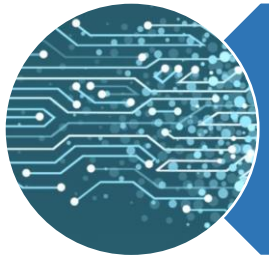
Literature Summary of Estuarine and Living Resources Studies Involving AI/ML

– *Jian Shen (VIMS) and Stephanie Schollaert Uz (NASA GSFC)*

1. Outlines



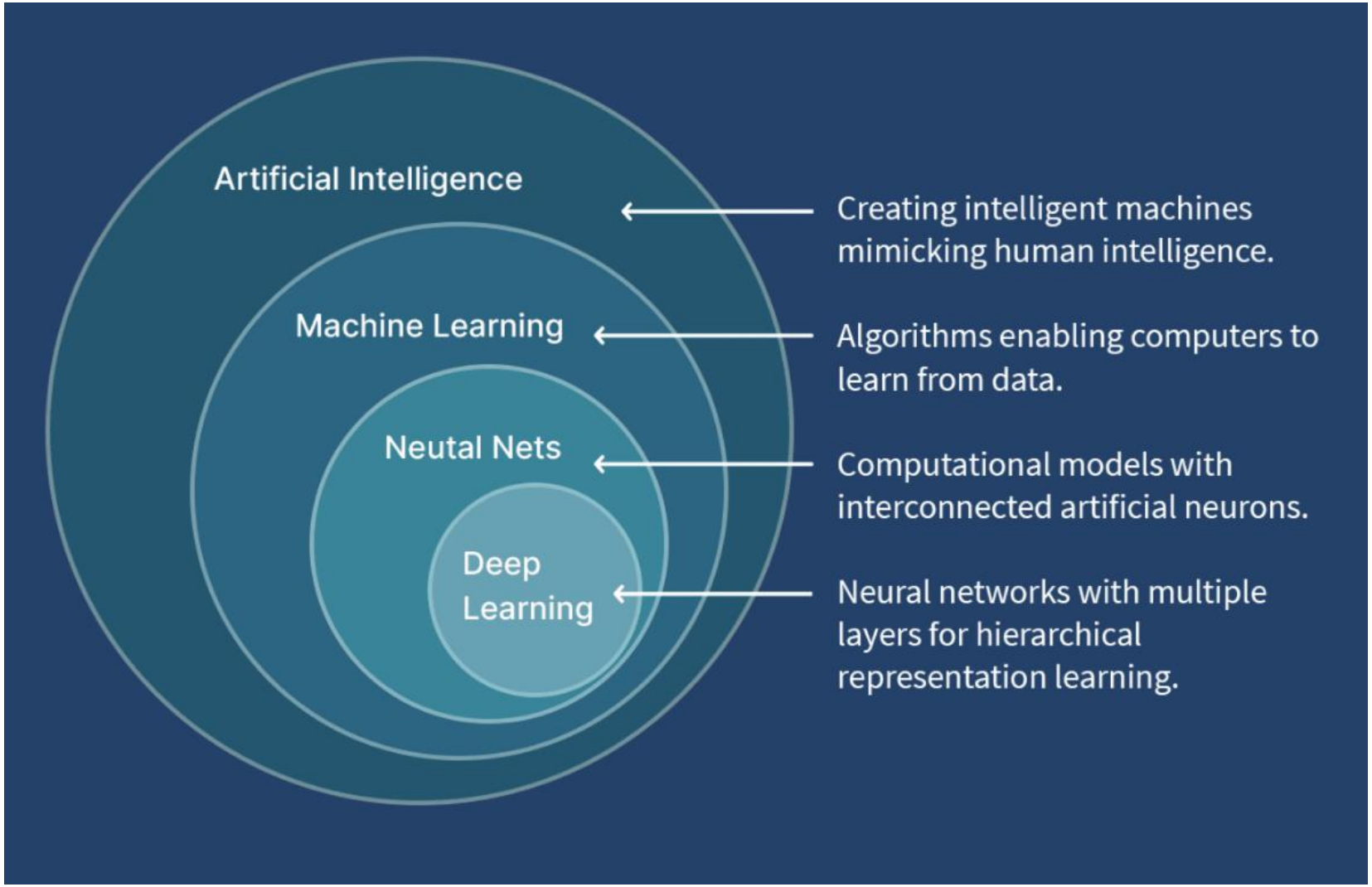
Methods Applied in the Bay

Applications

(Timeseries predictions)

Future Applications

2. Methods Applied in Chesapeake Bay



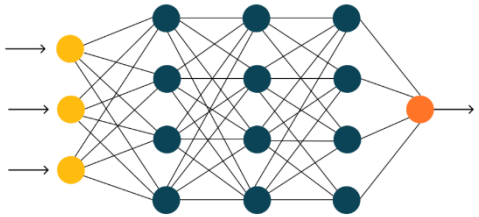
2. Methods

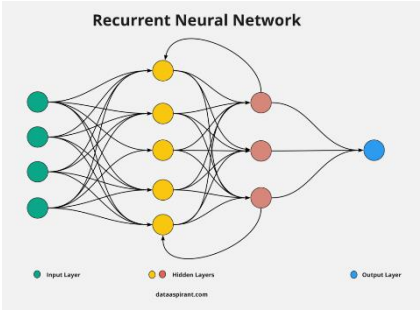
Supervise / unsupervised Learning

Regression, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forest (RF), Gradient Boosting (XGBoost, LightGBM, CatBoost), etc.

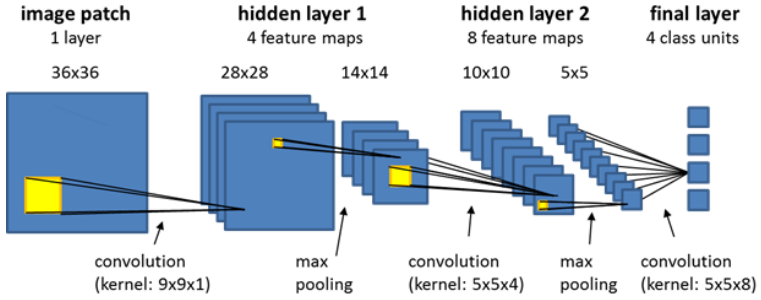
k-Means Clustering, Hierarchical Clustering, PCA (Principal Component Analysis)

