American Journal of Sciences and Engineering Research

E-ISSN– 2348–703X, Volume 8, Issue 5, 2025

A Multiple Linear Regression Analysis of Sea Surface Salinity Variability and Its Drivers (SSH, SST, Waves, Currents) in the Comoros Basin, Northwest of Madagascar

ABDALLAH Salim S.O.[*1,2], Sahoby LALAOHARISOA[1], SALIM Ahmed Ali[2], RATIARISON Adolphe A[1].

[1]*Atmosphere, Climate, and Ocean Dynamics laboratory, Doctoral School of Physics and Applications, University of Antananarivo, Antananarivo, Madagascar.*

[2]*Environmental and Climate Physics Laboratory, Faculty of Science and Technology, University of the Comoros, Moroni, Comoros.*

ABSTRACT: *The objective of this study is to model the average daily climatological salinity of the sea in the North of the Mozambique Channel. Our study is based on different climatic and oceanic variables taken from 1990 to 2020. The multiple linear regression model is used to analyze the interactions between these different variables. The explanatory variables taken into account are: sea surface height (SSH), sea surface temperature (SST), wave, and marine currents. The model obtained has a high quality of fit (R^2 = 0.985; adjusted R^2 = 0.984). All coefficients are statistically significant with a p-value < 0.05. The tolerance less than 0.1 attests to the absence of multicollinearity. The condition index (I = 2.70) is below the critical threshold of 30, which confirms the numerical stability of the model. In addition, the very low prediction error (MAPE = 0.06%) reflects the reliability of the proposed multiple linear regression model. All these results offer solid indications for the understanding of regional oceanographic dynamics and factors governing the variability of salinity in this area.*

Keywords: *Salinity, Granger causality test, multicollinearity, multiple linear regression.*

I. INTRODUCTION

The maritime sector, like the global community as a whole, is placing increasing importance on ocean change. This issue is of specific importance for the countries located around the Mozambique Channel, due to the complicity of oceanographic phenomena present in this area (Charles,C. et al 2020).

Understanding the dynamic processes at work in this region requires the analysis of key climatic and oceanic parameters. Among these variables, the salinity of the sea plays a key role.

In this study, we have the salinity as the main variable to examine its interactions with sea surface height (SSH), sea surface temperature (SST), marine currents and wave height, within the framework of heat exchanges at the surface, eddies and the dynamics of marine ecosystems.

We illustrate our analyses by the multiple linear regression approach.

II. MATERIAL AND METHODS

2.1. Presentation of the study area

Our study area extends between 16°S and 8°S in latitude, and between 40°E to 50°E in longitude, thus covering the North of the Mozambique Channel. This region notably includes water surrounding the archipelago of Comoros as well as the northwest coast of Madagascar. This region constitutes a strategic maritime zone due to

 Received-05-09-2025 Accepted- 13-09-2025

its complex oceanographic dynamics, which are driven by the interplay of currents, mesoscale eddies, and regional climate variability (Charles,C. et al 2020) (Obura, D. O. 2015).

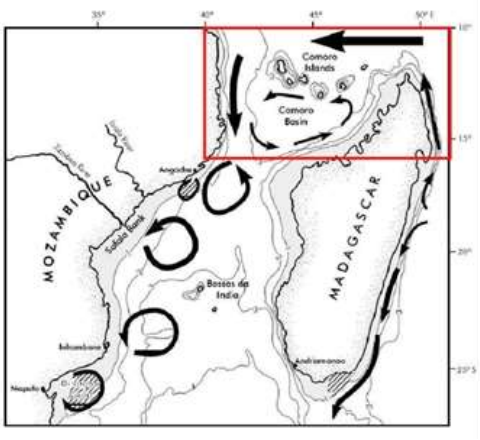

Figure 1: Study area

2.2. Presentation of study data

The data used in this study come from the sites Copernicus climate data store with a step of 0.25 and Copernicus marine data store with a step of 0.083. They include the following variables, all available as three-dimensional grids (longitude, latitude and time).

The datasets for this analysis were sourced from the following repositories: the Copernicus Climate Data Store (CDS) with a spatial resolution of 0.25 degrees, and the Copernicus Marine Data Store (CMDS) with a spatial resolution of approximately 0.083 degrees. All variables were retrieved as three-dimensional arrays spanning longitudinal, latitudinal, and temporal dimensions.

