Smoothing and Interpolating Noisy GPS Data with Smoothing Splines

Jeffrey J. Early* and Adam M. Sykulski†

*Corresponding author address: NorthWest Research Associates, 4118 148th Ave NE, Redmond, WA 98052, USA

E-mail: jearly@nwra.com

†Current affiliation: Data Science Institute and the Department of Mathematics and Statistics, Lancaster University, Bailrigg LA1 4YW, UK (email: a.sykulski@lancaster.ac.uk)

ABSTRACT

A comprehensive methodology is provided for smoothing noisy, irregularly sampled data with non-Gaussian noise using smoothing splines. We demonstrate how the spline order and tension parameter can be chosen *a priori* from physical reasoning. We also show how to allow for non-Gaussian noise and outliers which are typical in GPS signals. We demonstrate the effectiveness of our methods on GPS trajectory data obtained from oceanographic floating instruments known as drifters.

1. Introduction

In the summer of 2011 an array of floating ocean surface buoys (drifters) were deployed in the Sargasso Sea to assess the lateral diffusivity of oceanic processes (Shcherbina et al. 2015). Each drifter was equipped with a global positioning system (GPS) receiver recording locations every 30 minutes. Addressing the primary goal of understanding the physical processes controlling lateral diffusivity requires significant processing of the drifter positions, including removing the mean flow across all drifters, accounting for the large scale strain field, and analyzing the residual spectra for hints of a dynamical process. However, it quickly became clear that the GPS position data, which can have accurracies as low as a few meters (WAAS T&E Team 2016), was contaminated by outliers with position jumps of hundreds of meters or more. Prior to analysis, the position data requires removing the outliers, and interpolating gaps to keep the position data synchronized in time across the drifter array.

The basic problem is ubiquitous: 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 $\varepsilon_i \equiv x_i - x_{\text{true}}(t_i)$ with variance σ^2. The primary goal of *smoothing* is to find the true position $x_{\text{true}}(t_i)$ not contaminated by the noise, while the primary goal of *interpolating* is to find the true position $x_{\text{true}}(t)$ between observation times.

The approach taken here is to use smoothing splines. Our model for the 'true' path $x(t)$ is specified using interpolating b-splines $X^K(t)$ such that

$$x(t) = \sum_{i=1}^{N} \xi_i X_i^K(t), \tag{1}$$

where K is the order (degree $S = K-1$) of the spline. For N observations we construct N b-splines such that $x(t_i) = x_i$ for appropriately chosen coefficients ξ_i. To smooth the data we choose new

coefficients $\bar{\xi}_i$ that minimize the penalty function

$$\phi = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2 + \frac{\lambda_T}{t_N - t_1}\int_{t_1}^{t_N}\left(\frac{d^T x}{dt^T}\right)^2 dt, \tag{2}$$

for some tension parameter $\lambda_T \geq 0$. If $\lambda_T = 0$ then $\phi = 0$ and $\xi_i = \bar{\xi}_i$ because $x(t_i) = x_i$, but if $\lambda_T \to \infty$ then this forces the $x(t)$ to a T-th order polynomial (e.g., when $T = 2$, the model is forced to be a straight line because it has no second derivative). The resulting path $x(t)$ is known as a smoothing spline and was first introduced in modern form by Reinsch (1967), but according to De Boor (1978) the idea dates back to Whittaker (1923). Once S and T are chosen, the smoothing spline has one free parameter (λ_T) and its optimal value can be found by minimizing the expected mean square error when the true value of σ is known (Craven and Wahba 1979).

As a practical matter there are three issues that must be addressed before smoothing splines are applied to GPS data:

1. how do we choose S and T—and how do these choices affect the recovered power spectrum?

2. how do we modify the spline fit to accommodate the non-Gaussian errors of GPS receivers?

3. how do we identify and remove outliers?

To address these issues, but also serve as a practical guide to other practitioners, we start by reviewing B-splines in section 2 and introduce the canonical interpolating spline that is used as the underlying model for path $x(t)$ in (1). We also demonstrate the effect that choosing S has on the high-frequency slope of the power spectrum of the interpolated fit.

Section 3 takes a broad look at smoothing splines and the assumptions they make on the underlying process. Many of the ideas presented in this section are known to the statistics community, so here we present these ideas from a more physical perspective. We show that the penalty function

in (2) can be formulated as a maximum likelihood problem and applying tension is equivalent to assuming a Gaussian distribution on the tensioned derivative of the underlying process.

Section 4 uses ensembles from synthetic data designed to mimic the oceanographic data in order to test a number of choices that have to be made. We first establish that setting $T = S$ is a reasonable choice. We then show that the tension parameter can be chosen *a priori* (without optimization of the mean square error) when the *effective sample size* (which we define later) can be estimated from the data. This estimate for the effective sample size can then be used to reduce the coefficients, ξ^i, in the spline fit without increasing mean square error. Finally, we show how the effective sample size of the fit establishes the highest resolved frequency.

The second half of the manuscript addresses issues specific to GPS positions errors. In section 5 we discuss the assumptions of stationarity and isotropy required for bivariate smoothing splines. In section 6 we show that the GPS errors are not Gaussian distributed, but t-distributed, and we show how to modify the technique for a t-distribution. Finally, section 7 addresses how to modify the expected mean square error minimizer to make smoothing splines robust to outliers.

One of the major outcomes of this work is the implementation of Matlab classes for generating b-splines, interpolating splines, smoothing splines as well as a class specific to smoothing GPS data[1]. These classes are highlighted throughout the manuscript in their relevant sections.

2. Interpolating Spline

Assume that we are given N observations of a particle position (t_i, x_i) with no errors. The simplest possible form of interpolation would be a nearest neighbor method that assigns the position of the particle to the nearest observations in time. The resulting interpolated function $x(t)$ is a polynomial of *order* $K = 1$ (piecewise constant), shown in the top row of Fig. 1. The next level

[1] https://github.com/JeffreyEarly/GLNumericalModelingKit

of sophistication is to assume a constant velocity between any two observations and use that to interpolate positions between observations, second row of Fig. 1. This also means that we now have a piecewise constant function $\frac{dx}{dt}$ that represents the velocity of the particle, shown in the second row, second column of Fig. 1. This is a polynomial function of order $K = 2$.

It is slightly less obvious how to proceed to a polynomial of order $K = 3$. With N data points we can construct a piecewise constant acceleration (the second derivative) using the $N-2$ independent accelerations computed from finite differencing, but where to place *knot points* that define the boundaries of the regions and how to maintain continuity is slightly less clear. The approach taken here is to use B-splines.

a. B-Splines

A B-spline (or basis spline) of *order* K (*degree* $S = K - 1$) is a piecewise polynomial that maintains nonzero continuity across S knot points. The knot points are a nondecreasing collection of points in time that we will denote with τ_i. The basic theory is well documented in De Boor (1978), but here we will present a reduced version specifically tailored to our needs.

The m-th B-spline of order $K = 1$ is defined as

$$X_m^1(t) \equiv \begin{cases} 1 & \text{if } \tau_m \leq t < \tau_{m+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This is the rectangle function as shown in the first row, first column of Fig. 2. If we are given P knot points, then we can construct $P-1$ B-splines of order $K = 1$, although notice that if a knot point is repeated this will result in a spline that is zero everywhere. To represent an interpolating

function $x(t)$ for the N observations of a particle position (t_i, x_i) we define $N+1$ knot points as

$$\tau_m = \begin{cases} t_1 & m = 1, \\ t_{m-1} + \frac{t_m - t_{m-1}}{2} & 1 < m \leq N, \\ t_N & m > N. \end{cases} \tag{4}$$

This will create N independent basis functions that provide support for the region $t_1 \leq t \leq t_N$ (provided the last spline is defined to include the last knot point). The interpolating function $x(t)$ is defined as $x(t) \equiv X_m^1(t)\xi^m$ where the coefficients ξ^m are found by solving $X_m^1(t^i)\xi^m = x^i$. The result of this process is shown in Fig. 1 for 7 irregularly spaced data points.

All higher order B-splines are defined by recursion,

$$X_m^K(t) \equiv \frac{t - t_m}{t_{m+K-1} - t_m} X_m^{K-1}(t) + \frac{t_{m+K} - t}{t_{m+K} - t_{m+1}} X_{m+1}^{K-1}(t). \tag{5}$$

This recursion formula takes two neighboring lower order splines and ramps the left one up over its nonzero domain and ramps the right one down over its nonzero domain. The result of this process is to create splines that span across one additional knot point at each order, and maintain continuity across one more derivative. Examples are shown in Fig. 2.

Any knot points that are repeated T times will result in a total of $T-1$ splines of order one that are everywhere zero. This has the effect of introducing discontinuities in the derivatives for higher order splines. For our purposes, we will only use this feature to prevent higher order splines from crossing the boundaries. For $K = 2$ order splines we will use $N+2$ knot points at locations

$$\tau_m = \begin{cases} t_1 & m \leq 2, \\ t_{m-1} & 2 < m \leq N, \\ t_N & m > N. \end{cases} \tag{6}$$

This creates a knot point at every observation point, but repeats the first and last knot point. This has the effect of terminating the first and last spline at the boundary and creating N second order B-splines, $X_m^2(t)$. Once again the interpolating function $x(t)$ is defined as $x(t) \equiv X_m^2(t)\xi^m$ where the coefficients ξ^m are found by solving $X_m^2(t^i)\xi^m = x^i$. The second row of Fig. 1 shows an example.

