Detecting Sensor Drift at Clayoquot Slope

Team 52 Hz Whales

Christopher Chiu

Vu Nguyen (Liam)

Shu Han

Zirui Li (Grace)

Zixuan Deng

Ocean Networks Canada

University of Victoria

Agenda

- Introduction
- Key Issues
- Methodology
- Proposed Solutions & Results
- Summary
- Conclusion
- Future Work

University of Victoria

About Us

Christopher Chiu

MADS

Zirui Li (Grace)

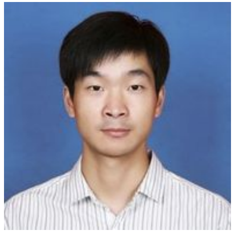
MADS

Zixuan Deng

MADS

Vu Nguyen (Liam)

MTIS



Shu Han

MTIS

University of Victoria

Introduction

- Stakeholder - Ocean Networks Canada
- Problem - Sensor drift in the Clayoqout Slope region
  - Conductivity, Temp, and Pressure

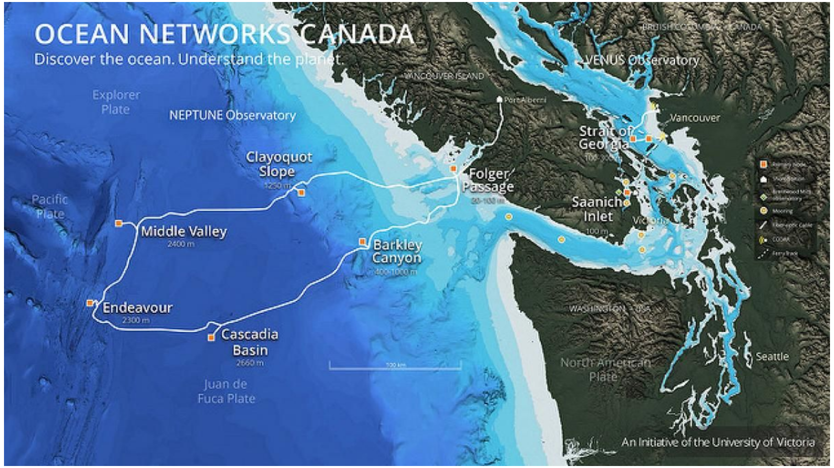
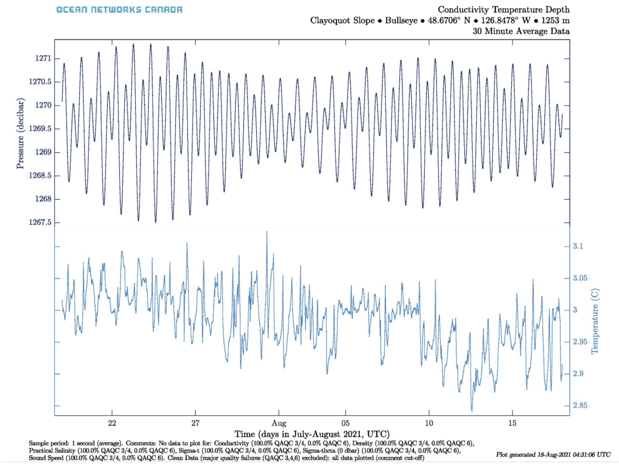


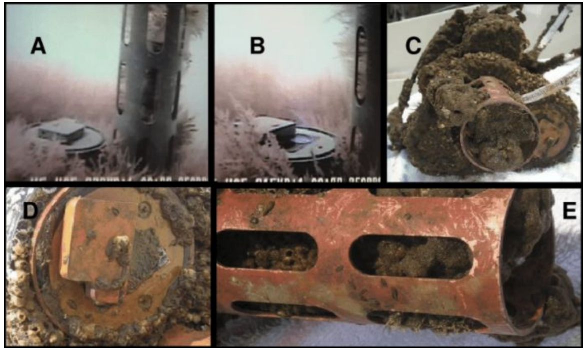
Image ref: https://www.oceannetworks.ca/



University of Victoria

Introduction

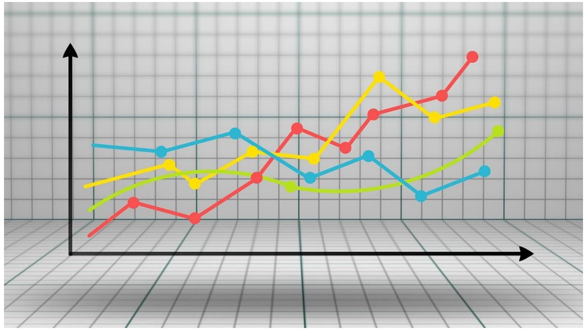
- Cause - Environmental and phsyical
  - Biofouling, Physical drift, Aging of components, Damaged, etc.



