Time Series Analysis for Predicting Hydroelectric Power Production: The Ecuador Case

Julio Barzola-Monteses [1,2,*] , **Mónica Mite-León** [1], **Mayken Espinoza-Andaluz** [3] ,
Juan Gómez-Romero [2] **and Waldo Fajardo** [2]

[1] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, 090514 Guayaquil, Ecuador; monica.miteleon@ug.edu.ec
[2] Department of Computer Science and Artificial Intelligence, Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación, Universidad de Granada, 18071 Granada, Spain; jgomez@decsai.ugr.es (J.G.-R.); aragorn@ugr.es (W.F.)
[3] Centro de Energías Renovables y Alternativas, Facultad de Ingeniería Mecánica y Ciencias de la Producción, Escuela Superior Politécnica del Litoral, ESPOL, 09-01-5863 Guayaquil, Ecuador; masespin@espol.edu.ec
* Correspondence: julio.barzolam@ug.edu.ec; Tel.: +593-99-9493575

check for updates

Abstract: Electrical generation in Ecuador mainly comes from hydroelectric and thermo-fossil sources, with the former amounting to almost half of the national production. Even though hydroelectric power sources are highly stable, there is a threat of droughts and floods affecting Ecuadorian water reservoirs and producing electrical faults, as highlighted by the 2009 Ecuador electricity crisis. Therefore, predicting the behavior of the hydroelectric system is crucial to develop appropriate planning strategies and a good starting point for energy policy decisions. In this paper, we developed a time series predictive model of hydroelectric power production in Ecuador. To this aim, we used production and precipitation data from 2000 to 2015 and compared the Box-Jenkins (ARIMA) and the Box-Tiao (ARIMAX) regression methods. The results showed that the best model is the ARIMAX $(1,1,1)$ $(1,0,0)_{12}$, which considers an exogenous variable precipitation in the Napo River basin and can accurately predict monthly production values up to a year in advance. This model can provide valuable insights to Ecuadorian energy managers and policymakers.

Keywords: hydroelectric power; production prediction; time series analysis; ARIMA; ARIMAX

1. Introduction

Between 1973 and 2016, the world gross electricity production increased from 6298 TWh to 25,082 TWh, an average annual growth rate of 3.3% [1,2]. In 1973, fossil sources generated 75.2% of electric power, while in 2016, they accounted for 65.3% of the production. Although the proportion of fossil fuels has decreased during this period, the absolute figures are very high due to the rising world population and per capita energy consumption. As a matter of fact, the amount of CO_2 emissions almost doubled, from 15,460 to 32,316 Mt. Hence, several proposals have recently emerged to incorporate CO_2 reduction criteria to coal/biomass combustion systems [3–5]. Complementarily, the role of green sources in the energy mix has gained relevance in the last decades.

Ecuador is located in a privileged zone with a huge potential in renewable energies: Sunbeams fall almost perpendicular, rainfalls are abundant, and the Andes mountains provide large hydroelectric, geothermal, and wind resources [6,7]. However, several impactful events, which occurred at the end of the last century, prevented the proper exploitation of such resources. Economic events, such as the beginning of oil extraction in 1972, the debt crisis in 1982, and natural disasters such as El Niño (1983, 1987, and 1998) and earthquakes (1987, 1995, and 1998) have caused a discontinuous and insufficient development and deployment of renewable energy systems in Ecuador [8].

This situation has been partially reverted in the last decades, since large hydroelectric projects have been introduced as a key element of the national development plan. Hydropower offers multiple advantages against other energy sources in Ecuador, since it is cheaper and cleaner despite the initial investment that it requires [9]. Carvajal et al. [10] showed that Ecuador's energy policy in the period 2007–2017 incentivized a doubling of its hydropower capacity.

However, as shown in Figure 1, the share of production originated by hydroelectric systems has declined from 1990 to 2015: In 1990, 79% of the total energy production came from hydroelectric sources, while in 2015, it was only 49%. This was caused by a misalignment between infrastructure development and exploitation policies, resulting in an increase of the use of fossil fuels despite their social, economic, and environmental costs. Given the considerable potential of Ecuadorian basins and the large hydropower capacity installed, there is a common consensus that the decreasing contribution of hydroelectric production should be reversed [11].
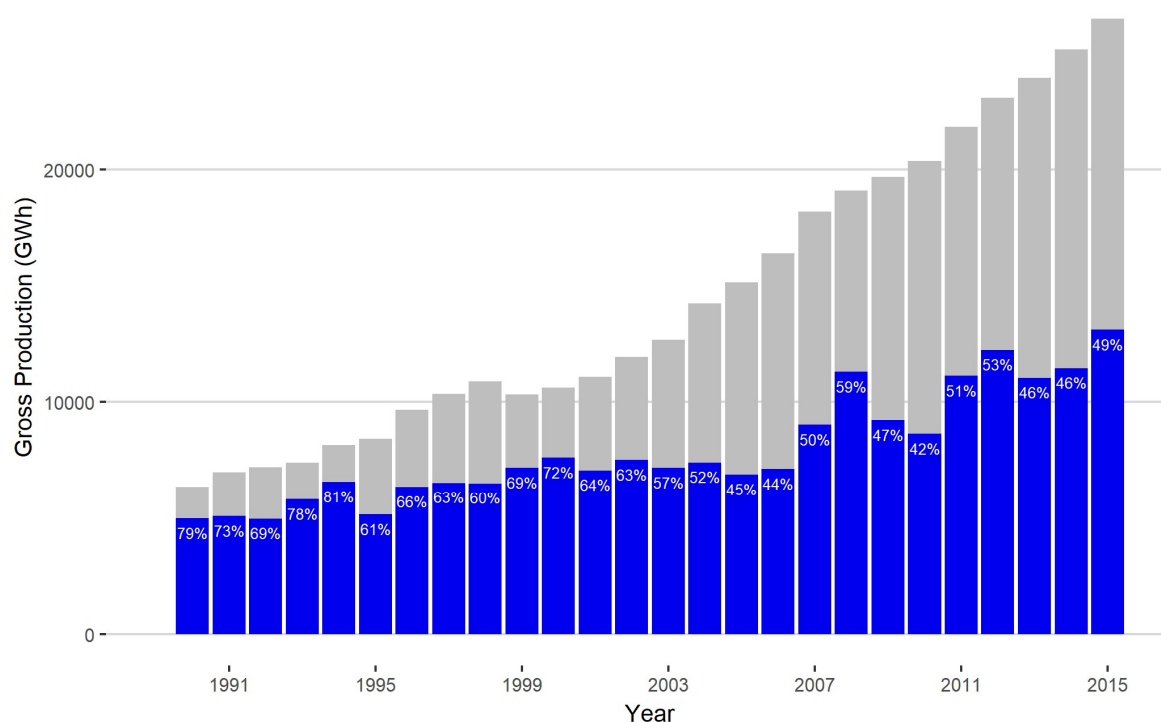

Figure 1. Annual development of Ecuador's electricity total generation. Blue bars represent the annual hydroelectric production with respect to the total generation.

Hence, Ecuador is currently implementing the National Master Plan for Electrification 2016–2025, which aims at covering more than 90% of the national electrical demand with hydroelectric sources [12]. The plan is already having an impact in Ecuador's energy production capability [13]. Ecuador ranked third in the global list of countries that added more energy capacity in 2016, following China and Brazil [14]. By the end of that year, the installed hydroelectric capacity in Ecuador (4400 MW) represented 58% of the total capacity (7587 MW). Generation from hydroelectric systems in 2016 amounted to 57.9% (15,814.72 GWh) of the total production (27,313.86 GWh). In 2017, the installed capacity of the Ecuadorian electrical system increased to 8036 MW [15]. This year, Ecuador achieved energy sovereignty for the first time, i.e., the electricity production satisfied the national electricity demand and even 210 GWh were exported. Despite this fact, the maximum demand used only 47% of the produced electricity, i.e., 53% of the energy that can be generated is not used [16,17]. As a consequence, the management and planning of energy in Ecuador has room for improvement.

Energy production forecasts are of great importance to the operators of the electrical system (to optimize the processes) and the decisionmakers (to define better policies and manage risks). Besides,

energy prices can be established considering the estimated future load demand. Therefore, it is crucial for an optimized energy system to accurately model the energy production and the energy demand to offer solutions at different levels of society [18].

The objective of this research work was to develop and apply forecasting models for predicting hydroelectric production in Ecuador from historical data. We focused on statistical time series analysis techniques, which have been successfully applied in the literature [19,20]. Since hydropower strongly depends on the meteorological conditions, rainfall data played a key role in our study. This data was easier to obtain than inflow, which was not available for the study. To the best of our knowledge, there is no similar study in the literature focused on the Ecuador case.

