On a Grey Box Modelling Framework for Nonlinear System Identification

T.J. Rogers, G.R. Holmes, E.J. Cross & K. Worden

Dynamics Research Group,
Department of Mechanical Engineering, University of Sheffield,
Mappin Street, Sheffield, S1 3JD, UK

Introduction

Data is provided for a cascaded tanks system with a hidden state (the top tank level), the input to the system is the pump voltage and the output is the voltage of a capacitive level sensor on the bottom tank

The system of interest is shown in the picture to the far right, provided (along with the data used) by M. Schoukens et. al. [1] as part of a workshop on nonlinear system identification held at VUB in Brussels in the spring of 2015

We aim to predict the level of the bottom tank given an arbitrarily varying input voltage to the pump

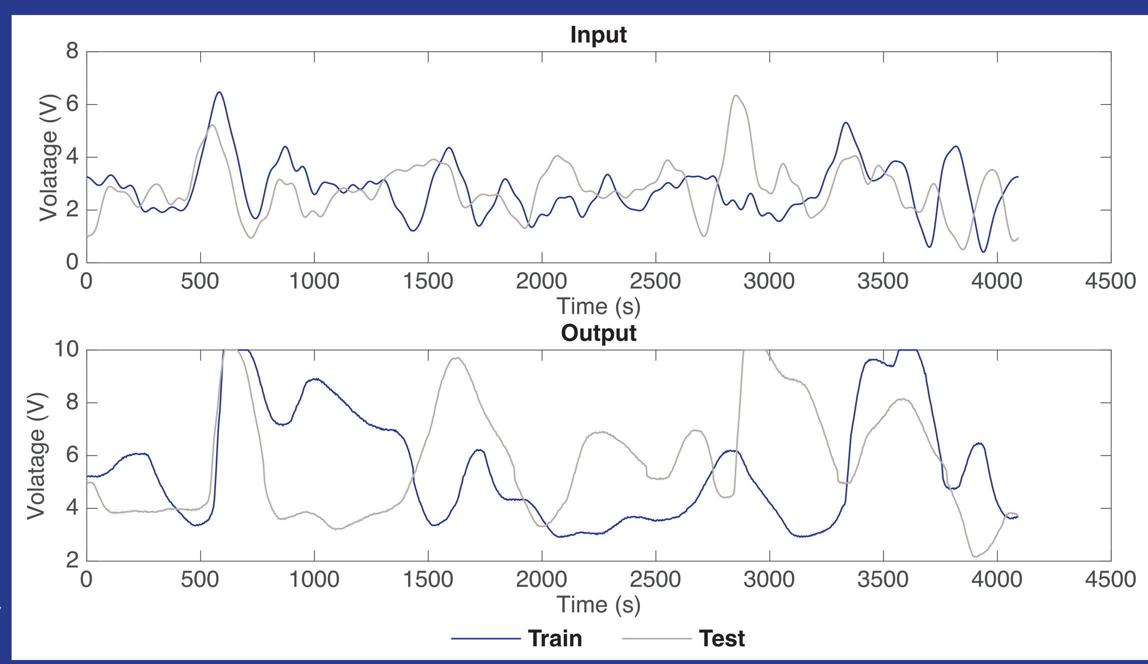
To do this we use a grey box modelling approach where physics is encoded in a parametric white box model and the residuals of this model are accounted for by a Gaussian Process (GP) regression model with a nonlinear auto-regressive (NARX) structure

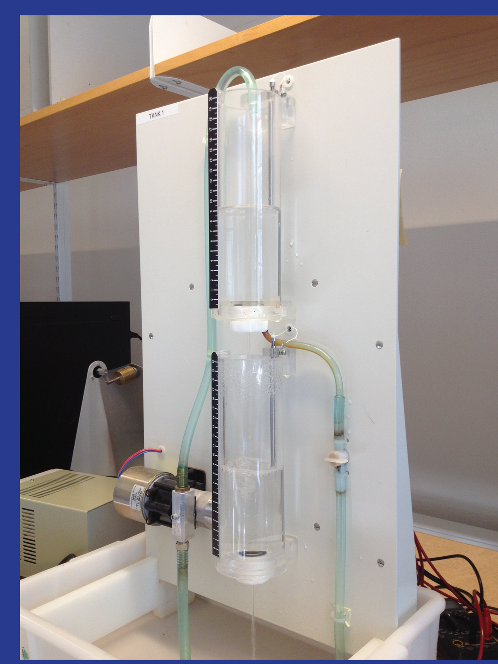
The quality of the model is assessed via a normalised mean square error (NMSE) the equation for which is shown below:

