pH ESTIMATION USING DNN

Di Wan[1], Pramod Thupaki[2], Gregor Reid[3]

[1]Institute of Ocean Sciences, Fisheries and Oceans Canada
[2]Hakai Institute, Victoria, Canada
[3]Department of Fisheries and Aquaculture, Shelburne, Canada

Fisheries and Oceans Canada
Pêches et Océans Canada

Hakai
Science on the Coastal Margin

NOVA SCOTIA
Fisheries and Aquaculture

Motivation

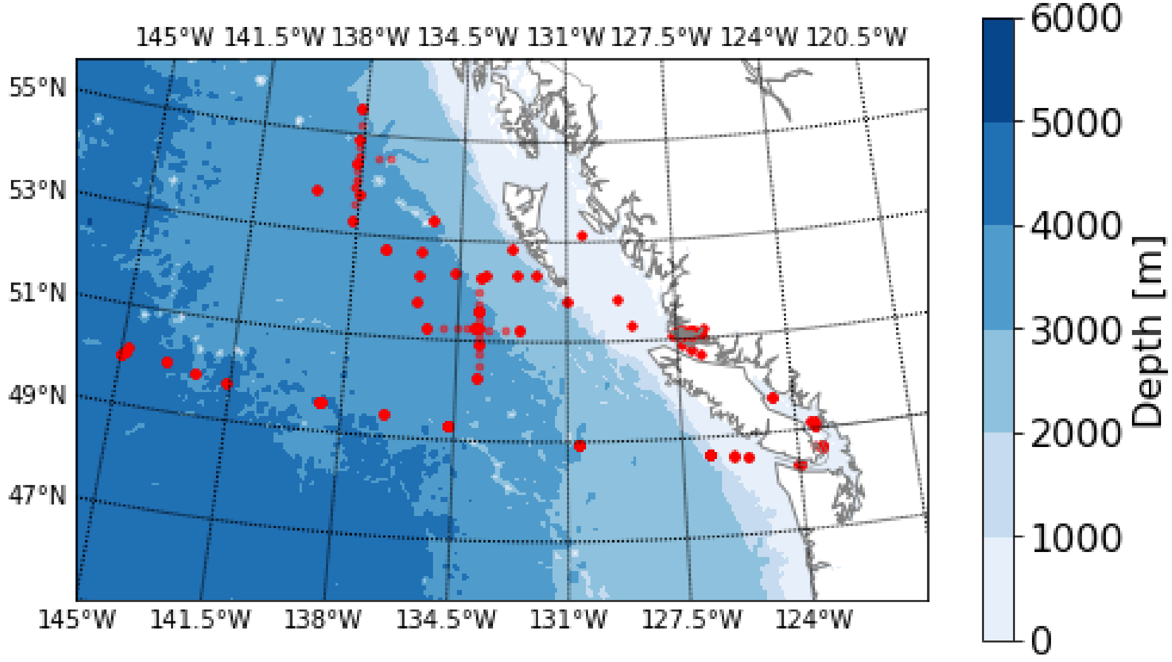
- Difficulty in measuring pH
- Ocean acidification vs coastal acidification
- High variability in pH in coastal regions: upwelling, coastal nutrient changes
- Current methods
    - Bottle: pH ~ (Temp, titration alkalinity, DIC)
    - Profile pH sensor (fast response, but requires more frequent calibration)
    - Moored pH sensors (allows less frequent calibration, but the sensors response time is longer)

