Mechanical Systems and Signal Processing 104 (2018) 188–223

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

On the confidence bounds of Gaussian process NARX models and their higher-order frequency response functions

K. Worden [a,*], W.E. Becker [b], T.J. Rogers [a], E.J. Cross [a]

[a] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, United Kingdom
[b] European Commission, Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra, VA, Italy

ARTICLE INFO

Article history:
Received 20 May 2017
Received in revised form 16 September 2017
Accepted 23 September 2017
Available online 4 November 2017

Keywords:
Nonlinear system identification
Gaussian processes
NARX models
Wave loading
Morison's equation

ABSTRACT

One of the most powerful and versatile system identification frameworks of the last three decades is the NARMAX/NARX[1] approach, which is based on a nonlinear discrete-time representation. Recent advances in machine learning have motivated new functional forms for the NARX model, including one based on Gaussian processes (GPs), which is the focus of this paper. Because of their nonparametric form, NARX models can only provide physical insight through their frequency-domain connection to Higher-order Frequency Response Functions (HFRFS). Because of the desirable properties of the GP-NARX form (no structure detection needed, natural confidence intervals), the analytical derivation of the HFRFs for the model is presented here for the first time. Furthermore, an algorithm for propagating uncertainty from the GP into the HFRF estimates is presented. A valuable by-product of the latter algorithm is a new test for nonlinearity, capable of detecting the presence of odd and even system nonlinearities. The new results are illustrated via two case studies; the first is based on simulation of an asymmetric Duffing oscillator. The second case study presents a validation of the new theory in the area of wave force prediction on offshore structures. This problem is one that has been considered by some of the authors before; the current paper takes the opportunity to highlight and correct a number of weaknesses of the original study in the light of modern best practice in machine learning.

1. Introduction

Over the last 30 years, one of the most versatile and enduring time series models used for nonlinear system identification has been the NARMAX (Nonlinear Auto-Regressive Moving Average with eXogenous inputs) model. The NARMAX model was introduced in 1985 [1,2] and has been the subject of constant interest and development since. (A comprehensive monograph on the theory and applications of the model recently appeared in [3]). The most general model form accommodates nonlinear discrete-time process *and* noise models. However, if the noise process can be assumed to be white Gaussian, the simpler NARX model can be adopted, and this will be the focus of this paper. The NARX model assumes a form whereby the current value of the system output is predicted using a nonlinear function F of previous inputs and outputs, i.e.

$$y_i = F(y_{i-1}, \dots, y_{i-n_y}; x_i, \dots, x_{i-n_x+1}) + \epsilon_i \tag{1}$$

[*] Corresponding author.
E-mail address: k.worden@sheffield.ac.uk (K. Worden).

[1] Nonlinear Auto-Regressive Moving Average with eXogenous inputs; NARX if the moving average noise model is omitted.

https://doi.org/10.1016/j.ymssp.2017.09.032
0888-3270/© 2017 The Authors. Published by Elsevier Ltd.

where the *residual sequence* ϵ_i is white Gaussian. The number of output (resp. input) lags is denoted n_y (resp. n_x). This formulation of a NARX model differs a little from the original [1,2], in that it allows the use of the present input x_i.

The earliest and still most common form of the NARX model adopts a multivariate polynomial (multinomial) expansion basis for the function F and learns the expansion coefficients/parameters by using linear (but advanced) least-squares methods. However, in general, *any* expansion basis which satisfies a universal approximation property can be used. This observation has led to various nonparametric NARX model forms based on machine learning, including Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) neural networks [4,5]. Most recently, variations on the NARX structure based on Polynomial Chaos Expansions (PCE-NARX) have emerged [6]. By utilising parameters that are random variables themselves, the PCE-NARX models are able to account for uncertainty in a natural way. The nonpolynomial variants of the NARX model have at least one very attractive feature, in that they bypass (or rather, usually ignore) the *structure detection* problem. One can think of the problem of establishing a 'traditional' NARX (or NARMAX) model in terms of two steps. The first step is *structure detection* i.e. determining which multinomial terms should be included in the model; the second is establishing the expansion parameters for the included terms i.e. *parameter estimation*. The nonparametric NARX models simply include all expansion terms consistent with certain *hyperparameters* of the model form e.g. number of nodes per layer in an MLP neural network. One then need only concern oneself with issues of including too many terms – leading to overfitting of models – and these issues can usually be addressed in a principled manner in a machine learning context [7].

Another fairly recent addition to the family of nonparametric NARX variants is one based on *Gaussian Processes* [8]. The GP-NARX[2] model form allows a number of potential advantages over the previously mentioned NARX variants, including a Bayesian framework encompassing the generation of natural confidence intervals for model predictions. (This accommodates uncertainty in a different way to the PCE-NARX models referred to earlier.) There is no intention here, to present a survey of the literature regarding GP-NARX/GP-AR models, the curious reader is instead directed towards the comparatively recent [10].

The main weakness of the nonparametric models is arguably in their inability to provide physical insight into the systems and processes under investigation – the expansion coefficients have no direct physical meaning. Of course, this problem was also present when polynomial NARX forms were used, but was overcome to a great extent by passing from the time-domain models to a *frequency-domain* representation. One of the most interesting and useful features of the NARX model is that, through a natural connection with the Volterra series [11], it allows the construction of Higher-order Frequency Response Functions (HFRFs) that allow one to visualise how different frequencies in the input to a nonlinear system interact in forming the output [12]. These objects are the natural nonlinear extension of the linear concept of a Frequency Response Function, which allows direct visualisation of resonant behaviour of dynamic systems. In the case of the polynomial form of the NARX model, the method for determining the HFRFs – the *harmonic probing* algorithm – proved to be a simple extension of the long-held algorithm for differential equations [13,14]. In the case of the neural network forms (MLP and RBF) of the NARX model, the harmonic probing algorithm could also provide closed form expressions for the HFRFS at the expense of a little more complicated algebra [15]. Other neural network structures – like the Time-Delayed Neural Network [16] – allow the calculation of Volterra *kernels* directly, rather than the HFRFs.

The main objective of the current paper is to provide a calculation for the HFRFs associated with the GP-NARX model. This is taken further, and an algorithm is also provided which can estimate confidence bounds on the HFRFs, based on the predictive uncertainty inherent in the GP algorithm. A useful by-product of the analysis is a new test for nonlinearity which can detect the presence of odd and even characteristics. The theory is validated via two case studies; the first is based on numerically simulated data from an asymmetric Duffing oscillator. The second case study develops a new application of the GP-NARX model in the area of *wave loading* on offshore structures. In the context of wave loading, it has long been known that the standard equation – Morison's equation – for the prediction of fluid loading forces on slender members [17], is inadequate outside a fairly narrow regime of wave conditions. There have been many attempts to improve on Morison's equation over the years, including one by the current first author, together with collaborators, based on polynomial NARMAX/NARX models [18]. The current paper updates that methodology, using the GP-NARX model. The advantages of the new formulation – including natural confidence bounds for predictions – are demonstrated, and the paper takes the opportunity to highlight and correct a number of weaknesses of the original study in the light of modern best practice in machine learning.

The layout of the paper is as follows: Section Two will provide a short summary of the relevant Gaussian process theory and how one can use it to define a NARX model. Section Three introduces a case study and shows how the GP-NARX model is applied in the context of a nonlinear Single-Degree-of-Freedom (SDOF) system. Section Four very briefly discusses the basic principles of the Volterra series and how it leads to the definition of HFRFs. Section Five presents the derivation of the HFRFs for the GP-NARX model, which are then computed for the case study system in Section Six. The new approach to computing uncertainty bounds for HFRFs is presented in Section Seven. The application of all the aforementioned theory to the wave loading problem for the Christchurch Bay Tower is presented and discussed in Section Eight. Finally, conclusions are presented.

[2] Within the machine learning community, the term GP-AR is sometimes used. This name makes complete sense in terms of the fact that the models are auto-regressive Gaussian processes; however it misses the fact that the terms AR or ARX in the time series literature usually refer to linear models. The term GP-NARX is preferred here as it indicates that the GP models are typically nonlinear. By appropriate choice of the GP covariance function, one could fit linear GP-AR models and the algorithm would then essentially be Bayesian linear regression [9].

2. Gaussian process NARX models

2.1. Gaussian processes

The *Gaussian Process* (GP) algorithm has its roots in the geostatistics community where it was developed as a tool for interpolating the profile of landscapes considered as random fields. This early work rests on the Masters thesis of Krige, which dates back to 1951 [19]; and the technique has long been known as *Kriging* in the geostatistics field. In more recent times, GPs were brought to the attention of the machine learning community by Neal [20] and Mackay [21], and consolidated in the recent book by Rasmussen and Williams [9]. The basic premise of the method is to perform inference over *functions* directly, as opposed to inference over *parameters* of functions.

For simplicity, the discussion here will assume that the system of interest has a single output variable.[3] Following the notation of [9] in the case of GPs in general, let $X = [\underline{x}_1, \underline{x}_2 \dots \underline{x}_N]^T$ denote a matrix of multivariate training inputs, and $\underline{y}$ denote the corresponding vector of training outputs. The input vector for a testing point will be denoted by the column vector $\underline{x}^*$ and the corresponding (unknown) output by y^*.

The regression relationship at the heart of the GP is,

$$y = f(\underline{x}) + \epsilon \tag{2}$$

where the 'noise' term ϵ is assumed to be a zero-mean random variable,

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{3}$$

and σ_n^2 – the noise variance – is a *hyperparameter* of the model (which requires estimation). It follows that,

$$\underline{y} \sim \mathcal{N}(\underline{f}, \sigma_n^2) \tag{4}$$

and f is essentially an unobserved *latent* variable of the approach.

A Gaussian process prior is formed by assuming a (Gaussian) distribution over *functions* for the latent f,

$$f(\underline{x}) \sim \mathcal{GP}(m(\underline{x}), k(\underline{x}, \underline{x})) \tag{5}$$

where $m(\underline{x})$ is the *mean function* and $k(\underline{x}, \underline{x}')$ is a positive-definite *covariance function.*

A defining property of the GP is that the density of a finite number of outputs from the process is multivariate normal. Using this fact with the known marginalisation properties of the Gaussian density allows one to consider the value of this function only at the points of interest: training points and predictions. Allowing $\underline{f}$ to denote the function values at the training points X, and f^* to denote the predicted function value at a new point $\underline{x}^*$, one has,

$$\begin{pmatrix} \underline{f} \\ f^* \end{pmatrix} \sim \mathcal{N}\left(\underline{0}, \begin{bmatrix} K(X,X) & K(X,\underline{x}^*) \\ K(\underline{x}^*,X) & K(\underline{x}^*,\underline{x}^*) \end{bmatrix}\right) \tag{6}$$

where a zero-mean prior has been used for simplicity (see [9] for a discussion), and $K(X,X)$ is a matrix whose i,j^{th} element is equal to $k(\underline{x}_i, \underline{x}_j)$. Similarly, $K(X, \underline{x}^*)$ is a column vector whose i^{th} element is equal to $k(\underline{x}_i, \underline{x}^*)$, and $K(\underline{x}^*, X)$ is the transpose of the same.

Since one is not interested in the unobserved variable $\underline{f}$, it can be marginalised (integrated out) from Eq. (4) [9], as the relevant integral,

$$p(\underline{y}) = \int p(\underline{y}|\underline{f})p(\underline{f})d\underline{f} \tag{7}$$

is over a multivariate Gaussian and has a closed-form solution. The result is the joint distribution for the training and testing target values for the observed y,

$$\begin{pmatrix} \underline{y} \\ y^* \end{pmatrix} \sim \mathcal{N}\left(\underline{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,\underline{x}^*) \\ K(\underline{x}^*,X) & K(\underline{x}^*,\underline{x}^*) + \sigma_n^2 \end{bmatrix}\right) \tag{8}$$

Once the joint distribution $p(\underline{y}, y^*)$ is converted into a conditional distribution $p(y^*|\underline{y})$, using standard results for the conditional properties of a Gaussian, one obtains the final expression [9],

$$y^* \sim \mathcal{N}(m^*(\underline{x}^*), k^*(\underline{x}^*, \underline{x}^*)) \tag{9}$$

where

$$m^*(\underline{x}^*) = k(\underline{x}^*, X)[K(X,X) + \sigma_n^2 I]^{-1}\underline{y} \tag{10}$$

[3] Notation is adopted in this paper where the inputs and outputs to *systems* are denoted by the symbols x and y respectively. The inputs and outputs to GPs are also denoted by x and y respectively. This decision has been made on the basis of maintaining consistency with the authors' previous works on both NARX models and GPs; it has been made in the belief that the context of any equations concerned removes any potential source of confusion.

is the *posterior predictive mean*, and,

$$k^*(\underline{x}^*,\underline{x}^*) = k(\underline{x}^*,\underline{x}^*) - K(\underline{x}^*,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,\underline{x}^*) + \sigma_n^2 \tag{11}$$

is the *posterior predictive variance*, again expressed for the observed y.

Thus the GP model provides a posterior distribution for the unknown quantity y^*. The mean from Eq. (9) can then be used as a 'best estimate' for a regression problem, and the variance can be used to define confidence intervals.

There remains the question of the choice of covariance function $k(\underline{x},\underline{x}')$. In practice, it is often useful to take a squared-exponential function of the form

$$k(\underline{x},\underline{x}') = \sigma_f^2 \exp\left(-\frac{1}{2l^2}||\underline{x}-\underline{x}'||^2\right) \tag{12}$$

although various other forms are possible (see [9]). Eq. (12) is the form adopted here. The covariance function involves the specification of two further *hyperparameters* σ_f^2 and l; together with σ_n^2, these can be optimised using an evidence framework [9]. Denoting the set of hyperparameters by $\underline{\theta}$, they can be found by maximising a function,

$$f(\underline{\theta}) = -\frac{1}{2}\underline{y}^T[K(X,X)+\sigma_n^2 I]\underline{y} - \frac{1}{2}\log|K(X,X)+\sigma_n^2 I| \tag{13}$$

which is equal to the log of the evidence, up to some constant. Since the number of hyperparameters in this case is small, the optimisation can be carried out simply by gradient descent, although it may be advantageous to use more powerful algorithms (an example using a population-based approach is presented in [22]).

2.2. GP-NARX models

The GP models discussed so far are essentially *static* maps, learning the relationship between point inputs and point outputs. However, it almost trivial to learn dynamical system behaviour, simply by adopting a NARX framework. The functional form in Eq. (1) is used with the function F represented by a GP. A slight variant of the NARX form which also uses the current input for prediction will be used here.

Once the GP-NARX model has been learned, there are various tests one can apply to assess the validity of the model. The most basic option is to compute *one step ahead* (OSA) predictions. In this case, using the training data, one computes the predictions for a given time using observed inputs and outputs up to that time, i.e.

$$y_i^* = F(y_{i-1},\dots,y_{i-n_y};x_i,\dots,x_{i-n_x+1}) \tag{14}$$

and compares the predicted and observed outputs.

It is always useful to have an objective measure of comparison, and the one used here will be the *Normalised Mean-Square Error* (NMSE) defined by,

$$\mathrm{NMSE}(\hat{y}) = \frac{100}{N\sigma_y^2}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{15}$$

Previous experience has shown that an NMSE of less than 5.0 indicates good agreement while one of less than 1.0 reflects an excellent fit.

Clearly, the OSA predictions are not a particularly stringent test of the model. A more demanding test is to compute the *Model Predicted Output* (MPO) defined by,

$$y_i^* = F(y_{i-1}^*,\dots,y_{i-n_y}^*;x_i,\dots,x_{i-n_x+1}) \tag{16}$$

and this test can be conducted on testing data as well as training data, which is an important consideration in the more general context of machine learning.[4]

Various *correlation functions* also provide a stringent means of validating models. The correlation function $\phi_{uv}(k)$ for two sequences of data u_i and v_i will be defined here by,

$$\hat{\phi}_{uv}(k) = \frac{\frac{1}{N-k}\sum_{i=1}^{N-k}u_i v_{i+k}}{\{E(u_i^2)E(v_i^2)\}^{\frac{1}{2}}} \quad k \geqslant 0 \tag{17}$$

with a similar expression for $k < 0$. The normalised form allows a simple expression for the 95% confidence interval for a zero result, namely $\pm 1.96/\sqrt{N}$.

For a linear system, it is argued in [14], that necessary conditions for model validity are,

$$\phi_{\epsilon\epsilon}(k) = \delta_{0k}; \quad \phi_{x\epsilon}(k) = 0 \quad \forall k \tag{18}$$

[4] Within other sectors of the system identification community, a different terminology is commonly adopted; it is often the case that computing the OSA predictions is simply referred to as *prediction*, while computing the MPO predictions is referred to as *simulation*. This is a little unfortunate, but the authors will continue with the OSA/MPO terminology to ensure consistency with their previous work.

The first condition is true only if the residual sequence ϵ_i is indeed a white noise sequence i.e. purged of structure; this is a test of any noise model whose job it is to reduce the residuals to white noise; a condition which is required for unbiased parameter estimates in the traditional system identification theory [23,24]. The second condition states that the residual signal is uncorrelated with the input sequence x_i i.e. the model has completely captured the component of the measured output which is correlated with the input. In the case of a nonlinear system it can be possible to satisfy the conditions above even if the model is invalid. It is shown in [14] that a more comprehensive test of the fitness of a nonlinear model requires the evaluation of three additional correlation functions. The extra conditions adopted here are,

$$\phi_{\epsilon(\epsilon x)}(k) = 0 \quad \forall k \geqslant 0; \quad \phi_{x^{2\prime}\epsilon}(k) = 0 \quad \forall k; \quad \phi_{x^{2\prime}\epsilon^2}(k) = 0 \quad \forall k \tag{19}$$

where the dash which accompanies x^2 above indicates that the mean has been removed.

The GP form for the NARX model has its advantages and disadvantages; two of the main issues will be discussed briefly here, with directions to the literature as to their possible means of solution.

The first problem is that the GP algorithm depends on the inversion of the covariance matrix K; this is an operation which costs $O(N^3)$ multiplications, where N is the number of training points. Slightly less costly is the prediction of new outputs with $O(N)$ multiplications needed for the predictive mean and $O(N^2)$ for the predictive variance. In fact, system identification with NARX models has traditionally been carried out with small training sets with a low number of thousands of data points, and this size of problem is typically feasible using a standard GP algorithm. However, if one wishes to move to larger training sets, the costs of computation can become prohibitive. This problem has led to the idea of *sparse Gaussian processes* which, as the name suggests, can establish models on reduced training sets [25]. One of the earlier methods is the so-called Fully Independent Training Conditional (FITC) model [26]. The FITC approach approximates the full GP by establishing M *pseudo-inputs* which are not restricted to be actual data points, but can be considered as hyperparamaters which can be learned. In the FITC approach, the computational complexity of establishing the GP is reduced to $O(M^2N)$; the cost of computing the predictive mean is reduced to $O(M)$ and that of the predictive variance is reduced to $O(M^2)$. Very recent work on *Variational Fourier Features* has offered the possibility that the cost of computation can be reduced from $O(M^2N)$ (for the sparse GP) to $O(MN)$ under some circumstances [27].