The contribution of this paper is twofold. First, it describes a methodology to model the production of hydroelectric plants through time series with the ARIMA univariate and ARIMAX bivariate approaches. Then, we present the application of this methodology to build and compare models for predicting production in Ecuador. Through the resulting model, the correlation between hydroelectric generation and precipitation is demonstrated and quantified. Furthermore, the paper shows that an ARIMAX model considering an exogenous variable (precipitation in one of the larger basins) can accurately predict monthly production values up to a year in advance. That implies that it is not necessary to use elaborated data to obtain a reliable prediction model. At the same time, this study improves the results obtained in a previous work which only considered ARIMA [21]. Accordingly, the methodology can be used to develop planning strategies in the energy sector in Ecuador. The applicability of the approach to other countries should be considered by future work.

The rest of the paper is organized as follow. Section 2 explains the methodology used in this work and Section 3 describes the data. Section 4 shows the experiments and a discussion of the results. Finally, Section 5 includes some concluding remarks and directions for future work.

2. Methods

2.1. Box-Jenkins and Box-Tiao Methods

Time series analysis comprises a wide collection of techniques for analyzing historical temporal data in order to extract meaningful features or characteristics of the data. A forecast model encodes a function that estimates the value of a prediction variable in the future from historical and other relevant data [22]. This prediction variable is usually also used in the input. In our case, the input and the prediction variable is energy production, and the additional variable is precipitation.

The accuracy of the model is measured as the difference between the predicted and the actual values. Learning a forecast model means fitting the model parameters to minimize this difference for a given dataset for which the input and the target values are known. It is expected that a forecast model will perform well with values outside the learning dataset. More details about time series analysis methods can be found in [23].

In the literature, we can find several methods to build forecast models from data. Among them, autoregressive models based on moving averages have proved effective in several problems [24]. In particular, there are applications of autoregressive models to predict renewable energy (wind, solar, hydro) production and demand [25–33].

Autoregressive models assume that observations are correlated through lagged linear relations and capture this correlation into the model. These models assume that the time series are stationary, i.e., the statistical properties, such as the mean, variance, and autocorrelation, are constant over time.

An autoregressive model of order p AR(p) is of the form:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t \tag{1}$$

where: Y_t is the value of the prediction variable at time t (energy production, in this case), ε_t is a random variable with mean 0 and variance σ^2_w (white noise), and φ_1 are the parameters of the models

(constants). If we define the backshift operator B as $B_{Yt} = Y_{t-1}$ (and recursively, $B^k{}_{Yt} = Y_{t-k}$), the equation can be rewritten as follows:

$$\varepsilon_t = (1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p)\, Y_t = \varphi_p(B)\, Y_t \tag{2}$$

The Box-Jenkins method identifies, adjusts, and checks a prediction model using an autoregressive integrated moving average (ARIMA) [34]. Box-Jenkins estimates the parameters of the model by applying the maximum likelihood. The model is formulated as follows:

$$\varphi_p(B)(1 - B)^d Y_t = \varphi_0 + \theta_q(B)\varepsilon_t \tag{3}$$

where $\varphi_p(B)$ is the autoregressive polynomial of order p, B is the backshift operator, Y_t is the forecast variable which represents a white noise process with normal distribution $N(0, \sigma^2)$, d is the dth difference operator, φ_0 is a numerical constant, $\theta_q(B)$ is the regular moving average (MA) polynomial of order q, and ε_t is the error of the model. These parameters (p, d, q) characterize the ARIMA model and must be identified *a priori* by analyzing the stationarity of the time series.

If the stationarity (S) is strong, ARIMA is extended to have two components: (1) One with regular structure ARIMA (p, d, q) that models the non-independence associated with the data, and (2) one with ARIMA structure (P, D, Q) that models the seasonality component. In this latter model, P is the autoregressive seasonal term, D is the seasonal term of difference, and Q corresponds to the seasonal term of moving average. A SARIMA model is formalized as follows:

$$\varphi_p(B)\Phi_P\,(B^s)(1 - B^s)^D(1 - B)^d\, Y_t = \varphi_0 + \theta_q(B)\,\Theta_Q\,(B^s)\varepsilon_t \tag{4}$$

where $\Phi_P\,(B^s)$ is the autoregressive polynomial of order p with seasonality, $\Theta_Q\,(B^s)$ is the regular moving average (MA) polynomial of order q with seasonality, D is the dth differential operator with seasonality, and B is the backshift operator with seasonality.

The Box-Tiao method extends Box-Jenkins by incorporating the observations of the covariates, also called explanatory or exogenous variables X (ARIMAX). The ARIMAX model is described mathematically as follows:

$$\varphi_p(B)(1 - B)^d Y_t = \varphi_0 + \Theta(B)X_t + \theta_q(B)\varepsilon_t \tag{5}$$

The difference with respect to Equation (3) is the term $\Theta(B)X_t$ is the observed one, while $\Theta(B)$ is the polynomial operator of the exogenous variable X_t.

2.2. Time Series Analysis Process

In order to obtain the ARIMA and the ARIMAX models, we followed the methodology depicted in Figure 2. First, the time series were transformed to comply with the requirement that they must be stationary. Once the series were stationary, the parameters of the series were estimated through correlogram functions, which were associated with statistical tests. Afterward, different models were estimated to select the one that best fit the training data. The selected model was validated through residue analysis and hypothesis testing. Then, this model was tested with additional unseen data to calculate prediction accuracy. These steps are described in the following subsections.
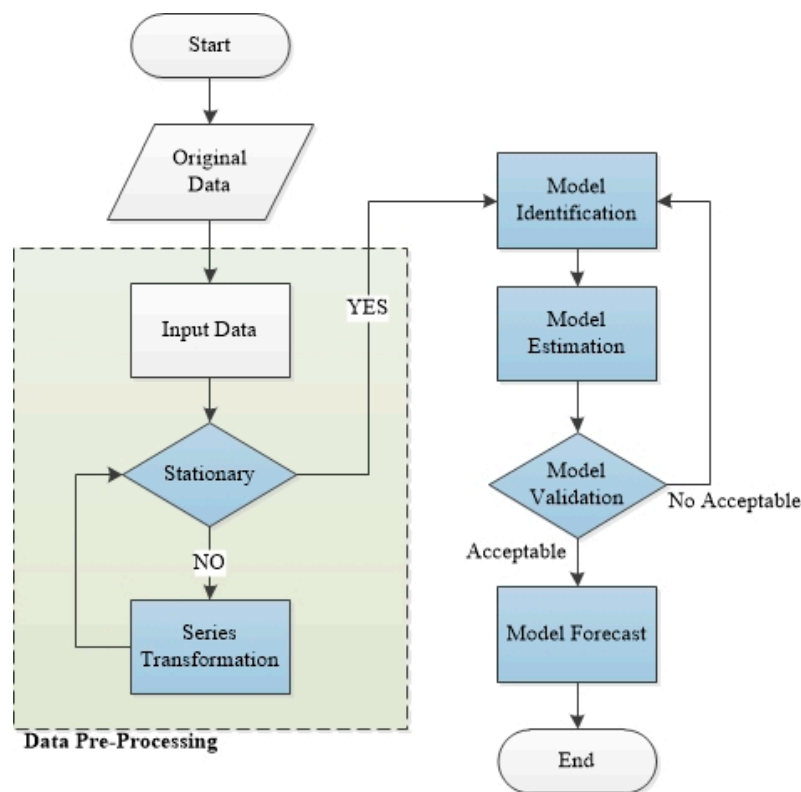
Figure 2. Flowchart of the time series analysis methodology.

2.2.1. Series Transformation

The assumption of stationarity must be verified before the identification of a suitable model. A time series can be considered stationary if the mean and variance are constant and if there are no significant trends and seasonal variations. Usually, logarithm and derivative operations are applied to eliminate the variability and decrease the trend, respectively [23,35].

2.2.2. Stationary Evaluation

The single autocorrelation function (ACF) and partial autocorrelation function (PACF) are correlogram functions to determine the degree of correlation between two consecutive values of the series. ACF and PACF provide an estimation of the *p* and *q* values of the ARIMA models [36,37].

First, the single ACF and PACF functions of the transformed series were calculated and plotted. These functions are useful for diagnosing the *p*, *d*, *q* components of the ARIMA models.

The ACF of order k ($k > 0$) of a stationary process (Y_t) measures the information that an observation of a period transmits directly to the observation k periods ahead, i.e., $(Y_1, Y_{1+k}), (Y_2, Y_{2+k}), \ldots (Y_{n-k}, Y_n)$. It is presented with the following equation:

$$\rho_k \equiv \frac{cov[Y_t, Y_{t+k}]}{Var[Y_t]^{\frac{1}{2}} \, Var[Y_{t+k}]^{\frac{1}{2}}} \equiv \frac{\gamma_k}{\gamma_o} \ (k = 1, 2, \ldots) \tag{6}$$

where γ_k is the self-covariance of order k, γ_o is the zero order self-covariance of a strictly stationary process, and $\gamma_o \equiv cov[Y_t, Y_t] \equiv var[Y_t]$.