$$NMSE = \frac{100}{N\sigma_y^2} \sum_{i=1}^{N} (y_i^* - y_i)^2$$

The inputs and outputs to the system for the training and testing data are also plotted to the right, it is clear to see that the system is dynamically complex and contains a hard nonlinearity due to the overflow of the tanks when the level exceeeds 10 V

Approaches with pure white box and black box philosophies are compared with a combination of the two approaches, referred to here as a grey box. The model is also tested for one step ahead (OSA) predictions, where the model only predicts a value one time step ahead, and the more demanding model predicted output, where the model is only given the initial conditions of the system





White Box

In the context of this research a white box model is defined as a fully parametric physics based model, this is important due to the lack of validation data and the small size of the training data set. If the models chosen accurately reflect the true physical process or processes governing the system, these white boxes will be immune to overfitting in the classical sense.

Here we also establish white box models with no noise terms, since the models will be carried forward into the grey box. Within the white box there is no capacity to differentiate noise and unknown physical processes, therefore, we leave all the unknown behaviour (noise and unmodelled physics) to be accounted for by the non-parametric modelling, in this case GP NARX, so as not to distort the residuals destroying structure the machine learning process is attempting to model.

Predictions are made with the model based on the state space equations which define the change of every state with respect to time, these can then be multiplied by the time step and added to the previous states, this is summarised below:

$$\begin{aligned} x_1(t+\Delta t) &= x_1(t) + \Delta t\, \dot{x}_1(t) \\ x_2(t+\Delta t) &= x_2(t) + \Delta t\, \dot{x}_2(t) \end{aligned}$$

Two models were considered for this case study, the first model (M1) was a simple application of Bernoulli's principle:

$$\dot{x}_1(t) = -k_1\sqrt{x_1(t)} + k_4 u(t) + w_1(t)$$

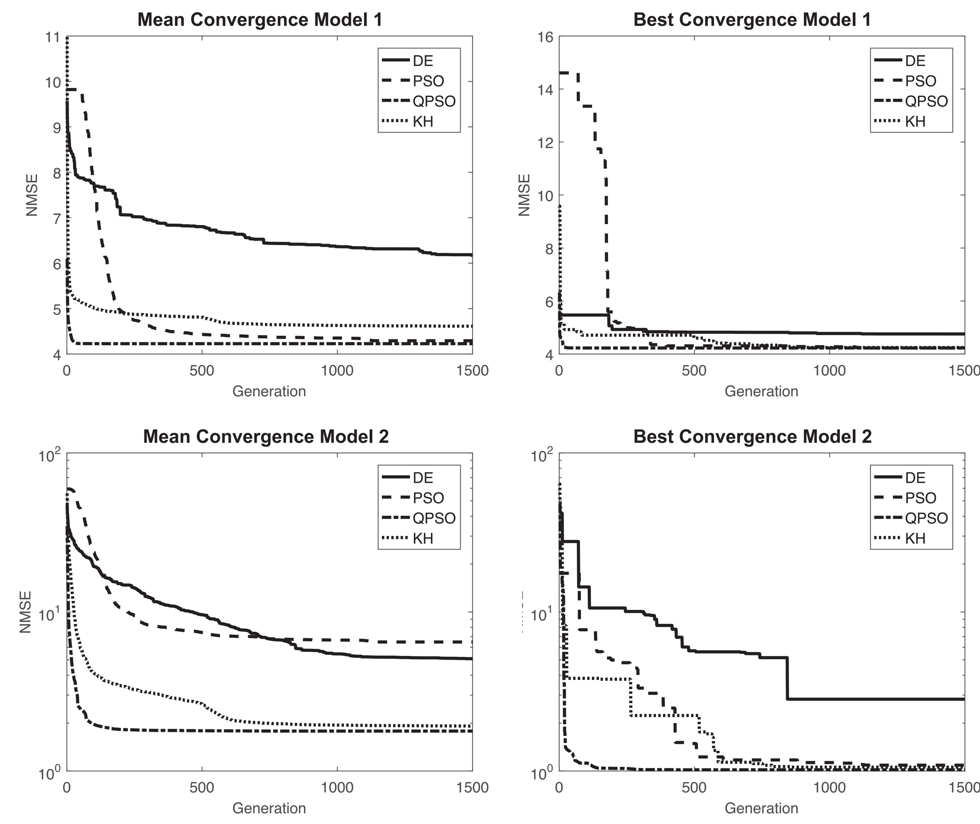
$$\dot{x}_2(t) = \begin{cases} k_1\sqrt{x_1(t)} - k_3\sqrt{x_2(t)} + w_2(t), & x_1(t) \leq 10 \\ k_1\sqrt{x_1(t)} - k_3\sqrt{x_2(t)} + k_5 u(t) + w_3(t), & x_1(t) > 10 \end{cases}$$

