SYMPOSIUM OP'24 ADVANCING OCEAN PREDICTION SCIENCE FOR SOCIETAL BENEFITS

Scan QR to Join LIVE POSTAL PRESENTATION
Nov 20, 2024
12:15 pm – 1:45 pm
4:30 pm – 6:30 pm

An approach for good modelling and forecasting of sea surface salinity in a coastal zone using machine learning LASSO regression models built with sparse satellite time-series datasets

Opeyemi Ajibola-James,[a,b,*] and Francis I. Okeke (Late),[a]
[a]University of Nigeria Enugu Campus, Faculty of Environmental Studies
Department of Geoinformatics and Surveying, Enugu, Nigeria
[b]Geo Inheritance Limited
Department of Data Science, Geoinformatics and Surveying, Port Harcourt, Nigeria
*Corresponding and First Author, Email: o.ajibolajames@geoinheritance.com, o.ajibolajames@gmail.com

1.0 Introduction

The risks of upstream seawater intrusion from coastal zones to the environment, food security and people's health, particularly in terms of evidence-based threats to optimum yield of sensitive plants such as paddy rice and horticultural crops (CGIAR-RCSA, 2016); and drinking water supply (Sneath, 2023), which are crucial for sustaining some 37% of the world's population living within 100 km of the coast (UNEP, 2024), are gradually becoming serious issues that require proactive environmental monitoring and good modelling approaches.

However, the temporal resolutions of relevant contemporary all-weather satellites that detect sea surface salinity (SSS) are unable to support real-time applications that can provide the required early warning information for mitigating such risks (Ajibola-James et al., 2023). The relatively low spatial resolution of the most relevant in situ salinity measurement by Argo floats (Kramer, 2002) exacerbated by their relatively scanty deployment along coastal zones; and the inaccurate salinity measurements that drift to higher values produced by over 60% of the floats between 2015 and 2019 make them relatively inefficient for mitigating such risks (Liu et al., 2024), particularly at a regional scale.

Our current practical knowledge of the efficiency of machine learning (ML) least absolute shrinkage and selection operator (LASSO) regression models built with relatively sparse all-weather satellite time-series datasets for achieving relatively accurate predictor variable selection, collinearity detection, and high SSS prediction accuracy that can provide early warning information for mitigating such risks is still limited. Consequently, **the objectives for this study are to**:

(i) determine the best parameter combination (PC) values for a relatively accurate ML LASSO regression model;
(ii) identify the best penalty and algorithm for constructing a relatively accurate L0-regularized regression (L0) model, and determine and validate potential predictor variables (PPVs) importance, and collinearity; and
(iii) predict and validate monthly SSS values for 12 months ahead (Jan.-Dec. 2021).

2.0 Methods

2.1 All-weather satellite and ancillary datasets acquisition

Table 1: Datasets utilized for the study

NASA's SMAP*
Sea surface salinity (sss) SSS error (sss_uncertainty) Wind speed (ws) High wind speed (hws) [via HYCOM_sss] Sea surface temperature (sst) (JPL, 2020) *Soil Moisture Active Passive
Copernicus & Earhdata
Absolute dynamic topography (adt) Sea level anomaly (sla) (CCCS, Undated) Precipitation (precip) (Huffman, 2019)

