Smoothing and Interpolating Noisy GPS Data with Smoothing Splines—Reviewer Response

Jeffrey J. Early & Adam M. Sykulski

October 7th, 2019

1 Comments from the editor

I am now in receipt of all reviews of your manuscript "Smoothing and Interpolating Noisy GPS Data with Smoothing Splines" and an editorial decision of Major Revision has been reached. The reviews are included below.

The reviewers have rather mixed impressions about your manuscript. They all agree that the manuscript is lengthy, and I encourage you to try to shorten it if possible. Also, please pay particular attention at how your method compares with other similar techniques. I invite you to submit a revised paper by Oct 14, 2019. If you anticipate problems meeting this deadline, please contact me as soon as possible at Volkov.JTECH@ametsoc.org to discuss an extended due date.

Thank you for inviting us to submit a revised manuscript. Please find our responses to each reviewer comment given below.

2 Reviewer 2

This is a very well written, rigorous manuscript on the details of mathematical/statistical smoothing (i.e. removal) of some very large GPS errors. Having designed and built over 500 drifting GPS oceanographic wave buoys of the years, my first impression is that they just need to engineer a better GPS solution. We never see such large outlier errors, nor such large variance on our data (whether you model it as Gaussian or t). Ensuring a good antenna, above sea level to avoid splashing water, and a good sky view so as to see as many satellites as possible (8+), the GPS errors they experience would not be present. Hence the need for such a complicated and computationally costly algorithm would not be needed. The use of GPS position and (especially) altitude should not be used. We use GPS chips that produce 3-axis Doppler velocities, accurate to 0.05 m/s, producing outliers one in a million samples. Outlier detection is simple by removing points beyond some multiple of the standard deviation, and performing a simple linear interpolation to fill it back in. But then again, my Master's Degree is in engineering, and my Bachelor's degree is in Mathematics, so I favor a more robust engineering design. That being said, this is certainly a publishable manuscript and I approve of it, I realize they may not have access to a better engineered system. Personally, I have no use for it, as my systems produce no such errors.

We absolutely agree that an engineering solution to this problem would be a much more satisfactory solution. As noted in the manuscript, one of the nine drifters showed no indications of outliers, suggesting that antenna configuration may be the culprit. That said, this issue appears to be fairly common in different GPS surface drifter experiments—it can be seen in drifters deployed as part of GLAD (Yaremchuk & Coelho, 2015) as well as the 2018 NISKINE cruise (unpublished). As neither of us are familiar with the engineering details, we are not sure why this issue hasn't been solved in these other contexts.

More broadly speaking, however, we believe our methodology applies in any setting with significant observational noise and outliers. For example, many ocean drifters from the Global Drifter Program use the ARGOS positioning system, which is even more prone to errors and outliers (Elipot et al, 2016). Outside of oceanography, GPS instruments are used in animal tracking for example, and here such data is also prone to significant errors and outliers (see for example, DeCesare et al, 2005) to which our methodology can be applied.

It's quite wordy (36 pages of text!), so I would challenge the authors to cut out about half of it, at their discretion, while keeping their message in tact. It's so long I really lost interest about half way through, and certainly didn't need to read the appendices. If JTECH wants to publish such as lengthy treatise, I'm all for it, as it's well done, but I think it's too long.

We have shortened sentences throughout the manuscript and the original text now falls within the 7500 word limit for JTECH at 7416. However, the new discussion section comparing to other methods brings the total to 7674.

My only specific edit would be to define the variables in equation (9): A, ω, λ. One can assume A is for amplitude, ω for frequency, but this is the first mention of λ (not λT), it could be interpreted as wavelength.

Thank you for spotting this. We rewrote this paragraph, including adding definitions for the parameters.

3 Reviewer 3

The submitted manuscript presents methods for smoothing and interpolating noisy, irregularly sampled data using smoothing splines with variable tension.

In particular, the authors demonstrate how to obtain optimum choices for adjustable parameters of the spline fit, notably the spline degree S and tension T, and how to deal with the non-Gaussian noise intrinsic to GPS position fixes obtained by surface ocean drifters. The work provides an extensive and careful description of the proposed methods, with numerical implementations also being made available by the authors. With the increasing number of deployments of GPS-tracked drifters in recent years, I believe that this study is of interest for the oceanography community and will be a valuable contribution to JTECH pending a Major Revision.

