Temporal Analysis and Prediction of Malaria Dynamics Using Meteorological Data in Southeast of Senegal: Suplementary informations

March 26, 2025

Contents

S1 Statistical description of meteorological data

The meteorological data used in this study were drawn from NASA platform from 2018 to 2022. These data were aggregated into weekly counts.

Variables	Mean	St.dev	Q1	Median	Q3	Min	Max
PRECTOTCORR : Weekly rainfall(mm)	4.6	8.3	0.0	0.1	6.5	0.0	56.1
days_with_precipitation : number of rainy days per week	3.3	3.2	0.0	2.5	7.0	0.0	7.0
WS10M_MAX : weekly mean of daily maximum wind speed at 10m (m/s)	5.0	2.0	4.1	5.1	5.9	2.5	8.1
WS10M_MIN :weekly mean of daily minimum wind speed at 10m(m/s)	1.6	0.6	1.1	1.5	2.1	0.6	3.5
WS10M: weekly mean of daily average wind speed at 10m (m/s)	3.2	0.8	2.5	3.3	3.7	1.6	5.2
WD10M :Weekly average of wind direction	168.7	63.9	101.1	187.6	223.9	53.6	268.7
T2M :weekly mean of daily average temperature (°C)	28.2	3.5	25.8	26.8	31.1	22.2	36.1
T2M_MIN :weekly mean of daily minimum (°C)	22.4	3.2	22.8	23.1	24.1	14.2	28.9
T2M_MAX : weekly mean of daily maximum temperature (°C)	34.7	4.9	30.2	33.5	39.9	28.0	44.1
TS : Weekly average of the Earth's surface temperature (°C)	29.0	4.6	25.8	26.8	32.8	22.1	38.9
PS : weekly mean of daily average atmospheric pressure (hPa)	99.4	0.15	99.3	99.5	99.5	98.9	99.7
QV2M : weekly mean of daily average specific humidity (%)	12.2	5.9	6.4	12.4	18.3	2.1	19.4
RH2M : weekly mean of daily average relative humidity (%)	54.7	28.9	23.1	61.2	85.0	10.2	91.3

Table 1: List of meteorological variables with their abbreviations and descriptive statistics. St.dev is the standard deviation, Min is the minimum, Q1 is the first quartile,Q3 is the third quartile dans Max is the maximum.

S2 Correlation analysis of meteorological variables

Correlations between variables play an important role in a descriptive analysis. The correlation matrix presented above was obteint by using Pearson correlation coefficient.
The correlation coefficients are represented by a color palette ranging from blue (positive correlations) to red (negative correlations). Values close to 1 are shown in blue hues, representing strong positive correlations, while values close to -1 are colored red, indicating strong negative correlations. Values close to 0 are represented in gray, signaling an absence of significant linear correlation.

