Platfform yr Amgylchedd Cymru
Environment Platform Wales

Better Water Quality for Wales Conference
27-29 June, 2023
USW Exchange, Newport and Online

An Artificial Intelligence Model for Bathing Water Quality Early Warning Systems

M.Y. Lam and R. Ahmadian

Funded by the Interreg Atlantic Area Program, EERES4WATER project (EAPA 1058/2018)

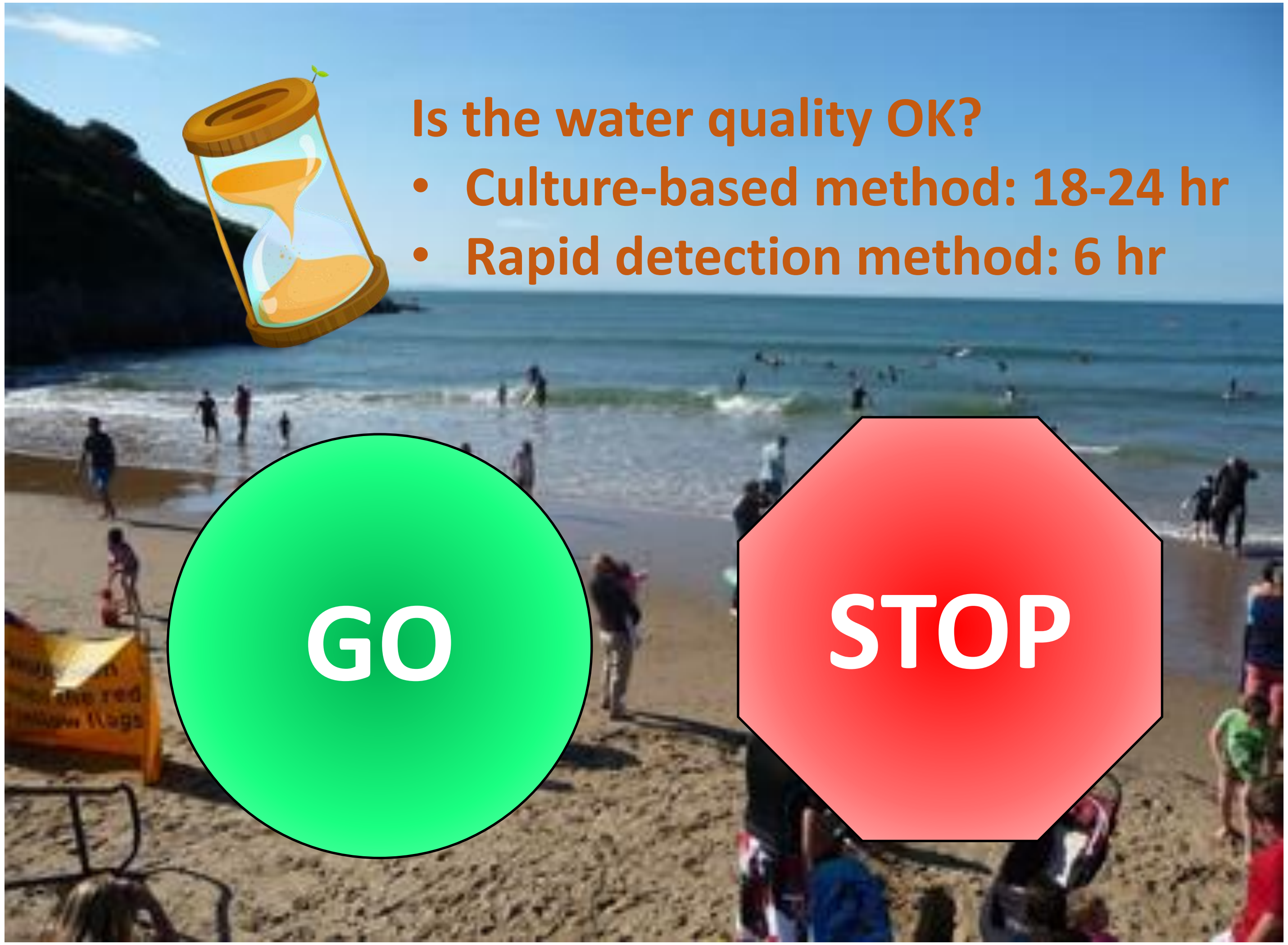
Cardiff University Prifysgol Caerdydd | eeres4water | European Union European Regional Development Fund | 1872 Prifysgol Aberystwyth University | smart coasts

Outline

- Need for real-time predictive models for bathing water quality early warning systems
- A novel Gamma-GA-ANN data-driven model
- Model testing at Swansea Bay, UK for predicting *Enterococci* and *E Coli*
- Conclusion

Public warning systems for bathing water quality

- Pathogen in bathing water cause public health problems
- Warning systems to inform the public about poor water quality
- Real-time prediction methods required
- This presentation introduces a novel predictive method



https://www.tourismforall.co.uk/news/read/2019/07/swansea-beach-adds-more-accessible-facilities-b77

