Atmos. Chem. Phys. Discuss., 11, C6813–C6818, 2011
www.atmos-chem-phys-discuss.net/11/C6813/2011/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series" *by* J. P. Musial et al.

Anonymous Referee #2

Received and published: 23 July 2011

General Comments
Tackling the issue of comparing the type of signal processing algorithms the authors describe here is a challenging topic. Particularly when an experiment to do so is designed around checking outcomes against known synthetically generated databases. The authors are to be commended for taking steps to do so.

The authors' objective, "to evaluate the performance of generally applicable methods that could be applied automatically to a large number of time series, without any human intervention," is certainly the Holy Grail of a large community of researchers. This paper is one more step in the right direction.

Specific Comments
The abstract is succinct and to the point. Not too long, but relatively comprehensive, if not totally inclusive. The Introduction sets the tone to capture the interest of the general reader, perhaps the non-specialist. There are useful comments on discretizing time series. The Introduction highlights the inadequacy of typical databases one routinely has to deal with. The authors mention the need for interpolating time series to new time references. On page 3 when the authors draw on the analogy with acoustics, I might favor adding radar, and have the sentence read "By analogy with such fields as acoustics and radar, the ...".

The motivation for the paper is well defined at the end of the Introduction. The point is to review several of the algorithms used in the literature that "might be of interest to a wider scientific audience."

Section 2.2 might best review in a few words what is meant by the term "Lagrange form," for the benefit of the general reader. Later in this section, the authors send the reader scurrying to the web (or, heaven for bid, the library) to look up Legendre and Tchebicheff polynomials, Empirical Orthogonal Functions (EOFs), Principal Component (or Factor) Analysis, Singular Spectrum Analysis and, of course, the Lomb-Scargle method.

Thus rather than a "review", I get the sense the authors intend this to be an inventory of background reading for the general audience.

Section 3.3.1. The paragraph might be amended to include the term ". when analyzing satellite data". For example:

– Uniformly distributed gaps. For each predefined percentage of missing data, a

random number generator U was used to iteratively select the location of the next data point to be removed from the series: xm =U[0;1] n. This situation might arise when the system of interest is occasionally unobservable, for instance due to the presence of clouds, when analyzing satellite data.

This is this is because the authors consider a variety of other types of data in their discussion.

Section 3.4 lists an interesting array of classes of time series to which the authors' methods might be applied. Again, I would have emphasize that as the authors segue into their error analysis in Section 4, it would be useful to see plots of actual examples of the synthetic time series. As for the synthetically generated functions, I find it somewhat difficult to visualize combinations of trigonometric functions having characteristic periods of 2π, when I am not also shown the nature of the data gaps, and the level of the added noise relative to the amplitude of the pure "signal".

Section 4 applies several processing methods to synthetically generated time series. I would find it quite useful to see time series plots of these types of data, to illustrate the general differences between the authors' random gap, winter gap and continuous gap data; as well as to illustrate extreme members of each of these classes that were used in the assessment. This becomes clearer when the reader drills down to the end of the paper and Figure 16, but I feel the potential impact of this earlier material is diminished without "seeing" the time series; with and without noise, with and without data gaps. When discussing the mean absolute error (MAE) in Figures 2 through 7, does the authors' assertion, "Each plot consists of 36 data points," imply that the contour plots are based on a set of 6x6 gridded and then interpolated values? This might be clarified.

In my view, the authors' presentation of the synthetically generated data is of limited use unless we can see the actual time series that were used. The latter

comment is not intended to detract from the authors' later discussion using actual time series, where they show quite clearly the actual time series and the results from the various algorithms.

I find it very useful that Figure 9 is reproduced at a respectable size, so that it is easy to read. Since the actual data points in Figure 9 are so critical to the point the authors are making, I feel it is appropriate to emphasize them more than shown; perhaps as solid black symbols, but then I suppose you would lose something in not seeing the overlap of the symbols as presently shown. Perhaps encouraging the editors to ensure the quality of the images to a level where the readers can enlarge each plot at their leisure would best serve the purpose.

Numerical Experiments on Actual Data

This material will have considerable reader interest, and in my view should be the centerpiece of the contribution. It needs to progress very carefully however. In my first quick read through, I am not certain as to how the data were decimated for the test: randomly, regularly or ...? Perhaps this was specifically explained and I just missed it.

I am not certain from inspecting the numerical experiments with actual time series described in Section 4, whether the respective time series themselves shown in Figures 9, 10 and 12 have been decimated with gaps prior to them being plotted in the figures. It would seem so from the figure to the right of the Mauna Loa CO_2 time series for 10% decimation. It looks like 1 in 10 points might have been dropped.

Frankly the use of dashed lines tends to suppress some of the information the authors are attempting to convey.

Especially in the cases of the FAPAR and Sun spot data, it would be interesting

to see how well each algorithm fits the original undecimated data by comparison with fitting the decimated cases.

For the quick reader, it would be a good idea to include the added noise level in the caption (as well as at the top of the graph, as is done) of Figure 16.

Interactive comment on Atmos. Chem. Phys. Discuss., 11, 14259, 2011.

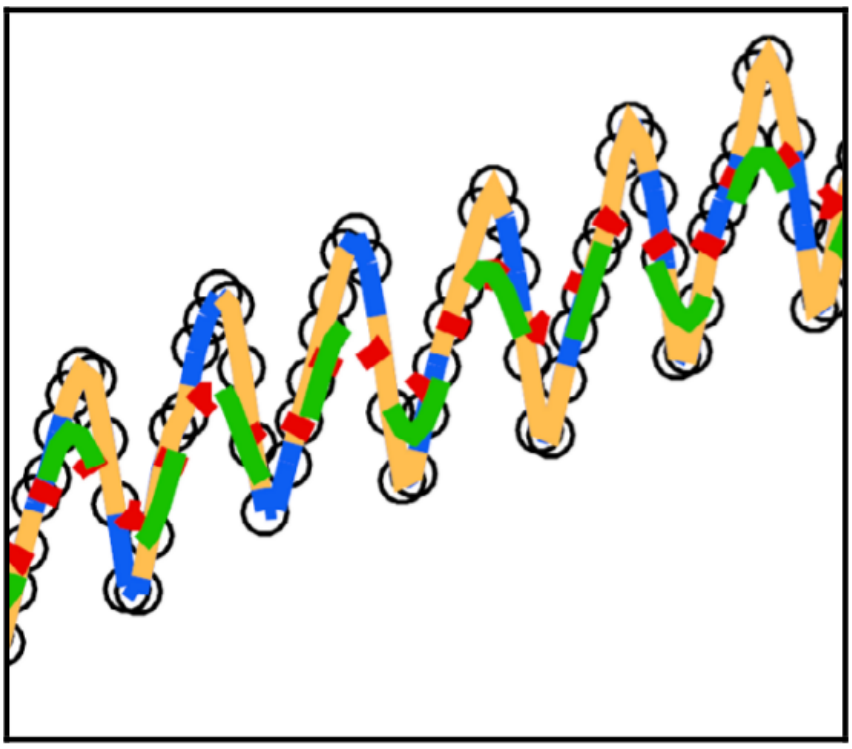

Fig. 1.