The manuscript rushes through concepts that I believe are key for the interpretation of the results, and many statements along the text need rephrasing or clarification. I also believe that it lacks a discussion about potential advantages/drawbacks of the presented methods relative to other procedures available in literature, which would be useful for informing readers about which methods are better suited for their needs. Finally, it is noted that the manuscript in its present form has 34 double-spaced pages, exceeding the 26-page limit stablished by the journal and thus requiring preliminary approval from the Chief Editor for publication with the page overage (personally, I have no objections to the paper length).

Please find below my major and minor comments.

Thank you for taking the time to make such detailed comments—it's very much appreciated.

3.1 Major comments

- **The manuscript exceeds the 26-page limit for JTECH submissions:** Excluding references, tables, figures, and the title and abstract pages, the manuscript currently has 34 double-spaced pages, thus exceeding the approximate 26-page limit stablished for JTECH submissions. According to the journal's formatting instructions, the authors should prepare a justification for the length of the manuscript on the submission's cover letter, and request the Chief Editor's approval for exceeding the page limit.

  We have shortened sentences throughout the manuscript and the original text now falls within the 7500 word limit for JTECH at 7416. After adding the discussion section comparing to other methods, the total length is now 7674 words. The paper curiously exceeds the approximate 26-page limit at 34 pages. We suspect that the extra white-space around the equations in draft format is responsible for the mismatch between the official word limit and the approximate page limit.

- **Discussion of advantages/caveats of proposed approach relative to other methods:** The manuscript currently lacks a discussion on the potential advantages and drawbacks of using smoothing splines following the proposed procedure relative to other processing methods described in literature, such as those applied to Global Drifter Program (GDP) drifters (Hansen and Poulain, 1996), to recent massive drifter deployments in the Gulf of Mexico (Yaremchuk and Coelho, 2015), and even for the generation of an GDP drifter dataset set at an hourly resolution in a previous paper by the authors (Elipot et al. 2016). I think that including such discussion could help informing readers of which methods are more suitable for their needs, and thus add value to this work. For example, it seems to me that the GPS position jumps described by the authors should probably be more severe in regions with intense wind and wave activity such as the Southern Ocean. Does that mean that their methods would outshine others for data collected in such regions? Please elaborate.

I also believe that the drifter dataset described in the study has been under-explored. The manuscript mentions that it contains position observations from 9 drifters, and yet results for only 2 of them (drifters 6 and 7) are presented. If within the reach of the authors, I suggest using the remaining observations to evaluate the proposed methods against other procedures described in the literature.

- Elipot, S., R. Lumpkin, R. C. Perez, J. M. Lilly, J. J. Early, and A. M. Sykulski (2016), A global surface drifter data set at hourly resolution, *J. Geophys. Res. Oceans*, 121
- Hansen, D., and Poulain, P.M. (1996), Quality control and interpolations of WOCE-TOGA drifter data. *J. Atmos. Ocean. Technol.*, 900–909.
- Yaremchuk, M., and E. F. Coelho (2015), Filtering drifter trajectories sampled at submesoscale resolution, *IEEE J. Oceanic Eng.*, 40(3), 497–505.

This is a very helpful comment and suggestion. We have added a *Discussion* section before the conclusions which we use to compare to the other methods used in the above three papers.

The Yaremchuk and Coelho (2015) paper is most similar the work done here. The methodology in that paper could be described as a subset of the approach taken here, where, rather than minimize the expected mean square error, they choose a smoothing parameter based on the expected values for acceleration and noise. This is similar to our approach of estimating an initial smoothing parameter, but because they do not include the effective sample size scaling factor, their methodology will generally over-tension (our tests confirm this). They handle outliers by using what is effectively an iterated least-square approach, using a weighting that is very similar to the Tukey biweight. So essentially, their penalty function is nearly identical to ours, but the choice of smoothing parameter is suboptimal for all the reasons described in the manuscript.

Both the Hansen and Poulian (1996) and Elipot et al. (2016) papers examine ARGOS tracked drifters. The methodology for handling outliers is the same in these approaches—a hard cutoff of 2.5 m/s (Figure 14 in Elipot et al.), which essentially discards data that appear as outliers. In our approach we never discard data, only unweight its importance. A cutoff for velocity would certainly work for removing outliers in GPS data as well, but our approach avoids this kind of 'magic number'.

