


Modeling a recurrent, hidden dynamical system using energy minimization and kernel density estimates

Trevor K. Karn, Steven Petrone, and Christopher Griffin

Applied Research Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

 (Received 9 April 2019; revised manuscript received 19 September 2019; published 29 October 2019)

In this paper we develop a kernel density estimation (KDE) approach to modeling and forecasting recurrent trajectories on a suitable manifold. For the purposes of this paper, a trajectory is a sequence of coordinates in a phase space defined by an underlying hidden dynamical system. Our work is inspired by earlier work on the use of KDE to detect shipping anomalies using high-density, high-quality automated information system data as well as our own earlier work in trajectory modeling. We focus specifically on the sparse, noisy trajectory reconstruction problem in which the data are (i) sparsely sampled and (ii) subject to an imperfect observer that introduces noise. Under certain regularity assumptions, we show that the constructed estimator minimizes a specific energy function defined over the trajectory as the number of samples obtained grows.

DOI: [10.1103/PhysRevE.100.042137](https://doi.org/10.1103/PhysRevE.100.042137)

I. INTRODUCTION

In this paper we propose an algorithm for modeling and forecasting a sparse, noisy, recurrent trajectory that lies entirely on a smooth Riemannian manifold embedded in an arbitrary dimensional Euclidean space. By *sparse*, we mean the signal may be subject to long gaps in observation; by *noisy* we mean the signal is sampled by an (unknown) imperfect observer. We will define *recurrent* precisely in the context of the underlying mathematical model; however, in general we mean the trajectory visits a neighborhood (or collection of neighborhoods) infinitely often. Examples of these trajectories include vehicle (ship, plane, and car) tracks, migration data (e.g., in birds, whales, and sharks), and some economics data subject to seasonality (e.g., detrended annual sales).

Our approach uses a combination of kernel density estimation (KDE) and energy minimizing inference for sparse trajectory reconstruction prior to model learning. Our goal in using a KDE is to construct distribution estimators rather than pointwise estimators with confidence intervals. That is, rather than using a traditional pointwise time-series forecasting method, our objective is to generate a sequence of probability distributions that can be used to generate an optimized pointwise estimator on demand. The methods proposed in this paper will generalize to smooth Riemannian manifolds in arbitrary dimensions; however, we will focus specifically on examples from compact subsets of \mathbb{R}^2 and the 2-sphere \mathbb{S}^2 as a representation of the Earth.

A. Related work

Our work extends and is related to the basic statistical problem of time-series modeling. Linear and nonlinear time-series modeling is a well-established field of statistics [1] and statistical process control [2]. Basic linear regression [3] and nonlinear regression [4] attempt to model observations $\{\mathbf{x}_t\}_{t=1}^N$ as functions of a variable $t \in \mathbb{R}$. In one dimension,

autoregressive integrated moving average models (ARIMA) extend these notions by allowing the model \mathbf{x}_t to vary as a function of past values and past shocks [1]. Seasonal ARIMA models extend this notion by adding seasonal periodicity [5]. Fractional ARIMA [6] models add short and long-range dependence, not expressible with classical ARIMA techniques. In particular, these nonlinear models are better able to express persistence and antipersistence. Finally (generalized) autoregressive heteroskedastic models (ARCH-GARCH) add heteroskedastic behavior to the error components of the time series, allowing globally stationary and locally nonstationary error terms to be analyzed [7]. Many of these techniques can be extended to vector-valued functions (of the type we consider). In particular, vector autoregressive models (VAR and VARIMA) [8] can be used to model time series of vector-valued functions. The most general models are the stochastic differential and difference equations that use Weiner and Lévy processes to model stochasticity [9] (chap. 1). Kernel-based approaches for forecasting stochastic dynamical systems modeled by (hidden) stochastic differential equations are considered extensively by Giannakis *et al.* [10,11]. In particular, in Refs. [10,11] the authors use a diffusion forecasting approach. The shift map of the stochastic process is expressed in a smooth basis of eigenfunctions. This is used to estimate the semigroup solution of the unknown stochastic differential equation without specific parameter estimation.

Grid-based methods that approximate the trajectory can be employed when the space of the time series is continuous but can be partitioned into a collection of discrete grid points and the trajectory modeled as a time series of these grid points. The work in Refs. [12–14] describes methods of using hidden Markov models (and in the case of Ref. [12], dynamic programming) to identify optimal estimators for the behavior of trajectories passing through the discretized state space. Reference [15] uses a multiresolution grid model and a continuous time model to construct a hybrid track estimator that attempts to retain the simplicity of a grid-based model without

sacrificing the accuracy of a continuous model. We note that many (but not all) of the approaches discussed are designed to generate pointwise forecasts with confidence regions, while our objective in using a KDE-based approach is to generate a sequence of probability distributions, which can be used to generate a pointwise forecast.

Forecasting dynamical systems, especially nonlinear dynamical systems, is a well-known problem in physics with applications to noise reduction and experimental data smoothing. Molecular trajectory modeling is considered in Refs. [16,17] using a variational approach with user supplied basis functions. This approach is in contrast to the standard Markov process approaches, which are more reminiscent of grid-based methods, already discussed. References [18–20] consider noise reduction in dynamical systems with Ref. [20] providing a fitting approach that is qualitatively similar to the work presented in this paper. Anomaly detection is considered in Ref. [21] with stated goals similar to those in Ref. [22] but applied to one-dimensional chaotic signals. Forecasting and nonlinear modeling is considered in Refs. [10,11,23–26]. In addition to this work, Ref. [27] applies stochastic hidden Markov models to fuzzy time-series forecasting. Fuzzy time-series forecasting is also considered in Ref. [28]. Reference [29] considers the problem of nonuniform state space reconstruction of chaotic time series. Chaotic time-series forecasting is also considered in Ref. [30], which uses an ant colony optimization algorithm to optimally embed a time series in an appropriate space. Joint continuous and discrete forecasting is considered in Ref. [31], while outlier detection of time series is considered in Ref. [32]. More recent work has applied multilayer perceptron neural networks to time-series forecasting [33,34].

Using a KDE for the purpose of modeling and forecasting recurrent trajectories has been studied by other authors in more restricted contexts. Pallotta, Vespe, and Bryan [22] use a kernel density estimation technique to model shipping routes using automatic identification system (AIS) data. They use the resulting distributions to identify anomalous behavior in ship routes. As they note, AIS data are exceptionally dense and can be used in real time to track ships.

Additionally, it is well known that a kernel density estimate can be used as a convolutional filter on noisy data. This was done in Ref. [35] in order to visualize streaming data from an aircraft. This is a trajectory in \mathbb{R}^3 , although Ref. [35] only considers the projection onto \mathbb{R}^2 . In both Refs. [22,35], the data are highly dense with minimal noise. This is not realistic in antagonistic situations or in cases where the trajectory cannot be observed with high fidelity. This occurs naturally when biologists observe animals in their natural habitat (e.g., see Ref. [36]). This paper considers situations in which the sampled trajectory is neither dense nor exhibits high signal-to-noise ratio. We contrast this to the work in [10,11], where the data are assumed to be more dense.

B. Paper organization

The remainder of this paper is organized as follows: In Sec. II we introduce notation and the underlying mathematical model to be used throughout the rest of the paper. In Sec. III we discuss our proposed modeling and forecasting algorithms.

Theoretical results on the algorithms are provided in Sec. IV. We present empirical results using synthetic and real-world data sets in Sec. V. Finally conclusions and future directions are presented in Sec. VI.

II. NOTATION AND PRELIMINARIES

A. Notation and assumptions

Let \mathbb{R} denote the real numbers, and $\mathbb{R}^+ = \{x \in \mathbb{R} : x \geq 0\}$. Let $(M, \mathbb{R}^+, \varphi)$ be a (hidden) dynamical system describing the motion over time (\mathbb{R}^+) of an (autonomous) particle on a d -dimensional, smooth Riemannian manifold M . The manifold M may be embedded in Euclidean space of dimension $m \leq 2d$, and such an embedding is guaranteed to exist by Whitney's strong embedding theorem. Throughout this paper, bold symbols will indicate positions on the manifold in an appropriate coordinate system; e.g., $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x} = \langle x_1, \dots, x_d \rangle$. We will often (without explicitly stating) identify the set of points in M with their image in \mathbb{R}^d under an appropriate chart.

If M is known, then our approach may be taken using M itself, e.g., using the KDE theory developed in Ref. [37]. In the case M is unknown, but the image of the embedding $M \hookrightarrow \mathbb{R}^m$ is known, our approach may be used taking \mathbb{R}^m to be the manifold of interest (even though the data may be drawn from a different underlying manifold). Determining M given data in \mathbb{R}^m is the fundamental problem in topological data analysis [38,39] and will not be explored further here.

Since our manifold is Riemannian, it may be endowed with an appropriate metric. For example, when $M \equiv \mathbb{R}^d$, the standard Euclidean metric is used; when $M \equiv \mathbb{S}^2$ (the 2-sphere), the Haversine metric is applicable. We denote distance between two points \mathbf{x}, \mathbf{y} in M as $d(\mathbf{x}, \mathbf{y})$. Again by our assumption of a Riemannian manifold, we have the existence of an inner product (positive-definite metric tensor), denoted $\langle \mathbf{x}, \mathbf{y} \rangle$. This should not be confused with a two-dimensional vector as the entries are vectors in bold rather than coordinates in standard typeface. We will use the inner product to quantify the degree to which (e.g., velocity) vectors located at \mathbf{x}, \mathbf{y} have similar heading. In Euclidean space, we would choose the usual dot product. Throughout the paper, we use \triangleq to denote equality by definition rather than derivation.

