State Space Models for Forecasting Water Quality Variables

An Application in Aquaculture Prawn Farming

Joel Janek Dabrowski
Data61, CSIRO
Brisbane, Queensland, Australia
joel.dabrowski@data61.csiro.au

Ashfaqur Rahman
Data61, CSIRO
Hobart, Tasmania, Australia
ashfaqur.rahman@data61.csiro.au

Andrew George
Data61, CSIRO
Brisbane, Queensland, Australia
Andrew.George@data61.csiro.au

Stuart Arnold
Agriculture and Food, CSIRO
Bribie Island, Queensland, Australia
Stuart.Arnold@csiro.au

John McCulloch
Data61, CSIRO
Hobart, Tasmania, Australia
John.Mcculloch@data61.csiro.au

ABSTRACT

A novel approach to deterministic modelling of diurnal water quality parameters in aquaculture prawn ponds is presented. The purpose is to provide assistance to prawn pond farmers in monitoring pond water quality with limited data. Obtaining sufficient water quality data is generally a challenge in commercial prawn farming applications. Farmers can sustain large losses in their crop if water quality is not well managed. The model presented provides a means for modelling and forecasting various water quality parameters. It is inspired by data dynamics and does not rely on physical ecosystem modelling. The model is constructed within the Bayesian filtering framework. The Kalman filter and the unscented Kalman filer are applied for inference. The results demonstrate generalisability to both variables and environments. The ability for short term forecasting with mean absolute percentage errors between 0.5% and 11% is demonstrated.

KEYWORDS

Bayesian filtering, Kalman filter, dissolved oxygen, pH, time series.

ACM Reference Format:
Joel Janek Dabrowski, Ashfaqur Rahman, Andrew George, Stuart Arnold, and John McCulloch. 2018. State Space Models for Forecasting Water Quality Variables: An Application in Aquaculture Prawn Farming. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219841

1 INTRODUCTION

In aquaculture farming, such as prawn farming, water quality is a vital entity to manage for the maximisation of productivity, quality, and health of the stock. Dissolved oxygen (DO) is possibly the most important water quality variable in aquaculture [6]. At low levels of DO, aquatic animals do not feed or grow well and are more susceptible to disease. High levels of mortality in the crop can occur due to anoxia and hypoxia if the DO reduces to extreme levels [21]. This is referred to as a "DO crash". Such crises will occur in a matter of hours during the night when algae is not producing oxygen through photosynthesis.

Crisis events such as those relating to extreme DO levels are a significant concern for farmers as the DO levels may not be monitored during critical times. Farmers generally measure dissolved oxygen twice a day; once in the early morning and once in the late afternoon. DO measurements are commonly made using a handheld probe [21]. The probe is dipped into each pond and measurements are recorded. More frequent monitoring of water quality parameters in general is a challenge due to high sensor costs, large farms, and limited staff. The purpose of the models presented in this study are to assist farmers with limited water quality data. This assistance is presented in the form of a model that is able to forecast water quality variables such as DO. Such forecasts could be used in an early warning system for predicting critical events such as low DO conditions. Farmers could respond, for example, by turning on additional aerators to increase DO levels.

One common approach to modelling water quality variables is to develop a complex ecosystem model. Such models take into account the various factors that may affect the water quality parameter of interest. For example, factors such as photosynthetic production, total respiration, oxygen exchanges with the atmosphere, and sediment oxygen demand are modelled [9, 26]. Such models are able to produce relatively accurate results. Due to the complexity of the models and the number of variables considered, they are naturally conformed to the particular ecosystem they model. Thus, without significant parameter adjustment, these models are not likely to generalise well to various aquaculture farming applications, environmental conditions, and variables. Furthermore, Link et al. [14] suggest that ecosystem models have not historically dealt with uncertainty in a robust manner, if even at all.

The second approach to modelling water quality variables is the application of data-driven models such as neural networks [18]. Such models do not require intricate knowledge of the ecosystem and environment as required by the ecosystem-based models. The models are based on the data provided to them. According to the curse of dimensionality, such models generally require larger amounts of data for training as the model complexity increases [17]. Since factors affecting water quality have nonlinear relationships with water quality variables, more complex models are required [5]. Data-driven models are trained to provide a prediction of some

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219841

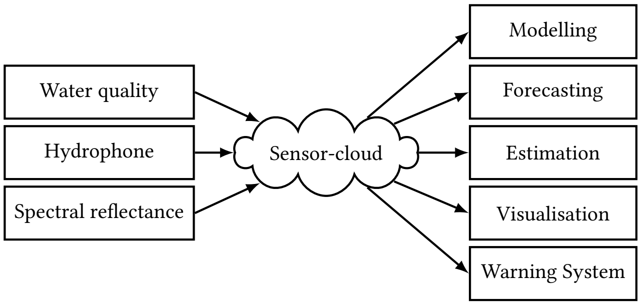

Figure 1: Aquaculture prawn farm decision support system.

unknown quantity given a set of measured inputs. Thus, to obtain a prediction, the set of current inputs are required. The models do have the ability to provide forecasts using methods such as windowing. Such forecasts are however based on previous data rather than a model of the underlying system that produces the data. For the model to learn the intricacies of the data dynamics, high frequency sampling is required to capture these dynamics.

The model presented in this study provides a means to model and forecast diurnal variables such as DO whilst providing levels of uncertainty of the predictions. The model lies between ecosystem-based model and data-driven model approaches. Though the model is generative, it is less complex than ecosystem models. The variable temporal dynamics that are incorporated into the model are inspired by data attributes. As a generative probabilistic model, it however does not completely rely on data. It is able to provide forecasting capability with limited data and is thus ideal in the prawn farming context. This allows for adaptability to various pond environments and variables. The model is however not constrained to aquaculture and is applicable to various other problems and applications involving seasonal time series data.

2 BACKGROUND AND CONTEXT

This study fits within a broader framework of the development of a decision support system for aquaculture prawn farms. A set of sensors have been deployed to monitor water quality related parameters in prawn ponds. The primary sensor is the YSI EXO2 Multiparameter Sonde which provides various measurements of water quality variables. Additionally, other sensors such as hydrophones and spectral reflectance sensors have also been deployed. The general framework of the decision support system is illustrated in Figure 1. In this study, modelling and forecasting of diurnal water quality variables is considered. Forecasts could be presented to farmers in the form of graphs or warnings. Forecasts can also be used as a basis for decision support systems.