- Target - Filter out the anomaly data

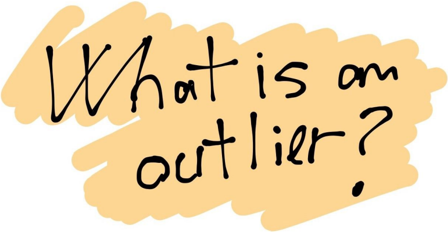
Image ref:
https://www.researchgate.net/figure/WQM-after-66-days-deployment-in-San-Luis-Bay-CA-A-ECO-with-copper-shutter-covering_fig4_224305743
https://amloceanographic.com/biofouling-control/

Key Issues



Multivariate
Time-Series dataset

(Conductivity, Pressure, and Temperature)



Ambiguity among anomalies

(Noise vs. Missing Data vs. Extreme Outliers)

Unclear ground truth labels

(QA/QC flags)

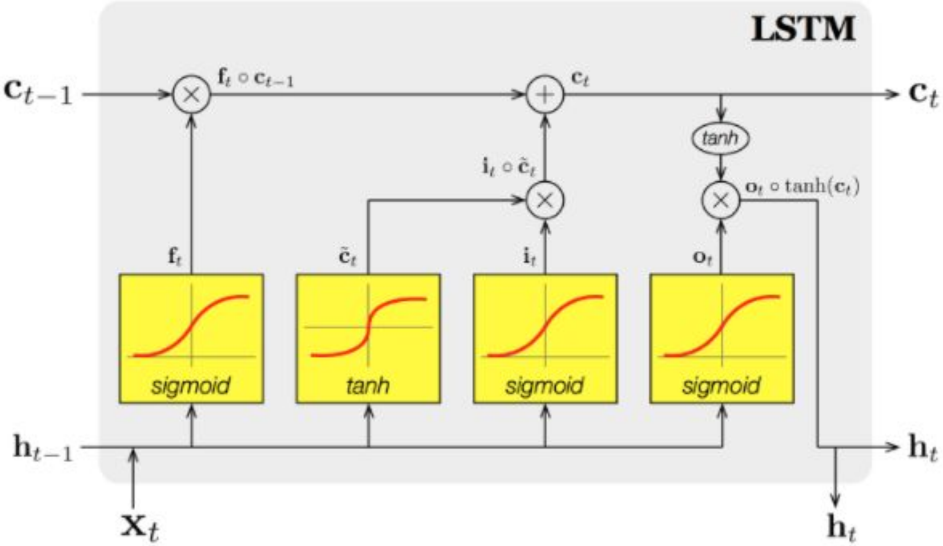
Methodology

- Prediction-based methods
  - LSTM
  - Prophet by Facebook
- Density-based methods
  - Local Outlier Factor from PyOD
- Proximity-based methods
  - KNN from sci-kit learn
  - Boxplots

University of Victoria

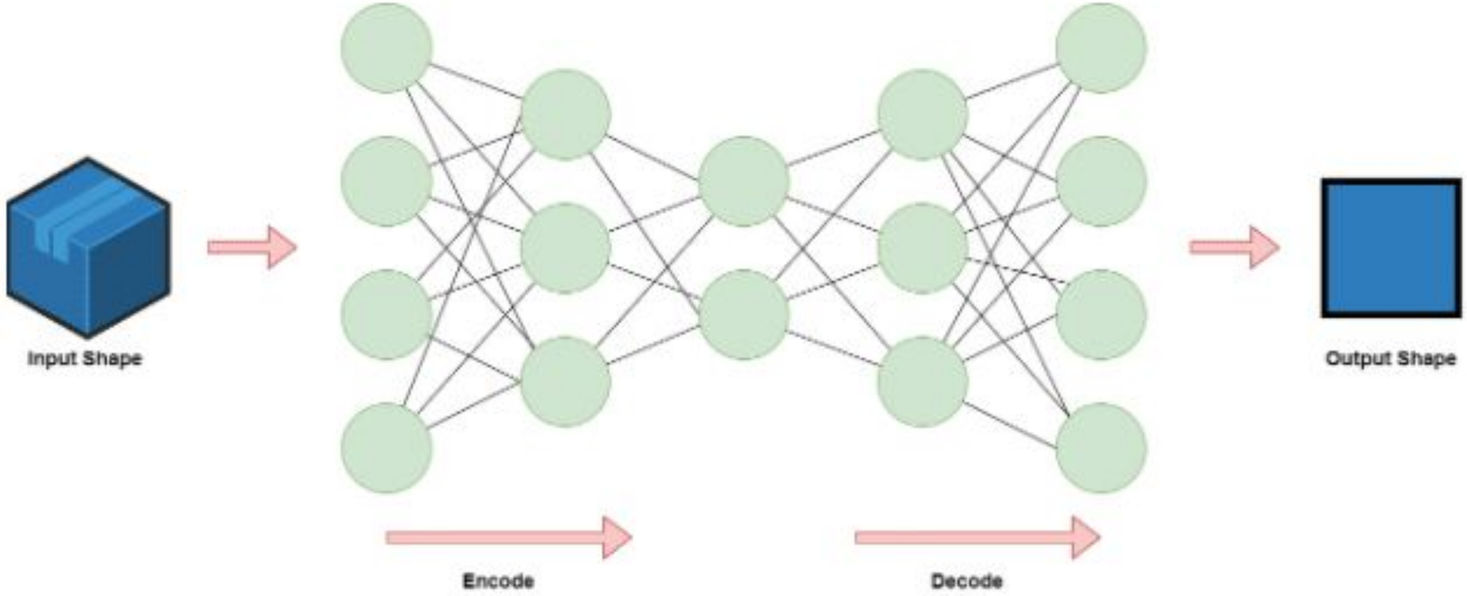
Long Short-Term Memory (LSTM)