The dynamical system we study is hybrid in the following sense: Fix a finite set $\mathcal{O} \subseteq M$. At any time t , either:

(1) There are positions $\mathbf{x}_0, \mathbf{x}_f \in \mathcal{O}$ and the function φ defines a subtrajectory $\mathbf{x}_t = \varphi_{\mathbf{x}_0}(t - t_0)$ in M so that

$$\begin{aligned} \varphi_{\mathbf{x}_0}(t - t_0) &\triangleq \arg \min_{\varphi} \int_{t_0}^{t_f} \mathcal{L}(\varphi, \dot{\varphi}, t) dt, \\ s.t. \mathbf{g}_t(\varphi, \dot{\varphi}) &\leq \mathbf{0}, \\ \varphi_{\mathbf{x}_0}(t_0) &= \mathbf{x}_0, \\ \varphi_{\mathbf{x}_0}(t_f - t_0) &= \mathbf{x}_f. \end{aligned} \tag{1}$$

where $\mathcal{L} : (\varphi, \dot{\varphi}, t) \mapsto r \in \mathbb{R}$ is a hidden energy function (Lagrangian) and $\mathbf{g}_t : (\varphi, \dot{\varphi}) \mapsto \mathbf{b} \in \mathbb{R}^m$ are hidden (possibly time parameterized) constraints.

(2) We have $\varphi(t) = \mathbf{x}_0 \in \mathcal{O}$. At some time $t + \tau$, a new $\mathbf{x}_f \in \mathcal{O}$ is chosen (possibly at random).

We assume the dynamical system is recurrent in the sense that the choice of \mathcal{O} is governed by an ergodic or periodic (hidden)

Markov chain with no transient states. Therefore, if $\varphi(t) = \mathbf{x}_0$, there is some $T < \infty$ so that $\varphi(t + T) = \mathbf{x}_0$.

In the problem we discuss, all relevant information about the dynamics, including θ , the Lagrangian \mathcal{L} and some (perhaps all) of the constraints \mathbf{g}_i are hidden. The assumption that $\varphi(t)$ is constructed from piecewise optimal paths is used to justify our method of inferring missing information in sparsely sampled data. For simplicity, in the remainder of this paper, we will assume that $\mathbf{g}_i(\varphi, \dot{\varphi})$ are time invariant and denote the constraint functions by \mathbf{g} . In the sequel, we assume data are sampled discretely via a sampling function $\eta : M \rightarrow \mathbb{R}^d$ (or $\eta : M \rightarrow \mathbb{R}^m$ as appropriate) and with unbiased noise to produce a sparse noisy signal:

$$\mathbf{x}_i = \eta(\varphi(t_i)) + \boldsymbol{\epsilon}_i, \tag{2}$$

here the $\boldsymbol{\epsilon}_i$ are unbiased noise vectors of appropriate dimension. In the sequel, we will elide the observation function η for the sake of clarity and identify $\varphi(t_i)$ with $\eta \circ \varphi(t_i)$. Whether we are using φ to mean a trajectory on M or its image in \mathbb{R}^d or \mathbb{R}^m will be clear from context.

We note, our approach is a parameter-free approximation method and our focus is *not* on estimating the distribution that describes $\boldsymbol{\epsilon}_i$, unlike, e.g., in the traditional Kalman filter estimation (see Ref. [40]).

In addition to the recurrence assumption, we assume that $\varphi(t)$ is piecewise smooth, and in particular at t_0 an instantaneous velocity can be constructed using initial conditions. In practice velocity is numerically approximated by a difference quotient. Finally, since we assume that $\varphi_{\mathbf{x}_0}(t - t_0)$ obeys a set of externally imposed constraints defined by $\mathbf{g}(\varphi, \dot{\varphi})$ in Eq. (1), we assume there is a feasible region $\Omega \subseteq M$ defined by $\mathbf{g}(\varphi, \dot{\varphi})$ and for all t , $\varphi(t) \in \Omega$. In particular, M may be convex as a set in \mathbb{R}^m , but Ω may not be, making the problem more challenging.

B. Techniques

We provide a brief overview of KDE in Euclidean space, which are a foundational element of our proposed algorithm. The interested reader may consult [41] for a more detailed overview of Euclidean KDE methods or Ref. [37] for KDE methods on a Riemannian manifold.

The KDE is a nonparametric estimate of the probability distribution of a set of data $\{x_i\}_{i=1}^N$. Formally, the univariate KDE \hat{f} is given by

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right),$$

with the requirement that the kernel K be a valid probability density function (PDF). Intuitively, one may think of a KDE as a mean of many probability distribution functions with a normalizing constant, $\frac{1}{h}$. The parameter $h > 0$ is referred to as the *bandwidth* and there are many well-established rules of thumb for choosing h given arbitrary data (see, e.g., Ref. [42]). The bandwidth controls the variance of the kernel and acts as a smoothing control on the resulting KDE.

The KDE can be generalized to arbitrary-dimensional Euclidean space. Let $\{\mathbf{x}_i\}_{i=1}^N$ be N data points in \mathbb{R}^m . Define

$\mathbf{x}_i = \langle x_{i_1}, \dots, x_{i_m} \rangle$. The KDE generalizes to:

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh_1 h_2 \dots h_m} \sum_{i=1}^N \left[\prod_{j=1}^m K\left(\frac{x_j - x_{i_j}}{h_j}\right) \right],$$

where K is a PDF defined on \mathbb{R}^d and h_j is the bandwidth parameter for the coordinate x_j . We will write the *bandwidth vector* $\mathbf{h} = \langle h_1, h_2, \dots, h_d \rangle \in (\mathbb{R}^+)^d$. The product defined inside the sum is called the *product kernel*.

1. Choice of kernel

One consideration which must be taken into account when using KDE methods is the choice of kernel. While the only restriction on K is that it is a valid PDF, there are a few canonical choices of kernel. The first, often used in image processing applications as the kernel of a convolutional filter on images, is the Gaussian or normal PDF. Moreover, the Gaussian kernel is used in both Refs. [22,35].

Another canonical choice is the Epanechnikov kernel. Defined as

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases},$$

the Epanechnikov kernel is a compactly supported parabola. In Ref. [43], Epanechnikov shows that the kernel minimizes relative global error in the following sense: Let f be the true distribution and assume that f is analytic. Suppose f is approximated by \hat{f} using Epanechnikov kernel. Then \hat{f} minimizes the functional:

$$\iint_{\mathbb{R}^2} \mathbb{E}[\hat{f}(\mathbf{x}) - f(\mathbf{x})]^2 d\mathbf{x}, \tag{3}$$

where \mathbb{E} is the classical expectation operator. The preceding result is derived using the calculus of variations with constraints and is independent both of distribution (subject to the requirement of analyticity) and constraints on the kernel function.

Equation (3) defines the mean integrated square error (MISE). Therefore, the Epanechnikov kernel minimizes MISE. Scott [41] notes "...the MISE error criterion has two different, though equivalent, interpretations: it is a measure of both the average global error and the accumulated pointwise error." The fact that the Epanechnikov kernel is a MISE minimizer makes it the natural choice of non-Gaussian kernel.

Moreover, the fact that the Epanechnikov kernel has compact support also implies that the inferred distribution has compact support (as the union of finitely many compact sets), which is sometimes desirable. In the context of our work we may implicitly constrain the velocity of the forecast by requiring the KDE have a compact support. The Gaussian kernel does not respect such a constraint, as it gives nonzero probability to the whole ambient space, rather than restricting the nonzero probability to a region that satisfies velocity constraints. As we also show in Sec. IV, a kernel with bounded support can also provide meaningful results regarding feasible regions.

2. KDE in the plane

In \mathbb{R}^2 , which is a sufficient for approximation of the Earth's surface over small regions, we may choose the Epanechnikov product kernel, given by the expression:

$$\hat{f}(x, y) = \frac{9}{16Nh_1h_2} \sum_{i=1}^N \left[1 - \frac{(x - x_i)^2}{h_1^2} \right] \left[1 - \frac{(y - y_i)^2}{h_2^2} \right].$$

This is our choice of kernel for the experimental results of Sec. VB.

3. KDE on the sphere

In Sec. V, we apply the proposed algorithm to cruise ships traveling on the surface of the Earth. While for this paper we consider a small enough region to approximate by the Earth as a plane, on larger scales this may lead to significant distortion. In such a case one might wish to approximate the Earth as a 2-sphere \mathbb{S}^2 . In this case, one could use the Kent distribution, the analog to the Gaussian distribution on a sphere, as the choice of kernel. Let λ be the longitude in degrees, and ϕ the latitude in degrees. The general formulation of the Kent distribution in spherical coordinates is

$$f(\lambda, \phi) = c(\kappa, \beta)^{-1} \exp(\kappa \cos \lambda + \beta \sin^2 \lambda \cos 2\phi).$$

Here κ is a parameter representing the *concentration* of the distribution, β is an analog of covariance, which Kent describes as the ‘‘ovalness,’’ and $c(\kappa, \beta)$ is a normalizing constant given by

$$c(\kappa, \beta) = \int_0^\pi \int_0^{2\pi} \exp(\kappa \cos x + \beta \sin^2 x \cos 2y) \sin x dy dx.$$

We make the simplifying assumption that the covariance of x, y is 0, which is equivalent to setting $\beta = 0$. Then we have $c(\kappa, 0) = 4\pi\kappa^{-1} \sinh \kappa$, simplifying the double integral above. A full description of the Kent distribution can be found in Ref. [44]. It is also possible to find a KDE on other compact Riemannian manifolds without boundary (see Refs. [37,45]). We do not further discuss this case, as Euclidean space is sufficient for our (and most other) applications.

