Supporting Information for "Statistical Modelling of the Annual Rainfall Pattern in Guanacaste, Costa Rica"

Rachel MacKay Altman, Ofir Harari, Nadya Moisseeva, Douw G. Steyn

1 Introduction

In this document, we provide details concerning the goodness-of-fit of our three models. We also provide estimates of the model parameters of secondary interest and plots of observed and predicted rainfall values that we did not include in the main manuscript.

2 Double Gaussian Model

2.1 Model diagnostics

Here we provide further details relating to the fit of the Double Gaussian model described in Section 3. Standardized residuals are computed as

$$r_{it} = \frac{\log(Y_{it} + 1) - \log(\hat{\lambda}_{it})}{\hat{\sigma}_t}.$$

Web Figure 1 shows the plots of the standardized residuals versus month of the year. The spread of the residuals is similar across months, especially in Nicoya, suggesting that our proposed variance structure is reasonable. However, the distribution of the residuals is clearly right-skewed (not normal) in the dry months. Web Figure 2, the plots of the residuals vs. the fitted values, tells a similar story—and reveals the large number of zero rainfall values (which form the diagonal pattern evident in the plots).

Web Figure 3 shows that autocorrelation in the residuals is small in magnitude and typically consistent with chance variation.

3 Regression Model with ARMA Errors

3.1 Model diagnostics

In this section, we describe the process and diagnostic tools that we used to specify the final form of the regression model with ARMA errors described in Section 3.

Figure 2 in the main manuscript shows the clear annual cycle of rainfall in Liberia and Nicoya. We treat this seasonality as deterministic, a choice that is supported by our exploration of an earlier model with stochastic seasonality that appeared to have unit roots.

Figure 2 also shows that the variance of monthly rainfall in month t of year i, Y_{it}, increases with its mean. Thus, we identify a transformation that stabilizes the variance. We consider models of the form

$$Y_{it}^{\lambda} = m_t + \gamma_1 \delta_{t1} x_{it1} + \gamma_2 \delta_{t2} x_{it1} + \gamma_3 \delta_{t3} x_{it1} + \gamma_4 x_{it2}/n + \epsilon_{it}, \tag{1}$$

We fit this model for different choices of λ using the method of least-squares (i.e., ignoring possible autocorrelation). We find that $\lambda = 0.25$ leads to a reasonably stable variance of the residuals for both the Liberia and Nicoya rainfall data. See the illustration of the residuals vs. month in Web Figure 4.

Next, we choose the model for ϵ_{it}. Web Figure 5 shows the estimated ACFs of the residuals when the ϵ_{it}'s are treated as independent. These plots suggest an AR(1) structure for the ϵ_{it}'s. We confirm the reasonableness of this choice by creating ACF plots of the residuals resulting from fitting this model (see Web Figure 6). That said, the model with independent and identically distributed errors also fits moderately well, reinforcing our argument that using the Double Gaussian model (which treats rainfall observations as independent conditional on knowledge of ONI) is a reasonable option in practice.

The QQ-plots of the residuals obtained by fitting the model with AR(1) errors (not included) shows, as in the case of the Double Gaussian model, that the distribution of the errors is reasonably approximated by the normal distribution in the wet months but not in the dry months.

3.2 Parameter Estimates and Predicted Values

Web Tables 1 and 2 provide the estimates of some of the parameters of the model in Web Equation 1 with AR(1) errors.

Web Figures 7–13 display the predicted, naive, and observed rainfall values over various time scales.

4 Tweedie Model

4.1 Model diagnostics

This section provides some plots that illustrate the goodness-of-fit of the Tweedie model presented in Section 3. Standardized residuals are computed as

$$r_{it} = \frac{Y_{it} - \hat{E}[Y_{it}]}{\widehat{SD}[Y_{it}]}.$$

Web Figure 14 presents the standardized residuals vs. month. The residuals are centred around 0, showing that the Tweedie model describes the mean rainfall each month very accurately. The residuals also have a relatively constant spread across months, suggesting that the variance structure implied by the Tweedie model provides a good description of the actual variance of the rainfall observations.

In Web Figure 15, we present the estimated autocorrelation functions (ACFs) of the standardized residuals of the Tweedie model. Interestingly, because we have included ONI in the model, the residuals are approximately uncorrelated. In constrast, when we exclude ONI, the residuals are clearly autocorrelated. In other words, ONI can essentially fully explain the autocorrelation in the monthly rainfall observations. As with the results based on the regression model with AR(1) errors, this finding serves to justify the assumption of the Double Gaussian model that rainfall observations are independent conditional on knowledge of ONI.

4.2 Parameter Estimates and Predicted Values

Web Tables 3 and 4 provide estimates of some the parameters of the Tweedie model.

Web Figures 16–22 display the predicted, naive, and observed rainfall values over various time scales.