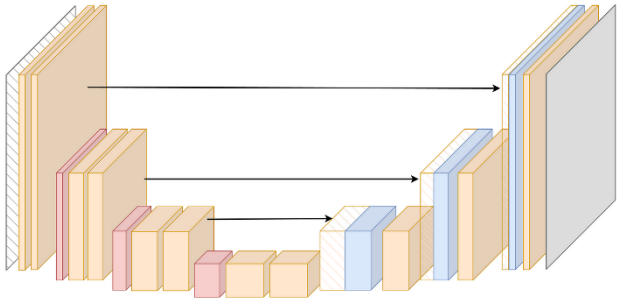
Neural network

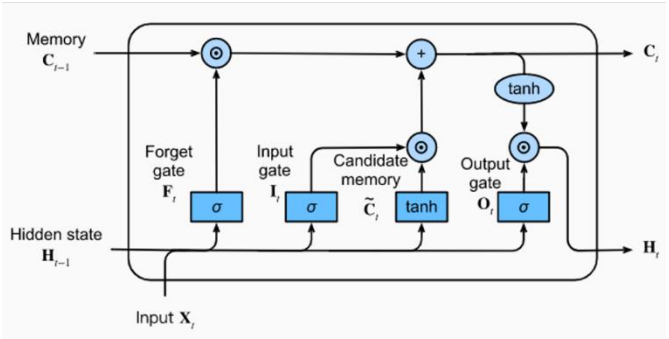




Deep leaning





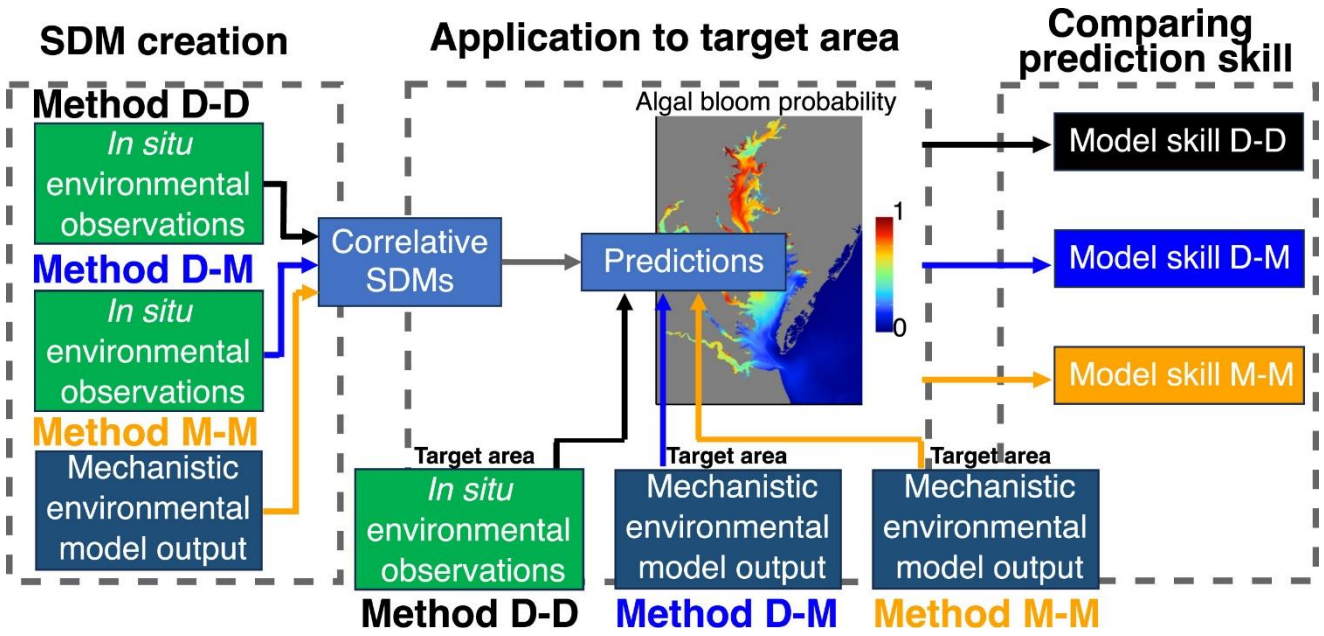


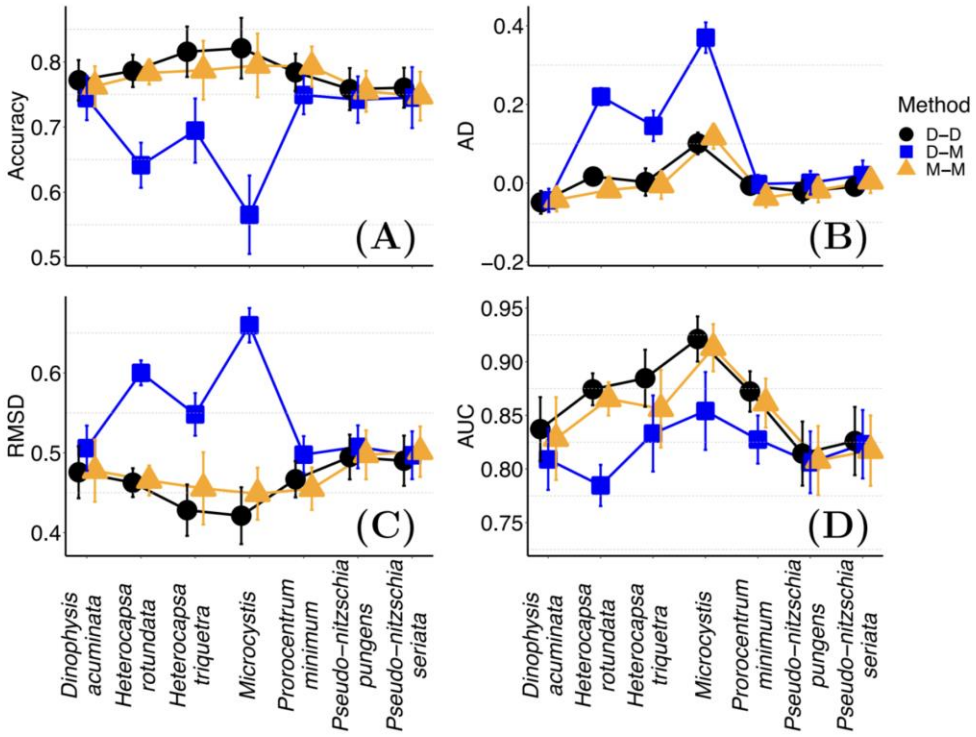
2. Applications: Ecosystem

- Species Distribution Model (Horemans et al., 2024, Ecological modeling)
  - Predict Species distribution based on environmental variables
    - Using numerical models (biases)
    - Using observation. (limited)

Water temperature (T)
Salinity (S)
Vertical gradient of salinity (gradS)
Apparent oxygen utilization (AOU)
pH
Dissolved inorganic nitrogen (DIN)
Total organic nitrogen (TON)
Solar irradiance at the water surface (swrad)[a]
Total water depth

Train SDMs using environmental mechanistic model outputs to improve prediction skill (generalized linear models)





2. Applications: Ecosystem

- Primary production (Scardi, 1996, Marine Ecology Progress Series)
  - Use environmental data to estimate primary production use NN model
  - $$PP = a + b \cdot BZ_p I_0$$

$$Ls(t) = \widetilde{Ls}(t,p) + Er(t),$$ → NN

Table 1 The data set from Harding et al. (1986) includes primary production (*PP*), surface irradiance (I_0), mean chlorophyll concentration (*B*), depth of photic zone (Z_p), light extinction coefficient (k_t) and station depth (*H*). The binary variable in the last column (*Bay*) represents the location of each station [Chesapeake (CB) or Delaware Bay (DB)] and was added in order to improve the performance of the artificial neural network. In Harding et al. (1986) the same result was achieved by calculating a different slope of the linear empirical model for each bay

Cruise	Stn	*PP* (g C m^{-2} d^{-1})	I_0 (E m^{-2} d^{-1})	*B* (mg chl m^{-3})	Z_p (m)	k_t (m^{-1})	*H* (m)	*Bay* (1 = CB, 0 = DB)
CB-1	I	0.101	28.0	1.3	2.6	1.80	6.0	1
Mar 82	II	0.877	3.1	5.6	6.0	0.77	7.5	1
	III	0.372	3.3	4.7	12.0	0.30	12.0	1
	IV	1.430	36.0	7.3	12.0	0.24	12.0	1
	V	0.317	33.0	4.2	11.0	0.15	11.0	1
CB-2	IV	0.476	16.0	13.0	1.7	2.80	5.0	1
Jul 82	III	1.780	58.0	9.2	3.4	1.40	7.0	1
	II	2.650	55.0	9.9	4.6	1.00	9.0	1
	I	0.611	52.0	4.7	8.1	0.57	11.0	1
CB-3	V	0.804	26.0	13.0	2.4	1.90	5.0	1
Oct 82	IV	0.954	18.0	10.0	4.7	0.98	8.0	1
	III	0.962	21.0	6.6	6.1	0.75	11.0	1
	II	0.603	20.0	3.7	9.5	0.49	12.0	1
	I	0.770	19.0	3.2	8.6	0.54	15.0	1
CB-4	V	0.089	36.0	3.7	1.3	3.40	6.0	1
Mar 83	IV	0.638	18.0	8.7	3.7	1.20	8.0	1
	III	1.910	41.0	12.0	4.6	1.00	9.0	1
	II	0.140	11.0	4.3	7.5	0.61	13.0	1
	I	0.369	34.0	3.9	6.6	0.70	12.0	1
DB-1	III	0.564	18.0	9.3	2.3	2.00	11.0	0
Nov 82	IV	0.110	10.0	6.2	1.3	2.70	12.0	0
	II	0.868	20.0	21.0	4.6	1.00	8.0	0
	I	0.323	24.0	2.9	5.3	0.86	25.0	0
DB-2	IV	0.021	6.6	5.5	1.1	4.30	8.0	0
Apr 83	III	0.171	12.0	7.8	1.8	2.50	10.0	0
	II	0.758	17.0	20.0	3.9	1.20	15.0	0
	I	0.326	23.0	6.4	6.1	0.68	19.0	0

Harding, et al. 1986

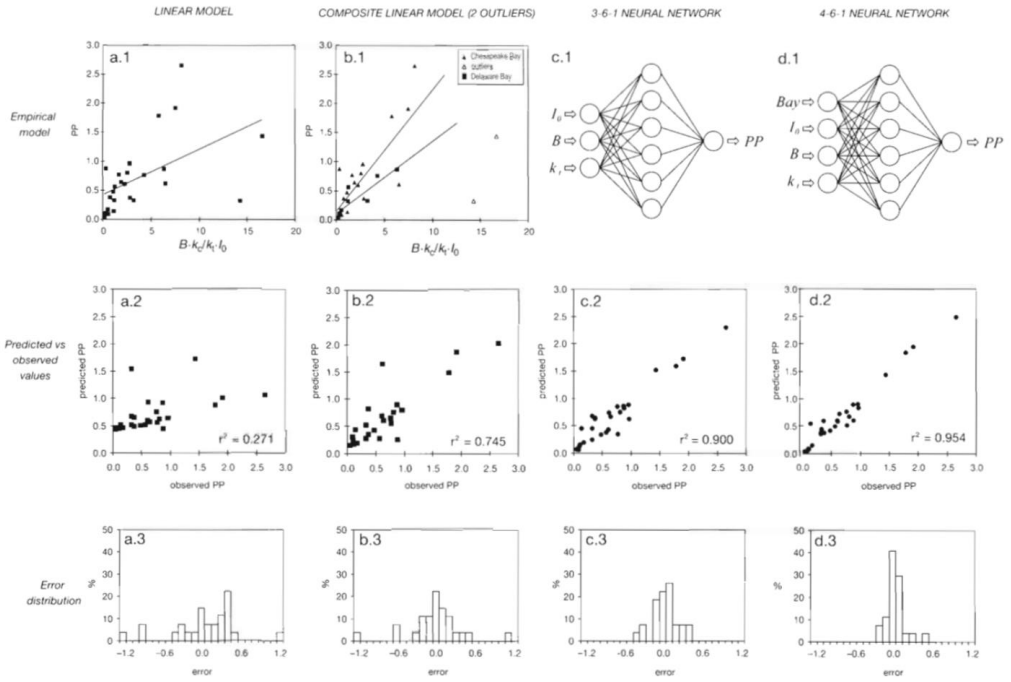
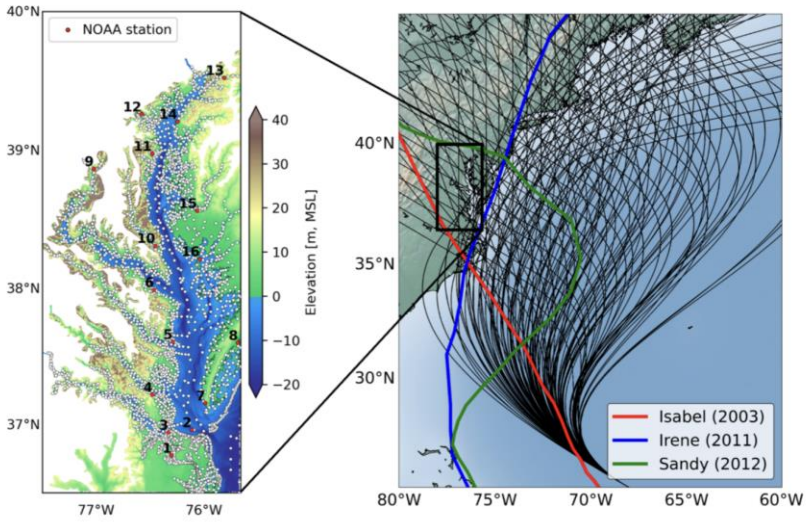


Fig. 1. Comparison between different empirical models of phytoplankton production (*PP*, g C m^{-2} d^{-1}). Models are based on surface irradiance (I_0, E m^{-2} d^{-1}), biomass (*B*, mg chl m^{-3}), light extinction coefficient (k_t, m^{-1}), light absorption by chlorophyll [k_c = 0.015 m^{-1}(mg chl m^{-3})$^{-1}$]. The second and fourth model (b & d) also take into account station location (*Bay*, Chesapeake Bay = 1 and Delaware Bay = 0). The neural network structures are simplified, as bias neurons are not shown. In the error distribution histograms, labels indicate the upper limit of each class