Many water quality parameters follow diurnal fluctuations [6]. These variables are typically affected by photosynthesis and respiration of organic matter in the ponds. During the day, organisms such as phytoplankton use solar radiation for photosynthesis. Through photosynthesis, carbon dioxide is absorbed and oxygen is released. At night photosynthesis ceases and carbon dioxide is continuously produced through respiration by organisms such as animals, phytoplankton, zooplankton, and bacteria. Carbon dioxide decreases during the day as it is absorbed through photosynthesis and increases at night. Carbon dioxide reacts with water to form carbonic acid. Increased carbonic acid results in a decrease in pH levels at night. Additionally, variables such nitrite, nitrate, ammonia levels fluctuate diurnally in response to changes in plant growth. Plant nutrient concentrations tend to decrease during the day as plants actively assimilate nutrients during daylight.

Water quality variables may also vary in a non-periodic way [6]. Such variations may be caused by biological activity such as algal blooms and weather-related variations.

3 RELATED WORK

Various ecosystem-based models have been proposed for modelling variables relating to water quality. Dissolved oxygen models are usually based on the mass balance equation [9, 15, 16, 26]. The equation states that the change in concentration is equal to the sum of production and exchange, minus the consumption. Each of these components require complex multivariable models. Such models often require precisely determined parameters pertaining to various physical, chemical, biological, and hydrological processes.

Related to the Bayesian filtering approach presented in this study, Pastres et al. [19] proposed the application of the extended Kalman filter for updating the estimates of the parameters of a DO-chlorophyll model. The model includes forcing variables such as meteorological data and water quality parameters. Lee et al. [13] applied a extended Kalman filter to produce forecasts of algal bloom and dissolved oxygen dynamics in a marine fish culture zone. A set of nine variables are modelled through a set of governing equations comprising thirty parameters. Kim et al. [12] coupled two numerical models using an ensemble Kalman filter for simulating algal bloom dynamics simulation in a large regulated river system. Huang et al. [11] used an ensemble Kalman filter to assimilate measured data into a spatial hydrodynamic-phytoplankton model for predicting dynamics of phytoplankton biomass. These models are relatively accurate. To achieve this accuracy various models, variables, and parameters are incorporated into the model. This however results in models that are confined to the particular application and are generally highly sensitive to parameter values. The model proposed in this study uses a model that is not dependent on ecosystem modelling and thus provides more flexibility to various applications.

A prominent data-driven method for modelling water quality data is the artificial neural network (ANN). Various applications have been presented in literature [1, 5, 8, 10, 20, 22]. The approach is to provide a set of input variables to the ANN. These input variables may include water quality variables, chemical concentrations, and meteorological conditions. The ANN is trained to predict the target output variable such as DO. Olyaie et al. [18] present a comparison of various machine learning based methods for predicting DO in the Delaware River. The machine learning methods presented are the ANN, linear genetic programming, and the support vector machine.

A form of temporal forecasting of water quality parameters using data-driven methods is performed by splitting the dataset in time [2, 3, 7, 23]. Data from a former period in time is used to train the ANN and the data from the latter period in time is used to test the ANN. By testing the model on the latter period of time, "temporal

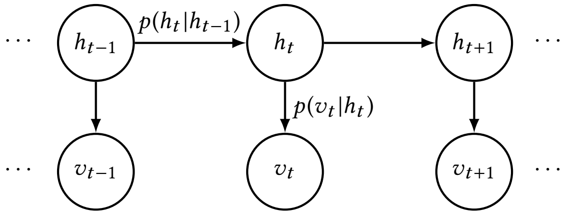

Figure 2: Graphical model of Markov assumptions.

forecasting” is performed [7]. In this approach however, the ANN is not operating as a generative model as it still requires inputs to provide the predictions. That is, nature of the data-driven model, the forecasts are data-driven. The model presented in this study is a probabilistic generative model and does not completely rely on data to provide forecasts.

4 BAYESIAN FILTERING

Consider a system comprising a latent or hidden variable h_t that evolves over time. A measurement can be made on the system which provides an observable or visible variable v_t. The observable variable is considered to have been emitted from the latent variable h_t. Assuming a first order Markov process, the graphical model describing this system is illustrated in Figure 2. The edges between the latent variables describe the transition distribution $p(h_t|h_{t-1})$. The edges between the latent and observable variables describes the emission distribution $p(v_t|h_t)$.

Forecasting relates to computing the conditional distribution $p(h_t|v_{1:t-1})$. This distribution can be computed in the context of filtering. The filtering operation involves computing the posterior conditional distribution $p(h_t|v_{1:t})$. This distribution describes of the current latent variable at time t given all the observations up to time t. To compute this conditional posterior Bayes rule may be applied to form a recursive algorithm involving the following prediction and update step [24]

$$p(h_t|v_{0:t-1}) = \int p(h_t|h_{t-1})p(h_{t-1}|v_{0:t-1})dh_{t-1} \tag{1}$$

$$p(h_t|v_{0:t}) \propto p(v_t|h_t)p(h_t|v_{0:t-1}) \tag{2}$$

The prediction step provides a means for forecasting.

4.1 State Space Representation

The model presented in Figure 2 can be represented in state space form. In this form, the latent variable h_t and observable variable v_t are vectors. The latent variable evolves over time according to some function $a()$ such that

$$h_t = a(h_{t-1}) + \eta_h \tag{3}$$

where $a()$ is referred to as the state transition function. In this study, the state noise process η_h is assumed to be distributed by a Gaussian, $\mathcal{N}(0, \Sigma_h)$. The observable variable is emitted from the latent variable according to some function $b()$ such that

$$v_t = b(h_t) + \eta_v \tag{4}$$

where $b()$ is referred to as the emission or measurement function. In this study, the measurement noise, η_v is assumed to be distributed by the Gaussian, $\mathcal{N}(0, \Sigma_v)$.

If the state transition and emission functions are linear, the system is referred to as the linear dynamic system (LDS). For the LDS, equations (3) and (4) are given by [4]

$$h_t = Ah_{t-1} + \eta_h \tag{5}$$

$$v_t = Bh_t + \eta_v \tag{6}$$

The matrix A is the state transition matrix and the matrix B is the emission matrix.

4.2 Kalman Filter

The Kalman filter is the Bayesian filter algorithm applied to the LDS. Let the updated estimate of the latent variable at time t be represented by a Gaussian with mean f_t and covariance F_t. In the prediction step, the Gaussian distributions described by (1) are parameterised by [4, 17]

$$\mu_h = Af_{t-1} \tag{7}$$

$$\mu_v = B\mu_h \tag{8}$$

$$\Sigma_{hh} = AF_{t-1}A^T + \Sigma_h \tag{9}$$

$$\Sigma_{vv} = B\Sigma_{hh}B^T + \Sigma_v \tag{10}$$

$$\Sigma_{vh} = \Sigma_{hh}B^T \tag{11}$$

In the update step described by (2), the mean f_t and covariance F_t are updated according to [4, 17]

$$K = \Sigma_{vh}\Sigma_{vv}^{-1} \tag{12}$$

$$f_t = \mu_h + K(v_t - \mu_v) \tag{13}$$

$$F_t = (I - KB)S_{hh} \tag{14}$$

where the matrix K is known as the Kalman gain matrix.