C. Deriving a point estimator and uncertainty regions with a KDE

Given a PDF (potentially a KDE) f on M , there are several ways to find a point estimator. The most obvious method is to compute the argument of expectation, $\mathbf{e} \in M$, such that $f(\mathbf{e}) = \mathbb{E}[f]$. However, if f is multimodal on M and there are constraints on the dynamics [see Eq. (1)], then it is possible $\mathbf{e} \notin \Omega$, where $\Omega \subseteq M$ is the feasible subset of M . In this case it is more useful to compute:

$$\tilde{\mathbf{x}} = \arg \max_{\mathbf{x} \in \Omega} f(\mathbf{x}), \quad (4)$$

as the point estimator. Depending on the numerical complexity, one can also define the conditional distribution $f_{\mathbf{x}|\Omega}$ and compute $\mathbf{e}, \tilde{\mathbf{x}}$ accordingly. We note that Eq. (4) may be a nonconvex optimization problem. In this case, it might be

simpler to compute the unconstrained optimization:

$$\tilde{\mathbf{y}} = \arg \max_{\mathbf{y} \in M} f(\mathbf{y}) \quad (5)$$

and either (i) accept that $\tilde{\mathbf{y}} \notin \Omega$ or (ii) project $\tilde{\mathbf{y}}$ onto the boundary of Ω . We discuss the limits of this approach in Sec. IV.

To find uncertainty regions, we compute the highest-density region, following the Monte Carlo technique established in Ref. [46]. Intuitively, the *highest-density region* (HDR) of a distribution function $f : M \rightarrow \mathbb{R}^+$ is the subset of M corresponding to the preimage of a horizontal ‘‘slice’’ of \mathbb{R}^+ where the slice includes the maximum of the PDF and continues extending down until the probability measure of the preimage of the slice meets a predetermined threshold.

More formally, given a PDF f with support $X \subseteq M$ define:

$$R(c) = \{\mathbf{x} \in X : f(\mathbf{x}) \geq c\}. \quad (6)$$

The $100\% \times (1 - \alpha)$ highest density region is the set $R(c_\alpha)$, where:

$$c_\alpha = \arg \max_c \int \cdots \int_{R(c)} f(\mathbf{x}) d\mathbf{x} \geq (1 - \alpha). \quad (7)$$

It is clear from this definition that $\tilde{\mathbf{y}} \in R(c_\alpha)$ for any $0 \leq \alpha < 1$. If we compute the HDR of the conditional distribution $f_{\mathbf{x}|\Omega}$, then the probability measure of $M \setminus \Omega$ is zero, and so $\tilde{\mathbf{x}} \in R(c_\alpha)$. This is also the case if the unconditional probability given to $M \setminus \Omega$ is sufficiently small.

Hyndman [46] proposes a Monte Carlo algorithm that samples the computed PDF (in our case the KDE) and uses the α quantile as an estimator for c_α . We use this algorithm in the sequel.

Last, since we consider a sequence of KDE's that advance in time, there is an implicit inclusion of the velocity of the object. Thus, we do not need to actually approximate or compute $\dot{\varphi}_{\mathbf{x}_0}(t - t_0)$ explicitly while generating a forecast. This stands in direct contrast to alternate approaches (e.g., Refs. [10,11,16,17]), which view forecasting as finding and solving a system of stochastic differential equations describing the motion of a particle.

III. ALGORITHM DESCRIPTION

In this section we motivate Algorithm 1, which forecasts a finite sequence of triples $\mathcal{F} = \{(\mathbf{p}_i, \hat{f}_i, R_i)\}_{i=N+1}^{N+\bar{q}}$, with $\bar{q} \in \mathbb{N}$, where \mathbf{p}_i estimates the value $\varphi_{\mathbf{x}_0}(t_i - t_0)$ and $\hat{f}_i : M \rightarrow \mathbb{R}$ is a distribution used to construct the HDR R_i that acts as an uncertainty region for \mathbf{p}_i . The algorithm takes as input the observation set $P = \{\mathbf{x}_i\}_{i=1}^N$, a time-indexed sequence of observed positions. Recall from Eq. (2), \mathbf{x}_i is observed with noise. Note that we do not require $t_i - t_{i-1} = t_{i+1} - t_i$ for the points in P .

The algorithm is broken into four stages:

(1) Identify a collection of points

$$H = \{\mathbf{x}_j \in P : j \in J\}$$

similar (defined by the metric and/or inner product on M) to the last known state of the particle. The set J is an index set of consecutive integers which respect the time series P , that is, $i_j < i_{j+1}$ for all j .

Algorithm 1. Forecasting Algorithm

Input: $P = \{\mathbf{x}_i\}_{i=1}^N, \epsilon > 0, \vartheta \in [0, 1], \Delta t, T, \hat{\mathcal{L}}, \mathbf{g}, \alpha \in [0, 1]$
Output: $\mathcal{F} = \{(\mathbf{p}_i, \hat{f}_i, R_i)\}_{i=N+1}^{N+\bar{q}}$
Initialize: $H \leftarrow \emptyset, \mathcal{P} \leftarrow \emptyset, \bar{\mathcal{P}} \leftarrow \emptyset, \bar{q} \leftarrow \lceil T/\Delta t \rceil$

- 1: **for** $\mathbf{x}_i \in P$ **do**
- 2: \triangleright Stage 1: Collect start points.
- 3: **if** $d(\mathbf{x}_i, \mathbf{x}_N) < \epsilon$ **and** $d(\mathbf{x}_{i-1}, \mathbf{x}_N) \geq \epsilon$ **and** $\delta(\mathbf{v}_i, \mathbf{v}_N) < \vartheta$ **and** $t_N - t_i > T$ **than**
- 4: $H \leftarrow H \cup \{\mathbf{x}_i\}$ {Retain index i in H .}
- 5: **end if**
- 6: **end for**
- 7: **for** $x_i \in H$ **do**
- 8: \triangleright Stage 2: Build sample paths.
- 9: $\mathcal{P} \leftarrow \mathcal{P} \cup \{P(\mathbf{x}_i, T)\}$
- 10: **end for**
- 11: **for** $Q \in \mathcal{P}$ **do**
- 12: \triangleright Stage 3: Densify all paths using Eq. (8).
- 13: $\bar{Q} \leftarrow \text{Densify}(Q)$
- 14: $\bar{\mathcal{P}} \leftarrow \bar{\mathcal{P}} \cup \{\bar{Q}\}$
- 15: **end for**
- 16: \triangleright Note: $\bar{\mathcal{P}} = \{\bar{Q}_1, \dots, \bar{Q}_{|H|}\}$. Also, for each $j \in \{1, \dots, |H|\}$, $\bar{Q}_j = \{\bar{\mathbf{x}}_i^{(j)}\}_{i=i_j}^{i_j+\bar{q}}$.
- 17: **for** $i \in \{N+1, \dots, N+\bar{q}\}$ **do**
- 18: \triangleright Stage 4: Compute \hat{f}_i and \mathbf{p}_i .
- 19: $X_i \leftarrow \emptyset$
- 20: **for** $j \in \{1, \dots, |H|\}$ **do**
- 21: $X_i \leftarrow X_i \cup \{\mathbf{x}_{i_j+i-N}^{(j)}\}$
- 22: **end for**
- 23: Compute \hat{f}_i using X_i and a kernel density estimate
- 24: Compute \mathbf{p}_i using Eq. (5)
- 25: Compute R_i using Eq. (6)
- 26: **end for**

(2) Extract subtrajectories of P corresponding to each identified point in H that (i) begin at the identified point and (ii) span an input forecast time. This set of subtrajectories is denoted \mathcal{P} .

(3) Densify the observed subtrajectories in \mathcal{P} to obtain $\bar{\mathcal{P}}$ using a line integral minimization on an estimator $\hat{\mathcal{L}}$ of \mathcal{L} . Each densified trajectory in $\bar{\mathcal{P}}$ is composed of points on M that are equally spaced in time.

(4) Let:

$$\bar{\mathcal{P}} = \{\{\bar{\mathbf{x}}_i^{(j)}\}_{i=i_j}^{i_j+\bar{q}} : j \in \{1, \dots, |H|\}\}$$

be the densified trajectories. For each $i \geq N$ (time index) use the set:

$$X_i = \{\bar{\mathbf{x}}_{i_j+(i-N)}^{(j)} : j \in \{1, \dots, |H|\}, i_j + i \leq N\}$$

to construct a KDE \hat{f}_i . Use the KDE to construct \mathbf{p}_i and an associated HDR representing an uncertainty region.

1. Metrics and tolerances

In Sec. II, we have already defined the distance $d(\mathbf{x}, \mathbf{y})$ and inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ on the manifold M . Given two velocity

vectors \mathbf{v}_i and \mathbf{v}_j , let the angle metric be

$$\delta(\mathbf{v}_i, \mathbf{v}_j) \triangleq 1 - \frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}.$$

This is just the standard cosine distance when $M \equiv \mathbb{R}^d$. In the absence of velocity data, we can use a finite difference method so that, e.g., in \mathbb{R}^d , the velocity vector of the last observed point in P is

$$\mathbf{v}_N \approx \frac{\mathbf{x}_N - \mathbf{x}_{N-1}}{t_N - t_{N-1}}.$$

More generally, if we are working in M , then we may use the language of velocity vectors on smooth manifolds (see, e.g., Ref. [47] or Chapter 3 of Ref. [48]) to approximate \mathbf{v}_N . It is in these details that we wish for M to be smooth. The details of this computation obfuscate the presentation of the proposed algorithm, thus we omit them and refer the readers to the provided references.

