Machine Learning ARIMA For Modelling and Forecasting Variations in Earth's Surface Phenomenon Using Sparse Time Series Satellite Data - A Case Study of Sea Surface Salinity in A Tropical Coast

OPEYEMI AJIBOLA-JAMES[1], FRANCIS I. OKEKE[2]
[1]*Geo Inheritance Limited, Port Harcourt, Rivers State, Nigeria*
[2]*Department of Geoinformatics and Surveying, University of Nigeria, Enugu Campus, Nigeria*

Abstract- ***The contemporary all-weather satellites observations of Earth's surface phenomena (ESP) are characterized by relatively sparse time series data that discourage their utilization in building efficient machine learning (ML) models for exploratory and predictive purposes. Additionally, data-poor areas usually have difficulties in meeting the multiple predictor variables requirement of building appropriate multivariate ML regression models. We utilized a relatively sparse sea surface salinity (SSS) dataset from the Soil Moisture Active Passive Mission (SMAP) satellite for this study. We determined the accuracy and variability of the relatively sparse SSS data. We built ML autoregressive integrated moving average (ARIMA) models; determined and validated the best model for modelling and forecasting ESP using the relatively sparse data as a case study. We show root mean squared differences, RMSDs (0.1279 and 0.1162 psu) for the modelling and forecasting data accuracy respectively. We show a standard deviation, SD (0.2528 psu) for the interannual SSS variability (iSSSv). We show the modelling accuracy with an R-squared, R2 (0.8345 psu) and its validation with a mean absolute percentage error, MAPE (0.7779%) for the best model. We show the best variant of the traditional SSS forecasts ("Lo") accuracy with root mean squared error, RMSE (0.5435 psu) and its validation with MAPE (1.5038%) for the best model. The results suggest relatively high modelling and prediction accuracy. The results imply that relatively sparse satellite time series data of at least 60 epochs can be integrated with a ML ARIMA model for modelling and forecasting variations in any ESP, regardless of the location.***

Indexed Terms- ***Earth's Surface Phenomenon, Sea Surface Salinity, Machine Learning Arima, VariationsModelling, Time Series Forecasting***

I. INTRODUCTION

Measurable physical, chemical and biological variables involved in the Earth's surface processes are generally considered as ESP, which are characterized by both spatial and temporal variations. The magnitude and frequency of such variations are usually driven by two or more relevant factors. In the case of variations in sea surface salinity (SSS) on a global spatial scale, evaporation, precipitation, river outflow, and melting ice are among the principal drivers (Dinnat et al., 2019). However, changes in SSS on a local spatial scale in the tropics, particularly along the Nigerian coastal zone have been innovatively linked to three physical oceanographic predictors, wind speed (WS), high wind speed (HWS), and sea level anomaly (SLA) in a recent study (Ajibola-James, 2023; Ajibola-James & Okeke, 2025). The risks of positive anomaly in ESP, including SSS in terms of upstream seawater intrusion induced by rising sea level (Zhou, 2011) and/or high tides have been linked to human and environmental health issues (CGIARCSA, 2016; Trung et al., 2016; Sneath, 2023; Ajibola-James, 2023). Since 1970s, both in situ and optical remote sensing approaches have been utilized for mapping and modelling spatial and/or temporal variations in ESP, particularly SSS using statistical methods

(Khorram, 1982; Qing et al., 2013). A few studies have integrated optical satellite SSS data (derived via indirect measurement) with machine learning (ML) methods to monitor and predict salt water intrusion to upstream (Nguyen et al., 2018); and estuary (Jiang et al., 2024). Contemporary radar satellite SSS data (derived via direct measurement) have also been integrated with machine learning (ML) methods due to the relative advantages of such synergism (Ajibola-James, 2023; Ajibola-James & Okeke, 2025). The all-weather capability of such radar satellite has helped to eliminate the problem of temporal data sparsity imposed on such optical SSS observation by cloud cover effects.