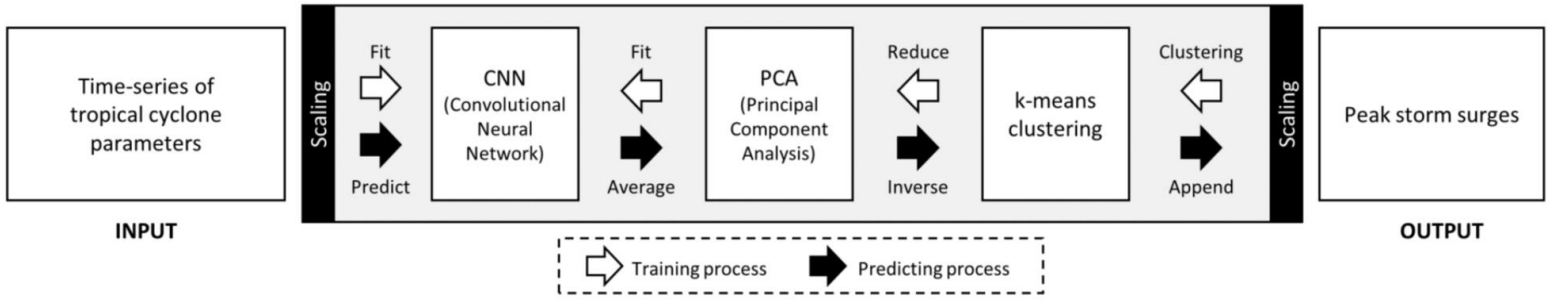
2. Applications: Storm surge

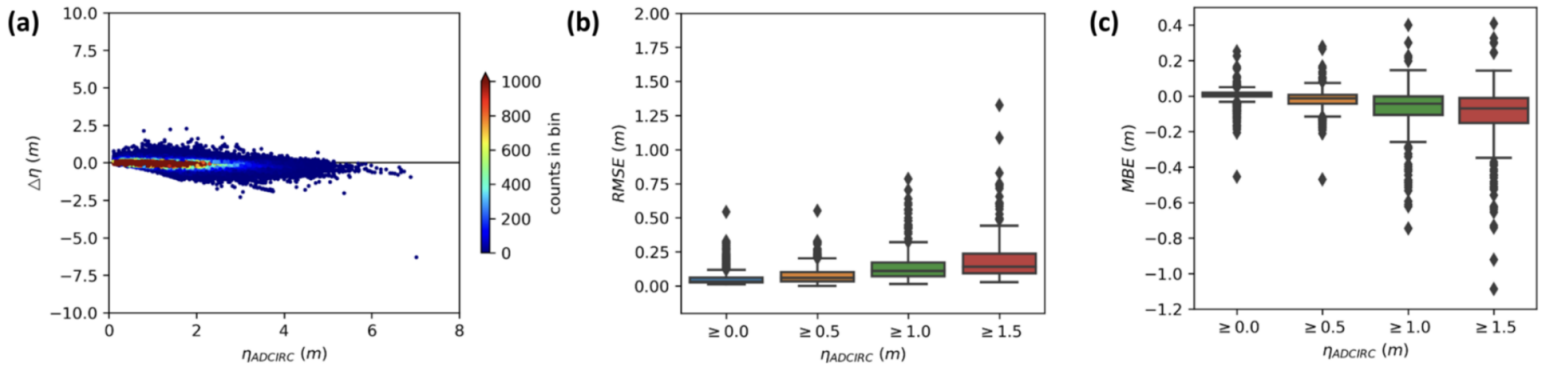
Peak storm surge (Lee, et al. 2024, Coastal Engineering)

Using storm surge simulations for 1050 synthetic tropical cyclones (TCs) in the US North Atlantic region (ADCIRC model)

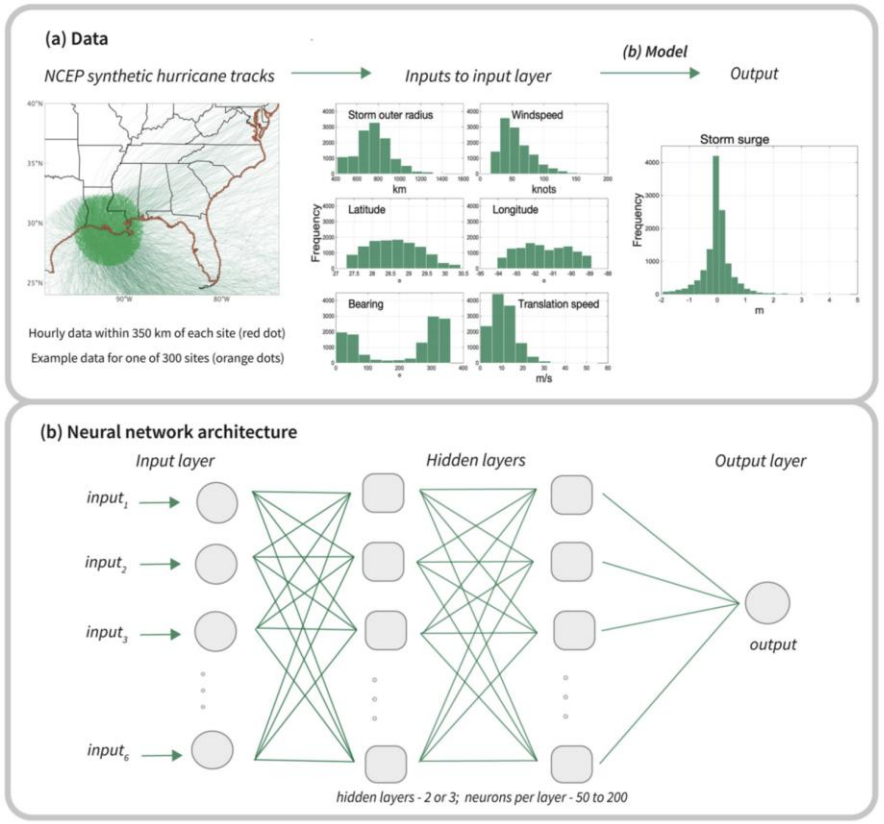


TC parameters (latitude LAT , longitude LON, heading direction ✓, central pressure Cp, radius of maximum winds Rmax, and translation speed Vf)





2. Applications: Storm surge



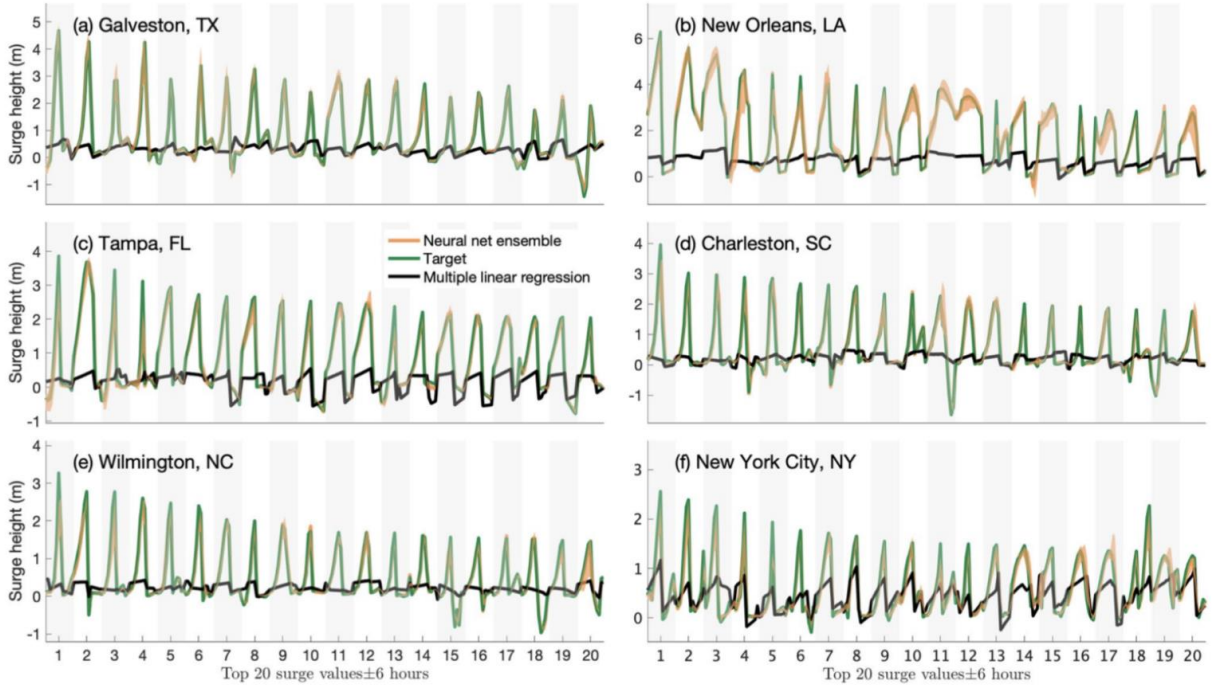
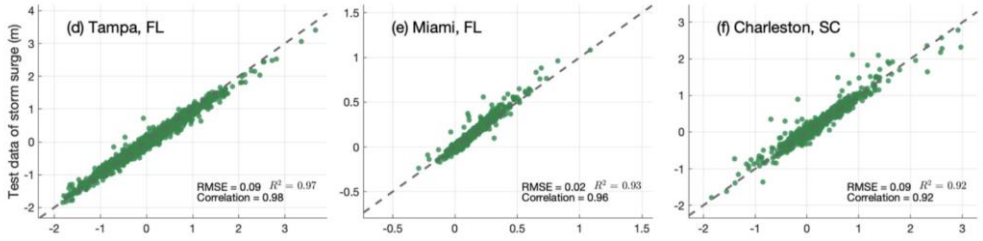


Figure 4. (a–f) Six hours before and after the 20 events with the largest surge heights at each sites. Green lines shows the target timeseries and orange colors denote the ANN models ensemble average (line) and range (shading) of predictions. The black lines show the multiple linear regression ensemble calculated using the same data input to the ANN models.

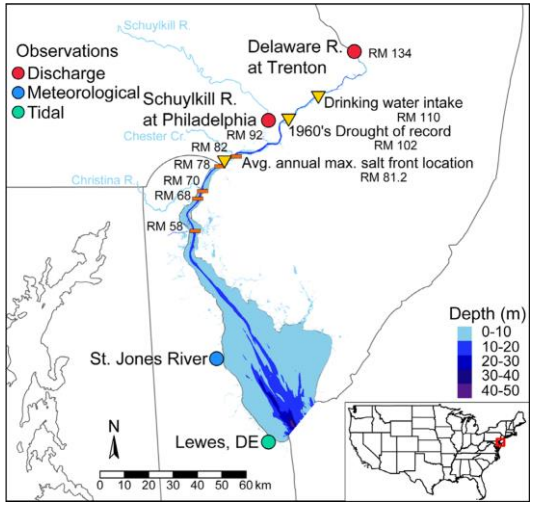
Use hourly wind generated by Holland (1980) wind model and train against ADCARC output for different track

Lockwood et al., JGR, 2022

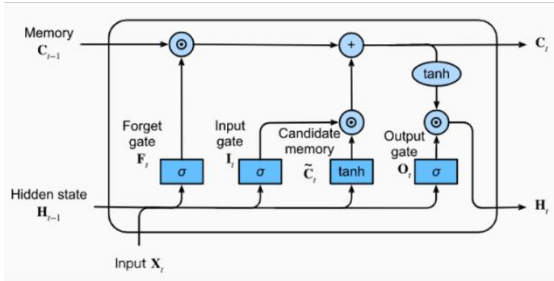


2. Applications: Saltwater intrusion

- 7-day averaged Saltwater intrusion (Gorski, et al. 2024. L&O)