Water quality prediction models

- Benefit aquaculture and water treatment
  - Evaluate environmental management strategies
  - Reduce energy use in water treatment: treat wastewater only to the required levels but not cleaner

Wastewater Treatment
Works (WWTW)

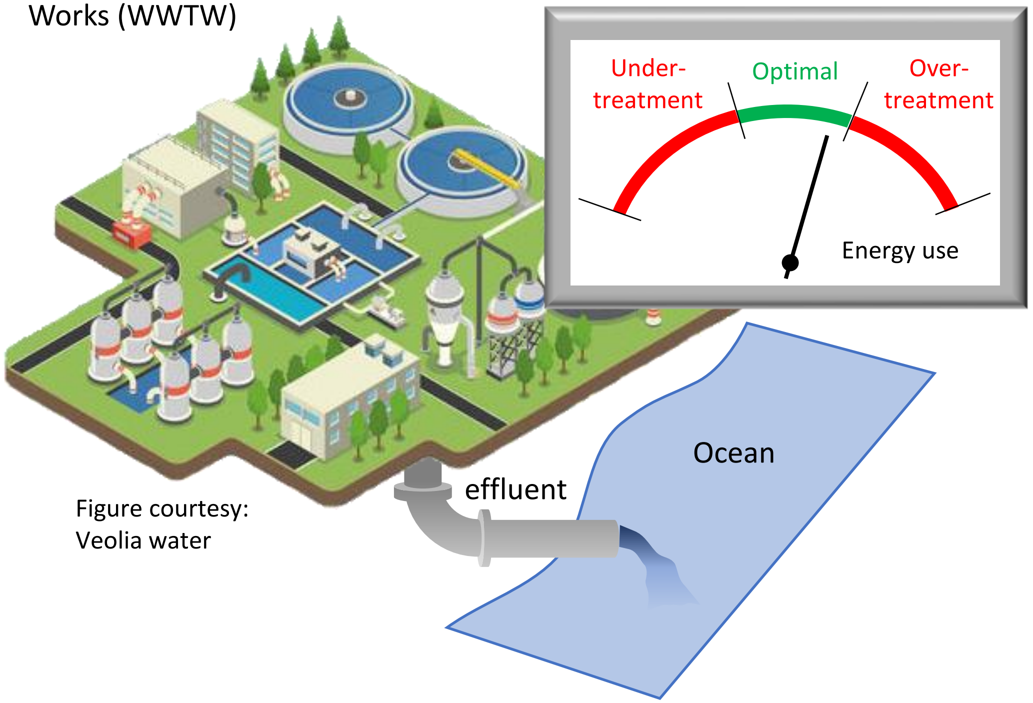
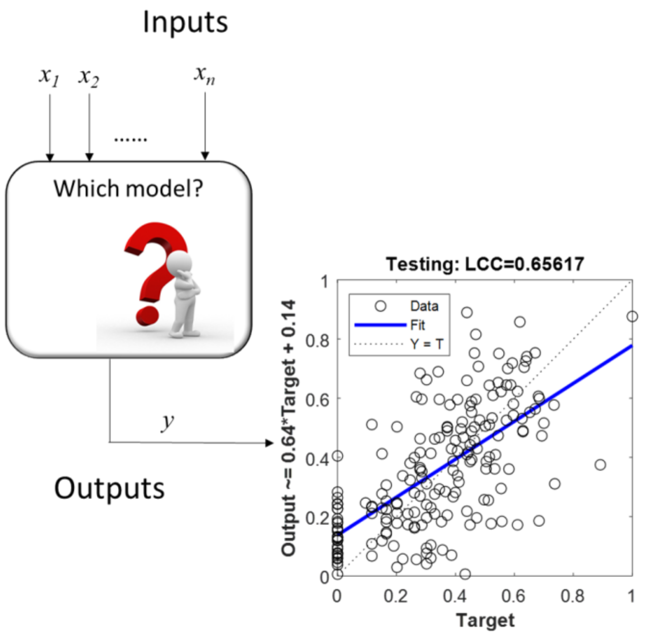