4.3 Unscented Kalman Filter

If the state transition function $a()$ or the emission function $b()$ are nonlinear, a closed form solution of the Bayesian filter is generally not available. A Gaussian passed through a linear function maintains its Gaussian form. However, a Gaussian passed through a nonlinear function is generally no longer Gaussian in form. The unscented Kalman filter passes a deterministic set of points referred to as sigma points through the nonlinear function [17, 25]. A Gaussian is fitted to these transformed sigma points. This Gaussian provides an approximation to the filtered posterior distribution.

Associated with the latent state vector h_t are a set of sigma points $\mathcal{X}_{h_t}$, mean weights w_{m_h}, and covariance weights w_{c_h}. Similarly, $\mathcal{X}_{v_t}$, w_{m_v}, and w_{c_v} are the sigma points, mean weights, and covariance weights associated with the observable vector v_t respectively. The unscented transform is applied to the sigma points which are passed through the state transition function $a()$ as follows [17, 25]

$$\mathcal{X}_{h_t}, w_{m_h}, w_{c_h} = UT\left[a\left(\mathcal{X}_{h_{t-1}}\right)\right]$$

where $UT[\,]$ is the unscented transform. The unscented transform is again applied to $\mathcal{X}_{h_t}$ which are passed through the state emission function $b()$ as follows [17, 25]

$$\mathcal{X}_{v_t}, w_{m_v}, w_{c_v} = UT\left[b\left(\mathcal{X}_{h_t}\right)\right]$$

With these distribution approximations, the UKF algorithm prediction equations are then given by [17, 25]

$$\mu_h = \sum_{i=0}^{2d} w_{m_h}^{(i)} a(\mathcal{X}_{h_t}^{(i)}) \tag{15}$$

$$\mu_v = \sum_{i=0}^{2d} w_{m_v}^{(i)} b(\mathcal{X}_{v_t}^{(i)}) \tag{16}$$

$$\Sigma_{hh} = \sum_{i=0}^{2d} w_{c_h}^{(i)} (\mathcal{X}_{h_t}^{(i)} - \mu_h)(\mathcal{X}_{h_t}^{(i)} - \mu_h)^T + \Sigma_h \tag{17}$$

$$\Sigma_{vv} = \sum_{i=0}^{2d} w_{c_v}^{(i)} (\mathcal{X}_{v_t}^{(i)} - \mu_v)(\mathcal{X}_{v_t}^{(i)} - \mu_v)^T + \Sigma_v \tag{18}$$

$$\Sigma_{vh} = \sum_{i=0}^{2d} w_{c_v}^{(i)} (\mathcal{X}_{h_t}^{(i)} - \mu_h)(\mathcal{X}_{v_t}^{(i)} - \mu_v)^T \tag{19}$$

With these distribution approximations, the following update steps for the UKF filter are defined [17, 25]

$$K = \Sigma_{vh}\Sigma_{vv}^{-1} \tag{20}$$

$$f_t = \mu_h + K(v_t - \mu_v) \tag{21}$$

$$F_t = \Sigma_{hh} - K\Sigma_{vv}K^T \tag{22}$$

5 MODEL FORMULATION

Consider the plot of DO over a period of ten days in Figure 3. From a data-analysis perspective, the data consists of an oscillatory function with a stochastic offset. The oscillation is due to diurnal fluctuations. The stochastic offset may be due to various influences such as variations in algae concentrations and external inputs into the pond such as water exchange. A sinusoid is used to model the oscillatory function and a constant velocity process is used to model the offset. The general model for the signal is given by

$$y_t = \alpha_t \sin(\omega t) + \gamma_t \tag{23}$$

where y_t is the model output, α_t is the sinusoid amplitude, ω is the frequency of the sinusoid, and γ_t is the offset model. With diurnal fluctuations, the period of the sinusoid is fixed at 24-hours. The constant velocity process implies that the first derivative of γ_t is constant.

At each time t, noisy sensor measurements on y_t are provided. In state space form, the variable y_t is represented by the measurement vector v_t. No measurements are made on the sinusoidal function and offset process. These variables are thus integrated into the latent vector h_t.

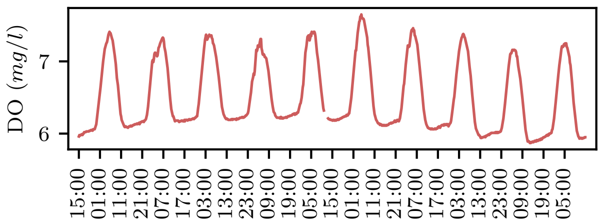

Figure 3: Plot of DO over a period of ten days

5.1 Linear-Gaussian Model

Assume that the state transition function and the emission function are linear. Furthermore, assume that the state noise process and the measurement noise are Gaussian distributed. To formulate (23) in state space form let $x_1 = \gamma_t$ and $x_2 = \alpha_t \sin(\omega t)$. The amplitude α_t is assumed to be constant in time such that $\alpha_t = \alpha$. With a constant velocity model, the first derivative of x_1 with respect to time is constant. The first and second derivatives of x_2 are given by

$$\dot{x}_2 = \omega\alpha\cos(\omega t)$$
$$\ddot{x}_2 = -\omega^2\alpha\sin(\omega t) = -\omega^2 x_2$$

The latent vector is formed by combining x_1 and x_2 and their derivatives in a vector. In state space form, the latent variable model is given by

$$\dot{h}_t = \Phi h_t$$
$$\begin{bmatrix}\dot{x}_1\\ \ddot{x}_1\\ \dot{x}_2\\ \ddot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & -\omega^2 & 0\end{bmatrix}\begin{bmatrix}x_1\\ \dot{x}_1\\ x_2\\ \dot{x}_2\end{bmatrix} + \eta_h \tag{24}$$

A key observation is that the model is independent of the sinusoid amplitude α.

For the discrete time system, $h_t = Ah_{t-1}+\eta^h$, the state transition matrix A can be derived using a Laplace transform or a Taylor series expansion [28]

$$A = e^{\Phi\Delta t} = I + \Phi\Delta t + \frac{(\Phi\Delta t)^2}{2!} + \frac{(\Phi\Delta t)^3}{3!} + \cdots \tag{25}$$

where Δt is the sample rate.