ML is a subset of artificial intelligence, and a method of data analysis that involves building systems (models and algorithms) that can learn from data without being explicitly programmed, identifying patterns, and making decisions with minimal human intervention (Ajibola-James, 2023). To make the model selection process simpler for forecasting, ML entails a variety of strategies for identifying patterns and relationships in the data (Chan-Lau, 2017). A notable advantage of ML models and algorithms is their increasing ability to handle the time component of relatively large amounts of data (complex structured, semistructured and unstructured datasets with several characteristics, including volume, velocity, veracity, value and validity) in predictive studies. A time series is a sequence of data collected over a specific period of time. The time scale component of a data series may be either every minute or hourly or daily or monthly or yearly. When only one of the time scales is involved, it is regarded as a single seasonality. Any situation involving datasets with more than one of the time scales, for example, hourly and daily or hourly, daily and monthly or daily, monthly, and yearly, is called multiple seasonality. Time series forecasting has become a significant part of ML since there are many prediction problems with time components (Ajibola-James, 2023).

In the parlance of ML, particular when it comes to ESP modelling, multivariate models are usually preferred to univariate models. The reasons for this preference are not far-fetched. Generally, the latter offer explicit information on the specific predictors in terms of their identity and relative contributions to the predictive power of such models. Additionally, the latter that fall in the category of sub-set selector models could help to alleviate the negative effects suffered by their predictive power and interpretability when improving fit by including a large number of independent variables. However, the increasing difficulties in meeting the multiple predictor variables requirement of building such multivariate (data-intensive) ML regression models; and accessing appropriate in situ data for such model validation in data-poor areas (such as tropical coasts including Africa) have been discouraging the exploitation of such ML models for modelling and prediction of ESP, particularly SSS till date. One of the most widely used ML methods for time series forecasting that has reasonably helped to overcome this barrier is ML ARIMA. In the applications of the ARIMA model, a widely used approach is known as the Box–Jenkins principle, which consists of three iterative steps, namely, model identification, parameter estimation, and diagnostic checking phases (Box & Jenkins, 1970). A relative advantage of a typical ML ARIMA model for time series modelling and prediction is that it does not require predictor (independent) variables to fit new (predicted) values. Thus, the amount of data input, data processing time, and computer hardware required for implementing it are relatively low (Ajibola-James, 2023). These are part of the motivations for its wide adoption, particularly in modelling and forecasting of health, financial and economic data (Renato, 2013; Zhirui & Hongbing, 2018; Nayak & Narayan, 2019; Khan & Gunwant, 2024; Yu et al., 2025).

The tropical coasts, particularly the Nigerian coastal zone, have been traditionally undersampled using appropriate in situ methods and are understudied using remote sensing techniques (Ajibola-James, 2023). More than often, such data-poor areas have difficulties meeting the multiple predictor variable requirement of building appropriate multivariate ML regression models. Despite the relative advantages of using ML ARIMA for modelling ESP, our knowledge of its accuracy in fitting new values when built with relatively sparse time series satellite data is still limited, particularly in such data-poor areas. Consequently, the objectives of this paper are to (i) determine the accuracy of relatively sparse SSS data

(Jan. 2016-Dec. 2020; and Jan.-Dec. 2021) for the study area; (ii) determine the interannual variability of such SSS data (Jan. 2016-Dec. 2020); and (iii) determine and validate a relatively accurate ML ARIMA model (Jan. 2016-Dec. 2020) and its SSS forecasts for 12 months (Jan.-Dec. 2021) as a case study for ESP.

II. STUDY AREA

The location adopted for this experimental study was the Nigerian coastal zone, which comprises the immediate maritime area (IMA) and the contiguous Exclusive Economic Zone (EEZ) and reaches approximately 200 nautical miles (370 km) offshore of the Nigerian continental shelf; this zone should not extend beyond the limits of approximately 350 nautical miles in accordance with the provisions of Article 76(8) of the 1982 United Nations Convention on the Law of the Sea (UNCLOS) (United Nations, undated). The IMA was established for the purpose of this study. The offset ranged from 58-100 km between the shoreline and the edge of the observation points in the contiguous EEZ (Figure 1). To significantly reduce the effect of the error associated with satellite SSS data acquisitions close to land masses on the data accuracy, as observed by Boutin et al. (2016), the IMA was excluded from the study area. The study area was restricted to 278 data observation points in the contiguous EEZ of approximately 295,027.4 km2 (Figure 1). In the area, the mean monthly rainfall ranges from approximately 28 mm in January to approximately 374 mm in September (Zabbey et al., 2019). Several rivers, including the Niger, Forcados, Nun, Ase, Imo, Warri, Bonny, and Sombreiro Rivers, discharge freshwater to the coastal region of Nigeria. Given the actual evaporation of 1,000 mm per annum, a total runoff of 1,700–2,000 mm, and an additional flow of 50–60 km^3 calculated for the water balance of the Niger system, a total of 250 km^3 per year eventually discharges into the Gulf of Guinea (Golitzen et al., 2005; Ajibola-James, 2023).