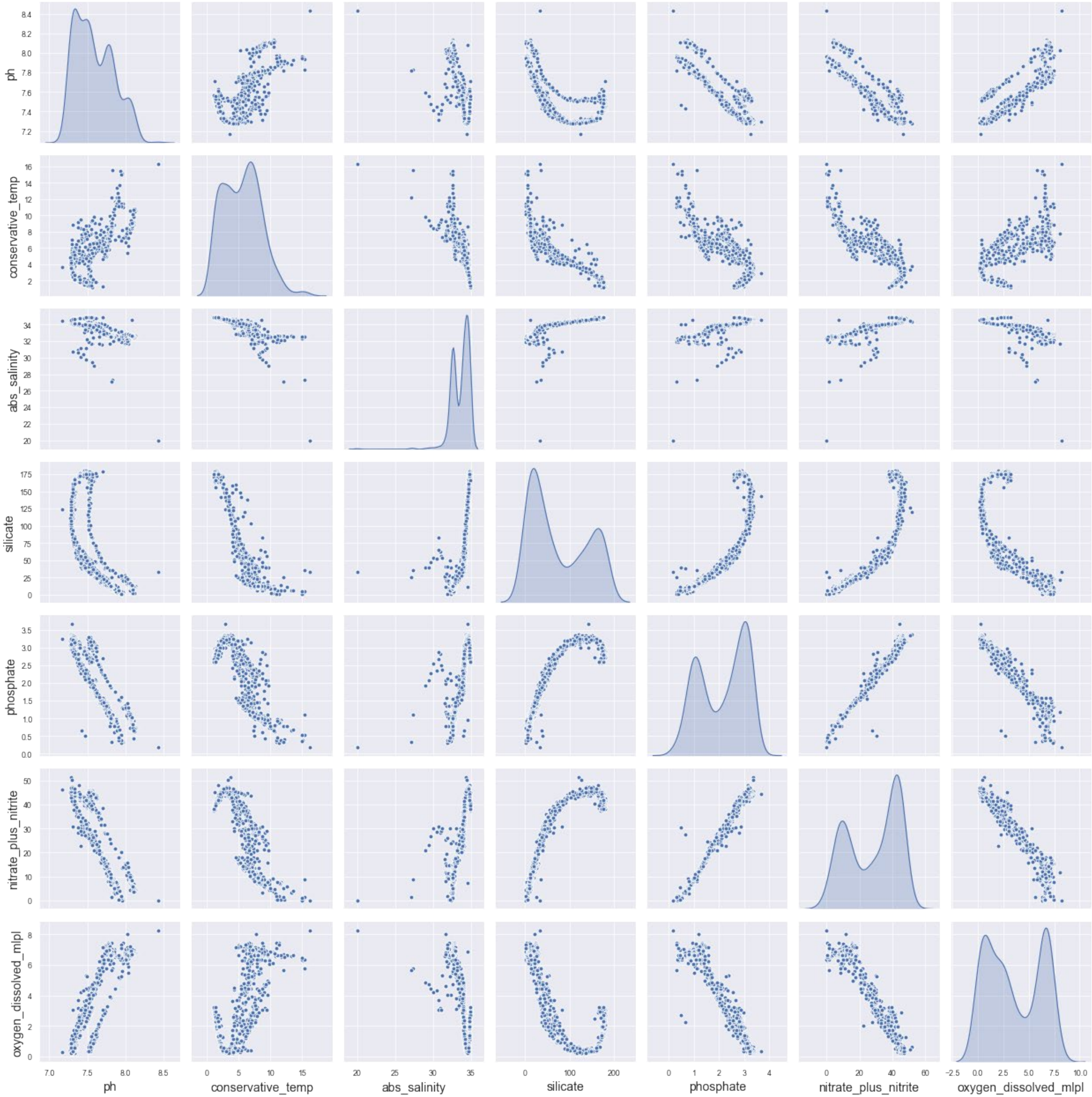
Objective

- Use easy to measure variables and not directly related to DIC or Alkalinity
- Accuracy ~0.01 – 0.001
- Achieve real-time or near real-time prediction (e.g. Argo SOCCOM biogeochemical floats)

Data Source

- 142 profiles from 2000 to 2018 that has pH values
- 2042 data points – 650 usable points
  - T, S, Phosphate, silicate, Nitrate+nitrite, DO, and pH
- Was not a trivial process to clean up the data