This process can be continued to higher and higher order B-splines. For splines that are of *even* order, we create $N+K$ knots points with

$$\tau_m^{K\text{-even}} = \begin{cases} t_1 & m \leq K, \\ t_{m-K/2} & K < m \leq N, \\ t_N & m > N, \end{cases} \tag{7}$$

and for splines that are *odd* order, we create $N+K$ knot points with

$$\tau_m^{K\text{-odd}} = \begin{cases} t_1 & m \leq K, \\ t_{m-\frac{K+1}{2}} + \frac{t_{m+1-\frac{K+1}{2}} - t_{m-\frac{K+1}{2}}}{2} & K < m \leq N, \\ t_N & m > N. \end{cases} \tag{8}$$

The knot points are chosen specifically to create N splines for the N data points such that the interpolated function $x(t)$ crosses all N observations (t_i, x_i). The path $x(t)$ is the *canonical interpolating spline of order K*. Examples are shown in Fig. 1.

The knot placements in (7) and (8) are equivalent to the *not-a-knot* boundary conditions described in De Boor (1978) and used in the cubic spline implementation in `Matlab`. In the usual formulation of the not-a-knot boundary condition, the knot positions do not change as a function of spline order, and therefore additional constraints have to be added at each order—especially the requirement that the highest derivative maintain continuity near the boundaries. In the formulation here, these constraints are implicit in (7) and (8).

b. Numerical implementation

The root class in our suite of Matlab classes is the `BSpline` class, which evaluates a complete B-spline basis set given a set of knot points. This class was used to generate Fig. 2.

The interpolating spline used to generate Fig. 1 is implemented in the `InterpolatingSpline` class—a sublcass of `BSpline`. This class generates interpolating splines of arbitrary order given a set of data points (t_i, x_i), thus generalizing the cubic spline command that is built in to Matlab.

c. Synthetic Data

Throughout this manuscript we generate synthetic data for both the signal and the noise. The velocity of the signal is generated from a Gaussian process known as the Matérn (Lilly et al. 2017). The spectrum of the Matérn is given by

$$S(\omega) = \frac{A^2}{(\omega^2 + \lambda^2)^{p/2}}, \tag{9}$$

with $p > 1$, which has finite amplitude at low frequencies and power-law fall off at high frequencies, two physically realistic properties observed, among other things, in ocean surface drifters (Sykulski et al. 2016).

For these experiments we choose values of $p = 2, 3, 4$ so that the high frequency spectrum is proportional to ω^{-2}, ω^{-3}, ω^{-4}. The Matérn is used to generate the *velocity* of the signal and integrated to get positions. Parameters are chosen such that the square root of velocity variance in each direction is $u_{\text{rms}} = 0.20$ m/s and the damping scale $\lambda^{-1} = 30$ minutes. These choices resemble the data from the drifters. Fig. 3 shows an example velocity spectrum of the signal with ω^{-2}.

The position data is contaminated with (white) Gaussian noise with $\sigma = 10$ meters, a value chosen to resemble GPS errors. In section a we consider noise generated from a t-distribution which more accurately reflects GPS errors.

For all of these experiments we use a range of *strides*, that is, subsampled versions of the underlying process as input into the spline fits. A stride of 100 indicates that the signal is subsampled to 1 every 100 data points. This lets us evaluate the quality of fit against different sampling rates.

d. Spline degree, S

We first examine a synthetic signal *uncontaminated* by noise, to examine the role of spline degree, S, on the interpolated fit. As noted in Craven and Wahba (1979), the degree of the spline sets its roughness. In terms of the power spectrum, this corresponds to the high frequency slope as can be seen in Fig. 3 which shows fits with $S = 1..4$. Setting $S = 1$ produces a high frequency fall off in the spline fit of ω^{-2}. Although this would appear to be a desirable feature when fitting to a process with slope ω^{-2}, the mean square error is consistently higher.

The bottom panel of Fig. 3 shows the coherence between the spline fit and the true signal. There is no discernible difference in coherence between spline fits with $S = 1..4$. The coherence quickly drops to near zero at the same frequency in all three cases. The implication here is that the spline fits are essentially producing noise at frequencies above the loss-of-coherence. 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). The conclusion here is that smoother is better: it is better to use an unnecessarily high order spline to avoid adding extra noise at high frequencies.

3. Smoothing Spline

A typical starting point for maximum likelihood is to establish the probability distribution function (PDF) of the errors, $\varepsilon_i \equiv x_i - x_{\rm true}(t_i)$. The canonical example in one-dimension (e.g., Press et al. (1992)) is to assume that the error in our position measurements are Gaussian i.i.d. and are therefore drawn from the following probability distribution

$$p_g(\varepsilon|\sigma_g) = \frac{e^{-\frac{1}{2}\frac{\varepsilon^2}{\sigma_g^2}}}{\sigma_g\sqrt{2\pi}}, \tag{10}$$

where σ_g is the standard deviation. This assumption alone places no assumptions on the signal itself, only on the structure of the noise.

The probability of the observed data given model $x(t)$ is

$$P = \frac{1}{\sigma\sqrt{2\pi}}\prod_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2\right], \tag{11}$$

where we have taken $\sigma = \sigma_g$.

Maximizing the probability function in (11) is also the same as minimizing its argument—up to a constant this is the log likelihood, called the penalty function

$$\phi = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2. \tag{12}$$

Stated in this way it is plain to see that this is the same as asking for the 'least-squares' fit of the errors.

a. Smoothing spline penalty function

The model used here will be the canonical interpolating spline of order K described in section 2. Of course, we have chosen our knot points such that the model intersects the observations and this certainly maximizes (11) (and minimizes (12)) because all the errors are zero, but the resulting distribution of errors (a delta function at zero) does not look anything like the assumed Gaussian

distribution. Thus, if we want the error distribution that we get out to look like that which we assumed, it is necessary to constrain the problem in some way.

The smoothing spline augments the penalty function of (12) by adding a global constraint on the m-th derivative of the resulting function as in (2). If $\lambda_T \to 0$ then this reduces to the least-squares fit in (12), but if $\lambda_T \to \infty$ then this forces the model to an T-th order polynomial.

To interpret the first term of (2), consider a motionless particle at true position x_0. Using the N relevant observations x_i, the *sample mean* $\bar{x} = \frac{1}{N}\sum x_i$ estimates the particle's position x_0. The unbiased *sample variance* estimates the variance of the noise, σ^2, and is given by $\hat{\sigma}^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2$, the expected value of which is $\left\langle \hat{\sigma}^2 \right\rangle = \left(1 - \frac{1}{N}\right)\sigma^2$.

Now consider the opposite extreme where the particle is moving so fast (or the observations are so sparse) that each observation is completely independent of its neighbors. In this case, each observation must be considered separately, so the sample mean at time t_i is just $\bar{x}_i = x_i$ (i.e., we are summing over the single relevant observation). In this scenario we cannot produce a sample variance, because there is only a single relevant observation at time t_i.

In practice, the number of relevant observations can be anywhere between 1 and N. Here we use the term *effective sample size*, denoted by n_{eff}, to describe the typical number of observations being used to estimate either the particle's position or the variance of the noise at any given time. In this context, the first term of (2) is proportional to an ensemble of multiple estimates of the sample variance

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{i=1}^{N}\left(x_i - x(t_i)\right)^2, \tag{13}$$

which is expected to scale as

$$\left\langle \hat{\sigma}^2 \right\rangle = \left(1 - \frac{1}{n_{\text{eff}}^{\text{var}}}\right)\sigma^2, \tag{14}$$

where $1 < n_{\text{eff}}^{\text{var}} \leq N$ is our definition of the effective sample size as determined from the sample variance. Revisiting the limiting cases, as $n_{\text{eff}}^{\text{var}} \to N$ the sample variance matches the true variance, but as $n_{\text{eff}}^{\text{var}} \to 1$, the sample variance vanishes.

There is a very simple physical interpretation for the second term in (2). Consider the case where $T = 1$ so that the smoothing spline is a constraint on velocity. When averaged over the integration time, the integral produces the root mean square velocity, u_{rms}, which means that the second term scales like u_{rms}^2. In general, where $x_{\text{rms}}^{(T)}$ is the root-mean-square of the T-th derivative, this means that λ_T scales like

$$\lambda_T = \left(1 - \frac{1}{n_{\text{eff}}^{\text{var}}}\right) \frac{1}{\left(x_{\text{rms}}^{(T)}\right)^2}. \tag{15}$$

The interpretation of the smoothing spline is therefore that the two terms are balanced by a relative weighting of the sample variance of the noise and mean square of the T-th derivative of the physical process. As will be discussed in section 4, both $x_{\text{rms}}^{(T)}$ and $n_{\text{eff}}^{\text{var}}$ can be estimated *a priori* and therefore a good initial estimate for λ_T can be made.

b. Smoothing spline maximum likelihood

The penalty function for the smoothing spline in (2) can be restated in terms of maximum likelihood under some conditions (see also chapter 3.8 in Green and Silverman (1994)). Assume that in addition to knowing about how the measurement errors are distributed like in (11), that we also know how the velocity of underlying physical process is distributed. For example, in geophysical turbulence it has been shown that the velocity probability distribution function is like the Laplace distribution (Bracco et al. 2000). To recover the smoothing spline, we need to consider the case where the velocity PDF is Gaussian. Stated as maximum likelihood, this means that at *any given instant* (not just the times of observation) we expected the model velocity to look Gaussian. We can discretize the problem by sampling the velocity Q times $t_q = t_1 + q\Delta t_q$, where $\Delta t_q = \frac{t_N - t_1}{Q-1}$ and

$q = 0..Q-1$. The maximum likelihood is thus stated as

$$P = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2\right] \cdot \prod_{q=1}^{Q} \frac{\sqrt{\gamma}}{x_{\text{rms}}^{(T)}\sqrt{2\pi}} \exp\left[-\frac{\gamma}{2}\left(\frac{x^{(T)}(t_q)}{x_{\text{rms}}^{(T)}}\right)^2\right], \tag{16}$$

which is simply the joint probability of the error distribution from (11) and the velocity distribution of the underlying physical process. We also include parameter γ for convenience in order to set the relative weighting between the two distributions, although it could be absorbed into the definition of $x_{\text{rms}}^{(T)}$. Writing (16) as a penalty function (after converting the product of exponentials into exponentials of sums), we have that

$$-\log P = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2 + \frac{\gamma}{2}\sum_{q=1}^{Q}\left(\frac{x^{(T)}(t_q)}{x_{\text{rms}}^{(T)}}\right)^2 + C, \tag{17}$$

where C is a constant. Setting $\gamma = \frac{N}{Q}$ and renormalizing the penalty function by $\frac{2}{N}$ (which has no effect on the location of its minimum), (17) can be written as

$$\phi = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - x(t_i)}{\sigma}\right)^2 + \frac{1}{t_N - t_1}\sum_{q=1}^{Q}\left(\frac{x^{(T)}(t_q)}{x_{\text{rms}}^{(T)}}\right)^2 \Delta t_q. \tag{18}$$

Apart from the discretization of the integral, (18) is the same as the penalty function for a smoothing spline (2).