The second model (M2) adds terms to M1 to account for losses in the system due to friction and geometrical losses.

$$\dot{x}_1(t) = -k_1\sqrt{x_1(t)} + k_5 x_1(t) + k_4 u(t) + w_1(t)$$

$$\dot{x}_2(t) = \begin{cases} k_1\sqrt{x_1(t)} - k_5 x_1(t) + k_6 x_2(t) - k_3\sqrt{x_2(t)} + w_2(t), & x_1(t) \leq 10 \\ k_1\sqrt{x_1(t)} - k_5 x_1(t) + k_6 x_2(t) - k_3\sqrt{x_2(t)} + k_5 u(t) + w_3(t), & x_1(t) > 10 \end{cases}$$

Four nonlinear optimisation schemes were used to determine the parameters of the two white box models, Differential Evolution (DE) [2], Particle Swarm Optimisation (PSO) [3], Quantum Particle Swarm Optimisation (QPSO) [4], and the Krill Herd (KH) [5].

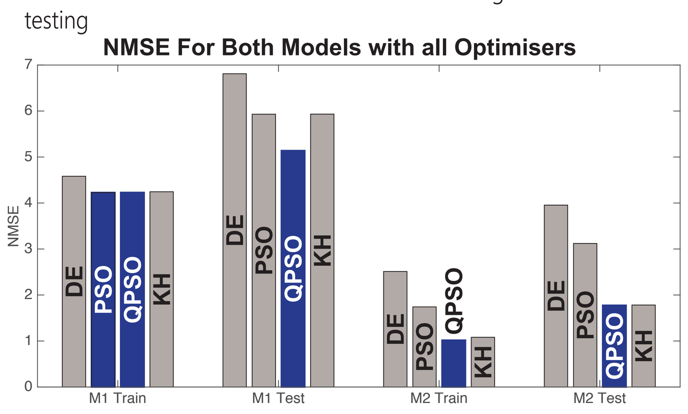


Each optimisation scheme was run 50 times to mitigate the effect of the stochastic components of each of the methods, the mean and best convergence curves for both M1 and M2 are shown to the left.

The QPSO method can be seen to have significant convergence speed advantages over the other methods, in addition to this it is capable of discovering and exploiting more successful areas of the parameter space than the other methods.

The results from predictions using all optimisation methods for both training and testing data are shown below. The additional terms added in the second model lead to far better performance in training and testing indicating it is not just overfitting the training data.

The lowest white box NMSE was 1.0174 for training and 1.7759 for testing



Black Box

Here we define black box models to be a modelling technique to be one where the model is entirey data driven providing no insight into the physical processes driving the observe behaviour. The method chosen for black box modelling in this reasearch is Gaussian Process NARX models.

NARX models attempt to model the output a system via a nonlinear combination of a set of input variables and output variables which are lagged such that the dynamics of the system are encoded in the model. NARX models introduce a number of hyperparameters which define the number of lags in the input and the output of the model this would usually be confirmed by a validation data set. In this case the lack of validation set means that these hyperparameters must be set using only the training data.

