Naval Facilities Engineering Systems Command
EXWC Engineering and Expeditionary Warfare Center

Development of a Tool to Predict Potential Future Environmental Violations

May 21, 2024

Hunter Klein
Kendrick White
NAVFAC EXWC

Distribution Statement: This information is Distribution A; Approved for public release; distribution is unlimited.

Presentation Outline

- Project Background
- Technical Approach
- Technical Progress
  - Database Survey
  - Model Development
  - Data Analysis
  - Tool development
- Future Efforts and Transition Plans
- Conclusions

Project Team

Team Member	Role
Hunter Klein	Principal Investigator
Kendrick White	AI/ML Technical Expert
Mark Savala	Web Application Technical Expert
Josh Fortenberry	CWA/SDWA Compliance Expert; Water Media Field Team

Background

- Project is an ongoing research, development, test, and evaluation (RDT&E) project
- Sponsored by Navy Environmental Sustainability Development to Integration (NESDI) Program
- Began through NESDI need for AI/ML computing to quantify Navy's regulatory compliance risk
- NESDI funding as proof-of-concept to develop and test relevant AI/ML applications towards benefitting environmental compliance

Project Overview

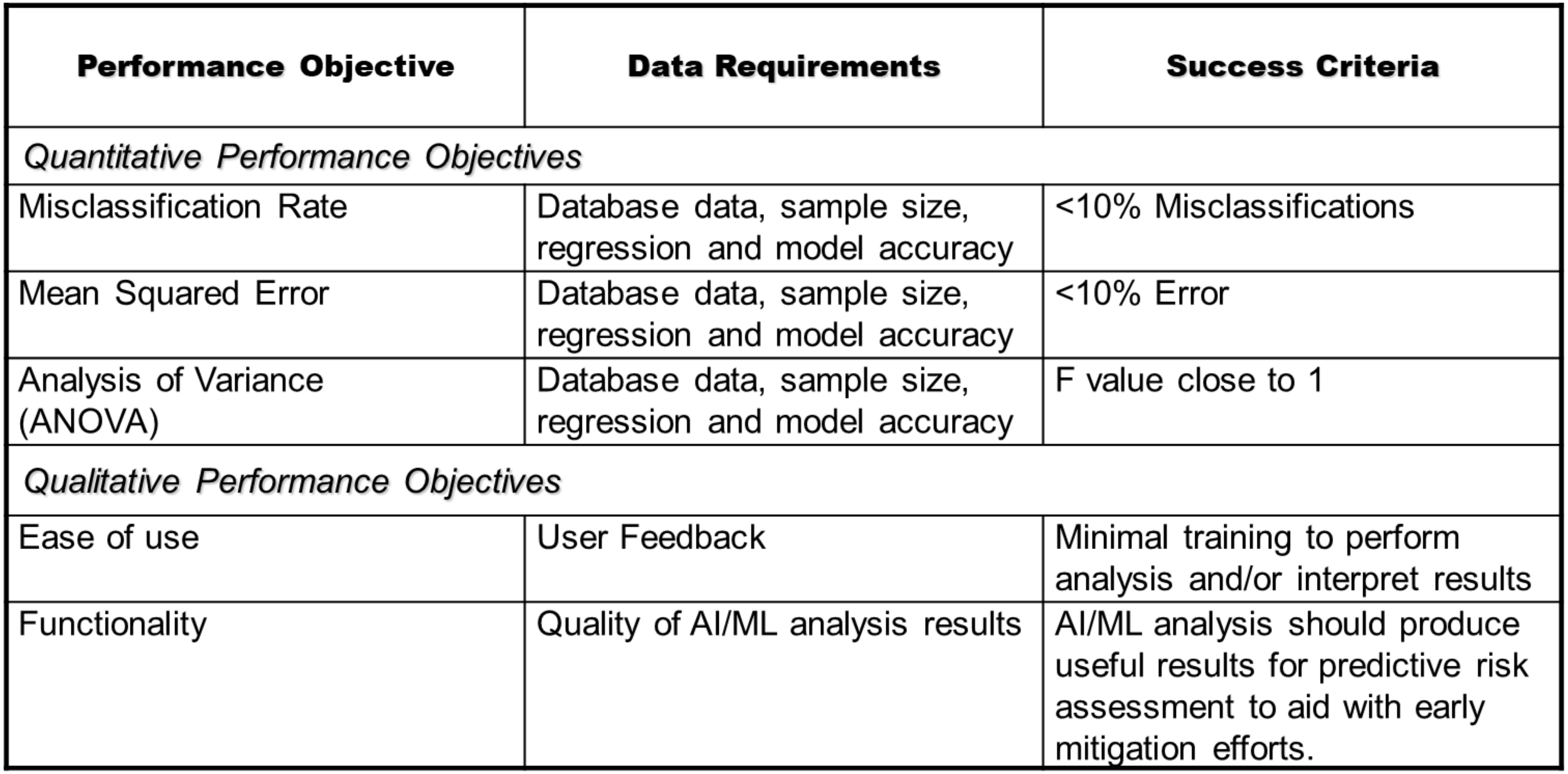
Objective

- Develop various models and a tool to analyze historic data to develop future projections for quantifying the risk of an exceedance or violation
- Support mitigation efforts to avoid potential noncompliance

Technical Approach

- Identify available environmental data for Navy installations
- Review and identify potential models that could meet objective
- Develop predictive models using available environmental data
- Compare and evaluate models based on performance criteria
- Develop basic front-end tool to display results and support future transition

Performance Criteria



Performance Objective	Data Requirements	Success Criteria
Quantitative Performance Objectives		
Misclassification Rate	Database data, sample size, regression and model accuracy	<10% Misclassifications
Mean Squared Error	Database data, sample size, regression and model accuracy	<10% Error
Analysis of Variance (ANOVA)	Database data, sample size, regression and model accuracy	F value close to 1
Qualitative Performance Objectives		
Ease of use	User Feedback	Minimal training to perform analysis and/or interpret results
Functionality	Quality of AI/ML analysis results	AI/ML analysis should produce useful results for predictive risk assessment to aid with early mitigation efforts.