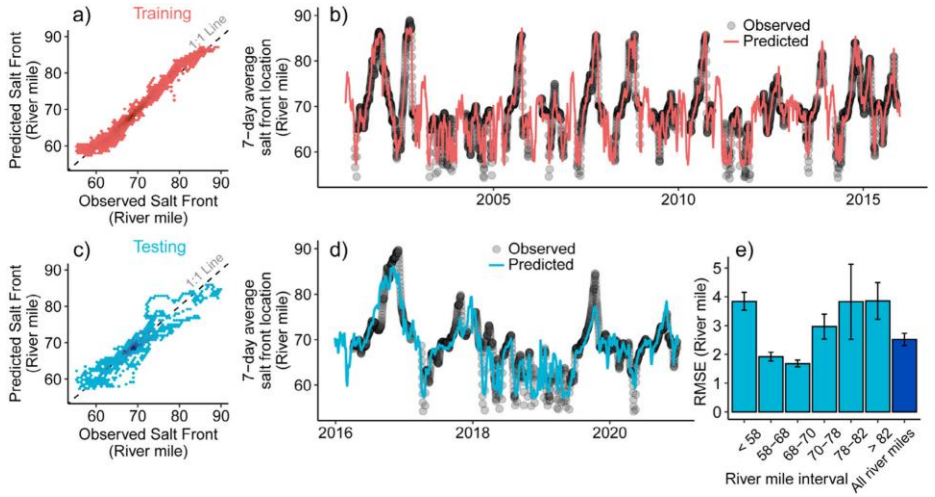


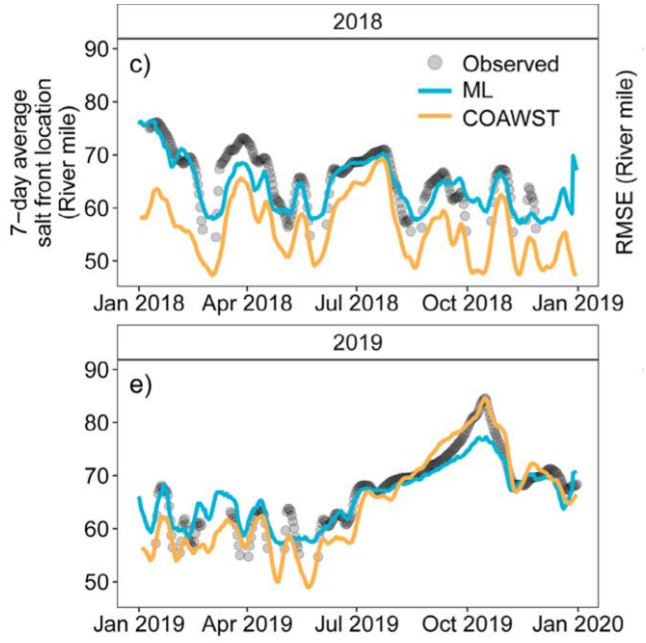
- Daily river discharge
- meteorological drivers
- tidal water level data
- Using past 365 days data



Observations for training

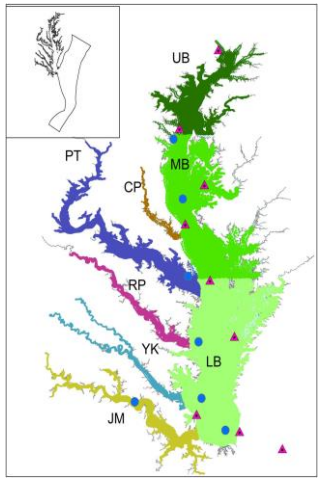
Good model performance
Better than 3D model





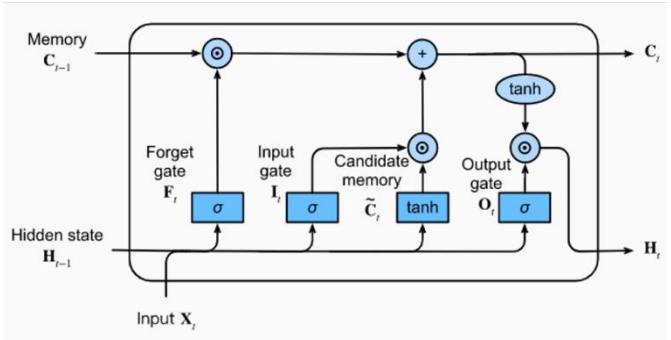
2. Applications: Wave

- Predict daily mean/maximum significant wave based on wind (Shen et al. 2024)

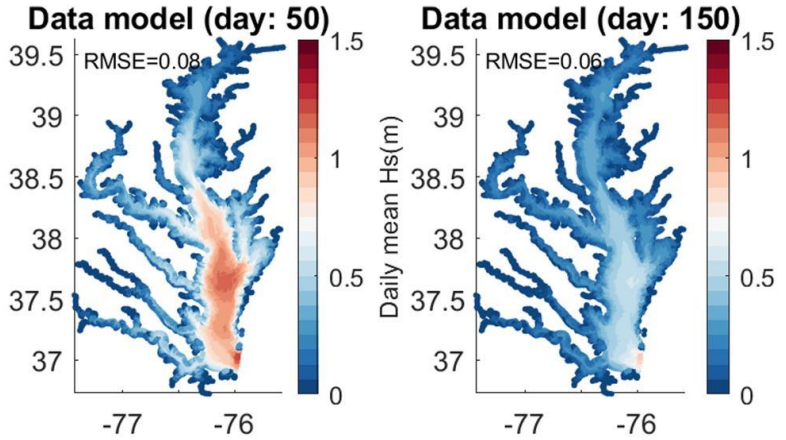


Daily Wind
9 stations

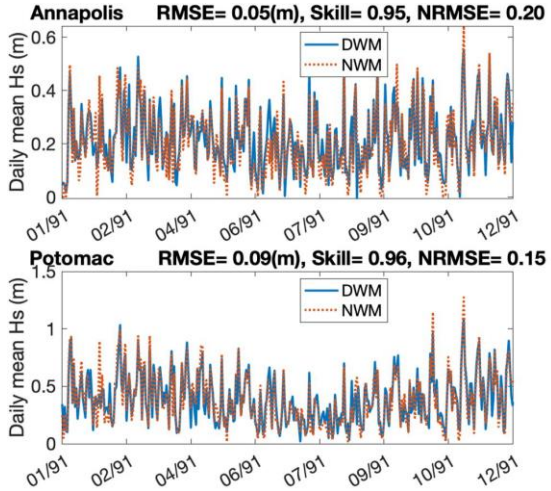
1991-1995



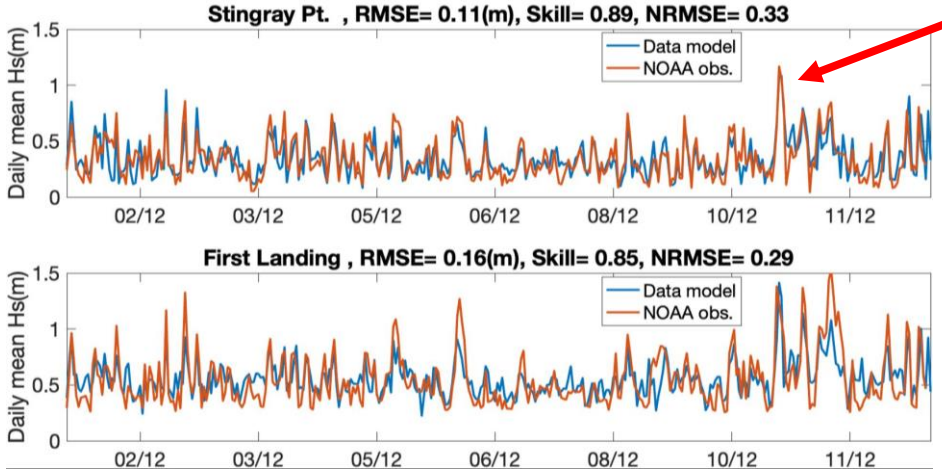
Trained by SCHISM model simulated wave



SCHISM



NOAA data



Hurricane Sandy (October 2012)

Advantage

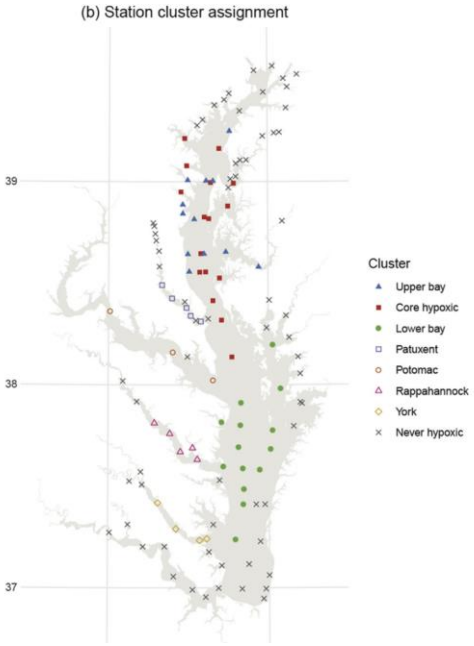
- Only inputs wind data
- It can conveniently do forward and backward simulations of wave
- Easy to access climate change

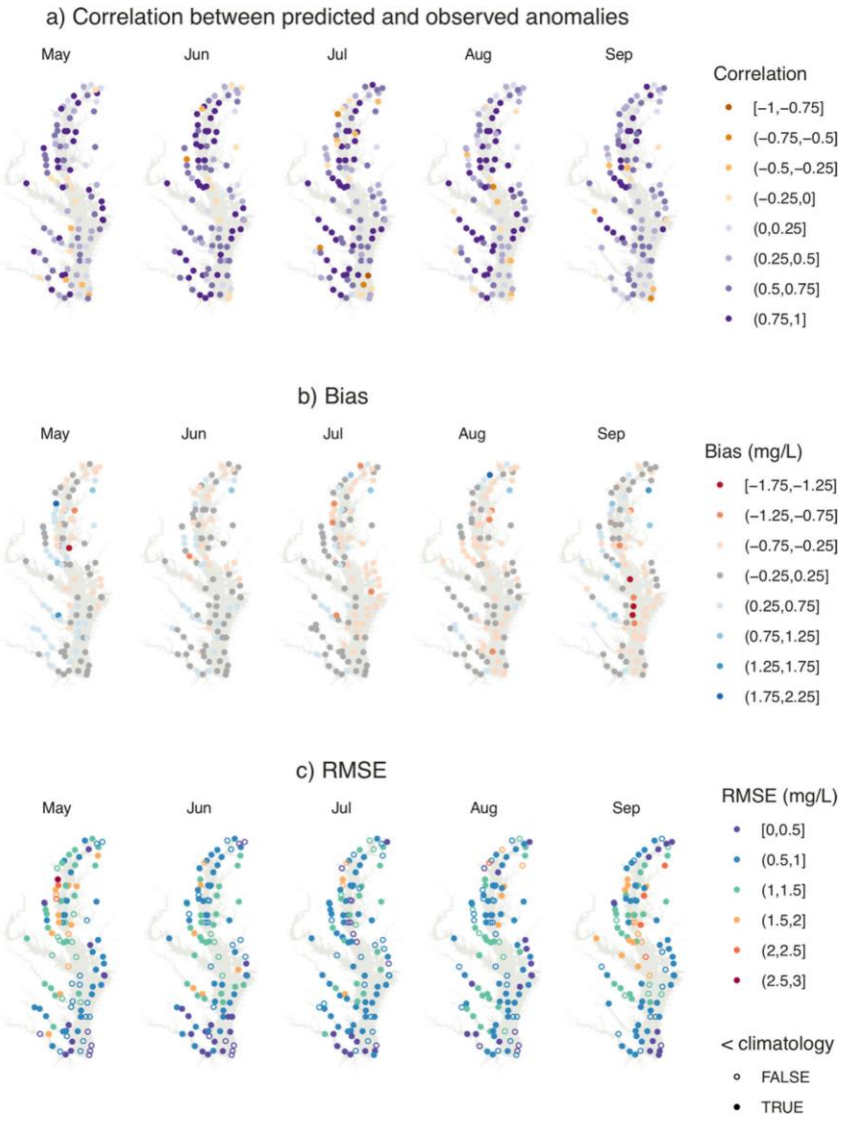
2. Applications: DO

- Column minimum DO (Rossa and Stock, ECSS 2019)
- Model tree (Cunist package for R)

Table 1
Variables used as inputs to the mechanistic dissolved oxygen model.