Monthly Mean of each data

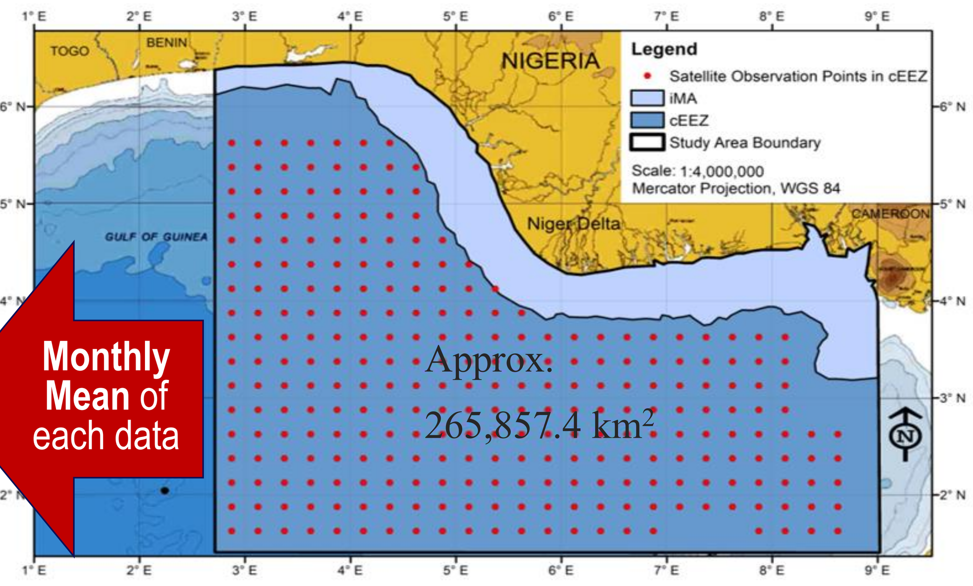

Figure 1: Map of the study area showing the 278 data points (in red) of each variable observation, sss, ws, hws, sst, adt, sla and precip (Jan. 2016-Dec. 2020); sss and sss_uncertainty (Jan.-Dec. 2021)
Source of basemap: Anyikwa & Martinez (2012)
Modification: Authors (2024)

2.2 Data preparation

Extraction & Transformation (Python libraries/Spyder IDE: .nc4 and .nc ➡ .CSV .CSV .CSV).
Cleaning (Python libraries/the IDE: Null Values & Outliers; RStudio: Monthly Mean of data).
Partitioning into **Training** (Jan. 2016-Dec. 2020) and **Validation** (Jan.-Dec. 2021) **datasets**.
Determination of SMAP SSS error with *sss_uncertainty* – *Jan.-Dec 2021* (MS Excel: RMSD).

2.3 Parameterization and ML LASSO regression model development in RStudio

- **Identified the possible lookback (LB) & h-step-ahead (H)** PCs using some data in Table1.
- **Built 6 ML models** with ***ForecastML*** & **determined the best LB & H** with R^2 & **MAPE.**

2.4 Determination of PPVs importance and collinearity in RStudio

- **Adopted L0L2 penalty,** and **Cyclic Coordinate Descent & Partial Swap-Inescapable (CD-PSI) algorithm** for **building 6 L0 models** with ***L0Learn*** (Hazimeh & Mazumder, 2020).

2.5 Experimental validation of PPVs importance and collinearity in RStudio

- **Built 7 ML LASSO regression models** with ***ForecastML*** using **the best PC** and **7 variants of PPVs** to **forecast monthly SSS** (Jan.-Dec. 2021) in **a series of experiments A to G.**

2.6 Prediction of SSS and validation of the SSS forecast accuracy in RStudio

- **Adopted the best LASSO model (highest R^2)** for the **SSS prediction** (Jan.-Dec. 2021).
- **Validated the predicted monthly SSS** with the **satellite observed monthly SSS** in 2021 over the coast by computing the **RMSE** and **MAPE** with ***MLmetrics*** **library**.

3.0 Results

3.1 RMSD between the satellite SSS and in situ SSS using the sss_uncertainty

- The preliminary result shows that the root-mean-square difference (RMSD) of the SSS (Jan.-Dec. 2021) over the study area (Figure 1) is **0.116207 practical salinity unit (psu)**.