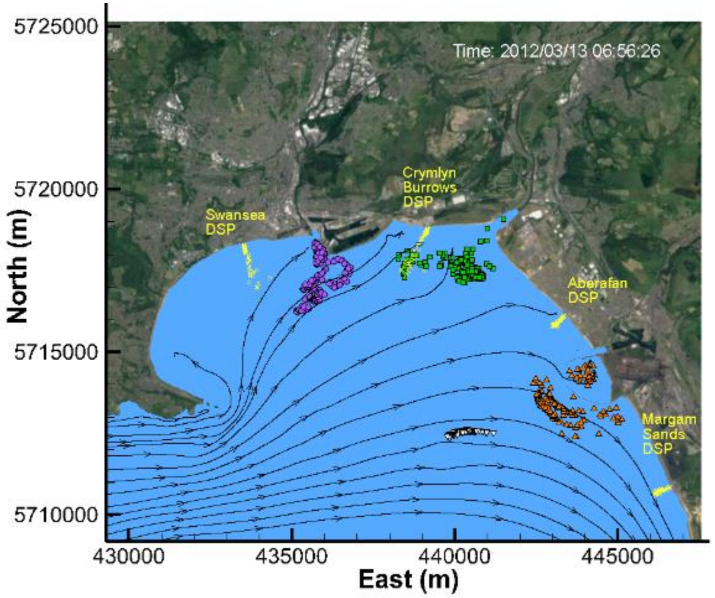
Figure courtesy:
Veolia water

Types of water quality models

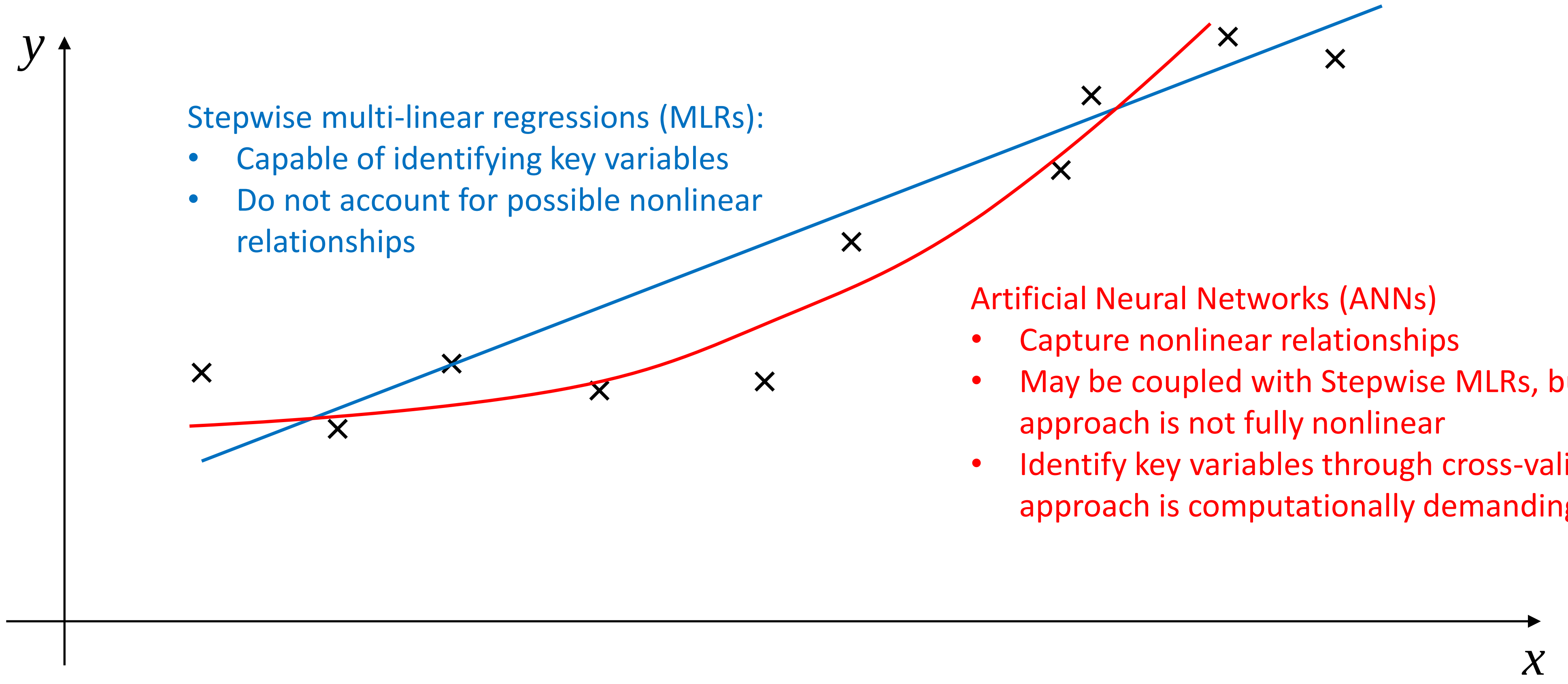
- Artificial Intelligence (AI) models
  - Establish relationship between explanatory variables and bacteria concentrations from measured data
  - Require less computational power – timely prediction



- Hydro-environmental models
  - Solve the flow and bacterial equations
  - Require (i) detail site knowledge, and (ii) high computational power
  - Provide understanding transport and fate of FIO