Gaussian Processes provide a powerful nonparametric machine learning technique [6] which can be defined as a distribution over functions. Predictions are made by defining a joint distribution between the training data and the predictive inputs, this asserts that there is an underlying multivariate distribution from which the training and testing data is drawn from. The covariance matrix between the inputs and outputs is defined by means of a covariance function to avoid having to learn every parameter in the matrix. The process of training the Gaussian Process model involves the optimisation of the hyperparameters of the covariance function by maximising the marginal likelihood of the output with respect to the input data points and the covariance function hyperparameters.

The process of training the GP NARX model is the same as it would be for a static GP, however, the structure of the NARX model introduces a dependency in the input on the output of the model. Therefore, there are two test cases for the GP NARX model as opposed to the single test case in the static GP. The first of these test cases is the one step ahead model, where the inputs to model are lagged versions of the system inputs and also the lagged true outputs of the system.

We can represent this model in the form shown below:

$$y_i^* = F(y_{i-1}, \ldots, y_{i-n_y}; x_{i-1}, \ldots, x_{i-n_x})$$

A more rigorous test of the model is to use the model predicted output where the lagged outputs used are those generated by the previous model predictions, in this way multiple steps ahead can be predicted which is a much more stressful task. The difference between this and the OSA case is expressed in the following equation:

$$y_i^* = F(y^*_{i-1}, \ldots, y^*_{i-n_y}; x_{i-1}, \ldots, x_{i-n_x})$$

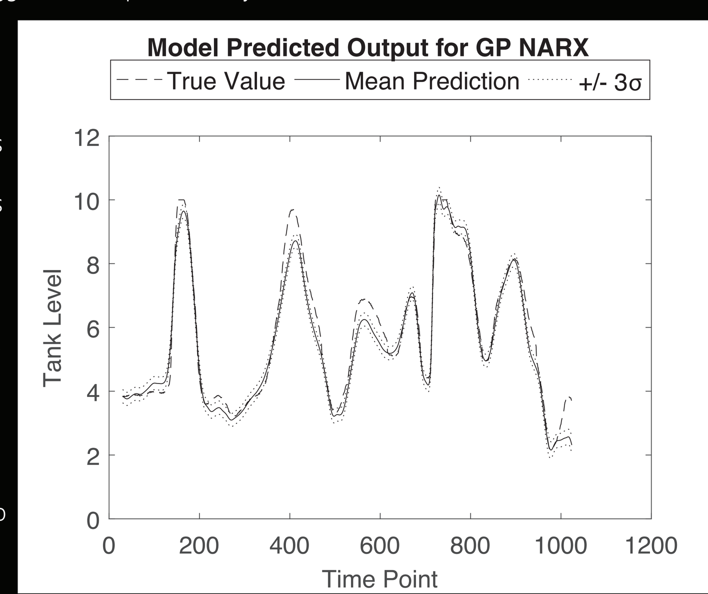
The NARX configuration also introduces extra hyperparameters to the model, these would usually be determined via a validation set. However, in this case the lack of a validation set and the small size of the training data set means that these hyperparameters must be determined from the training data alone.

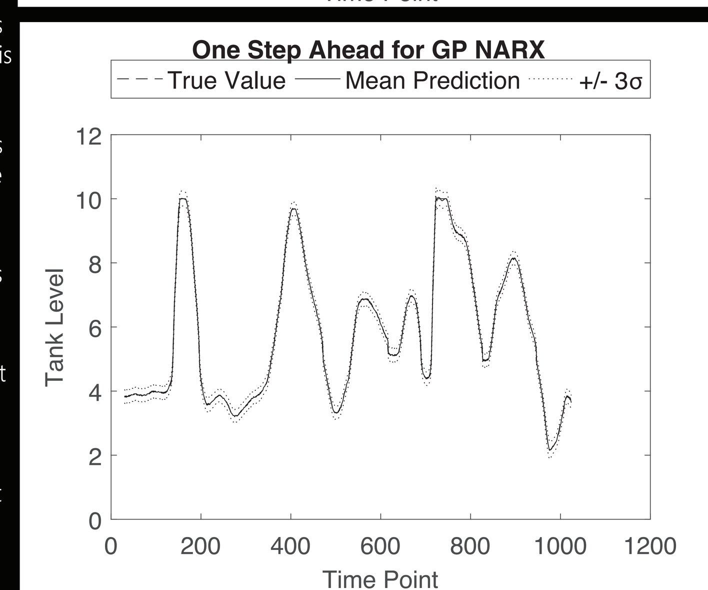
The benefits of training GP models in a Bayesian manner, via the optimisation of the marginal likelihood is that the quality of fit of the model is traded off against the complexity of the model to provide an automatic resistance to overfitting. This helps to ensure that the number of lags which are chosen in the NARX model is minimised.

