Artificial Intelligence-Enabled QC for Water Level Observations

IOOS DMAC Annual Meeting 2023

Current Contributors:

Jimmy Spore, Lindsay Abrams, Greg Dusek, Hassan Moustahfid (NOAA CO-OPS/IOOS)

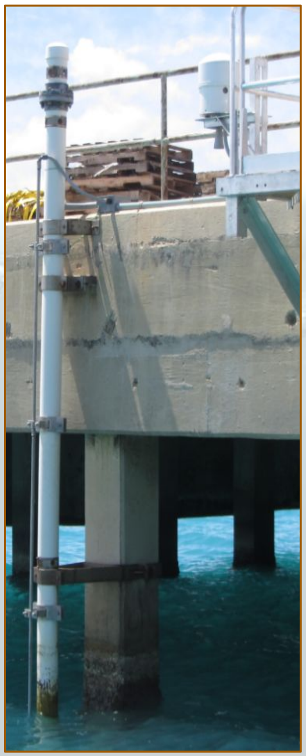
Evan Krell, Philippe Tissot, Felimon Gayanilo (Texas A&M Corpus Christi)

Overall Goal

Develop an automated AI-enabled quality control application to identify and correct bad water level data across our network based on CO-OPS historical data. Also aim to build this out as a community tool.

