A Machine Learning Approach to the Prediction of Tidal Currents.

Dripta Sarkar, Michael Osborne, and Thomas Adcock

Department of Engineering Sciences, University of Oxford, Parks Road, Oxford OX13PJ, United Kingdom

ABSTRACT

We propose the use of techniques from Machine Learning for the prediction of tidal currents. The classical methodology of harmonic analysis is widely used in the prediction of tidal currents and computer algorithms based on the method have been used for decades for the purpose. The approach determines parameters by minimizing the difference between the raw data and model output using the least squares optimization approach. However, although the approach is considered to be state-of-the-art, it possesses several drawbacks that can lead to significant prediction errors, especially at locations of fast tidal currents and 'noisy' tidal signal. In general, careful selection of tidal constituents is required in order to achieve good predictions, and the underlying assumption of stationarity in time can restrict the applicability of the method to particular situations. There is a need for principled approaches which can handle uncertainty and accommodate noise in the data. In this work, we use Gaussian process, a Bayesian non-parametric technique, to predict tidal currents. The overall objective is to take advantage of the recent progress in machine learning to construct a robust yet efficient algorithm. The development can specifically benefit the tidal energy community, aiming to harness energy from location of fast tidal currents.

INTRODUCTION

Tidal waves are produced by a combination of the gravitational forces of the sun and the moon. Prediction of tidal currents are necessitated by practical requirements like navigation, protection from flooding, coastal management to recent developments of energy extraction. The earliest known tidal predictions dates back to the eleventh century of that of a bore on the Qiantang river in China. There have been many advances in methodologies for tidal analysis since then. The most widely accepted and used method is that of the harmonic analysis, where the observed tidal variations are considered as a resultant of various periodic components of known frequencies, with the amplitudes and phases determined using the least-squares fitting procedure. Computer codes based on harmonic analysis have been used for decades for the prediction of tidal heights (1-D) and currents (2-D). The first codes were based on uniformly sampled raw data, linearized times for nodal corrections. Over the years various advances have been made to harmonic analysis approach and Foreman et. al. (2009) recently incorporated the nodal and astronomical arguments at exact times and inferences directly in the matrix formulation before the least squares optimization is performed. As such, the corrections and the inference directly influence all the constituents rather than any specific reference one. Most cases of the tidal current analysis assumes that the process is stationary in time. Investigations of nearly stationary data records provide valuable understanding of the tidal dynamics and the harmonic analysis approach is widely used for the purpose of predictions. However, there are several shortcomings of this methodology. Selection of appropriate tidal constituents is a big challenge, the mis-specification of which may lead to overfitting of data or can compromise with the numerical stability of the matrix inversion required to obtain the coefficients, if the frequencies are considered at locations where energy is absent. Appropriate modelling of noise is another issue. In tidal analysis, signals which do not contribute to the tidal variations are classified as 'noise'. In reality, there can be cases where the non-tidal signal is much stronger than the tidal e.g. the occurrence of a stormy event. It is difficult to incorporate such effects in the tidal harmonic analysis formulation. Another aspect of the tidal harmonic analysis tools is that the confidence intervals are generated for the current ellipse parameters (the amplitude and Greenwich phase) through a non-linear mapping from the model parameters. However, in a lot of practical applications it is more useful to generate confidence interval estimates directly in the time domain.

In this work, we present an alternate approach to predict tidal currents using probabilistic machine learning techniques for time-series analysis, which brings robust, stable, computationally practical and principled approaches for handling uncertainty, and can naturally handle the challenges of real world data (Roberts et al.; 2013). Data analysis has been performed with the aid of principled Bayesian approaches in the fields of geostatistics where it is known as kriging (Matheron; 1973), mete-

orology (Thompson; 1956), spatial statistics (Ripley; 2005), machine learning (Rasmussen and Williams; 2006). Gaussian processes, a Bayesian non-parametric approach, have been shown to be well-suited in solving a variety of time-series modelling problems (Roberts et al.; 2013) and in this work we pursue this methodology to model tidal current data. In the next sections, brief introductions to tidal harmonic analysis and Gaussian process regression are provided, followed by results and discussion.

TIDES AND HARMONIC ANALYSIS

Theory based on potential fields, which gives the forces due to the tidal generating bodies (the sun and the moon), predict the existence of hundreds of unique tidal constituents (frequencies). Each of such constituent can be expressed as a linear combination of the rate of change