The measurement vector v_t is given by

$$v_t = Bh_t + \eta^v$$
$$\begin{bmatrix}y_t\end{bmatrix} = \begin{bmatrix}1 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}x_{1_t}\\ \dot{x}_{1_t}\\ x_{2_t}\\ \dot{x}_{2_t}\end{bmatrix} + \eta^v \tag{26}$$

Equations (24) and (26) are linear equations. The KF is applied for inference in this model.

5.2 Nonlinear-Gaussian Model

The linear-Gaussian model assumes that the sinusoid amplitude α_t is constant in time. To allow for a time varying amplitude, α_t may be integrated into the state vector, h_t. Let $x_1 = \gamma_t$ and $x_3 = \alpha_t$ and $x_4 = \sin(\omega t)$. Assuming a constant velocity models for x_1 and x_3, the system dynamics are given by

$$\dot{h}_t = \Phi h_t$$
$$\begin{bmatrix}\dot{x}_1\\ \ddot{x}_1\\ \dot{x}_3\\ \ddot{x}_3\\ \dot{x}_4\\ \ddot{x}_4\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & -\omega^2 & 0\end{bmatrix}\begin{bmatrix}x_1\\ \dot{x}_1\\ x_3\\ \dot{x}_3\\ x_4\\ \dot{x}_4\end{bmatrix} \tag{27}$$

For the time invariant system, $h_t = Ah_{t-1} + \eta^h$, the state transition matrix A can be approximated using (25). The emission function is given by

$$v_t = b(h_t) + \eta^v$$
$$y_t = x_{1_t} + x_{3_t} x_{4_t} + \eta^v \tag{28}$$

The state transition function $a()$ is a linear operation involving matrix A. With the product of x_{2_t} and x_{3_t}, the emission function is however not linear. The UKF is applied for inference in this model.

6 RESULTS

6.1 Dataset

The dataset used in this study consists of DO and pH readings taken from a YSI EXO2 Multiparameter Sonde sensor [27]. The sensors were deployed in two different prawn ponds in Queensland, Australia. The dataset spans 45 days. With 15 minute intervals between sensor readings, one ponds dataset consists of 4320 samples. The first pond is a large 0.18ha grow-out pond and the second pond is a small 0.022ha nursery pond. The data acquired from the grow-out pond are plotted in Figure 4. The data acquired from the nursery pond are plotted in Figure 5.

There is missing data in the dataset for the grow-out pond. Furthermore, sensors are periodically removed from the water for cleaning. This results in outlier samples in the pH data. These are indicated by the spikes in data in Figure 5.

6.2 Methodology

To demonstrate the modelling and forecasting capability of the model, the dataset is reduced to three sensor samples per day. This simulates the farmers acquiring samples at various times of the day using a handheld probe. Between samples, the model is required to provide a forecast of the water quality variable every 15 minutes. When a new sensor sample is provided, the model parameters are updated. Thus, with 15 minute intervals over 24 hours, a total of 96 forecasts are provided by the model. Only 3 of these 96 forecasts are updated with sensor measurements.

To demonstrate a 5-day forecasting ability of the model, the model is not provided with any updates in the last five days of the evaluation. The model is thus required to forecast these values with no updates. This provides an indication of the model's ability to forecast the water quality variable into the future.

To demonstrate the ability of the model to adapt to various environments the same DO and pH model parameters are used for both pond datasets. Given that the two ponds differ significantly in their size and purpose, the dataset differ significantly.

To provide an evaluation of the error between the posterior filtered result and the measured data, the absolute percentage error is computed. This absolute percentage error for sample at time t is given by

$$\epsilon_t = 100 \left| \frac{v_t - \mu_v}{v_t} \right| \tag{29}$$

where v_t is the sensor data sample at time t and μ_v is the mean of filtered prediction of the water quality variable at time t. The mean absolute percentage error over the entire dataset is given by

$$\mathbb{E}[\epsilon] = \frac{1}{T} \left| \sum_{t=0}^{T} \frac{v_t - \mu_v)}{v_t} \right| \tag{30}$$

6.3 Linear-Gaussian Model Results

The modelling and forecasting results for the linear-Gaussian model are illustrated in Figure 6 and Figure 7. The sensor samples that are provided as observations to the model are indicted by black diamond markers. The thin red curve is a plot of the model's posterior filtered mean. The shaded blue region provides an indication of the standard deviation of the variance Σ_{vv}. This region illustrates the uncertainty of the model in its forecasts. The thick black curve is a plot the complete dataset. This curve is provided as ground truth for the visual evaluating the model performance.

The results demonstrate that the model is able to track the diurnal fluctuations and the random offset in both ponds. The initial tracking of the data is best in the grow-out pond. The model parameters were initialised particularly for this pond. The dynamics in the nursery pond are more exaggerated than those in the grow-out pond. The model however adapts to these variations. The model performs well in the 5-day forecasting in both ponds. There is an overshoot in the DO forecast for the grow-out pond. The 5-day forecasts in general continue with the sinusoid amplitude and trajectory that was last estimated. This does however imply that the forecast

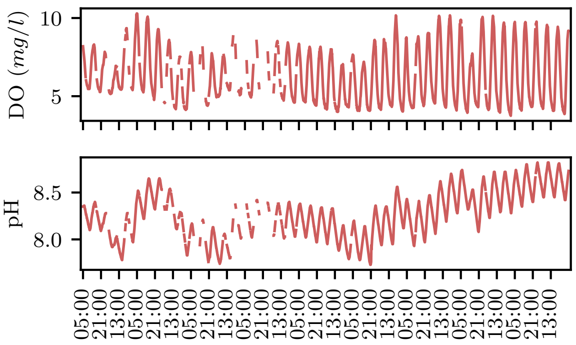

Figure 4: Large grow-out pond DO dataset (pond 1)

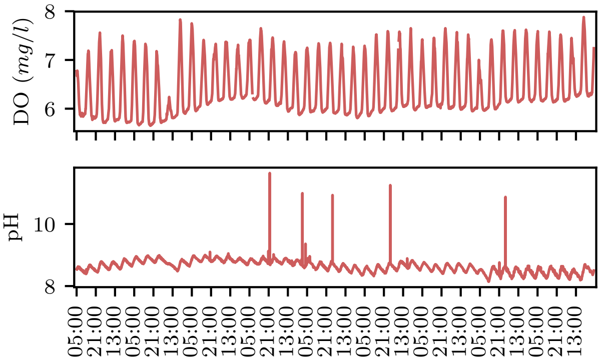

Figure 5: Small nursery pond DO dataset (pond 2)

