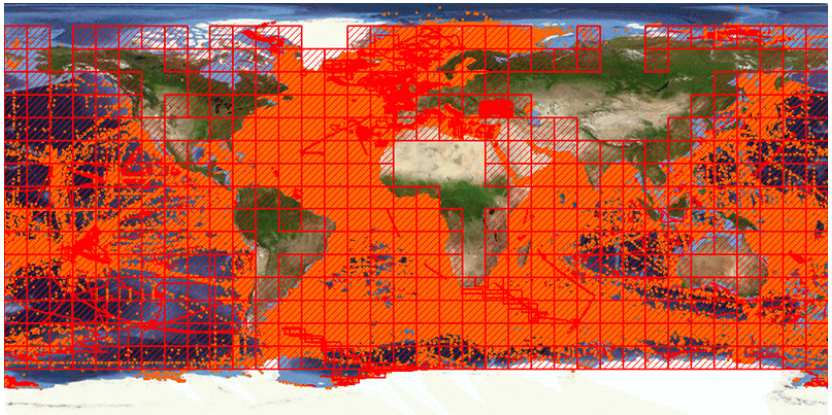
ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG

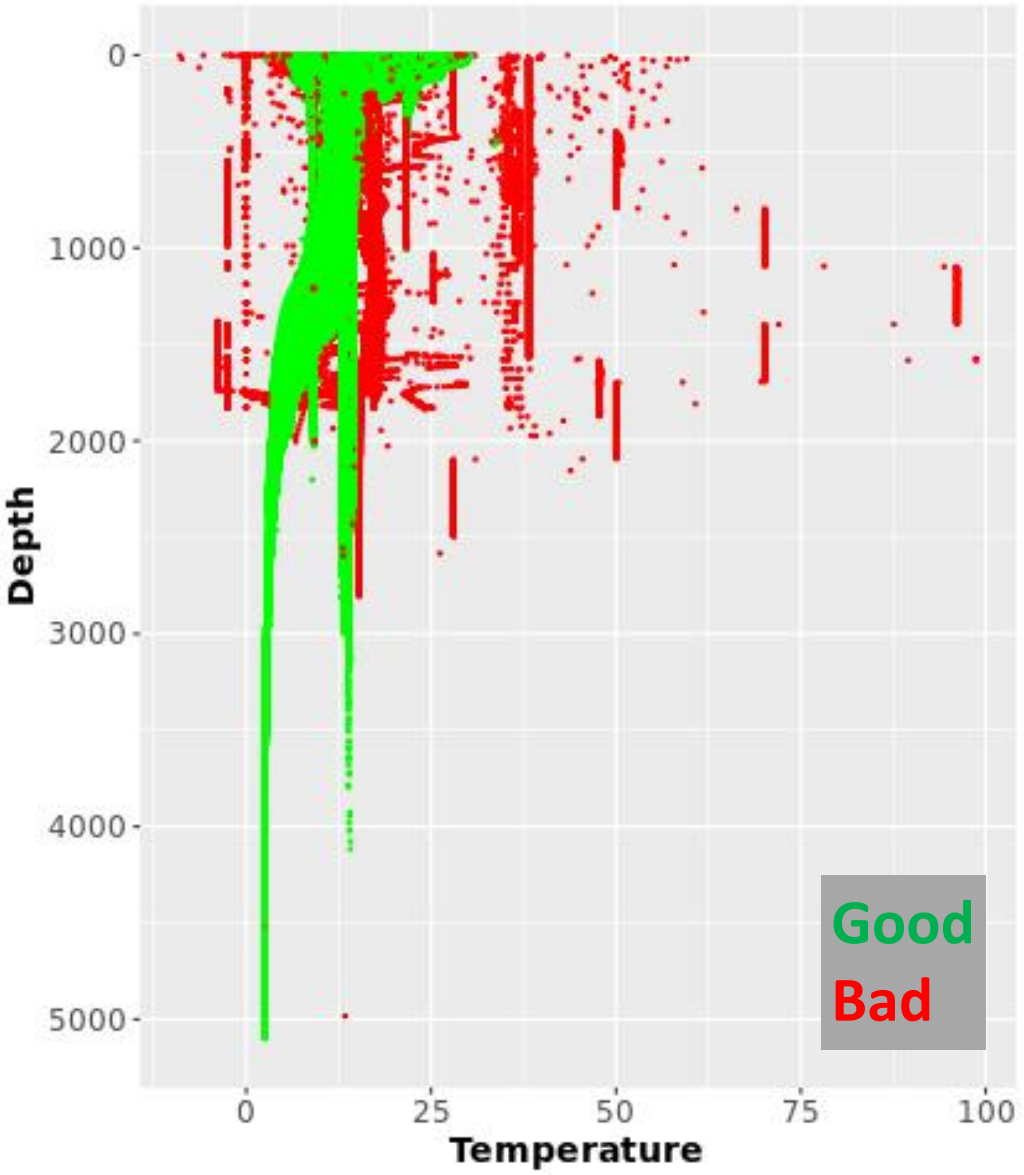
Automation of Quality Control for Global Ocean Data