The second problem with the GP-NARX formulation relates to noise on the training data. The standard formulation of the GP algorithm assumes that the training inputs are noise-free and that the noise on the outputs is Gaussian with constant variance as in Eq. (2). This can be an issue if one is attempting multi-step ahead predictions with a GP-NARX model; because of the feeding back of the output predictions, the outputs *become* inputs and carry their predictive uncertainty with them. The most principled approach to this problem is adopt the Bayesian approach of marginalising or integrating over the input noise distributions; however, even if these were known with complete accuracy, the computation would be intractable. One of the first comprehensive studies of this problem appears to have been the work leading to the thesis [28]. Closed-form approximate solutions for the predictive mean and variance in the presence of input noise can be found in [29]. An interesting recent paper [30], uses an approximation to convert input noise into output noise.

In the case studies presented here (for reasons discussed in the relevant sections), the issues referred to above have been ignored without (it is believed), damage to the results; however, in other engineering problems they will likely need to be addressed.

3. Case study – an asymmetric Duffing oscillator

In order to illustrate the use of the GP-NARX formulation, data simulated from a Duffing oscillator data system will be used. In the asymmetric case when a quadratic stiffness is present, the relevant equation of motion is,

$$m\ddot{y} + c\dot{y} + ky + k_2y^2 + k_3y^3 = x(t) \tag{20}$$

Data were simulated here by integrating the equation of motion using a fourth-order fixed-step Runge–Kutta algorithm [31]. The parameters adopted were $m = 1$, $c = 20, k = 10^4, k_2 = 10^7$ and $k_3 = 5 \times 10^9$. The excitation used was a zero-mean white Gaussian random sequence with a standard deviation of 2.0. The time step used was $\Delta t = 0.001$ seconds corresponding to a sampling frequency of 1 kHz. Noise was added to the data to introduce an element of reality to matters; initially in this case, Gaussian noise of 1% RMS of the signal was added to the response time data. As a more severe test of the prediction capability of the model, the results presented here are for an independent test set of data, also comprising 1000 samples of data from the system at the same level of excitation as the training data and with the same amplitude of added noise.

As discussed in Section 2.1, there are three hyperparameters for the simple GP formulation used here; these were determined here by using a conjugate gradients algorithm in order to maximise the log marginal evidence in Eq. (13). Before doing this, it was necessary to establish the number of input and output lags needed in the model; these numbers are hyperparameters of the GP-NARX model. A quick search using the errors on a validation set gave the values $n_x = n_y = 3$. Once the lag numbers were established, the GP-NARX model was fitted and the optimal GP hyperparameters were found to be: $\sigma_f^2 = 757.4, l = 9.581$ and $\sigma_n^2 = 3.057 \times 10^{-4}$. To improve the conditioning of the estimation process, all data were standardised before the computation, the scales for the data were reintroduced after predictions were made.

The first set of results presented are for the OSA predictions on the test data set from the trained GP as shown in Fig. 1. The NMSE value for the predictions was 0.028, indicating an excellent result. The discrepancies in the predictions are barely visible, and the confidence intervals (±3 standard deviations) are very close to the data.

As discussed above, the MPO predictions provide a more stringent test and these are shown in Fig. 2. The corresponding NMSE in this case was 3.44, which still indicates a good fit.

As discussed in [32], the confidence intervals are still very small and do not accommodate the observed prediction errors. This is because not all of the uncertainty has been accounted for. In the predictions so far, the predicted outputs have been fed back into the model in order to form the MPO. This means that the only uncertainty accounted for in the predictions is the *parameter* uncertainty. In order to take a proper Bayesian viewpoint, one should allow for the fact that each prediction is actually a sample from a distribution; this distribution being determined by the parameter distribution. To account for this, during a prediction run, at each instant i the prediction y_i^* was sampled from the distribution specified by the predictive mean and covariance as specified by Eqs. (10) and (11). One such run generates a single realisation of the prediction process, in order to accumulate information about the distribution of predictions with state estimation taken into account, a Monte Carlo approach was adopted here with 25 different runs conducted. Fig. 3 shows the 25 realisations of the predictions.

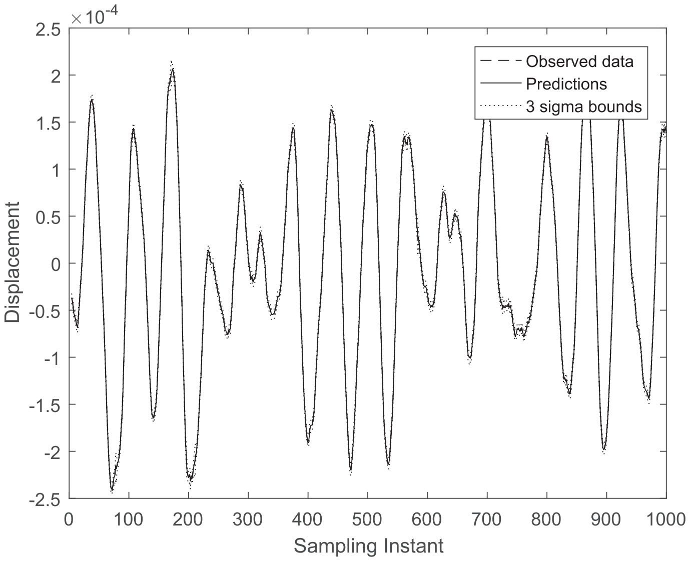


Fig. 1. OSA predictions for GP-NARX model of Duffing oscillator data.

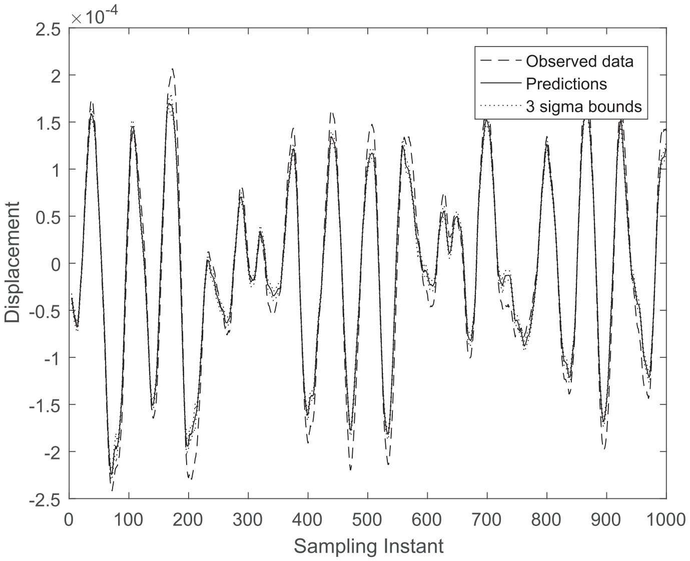


Fig. 2. MPO predictions for GP-NARX model of Duffing oscillator data.

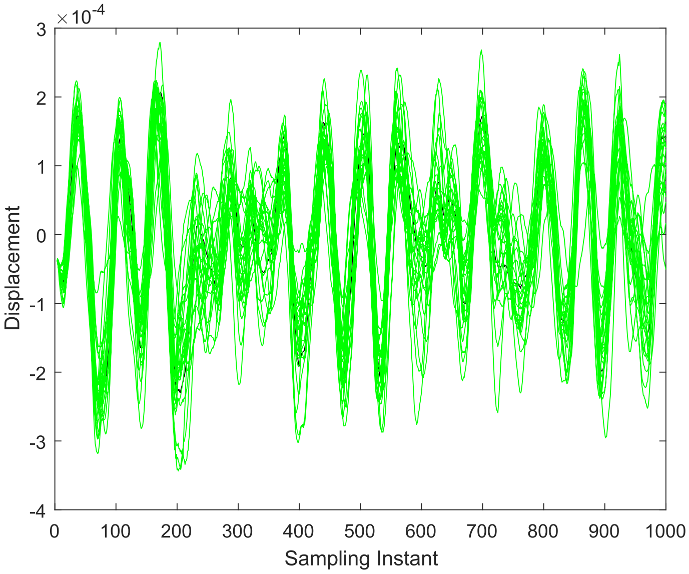


Fig. 3. MC realisations of predictions for GP-NARX model of Duffing oscillator data.

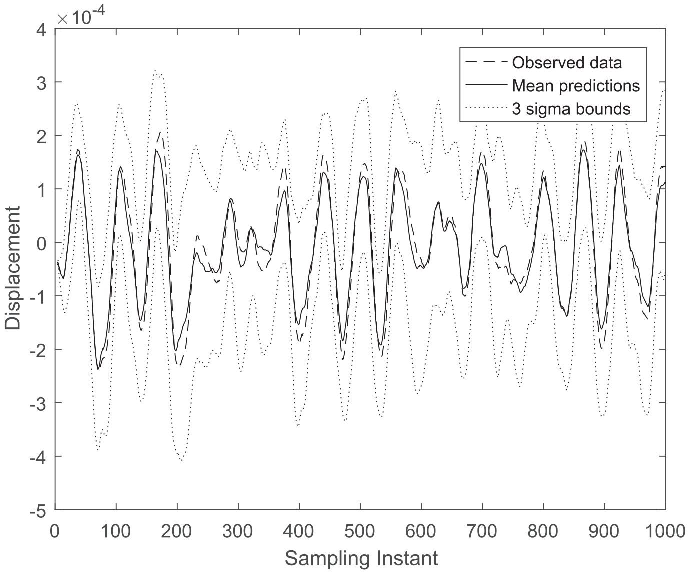


Fig. 4. MC predictions for GP-NARX model of Duffing oscillator data.

There is clearly a great deal more uncertainty associated with the predictions now. From the MC realisations, one can estimate a mean prediction and determine $\pm 3\sigma$ confidence bounds, and the result of the analysis for the case here is shown in Fig. 4. The confidence intervals are now a more appropriate assessment of the predictive capability of the model. This exercise shows clearly that the dominant contribution to uncertainty in the predictions is not the direct component from the parameter uncertainty, but the indirect component due to state estimation from the uncertain parameters.

Having established a benchmark data set and illustrated the GP-NARX performance, the paper continues by developing the theory for the HFRFs and illustrating that theory on the benchmark data.

4. The Volterra series and higher-order FRFs

In the time-domain analysis of linear dynamical systems, the *impulse response function* $h(\tau)$ is known to characterise the system completely. For such a system, excited by an input signal $x(t)$, the response $y(t)$ is given by the convolution integral (sometimes called *Duhamel's integral*),

$$y(t) = \int_{-\infty}^{\infty} d\tau \, h(\tau)x(t-\tau) \tag{21}$$

This relationship was extended to nonlinear systems by Volterra [33] in the early part of the last century; the output of a nonlinear system was shown to require additional higher-order contributions such that the total response, $y(t)$, is given by,

$$y(t) = y_0 + y_1(t) + y_2(t) + y_3(t) + \ldots + y_n(t) \tag{22}$$

where y_0 is a constant, and the general term is given by,

$$y_n(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} d\tau_1 d\tau_2 \ldots d\tau_n \, h_n(\tau_1, \tau_2 \ldots \tau_n)x(t-\tau_1)x(t-\tau_2)\ldots x(t-\tau_n) \tag{23}$$

This is essentially a generalisation of the standard Taylor series to the case of *functionals* i.e. mappings between functions. The generalised coefficients of the series h_n, are the n^{th}-order *Volterra kernels*, and these can be thought of as multi-dimensional, or higher-order, impulse response functions [11]. The series provides a representation of a given functional or system $y(t) = S[x(t)]$, which is insensitive to the input $x(t)$, provided that the system is time-invariant and contains only analytic nonlinearities [34].

The Volterra series is thus a time-domain representation for nonlinear systems. As in the case of linear systems, a dual frequency-domain representation exists which can give a clearer perspective of system behaviour in some respects. For a linear system, the frequency-domain representation is obtained simply by taking the Fourier transform of Eq. (21); the result is well known,

$$Y(\omega) = H(\omega)X(\omega) \tag{24}$$

where

$$H(\omega) = \int_{-\infty}^{\infty} d\omega \, e^{-i\omega t}h(t) \tag{25}$$

is the system *Frequency Response Function* (FRF), and $Y(\omega)$ and $X(\omega)$ have similar definitions. By direct extension of the linear case, the higher-order FRFs (HFRFs) $H_n(\omega_1, \ldots, \omega_n)$, can be defined as the multi-dimensional Fourier transforms of the kernels,

$$H_n(\omega_1, \ldots, \omega_n) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} d\tau_1 \ldots d\tau_n h_n(\tau_1, \ldots, \tau_n)e^{-i(\omega_1\tau_1 + \ldots + \omega_n\tau_n)} \tag{26}$$

with inverses,

$$h_n(\tau_1, \ldots, \tau_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} d\omega_1 \ldots d\omega_n H_n(\omega_1, \ldots, \omega_n)e^{+i(\omega_1\tau_1 + \ldots + \omega_n\tau_n)} \tag{27}$$

The frequency-domain dual of expression (22) then follows directly as,

$$Y(\omega) = Y_1(\omega) + Y_2(\omega) + Y_3(\omega) + \ldots \tag{28}$$

where

$$Y_1(\omega) = H_1(\omega)X(\omega) \tag{29}$$

$$Y_2(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega_1 H_2(\omega_1, \omega - \omega_1)X(\omega_1)X(\omega - \omega_1) \tag{30}$$

$$Y_3(\omega) = \frac{1}{(2\pi)^2} \times \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d\omega_1 d\omega_2 H_3(\omega_1, \omega_2, \omega - \omega_1 - \omega_2)X(\omega_1)X(\omega_2)X(\omega - \omega_1 - \omega_2) \tag{31}$$

and so on.

The interpretation of these quantities is well established, a description can be found in [12]. If the equations of motion are known for a system, the method of harmonic probing can be used in order to compute the HFRFs [13]. Harmonic probing for the Gaussian process NARX models is discussed in the next section and the technique is illustrated using the Duffing oscillator system in Eq. (20); the exact results in this case are well known as [12],

$$H_1(\omega) = \frac{1}{-m\omega^2 + ic\omega + k} \tag{32}$$

$$H_2(\omega_1,\omega_2) = -k_2 H_1(\omega_1)H_1(\omega_2)H_1(\omega_1+\omega_2) \tag{33}$$

and,

$$\begin{aligned} H_3(\omega_1,\omega_2,\omega_3) = -\frac{1}{6}H_1(\omega_1+\omega_2+\omega_3) \\ \times\{4k_2(H_1(\omega_1)H_2(\omega_2,\omega_3)+H_1(\omega_2)H_2(\omega_3,\omega_1)+H_1(\omega_3)H_2(\omega_1,\omega_2))+6k_3H_1(\omega_1)H_1(\omega_2)H_1(\omega_3)\}\end{aligned} \tag{34}$$

for the first three HFRFs.

In order to see the important structure in the HFRFs, it is often sufficient to plot only the leading diagonal i.e. $H_2(\omega,\omega)$. This format also allows simple comparisons between the functions.

5. Harmonic probing of the GP-NARX model

If the governing equations of motion are known, the HFRFs of a system can be obtained analytically by the use of the *harmonic probing* algorithm, introduced by Bedrosian and Rice [13]. Although this was originally designed for continuous-time systems, the algorithm was extended to the type of discrete-time systems considered here by Billings and Tsang [4].

Before proceeding, it is necessary to determine the explicit form of the GP-NARX model. First of all, one observes, following [9], that the GP is essentially an expansion in terms of basis functions fixed by the covariance kernel and the training data, the predicted output y^* corresponding to a new input $\underline{x}^*$ is given by,

$$y^* = \sum_{i=1}^{N} a_i k(\underline{x}^*, \underline{x}_i) \tag{35}$$

where according to Eq. (10),

$$\underline{a} = [k(X,X) + \sigma_n^2 I]^{-1}\underline{y} \tag{36}$$

and this is fixed by the training data. If one adopts the squared exponential covariance function of (12), one arrives at the GP-NARX form,

$$y_i = \sigma_f^2 \sum_{j=1}^{N} a_j \exp\left\{-\frac{1}{2l^2}\left[\sum_{k=1}^{n_y}(y_{i-k}-v_{jk})^2 + \sum_{m=0}^{n_x}(x_{i-m}-u_{jm})^2\right]\right\} \tag{37}$$

where the matrix $V = \{v_{ij}\}$ is formed from the first n_y columns of the matrix X and $U = \{u_{ij}\}$ is formed from the remaining $n_x + 1$ columns of X.

Note that this expression is essentially that of the radial-basis function neural network considered in [15]; this means that the HFRFs derived in that paper are applicable here. However, the analysis here presents a more direct approach in terms of homogeneous ARX and NARX model coefficients at each polynomial order; the expressions here also correct some typographical errors in [15]. The first issue which arises is that the function in (37) must be expanded as a polynomial in order to apply harmonic probing. As observed in [15], direct expansion means that the term of order n will contain powers of all orders up to n and this makes it impossible to group linear terms etc. The solution is simple, a trivial rearrangement yields the more amenable form,

$$y_i = \sigma_f^2 \sum_{j=1}^{N-p} a_j \gamma_j \exp\left\{-\frac{1}{2l^2}\left[\sum_{k=1}^{n_y}(y_{i-k}^2 - 2v_{jk}y_{i-k}) + \sum_{m=0}^{n_x}(x_{i-m}^2 - 2u_{jm}x_{i-m})\right]\right\} \tag{38}$$

where

$$\gamma_j = \exp\left\{-\frac{1}{2l^2}\left[\sum_{k=1}^{n_y} v_{jk}^2 + \sum_{m=0}^{n_x} u_{jm}^2\right]\right\} \tag{39}$$

Now, the basis of the harmonic probing method is to examine the response of the system to certain very simple inputs. In order to identify $H_1(\omega)$, for example, the system is ‘probed’ with the single harmonic,

$$x_i^p = e^{i\Omega t} \tag{40}$$

Substituting this expression into the Volterra series (22), the corresponding response is [12],

$$y_i^p = H_1(\Omega)e^{i\Omega t} + H_2(\Omega,\Omega)e^{2i\Omega t} + H_3(\Omega,\Omega,\Omega)e^{3i\Omega t} + \dots \tag{41}$$

Now, consider the consequences of substituting the expressions (40) and (41) into the GP function (38) and expanding it as a polynomial. None of the higher-order terms in (41) can combine in any way to generate a component at the fundamental frequency of excitation Ω. As a result, if the coefficient of $e^{i\Omega t}$ is extracted from the resulting expression, the *only* HFRF which can appear is $H_1(\Omega)$; thus the expression can be rearranged to give an analytical expression for H_1. In fact, one need only consider the linear terms in the expansion in order to extract H_1, so one essentially considers the ARX model,

$$y_i = \sigma_f^2 \sum_{j=1}^{N-p} \frac{a_j \gamma_j}{l^2} \left\{ \sum_{k=1}^{n_y} v_{jk} y_{i-k} + \sum_{m=0}^{n_x} u_{jm} x_{i-m} \right\} \tag{42}$$

Changing the order of summation here results in the standard ARX form,

$$y_i = \sum_{j=1}^{n_y} \alpha_j y_{i-j} + \sum_{j=0}^{n_x} \beta_j x_{i-j} \tag{43}$$

where

$$\alpha_j = \frac{\sigma_f^2}{l^2} \sum_{i=1}^{N-p} a_i \gamma_i v_{ij} \tag{44}$$

$$\beta_j = \frac{\sigma_f^2}{l^2} \sum_{i=1}^{N-p} a_i \gamma_i u_{ij} \tag{45}$$

Harmonic probing of this expression is straightforward; one substitutes the probing expressions (40) and (41) into (43) and collects together all the coefficients of $e^{i\Omega t}$. In doing this, account must be taken of the effect of time-delays on the harmonic signals, this is straightforward to compute as,

$$x_{i-k} = \Delta^k x_i = \Delta^k e^{i\Omega t} = e^{-ki\Omega\Delta t} e^{i\Omega t} \tag{46}$$

$$y_{i-k} = \Delta^k y_i = \Delta^k H_1(\Omega) e^{i\Omega t} = e^{-ki\Omega\Delta t} H_1(\Omega) e^{i\Omega t} \tag{47}$$

where Δ is the backward shift operator. The result of the calculation is,

$$H_1(\Omega) = \frac{\sum_{j=0}^{n_x} \beta_j e^{-ij\Delta t\Omega}}{1 - \sum_{j=1}^{n_y} \alpha_j e^{-ij\Delta t\Omega}} \tag{48}$$

with the α_j and β_j as defined in Eqs. (44) and (45).