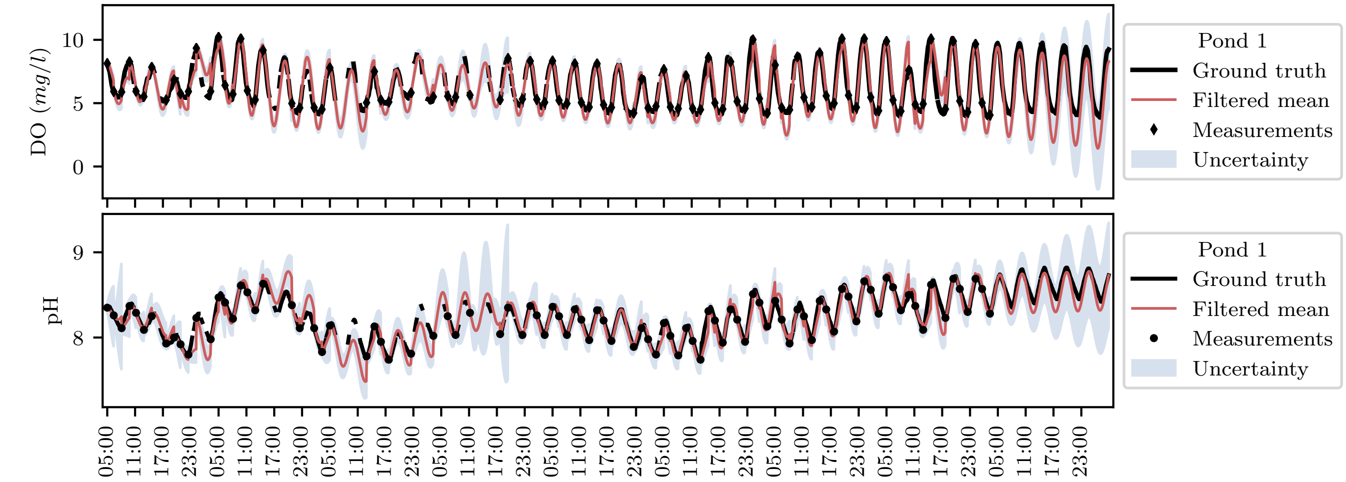

Figure 6: Kalman filtered results in the grow-out pond (pond 1).

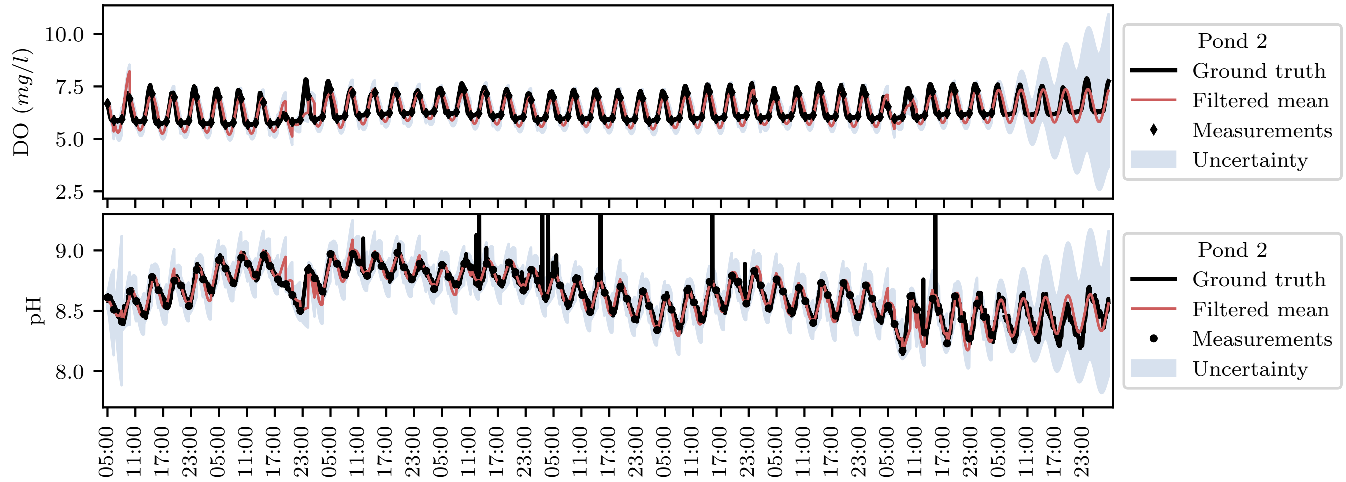

Figure 7: Kalman filtered results in the nursery pond (pond 2)

will not include deviations in trajectory. Long term forecasts are thus less likely to be accurate if dynamics change. This is further implied by the increasing levels of uncertainty of the forecasts as time increases.

An example of the inferred latent variables is illustrated in Figure 8. The filtered posterior mean with its standard deviation are plotted in the top figure. The offset model $x_1 = \gamma_t$ is plotted in the middle figure. As expected, the offset provides an approximation of the waveform mean. The sinusoidal model x_2 is plotted in the lower figure of Figure 8. The frequency of the oscillation remains constant however the amplitude varies. The linear-Gaussian model is independent of the sinusoidal amplitude. It is thus able to track the varying amplitudes in the data as the sensor observations are presented. This property is highly valuable in modelling such data. In forecasting however, the amplitude remains constant. This is especially evident in the 5-day forecast.

The absolute percentage error given by (29) is plotted in Figure 9 and Figure 10. As illustrated in Figure 9, the error in the grow-out pond increases with the 5-day forecasts. It remains low in the nursery pond forecast. The average values of the absolute percentage error over the entire set, is given by (30). The results are presented in Table 1. The error for the DO is higher than the pH error. This is primarily due to the flattened troughs in the DO data. The sinusoidal model overshoots these troughs resulting in increased error. The model is able to track the are more sinusoidal pH waveforms. The error for the pH is less than 1% for both ponds. This low error demonstrates a high level of accuracy.

6.4 Nonlinear-Gaussian Model Results

The modelling and forecasting results for the nonlinear-Gaussian model are illustrated in Figure 11 and Figure 12. In these results, a time varying sinusoid amplitude is modelled. The model requires several samples to align itself to the DO data in the nursery pond.