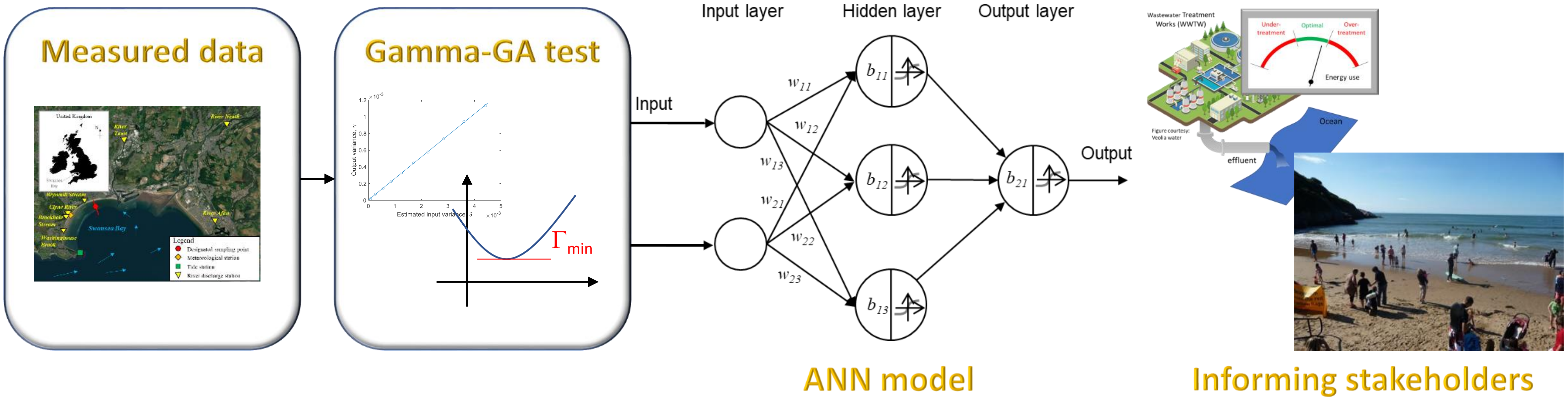
Linear and nonlinear data driven models



- Propose a Gamma test-Genetic Algorithm-ANN (Gamma-GA-ANN) model for full nonlinear variable identification and water quality prediction

Gamma-GA-ANN model

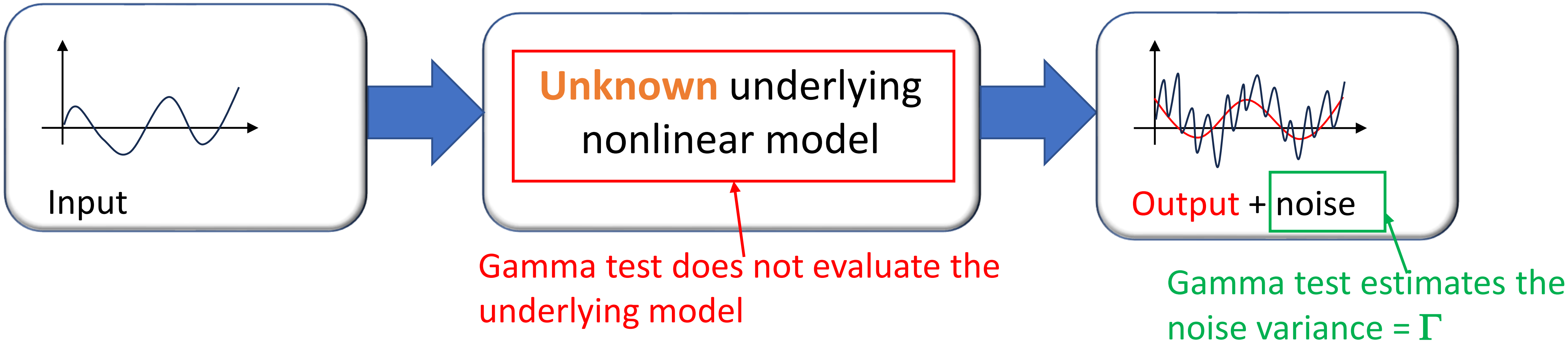
- Method outline:
  - Obtain data with sensors from the site
  - Apply Gamma-GA tests to identify key variables governing FIO concentrations
  - Predict FIO concentrations by ANN models
  - Inform water treatment operators and swimmers of impending poor water quality



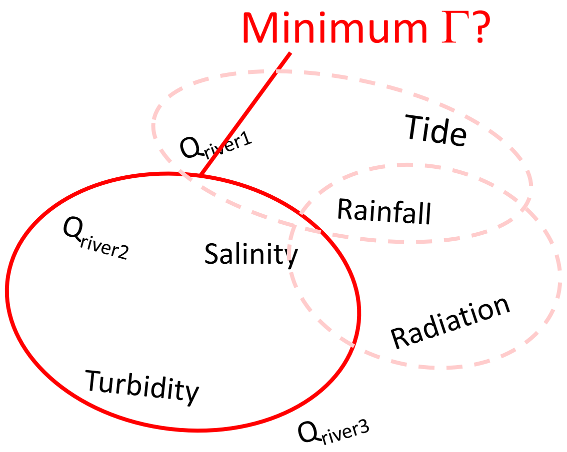
European Union
European Regional Development Fund