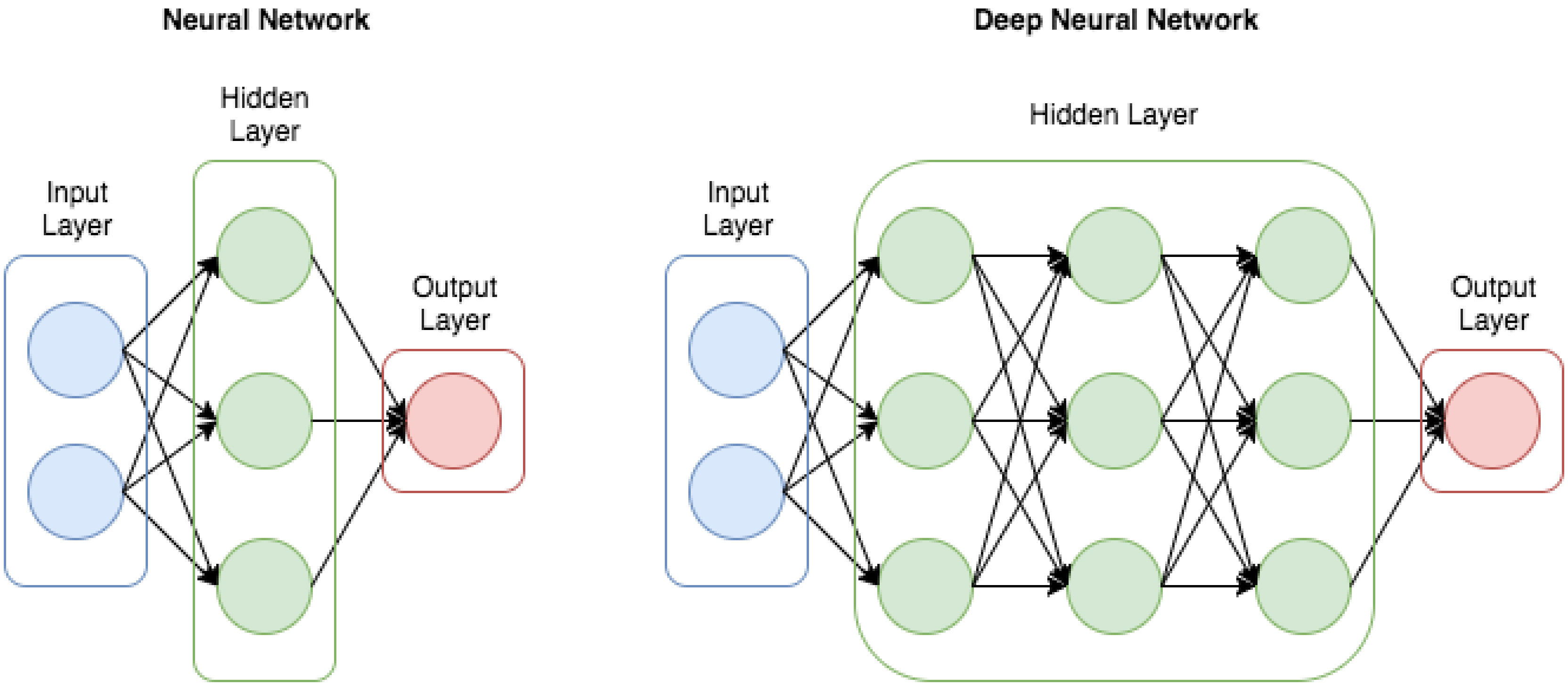



ph
8.4
8.2
8.0
7.8
7.6
7.4
7.2
conservative_temp
16
14
12
10
8
6
4
2
abs_salinity
34
32
30
28
26
24
22
20
silicate
175
150
125
100
75
50
25
0
phosphate
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
nitrate_plus_nitrite
50
40
30
20
10
0
oxygen_dissolved_mlpl
8
6
4
2
0
7.0
7.5
8.0
8.5
ph
0
5
10
15
conservative_temp
20
25
30
35
abs_salinity
0
100
200
silicate
0
1
2
3
4
phosphate
0
20
40
60
nitrate_plus_nitrite
-2.5
0.0
2.5
5.0
7.5
10.0
oxygen_dissolved_mlpl

METHODS

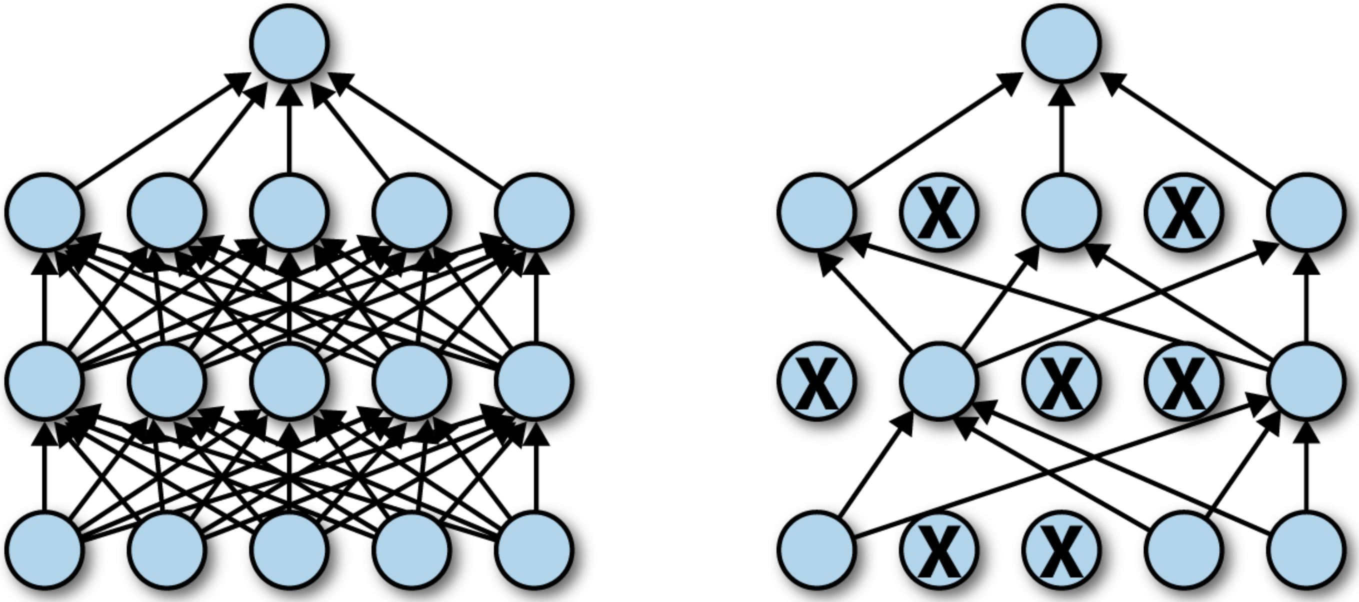
- Deep neural network (Validation split = 0.2)
- Stochastic Gradient Descent
- Apply dropout nodes to prevent overfitting
- Variables used: pH, T, S, phophate, nitrate+nitrite, silicate, DO
- Linear activation function is used – linear and fast; modification can be made