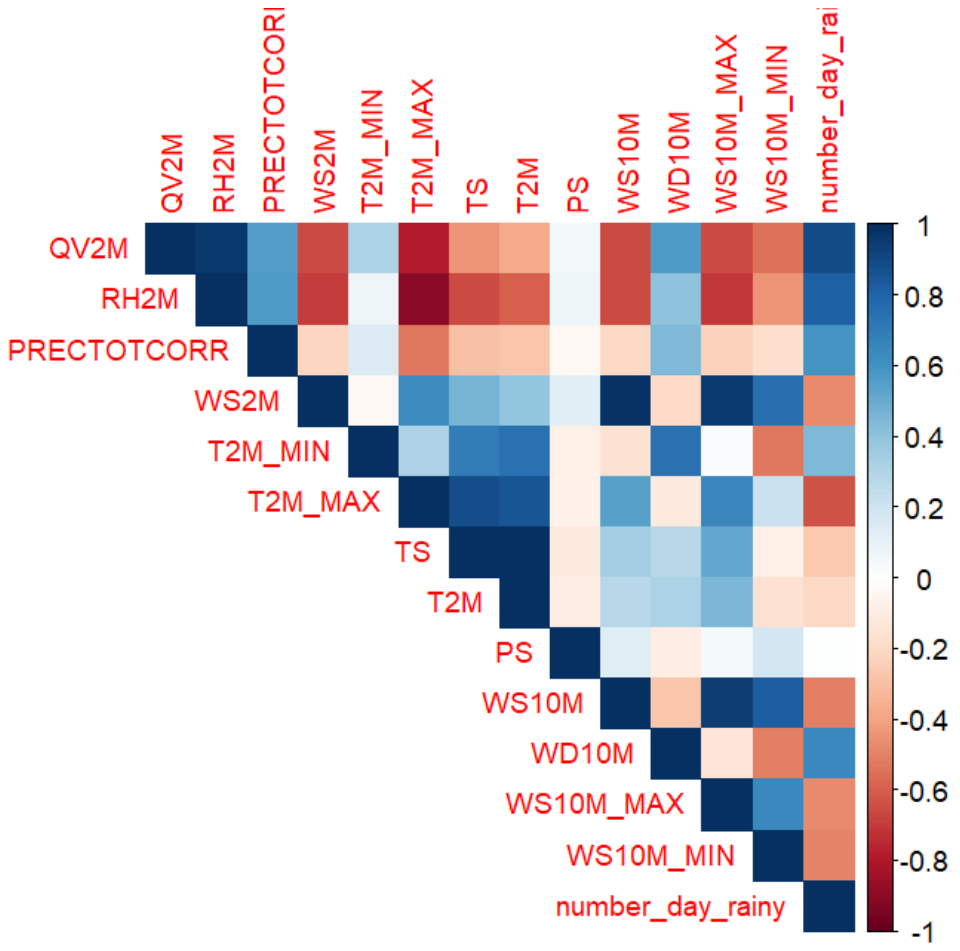

Figure 1: Correlation matrix of the meteorological variables.

S3 Principale Componsent Analysis

The table below presents the contributions of each variable as well as the correlation coefficients with the principal components derived from the principal component analysis.we have the coefficients that indicate the association between a variable and the principal component, the sign indicates the direction of this association The first principal component is associated among other things with a high maximum temperature (T2M_MAX) but also with high maximum wind speeds; Furthermore, it is associated with low relative and specific humidity and low precipitation frequencies The second component is first associated with high minimum temperatures (at 10m) and the earth's surface temperature and positively associated with the wind direction The third component is mainly associated with an increase in precipitation intensity.

Variables	Dim.1		Dim.2		Dim.3	
	Contribution	correlation	Contribution	correlation	Contribution	correlation
PRECTOTCORR	4.23 %	-0.55	0.43%	0.13	19.61%	0.59
WS10M _MAX	10.12%	0.85	0.07%	-0.05	12.52%	0.47
WS10M_MIN	4.48%	0.57	11.44 %	-0.67	8.89 %	0.40
T2M	5.43%	0.62	15.26%	0.77	0.06%	-0.03
T2M_MIN	0.01%	-0.02	23.81%	0.96	1.78%	0.18
PS	3.47%	-0.50	11.09%	-0.66	3.85%	0.26
QV2M	11.66%	-0.91	1.61%	0.25	2.79%	0.22
T2M_MAX	11.61%	0.91	3.14%	0.35	1.38%	-0.16
WS10M	9.22%	0.81	1.76%	-0.26	14.68%	0.51
WS2M	9.98%	0.85	0.45%	-0.13	14.31%	0.50
days_with_precipitation	8.71 %	-0.79	3.06%	0.34	9.07%	0.40
WD10M	1.64 %	-0.34	14.57 %	0.75	9.10%	0.40
TS	6.53%	0.68	13.29%	0.72	0.01%	-0.01
RH2M	12.91%	-0.96	0.01%	0.02	1.97%	0.19

Table 2: Table showing the contributions of each variable and the correlation coefficients with the principal components derived from the principal component analysis.