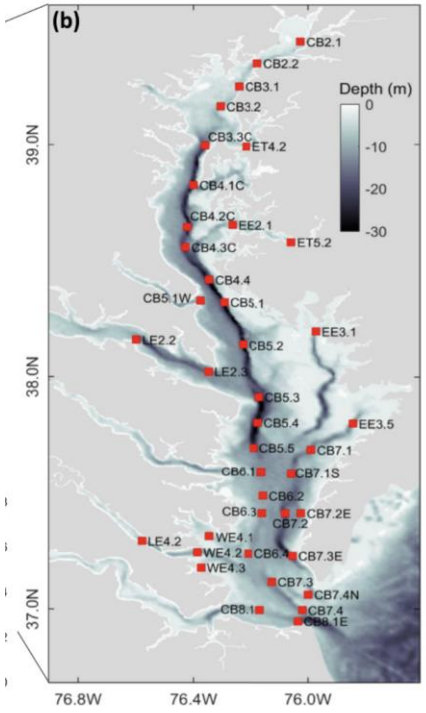
Abbreviation	Input variable	Data source
L5	TN load from Susq. River, total over previous 5 months	USGS
W_{spring}	Mean wind along NE/SW axis, Feb–Apr	NDBC
$\bar{T}$	Column-mean temperature anomaly, forecast month	CBP
MSL	Mean sea level anomaly, forecast month	PSMSL
$\Delta\rho$	Vertical density difference anomaly, forecast month	CBP
M	Forecast month	
H	Forecast hour	
D	Profile bottom depth	CBP
X	Longitude	CBP
Y	Latitude	CBP

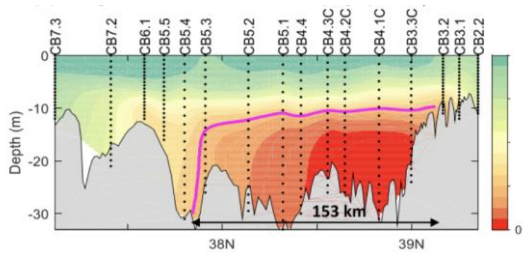


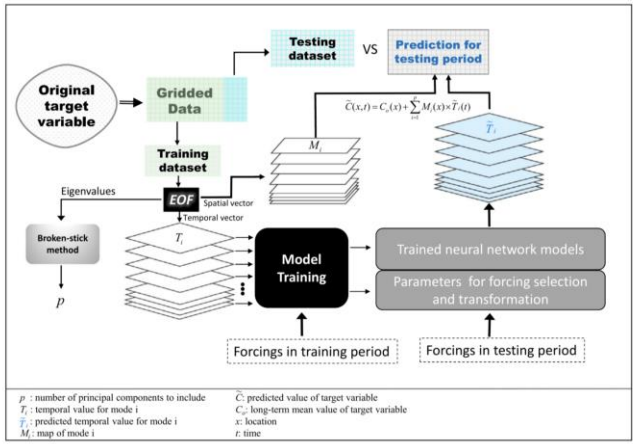


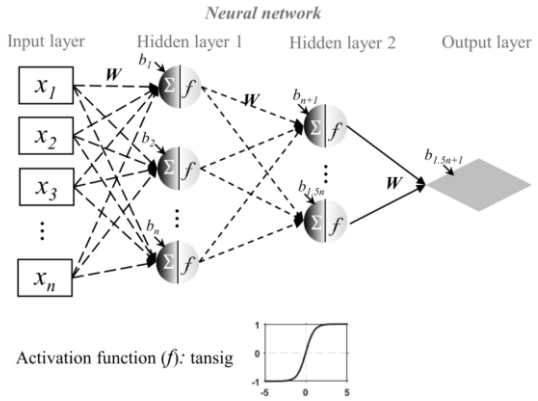
2. Applications: DO

- Prediction of DO (Xu et. 2020, Water Resource Research









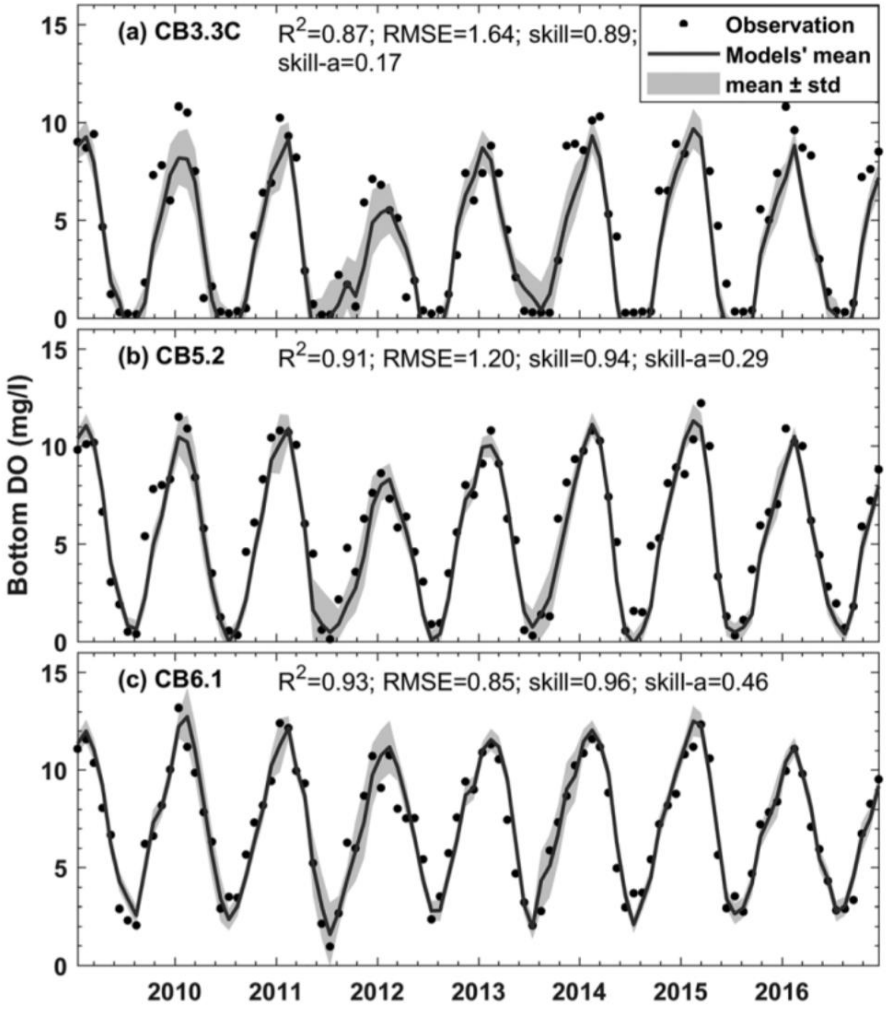
- Drive by all external forcing: flow, N,P loadings, wind, air temperature, heat flux
- Use parameter transformation
- Be able to be used for management

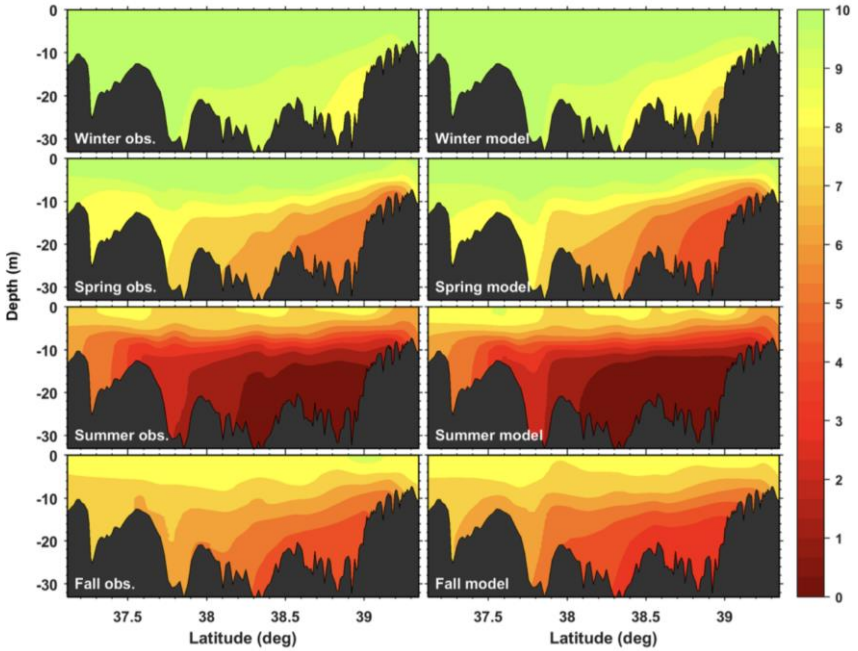
Table 1
A List of the Transformation Options in the Data-Driven Model

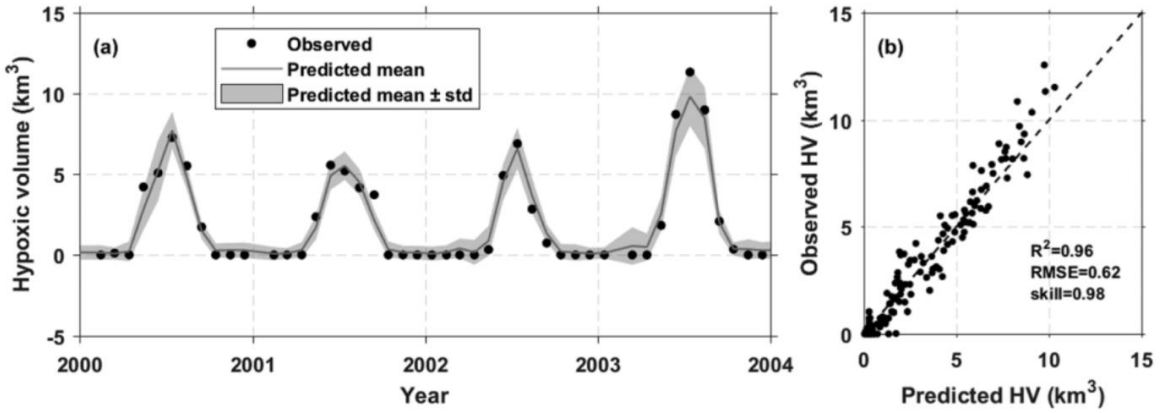
Transformation	Subtypes	Formula
Time-lag transformation	1–7	$\phi(t) = x(t\text{-lag})$, with lag ranging within 0, 10, ... 60 days
Accumulative transformation	1–13	$\phi(t) = \text{mean}(x(\tau))$, where $\tau \in [t1 - acc, t2]$, with acc ranging from 0 to 120. *t1* and *t2* are the beginning and end of each month; t1 = t − 15 and t2 = t + 15
Regular transformation	1	$\phi = x$
	2	$\phi = \log(x)$
	3	$\phi = 1/x$
	4	$\phi = \exp((x\text{-mean}(x))/\text{std}(x))$
	5	$\phi = x/(p50 + x)$, also known as Monod-type filter
	6	$\phi = x/(p75 + x)$
	7	$\phi = x/(p25 + x)$
	8	$\phi = (x\text{-mean}(x))/\text{std}(x)$

Note. x = forcing variable; ϕ = transformed forcing variable; t = time; std(x) = the standard deviation of x; mean(x) = the mean value of x; P25, P50, P75 = the 25, 50, and 75 percentile of x.