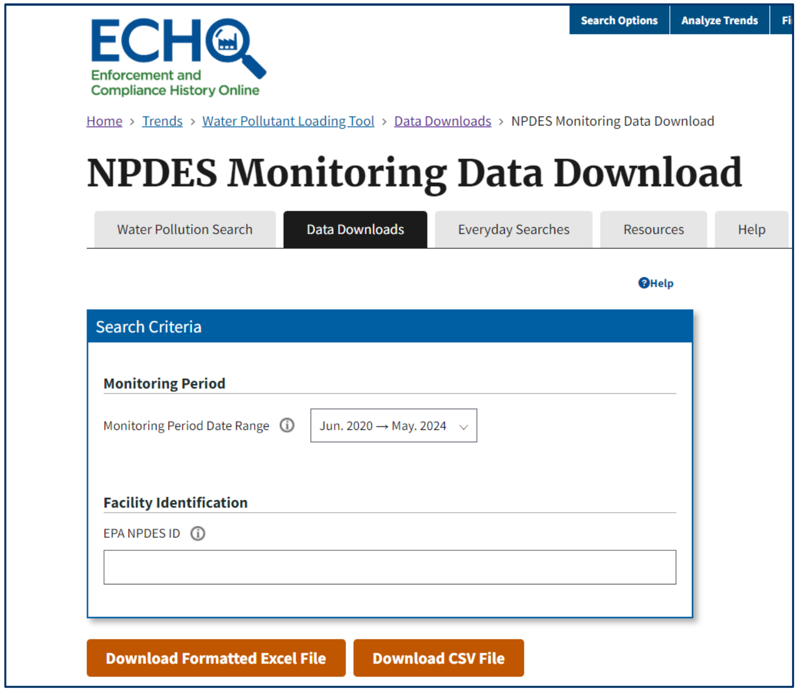
Database Survey

Overview

- Identify data available through environmental databases
- Determine feasibility of downloading, formatting, and processing the existing data for use by the models

Results

- NAVFAC has internal data focused primarily on exceedances only
- Installation-level monitoring data may be available, but formatting could vary
- ECHO has public data for NPDES, SDWA, CAA, Hazardous waste
- NPDES data seemed most comprehensive for Navy installations



Home › Trends › Water Pollutant Loading Tool › Data Downloads › NPDES Monitoring Data Download

NPDES Monitoring Data Download

Search Criteria

Monitoring Period

Monitoring Period Date Range ⓘ Jun. 2020 → May. 2024

Facility Identification

EPA NPDES ID ⓘ

NPDES Data

Common Permit Types

- Industrial Stormwater
  - Stormwater runoff from industrial facilities
- General Permits
  - Groups similar facilities
- Individual Permits
  - Single facilities; includes WWTFs

Benefits

- Historic data going back to 2007
- Majority of Navy installations have at least 1 permit
- Standardized formatting
- Small CSV file downloads