S4 Cluster Analysis and meteorological variable variation based on PCA Results

After performing principal component analysis (PCA), a clustering analysis was applied to the PCA results. The graphs below illustrate the variation of the variables across the different identified clusters

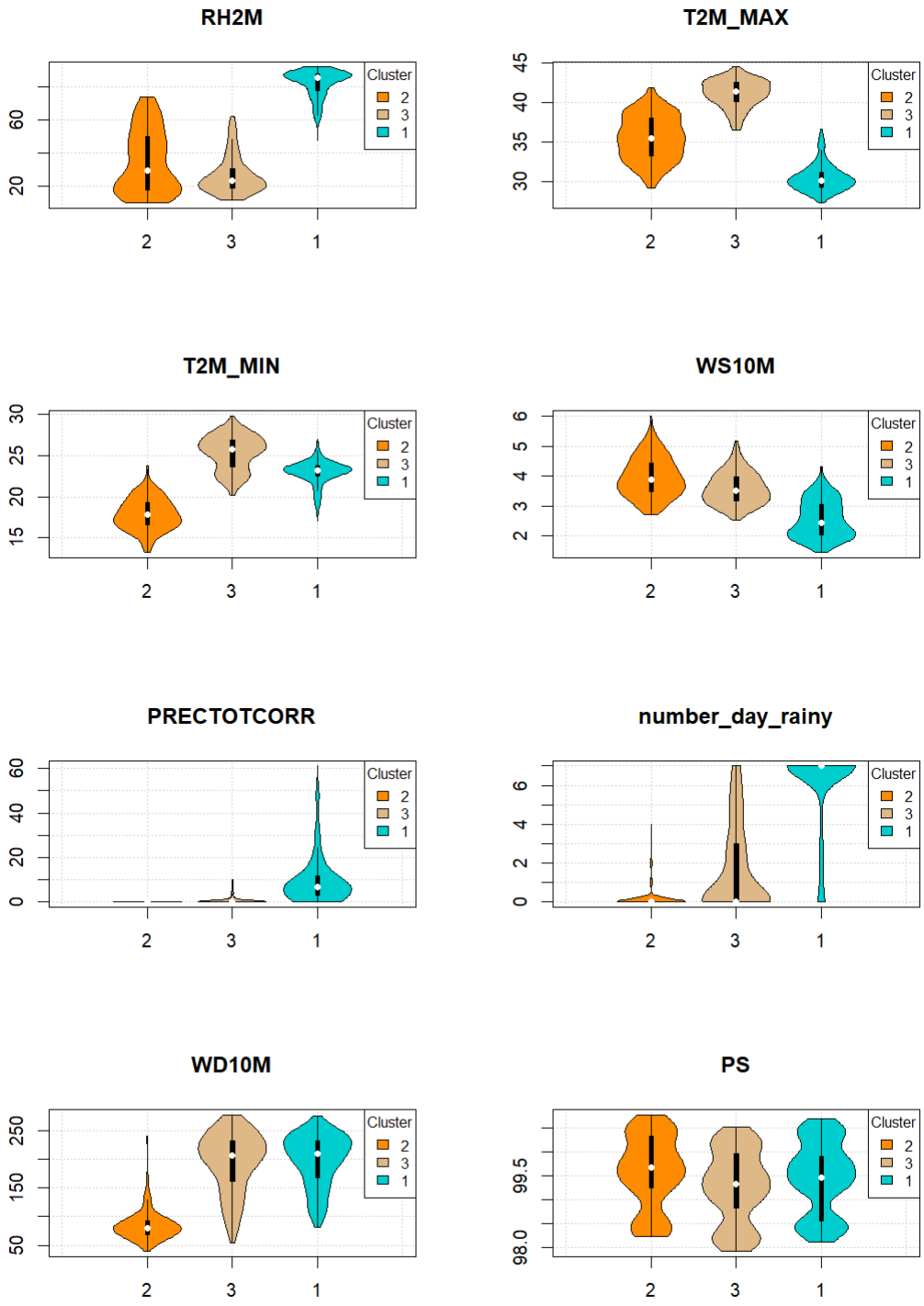

Figure 2: Variation on environnemental variables according cluster

S5 Anova test analysis

	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$
cluster & Dim1	2	5845	2922.5	2401	$< 2e-16$***
cluster & Dim1	2	2751.7	1375.8	1659	$< 2e-16$***
cluster & Dim1	2	45.6	22.809	13.22	$2.15e-06$ ***