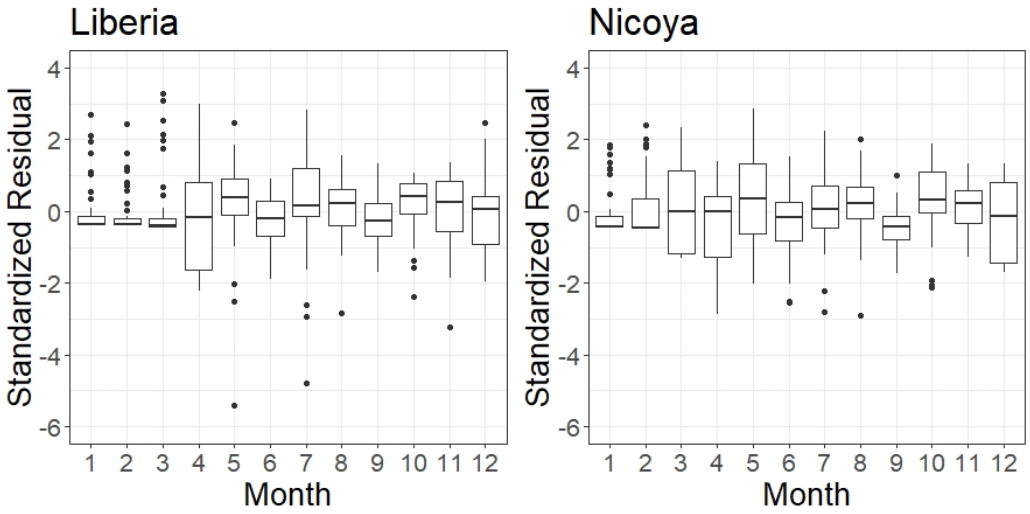

Figure S1: Standardized residuals versus month of the year for the DG model

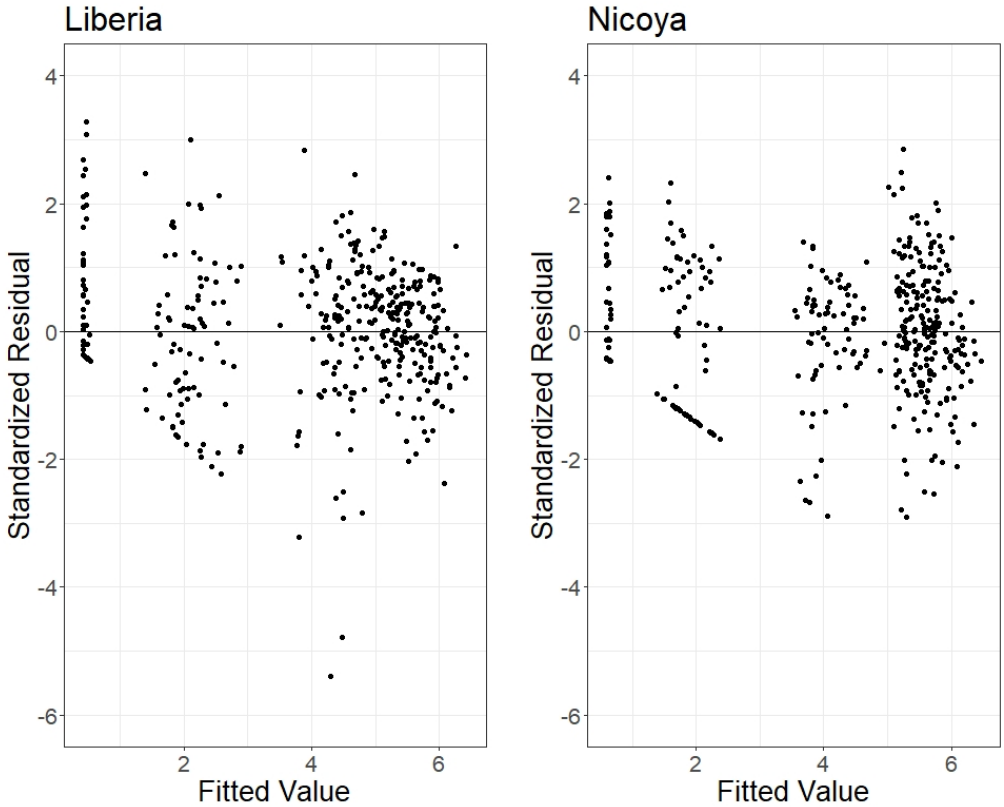

Figure S2: Standardized residuals versus fitted values for the DG model

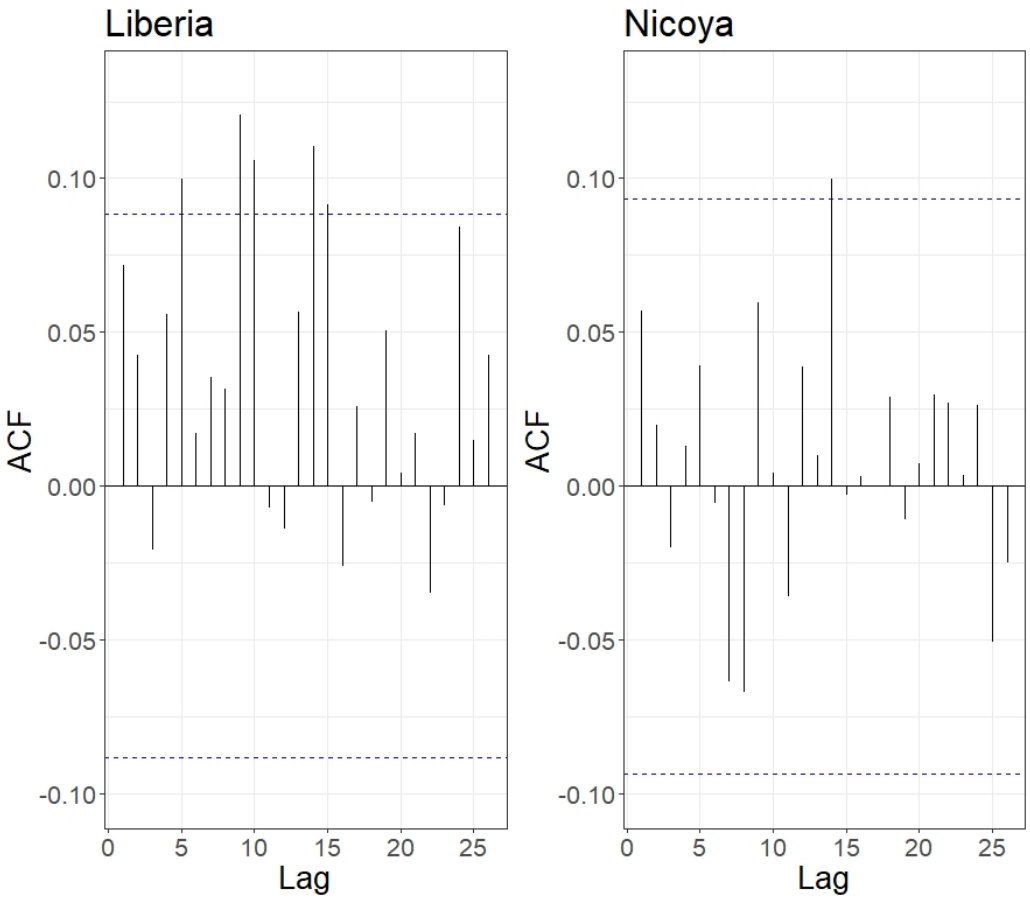

Figure S3: Standardized residual autocorrelation versus lag in months for the DG model

Table S1: Estimates and standard errors of the parameters of the regression model with AR(1) errors in Web Equation 1 fit to the Liberia data.

Parameter	Estimate	Standard error	p-value
m_1	0.429	0.120	0.000
m_2	0.337	0.120	0.005
m_3	0.453	0.120	0.000
m_4	1.640	0.120	0.000
m_5	3.671	0.121	0.000
m_6	3.852	0.121	0.000
m_7	3.257	0.121	0.000
m_8	3.729	0.121	0.000
m_9	4.230	0.121	0.000
m_{10}	4.215	0.121	0.000
m_{11}	2.953	0.121	0.000
m_{12}	1.471	0.121	0.000
γ_1	-0.104	0.049	0.032
γ_2	-0.380	0.108	0.000
γ_3	-0.303	0.074	0.000
γ_4	0.014	0.115	0.902
Lag-1 Coefficient	0.076	0.045	-
Residual Variance	0.459	-	-

Table S2: Estimates and standard errors of the parameters of the regression model with AR(1) errors in Web Equation 1 fit to the Nicoya data.

Parameter	Estimate	Standard error	p-value
m_1	0.494	0.138	0.000
m_2	0.688	0.146	0.000
m_3	1.245	0.140	0.000
m_4	2.356	0.139	0.000
m_5	4.001	0.140	0.000
m_6	4.097	0.139	0.000
m_{71}	3.859	0.139	0.000
m_8	4.163	0.139	0.000
m_9	4.413	0.140	0.000
m_{10}	4.419	0.140	0.000
m_{11}	3.212	0.140	0.000
m_{12}	1.238	0.141	0.000
γ_1	-0.025	0.055	0.646
γ_2	-0.152	0.121	0.212
γ_3	-0.239	0.083	0.004
γ_4	-0.105	0.136	0.441
Lag-1 Coefficient	0.100	0.048	-
Residual Variance	0.538	-	-

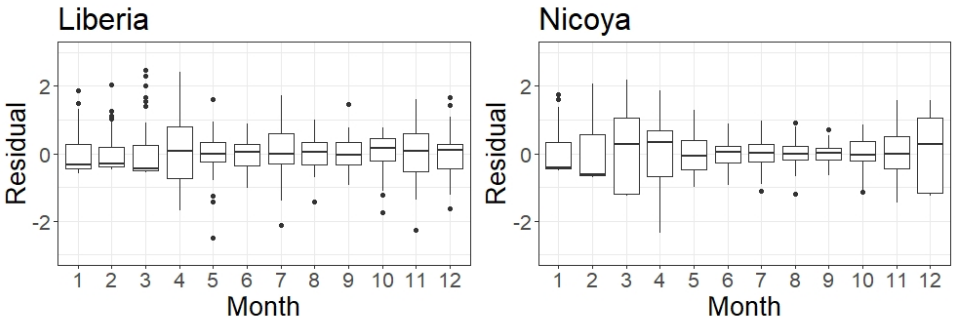

Figure S4: Residuals (obtained by fitting the model in Web Equation 1 with $\lambda = 0.25$ and ϵ_{it} treated as independent and identically distributed) vs. month of the year.

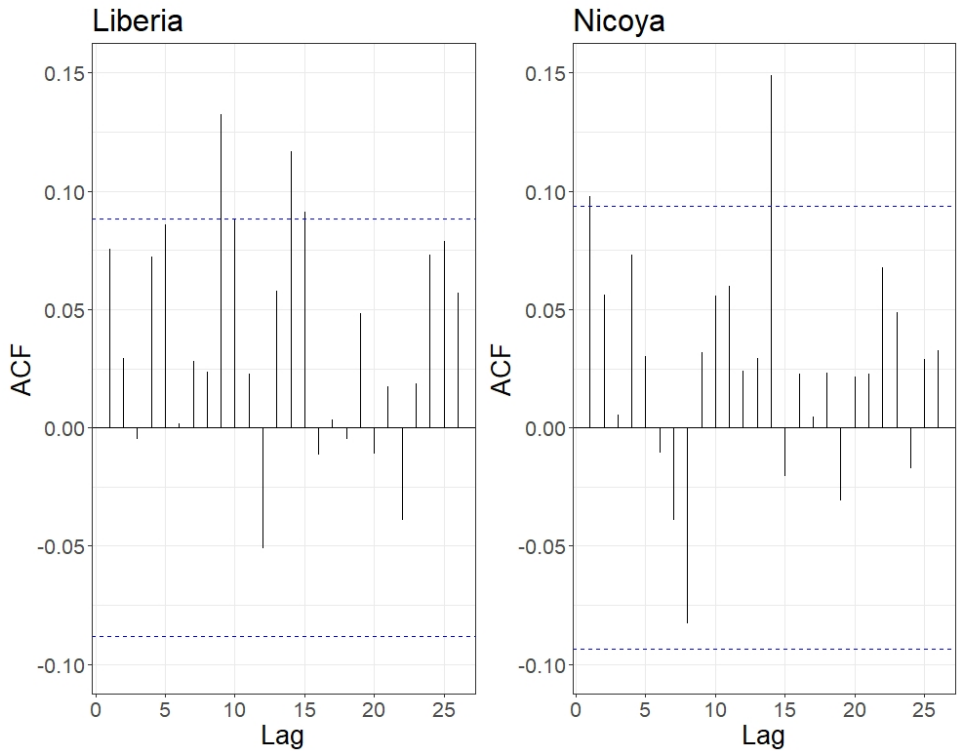

Figure S5: Residual autocorrelation versus lag in months for the model in Web Equation 1 fit to the data from both stations: ϵ_{it} treated as independent and identically distributed.

