Reply to Anonymous Referee #1

We are grateful to the referee for her/his constructive and thorough criticism and suggestions to our manuscript. Please, find below a detailed point-by-point reply (*referee's comment in italic*).

- This manuscript needs a lot of improvement in terms of the grammar and writing. There are many awkward phrases and incorrect word choices that need to be improved (a subset are listed below). The paragraph structures are also in need of modification (many paragraphs contain only 1 or 2 sentences).

We will improve grammar and writing of the manuscript by contacting a native English speaker/writer. Thank you for pointing out the subset of incorrect word choices. We will also modify the paragraph structures to avoid too small paragraphs.

-Section 2: I'd be helpful to include more detail for the preprocessing algorithms: shuffle and bitshuffle. Also this section in general needs improvement. It's a bit "choppy" to read (needs smoother and better transitions between topics) and feels like more details would be helpful on the methods (especially the ones that the digit rounding algorithm builds on).

More details will be added on the shuffle and bitshuffle algorithms. More details will also be added on the bit-grooming and decimal rounding algorithm, algorithm on which the digit rounding algorithm is built. We will do what is needed to improve this section in general.

-Section 4: Why does using the synthetic data in 4.1 to assess performance make sense - it seems unrelated to the application area of interest. I'd argue that the metrics used in 4.1 are really minimal requirements as well. Also take care when referring to "performance" as it is overloaded term...do you mean speed or effectiveness (it's used both ways)

The objective of using synthetic data was to control the data parameters, such as the SNR, to be able to assess the impact of these parameters on the compression ratios. The results are not reported in this paper which rather focuses on providing a comparison of the compression ratio and speed of different algorithms. It has also been chosen to present only the minimal set of relevant metrics to avoid overloading the paper. We will be more rigorous and replace the term "performance" by "compression ratio" or "compression speed" in the text.

*-fpzip is a fast and effective lossless method that would have been nice to compare (I *think* there is an fpzip filter available). Also I believe that any hdf5 filter can be accessed through NetCDF4 (see last sentence in conclusion) - consider contacting the Unidata folks.*

Thank you for the suggestion. Indeed a HDF5 filter is accessible for fpzip. However, many lossless compression algorithms exist. In our paper, we chose to evaluate the most "popular", i.e. the lossless compression algorithms the most used in applications. Thank you for pointing this evolution of the NetCDF-4 library: from version 4.6.0 - January 24, 2018, NetCDF fully supports HDF5 dynamic filters. The text of the paper will be modified so as to provide the example usage using the new NetCDF-4 features.

-Comments on doing compression in parallel?

We do not consider running compression algorithm in parallel in this work and will make it clear in the manuscript. It is a possible extension of this study.

-When reading the conclusion, it's hard to see what the main contributions of this paper are. It's fairly well known already that preprocessing of scientific data (e.g., bit shuffle or shuffle) improves lossy compression. Also the statements in the conclusion aren't specific to a particular type of data set, but are presented as more general conclusions.

Given that the effectiveness and performance (speed) of lossy and lossless compression are very data, application, and variable dependent, the general statements here are not well justified by the small sample of data in the paper. I'd suggest focusing the paper more heavily on the data in Section 5 (if it's of interest) and tailoring the discussion in that manner. Or maybe the focus was to be more on speeds than quality, in which case it's be important to work to get sz and fpzip working, particularly via netcdf-4...

Thank you for the suggestion that will help highlighting the main contributions of our work. As suggested, the paper will be reworked to focus more on the application to the CFOSAT and SWOT datasets. We will also avoid general statements but attach our conclusions to our application case.

————————————- Specific items: ————————————-

-p2, line 20: note that fpzip can also be lossless

Thank you for the remark. Fpzip will be presented both as a lossless and lossy compression algorithm.

-p. 8: discussion of figure 5: is the width of the bars related to the compression levels? (e.g. line 20 statement is unclear)

No. All the bars have the same width. Each vertical bar represents a compression level. For instance, the 9 compression levels of Deflate are represented by 9 vertical bars. This will be clarified in the text p.8.

-p.8, lines 28-29: Why is this the case? (Add some discussion beyond describing the figure.)

These lower compression/decompression speeds are not well understood and would require further investigation to be fully understood. It might be related to HDF5 chunking. Indeed, HDF5 split the data into chunks of small size that are independently compressed. This allows HDF5 to improve partial I/O for big datasets but can sometimes reduce the compression/decompression speeds. This discussion will be added to the text.

-p.8, line 20: I feel like the parameters should be better explained for –filter so that the reader can try them more easily. For example, what does the "32017" mean? I think that the following 0 is for sz, but this is not stated either.

We will add the meaning of each parameters. Each HDF5 filter is identified by a unique ID. "32017" is the identifier of Sz filter. The following "0" is the number of filter parameters. In the case of Sz, the filter does not have any parameter to set. That is why there are 0 parameters. Sz compressor is configured via the sz.config file. The same explanations will be added for the other filters used in the paper.

-p. 10, line 6: re: "experiments have shown" - whose or which experiments (cite?)