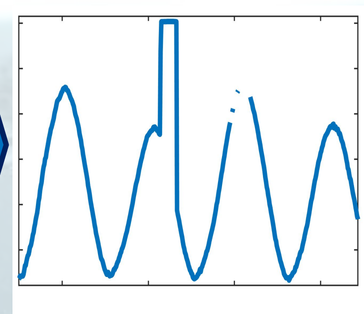
Current Water Level QC Workflow

Primary + Backup water levels



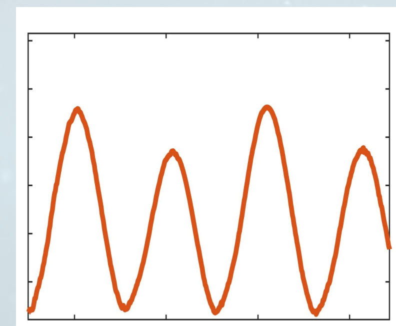
Ingestion QC

Preliminary Data



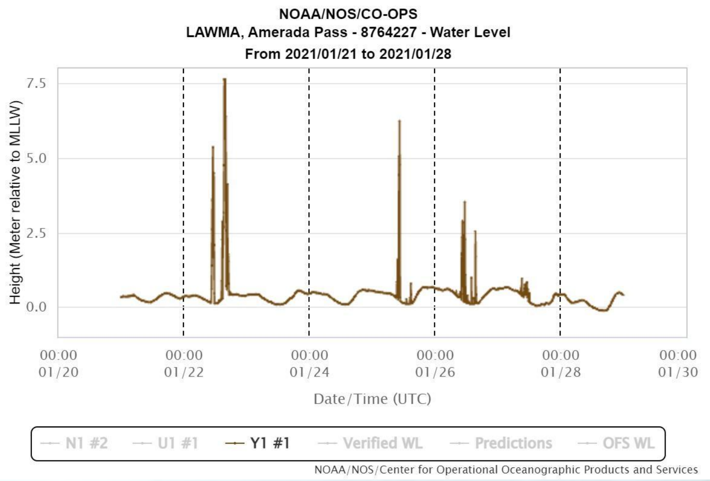
Manual processing and verification

Verified data

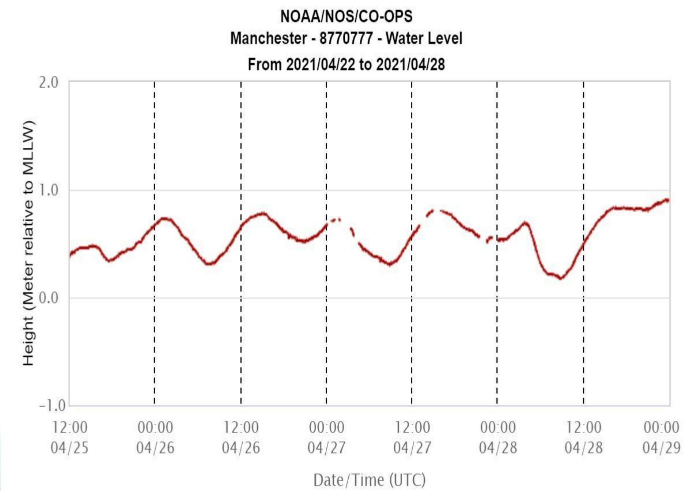


Common Data Quality Issues

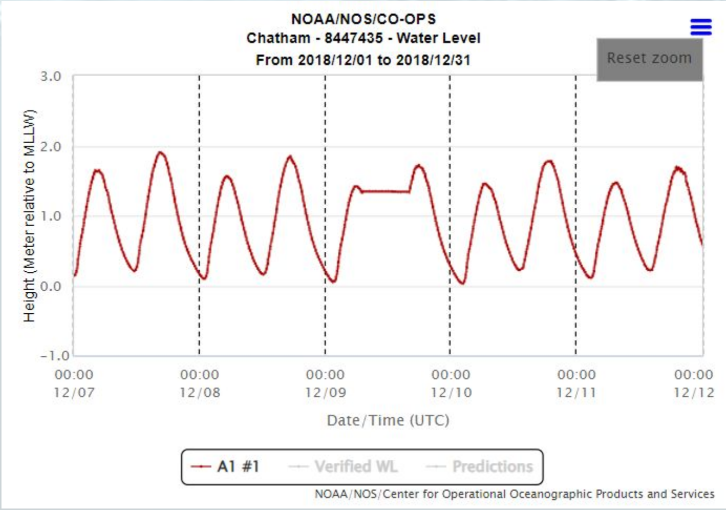
Spikes



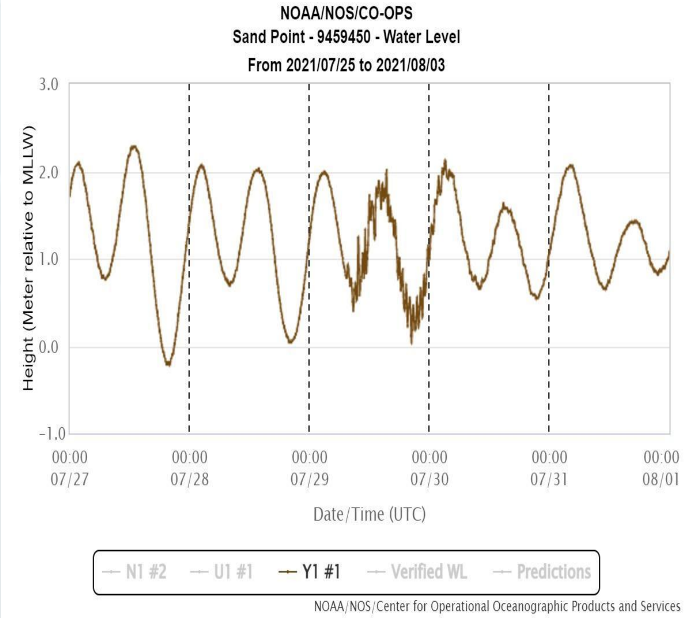
Missing



Flats



Rare Events (this is good data)



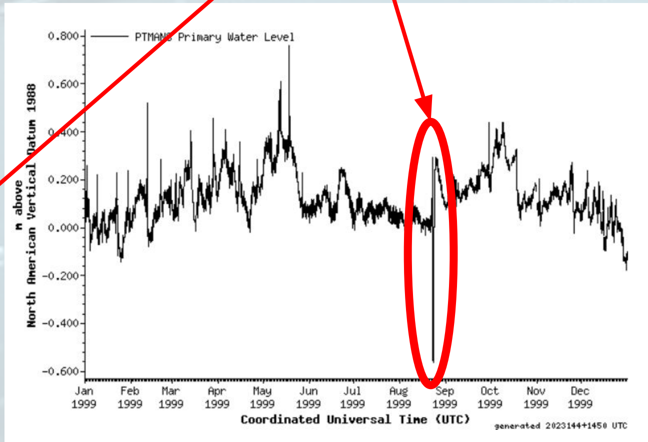
Idealized AI-Enabled QC Workflow

(1) Data preprocessing

(2) Classify good and bad water level data using AI model (trained with verified data)

(3) Correct bad data automatically using secondary algorithm (future development)