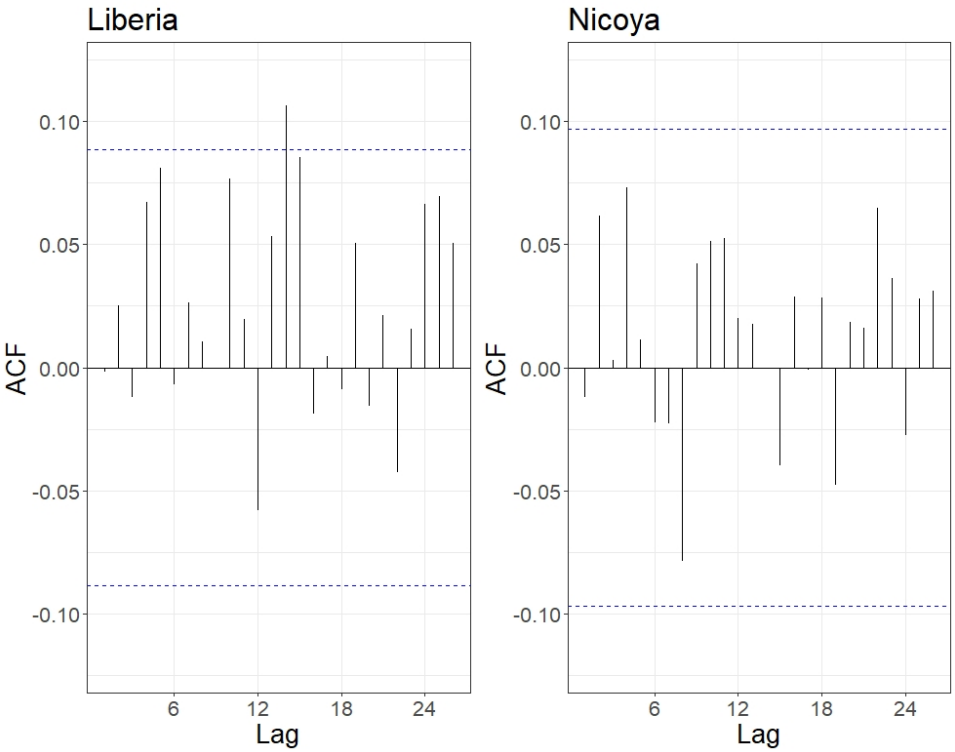

Figure S6: Residual autocorrelation versus lag in months for the model in Web Equation 1 fit to the data from both stations: ϵ_{it} described by an AR(1) process.

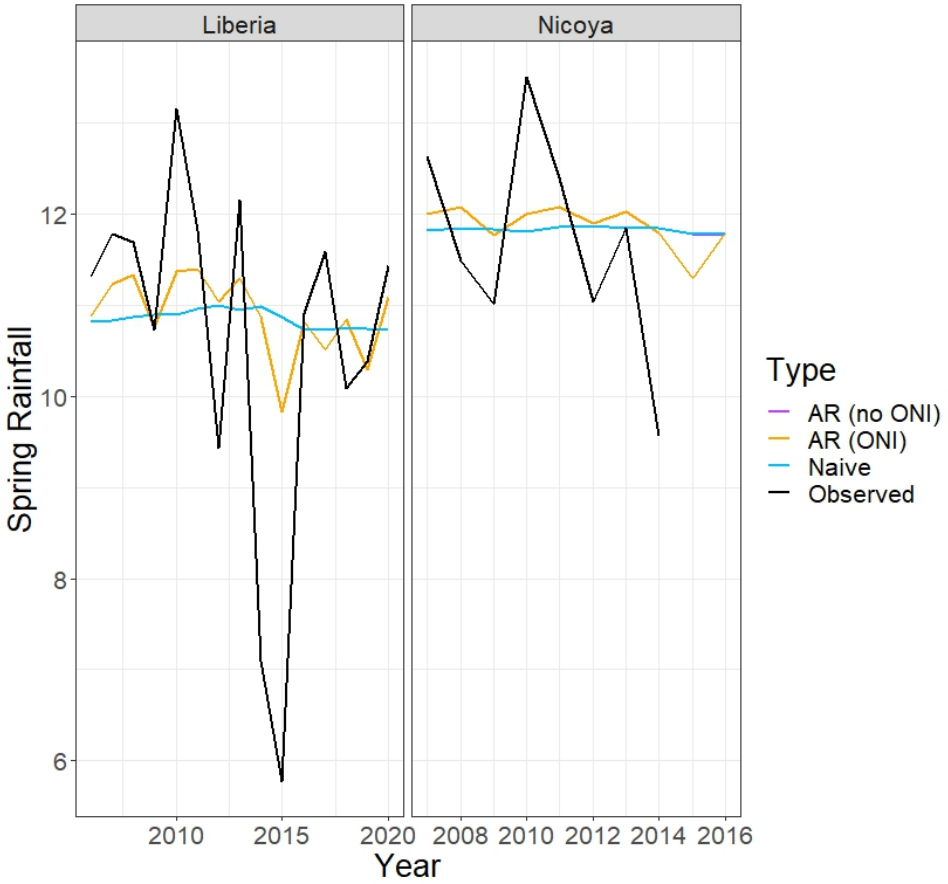

Figure S7: Predicted (using the regression model with AR(1) errors and naive approach) and observed total spring rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. Rainfall is reported on the scale $Y^{0.25}$. The naive predictions overlap with those based on the regression model without ONI.

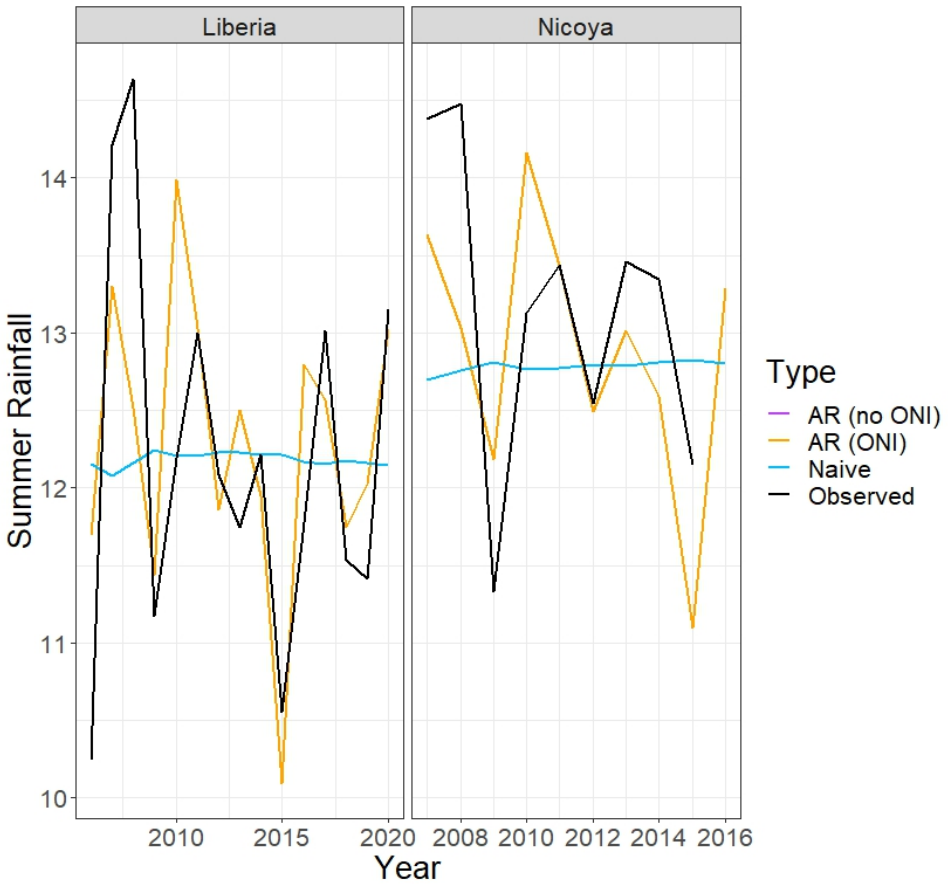

Figure S8: Predicted (using the regression model with AR(1) errors and naive approach) and observed total summer rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. Rainfall is reported on the scale $Y^{0.25}$. The naive predictions overlap with those based on the regression model without ONI.