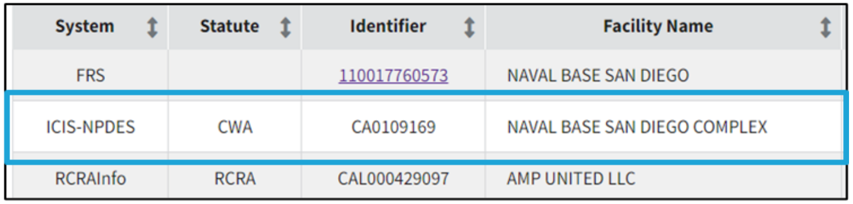
Source: USEPA - JBPHH WWTP, 2012

Source: NAVFAC HI - JBPHH WWTP Fact Sheet, 2020

NPDES Data

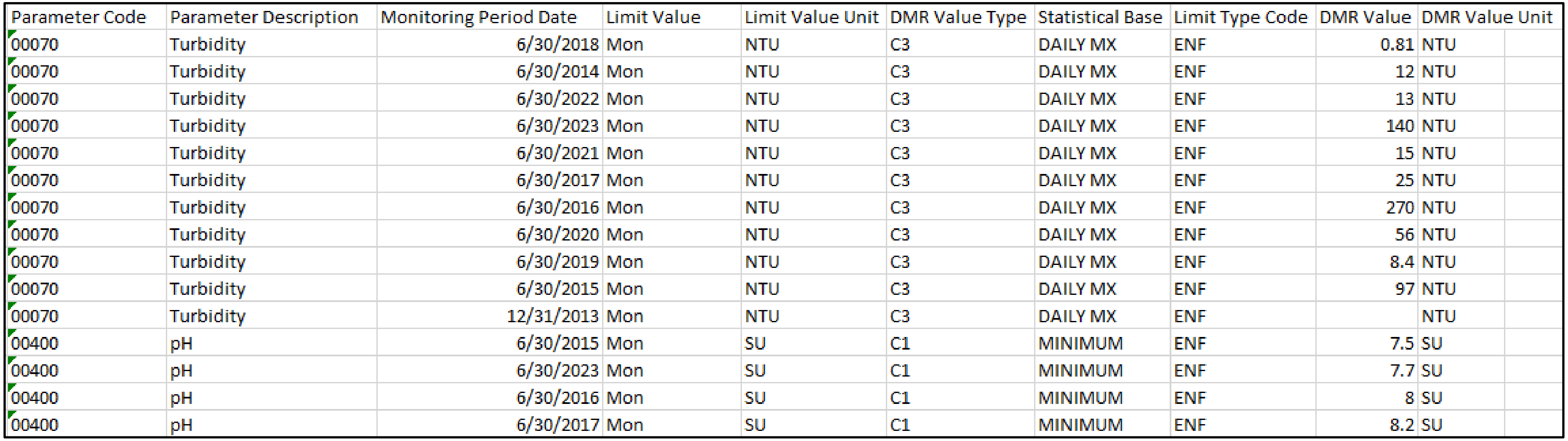
Limitations and Constraints

- Can be difficult to identify facilities covered under general permits
- Data typically is only reported once a month



System	Statute	Identifier	Facility Name
FRS		110017760573	NAVAL BASE SAN DIEGO
ICIS-NPDES	CWA	CA0109169	NAVAL BASE SAN DIEGO COMPLEX
RCRAInfo	RCRA	CAL000429097	AMP UNITED LLC

NPDES Data Example



Parameter Code	Parameter Description	Monitoring Period Date	Limit Value	Limit Value Unit	DMR Value Type	Statistical Base	Limit Type Code	DMR Value	DMR Value Unit
00070	Turbidity	6/30/2018	Mon	NTU	C3	DAILY MX	ENF	0.81	NTU
00070	Turbidity	6/30/2014	Mon	NTU	C3	DAILY MX	ENF	12	NTU
00070	Turbidity	6/30/2022	Mon	NTU	C3	DAILY MX	ENF	13	NTU
00070	Turbidity	6/30/2023	Mon	NTU	C3	DAILY MX	ENF	140	NTU
00070	Turbidity	6/30/2021	Mon	NTU	C3	DAILY MX	ENF	15	NTU
00070	Turbidity	6/30/2017	Mon	NTU	C3	DAILY MX	ENF	25	NTU
00070	Turbidity	6/30/2016	Mon	NTU	C3	DAILY MX	ENF	270	NTU
00070	Turbidity	6/30/2020	Mon	NTU	C3	DAILY MX	ENF	56	NTU
00070	Turbidity	6/30/2019	Mon	NTU	C3	DAILY MX	ENF	8.4	NTU
00070	Turbidity	6/30/2015	Mon	NTU	C3	DAILY MX	ENF	97	NTU
00070	Turbidity	12/31/2013	Mon	NTU	C3	DAILY MX	ENF		NTU
00400	pH	6/30/2015	Mon	SU	C1	MINIMUM	ENF	7.5	SU
00400	pH	6/30/2023	Mon	SU	C1	MINIMUM	ENF	7.7	SU
00400	pH	6/30/2016	Mon	SU	C1	MINIMUM	ENF	8	SU
00400	pH	6/30/2017	Mon	SU	C1	MINIMUM	ENF	8.2	SU

WQI Calculation

- Provides simplified single output to track overall water quality trends
  - Also tracking and forecasting individual parameters
- Plan to further develop WQI model to include more parameters

Simple WQI model *(from Purdue University)*

$$\boxed{\text{WQI} = \text{TEMP}*(\text{BOD} + \text{TSS} + \text{DO} + \text{COND})}$$

- TEMP - Temperature index *(range 0-1)*
- BOD - Biochemical Oxygen Demand index *(out of 30)*
- TSS - Total suspended solids index *(out of 25)*
- DO - Dissolved Oxygen index *(out of 25)*
- COND - Conductivity index *(out of 20)*
- WQI is out of 100, with 100 being EXCELLENT and 0 being POOR

*WQI = Water Quality Index

Identification of Potential Models

Desired Capabilities

- Track potential trends and correlations using available CSV files
- Potential to incorporate external factors from other data sources (e.g. NOAA weather data)
- Forecast future values for individual parameters and overall WQI

Potential Suitable Models

- Fuzzy Inference Systems
- Hidden Markov Models
- Linear Autoregressive Integrated Moving Average Models
- Recurrent Neural Networks

Fuzzy Inference Systems

Description

- Artificial Neuro-Fuzzy Inference Systems (ANFIS) perform IF-THEN inferences using fuzzy logic.
- Outputs are numerical quantities between 0 and 1, that represent the degree to which the inference is true.
- The degree of truth is learned by an Artificial Neural Network (ANN).
- For example, the statement “If x is above THRESHOLD, then y is above level B” would be learned by exposing an ANN to many instances of pairs (x, y). After training, the degree of truth for new instances (x, y) would be estimated.

Advantages

- ANFIS is attractive in that hard boundaries can be avoided and varying imprecision are incorporated
- Highly recommended in WQI literature.

Limitations (why not used)

- In the absence of ground truth WQI estimates, training an ANFIS system is not possible with supervised learning.

Hidden Markov Model

Description

- A Hidden Markov Model (HMM) assumes that there is an underlying unobserved system that is driving the measured observations. At every point in time, the system is described by an underlying state X_t, and given X_t there is a probability that a particular observation Y_t is observed.
- For a sequence of observations $Y_1, Y_2, Y_3, \ldots, Y_t$ we attempt to infer the sequence of states $X_1, X_2, X_3, \ldots, X_t$ that produced those observations. It is often assumed that there are only a finite number of different possible states.