Table 3: Anova result

Tuckey test analysis

DIM1				
Clusters	diff	lwr	upr	p adj
2-1	4.5790133	4.37872784	4.7792988	0.0000000
3-1	4.8893094	4.69912368	5.0794951	0.0000000
3-2	0.3102961	0.08975458	0.5308376	0.0028487
Dim2				
cluster	diff	lwr	upr	p adj
2-1	-2.590559	-2.755877	-2.425242	0
3-1	1.870731	1.713750	2.027712	0
3-2	4.461290	4.279253	4.643327	0
Dim3				
Clusters	diff	lwr	upr	p adj
2-1	-0.1673381	-0.4058430	0.07116693	0.2265770
3-1	-0.4955900	-0.7220680	-0.26911210	0.0000010
3-2	-0.3282520	-0.5908783	-0.06562560	0.0095853

Table 4: Tukey test analysis

S6 Model Summary and Statistical Overview

For each model, we provide the summary tables of the GAM models, including the linear and nonlinear effects, with the associated p-values for each term. The p-values help assess the statistical significance of the effects, allowing us to determine which variables have a significant impact on the model results.

S6.1 Model with Principal Components as predictors (Model1)

Table 5: Parametric coefficients

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.278e+00	3.963e-02	-32.249	$< 2e-16$ ***
POPULATION	-1.297e-06	8.287e-07	-1.566	0.117848
DISTRICTKedougou	-1.633e-02	5.536e-02	-0.295	0.768032
DISTRICTSalemata	-1.337e-01	3.852e-02	-3.471	0.000549 ***
DISTRICTSaraya	1.821e-01	2.437e-02	7.469	2.22e-13 ***

Table 6: Approximate significance of smooth terms

	edf	Ref.df	F	p-value
s(MONTH)	8.3820478	10	11.606	$< 2e-16$***
s(lag2_dim1)	5.0506258	9	14.417	$< 2e-16$***
s(lag2_dim2)	8.4246108	9	6.674	$< 2e-16$ ***
s(lag14_dim3)	0.0005609	9	0.000	0.68
s(Year)	2.6854890	3	12.230	$< 2e-16$ ***

S6.2 Model with Variables Selected from Principal Components (model2)

Table 7: Parametric coefficients

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$-1.254e^{+00}$	$4.105e^{-02}$	-30.560	$< 2e-16$ ***
POPULATION	$-1.340e^{-06}$	$8.451e^{-07}$	-1.586	0.11318
DISTRICTKedougou	$-2.471e^{-02}$	$5.644e^{-02}$	-0.438	0.66168
DISTRICTSalemata	$-1.532e^{-01}$	$4.084e^{-02}$	-3.750	0.00019 ***
DISTRICTSaraya	$1.663e^{-01}$	$2.567e^{-02}$	6.477	$1.68e^{-10}$ ***

Table 8: Approximate significance of smooth terms

	edf	Ref.df	F	p-value
s(MONTH)	8.187	10	9.850	$< 2e-16$ ***
s(lag2_RH2M)	5.645	9	14.894	$< 2e-16$***
s(lag2_T2M_MIN)	4.275	9	2.183	0.000171 ***
s(lag14_PRECTOTCORR)	5.007	9	1.242	0.037313 *
s(Year)	2.879	3	15.129	$< 2e-16$ ***

S6.3 Model with Principal Components and Interactions (model3)

Table 9: Parametric coefficients

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$-1.274e^{+00}$	$3.999e^{-02}$	-31.871	$< 2e-16$ ***
POPULATION	$-1.601e^{-06}$	$8.457e-07$	-1.893	0.05872 .
DISTRICTKedougou	$1.024e^{-03}$	$5.649e^{-02}$	0.018	0.98554
DISTRICTSalemata	$-1.333e^{-01}$	$3.859e^{-02}$	-3.455	0.00058 ***
DISTRICTSaraya	$1.891e^{-01}$	$2.463e^{-02}$	7.679	$4.98e-14$ ***

Table 10: Approximate significance of smooth terms

	edf	Ref.df	F	p-value
s(lag2_dim1)	5.4940216	9	15.370	$< 2e^{-16}$***
s(lag2_dim2)	8.2523970	9	6.407	$< 2e^{-16}$ ***
s(Dim3):cluster1	0.9771530	9	0.333	0.0444 *
s(Dim3):cluster2	1.2052057	9	0.286	0.1026
s(Dim3):cluster3	0.0001985	9	0.000	0.9726
s(MONTH)	8.3515596	10	11.263	$< 2e^{-16}$ ***
s(Year)	2.4673571	4	8.523	$< 2e^{-16}$ ***

S7 Model Comparison

To choose the best model among the three GAMs, you can compare the following criteria:the GCV score, adjusted R^2 and deviance explained. The optimal model will have the lowest GCV or Cp value (indicating better generalization), the highest adjusted R^2 (indicating better fit), and the highest deviance explained (reflecting the proportion of variability in the data captured by the model). The GCV score for the third model is much smaller than the GCV score for the first and the second model.This model explained 97.7% of the variance, with an explained deviance of 86.3%.