LSTM is commonly used for solving sequence prediction/detection problems, such as detecting and predicting anomaly in sales by finding patterns in stock markets' data. It is a special kind of Recurrent Neural Network (RNN).



University of Victoria

LSTM Autoencoder



University of Victoria

LSTM

Result

```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_16 (LSTM)               (None, 128)               66560
_________________________________________________________________
dropout_16 (Dropout)         (None, 128)               0
_________________________________________________________________
repeat_vector_8 (RepeatVecto (None, 7, 128)            0
_________________________________________________________________
lstm_17 (LSTM)               (None, 7, 128)            131584
_________________________________________________________________
dropout_17 (Dropout)         (None, 7, 128)            0
_________________________________________________________________
time_distributed_8 (TimeDist (None, 7, 1)              129
=================================================================
Total params: 198,273
Trainable params: 198,273
Non-trainable params: 0
_________________________________________________________________
```

University of Victoria

LSTM

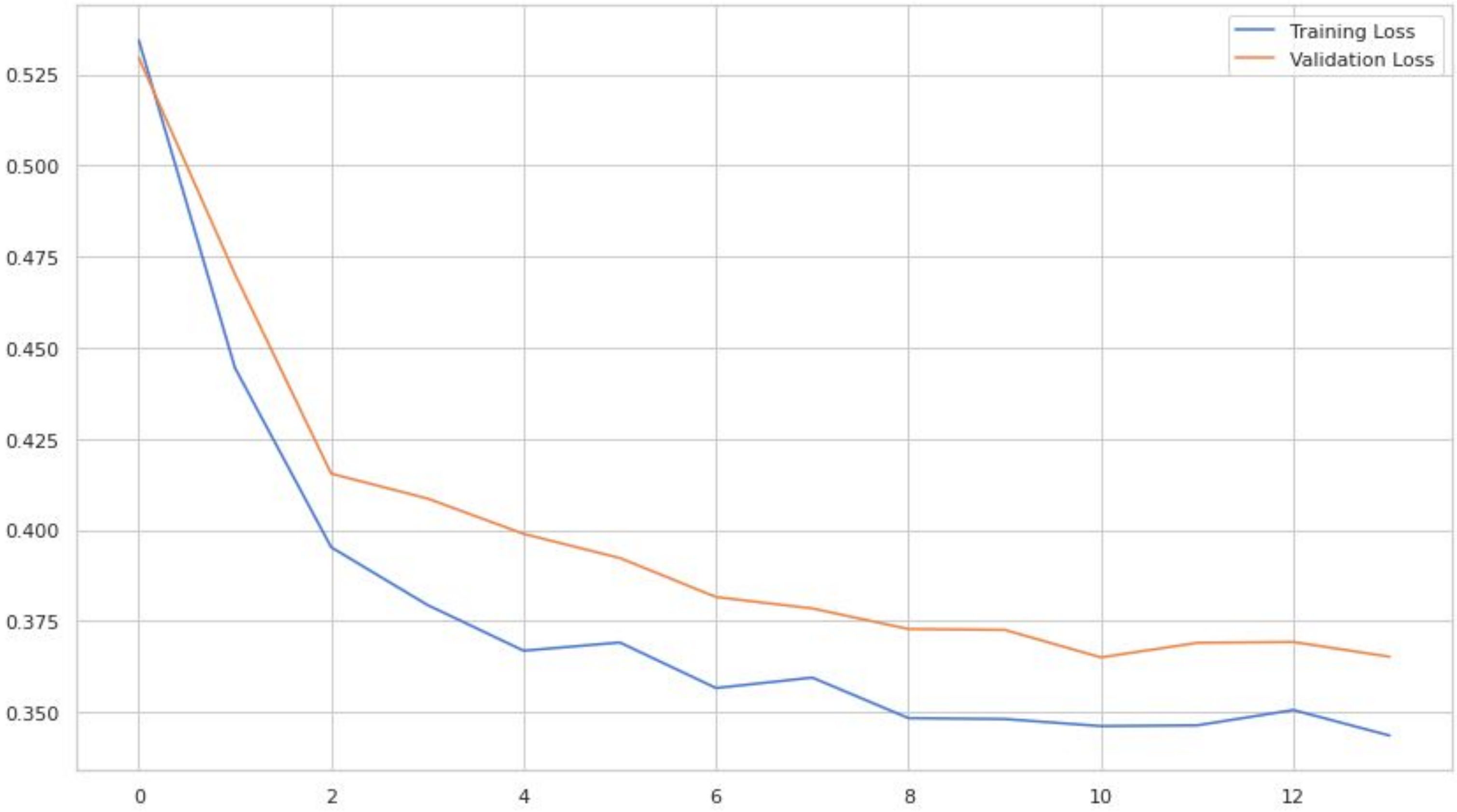
Temperature Feature

Data range: Jul 2011 - Nov 2015

University of Victoria

LSTM

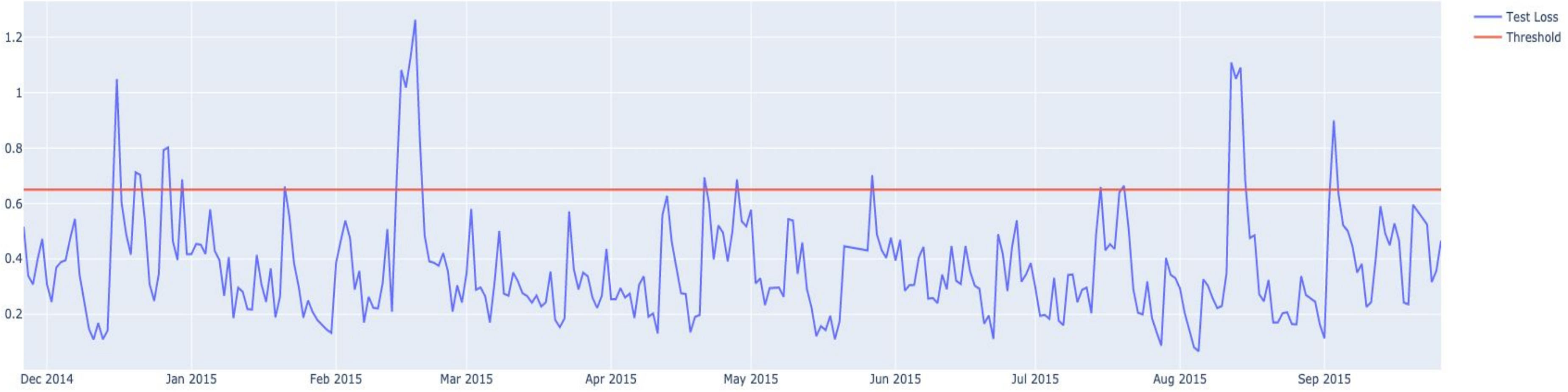
MAE
Result



University of Victoria

LSTM

Test data: Dec 2014 - Nov 2015



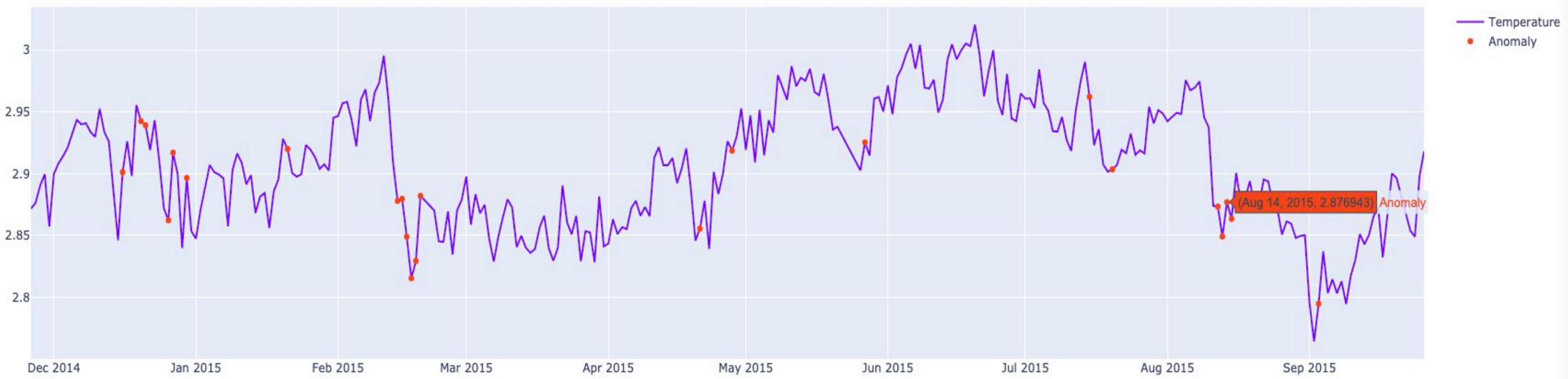
LSTM

Result

	Time	Temperature	loss	threshold	anomaly
1248	2014-12-16	-0.035957	1.098296	0.65	True
1252	2014-12-20	0.713784	0.723746	0.65	True
1253	2014-12-21	0.652169	0.721569	0.65	True
1258	2014-12-26	-0.742521	0.827854	0.65	True
1259	2014-12-27	0.250214	0.841219	0.65	True
1262	2014-12-30	-0.121874	0.737595	0.65	True
1284	2015-01-21	0.303742	0.662332	0.65	True