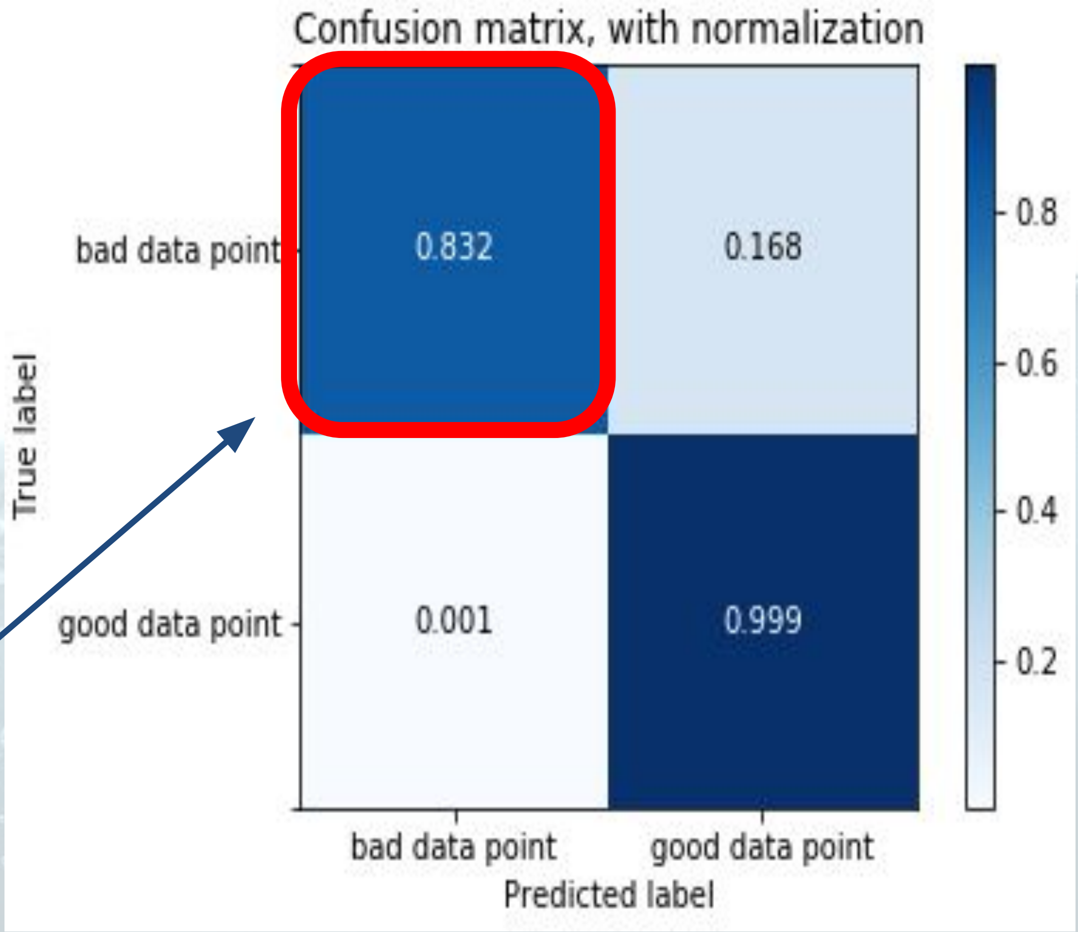
Good or bad data point?