The PACF measures the information that an observation of a period transmits directly to the observation k periods ahead, eliminating the information $Y_{t-1}, \ldots Y_{t-k+1}$ that both contain.

The PACF is denoted by φ_{kk}, for k = 1, 2, … and represented with the following equations:

$$\phi_{11} = corr([Y_{t+1}, Y_t] \equiv \rho_1 \tag{7}$$

$$\varnothing_{kk} \equiv corr\left(Y_{t+k} - \hat{Y}_{t+k}, Y_t - \hat{Y}_t\right), \; k \geq 2 \tag{8}$$

The first-order autocorrelation was formally tested with the Durbin-Watson statistic test [38], which measures the linear association between adjacent residuals. The null hypothesis for this case is that the series presents autocorrelation.

Second, the Augmented Dickey Fuller (ADF) test, also called the unit root test, was used to test the stationarity properties of the series [39,40].

2.2.3. Model Identification

Once the parameters (p, d, q) and (P, D, Q) were identified in previous stages, auto-regressive (AR), moving average (MA), seasonal auto-regressive (SAR), and eXogenous regressive (XREG) models were determined. The Equations (3)–(5) were used to identify the model that best fits each series.

2.2.4. Model Estimation

After the models were designed and implemented, the best model was selected considering penalized likelihood criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) [41]. The AIC and BIC were used here to compare of models' performance [23]. The AIC and BIC are described mathematically as follows:

$$AIC = log \, \hat{\sigma}_k^2 + \frac{n + 2k}{n} \tag{9}$$

$$BIC = log \, \hat{\sigma}_k^2 + \frac{k \log n}{n} \tag{10}$$

where $\hat{\sigma}_k^2$ is the maximum likelihood estimate of variance, n is the number of dates, and, k is the number of parameters of the model.

The BIC strongly penalizes the number of involved parameters. High values of AIC mean that the observed data does not fit the models, while lower values indicate strong evidence that the observed data fit the models. Similarly, lower values of BIC indicate better fitting of the models.

2.2.5. Model Validation

In order to evaluate the prediction accuracy of the models of the Box-Jenkins and Box-Tiao, a residual analysis was performed. Before that, a first plot was done to analyze if whether there were atypical errors which indicated the need of an intervention.

In order for the ARIMA and ARIMAX models to be viable for the adjustments of the observed data, the error term ε_t in Equations (1)–(3) should behave as white noise, i.e., zero mean, a constant variance, and no correlation. In addition, the term ε_t must follow a normal distribution. To check these assumptions, statistical tests should be applied to the residuals, as in [23,26].

The most widely used model is the Box-Ljung test [36], with the null hypothesis being that the series is uncorrelated. Accordingly, a Box-Ljung test was applied to the (squared) residuals to verify that the variance is constant. Afterward, the Jarque-Bera test [42] was applied to verify that the residuals were normal. In our experiments, these tests were applied to the significance level of $\alpha = 0.05$.

2.2.6. Model Forecast

Once the ARIMA and ARIMAX models have been validated, future time series values can be forecasted. In our case, energy production was forecasted for a 12-month horizon corresponding to the year 2015, with a typical 95% confidence interval (CI) [22].

2.3. Measures of Accuracy

In order to evaluate the quality of forecasted data of the proposed models, statistical analysis of errors was used. Three typical standards were selected in this research work: Mean absolute error (MAE), mean absolute percentage error (MAPE), and mean absolute scaled error (MASE).

The MAE represents the average error value between the observed and the adjusted series. The MAE is described mathematically as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| v_{adj} - v_{obs} \right| \tag{11}$$

where v_{adj} represents the individual value of the forecasted time series, v_{obs} corresponds to the individual value of the observed time series, and n is the order of the series. In this study, the MAE is measured in gigawatt hours (GWh) since the predicted variable is hydroelectric production.

The mean absolute percentage error (MAPE) is another statistical parameter considered in this study. The advantage of using this parameter is that it uses percentages (%) to show the data, which allows an easy and quick evaluation of the predicted model [30]. The MAPE is described mathematically as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{v_{adj} - v_{obs}}{v_{obs}} \right| \times 100 \tag{12}$$

Finally, the mean absolute scaled error (MASE) was considered in the analysis. This type of error is independent of the data. Lower values of scaled error (q_t) result in better forecasts [43,44]. The MASE is described mathematically as follows:

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^{n} |Y_i - Y_{i-1}|} \tag{13}$$

$$\text{MASE} = mean \left(|q_t| \right) \tag{14}$$

where e_t represents the error between adjusted and observed values.

3. Data

Data for the current study, corresponding to the 2000–2015 period, have been obtained from official and governmental institutions of Ecuador, namely the Electricity Regulation and Control Agency (ARCONEL) [11] and the National Institute of Meteorology and Hydrology (INAMHI) [45]. The modeling analysis considers the period from 2000 to 2014, and 2015 data were used to validate the predictivity of the model.

In this section, we present information about the Ecuadorian hydroelectric system, the country's hydrographic basins, the regions and the dataset used in this work, and the time series used to build the models.

3.1. Ecuadorian Hydroelectric System

The hydroelectric system in Ecuador is concentrated in a relatively small group of stations. In 2016, the 13 largest power stations provided 89.57% of the total hydroelectric generation. Table 1 shows the contributions of hydroelectric generation grouped by management company and geographical region of watersheds in detail.

Coca Codo Sinclair is the largest hydroelectric power plant in the country. Over the next years, this plant is expected to have 1500 MW of power and satisfy 35% of the country demand. The socioeconomic and environmental impacts of this project have been largely studied [46–48], but its management and planning has received less attention.

Table 1. Electric production of the country's 13 largest hydropower stations during 2016.

Electric Company	Hydropower Station	Watershed	Production [GWh]	Percentage of Total Production [%]
CELEC [49]–Hidropaute [50]	Mazar Molino Sopladora	Santiago	6851.61	43.32
CELEC–Coca Codo Sinclair [51]	Quijos Manduriacu Coca Codo Sinclair	Napo	3264.01	20.64
CELEC–Hidroagoyán [52]	Agoyán Pucará San Francisco	Pastaza	2413.73	15.26
CELEC–Hidronación [53]	Baba	Guayas	1209.89	7.65
Elecaustro [54]	Ocaña 1 Machángara Saucay	Santiago	427.99	2.71
TOTAL				89.58

3.2. Rainfall and Watersheds in Ecuador

The estimated water potential, at the level of river basins and sub-basins, is about 15,000 m^3/s distributed on the Ecuadorian continental surface. Its potential is geographically distributed in two regions: Amazon (east) and Pacific (west), with a flow capacity of 71% and 29%, respectively, according to the National Water Resources Council (CNRH) [55]. Twenty-four watersheds are inside the Pacific area and seven watersheds are inside the Amazon area. The area of the latter is 131,726 km^2, which corresponds to 51% of the country's hydrographic system.

In Ecuador, the following values of the hydroelectric potential have been identified: (i) Medium theoretical hydroelectric potential, with an estimated average monthly flow of 91,000 MW; (ii) technically feasible potential: 31,000 MW (in 11 watersheds); and (iii) economically feasible potential: 22,000 MW (in 11 watersheds). Currently, Ecuador has used 24.55% of the economically feasible potential, i.e., 5401 MW [12].

3.3. Regions of Study and Dataset

Evidently, water is the main resource to produce power hydroelectric. Therefore, water inflow is a good predictor of hydroelectric performance [20,21]. In this study, we used precipitation data as a proxy for hydropower stations water inflow. Precipitation data are relevant because the regions of the study have a considerable rainfall variability, since they are located within a region with warm-humid climate [56]. Besides, precipitation data are typically used in weather derivatives [57], which are financial instruments that cover the effects of adverse or unexpected weather conditions. Consequently, they are a usual input in the decision-making process.

Note that the precipitation value is not an estimation of the inflow, but a potential predictor. In the literature, there are other works using streamflow data collected from sensors installed in the reservoirs [58,59]. Such approaches provide more accurate water inflow estimations but are also more costly to implement at a larger scale. At the time of this study, these data were not available in Ecuador. An alternative to consider in the future is to use stochastic models to generate this data [60,61].

We considered precipitation data from the three main Amazon watersheds, namely:

1. Napo, with an area of 59,505 km^2 fed by 15 rivers.
2. Pastaza, with an area of 23,190 km^2 fed by 11 rivers.
3. Santiago with an area 24,920 km^2 fed by four rivers.

The three basins add up an area of 107,615 km^2, which represent about 82% of the Amazon hydrological system and 42% of the country's total. They contain the 12 largest hydroelectric stations, generating about 82% of the hydropower production, as shown in Table 1.