1308	2015-02-14	-0.459072	0.705486	0.65	True
1309	2015-02-15	-0.428728	1.122368	0.65	True
1310	2015-02-16	-0.983811	1.035288	0.65	True
1311	2015-02-17	-1.595199	1.162947	0.65	True
1312	2015-02-18	-1.342025	1.307322	0.65	True
1313	2015-02-19	-0.384844	0.846507	0.65	True
1374	2015-04-21	-0.865907	0.745085	0.65	True
1381	2015-04-28	0.279864	0.685080	0.65	True
1406	2015-05-27	0.400806	0.717409	0.65	True
1455	2015-07-15	1.070477	0.695192	0.65	True
1459	2015-07-19	-0.029993	0.651793	0.65	True
1460	2015-07-20	0.005123	0.678336	0.65	True
1483	2015-08-12	-0.541803	1.141767	0.65	True
1484	2015-08-13	-0.981683	1.070405	0.65	True
1485	2015-08-14	-0.477323	1.126406	0.65	True
1486	2015-08-15	-0.722262	0.684163	0.65	True
1504	2015-09-02	-2.522310	0.668894	0.65	True
1505	2015-09-03	-1.969966	0.947878	0.65	True

LSTM

Test data: Dec 2014 - Nov 2015



Prophet - Definition

PROPHET

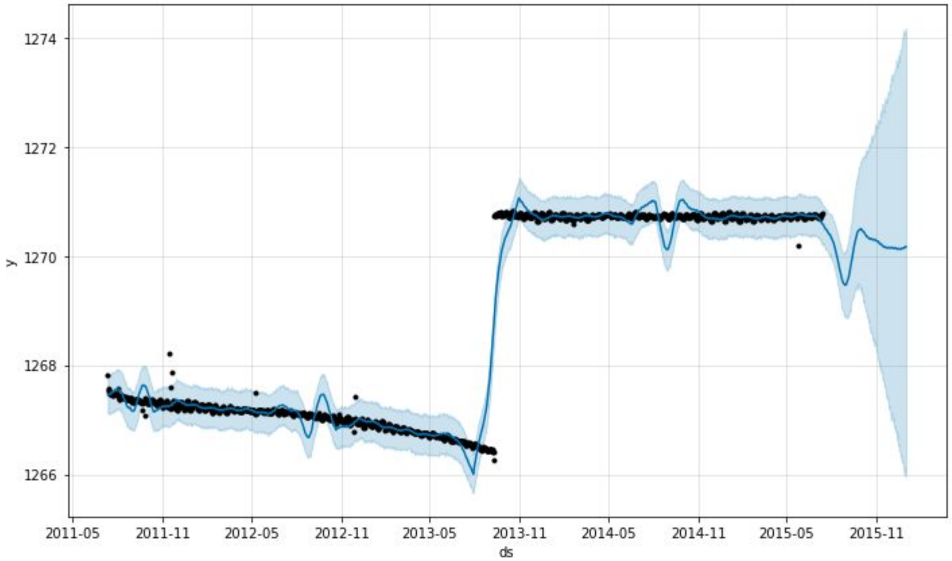
FACEBOOK

Prophet is an open-source time series forecasting tool developed by Facebook. It is a procedure for **forecasting time series data** based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have **strong seasonal effects and several seasons of historical data**. Prophet is robust to missing data and shifts in the trend and typically handles outliers well [13]. Prophet allows analysts to create a forecasting task more conveniently and directly; it is user-friendly for non-programming users.

Prophet - Implementation

```
# Trian with Temperature data, change y with ‘pressure_y’ when training Pressure Data
dict_pro = {'ds': ds_array,
            'y': temp_y
            }
df = pd.DataFrame(dict_pro)
m = Prophet()
m.fit(df)
# Generate a prediction dataframe for next 172 days
future = m.make_future_dataframe(periods=172)
forecast = m.predict(future)
forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']]
# Plot the forecast diagram
fig = m.plot(forecast)
```