$\tau =$ lunar time (24.8 hr)
$s =$ mean longitude of the moon (27.32 day),
$h =$ mean longitude of the sun (365.24 day)
$p =$ longitude of the perigee (8.85 yr)
$N' =$ negative longitude of the ascending node (18.61 yr)
$p' =$ longitude of the perihelion (21000 yr)

where the mean longitude is the ecliptic longitude of the orbiting body if the orbit is circular, perigee is the point in the orbit of the moon at which it is nearest to the earth, perihelion is the point in the orbit of the earth at which it is closest to the sun, ascending node is the point at which an orbit crosses the elliptic plane going north. The effect of the perihelion (p') is usually ignored in all computations as it is almost constant over historical time (Pawlowicz et al.; 2002). The lunar declination is governed by the plane of motion of the moon which is inclined at an angle of $5°09'$ to the plane of the ecliptic and the inclination of this plane changes with a period of 18.61 years, with the ascending node performing one complete backward motion along the ecliptic over this period of time. The tidal constituents depending on the lunar declination have a pronounced 18.61 year modulation. Nonlinear interaction of the astronomical tidal components produces secondary tides known as overtides or compound tides. And because nonlinear interactions are predominantly produced in the shallow water region, they are often referred to as shallow water tides.
The expression for the potential due to a tidal generating body (which can be the sun, moon) at a location on Earth with longitude λ and latitude θ is given by

$$V(\lambda,\theta) = \frac{3GM\rho^2}{4R_0^3}\Big(\frac{R_0}{R}\Big)^3\Bigg[\frac{4}{3}\Big(\frac{1}{2}-\frac{3}{2}\sin^2\theta\Big)\Big(\frac{1}{2}-\frac{3}{2}\sin^2\delta\Big) + \sin 2\theta \sin 2\delta \cos H + \cos^2\theta\cos^2\delta\cos 2H\Bigg] \quad (3.1)$$

where δ is the declination of the tidal generating body, $H = \omega_0 t + \lambda - A$ is referred to as the hour angle, A is the right ascension of the tidal generating body or the apparent longitude with respect to Υ, the celestial reference point/origin. The coefficient $3GM\rho^2/4R^3$ is referred to as the Doodson constant. On expansion of the trigonometric terms, the potential in (3.1) can be expressed as

$$V(\lambda,\theta) = \sum_{i=1}^{3} V_i(\lambda,\theta) \quad (3.2)$$

where

$$V_i(\lambda,\theta) = DG_i \sum_j C_j \cos(\sigma_j t + i\lambda + \theta_j) \quad (3.3)$$

with

$$G_0 = (1-3\sin^2\theta)/2, \quad G_1 = \sin 2\theta, \quad G_2 = \cos^2\theta,$$

where D is the Doodson constant and C_j is the amplitude of the component (see Hendershott; 2005). The harmonic frequency σ_j is a linear combination of the angular velocity of the Earth's rotation ω and the sum and difference of the angular velocities ω_k, which are the five fundamental astronomical frequencies, having the largest effect modifying the potential. The gravitational force vectors due to the tidal generating bodies are obtained as the gradient of the scalar potential: $F = -\nabla V$. An elegant decomposition of the tidal constituents into groups with similar frequencies and spatial variability was developed by (Doodson; 1921) . Using Doodson's expansion, each constituent of the tide has a frequency

$$f = n_1 f_1 + n_2 f_2 + n_3 f_3 + n_4 f_4 + n_5 f_5 + n_6 f_6 \quad (3.4)$$

where n_i are the Doodson number with $n_1 = 1, 2, 3$ and n_2 to n_6 are between -5 and $+5$. To avoid negative numbers, Doodson added five to n_2 to n_6. Each tidal constituent has a Doodson number.

The Equilibrium tidal analysis considers the free surface to be a level surface due to the influence of the tidal forces and the Earth's gravity The observed tides are usually larger than the Equilibrium Tide due to the dynamic response of the ocean to tidal forces, however their frequencies are the same. The Equilibrium analysis gives an estimate of the importance of the tidal constituents, which facilitates the selection of appropriate constituents for the harmonic analysis. The amplitude ratios and phase difference between constituents are also used for certain computations in the latter. The satellites (nodal, in case of just moon) corrections are computed from the equilibrium response. The frequencies of the satellites are very close to that of the main constituents, and it is standard to consider the true amplitude ratios and phase differences to be the same as in case of the equilibrium analysis.

Let us consider a time series: $y(t)$, $t = t_1, t_2, ..., t_M$, where the observation times are regularly spaced at an interval Δt. The model equation with N constituents can be expressed as

$$y(t) = \sum_{k=1}^{N}\Big(a_k \exp^{i\omega_k t} + a_{-k}\exp^{-i\omega_k t}\Big) + b_0 + b_1(t - t_{ref}) \quad (3.5)$$

where b_0 is the mean value, the second term with coefficients b_1 indicate the trend, while the term inside the summation indicate

the variation of the constituents. The objective is to determine the coefficients which minimizes the error between the raw data and the model output. The least square error fit is used for this purpose such that the coefficients minimizes the relation

$$E = \sum_m |x(t_m) - y(t_m)|^2 = ||Ta - y||^2 \tag{3.6}$$

where $y = [y(t_1), y(t_2), ..., y(t_m)]'$, $a = [a_1, a_2, ..., a_N, a_{-1}, a_{-2}, ..., a_{-N}, b_0, b_1]$ and T is a $M \times 2N+2$ of linear and sinusoidal basis functions evaluated at the observation times. Given sufficient amount of raw data available, the number of observation points M are more than the number of unknown coefficients, the standard method of obtaining the solution is the ordinary least squares and can be obtained as

$$a = (T^T)^{-1} T^H y \tag{3.7}$$

The standard parameters are then computed as

$$\begin{aligned} L_k &= |a_k| + |a_{-k}| \\ l_k &= |a_k| - |a_{-k}| \\ \theta_k &= \Big(\frac{ang(a_k) + ang(a_{-k})}{2}\Big) \\ g_k &= v_k - \Big(\frac{ang(a_k) - ang(a_{-k})}{2}\Big) \end{aligned}$$

where L_k and l_k are the semi-major and semi-minor axis of the ellipse respectively, θ_k is the angle of inclination of the northern semi-major axis counter-clockwise from due east, and g_k is the Greenwich phase (see Foreman; 1978).