Following the aim of the paper, we wanted to improve production forecasts by considering a minimum but significant amount of additional data, i.e., precipitation data. Precipitation data corresponding to the period 2000–2015 were obtained from pluviometric stations (PS) located inside the basins and managed by the INAMHI: Napo (13 PS), Pastaza (27 PS) and Santiago (28 PS). Figure 3 shows the geographical location of each PS inside their corresponding watersheds.
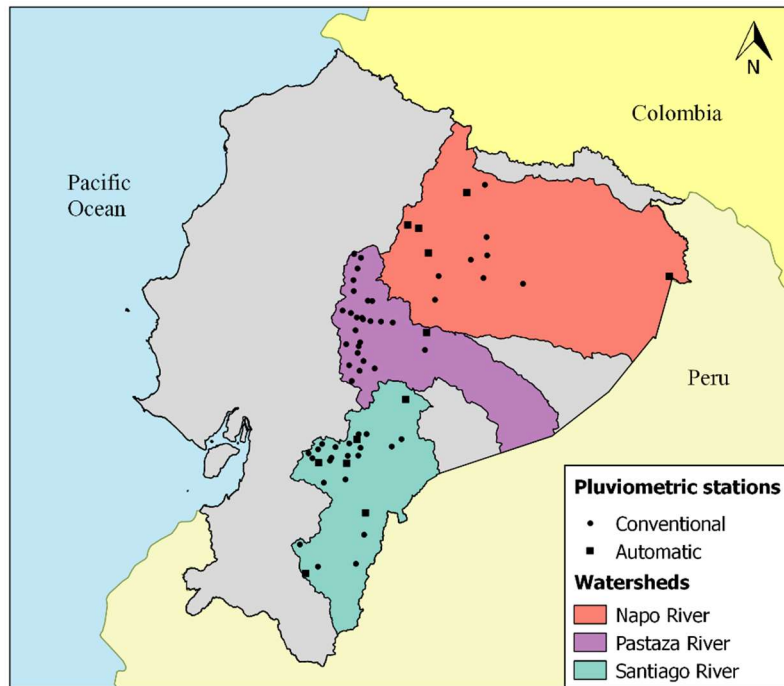
Figure 3. Geographical location of conventional and automatic pluviometric stations.

The source data included a single data point per station and month, representing the total rainfall measured by the station in the month. To obtain a single value per basin and month, we averaged these values for each basin. In this way, we have a single value for each basin representing a rough approximation of the expected monthly rainfall at any randomly selected point inside the basin. Furthermore, we assumed that basins are independent and thus aggregated the three monthly values to obtain the overall rainfall that can be potentially used to produce energy. Certainly, this entails a considerable simplification of the problem, and other aggregation methods could be considered.

Figure 4 shows the water shortage that normally occurs on the Amazon region from October to March [62]. The aggregated expected rainfall in the Napo, Pastaza, and Santiago basins is represented by the blue bar chart. The average of the total precipitation during these two years was 466.5 mm (black line). Precipitation in the months of October to March, in general, is below the average of the total precipitation since these months correspond to the Ecuadorian dry season. This drought influences the hydroelectric systems, which is evident in the curve of hydroelectric production.
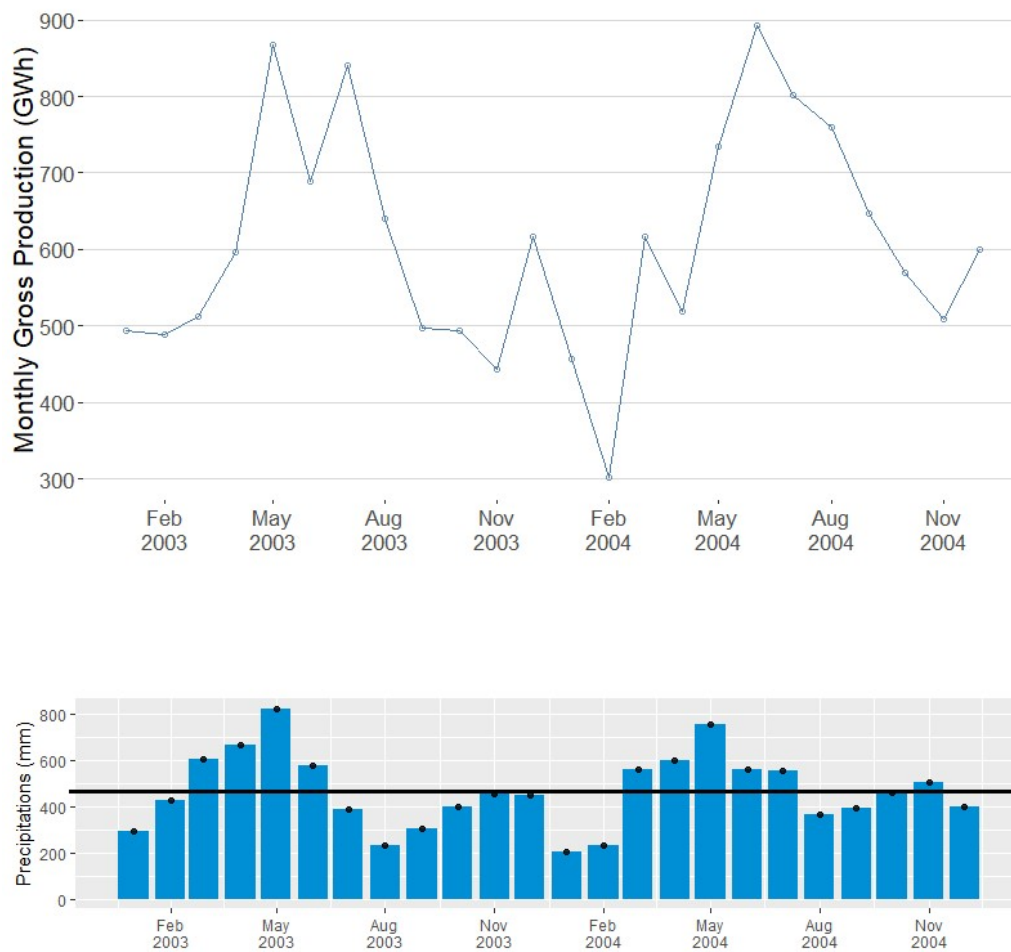
Figure 4. Monthly hydroelectric production and total precipitation of the three considered watersheds during 2003–2004.

3.4. Time Series

The following five series have been considered during the period 2000–2014:

1. Monthly gross production (MGP) of hydroelectric systems [GWh];
2. Average monthly precipitation (AMP) in Napo watershed [mm];
3. AMP in Pastaza watershed [mm];
4. AMP in Santiago watershed [mm];
5. Total average monthly precipitation (TAMP) in the three considered watersheds [mm]

To know the behavior and better fit in the analysis of the modeling, an analysis of each one of the five series was performed. Correlations between pairs of series were also studied, in particular, MGP-AMP$_{Napo}$, MGP-AMP$_{Pastaza}$, MGP-AMP$_{Santiago}$, and MGP-TAMP.

4. Experiments and Discussion

4.1. Exploratory Data Analysis and Series Transformation

Figure 5 shows the AMP of the Napo, Pastaza, and Santiago watersheds and MGP series during the period 2000–2014. All series show a direct correlation, i.e., as precipitation increases or decreases in watersheds, the production of hydropower increases or decreases. Also, the AMP of the watersheds is correlated with the MGP series. Moreover, the MGP series presents variability and trend over time, in contrast to AMP series that show variability but no trend through the time.
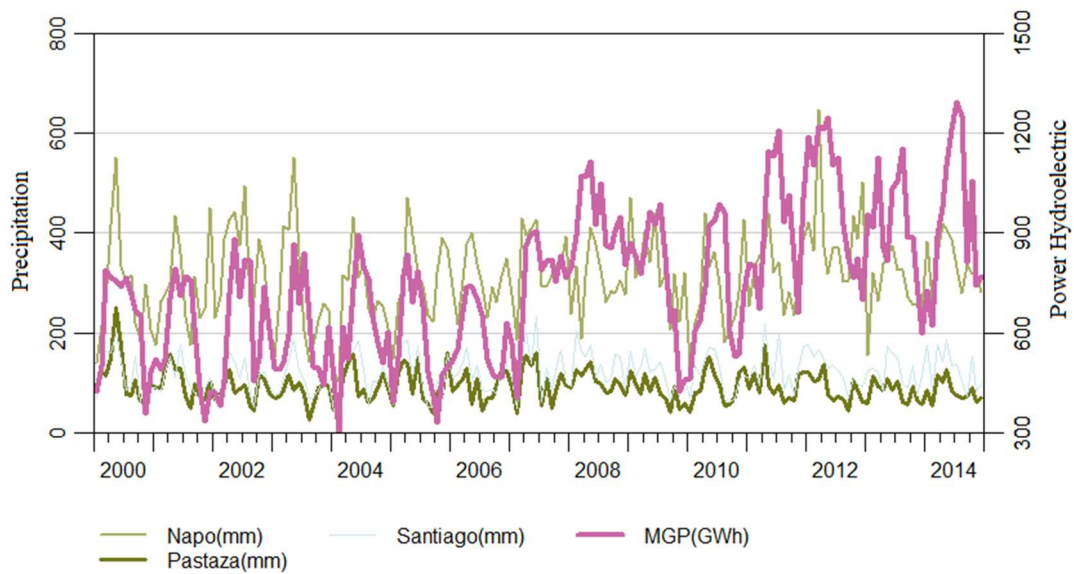
Figure 5. Time series of precipitation [mm] at river basins and hydroelectric production [GWh] in Ecuador, 2000–2014.