Serdar Demirel, Sebastian Mieruch,
Mentor: Steffen Seitz

Alfred-Wegener-Institut,
Bremerhaven, Germany

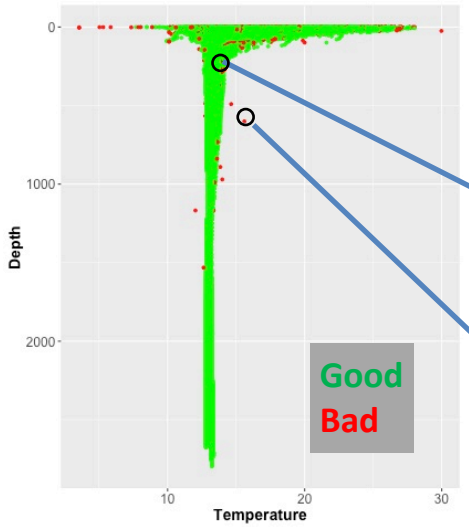
Reminder




0
1000
2000
3000
4000
5000
Depth
0
25
50
75
100
Temperature
Good
Bad

- **Which network architecture?**
  - **MLP**
- **How to deal with imbalanced dataset?**
  - **Create ROC curve and tune classification threshold**
- **How to optimize network skill?**
  - **Use large dataset**
  - **Overfitting, epoch, loss etc.**
  - **ResNet (to be done)**

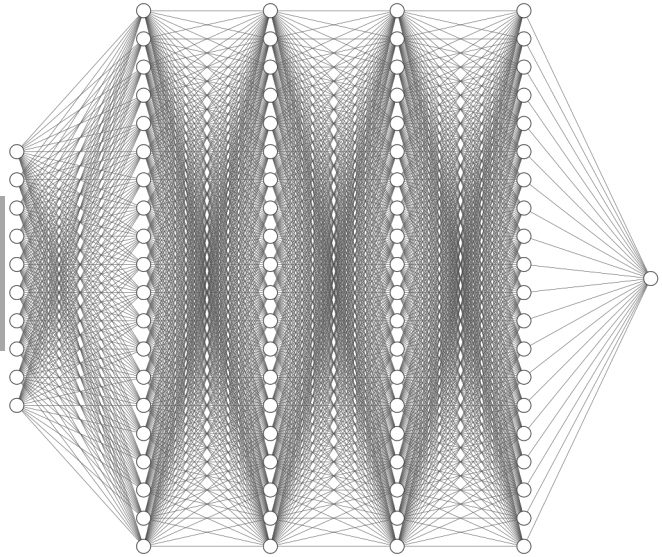
Model Architecture and pre-processing



One sample
14 Features

One sample
14 Features

Features



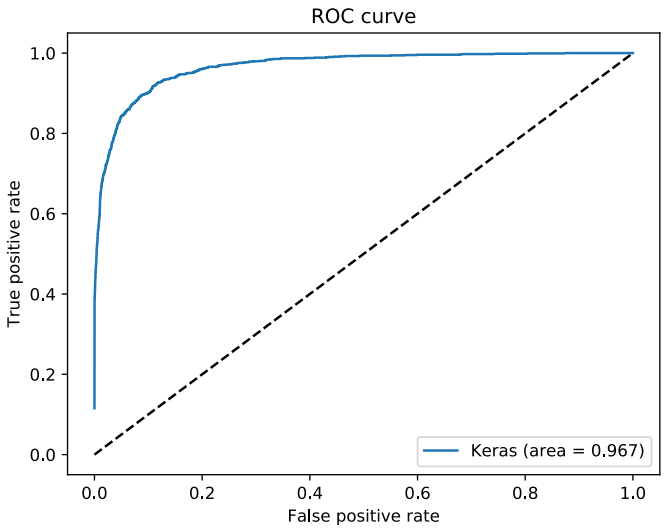
Flag;
Good or
Bad

Every sample has consistent feature counts.
Missing values are chosen as a unphysical value > -10

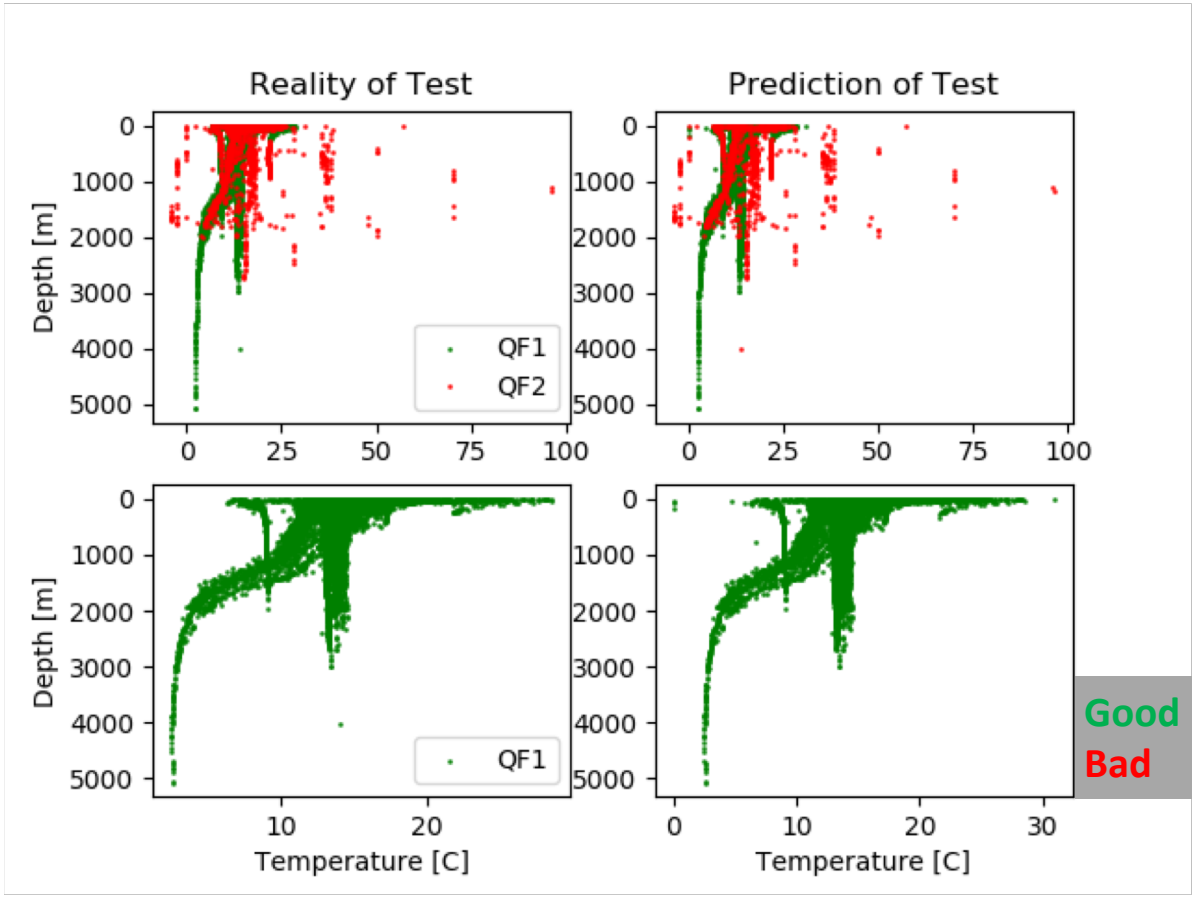
Features:

- Depth,
- Temperature,
- Latitude,
- Longitude,
- and some others

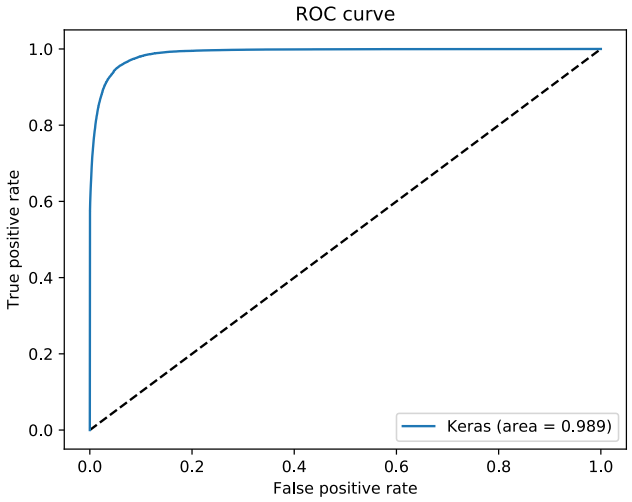
Dataset with 50 thousand samples (Mediterranean and black sea)



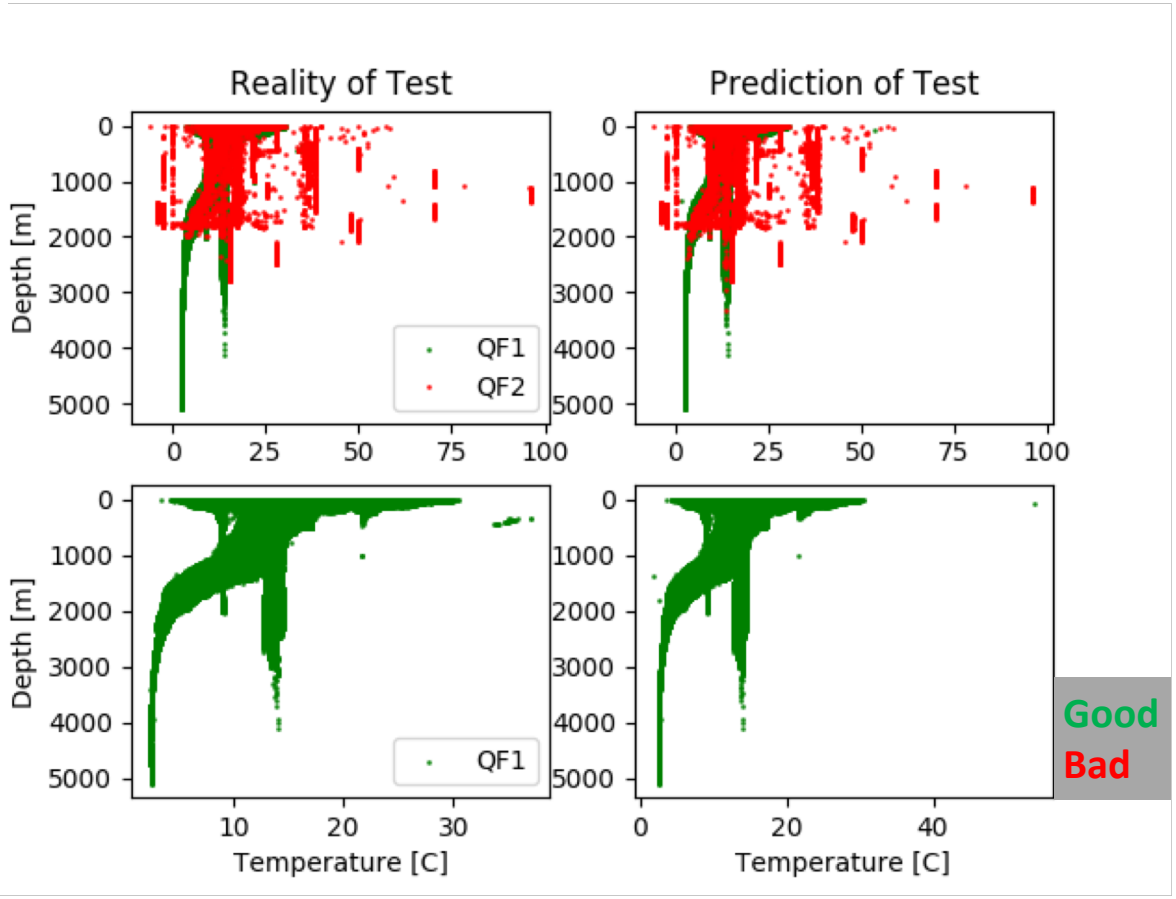
AUC area = 0.967



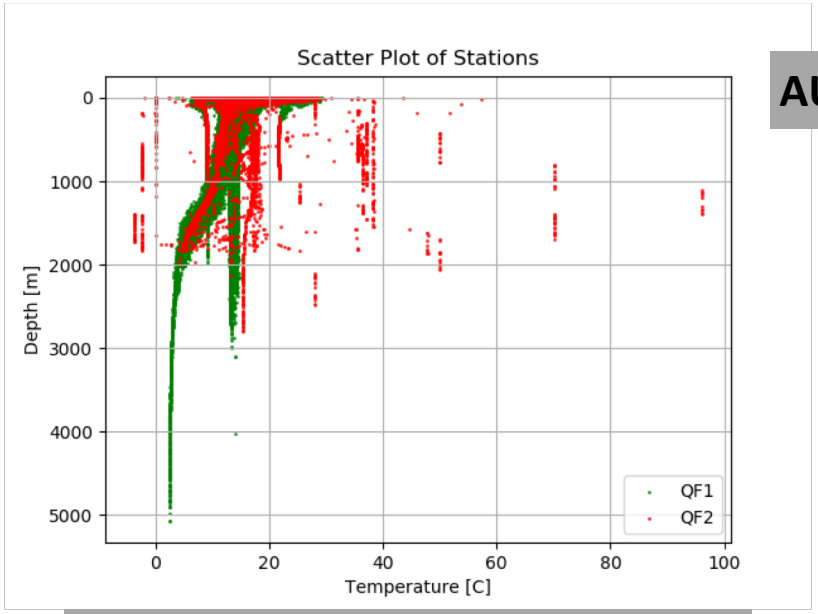