- Sea surface temperature (SST) in Kelvin (K)
- The height of the sea surface (SSH) in meter (m)
- Sea surface salinity (SSS) in PSU,
- Wave speed in m/s
- The marine current (at the sea surface) in m/s
- The height of the wave in meters (m)
- The atmospheric pressure (Pa) taken at sea level
- Zonal wind (U, in m/s): component of the wind in the east-west direction.
- Southerly wind (V, in m/s): component of the wind in the north-south direction.

These are daily series taken from January 1, 1990 to December 31, 2020 in netcdf (nc) form.

2.3. Data preprocessing

Before integrating the variables into the multiple linear regression model, we carried out several preprocessing steps to ensure the reliability of the analyses, notably the calculation of climatological averages, the stationarity test, differentiation, and the causality test.

2.3.1. Calculate daily climatological spatiotemporal averages

To mitigate seasonal fluctuations and have exploitable long-term climate trends, we calculated daily climatological averages. These averages are calculated considering both the spatiotemporal dimensions of study data ranging from 1993 to 2020 (Niriko , H. et al, 2025).

$$\bar{X} = \frac{1}{N_{lat} N_{lon} N_t} \sum_{i=1}^{N_{lat}} \sum_{j=1}^{N_{lon}} \sum_{t=1}^{N_t} X_{i,j,t}$$

Where $\overline{X}$: the daily space-time average

N_{lat} , N_{lon} : Spatial dimension

N_t : Total number of time periods

$X_{i,j,t}$: Value of the spatiotemporal variable

2.3.2. Stationarity test

Following the calculation of the means of the spatio-temporal variables, we used the Dickey-Fuller Augmented (ADF) test to verify the stationarity of our time series.

It is a widely recognized statistical tool for estimating the presence of a unit root in a time series which is a key indicator of the non-stationarity of the series.

Hypotheses:

H_0: the series is non-stationary.

H_1: the series is stationary.

2.3.3. Differentiation

The Dickey-Fuller Augmented (ADF) test previously performed proves the non-stationarity of our data (p-value > 0.05). Therefore, the series are differentiated up to order 2 to ensure their stationarity

This differentiation operation stabilizes the average of the series by eliminating trends, thus making the data suitable for more advanced statistical analyses.

2.3.4. Granger's causality

The Granger causality test is a statistical test used in time series analysis to establish whether there is a causal relationship between two variables (Oudra M, A. et Dada, I. 2019).

Hypothesis:

H_0: X does not cause Y

H_1: X has a causal effect on Y

For the model, we select only variables with significant relationships (p-values less than 0.05).

2.4. Analysis of the linear dependence between explanatory variables

In this phase, we will study the linear dependence between the explanatory variables used for our model. To do this, five complementary approaches were used:

2.4.1. Correlation matrix

The correlation matrix R allows a first visualization of linear links between variables. High absolute value correlation coefficients (> 0.7) may indicate potential redundancy; this is determined by (R. Bourbonnais 2015):

$$R_{i,j} = \frac{cov\left(X_i X_j\right)}{\sigma_{X_i}\sigma_{X_j}}$$

Where $R_{i,j}$: The correlation matrix

$\text{cov}(X_i, X_j)$: Covariance between variables X_i and x_j

$\sigma_{X_i}\sigma_{X_j}$: Standard deviations of the variables X_i and X_j

2.4.2. Auxiliary regressions (between explanatory variables) and tolerance calculations

Each explanatory variable is regressed on the others, making it possible to evaluate its auxiliary R^2, from which we deduce the tolerance by (Berrouyne, M.):

$$Tol = 1 - R^2$$

A tolerance below 0.1 is generally considered an indicator of critical multicollinearity.

Tolerance	R^2 auxiliary	Level of multicollinearity	Interpretation
> 0.25	< 0.75	Low	No worries
0.1 – 0.2	0.8 – 0.9	Moderate	To be watched, but often acceptable
< 0.1	> 0.9	Higher	Serious problem: very redundant variable

Table 1: Tolerance interval (Belzile, L.,)

2.4.3. KLEIN test

This is not a test per se but rather a simple indicator to quickly identify problematic situations. It is based on the comparison of R^2 main and R^2 auxiliary.

The hypotheses according to Klein:
If R^2 main >R^2 auxiliary: no multicollinearity
If R^2 main <R^2 auxiliary: existence of multicollinearity
In case of collinearity, we eliminate the highly correlated variables.