The extraction of H_2 is a little more complicated, this requires probing with two independent harmonics, so,

$$x_i^p = e^{i\Omega_1 t} + e^{i\Omega_2 t} \tag{49}$$

The standard computation using based on (22) shows that the corresponding response is [12],

$$y_i^p = H_1(\Omega_1) e^{i\Omega_1 t} + H_1(\Omega_2) e^{i\Omega_2 t} + 2H_2(\Omega_1, \Omega_2) e^{i(\Omega_1+\Omega_2)t} + \dots \tag{50}$$

The argument proceeds as for H_1; if these expressions are substituted into the GP function (38), the only HFRFs to appear in the coefficient of the sum harmonic $e^{i(\Omega_1+\Omega_2)t}$, are H_1 and H_2, where H_1 is already known from Eq. (48). As before, the coefficient can be rearranged to give an expression for H_2 in terms of the GP parameters and H_1. The only terms in the expansion of (38) which are relevant for the calculation are those at first and second-order. The calculation is straightforward but tedious and yields,

$$H_2(\Omega_1, \Omega_2) = \frac{A + B + C}{D} \tag{51}$$

where

$$A = \sum_{k=1}^{n_y} \sum_{l=1}^{n_y} \alpha_{kl} H_1(\Omega_1) H_1(\Omega_2) \left(e^{-i\Omega_1 k\Delta t} \cdot e^{-i\Omega_2 l\Delta t} + e^{-i\Omega_2 k\Delta t} \cdot e^{-i\Omega_1 l\Delta t} \right) \tag{52}$$

$$B = \sum_{k=1}^{n_y} \sum_{l=0}^{n_x} \beta_{kl} \left(H_1(\Omega_1) e^{-i\Omega_1 k\Delta t} \cdot e^{-i\Omega_2 l\Delta t} + H_1(\Omega_2) e^{-i\Omega_2 k\Delta t} \cdot e^{-i\Omega_1 l\Delta t} \right) \tag{53}$$

$$C = \sum_{k=0}^{n_x} \sum_{l=0}^{n_x} \gamma_{kl} \left(e^{-i\Omega_1 k\Delta t} \cdot e^{-i\Omega_2 l\Delta t} + e^{-i\Omega_2 k\Delta t} \cdot e^{-i\Omega_1 l\Delta t} \right) \tag{54}$$

and,

$$D = 1 - \sum_{k=1}^{n_y} \alpha_k e^{-i(\Omega_1+\Omega_2)k\Delta t} \tag{55}$$

The coefficients in the above expressions are given by,

$$\alpha_{jm} = \frac{\sigma_f^2}{4l^4}\sum_{i=1}^{N-p} a_i \gamma_i v_{ij} v_{im} - \delta_{jm}\frac{\sigma_f^2}{2l^2}\sum_{i=1}^{N-p} a_i\gamma_i \tag{56}$$

$$\beta_{jm} = \frac{\sigma_f^2}{2l^4}\sum_{i=1}^{N-p} a_i \gamma_i v_{ij} u_{im} \tag{57}$$

$$\gamma_{jm} = \frac{\sigma_f^2}{4l^4}\sum_{i=1}^{N-p} a_i \gamma_i u_{ij} u_{im} - \delta_{jm}\frac{\sigma_f^2}{2l^2}\sum_{i=1}^{N-p} a_i\gamma_i \tag{58}$$

where δ_{jm} is the standard Kronecker delta.

Derivation of H_3 is considerably more lengthy and requires probing with three harmonics, the expression is not given here for reasons of space. The following results section of this paper will present examples of these calculations for H_1 and H_2.

6. HFRF results for the Duffing oscillator case study system

In this section, the HFRFs for the asymmetric Duffing oscillator system of Eq. (20) are estimated from a GP-NARX model fitted to the simulated data. In fact, a subtlety suggests a reanalysis of the data. Usually, if input and output data are known to be corrupted by noise, best practice demands that a NARMAX model with nonlinear noise model is fitted to the data in order to avoid the possibility of bias in the parameter estimates [12]. However, when the HFRFs are to be computed, the noise model is discarded, leaving a NARX model for harmonic probing. In the case of the GP model, the noise variance σ_n^2 is essentially built into the parameter estimates as it effectively acts as a regularisation parameter in inverting the $K(X,X)$ matrix to form the parameters $\underline{a}$. Furthermore, the objective here is to compare the HFRF estimates with exact forms derived from Duffings equation. As the effect of noise on the HFRFs will be considered a little later, a data set was analysed here where only 0.001% noise was added to the Duffing response data.

As before, the GP hyperparameters were estimated by maximising the log marginal evidence, in this case the results were: $\sigma_f^2 = 129.2, l = 8.027$ and $\sigma_n^2 = 5.54 \times 10^{-11}$. The model gave an OSA error of 9.4×10^{-7} and an MPO error of 0.001. The comparisons between predicted and measured response are not given as the curves are not distinguishable given the accuracy of the predictions. However, it is meaningful to give comparisons between the exact HFRFs, given by Eqs. (32) and (33), and those estimated from the GP. Fig. 5 shows a comparison between the exact and estimated $H_1(\omega)$; it is clear that the estimate is very accurate indeed.

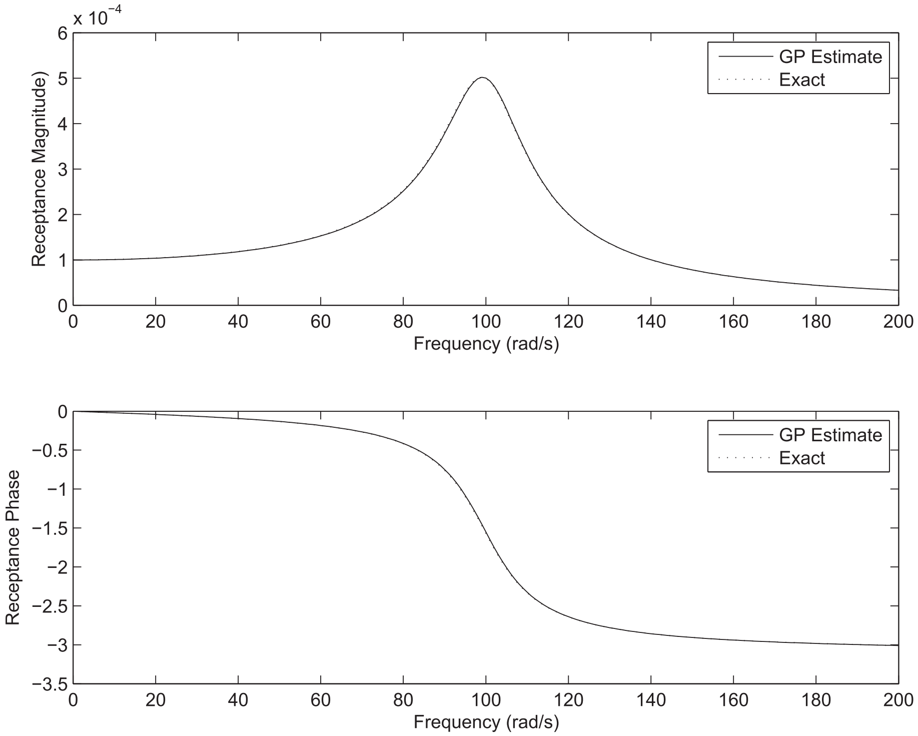


Fig. 5. Gaussian Process estimate of $H_1(\omega)$ compared to exact result.

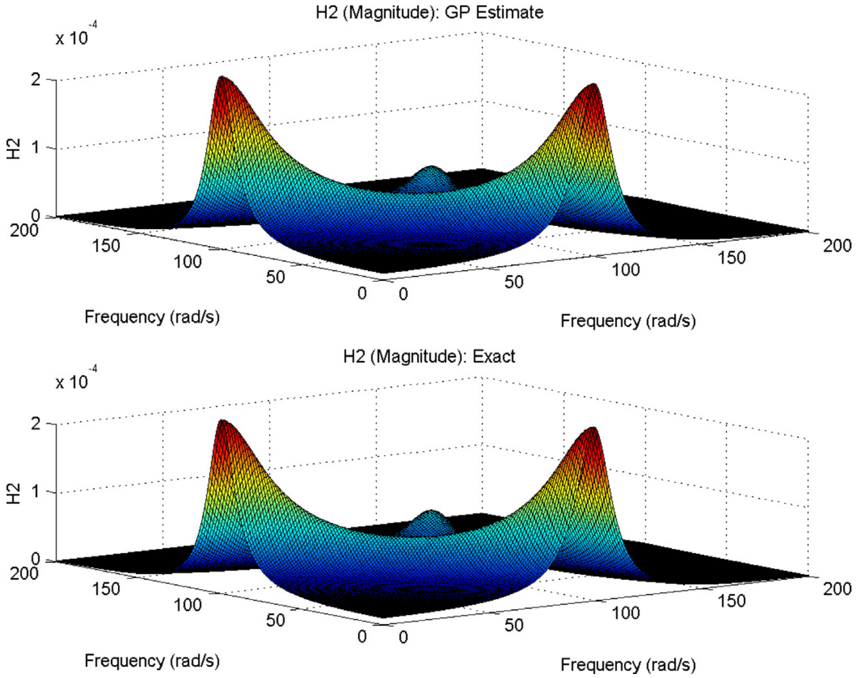


Fig. 6. Gaussian Process estimate of $H_2(\omega_1, \omega_2)$ magnitude compared to exact result.

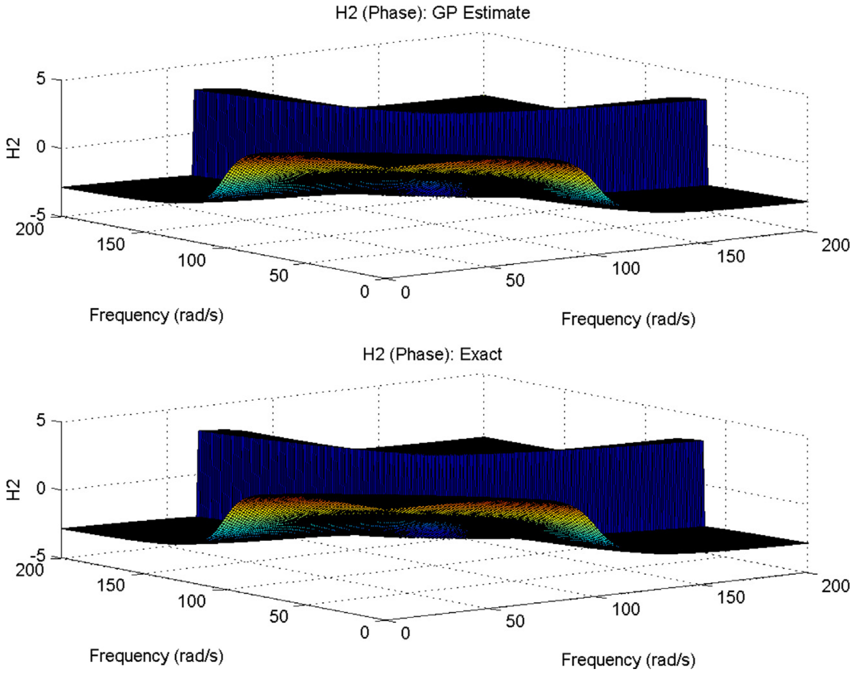


Fig. 7. Gaussian Process estimate of $H_2(\omega_1, \omega_2)$ phase compared to exact result.

Figs. 6 and 7 show comparisons between the exact and estimated H_2 functions in terms of magnitude and phase respectively. Because a direct visual comparison is subjective when the surfaces are displayed, the exact and estimated diagonals $H_2(\omega, \omega)$ are shown in Fig. 8, the accuracy of the estimates is clearly excellent.

The results presented here are much better than those presented in [15]; however, this is not perhaps a fair comparison as the neural network hyperparamaters in the earlier reference were not optimised using a fully principled approach according to current practice in machine learning.

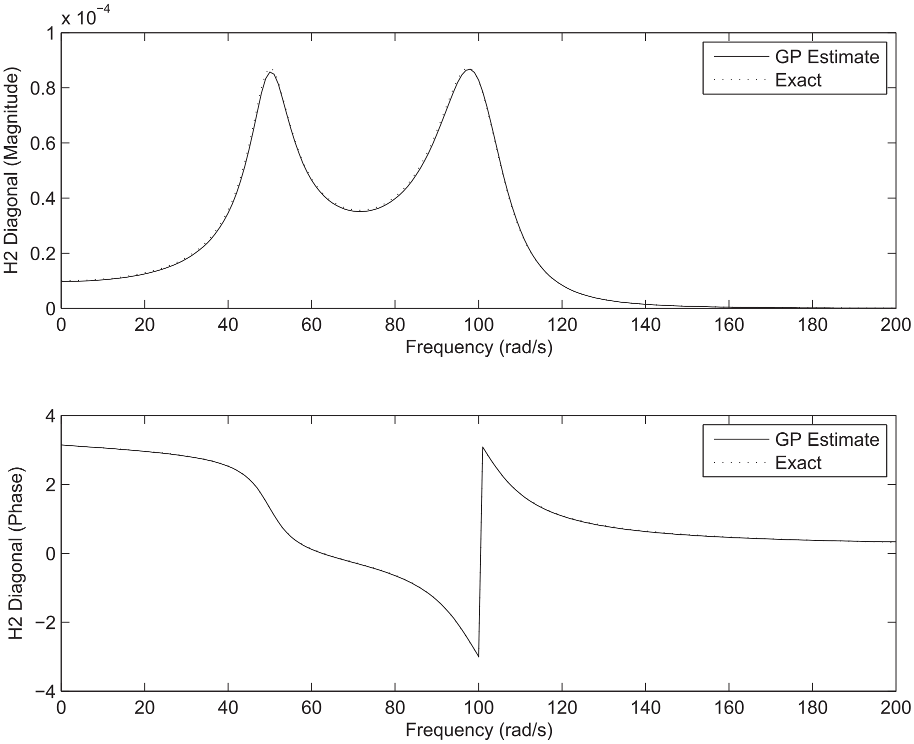


Fig. 8. Gaussian Process estimate of $H_2(\omega,\omega)$ magnitude and phase compared to exact result.

7. Uncertainty bounds on HFRFs

7.1. Proposed algorithm

In the GP-NARX framework presented so far, the potential of the Gaussian process has not been fully exploited, because the GP provides a stochastic fit to a set of training data, but so far this stochasticity has only been propagated as far as the time domain. Ideally one would like to propagate the uncertainty all the way to the end product of the exercise – the HFRFs.

To achieve this end, a Monte Carlo (MC) framework is adopted here. It is assumed that the best estimate of uncertainty in the time domain is represented by the Monte Carlo predictions in Fig. 3. Importantly, each of the lines in this figure represents a draw from a single underlying Gaussian process, which has a single set of α_j and β_j coefficients. In order to propagate the uncertainty to the frequency domain, the approach taken here is to treat each Monte Carlo draw as a new set of training data, and fit a new GP-NARX model to each draw. Each new GP-NARX model comes with its own α_j and β_j, each of which can be used to generate a draw of the HFRFs. The resulting distributions over HFRFs can be used to build mean HFRFs and confidence intervals.

Let $y^*_{r,i}, i = 1,\dots,N, r = 1,\dots,R$ represent the GP prediction at x_i, corresponding to the r^{th} Monte Carlo draw. For each of the R Monte Carlo draws, a new GP-NARX model is fitted, such that,

$$y^{**}_{i,r} = F(y^*_{r,i-1},\dots,y^*_{r,i-n_y}; x_i, x_{i-1},\dots,x_{i-n_x}) \tag{59}$$

where $y^{**}_{i,r}$ represents the GP-NARX prediction corresponding to the r^{th} MC draw at x_i. Notice that the GP is fitted using the OSA approach, because the interest is only in obtaining the corresponding α_j and β_j parameters. To simplify matters, it is also assumed that the hyperparameters of each GP are all equal to the hyperparameters estimated in the original GP.

From here, the procedure is straightforward. For the r^{th} GP fit, the corresponding $\alpha_{j,r}$ and $\beta_{j,r}$ values can be obtained using the expressions in (44) and (45), via the parameters of the GP. These can then be translated into a draw from the HFRF, resulting in a total of R samples from the distribution of the HFRF. This easily yields statistics of the distribution, such as pointwise means, modes and confidence intervals. In this paper, moderate values of $R = 50$ and in some cases $R = 100$ are chosen, since the objective is mainly to illustrate the methodology. However the computational cost of generating draws of the HFRFs is relatively low, so the sample size can easily be increased to higher numbers. To ensure accuracy, one could even sequentially increase the sample size and monitor the convergence of the statistics of interest.

7.2. Illustration on the Duffing oscillator case study

In order to illustrate the use of the GP-NARX formulation, the data simulated from the asymmetric Duffing oscillator system with added noise of 1% on the response will be used. The only difference in the GP-NARX analysis will be that 50 MC realisations/draws were generated.

Fig. 9 shows the 50 realisations of the Monte Carlo draws from the GP-NARX model. Since these are OSA predictions for each realisation with a high density of training data, the GP fits almost perfectly to each data series. Collectively, there is a moderate degree of uncertainty in fitting the GP-NARX model to the underlying training data, as evidenced by the spread of fits. As noted earlier, the dominant contribution to uncertainty in the predictions is not the direct component from the parameter uncertainty, but the indirect component due to state estimation from the uncertain parameters.

In order to propagate the uncertainty in Fig. 3 into the frequency domain, a new GP-NARX model is fitted to each of the 50 Monte Carlo draws, yielding draws of the HFRF by harmonic probing. At each frequency, the mean and standard deviation can be estimated over the 50 draws. The result is in Fig. 10 which plots the mean and 99% confidence intervals. It is evident

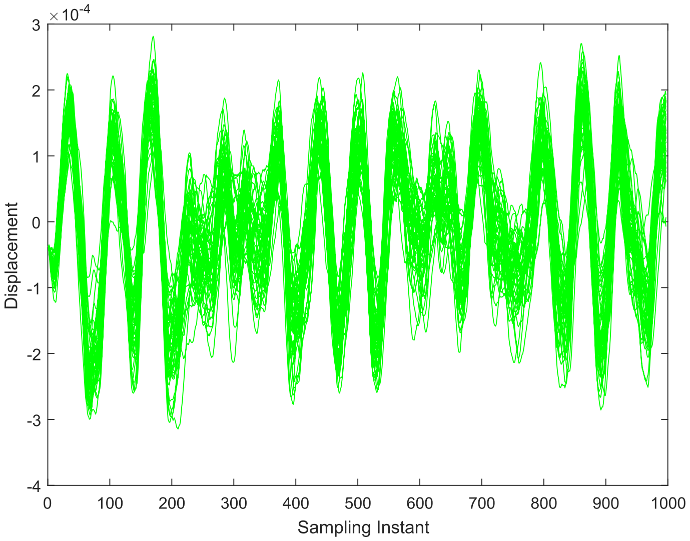


Fig. 9. MC realisations of predictions for GP-NARX model of Duffing oscillator data.