Dataset with 5 million samples (Mediterranean and Black Sea)



AUC area = 0.989

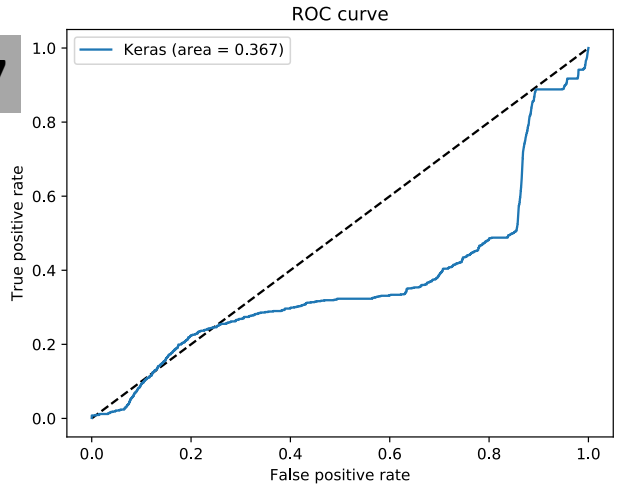


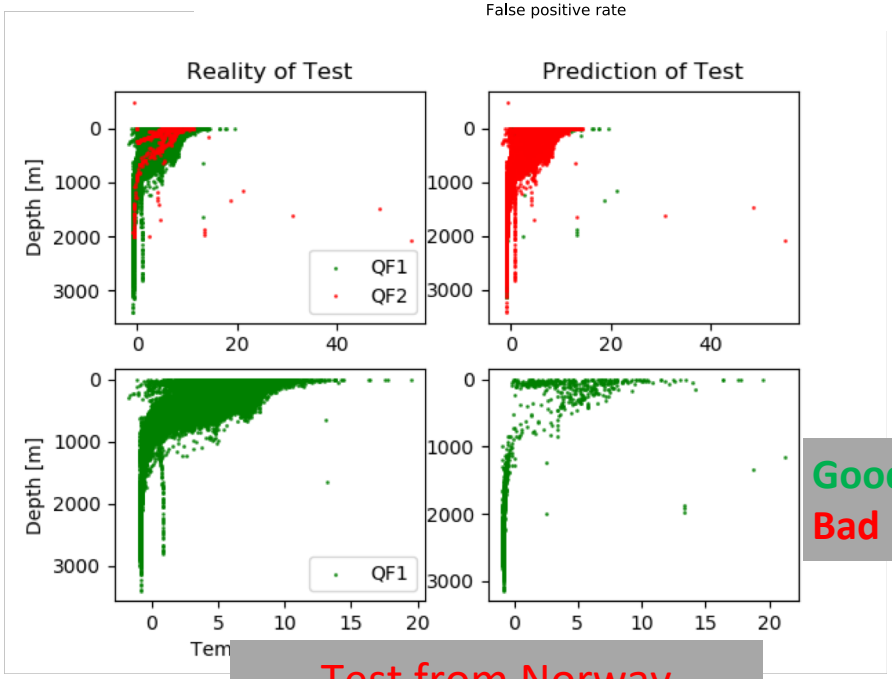
Testing neural network with unknown data from different region (50k samples)



