Data Driven Study of Estuary Hypoxia

Md Monwer Hussain
University of New-Brunswick
Fredericton, New-Brunswick, Canada
Monwer.305@unb.ca

Guillaume Durand
National Research Council Canada
Moncton, New-Brunswick, Canada
Guillaume.Durand@nrc-cnrc.gc.ca

Michael Coffin
Department of Fisheries and Oceans Canada
Moncton, New-Brunswick, Canada
Michael.Coffin@dfo-mpo.gc.ca

Julio J. Valdés
National Research Council Canada
Ottawa, New-Brunswick, Canada
Julio.Valdes@nrc-cnrc.gc.ca

Luke Poirier
Department of Fisheries and Oceans Canada
Moncton, New-Brunswick, Canada
Luke.Poirier@dfo-mpo.gc.ca

Abstract

This paper presents a data driven study of dissolved oxygen times series collected in Atlantic Canada. The main motivation of presented work was to evaluate if machine learning techniques could help to understand and anticipate hypoxic episodes in nutrient-impacted estuaries, a phenomenon that is exacerbated by increasing temperature expected to arise due to changes in climate. A major constraint of the analysis was limiting ourselves to a single variable, the dissolved oxygen time series. Our preliminary findings show that recurring neural networks, and in particular LSTM, may be capable of predicting short horizon levels while traditional analyses are adequate for longer range hypoxia prevention.

1 Introduction

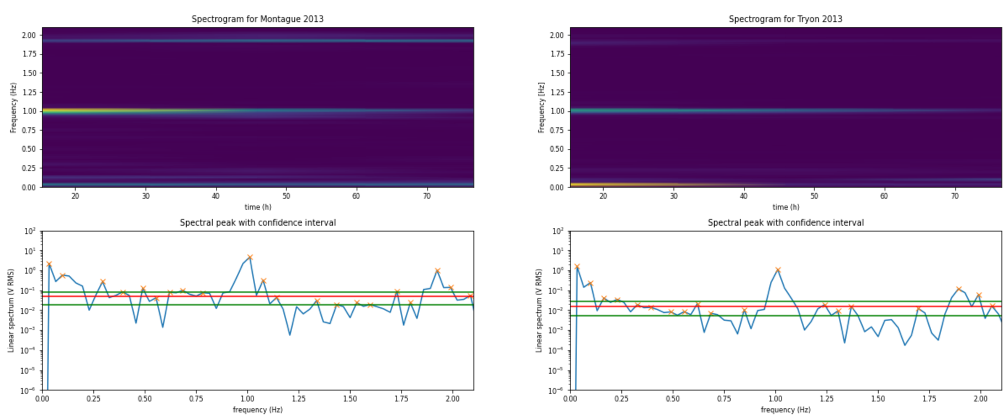
Nutrient loading to waterways is increasing worldwide due to anthropogenic activities, specifically increased land-use, agriculture, industry and animal waste. As a result, simple algal species capable of outcompeting vascular plants for nutrients and space become dominant, i.e., the addition of otherwise limiting nutrients results in a eutrophic environment. There are myriad consequences to this change as both habitat and water quality decline and in severe cases can result in mass mortality of fauna. Given that eutrophication is a process and is on a continuum, it is imperative to identify where a particular system is on this continuum. Dissolved oxygen (DO) has proven effective at delineating trophic status of estuaries and is a key component of ecosystem function, as it reflects oxygen production through photosynthesis, oxygen consumption through respiration (including decomposition of organic matter), is critical for animal health, and its absence (anoxia) is a telltale symptom of a eutrophic environment. Unfortunately, “there is no other environmental variable of such ecological importance to coastal ecosystems that has changed so drastically in such a short period of time as a result of human activities as dissolved oxygen” [7]. Further, projected increases in temperature are a harbinger of increased hypoxic/anoxic episodes as dissolved oxygen solubility declines with temperature, not to mention increased oxygen demand as metabolism increases concomitantly. Instruments for monitoring dissolved oxygen optically are relatively inexpensive and accurately record data without drift. While most eutrophication monitoring programs rely on a suite of environmental indicators, analysis of high

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.