As input to Algorithm 1, we take two parameters $\epsilon > 0$, which is a tolerance on $d(\cdot, \cdot)$ and ϑ , which is a tolerance on $\delta(\cdot, \cdot)$. These function as hyperparameters in our proposed algorithm.

2. Forecast duration

Define $P(\mathbf{x}_i, T)$ to be the forward time restriction of P beginning with $\mathbf{x}_i \in P$ and including all points \mathbf{x}_k so that $t_k - t_i \leq T$. That is,

$$P(\mathbf{x}_i, T) = \{\mathbf{x}_k \in P : k \geq i \wedge t_k - t_i \leq T\}.$$

In Algorithm 1, T is the duration of the forecast and is an input parameter.

3. Sampling period

For sparse track reconstruction and forecast generation, we require a sampling frequency. The sampling frequency is a value Δt so that if $\bar{Q} = \{\bar{\mathbf{x}}_i^{(j)}\}_{i=i_j}^{i_j+\bar{q}} \in \bar{\mathcal{P}}$ is a densified trajectory corresponding to some subtrajectory $Q \in \mathcal{P}$, then for all $i \in \{i_j, \dots, i_j + \bar{q}\}$: $t_{i+1} - t_i = \Delta t$, where $\bar{\mathbf{x}}_i^{(j)} \in \bar{Q}$. This sampling period gives resolution to intermediate points of prediction but does not affect predictions made at any given point. It is now straightforward to see that $\bar{q} = \lceil T/\Delta t \rceil$.

4. Track densification

Suppose $\mathbf{x}_i \in H$ and $P(\mathbf{x}_i, T)$ must be densified; i.e., there is some pair $\mathbf{x}_j, \mathbf{x}_{j+1} \in P(\mathbf{x}_i, T)$ so that $t_{j+1} - t_j > \Delta t$. (Note: If $P(\mathbf{x}_i, T)$ is too dense, then it is trivial to downsample it to make it sparser.) If approximations $\hat{\mathcal{L}}$ and $\hat{\mathbf{g}}$ are available, then it is trivial to solve (numerically):

$$\begin{aligned} & \min_{\varphi} \int_{t_j}^{t_{j+1}} \hat{\mathcal{L}}(\varphi, \dot{\varphi}, t) dt, \\ & \text{s.t. } \hat{\mathbf{g}}(\varphi, \dot{\varphi}) \leq \mathbf{0}, \\ & \varphi(t_j) = \mathbf{x}_j, \varphi(t_{j+1}) = \mathbf{x}_{j+1}. \end{aligned} \quad (8)$$

The resulting solution can be used to provide an estimated track of arbitrary density. If $\hat{\mathcal{L}}$ is not already available, then it

is straightforward to define a Gaussian well function:

$$\hat{\mathcal{L}}_G(\mathbf{x}) \triangleq \sum_{j=1}^N \left\{ 1 - \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left[-\frac{d(\mathbf{x}, \mathbf{x}_j)^2}{2\sigma_j^2} \right] \right\} \quad (9)$$

or a least-squares function:

$$\hat{\mathcal{L}}_{LS}(\mathbf{x}) \triangleq \sum_{j=1}^N \frac{1}{\sigma_j} d(\mathbf{x}, \mathbf{x}_j)^2 \quad (10)$$

and solve the constrained line integral minimization problem:

$$\begin{aligned} \min_{\varphi} \int_{\varphi} \hat{\mathcal{L}}_*(\mathbf{x}) d\mathbf{x}, \\ \text{s.t. } \hat{\mathbf{g}}(\varphi, \dot{\varphi}) \leq \mathbf{0}, \\ \varphi(t_j) = \mathbf{x}_j, \varphi(t_{j+1}) = \mathbf{x}_{j+1}. \end{aligned} \quad (11)$$

Here $*$ indicates the choice of Lagrangian. Using Eq. (9) in Eq. (8) causes inferred trajectories to follow historical trends, since the center of the Gaussian well yields the minimal energy, while using Eq. (10) causes the path to minimize the square error. One benefit to Eq. (10) is it has simpler theoretical properties as we show in the sequel. On the other hand, when using Eq. (9), if σ_i is an increasing function of t_i , then this is a continuous variant of pheromone routing [49,50].

Constraint inference is more difficult. In practical situations (e.g., ship tracks) there are obvious constraints in play, like land avoidance (see Sec. V for examples). For the remainder of this paper, we assume that the constraint function \mathbf{g} (or at least $\hat{\mathbf{g}}$, and hence the feasible region Ω) is known and consider constraint estimation as future work. Algorithm 1 shows the pseudocode for the proposed algorithm.

IV. THEORETICAL RESULTS

If the Lagrangian \mathcal{L} is stationary, then we can show that the optimal solution to the problem given in Eq. (11) is asymptotically $\varphi(t)$ when $t \in [t_0, t_f]$ and $\varphi(t) \equiv \varphi_{\mathbf{x}_0}(t - t_0)$ as the sampling rate increases. Assume $\varphi_{\mathbf{x}_0}(t_f - t_0) = \mathbf{x}_f \in \mathcal{O}$. Further assume we have $n \in \mathbb{N}$ observations of the continuous path connecting \mathbf{x}_0 with \mathbf{x}_f denoted as $\{\mathbf{x}^{(i)}(t)\}_{i=1}^n$ with t_i representing the time at which position is observed. We are considering the asymptotic case when the sampling rate is infinite [i.e., $\mathbf{x}^{(i)}$ can be thought of as a function from $[t_0, t_f] \rightarrow M$], so:

$$\mathbf{x}^{(i)}(t) = \varphi_{\mathbf{x}_0}(t - t_0) + \boldsymbol{\epsilon}_t^{(i)} = \varphi(t_i) + \boldsymbol{\epsilon}_t^{(i)}. \quad (12)$$

Assuming we use $\hat{\mathcal{L}}_{LS}$ as our estimation for \mathcal{L} then we solve:

$$\min_{\gamma} \int_{t_0}^{t_f} \sum_{i=1}^n d[\gamma(t), \mathbf{x}^{(i)}(t)]^2 \|\gamma'(t)\| dt,$$

subject to \mathbf{g} . Here $\|\gamma'(t)\|$ accounts for the length of γ on M so that geodesic trajectories are preferred. At any time instant t , the value $\gamma(t)$ minimizing $d(\gamma(t), \mathbf{x}^{(i)}(t))^2$ is

$$\gamma^*(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}(t) = \varphi(t) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{\epsilon}_t^{(i)}$$

from Eq. (12). (To see this, note the integrand is simply the energy function for a mechanical equilibrium point.) We assumed $\boldsymbol{\epsilon}_t^{(i)}$ was unbiased. Therefore, as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{\epsilon}_t^{(i)} \rightarrow \mathbf{0}$$

and $\gamma^*(t) \rightarrow \varphi(t)$. We ignored constraints $\mathbf{g}(\gamma, \dot{\gamma}) \leq \mathbf{0}$ only because we can see that φ must satisfy these constraints and, therefore, asymptotically so will γ . A similar argument can be made for $\hat{\mathcal{L}}_G$, but it is not as clean, due to the additional exponential function in the Gaussian.

Using the above results, we see that the proposed technique for filling in missing information in our discretely sampled noisy signal is (in some sense) an optimal one, assuming a stationary Lagrangian. In the case of nonstationarity, the problem is more difficult, and hence the use of heteroskedastic variances σ_i [see Eq. (9)] related to the time of the observation.

The inferred point predictor given in Eq. (5) is simple to implement but does not take constraints into consideration. Supposing we know the true feasible region Ω , we quantify how far outside Ω a point predictor \mathbf{p}_i could be. This can be used to determine whether it is appropriate to go through the effort of computing Eq. (4) or to simply use Eq. (5).

As before, let $\Omega \subseteq M$ be the feasible region for the trajectory $\varphi_{\mathbf{x}_0}(t - t_0)$. Without loss of generality, assume Ω is a proper subset of M , so that the feasible region is nontrivial. Let $Y = M \setminus \Omega \neq \emptyset$ be the infeasible or *forbidden* region. Denote by $\bar{\cdot}$ the topological closure of a set and denote by ∂ the topological boundary of a set.

We show that feasible regions (and hence forbidden regions) are (partially) inferred as a part of Algorithm 1. To do this, we will use the *Hausdorff distance*, defined on the power set 2^M of M by $\rho : (2^M)^2 \rightarrow \mathbb{R}^+$ by

$$\rho(S_1, S_2) = \inf\{d(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in S_1, \mathbf{y} \in S_2\}.$$

That is, ρ is the smallest distance between points in S_1 and S_2 . When we write ρ with a set and a single point $\mathbf{x} \in M$, we will abuse notation and understand $\rho(S_1, \mathbf{x}) \triangleq \rho(S_1, \{\mathbf{x}\})$ so that the singleton $\{\mathbf{x}\} \in 2^M$.