Selection of constituents

Tidal constituents are chosen from a list of 146 constituents (45 astronomical and 101 shallow water constituents). Deciding on which constituent to select is itself a big challenge. The Rayleigh criterion (Godin; 1972) states that a time series of minimum length T is required to distinguish between constituents with a frequency difference of T^{-1}. An additional criterion is required for deciding the order of inclusion. Suppose, two particular tidal constituents (their frequency difference) doesn't satisfy the Rayleigh criterion, then the constituent with larger amplitude from the equilibrium tidal analysis is considered for inclusion in the harmonic analysis. The automated selection algorithm developed by (Foreman; 1977) is widely used, where constituents are selected from a basis of all the 45 astronomical and 24 of the most important shallow water constituents. Constituents which do not satisfy the Rayleigh criterion are ignored. However, if nodal corrections are completely avoided so as to directly include the satellite and major constituents, the analysis would involve the evaluation of 529 unknown constituents (see e.g. Foreman et al.; 2009). The latter also showed with a time series of length 19 years that the methodology with nodal correction (satellite amplitude ratios and phases based on tidal potential theory) gave the same accuracy as that with approach considering all 529 constituents. Increasing the number of constituents can enhance the representation of the data (by overfitting), with no better predictions or correspondence of the results to the physical processes that are analysed.

The harmonic analysis techniques are not suitable for non-stationary processes unless they include frequencies of the non-tidal processes. Theoretically it can be possible to perform such an analysis, but it requires the knowledge of the exact frequencies of the non-tidal processes, which in reality is a big challenge. If the non-tidal frequencies are not included, then the non-linearity of the least-squares methodology can result in the interaction of the tidal and non-tidal frequencies, while the numerical stability of the matrix inversion can be jeopardised if frequencies are considered at locations where no energy is present (Jay and Flinchem; 1999).

Inference

At short length scales of records, the frequency resolution of the classical harmonic analysis approach deteriorates and in cases dissimilar constituents become unresolvable. If certain important constituents are not included directly in the analysis, they can be included indirectly by inferring their major and minor semi-axis lengths and Greenwich phase lags from the neighbouring constituents that are included. Inference has the effect of reducing any periodic behaviour of the ellipse parameters and the phase of the constituents used for inference as it removes the interaction from the neighbouring inferred constituent (Foreman and Henry; 1989).

Estimation of Confidence Intervals

In the classical harmonic analysis approach, the confidence intervals of the current ellipse parameters, which are non linear functions of the model parameters, are generated using either a Monte-Carlo uncertainty propagation method or a linearisation approach. the correlations between the model parameters which are complex coefficients are computed using complex bi-variate normal statistics. The confidence intervals of the constituents for the colored noise case are derived by estimating the actual residual spectrum at the appropriate frequencies and then using them to scale the individual elements of the covariance matrix obtained from the white noise case (see Codiga; 2011).

GAUSSIAN PROCESS REGRESSION

Gaussian processes represent a non-parametric approach to modelling unknown functions, and have been widely used in solving a variety of regression problems (Ghahramani; 2015; Rasmussen and Williams; 2006). The modelling framework considers a prior distribution directly in the space of functions. A Gaussian process is described by its mean and covariance function, similar to the description of a gaussian distribution in terms of its mean and covariance matrix (MacKay; 2003). Consider a dataset with n observations $\mathcal{D} = \{(x_i, y_i)|i = 1, 2, .., n\}$ and the objective is to make predictions for new inputs x_*. The outputs are generated by a latent function $f(x)$ with the addition of Gaussian white noise of constant variance (σ_n^2)

$$y_j = f(x_j) + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_n^2). \tag{4.1}$$

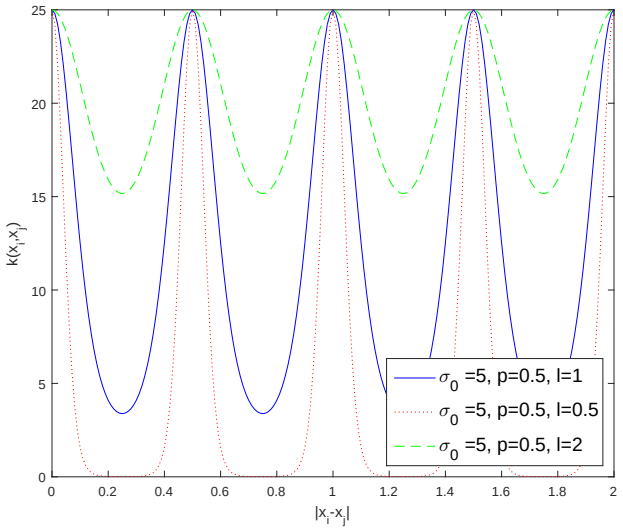

FIGURE 1 The figure plots the variations of a periodic kernel function for different values of the length scale l and fixed signal variance σ_0 and period p.