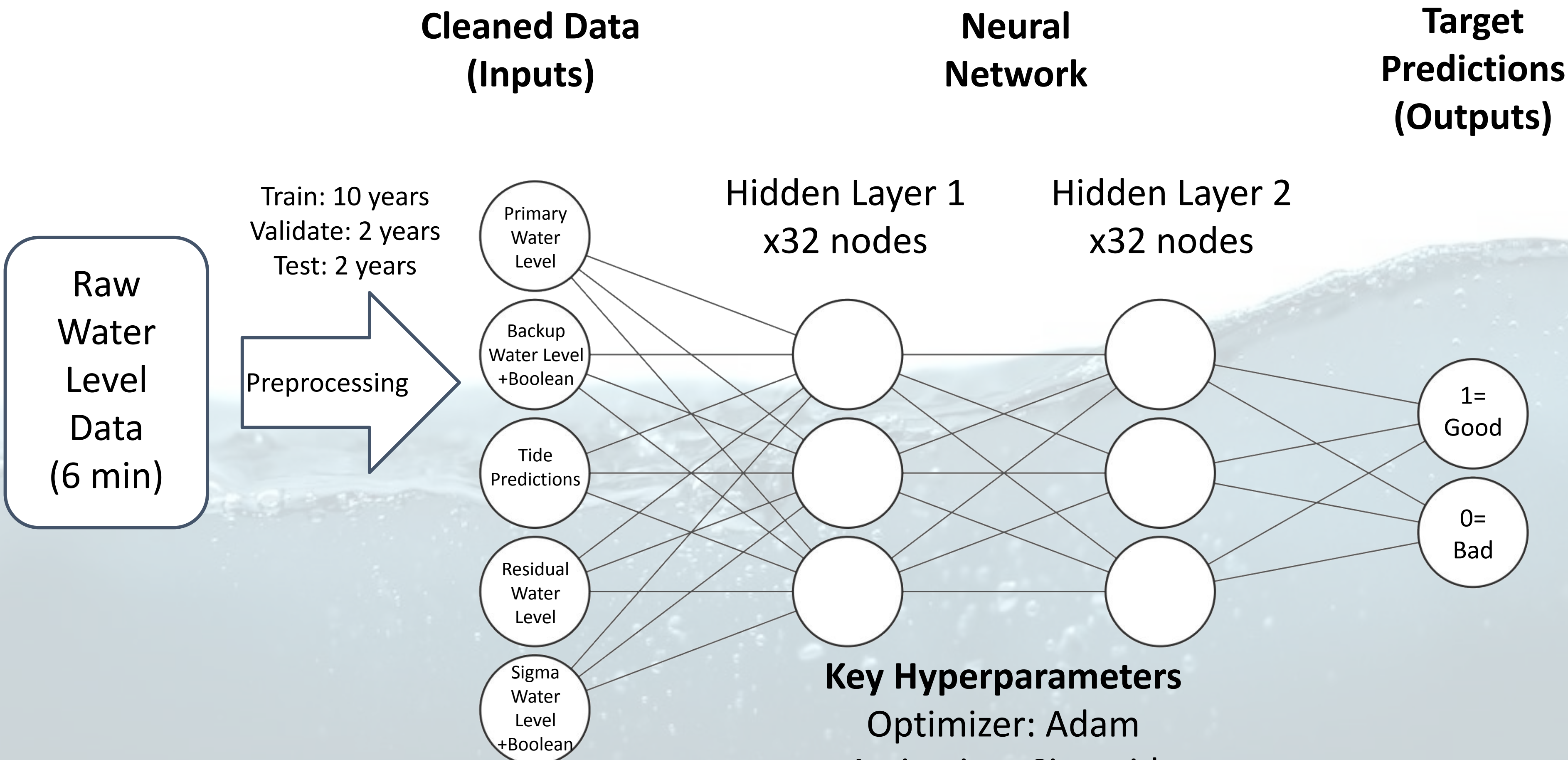
STATION ID	DATE TIME	SENSOR USED PRIMARY	PRIMARY	PRIMARY SIGMA	PRIMARY SIGMA TRUE	BACKUP	BACKUP TRUE	PREDICTION	TARGET
8557380	2017-01-01 00:00:00	A1	-0.324	0	1	-0.31088	1	-0.188	1

Classification Model Testing

Model	Accuracy
Logistic Regression	98.9
Random Forrest	99.7
Gradient Boost	99.7
Neural Net	99.7