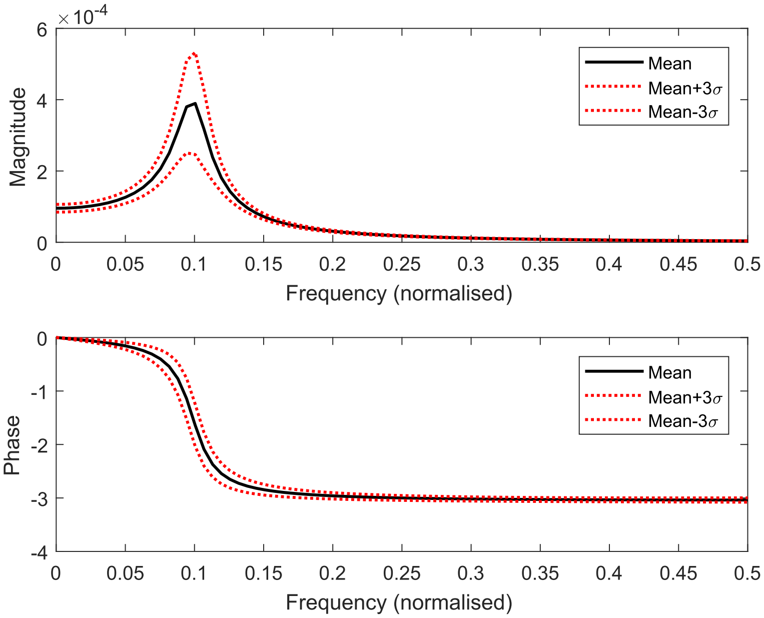
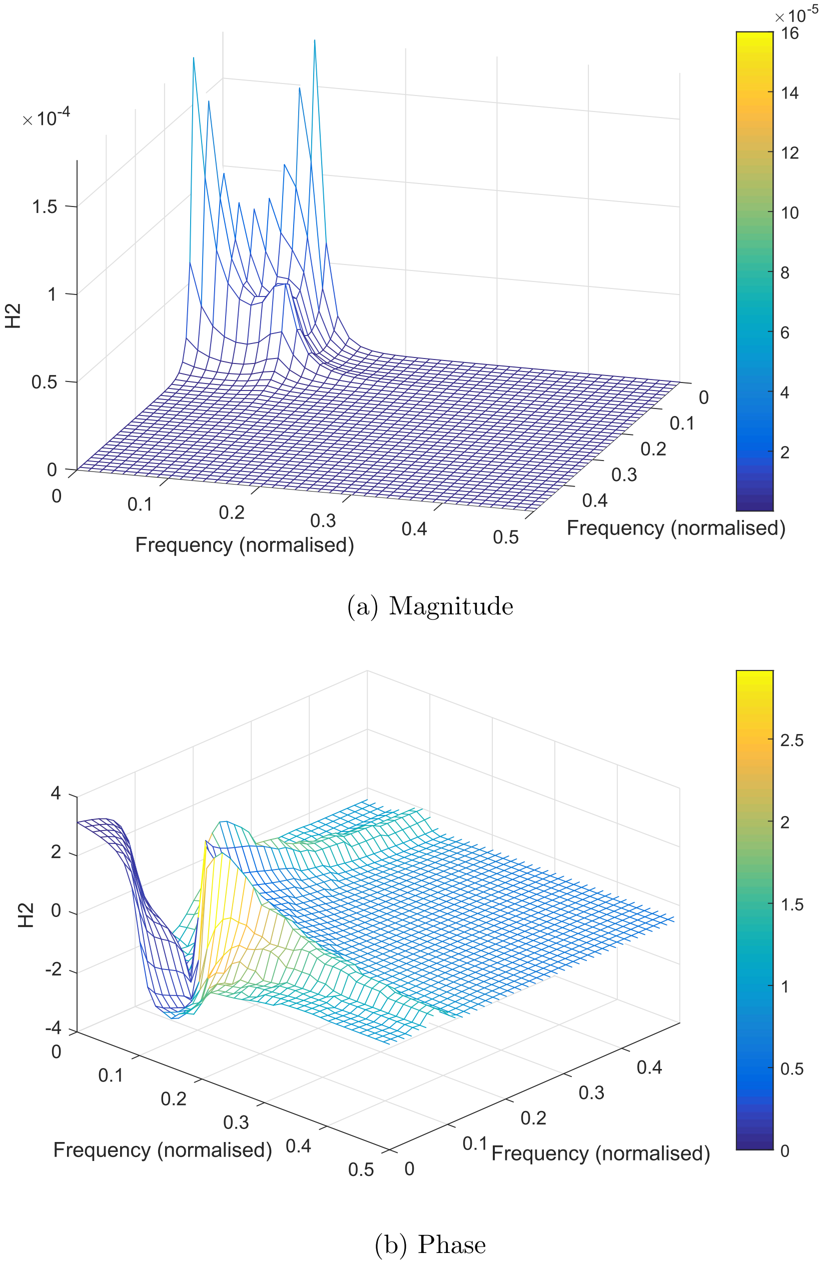


Fig. 10. H1 plotted with mean values over $R = 50$ draws from the GP, and 99% confidence intervals.

that the uncertainty is actually quite narrow in the magnitude plot: although the time domain fits show some scatter, the underlying frequencies contained within each signal are quite similar. In the phase plot, the uncertainty is also quite small.

Fig. 11a shows the mean H_2 HFRF over the 50 draws, with the uncertainty (variance) at each point represented by the colour map. The response is flat at normalised frequencies above about 0.2, and the uncertainty is correspondingly very low. At the lowest frequencies the magnitude increases sharply and the uncertainty is also considerably higher. This is essentially a reflection of what is already visible in Fig. 10, with the additional information that the interaction between the two frequencies is somewhat negligible. The corresponding phase plot in Fig. 11b shows a similar picture: low uncertainty over the large range of frequencies explored, but higher uncertainty in the high-phase regions at the lowest frequencies.

In order to see the important structure in the HFRFs, it is often sufficient to plot only the leading diagonal i.e. $H_2(\omega, \omega)$: see Fig. 12. This format also allows simple comparisons between the functions. This gives extra information over the H_1 plots alone, because the interaction between ω_1 and ω_2 is clearly visible in the first peak of the plot. The confidence intervals,



(a) Magnitude

(b) Phase

Fig. 11. H2 magnitude and phase plotted with mean values over $R = 50$ draws from the GP; colour map indicating standard deviation.

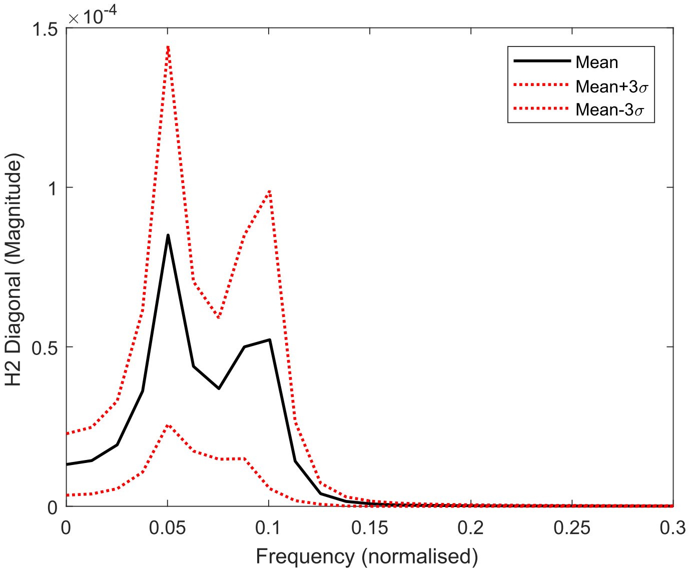


Fig. 12. H2 magnitude at $\omega_1 = \omega_2$: mean values over $R = 50$ draws from the GP and 99% confidence intervals.

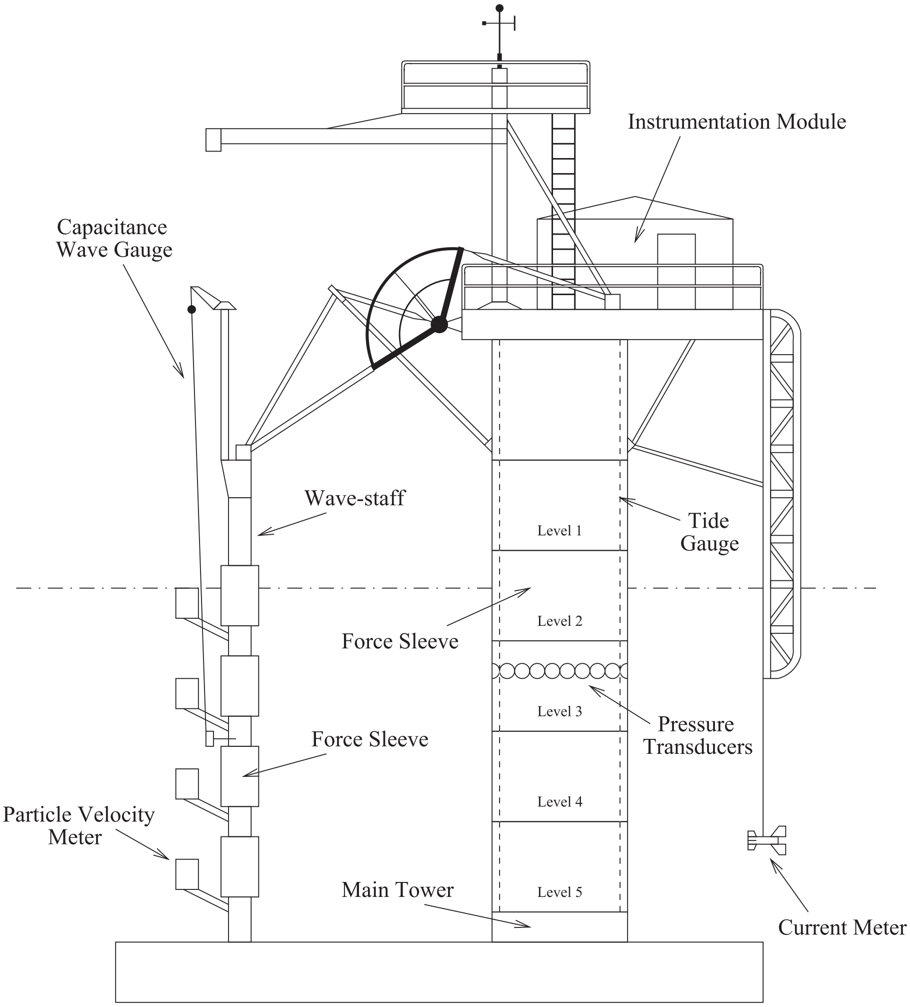


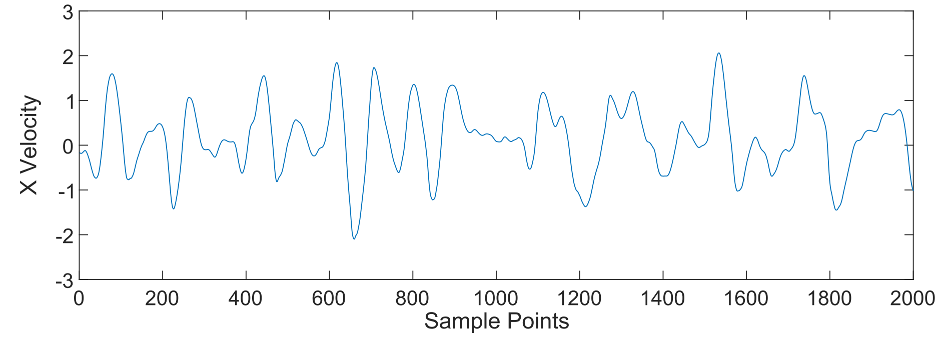
Fig. 13. Schematic of Christchurch Bay Tower.

while wide, reveal that this first peak is not merely due to a spurious fit of the NARX model, because it is present in some way over all Monte Carlo iterations.

Having validated the new theory for the HFRFs and associated uncertainty bounds on simulated data, the ideas will now be applied in a more challenging context i.e. on data associated with wave forces on offshore structures, acquired in a random sea.

8. Wave loading on offshore structures

Since its introduction in 1950 [17], the Morison equation has provided the main means of predicting wave forces on slender cylinders incorporated in offshore structures. In the usual notation,



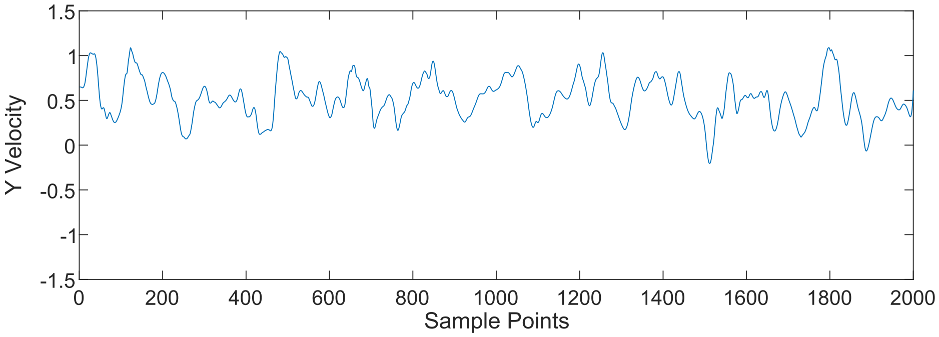


Fig. 14. X and Y-direction velocities for sample of CBT data chosen for training.

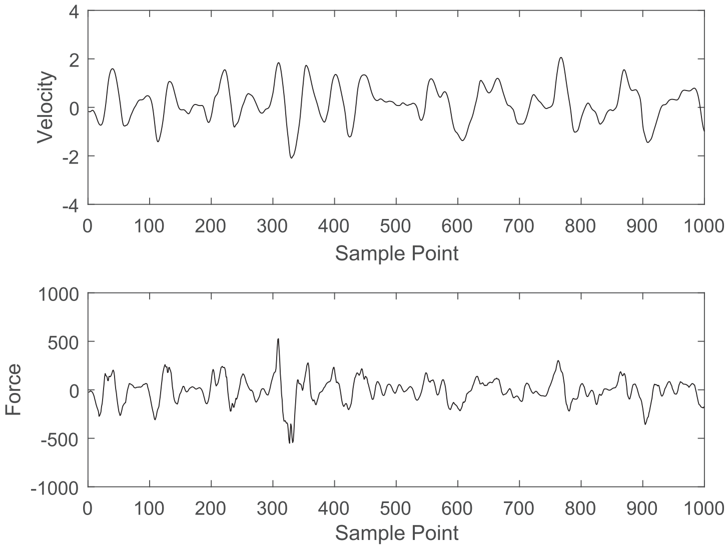


Fig. 15. CBT velocity and force data: training set.

Table 1
Parameter table for Morison model of CBT data from [18].

Model term	Parameter	St.Dev.
u_i	0.88080e+03	0.20344e+02
u_{i-1}	−0.84593e+03	0.20008e+02
u_i^3	0.33983e+02	0.21913e+01

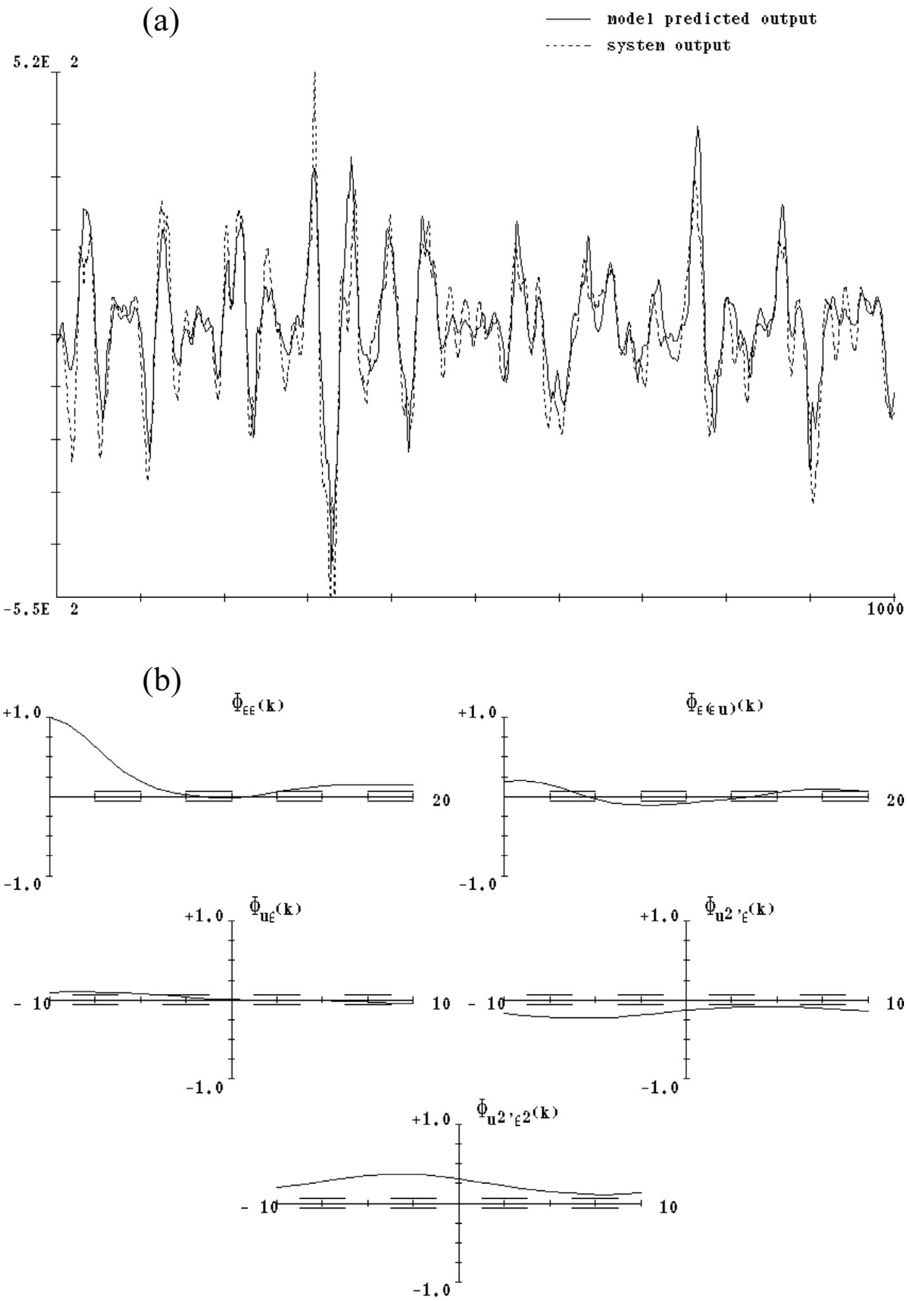


Fig. 16. Results reproduced from [18]: (a) MPO predictions for discrete Morison fit to CBT training data; (b) corresponding correlation test results.

$$F(t) = \frac{1}{2}\rho D C_d u|u| + \frac{1}{4}\pi\rho D^2 C_m \dot{u} \tag{60}$$

where $F(t)$ is the force per unit axial length, $u(t)$ is the instantaneous flow velocity, ρ is water density and D is the diameter of the cylindrical member of interest. The equation is motivated by basic fluid physics and encodes simple processes associated

Table 2
Parameter table for NARMAX model of Christchurch Bay data.