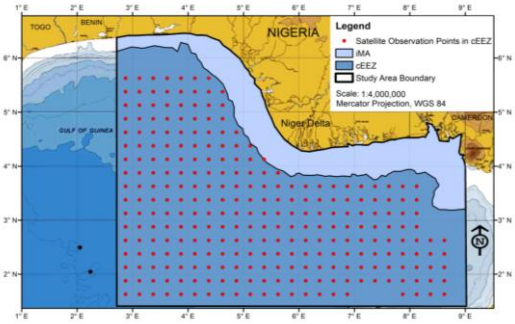

Figure 1: Map of the Study Area showing the 278 Points (in red) of SMAP Satellite SSS Data Observations (January 2016-December 2021).

Source: Anyikwa & Martinez (2012) and Modification: Authors (2024)

III. MATERIALS AND METHODS

A. Satellite Observations and Map

The monthly SMAP satellite SSS and the Uncertainty time series datasets utilized for this particular study were retrieved from NASA's SMAP online repository managed by NASA's Joint Propulsion Laboratory, JPL (JPL, 2020) in network Common Data Form-4 (netCDF-4) file format. The characteristics of the cleaned datasets analysed for the study are presented in Table 1. The base map of the study area was sourced from Anyikwa & Martinez (2012) and modified as appropriate (Figure 1).

Table 1: Feature Datasets utilized for the Study

Data Name	Data Variable	Observation (Obs.) Period	Time Scale	Spatial Resolution	Obs. per Time	Total Obs.	Data Purpose
SMAP	SSS; SSS Uncertainty	Jan. 2016 to Dec., 2020	Monthly	0.25° (Lat.) × 0.25° (Lon.)	278	16680	Modelling
S	SSS;	Jan.	Mo	0.25	27	33	Vali

MAP	SSS Uncertainty	to Dec. 2021	nthly	° (Lat.) × 0.25 ° (Lon.)	8	36	dation

B. Data Preparation

The appropriate data preparation tasks (data extraction, cleaning and selection) were implemented using scripted procedures. The dataset was automatically extracted from the netCDF-4 file into comma-separated Microsoft (MS) Excel (.csv) file by executing a python 3.10.2 script in Spyder IDE (Integrated Development Environment) 5.2.2 software. The data cleaning involved rigorous supervised-automatic deletion of the observation records with null values (redundant empty records that have no relevance to the study area, which might be created in the process of the data file transformation) and outliers induced by radio frequency interference (RFI) and land contamination in the dataset stored in the .csv file. This was achieved through appropriate tasks. First, automatic deletion of null values by executing a python script with libraries pandas, numpy, csv and xarray in the IDE. Second, visual identification and verification of outliers by overlaying each of the monthly SSS observations in the .csv files on the Google Earth Pro online to ascertain their proximity to land and tendency for land contamination. Third, automatic deletion of the predetermined outliers by using their concatenated location coordinates as criteria for executing a python script with the same libraries and IDE that was utilized in (a) above. A total of 278 appropriate satellite observation points, which constitute the study area (Figure 1) were selected for analysis in this stud by executing a python script in the IDE. The points were imported and merged with the base map using the overlay function in ArcMap 10.4.1 (Ajibola-James, 2023).

C. Data Accuracy and Variability

The accuracy of the satellite SSS datasets for the modelling and validation were computed in MS Excel software by using the SSS Uncertainty datasets (the difference between in situ SSS and satellite SSS) (Table 1). To compute the accuracy of the modelling data, the SSS uncertainty data of 16680 observation points were uploaded to column A in Excel to produce the formula A2:A16681 for computing the sum square (SUMSQ) in cell C2, which was given by the formula SUMSQ (A2:A16681). The mean squared difference (MSD) given by formula =(C2/16680) was computed in cell D2, while the RMSD was finally computed by using formula =SQRT(D2). The same procedure was replicated for computing the accuracy of the forecasting data using 3336 observation points.

The interannual variability of the SSS data was determined by utilizing the MLmetrics library to compute the SD, a universal measure of variability in R 4.1.3/R-studio 2022.02.3-492 software. After the mean annual SSS values for 2016 to 2020 were uploaded to the software by running data_obs_sss <- read.csv(file.choose(), header = TRUE, stringsAsFactors = FALSE), the dataframe produced (Table 2) by running data_sss <- data_obs_sss[, c("year", "sss")] was vectorized by running sss_2016_2020 <- data_sss$sss. The SD was finally computed by running sd (sss_2016_2020).