The three series of precipitation data show stationarity and seasonality. The hydroelectric production series that does not show stationarity but seasonality. As part of the study, the hydroelectric production series was transformed to make it stationary.

Figure 6 depicts the cross-correlation function (CCF) between two series at lag *k*, estimated as follows: If the simple cross-correlation is larger than the standard deviation error, CCF is considered significantly different from zero. When the joint series show correlation, it is necessary to separate the linear association to make them stationary. In the separation step, the residuals of the model are transformed into white noise in a process known as (pre-)whitening. In ARIMA models, through the correlogram functions ACF-PACF, the possible order values (*p*, *d*, *q*) are diagnosed to eliminate the trend and variability. In contrast, the dynamic relationship between two series is analyzed through the cross-correlation function adjusted to pre-whitened residuals [34], to estimate the best ARIMAX model.
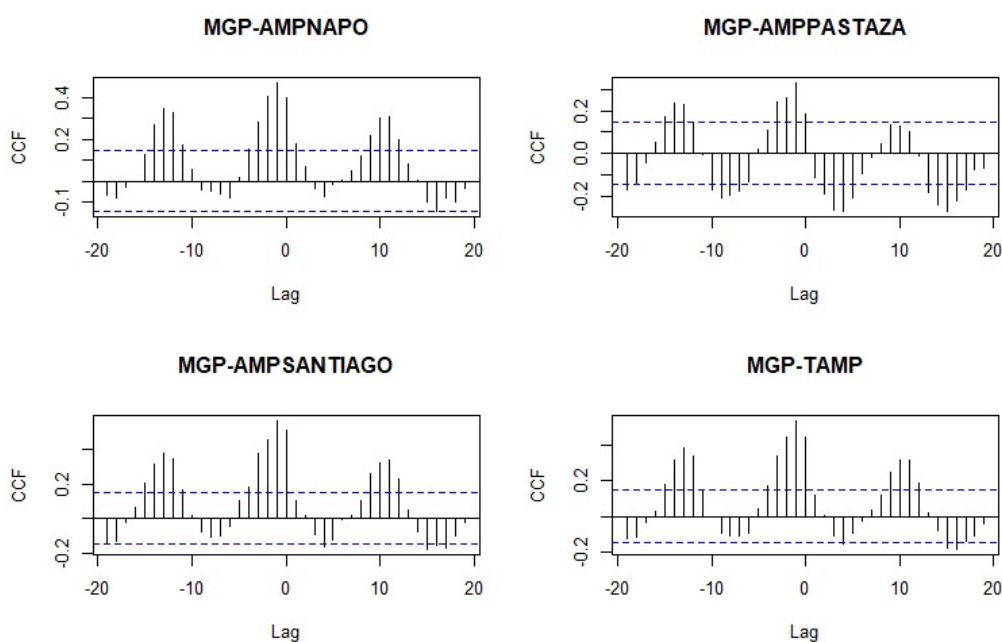

Figure 6. Cross-correlation function between precipitation and power hydroelectric.

4.2. Hydroelectric Energy Generation Modeled with ARIMA

Now we can proceed to the model identification step described in Figure 2. First, we focused on the ARIMA model.

Figure 7a presents the autocorrelation functions ACF and PACF used to evaluate stationary after pre-whitening. Both functions decay exponentially in a delay or lag of 5, which is significant with seasonal frequencies suggested by the SARIMA model. Figure 7b shows the same functions applied to the series of differences. The value of the Durbin-Watson statistic parameter obtained is 1.98, so the series presents evidence of weak positive autocorrelation.
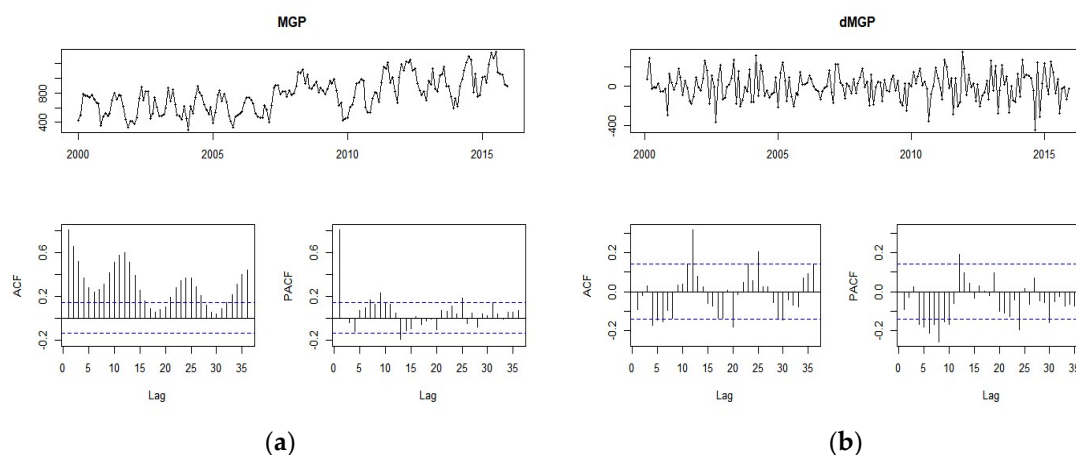

(**a**)   (**b**)

Figure 7. ACF and PACF autocorrelation functions (see Section 2.2.2). (**a**) Correlogram functions of the monthly gross production (MGP) series; (**b**) correlogram functions of the differenced MGP series.

For the analysis of the stationarity of the MGP series, we obtained the ADF statistics, whose value was t-ADF = −8.77 with a *p*-value of 0.01. There is significant evidence that the series is stationary and has no unit roots, with a maximum delay of 13 months. To determine the variants, the regression test was used to derive the ADF test which resulted to be significant with a value $t = -15.06$ at $\alpha = 0.05$ of significance level. This means that the series shows a random walk with mean zero.

Once we know that the series is stationary, we can step into the model identification stage. Table 2 shows a summary of the results obtained by applying the Box-Jenkins method to the MGP series. Figure 8 depicts comparison between actual and predicted values. The model minimizing AIC and BIC was ARIMA $(1,1,1) \times (0,0,1)_{12}$ with random walk. The model was confirmed to satisfy the assumptions of the SARIMA model. The validation results with the 2015 data were within the designated confidence interval.

Table 2. Box-Jenkins method applied in MGP series.

Model Identification			
	AR(1)	*MA(1)*	*SAR(1)*
Coefficients	0.65	−0.98	0.38
Standard Deviation	0.06	0.02	0.07
Model Estimation			
Penalized Likelihood Criteria		**AIC**	**BIC**
Values		2392.85	2405.86
Model Validation			
The model presents atypical values?			NO
Significance level			$\alpha = 0.05$

Table 2. *Cont.*

| Box-Ljung Test		
H$_0$	***p*-Values**	*Validation*
MGP series is uncorrelated	0.735	There is significant evidence that MGP series is uncorrelated
Box-Ljung Test with Squared Residuals		
H$_0$	***p*-Values**	*Validation*
MGP series residuals have constant variance	0.957	There is significant evidence that MGP series residuals have constant variance
Jarque-Bera Test		
H$_0$	***p*-Values**	*Validation*
MGP series residuals have constant variance	0.215	There is significant evidence that MGP series residuals are normal
Equation of the Model		
$X_t = 0.65 X_{t-1} + \varepsilon_t - 0.98 \varepsilon_{t-1} + 0.38\, X_{t-12}$		
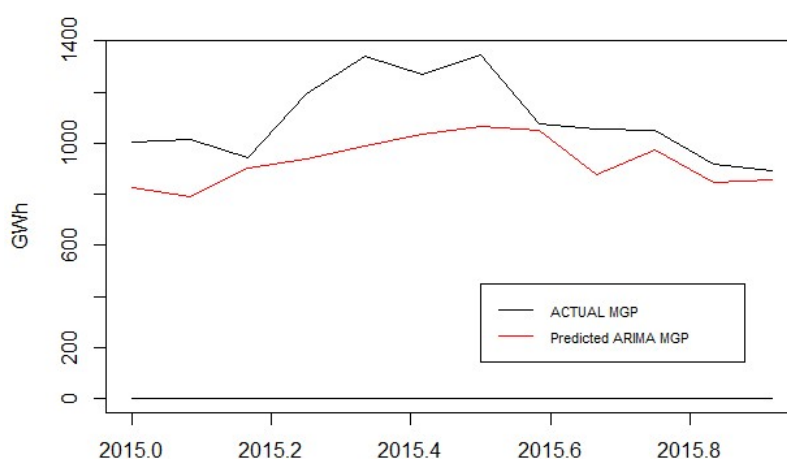
Figure 8. Comparison between actual and predicted values with the ARIMA model.