The approach taken in Elipot et al. requires parameter choices for weighting nearby observations that aren't generalizable from the ARGOS data to the GPS data.

The most interesting connection between these different approaches is that kriging used by Hansen and Poulian (1996) is much more closely related to splines than one might first imagine. Handcock et al. (1994) point out that smoothing splines and kriging are really just two specific parameter choices from a more general formulation of splines. In fact, the generalization is to assume the physical process model follows a Matern-like covariance structure—exactly the model we used to generate synthetic trajectories.

- **Practical examples of the numerical implementation of the described methods:** I downloaded the Matlab classes developed by the authors at https://github.com/JeffreyEarly/GLNumericalModelingKit. I think that the readers of this study would appreciate if practical examples for processing the set of drifter observations described in the study were also made available.

  Yes, thank you for the nudge. Beyond the existing Readme.md, we will continue to add more documentation as the paper gets closer to publication.

3.2 Minor comments

- It is often unclear which Sections are referenced along the text (e.g. lines 146, 286, 405, 410, 416, 486, 493, 514, 515, 583, etc), specifically because the Subsection letter is cited unaccompanied from the Section number. I think that this may be caused by a bug on the AMS Latex package, considering that this issue is absent in the arXiv version of the paper. Please proof-read the manuscript, correcting this issue where needed;

  Fixed.

- A similar problem is observed on how the Appendices are referenced (e.g. lines 248, 251, 335, etc), which are shown with a number instead of a capital letter. Please correct all occurrences of this issue.

  Fixed.

- Caption of Figure 1: Please expand the first period as "An example of interpolating between 7 data points using spline functions of order K";

  Good suggestion, fixed.

- Line 101: For clarity, please modify this sentence as "All higher order ($K > 1$) B-splines are defined by recursion";

  Changed.

- Equation (9): Please explicitly define all the equation's variables (ω, λ, p and A) in the text immediately following the equation.

  Added.

- Lines 140-141: It is mentioned that the Matérn spectrum is used to the generate the velocity of the signal. Please provide further details on how this is done.

  We point to the reference Lilly et al. (2017), which explicitly spells out the procedure for generating the Matern process using the Cholesky decomposition. We also now point directly to `jLab` for generating such trajectories.

  - In particular, Equation (9) for me suggests that the spectrum varies solely as a function of frequency ω, using prescribed magnitude A and damping scale λ. Is that so, or is A instead obtained stochastically? Also, is the signal speed obtained simply by taking the inverse Fourier transform of (9)? Please elucidate.

  We rewrote the paragraph and, hopefully, clarified this a bit better.

- Line 140: It is also mentioned that the Matérn spectrum is used to obtain the velocity of the signal. However, considering that you are representing the evolution of a particle along a single positive spatial dimension, wouldn't the term speed thus be more suitable, in this case?

  The signal generated is bivariate (and thus a velocity)—we clarified this in the text.

- Line 141: It is mentioned that the velocities are integrated to obtain the positions. Please state at which time increment the positions are computed;

  Added this in the rewritten paragraph.

- Lines 148-150: It is unclear for me how the quality of the fit is assessed. Is the interpolated versions of the subsampled data evaluated against the original (not-subsampled) signal by computing the RMS error between them? If so, please rephrase these lines to make this message go through more clearly.

  Yes. Fixed.

- Figure (3):

  - Top panel y-axis label: I think that stating the speed power spectral density units as (m^2/s) is a bit counter-intuitive. I suggest changing it to [(m/s)^2/cpm];

    We chose m^2/s because the broader context in which the Latmix experiment is analysed is to study the assessment of lateral diffusivity, which has units of m^2/s. Because of the connection between the velocity spectrum and diffusivity, the 'Lagrangian statistics' literature uses these units.

  - Bottom panel legend: The RMSE units is set to meters, in contrast with the top panel, that shows the power spectral density for speed. Please double-check what time-series is used in the coherency analysis. I suggest using the velocity time-series, for consistency with the top panel;

    Correct, the velocity time series is used for both panels. We changed the word signal to velocity and added the following sentence: *In analyzing the quality of fits, we always use velocities when computing the power spectrum, but report mean square errors from the positions.*

  - Caption: "…shows the velocity spectrum of the signal (black)." To make the caption self-explanatory, please explain that the signal correspond to the synthetic Lagrangian velocities built using the Matérn spectrum.