There is an important special case when tension is applied at the same order as the spline, $T = S$. In this case the spline is piecewise constant for $x^{(T)}$ with exactly $N - T$ unique values. The parameter $\gamma = \frac{N}{N-T} \approx 1$ and (16) can be simplified. This case is appealing because only the $N - T$ unique values of the derivative $x^{(T)}$ that can be computed from N data points are being used for tension, which is not the case when $T < S$.

This maximum likelihood perspective shows that adding tension to the penalty function is equivalent to assuming that one of the higher order derivatives in the model (e.g., velocity if $T = 1$) is

Gaussian. This is therefore making an assumption about the underlying *physical process* of the model. This is in contrast to the first term which is entirely a statement about *measurement noise*.

As an aside, writing the smoothing spline as a maximum-likelihood condition (16), suggests that if the underlying physical process has a non-zero mean value in tension, the fit will not behave as expected. However, smoothing splines can be easily modified to accommodate a mean value in tension, as shown in appendix 8.

c. Optimal parameter estimation

For a given choice of T and λ_T, the minimum solution to (2) can be found analytically (see Teanby (2007) and our appendix 8). Once the solution is found the smoothing matrix $\mathbf{S}_\lambda$ is defined as the matrix that takes the observations $\mathbf{x}$ and maps them to their smooth values, $\hat{\mathbf{x}} = \mathbf{S}_\lambda \mathbf{x}$.

The free parameter λ_T is a relative weighting between the two terms in (2) and choosing its optimal value can be done by minimizing the expected mean square error (Craven and Wahba 1979),

$$\mathrm{MSE}(\lambda) = \frac{1}{N}||(\mathbf{S}_\lambda - I)\mathbf{x}||^2 + \frac{2\sigma^2}{N}\mathrm{Tr}\mathbf{S}_\lambda - \sigma^2, \tag{19}$$

where $||\cdot||^2$ is the Euclidean norm and Tr indicates the trace.

It is worth noting that a fair amount of the literature on smoothing splines is devoted to minimizing the mean square error when the variance, σ^2, is *not* known. For example, Craven and Wahba (1979) and Wahba (1978) use cross-validation to estimate σ and minimize the mean square error. Recent work comparing different estimators shows that no single technique appears to be optimal (Lee 2003). For our application however, the errors in GPS data can be relatively easily established, as shown in section 6.

The mean square error in (19) is a combination of the sample variance and the variance of the mean. As already discussed in the context of the penalty function ϕ in section a, the first term in (19) is an ensemble of sample variances, and therefore by combining (13), (14) and (19) we obtain

$$\left(1-\frac{1}{n_{\text{eff}}^{\text{var}}}\right)\sigma^2 = \frac{1}{N}||\left(\mathbf{I}-\mathbf{S}_\lambda\right)\mathbf{x}||^2. \tag{20}$$

The second term in (19) is proportional to twice the squared standard error, i.e., the variance of the sample mean. As discussed in Teanby (2007), the quantity $\mathbf{S}_\lambda\Sigma$ is the covariance matrix with the squared standard error along the diagonal and thus the mean squared standard error is given by $\frac{1}{N}\text{Tr}\left(\mathbf{S}_\lambda\Sigma\right)$. The variance of the sample mean is known to scale inversely with the number of samples being used to estimate the mean. Thus, we use this to define the effective sample size of the variance of the mean, $n_{\text{eff}}^{\text{SE}}$ with

$$\frac{\sigma^2}{n_{\text{eff}}^{\text{SE}}} = \frac{1}{N}\text{Tr}\left(\mathbf{S}_\lambda\Sigma\right). \tag{21}$$

Taking the measures of effective sample size as functions of λ, the mean square error can be expressed by combining (19)–(21) such that

$$\text{MSE}(\lambda) = 2\frac{\sigma^2}{n_{\text{eff}}^{\text{SE}}} - \frac{\sigma^2}{n_{\text{eff}}^{\text{var}}}. \tag{22}$$

If one assumes that $n_{\text{eff}}^{\text{var}} = n_{\text{eff}}^{\text{SE}}$, then the expected mean square error from (19) is equal to σ^2/n_{eff}. Although not shown here, in an empirical analysis we find that $n_{\text{eff}}^{\text{var}}$ and $n_{\text{eff}}^{\text{SE}}$ are approximately equal, although $n_{\text{eff}}^{\text{var}}$ becomes highly variable when $n_{\text{eff}}^{\text{SE}}$ approaches 1.

These measures of effective sample size can be used to estimate the value of λ_T necessary for optimal tension without minimizing the expected mean square error. Note that the definition of effective sample size used here is related to, but not the same as, the notion of degrees-of-freedom used in Cantoni and Hastie (2002) and references therein.

4. Spline order, tension order, and the spectrum

With a model path (1), a penalty function (2), and a minimization condition (19), we have all the primary pieces to create a smoothing spline interpolant to the data. However, there are a number of choices that still have to be made. In this section we use synthetically generated data to represent our physical process, and contaminate the process with Gaussian noise as described in section c. We use this synthetically generated data to test our ability to recover the signal and examine the effects of changing the spline and tension order on the mean square error and the resulting spectrum.

The results of this section are empirical, and it is important to acknowledge upfront that any conclusions reached *may* depend on our particular choice of physical model that generates the signal which has been chosen to resemble the oceanographic data of interest. Nevertheless, our expectation is that the conclusions here are 'O(1)' correct, and applicable, at least, to our GPS tracked drifter dataset.

a. Tension degree, T

Given a smoothing spline of degree S, the tension in the penalty function (2) can be applied at any degree $T \leq S$. We use the synthetic data for the three different slopes to empirically establish the relationship between the tension degree, T and the spline degree, S.

For $S = 1 \ldots 5$ and all $T \leq S$ we minimize the mean square error against the true values. The minimization is performed for 200 ensembles of noise and signal with three slopes (ω^{-2}, ω^{-3}, ω^{-4}) and 5 different strides. For a given slope, stride, and realization of noise, we identify the minimum mean square error across S and T and compare all other values of S and T as a percentage increase relative to that minimum. After aggregating across slopes, strides, and ensembles, the 68% confidence range is shown in table 1.

The results in table 1 show that while setting $T = S$ may not always be optimal, it is never significantly worse than the optimal choice. Thus, for the remainder of the manuscript, we will always take $T = S$. This choice is the same as the special case highlighted in section 3.

b. Loss of coherence

The loss-of-coherence defines the time scale below which the smoothing spline is not providing useful information. A reasonable hypothesis is that this scale is related to the effective sample size, n_{eff} because the effective sample size indicates how many points are being used to estimate the true value. Therefore the loss-of-coherence occurs at the *effective Nyquist* which we define as

$$f_s^{\text{eff}} \equiv \frac{1}{2n_{\text{eff}}\Delta t}. \tag{23}$$

In practice, we use $n_{\text{eff}}^{\text{SE}}$ because it is less variable than $n_{\text{eff}}^{\text{var}}$ for values near 1 and is the more direct measure of how many points are being used to estimate the model path. Fig. 4 shows the power spectrum and coherence of optimal tension fits for three different strides of the data. In all three cases (23) indicates almost exactly where the coherence drops below 0.5.

c. Reduced spline coefficients

One practical consideration when working with large datasets is that the computational cost of creating the spline fit may be limited by the rate of solving for the spline coefficients. It is therefore beneficial to reduce knot points (and therefore total splines) where possible. A reasonable hypothesis is to suppose that when the effective sample size is large, as measured by (21), that we may be able to avoid placing a knot point at every data point—essentially 'skipping' data points.

To test this idea, we find the optimal fit over a range of different strides (which varies the effective sample size) and increase the number of knot points that are skipped until the mean square error starts to rise. We find that we can safely skip $\max(1, \text{floor}(2n_{\text{eff}}/3))$ knot points without sacrificing

any precision. In fact, as can be seen in table 2, in some cases the optimal mean square error improves with fewer knot points. The 'full dof' column indicates a fit where one knot point is created for every observation point, whereas the 'reduced dof' indicates a fit where the number of knot points is reduced.