Training dataset comes from Mediterranean and Black Sea

AUC area = 0.367



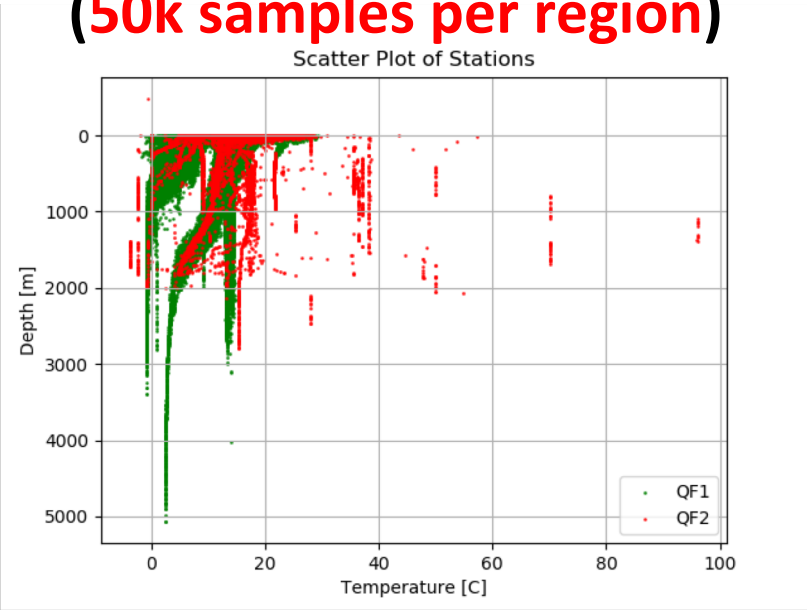


Good
Bad

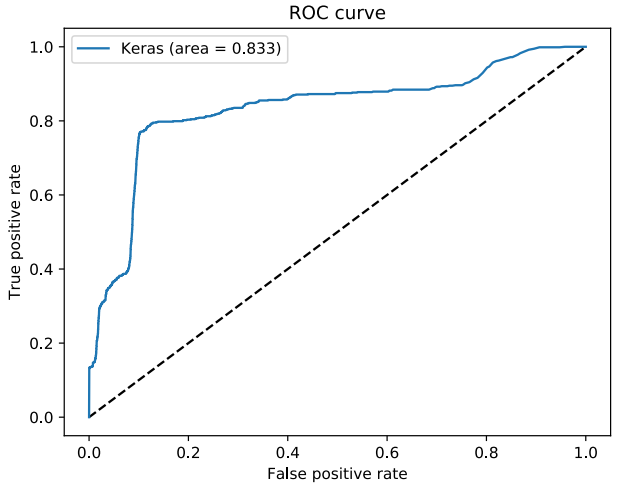
Note that: Norway is a totally different region compare to Mediterranean and Black Sea.

Test from Norway

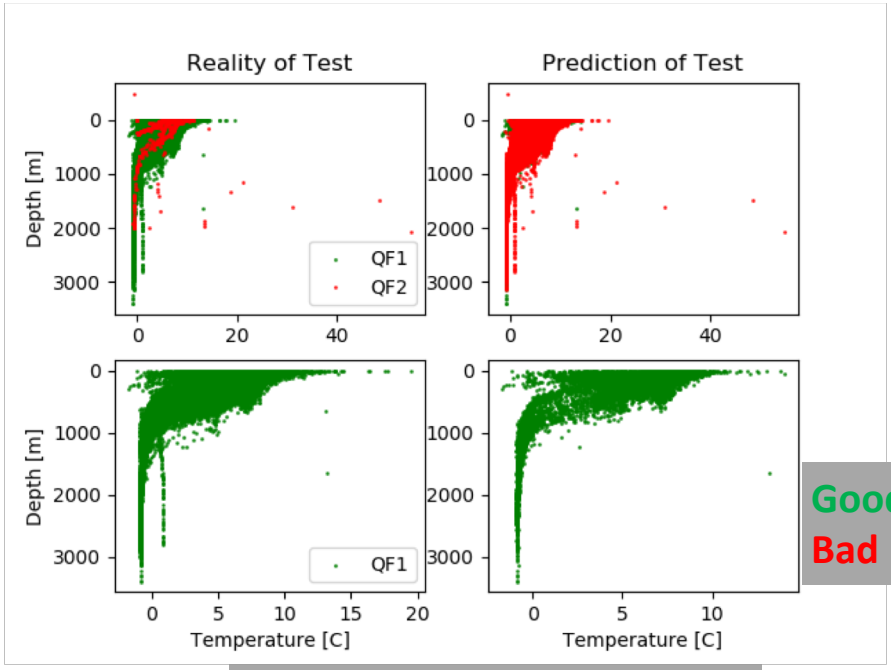
Testing neural network with unknown data from the same region (50k samples per region)



AUC area = 0.833



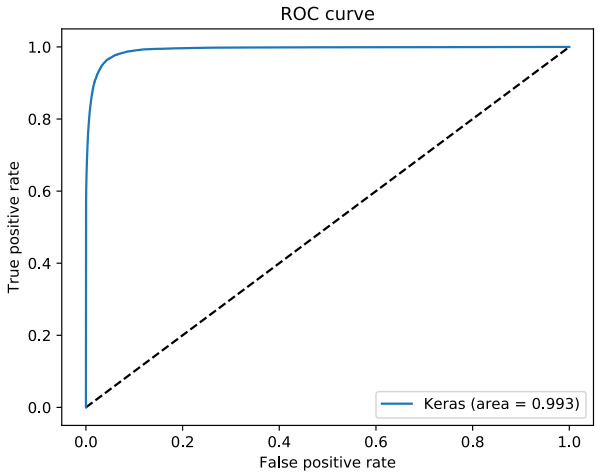
Training dataset comes from Mediterranean, Black Sea and Norway