Gamma - GA test

FIO concentration in coastal water: An input-system-output representation

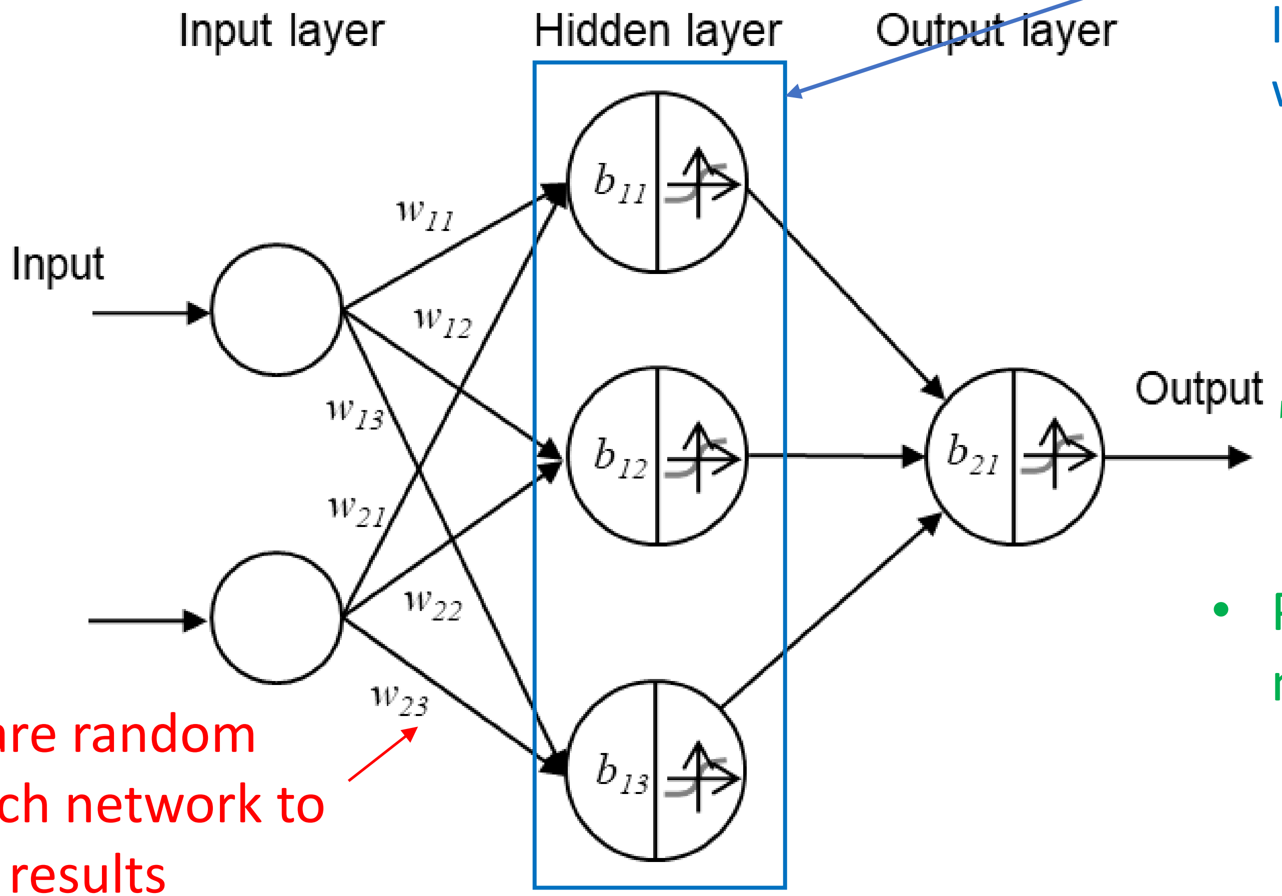


- Variable identification: Gamma-GA test
  - The key variables are the variables that give the smallest Γ.
  - Conduct Gamma tests to all possible variable combinations → computationally demanding.
  - Genetic Algorithm (GA) is employed



