Nonlinear Processes
in Geophysics

Interactive comment on "Efficient Bayesian inference for ARFIMA processes" *by* T. Graves et al.

Anonymous Referee #1

Received and published: 13 July 2015

A. Summary

The authors present a novel approach that involving Bayesian inference to the earlier existing problem of inferring the nature of long range dependence (LRD) from time series data. The exposition of their method in the manuscript is mathematically thorough and consistent. The authors consider several relevant theoretical examples to illustrate the features of their approach as well as highlight the relatively better performance of their method as compared to existing techniques that seek to address this problem. The method was finally applied to the Nile river minima level time series data and physically reasonable results were obtained. Furthermore, the authors shortly mention the possible extensions of the method and its potential future applications.

B. Recommendation

In my opinion, the method presented in this paper is a novel way of applying Bayesian reasoning thoroughly to the chosen problem of inferring LRD in time series data. The authors have done a thorough job of grounding their approach on a rigorous mathematical foundation and presenting the theory unambiguously and carefully.

That being said, the chosen geo-scientific example in the manuscript, viz. the Nile river minima time series was given only a short, cursory treatment and no effort was made to infer the results obtained by the method in this case. Given the focus of NPG and its target audience of geoscientists, I feel that the authors could move some of the mathematically intensive sections of their method to an appendix and thereafter devote more space and attention to the application of the method to the Nile data and how the method relates to physically understandable features of the Nile river system.

I thus recommend publication in NPG only after a few major shifts in presentation and focus of the manuscript, which I hope the authors can address with a major revision. Besides these points, there are certain technical issues with the presentation that I list out below. I hope that these suggestions help the authors to better their manuscript.

C. General Comments

1. A first major point that could be improved is the focus given to the application of the method developed by the authors to real data. In the current version, the manuscript gives much space to the theory and constructed, illustrative examples which, even though necessary and absolutely crucial, should ideally in the end lead to an equally detailed example from the real-world with a discussion on how the method performs in that case. I feel that, given the particular focus of NPG, which is at the juncture of theory and application, it is important to properly discuss the relevant real-world implications of the results obtained when applying the method to the Nile river minima for instance. This is lacking in the current version. One such difficulty for me in understanding the final application was that I was unable to relate the parameters p, d, and q to the physical system that gave rise to the Nile data set.

2. Another main issue for me was that there were discrepancies between the title, abstract, and the numerical examples in the text in terms of the main focus. The title indicated that the manuscript is about "Bayesian inference for ARFIMA processes", the abstract suggested that the main focus was LRD inference, and finally the text devoted quite some space to "short memory" as well. The authors should take care to clearly state what are their objectives (which they seem to do even now) and thereafter, relate every new topic/application/idea later in the text to the stated objectives (which is lacking).

3. A last major issue that I have with the manuscript is that the Tables and Figures do not have adequate captioning. The table captions do not indicate what are the various parameters represented and for what kind of numerical experiments. The figures do not have a clear "visual" legend and neither do the captions indicate what kind of numerical analysis gave rise to the figure. I feel that this is an extremely crucial part in communicating results.

D. Specific / Technical comments

TITLE: Perhaps the title is a too vague. Maybe the authors can consider giving a more informative title?

P574, L7: "Rather than Mandelbrot's fractional Gaussian noise ..." There is no clear comparison of the method from this study to Mandelbrot's fractional noise. What is the purpose of mentioning it here?

P574, L15: "We illustrate our new methodology on the Nile water level data..." What are the main results? How do they enhance our understanding of the Nile river data and LRD?

P574, L26: "A standard definition..." even so, please cite a reference for this definition.

P576, L1: "It would offer the ability to marginalize out aspects of a model apparatus and data, such as short memory..." Perhaps this idea is key to the extensions of short

memory later in the text? If so, the authors should address this is more detail here and later as well.

P576, L5: ARFIMA has not been defined before this in the main text (only in the abstract).

P576, L15: Why is it necessary to include/Cite statements about something being "too hard to work with"?

P576, L18: "many of the above drawbacks..." It is not clear to me how many drawbacks were mentioned before this line.

P577, L18: "2 Time series definitions and the ARFIMA model" I personally feel that barring the definition of ARFIMA processes, the rest can moved to an appendix.

P577, L22: You defined covariance function \gamma (k) as "Cov(X_t, X_{t + k})" but what is "Cov". This definition is ambiguous and unclear.

P577, L25: "the "backshift" operator" - Here and later, the use of quotes for terms and definitions are a bit distracting. Maybe the authors can find some other way of emphasis?

P579, L11: I understand that this is the LRD parameter "d" being referred to here, and which was defined in the Introduction. But I find it hard to relate the $|d| < 1/2$ statement to the preceding equation, i.e., Eq. 5.

P580, L15: "Choosing $p = q = 0$ recovers FI(d) $\equiv$ ARFIMA(0, d , 0)". Is this the definition of ARFIMA? If so, maybe it is possible to add a line stating this more clearly?

P582, L1: "3 Likelihood evaluation for Bayesian inference". I feel it might be better if the authors added a few words here about the interpretation of the parameters p, d, and q of the ARFIMA process here (even at the cost of repetition) before starting with the model inference part. Also, why is the starting point of the inference is a ARFIMA(0, d, 0) process and not some other value of p and q? What kind of a process is this?

P582, L9: "causal" in what sense? Maybe briefly mention here.

P582, L24: It might be ambiguous to infer AR(P) dependence from a time series of length n = P.

P588, L5: "with chains moving between and within models..." It is unclear what this means for someone unfamiliar with numerical techniques of Bayesian likelihood estimation and MCMC methods.

P588, L7: What is "RJ"?

P588, L9: What is "FEXP"?

P589, L20: Does "MVN" mean Multivariate Normal"?

P591, L11: What is a transdimensional move?

P593, L3: The authors should maybe add one more line on why they choose the Poisson distribution here, it is not clear from the parsimony argument they mention.

P595, L22: Is it not possible to show the results in a graphical way, such as a histogram instead of a table?

P596, L9: Figure 1 legend is unclear. If I understand correctly, for each value of d_l, there are 1024 estimates and thus 1024 "x" markers on the vertical axis - is this correct? I suggest to use better captions to remove such ambiguities.

P600, L8: For the sake of reproducibility, I suggest that the authors state the source of the data and from where it was obtained. Also, maybe they should devote a few lines on the nature of the data, and preprocessing of the data such as removal of missing values, outliers and the like.

P601, L11: Reference missing to the dissertation cited here.