Models	GCV.Cp	Adjusted R^2	Deviance explained
Model 1	12.146	0.976	86.1%
Model 2	12.331	0.976	86%
Model3	12.068	0.977	86.3%

Table 11: Comparison table of models for each district, using the GCV.Cp,adjusted R^2 and explained deviance criterion to evaluate the model's performance and complexity

Comparing these different measures, we can conclude that model 3 is more suitable.
This model, built using principal components with an interaction between meteorological clusters.

S8 Model Evaluation and Diagnostics

To assess the quality of the best GAM model fit, we analyzed the correlation between the model residuals using Autocorrelation Function (ACF) as well as the Partial Autocorrelation Function (PACF), and also performed a gam.check to verify the residuals conformity with the model assumptions and ensure that the model has properly captured the temporal structures in the data.

S8.1 Model Evaluation and Diagnostics Based on ACF and PACF analysis

The analysis of the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) of the residuals helps to check for the presence of unmodeled residual autocorrelations. The ACF and PACF of the residuals show no significant correlations, it suggests that the residuals are independent, indicating that the model has effectively captured the relationships present in the data.

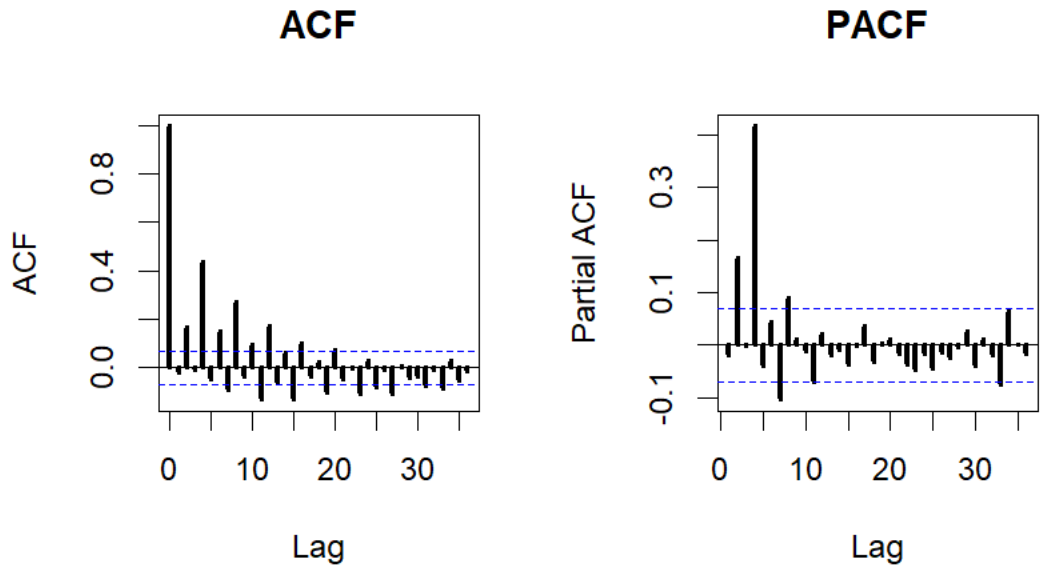

Figure 3: The ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) curves of the residuals of the best model.