Table 2: Dataframe for Computing Interannual Variability in SSS

Year	SSS
2016	33.15872
2017	33.12886
2018	32.79823
2019	32.55897
2020	33.02366

D. Autoregressive Integrated Moving Average Model and Algorithm

The ML ARIMA models and algorithms were built primarily with the forecast library 8.17.0 in R 4.1.3/R-studio 2022.02.3-492 software. The modelling SSS data was characterised by 60 monthly epochs (Jan. 2016-Dec. 2020), which was utilized for training the ML ARIMA models to forecast variations in SSS for 12 months ahead. Other complimentary libraries, such as tseries and

MLmetrics, were also used in this process. Model fitting and selection were achieved with the auto.arima() function. The function helped to determine the best model for given input data based on relevant model evaluation criteria. At the inception of the ML modelling task, the dataframe, df, containing the 60 monthly epochs of the SSS data was transformed from "function" to “time series” to satisfy one of the basic assumptions of the ARIMA model.

Appropriate visual and metric approaches were leveraged to assess the stationarity of the time series data. The former involved the inspection of autocorrelation function (ACF) and partial autocorrelation function (PACF) plot patterns (Fattah et al., 2018; Hyndman & Athanasopoulos, 2021), while the latter involved hypothesis testing using augmented Dickey-Fuller (ADF) test metrics (Cheung & Lai, 1995; Ajibola-James, 2023). The following hypotheses and decision rules were adopted for the ADF test:

H_0: Nonstationary
H_1: Stationary

where H_0 is the null hypothesis and H_1 is the alternative hypothesis.

If the p value is ≤ 0.05, H_0 is rejected to support H_1.

Given that the computed p value = 0.1769, which is > 0.05, H_0 of Nonstationary was accepted to reject H_1 of Stationary. To achieve “stationarity”, another basic assumption of the ARIMA model, first-order differencing was applied to the data. The ADF test metric was repeated to reassess the output of the differenced data. Given that the computed p value = 0.01, which is < 0.05, H_0 of Nonstationary was rejected to accept H_1 of Stationary. The best ARIMA model together with the most appropriate parameters were identified using the auto.arima function; and determined with the allied Akaike information criterion (AIC) minimization. The residual of the best ML ARIMA model was also assessed for randomness with the Ljung-Box test based on the following hypotheses and decision rule:

H_0: No white noise
H_1: White noise
If the p value is ≥ 0.05, H_0 is rejected (Hyndman & Khandakar, 2008).

Given that the computed p value = 0.4522, which is > 0.05, H_0 of Nonstationary (No white noise) was rejected to accept H_1 of Stationary (White noise). The best ARIMA model characterised by White noise was used as input for building the user-defined ML ARIMA model to forecast SSS 12-months ahead (Jan.-Dec. 2021). The traditional variants of the SSS forecasts (Jan.-Dec. 2021) by the best ARIMA model are characterised by “Forecast”, “Lo 95”, and “Hi 95”. The graph of the three variants of the SSS forecast values by the model was generated in the Excel.

E. Determination and Validation of ARIMA Model Accuracy for Modelling and Forecasting SSS

The accuracy of the built ML ARIMA model for modelling variations in SSS was computed by using the R^2 performance metric, which represents the amount of variation explained by the ML model. The forecasting accuracy was determined with the RMSE, a measure of accuracy that reveals the magnitude of the difference between the predicted and observed values. The validation of the modelling and forecasting accuracy of the best ML model in relation to error estimation, which is also known as residual variation, was also computed in terms of MAPE, a good measure of the absolute percentage difference between predicted and observed values. In general, the greater the R^2 value is, the greater the amount of variation explained by the ML model. Conversely, lower values of MAPE and RMSE indicate relatively good accuracy of forecasts made by the model. In terms of the interpretation of the error metrics in real-world applications, the MAPE seems to be the most versatile because it is usually computed in percentage (%) units. In addition, what should be considered an acceptable accuracy level seems to be properly documented for the MAPE. In this regard, a MAPE less than 10% is considered to indicate “high prediction accuracy” (Lewis, 1982; Ajibola-James, 2023). It should be underscored that the true test of an ML time series model’s performance is in accurately forecasting new values. This is usually

determined by the value of its performance metrics in forecasting new target values that are not included in the model's training datasets.