It should be noted, however, that during the training step only lagged versions of the true outputs are used. This means that the model is trained primarily for the one step ahead prediction case this would not be a problem if the model was making zero noise perfect predictions as the predicted outputs would be the same as the true outputs. However, the astute observer will notice that the the NARX model also violates one of the assumptions of the Gaussian Process model: that there is no noise on the inputs to the model. Methods exist in the literature which could alleviate the issue caused by this.

The performance of the black box can be seen to be very good in the OSA case with an NMSE of 0.0565, the MPO predictions have a higher NMSE of 4.6174. This is in line with the expected issues with training the GP NARX model with only the true lagged results.

These results do show that the GP NARX model has a great deal of potential in nonlinear system identification. One of the key strengths of the model is that the method automatically returns a set of confidence intervals as each prediction returns a marginal distribution with respect to the test inputs, the training input and the training outputs which is Gaussian at every point. This prompts a move towards using the posterior likelihood of the predictions as a measure of model fit as opposed to the use of NMSE which only compares the point estimates of the model and does not include the information that is encoded in the confidence intervals.





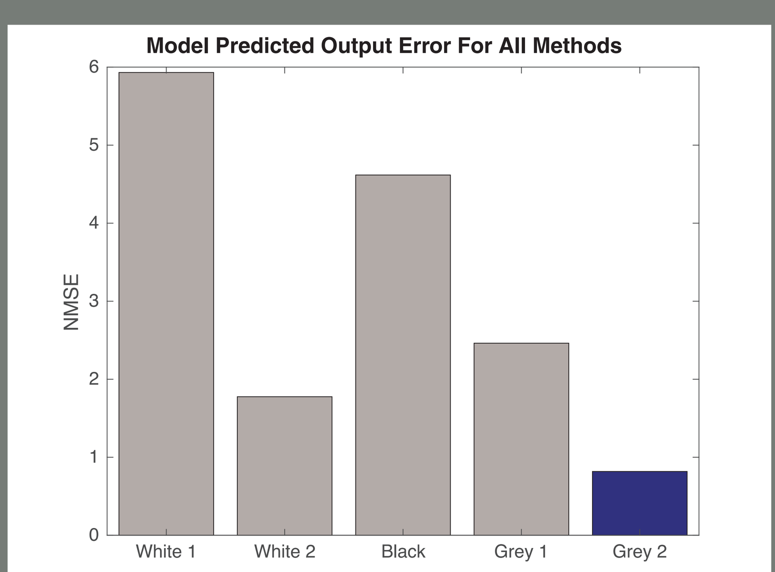
Grey Box

The grey box method is a combination of the two methods detailed above, this aims to encode the prior information of a physics based white box into a black box method (in this case the GP NARX model as used in the black box modelling). There are two ways in which the prior informattion could be encoded: either the black box model could be used to fit the residuals from the white box, this is shown below as model type A. The other method is to use the outputs of the white box as an informative inputs to the black box model, this includes the estimates of the hidden state in the white box. This model is shown as model B below.

$$\text{A)}\quad \mathbf{y_p} = \overbrace{f(X)}^{\text{White box}} + \overbrace{\epsilon(X)}^{\text{Black box}} \qquad \text{B)}\quad \mathbf{y_p} = \underbrace{g(X, \overbrace{f(X)}^{\text{White Box}})}_{\text{Black Box}}$$

In the case study presented here model B performs much better than model A; it is suspected that, due to the accuracy of the white box model, there is a low signal to noise ratio in the residuals of the white box model which makes it difficult for the black box method to fit. However, when the outputs of the white box model are used as additional informative inputs to the black box they can better define the state space of the model which leads to improved predictions. This method relies on the ability of the white box models to extrapolate where the training data has not fully defined the state space.