4.3. Hydroelectric Energy Generation Modelled with ARIMAX

Regarding the ARIMAX model, the amount of precipitation at each watershed was considered as the exogenous variable. The steps described in Section 2 produced the results presented below.

Figure 9 shows the residual diagnosis of each pair of series. The following characteristics were found: (i) Each series did not present outliers; (ii) each pair of series was uncorrelated; (iii) ACF lags were not significant; and iv) the p-value of the Ljung-Box test was greater than 0.05, which indicates that the squared residuals were uncorrelated over time, i.e., the standardized residuals were independent.

Table 3 shows the coefficients of the best ARIMAX models for hydroelectric production with the exogenous variable. The model with the lowest AIC and BIC is the pair of MGP-AMP$_{SANTIAGO}$ series. Figure 10 shows the forecasts of each ARIMAX model for the year 2015.
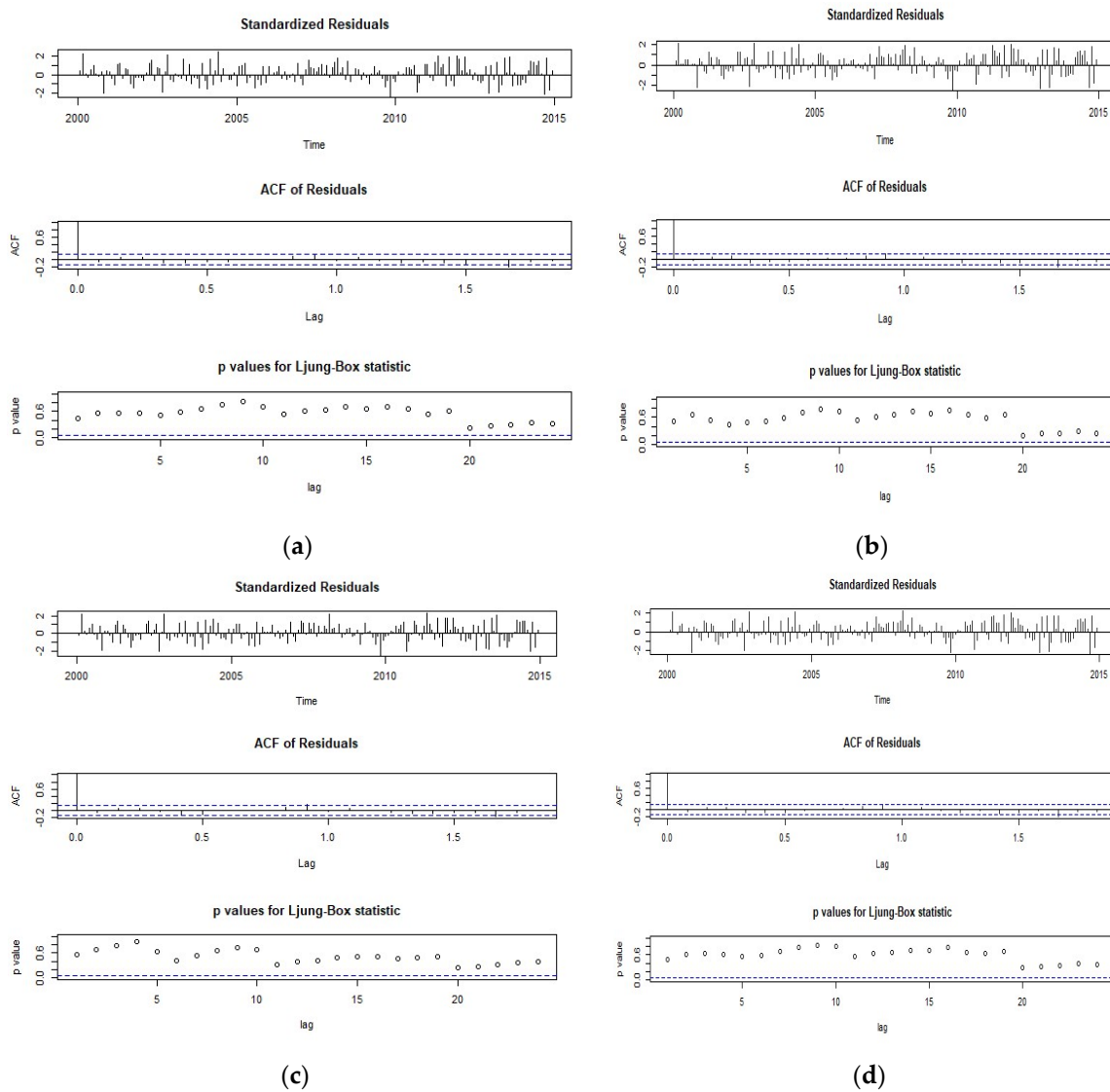
Figure 9. Residual diagnosis of each pair of series. (**a**) Correlogram MGP-AMP$_{PASTAZA}$ average monthly precipitation AMP$_{PASTAZA}$; (**b**) series MGP-AMP$_{NAPO}$; (**c**) Correlogram MGP-AMP$_{SANTIAGO}$; (**d**) series MGP-TAMP.
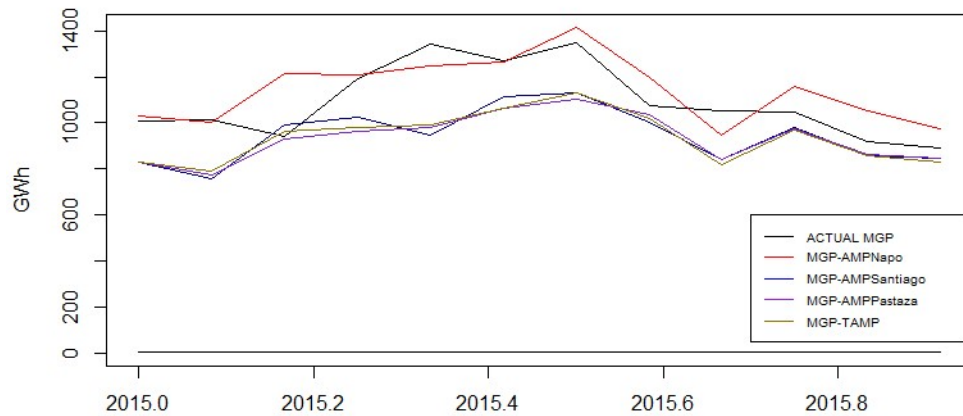

Figure 10. Comparison between actual and predicted values 2015 with ARIMAX Models.

Table 3. Box-Tiao method applied in MGP series with exogenous variable.

Model Identification				
Models	**AR(1)**	**MA(1)**	**SAR(1)**	**XREG**
MGP-AMP$_{PASTAZA}$ **ARIMAX (1,1,1)(1,0,0)$_{12}$**				
Coefficients	0.66	−0.96	0.40	0.80
Standard Deviation	0.06	0.02	0.07	0.28
MGP-AMP$_{NAPO}$ **ARIMAX (1,1,1)(1,0,0)$_{12}$**				
Coefficients	0.66	−0.97	0.38	0.22
Standard Deviation	0.06	0.02	0.08	0.10
MGP-AMP$_{SANTIAGO}$ **ARIMAX (1,1,1)(1,0,0)$_{12}$**				
Coefficients	0.66	−0.97	0.38	1.21
Standard Deviation	0.07	0.02	0.08	0.21
MGP-AMP **ARIMAX (1,1,1)(1,0,0)$_{12}$**				
Coefficients	0.65	−0.96	0.37	0.25
Standard Deviation	0.07	0.02	0.08	0.06
Model Estimation				
Penalized Likelihood Criteria	**MGP-AMP$_{PASTAZA}$**	**MGP-AMP$_{NAPO}$**	**MGP-AMP$_{SANTIAGO}$**	**MGP-AMP**
AIC	2237.40	2237.85	2213.66	2230.85
BIC	2253.34	2256.36	2229.60	2246.78

Model Validation			
Models	**Outliers**	**Assumption Tests**	**p-Values**
MGP-AMP$_{PASTAZA}$	AO and IO were not detected	Box-Ljung Box-Ljung test with squared residuals Jarque-Bera	0.44 0.74 0.43
MGP-AMP$_{NAPO}$	AO and IO were not detected	Box-Ljung Box-Ljung test with squared residuals Jarque-Bera	0.50 0.70 0.38
MGP-AMP$_{SANTIAGO}$	AO and IO were not detected	Box-Ljung Box-Ljung test with squared residuals Jarque-Bera	0.55 0.83 0.35
MGP-TAMP	AO and IO were not detected	Box-Ljung Box-Ljung test with squared residuals Jarque-Bera	0.56 0.69 0.06

| Equation of the Models	
MGP-AMP$_{PASTAZA}$	$X_t = 0.66X_{t-1} + \varepsilon_t - 0.96\varepsilon_{t-1} + 0.40X_{t-12} + 0.80X_{REG}$
MGP-AMP$_{NAPO}$	$X_t = 0.66X_{t-1} + \varepsilon_t - 0.97\varepsilon_{t-1} + 0.38X_{t-12} + 0.22X_{REG}$
MGP-AMP$_{SANTIAGO}$	$X_t = 0.66X_{t-1} + \varepsilon_t - 0.97\varepsilon_{t-1} + 0.38X_{t-12} + 1.21X_{REG}$
MGP-TAMP	$X_t = 0.65X_{t-1} + \varepsilon_t - 0.96\varepsilon_{t-1} + 0.37X_{t-12} + 0.25X_{REG}$

4.4. Comparison and Discussion

Table 4 shows the MAE, MAPE, and MASE of the forecasts of each model for the year 2015. It can be seen that the best results (lower error) were obtained by the ARIMAX model of MGP-AMP$_{NAPO}$. This is corroborated in Figure 10, which indicates that MGP-AMP$_{NAPO}$ was better.