Neural Network Approach



Key Hyperparameters

Optimizer: Adam

Activation: Sigmoid

Loss Function: Binary Cross-entropy

Epochs: 30

Batch Size: 256

Methods

Language

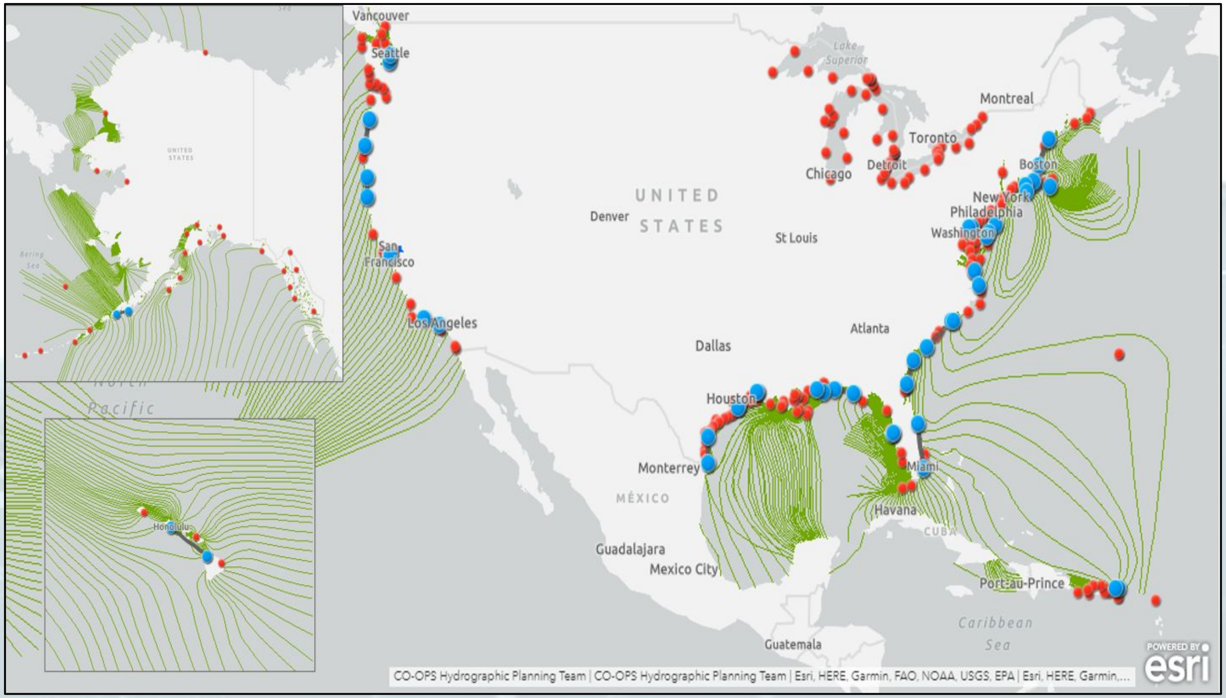
python

Libraries

TensorFlow

Keras

scikits learn



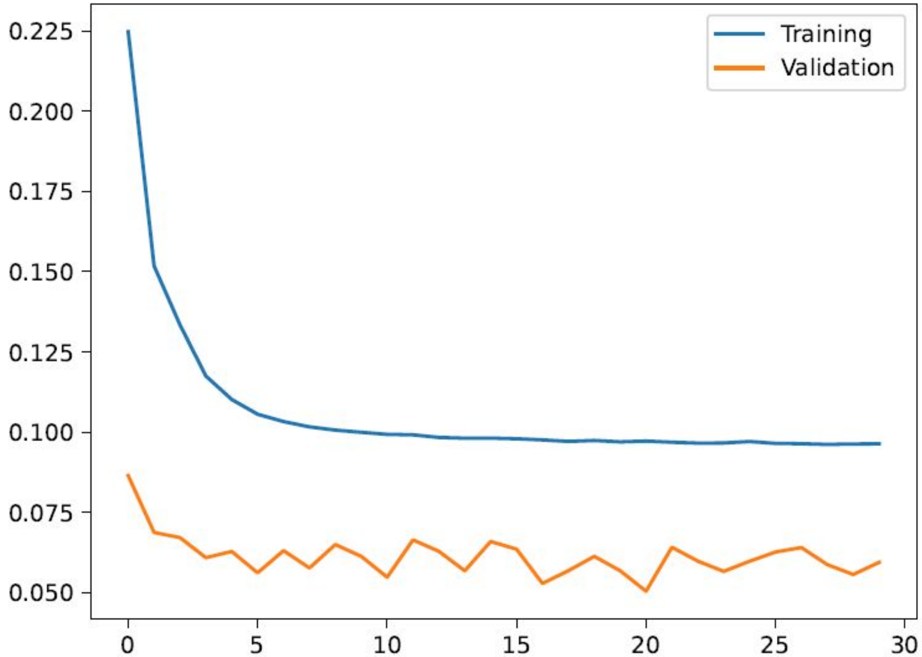
Map of NOAA water level stations (red markers) and the stations that were included in the training dataset for the model (blue markers). Green lines are cotidal lines.