frequency dissolved oxygen time series data provides a more parsimonious approach and comparable accuracy for nutrient impacted estuaries and bays in Atlantic Canada [2]. The analytical tools used for dissolved oxygen time-series data thus far have involved substantial data loss through the establishment of biologically relevant metrics that reflect symptoms of eutrophication, i.e., “high” or “low” oxygen [1]. This prior approach was prescriptive but was only useful for classifying the status of estuaries and bays at an annual time-step and was therefore insufficient for predicting future dissolved oxygen conditions at any temporal scale. Herein we apply Machine Learning to model DO evolution over time and evaluate if models are capable of anticipating anoxia. For developing nations in particular, this methodology presents a potentially powerful tool for predicting water quality in nutrient-impacted environments enabling proactive decision-making towards sustainable fisheries.

2 Data Exploration

The study utilized optical dissolved oxygen loggers installed in multiple Atlantic Canada estuaries to collect hourly oxygen data from early July to late September. The data set contained a total of 62 time series from 2013 to 2020 from which 37 without missing data were retained for analyses. Time series of Montague and Tryon sites for summer 2013 are represented in Figure 2a and 2b. All the 37 time series are characterized by highly variable data, with dissolved oxygen ranging from anoxic (0 mg/L) to supersaturated (> 10 mg/L) within hours at nutrient-impacted sites but hovering closer to 8 mg/L at sites with lower nutrient loading. An example of a site with minimal nutrient impact and few hypoxic instances would be Tryon (2013) which is in direct contrast to a highly nutrient impacted site Montague (2013). In an attempt to understand the significant variations of DO levels, we conducted a frequency analysis.



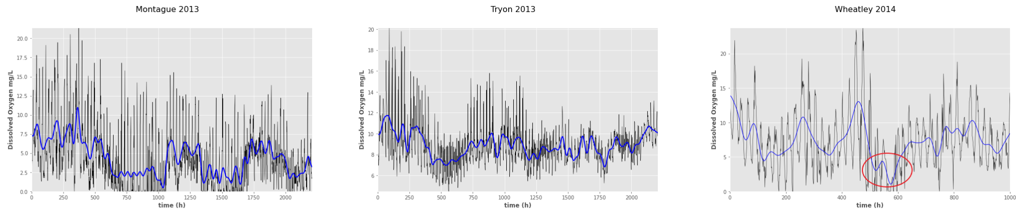
(a) Montague 2013 (Bad Site)   (b) Tryon 2013 (Good Site)

Figure 1: Spectrograms with significant spectral peaks for Montague and Tryon 2013

Figure 1 highlights spectrograms and spectral graphs of Montague and Tryon (2013). A spectrogram is a visual representation of the signal strength of a signal over time at various frequencies present in a waveform. A spectrogram is a two-dimensional visual representation with time as horizontal axis and frequency as vertical axis. There is a 3rd variable or dimension which is represented by color. In our example, the more reddish the color is, the more energy or intensity there is in the area.

In Figure 1, spectrograms were created using 24 hours as sample frequency. If we look at both plots in Figure 1, we can see there are most visible (significant) cycles at 2 Hz and 1 Hz. The 1 Hz refers to a 24 hours cycle while the 2 Hz refers to a 12 hours cycle. We assumed that these cycles may be respectively related to tides and days and that they could explain the choppiness of DO fluctuations.

Using a low-pass filter, it is possible to remove all cycles with periods below 24 hours (above 1 Hz). The filtered time series are in blue in Figure 2. As expected, time series shapes are smoother and easier to visually interpret. For instance, it may be possible to identify suspicion of hypoxia episodes by defining an amount of time under a specific threshold as proposed by [2].



(a) Montague 2013 (b) Tryon 2013 (c) Wheatley 2014 (subset)

Figure 2: (2a, 2b), Effect of low-pass filter (blue) on Montague and Tryon 2013 time series (black). (2c) Subset of the original Wheatley 2014 time series highlighting a documented hypoxia episode (red)

3 Preliminary Results

We conducted a significant number of analyses using supervised and unsupervised techniques. We describe in the following some that yielded interesting findings.