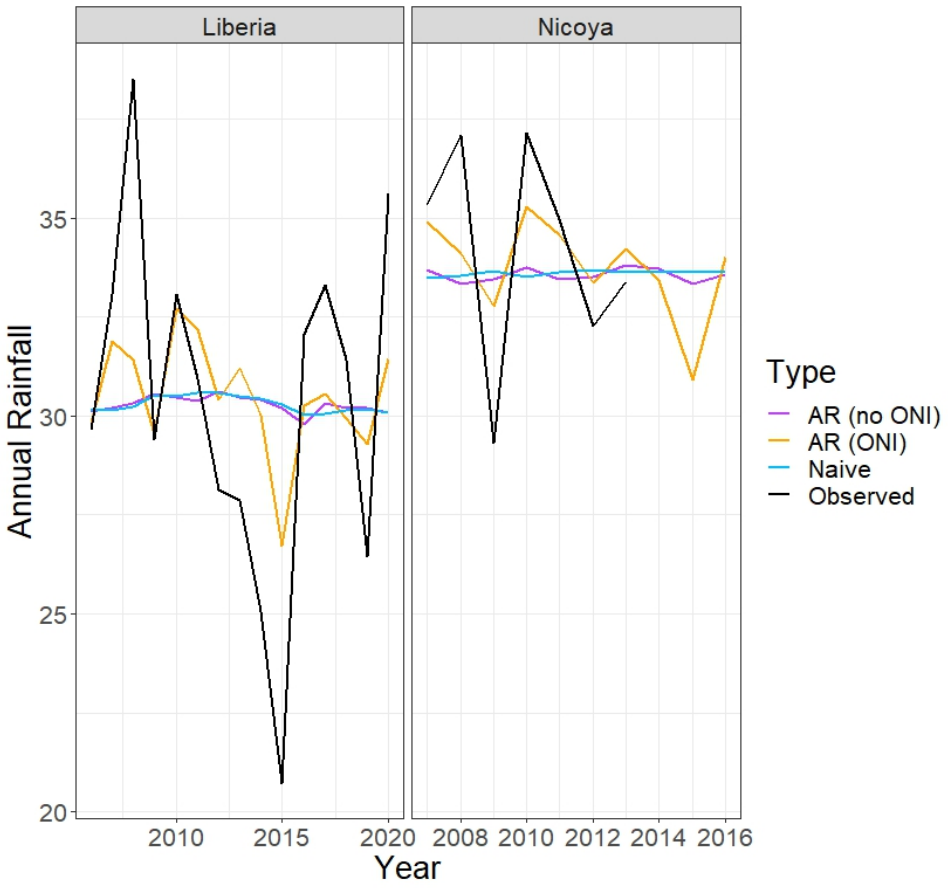

Figure S9: Predicted (using the regression model with AR(1) errors and naive approach) and observed total annual rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. Rainfall is reported on the scale $Y^{0.25}$.

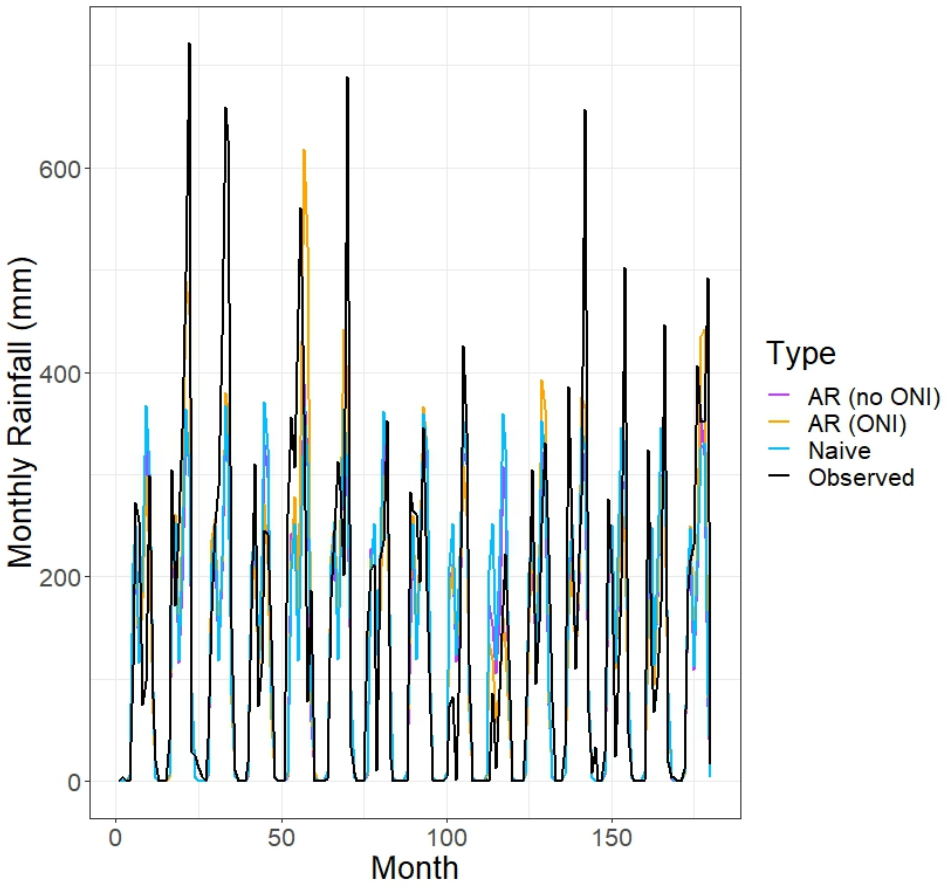

Figure S10: One-month-ahead-predicted (using the regression model with AR(1) errors and naive approach) and observed monthly rainfall values at the Liberia station from 2006 to 2020. "Month" is the number of months since January 2006.

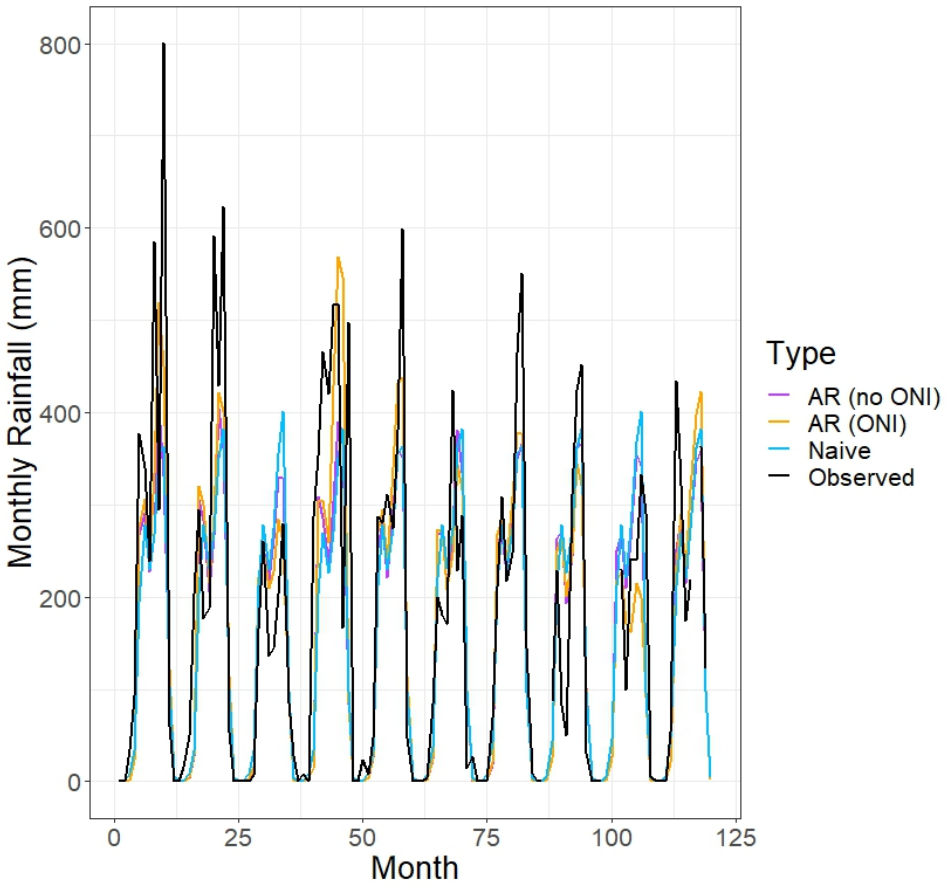

Figure S11: One-month-ahead-predicted (using the regression model with AR(1) errors and naive approach) and observed monthly rainfall values at the Nicoya station from 2007 to 2016. "Month" is the number of months since January 2007.