Let Y be a fixed forbidden region with closure and boundary denoted as above. Assume the prediction point \mathbf{p}_i is computed with the unconstrained Eq. (5) and the Epanechnikov kernel $K(\mathbf{x})$ with bandwidth vector \mathbf{h} . Recall d is the dimension of M , and let $\|\cdot\|$ be the Euclidean metric on \mathbb{R}^d . Then:

$$\mathbf{p}_i \in \Omega \cup \left\{ \mathbf{m} \in M : \rho(\partial\bar{Y}, \mathbf{m}) \leq \|\mathbf{h}\| + \max_{1 \leq i \leq N} \{\|\boldsymbol{\epsilon}_t^{(i)}\|\} \right\}. \quad (13)$$

In other words, the distance from any prediction point to the boundary (of the closure) of the forbidden region is at most the magnitude of the worst-case noise plus the magnitude of the bandwidth $\|\mathbf{h}\|$. If $\mathbf{p}_i \in \Omega$, then this is trivial, so we consider the case when $\mathbf{p}_i \in Y \subseteq M$.

Let $K_{\mathbf{h}}(\mathbf{x})$ be the shifted Epanechnikov product kernel

$$K_{\mathbf{h}}(\langle x_1, x_2, \dots, x_d \rangle) = K \left(\frac{x_1 - x_1^j}{h_1}, \frac{x_2 - x_2^j}{h_2}, \dots, \frac{x_d - x_d^j}{h_d} \right)$$

centered at $\mathbf{x}^{(j)} = \langle x_1^j, x_2^j, \dots, x_d^j \rangle$. Then the support of $K_{\mathbf{h}}(\mathbf{x})$ is the parallelepiped:

$$[x_1^j - h_1, x_1^j + h_1] \times [x_2^j - h_2, x_2^j + h_2] \times \dots \times [x_d^j - h_d, x_d^j + h_d].$$

The support of \hat{f}_i (the i th estimated distribution) is the union of the supports of the individual kernels centered at $\mathbf{x}_i^{(j)}$ for $1 \leq j \leq |H|$, hence there is some point $\mathbf{x} \in \{\mathbf{x}_i^{(j)}\}_{j=1}^{|H|}$ such that $d(\mathbf{x}, \mathbf{p}_i) \leq \|\mathbf{h}\|$. For the right discrete time point t , $\mathbf{x} = \varphi(t) + \epsilon_t$ is perturbed by at most $\max_{1 \leq i \leq N} \{\|\epsilon_t\|\}$, then

$$\mathbf{x} \in \Omega_\epsilon \triangleq \Omega \cup \{\mathbf{m} \in M : \rho(\partial\bar{Y}, \mathbf{m}) \leq \max_{1 \leq i \leq N} \{\|\epsilon_t\|\}\},$$

since $\varphi(t) \in \Omega$.

To maximize

$$\sup_{\mathbf{y} \in (\text{supp}\hat{f}_i) \cap Y} \rho(\Omega_\epsilon, \mathbf{y}),$$

that is, to have conditions which allow \mathbf{p}_i to be as far, away from the boundary of (the closure) of Y , while not being in Ω , we need \mathbf{x} to minimize $\rho(Y \setminus \Omega_\epsilon, \mathbf{x})$. The closest that \mathbf{x} could be to $Y \setminus \Omega_\epsilon$ without being in it is if $\mathbf{x} \in (\partial\bar{Y} \setminus \Omega_\epsilon) \cap \Omega_\epsilon$, the boundary of $\bar{Y} \setminus \Omega_\epsilon$. This, of course, assumes that $Y \setminus \Omega_\epsilon$ is open—if it were closed, then $\rho(Y \setminus \Omega_\epsilon, \mathbf{x})$ is strictly positive and our argument still holds. We now see that $\rho[(\partial\bar{Y} \setminus \Omega_\epsilon), \mathbf{p}_i] \leq \|\mathbf{h}\|$, which implies Eq. (13) via the triangle inequality. Assuming that the noise is sufficiently small to allow all observations to be in the feasible region Ω , then $\|\mathbf{h}\|$ alone serves as an upper bound on the distance inside Y at which \mathbf{p}_i may appear.

It should be noted that the essence of this argument extends to any Kernel whose support is bounded. We also note that if the topological diameter of a forbidden region Y' is smaller than $\|\mathbf{h}\|$, then this property does not prohibit \mathbf{p}_i from being at any point of Y' .

V. EXPERIMENTAL RESULTS

We discuss two sets of experiments to test Algorithm 1. In the first experiment, we forecast two cruise ships over several days (Carnival line’s *Freedom* and *Dream*) to evaluate the performance of Algorithm 1 in a real-world context. In the second experiment, we evaluate the efficacy of Algorithm 1 with two synthetic data sets. Use of a synthetic data set allows us to more closely control the underlying dynamical system and provides a method for exploring potential limitations of the proposed technique.

We use two error metrics to evaluate the algorithm: *absolute pointwise error* (APE), and *percentage in highest density region* (%HDR). Let \mathbf{x}_i be the true position of particle s at time t_i . We create a forecast \mathcal{F} using Algorithm 1 (including information prior to t_i) and obtain prediction point \mathbf{p}_i for time t_i . The APE function is defined as the distance $\text{APE}(t_i) \triangleq d(\mathbf{x}_i, \mathbf{p}_i)$. As noted, we can construct HDR R_i at t_i . Let:

$$\chi_{R_i}(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in R_i \\ 0 & \text{otherwise} \end{cases}$$

be an indicator function, and then define

$$\% \text{HDR}(\mathcal{F}) \triangleq \frac{1}{q} \sum_{i=N+1}^{N+q} \chi_{R_i}(\mathbf{p}_i).$$

APE tells us how far off the pointwise forecast is while %HDR tells if the true position is in the derived uncertainty region. We compute mean and standard deviation of APE for an entire forecast to give a global error metric for the forecast as a whole.

A. Ship track forecasts

Cruise ships exhibit highly recurrent behavior as they travel from port to port, according to a list of destinations which appeal to tourists. Cruise ships also use AIS to give their positions at a high sampling rate with low noise. This makes them excellent subjects on which to test our algorithm, as we can downsample a portion of a known track and generate noise to create training data. After creating a forecast from this sparse, noisy training data, we can then use the remainder of the track as high-resolution ground truth for an error analysis of the forecast.

In order to test our algorithm with this data, we used (approximately) two years of positional data on the Carnival line cruise ships *Dream* (from December 2011–July 2012) and *Freedom* (from March 2012–June 2013). The data was taken from Ref. [51] under fair use. These data were densely sampled, giving a location for each ship on average about once every 15 min. The first 80% of the historical trajectory of each ship was used as the historical data for “training” the KDE model and the last 20% was used as an unseen track on which to test.

We downsampled the training data (the first 80% for each ship) to give one position every day with exactly 24-h intervals, while retaining the resolution of the unseen track. Since there were usually not AIS positions at exactly 24 h of time difference, we linearly interpolated between the nearest known position before the 24-h mark and nearest known position after the 24-h mark. The choice of linear interpolation is “wrong” in the sense that we are working on an oblate spheroid as the manifold but served the purpose of introducing noise into the training data. The result was a sparse noisy track; this was the desired condition for our historical data.

Define the *diameter* of a geographic data set to be the greatest distance between any two points in the data set. To give geographic context to our results, we note that the diameter of *Dream*’s history is 1439.1 nautical miles (NM). Additionally, over the course of the entire history, *Dream* traveled at least 144,234 NM. Similarly the diameter of *Freedom*’s history is 1351.8 NM, and it traveled at least 115,773 NM over its history.

We generated several forecasts with different parameters. In particular, we considered forecast windows of 1 week with 15-min resolution, and with input search radii of 10, 20, and 40 NM. In each case we chose a bandwidth of 1.5 degrees of latitude-longitude or 90 NM. The bandwidth was chosen by trial-and-error to yield a smooth forecast. We consider the problem of automated bandwidth selection as a problem for future work.

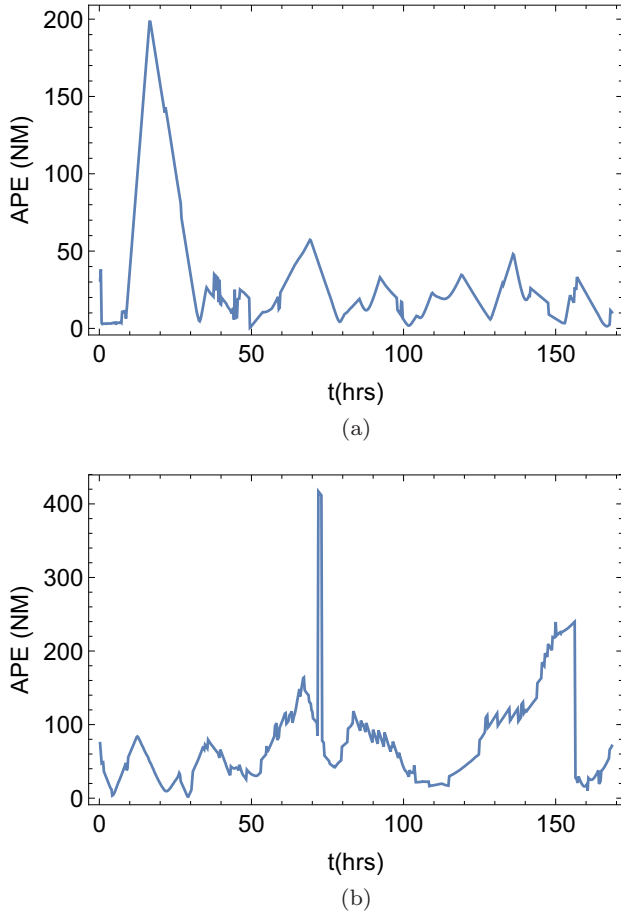


FIG. 1. Error plots of pointwise error at 15-min intervals for a 7-day forecast of *Dream* (a) and *Freedom* (b) with search radii of 10 NM and 40 NM, respectively.