Distance Method/ Cluster Method	DTW	Area	PCM	Frechet Distance	Curve Length
Optics	0.706	0.699	1.278	1.628	1.349
HDBscan	0.706	0.699	0.783	0.680	1.048
Spectral Clustering	0.999	1.000	1.272	0.957	0.953
Gaussian Mixture	0.589	0.587	0.716	0.569	0.653
K-Means	0.700	0.696	0.718	0.569	0.653
K-Medoids	1.035	0.830	1.088	0.725	1.421

Table 1: Davies Bouldin scores for different clustering methods and similarity measures (Lower values indicating better clustering)

Clustering was applied on the data set using the distance metrics and the methods listed in Table 1. The best combination using Davies Bouldin score [4] (supported by Silhouette scores) was obtained for Gaussian mixture on Area Distance. Three clusters were obtained as summarized in Table 2.

Gaussian Mixture Clustering on Area distance		
Cluster-1	Cluster-2	Cluster-3
Bideford 2014	Kildare 2014	Enmore 2014
Mill 2014	Kildare 2013	Bideford 2013
Wheatley 2014	**Montague 2013**	Bouctouche 2013
Mill 2013	Stanley 2013	Dunk 2013
Souris 2013	Kildare 2018	Enmore 2013
Wheatley 2013	Hunter River 2019	Tatamagouche 2013
Wilmot 2013		**Tryon 2013**
Wheatley 2018		West 2013
Covehead 2019		Pugwash 2018
Mill 2019		River John 2018
Wilmot 2019		etc...

Table 2: Cluster information for Gaussian Mixture Model on Area Distance

A first finding was that the cluster distribution may be explained through similar frequency of anoxic episodes. Cluster-3 is the largest and it contains 20 of the 37 series and is populated by time series characterized by normoxia, and/or a lack of incidence of anoxia, i.e., they are good sites. Cluster-2 contains time series where anoxia occurs frequently, they are nutrient-impacted sites. Finally Cluster-1 contains sites where anoxia occurs periodically but not as frequently as Cluster-2, perhaps in the process of improving towards Cluster-3 or further declining towards Cluster-2. Time series prediction was accomplished using ARIMA [5] "Auto Regressive Integrated Moving Average", which is a common tool for this type of analysis. Specifically, we compare results between ARIMA and Long Short-Term Memory (LSTM) neural networks [6] in Table 3. LSTM is a special kind of Recurrent Neural Network (RNN) which solves the Long Term Dependency problem that occurs with RNN. We have used min-max scalar to pre-process our data when using LSTM. We used the same model composed of two hidden layers of 20 neurons with ADAM optimizer and "mean squared error" as loss function. Training size and testing were 1300 and 309 hours, respectively, and we used the last 24 hours of observations as input to predict the subsequent 12 hours. For ARIMA we used the whole time series (1609 points) minus the prediction horizon for training. In our experiment, we were able to observe that ARIMA performs well in time series that were low-pass filtered (i.e., where all cycles occurring within 24 hours, like tides and daylight, were removed) but performed very poorly on raw time series as illustrated in Table 3. We increased ARIMA prediction horizon to 48 hours on low-pass filtered data to highlight that its performance metrics almost compare to LSTM on a 12 hours prediction horizon. A 12 hours prediction horizon for ARIMA on filtered data would lead to almost perfect results. However, on unfiltered data, and in particular for Tryon (2013), performances are poor as ARIMA does not even follow the trend (Pears. -.52).

Data	Method	Time series	Prediction Horizon	RMSE	MAE	Pearson
Unfiltered Data	LSTM	Wheatley 2014	12h	2.38	1.87	0.76
		Tryon 2013		1.52	1.09	0.50
		Montague 2013		2.82	1.88	0.55
	ARIMA	Wheatley 2014	12h	3.52	3.77	0.78
		Tryon 2013		1.81	1.92	-0.52
		Montague 2013		4.25	5.96	0.21
Filtered Data	LSTM	Wheatley 2014	12h	0.34	0.24	0.99
		Tryon 2013		0.14	0.09	0.96
		Montague 2013		0.35	0.26	0.97
	ARIMA	Wheatley 2014	48h	0.37	0.59	0.87
		Tryon 2013		0.12	0.17	1
		Montague 2013		0.36	0.56	0.66

Table 3: Prediction results of LSTM and ARIMA on unfiltered and low-pass filtered data. RMSE is the Root Mean Squared Error and MAE is the Mean Absolute Error.