The power of the grey box model lies in the ability of the model to provide a best of both worlds alternative to the black or grey box methods. The final test of the method was to see if an improvement in the accuracy of the white box portion of the method would follow on to an improved performance in the black box method.

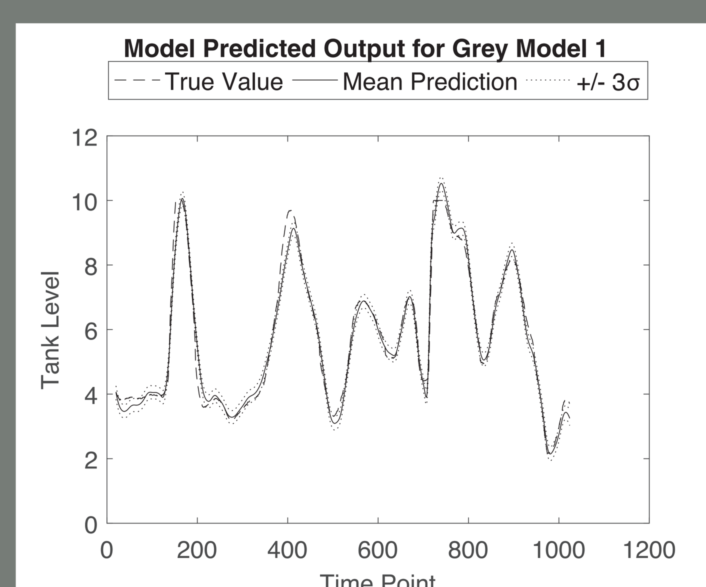


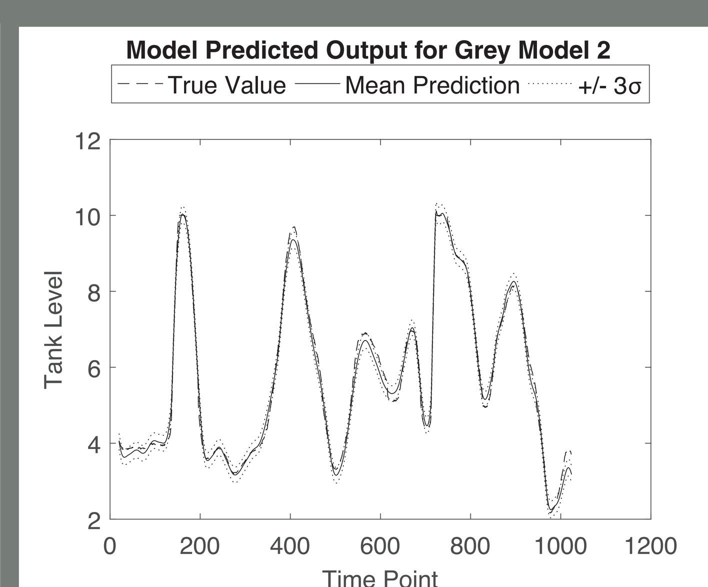
The outputs from the two grey box models are shown to the right for the one step ahead and the mode predicted output cases. The first grey box model is using the outputs of the first white box model M1 and the second grey box model is using the outputs of M2. The grey box also shows the same behaviour of better performance in the one step ahead case than the model predicted output. It can be observed that the portions of the test data that cannot be well fitted in the MPO case in the black box also are the largest sources of error in the grey box. This would suggest that the areas of the state space relating to this output behaviour is not well explored by the training data.