IV. RESULTS AND DISCUSSION

A. Data Accuracy

The accuracies of the relatively sparse SSS data over a geographical area of approximately 6.5° × 4.5° in terms of the RMSD are 0.1279 psu and 0.1162 psu for the modelling dataset and forecasting dataset, respectively. The two RMSD values show a relatively high level of accuracy exceeding the SMAP missions' accuracy requirement of 0.2 psu by substantial margins of approximately 36.05% and 41.90%, respectively. It should be noted that relatively high accuracy was achieved by the rigorous supervised automatic data cleaning approach, which primarily involved deletion of the outliers induced by RFI and land contamination in the satellite dataset. This implies that the data preparation technique can reasonably affect the accuracy of the input dataset in a modelling and predictive study.

B. Interannual Variability

The interannual variability in the SSS data in terms of the SD shows 0.2528 psu. This suggests that the iSSSv is relatively stable (predictable) given that it is approximately 74.72% less than 1 SD. This result shows that the dataset could be considered a viable input for the ARIMA model. However, to achieve the most appropriate level of "stationarity", a basic assumption of the ARIMA model, the first-order differencing was applied to the data as earlier mentioned in section 3.4. This implies that the order of differences that would be taken in a given input data for ML ARIMA modelling is a function of the SD value. Consequently, data variability assessment using the SD value should be considered an essential aspect of exploratory data analysis (EDA) in the process of building ML ARIMA models. Additionally, the result of the iSSSv also implies that its major drivers in terms of the physical oceanographic predictors (WS, HWS, and SLA) are relatively stable (predictable) for the period in question (2016-2020).

C. Determination and Validation of the Best ARIMA Model

The best ML ARIMA model with the most appropriate parameter that scored the minimum AIC value of 81.80972 was ARIMA(0,1,2)(0,1,1)[12] (Table 3). The result of the preliminary automatic model selection task show that the AIC metric is an efficient approach for determining the best ML ARIMA model and the most appropriate parameters. The associated terms, coefficients and p-values are ma1, -0.3858, 0.0000; ma2, -0.2633, 0.0000; and sma1, -0.7099, 0.0000 respectively. The result of the z-test performed on the terms of the best model, which shows a p-value of 0.0000 (< 0.05) for the three terms implies that all the terms utilized in building the best ARIMA models are statistically significant. The result also imply that they have meaningful impact on the time series. The result of the modelling accuracy assessment performed with R^2 is 0.8345, while the result of its validation with MAPE is 0.7779%. The relatively high R^2 value shows that the ML ARIMA model explained a relatively large amount of variation, while the relatively low MAPE value shows that the ML ARIMA model has a relatively high modelling accuracy.

Table 3: Determination of the Best ML ARIMA Model with auto.arima Function

ARIMA Model	AIC
ARIMA(2,1,2)(1,1,1)[12]	: Inf
ARIMA(0,1,0)(0,1,0)[12]	: 93.4323
ARIMA(1,1,0)(1,1,0)[12]	: 87.02665
ARIMA(0,1,1)(0,1,1)[12]	: 81.91669
ARIMA(0,1,1)(0,1,0)[12]	: 88.61325
ARIMA(0,1,1)(1,1,1)[12]	: Inf
ARIMA(0,1,1)(1,1,0)[12]	: 82.80056
ARIMA(0,1,0)(0,1,1)[12]	: 87.21183
ARIMA(1,1,1)(0,1,1)[12]	: 82.69393
ARIMA(0,1,2)(0,1,1)[12]	: 81.80972
ARIMA(0,1,2)(0,1,0)[12]	: 89.46024
ARIMA(0,1,2)(1,1,1)[12]	: Inf
ARIMA(0,1,2)(1,1,0)[12]	: 83.67727
ARIMA(1,1,2)(0,1,1)[12]	: 83.56041
ARIMA(0,1,3)(0,1,1)[12]	: 83.48725
ARIMA(1,1,3)(0,1,1)[12]	: Inf