3.2 Parameterization of ML LASSO regression models

Table 2: Performance of the 6 possible LB and H PCs in the time series forecasting of SSS with the ML LASSO models

PCs Values	PPVs	R^2	RMSE (psu)	MAE (psu)	MAPE (%)	PCs Performance Position
LB:36, H:36	ws, hws, sst, adt, sla, precip	0.55423	0.79365	0.57143	1.76698	6th
LB:36, H:24	ws, hws, sst, adt, sla, precip	0.56596	0.78314	0.56587	1.74519	5th
LB:36, H:12	ws, hws, sst, adt, sla, precip	0.59804	0.75364	0.55023	1.69723	4th
LB:24, H:24	ws, hws, sst, adt, sla, precip	0.73891	0.58156	0.40802	1.25989	2nd
LB:24, H:12	ws, hws, sst, adt, sla, precip	**0.77123**	**0.54437**	**0.36248**	**1.12097**	**1st**
LB:12, H:12	ws, hws, sst, adt, sla, precip	0.60541	0.68216	0.53759	1.65196	3rd

3.3 Determination of PPVs importance and collinearity

Table 3: Determination of PPVs importance and collinearity using the L0 models with L0L2 & CD-PSI

L0 Maximum Support Size	PPVs Importance in Descending Order	Coefficient	Intercept
maxSuppSize = 6	V1 (ws) V2 (hws) V5 (sla) V4 (adt) V3 (sst) V6 (precip)	0.06453669 -0.03312278 -1.45196918 -1.45194054 -0.03004772 0.07919975	42.91275984
maxSuppSize = 5	V1 (ws) V2 (hws) V5 (sla) V4 (adt) V3 (sst)	0.06456058 -0.03330373 -1.44796925 -1.44796835 -0.02981000	42.86146292
maxSuppSize = 4	V1 (ws) V2 (hws) V5 (sla) V4 (adt)	0.14046467 -0.06784094 -2.72801923 -2.72839010	34.52565342
maxSuppSize = 3	V1 (ws) V2 (hws) V5 (sla)	0.3885072 -0.1784569 -6.2265858	33.1928511
maxSuppSize = 2	V1 (ws) V2 (hws)	0.5631397 -0.2651062	32.5008548
maxSuppSize = 1	V1 (ws)	0.07405407	32.68467276

Perfect collinearity and relative importance (RI) detected

Perfect collinearity and RI detected

Perfect collinearity and RI Detected

Most important predictors selected

3.4 Experimental validation of PPVs importance and collinearity

Table 4: Evidence-based validation of PPVs importance and collinearity with ML LASSO models

Experiment	PPVs	R^2	RMSE	PPVs Performance Position
A	V1 (ws), V2 (hws), V5 (sla) V4 (adt), V3 (sst), V6 (precip)	0.77123	0.5443722	6th
B	V1 (ws), V2 (hws), V5 (sla) V4 (adt), V3 (sst)	0.8189632	0.4842613	5th
C	V1 (ws), V2 (hws), V5 (sla) V4 (adt)	0.8239762	0.4775096	1st
D	V1 (ws), V2 (hws), V5 (sla)	0.8239762	0.4775096	1st
E	V1 (ws), V2 (hws), V4 (adt)	0.8239761	0.4775098	2nd
F	V1 (ws), V2 (hws)	0.8223169	0.4797549	3rd
G	V1 (ws)	0.8216164	0.4806997	4th

Perfect collinearity validated

Most important predictors validated with the BEST MODEL

RI validated

3.5 Prediction of SSS and validation of the SSS forecast accuracy

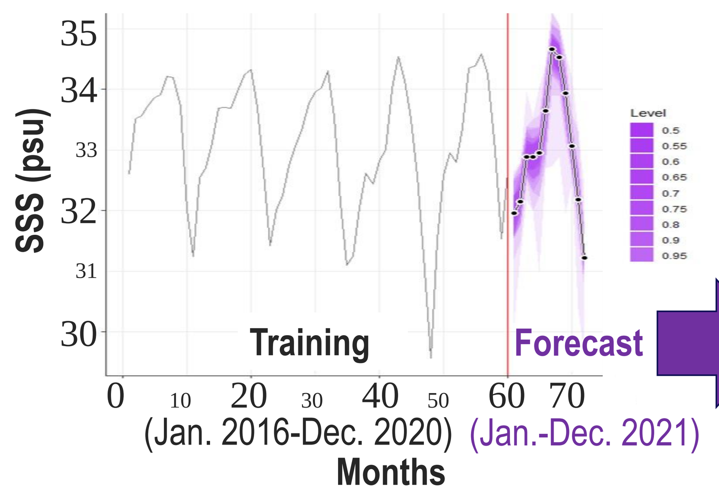

Figure 2: Monthly SSS forecast plot (Jan.-Dec. 2021) with the best ML LASSO model