Prophet - Pressure Prediction



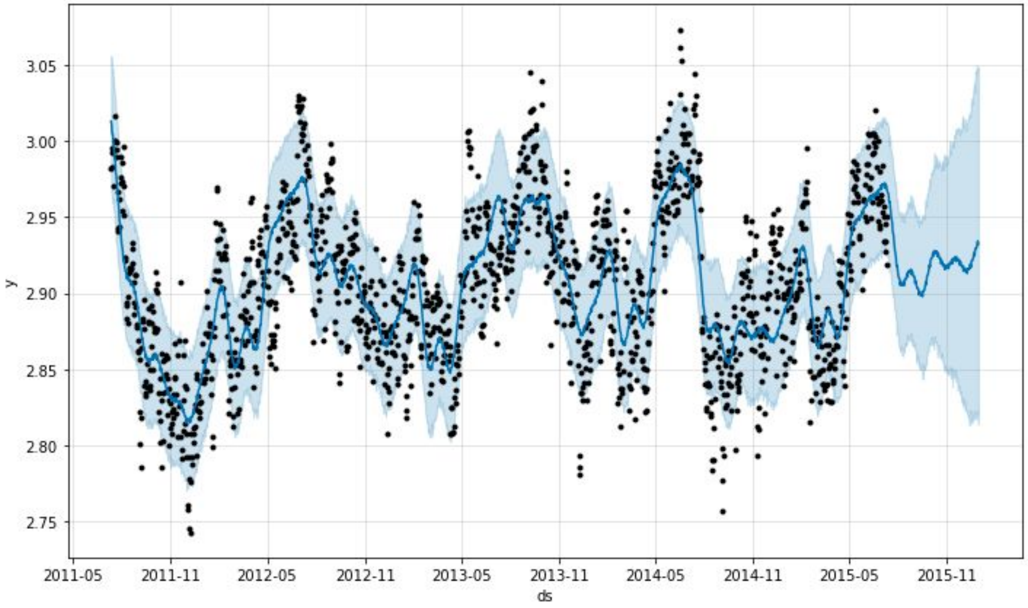
	ds	yhat	yhat_lower	yhat_upper
0	2011-07-12	1267.456168	1267.106491	1267.827591
1	2011-07-13	1267.465097	1267.106449	1267.811958
2	2011-07-14	1267.462862	1267.098548	1267.783808
3	2011-07-15	1267.463398	1267.116975	1267.808191
4	2011-07-16	1267.468520	1267.107512	1267.823624
...	...	...	...	...
1620	2015-12-27	1270.142201	1265.734685	1274.488440
1621	2015-12-28	1270.149851	1265.836505	1274.721739
1622	2015-12-29	1270.166057	1265.718876	1274.681745
1623	2015-12-30	1270.174610	1265.771354	1274.821504
1624	2015-12-31	1270.172119	1265.621072	1274.813690

1625 rows × 4 columns

University of Victoria

Historical Samples Range: 2011-07-12 to 2015-07-12
Prediction Range: 2015-07-12 to 2015-12-31

Prophet - Temperature Prediction



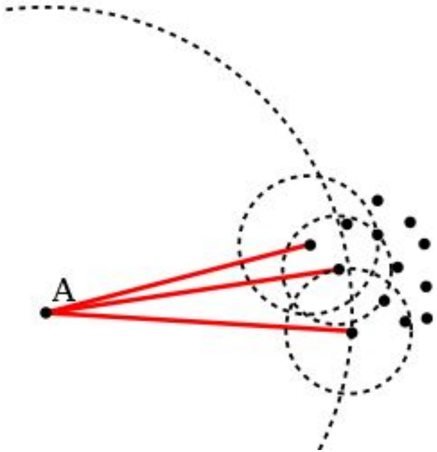
	ds	yhat	yhat_lower	yhat_upper
0	2011-07-12	3.012963	2.967109	3.056031
1	2011-07-13	3.011066	2.967106	3.052671
2	2011-07-14	3.005131	2.959509	3.046324
3	2011-07-15	3.004860	2.962892	3.047780
4	2011-07-16	3.001371	2.959074	3.041897
...	...	...	...	...
1620	2015-12-27	2.929596	2.817314	3.044462
1621	2015-12-28	2.930403	2.821874	3.049795
1622	2015-12-29	2.932580	2.818375	3.046258
1623	2015-12-30	2.934303	2.823418	3.048223
1624	2015-12-31	2.932194	2.813871	3.048678

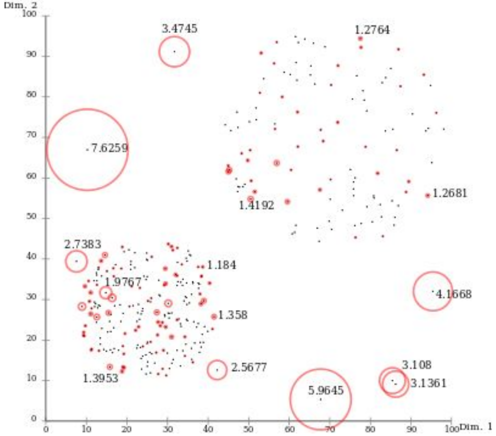
1625 rows × 4 columns

Historical Samples Range: 2011-07-12 to 2015-07-12
Prediction Range: 2015-07-12 to 2015-12-31