This means that when handling large datasets, we can reduce the number of splines being used if the effective sample size is large, and we can simply 'chunk' the data (split into multiple independent pieces) when the effective sample size is small.

d. Interpolation condition

To estimate the value of λ_T from (15), we require an estimate of the mean square value of a derivative of the process, $x_{\text{rms}}^{(T)}$ as well as an estimate of the effective sample size, n_{eff}. Assuming one can make an estimate of $x_{\text{rms}}^{(m)}$ from the signal (see appendix 8), we just need a method for estimating the effective sample size.

We argue that the effective sample size should vary based on the relative size of the measurement errors to the speed of motion. For example, if the position errors are only 1 meter, but a particle typically travels 10 meters between measurements, then it is hardly justifiable to increase the tension so that the smoothing spline misses the observation points by 1 meter. There is not enough statistical evidence to suggest that the particle didn't go right through the observation point. On the other hand, if the position errors are 1 meter, but the particle typically travels 10 centimeters between measurements, nearby measurements provide more information about the particle's true position during that time, so our estimate of the particle's true position is closer to a mean of the nearby observations.

This idea can be made more rigorous by noting that one would consider change in position, Δx, statistically significant if it exceeds the position errors σ by some factor. Assuming the physical

process has a characteristic velocity scale, u_{rms}, we use this concept to define Γ as

$$\Gamma \equiv \frac{\sigma}{u_{\text{rms}}\Delta t}, \tag{24}$$

where Δt is the typical time between observations. This argument suggests that the effective sample size should be proportional to Γ, i.e.,

$$n_{\text{eff}}^{\Gamma} = \max\left(1, C \cdot \Gamma^{m}\right) \tag{25}$$

where C and m are unknown constants, and we prevent the effective sample size from dropping below 1. Intuitively this means that as long as the particle does not move too far between observations, nearby observations help to estimate the true position of the particle.

To test the relationship between Γ and the effective sample size, we compute the optimal smoothing spline for a range of values of Γ (created by sub-sampling the signal) for the three different slopes (ω^{-2}, ω^{-3}, ω^{-4}). The value $n_{\text{eff}}^{\text{SE}}$ is computed from the optimal solution for 50 ensembles and shown in Fig. 5. The fits are remarkably good, but depend on the slope of process. Processes with shallower slopes (rougher trajectories) provide a smaller effective sample size for a given value of Γ.

Using the interpolation condition Γ to estimate the effective sample size, we set $n_{\text{eff}}^{\Gamma} = 14 \cdot \Gamma^{0.71}$, the empirically determined best fit for slope ω^{-3}. For all spline fits then, we use

$$\lambda_T^{\text{initial}} = \left(1 - \frac{1}{n_{\text{eff}}^{\Gamma}}\right) \frac{1}{\left(x_{\text{rms}}^{(T)}\right)^2} \tag{26}$$

as an initial estimate for the optimal smoothing parameter where both $x_{\text{rms}}^{(T)}$ in (26) and u_{rms} in (24) are estimated using the method described in appendix 8.

The scaling law for n_{eff}^{Γ} can be found analytically. Let the position observations be given by x_i where

$$x_i = u_{\text{rms}} i\Delta t + \varepsilon_i \text{ where } \varepsilon_i = \mathcal{N}(0, \sigma). \tag{27}$$

If the effective sample size is $\langle n \rangle$, then the particle changes position by $\langle n \rangle u_{\text{rms}} \Delta t$ between samples. Applying the two-sample z-test, two positions will be considered different for $z > z_{\min}$ where

$$z = \frac{\langle n \rangle u_{\text{rms}} \Delta t}{\sqrt{\frac{\sigma^2}{\langle n \rangle} + \frac{\sigma^2}{\langle n \rangle}}} \quad \Rightarrow \quad \langle n \rangle = \left(\frac{z\sigma\sqrt{2}}{u_{\text{rms}} \Delta t} \right)^{\frac{2}{3}}. \tag{28}$$

The power law in (28) matches the empirically derived power laws shown in Fig. 5 and suggests that m in (25) should be $m = 2/3$. This also suggests that the coefficient C in (25) can be related to z, a measure of statistical significance.

e. Optimal fits

Table 2 summarizes the key results of this section by applying a smoothing spline with with $S = 3$ to the 200 ensembles of the noise and signal with three different slope (ω^{-2}, ω^{-3}, ω^{-4}) and five different strides. 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 to minimize the mean square error—this is the lower bound. The fourth column shows average increase in mean square error when reducing the number of spline coefficient as documented in section c. There is almost no change in mean square error and therefore all subsequent methods (whether blind or unblind) use this technique. The fifth column uses (26) from section d to provide a (blind) initial guess of the tension parameter. Here the results are mixed—a typical increase in mean square error is about 30-50% when the effective sample size is large. While this might seem large, this is a small fraction of the total variance of the noise, e.g., an optimal mean square error of 6 m^2 increase to 8 m^2 when the total variance is 100 m^2. When the data sets are small (and computation time is not a limiting factor), nearly optimal fits can be found using (19), as shown in the last column of the table.

f. Numerical implementation

The numerical implementation of the methods in this section are available in the `SmoothingSpline` class which subclasses `BSpline`. This class is initialized with three required parameters: a set of data points (t_i, x_i) and a distribution (specifically a normal distribution for the results in this section). The initial value of λ_T is chosen using (26). The `SmoothingSpline` class implements a `.minimize()` method which takes any function of the spline as an argument (such as (19)), and minimizes the function by varying λ_T.

5. Bivariate smoothing splines and stationarity

Up to this point we have considered univariate data, (t_i, x_i), but GPS position data is fundamentally bivariate. The term 'bivariate' in the context of splines is often used to denote splines defined on two independent variables—however, in this context we define bivariate to mean two dependent variables (e.g., x and y) and one independent variable (e.g., t).

The trivial approach to working with such bivariate data is to treat each direction independently—i.e., minimize λ_T^x and λ_T^y independently of each other. However, it is often the case that the underlying physical process is isotropic. In the context of the maximum likelihood formulation of smoothing splines (18), this means that we expect $x_{\text{rms}}^{(T)}$ (the rms value of the tensioned variable) to be the same in all directions (invariant under rotation). This however does *not* mean that λ_x should necessarily equal λ_y. To be explicit, if

$$\begin{aligned}\lambda_T^x &= \left(1 - \frac{1}{n_{\text{eff}}^x}\right) \frac{1}{\left(x_{\text{rms}}^{(T)}\right)^2}, \\ \lambda_T^y &= \left(1 - \frac{1}{n_{\text{eff}}^y}\right) \frac{1}{\left(y_{\text{rms}}^{(T)}\right)^2},\end{aligned} \tag{29}$$

then even if $x^{(T)}_{\text{rms}} = y^{(T)}_{\text{rms}}$, the effective sample sizes n^x_{eff} and n^y_{eff} will not necessarily be equal if there is any mean velocity because, as shown in section d, the effective sample size depends on velocity.

Therefore to assume isotropy in λ_T and use a bivariate smoothing spline, the mean velocity from the underlying process must be removed. What qualifies as mean and fluctuation rarely has a clear answer, but a reasonable option is letting a polynomial of degree $T+1$ define the mean. This has the added benefit of removing a constant non-zero tension value, which as shown in section b, changes the problem formulation.

It is worth noting that it is not actually isotropy that requires removing the mean velocity, but in fact stationarity. The effective sample size is shown to be dependent on rms velocity, so if the velocity varies in time, then the optimal effective sample size will need to vary as well. This means that not only do smoothing splines require stationarity in the tensioned variable $x^{(T)}$ as shown in section b, but they also require stationarity in the velocity $x^{(1)}$ to be effective. This last requirement can be solved by either removing the mean (as we have suggested), or segmenting observations into pseudo-stationary chunks.

a. Assessing errors

Removing the mean or some other low-passed version of the data means that the total smoothing matrix will be some combination of the low-passed and high-passed smoothing matrices. Once this matrix is computed, it can be used to compute the standard errors.

We first create a low pass filter to capture the *mean* component of the flow using a simple polynomial fit,

$$\bar{\mathbf{x}} = \bar{\mathbf{S}}\mathbf{x} \tag{30}$$

and then define the residual as our stationary part,

$$\mathbf{x}' \equiv \mathbf{x} - \bar{\mathbf{x}}. \tag{31}$$

We now compute the smoothing spline as usual on the residual,

$$\mathbf{x}'_\lambda = \mathbf{S}_\lambda \mathbf{x}' \tag{32}$$

So the total, smoothed path is

$$\begin{aligned} \hat{\mathbf{x}} =& \bar{\mathbf{x}} + \mathbf{x}'_\lambda = \bar{\mathbf{S}}\mathbf{x} + \mathbf{S}_\lambda \left(\mathbf{x} - \bar{\mathbf{S}}\mathbf{x}\right) = \left(\bar{\mathbf{S}} + \mathbf{S}_\lambda - \mathbf{S}_\lambda \bar{\mathbf{S}}\right) \mathbf{x} \\ \equiv& \mathbf{S}_T \mathbf{x} \end{aligned} \tag{33}$$

From this we can compute the covariance matrix and the standard error.

b. Numerical implementation

The `BivariateSmoothingSpline` class is initialized with data (t_i, x_i, y_i) and a distribution. For a spline of degree $S = T$, a spline of degree $S+1$ is used to remove the mean in each direction. In the case of a normal distribution, this is simply a least squares polynomial fit. By assumption, the residual data ($\mathbf{x}'$, $\mathbf{y}'$ in the notation above) is stationary and isotropic, so the tension parameter λ_T is applied equally to spline fits in the two directions. Minimization is performed on the sum of the expected mean square error in both directions.

6. GPS data set

The primary dataset considered here will be nine surface drifters that were deployed in the Sargasso Sea in the summer of 2011 (Shcherbina et al. 2015). In the past, such drifters used the Argos positioning system which has significantly poorer temporal coverage and position accuracy (Elipot et al. 2016), but recently the majority of surface drifters have employed GPS receivers and transmitted their data back through Argos or Iridium satellites.