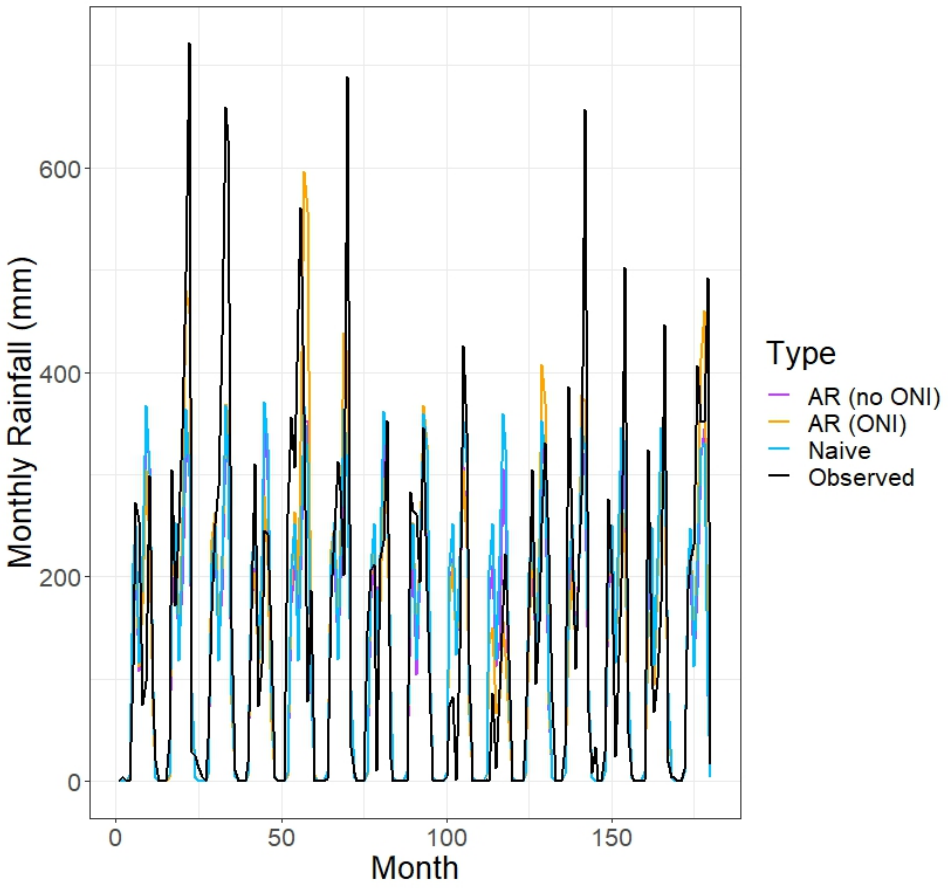

Figure S12: Three-month-ahead-predicted (using the regression model with AR(1) errors and naive approach) and observed monthly rainfall values at the Liberia station from 2006 to 2020. "Month" is the number of months since January 2006.

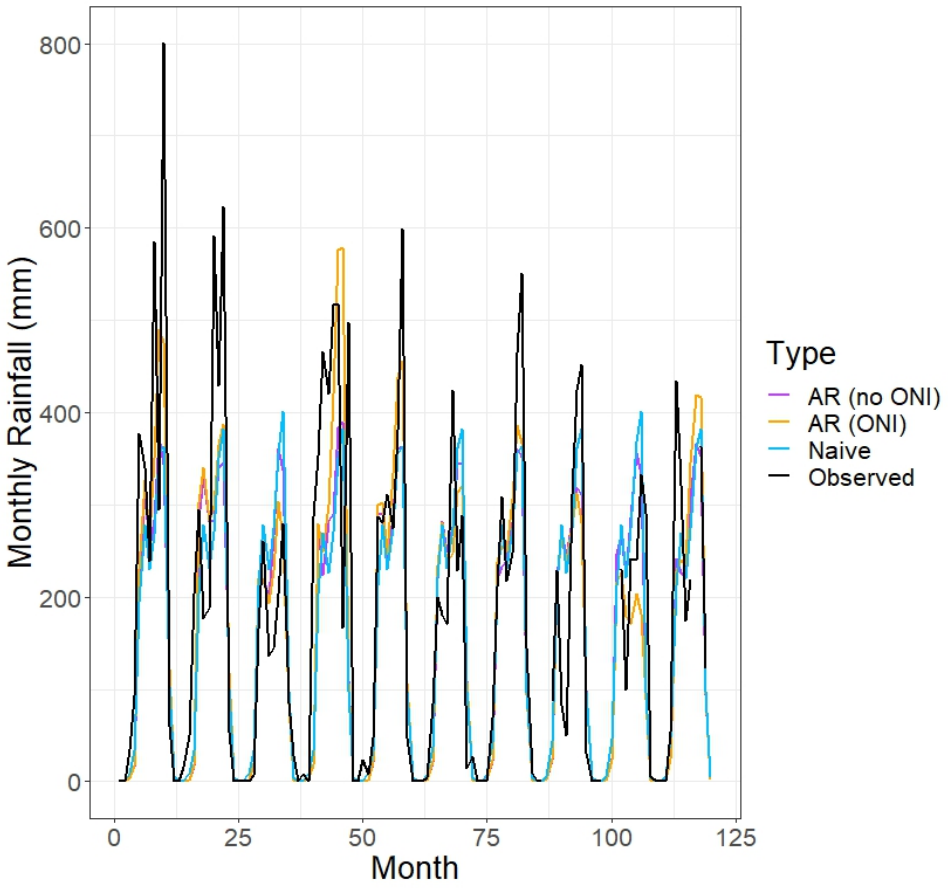

Figure S13: Threee-month-ahead-predicted (using the regression model with AR(1) errors and naive approach) and observed monthly rainfall values at the Nicoya station from 2007 to 2016. "Month" is the number of months since January 2007.

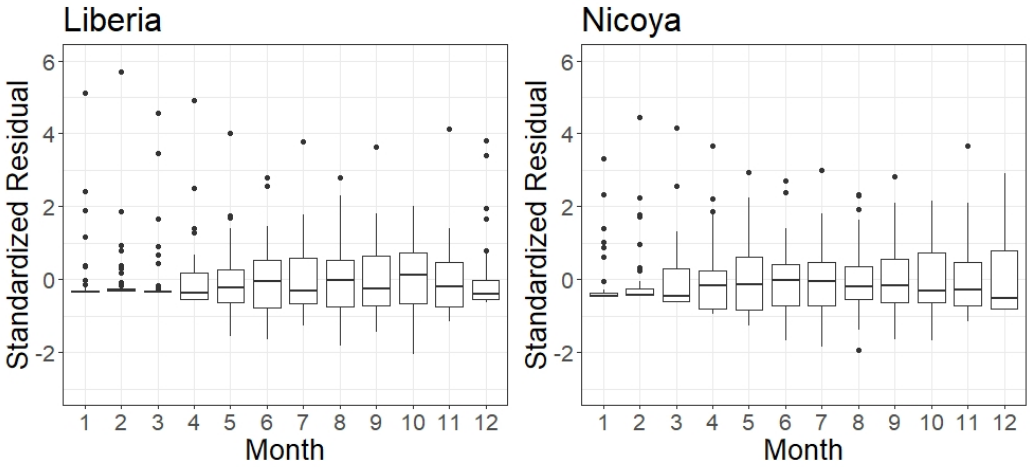

Figure S14: Standardized residuals versus month of the year for the Tweedie model