2.4.4. Farrar–Glauber test

Farrar and Glauber (in 1968) formalized a multicollinearity test. It is a more rigorous statistical test based on the correlation matrix, more particularly from its eigenvalues.
The first step is to calculate the determinant of the matrix of correlation coefficients between the explanatory variables, then apply the hypotheses in the second step and finally transform the determinant of R (R. Bourbonnais 2015).
The test statistic follows a law of χ^2 and allows to judge if the set of variables is linearly interdependent. Its formula is:

$$\chi^2 = -(n-1-\frac{2p+5}{6})\ln D$$

Where *p*: number of variables
n: number of days
D: determinant of the correlation matrix *R*
The Farrar-Glauber test governed with the following assumptions:
H_0: *D* = 1, there is no collinearity
H_1: *D* < 1, there is a presence of collinearity

2.4.5. The condition index

Another approach used to detect a multicollinearity problem is to analyze the condition index. This condition index is equal to the square root of the ratio between the highest eigenvalue λ_{max} and the lowest eigenvalue λ_{min} (Berrouyne, M) (Belzile, L.,).

$$I = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

When this ratio is greater than 30, it indicates strong multicollinearity.

Condition index	Level of multicollinearity	Interpretation
< 10	Low / negligible	No action required
10 – 30	Moderate	Pay attention to certain variables
> 30	High / severe	Serious risk: redundancy, instability of the model
> 100	Very severe	Model probably unstable, to be corrected

Table 2: Interval of the condition index (Faster Capital.)

2.5. Multiple linear regression

Multiple linear regression is a statistical method used to describe and model the variations of an endogenous variable (variable to be explained) associated with the variations of several exogenous variables (explanatory variables) (Berrouyne, M.).

It is written in the form: $y = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \varepsilon$

Where: $\beta_0, \beta_1, \beta_2 ..... \beta_n$: parameters to be estimated

$X_1, X_2, X_3 .... X_n$: Explanatory variables

ε : Riesidus

Then, I collected and categorized the data. I certainly compared data of 3 different tools mentioned above.

III. RESULT

After the different stages of analysis and processing of our data, we present the results.

3.1. Causality

The analysis of the causality test applied to oceanic variables makes it possible to distinguish two types of dynamic relationships:

- ***A unidirectional causality***

Salinity → ***pressure*** *(p-value = 0): the pressure causes the salinity*

Wind U → ***salinity*** *(p-value = 0): the wind U has a significant unidirectional causal effect towards salinity.*

Wind V → ***salinity*** *(p-value = 0): there is a significant causality of wind V to salinity.*

Salinity has no causal effect on pressure and wind (U and V), but on the other hand these variables have causal effects on salinity.

- ***A bidirectional causality***

Salinity ↔ ***marine current*** *(the P-values = 0): salinity and marine current have a bidirectional causal relationship.*

SSH ↔ ***salinity*** *(p-values = 0), salinity and SSH have a bidirectional causal relationship.*

SSS ↔ ***SST*** *(the p-values = 0): salinity and temperature are mutually causal*

Wave ↔ ***salinity*** *(p-values = 0; 2.259e-09): salinity and wave have a bidirectional causal relationship.*

- ***Variables selected***

After the Granger causality test, we take only those variables that have a bidirectional causal relationship with salinity (the SSH, the SST, the marine current and the wave).

These variables show significant causal relationships, making the model relevant.

3.2. Analysis of the linear dependence between explanatory variables

Using the Granger causality test, we selected variables with a significant bidirectional relationship with salinity (SSH, SST, wave height and sea current). A redundancy analysis was then conducted to detect the possible presence of multicollinearity between the explanatory variables of the multiple linear regression model. Multicollinearity can indeed lead to an instability of the estimated coefficients, as well as a loss of statistical significance.

3.2.1. Correlation between explanatory variables

Our correlation matrix (fig.2) shows links between the explanatory variables, more specifically between SST and other variables, with coefficients ranging up to.

Figure 2: Correlation matrix

3.2.2. KLEIN test

The empirical test of Klein is applied to compare the determination coefficient (R^2) of the main model (salinity ~ SSH + SST + wave + current) with those of the auxiliary regressions (each explanatory variable regressed on the others).

R^2 main: 0.9846

The results are represented in table 3.

Explanatory variables	R^2 auxiliary	Interpretation according to Klein
SSH	0.2674	No multicollinearity detected
SST	0.6684	No multicollinearity detected
Wave	0.6963	No multicollinearity detected
Current	0.5578	No multicollinearity detected

Table 3: Auxiliary determination coefficients of the explanatory variables

No auxiliary R^2 exceeds the main R^2, which highlights the absence of redundancy concerning according to this criterion.