Regional Model Training Results

Model	True Positives (Bad Data)	False Negatives (Bad Data)
HI	0.984	0.016
Northeast Coast	0.931	0.069
Southwest Coast (CA)	0.915	0.085
East Coast	0.9	0.1
Southeast Coast	0.889	0.111
AK	0.874	0.126
Atlantic	0.873	0.127
West Coast	0.872	0.128
Pacific (WC, AK & HI)	0.871	0.129
All Stations (50 total)	0.85	0.15
Gulf of Mexico	0.826	0.174
USVI	0.795	0.205
Northwest Coast (OR & WA)	0.672	0.328

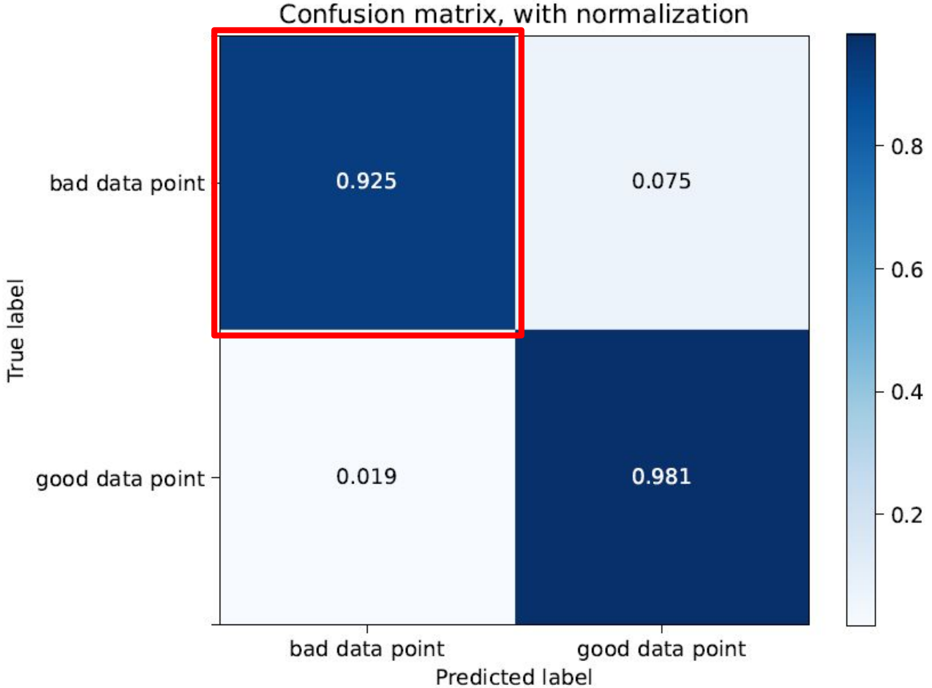
Good Performance: Northeast regional model