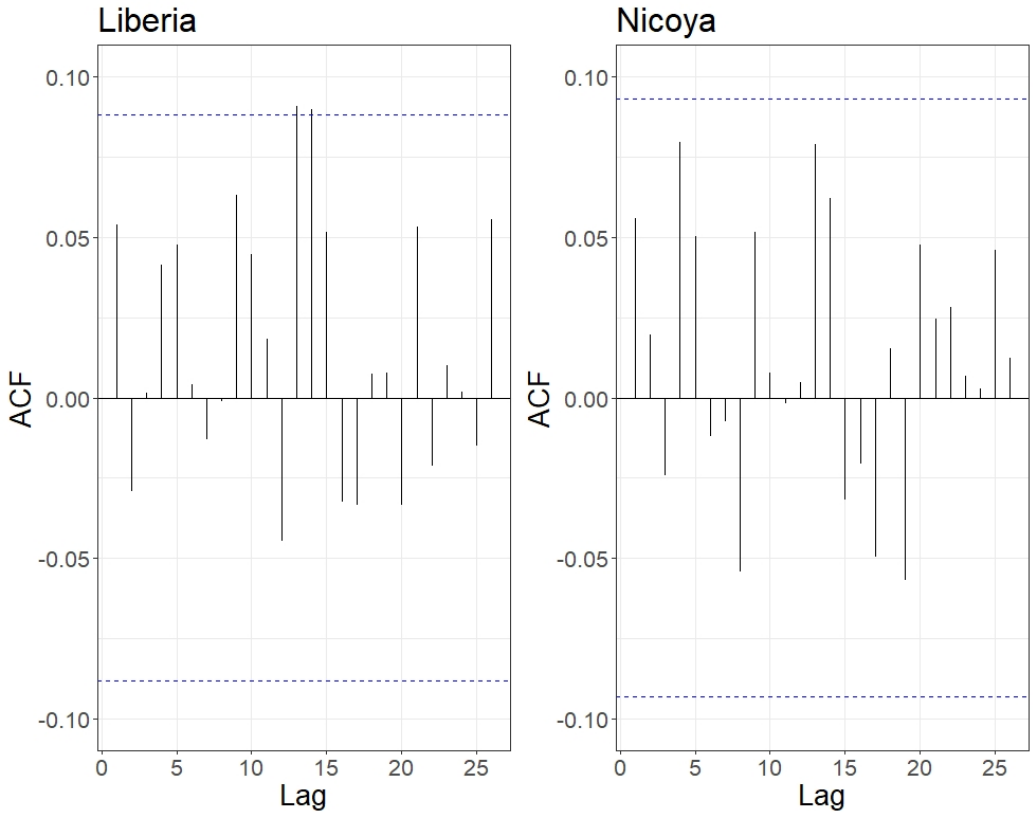

Figure S15: Estimated autocorrelation functions of the standardized residuals of the Tweedie model

Table S3: Estimates and standard errors of some of the parameters of the Tweedie model fit to the Liberia data

Parameter	Estimate	Standard error	p-value
a_1	0.722	0.447	0.106
a_2	0.246	0.501	0.624
a_3	1.586	0.456	<0.001
a_4	3.264	0.283	<0.001
a_5	5.368	0.114	<0.001
a_6	5.450	0.088	<0.001
a_7	4.962	0.133	<0.001
a_8	5.336	0.096	<0.001
a_9	5.820	0.092	<0.001
a_{10}	5.826	0.087	<0.001
a_{11}	4.646	0.145	<0.001
a_{12}	2.393	0.250	<0.001
ϕ_1	9.909	-	-
ϕ_2	10.927	-	-
ϕ_3	13.155	-	-
ϕ_4	7.924	-	-
ϕ_5	0.384	-	-
ϕ_6	0.165	-	-
ϕ_7	0.574	-	-
ϕ_8	0.222	-	-
ϕ_9	0.190	-	-
ϕ_{10}	0.154	-	-
ϕ_{11}	0.707	-	-
ϕ_{12}	4.735	-	-

Table S4: Estimates and standard errors of some of the parameters of the Tweedie model fit to the Nicoya data.

Parameter	Estimate	Standard error	p-value
a_1	1.090	0.369	0.003
a_2	1.650	0.381	<0.001
a_3	2.919	0.269	<0.001
a_4	4.204	0.183	<0.001
a_5	5.688	0.104	<0.001
a_6	5.736	0.082	<0.001
a_7	5.499	0.083	<0.001
a_8	5.801	0.083	<0.001
a_9	6.001	0.068	<0.001
a_{10}	6.047	0.084	<0.001
a_{11}	4.927	0.142	<0.001
a_{12}	2.617	0.214	<0.001
ϕ_1	7.855	-	-
ϕ_2	10.855	-	-
ϕ_3	9.523	-	-
ϕ_4	7.608	-	-
ϕ_5	0.292	-	-
ϕ_6	0.140	-	-
ϕ_7	0.147	-	-
ϕ_8	0.146	-	-
ϕ_9	0.063	-	-
$\phi_1 0$	0.149	-	-
$\phi_1 1$	0.636	-	-
$\phi_1 2$	4.954	-	-

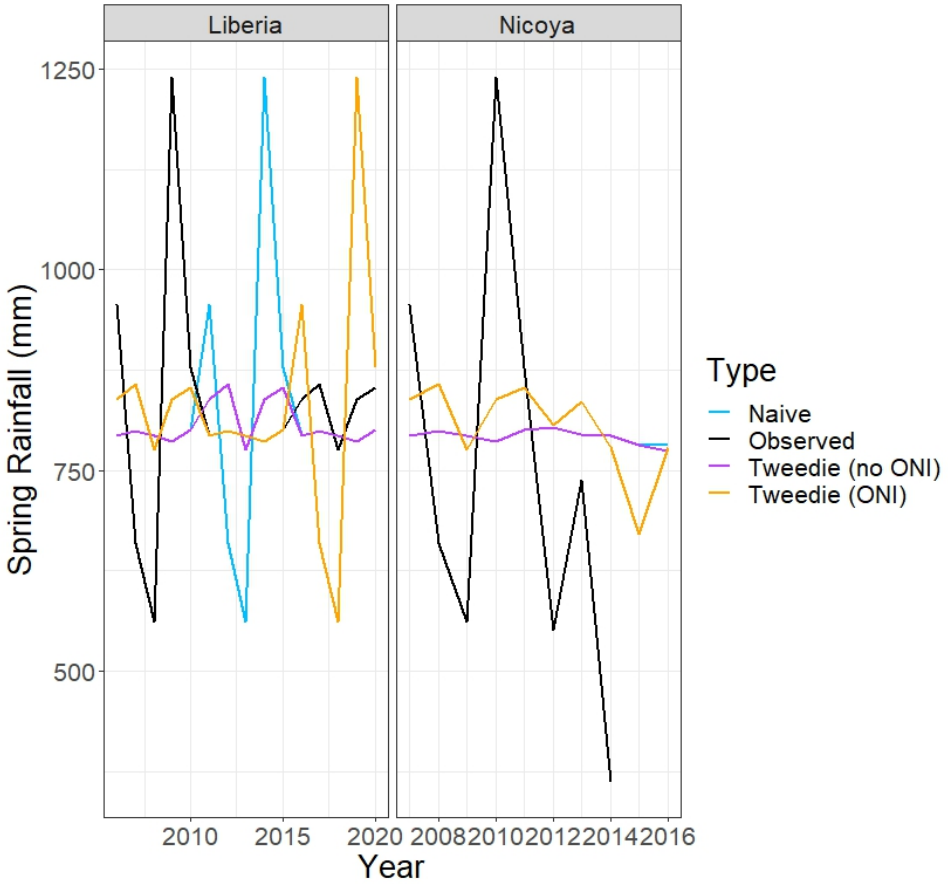

Figure S16: Predicted (using the Tweedie and naive approaches) and observed total spring rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. At the Nicoya station, the naive predictions largely overlap with those based on the Tweedie model without ONI.