The GPS receiver sits on the surface drifter and collects position data, but because of atmospheric conditions or ocean waves, the receivers are sometimes unable to obtain a position, or when they do, it is highly inaccurate. Thus, despite nominal accuracies of a few meters, it is often the case that some positions are off by more than 1000 meters, as can be seen in Fig. 8. Applying a smoothing spline fit using the methodology in section 3 produces an extremely poor fit, with clear overshoots to bad data points.

a. GPS error distribution

We characterize the GPS errors by considering data from a motionless GPS receiver allowed to run for 12 hours. The specific GPS receiver used for this test was not the same as the one used for the drifters (because it was no longer available) but should produce errors similar enough for this analysis.

The position recorded by the motionless GPS are assumed to have isotropic errors with mean zero, which means that the positions themselves are the errors. The probability distribution function (PDF) of the combined x and y position errors are shown in Fig. 6.

The error distribution is first fit to a zero-mean Gaussian PDF (10). The maximum likelihood fit is found by simply computing the standard deviation of the sample, which is found to be $\sigma \approx 10$ meters and shown as the gray line in Fig. 6. However, it is clear the error distribution shows much longer tails than the Gaussian PDF.

The Student t-distribution is a generalization of the Gaussian that produces longer tails and is defined as

$$p_s\left(\varepsilon|\nu,\sigma_s^2\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma_s\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{\varepsilon^2}{\sigma_s^2\nu}\right)^{-\frac{\nu+1}{2}}, \tag{34}$$

where the σ_s parameter scales the distribution width and the ν parameter sets the number of degrees of freedom. The variance is $\sigma^2 = \sigma_s^2\frac{\nu}{\nu-2}$ and only exists for $\nu > 2$. The t-distribution is

equivalent to the Gaussian distribution when $\nu \to \infty$. We find the best fit t-distribution to the data by minimizing the Anderson-Darling test. The best fit with parameters $\sigma_s \approx 8.5$ meters and $\nu \approx 4.5$ is shown as the black line in Fig. 6. Different choices in GPS receivers and using the Kolmogorov-Smirnoff test results in very similar parameters, i.e., $\sigma_s \approx 8-10$ meters and $\nu \approx 4-6$.

The *position* error distributions also imply a combined *distance* error distribution by computing $\varepsilon_d = \sqrt{\varepsilon_x^2 + \varepsilon_y^2}$ and is shown in the lower panel of Fig. 6. For two independent Gaussian distributions this results in a Rayleigh distribution,

$$p_r(\varepsilon_d|\sigma_g) = \frac{\varepsilon_d}{\sigma_g^2} e^{-\frac{1}{2}\frac{\varepsilon_d^2}{\sigma_g^2}}. \tag{35}$$

The distance distribution for two t-distributions is computed numerically and is shown in the bottom panel of Fig. 6 on top of the actual distance errors. Approximately 95% of distance errors are within 30 meters.

Fig. 7 shows the autocorrelation function of the GPS position errors. We find a rough empirical fit to be $\rho(\tau) = \exp(\max(-\tau/t_0, -\tau/t_1 - 1.35))$ where $t_0 = 100$ seconds and $t_1 = 760$ seconds, which reflects an initially rapid fall off in correlation, followed by a slower decline. The smallest sampling interval of the GPS drifters in question is 30 minutes and therefore it is safe to assume the errors are uncorrelated for our purposes. Although the drifter sampling rate allows us to avoid further discussion of the autocorrelation function of GPS errors, accounting for autocorrelation is a relatively easy extension (and in fact, already implemented in the code).

The smoothing spline algorithms described in section 3 are modified to use the t-distribution as described in section 8. Table 3 shows that the conclusions reached for Gaussian data in section 3 still apply with t-distributed data.

7. Minimization with Outliers

The goal here is to find a smooth solution in the presence of outliers—points that do not appear to be of the known error distribution for the GPS receiver shown in section a. These points are obviously problematic as can be seen in Fig. 8, where individual data points jump hundreds of meters and even several kilometers away from its neighbors. Errors of this size are inconsistent with the noise analysis of the preceding section, so the goal here is to find a model path $x(t)$ robust to this uncharacterized noise. What makes outliers 'obvious' to the eye is that they appear as unexpectedly large motions, inconsistent with most of the other motion for that path. In this sense, the smoothing spline formulation is a good one as it assumes the motion at some order (e.g., acceleration) is Gaussian, as shown in section b. Interestingly, in the nine drifters we are analyzing here, one drifter shows no obvious outliers, suggesting the issue may be related to how the antenna is configured. This particular drifter serves as a useful point of comparison.

Minimizing with the expected mean square error (19) produces a fit so poor that it is not worth showing. Because outliers add enormous amounts of variance, the expected mean square error vastly under tensions the spline—essentially chasing every outlier shown in Fig. 8. Because some of the noise is uncharacterized, this suggests using a method such as cross-validation might be effective. The orange line in Fig. 8 uses a smoothing spline fit, assuming Student t-distributed errors, but minimized with cross-validation. This fit performs relatively well, but compared with the drifter 7, it is clear that it still chases some outliers. The goal in this section is to develop a method robust to outliers in cases where we know something about the noise.

The basic problem formulation is as follows: we define a new 'robust distribution', p_{robust}, that includes the known noise distribution, p_{noise}, plus an unknown (or assumed) form of an outlier

distribution, p_{outlier},

$$p_{\text{robust}}(\varepsilon) = (1-\alpha)\cdot p_{\text{noise}}(\varepsilon) + \alpha \cdot p_{\text{outlier}}(\varepsilon). \tag{36}$$

We consider a t-distribution for p_{noise} with parameters found from the GPS errors in section a. The distribution of p_{outlier} is also set to be a t-distribution, but with $\nu = 3$ and $\sigma = 50\sigma_{\text{gps}}$ which roughly matches the total variance of the observed outliers. In our tests we varied α from 0 up to 0.25, approximately the range of observed outliers from the drifter data sets.

Throughout our attempts to smooth the noisy GPS data we tried many different approaches to modifying smoothing splines for robustness to outliers, but ultimately found that enormous gains in accuracy are made by simply discarding outliers while minimizing the expected mean square error (19). The results of this approach are shown in section a, but we also document our methodology to reliably estimate the outlier distribution in section b.

a. Robust minimization

The whole problem with outliers is that we do not know their distribution, so minimizing the expected mean square error using (19) with the expected variance from the robust distribution defined in (36) cannot possibly work. Outliers add extra variance, and will therefore cause the spline to be under tensioned (λ_T too small). The key concept behind our method is to simply exclude the outliers from the calculation of (19), where outliers are defined as points unlikely to arise with the known noise distribution. The *ranged expected mean square error* thus replaces σ^2 with,

$$\sigma_\beta^2 = \int_{\text{cdf}^{-1}(\beta/2)}^{\text{cdf}^{-1}(1-\beta/2)} z^2 p_{\text{noise}}(z)\, dz \tag{37}$$

and discards all rows (and columns) of $\mathbf{S}_\lambda$ where $(\mathbf{S}_\lambda - I)\mathbf{x} < \text{cdf}^{-1}(\beta/2)$ or $(\mathbf{S}_\lambda - I)\mathbf{x} > \text{cdf}^{-1}(1-\beta/2)$.

To test this approach we generate data as before, but now also let a certain percentage of outliers (α) be generated with an outlier distribution following (36). We consider five different values of $\beta = \left[\frac{1}{50}, \frac{1}{100}, \frac{1}{200}, \frac{1}{400}, \frac{1}{800}\right]$ as well as $\beta = 0$, which is just (19). Tests across a number of ensembles with outlier ratios $\alpha = [0.0, 0.05, 0.10, 0.25]$ we find that $\beta = \frac{1}{100}$ is overall the best choice.

b. Full tension solution and outlier distribution

The *full tension* solution is defined as the maximum allowable value of λ given the known noise distribution. That is, the spline fit is pulled away from the observations so that the distribution of observed errors $(x_i - x(t_i))$ matches the expected distribution $p_{\text{noise}}(\varepsilon)$. In cases where the effective sample size n_{eff} is large, the full tension solution will approximately match the optimal (minimal mean square error) solution. In cases where the effective sample size is small, the full tension solution is more akin to a low-pass solution (because increasing λ is equivalent to decreasing $x^{(T)}_{\text{rms}}$).

In the simplest case where there are no outliers, the full tension solution can be found by requiring that the sample variance match the variance of $p_{\text{noise}}(\varepsilon)$. When outliers are present, a more robust method of estimation is required. After some experimentation, we found that the most reliable method of achieving full tension is to minimize the Anderson-Darling test of $p_{\text{noise}}(\varepsilon)$ on the interquartile range of observed errors. In fact, we found that this method can be used to estimate the outlier distribution and further refine both the full tension solution and the range over which the expected mean square error is computed.

The outlier distribution is estimated in the following fashion. We first assume that the outlier distribution follows a t-distribution with $\nu = 3$ and that $\alpha < 0.5$. If the spline is in full tension, then the observed total variance can be used to find σ_o for the outlier distribution. From (36) it

follows that,

$$\mathrm{var}_{\mathrm{total}} = (1-\alpha)\mathrm{var}_{\mathrm{noise}} + \alpha 3\sigma_o^2 \tag{38}$$

which, given some α, can be solved for σ_o. Our method considers 100 different values of α logarithmicaly spaced from 0.01 to 0.5 and chooses the value which minimizes the Anderson-Darling test.