3.2.3. Farrar-Glauber test

The global Farrar–Glauber test reveals a χ^2 statistic of 745.73 with a null p-value, confirming the presence of significant global multicollinearity between explanatory variables.

3.2.4. Auxiliary regressions and tolerance calculations

Auxiliary regressions allowed to calculate the R^2 and the tolerance of each explanatory variable:

Variables	R^2 auxiliary	Tolerance =1–R^2	Interpretation
SSH	0.267	0.733	Acceptable
SST	0.668	0.332	Moderate
Wave	0.696	0.304	Moderate
Current	0.558	0.442	Acceptable

Table 4: Auxiliary R^2 results and tolerance

All tolerances are greater than 0.25, multicollinearity is excluded. SST and Wave variables show a moderate correlation, but they remain within an acceptable range for multiple linear regression.

Analysis of the correlation matrix, Klein test, Farrar-Glauber test, tolerances and condition index indicate moderate but tolerable multicollinearity in the model. No questioning of the validity of the multiple linear regression model, although the SST variables and the wave show a more marked interdependence with the other variables. These results allow the model to be maintained while taking particular vigilance in the interpretation of the coefficients associated with SST and wave.

3.3. Multiple linear regression model

Once the absence of severe multicollinearity was validated, surface salinity was modeled using a multiple linear regression taking into account the following explanatory variables: sea surface height (SSH), surface temperature (SST), height of the waves and sea current.

The estimated model is of the form:

$$salinity = \alpha_0 + \alpha_1 \times SSH + \alpha_2 \times SST + \alpha_3 \times wave + \alpha_4 \times current$$

With α_0 : the intercept

$\alpha_1, \alpha_2, \alpha_3 and \alpha_4$: Coefficients estimated consecutively for SSH, SST, wave and sea current.

The following table 5 shows the estimated coefficients for each variable included in our model as well as their associated statistics, such as:

Estimated value: it refers to the effect of each independent variable on salinity

Standard error: measures the variability of the coefficient estimate

Statistical test: allows to experiment if a coefficient is significantly different from zero

P-values: shows the degree of statistical significance of any variable.

A value P less than 5% is usually set as significant.

Coefficients	Estimation	Standard error	T-stat	P-Value
α_0	35.207	0.080835	435.54	0
α_1(SSH)	10.015	0.10242	97.783	0
α_2(SST)	-0.1576	0.001989	-79.235	0
α_3(Wave)	-0.10589	0.013966	-7.5819	0
α_4(Current)	0.72472	0.043384	16.705	0

Table 5: Estimation coefficients of the explanatory variables

3.3.1. Coefficient analysis

All coefficients are statistically significant (p-value = 0).

The model indicates that:

SSH (sea level) and currents contribute to increasing salinity (positive coefficient).

SST (temperature) and waves have a reducing effect on salinity (negative coefficient).

This can be interpreted physically:

A sea level rise (SSH) can indicate an influx of more salty water.

High temperatures can promote dilution (e.g., melting ice, precipitation).

Sea currents can carry salt water from other areas.

The waves, via mixing or the supply of fresh water on the surface (rain), can locally reduce the salinity.....

3.3.2. Importance of explanatory variables

Figure 3 illustrates the relative influence of different explanatory variables on salinity variability, based on the absolute value of their coefficients. We note that SSH clearly dominates the whole, its effect being much more marked than that of other factors. On the other hand, sea surface temperature (SST), wave height and sea current appear as less significant variables, with relatively moderate impacts.

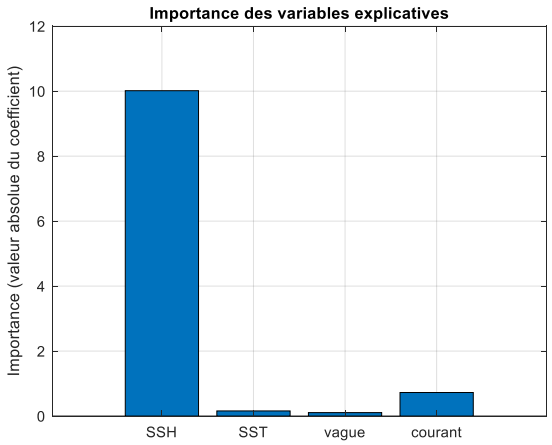

Figure 3: Importance of explanatory variables

3.4. Residue analysis

The analysis of residues constitutes an essential step in the evaluation of the quality and robustness of our model. The residues, defined by $\varepsilon = y_i - \hat{y}_i$ represent the difference between observed values and predicted values (Berrouyne, M).