DEEP NEURAL NETWORK



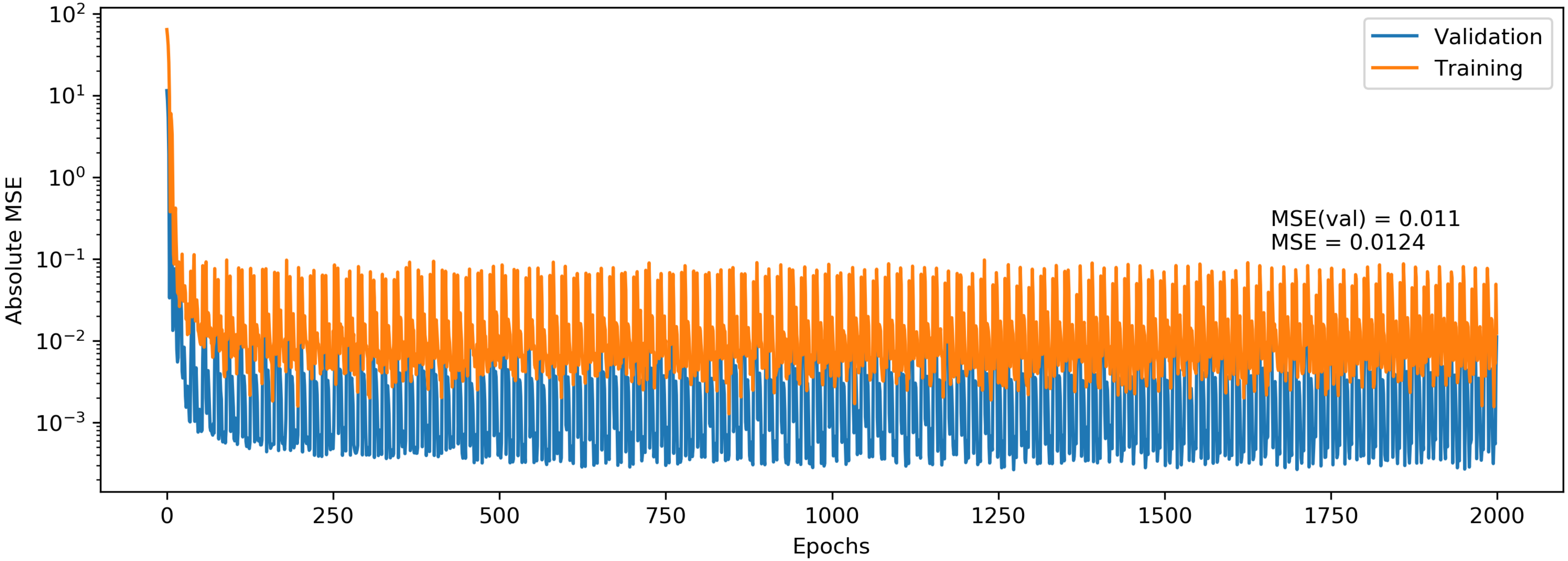
Dropout: prevent overfitting

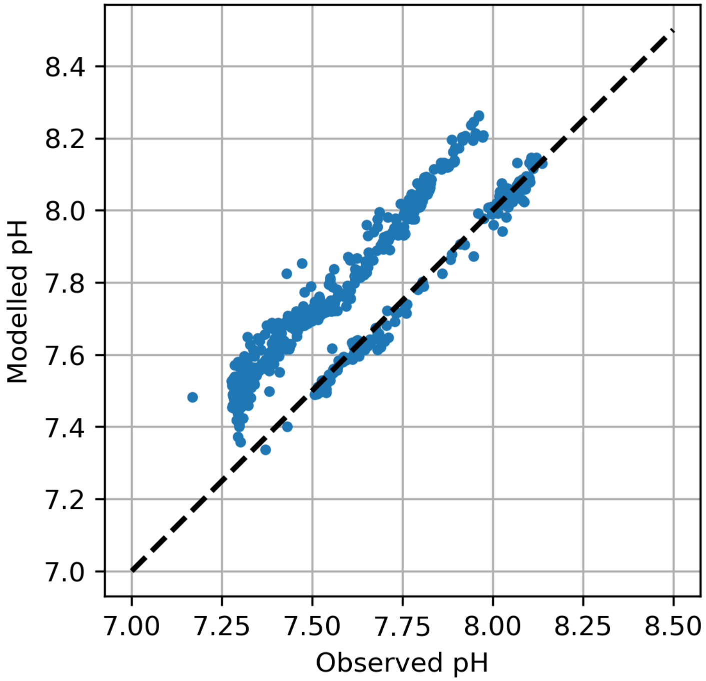
- 2 layers, 32 nodes per layer, 50% dropout