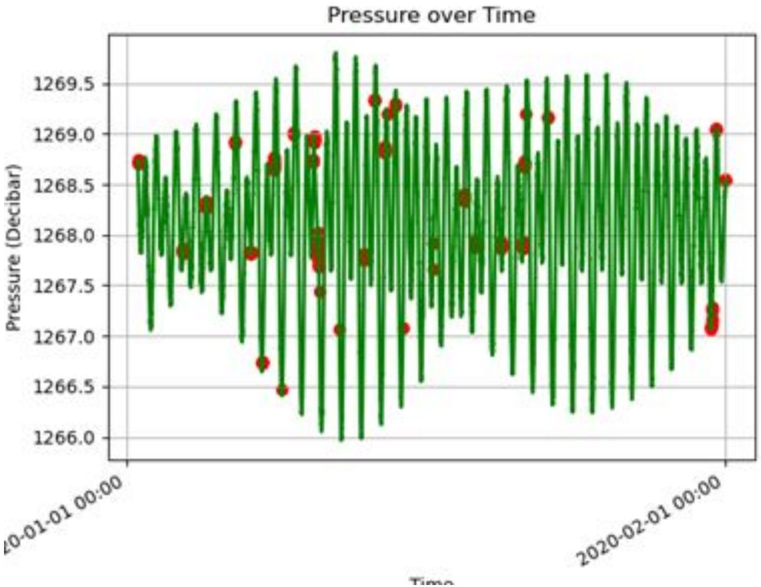
Local Outlier Factors

"The local outlier factor is based on a concept of a **local density**, where locality is given by k nearest neighbors, whose distance is used to estimate the density. By **comparing the local density of an object to the local densities of its neighbors**, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be **outliers**" [14]