2. Applications: DO

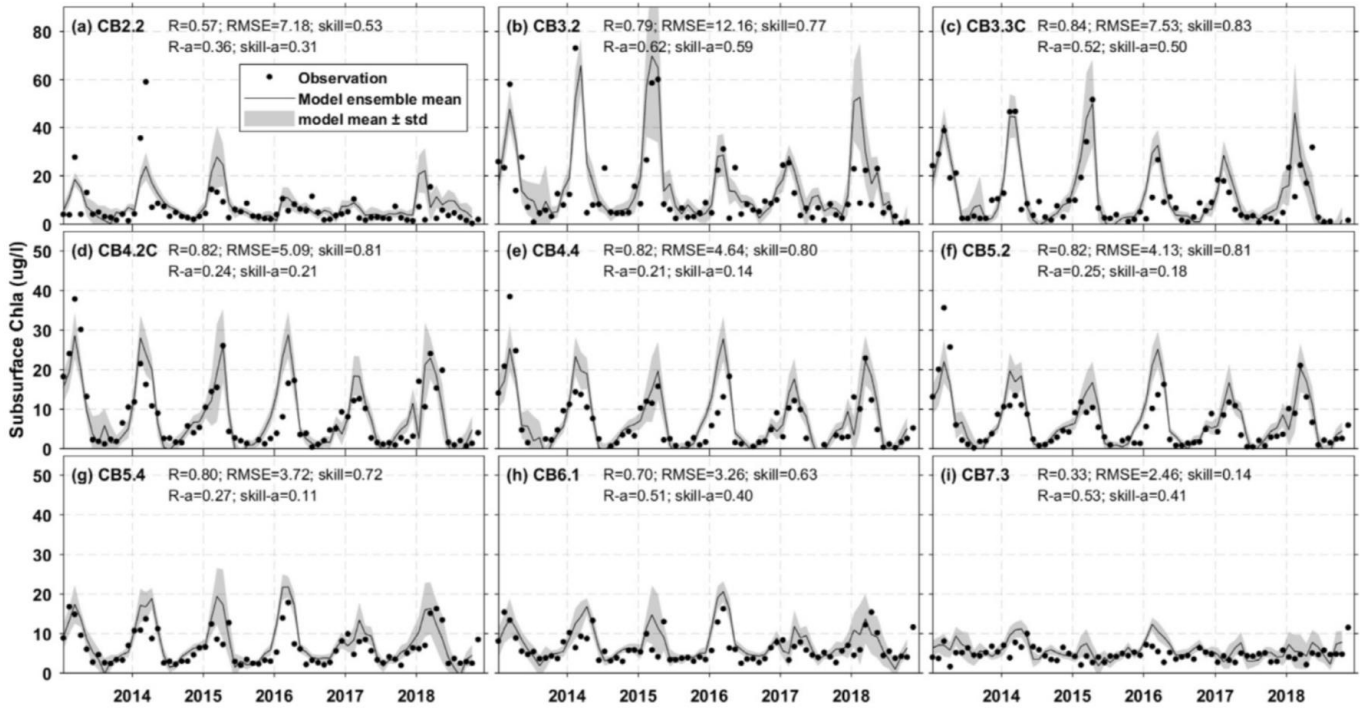


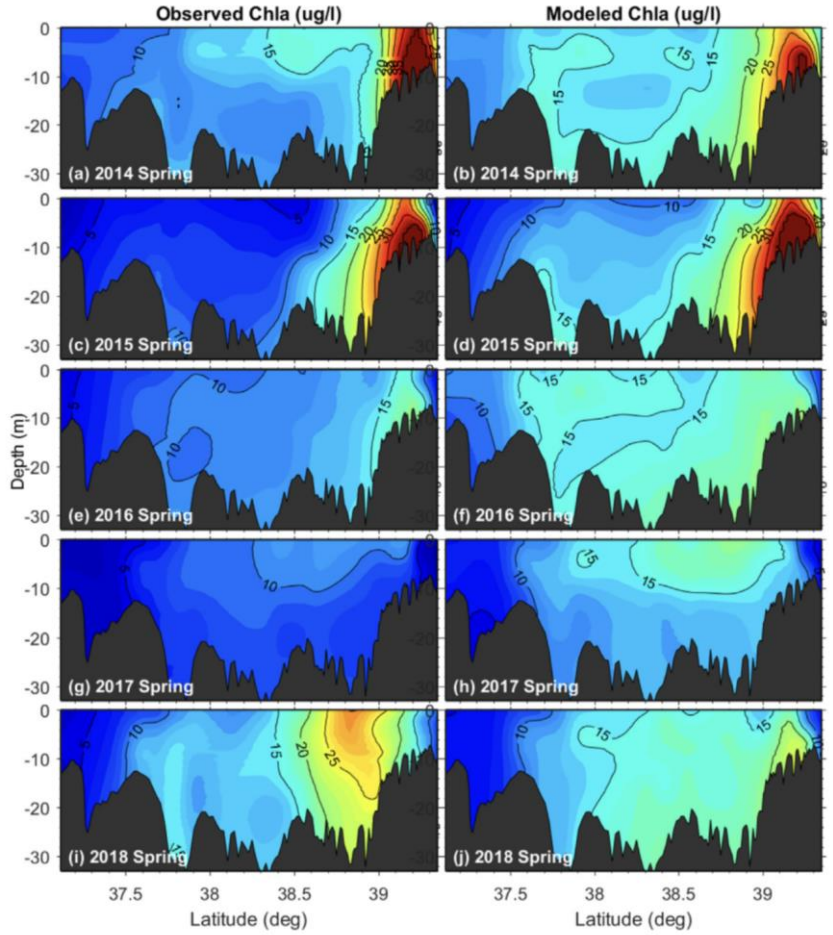




2. Applications: Phytoplankton

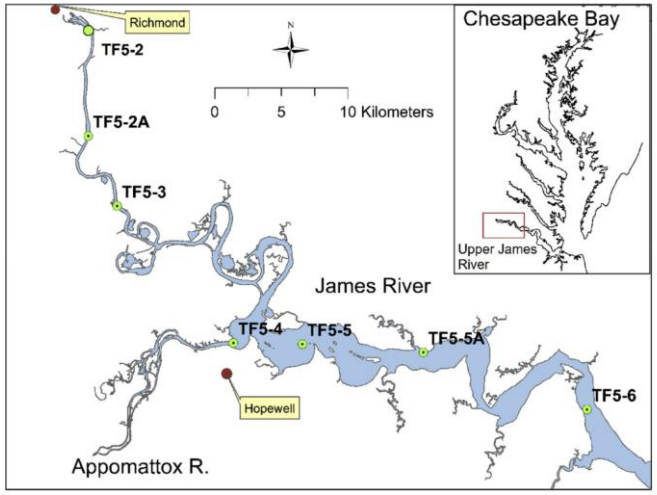
- Predict phytoplankton (Xu and Shen, 2021. Ocean Modeling

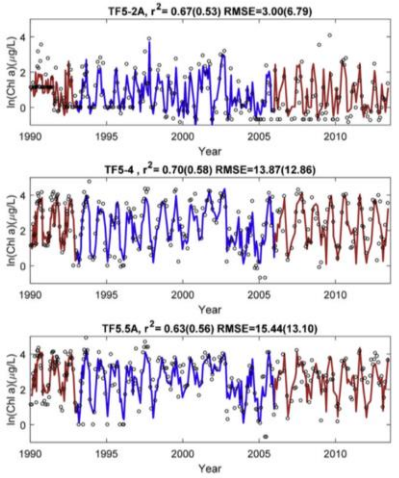


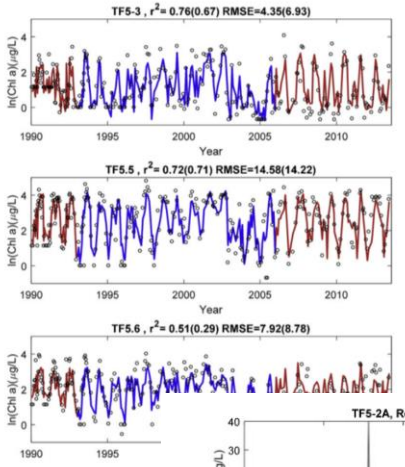


2. Applications: Phytoplankton

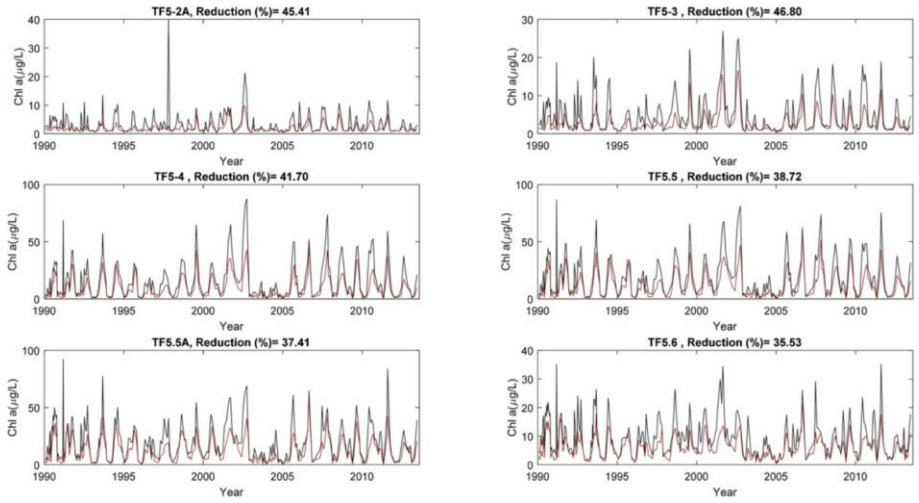
- Can we use ML model for management?
- Predicate phytoplankton (Shen et al., 2019, Ecological modeling