(a) Standard Neural Net

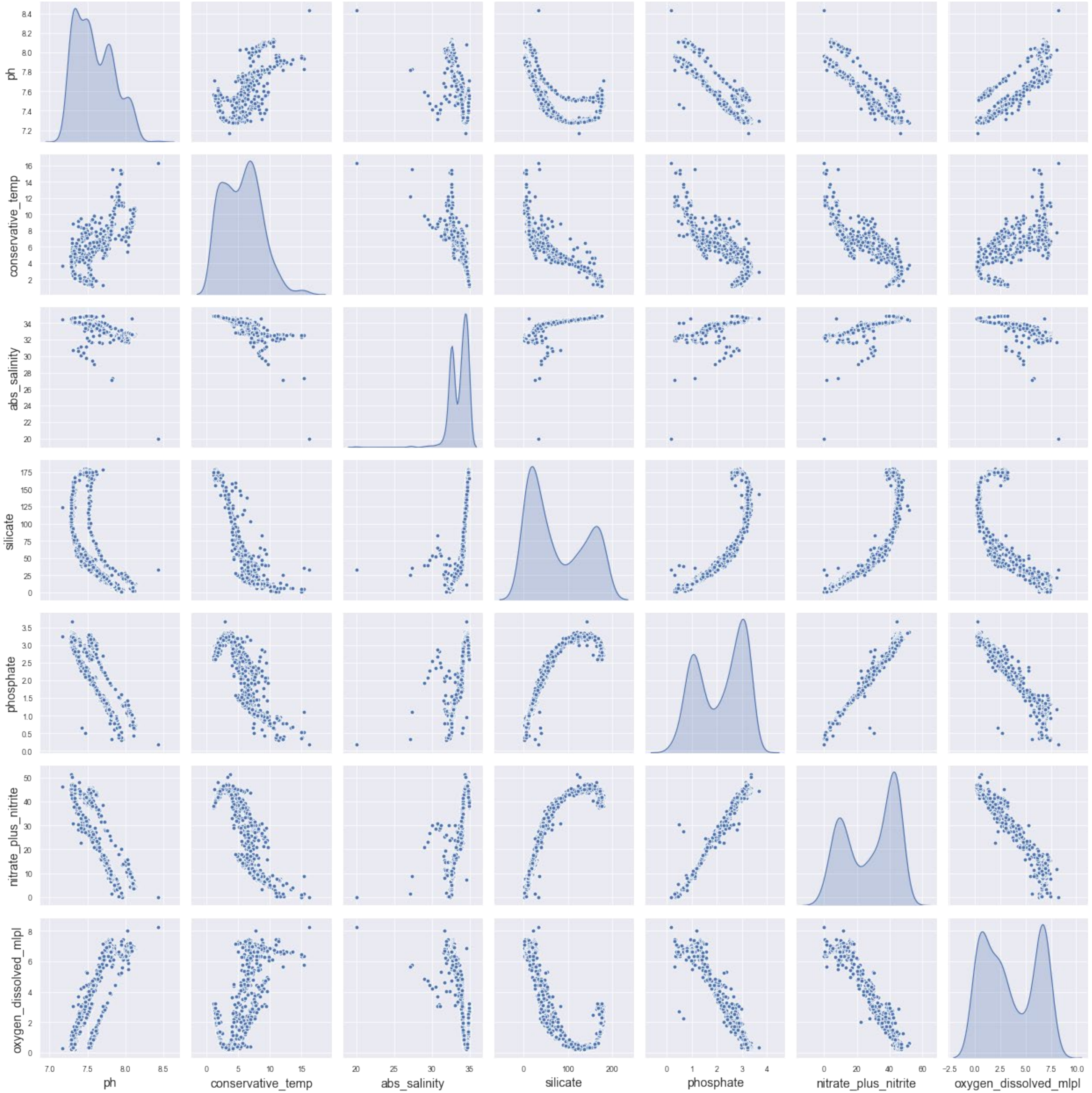
(b) After applying dropout



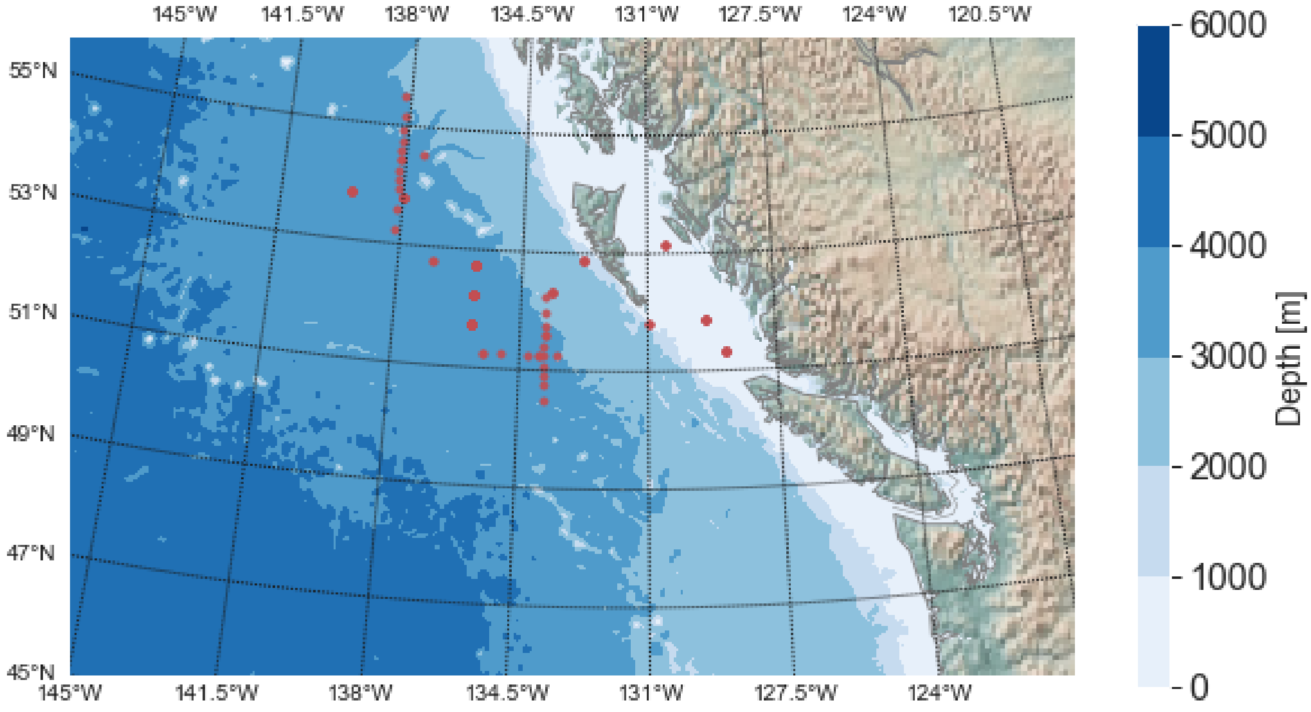


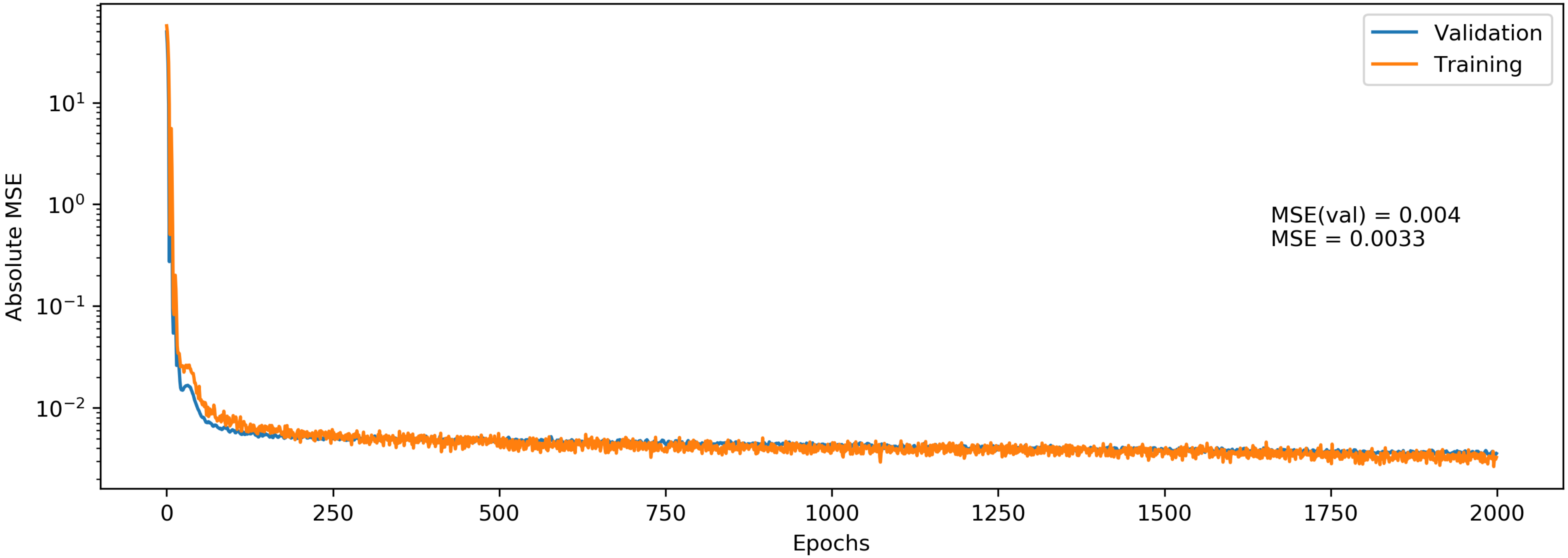
- MSE ~ 0.01 for training;
- MSE ~ 0.01 for validation
- Why is validation error more than training error??