Good
Bad

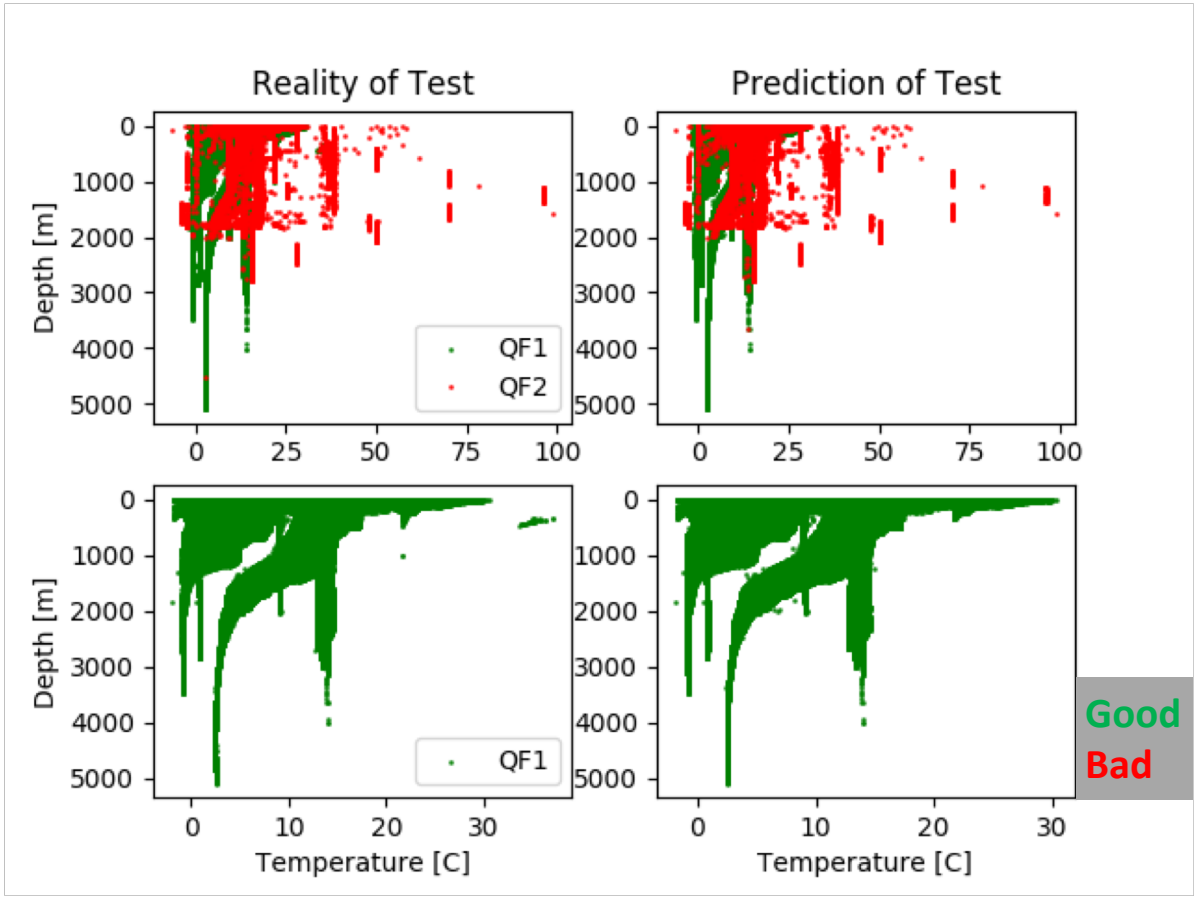
Test from Norway

Dataset with 11 million samples (Mediterranean, Black Sea, and Norway)



AUC area = 0.993

Large data was necessary to get the job done as good as for one region only.



Training on the World dataset, seems like a harder task. So that advanced architectures may be needed in the future implementation. -> ResNET?

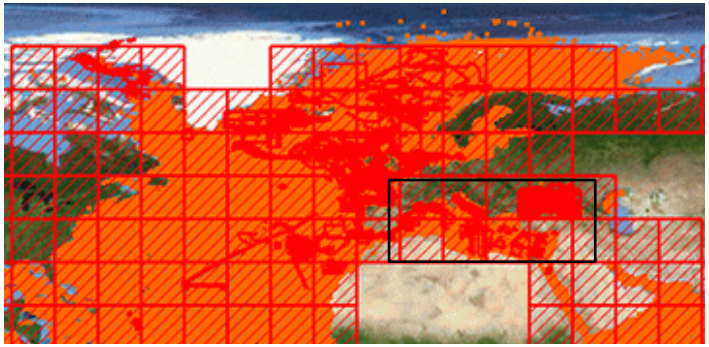
- **Next steps**
  - **Writing a paper about our study on the Mediterranean dataset**

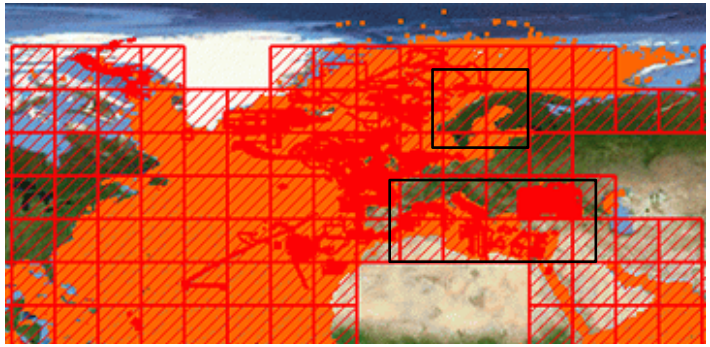
- **Future**
  - **More optimization**
  - **Apply on global ocean dataset**
  - **Develop an operational system for automated QC**
  - **Include in the SeaDataNet QC workflow**
  - **Publish as an open online service**