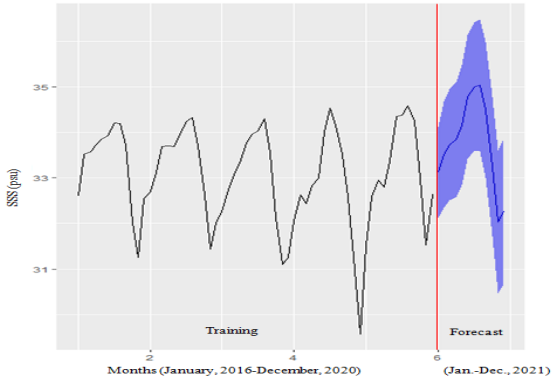

Figure 2: Modelling (Jan. 2016-Dec. 2020) and Forecasting (Jan.-Dec. 2021) of SSS Variations using the Best ML ARIMA Model

In Figure 2, the "Training" side shows the result of using 60 monthly epochs (Jan. 2016-Dec. 2020) of the data to train the best ML ARIMA model for modelling variations in the SSS, while the adjoining "Forecast" side shows the result of the 12 monthly epochs (Jan.-Dec. 2021) of the SSS forecast.

D. Determination and Validation of Forecasting Accuracy of the Best ARIMA Model

The results of the traditional variants of the SSS forecasts, "Forecast", "Lo 95", and "Hi 95" (Jan.-Dec. 2021) by the best ARIMA model are presented in Figure 3. The results of the "Forecast", and "Hi 95" variants in relation to the observed SSS show over-estimation of the SSS values for the entire period (Jan.-Dec. 2021). However, the "Hi 95" variant shows a higher level of the over-estimation. In relation to the observed SSS, the "Lo 95" variant shows the lowest level of over-estimation for 10 months, under-estimation for 1 month (Jun), and precise-estimation for 1 month (Oct.).

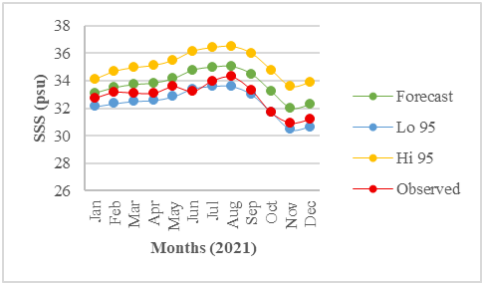

Figure 3: Observed SSS and the Traditional Variants of the SSS Forecasts (Jan.-Dec. 2021) by the Best ARIMA Model

In Table 4, the accuracy of the traditional variants of the SSS forecasts by the best ARIMA model are presented in Table 4. Given that the RMSE is relatively difficult to interpret for such applications due to the squared nature of the measured error, the MAPE was utilized to validate the forecasting accuracy. The relatively low MAPE, which ranges from 1.5038 to 6.9260% (less than the 10% upper limit of high prediction accuracy benchmark) show that the best ARIMA model has a relatively high forecasting accuracy. This also suggests that the "Lo 95" variant, which shows the lowest level of over-estimation for 10 months (Figure 3) and the lowest MAPE of 1.5038% (Table 4) is validated to offer the most accurate traditional SSS forecast values.

Table 4: Accuracy of the Traditional Variants of the SSS Forecasts by the Best ARIMA Model

Traditional Forecasts Variants	RMSE (psu)	MAE (psu)	MAPE (%)
Forecast	0.9850	0.9041	2.7665
Lo 95	0.5435	0.4958	1.5038
Hi 95	2.3283	2.2673	6.9260

CONCLUSION

The use of a relatively sparse satellite time series SSS data from a tropical coast, the Nigerian coastal zone as a case study for ML ARIMA for modelling and forecasting variations in ESP yields encouraging results. This imply that relatively sparse satellite time series data from at least 60 epochs (hourly, daily, weekly monthly or yearly) can be productively utilized for building a relatively accurate ML ARIMA model for modelling and forecasting variations in any ESP in any geographical area. A relative advantage of the time series model is that it does not require predictor (independent) variables to model variation and fit new (forecast) values. In this regard, the costs (in terms of the amount of data input, data processing time, and computer hardware) of implementing it are relatively low and affordable. The variation modelling accuracy that was validated with a MAPE of 0.7779% is more than 2 times greater (better) than the forecasting accuracy with a MAPE of 2.7670%. It should be underscored that such a difference in accuracy (in which the accuracy

of the former exceeds that of the latter) is a normal experience in such applications of ML models because the observed data utilized for validating the accuracy of the latter are relatively new to the ML model. The synopses of our key findings are subsequently presented.