After performing the energy minimization step of Algorithm 1, we have densely sampled paths. Error metrics for forecasting are plotted in Fig. 1. The trajectories of cruise ships change frequently (e.g., as a result of stop over at ports of call). By way of comparison, we note that course change dynamics would have to be known *a priori* when using, e.g., a Kalman filter.

Table I shows summary statistics for the cruise ship forecasting experiments. For *Dream*, as we increase the search radius, we see an increase in the average error and standard deviation of error and a decrease in percent in HDR. For *Freedom*, the average error went down from 10 NM to 20 NM to 40 NM. The standard deviation also only marginally increased and %HDR increased significantly.

It is curious that the examples exhibit different behavior as the search radius grows. One possible explanation for this phenomenon is that for *Dream* with radius 10 NM, the prediction in the first several hours is less than 10 NM, while for *Freedom* with radius 10 NM, there are predictions made in the first hour with error greater than 10 NM. By declaring a search radius we are saying two points are in the same location if they are sufficiently close; i.e., we are creating an equivalence class on the observed data. Thus, we cannot expect our error to be smaller than our search radius. If the error is initially smaller than the search radius, then expanding the search radius would add extra data, which would contribute to a less accurate prediction. On the other hand, if the error is greater than the search radius, then increasing the radius does not add data that is further away than the error, and so it might improve the forecast.

At a higher level, this example provides relevant information about the proposed method. First, there is not necessarily a single best initial search radius—it is context dependent; i.e., a parameter of the model that must be fit. Second, it validates our choice of HDR as uncertainty region, because our worst prediction (*Freedom*, 10 NM), was still within the HDR 71.7% of the time. Moreover, Table I shows that greater error and standard deviation of error corresponds to a lower %HDR. It is true that the HDR depends on the bandwidth parameter, but with a properly chosen bandwidth, we have a reasonable uncertainty region.

Another positive aspect of our forecast is how it treats land. For the most part, the forecast respects the fact that it must remain in the water, and in the few cases where the forecast does go over land, it is only over small islands or tips of peninsulas, which may not be considered by the energy minimization step [52]. This is important to note because we did not give as an input the location of landmasses in the statistical model. Not only does it respect navigational constraints in this manner, but we also see two interesting patterns in the *Freedom* forecast. When the ship goes around the Bahamas, the true track goes west of the islands, while the forecast goes mostly east of the islands. In this sense, we have a valid navigational pattern given by the forecast. A similar occurrence happens as the ship proceeds southeasterly, north of Puerto Rico. The forecast goes south in between Hispaniola and Puerto Rico *avoiding both landmasses*, before proceeding northwest at which point the error goes down to around 50 NM. While the forecast was wrong, it gave valid outputs with all knowledge of land contained in the estimate $\hat{\mathcal{L}}$ and upsampling step, and no knowledge of land in computing \hat{f}_i or \mathbf{p}_i . This indicates if the data were not sparse (e.g., streaming

TABLE I. Table of results for cruise ship data.

Ship	Search radius	Average error (NM)	Standard deviation of error (NM)	% in HDR
<i>Dream</i>	10	35.9	43.9	90.9
<i>Dream</i>	20	65.6	125.5	89.7
<i>Dream</i>	40	74.8	128.5	88.7
<i>Freedom</i>	10	93.9	66.7	71.7
<i>Freedom</i>	20	90.5	68.3	76.3
<i>Freedom</i>	40	85.9	69.9	78.9

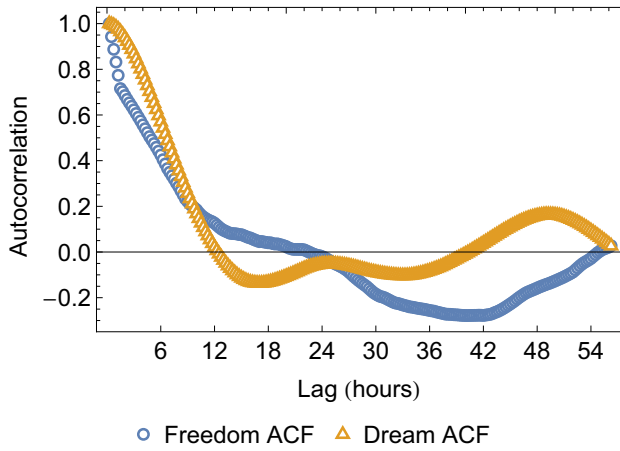


FIG. 2. Autocorrelation function of error, with offset of 24 h. Dashed line indicates critical value for statistical significance at ± 0.140 .

AIS positions), then the forecast would have avoided land without explicitly computing an approximation \mathcal{L} .

Examining the plots in Fig. 1 more closely, we do not see a general trend in error with respect to time. For *Dream* the worst error occurs within the first 2 days of prediction and is almost entirely temporal. For *Freedom* the worst error is in the last 2 days and is almost entirely spatial. A traditional forecaster such as a Kalman or particle filter would be expected to have increasing error over time. Finally, we note that the error seems to be periodic. More specifically, it seems to cycle roughly in relation to days. It is possible that this is an artifact of our daily downsampling while preprocessing the historical data. This claim is supported by computing the autocorrelation function for the error time series, which can be seen in Fig. 2. The plot shows statistically significant autocorrelation around the 48-h period for both forecast, as well as for *Dream* at around the 80- and 125-h marks.

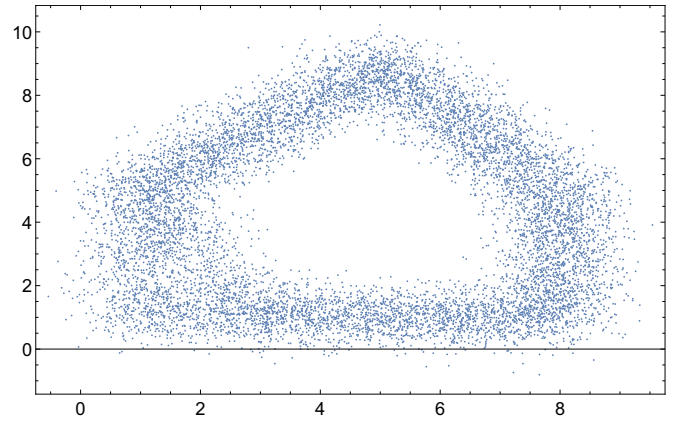
To summarize: Algorithm 1 gives a very reasonable forecast, with only minimal information about the manifold of interest. It respects navigational constraints and does not appear to lose accuracy over time in any general way.

B. Synthetic data forecasts

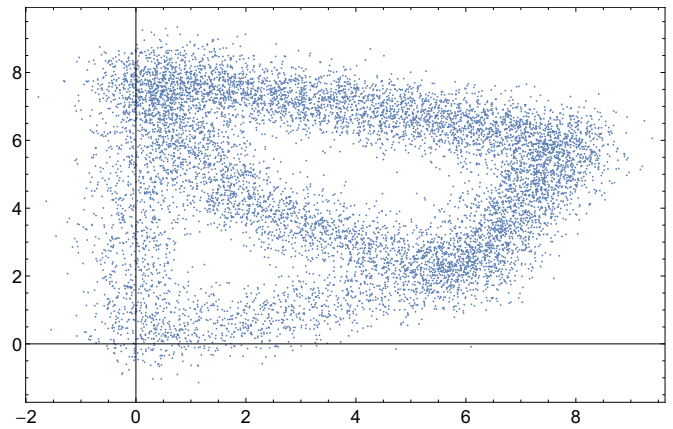
In order to have more control over both the training and testing data set (thereby guaranteeing the data meets our assumptions), we created two synthetic histories of trajectories. They were trajectories on \mathbb{R}^2 , and included loiter points and followed the same underlying model, with a different geometry.

1. Plane trajectories

The two synthetic histories of trajectories on \mathbb{R}^2 consisted of 10 000 data points with Gaussian noise, shown in Figs. 3(a) and 3(b). The generated tracks included six and five loiter points, respectively, including a bifurcating trajectory where, after reaching the top leftmost loiter point the particle uniformly chose to go toward the center of the system or proceed due south. The purpose of the loiter points is to understand how the algorithm treats speed implicitly, while the purpose



(a)



(b)

FIG. 3. Synthetic trajectories moving between loiter points.

of the bifurcation is to see how the forecasting algorithm deals with such phenomena.

The deterministic dynamical system which we used to generate the trajectories has the particle proceed at a constant velocity, in this case 1 spatial unit per time unit. It proceeds on a line from the first loiter point in a list to the second and when it comes within a specified distance of the second, it proceeds to the third, and so on until it has reached the final point in the list. We chose to use 0.1 units for the specified distance. On reaching the final point, the particle proceeds again to the first point in the list (i.e., it is recurrent). At some of the points we provided two options. The option actually taken was chosen uniformly at random from the two. The choice was made each time the particle left the previous loiter point, so we would expect that when two options are given, a trajectory passes through each point roughly half of the time.

For the first synthetic trajectory, the following points served as loiter points:

$$\begin{matrix} (1, 1), & (3, 1), & (8, 1), \\ (8, 5), & (5, 9), & (1, 5). \end{matrix}$$

The trajectory started at (1,1) and proceeded to (8,1). On reaching (1,5), the choice was made (uniformly at random) to proceed to either (1,1) or (3,1).

TABLE II. APE results for forecasts of first and second trajectories.