    Good suggestion; fixed.

- Lines158-165: I don't think that many of your readers will be immediately familiar with magnitude-square coherency estimates. Please briefly explain how to interpret this statistic, after introducing it in the text.

  Added: *A coherence of 1 indicates that the signals are perfectly matched at a given frequency, while a coherence of 0 indicates that the signals are unrelated.*

- Lines 161-163: "This is why the shallower slopes (with more variance at high, incoherent frequencies) have a larger mean square error than the steeper slopes (with less variance at high, incoherent frequencies)". I was confused by this sentence, as it initially lead me to believe that the errors were computed using the interpolated signals preliminarily high-pass filtered to isolate frequencies higher than the Nyquist frequency of the strided data (!!). Please reformulate it, for clarity.

  Thank you for seeking clarification here. The signal has not been high-pass filtered to calculate errors. The error we measure is an overall measure with respect to the original signal, and is thus calculated at all frequencies up to the Nyquist of the raw data. We have amended the text to be clear that we are referring to the *overall* mean square error with regard to the true signal.

- Line 169: The i.i.d acronym is used, which refers to an independent and identically distributed random variable. While this acronym might be common place for the statistics community, I think that a large fraction of the JTECH audience will not be familiar with it. Please reformulate the sentence on this line avoiding the use of this acronym.

  Very true. We removed the acronym and stated the assumptions more explicitly.

- Lines 175-176: "– up to a constant this is the log likelihood". Unclear (I think there is a missing comma after "constant"). Please reformulate, for clarity.

  Made the 'up to a constant' parenthetical.

- Lines 184-185: The tone of this sentence seems excessively informal, and its construction can obscure the message the authors want to convey. Please reformulate, for clarity.

  Reworded: *Thus, it is necessary to add additional constraints the problem if the assumed error distribution is to be recovered.*

- Equations (19) and (20): the variable "I" is undefined in text;

  Fixed.

- Caption of Figure 4: Please state what the vertical dashed lines correspond to;

  Whoops, thanks.

- Line 315: If the vertical dashed lines in Fig. (4) correspond to f^{eff}, than stating that this parameter "indicates almost exactly where the coherence drops below 0.5" is a stretch, as it visually intercepts values closer to 0.6 for the yellow and orange lines and to 0.7 for the blue. Please reformulate this sentence;

  Changed the word 'exact' to 'approximate'

- Line 326: Table 2 does not include a "Full dof" column. Please double-check both the text and the information provided in Table 2;

  Fixed and reworded.

- Line 326: The acronym "dof" is undefined. I assume it refers to "degrees of freedom", but I think this should be explicitly stated along the text;

  Fixed.

- Lines 355-356: Please add "spectral" to the sentence "... for three different spectral slopes ...";

  Fixed.

- Line 378 and Table 2: It is unclear what is meant by "blind" and "unblind" methods. Please define this along the text;

  Added a parenthetical note: *The second and third columns show the effective sample size and average mean square error when the smoothing spline is applied using the true values (i.e., 'unblinded') to minimize the mean square error—this is the lower bound.*

- Lines 437-441: The text here described the drifter GDP dataset: I believe it should further include information on (a) the sampling rates of the GPS receivers, (b) the length of the used records, and (c) the temporal increment the data was interpolated to.

  Our introductory paragraph is perhaps misleading—we're not using GDP drifters here, and so haven't studied the current sample rates, etc.

- Lines 471-473: Here, it is mentioned that two t-distributions are shown in the bottom panel of Fig. 6. However, this contrasts with Fig. 6 caption, which mentions that the gray line correspond to a Gaussian distribution, and the black to a t-distribution. Please double-check both the caption and the text, and remove the ambiguities.

  The plot shows the best fit assuming Gaussian-distributed noise (gray) and t-distributed noise (black), as indicated in the caption. We have adjusted the text in the main body to make this clearer, thank you for this spot.

- Lines 474-478: From the information provided by this paragraph alone, it is unclear for me why one can conclude that it is safe to assume that the 30-min sampling interval leads to statistically independent observations. Please rephrase this sentence for clarity, perhaps further mentioning that the autocorrelation function becomes statistically undistinguishable from zero within 30 min;

  Fixed.

- Line 498: I suggest substituting the sentence "vastly under tensions the spline" to "vastly underestimates the spline tension";

  Fixed.