Model term	Parameter
F_{i-1}	0.198e+01
F_{i-2}	−0.126e+01
F_{i-3}	0.790e−01
F_{i-4}	0.395e+00
F_{i-5}	−0.328e+00
F_{i-6}	0.111e+00
u_i	0.119e+03
u_{i-1}	−0.300e+03
u_{i-2}	0.323e+03
u_{i-3}	−0.155e+03
u_{i-4}	−0.946e+01
u_{i-5}	0.273e+02
F_{i-3}^2	−0.193e−03
$F_{i-2}F_{i-5}$	0.137e−03
F_{i-1}^3	−0.232e−05
$F_{i-1}^2F_{i-4}$	−0.193e−05
$F_{i-1}u_{i-4}^2$	−0.221e+00
$F_{i-4}u_i^2$	0.188e+00
$F_{i-3}u_iu_{i-4}$	−0.457e+00
$F_{i-2}u_{i-3}^2$	0.466e+00
$F_{i-1}F_{i-2}u_i$	−0.731e−03
$F_{i-1}^2u_{i-4}$	0.482e−03
$u_{i-3}u_{i-4}^2$	0.437e+02
$u_iu_{i-4}^2$	0.158e+03
$u_{i-1}u_{i-4}^2$	−0.196e+03
F_{i-2}^3	0.101e−04
$F_{i-1}F_{i-2}^2$	−0.222e−04
$F_{i-1}^2F_{i-3}$	0.483e−05
$F_{i-1}^2F_{i-2}$	0.120e−04

with fluid inertia and drag. The dimensionless drag and inertia coefficients C_d and C_m depend on the characteristics of the flow. In general the main dependence is taken to be on *Re*, the Reynolds number, and *KC*, the Keulegan-Carpenter number, although these parameters do not have generally-accepted definitions in random or directional waves. The coefficients C_d and C_m are usually obtained by applying least-squares procedures to measured force, velocity and acceleration data.

It has long been considered obvious that Morison's equation only really represents the main trends in measured data; various characteristics of flows are not well represented at all. In particular, Morison's equation breaks down quite badly if substantial vortex-shedding is present [35]. A rather serious consequence of the weaknesses of Morison's equation is that peak forces can be under-predicted. Furthermore, because the equation does not capture higher harmonic behaviour in periodic or near-periodic flows very well, it is not very helpful in the determination of fatigue lives of structural elements. There have been attempts to capture the higher-harmonic behaviour in extensions of the Morison equation e.g. in [36]; however, it is probably safe to say that there is no universally-accepted simple extension.

While the offshore industry has pursued other means of predicting wave forces e.g. from complex CFD models, and this has helped, it would still be a boon if a 'simple' predictive model could be found which improved on Morison. Some time ago, one of the current authors investigated the idea that system identification techniques could produce data-based models able to improve on Morison. One of the results of this programme was the paper [18], in which NARMAX and NARX models were considered as a potentially promising structure. While the paper made a number of interesting discoveries, it could be argued that the level of rigour of the regression technique falls short of modern machine learning theory and practice [7].

The objective of this part of the current paper is to revisit the wave loading problem, bringing to it the much more modern machine learning technology associated with the GP-NARX model, and a more rigorous approach to learning from data. The GP-NARX model will show some clear advantages over the polynomial forms used in [18]. Before moving on to the analysis of the data, the next two sections will summarise the context of the original problem and briefly revisit the original results so they can be used as a meaningful basis for comparison with the new results from GP-NARX.

8.1. The Christchurch Bay Tower

The data considered in this paper were acquired during a comprehensive campaign of experimental study on a substantial structure in a real directional sea environment off the south coast of the UK. A schematic of the structure – the Christchurch Bay Tower (CBT) – is shown in Fig. 13. Only a brief description of the structure will be given here; it is described in considerably more detail in [37].

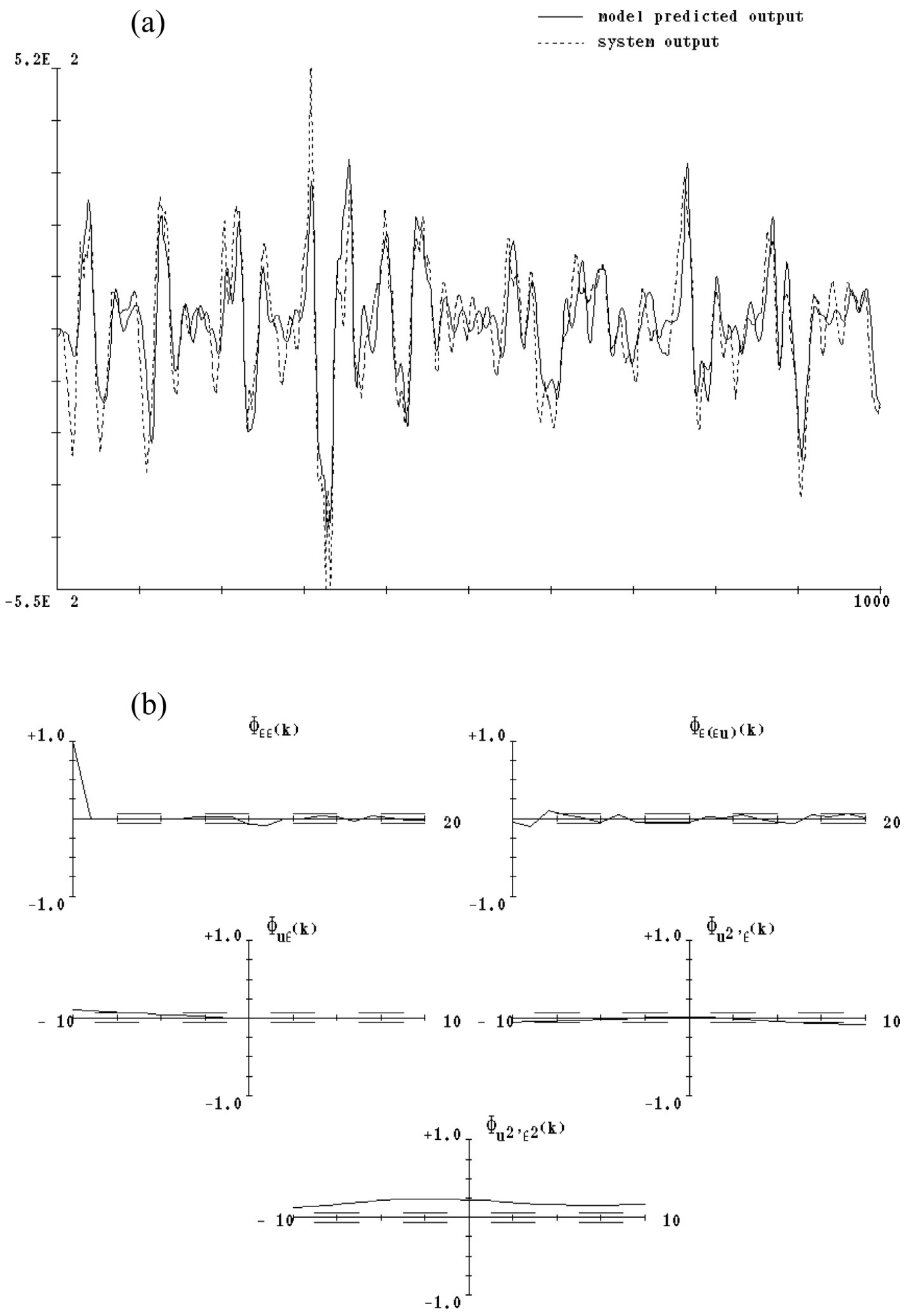


Fig. 17. Results reproduced from [18]: (a) MPO predictions for polynomial NARMAX fit to CBT training data; (b) corresponding correlation test results.

The tower was instrumented with pressure transducers and velocity meters. The data considered here were measured on the small diameter wave staff (Morison's equation is only really appropriate for slender members) illustrated to the left of Fig. 13. Substantial wave heights were observed in the tests (up to 7 m) and the sea was directional with a prominent current. The velocities were measured with calibrated perforated ball meters attached at a distance of 1.228 m from the cylinder axis. This will not give the exact velocity at the centre of the force sleeve unless waves are unidirectional with crests parallel to the line joining the velocity meter to the cylinder. This is called here the Y-direction and the normal to this, the X-direction. The waves are however always varying in direction so data was chosen here from an interval when the oscillatory velocity in the X-direction was maximal relative to that in the Y direction. A sample of 2000 points of velocity fitting this criterion is shown in Fig. 14. It can be seen that the current is mainly in the Y-direction. In this case the velocity ball is upstream of the cylinder and interference by the wake on the ball will be as small as possible with this arrangement.

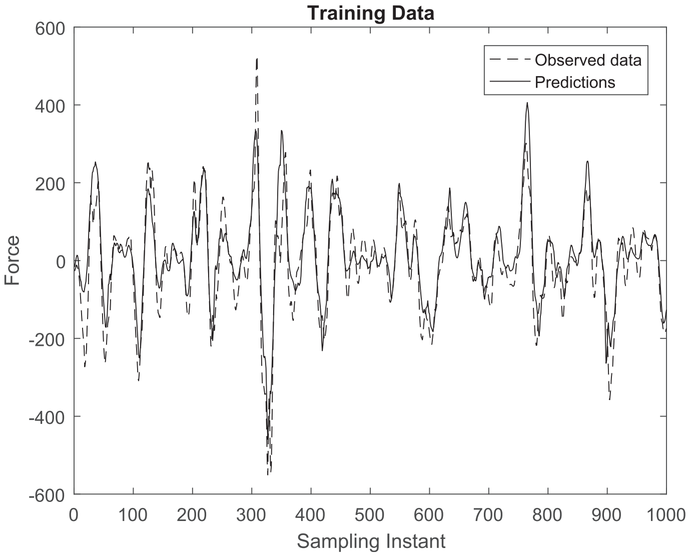


Fig. 18. Morison equation predictions compared to measured CBT data: training set.

The velocity (input) and force (output) samples for the training data (subsampled by taking alternative points) are shown in Fig. 15. One of the issues with the paper [18] in terms of accepted machine learning practice now, was that conclusions were drawn on the basis of fits to *only* training data. In modern terms, this is a mistake; it is now regarded as essential that the generalisation capacity of the model be checked by considering predictions on an independent testing set. In fact, for any model containing hyperparameters [7], one requires three independent data sets: a training set (for parameter estimation), a validation set (for optimising any hyperparameters) and a testing set (as a final judgement on generalisation). In order to select the necessary data for the GP-NARX training data without introducing too much directionality, a 1000 points of data immediately before the training data were chosen as a testing set, while 1000 points immediately after were selected as a validation set, if required. Both sets were again obtained by subsampling 2000-point sets.

8.2. Previous results for the CBT data

An objective of this part of the paper is to compare the efficacy of the GP-NARX model to the that of the polynomial NARMAX models fitted in [18]. In that reference, the discrete form of Morison's equation was fitted to the data to serve as a basis for comparison. The estimated coefficients are presented in Table 1 together with their estimated standard deviations, The *NMSE* for the model was 21.43 which indicated fairly poor agreement with reality. The MPO predictions for the model are shown in Fig. 16, together with the correlation tests. (Up to now, the correlation tests have not been used because the models of the Duffing oscillator system analysed previously were so accurate, that their validity was adequately expressed by the low OSA and MPO prediction errors; on the CBT, the correlation tests will prove crucial.) The natural conclusion was that the model was inadequate. In particular, the correlation tests show many excursions outside the confidence interval; the linear correlation tests show that the residual auto-correlation is far from a delta function. This clearly shows that there are relevant physical processes uncaptured by the Morison model, which are escaping into the residual sequence.

The data were then analysed using a polynomial NARMAX model based on an orthogonal least-squares algorithm incorporating a structure detection method to determine which terms should be included in the model [38]. A linear noise model was included. The resulting model is given in Table 2.[5]

A very complex model was obtained which includes many terms with no clear physical interpretation. The only point of real physical interest is that the coefficients of the terms u_{i-1} and u_{i-2} are roughly equal in magnitude and opposite in sign, and this could be interpreted as the model attempting to form a term linear in the acceleration. It was recognised at the time that this model was probably over-complex and could have been improved by careful optimisation. However, it was considered sufficient at the time to illustrate the main points of the argument for the superiority of the NARMAX model. The explanation supporting the conclusion that the NARMAX model was superior to Morison was the that inadequacy of Morison's equation was due to gross vortex shedding effects which could even be observed in simplified experimental conditions [35]. It was assumed that the nonparametric NARMAX model was capturing some of the missing physics. It is still reasonable to offer this explanation. The NARMAX model predicted output and correlation tests are shown in Fig. 17. Although the validity tests showed a great deal of improvement, the model predicted output was actually a little worse. This is still understand-

[5] The coefficients of the noise model terms are not included in the table. The noise terms are only included in the model as a means of reducing parameter estimation bias, they are not used in making predictions or in computing the HFRFs.

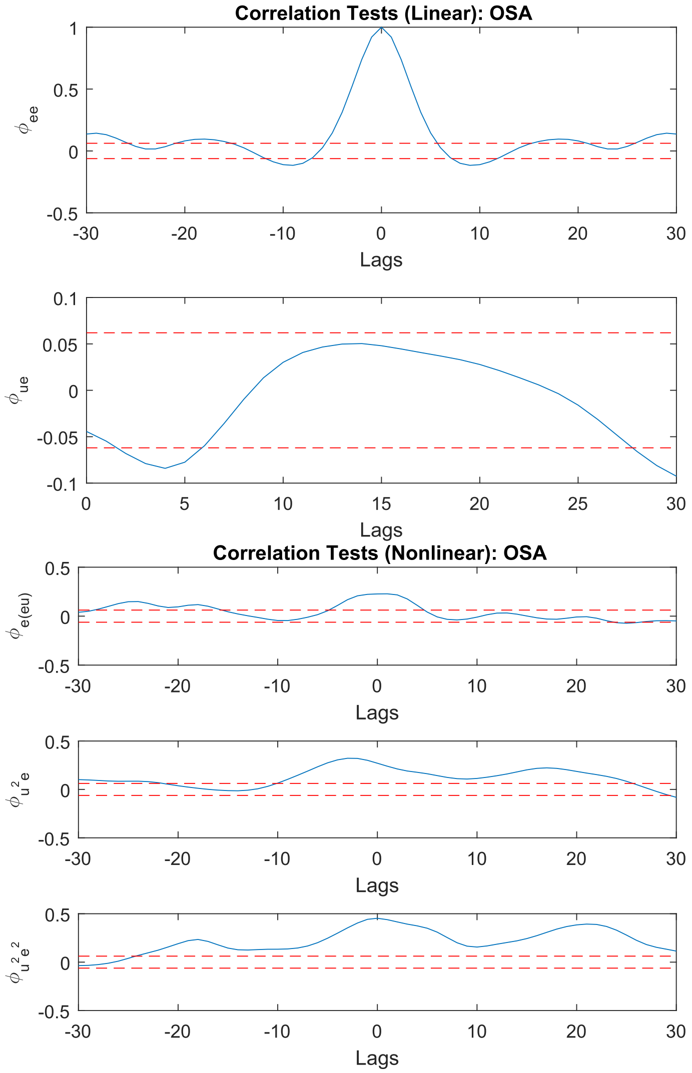


Fig. 19. Correlation tests for Morison model on CBT training data.

able; one of the effects of correlated noise (indicated by the function $\phi_{\epsilon\epsilon}$ in Fig. 16) is to bias model coefficients so that the model fits the data rather than the underlying system. In this case the Morison model predicted output is probably accounting for some of the system noise; this is clearly incorrect. When the linear noise model was added in the NARMAX model to reduce the noise to a white sequence, the unbiased model no longer predicted the noise component and the model predicted output subsequently appeared to represent the data less well. The fact that the final correlation function in Fig. 17(b) still indicated problems with the model can possibly be attributed to the time-dependent phase relationship between input and output resulting from the directional variations of the sea state; the data is actually *nonstationary*. Having summarised the previous work, the paper now moves on to the analysis based on GP-NARX.

8.3. Results from GP-NARX

In this section, results are presented for wave-loading predictions based on the nonparametric GP-NARX form. Note that there will be no attempt to fit a structured noise model in this analysis, so the algorithm truly is NARX, rather than NARMAX. First, it is worth looking back on the Morison analysis from [18]. The analysis there, featured a discrete form of the equation with an additional parameter, rather than the traditional continuous-time form of Morison as presented in Eq. (60). As the acceleration data are actually available from the CBT programme, it is of interest to see how the traditional Morison's equation performs. Morison's equation was fitted to the CBT training data using a basic least-squares parameter estimation, implemented via a pseudo-inverse solution of the normal equations. When this was carried out, the predictions on the train-

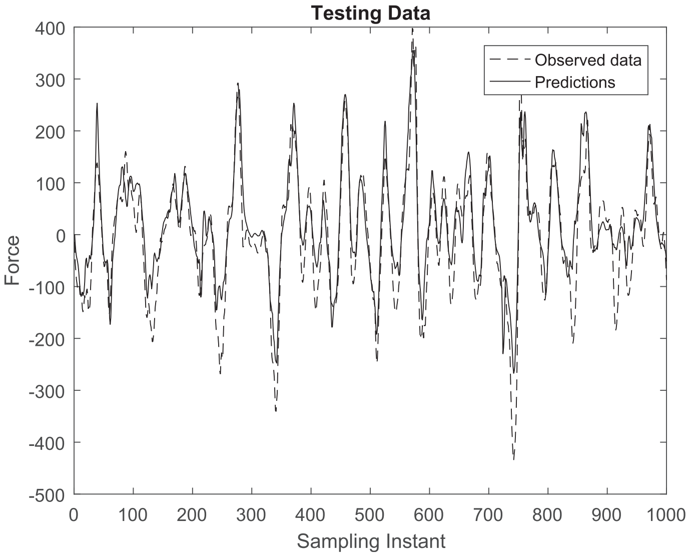


Fig. 20. Morison equation predictions compared to measured CBT data: testing set.