Artificial Neural Network

- Feed forward network



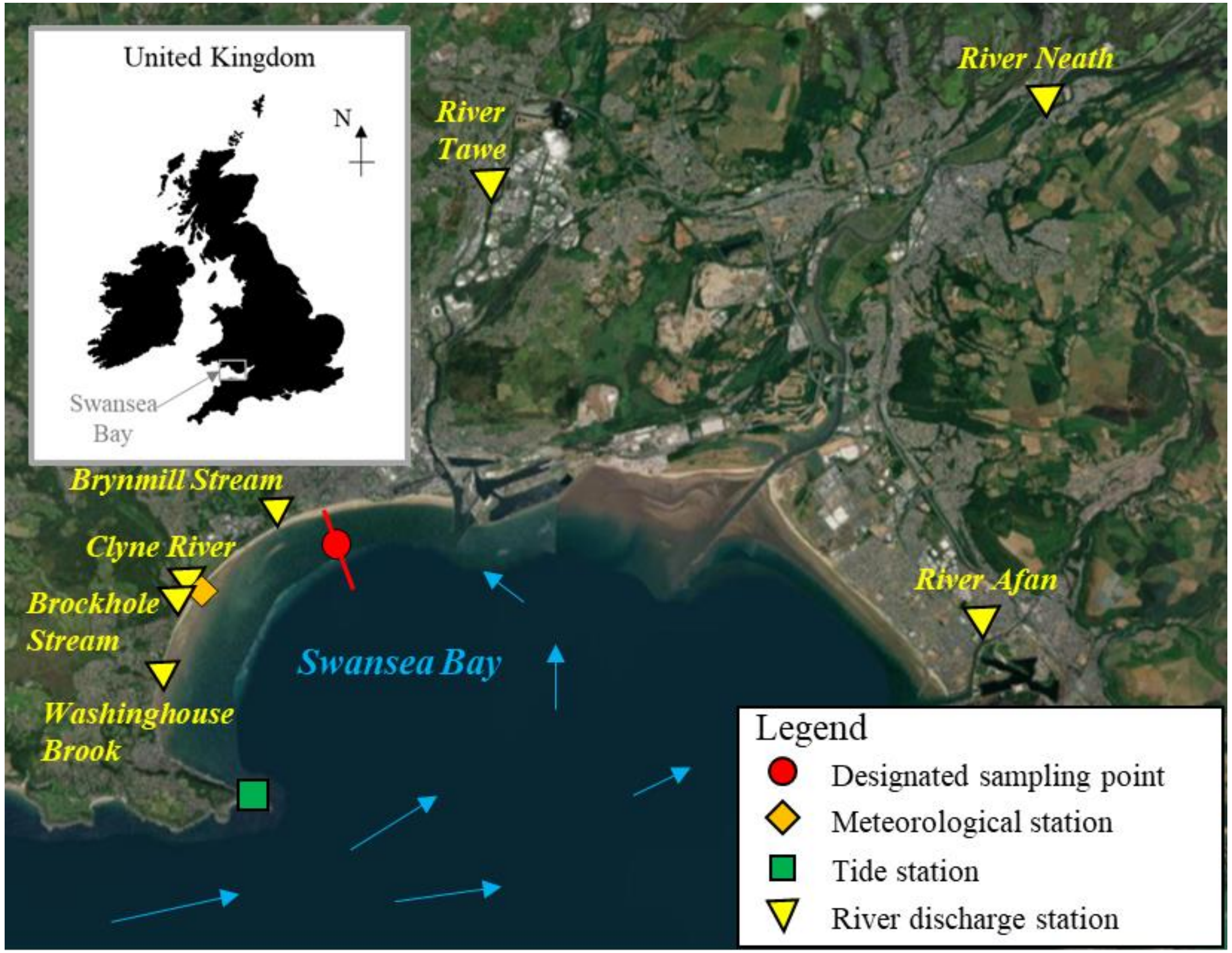
- One hidden layer was sufficient in our study
- Increased the number of hidden layer nodes stepwise and stop when overfitting occurs

- Performance function: mean square error (MSE)

- Initial weights are random
- 300 runs for each network to ensure optimal results

Field data from Swansea Bay, UK

- Popular beaches such as Swansea Beach
- Sampling period: bathing season of 2011
- FIO Sampling interval: 30 min
- Total number of data:
  - 204 variables (including time-lagged data) x 949 time instants = 193596 data
- Remove redundant variables with collinearity test:
  - 23 variables were retained