Time	Prediction error 1st test	In 70% HDR?	Prediction error 2nd test	In 95% HDR?
0.5	0.162085	Yes	0.148339	Yes
1.	0.261795	Yes	0.323111	Yes
1.5	0.268559	Yes	0.194149	Yes
2.	0.280586	Yes	0.279196	Yes
2.5	0.162447	Yes	0.407433	Yes
3.	0.00604629	Yes	0.166779	Yes
3.5	0.223472	Yes	0.32591	Yes
4.	0.284221	Yes	0.26897	Yes
4.5	0.207216	Yes	0.396418	Yes
5.	0.146127	Yes	0.291473	Yes
5.5	0.379875	Yes	0.29377	Yes
6.	0.568606	Yes	0.670763	Yes
6.5	0.543305	Yes	0.529087	Yes
7.	0.492582	Yes	0.230524	Yes
7.5	0.584862	Yes	0.120457	Yes
8.	1.02672	Yes	0.240666	Yes
8.5	0.702077	Yes	0.407801	Yes
9.	0.32379	Yes	0.548738	Yes
9.5	0.48308	Yes	0.292232	Yes
10.	0.402685	Yes	0.374352	Yes
10.5	0.329985	Yes	0.325699	Yes

For the second synthetic trajectory, the following points served as loiter points:

$$(6, 2), (8, 6), (0, 8), \\ (0, 0), (2, 4).$$

The trajectory started at (0,0) and proceeded to (6,2). On reaching (0,8), the choice was made between (0,0) or (2,4) [after which time the particle proceeded to (6,2)].

We recorded the position of the forecast at time intervals of 0.05 units and added mean zero, normally distributed noise with standard deviation 0.5 units. We then generate a mask of 0's and 1's from a binomial distribution with parameter 0.1 in order to retain (on average) 10% of the data. After downsampling the noisy data to roughly 10% of its original density, we used Mathematica's `Interpolate` function in order to linearly interpolate the remaining points. The `Interpolate` function returns a Mathematica interpolating function. We then plugged in the original time points (from $t = 0$ to $t = 500$, sampled at time steps of 0.05 units) to this new interpolating function to return the upsampled points from the downsampled 10% of original data. The interpolating function allows us to upsample and thus have equally (temporally) spaced data while at the same time starting with sparse data. The upsampled data was used to generate KDEs. Since the trajectories were piecewise linear, this is consistent with solving Eq. (11).

The data used to evaluate the performance of our algorithm (i.e., the test trajectories) was generated from the same dynamical system but was neither downsampled to 10% of its original density nor did it have noise added. This is because we do not want the noise of the test path to be a confounding factor when evaluating the pointwise error.

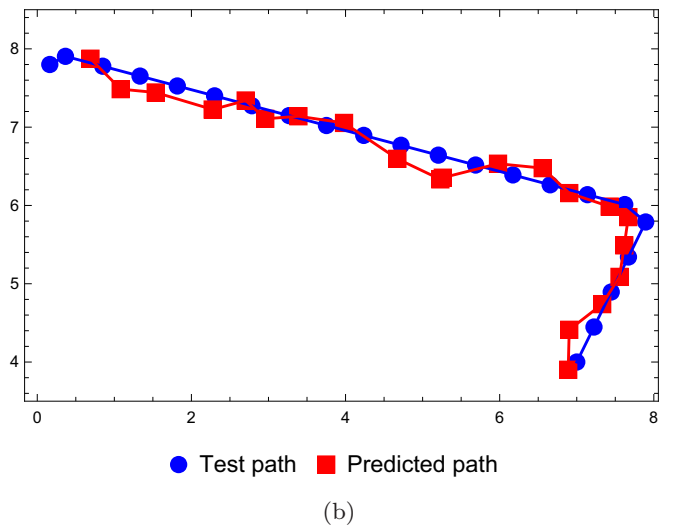
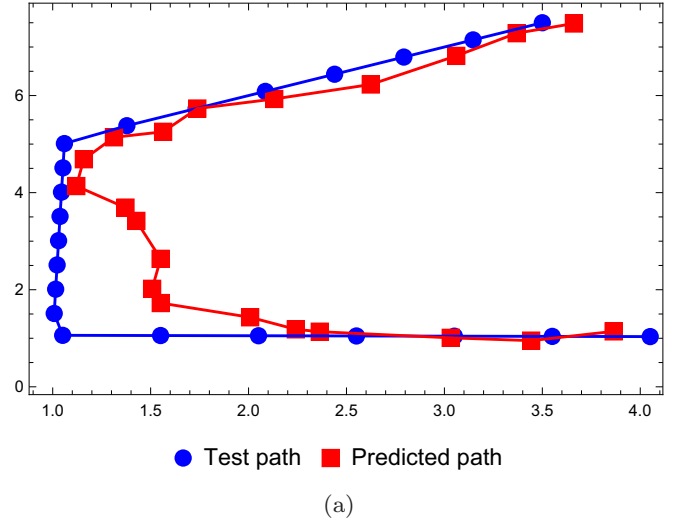


FIG. 4. Synthetic trajectories and forecasts moving between loiter points.

For both unique trajectories, we forecasted a length $\bar{q} = 20$ trajectory $\{\mathbf{x}_i\}_{i=N+1}^{N+\bar{q}}$ for the dynamical system that generated the first $N = 10\,000$ data points, with $\mathbf{x}_{N+1} = (3.5, 7.5)$. In reporting (see Table II) we consider each index to be a half-unit of time, so that \mathbf{x}_{N+1} corresponds to reporting time $t = 0.5$ and $\mathbf{x}_{N+\bar{q}}$ corresponds to time $t = 10.5$. We note that $\{\mathbf{x}_i\}_{i=N+1}^{N+\bar{q}}$ was not used to train the model (in the sense of contributing to the data that was used to build a KDE). These tracks are shown plotted with circles in Figs. 4(a) and 4(b), where the points represent the actual $\{\mathbf{x}_i\}$ and are connected linearly in both space and time to give a position for any $t \in [t_{N+1}, t_{N+\bar{q}}] \subseteq \mathbb{R}^+$.

After generating training data and a ground-truth trajectory for comparison to predictions, we compute a prediction $\{\mathbf{p}_i\}_{i=N+1}^{N+\bar{q}}$. We plot the predictions as squares in Figs. 4(a) and 4(b).

In Fig. 4(a), we see that the overall shape of the prediction and ground truth are roughly similar, and they are extremely similar in Fig. 4(b). In particular, the predicted path and

TABLE III. Error between predicted point at time t and the nearest true point ignoring time for the trial in Fig. 4(b).

Time	Min. Distance 1st Test	1st Test Less than $\epsilon = .25$	Min. Distance 2nd Test	2nd Test Less than $\epsilon = .25$
0.5	0.162	Yes	0.148	Yes
1.	0.262	No	0.311	No
1.5	0.269	No	0.194	Yes
2.	0.281	No	0.158	Yes
2.5	0.162	Yes	0.234	Yes
3.	0.006	Yes	0.167	Yes
3.5	0.223	Yes	0.237	Yes
4.	0.284	No	0.234	Yes
4.5	0.207	Yes	0.24	Yes
5.	0.146	Yes	0.291	No
5.5	0.38	No	0.294	No
6.	0.569	No	0.18	Yes
6.5	0.543	No	0.301	No
7.	0.493	No	0.231	Yes
7.5	0.585	No	0.12	Yes
8.	0.832	No	0.105	Yes
8.5	0.592	No	0.177	Yes
9.	0.232	Yes	0.292	No
9.5	0.207	Yes	0.292	No
10.	0.039	Yes	0.186	Yes

ground truth are very close up through the predictions made after the ground truth leaves the loiter regions at (1,5), and (8,6), respectively.

In Table III we illustrate the distance between every predicted point \mathbf{p}_i and the *closest* point on the true trajectory. This illustrates the impact of speed-estimation error rather than spatial error; i.e., this distance ignores the time at which a forecast expects the particle to be in a certain location.

In Table III and Fig. 5, we see that mean distance between a forecast point and any true point is less than ϵ for 40% of the time in Test 1 and 70% of the time for Test 2. We draw attention to these facts because by choosing ϵ , as we noted, we are essentially constructing an equivalence relation on the data. The larger error in Test 1 is most likely caused by the branching nature of the behavior near the forecast region, causing the center of mass in the KDE to move between

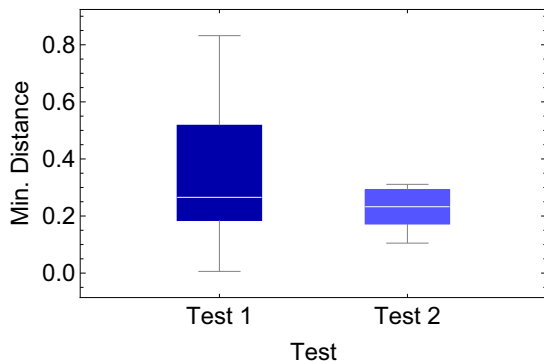


FIG. 5. Box and whisker chart illustrating the properties of the error distribution of data in Table III.

the two possible branches. This can be corrected by using a multihypothesis branching forecast approach [53] that seeks multiple local maxima in the KDE. We note that it is relatively straightforward to construct such a multihypothesis pointwise forecast that would provide branching paths following multiple local maxima of the KDE and this is considered in future work.