With an estimate for $p_{\mathrm{robust}}(\varepsilon)$, the full tension solution can be refined by now minimizing the Anderson-Darling test of $p_{\mathrm{robust}}(\varepsilon)$ on the interquartile range of observed errors. This iterative process converges quite quickly on a good estimate for the outlier distribution and the full tension solution.

c. Extension to bivariate data

The strategies in this section are relatively easily extended to bivariate data. All error distributions are assumed isotropic, and thus the outlier distribution can be estimated by including the errors from both independent directions. The ranged expected mean square error calculation defined in section a uses the *distance* of the error for its cutoff in order to remain invariant under rotation.

Application of this methodology to one of the GPS drifters (drifter 6) is shown in Fig. 8. Although it is impossible to know exactly how well the smoothing spline fit performed, comparison with drifter 7 (with no apparent outliers) suggests that our methodology successfully avoids chasing outliers.

d. Numerical implementation

The `GPSSmoothingSpline` inherits from the `BivariateSmoothingSpline` class and assumes the errors follow the t-distribution found in section a. The class also projects latitude and longitude using a transverse Mercator projection with the central meridian set to the center of the dataset.

8. Conclusions

The methodology manuscript solves our initial problem of finding smoothed, interpolated positions from our noisy GPS drifter dataset with outliers. For signals similar to the Matérn process, we found that

1. the spline degree S should be set to a value higher than the high frequency slope of the process (section 2)

2. the tension degree T can be set to $T = S$ (section 4), and

3. the optimal tension parameter can be estimated *a priori* (also section 4).

For the GPS data in particular, there appear to be three key steps for using smoothing splines to achieve these results:

1. using a t-distribution for the noise (section 6),

2. removing the mean velocity to make the bivariate data stationary (section 5), and

3. using the ranged expected mean square error for robustness to outliers (section 7).

The effective Nyquist identified in section b indicates that the power spectrum for the GPS drifters resulting from the smoothed fit is valid up to about half the nominal sampling rate.

Acknowledgments. Thanks to Miles Sundermeyer whose drifters were used in this analysis. This work was funded by ONR through the Scalable Lateral Mixing and Coherent Turbulence Departmental Research Initiative (LatMix) and National Science Foundation award 1658564.

APPENDIX A

Numerical implementation

The B-splines are generated using the algorithm described in De Boor (1978) with knot points determined by (7) and (8). The matrix $\mathbf{X}$ with components X^i_m denotes the m-th B-spline at time t_i. In this notation the column vector ξ^m represents the coefficients of the splines such that positions at time t_i are given by $\hat{x}^i$ where $\hat{x}^i = X^i_m \xi^m$.

The smoothing spline condition given in (16) can be augmented to include a nonzero mean tension, μ_u,

$$\phi = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - x(t_i)}{\sigma_i}\right)^2 + \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{u(t_q)-\mu_u}{\sigma_u}\right)^2, \tag{A1}$$

where we have taken $T = 1$ for this calculation. The discretized penalty function is

$$\phi = [\mathbf{x} - \mathbf{X}\xi]^{\mathrm{T}}\Sigma^{-1}[\mathbf{x} - \mathbf{X}\xi] + \lambda_1 [\mathbf{V}\xi - \mu]^{\mathrm{T}}[\mathbf{V}\xi - \mu], \tag{A2}$$

where Σ denotes the covariance matrix describing the measurement errors and we absorbed several constants into λ_1. To find the coefficients that minimize this function, we take the derivative with respect to ξ, set it to zero, and solve for ξ,

$$\xi = \left[\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X} + \lambda_1 \mathbf{V}^{\mathrm{T}}\mathbf{V}\right]^{-1}\left[\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \mu\lambda_1 \mathbf{V}^{\mathrm{T}}\iota\right], \tag{A3}$$

where ι is a vector of 1s. The operation $\mathbf{V}^{\mathrm{T}}\iota$ essentially integrates the m-splines and results in a column vector with the integrated values.

We define the smoothing matrix as the linear operator that takes observations $\mathbf{x}$ to their smoothed values $\hat{\mathbf{x}}$,

$$\hat{\mathbf{x}} = \mathbf{S}_\lambda \mathbf{x}. \tag{A4}$$

From this definition and (A3),

$$\mathbf{S}_\lambda \equiv \mathbf{X}\left[\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X} + \lambda_1 \mathbf{V}^{\mathrm{T}}\mathbf{V}\right]^{-1}\mathbf{X}^{\mathrm{T}}\Sigma^{-1}, \tag{A5}$$

when $\mu = 0$.

APPENDIX B

Iteratively reweighted least squares

In practice it is challenging to use the t-distribution directly because it does not result in a linear solution for the coefficients as in (A3). One method around this issue is to use a search algorithm to directly look for the maximum values. Alternatively, one can use the iteratively reweighted least squares (IRLS) method.

The idea with IRLS is to reweight the coefficients of the Gaussian, σ_g in (10), so that the resulting distribution looks like the desired distribution, e.g., (34). Recalling that $\varepsilon_i \equiv x_i - x(t_i, \xi)$, the minimization condition that $\frac{dp_g}{d\xi} = 0$, implies that

$$\frac{\varepsilon_i}{\sigma_g^2}\frac{\partial x(t_i, \mathbf{x})}{\partial \xi} = 0, \tag{B1}$$

for the Gaussian distribution, whereas for the t-distribution this implies that,

$$\frac{\varepsilon_i}{\sigma_s^2}\frac{\nu+1}{\nu}\left(1 + \frac{\varepsilon_i^2}{\nu\sigma_s^2}\right)^{-1}\frac{\partial x(t_i, \mathbf{x})}{\partial \xi} = 0. \tag{B2}$$

This means that one can set

$$\sigma_g^2 = \sigma_s^2 \frac{\nu}{\nu+1}\left(1 + \frac{\varepsilon_i^2}{\nu\sigma_s^2}\right), \tag{B3}$$

to get a matching distribution. Of course, this is only true if ε_i is already known, which initially it is not. So the method becomes iterative—one starts with ε_i determined from the Gaussian fit, then determine a new ε_i after reweighting σ_g. This method iterates until σ_g stops changing. We can rewrite (B3) as a function of ε_i,

$$w_s(\varepsilon_i) = \sigma_s^2 \frac{\nu + \frac{\varepsilon_i^2}{\sigma_s^2}}{\nu + 1}. \tag{B4}$$

From (B4) it is clear that if $\varepsilon_i < \sigma_s$ then it will be reweighted to a smaller value, essentially making the observation point more strongly weighted. On the other hand, if $\varepsilon_i > \sigma_s$, then its relative weighting will decrease, and it will be treated more as an outlier.

More generally, the weight function $w(z)$ for a pdf $p(z)$ is found by setting $-\partial_z \log p(z)$ equal to $-\partial_z \log p_g(z)$ of a Gaussian pdf where $w(z)$ replaces σ_g^2, and then solving for $w(z)$. The result is that,

$$\frac{z}{w(z)} = -\frac{\partial_z p}{p} \quad \Rightarrow \quad w(z) = -z\frac{p}{\partial_z p}. \tag{B5}$$

Note that the same strategy could be used to reshape the pdf of a Gaussian to match the desired distribution, but here we simply match the minimization conditions of the pdfs.

As a point of reference, Tukey's biweight is given by,

$$\psi(z) = \begin{cases} \frac{z}{\sigma_{tb}^2}\left(1 - \frac{z^2}{c^2\sigma_{tb}^2}\right)^2 & |z| < c \cdot \sigma_{tb} \\ 0 & \text{else,} \end{cases} \tag{B6}$$

which, as a weight function is,

$$w_{tb}(\varepsilon_i) = \frac{z}{\psi(z)}. \tag{B7}$$

In a practical sense, the Σ^{-1} of (A5) is replaced with the diagonal matrix $W \equiv \mathrm{diag}(1/w(\varepsilon_i))$ populated with the reweighted values for each observation such that,

$$\mathbf{S}_\lambda \equiv \mathbf{X}\left[\mathbf{X}^{\mathrm{T}} W \mathbf{X} + \lambda_1 \mathbf{V}^{\mathrm{T}}\mathbf{V}\right]^{-1} \mathbf{X}^{\mathrm{T}} W. \tag{B8}$$

This operator is again used to compute the standard error from the variances, $\mathbf{S}_\lambda \Sigma$, where the variance is assumed to be $\sigma_s^2 \frac{\nu}{\nu-2}$ for each observation when using a t-distribution.

The reality is that the smoothing spline solution *does* depend on the initial value of $w(\varepsilon_i)$ used in the IRLS method. That said, we find that for uniform initial weightings (e.g., all values start with the square root of the variance), the differences are not statistically significant from other initial values.

APPENDIX C

Estimating the variance of the signal

The method in this paper depends on good estimates of the root-mean-square velocity, $u_{\rm rms}$, of the signal in order to determine the effective degrees of freedom, as well as the variance of the tensioned derivative. The approach taken here is to compute the power spectrum of the signal at the derivative of interest, and sum the variance that is statistically significantly greater than the expected variance of the noise.