Advantages

- HMMs have shown excellent performance in predicting the movement of asset returns, as well as speech recognition and language processing

Limitations (why not used)

- Assumes observations are independent. Assumes linear relation between underlying state and observations.

Linear ARIMA Models

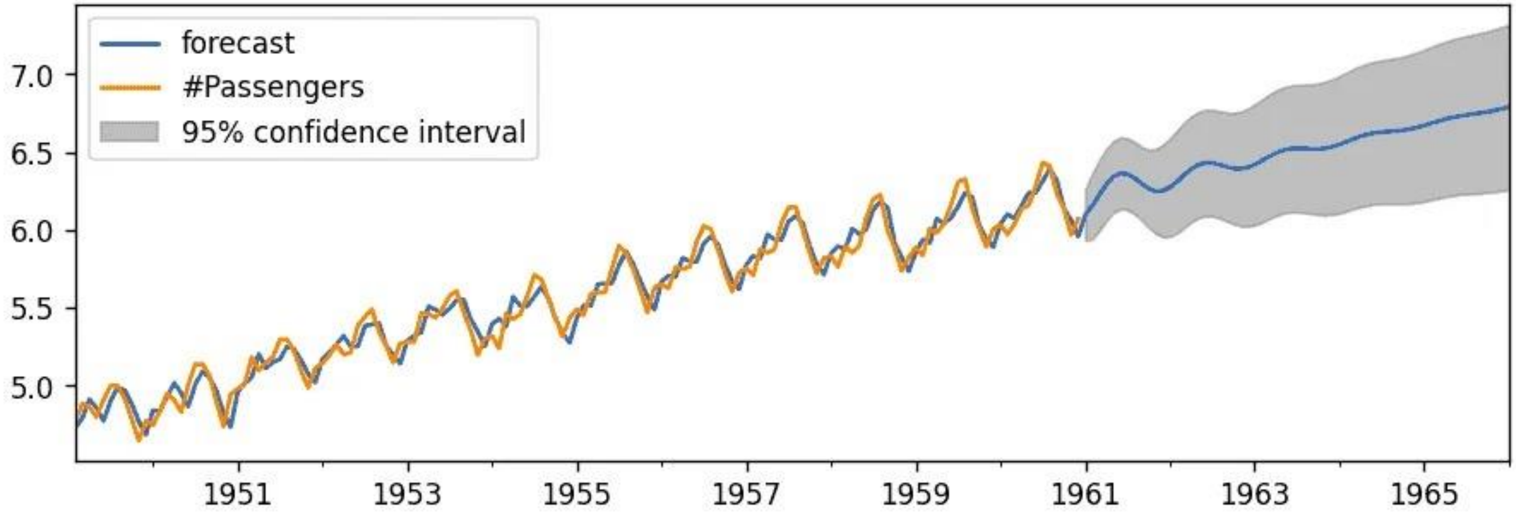
Description

- Autoregressive Integrated Moving Average (ARIMA) models are a general class of time series models in which it is assumed that current observations can be explained by a linear combination of past observations and random inputs, or "shocks", originating from outside the system.
- The equation for an ARMA model is of the form

$$\boxed{\begin{gathered} Y_t = A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_q Y_{t-q} + B_1 e + B_2 e + \cdots + B_p e_{t-p} \\ = A_0 + \sum A_i Y_{t-i} + \sum B_i e_{t-i} \end{gathered}}$$

  - Where the constant A_0 is the system average, $\sum A_i Y_{t-i}$ is the autoregressive component, and $\sum B_i e_{t-i}$ is the moving average component.
  - The autoregressive component represent how well the system can be explained its past values.
  - The moving average component represents how much of the system is explained by outside inputs

Linear ARIMA Models



Source: Using Arima Model and Python for Time Series Forecasting. Barshir Alam, 2022

Advantages

- Interdependence amongst parameters is readily apparent
- Model is easily interpretable
- Simple model that can be developed quickly

Potential Limitations

- Environmental monitoring data is likely not linear nor statistically stable

Recurrent Neural Networks

Description

- Recurrent Neural Networks (RNNs) are generalizations of mapping Artificial Neural Networks (ANNs) that can associate entire sequences of input vectors with an output vector sequence. That is, they can map sequences to a single output, or vice versa. This capability allows RNNs to capture serial dependence amongst input values.