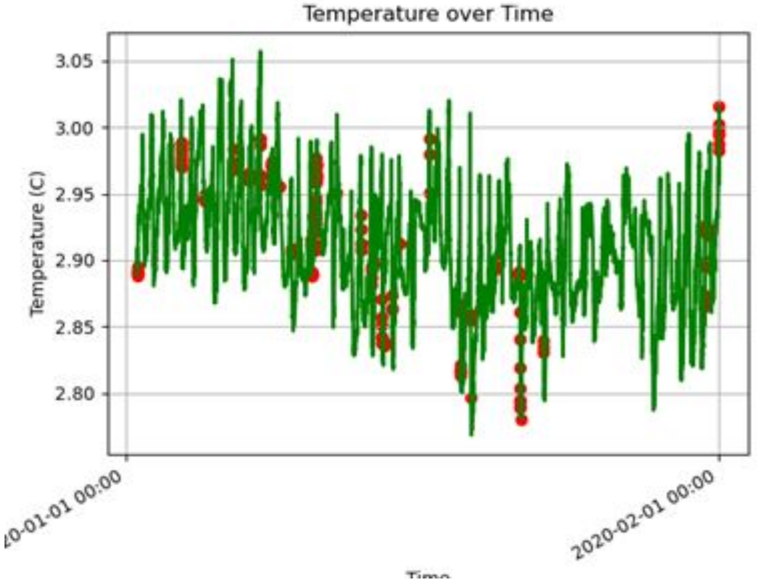


LOF - Pressure Prediction



University of Victoria

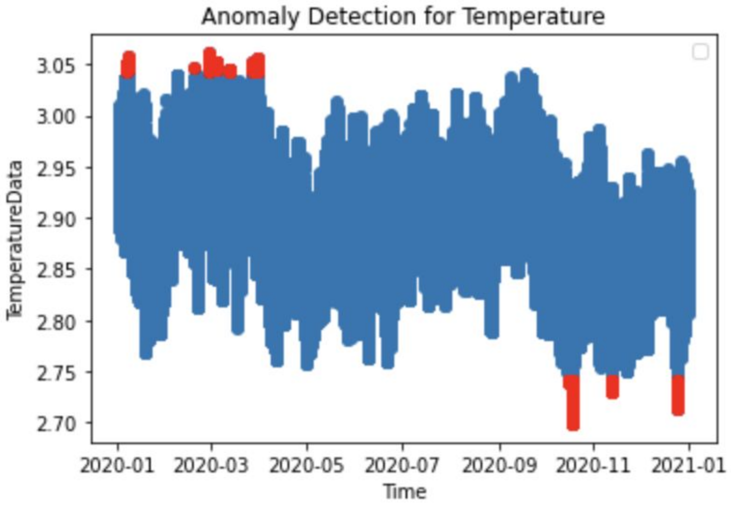
LOF - Temperature Prediction

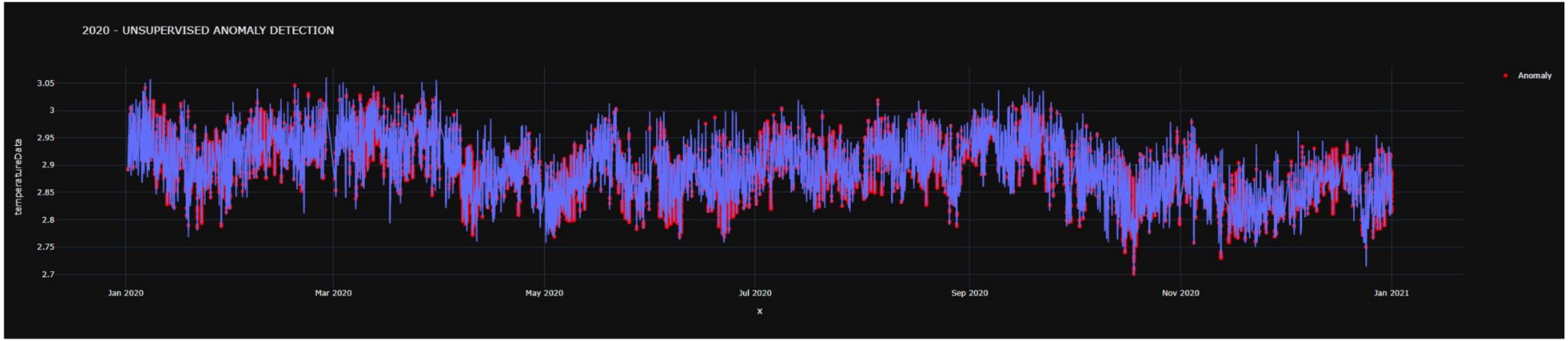


University of Victoria

Bonus approaches

- KNN - tried but failed
- Boxplot





Summary

	Pros	Cons
LSTM	- Good for time series data analysis - Fast	- Hard to define “abnormaly”. High rate of false positive
Prophet	- Designed for analyzing time series data, accurate and fast - Robust to handle missing data and outliers - Easy to implement	- Incorrect prediction if when using small historical dataset - Only linear model and logistic growth model available - Hard to prepare the environment. It requires Python >= 3.7; Linux or macOS system; x86-64 CPU and C++ compiler: gcc >= 9.0 or clang >= 10.0
LOF	- Unsupervised model - Good at detecting outliers based on a local neighbourhood	- Large Computational Overhead - Requires strong assumptions - Hard to evaluate results

University of Victoria

Conclusion

After implementing 3 approaches and comparing the results, we get these conclusions:

- Both LSTM and Prophet are better solutions.
- Prophet is the easiest one to implement.
- Prophet can handle collective outliers.

Future Work

- For the consideration of training time, we are using 1/60 of data, this could be improved when having enough time and more powerful machine for training.
- Have a more detailed check with ground truth, figure out the performance of our model and try to improve it.
- The definition for anomaly data is ambigious.
- LSTM could improve to use interactive thresh hold

University of Victoria

References

[1] “Introduction to clayoquot slope,” https://www. oceannetworks.ca/introduction-clayoquot-slope, 2013, Accessed: 2021-04-30.
[2] Dilumie Abeysirigunawardena, Marlene Jeffries, Michael G Morley, Alice OV Bui, and Maia Hoeberechts, “Data quality control and quality assurance practices for ocean networks canada observatories,” in OCEANS 2015-MTS/IEEE Washington. IEEE, 2015, pp. 1–8.