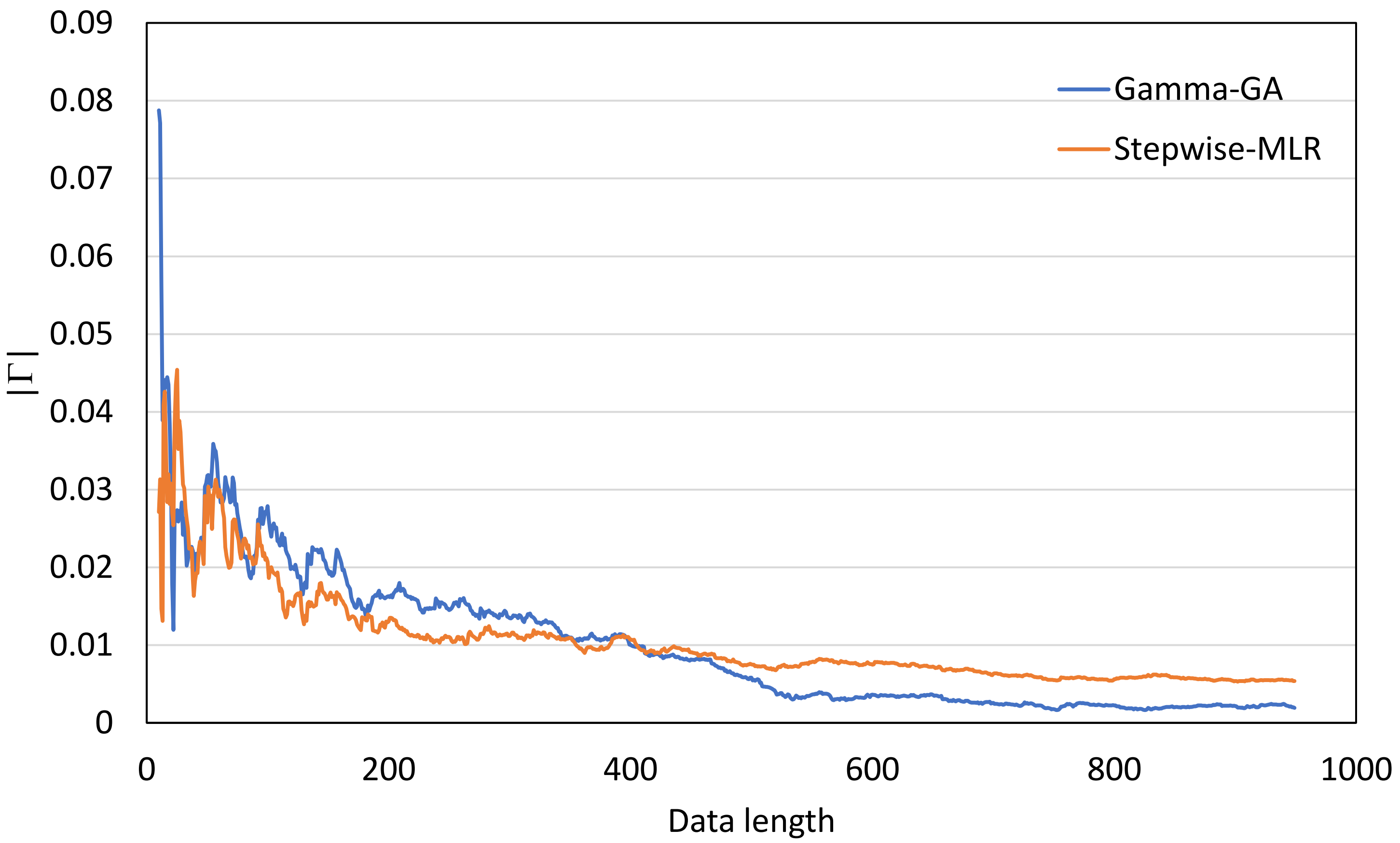


Variables selected by Gamma-GA test

- Tide level and Wind were always selected
    - Consistent with previous mechanistic and AI model results.
- Streamflow was included by the Gamma-GA test for *Enterococci* only
    - River flows are known to be important FIO sources, but
    - Flow rate alone does not fully characterize the effect of rivers

Variables identified from the correlation analysis	Enterococci		E Coli	
	Gamma test	Stepwise linear analysis	Gamma test	Stepwise linear analysis
Streamflow [lag 10 h]	1	0	0	1
Mumbles Level [lag 2 h]	0	0	0	0
Mumbles Level [lag 4 h]	1	0	1	1
Mumbles Level [lag 6 h]	1	1	1	0
Global Radiation [lag 2 h]	0	1	0	1
Global Radiation [lag 4 h]	0	0	1	0
Global Radiation [lag 6 h]	1	0	0	0
Temperature [lag 2 h]	1	0	0	1
Temperature [lag 6 h]	1	0	1	0
Relative Humidity [lag 2 h]	0	1	0	1
Relative Humidity [lag 8 h]	1	0	1	0
Cum. of Rain [lag 2 h]	0	0	0	0
Cum. of Rain [lag 3 h]	0	0	0	0
Cum. of Rain [lag 4 h]	0	0	0	0
Cum. of Rain [lag 6 h]	0	0	0	0
Cum. of Rain [lag 8 h]	0	0	0	0
Cum. of Rain [lag 10 h]	0	0	0	0
Cum. of Rain [lag 12 h]	0	1	0	0
Wind Speed N [lag 2 h]	0	1	1	1
Wind Speed N [lag 6 h]	0	0	1	0
Wind Speed N [lag 10 h]	0	1	0	1
Wind Speed E [lag 2 h]	1	1	1	1
Wind Speed E [lag 10 h]	0	1	0	0

M-test



- Determine the necessary data length for successful AI model development
  - Gamma-GA test selected variables achieved better |Γ| when data length exceeds 500.
- The data were divided into training, validation and testing sets. The training set had more than 500 data points.