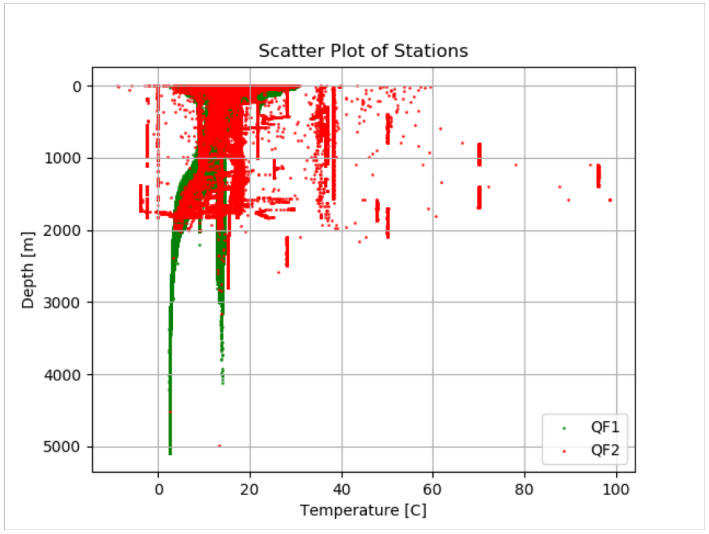
- **Thank you for your attention**

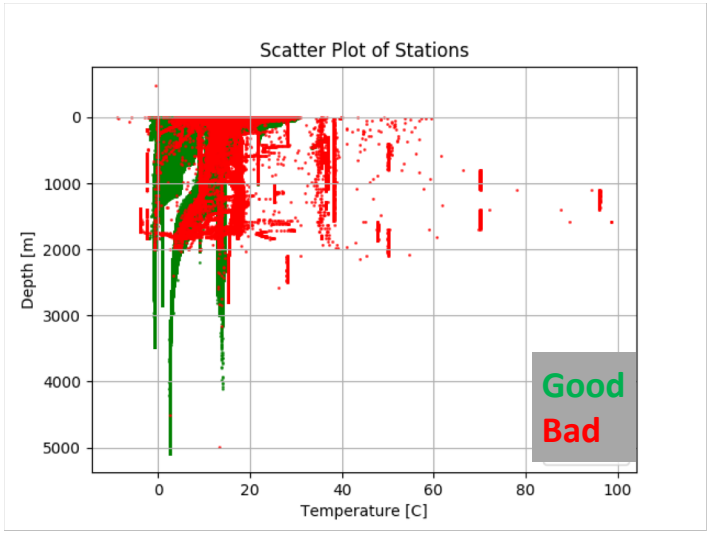
Main issues that are solved at Hackathon?

- Our biggest problem to find a strategy to handle with imbalanced dataset (~5 million sample labelled as good, ~50 thousand)
  - Increasing the accuracy of minority classification
- Increasing the dataset rapidly from 5 million to 11 million
  - Including several regions to the dataset