Using the membership in the HDR as a measure of performance, we see that it is a good way to measure the temporal accuracy of a forecast but not necessarily the spatial accuracy. This fact is exemplified by trajectory two. We also note that the HDR does not represent a “confidence region” in the frequentist sense of the term but instead has a more Bayesian flavor.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

The previous literature on modeling and forecasting using KDEs has focused on well sampled data. In this paper, we develop Algorithm 1 to generate forecasts for the cases when we have a sparse, noisy data drawn from a recurrent trajectory. This approach offers an alternative to existing forecasting methods in cases when data are sparse and noisy, since it requires significantly fewer data. The energy minimization technique we use allows us to “connect the dots” between existing data points in an intelligent way, so that we can sample data as finely as we wish.

Algorithm 1 has several useful features. Traditional forecasters use some approximation or knowledge of speed, which is then projected forward in time. Our forecasting algorithm treats speed implicitly as we develop time-indexed probability distributions over a smooth Riemannian manifold. This simplifies the computation and makes the algorithm path independent. In practice, this allows us to pick specific times t_i at which to provide a forecast position \mathbf{p}_i and uncertainty function f_i rather than having the requirement of computing intermediate steps to predict the whole path.

The algorithm also has nice theoretical properties. In particular, the optimal solution of a minimal energy trajectory is the asymptotic limit of the prediction as the sampling rate goes to infinity. Moreover, with a reasonable choice of parameters, it respects the existence of forbidden regions, albeit with a “fuzzy” boundary. With reasonable assumptions on (or knowledge of) the noise, we can easily quantify the fuzziness of the boundary.

When we assume but cannot guarantee that the governing energy function is the same, (as in the cruise ship forecasts) we still get reasonably good results. When we *can* guarantee that the underlying energy function is the same (as in our synthetic forecast) the algorithm performs quite well. It gives forecasts, with error that is mostly temporal rather than spatial. When the error is spatial, it is as a result of a valid path which was not taken. In both the spatial and temporal error cases we still have an uncertainty region for each point, given by the highest density region of the KDE \hat{f} , and experimental evidence supports our claim that this makes sense as an uncertainty region.

While the algorithm performs well, there are certainly improvements which can be made. In particular, we note

that often times the error is temporal rather than spatial. That is, the predicted point is not off because the predicted trajectory has high error in space, but rather because the time is off. A next step would be to consider how to improve this weakness in the algorithm. Additionally, for our ship forecast, we heuristically tuned the bandwidth to give a smooth path. It would be desirable to find a nonheuristic way to find the appropriate bandwidth, as the existing bandwidth selection rules (particularly Scott's and Silverman's rules of thumb) give a track with high error. Finally, improving our handling

of bifurcating tracks when constructing \mathbf{p}_i may also improve the resulting forecasts.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research through Naval Sea Systems Command DO 0451 (Task 23875). We thank Brady Bickel, Eric Rothoff, and Douglas Mercer for helpful conversations during the development of this paper.

-
- [1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, New York, 2008).
- [2] E. del Castillo, *Statistical Process Adjustment for Quality Control* (Wiley Interscience, New York, 2002).
- [3] R. Hogg and E. Tanis, *Probability and Statistical Inference*, 7th ed. (Pearson/Prentice-Hall, Upper Saddle River, NJ, 2006).
- [4] G. A. F. Serber and C. J. Wild, *Nonlinear Regression* (John Wiley & Sons, New York, 2003).
- [5] E. Ghysels and D. R. Osborn, *The Econometric Analysis of Seasonal Time Series* (Cambridge University Press, Cambridge, 2001).
- [6] J. R. M. Hosking, *Biometrika* **68**, 163 (1981).
- [7] R. F. Engle, *J. Econ. Perspect.* **15**, 157 (2001).
- [8] R. S. Hacker and A. Hatemi-J, *J. Appl. Stat.* **35**, 601 (2008).
- [9] B. K. Oksendal, *Applied Stochastic Control of Jump Diffusions* (Springer, Berlin, 2007).
- [10] T. Berry, D. Giannakis, and J. Harlim, *Phys. Rev. E* **91**, 032915 (2015).
- [11] D. Giannakis, *Appl. Comput. Harm. Anal.* **47**, 338 (2019).
- [12] F. Martinerie and P. Forster, in *Proceedings of the ICASSP Conference* (IEEE, San Francisco, CA, USA, 1992).
- [13] L. R. Rabiner, *Proc. IEEE* **77**, 257 (1989).
- [14] R. L. Streitt and R. F. Barrett, *IEEE Trans. Acoust. Speech Sign Process.* **38**, 586 (1990).
- [15] C. Griffin, R. R. Brooks, and J. Schwiier, *IEEE Trans. Autom. Contr.* **56**, 1926 (2011).
- [16] F. Nüske, R. Schneider, F. Vitalini, and F. Noé, *J. Chem. Phys.* **144**, 054105 (2016).
- [17] F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, *J. Chem. Theory Comput.* **10**, 1739 (2014).
- [18] E. J. Kostelich and J. A. Yorke, *Phys. Rev. A* **38**, 1649 (1988).
- [19] E. J. Kostelich and T. Schreiber, *Phys. Rev. E* **48**, 1752 (1993).
- [20] T. Schreiber, *Phys. Rev. E* **47**, 2401 (1993).
- [21] Y. Cao, W.-w. Tung, J. B. Gao, V. A. Protopopescu, and L. M. Hively, *Phys. Rev. E* **70**, 046217 (2004).
- [22] G. Pallotta, M. Vespe, and K. Bryan, *Entropy* **15**, 2218 (2013).
- [23] P. E. McSharry and L. A. Smith, *Phys. Rev. Lett.* **83**, 4285 (1999).
- [24] H. U. Voss, *Phys. Rev. Lett.* **87**, 014102 (2001).
- [25] C. M. Danforth and J. A. Yorke, *Phys. Rev. Lett.* **96**, 144102 (2006).
- [26] H. L. D. de S. Cavalcante, M. Oriá, D. Sornette, E. Ott, and D. J. Gauthier, *Phys. Rev. Lett.* **111**, 198701 (2013).
- [27] S. Li and Y. Cheng, *IEEE Trans. Syst. Man Cybernet. B* **40**, 1255 (2010).
- [28] K. Huarng, T. H. Yu, and Y. W. Hsu, *IEEE Trans. Syst. Man Cybernet. B* **37**, 836 (2007).
- [29] M. Han, W. Ren, M. Xu, and T. Qiu, *IEEE Trans. Cybernet.* **49**, 1885 (2019).
- [30] M. Shen, W. Chen, J. Zhang, H. S. Chung, and O. Kaynak, *IEEE Trans. Cybernet.* **43**, 790 (2013).
- [31] E. Ramasso, M. Rombaut, and N. Zerhouni, *IEEE Transact. Cybernet.* **43**, 37 (2013).
- [32] F. Rasheed and R. Alhaji, *IEEE Trans. Cybernet.* **44**, 569 (2014).
- [33] R. Rico-Martínez, K. Krischer, I. Kevrikidids, M. Kube, and J. Hudson, *Chem. Eng. Commun.* **118**, 25 (1992).
- [34] M. M. Rahman, M. M. Islam, K. Murase, and X. Yao, *IEEE Trans. Cybernet.* **46**, 270 (2016).
- [35] O. D. Lampe and H. Hauser, in *Proceedings of the 2011 IEEE Pacific Visualization Symposium* (IEEE, Piscataway, NJ, USA, 2011), pp. 171–178.
- [36] Oearch (2018) [<https://www.oearch.org/tracker/>].
- [37] A. le Brigant and S. Puechmorel, *Entropy* **21** (2019).
- [38] R. Ghrist, *Bull. Am. Math. Soc.* **45**, 61 (2007).
- [39] G. Carlsson, *Bull. Am. Math. Soc.* **46**, 255 (2009).
- [40] R. E. Kalman, *Trans. ASME J. Basic Eng.* **82**, 35 (1960).
- [41] D. W. Scott, *Multivariate Density Estimation*, 2nd ed., Wiley Series in Probability and Statistics (John Wiley & Sons, Inc., Hoboken, NJ, 2015).
- [42] M. C. Jones, J. S. Marron, and S. J. Sheather, *J. Am. Stat. Assoc.* **91**, 401 (1996).
- [43] V. Epanechnikov, *Theory Probab. Appl.* **14**, 153 (1969).
- [44] J. T. Kent, *J. Roy. Stat. Soc. Ser. B* **44**, 71 (1982).
- [45] B. Pelletier, *Statist. Probab. Lett.* **73**, 297 (2005).
- [46] R. J. Hyndman, *Am. Stat.* **50**, 120 (1996).
- [47] N. Boumal, Discrete curve fitting on manifolds, Master's thesis, Université catholique de Louvain, 2010.
- [48] J. M. Lee, *Introduction to Smooth Manifolds*, 2nd ed., Graduate Texts in Mathematics, Vol. 218 (Springer, New York, 2013).
- [49] K. M. Sim and W. H. Sun, *IEEE Trans. Syst. Man Cybernet. A* **33**, 560 (2003).
- [50] X. Hu, J. Zhang, H. S. Chung, Y. Li, and O. Liu, *IEEE Trans. Syst. Man Cybernet. B* **40**, 1555 (2010).
- [51] Live Marine Information [sailwx.info].
- [52] This step was carried out using a gridded representation of the Earth using a numerical approximation of the line integral.
- [53] C. Griffin, R. R. Brooks, and J. Schwiier, in *Distributed Sensor Networks: Sensor Networking and Applications (Volume Two)*, Vol. 2, edited by S. S. Iyengar and R. R. Brooks (CRC Press, Boca Raton, FL, 2016), Chap. 39.