- Line 499: "... this suggests that using a method...";

  Fixed.

- Lines 572-573: For clarity, I suggest substituting the sentence "For signals similar to the Matérn process ..." to "For signals with spectral characteristics compatible with those of the Matérn process".

  Changed to 'More generally, for signals with second-order structure similar to a Matérn process...'

4 Reviewer 4

This paper tries to deal with the noisy GPS data, using the smoothing splines. When I try to read this paper, it is difficult for me to locate the key points of this study.

We have shortened and improved various sentences to improve the readability of the manuscript and we hope the contributions are now clearer. In addition, we have a new section (Section 8) with contrasts our method with competitors and puts our contributions in a wider context. Our key contribution is that we comprehensively detail how smoothing splines can be used to smooth and interpolate GPS data where we show how the parameters can be automatically set using correct physical reasoning, while simultaneously providing automated methods to deal with outliers (where we provide open-source code for the benefit of practitioners).

In the abstract, the authors tell me what they do in this work, but they should be specified. How the spline order and tension parameter are chosen,

We are not sure if the reviewer is asking for these details to be in the abstract or main body. If in the main body, then we have already specified exactly how the spline order should be set in Section 2d, and the tension parameter in Section 3c. We refer the reviewer to these sections. If in the abstract then we feel this level of detail is too much for the abstract and would make it too lengthy and difficult to explain without properly setting notation and the setup first.

...and why this method is effective rather than other methods, when solving the noisy GPS data.

Thank you for this suggestion. As we just mentioned, we have included a new discussion section (Section 8) which contrasts our method with alternative approaches.

The sections talking about the smoothing splines are too long, and if I write this paper, I will start with showing the noisy GPS data, analyze them, show optional methods, make the optimal choice and show the final satisfying results.

Thank you for this suggestion. It is true that perhaps this paper could be written in a different ordering. The key motivation for our ordering is that we want our methods to be as accessible as possible to a wider community. GPS signals are everywhere and our paper is deliberately written such that it is not solely focused on oceanographic GPS drifter data from one experiment. By providing a thorough mathematical analysis of smoothing splines we are able to arrive at optimal parameter choices which apply in any setting, and are mathematically sound and robust. The danger of starting with the data is that the chosen method becomes ad-hoc and over-fitted to one data set. We develop our methodology from first principles, and then test on a real-world application afterwards—this is the usual approach in methodologically-driven papers, and this is why we have kept with the structuring of the paper that we originally had.

Obviously, the authorship put little attention to the noisy data.

We are not sure we understand this point. The paper is fundamentally driven on noisy observations. The fact that observations are noisy is stated up front in the very first equation of the paper, in paragraph 2 of the introduction, where we state that "observations from GPS receivers return observed positions x_i at times t_i that differ from the true positions $x_{\text{true}}(t_i)$ by some noise $\epsilon_i \equiv x_i - x_{\text{true}}(t_i)$ with variance σ^2." This is the centre of our attention when building our methodology. Furthermore, we explore the possibility of heavy-tailed noise and large outliers in Sections 6 and 7 respectively.

Obviously, the structure of this paper is not well organized. Spline methods are mature methods to filter the noisy data, this paper is not novel and the application is not so satisfying.

Spline methods are indeed mature methods and widely used. This is why a paper outlining how parameters can be chosen *a priori* is especially significant and novel. Furthermore we propose novel methods to deal with non-Gaussian noise and outliers. The application is very important to the oceanographic community. The subject of smoothing and interpolating oceanographic drifter data has received wide attention (Elipot et al (2016), Hensen and Poulain (1996) and Yaremchuk and Coelho (2015)) and we have made clear our novelty and contribution with respect to these papers in the new discussion section (Section 8).

In terms of the GPS data in Fig 8, where are the noisy data? Or the noisy data are not exaggerated allowing to see? More detailed work is necessary in the real noisy data.

This figure represents a real data example and the significant presence of noise in data such as this is well documented in the literature and clear to see from the figure. For example, a linear interpolant assuming no noise would produce a very jagged path that is clearly not realistic or close to the true path. We have not over or under exaggerated the noise in the figure, and we do not see the rationale for doing so given this is real data. In terms of further analysis, indeed we encourage other practitioners to implement these procedures on other GPS data sets, and this is why we have structured the paper as we have, and provided online code to help readers to do this.