Loss



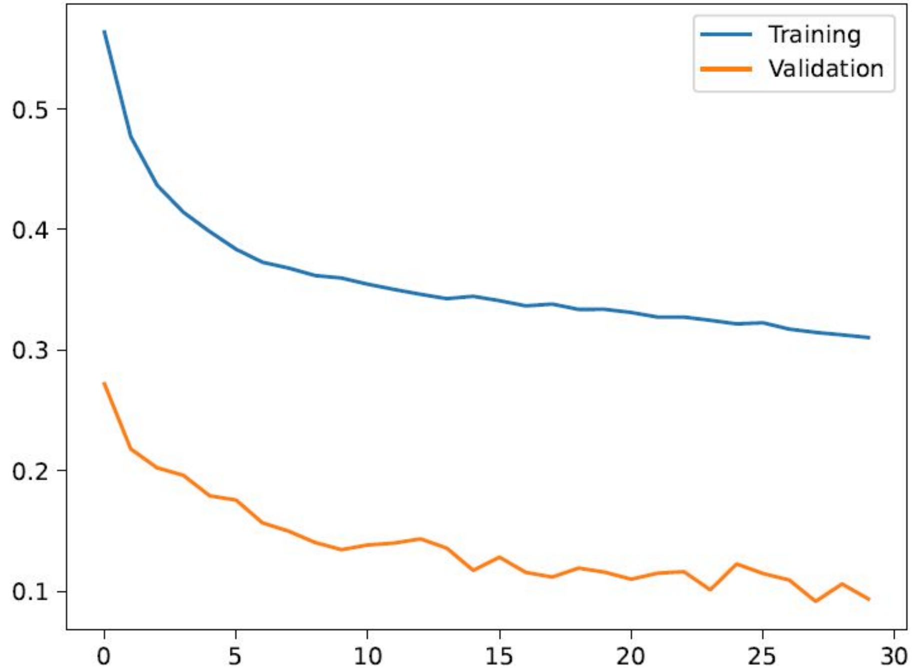
Epochs

Validation



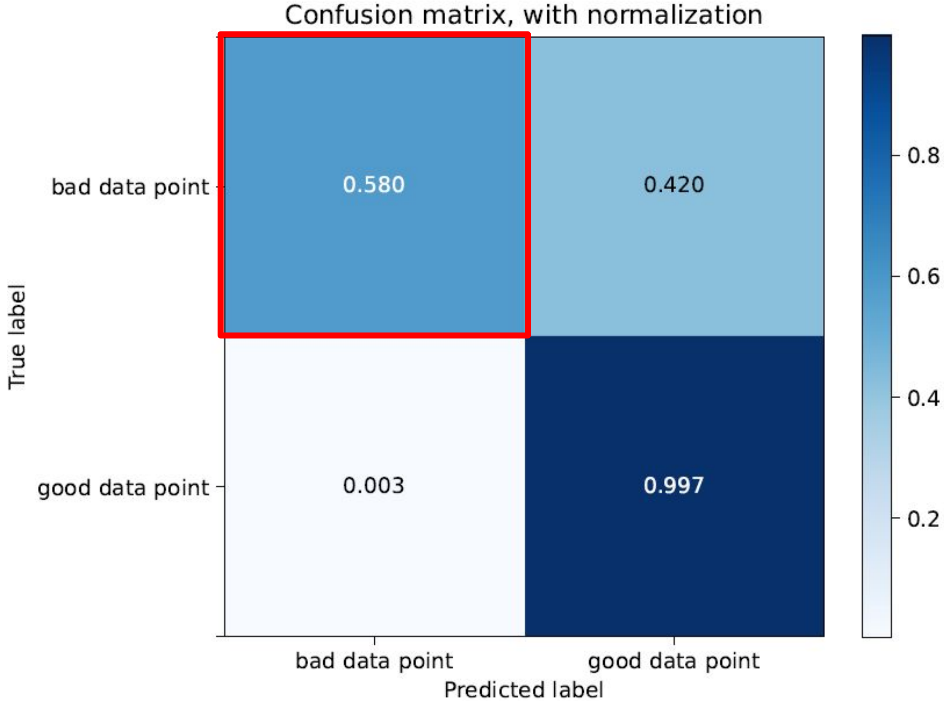
Bad Performance: Northwest regional model