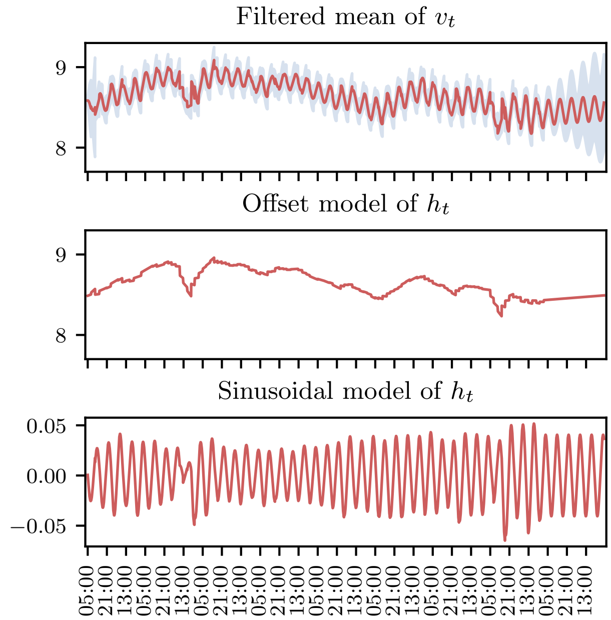

Figure 8: Plot of the UKF filtered posterior mean f_t and elements from the state vector h_t for the nursery pond.

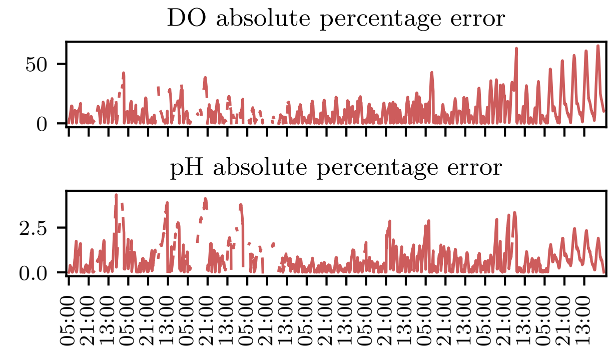

Figure 9: Linear-Gaussian model absolute percentage error given by (29) plot for the grow-out pond (pond 1)

	Pond 1	Pond 2
DO	10.74	4.42
ph	0.75	0.52

Table 1: Linear-Gaussian model mean absolute percentage error given by (30)

High levels of uncertainty of the forecasts during this period is indicated by the high values in the variance. The 5-day forecasts of the nonlinear-Gaussian models appear more aligned to the data than those of the linear-Gaussian model.

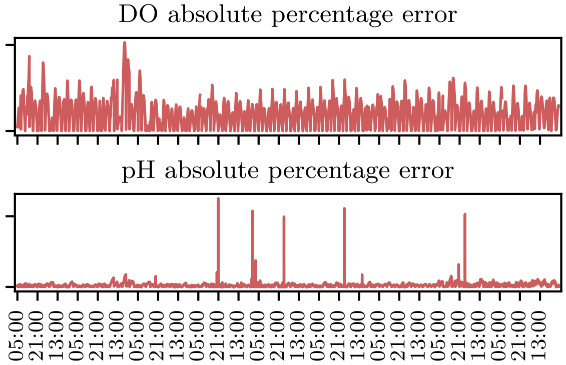

Figure 10: Linear-Gaussian model absolute percentage error given by (29) plot for the nursery pond (pond 2)

	Pond 1	Pond 2
DO	10.46	4.89
ph	0.82	0.57

Table 2: Nonlinear-Gaussian model mean absolute percentage error given by (30)

The latent variables are illustrated in Figure 13. The filtered posterior mean with its standard deviation are plotted in the top figure. The offset model $x_1 = \gamma_t$ is plotted in the second figure. The amplitude model $x_3 = \alpha_t$ is plotted in the third figure. The sinusoidal model $x_4 = sin(\omega t)$ is plotted in the lower figure of Figure 13. The sinusoidal model amplitude remains relatively constant over time. The amplitude model provides the means to vary the sinusoid amplitude. In the linear-Gaussian model results, the amplitude of the sinusoidal model varies significantly over time.

The average values of the absolute percentage error for the ponds are presented in Table 2. The error levels for the nonlinear-Gaussian model are generally higher than those of the linear-Gaussian model. This indicates that the linear-Gaussian model performs better than the nonlinear-Gaussian model.

7 CONCLUSION

In this study, a linear and nonlinear Gaussian state space model are presented. Inference in these models was performed using the FK and UKF respectively. The purpose of the approach is to provide a means to model and forecast diurnal water quality parameters with limited data. The common approaches to water quality parameter modelling found in literature are ecosystem-based models or data-driven models. The model presented provides an alternative to these approaches by providing a data-inspired generative model.

Overall, the results show that the linear-Gaussian model outperforms the nonlinear-Gaussian model. The error results are lower and the modelling is more aligned to the data. The 5-day forecasts are slightly more accurate in the nonlinear-Gaussian model. The general poorer performance of the nonlinear-Gaussian model may relate to the lack of observations presented. The curse of dimensionality suggests that with an increase in a model's parameter

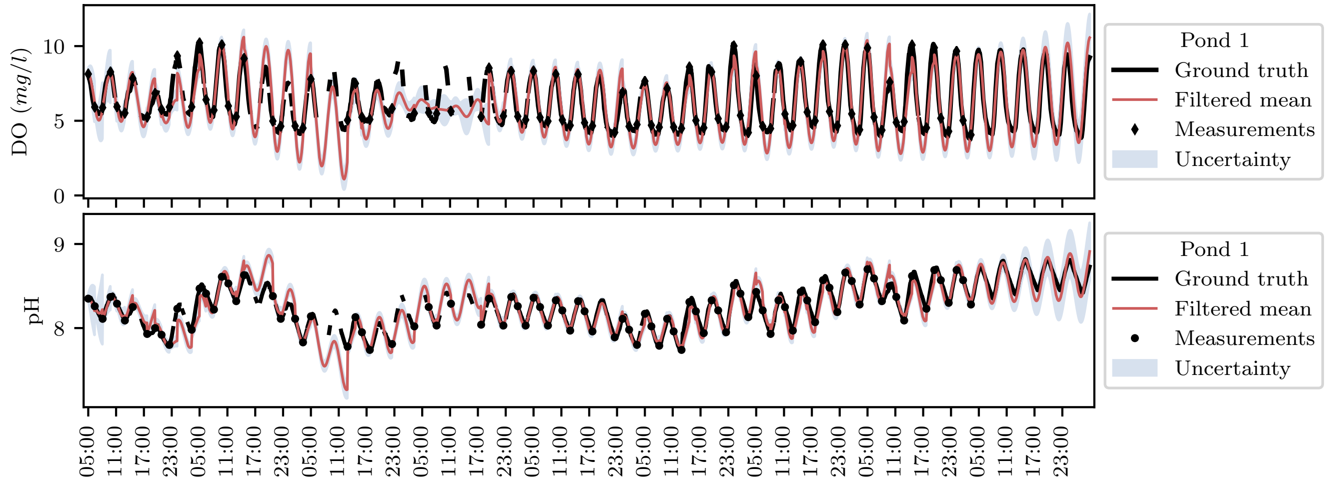

Figure 11: Unscented Kalman filtered results in the grow-out pond (pond 1).