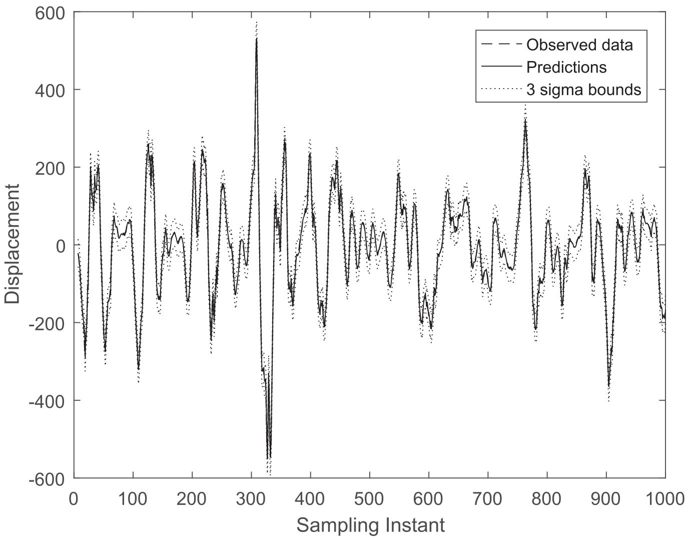


Fig. 21. GP-NARX OSA predictions compared to measured CBT data: training set.

ing data (there is of course no distinction between OSA and MPO predictions for the continuous-time model) a normalised MSE of 20.46 was obtained; this actually improves a little on the discrete-time version discussed in the last section. This is probably due to the aforementioned error on the velocity data, which when differenced, would lead to an erroneous acceleration estimate. The predictions are shown in Fig. 18.

When one considers the correlation tests for the model (Fig. 19), one sees that they are no better than those obtained for the discrete-time equation. In particular, the autocorrelation of residuals is again, far from a delta-function. Following current best practice in machine learning, the model was then used to predict on the independent testing set. Somewhat surprisingly, a very similar NMSE is obtained i.e. 19.52; the results are as shown in Fig. 20.

The analysis now moves to the GP-NARX model. One of the failures of [18] in terms of rigour, was the failure to recognise that the lag numbers n_x and n_y are actually *hyperparameters*. If [18] had actually used an independent testing set, it is possible that this would have been recognised; however, the opportunity was missed. One correct approach to determining n_x and n_y would be by cross-validation on the independent validation set; however, for the sake of brevity of the current paper, only an $n_x = 5, n_y = 6$ model will be given, in order that a direct comparison can be made with the polynomial NARMAX model summarised in Table 2. When a model of this type was fitted to the training data, an NMSE of 0.77 was obtained for the OSA

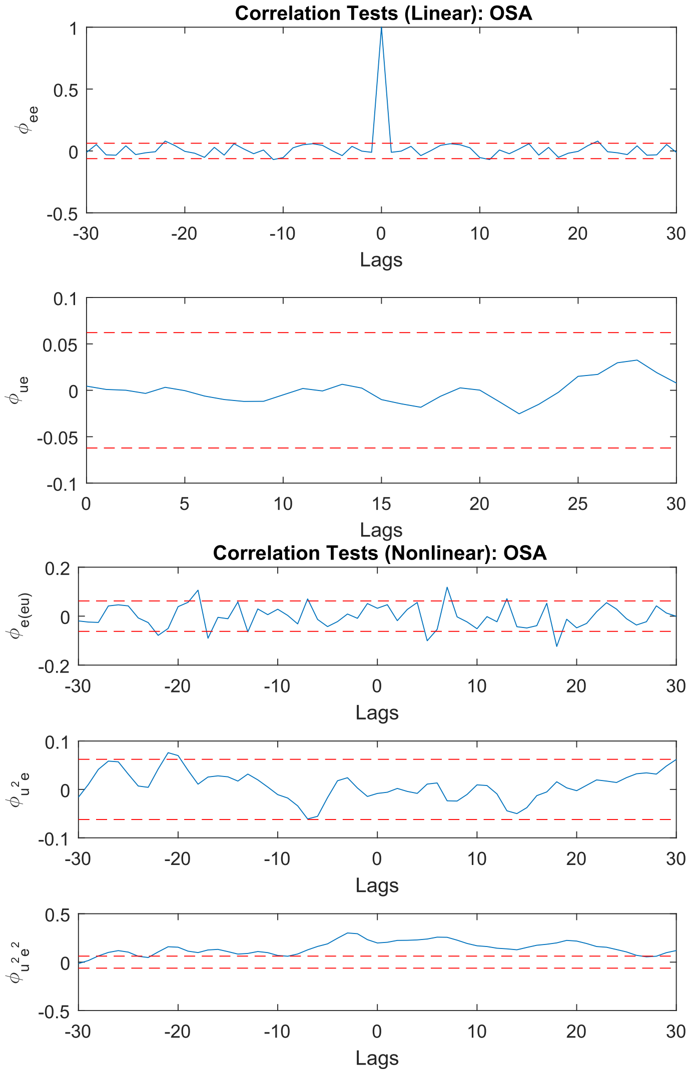


Fig. 22. Correlation tests for GP-NARX model on CBT training data.

predictions and one of 18.01 for the MPO predictions. The OSA and MPO predictions are compared to the measured data in Figs. 21 and 23 respectively. Furthermore, the correlation tests for the model show a very marked improvement over Morison (Fig. 22). The correlation tests for the GP-NARX are at least as good as those for polynomial NARMAX in [18], despite the fact that the GP-NARX form does not incorporate a noise model. One is perhaps forced to speculate that the noise model obtained in [18] did not, in fact, accomplish very much.

Another point to note is that the GP-NARX predictions are naturally accompanied by confidence intervals, and that over most of the data, even when the detailed predictions are in error, the confidence intervals include the measured data. This is a major advantage of the GP-NARX form. While it would be possible to obtain confidence intervals for the polynomial NARX models, it would require Monte Carlo analysis and would potentially be computationally labourious (see e.g. [39]).

As discussed earlier, the confidence intervals do not quite accommodate the observed prediction errors. This because not all of the uncertainty has been accounted for. In the predictions so far, the predicted outputs have been fed back into the model in order to form the MPO. This means that the only uncertainty accounted for in the predictions is the *parameter* uncertainty. In order to show the full predictive uncertainty, one should carry out a Monte Carlo analysis as before, sampling from the predictive distributions at each time step. Fig. 24 shows 10 MC realisations of the predictions superimposed on the measured data.

More uncertainty is associated with the predictions now. From the MC realisations, one computes the mean predictions and determines the $\pm 3\sigma$ confidence bounds, and the result of the analysis for the case here is shown in Fig. 25. The confidence intervals are now a more appropriate assessment of the predictive capability of the model. One can see here a major

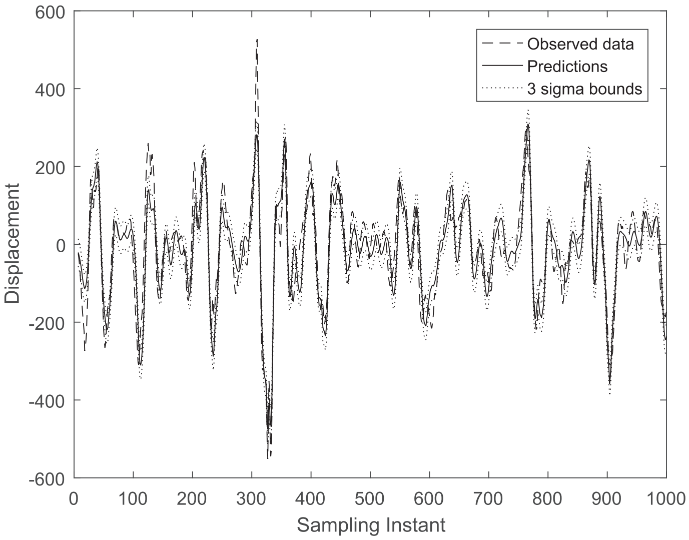


Fig. 23. GP-NARX MPO predictions compared to measured CBT data: training set.

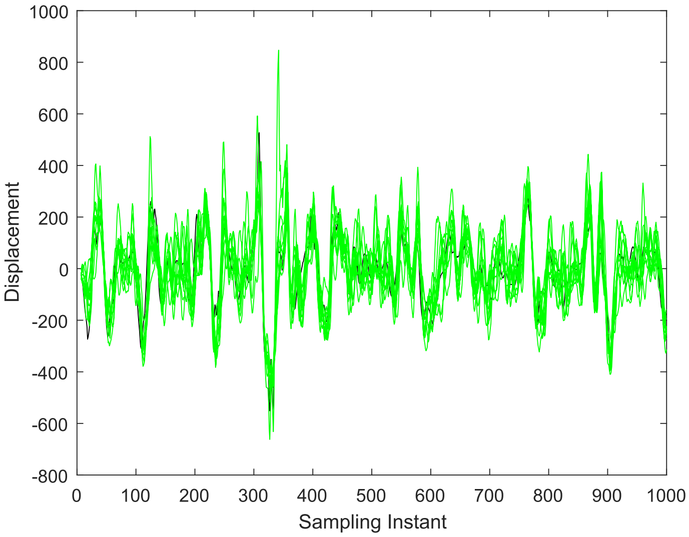


Fig. 24. 10 MC realisations (green) of the GP-NARX MPO predictions (incorporating full predictive uncertainty) compared to measured CBT data (black): training set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

advantage of the GP-NARX formulation; even though the model does not accurately capture all of the behaviour, it is clearly realistic about the amount of uncertainty associated with its predictions.

So far, things seem to be going rather well. The GP-NARX model appears to match or improve on the polynomial NARMAX model in raw terms, but also has the marked advantages of: (a) not requiring any structure detection – the form is nonparametric, and (b) providing confidence intervals that appear to represent the true predictive uncertainty of the model. Unfortunately, it now becomes necessary to confront one of the limitations of the analysis in [18], and things will temporarily unravel a little.

In keeping with modern machine learning practice (and, if truth be told, the more careful practices in system identification at the time of [18]), the model obtained above *must* be evaluated on an independent testing set. As mentioned in Section 8.1, such a testing set has been set aside. When the GP-NARX model is used to make predictions on the testing set, the OSA and MPO NMSEs rise to 1.03 and 27.04 respectively. This represents a marked degradation on the MPO error compared to the training set. The predictions are shown in Figs. 26 and 28. This is a disappointment and highlights a breakdown in best practice in [18]. Furthermore, the correlation tests also degrade somewhat on the testing set (Fig. 27). However, a closer look provides consolation. First of all, although the correlation tests on the testing set do degrade, they are still a considerable

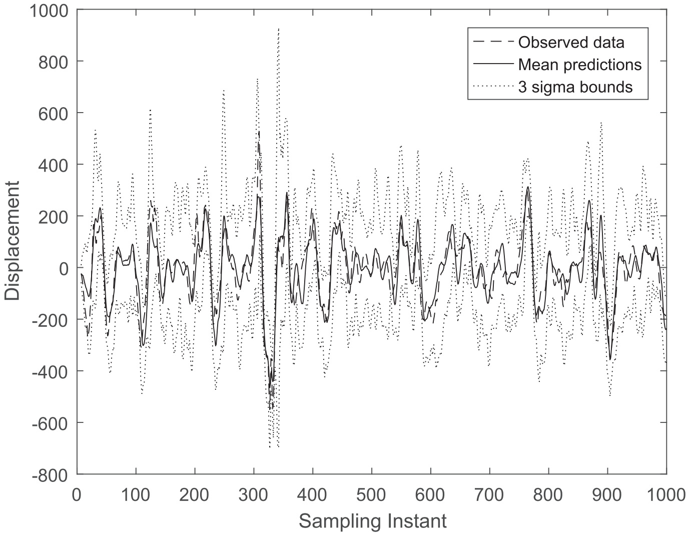


Fig. 25. GP-NARX MPO predictions with confidence bounds incorporating full predictive uncertainty, compared to measured CBT data: training set.

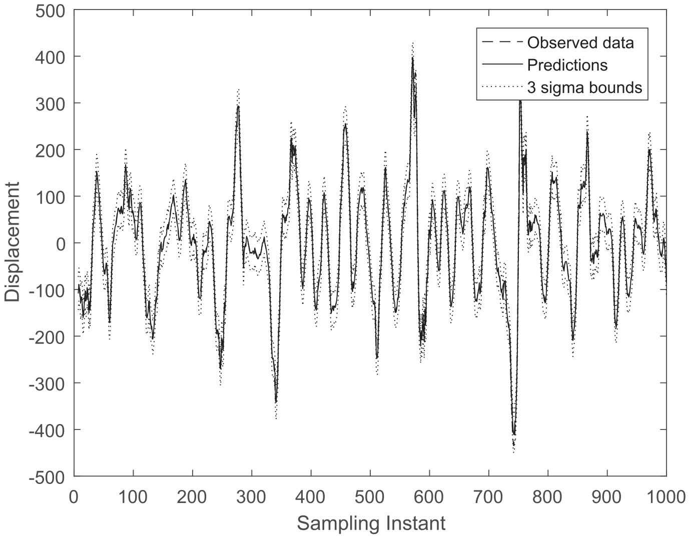


Fig. 26. GP-NARX OSA predictions compared to measured CBT data: testing set.

improvement on those from the Morison equation; in fact the autocorrelation of the residuals is much better. Secondly, if one considers the confidence intervals; when the predictions degrade on the testing set, the confidence intervals appear to expand somewhat to express reduced confidence. The latter feature is a real advantage of the GP-NARX formulation; the effect can be seen in Figs. 29 and 30.

8.4. HFRFs for the CBT data

The analysis can now move on to the estimation of the HFRFs and their confidence bounds for the CBT data, using the new theory. First of all, results for H_1 and H_2 without uncertainty analysis will be presented. Fig. 31 shows the estimated $H_1(\omega)$ function. The frequencies in the figure (and in subsequent figures) are normalised such that the Nyquist frequency is π rad/s. The results are very similar to those obtained in [18].

Figs. 32 and 33 show, in contour map form, the magnitude and phase respectively, of the $H_2(\omega_1, \omega_2)$ HFRF. The frequency ranges have been restricted to one half of the Nyquist frequency because the of the nonlinear wrap-around effect present in HFRFs of order two and higher [12].

Finally, Fig. 34 shows the diagonal $H_2(\omega, \omega)$ estimated from the data.

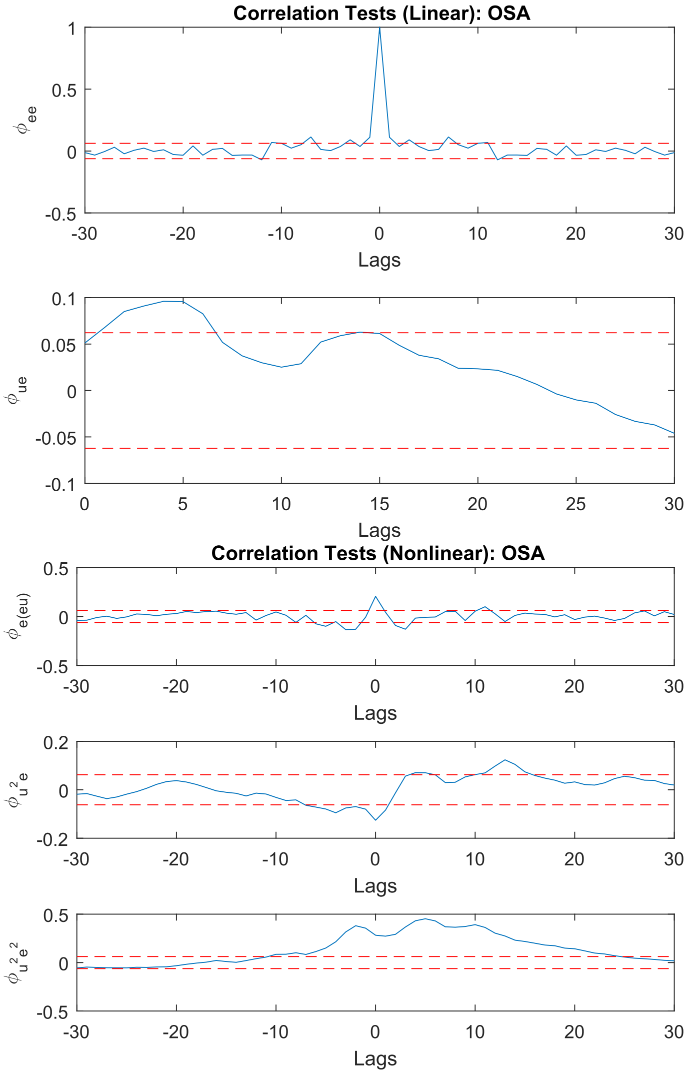


Fig. 27. Correlation tests for GP-NARX model on CBT testing data.

So far, there are no particular surprises; the results are broadly similar to what was observed in [18]. However, the new approach to uncertainty propagation can now be applied in order to investigate confidence intervals on the CBT HFRFs. The algorithm was applied exactly as it was in the case of the Duffing oscillator case study earlier; the only real difference being that 100 draws were taken for the MC samples, the reason for this will be explained a little later. Fig. 35 shows the 100 draws for the $H_1(\omega)$ function, and Fig. 36 shows the corresponding mean H_1 function with its associated $\pm 3\sigma$ confidence intervals.

It is important to mention something which occurred in generating the 100 MC realisations for the CBT GP-NARX models. In a number of the realisations, the response suddenly jumped to a very high amplitude and oscillated with a very high frequency. Fig. 37 shows one such realisation. At first this was a concern, but it was rationalised as follows. The GP-NARX model is a highly-nonlinear dynamical system, and as such, it is likely to have multiple equilibria. Because of this, it is likely that some MC realisations may generate sequences of responses that allow an escape into the domain of attraction of one of these equilibria. This would only be a concern if one of the escapes manifested in predictions on measured data, rather than in the MC realisations. in fact, this did not occur, but to make a more careful check, the fitted model was used to make predictions on the entire CBT run 53 data, which amounted to 12416 data points. As anticipated, no escapes occurred on the MPO predictions; the overall NMSEs were 0.74 for the OSA predictions, and 23.2 for the MPO. These are a little higher than on the training data, but no worse than those obtained on the 1000-point testing set discussed earlier.

Fig. 38 shows a period of the predictions where one of the realisations generates an escape; in general, these appear to occur when there is some kind of peak in the training data, combined with a few high random numbers in succession. This seems to be enough that the $||x - x'||$ term in the covariance function is very large, i.e. the distances between the test point and the training points are large, which makes the covariance go close to zero. It seems that once the prediction has passed

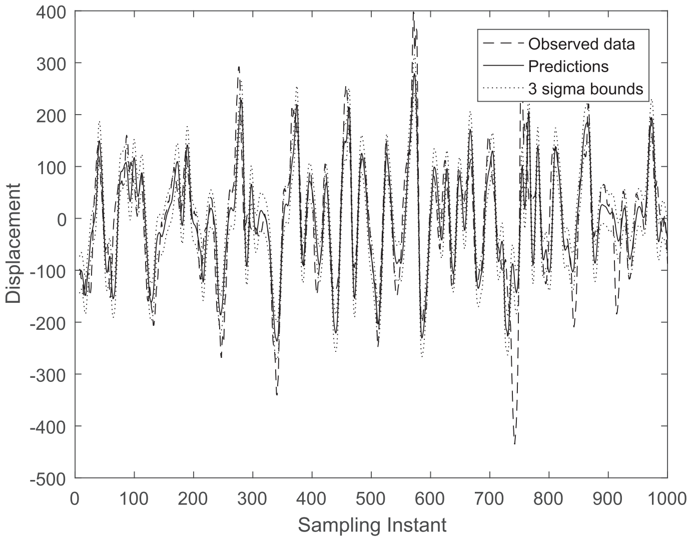


Fig. 28. GP-NARX MPO predictions compared to measured CBT data: testing set.