It is based on our own experiments that haven't been published. The sentence will be reworked as follows: "We have found that Shuffle or Bitshuffle preprocessing do not increase the compression ratio when applied after Sz. We have also found that and Bitshuffle provide lower compression ratio than Shuffle when applied after Bit Grooming. That is why only Shuffle is applied after Bit Grooming."

-Table 5 : Why is the speed faster for 1d?

As previously, the lower compression/decompression speeds obtained with the dataset s3D are not well understood and might be related to HDF5 chunking. This discussion will be added to the text.

-Section 4.4.1, line 21-24: Any idea why you get these results?

Sz performs better on smooth signals since it makes use of a prediction step. The signal s1 being highly noisy, Sz prediction might often fail. This can explain the lower compression ratio on the signal s1. On the contrary, Bit-grooming does not makes any prediction. This can explain why it achieves better compression than Sz on the signal s1. This hypothesis will be added to the text.

-Section 4.4.1, last sentences: It's unclear to me what the value of these synthetic data sets is - especially given the statement on p. 11, line 3, about the dependence on the dataset

As suggested previously, the paper will be reworked to focus more on the application to CFOSAT and SWOT datasets without drawing general conclusions based on the results obtained on the synthetic datasets.

-page 11, line 9-10: I'd include characteristics of the data (e.g., maximum abs. value) earlier in the text when the two datasets are introduced.

Your suggestion will be taken into account: the characteristics of the data will be introduced in section 4.2.

-page 11, line 24: I don't see relative error mentioned in Table 6 - it seems to just be absolute error

The text will be modified to make it clearer: "…all three algorithms respect the maximum absolute error of 0.5 which, for the signal s1, corresponds to a relative error of 0.00424."

-p.11-12: Need more of a discussion of the results in Figure 7. For 3D, it looks like bit grooming and digit rounding are similar - I don't see a clear advantage.

More discussion on the results will be added to the text. For the s3D you are right, there is no clear advantage. It is written "the Digit Rounding algorithm provides compression performance very closed to the one of the Bit Grooming algorithm".

-p.12, lines 16-17: SZ compression can be controlled with an absolute error bound, so why is the relative error bound adjusted to get the desired abs. error?

The objective was to see if Sz compression configured with a relative error bound respect the error bound specified. As the digit rounding and bit-grooming algorithm can only be configured on a number of significant digits, they can only "produce" absolute error in 0.5, 0.05, 0.005, etc. In order to be able to compare Sz configured with a relative error bound with those algorithms, we have configured the relative error bound to obtain a maximum absolute error of 0.5. These explanations will be added to the text.

-Section 5.1: It is disappointing not to have SZ results on the real data of interest. Were the SZ authors contacted? I would think that they could have helped resolve this issue.

Yes, we had some exchanges. The issue is still under investigation.

-p. 15, line 18: "which only a few attributes may be missing" - It's unclear what this means. It's super helpful to really detail the data being compressed so that one can make sense of the results.

Details on the datasets will be added to the text.

-p. 14, line 30: Please share more specific information about the precision required by the scientists for the data. Again, more information is useful for interpreting results.

The configuration and the precision of each variable will be made available.

-Section 5 seems like it should be the highlight of the paper as here we are seeing the results on the real data. But it feels like more detail is needed on the data and more discussion of the implications of the results.

Section 5 will be developed to add more details on the data and more discussion on the results obtained.

——————————— Typos, etc.: ———————————

-abstract, line 7: incorrect use of "imposes"

-p.1, line 22: "quite spread"=> "quite prevalent" or "quite popular" , "widely spread" => "widely used"

-p.1., line 26: "reduce significantly" => "significantly reduce"

-p.1. line 27: This sentence (that continues to page 2) is too long.

-p.2, line 3: "can afford for" is awkward

-p2, lines 10-24: this region is 5 paragraphs

-p.3, lines 3-4: awkwardly worded

-p. 3, line 9: one sentence paragraph

-p.3, line 12: missing "," after "Deflate"

-p.3 line 14: not sure what is meant by "new concurrent"

-p.3, line 13: "widely spread" => "widely used"

-p.3, line 16: awkward sentence: "This allows Deflate achieving rather high compression ratios"

-P.3, section 3: again, there are too many tiny paragraphs

-p.4, line 7: "are of same interest" is awkward

-p.4, line 21: Table number is not given

-p.5, line 15: awkwardly worded

-p.5, line 24: One sentence paragraph

-p.7, section 4.2: define f_s, f_ech

-p.7, line 22: "use embarks" is awkward

-p.8, line 7: "declined" doesn't make sense

-p.8, line 14: "embark" - incorrect usage

-p.10, line 10: "This correspond corresponding" needs to be fixed

-p. 10, line 24: Note sure I'd use "performances" here as earlier it was used to indicate speed.

-p. 15, line 21: "ration" => "ratio"

-p.15, line 14: another one sentence paragraph

-p.16, line 9: "Extends to this work" - awkwardly worded

Response: we thank you for highlighting typos that will help us to improve the manuscript.