Conversely, LSTM seems capturing more adequately unfiltered time series chopiness as illustrated also by Figure 3. The model follows the observed trend but has difficulties to anticipate amplitudes. One explanation could be that LSTM manages, to some extent, at exploiting tides and days cycles information contained in the time series.

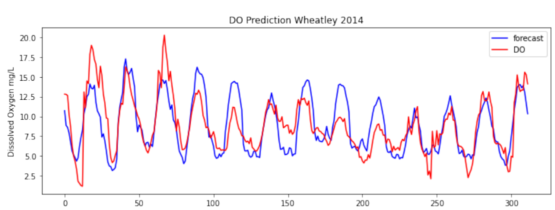


Figure 3: Prediction for Wheatley 2014 (Unfiltered)

4 Conclusion

As expected, the trade-off between noise and useful information is not obvious. DO measurement can be altered by several external factors. Tides and days cycle are obvious but others remain to be investigated. For instance, while the DO probes are reliable drifting algae could smother the probe resulting in artificially low DO measurement.

This trade-off needs also to be questioned through the lens of its applicability. Anoxia is unequivocally detrimental to animal health but the severity of the consequences of anoxia depend on both the state of the system prior to its occurrence, the duration of the episode, and synergistic effects from other stressors, e.g., temperature, disease, etc. For these reasons it isn't recommended that a biological outcome, like animal mortality, be the endpoint of interest but the occurrence of anoxia itself. The capacity to predict intermittent anoxia could help identify timing of fisheries and/or aquaculture husbandry to avoid additional stress on animals, particularly in areas where temperature is increasing.

For this purpose, we have shown, that LSTM gives relatively good results in filtered signals but may be performing less accurately than more traditional methods, like ARIMA. We also observed that LSTM produced somewhat monotonous sequences as predictions on such data. For instance, predicting the last known value for the whole prediction frame may provide similar MAE and RMSE than LSTM on filtered data. Conversely, LSTM showed superior performance on the raw time series which is corroborated by recent work using LSTM to fill data gaps in coastal time series measurements, including DO [3].

Finally, this preliminary study was limited to an univariate approach, modelling time series separately. While bringing interesting developments in the understanding of dissolved oxygen concentrations, it did not explore the relationship between sites within a given cluster. Consequently, as a natural development, we plan to evaluate multivariate approaches like convolutional LSTM networks [9, 8].

References

[1] Michael Coffin, Simon Courtenay, Christina Pater, and Michael van den Heuvel. An empirical model using dissolved oxygen as an indicator for eutrophication at a regional scale. *Marine Pollution Bulletin*, 133, 05 2018.

[2] Michael Coffin, Luke Poirier, Jeff Clements, Eva Dickson, Thomas Guyondet, and Michael van den Heuvel. Dissolved oxygen as a Marine Environmental Quality (MEQ) measure in upper estuaries of the southern Gulf of St Lawrence: Implications for nutrient management and eelgrass (Zostera marina) coverage. Technical report, Department of Fisheries and Oceans Canada Can. Sci. Advis. Sec. Res. Doc. 2021/056. iv +27, 2021.

[3] Steefan Contractor and Moninya Roughan. Efficacy of feedforward and lstm neural networks at predicting and gap filling coastal ocean timeseries: Oxygen, nutrients, and temperature. *Frontiers in Marine Science*, 8:368, 2021.

[4] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI*, 1(2):224–227, 1979.

[5] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis, Forecasting and Control*. Prentice Hall, 1976.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Dan Laffoley and John M Baxter. *Ocean deoxygenation: Everyone's problem-Causes, impacts, consequences and solutions*. IUCN Gland, 2019.

[8] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 802–810, Cambridge, MA, USA, 2015. MIT Press.

[9] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. volume 34, pages 914–921, Apr. 2020.