With: y_i : Actual values

$\hat{y}_i$: Predicted values

They allow to check if the model corresponds well to the data and to identify possible errors or unusual behaviors. To this end, two graphical analyses were mobilized: the presentation of point clouds thus ensuring the homogeneity of the error variance and the independence of the residuals and the histogram of the residuals to examine their distribution.

3.4.1. Point cloud presentation

Residuals appear globally centered around zero, which is expected in a well-fitting model. There is no obvious linear relationship between the residuals and observations, indicating an absence of systematic bias.

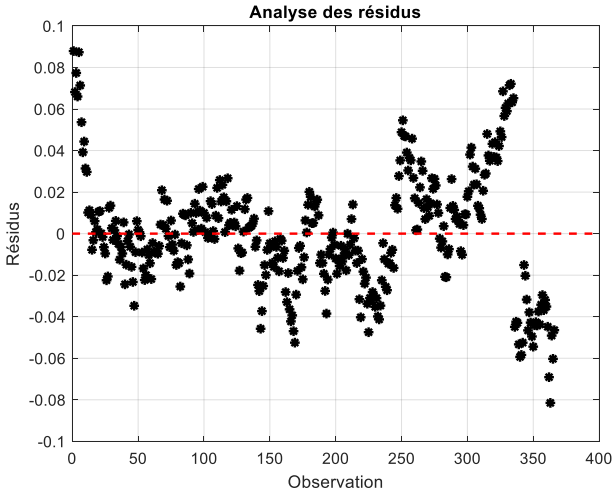

Figure 4: Residue representation

3.4.2. Residue normality test

The histogram is centered around 0 having a residue between -0.08 and 0.08. Its general form is symmetrical and bell-shaped, which drives the normal approximative of residues (fig.5).

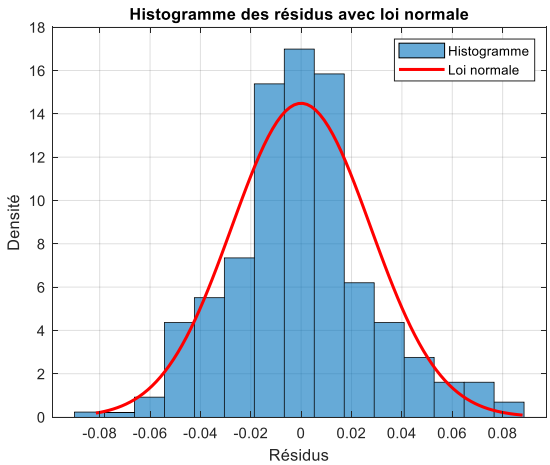

Figure 5: Residue Histogram

3.4.3. Standardized Kolmogorov-Smirnov test

The MAPE (Mean Absolute Percentage Error) is one of the most widely used metrics for model prediction accuracy. It measures the average magnitude of the error produced by the model or the average difference between the predictions. It is determined by (C. Davide, J. W. Matthijs et J. Giuseppe 2021) (Hébert-pinard, C. 2023):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

With: y_i : Actual values

$\hat{y}_i$: Forecast values

n: Observation numbers

In order to have an overall idea of the quality of the linear fit, we define R^2 the determination coefficient and its adjustment for measuring the part of the total variation of *Y* explained by the regression model on *X*.
In our study, R-Squared (R^2) 0.985 and Adjusted R-Squared 0.984 are substantially equal to 1, the points are aligned on the line, the linear relation explains a reasonable adjustment.
MAPE = 0.06% (<10%), forecast errors are small in proportion to actual values, which makes model estimates reliable (C. Davide, J. W. Matthijs et J. Giuseppe (2021) (Hébert-pinard, C. 2023).

3.4.4. Figure representation from the model

Figure 6 represents the curve derived from the model and the values studied.
Predicted values curve (blue): it represents the estimation of the values from our model.
Curve of observed values (red): it represents the measured values at each time interval.