- Not enough data?
- Clearly 2 patterns

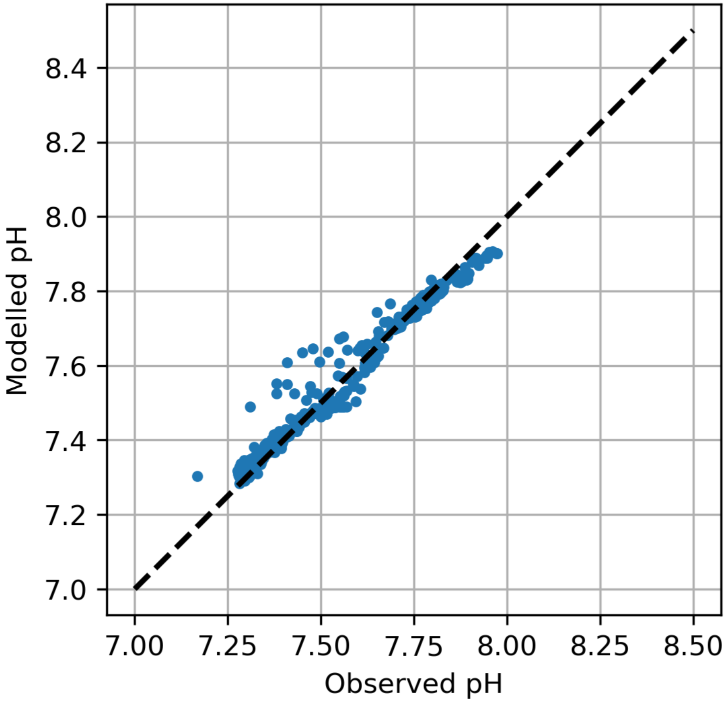
2 patterns


ph
conservative_temp
abs_salinity
silicate
phosphate
nitrate_plus_nitrite
oxygen_dissolved_mlpl
8.4
8.2
8.0
7.8
7.6
7.4
7.2
16
14
12
10
8
6
4
2
34
32
30
28
26
24
22
20
175
150
125
100
75
50
25
0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
50
40
30
20
10
0
8
6
4
2
0
7.0
7.5
8.0
8.5
0
5
10
15
20
25
30
35
0
100
200
0
1
2
3
4
0
20
40
60
-2.5
0.0
2.5
5.0
7.5
10.0
ph
conservative_temp
abs_salinity
silicate
phosphate
nitrate_plus_nitrite
oxygen_dissolved_mlpl