Given a process observed with values x_n at times $t_n = n\Delta$ where $n = 1..N$, we estimate the mean of its m-th derivative by performing a least squares fit to the polynomial $\bar{x}_n \equiv p_m t_n^m + p_{m-1} t_n^{m-1} + .. + p_0$. The *detrended* time series is then defined as $\tilde{x}_n \equiv x_n - \bar{x}_n$. The power spectrum of this time series is given by

$$S_{\rm signal}(f_k) = \frac{\Delta}{N} \left| \sum_{n=0}^{N-1} x_n e^{-2\pi i f_k t_n} \right|^2 , \tag{C1}$$

where the frequencies f_k are given by $f_k = \frac{k}{N\Delta}$. By Plancherel's theorem,

$$\sum_{k=0}^{N-1} S(f_k) \cdot \frac{1}{N\Delta} = \frac{1}{N\Delta} \sum_{i=0}^{N-1} x_i^2 \Delta. \tag{C2}$$

The power spectrum of the m-th derivative of the process is computed as

$$S_{\rm signal}^{(m)}(f_k) = (2\pi f_k)^{2m} \cdot S(f_k). \tag{C3}$$

Note that it is important to detrend the signal prior to computing the derivative because, by assumption, the signal is periodic and has no secular trend.

The noise, ε_i, has total variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2$. Because the noise is assumed to be uncorrelated, the variance distributes evenly across all frequency. The spectrum of the noise is therefore

$$S_{\text{noise}}(f_k) = \sigma^2\Delta, \tag{C4}$$

which immediately can be seen to satisfy Plancherel's theorem (C2). The m-th derivative of the noise has the power spectrum

$$S^{(m)}_{\text{noise}}(f_k) = \sigma^2\Delta(2\pi f_k)^{2m}. \tag{C5}$$

The technique used here sums the variance of the signal for a given frequency if it exceeds the expected variance of the noise at the frequency by some threshold. The estimate of power at each frequency follows a χ^2 distribution with 2 degrees-of-freedom, so we choose the threshold based on the 95-th percentile of the expected distribution. And thus,

$$x^{(m)}_{\text{std}} = \sum_{k=0}^{N-1} S^{(m)}_{\text{signal}}(f_k)\cdot\left(S^{(m)}_{\text{signal}}(f_k) > qS^{(m)}_{\text{noise}}(f_k)\right)\cdot\frac{1}{N\Delta}, \tag{C6}$$

where $q \approx 20$ for the 95-percent confidence.

References

Bracco, A., J. H. LaCasce, and C. Pasquero, 2000: The velocity distribution of barotropic turbulence. *Phys. Fluids*, **12 (10)**, 2478.

Cantoni, E., and T. Hastie, 2002: Degrees-of-freedom tests for smoothing splines. *Biometrika*, **89 (2)**, 251–263.

Craven, P., and G. Wahba, 1979: Smoothing noisy data with spline functions. *Numer. Math.*, **31 (4)**, 377–403.

De Boor, C., 1978: *A practical guide to splines*, Vol. 27. Springer-Verlag New York.

Elipot, S., R. Lumpkin, R. Perez, J. J. Early, and A. M. Sykulski, 2016: A global surface drifter dataset at hourly resolution. *J. Geophys. Res. Oceans*.

Green, P. J., and B. W. Silverman, 1994: *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.

Lee, T. C. M., 2003: Smoothing parameter selection for smoothing splines: a simulation study. *Comput. Stat. Data Anal.*, **42 (1-2)**, 139–148.

Lilly, J. M., A. M. Sykulski, J. J. Early, and S. C. Olhede, 2017: Fractional Brownian motion, the Matérn process, and stochastic modeling of turbulent dispersion. *Nonlin. Processes Geophys.*, **24 (3)**, 481–514.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in C*. 2nd ed., The Art of Scientific Computing, Cambridge University Press.

Reinsch, C. H., 1967: Smoothing by spline functions. *Numer. Math.*, **10 (3)**, 177–183.

Shcherbina, A. Y., and Coauthors, 2015: The LatMix Summer Campaign: Submesoscale Stirring in the Upper Ocean. *Bull. Amer. Meteor. Soc.*, **96 (8)**, 1257–1279.

Sykulski, A. M., S. C. Olhede, J. M. Lilly, and E. Danioux, 2016: Lagrangian time series models for ocean surface drifter trajectories. *J. R. Statist. Soc. C*, **65 (1)**, 29–50.

Teanby, N. A., 2007: Constrained Smoothing of Noisy Data Using Splines in Tension. *Math Geol*, **39 (4)**, 419–434.

WAAS T&E Team, 2016: Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report. Tech. Rep. 92, William J. Hughes Technical Center.

Wahba, G., 1978: Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc. B*.

Whittaker, E. T., 1923: On a New Method of Graduation. *Proceedings of the Edinburgh Mathematical Society*, **41 (01)**, 63–75.

LIST OF TABLES

TABLE 1. 68th percentile range of increase in mean square error from the optimal fit

	T				
S	1	2	3	4	5
1	33.8-80.3%				
2	14.0-75.1%	0.8-12.1%			
3	17.1-77.5%	1.0-13.1%	0.0-4.5%		
4	22.8-81.9%	1.0-14.5%	0.0-4.6%	0.0-6.3%	
5	27.6-91.4%	0.8-15.4%	0.0-4.6%	0.0-6.1%	0.0-12.8%

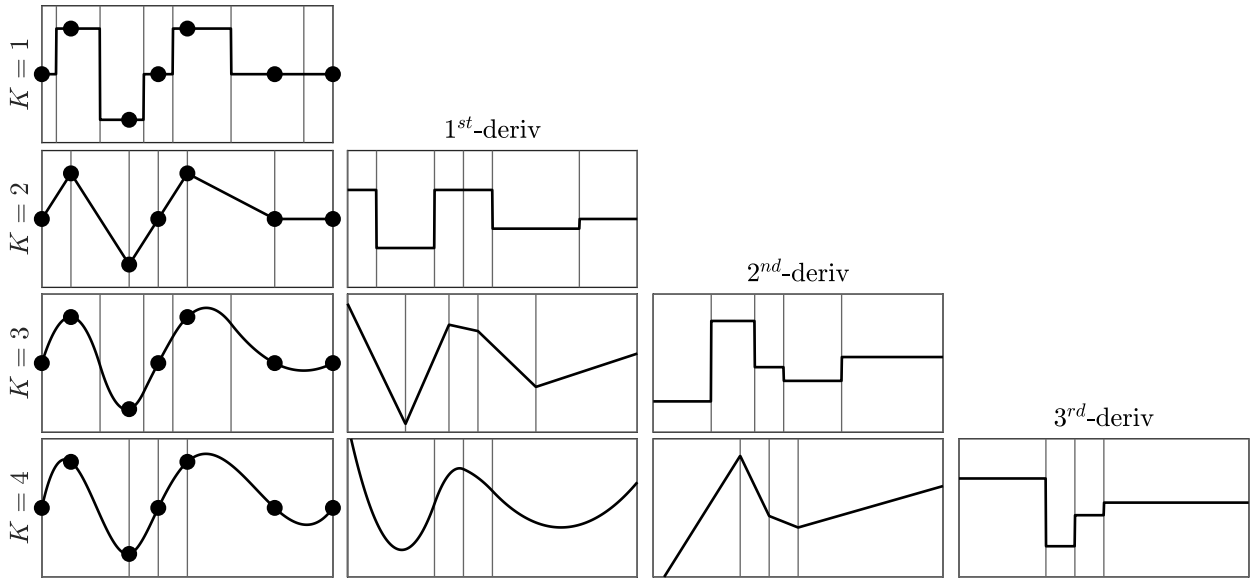
TABLE 2. Mean square error and effective sample size for a range of strides and smoothing spline methods.

stride	n_{eff}	optimal mse	reduced dof	blind initial	expected mse
ω^{-2}					
1	8.6	11.5 m^2	0.1%	56.4%	7.4%
2	4.9	20.4 m^2	0.0%	36.3%	2.8%
4	2.9	34.2 m^2	0.1%	20.0%	1.7%
8	1.7	55.9 m^2	0.0%	5.6%	1.0%
16	1.2	81.8 m^2	0.0%	3.6%	0.5%
ω^{-3}					
1	12.5	7.64 m^2	-0.1%	38.6%	6.4%
2	7.1	13.4 m^2	-0.1%	20.4%	3.5%
4	4.1	23.5 m^2	-0.0%	9.8%	2.2%
8	2.3	41.8 m^2	0.0%	1.7%	1.2%
16	1.4	67.9 m^2	0.0%	9.6%	0.6%
ω^{-4}					
1	15.6	5.69 m^2	-0.1%	33.8%	7.9%
2	9.0	10.5 m^2	-0.1%	18.6%	5.1%
4	5.0	18.6 m^2	-0.0%	8.6%	2.4%
8	2.8	33.2 m^2	0.0%	3.2%	1.5%
16	1.6	57.6 m^2	0.0%	15.4%	0.8%

TABLE 3. Same as table 2, but with noise following a t distribution.

stride	n_{eff}	optimal mse	reduced dof	blind initial	expected mse
ω^{-2}					
1	8.2	11.8 m^2	0.3%	66.7%	7.7%
2	4.7	20.9 m^2	0.3%	47.3%	6.6%
4	2.8	38.0 m^2	0.1%	24.2%	4.4%
8	1.6	66.3 m^2	0.0%	8.2%	9.3%
16	1.2	101. m^2	0.0%	8.1%	3.7%
ω^{-3}					
1	12.1	7.51 m^2	-0.1%	36.2%	8.8%
2	6.8	13.4 m^2	-0.1%	22.8%	7.0%
4	3.9	26.0 m^2	-0.0%	11.5%	3.8%
8	2.2	47.5 m^2	0.0%	2.2%	3.2%
16	1.3	82.5 m^2	0.0%	12.6%	8.5%
ω^{-4}					
1	14.9	6.01 m^2	-0.2%	35.3%	9.0%
2	8.6	10.5 m^2	-0.2%	24.8%	7.0%
4	4.8	19.1 m^2	-0.1%	7.8%	4.6%
8	2.7	36.4 m^2	0.0%	3.2%	2.7%
16	1.6	69.1 m^2	0.0%	18.9%	11.5%

LIST OF FIGURES



730 FIG. 1. An example of interpolating between 7 data points. The data points are shown as circles, and the
731 interpolated function is shown as solid black lines. We show four different orders of interpolation $K = 1..4$
732 (rows) and their nonzero derivatives (columns). The thin vertical grey lines are the knot points.