- The data preparation technique made the RMSD of the SSS data for modelling and forecasts validation to show relatively high level of accuracy that exceeded the SMAP mission's accuracy requirement of 0.2 psu by considerable margins of about 36.05% and 41.90% respectively.
- The iSSSv, and its major physical oceanographic predictors (WS, HWS, and SLA) were relatively stable (predictable) for the period (2016-2020).
- The "Lo 95" variant made the most accurate traditional SSS forecasts (Jan.-Dec. 2021) with the lowest validation MAPE of about 1.5038% (approximately 6 times less than the 10% upper limit of high prediction accuracy benchmark).

RECOMMENDATIONS

Considering the relatively high accuracy of the ML ARIMA model coupled with its relatively low costs of implementation, the following are highly recommended.

a. The ML model and its algorithm should be updated and adopted by stakeholders (particularly government agencies and aquatic entrepreneurs) as early warning decision support tools that will enable them to provide proactive and sustainable preventive measures to any current and future risks that may be posed by any ESP to humans and the environment. For example, the ML model built with SSS data can serve as a decision support tool for providing early warning information on the risk of upstream seawater intrusion to the drinking water supply, people's health, sensitive plants such as rice and horticultural crop yield, and the environment.
b. Further studies on the comparative assessment of the best ML ARIMA model with the one utilizing a relatively small or large number of monthly mean SSS satellite observations should be encouraged.
c. Additionally, appropriate local and global funding that will facilitate prompt execution of the recommendations in (1) and (2) above should be equitably provided to reliable but relatively marginalized individual researchers, private research organizations and private/public research institutions in the geospatial, AI/ML, and related industries in Africa, particularly in Nigeria as soon as possible. This will encourage the development of realistic low-cost univariate ML models built with remotely sensed data.

ACKNOWLEDGEMENTS

Opeyemi Ajibola-James: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. Francis Okeke: Supervision.

Funding: This research received no external funding.
Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used was obtainedfromJPLvia https://doi.org/10.5067/SMP50-3TMCS accessed on 10 July 2022.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Ajibola-James, O. (2023). Assessment of sea surface salinity variability along Nigerian coastal zone using machine learning – 2012-2021 [Doctoral thesis, University of Nigeria, Nsukka, Enugu Campus].

[2] Ajibola-James, O., & Okeke, F. I. (2025). 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. *Advances in Space Research.* https://doi.org/10.1016/j.asr.2025.07.077

[3] Anyikwa, O. B., & Martinez, N. (2012). Continental Shelf Act, 2012. The International Maritime Law Institute, IMO. https://imli.org/wp-content/uploads/2021/03/Obiora-Bede-Anyikwa.pdf

[4] Boutin, J., Chao, Y., Asher, W. E., Delcroix, T., Drucker, R., Drushka, K., Kolodziejczyk, N., Lee, T., Reul, N., Reverdin, G., Schanze, J., Soloviev, A., Yu,, L., Anderson, J., Brucker, L., Dinnat, E., Santos-Garcia, A., Jones, W., Maes, C., Meissner, T., Tang, W., Vinogradova, N., & Ward, B. (2016). Satellite and in situ salinity: understanding near-surface stratification and subfootprint variability. Bulletin of the American Meteorological Society, 97(8), 1391–1407. https://doi:10.1175/bams-d-15-00032.1

[5] Box, G.E.P. & Jenkins, G. (1970). Time series analysis, forecasting and control. San Francisco: Holden-Day.

[6] CGIAR Research Centers in Southeast Asia. (2016). The drought and salinity intrusion in the Mekong River Delta of Vietnam. https://cgspace.cgiar.org/rest/bitstreams/78534/retrieve/

[7] Chan-Lau, J. A. (2017). Lasso Regressions and Forecasting Models in Applied Stress Testing. International Monetary Fund (IMF) Working Paper, WP/17/108. https://www.imf.org/~/media/Files/Publications/WP/2017/wp17108.ashx

[8] Cheung, Y.-W., & Lai, K. S. (1995). Lag order and critical values of the Augmented Dickey-Fuller test. Journal of Business & Economic Statistics, 13(3), 277–280. https://doi.org/10.2307/1392187

[9] Dinnat, E. P., Le Vine, D. M., Boutin, J., Meissner, T., & Lagerloef, G. (2019). Remote sensing of sea surface salinity: Comparison of satellite and in situ observations and impact of retrieval parameters. Remote Sensing, 11(7). https://doi.org/10.3390/rs11070750