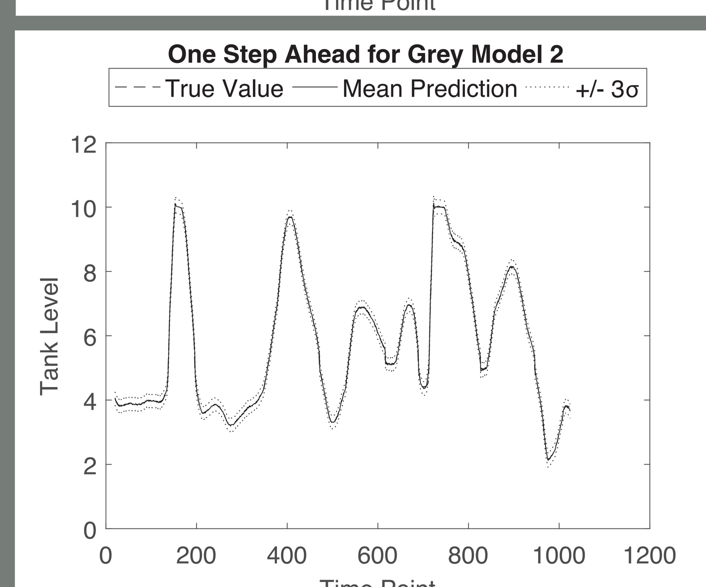
The NMSE from all of the models, is shown to the right. It is clear that the significant improvement from the additional terms in the second white box model (M2) is carried through into the grey box model. It is interesting that the performance of the second white box model is better than that of the black box or even the first grey box model. This would seem to suggest that the addition of prior knowledge into the physical behaviour of the system of interest can be of great advantage even when applying advanced machine learning techniques.

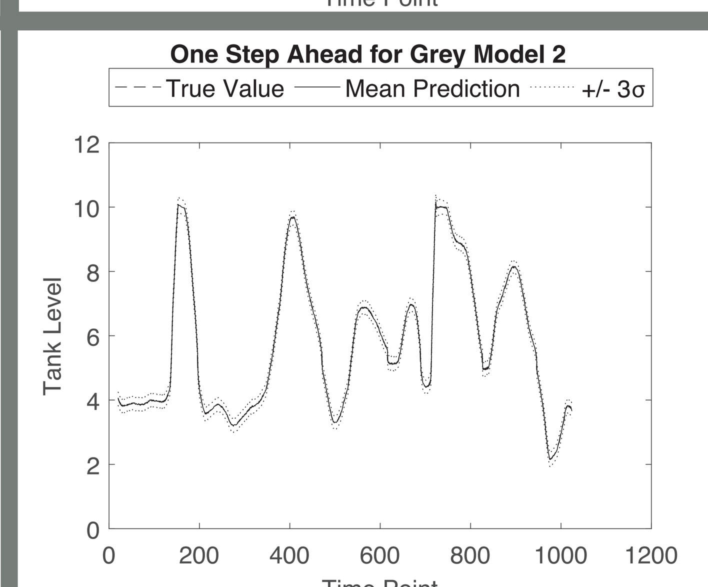
It is also seen that the reduction in error is proportionally similar for both of the grey box models which were tested. This leads to further insight into the behaviour of the system, in that it implies that structure remains in the residuals of the white box, therefore, there is a causal relationship remaining between the inputs and the outputs to the GP NARX model. Which itself implies that there must be an improved physical model which could be constructed.

Presented here is a powerful nonlinear system identification method which is very capable in handling this dynamic system with only a small number of training points. The method achieves a best NMSE of 0.0442 in the OSA case and 0.8178 for the MPO case.









Conclusions

- The combination of white and black box models into a grey box framework can lead to significant performance improvements due to the encoding of prior knowledge into the model.
- This improvement would indicate that modern engineering methods should be keen to adopt state of the art machine learning techniques, however, physical models can still lead to more informative models which achieve lower error.
- The use of a normalised mean square error (NMSE) may not be the most effective way to assess Gaussian Process models as it does not take account of the information contained in the confidence intervals
- It is worth encoding as much information in the white box as possible since this will lead to better performance of the machine learning method along with the improved error in the white box

[1] Schoukens M., Mattsson P., Wigren T., and No el J.P., "Cascaded tanks benchmark combining soft and hard nonlinearities," 2016.
[2] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," Journal of Global Optimization, vol. 11, no. 4, pp. 341–359, 1997.
[3] K. J. Eberhart, R.C. et al., "A new optimizer using particle swarm theory," in Proceedings of the sixth international symposium on micro machine and human science, vol. 1. New York, NY, 1995, pp. 39–43.
[4] J. Sun, B. Feng, and W. Xu, "Particle swarm optimization with particles having quantum behavior," in Congress on Evolutionary Computation, 2004.
[5] A. Gandomi and A. Alavi, "Krill herd: a new bio-inspired optimization algorithm," Communications in Nonlinear Science and Numerical Simulation, vol. 17, no. 12, pp. 4831–4845, 2012
[6] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning. MIT Press, 2006.