated and robust package which is Ristretto.

3. The Comparative results with the existing techniques are necessary to prove the efficiency of the proposed method

We agree with this assessment. However, there currently are no techniques being used for such massive downsampling of data for porting to laptop/desktop architecture. Indeed, the current methods being used are the standard PCA/EOF/POD architecture which are computationally intractable unless used on HPC architectures. More succinctly: we have not found evidence in this field of application of other techniques (for comparison with randomized algorithms) being used for massively downsampling the data for producing dominant correlated features of the data. As for a comparison and evaluation of randomized methods, this has already been done and cited in the current work (See the citation to Erichson et al 2016).

As for distributed computing and HPC computing architectures, this is exactly what we are trying to avoid. Not only is one required to learn how to parse the computations to an HPC architecture, but few practitioners have access to such computing. Moreover, why would one do this when one can simply downsample through the randomized algorithm to achieve the desired results on laptop/desktop computing using standard python code. This is fundamentally a desktop/laptop enabled code by design. For data up to 10Tb, the data can be easily read into fast memory for producing the diagnostic features. Data that far exceeds 10Tb may require HPC architectures, but this is a significantly different world of computation than what is targeted here. Regardless, even for such enormous data sets, HPC computations can be greatly enhanced with the randomized algorithms.

4. Add some more related works developed in the recent years in the introduction section.

We have added a couple of references to very recent randomized tensor decomposi-

tions (see response to reviewer 2, point 3). There are several directions that people have taken randomized algorithms, but they feel very far afield and perhaps inappropriate for the current application in global chemistry. But the randomized tensor decomposition work is in line with the application advocated.

5. Any pictorial representation or block diagram need to clearly figure out your whole work.

We have modified Fig. 1 in order to demonstrate the algorithmic architecture on the real-world data. We hope this improvement will more clearly show how the algorithms is used.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-308, 2018.