S8.2 Model Evaluation and Diagnostics Based on gam.check

The table below presents the results from the gam.check function, including the k' (effective degrees of freedom), EDF (estimated degrees of freedom), k-index, and p-values for each smooth term in the model. These values help assess the model's fit, with the p-values indicating the significance of each smooth term.

	k'	edf	k-index	p-value
s(lag2_dim1)	9	$5.49e^{+00}$	1.00	0.60
s(lag2_dim2)	9	$8.25e^{+00}$	1.02	0.82
s(Dim3):cluster1	9	$9.77e^{-01}$	1.00	0.57
s(Dim3):cluster2	9	$1.21e^{+00}$	1.00	0.59
s(Dim3):cluster3	9	$1.99e^{-04}$	1.00	0.60
s(MONTH)	10	$8.35e^{+00}$	1.02	0.76
s(Year)	4	$2.47e^{+00}$	1.01	0.71

The gam.check also provides four plots:

- The QQ plot shows the normality of the residuals. The points should follow a straight line, indicating that the residuals are normally distributed.
- The Residual vs Fitted values plot shows the distribution of the residuals relative to the predicted values. A good fit is indicated by a random distribution of the residuals around zero.
- The Residual histogram visualizes whether the residuals follow a symmetric distribution, which is another sign of a good model fit.
- The Fitted values vs Response plot visualizes how the predicted values (fitted values) compare to the actual observed values (response). A good model fit will be indicated by points close to the identity line ($y = x$).

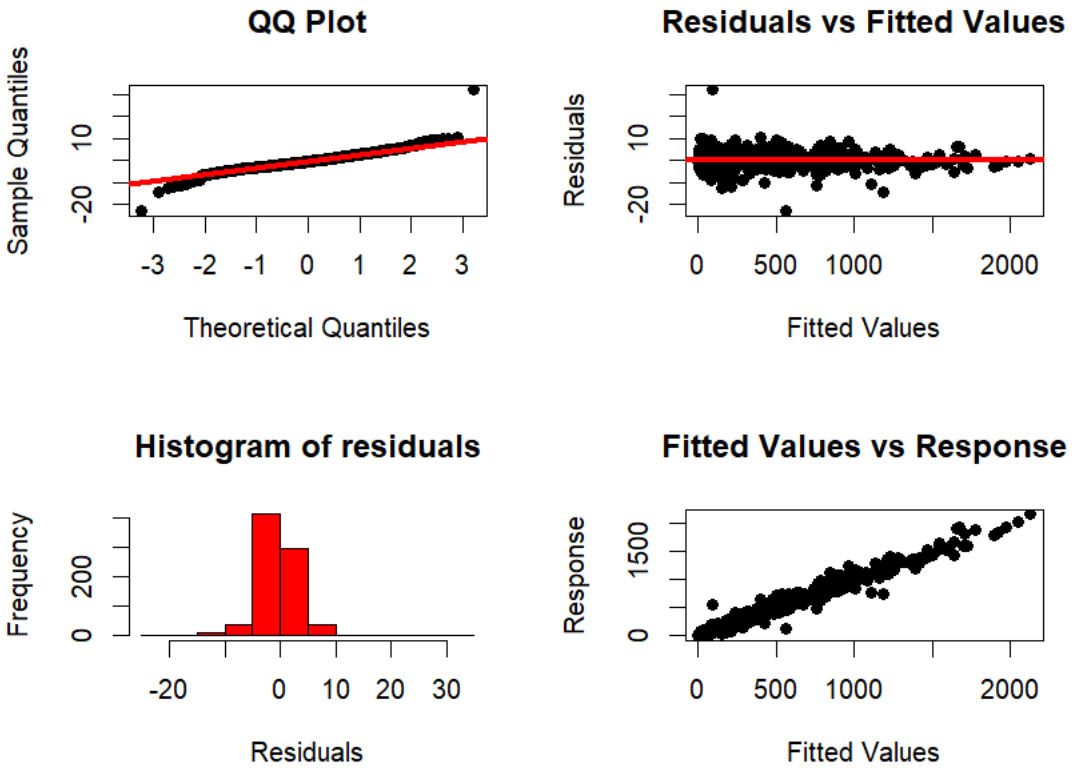

Figure 4: Diagnostic plots from the gam.check function).

These results, combined with the absence of residual correlations and the normal distribution of the residuals, suggest that the model is well-fitted and accurately captures the relationships in the data

S9 Evaluation of Prediction and Fit Errors for the Models

The table below presents the prediction errors for each model, using the mean absolute error (MAE) and root mean squared error (RMSE) metrics

Models	MAE	RMSE
Model1	278.1185	278.0591
Model2	278.205	278.1487
Model3	278.0191	277.9502

Table 12: Table of prediction errors for each model, measured using the mean absolute error (MAE) and root mean squared error (RMSE)