Advantages

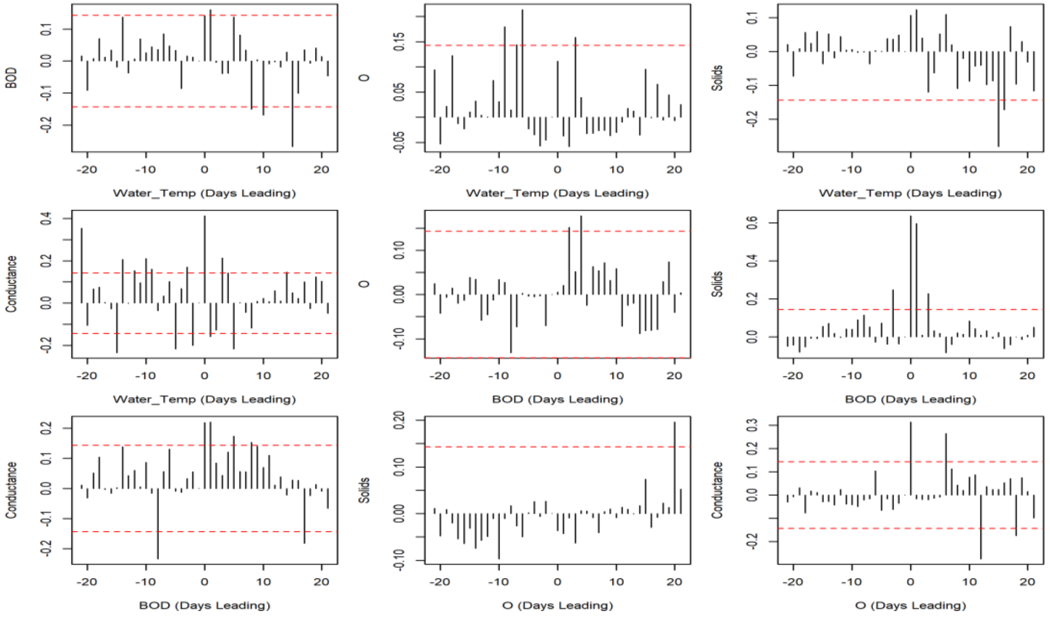
- Mapping entire sequences allows RNNs to capture serial dependence amongst input values. This for smaller, simpler machine learning models, and more efficient use of data.
- ANNs are flexible in that they can approximate any nonlinear function.

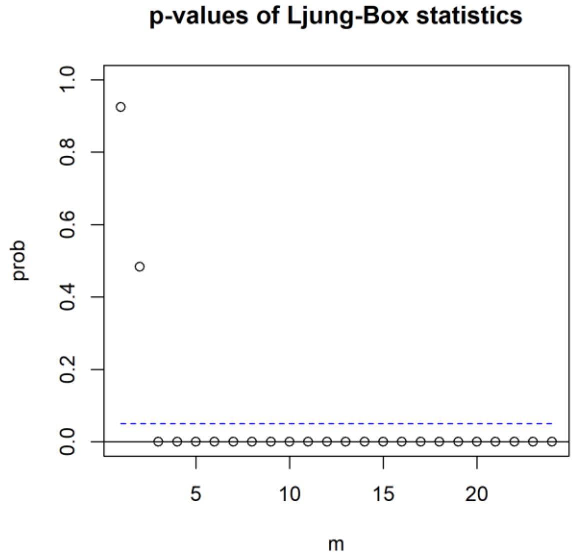
Potential Limitations

- The layering of ANNs does facilitate analysis of how inputs interact. Their flexibility can lead to poorer generalization to new inputs.

AI/ML Model Development

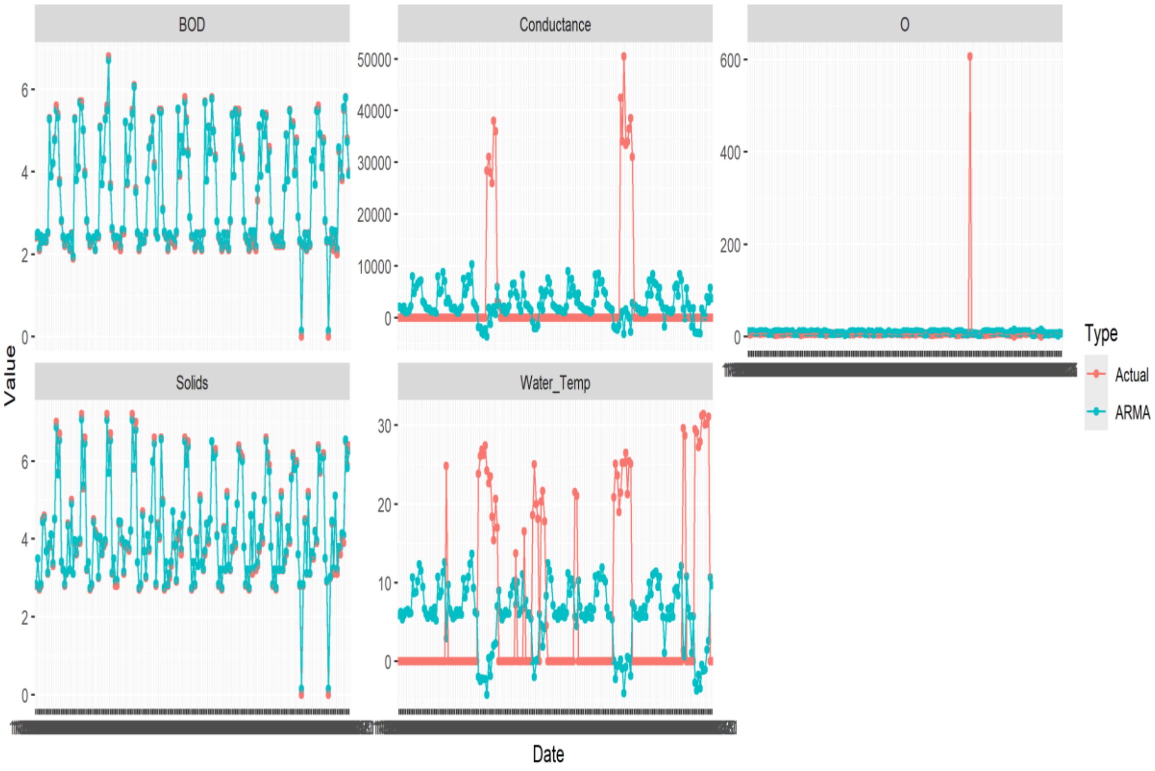
- Parameter data is a multivariate time series. Imperative to analyze serial- and cross-correlations amongst parameters to inform model development.
- Multivariate Partial Autocorrelation Function indicates possible time lags amongst parameters.
- Box-Ljung statistic suggests order 1 lag for linear model