Models	MAE (GWh)	MAPE (%)	MASE
MGP-ARIMA	101.08	14.32	0.68
MGP-AMP$_{PASTAZA}$ ARIMAX	71.13	10.21	0.55
MGP-AMP$_{NAPO}$ ARIMAX	70.81	10.17	0.57
MGP-AMP$_{SANTIAGO}$ ARIMAX	91.63	13.31	0.62
MGP-TAMP ARIMAX	97.00	13.88	0.65

Figure 11 visually shows the forecast of hydroelectric production in Ecuador for 2015, considering ARIMA and ARIMAX models. A better forecast was obtained with the ARIMAX models of MGP-AMP$_{NAPO}$.
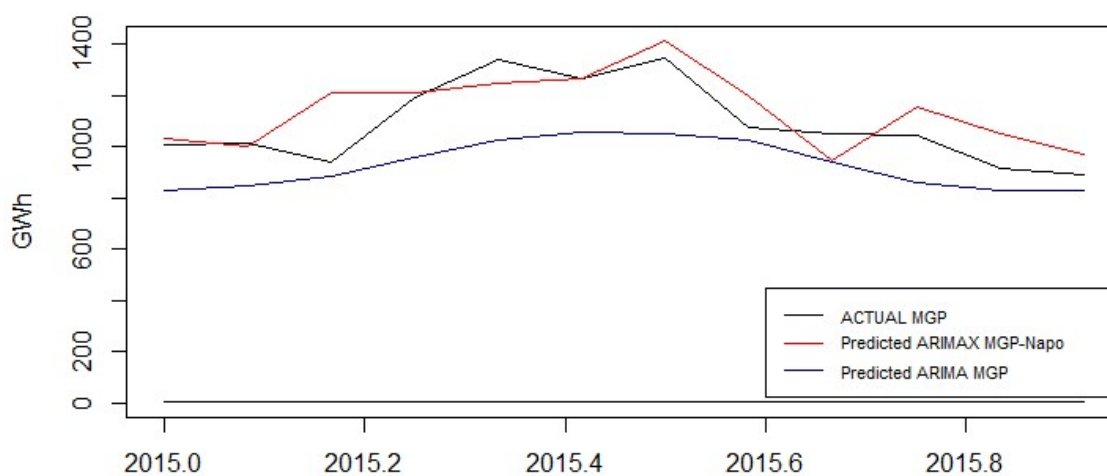
Figure 11. Predicted values comparison of ARIMA and ARIMAX models.

Although the ARIMAX model of MGP-AMP$_{SATIAGO}$ has the best estimation model considering the AIC and BIC values, ARIMAX model MGP-AMP$_{NAPO}$ has the lower forecast error values. Hence, ARIMAX model MGP-AMP$_{NAPO}$ was selected.

Figure 11 depicts that ARIMAX MGP-AMPNAPO $(1,1,1)(1,0,0)_{12}$ model presented a better predictivity of gross monthly hydroelectric energy production in Ecuador. This model estimated that the occurrence of the production is for the parameter $p = 1$ and a suitable moving average of $q = 1$ for the regular component. At the same time, for seasonal component, $P = 1$ and the strong seasonality established as $S = 12$. The exogenous variable are precipitation values at Napo watershed. The value of MAE is 70.81 GWh, MAPE is 10.17%, and MASE is 0.57.

5. Conclusions

This work shows that the ARIMAX models with an exogenous variable have better performance than the univariate ARIMA models for predicting hydroelectric energy production in Ecuador. Our analysis of the production and the precipitation of the Napo, Pastaza, and Santiago basins (alone and aggregated) yielded that the ARIMAX MGP-AMP$_{NAPO}$ $(1,1,1)(1,0,0)_{12}$ model adequately adjusts the data of the hydroelectric energy production in Ecuador series.

This research helps to describe and predict hydropower generation, particularly in Ecuador. Results obtained with the proposed model can be useful for the organizing and planning of the electric sector, which are important for the energy policymaker sector. The methodology can be also extended to use in other cases.

Further studies to improve the accuracy of the predictions can be performed including bioinspired techniques and optimization algorithms. Besides, we plan to study the impact of incorporating additional data sources to the model, e.g., synthetic streamflow data, as mentioned in Section 3.3, or

other weather parameters, like wind, rain, etc. The applicability of the approach to other countries remains as future work.

References

1. International Energy Agency (IEA). Key World Energy Statistics 2018. Available online: https://webstore.iea.org/key-world-energy-statistics-2018 (accessed on 26 February 2019).

2. International Energy Agency (IEA). Electricity Information: Overview 2018. Available online: https://webstore.iea.org/electricity-information-2018-overview (accessed on 26 February 2019).

3. Sher, F.; Pans, M.A.; Afilaka, D.T.; Sun, C.; Liu, H. Experimental investigation of woody and non-woody biomass combustion in a bubbling fluidised bed combustor focusing on gaseous emissions and temperature profiles. *Energy* **2017**, *141*, 2069–2080. [CrossRef]

4. Sher, F.; Pans, M.A.; Sun, C.; Snape, C.; Liu, H. Oxy-fuel combustion study of biomass fuels in a 20 kWth fluidized bed combustor. *Fuel* **2018**, *215*, 778–786. [CrossRef]

5. Zhang, Y.; Fang, Y.; Jin, B.; Zhang, Y.; Zhou, C.; Sher, F. Effect of Slot Wall Jet on Combustion Process in a 660 MW Opposed Wall Fired Pulverized Coal Boiler. *Int. J. Chem. React. Eng.* **2019**, *17*, 1–13. [CrossRef]

6. Cevallos-Sierra, J.; Ramos-Martin, J. Spatial assessment of the potential of renewable energy: The case of Ecuador. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1154–1165. [CrossRef]

7. Barzola-Monteses, J.; Espinoza-Andaluz, M. Performance Analysis of Hybrid Solar/H2/Battery Renewable Energy System for Residential Electrification. In Proceedings of the 10th International Conference Application Energy, Hong Kong, China, 22–25 August 2018; Elsevier: Hong Kong, China, 2018.

8. UASB-DIGITAL. ¿Es sustentable la política energética en el Ecuador? Available online: http://hdl.handle.net/10644/3036 (accessed on 11 March 2019).

9. Buñay, F.; Pérez, F. Comparación de Costos de Producción de Energía Eléctrica Para Diferentes Tecnologías en el Ecuador. Bachelor's Thesis, Universidad de Cuenca, Cuenca, Spain, 2012.

10. Carvajal, P.E.; Li, F.G.N.; Soria, R.; Cronin, J.; Anandarajah, G.; Mulugetta, Y. Large hydropower, decarbonisation and climate change uncertainty: Modelling power sector pathways for Ecuador. *Energy Strateg. Rev.* **2019**, *23*, 86–99. [CrossRef]

11. Agencia de Regulación y Control de Electricidad (ARCONEL). Available online: https://www.regulacionelectrica.gob.ec/ (accessed on 3 January 2019).

12. Ministerio de Electricidad y Energía Renovable (MEER). *Plan Maestro de Electricidad 2016–2025*; ARCONEL: Quito, Ecuador, 2016; pp. 1–440.

13. Agencia de Regulación y Control de Electricidad (ARCONEL). *Estadística Anual y Multianual del Sector Eléctrico Ecuatoriano (2016)*; ARCONEL: Quito, Ecuador, 2016.

14. International Hydropower Association (IHA). *Hydropower Status Report*; IHA: London, UK, 2017.

15. Ministerio de Electricidad y Energía Renovable (MEER). *Rendición de cuentas MEER 2017*; MEER: Quito, Ecuador, 2017; pp. 1–39.

16. Ordoñez, L. Centro de Investigación y Capacitación Eléctrica. Available online: https://cicecuador.org/ (accessed on 2 January 2019).