Management scenario:
Reduce nutriment by 50%



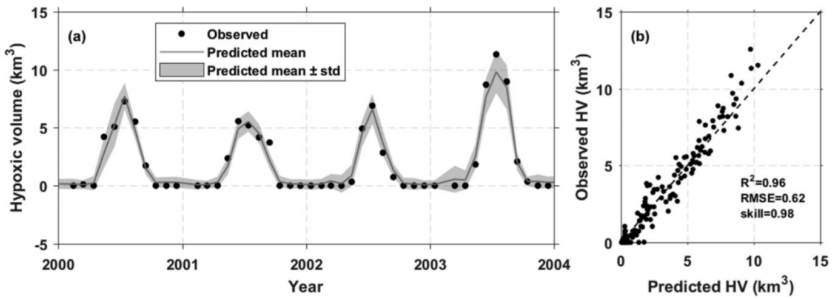
- Input parameters: watershed model outputs (flow, nutrients, temperature)
- Support vector machine LS-SVM (project to high dimension)
- Parameter transformation
- Without use temperature as an independent variable

$$TN_{new} = \frac{TN}{H_{TN} + TN}\theta^{T-20}$$

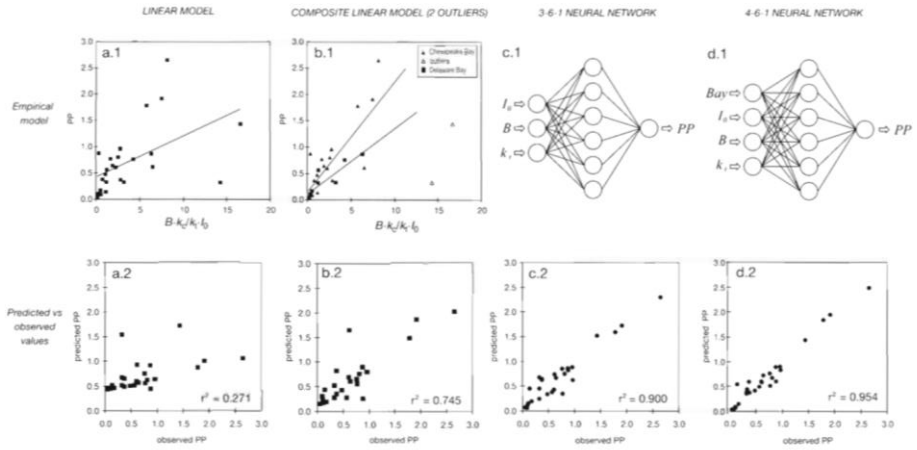
3. Future Applications: Predication/Management

- Using ML to predict future
- Applying it for management

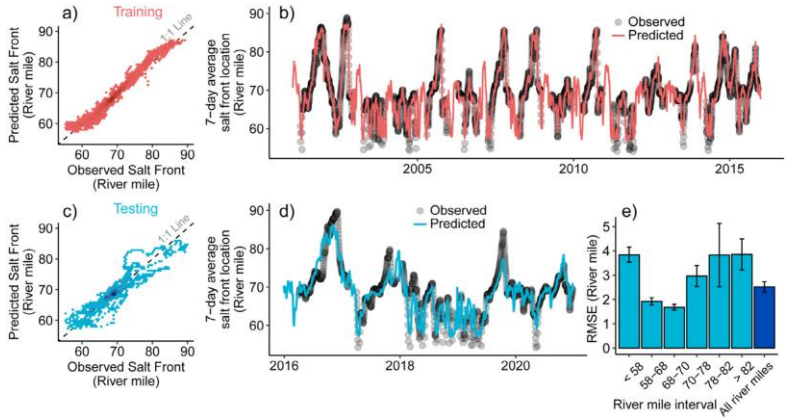
Hypoxia volume



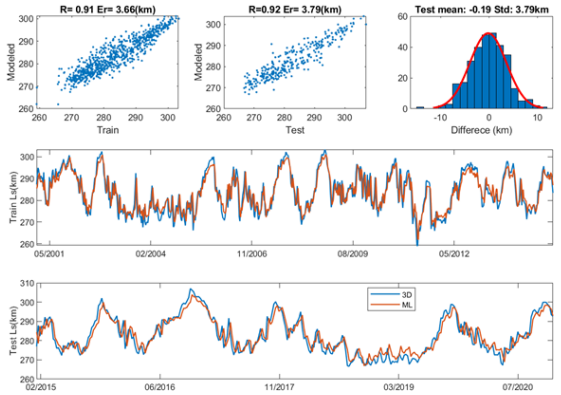
Ecosystem status



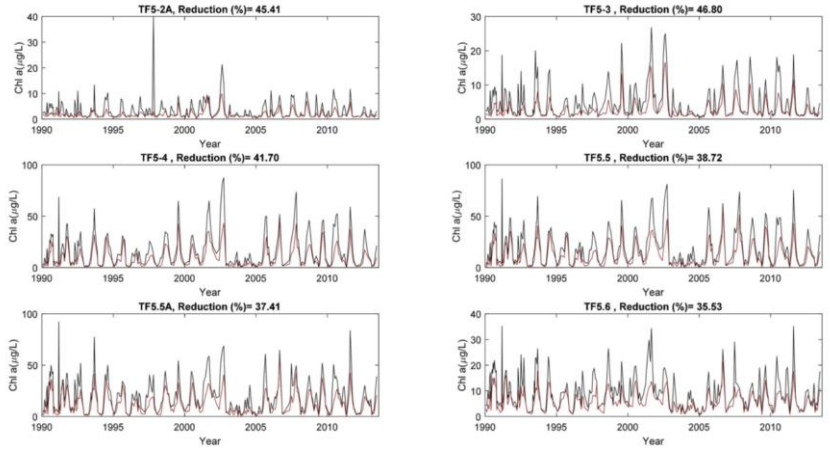
Saltwater intrusion



7-day forward prediction of saltwater intrusion of the Bay

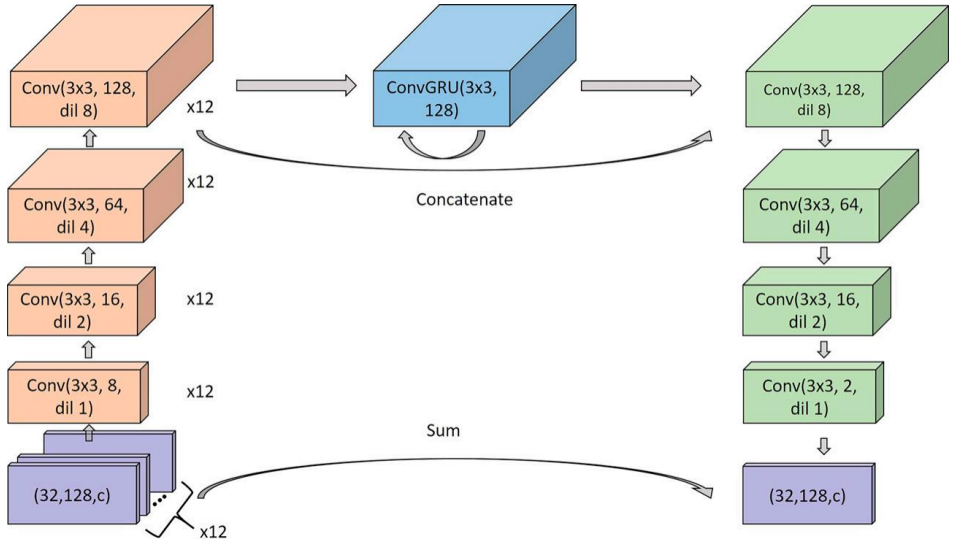


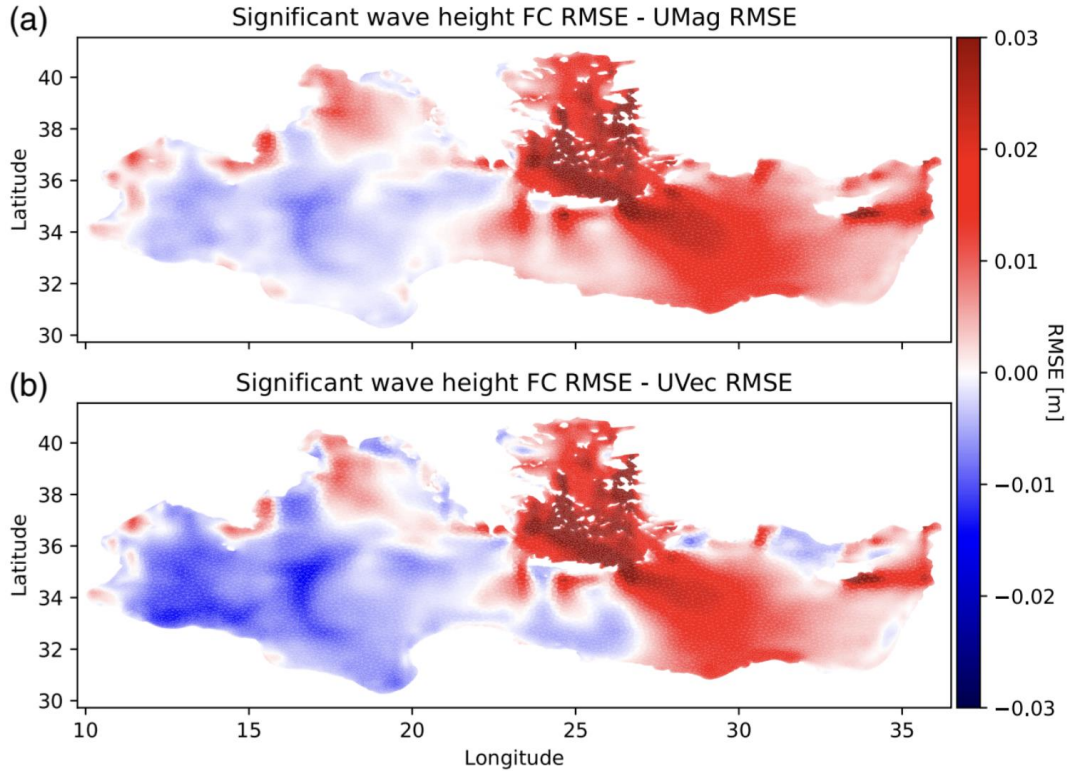
Nutrient reduction



3. Future Applications

- Using a data-driven deep learning model to improve wind forecasting accuracy, and improving wave forecasting (Yevnin & Tiked, 2022, PO)
- Can be used to link observations and numerical model)