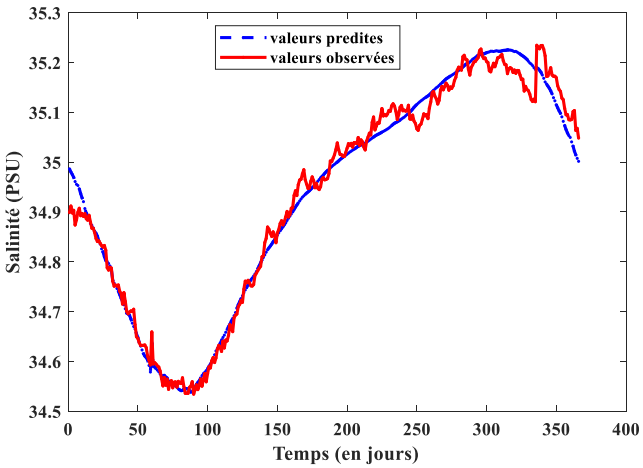

Figure 6: Representations of the model

This representation shows the performance of the salinity prediction model over time, which gives a well-adapted agreement between the model and the actual parameters.

IV. DISCUSSION

The model noted the satisfactory overall results of multicollinearity We obtained a condition index of 2.70 below the critical threshold 30, which attests to the numerical stability of the model. The presence of moderate multicollinearity is detected by the different tests conducted, however it does not compromise the validities of the estimates.
This interdependence remains between the explanatory variables more precisely between SST and wave height (r = -0.78) and SST and marine current (r = -0.71), they are below the critical thresholds conventionally accepted.
The Klein test showed no significant redundancy between the explanatory variables, and the calculated tolerance values indicate a low multicollinearity.
It is observed that all tolerances are above the critical threshold of 0.25, which allows excluding the presence of a worrying multicollinearity in the model.
The Farrar-Glauber test shows that there is no excessive dependence between the predictors. This confirms that multicollinearity has no significant effect on the model. Combined with tolerance, this diagnosis strengthens the reliability of regression estimates.
In our analysis, the condition indices are well below the critical threshold of 30. This shows that multicollinearity does not present a problem and does not hinder the interpretation of the coefficients. Consistent with the Farrar-Glauber test, as well as with tolerance, the condition index confirms the statistical strength of the model and the relevance of the variables selected.

V. CONCLUSION

We modeled the variation of the daily mean salinity in our study area using multiple linear regression. We proceeded with the causality test for the selection of explanatory variables used in the model, followed by the other treatments, namely multicollinearity, for the confirmation of these variables.

These processes minimize inconsistencies to have a statistical indicator (R^2= 0.985), and adjusted statistical index Adjusted R-Squared R2 adjusted = 0.984 reflecting a low value of the residuals.

The percentage of error is estimated by MAPE = 0.06%, which affirms the good accuracy of the model.

The results of our model offer precise quantitative and qualitative indicators to improve the analyses of different climate variables in our study area.

VI. REFERENCES

1. Charles,C. et al (2020). «Intermediate and deep ocean current circulation in the Mozambique Channel: New insights from ferromanganese crust Nd isotopes» Marine geology, vol. 420, n° %1106356, pp. 1-13.
2. Obura, D. O. (2015). The Northern Mozambique Channel, A Background Document: WWF International and CORDIO East Africa.
3. Niriko , H. et al,(2025). «Modélisation de la hauteur de vague au Sud et Sud-Est de Madagascar,» IJPSAT, vol. 49, pp. 46-56.
4. Oudra M, A. et Dada, I. (2019). «Cointégration et Causalité entre Gouvernance et Croissance Économique : Cas du Maroc,» Revue du Contrôle de la Comptabilité et de l'Audit, vol. 4, n° %12, pp. 260-296.
5. R. Bourbonnais (2015). Économétrie : Cours et exercices corrigés, 5 rue Laromiguière, 75005 Paris: DUNOD.
6. Berrouyne, M. Analyse de la regression support de cours Aspects théorique et pratique, I NSEA,Avenue allal al Fassi B.P : 6215: Rabat instituts.
7. Belzile, L.,«github,» [En ligne]. Available: https://lbelzile.github.io/MATH60604-diapos/MATH60604_d2h_colinearite.pdf. [Accès le 16 Août 2025].
8. Faster Capital. [En ligne]. Available: https://www.fastercapital.com/content/Condition-Index--Conditioned-for-Analysis--Using-the-Condition-Index-to-Measure-Multicollinearity.html. [Accès le 16 Août 2025].
9. C. Davide, J. W. Matthijs et J. Giuseppe (2021). «The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, » Camputer Science, n° %1623, p. 18.
10. Hébert-pinard, C. (2023). Analyse de l'impact des variables météorologiques sur la prévision de la demande énergétique au Québec, Montréal: Université du Québec.