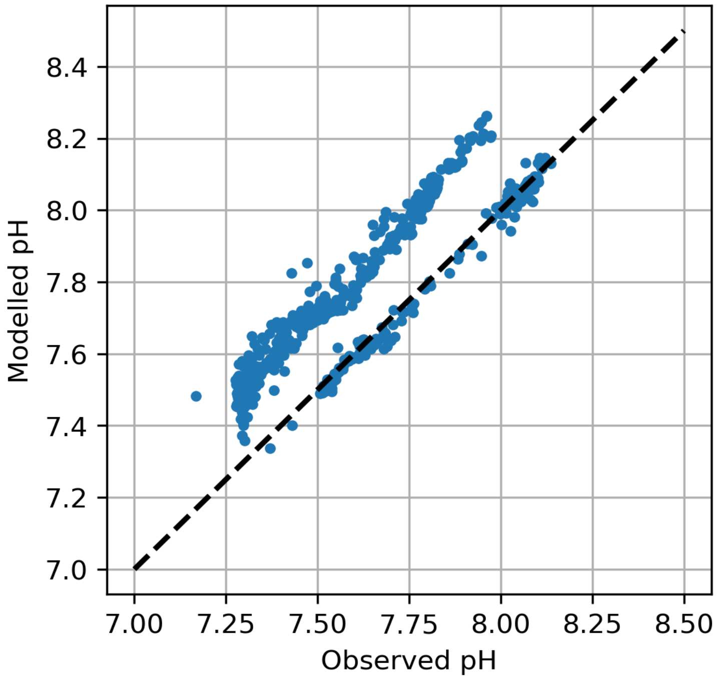
- Manual investigation
- 2001-2003
- 4 cruises, 124 data points

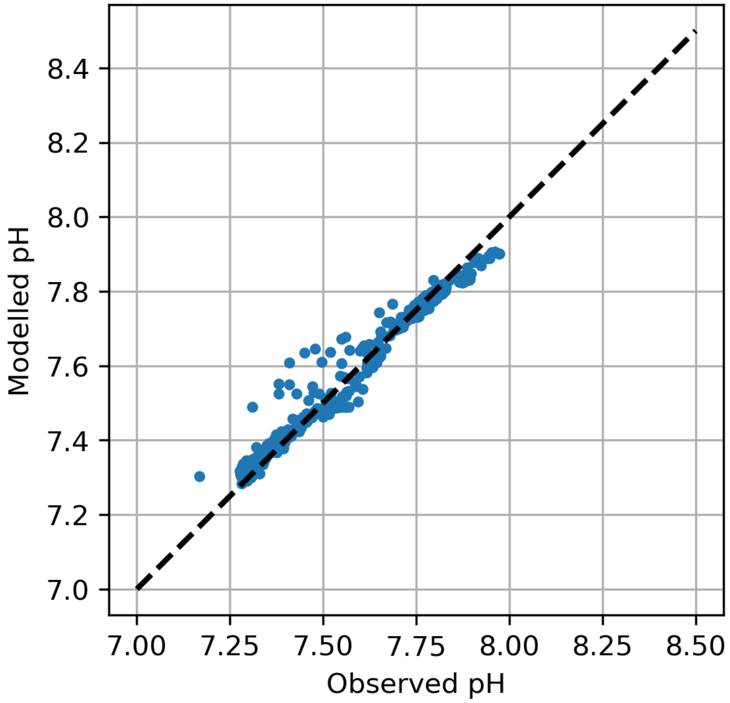






- MSE ~ 0.0033 for training;
- MSE ~ 0.004 for validation
- R2 = 0.98





- MSE ~ 0.0033 for training;
- MSE ~ 0.004 for validation
- R2 = 0.98
- How should we explain the 2 groups? Is it related to bad observations or cool findings?

Summary

- Why is validation error more than training error?
    - Insufficient data?
- More data is needed to aim for better accuracy
- How should we explain the 2 groups? Is it related to bad observations or cool findings?
- Can we use ML to spot cool science and/or QA/QC?

Acknowledgment

- DFO IOS Data group
- Lisa Miller
- Charles Hannah