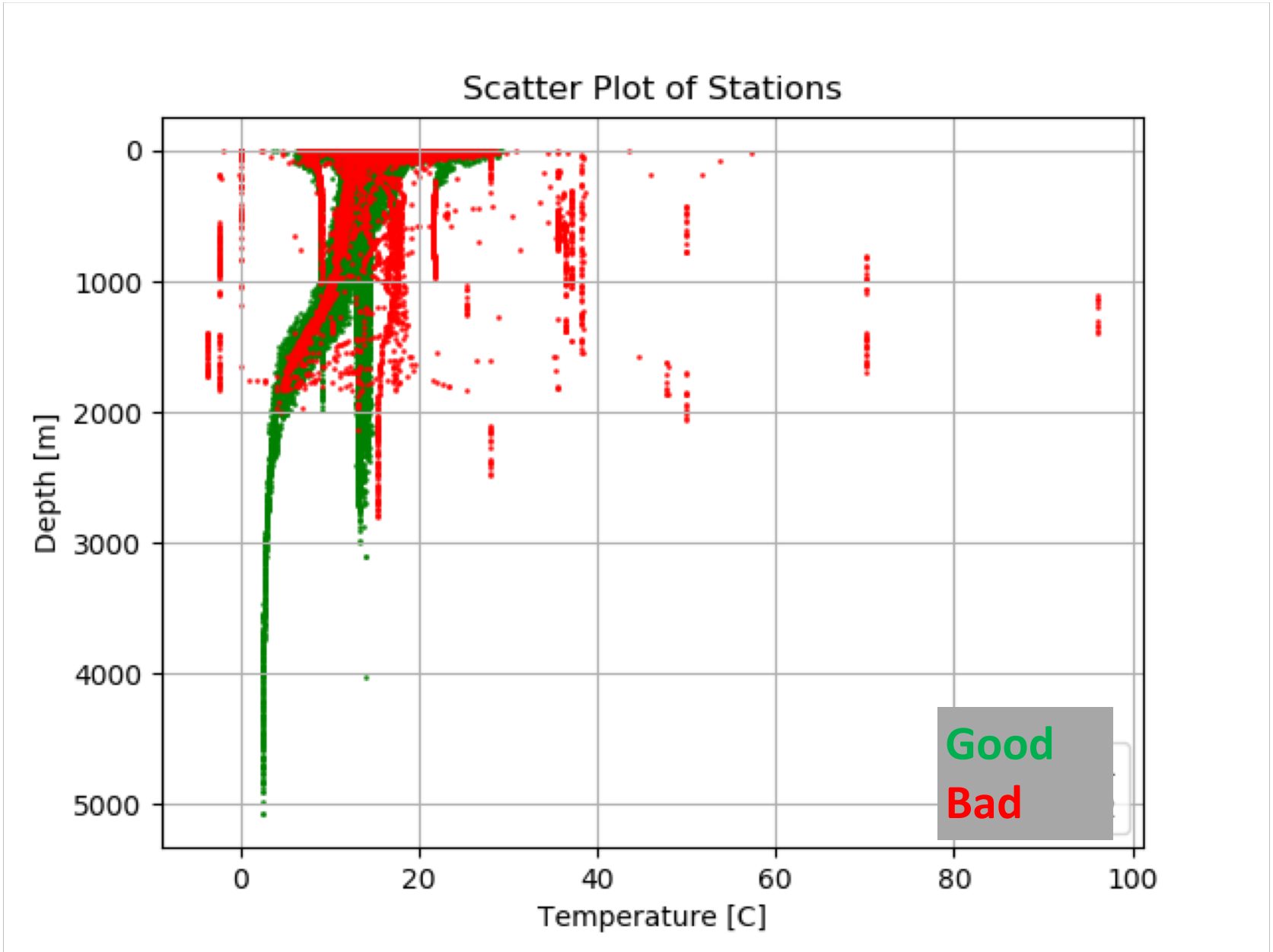




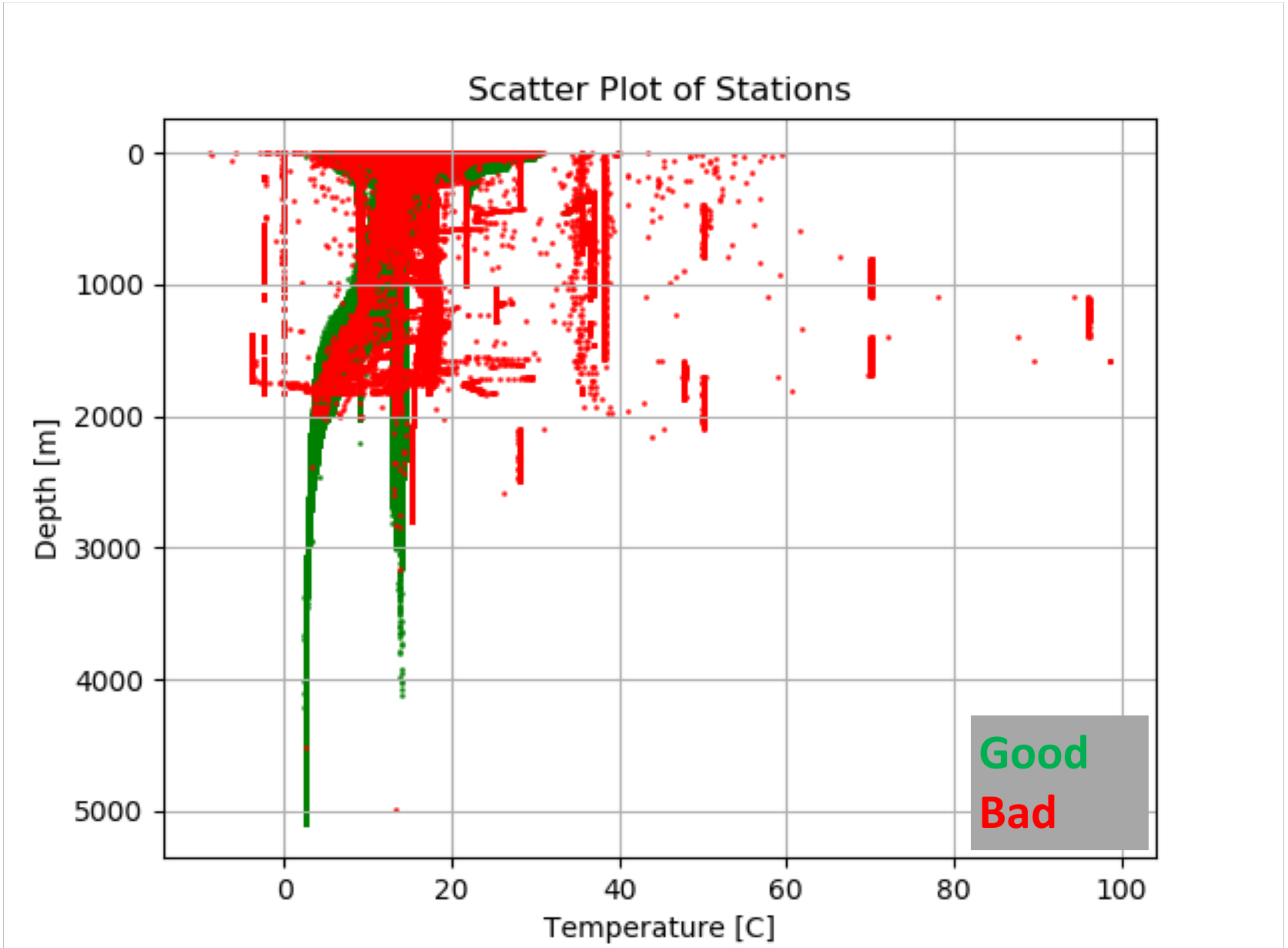




- **Dataset with 50 thousand samples (Mediterranean and black sea)**



- **Dataset with 5 million samples (Mediterranean and black sea)**



- **Dataset with 11 million samples (Mediterranean, Black Sea, and Norway)**