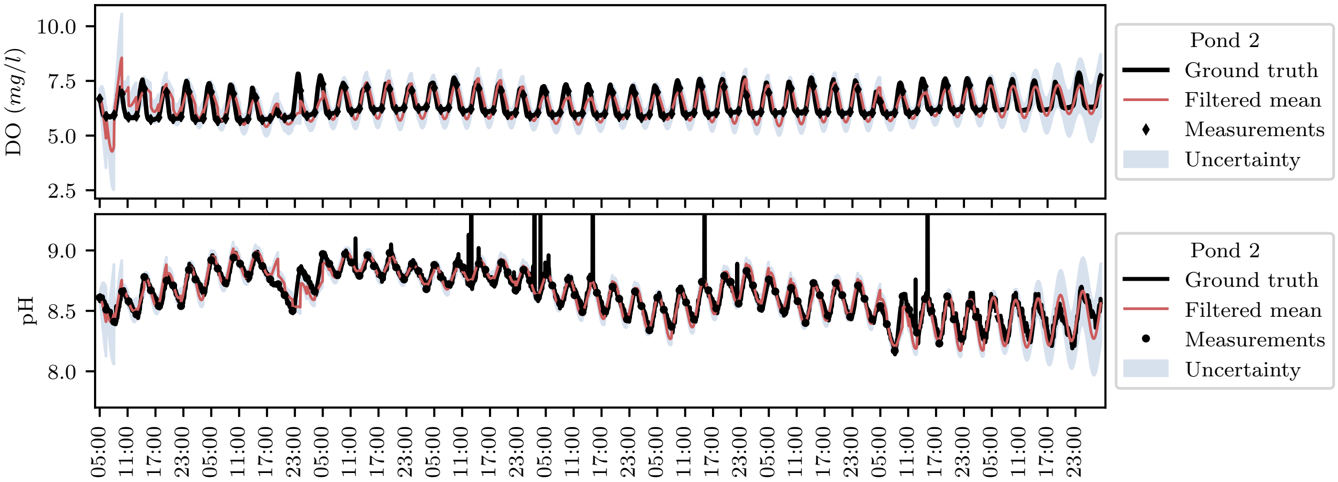

Figure 12: Unscented Kalman filtered results in the nursery pond (pond 2)

space, more data is required to fit the model. Given that prawn farmers sample data relatively infrequently, the less complex, linear-Gaussian model is thus favoured.

The models provide a means to produce relatively accurate forecasts. With the constant velocity process assumptions, these forecasts are however limited to linear trends. The model can be made more complex by, for example assuming a constant acceleration process. Furthermore, more complex waveforms could be modelled by including additional Fourier series components. The curse of dimensionality suggests that this however is not likely to be successful with limited data. The linear trends and sinusoids have however shown to perform well in the considered application.

One of the key features of the models is that they provide a level of uncertainty on their forecasts. The results demonstrate that the uncertainty increase the further into the future the models predict.

The models are demonstrated to adapt to different ponds and variables. The same model was used over differing ponds and water quality variables. DO and pH have different ranges of values and thus initial model parameters were adjusted for these datasets. The same parameters were however used over the different ponds. It is well known that even adjacent ponds can differ significantly in dynamics [6]. This capability is thus highly useful in the prawn pond farming context.

The models are based on the structure of the data rather than the underlying processes that generates the data. This allows flexibility for application of the model to problems other than aquaculture. In general, the model is applicable to data with an additive decomposition of trend, seasonal, and stochastic components.

In future work, the inclusion of various other sensor data could be included in the model. For example, if the chlorophyll level provides an indication of the phytoplankton concentration, it may provide an indication of the waveform offset level γ_t. Such inclusions may however not be practically relevant as farmers generally do not routinely make measurements on phytoplankton biomass [6]. In

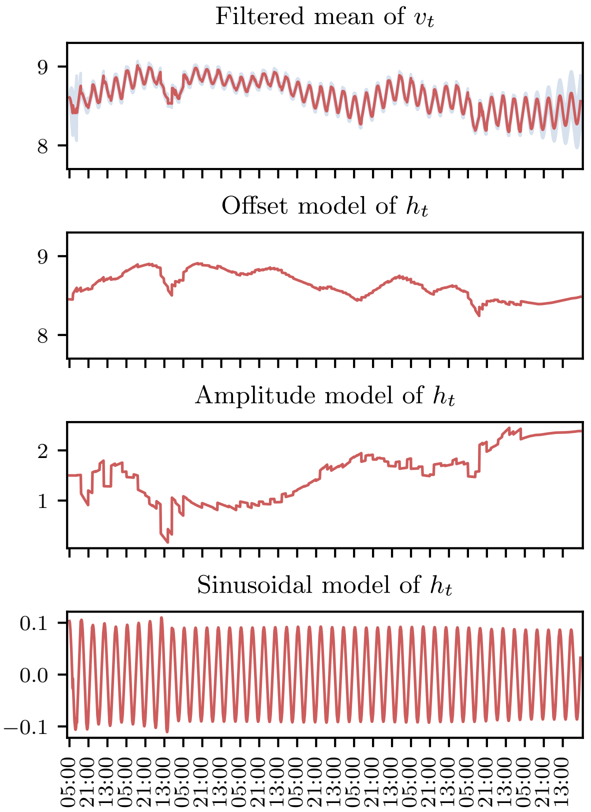

Figure 13: Plot of the UKF filtered posterior mean f_t and elements from the state vector h_t for the nursery pond.

the future research that is being conducted at the CSIRO, these models will be incorporated within the framework of a prawn farm decision support system.

REFERENCES

[1] A.A. Masrur Ahmed. 2017. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *Journal of King Saud University - Engineering Sciences* 29, 2 (2017), 151 – 158. https://doi.org/10.1016/j.jksues.2014.05.001

[2] Davor Antanasijević, Viktor Pocajt, Dragan Povrenović, Aleksandra Perić-Grujić, and Mirjana Ristić. 2013. Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. *Environmental Science and Pollution Research* 20, 12 (01 Dec 2013), 9006–9013. https://doi.org/10.1007/s11356-013-1876-6

[3] Murat Ay and Ozgur Kisi. 2012. Modeling of Dissolved Oxygen Concentration Using Different Neural Network Techniques in Foundation Creek, El Paso County, Colorado. *Journal of Environmental Engineering* 138, 6 (2012), 654–662. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000511 arXiv:https://ascelibrary.org/doi/pdf/10.1061/(ASCE)EE.1943-7870.0000511

[4] D. Barber. 2012. *Bayesian Reasoning and Machine Learning.* Cambridge University Press.

[5] Nikita Basant, Shikha Gupta, Amrita Malik, and Kunwar P. Singh. 2010. Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water âĂŤ A case study. *Chemometrics and Intelligent Laboratory Systems* 104, 2 (2010), 172 – 180. https://doi.org/10.1016/j.chemolab.2010.08.005

[6] Claude E Boyd and Craig S Tucker. 1998. *Pond aquaculture water quality management.* Springer US. https://doi.org/10.1007/978-1-4615-5407-3

[7] Anita Csábrági, Sándor Molnár, Péter Tanos, and József Kovács. 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. *Ecological Engineering* 100 (2017), 63 – 72. https://doi.org/10.1016/j.ecoleng.2016.12.027

[8] Emrah Dogan, BÃijlent Sengorur, and Rabia Koklu. 2009. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management* 90, 2 (2009), 1229 – 1235. https://doi.org/10.1016/j.jenvman.2008.06.004

[9] Vincent Ginot and Jean-Christophe Hervé. 1994. Estimating the parameters of dissolved oxygen dynamics in shallow ponds. *Ecological Modelling* 73, 3 (1994), 169 – 187. https://doi.org/10.1016/0304-3800(94)90061-2

[10] Jianxun He, Angus Chu, M. Cathryn Ryan, Caterina Valeo, and Beryl Zaitlin. 2011. Abiotic influences on dissolved oxygen in a riverine environment. *Ecological Engineering* 37, 11 (2011), 1804 – 1814. https://doi.org/10.1016/j.ecoleng.2011.06.022

[11] Jiacong Huang, Junfeng Gao, Jutao Liu, and Yinjun Zhang. 2013. State and parameter update of a hydrodynamic-phytoplankton model using ensemble Kalman filter. *Ecological Modelling* 263 (2013), 81 – 91. https://doi.org/10.1016/j.ecolmodel.2013.04.022

[12] Kyunghyun Kim, Minji Park, Joong-Hyuk Min, Ingu Ryu, Mi-Ri Kang, and Lan Joo Park. 2014. Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. *Journal of Hydrology* 519 (2014), 2810 – 2821. https://doi.org/10.1016/j.jhydrol.2014.09.073

[13] Joseph H. W. Lee, J. Q. Mao, and K. W. Choi. 2009. The Extended Kalman Filter for Short Term Prediction of Algal Bloom Dynamics. In *Advances in Water Resources and Hydraulic Engineering.* Springer Berlin Heidelberg, Berlin, Heidelberg, 513–517.

[14] J.S. Link, T.F. Ihde, C.J. Harvey, S.K. Gaichas, J.C. Field, J.K.T. Brodziak, H.M. Townsend, and R.M. Peterman. 2012. Dealing with uncertainty in ecosystem models: The paradox of use for living marine resource management. *Progress in Oceanography* 102 (2012), 102 – 114. https://doi.org/10.1016/j.pocean.2012.03.008 End-to-End Modeling: Toward Comparative Analysis of Marine Ecosystem Organization.

[15] Zhimin Lu and Raul H Piedrahita. 1996. *Stochastic Modeling of temperature and dissolved oxygen in stratified fish ponds.* Technical Report. Department of Biological and Agricultural Engineering University of California, Davis, USA. Thirteenth Annual Technical Report.

[16] Heidi Ina Madsen, Jes Vollertsen, and Thorkild Hvitved-Jacobsen. 2007. Modelling the oxygen mass balance of wet detention ponds receiving highway runoff. In *Highway and Urban Environment*, Gregory M. Morrison and Sébastien Rauch (Eds.). Springer Netherlands, Dordrecht, 487–497.

[17] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective.* MIT press.

[18] Ehsan Olyaie, Hamid Zare Abyaneh, and Ali Danandeh Mehr. 2017. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. *Geoscience Frontiers* 8, 3 (2017), 517 – 527. https://doi.org/10.1016/j.gsf.2016.04.007

[19] R Pastres, S Ciavatta, and C Solidoro. 2003. The Extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. *Ecological Modelling* 170, 2 (2003), 227 – 235. https://doi.org/10.1016/S0304-3800(03)00230-8 ISEM The third European Ecological Modelling Conference.

[20] Vesna Ranković, Jasna Radulović, Ivana Radojević, Aleksandar Ostojić, and Ljiljana Čomić. 2010. Neural network modeling of dissolved oxygen in the GruÅ¿a reservoir, Serbia. *Ecological Modelling* 221, 8 (2010), 1239 – 1244. https://doi.org/10.1016/j.ecolmodel.2009.12.023

[21] Chris Robertson (Ed.). 2006. *Australian prawn farming manual: health management for profit.* Queensland Department of Primary Industries and Fisheries (QDPI&F).

[22] Bernhard H. Schmid and Jari Koskiaho. 2006. Artificial Neural Network Modeling of Dissolved Oxygen in a Wetland Pond: The Case of Hovi, Finland. *Journal of Hydrologic Engineering* 11, 2 (2006), 188–192. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:2(188) arXiv:https://ascelibrary.org/doi/pdf/10.1061/(ASCE)1084-0699(2006)11:2(188)

[23] Kunwar P. Singh, Ankita Basant, Amrita Malik, and Gunja Jain. 2009. Artificial neural network modeling of the river water qualityâĂŤA case study. *Ecological Modelling* 220, 6 (2009), 888 – 895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

[24] S. Thrun, W. Burgard, and D. Fox. 2005. *Probabilistic Robotics.* MIT Press.

[25] E. A. Wan and R. Van Der Merwe. 2000. The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373).* 153–158. https://doi.org/10.1109/ASSPCC.2000.882463

[26] Zhen Xu and Y. Jun Xu. 2016. A Deterministic Model for Predicting Hourly Dissolved Oxygen Change: Development and Application to a Shallow Eutrophic Lake. *Water* 8, 2:41 (2016). https://doi.org/10.3390/w8020041

[27] YSI. [n. d.]. EXO2 Multiparameter Sonde. ([n. d.]). Retrieved February 9, 2018 from https://www.ysi.com/EXO2

[28] P. Zarchan and H. Musoff. 2000. *Fundamentals of Kalman Filtering: A Practical Approach.* Number v. 190, pt. 1 in Fundamentals of Kalman Filtering: A Practical Approach. American Institute of Aeronautics and Astronautics, Incorporated.