Model Comparison

		Key variable identification	
		Gamma-GA test	Stepwise linear (SL) regression
Prediction model	ANN	GG-ANN model	SL-ANN model
	SL	GG-Linear model	SL-Linear model

- Models that use ANN gave a superior predictive performance compared to linear regression models
- Gamma-GA-ANN model gave better prediction results for *Enterococci*

Enterococci

Realization 1						
	MSE			R^2		
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0074	0.0157	0.0214	0.8369	0.6654	0.5361
SL-ANN	0.0210	0.0232	0.0260	0.5400	0.5079	0.4357
GG-Linear	0.0357		0.0399	0.2224		0.1348
SL-Linear	0.0311		0.0328	0.3235		0.2883
Realization 2						
	MSE			R^2		
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0134	0.0172	0.0227	0.7177	0.6025	0.4993
SL-ANN	0.0257	0.0194	0.0295	0.4542	0.5518	0.3611
GG-Linear	0.0368		0.0352	0.2021		0.2246
SL-Linear	0.0312		0.0322	0.3229		0.2895
Realization 3						
	MSE			R^2		
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0071	0.0188	0.0199	0.8292	0.6457	0.6156
SL-ANN	0.0192	0.0243	0.0225	0.5385	0.5418	0.5699
GG-Linear	0.0359		0.0393	0.1944		0.2403
SL-Linear	0.0320		0.0293	0.2812		0.4337

Model Comparison

- GG-ANN model did not always give better prediction for *E Coli* compared to SL-ANN
- Explanation:
  - This *Enterococci* data has more extreme values (17.8% of the data) compared to the *E Coli* data (8.9% of the data); and
  - GG-ANN model is better for capturing extreme values

E Coli

	MSE			R^2		
Realization 1						
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0067	0.0159	0.0214	0.8396	0.6077	0.5312
SL-ANN	0.0096	0.0154	0.0174	0.7705	0.6198	0.6177
GG-Linear	0.0325		0.0342	0.2166		0.2482
SL-Linear	0.0297		0.0305	0.2838		0.3290
Realization 2						
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0119	0.0178	0.0205	0.7208	0.5939	0.4801
SL-ANN	0.0113	0.0135	0.0188	0.7370	0.6914	0.5221
GG-Linear	0.0331		0.0315	0.2294		0.1996
SL-Linear	0.0299		0.0295	0.3041		0.2484
Realization 3						
	Training	Validation	Testing	Training	Validation	Testing
GG-ANN	0.0073	0.0146	0.0196	0.8066	0.6747	0.6337
SL-ANN	0.0091	0.0155	0.0186	0.7582	0.6539	0.6501
GG-Linear	0.0321		0.0363	0.1873		0.3181
SL-Linear	0.0297		0.0310	0.2501		0.4167

Performance table under EU rBWD classification

Gamma-GA-ANN models

		Observed: Not poor	Observed: Poor	
Predicted	Not poor	146	11	93%
Predicted	Poor	14	18	56%
		91% **Specificity**	62% **Sensitivity**	87%

SL-ANN models

		Observed: Not poor	Observed: Poor	
Predicted	Not poor	155	22	88%
Predicted	Poor	5	7	58%
		97% **Specificity**	24% **Sensitivity**	86%

Enterococci, Realization 1, testing set

Gamma-GA-ANN models

		Observed: Not poor	Observed: Poor	
Predicted	Not poor	170	6	97%
Predicted	Poor	8	5	38%
		96% **Specificity**	45% **Sensitivity**	93%

SL-ANN models

		Observed: Not poor	Observed: Poor	
Predicted	Not poor	172	8	96%
Predicted	Poor	6	3	33%
		97% **Specificity**	27% **Sensitivity**	93%

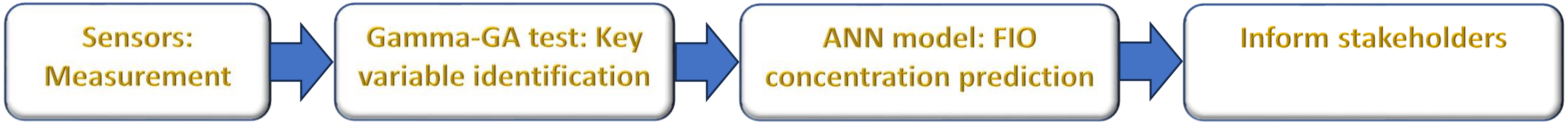
E Coli, Realization 3, testing set

- GG-ANN model improved the sensitivity for FIOs
  - Consistent with previous literature that nonlinear model captures better the extreme values

Specificity: % of NOT false alarm
Sensitivity: % of poor water quality identified
Overall accuracy: % of correct identification

Conclusion

- A data-driven GG-ANN model has been developed for FIO concentration prediction without unnecessary input variables



- GG-ANN model performance was evaluated at Swansea Bay, UK
  - Better predicted *Enterococci* for all three sets and most of the training sets for *E Coli*
  - Better in identifying events of poor water quality
  - Suitable for bathing water warning applications

https://www.tourismforall.co.uk/news/read/2019/07/swansea-beach-adds-more-accessible-facilities-b77

Thank you for listening

- Contact
  - M. Y. Arthur Lam: lamM7@cardiff.ac.uk
  - R. Ahmadian: AhmadianR@cardiff.ac.uk