Loss



Epochs

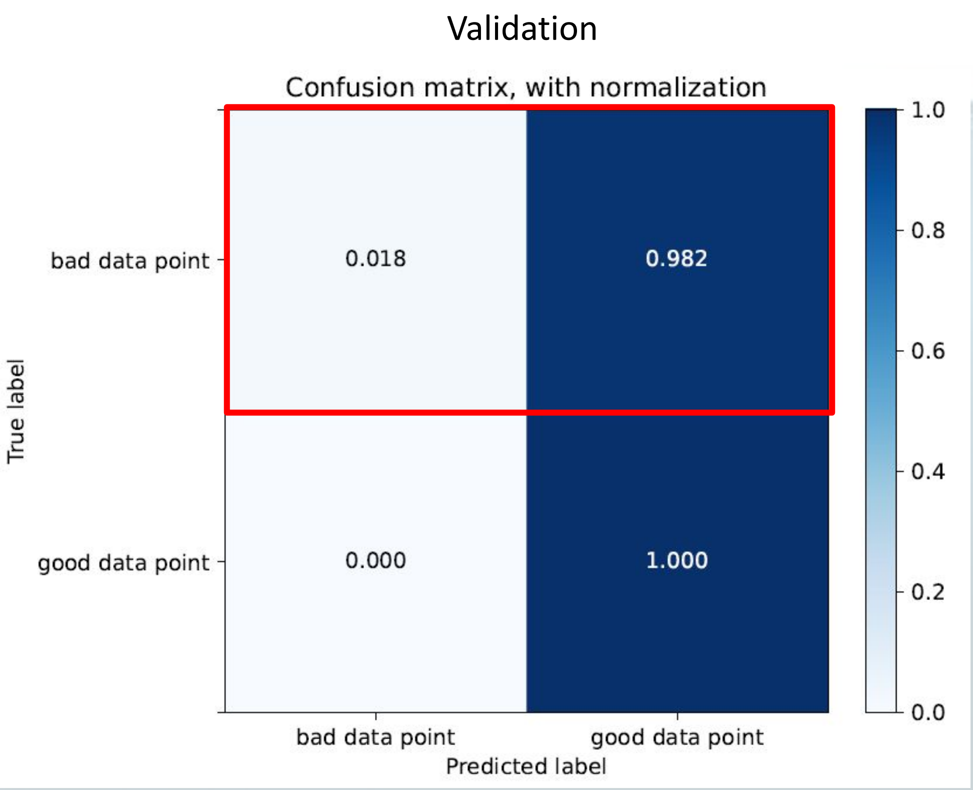
Validation



The Culprit- Bad backup data for Tacoma, WA?

Bad backup data in training resulting in model overfitting?

Need to be consistent in how “bad” backup data is handled when training

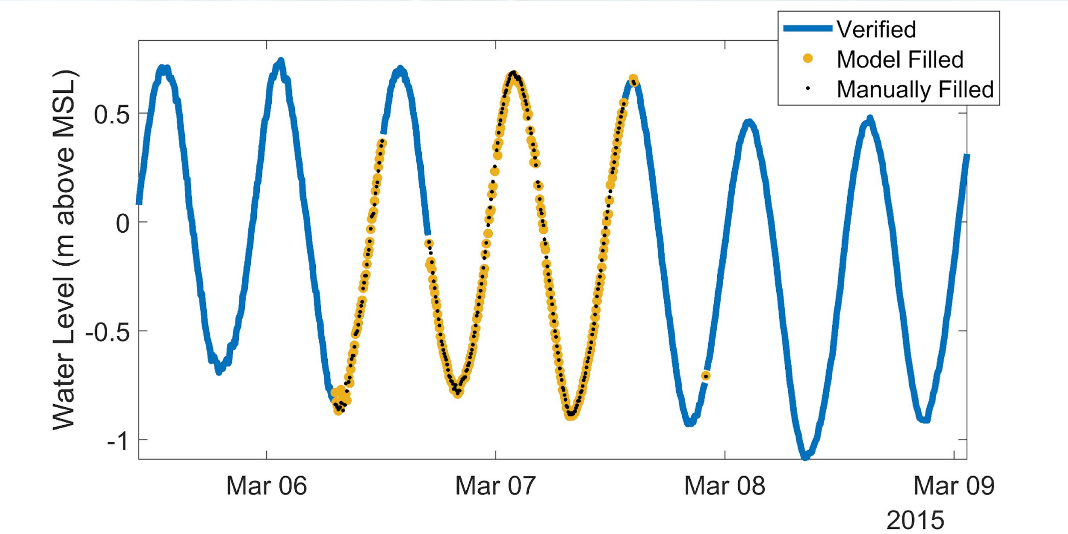


Main Takeaways

- Machine Learning (ML) approaches can be used to accurately classify good/bad water level data.
- Neural net model is best approach from our testing so far.
- Quality of data used for training is very important (highlights importance of **AI-ready data**).
- GPUs can be used to perform many experiments at once and will accelerate AI/ML research.
- An automated AI-enabled QC application would substantially reduce CO-OPS resource requirements and could be used as a community tool.

Next Steps

- Troubleshoot issues with AI model for classifying bad water level data
- Bad data is rare: more data preprocessing/training while ensuring rare cases are present
- Develop and test algorithms to correct bad water level data- more complicated and will require more work/external collaboration
- Initial results are promising (using regression model) :



Accelerating our project

- Advanced project at OpenACC Hackathon (UF) with NVIDIA mentors (Ryan Simpson & others) and collaborators from IOOS & Texas A&M-Corpus Christi & used UF’s HPC (HiPerGator)
- Convert to Parquet data format to work with RAPIDS library (much faster performance)
- Looking to set up CO-OPS cloud server with GPU time

Questions for You

Have you employed automatic QC measures?

Would you be interested in a QC tool that automatically identifies **and** corrects bad data?

What gap filling algorithms do you currently use? Recommendations on AI/ML methods for gap filling?

Opportunities to collaborate? Can we use your water level datasets to test our model?

Contact Information:
Jimmy Spore (james.spore@noaa.gov)
Lindsay Abrams (lindsay.abrams@noaa.gov)