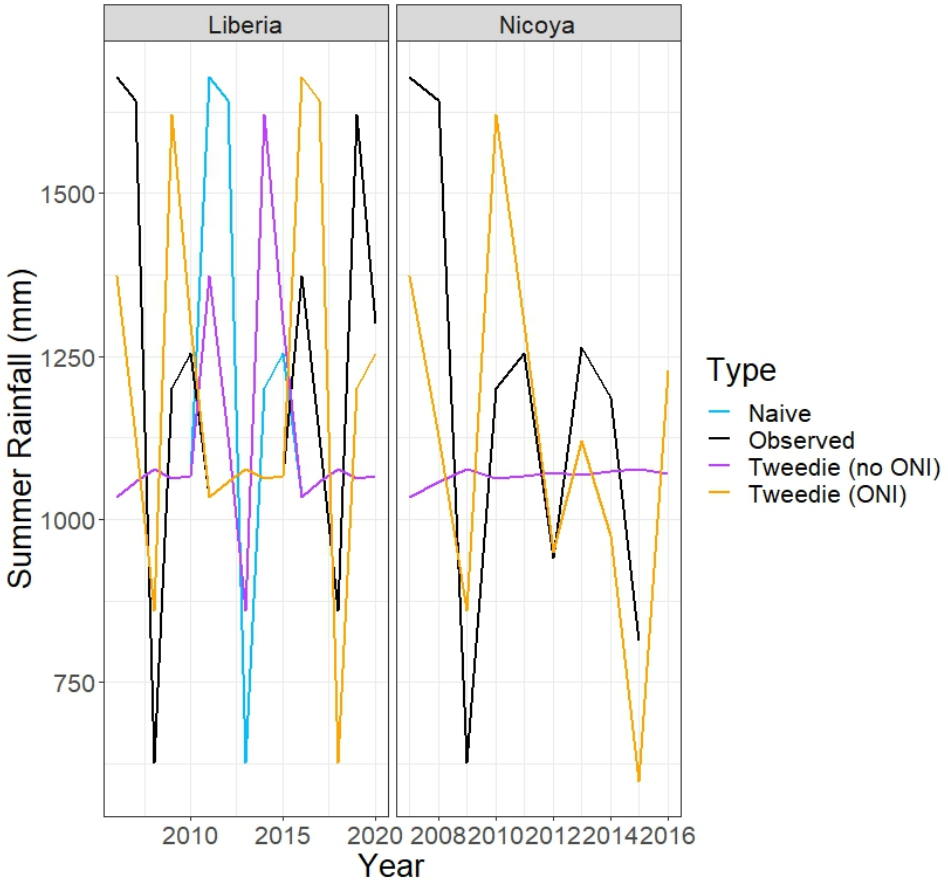

Figure S17: Predicted (using the Tweedie and naive approaches) and observed total summer rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. At the Nicoya station, the naive predictions overlap with those based on the Tweedie model without ONI.

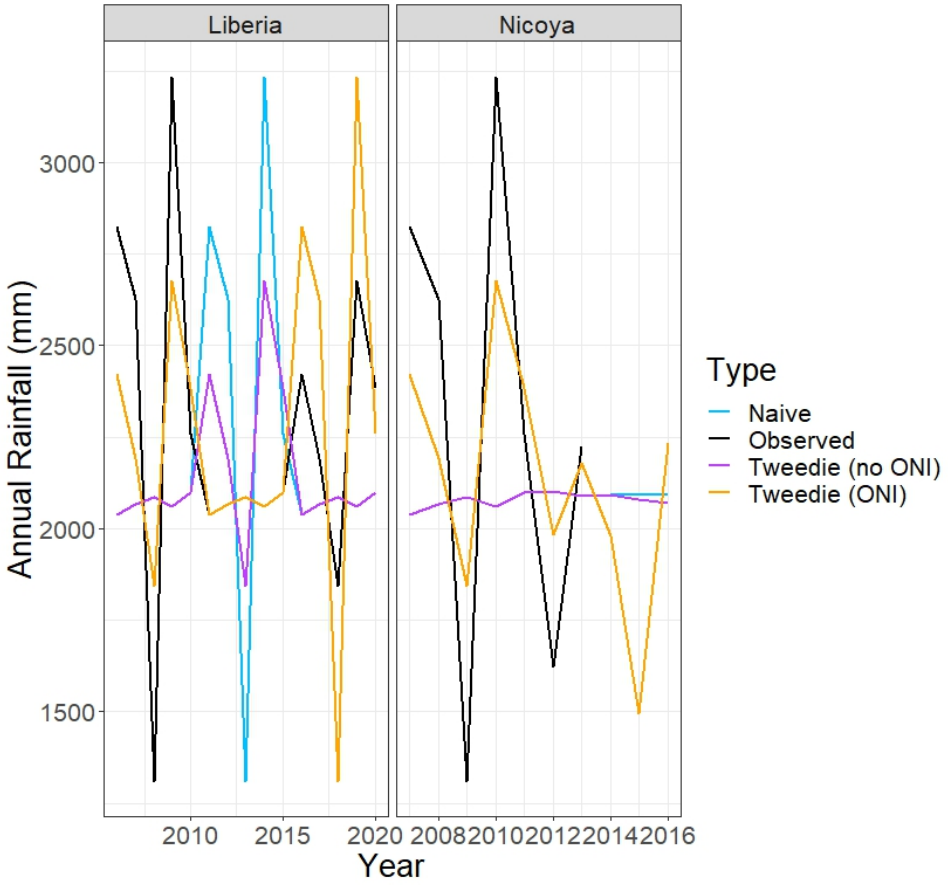

Figure S18: Predicted (using the Tweedie and naive approaches) and observed total annual rainfall values at the Liberia station (left) from 2006 to 2020 and at the Nicoya station (right) from 2007 to 2016. At the Nicoya station, the naive predictions largely overlap with those based on the Tweedie model without ONI.

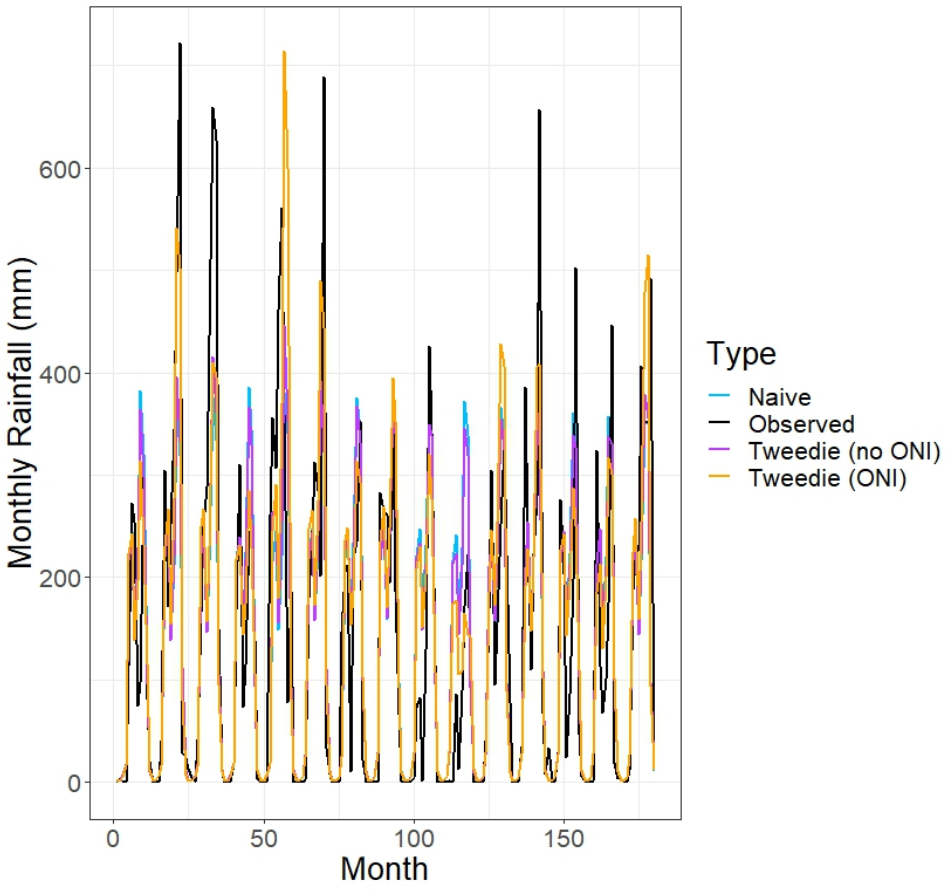

Figure S19: One-month-ahead-predicted (using the Tweedie and naive ap-proaches) and observed monthly rainfall values at the Liberia station from 2006 to 2020. "Month" is the number of months since January 2006.