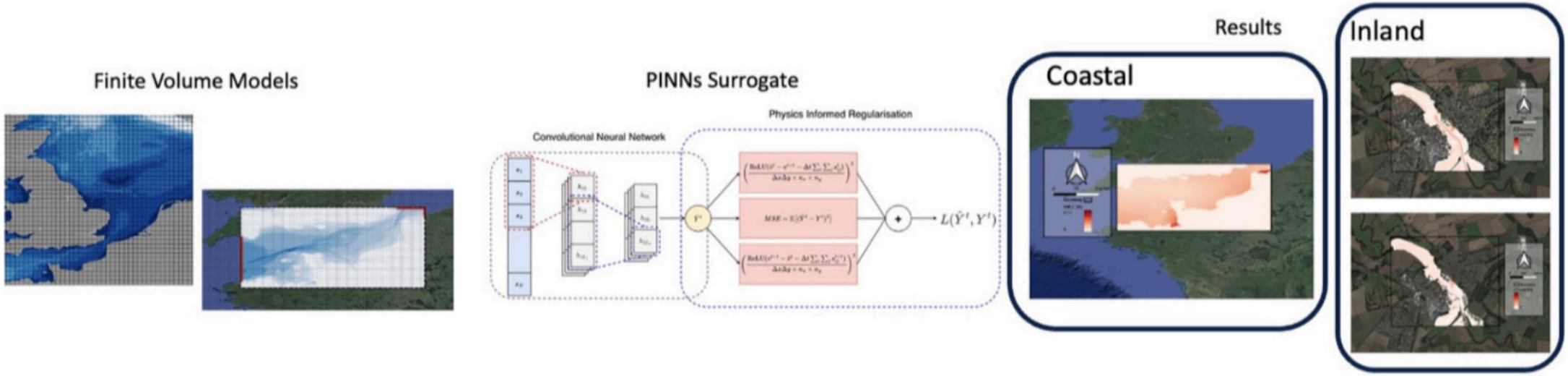


Applications:

Encode environmental; variables
Construction of hypoxia volume

3. Future Applications:
Physics-informed neural networks

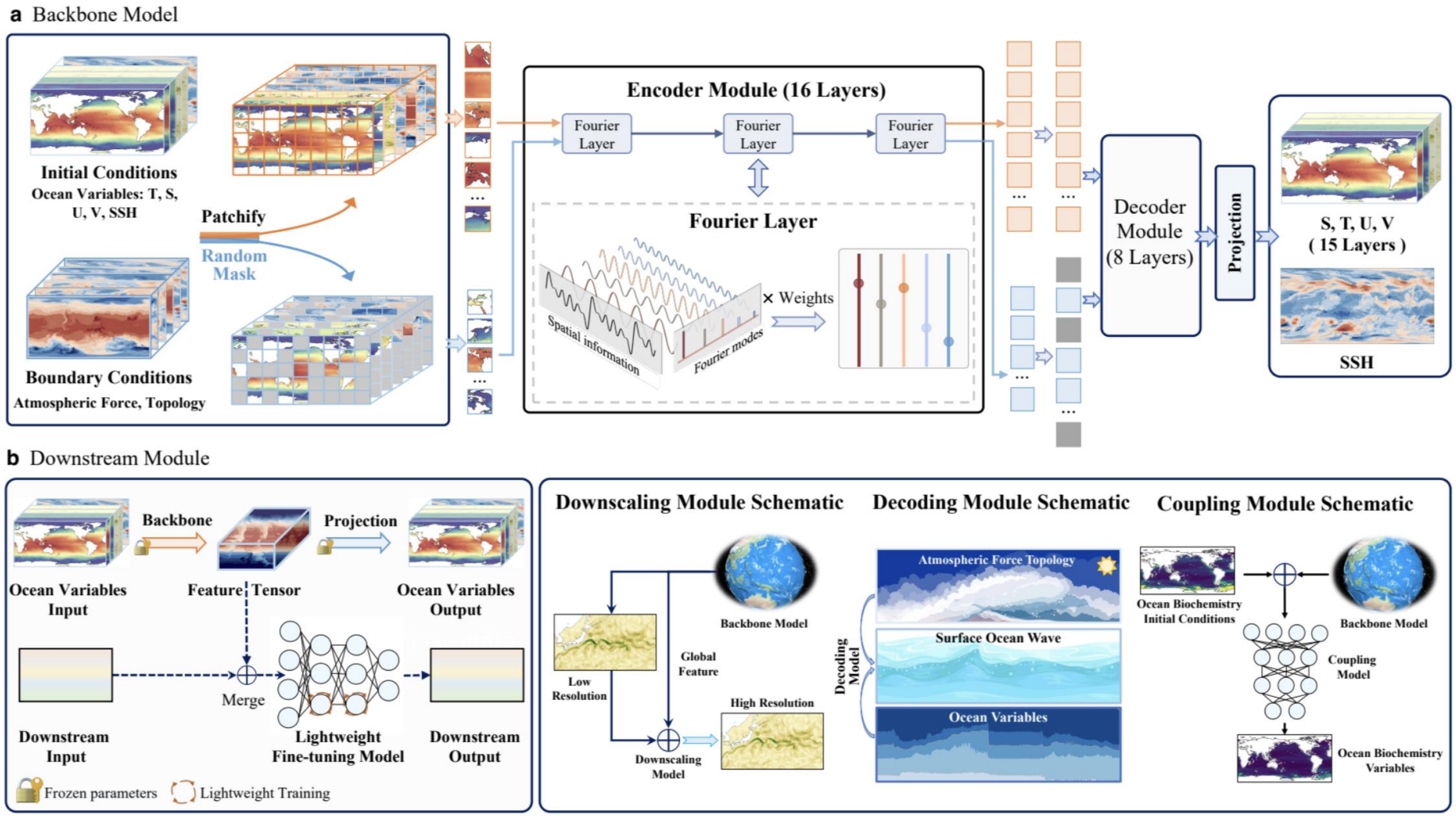
- One of the challenge of application of ML in estuary is high variations
- Physics-Informed Neural Network-based surrogate model for hydrodynamic simulators governed by Shallow Water Equations.
  Donnelly et al. Science of the Total Environment, 2024



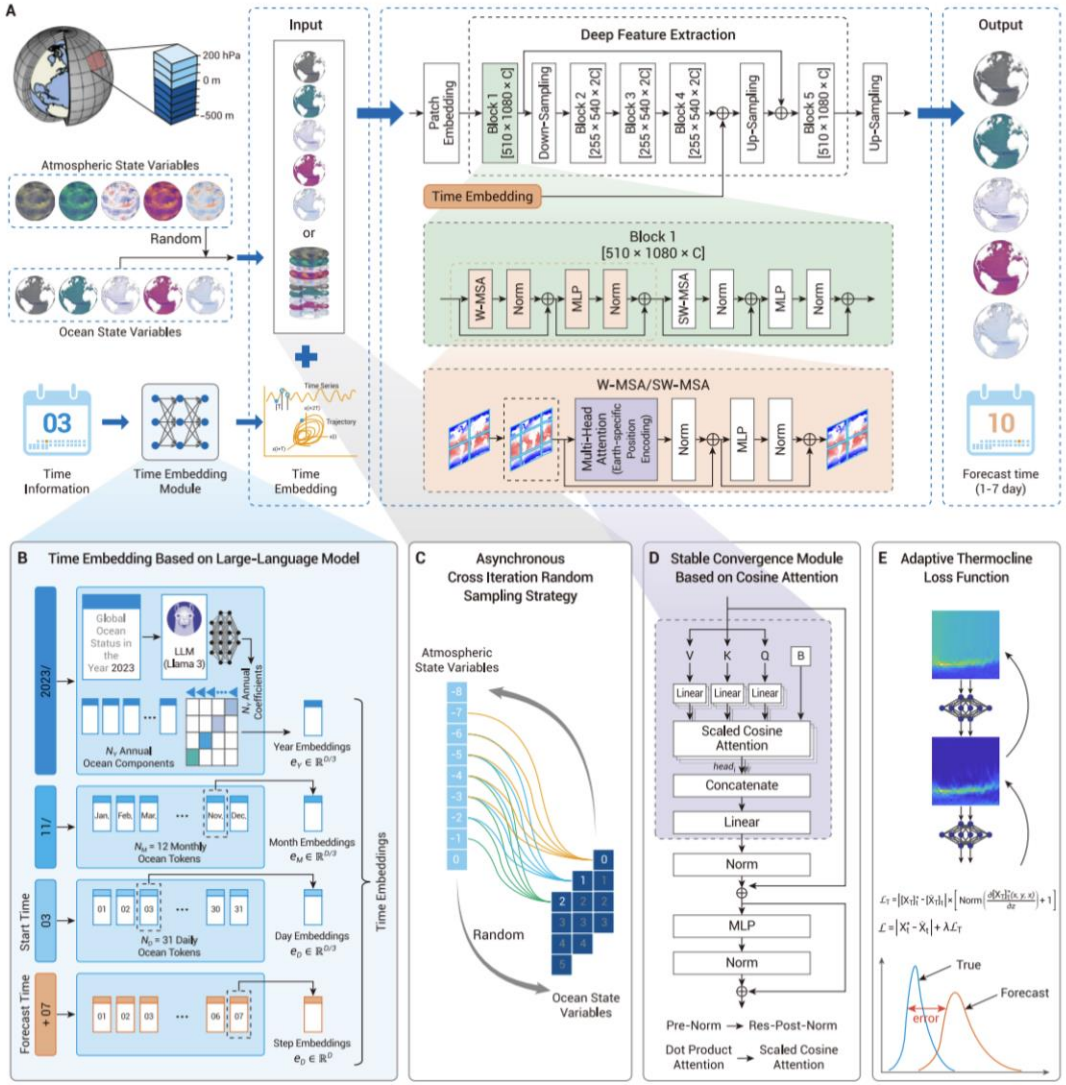
$$\nabla \cdot \boldsymbol{u} = \mathbf{0} \longrightarrow \Delta \boldsymbol{v} = \Delta(\boldsymbol{s} - \boldsymbol{q}) \longrightarrow \mathrm{ReLU}\left(\hat{v}^{(t)} - v^{(t+1)} - \Delta t \sum_i \sum_i s_{ij}^t\right) + ReLU\left(v^{(t+1)} - \hat{v}^{(t)} - \Delta t \sum_i \sum_i s_{ij}^t\right)$$

3. Future Application: Digital twin

- AI-GOMS model (Xiong et al., 2023, arXiv)



3. Future development: Digital twin



- (A) Overall system pipeline integrating Time Embedding and deep feature extraction
- (B) LLMbased Time Embedding module,
- (C) Asynchronous Cross-Iterative Random Sampling Strategy,
- (D)Ocean self-attention module based on cosine attention, and
- (E) Adaptive loss function for thermocline forecasts. W-MSA means window-based multi-head self-attention module.

Yang et al., 2024. arVix

Conclusions

- Machine learning (ML) methods have been utilized in the Bay for a variety of applications.
- Given the large amounts of observational data and numerous numerical models available, ML has significant potential for improving forecasting and management efforts in the Bay.