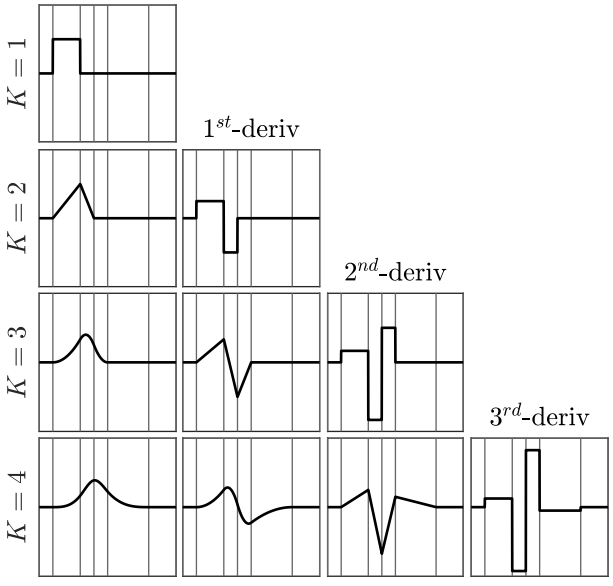


FIG. 2. B-splines and derivatives (columns) for orders $K = 1..4$ (rows).

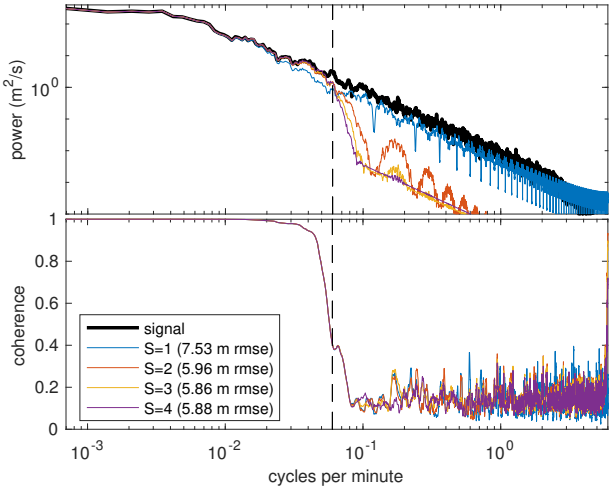


FIG. 3. The upper panel shows the velocity spectrum of the signal (black). The blue, red, and orange lines show the spectrum of the interpolating spline fit to the data with a stride of 100 for $S = 1..4$, respectively. The dashed vertical line denotes the Nyquist frequency of the strided data. The bottom panel shows the coherence between the smoothed signals and the true signal.

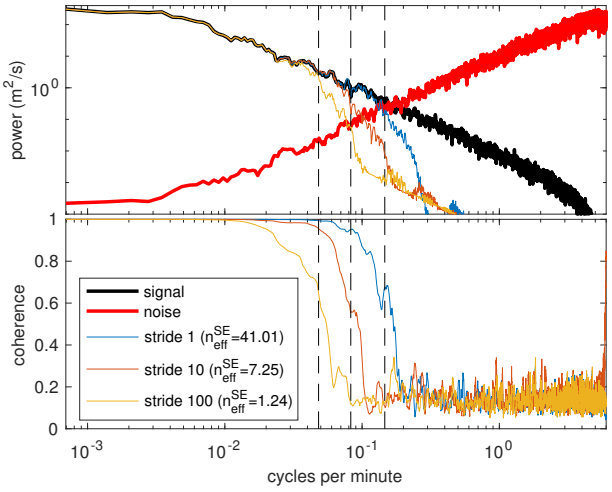


FIG. 4. The upper panel shows the uncontaminated velocity spectrum of the signal (black) and velocity spectrum of the noise (red). The observed signal is the sum of the two. The blue, red, and orange lines show the spectrum of the smoothing spline best fit to the observations with all, 1/10th and 1/100th the data, respectively. The bottom panel shows the coherence between the smoothed signals and the true signal.

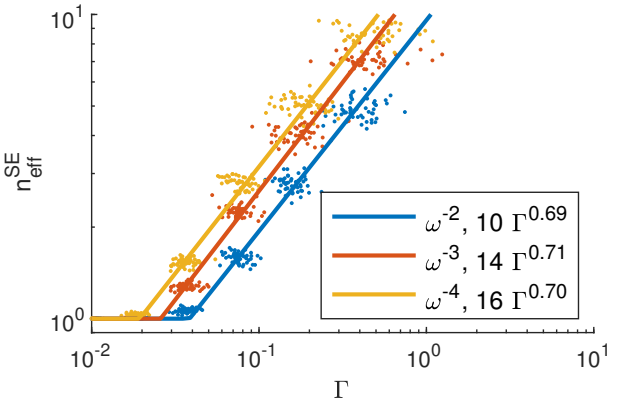
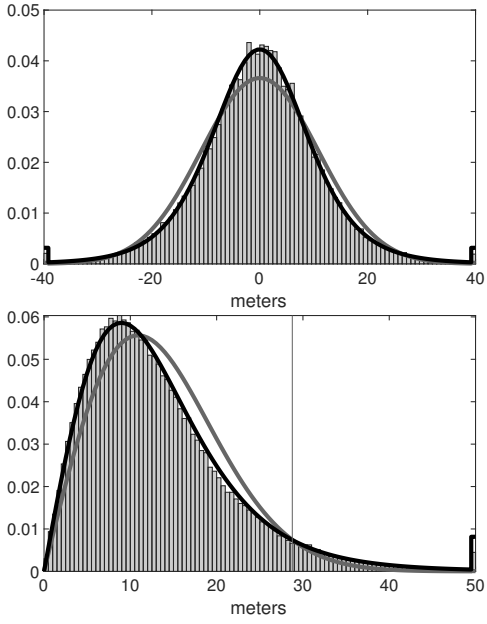


FIG. 5. Effective sample size from the standard error vs Γ



741 FIG. 6. The top panel shows the position error distribution of the motionless GPS. The gray/black lines are
742 the best fit Gaussian/t-distributions respectively. The bottom panel shows the distance error distribution with the
743 corresponding expected distributions from the Gaussian and t-distribution. The vertical line in the bottom panel
744 shows the 95% error of the t-distribution.

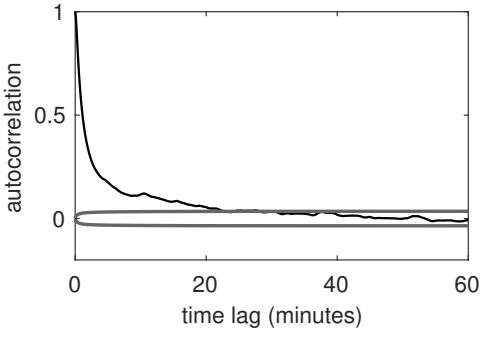


FIG. 7. The autocorrelation function of the GPS positioning error with 99% confidence intervals shown in gray. The correlation at drifter sampling period of 30 minutes is indistinguishable from zero.

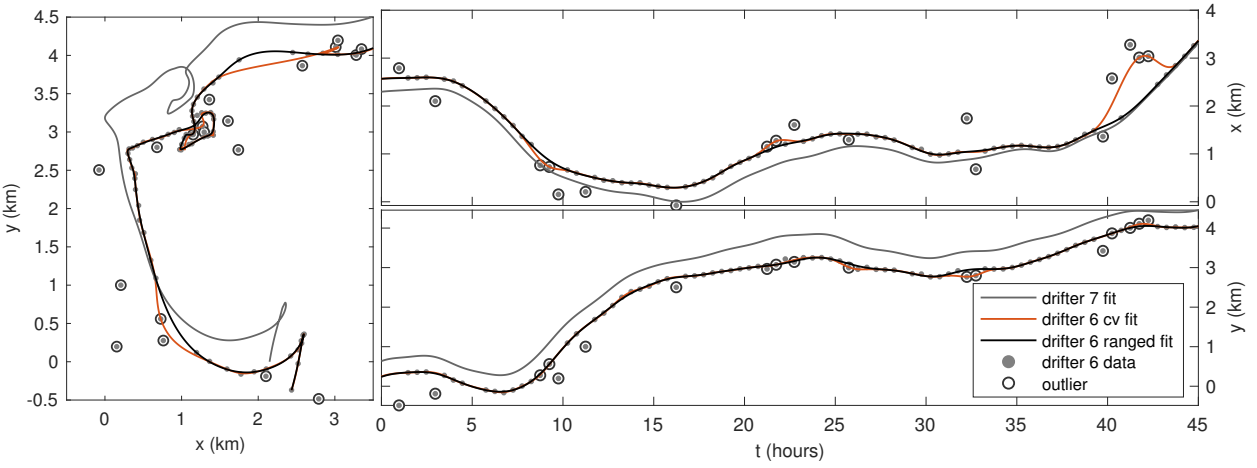


FIG. 8. GPS position data for a 40 hour window from drifter 6. The points are the recorded positions and the black line is the optimal fit using the ranged expected mean square error. Data points with less than 0.01% chance of occurring are highlighted and deemed outliers. The light grey line is the is optimal smoothing spline fit for drifter 7, which has no apparent outliers and was released a few hundred meters from drifter 6. The orange line is the smoothing spline fit assuming t-distributed errors, but using cross-validation to minimize λ_T