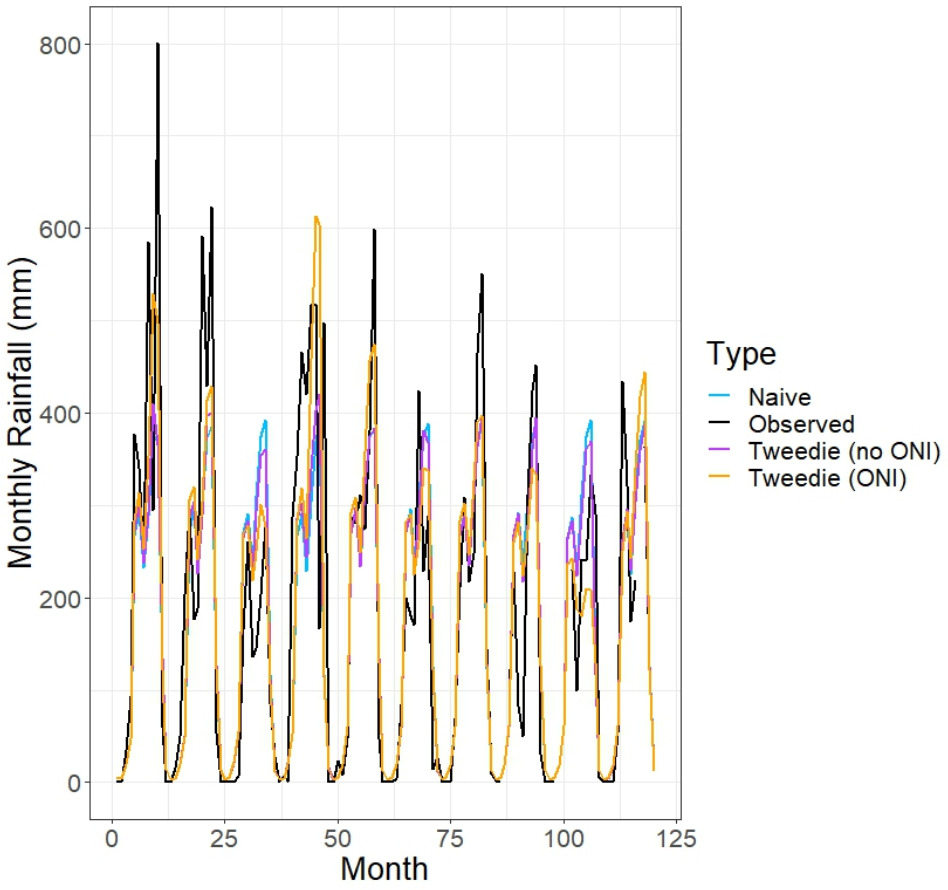

Figure S20: One-month-ahead-predicted (using the Tweedie and naive ap-proaches) and observed monthly rainfall values at the Nicoya station from 2007 to 2016. "Month" is the number of months since January 2007.

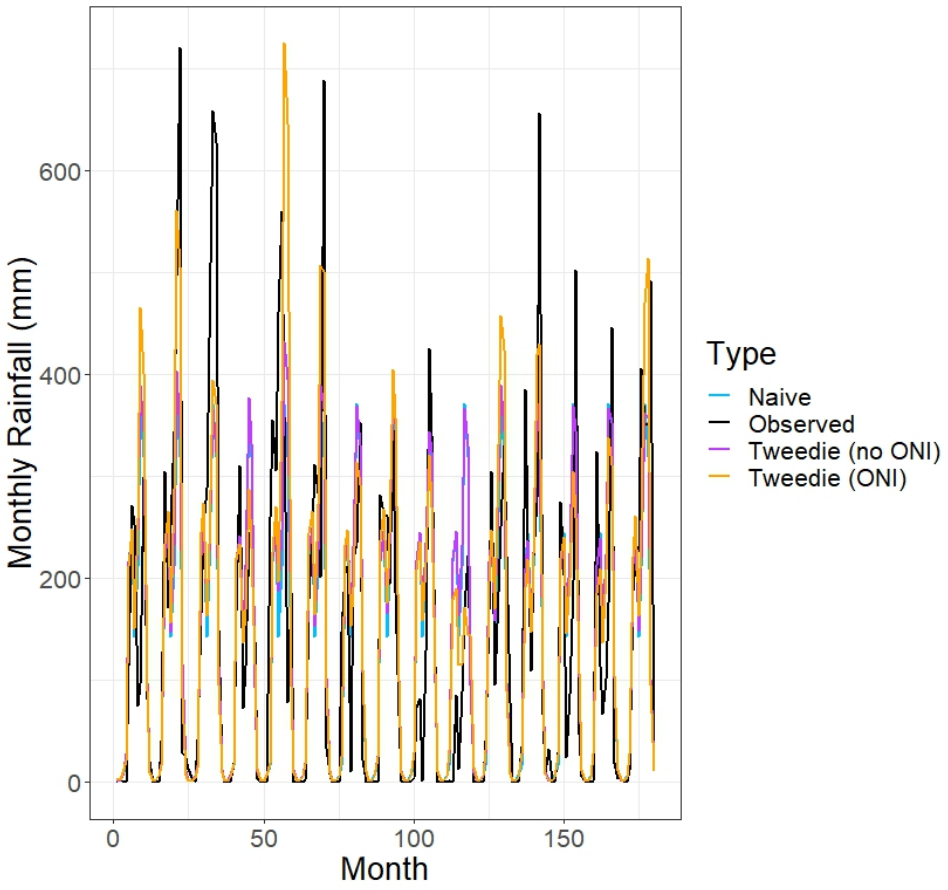

Figure S21: Three-month-ahead-predicted (using the Tweedie and naive ap-proaches) and observed monthly rainfall values at the Liberia station from 2006 to 2020. "Month" is the number of months since January 2006.

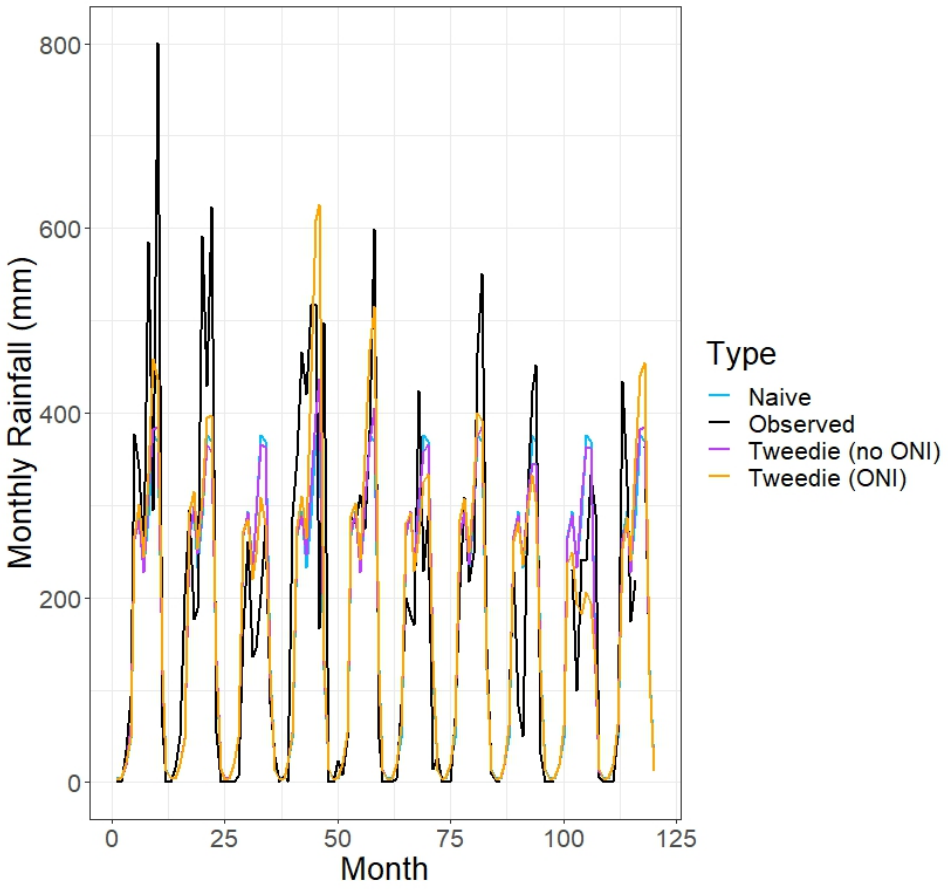

Figure S22: Threee-month-ahead-predicted (using the Tweedie and naive ap-proaches) and observed monthly rainfall values at the Nicoya station from 2007 to 2016. "Month" is the number of months since January 2007.