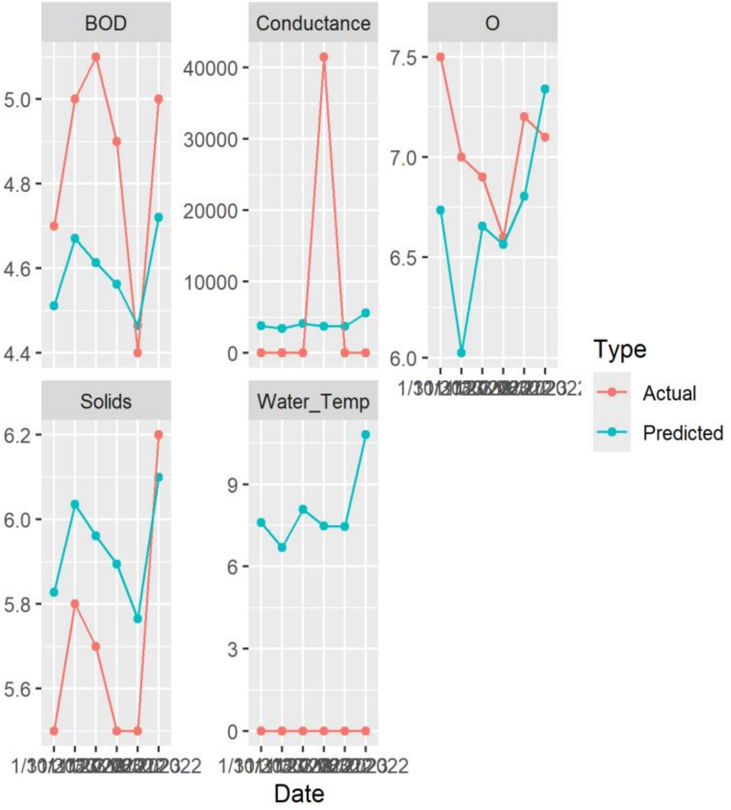




ARIMA Model Development

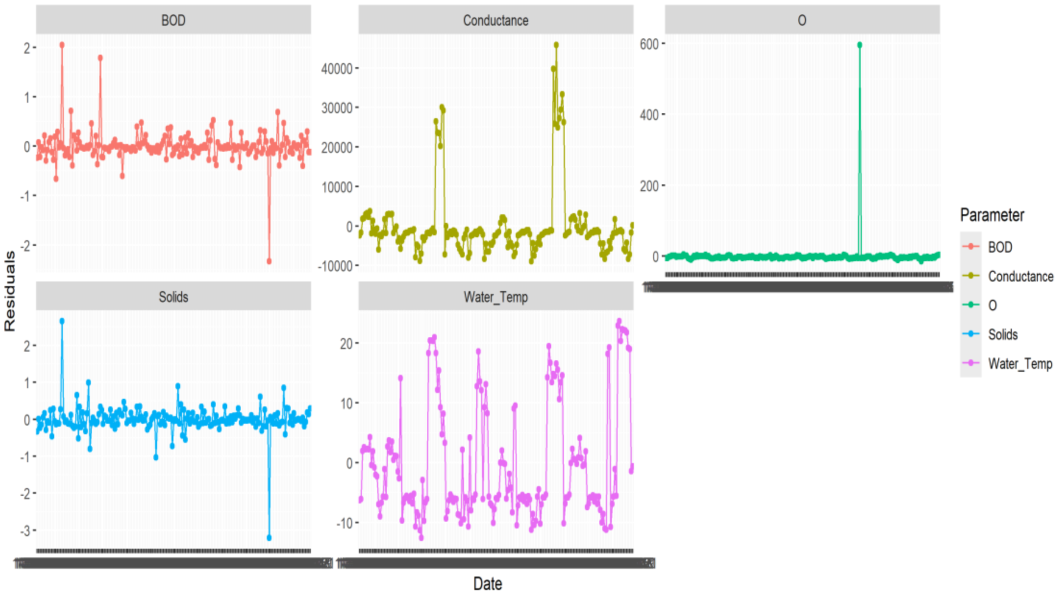
- Order 1 Autoregressive model, AR(1), was fit to each facility record using Maximum Likelihood Estimation.