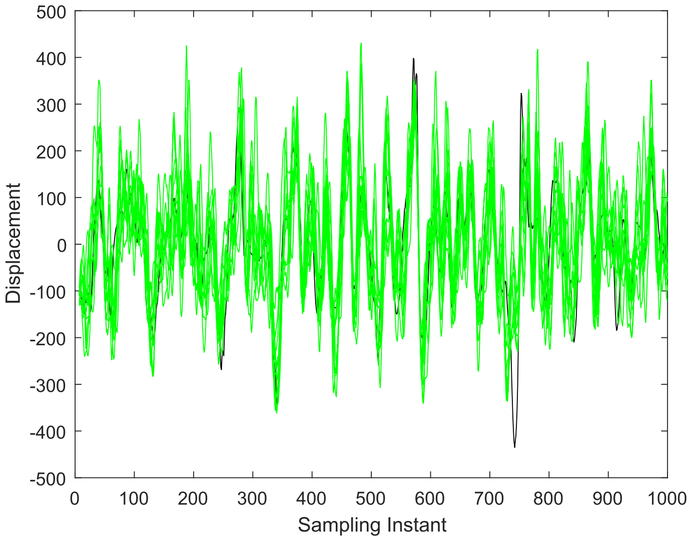


Fig. 29. 10 MC realisations (green) of the GP-NARX MPO predictions (incorporating full predictive uncertainty) compared to measured CBT data (black): testing set. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

some threshold, it basically forgets about the lagged y information in the time series. In order to investigate a little further, models were fitted with higher numbers of y lags, based on the hypothesis that this would make the measured output behaviour harder to forget. This hypothesis was supported by the small number of possibilities considered: if the number of y lags was increased to eight, four draws from 50 generated an escape; when the number of lags was increased to ten, only 3 escapes in 50 occurred. This is by no means conclusive, but in many ways it does not matter too much in practice. Having established that escapes can occur in the MC process which are associated with a non-physical state of the GP-NARX, and further having established that the escapes to not occur on measured data, a simple solution to the problem appears to be discarding any draws containing an escape as the remaining draws will then be representative of the (physical) equilibrium of interest. This was the approach taken in generating the confidence bounds for the CBT data here.

The second point here, is that it is important to take a little care in how the confidence intervals are generated for the HFRF plots. Consider the magnitude plot in particular. It is incorrect to generate confidence intervals based on Gaussian statistics on the magnitude draws, for the simple reason that the FRF magnitude is positive semi-definite and thus cannot have a Gaussian density for a frequency line. One must estimate the confidence intervals on the real and imaginary parts of the HRFs separately, and then convert to magnitude and phase. Considering this fact leads one to realise that one can quite

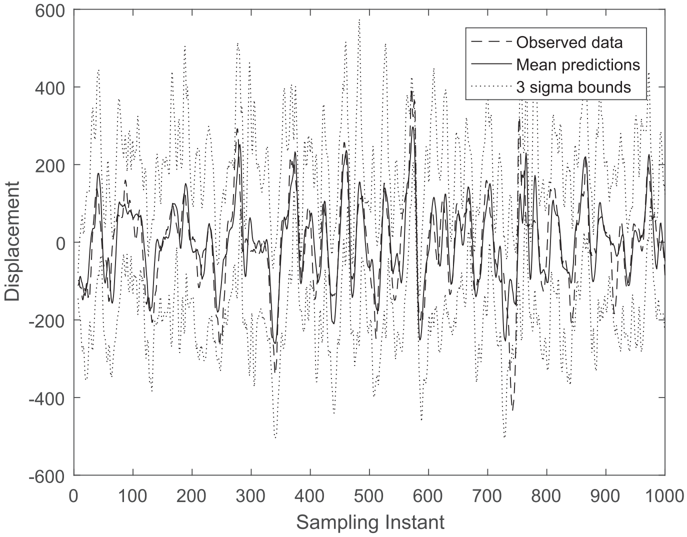


Fig. 30. GP-NARX MPO predictions with confidence bounds incorporating full predictive uncertainty, compared to measured CBT data: testing set.

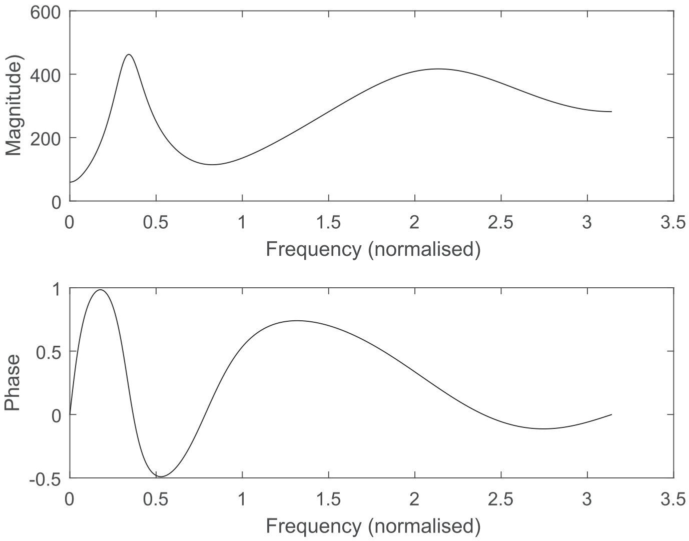


Fig. 31. H_1 for the CBT training data, estimated from the GP-NARX model.

happily use other statistics in order to estimate confidence bounds e.g. more robust statistics such as percentiles. The advantage of percentiles is that they do not make the assumption of any particular probability distribution. Fig. 39 shows confidence bounds from taking the first and 99^{th} percentiles on each frequency line. Incidentally, this was the reason for increasing the number of MC samples to 100 for this example, and clearly higher percentiles and increased accuracy could be obtained by increasing the sample size still further. In fact the number of draws can easily be increased e.g. to 1000 and above because the computational cost of doing so is quite low.

Moving on to H_2, Fig. 40 shows the 100 MC draws from the diagonal $H_2(\omega,\omega)$, and Fig. 41 shows the mean of the distribution and the $\pm 3\sigma$ confidence intervals. One immediately observes that there is greater uncertainty associated with this quantity than with the H_1 FRF in Fig. 36; this is expected, and is because the H_2 function is composed of multiples of the H_1 function at different frequencies. In the case of the Duffing oscillator, this is clearly shown in Eq. (33); however, this is a quite general property of HFRFs as discussed in [12]. The increased uncertainty in estimated HFRFs as their order increases is also found in [39], which takes a different approach to MC analysis used to estimate confidence bounds on HFRFs.

Fig. 41 also illustrates a very useful by-product of this analysis. Note that between the normalised frequencies of around 0.1 and 0.25, the lower bounds of the confidence intervals are above zero. This can be interpreted as saying that there is 99.7% certainty that there is a non-zero $H_2(\omega,\omega)$; this in turn, means that there is 99.7% certainty of an even nonlinear effect.

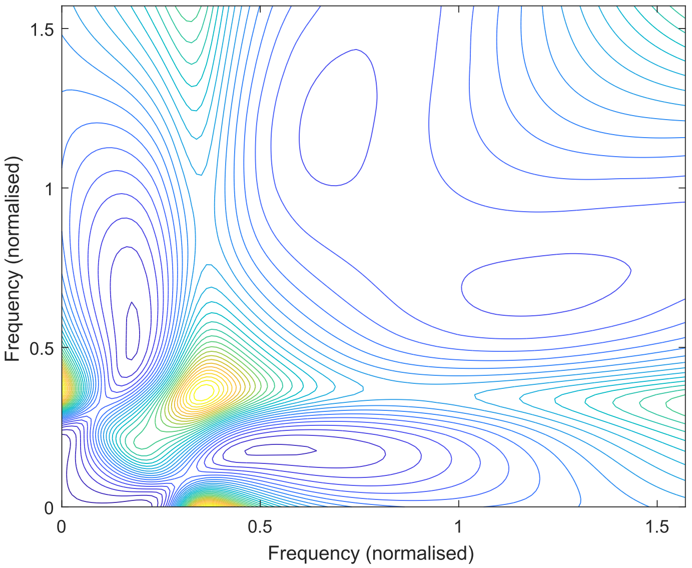


Fig. 32. H_2 magnitude for the CBT training data, estimated from the GP-NARX model.

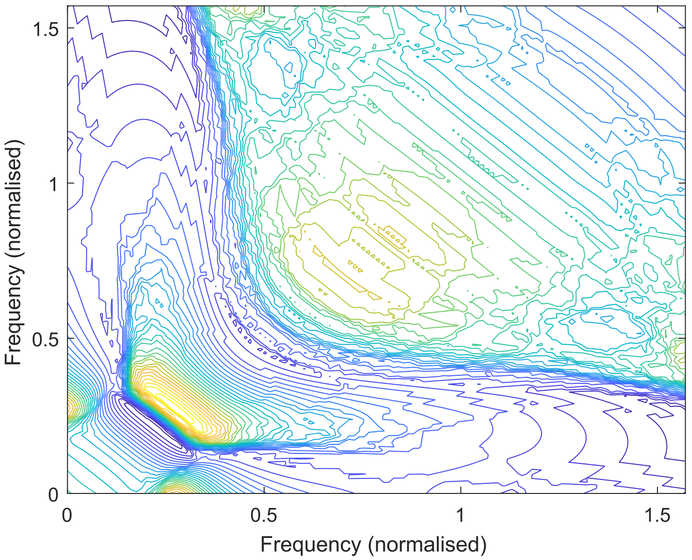


Fig. 33. H_2 phase for the CBT training data, estimated from the GP-NARX model.

The figure serves as a detector, not only of nonlinearity, but of even nonlinearity. Furthermore, the figure establishes which frequency range the nonlinearity is clearly detected over. (Note that, if the confidence bounds extended down to zero across the whole range, this would not establish linearity, but could neither be interpreted as detection of nonlinearity at the given confidence level.) Using the diagonal $H_3(\omega, \omega, \omega)$ in the analysis would similarly establish a detection test for odd nonlinearity. In terms of the information it delivers, the proposed test for nonlinearity bears some resemblance to that of Schoukens and co-workers [40–42], which is based on FRF distortions. However, there would be pros and cons of the two approaches. A disadvantage of the new test is that it does not have anything like the discriminatory power of the Schoukens test; advantages could be that the new test does not depend on specifically designed excitations (multisines) and does not depend on averaging, so can be used on small time data sets.

It is interesting physically, that the analysis here detects a quadratic nonlinearity. Quite apart from the fact that Morison's equation has an odd nonlinearity in the velocity, one might expect that quite generally the wave force on a member should be symmetrical in terms of wave direction. However, this is to forget that a current is present in the CBT case, and even though it is predominantly in the Y-direction, it is clearly sufficient to introduce an even component in the drag nonlinearity.

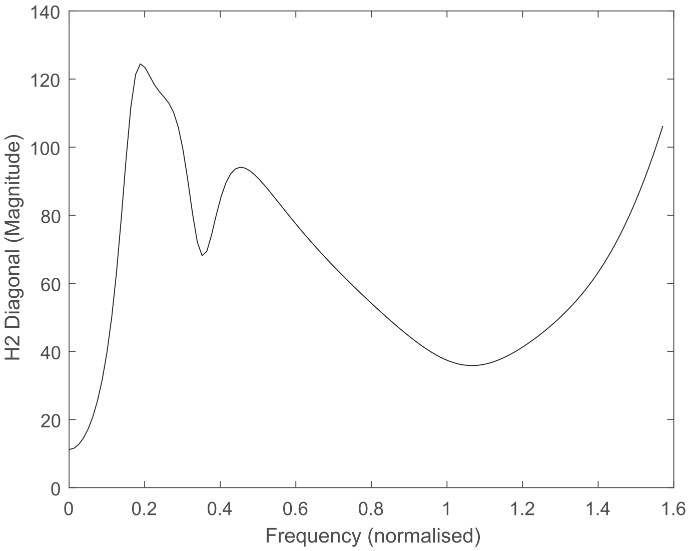


Fig. 34. H_2 magnitude diagonal for the CBT training data, estimated from the GP-NARX model.

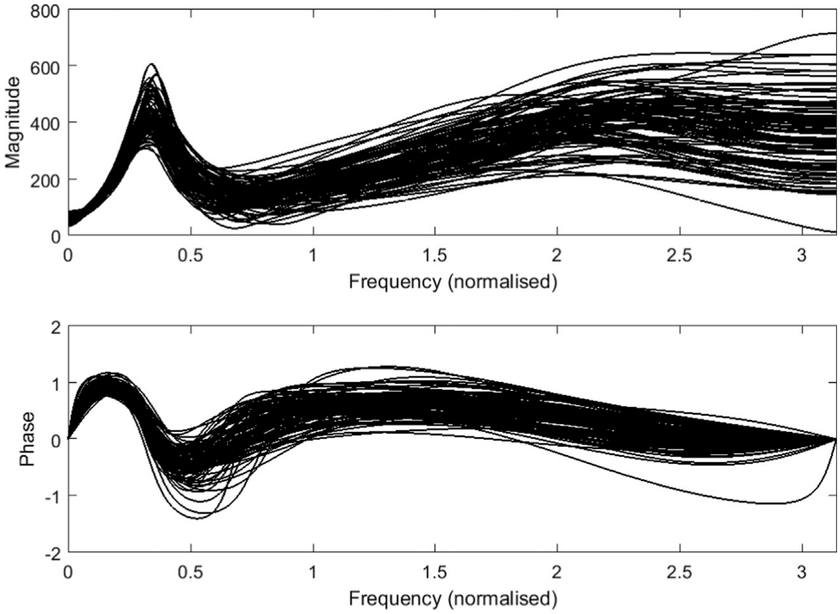


Fig. 35. 100 MC draws from the H_1 distribution obtained from the GP-NARX analysis of the CBT data.

As discussed earlier, different statistics can be used to establish the confidence bounds; Fig. 42 shows the mean diagonal H_2 together with confidence limits corresponding to the first and 99^{th} percentiles. There are two interesting features to this figure; the first is that the more robust statistics used, also allow a clearer accommodation of skewness in the confidence intervals for each frequency line. The second observation is that, using these statistics, the even nonlinear effect is detected at the 98% confidence level, across the entire frequency range.

8.5. Discussion of CBT results

The results presented here show good agreement with the previous analysis in [18], where agreement could be expected. Apart from that, the new approach here, based on the GP-NARX model has a number of clear advantages; foremost among these are the avoidance of the structure detection test and the principled Bayesian framework that allows the computation of confidence intervals on predictions. However, it should be mentioned that Bayesian analysis of polynomial NARX models is

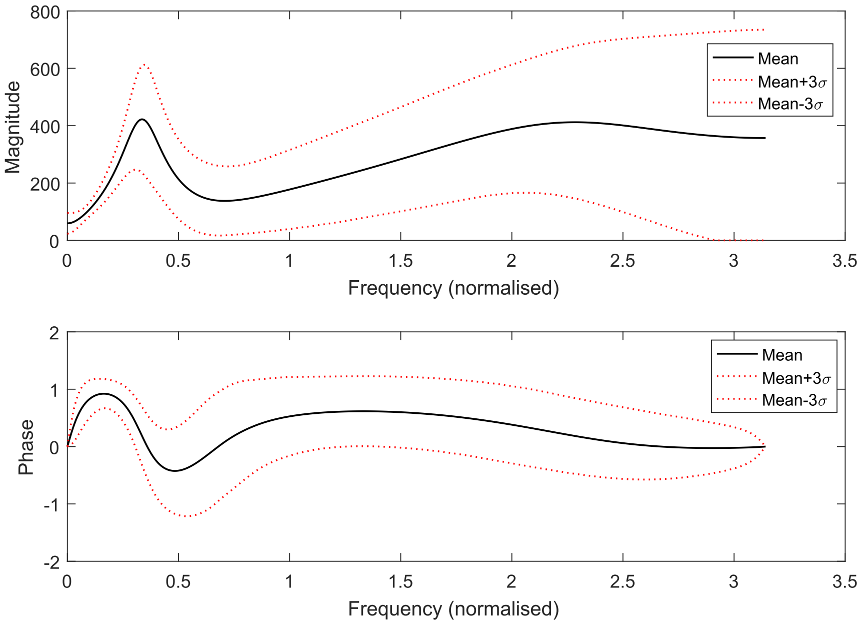


Fig. 36. Mean $H_1(\omega)$ estimated from the GP-NARX analysis of the CBT data, together with $\pm 3\sigma$ confidence intervals.

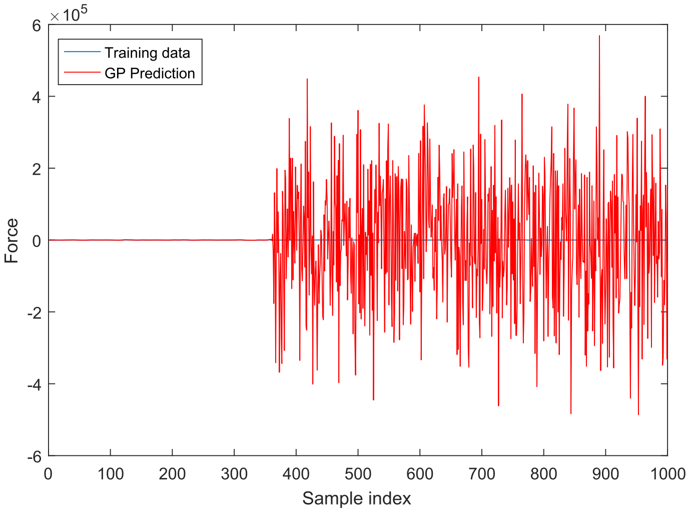


Fig. 37. MC realisation of the GP-NARX predictions showing an escape to a non-physical high-amplitude high-frequency solution.

possible; interesting recent references are [43,44]. In the former paper, a Reversible-Jump Markov Chain Monte Carlo (RJ-MCMC) approach is taken [45]; the approach allows structure detection via the insertion and deletion of polynomial terms via birth/death moves in the RJ-MCMC algorithm. The approach is quite demanding computationally, but does allow the fitting of a noise model. In contrast, the approach taken in [44] is conceptually simpler and makes use of an algorithm for Bayesian parameter estimation in generalised linear models [46]. The simpler approach accomplishes structure detection via an *Automated Relevance Determination* (ARD) on the polynomial model terms, but cannot fit a noise model.

Perhaps the most clear restriction imposed in the analysis of this paper, was to a Single-Input–Single-Output (SISO) model. While the Christchurch Bay Tower was certainly not placed in the most directional sea imaginable, the effects of direction are clearly visible in the data. The effects were reduced as far as possible here, by optimising the training data so that a window was chosen which had the largest relative proportion between the X-direction variance in the velocity and the Y-direction variance. (This choice was not taken solely to produce a SISO problem, it was also because, if the waves are not propagating in the X-direction, the velocity meter is not at the same point in the wave front as the wave staff.) The immediate effect of choosing the training data in this manner is to ensure that any validation or testing data will be *more*

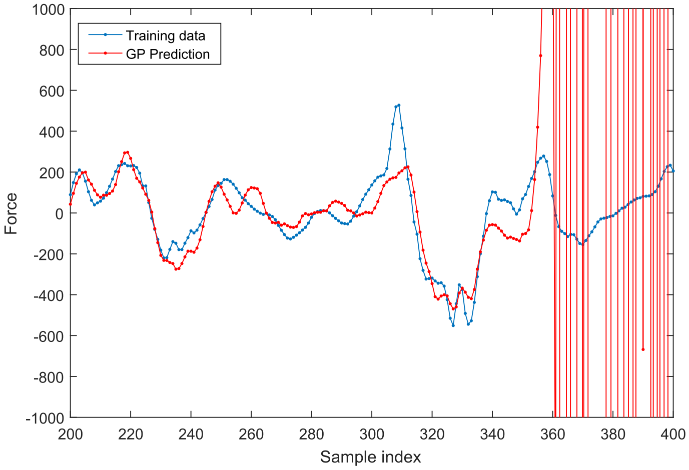


Fig. 38. Zoom of Fig. 37 showing detail of escape point.