[10] Fattah, J., Ezzine, L., Aman, Z., el Moussami, H., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10. https://doi.org/10.1177/1847979018808673

[11] Golitzen, K. G. (Ed.), Andersen, I., Dione, O., Jarosewich-Holder, M., & Olivry, J. (2005). The Niger River Basin: A vision for sustainable management. World Bank, Washington, DC. https://doi.org/10.1596/978-0-8213-6203-7

[12] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(1),1–22.https://doi.org/10.18637/jss.v027.i03

[13] Hyndman, R.J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice, 3rd edition, OTexts. Melbourne, Australia. Retrieved August 31, 2022, from https://otexts.com/fpp3/

[14] Jiang, D., Dong, C., Ma, Z., Wang, X., Lin, K., Yang, F., & Chen, X. (2024). Monitoring saltwater intrusion to estuaries based on UAV and satellite imagery with machine learning models. Remote Sensing of Environment, 308, 114198. https://doi.org/10.1016/j.rse.2024.114198

[15] Joint Propulsion Laboratory. (2020). JPL CAP SMAP Sea Surface Salinity Products (PO.DAAC; Version V5.0) [Dataset]. JPL, CA, USA. Retrieved July 10, 2022, from https://doi.org/10.5067/SMP50-3TMCS

[16] Khan, I., & Gunwant, D. F. (2024). Revealing the future: an ARIMA model analysis for predicting remittance inflows. Journal of Business and Socio-economic Development. https://doi.org/10.1108/JBSED-07-2023-0055

[17] Khorram, S. (1982). Remote sensing of salinity in the San Francisco Bay Delta. Remote Sensing of Environment, 12(1), 15–22. https://doi.org/10.1016/0034-4257(82)90004-9

[18] Kotu, V., & Deshpande, B. (2019). Time Series Forecasting. Data Science, Elsevier, 395–445. https://doi.org/10.1016/B978-0-12-814761-0.00012-5

[19] Lewis, C. D. (1982). Industrial and business forecasting methods: A radical guide to exponential smoothing and curve fitting. London: Butterworth Scientific.

[20] Nayak, MDP., & Narayan, K. (2019). Forecasting Dengue Fever Incidence Using

ARIMA Analysis. International Journal of Collaborative Research on Internal Medicine & Public Health, 11(3):924-932.

[21] Nguyen, P. T. B., Koedsin, W., McNeil, D., & Van, T. P. D. (2018). Remote sensing techniques to predict salinity intrusion: Application for a data-poor area of the coastal Mekong Delta, Vietnam. International Journal of Remote Sensing, 39(20), 6676–6691. https://doi.org/10.1080/01431161.2018.1466071

[22] Qing, S., Zhang, J., Cui, T., & Bao, Y. (2013). Retrieval of sea surface salinity with MERIS and MODIS data in the Bohai Sea. Remote Sensing of Environment, 136, 117–125. https://doi.org/10.1016/j.rse.2013.04.016

[23] Renato, CS. (2013). Disease management with ARIMA model in time series. Einstein. 11(1):128-31.

[24] Sneath, S. (2023, September 23). Louisiana: New Orleans declares emergency over saltwater intrusion in drinking water. The Guardian. https://www.theguardian.com/us-news/2023/sep/22/louisiana-drought-drinking-water-mississippi-river-saltwater-new-orleans

[25] Trung, N. H., Hoanh, C. H., Tuong, T. P., Hien, X. H., Tri, L. Q., Minh, V. Q., Nhan, D. K., Vu, P. T., & Tri, V. P. D. (2016). Climate Change Affecting Land Use in the Mekong Delta: Adaptation of Rice-Based Cropping Systems (CLUES) Theme 5: Integrated Adaptation Assessment of Bac Lieu Province and Development of Adaptation Master Plan. https://www.researchgate.net/publication/301612048_Climate_change_affecting_land_use_in_the_Mekong_Delta_Adaptation_of_rice-based_cropping_systems_CLUES_ISBN_978-1-925436-36-5

[26] United Nations. (n.d.). United Nations Convention on the Law of the Sea. https://www.un.org/depts/los/convention_agreements/texts/unclos/unclos_e.pdf

[27] Yu, W., Cheng, X., & Jiang, M. (2025). Exploitation of ARIMA and annual variations model for SAR interferometry over permafrost scenarios. IEEE Journal, 1-16. https://doi.org/10.1109/JSTARS.2025.3550748