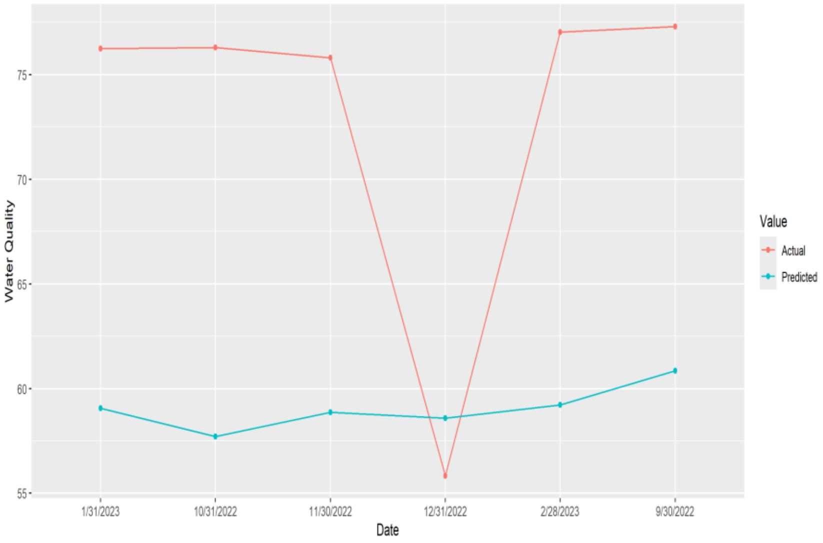




ARIMA Fit Diagnostic

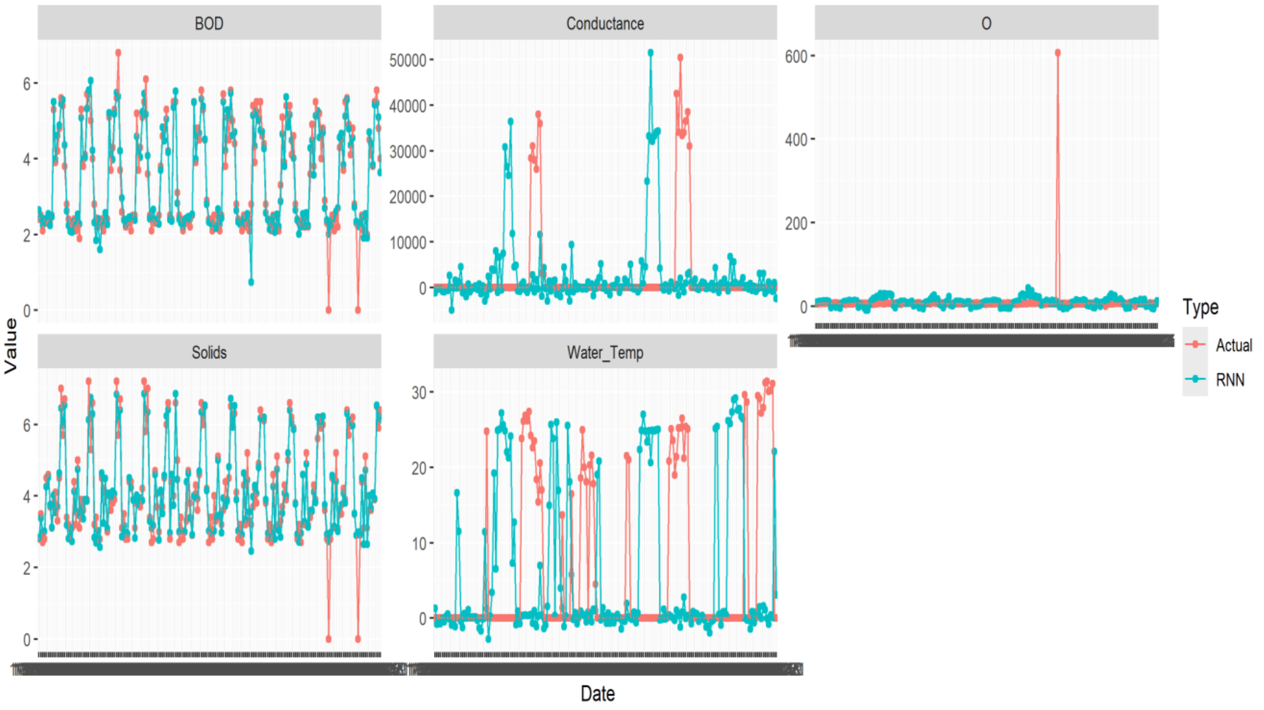
- Residuals from AR(1) fit for some parameters show slight skew.
- Parameter shrinkage could be employed to tighten the model.