A prior distribution is considered on the latent functions such that for a given set of training points $x = \{x_1, x_2, ..., x_n\}$, the corresponding vector of function evaluations is given by the distribution $f|x \sim \mathcal{N}(0, K_{ff})$, and the properties of the prior are dictated by the choice of the covariance function $K_{ff}(i,j) = \mathbb{E}[f(x_i)f(x_j)]$ (with zero prior mean function). The specification of the prior is important as it fixes the properties of the function considered for inference (Rasmussen and Williams; 2006). Various forms of the kernel function are known and they can be combined in numerous way (addition, multiplication), depending upon the problem, to model complex data. Designing an appropriate kernel function is an important aspect of a machine learning problem. The exponentiated quadratic is one kernel function which is ubiquitously used in solving a wide variety of problems

$$K_q(i,j) = \sigma_{f,q}^2 \exp\left(\frac{|x_i - x_j|^2}{l_q^2}\right). \tag{4.2}$$

However, it is a very smooth (infinitely differentiable) function, which makes it unsuitable in modelling practical data. For such cases, the *Matérn* covariance function is considered to be more appropriate

$$K_q(i,j) = \sigma_{f,q}^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{|x_i - x_j|}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{|x_i - x_j|}{l}\right) \tag{4.3}$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind. For the particular problem of tidal current prediction, given the harmonic nature of the variations, a good choice of a kernel function would be a periodic kernel function

$$K_q(i,j) = \sigma_{f,q}^2 \exp\left(\frac{-2}{l_q^2}\sin^2\left(\frac{\pi|x_i - x_j|}{p_q}\right)\right) \tag{4.4}$$

and given the multiple number of distinct periodic tidal constituents, addition of a finite number of periodic kernel functions is considered, such that

$$K_{ff} = \sum_{q=1}^{M} K_q. \tag{4.5}$$

The variation of a periodic kernel function with same frequency and signal variance but different length-scales is illustrated in figure 1. The joint distribution of the training points and test targets is given by

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{ff} + \sigma_n^2 I_n & k_{f*} \\ k_{f*}^T & k_{**} + \sigma_n^2 \end{bmatrix}\right)$$

where K_{ff} is the covariance matrix expressing the correlations between all the training points, k_{f*} is the vector of covariance between the training points and the test target, k_{**} is the prior variance and y_* are the actual output values at the test locations. Finally, the predictive distribution is obtained as

$$y_*|x_*, y, x \quad \sim \quad \mathcal{N}(\mu_*, \sigma_*), \tag{4.6}$$

where the mean (μ_*) and the variance (σ_*) are expressed as

$$\begin{aligned} \mu_* &= k_{f*}^T (K_{ff} + \sigma_n^2 I_n)^{-1} y \\ \sigma_* &= k_{**} - k_{f*}^T (K_{ff} + \sigma_n^2 I_n)^{-1} k_{f*}. \end{aligned} \tag{4.7}$$

All the unknown hyperparameters of the problem are denoted by θ and they are determined by maximizing the log marginal likelihood (type-II maximum likelihood)

$$\log p(y|\theta) = -\frac{1}{2} y^T (K + \sigma_n^2 I_n)^{-1} y - \frac{1}{2}|K + \sigma_n^2 I_n| - \frac{n}{2}\log(2\pi). \tag{4.8}$$

More sophisticated techniques like Markov chain Monte Carlo are available for determining the hyperparameters, however they are computationally expensive.

Multi-Output

The formulation described until now does not account for the correlations between the output variables, which in the case of tidal current prediction are the horizontal velocities u and v. In the GP framework, the problem of multiple output reduces to the specification of an appropriate covariance function, which while being positive semi-definite, captures the dependencies between the data points across all the outputs (Alvarez and Lawrence; 2011). The linear model of coregionalization represents the covariance function as a product of the contribution of two covariance functions, where one of them (the coregionalization matrix) models the dependence between functions independent of the input vector, and the other covariance function models the input dependence independently of the particular set of functions. The covariance function for multiple output can be expressed as

$$K_{ff} = \sum_{q=1}^{M} \Upsilon_q \otimes K_q \tag{4.9}$$

where,

$$K_q(i,j) = \exp\left(\frac{-2}{l_q^2}\sin^2\left(\frac{\pi|x_i - x_j|}{p_q}\right)\right) \tag{4.10}$$

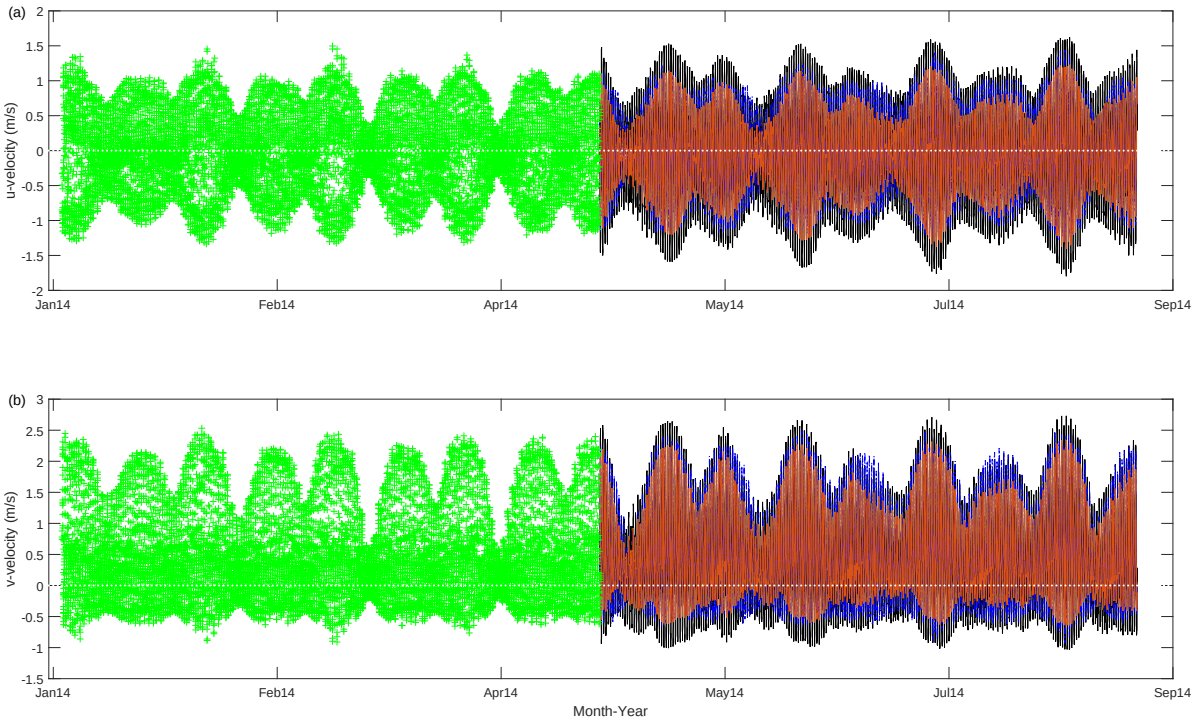

FIGURE 2 The figure shows the variations of the horizontal velocities u and v with time. Data for the first 120 days of the year 2014 are used to predict the tidal current variations for the next 120 days. The green $'+'$ markers indicate the training points, the red solid line indicates the mean of the predictions, the black solid line containing the region in grey indicates the 95% confidence interval and the blue dotted line indicates the actual values. The white dotted line across the centre indicates zero velocity. The figure provides a good macroscopic view of the tidal prediction, however the details are obscured due to the vast amount of information it contains. An addition figure is provided next illustrating a small section of the v velocity shown in (b). The computations were performed using the FullGP formulation.

and the matrix Υ_q is assumed to be of spherical parametrization kind (Pinheiro and Bates; 1996; Osborne et al.; 2012) with

$$\Upsilon_q = \text{diag}(e_q) S_q^T S_q \text{diag}(e_q). \tag{4.11}$$

Note, in (4.11), e gives a description for the length scale of each output variable

$$\text{diag}(e_q) = \begin{pmatrix} l_{u,q} & 0 \\ 0 & l_{v,q} \end{pmatrix} \tag{4.12}$$

and S is an upper triangular matrix, the i th column of which is associated with particular spherical coordinates of points of $\Re^i$:

$$S_q = \begin{pmatrix} 1 & \cos\theta_q \\ 0 & \sin\theta_q \end{pmatrix} \tag{4.13}$$

The final form of the coregionalization matrix is given by

$$\Upsilon_q = \begin{pmatrix} l_{u,q}^2 & l_{u,q} l_{v,q} \cos\theta_q \\ l_{u,q} l_{v,q} \cos\theta_q & l_{v,q}^2 \end{pmatrix} \tag{4.14}$$

This method will be referred to as `FullGP` in the analysis following.

Sparse Spectral representation (SSGP)

Shortcomings of the naïve `FullGP` implementation include its high computational costs $O(n^3)$ and memory requirements $O(n^2)$, which prohibits its application to problems with large number of data points. Several methods have been proposed to address the issue (see e.g. Snelson and Ghahramani; 2005; Walder et al.; 2008). The goal is to obtain an algorithm which reduces the computational complexity while still retaining the predictive accuracy. The sparse spectral GP (SSGP) developed by (Lázaro-Gredilla et al.; 2010) is a powerful and effective algorithm, reducing the computational costs to $O(nm^2)$ and memory requirements to $O(nm)$, where m is the number of spectral points which is typically much less than n (no. of data points). The method makes use of the Weiner-Khitchine theorem which states that the power spectrum and auto-correlation of the random process constitute a Fourier pair, and the Bochner's theorem stating that any stationary covariance function can be represented as a Fourier transform of a positive finite measure. The resultant integral is approximated using a Monte-Carlo approach, by considering the average of a finite set of samples corresponding to frequencies which are known as the spectral points. Learning of these spectral points is equivalent to learning the kernel function.

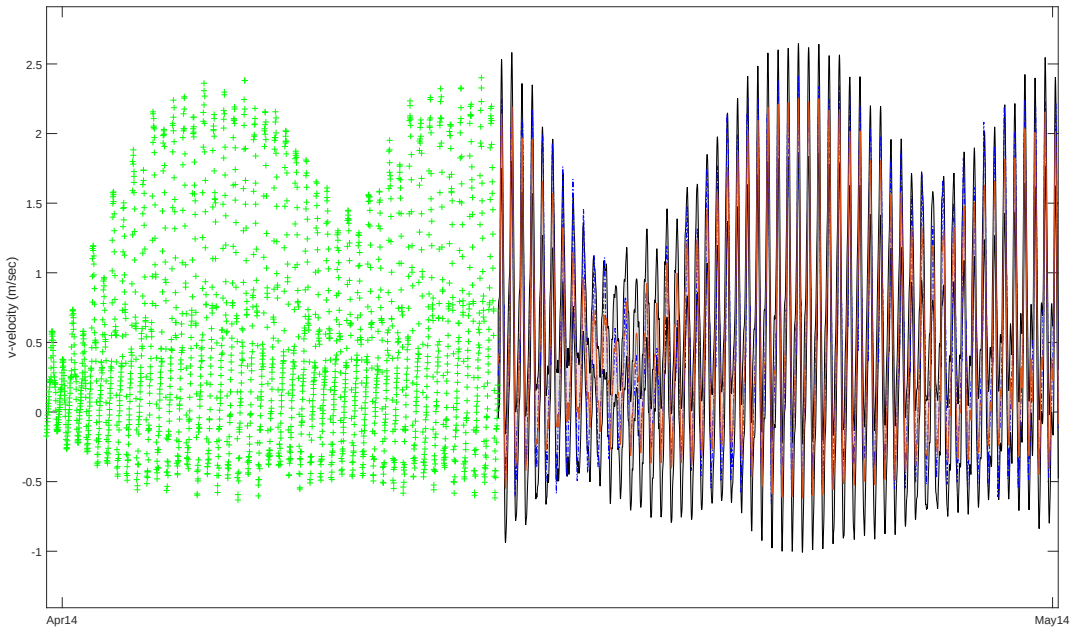

FIGURE 3 The figure shows the variation of a small section of the v velocity shown in figure 2(a). The region plotted includes the transition from training to prediction zone. The green '+' indicate the training data points, the red line indicates the mean and grey region indicates the 95% confidence interval and the blue dotted line plots the actual data values in the prediction zone.

In SSGP, the covariance function is expressed as

$$k(x_i, x_j) = \frac{\sigma_0^2}{m} \sum_{q=1}^{m} \cos(2\pi s_q^T (x_i - x_j)) \tag{4.15}$$

where, s_q are the spectral frequencies approximating any stationary covariance function, m is the number of spectral points and σ_0^2 is the variance.
The mean, variance and the log marginal likelihood are then expressed as

$$\mu_* = \phi(x_*)^T A^{-1} \Phi_f y, \qquad \sigma_* = \sigma_n^2 + \sigma_n^2 \phi(x_*)^T A^{-1} \phi(x_*), \tag{4.16}$$

$$\log p(y|\theta) = -\frac{[y^T y - y^T \Phi_f^T A^{-1} \Phi_f y]}{2\sigma_n^2} - \frac{1}{2} \log |A| + m \log \frac{m\sigma_n^2}{\sigma_0^2} - \frac{n}{2} \log 2\pi\sigma_n^2 \tag{4.17}$$

where

$$\phi(x) = [\cos(2\pi s_1^T x) \sin(2\pi s_1^T x) .... \cos(2\pi s_m^T x) \sin(2\pi s_m^T x)]^T, \tag{4.18}$$

$\Phi_f = [\phi(x_1), .., \phi(x_n)]$ is the design matrix and $A = \Phi_f \Phi_f^T + (m\sigma_n^2/\sigma_0^2) I_{2m}$.

RESULTS AND DISCUSSION

The data analysed in this paper is taken from a depth-averaged model of tides in the Pentland Firth. The model was tuned and compared to field data in Adcock et al. (2013) although only limited measurements were available. The model was forced with eight tidal constituents - - M_2, S_2, N_2, K_2, MU_2, NU_2, O_1, K_1, and the domain extends to the continental shelf to the west of the Pentland Firth and an approximately equal distance to the east. The Pentland Firth, Scotland is considered to be a prime location for installation of large arrays of tidal turbines (Adcock et al.; 2014). The analysis has been performed on tidal current data for approximately 8 months (240 days). Data for first 120 days are used as input for training, while the remaining 120 days are used for validation. Five randomly chosen locations in the Pentland Firth region are considered for the analysis (see Table 1).

TABLE 1 Longitude and latitude of the five chosen locations

Case no.	Longitude (°)	Latitude (°)
1	-3.1921	58.7368
2	-3.1280	58.6540
3	-3.0013	58.6839
4	-3.0436	58.7791
5	-3.0889	58.7160

The classical harmonic analysis is performed using the state-of-the-art Unified Tidal analysis toolbox (UTide) developed

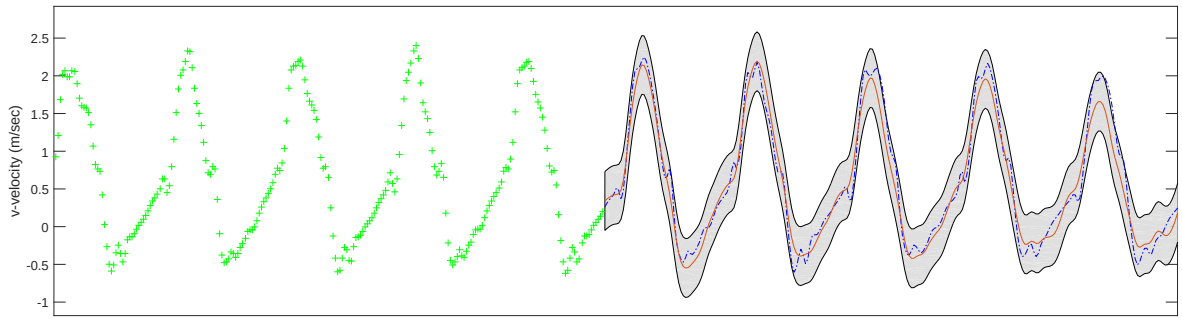

FIGURE 4 A zoomed section of figure 3, clearly showing the predicted mean (red) and confidence interval (grey), along with the periodic nature of the variations.

by (Codiga; 2011). Computations for all the three approaches (UTide, `FullGP`, SSGP) are performed with the same no. of tidal constituents, chosen by the auto-selection algorithm developed by (Foreman; 1977). The frequencies of the periodic covariance functions in the `FullGP` formulation and the spectral points in SSGP are specified to be the frequencies determined by the auto-selection procedure (Foreman; 1977). Figure 2 plots the variation of the horizontal velocities (u and v) versus time for one of the locations obtained using FullGP. The mean and confidence margins of the predictions agree well with the data. Due to the vast amount of information contained in figure 2, the short-term variations and the details are not properly visible. Figure 3 presents a short section of the figure 2(b) illustrating the details of the input, output parameters and the short term variations of the u velocity. A comparison of the root mean squared error (RMSE) in the predictions is shown in Table 2 . The predictions from the SSGP have the largest RMSE, however those from UTide and `FullGP` are quite similar to each other, with the `FullGP` giving slightly better predictions in general (average) because of more accuracy in its u predictions. Note, the SSGP formulation does not consider the cross correlations between the two velocities (u and v). This is possibly one of the reasons leading to the increase in the root mean squared error. Also, the computations for the classical tidal harmonic analysis include nodal corrections, which although small (due to short time-span considered), can contribute in lowering RMSE. The auto-selection algorithm choose 35 tidal constituents for the input data considered in this analysis, with a minimum threshold factor of one in the Rayleigh criterion. However, the Rayleigh criterion is considered to be overly conservative in case of strongly tidal signal, rejecting constituents that may be well resolved from each other (Codiga; 2011). In order to investigate this aspect, computations are performed with 59 tidal constituents using the UTide and SSGP formulation. A comparison of the RMSE from the two methods with 59 constituents is presented in Table 3. There is a general reduction in the error from both the methods compared to the previous case (with 35 constituents). However the magnitudes of the RMSE from the SSGP in this case (with larger number of tidal constituents) are similar to those from the UTide. Some computations were also performed with even higher number of constituents including some low frequencies. An obvious increase in the RMSE was observed in both the methods (UTide and SSGP), however the magnitude of the RMSE was much more with the harmonic analysis, suggesting that the latter is more sensitive to the over-specification of the tidal constituents than SSGP.

TABLE 2 Comparison of root mean squared error in (m/s) with constituents chosen by the auto-selection criterion

Velocity	UTide	`FullGP`	SSGP
u_1	0.2654	**0.2370**	0.3669
v_1	**0.1612**	0.1721	0.2794
u_2	0.3048	**0.2817**	0.3530
v_2	**0.0592**	0.0613	0.0954
u_3	0.3115	**0.2969**	0.3515
v_3	0.2904	**0.2780**	0.3221
u_4	**0.1864**	0.1881	0.1920
v_4	**0.2096**	0.2190	0.2393
u_5	0.2988	**0.2582**	0.3954
v_5	0.1915	**0.1829**	0.2265

TABLE 3 Comparison of root mean squared error (m/s) with 59 tidal constituents

Velocity	UTide	SSGP
u_1	**0.1932**	0.2329
v_1	0.1560	**0.1549**
u_2	**0.2303**	0.2323
v_2	**0.0570**	0.1583
u_3	**0.2734**	0.2815
v_3	**0.2531**	0.2592
u_4	0.1813	**0.1755**
v_4	**0.1829**	0.1875
u_5	**0.2255**	0.2314
v_5	**0.1681**	0.1744