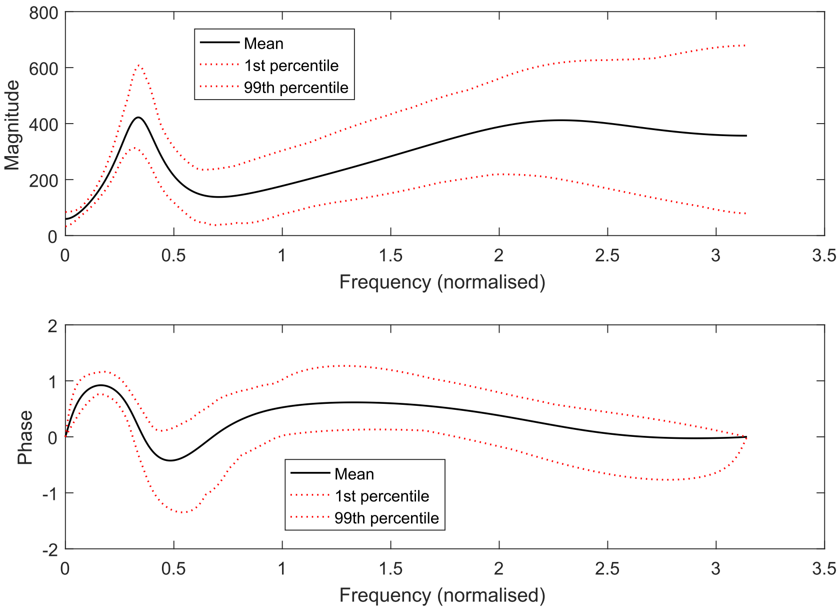


Fig. 39. Mean $H_1(\omega)$ estimated from the GP-NARX analysis of the CBT data, together with confidence intervals from first and 99^{th} centiles.

directional; this is presumably one reason why there is a noticeable degradation in going from the training data to the testing data in terms of model predicted outputs. Fortunately, it is a straightforward matter to generalise the GP-NARX structure to provide a Multi-Input–Single-Output (MISO) structure. Furthermore, as both X- and Y-direction data are available from the CBT experiment, for velocities, accelerations and forces, it is a simple matter to benchmark the MISO structure on the wave loading problem. One imagines, that the most powerful model paradigm would provide a MIMO structure allowing the prediction of correlated X- and Y-direction forces. While this latter option does not appear to be entirely straightforward for Gaussian processes, other algorithms, like GP regression networks [47], could rise to the challenge. Perhaps what is indicated here, is a backward step to the simplified environment of a wave flume (data from the De Voorst facility were analysed in [18] along with the CBT data); however, even in that situation, vortex shedding can and will produce out-of-line forces which would present problems to the SISO approach.

One can also think of the directionality as a type of nonstationarity of the data, so it is then natural to think that perhaps one can exploit heteroskedastic versions of the Gaussian process [48,49] in the formulation of NARX models.

One of the issues not considered here, is that of hyperparameters. For the sake of brevity of the current paper, a single specification of the lag numbers n_y and n_x was considered here in order to facilitate comparison with the results of [18]. In reality, the hyperparameters should be determined by cross-validation on an independent data set. A more satisfactory

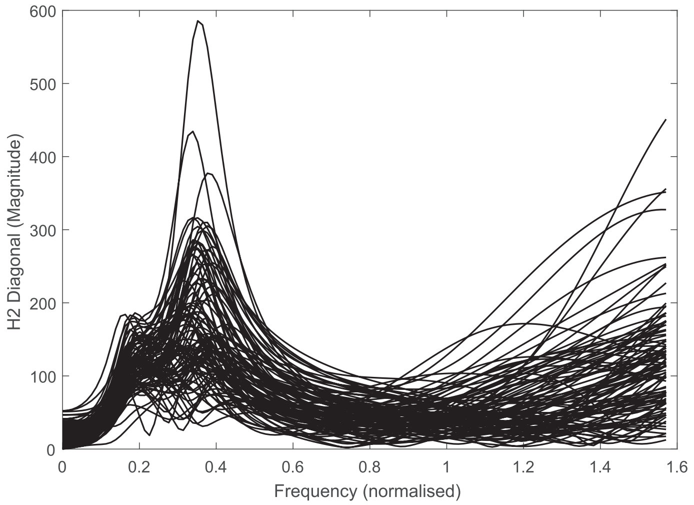


Fig. 40. 100 MC draws from the $H_2(\omega,\omega)$ diagonal distribution obtained from the GP-NARX analysis of the CBT data.

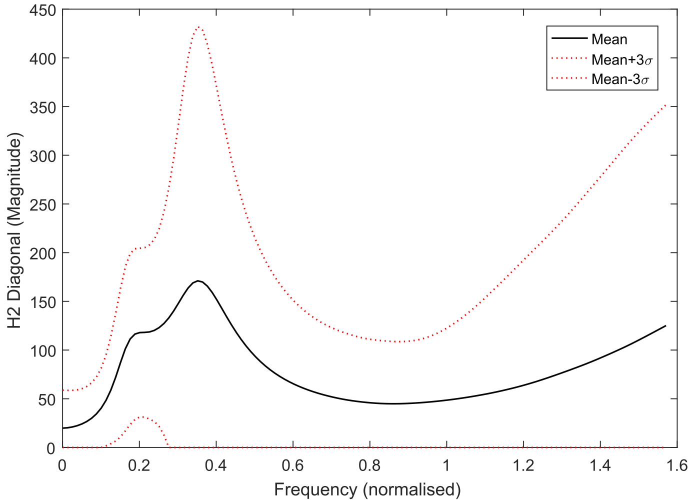


Fig. 41. Mean diagonal $H_2(\omega,\omega)$ estimated from the GP-NARX analysis of the CBT data, together with $\pm 3\sigma$ confidence intervals.

solution still would be to marginalise over the lag numbers via a weighted average of models. One might also expect that fixing hyperparameters on the basis of cross-validation on an independent data set might would provide some defence against the fragility created by choosing the training set in such a particular manner as here.

The desirability of estimating the HFRFs does not constrain the approach to SISO models; in fact, a harmonic probing algorithm for MISO models has been available for some time [50] and is currently being applied to the extraction of HFRFs for multi-input GP-NARX models.

9. Conclusions

The main aim of this paper has been to present analytical expressions for the HFRFs of Gaussian process NARX models (specific to the squared-exponential covariance function). The expressions have been validated on simulated data from an SDOF nonlinear system where the theoretical results are known. The excellent agreement between the exact HFRFs and those extracted from the GP-NARX model fitted to simulated data confirm that the expressions are correct. Perhaps more importantly, the results show that it is possible to obtain accurate estimates of system HFRFs by using GP-NARX models. This in itself is not a surprise as previous work had shown that the HFRFs could be extracted from neural network NARX models learnt from data. However, the GP form of the NARX model has a major advantage over the previous crisp neural network models in allowing the computation of confidence intervals for predictions, in turn leading to a means of establishing

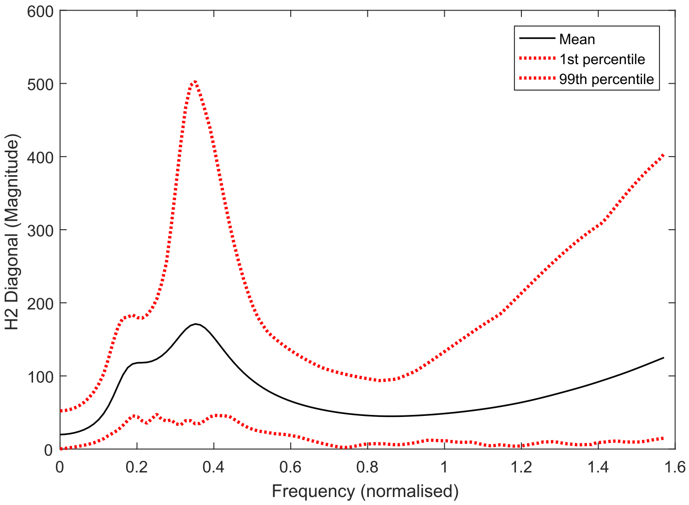


Fig. 42. Mean diagonal $H_2(\omega,\omega)$ estimated from the GP-NARX analysis of the CBT data, together with confidence intervals from first and 99th centiles.

confidence intervals for the HFRFs, which is also introduced and validated in this paper. The nonparametric class of NARX models in general has an advantage over multinomial NARX, in not requiring a structure detection step.

The second aim of the paper was to show how the new methods could be applied to a challenging real problem – that of wave force prediction. GP-NARX models were fitted to wave force data from the Christchurch Bay Tower, a structure which experiences the forces imposed by a random directional sea. The results showed an improvement over a previous analysis conducted using polynomial NARX models and the opportunity was taken to correct some of the minor errors of judgement in the previous analysis concerning the rigour of the machine learning practice. An interesting by-product of the analysis here became apparent in the analysis of the CBT data; the method provides a test for detecting nonlinear systems, capable of resolving odd and even nonlinearity.

There remain a number of outstanding issues. Among the simpler ones to deal with is the fact that the results here are specific to the squared-exponential kernel for the GP. While this is the most commonly used kernel, it is not the most versatile, and other kernels are under investigation. A more serious issue concerns input noise for the GP formulation, as the NARX form means that the GP output noise feeds back into its input, and thus violates one of the defining properties of the GP algorithm. In order to keep the current paper to a manageable length, this issue was not addressed here. Separate justifications for this can be given for the two case studies. In the case of the Duffing oscillator example, the objective was to validate the expressions for the higher-order FRFs, so simulations with low noise were carried out. In the case of the CBT data, the low noise assumption is not justified; however, the results here show an improvement on previous analysis where a noise model was fitted to the data. In any case, the noise in the CBT case is almost certainly correlated (partly resulting from phase variations in the velocity measurements) and would still present a challenge to existing methods; further work is needed, and is ongoing.

With regards to the wave force analysis a number of outstanding issues remain. First of all, there remains the problem of hyperparameter estimation. In the analysis of this paper the numbers of input and output lags in the GP-NARX models were fixed in order to allow comparison with previous analysis; these should generally be established in a principled manner e.g. by cross-validation. This analysis is currently being carried out as part of a larger programme of work on the CBT data, which will also extend the GP-NARX models and their associated HFRFs to the Multi–Input–Multi–Output (MIMO) case.

Acknowledgements

The support of the UK Engineering and Physical Sciences Research Council (EPSRC) through grant reference numbers EP/J016942/1 and EP/K003836/2 is gratefully acknowledged. In addition, the authors are grateful to Ramboll Oil and Gas, Denmark for financial support for Tim Rogers; they would also like to specifically thank Dr. Ulf Tyge Tygeson for various useful discussions regarding wave loading. Finally, the authors would like to thank Dr. James Hensman, formerly of the University of Sheffield Centre for Translational Neuroscience for a number of interesting and useful discussions, and Graeme Manson (University of Sheffield) for detecting a critical bug in the HFRF code.

References

[1] I. Leontaritis, S. Billings, Input-output parametric models for nonlinear systems, part I: deterministic nonlinear systems, Int. J. Control 41 (1985) 303–328.

[2] I. Leontaritis, S. Billings, Input-output parametric models for nonlinear systems, part II: stochastic nonlinear systems, Int. J. Control 41 (1985) 329–344.

[3] S. Billings, Nonlinear System Identification: NARMAX, Methods in the Time, Frequency, and Spatio-Temporal Domains, Wiley-Blackwell, 2013.
[4] S. Billings, H. Jamaluddin, S. Chen, Properties of neural networks with applications to modelling non-linear dynamical systems, Int. J. Control 55 (1992) 193–1350.
[5] S. Chen, S. Billings, C. Cowan, P. Grant, Practical identification of NARMAX models using radial basis functions, Int. J. Control 52 (1990) 1327–1350.
[6] M. Spiridonakos, E. Chatzi, Metamodeling of dynamic nonlinear structural systems through polynomial chaos NARX models, Comput. Struct. 157 (2015) 99–113.
[7] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1998.
[8] R. Murray-Smith, T. Johansen, R. Shorten, On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures, in: Proceedings of European Control Conference, 1999, pp. BA–14.
[9] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.
[10] J. Kocijan, Dynamic GP models: an overview and recent developments, in: Proceedings of 6th International Conference on Applied Mathematics, Simulation and Modelling, 2012, pp. 38–43.
[11] M. Schetzen, The Volterra and Wiener Theories of Nonlinear Systems, John Wiley Interscience Publication, 1980.
[12] K. Worden, G. Tomlinson, Nonlinearity in Structural Dynamics: Detection, Modelling and Identification, Institute of Physics Press, 2001.
[13] E. Bedrosian, S. Rice, The output properties of Volterra systems driven by harmonic and Gaussian inputs, Proc. IEEE 59 (1971) 1688–1707.
[14] S. Billings, K. Tsang, Spectral analysis for nonlinear systems, part I: parametric non-linear spectral analysis, Mech. Syst. Signal Process. 3 (1989) 319–339.
[15] J. Chance, K. Worden, G. Tomlinson, Frequency domain analysis of NARX neural networks, J. Sound Vib. 213 (1997) 915–941.
[16] J. Wray, G. Green, Calculation of the Volterra kernels of nonlinear dynamic systems using an artificial neural network, Biol. Cybern. 71 (1994) 187–195.
[17] J. Morison, M. O'Brien, J. Johnson, S. Schaf, The force exerted by surface waves on piles, Petrol. Trans. 189 (1950) 149–157.
[18] K. Worden, P. Stansby, G. Tomlinson, S. Billings, Identification of nonlinear wave forces, J. Fluids Struct. 8 (1994) 19–71.
[19] D. Krige, A statistical approach to some mine valuations and allied problems at the Witwatersrand Master's thesis, University of Witwatersrand, 1951.
[20] R. Neal, Monte Carlo implementation of Gaussian process models for Bayesian regression and classification, Tech. Rep. Available from: <arXiv:Physics/9701026>, 1997.
[21] D. Mackay, Gaussian processes – a replacement for supervised neural networks. Lecture notes for tutorial, in: International Conference on Neural Information Processing Systems, 1997.
[22] T. Rogers, G. Manson, K. Worden, E. Cross, On the choice of optimisation scheme for Gaussian process hyperparameters in SHM problems, in: Proceedings of 11th International Workshop on Structural Health Monitoring, Stanford University, Palo Alto, CA, 2017.
[23] L. Ljung, System Identification: Theory for the User, Prentice Hall, Englewood Cliffs, 1987.
[24] T. Söderstrom, P. Stoica, System Identification, Prentice Hall, London, 1988.
[25] J. Quinonero-Candelo, C. Rasmussen, A unifying view of sparse approximation Gaussian process regression, J. Mach. Learn. Res. 6 (2005) 1939–1959.
[26] E. Snelson, Z. Ghahramani, Sparse Gaussian processes using pseudo-inputs, in: Advances in Neural Information Processing Systems, MIT Press, 2006.
[27] J. Hensman, N. Durrande, A. Solin, Variational Fourier features for Gaussian processes, Tech. Rep. Available from: <arXiv:Stats.ML/1611.06740v1>, 2016.
[28] A. Giraud, Approximate methods for propagation of uncertainty with Gaussian process models, Ph.D. thesis, University of Glasgow, 2004.
[29] J. Quinonero-Candelo, A. Giraud, C. Rasmussen, Prediction at an uncertain input for Gaussian processes and relevance vector machines – application to multiple-step ahead time series forecasting, Tech. Rep., Department of Informatics and Mathematical Modelling, Technical University of Denmark, 2003.
[30] A. McHutchon, C. Rasmussen, Gaussian process training with input noise, in: Proceedings of 24th International Conference on Neural Information Processing Systems, 2011.
[31] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes: The Art of Scientific Computing, third ed., Cambridge University Press, 2007.
[32] K. Worden, J. Hensman, Parameter estimation and model selection for a class of hysteretic systems using Bayesian inference, Mech. Syst. Signal Process. 32 (2011) 153–169.
[33] V. Volterra, Theory of Functionals and of Integral and Integro-Differential Equations, Dover, 1959.
[34] G. Palm, T. Poggio, The Volterra representation and the Wiener expansion: validity and pitfalls, SIAM J. Appl. Math. 33 (1997), XX–YY.
[35] E. Obasaju, P. Bearman, J. Graham, A study of forces, circulation and vortex patterns around a circular cylinder in oscillating flow, J. Fluid Mech. 196 (1988) 467–494.
[36] T. Sarpkaya, Resistance in unsteady flow – search for a model. Tech. Rep., Naval Postgraduate School, Department of Mechanical Engineering, vol. 93943, Monterey, CA, 1978.
[37] J. Bishop, Aspects of large scale wave force experiments and some early results from Christchurch Bay. Tech. Rep. NMI-R57, National Maritime Institute, 1979.
[38] S. Chen, S. Billings, W. Luo, Orthogonal least squares methods and their application to non-linear system identification, Int. J. Control 50 (1989) 1873–1896.
[39] K. Worden, Confidence bounds for frequency response functions from time series models, Mech. Syst. Signal Process. 12 (1998) 559–569.
[40] J. Schoukens, Y. Rolain, J. Swevers, J.D. Cuyper, Simple methods and insights to deal with nonlinear distortions in FRF measurements, Mech. Syst. Signal Process. 14 (2000) 657–666.
[41] J. Schoukens, R. Pintelon, Y. Rolain, T. Dobrowiecki, Frequency response function measurements in the presence of nonlinear distortions, Automatica 37 (2001) 939–946.
[42] K. Vanhoenacker, T. Dobrowiecki, J. Schoukens, Design of multisine excitations to characterize the nonlinear distortions during FRF measurements, IEEE Trans. Instrum. Meas. 50 (2001) 1097–1102.
[43] T. Baldacchino, S. Anderson, V. Kadirkamanathan, Computational system identification for Bayesian NARMAX modelling, Automatica 49 (2013) 2641–2651.
[44] W. Jacobs, T. Baldacchino, S. Anderson, Sparse Bayesian identification of polynomial NARX models, in: Proceedings of the 17th IFAC Symposium on System Identification, Beijing, China, 2015.
[45] P. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (1995) 711–732.
[46] J. Drugowitsch, Variational Bayesian inference for linear and logistic regression, Tech. Rep. Available from: <arXiv:stat.ML/1310.5438>, 2014.
[47] A.G. Wilson, D. Knowles, Z. Ghahramani, Gaussian process regression networks, in: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, 2012.
[48] R. Gramacy, tgp: An R package for Bayesian nonstationary semiparametric nonlinear regression and design by treed Gaussian process models, J. Stat. Softw. 19 (2007).
[49] M. Lázaro-Gredilla, M. Titsias, Variational heteroscedastic Gaussian process regression, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 2011.
[50] K. Worden, G. Manson, G. Tomlinson, A harmonic probing algorithm for the multi-input Volterra series, J. Sound Vib. 201 (1997) 67–84.