17. Operador Nacional de Electricidad (CENACE). Available online: http://www.cenace.org.ec/docs/InformacionOperativa.htm (accessed on 2 January 2019).

18. Elamin, N.; Fukushige, M. Modeling and forecasting hourly electricity demand by SARIMAX with interactions. *Energy* **2018**, *165*, 257–268. [CrossRef]

19. Koivisto, M.; Das, K.; Guo, F.; Sørensen, P.; Nuño, E.; Cutululis, N.; Maule, P. Using time series simulation tools for assessing the effects of variable renewable energy generation on power and energy systems. *WIREs Energy Environ.* **2019**, *8*, 1–15. [CrossRef]

20. Dehghani, M.; Riahi-Madvar, H.; Hooshyaripor, F.; Mosavi, A.; Shamshirband, S.; Zavadskas, E.K.; Chau, K. Prediction of hydropower generation using grey Wolf optimization adaptive neuro-fuzzy inference system. *Energies* **2019**, *12*, 289. [CrossRef]

21. Mite-León, M.; Barzola-Monteses, J. Statistical Model for the Forecast of Hydropower Production in Ecuador. *Int. J. Renew. Energy Res.* **2018**, *10*, 1130–1137.

22. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed.; OTexts: Melbourne, Australia, 2018.

23. Shumway, R.H.; Stoffer, D.S. *Time Series Analysis and Its Applications with R Example*, 4th ed.; Springer: Pittsburgh, PA, USA, 2017.

24. Maçaira, P.M.; Marcio, A.; Thomé, T.; Luiz, F.; Oliveira, C. Time series analysis with explanatory variables: A systematic literature review. *Environ. Model. Softw.* **2018**, *107*, 199–209. [CrossRef]

25. Raza, M.Q.; Nadarajah, M.; Ekanayake, C. On recent advances in PV output power forecast. *Sol. Energy* **2016**, *136*, 125–144. [CrossRef]

26. Camelo, N.; Sérgio, P.; Bosco, J.; Leal, V.; Cesar, P.; Carvalho, M.D. A hybrid model based on time series models and neural network for forecasting wind speed in the Brazilian northeast region. *Sustain. Energy Technol. Assess.* **2018**, *28*, 65–72.

27. Soman, S.S.; Zareipour, H.; Malik, O.; Mandal, P. A review of wind power and wind speed forecasting methods with different time horizons. In Proceedings of the North American Power Symposium 2010, Arlington, TX, USA, 26–28 September 2010; pp. 1–8.

28. Kim, D.; Hur, J. Short-term probabilistic forecasting of wind energy resources using the enhanced ensemble method. *Energy* **2018**, *157*, 211–226. [CrossRef]

29. Lee, Y.; Hur, J. A simultaneous approach implementing wind-powered electric vehicle charging stations for charging demand dispersion. *Renew. Energy* **2019**, *144*, 172–179. [CrossRef]

30. Camelo, H.; Lucio, P.S.; Leal Junior, J.B.V.; Marques de Carvalho, P.C.; Glehn dos Santos, D. Innovative hybrid models for forecasting time series applied in wind generation based on the combination of time series models with artificial neural networks. *Energy* **2018**, *151*, 347–357. [CrossRef]

31. Inman, R.H.; Pedro, H.T.C.; Coimbra, C.F.M. Solar forecasting methods for renewable energy integration. *Prog. Energy Combust. Sci.* **2013**, *39*, 535–576. [CrossRef]

32. Pektas, A.O.; Cigizoglu, H.K. ANN hybrid model versus ARIMA and ARIMAX models of runoff coefficient. *J. Hydrol.* **2013**, *500*, 21–36. [CrossRef]

33. Suganthi, L.; Samuel, A.A. Energy models for demand forecasting—A review. *Renew. Sustain. Energy Rev.* **2012**, *16*, 1223–1240. [CrossRef]

34. Box, G.E.P.; Jenkings, G.; Reinsel, G. *Time Series Analysis*; Prentice Hall: New Jersey, NJ, USA, 1994.

35. Cryer, J.D.; Chan, K.S. *Time Series Analysis with Applications in R*, 2th ed.; Springer: Iowa City, IA, USA, 2008.

36. Box, G.E.P.; Pierce, D.A. Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models. *J. Am. Stat. Assoc.* **1970**, *65*, 1509–1526. [CrossRef]

37. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, Revised ed.; Holden-Day: Oakland, CA, USA, 1976.

38. Kramer, W. Durbin-Watson Test. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Kragujevac, Serbia, 2011.

39. Dickey, D.A.; Fuller, W.A. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica* **1981**, *49*, 1057–1072. [CrossRef]

40. Said, S.E.; Dickey, D.A. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **1984**, *71*, 599–607. [CrossRef]

41. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]

42. Pfaff, B. VAR, SVAR and SVEC Models: Implementation Within R Package Vars. *J. Stat. Softw.* **2008**, *27*, 1–32. [CrossRef]

43. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast* **2006**, *22*, 679–688. [CrossRef]

44. Franses, P.H. A note on the Mean Absolute Scaled Error. *Int. J. Forecast* **2016**, *32*, 20–22. [CrossRef]

45. Instituto Nacional de Meteorología e Hidrología (INAMHI). Available online: http://www.serviciometeorologico.gob.ec/ (accessed on 3 January 2019).

46. Rodríguez Carreño, M.G. Impacto del Proyecto Hidroeléctrico Coca Codo Sinclair en la Matriz Energética Como Generador de Oportunidades del Ecuador en el Mundo. Bachelor's Thesis, Universidad del Azuay, Cuenca, Spain, 2015.

47. Chen, Y. Impacto Socio-Económico del Proyecto Hidroeléctrico Coca Codo Sinclair, Construido por la Empresa China Sinohydro, Para la Economía Ecuatoriana. Bachelor's Thesis, Pontificia Universidad Católica del Ecuador, Quito, Ecuador, 2015.

48. López, V. El proyecto hidroeléctrico Coca Codo Sinclair y la gobernanza energética en la Amazonía ecuatoriana. *Let. Verdes* **2011**, *8*, 1–3. [CrossRef]

49. Corporación Eléctrica del Ecuador (CELEC). Available online: https://www.celec.gob.ec/ (accessed on 5 March 2019).

50. CELEC—Hidropaute 2019. Available online: https://www.celec.gob.ec/hidropaute/ (accessed on 5 March 2019).

51. CELEC—Coca Codo Sinclair 2019. Available online: https://www.celec.gob.ec/cocacodosinclair/ (accessed on 5 March 2019).

52. CELEC—Hidroagoyán 2019. Available online: https://www.celec.gob.ec/hidroagoyan/ (accessed on 5 March 2019).

53. CELEC—Hidronación 2019. Available online: https://www.celec.gob.ec/hidronacion/ (accessed on 5 March 2019).

54. Electro Generadora del Austro (ELECAUSTRO). Available online: http://www.elecaustro.com.ec/ (accessed on 5 March 2019).

55. Consejo Nacional de Recursos Hídricos (CNRH). *División Hidrográfica del Ecuador*; Ministerio del Ambiente: Quito, Ecuador, 2002.

56. Laraque, A.; Loup Guyot, J.; Pombosa, R. *Hidroclimatología del Oriente e Hidrosedimentología de la Cuenca del Napo. La Cuenca Oriente Geol. y Petróleo*; Institut Français d'études Andines: Quito, Ecuador, 2004; pp. 131–351.

57. Hernández, J.; Carvajal-Serna, F. Cobertura al riesgo ante la variabilidad hidrológica en una central hidráulica a filo de agua usando derivados climáticos. *Lect. Econ.* **2017**, *87*, 191–222. [CrossRef]

58. Aronica, G.T.; Bonaccorso, B. Climate change effects on hydropower potential in the Alcantara River basin in Sicily (Italy). *Earth Interact* **2013**, *17*, 1–22. [CrossRef]

59. Can, I.; Tosunoğlu, F. The estimation of the firm hydroelectric potential of the Çoruh Basin, Turkey. *Energy Sources Part A Recover Util. Environ. Eff.* **2012**, *34*, 1609–1618. [CrossRef]

60. Dashora, I.; Singal, S.K.; Srivastav, D.K. Streamflow prediction for estimation of hydropower potential. *Water Energy Int.* **2015**, *57*, 54–60.

61. Zhao, T.; Zhao, J. Forecast-skill-based simulation of streamflow forecasts. *Adv. Water Resour.* **2014**, *71*, 55–64. [CrossRef]

62. Cevallos Escobar, J.G.; Macas Díaz, D.F. Análisis de la Complementariedad Hidrológica de las Vertientes del Amazonas y del Pacífico en el Ecuador Considerando las Nuevas Centrales Hidroeléctricas Proyectadas Hasta el Año 2016. Ph.D. Thesis, Escuela Politécnica Nacional, Quito, Ecuador, 2012.