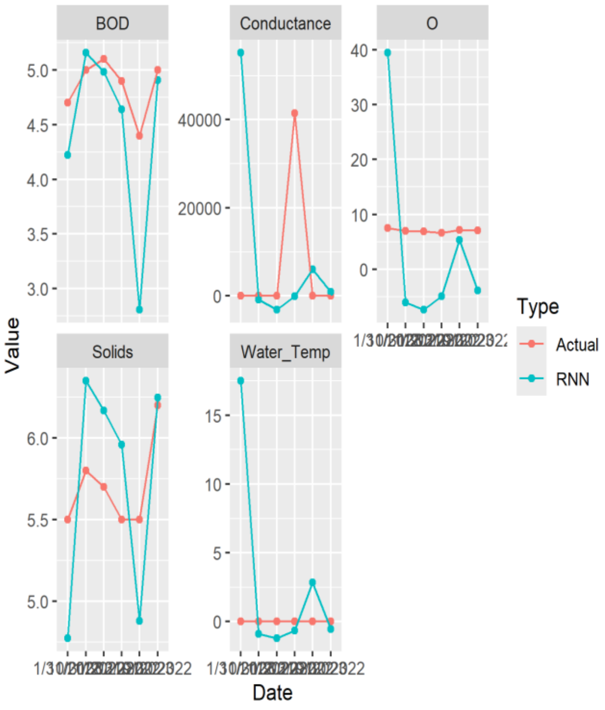




RNN Model Development

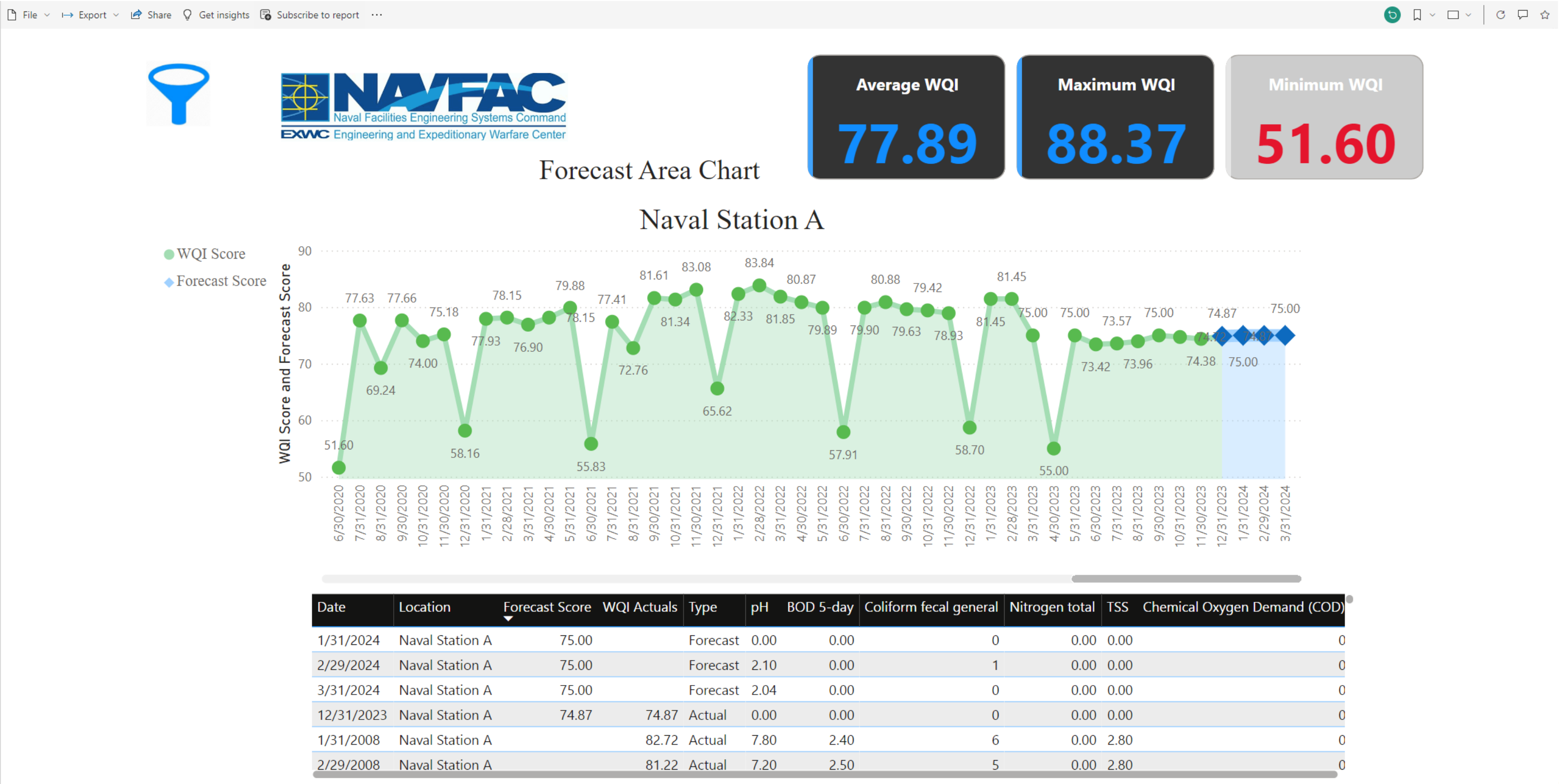
- An order-2 Recurrent Neural Network (RNN) was fit to each facility record.
- Five (5) layers, each with 5 neurons.





Web-Based Tool Development

- Web application being developed to allow end users to view data models and prediction outputs quickly and easily
- Using Power BI as part of Navy’s Microsoft tool package
- Serves as basis for future transition and integration efforts



Average WQI
77.89

Maximum WQI
88.37

Minimum WQI
51.60

Date	Location	Forecast Score	WQI Actuals	Type	pH	BOD 5-day	Coliform fecal general	Nitrogen total	TSS	Chemical Oxygen Demand (COD)
1/31/2024	Naval Station A	75.00		Forecast	0.00	0.00	0	0.00	0.00	0
2/29/2024	Naval Station A	75.00		Forecast	2.10	0.00	1	0.00	0.00	0
3/31/2024	Naval Station A	75.00		Forecast	2.04	0.00	0	0.00	0.00	0
12/31/2023	Naval Station A	74.87	74.87	Actual	0.00	0.00	0	0.00	0.00	0
1/31/2008	Naval Station A		82.72	Actual	7.80	2.40	6	0.00	2.80	0
2/29/2008	Naval Station A		81.22	Actual	7.20	2.50	5	0.00	2.80	0

Technology Transition and Future

Remaining with Project

- Further model developments and performance evaluation
- Further development of web-based tool for Navy and/or DoD use

Future Work

- Integration of tool into Navy share-sites for end-users to access through established Navy tools
- Application of models with other environmental data-types (e.g. air quality, drinking water, etc)
- Potential for utilizing more comprehensive internal monitoring data that installations may have

Conclusions

Recap

- Development of web-based tool and AI/ML modeling approach to predict noncompliance risk
- RNN and ARIMA models are currently the primary focus for predicting NPDES water pollution discharge
- Data analysis in progress but preliminary results are promising
  - Initial models reached 0.28 mean squared error
  - Performance can be improved with model tuning

Benefits

- Tool outputs and predictions should aid compliance risk identification
- Supports early mitigation efforts to avoid exceedances and noncompliance

Contact Information

For additional information, please contact

Hunter Klein

hunter.w.klein.civ@us.navy.mil

Kendrick White

kendrick.l.white.civ@us.navy.mil

Q/A