[3] S. Bacon, L. R. Centurioni, and W. J. Gould, “The evaluation of salinity measurements from palace floats,” Journal of Atmospheric and Oceanic Technology, vol. 18, no. 7, pp. 1258 – 1266, 2001.
[4] Eugine Kang, “Long short-term memory (lstm): Concept,” https://medium.com/@kangeugine/ long-short-term-memory-lstm-concept-cb3283934359, Accessed: 2021-04-30.
[5] Sagar Sharma, “Activation functions in neural networks,” https://towardsdatascience.com/ activation-functions-neural-networks-1cbd9f8d91d6, Accessed: 2021-04-30.
[6] Xuguang Liu, “A real-time detection method for abnormal data of internet of things sensors based on mobile edge computing,” https://www.hindawi.com/journals/ mpe/2021/6655346/, Accessed: 2021-04-30.
[7] Mohammad Esmalifalak, “Anomaly detection with knn,” https://www.youtube.com/watch?v=RwmttGrJs08, Accessed: 2021-04-30.
[8] Mahbubul Alam, “k-nearest neighbors (knn) for anomaly detection,” https://towardsdatascience.com/ k-nearest-neighbors-knn-for-anomaly-detection-fdf8ee160d13, Accessed: 2021-04-30.
[9] Mayank Tripathi, “Knowing all about outliers in machine learning,” https:// datascience.foundation/sciencewhitepaper/ knowing-all-about-outliers-in-machine-learning, Accessed: 2021-04-30.
[10] Naysan Saran, “Interquartile range (iqr) to detect outliers,” https://naysan.ca/2020/06/28/ interquartile-range-iqr-to-detect-outliers/, Accessed: 2021-04-30.
[11] “Box plot diagram to identify outliers,” https://www.whatissixsigma.net/ box-plot-diagram-to-identify-outliers/, Accessed: 2021-04-30.
[12] “1.5. box plots,” https://learnche.org/pid/ data-visualization/box-plots, Accessed: 2021-04-30.
[13] “Prophet” , “https://facebook.github.io/prophet/”
[14] “Local Outlier Factor”, “https://en.wikipedia.org/wiki/Local_outlier_factor”

Thank you for listening! 😁

Questions?

University of Victoria