CONCLUSION

The problem of tidal current prediction is formulated in the framework of Bayesian Gaussian processes. The unknown hyperparameters of the covariance functions, expressing correlations between data points, are determined using an optimization procedure which aims to maximize the log marginal likelihood. Given the harmonic nature of the tidal variations with deterministic frequencies, periodic covariance functions were used in this analysis. The full (complete) version of the Gaussian process produces a slightly better overall accuracy (average of mean squared errors is the least) in the predictions than the harmonic analysis. However the approach is computationally expensive, and to address the issue an alternative SSGP technique is explored which is found to produce similar predictive accuracy (slightly worse) as the harmonic analysis with larger number of tidal constituents. The probabilistic machine learning approach can handle uncertainty and noise, and generates confidence intervals directly in the time-domain (unlike harmonic analysis), which would be useful in practical applications. Several aspects could be investigated in the future including the consideration of more complex kernel functions (rougher periodic covariance function) which could appropriate for real-world data analysis. It is envisaged that the ideas and approaches presented in this work could be useful in the prediction and analysis of ocean waves in general.

The results presented here are preliminary outputs of the ongoing investigation. More detailed analysis will be presented in the near future.

ACKNOWLEDGEMENTS

This work is supported by Engineering and Physical Sciences Research Council, UK.

REFERENCES

Adcock, T. A., Draper, S., Houlsby, G. T., Borthwick, A. G. and Serhadlıoğlu, S. (2013). "The available power from tidal stream turbines in the Pentland Firth", **469**(2157): 20130072.

Adcock, T. A., Draper, S., Houlsby, G. T., Borthwick, A. G. and Serhadlıoğlu, S. (2014). "Tidal stream power in the Pentland Firth–long-term variability, multiple constituents and capacity factor", *Proc IMechE Part A: J of Power and Energy* pp. 1–8.

Alvarez, M. A. and Lawrence, N. D. (2011). "Computationally efficient convolved multiple output gaussian processes", *The Journal of Machine Learning Research* **12**: 1459–1500.

Codiga, D. (2011). "Unified Tidal Analysis and Prediction Using the UTide Matlab Functions." technical report 2011-01. graduate school of oceanography, university of rhode island, narragansett, ri. 59pp, *ftp://www.po.gso.uri.edu/pub/downloads/codiga/pubs/2011Codiga-UTide-Report.pdf* .

Doodson, A. T. (1921). "The harmonic development of the tide-generating potential", *Proceedings of the Royal Society of London. Series A* pp. 305–329.

Foreman, M. (1977). "Manual of tidal heights analysis and prediction", *Pacific Marine sciences Report 77-10, Institute of Ocean Sciences, Patricia Bay* .

Foreman, M. (1978). "Manual for Tidal Currents Analysis and Prediction".

Foreman, M., Cherniawsky, J. and Ballantyne, V. (2009). "Versatile harmonic tidal analysis: Improvements and applications", *Journal of Atmospheric and Oceanic technology* **26**(4): 806–817.

Foreman, M. and Henry, R. (1989). "The harmonic analysis of tidal model time series", *Advances in water resources* **12**(3): 109–120.

Ghahramani, Z. (2015). "Probabilistic machine learning and artificial intelligence", *Nature* **521**(7553): 452–459.

Godin, G. (1972). "The analysis of tides", *University of Toronto Press* .

Hendershott, M. (2005). "Lecture 1: Introduction to ocean tides", *2004 Program of Study: Tides, Woods Hole Oceanographic Institution* .

Jay, D. A. and Flinchem, E. P. (1999). "A comparison of methods for analysis of tidal records containing multi-scale non-tidal background energy", *Continental Shelf Research* **19**(13): 1695–1732.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E. and Figueiras-Vidal, A. R. (2010). "Sparse spectrum Gaussian process regression", *The Journal of Machine Learning Research* **11**: 1865–1881.

MacKay, D. (2003). *"Information theory, inference and learning algorithms"*, Cambridge University Press.

Matheron, G. (1973). "The intrinsic random functions and their applications", *Advances in applied probability* pp. 439–468.

Osborne, M. A., Roberts, S. J., Rogers, A. and Jennings, N. R. (2012). "Real-time information processing of environmental sensor network data using bayesian gaussian processes", *ACM Transactions on Sensor Networks (TOSN)* **9**(1): 1.

Pawlowicz, R., Beardsley, B. and Lentz, S. (2002). "Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE", *Computers & Geosciences* **28**(8): 929–937.

Pinheiro, J. C. and Bates, D. M. (1996). "Unconstrained parametrizations for variance-covariance matrices", *Statistics and Computing* **6**(3): 289–296.

Rasmussen, C. E. and Williams, C. K. (2006). "Gaussian processes for machine learning", *the MIT Press* .

Ripley, B. D. (2005). *"Spatial statistics"*, Vol. 575, John Wiley & Sons.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. and Aigrain, S. (2013). "Gaussian processes for time-series modelling", *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **371**(1984): 20110550.

Snelson, E. and Ghahramani, Z. (2005). "Sparse Gaussian processes using pseudo-inputs", *Advances in neural information processing systems*, pp. 1257–1264.

Thompson, P. D. (1956). "Optimum smoothing of two-dimensional fields", *Tellus* **8**: 384–393.

Walder, C., Kim, K. I. and Schölkopf, B. (2008). "Sparse multiscale Gaussian process regression", *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 1112–1119.