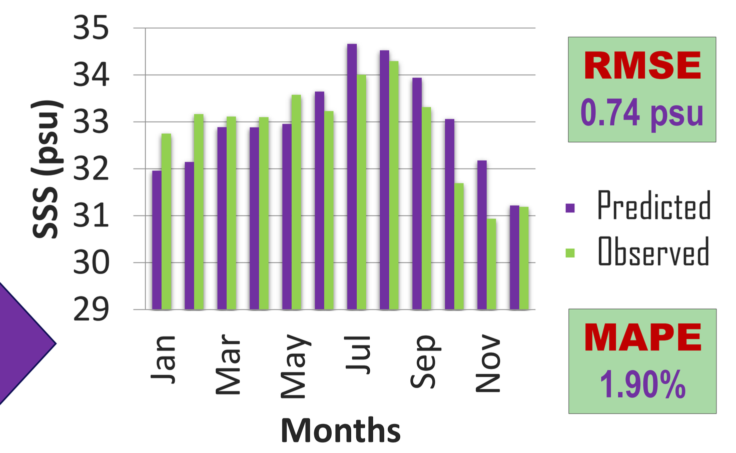

Figure 3: Bar chart of the SSS forecast and SMAP satellite observed SSS, and the SSS forecast accuracy metrics (Jan.-Dec. 2021)

4.0 Discussions

The result of the RMSD (**0.12 psu**) between the satellite SSS and in situ SSS exceeds the SMAP satellite mission's accuracy of **0.2 psu** by a substantial margin of about **41.9%**. This **implies credible validation data for the SSS forecast**; and **credible data preparation method** that involved rigorous supervised-automatic deletion of observation records with null values, and outliers induced by the radio frequency interference (RFI) and land contamination. **The results** of the evidence-based approach for determining and validating the best PC (Table 2), the most important PPVs combination, and collinearity (Table 3 & 4), which **produced the most accurate ML LASSO model (R^2 = 0.8239762)** (Table 4) **that predicted the SSS** (Jan.-Dec. 2021) **at a relatively high accuracy level** (Figure 2), RMSE of 0.74 psu and **MAPE of 1.90%, about 5 times less than 10% limit** (Lewis, 1982) (Figure 3) **have the following implications:**

- ❑ **Accuracy of such a ML LASSO regression model depends largely on evidence-based** success of parameter values selection, most important PPVs selection, collinearity detection tasks; **evidence-based** accuracy of the algorithms involved in each of the tasks; and the accuracy of satellite data utilized for the model building and forecast values validation.
- ❑ **L0-regularized regression models with L0L2 and CD-PSI are relatively efficient for PPVs' RI detection, most important PPVs selection and collinearity detection.**
- ❑ **Performance** of such a **ML LASSO** model can be **optimized** with such **L0-regularized** model.
- ❑ The results are **consistent** with the claim of Hazimeh & Mazumder (2020) on **L0** performance.

5.0 Conclusion

As demonstrated by the results of this study, **a good approach** for using relatively sparse satellite time-series datasets of 60 epochs (monthly scale) to build a relatively accurate ML LASSO regression model for useful SSS forecasting **should begin with rigorous** supervised-automatic deletion of observation records with null values and outliers, **followed by unbiased** selection of appropriate parameter values, **evidence-based** identification of important predictor variables and collinearity assessment. **This good modelling and forecasting approach can be adopted by the stakeholders for replicating the relatively high SSS prediction accuracy to provide useful early warning information for proactive monitoring and mitigation of the risks of upstream seawater intrusion from coastal zones**, particularly to **people's health** (drinking water supply) and **food security** (crops' yield) in coastal areas.

References

- Visit: https://references.aiwaapp.live or Scan QR: