Conditional seasonal markov-switching autoregressive model to simulate extreme events: Application to river flow

Bassel Habeeb, Emilio Bastidas-Arteaga, Mauricio Sánchez-Silva, You Dong

Please cite this paper as: Habeeb, B., Bastidas-Arteaga, E., Sánchez-Silva, M., & Dong, Y. (2024). Conditional seasonal markov-switching autoregressive model to simulate extreme events: Application to river flow. *Environmental Modelling & Software*, *178*, 106066. https://doi.org/10.1016/j.envsoft.2024.106066

Conditional Seasonal Markov-Switching Autoregressive Model to Simulate Extreme Events: Application to River Flow

Bassel Habeeb[a], Emilio Bastidas-Arteaga[a, *], Mauricio Sánchez-Silva[b], You Dong[c]

[a] *Laboratory of Engineering Sciences for the Environment UMR CNRS 7356, La Rochelle University, France.* bassel.habeeb@univ-lr.fr, ebastida@univ-lr.fr
[b] *Department of Civil & Environmental Engineering, University of Los Andes, Colombia.* msanchez@uniandes.edu.co
[c] *Department of Civil & Environmental Engineering, Hong Kong Polytechnic University, Hong Kong.* you.dong@polyu.edu.hk

Abstract

Extreme events have the potential to significantly impact transportation infrastructure performance. For example, in the case of bridges, climate change impacts the river discharge, hence scouring patterns, which in turn, affects the bridge foundation stability. Therefore, extreme events (river flow) forecasting is mandatory in bridge reliability analysis. This paper approaches this river flow forecasting problem by developing a Markov-Switching Autoregressive model coupled with a conditional hidden seasonal Markov component. In addition, the proposed model is also combined with the deep machine learning neural networks method to forecast river flow from a dataset or from simulations. The proposed method is illustrated by using realistic data: historic river flow values of the Thames River. The results indicate that the proposed model well represented the extreme events within the dataset. In terms of river flow forecasting, the results indicate that the forecasts improve when the training period changes from 20 years to 40 years.

Keywords: Seasonal Markov-Switching Autoregressive model; Conditional hidden seasonal Markov process; River flow forecasting; Extreme events; Recurrent Neural Networks.

Software and data availability:
Software name: SMSAR_model
Developer: Bassel Habeeb
First year available: 2023
Hardware requirements: PC/Mac/Linux
Software requirements: R statistical environment and language
Program language: R
Program size: 71.8 KB
Availability: https://github.com/ebastidasarteaga/SMSAR_model
License: GPL-3.0
Data for benchmarking: https://github.com/ebastidasarteaga/SMSAR_model/blob/main/UKdatabase.xlsx
Size of archive: 47.6 KB

[*] Corresponding author: Emilio Bastidas-Arteaga, Phone: +33 (0)5 86 56 22 32, email: ebastida@univ-lr.fr

Notation and Acronyms:

AR:	Autoregressive.
ARIMA:	Autoregressive Integrated Moving Average.
c:	Constant term of the Autoregressive process.
E:	Severe values limit.
Ev_t:	Months of interest effect value.
HSM:	Conditional Hidden Seasonal Markov component.
I and J:	Markovian regimes states.
k:	Markovian state.
M:	Switching regimes number.
MES_y:	Monthly effect based on the yearly seasonality.
ML:	Machine learning.
Moi_t:	Months of interest increase rate.
MS:	Markov-Switching.
MSAR:	Markov-Switching Autoregressive.
NHMSAR:	(Non)- Homogeneous Markov-Switching Autoregressive.
NN:	Deep machine learning Neural Networks forecasting method.
$n_{t_{Moi}}$:	Number of months of interest.
p:	Autoregressive order.
Q:	Upper quartile of the dataset.
$\mathcal{S}_m$:	Seasonal transition probability matrix.
SMSAR:	Seasonal Markov-Switching Autoregressive.
$\mathcal{S}_t$:	Hidden Markov chain.
S_y:	Yearly seasonality.
S_{yR}:	Yearly seasonality remainder.
t:	Months of the year.
t_{Moi}:	Months of interest.
t_y:	Yearly time series.
WN:	White Noise.
X_t:	Time state.
Y_t:	Monthly time series state.
Z:	Lower quartile of the dataset.
φ:	Autoregressive model parameter.
Ψ:	Current state of the probability transition matrix.
ε_t:	Errors.
μ:	Mean value.
ε_t:	Random white noise.
σ:	Standard deviation.
Ω:	Transition state of the probability transition matrix.
σ^2:	Variance.

1. INTRODUCTION

The lifetime of infrastructure assets depends on a variety of aspects, which include construction materials, the evolution of the demand, changes in the surrounding environment, as well as design assumptions, maintenance, and operation policies. All physical systems change and suffer a loss in capacity through time due to progressive deterioration (corrosion, fatigue, creep, etc.) [1–3] or shock-based deterioration (extreme events) [4]. This decrease in capacity yields an increase in the likelihood of failure and therefore, regular inspections and maintenance are essential to ensure adherence to the necessary safety standards [5–7]. In the case of bridges crossing rivers, scouring poses a prevalent threat to the structure's integrity. This phenomenon primarily occurs during periods of intense flooding, when the escalated forces of water displace sediments from the riverbed surrounding bridge foundations. It emerges as a significant contributor to the loss of structural capacity and potential failures of bridges crossing rivers [8–10]. For instance, scouring is responsible for 60 percent of bridge failures in the United States, resulting in an annual maintenance cost of $30 million [11].

Extreme weather events such as droughts and floods are expected to increase in frequency and intensity in some places due to climate change, which in turn will impact the infrastructure in several ways [12]. In the case of bridges crossing rivers, climate change may have a direct impact on the river discharge and, consequently, may modify the bridge scouring patterns since precipitation patterns could become more volatile and unpredictable, causing floods. Accordingly, forecasting river flow is essential for identifying potential problems such as droughts or floods, formulating mitigation plans, and increasing resilience. In addition, this is also important to make informed decisions about water allocation and management.

Advancements in modelling capabilities have made it possible to generate flood simulations [13] and deploy comprehensive flood forecasting systems, which are large-scale systems based on differential equations, these systems play a pivotal role in informing warnings about potential floods [14]. Hydrological modelling systems, such as HEC-RAS and HEC-HMS can be used in investigating the multifaceted impacts of flooding events, encompassing phenomena such as, scour, sediment transport, and water quality [15]. Forecasting river flow is mostly based on climate models; however, it is essential to highlight the inherent uncertainties stemming from the climate models' variability, forcing factors, aerosols, jet stream impact, resolution, etc. On the other hand, time series forecasting methods may be a preferable choice for short-term forecasts since they are based on actual data observed directly from the system.

Hydrological time series models make forecasts based on historical data, from which it is possible to extract statistical information and associated seasonal and cyclic patterns [16,17]. For example, river flow discharge values can be simulated with non-linear models [18], however, those models are not able to simulate the long-term dependence of river flow. A quantum leap within the long-term dependence occurred by considering the fractionally integrated process within the Autoregressive Integrated Moving Average (ARIMA) model, this is achieved by inserting the long-term dependence into the Box & Jenkins framework, and the resulting model is the Autoregressive Fractional Integrated Moving Average (ARFIMA) model. This model can simulate the river flow, as it captures the long-term memory and the non-stationary nature of the system. Nevertheless, even with seasonal effects, this model shows inaccurate hydrological forecasts due to the conservation of the mean value in the analysis. For example, in the case study of the Tisza River [19], the ARFIMA model with seasonal effects failed to simulate the river flow hydrological fact, i.e., high river flow values for a short time interval are followed by low river flow values for a long time interval.

The regime-switching technique can be used to improve the model performance to simulate the river flow hydrological fact, i.e., high river flow values for a short time interval are followed by low river flow values for a long time interval. This is achieved by defining regimes equivalent to wet and dry periods [20–22]. Consequently, regime-switching models are a preferable choice within the framework of hydrological time series analysis. The regime-switching process improves the concept of stochastic processes within the context of time series analysis by using the hidden Markov process [23]; this process is a type of probabilistic model used to describe a sequence of events or observations where the state is not directly visible, but the output, dependent on the state, is visible, i.e., changes in mean and variance. The Markov-Switching (MS) process was proposed to capture the discrete changes as a result of the volatility behavior detected in economic applications [24].

The Markov-Switching Autoregressive (MSAR) model is an improvement to the Markov-Switching (MS) process, providing a more comprehensive method for identifying structural breaks in the time series dynamics by embedding an autoregressive component. This helps to describe the time evolution of the series more accurately, in which the switching between the autoregressive processes is regulated by a hidden Markov chain. Accordingly, the MSAR model was first approached within the context of statistical climate analysis in 2007, as climatic conditions depend on past patterns [25], in which the

variations within the MSAR model were associated with periodic severe volatility [26]. Further investigations assured the ability of the MSAR model to accurately represent climatic changes [27,28]. In addition, the significance of the regime-switching process arose in modeling the seasonal accumulation of the river flow hydrological fact, i.e., high river flow values for a short time interval are followed by low river flow values for a long time interval [29]. However, the main limitation within the latest regime-switching models when simulating a hydrological time series dataset is the low quality of the simulations over a long period (hundreds of years), especially its ability to simulate extreme events/severe values. In other words, the Markov-Switching approach has neither been investigated in simulating a long-time series dataset nor extreme events.

Recently, Machine Learning (ML) methods have been applied within the context of hydrological time series with great success due to their ability to make forecasts, in which the performance of ML methods has surpassed traditional mathematical models by effectively addressing complex regression mathematical problems [30,31]. In this context, the artificial intelligence Neural Network (NN) system is a deep learning method composed of interconnected neurons, which process and transmit information from the historical dataset to solve complex problems [32]. The NN systems, i.e., Recurrent, Random forest, Convolutional, and Multiple perceptions, have shown success within the context of hydrological time series forecasts [33–35]. The integration of a regime-switching process into the Recurrent NN system has been shown to improve the precision of the forecasts, as the regime-switching process helps the NN system to better recognize and respond to changes in the data [36]. In addition, the length of the historical patterns fed into the model seems to impact the accuracy of the forecasts and should be further investigated [37,35].

Lately, time series forecasting models based on Short-term and Long-term memory processes were not able to simulate extreme events/severe values of a hydrological time series. In addition, the impact of the training period requires further investigation in the framework of hydrological time series data (river flow), and whether a fitted simulation can provide more accurate forecasts than forecasting directly the hydrological dataset.

This paper overcomes the former limitations within the literature regarding modeling extreme river flow discharge values. This paper proposes a novel stochastic framework to simulate extreme values and forecast hydrological time series data. The objectives of this paper are outlined as follows.

- Simulate the hydrological facts of rivers, i.e., high river flow values for a short time interval are followed by low river flow values for a long time interval, and recognize the dynamic pattern of the flow.
- Simulate extreme river flow values.
- Investigate the impact of the training period on the pattern recognition of the proposed SMSAR model forecasts using deep machine learning methods.
- Investigate whether forecasting the database or forecasting the fitted SMSAR model simulation provides more accurate results using deep machine learning methods.

In this paper, the proposed Seasonal Markov-Switching Autoregressive (SMSAR) model is initiated as MSAR model and is then coupled with a conditional Hidden Seasonal Markov (HSM) component to enhance its performance. Afterward, the SMSAR simulation and the dataset of a given period are then used for training the deep machine learning Neural Networks (NN) method, i.e., Recurrent NN system to evaluate whether forecasting the database or forecasting the fitted proposed simulation provides more accurate results and to investigate the impact of the training period on the pattern

recognition of the forecasts. All methodological developments are illustrated with the hydrological time-series analysis of the Thames River, and the results are compared with those of the MSAR model.

This paper is organized as follows. Section 2.1 provides a flow diagram of the methodology and an overview of the numerical phases considered for simulating the SMSAR model, severe values, and forecasting; these numerical phases are also illustrated by a pseudo code algorithm in Appendix. 1. Section 2.2 provides an in-depth illustration of the SMSAR model. Section 2.3 sets out the threshold for determining severe values. Section 3 describes the case study that will be analyzed and discussed in Section 4.1 to assess the quality of the proposed SMSAR model. Section 4.2 investigates the ability of the SMSAR model in simulating severe values. Section 4.3 investigates using the Recurrent Neural Networks method; whether forecasting the dataset or forecasting the fitted simulation of the proposed SMSAR model provides better results. In addition, this section also investigates the training period's impact on pattern recognition.

2. METHODS

2.1. General description

This paper proposes a Conditional Seasonal Markov-Switching Autoregressive model to simulate severe river flow values and to predict potential flooding by using deep machine learning Recurrent NN method. The pseudo-algorithm in Appendix. 1 provides a detailed illustration of the simulation methodology. An overview of this methodology is presented in Fig. 1 to visualize the sequence of processes involved in this study and illustrated as follows.

The time series analysis starts with the historical monthly river flow values of the Thames River as input. A structural break test is performed to ascertain the suitability of the Markov-Switching Autoregressive process for simulating the dataset. If the test indicates a valid fit, the proposed framework proceeds; otherwise, alternative models capable of simulating the long-term dependence of hydrological time series, such as the SARFIMA model can be considered.

Within the proposed framework, two regimes are determined to describe the transition between the latent regimes equivalent to wet and dry periods within the Markov-Switching process (Section 2.2.2), in which the evolution of the dynamics of each regime is controlled by the order of the autoregressive component (Section 2.2.1), which is estimated by using the Akaike Info Criterion (AIC) method. Subsequently, the simulation employs a conditional Hidden Seasonal Markov component (Section 2.2.3) to enhance the accuracy of both simulations and forecasts. This component consists of state and condition phases. The state phase defines the months of interest and provides their effect value. In parallel, the condition phase determines the years of seasonality and allocates the months of interest effect value to the years of seasonality.

The improvements of the proposed SMSAR model are demonstrated through a comprehensive analysis (Section 4.1). Subsequently, the SMSAR model's efficacy in simulating extreme river flow values is investigated (Section 4.2), followed by an evaluation of its impact on improving the accuracy of predictions (Section 4.3).

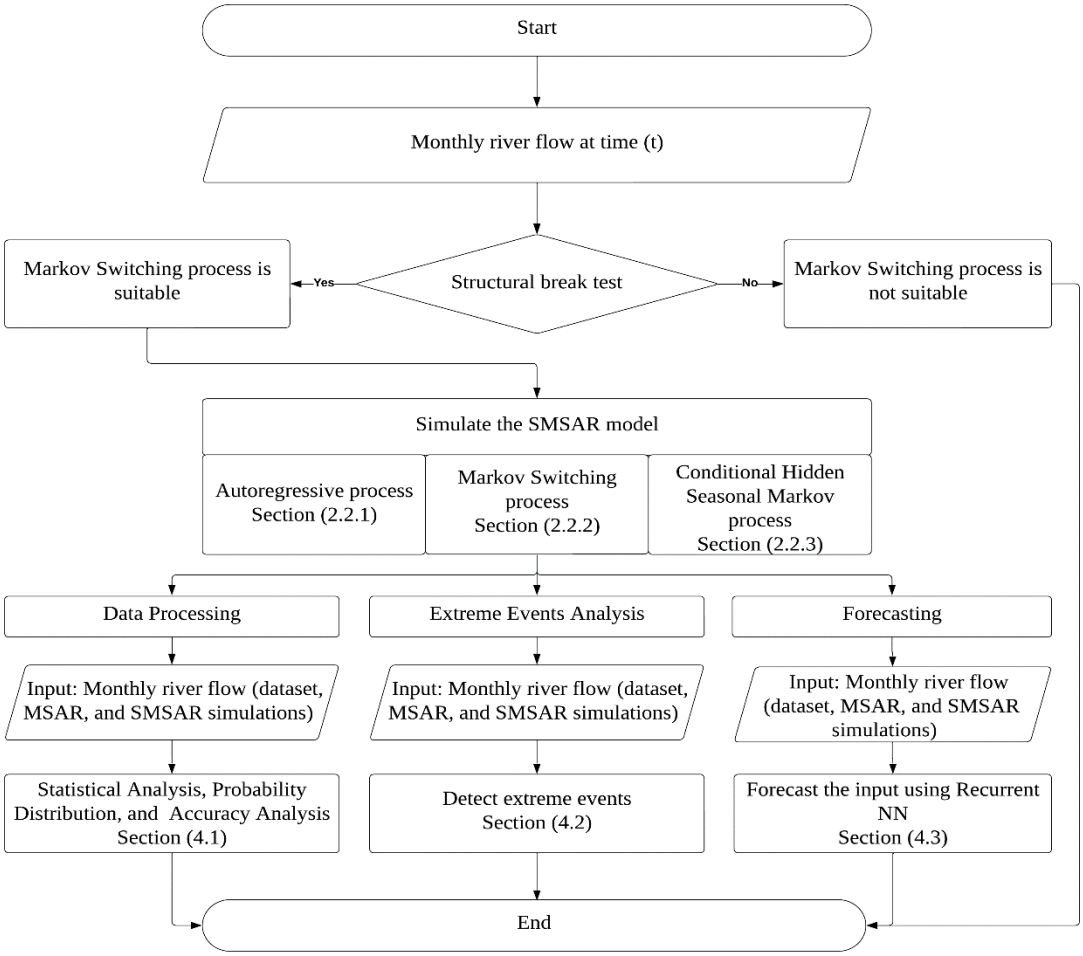

Fig. 1. Flow diagram of the methodology.

2.2. SMSAR Model

The proposed Seasonal Markov-Switching Autoregressive (SMSAR) model is a discrete-time process based on three components that will be detailed below, i.e., Hidden Seasonal Markov process (HSM), Markov-Switching process (MS), and Autoregressive process (AR). The analysis of the MSAR model components is carried out by the NHMSAR package in R Statistical Software with the addition of the proposed Hidden Seasonal Markov component [38].

The SMSAR model dynamics are based on the statistical information of the historical dataset, in which the dynamics of the model are the result of the autoregressive process. Furthermore, the dynamics variation is based on the Markov-Switching process which is significant in the simulation of consecutive events since each regime includes the characteristics of a specific time domain. In a sense, two regimes are used to present the behavior of high and low river flow values. In addition, the proposed Hidden Seasonal Markov component aims to enhance the ability of the MSAR model to simulate extreme values by computing the months of interest and the years of seasonality, in which the months of interest effect value will be applied at those time states, i.e., years of seasonality.

2.2.1. Autoregressive component

The Autoregressive process [39] is based on the assumption that the current values X_t of a time series are dependent on past values. This dependence offers profound insights into capturing the evolving

dynamics over time and is governed by parameters of the autoregressive process (Eq. (1)) to represent the decomposed components of the past values, i.e., level, trend, pattern, and random noise. The autoregressive order p indicates the number of previous observations taken into account to estimate the current value by quantifying the extent of their influence. In order to determine the optimal order, the model selection technique, i.e., Akaike's information criteria is employed, balancing model complexity and goodness of fit [40]. The autoregressive process of order p is symbolized as *AR(p)* at time t, writes:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \,;\; \varepsilon_t \sim\; WN(0, \sigma_\varepsilon^2\,) \tag{1}$$

where φ is a parameter of the model, c is the constant term of the process, ε_t is a random white noise WN which is a purely random process with a zero mean value and variance σ_ε^2.

2.2.2. Markov switching component

The Markov switching process [41] offers a robust framework for analyzing the evolving dynamics of a time series over time across a finite set of unobserved states that are characterized by different values of mean and variance by identifying different regimes in the data and estimating the transition probabilities between these regimes over time. This process is particularly valuable for modeling phenomena where the underlying dynamics are in transitions between different regimes [42], such as economic expansions and recessions, shifts in financial markets, or hydrological facts of rivers, i.e., high river flow values for a short time interval are followed by low river flow values for a long time interval. The transition probability state of the hidden Markov chain (Eq. (2)), writes:

$$P(\mathcal{S}_t = I \mid \mathcal{S}_{t-1} = J) = P(I|\,J); \mathcal{S}_t \,\epsilon\, \{1, \dots, M\} \tag{2}$$

where M is the number of regimes, $\mathcal{S}_t$ is the Hidden Markov chain state space, I and $J \,\epsilon\, \mathcal{S}_t$ and represent the Markovian regimes.

The MSAR model [43] is a discrete-time process with two components $\{X_t, \mathcal{S}_t\}$, where X_t (Eq. (1)) is the autoregressive process which indicates the evolving dynamics of the time series within each regime and $\mathcal{S}_t$ (Eq. (2)) is the Hidden Markov chain state space that incorporates the transition between the regimes. In this particular application, $\{X_t\}$ denotes the evolving dynamics of the maximum monthly river flow discharge and $\mathcal{S}_t \,\epsilon\, \{\text{Regime 1}, \text{Regime 2}\}$ represents the latent two regime states of low and high river flow values, respectively, as presented in Fig. 2.

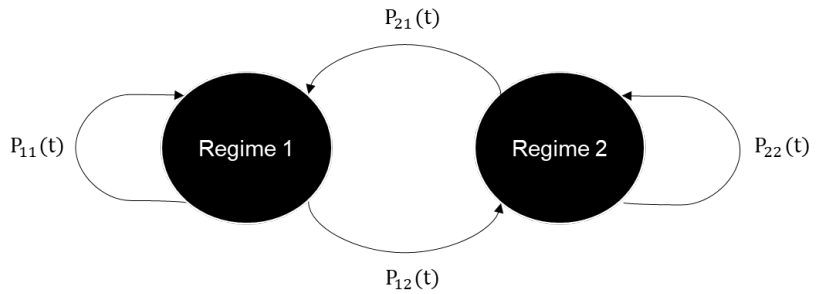

Fig. 2. Transition probability diagram.

The MSAR model general formula (Eq. (3)) explicates the stochastic relationship of the regimes switching Markovian chain (Eq. (2)) with the autoregressive component (Eq. (1)). In this application, the MSAR model (Eq. (4)) presents the dynamics of a time series with an autoregressive model of

order two, in which the switching between the two regimes of low and high river flow is controlled by a hidden Markov chain, writes:

$$Y_t|\, Y_{t-1} = \sum_{k=1}^{M} P(\mathcal{S}_t)\left(\sum_{i=1}^{p} \varphi_{i|k} Y_{t-i}\right) \tag{3}$$

$$(Y_t|\, Y_{t-1}|\, \mathcal{S}_t): \begin{cases} Y_t = c_1 + \varphi_1^1 X_{t-1} + \varphi_2^1 X_{t-2} + \varepsilon_t, & when\ \mathcal{S}_t = 1 \\ Y_t = c_2 + \varphi_1^2 X_{t-1} + \varphi_2^2 X_{t-2} + \varepsilon_t, & when\ \mathcal{S}_t = 2 \end{cases} \tag{4}$$

where Y_t is the monthly time series state, k is a Markovian regime state $\epsilon\, \mathcal{S}_t$, φ_i^M represents the autoregressive coefficient of order i for regime M, and c_M is the constant term of the process for regime M.

2.2.3. Conditional Hidden Seasonal Markov component

The conditional Hidden Seasonal Markov component is a bivariate stochastic discrete-time process that improves the modeling framework by incorporating supplementary information enabling a more comprehensive simulation of complex dependencies and underlying dynamics inherent in the time series data [44]. In this paper, this component accommodates influences that may not be fully captured within the states of the Markov-switching process (Section 2.2.2) by introducing an additional transitional probability matrix to account for anomalies to simulate high river flow values. The structure of this component consists of state and condition phases. The state phase detects the months of interest and provides the months of interest effect. Then, the condition phase defines the years of seasonality and allocates the months of interest effect to the years of seasonality. This component is demonstrated as follows.

The state phase considers the positive variance percentage ratio of the months (Eq. (5)) as the indicator of determining the months of interest. In this way, the actual results of the months of interest exceed the expected result based on the entire months, indicating a sign of growth. In addition, the state phase computes the effect value of the months of interest (Eq. (6)) based on the adjusted expected mean of the months of interest accordingly to the variance percentage increase value. This is carried out to compute and relate the mean value of the months of interest to the spread of the data between different monthly distributions. The state phase writes:

$$Moi_t = \left[\frac{\mu(\sum_{t_y=1}^{n} {\sigma^2}_{(t)})}{\mu\left(\sum_{t_y=1}^{n}(\sum_{t=1}^{12} {\sigma^2}_{(t)})\right)} \times 100\right] - 100 > 0;\ \ \left(t \,\middle|\, t_y\right) \epsilon \{1, \dots, 12\} \tag{5}$$

$$Ev_t = \mu\left(\sum_{t_y=1}^{n} t_{Moi}\right) \times \left[100 - \frac{(Moi_t|t_{Moi})}{100}\right];\ \ t_y \,\epsilon\, \{1, \dots, n\} \tag{6}$$

where Moi_t is the months of interest increase rate, t are the months of the year, σ^2 is the variance, t_y is the yearly time series, Ev_t is the months of interest effect value and t_{Moi} are the months of interest based on the months of interest increase rate.

The condition phase determines the years of seasonality and allocates the months of interest to the years of seasonality based on their variability. First, the condition phase determines the years of seasonality (Eq. (7)) based on the yearly variance by stating that, when the variance of the data in a single year exceeds the average, then that year is considered a year of seasonality. This is an indication that the year has experienced extreme values, and the monthly data distribution is more spread.

$$S_y = \sigma^2(t_y) > \mu \sum_{t_y=1}^{n} \sigma^2{}_{(t_y)} \tag{7}$$

where S_y are the years of seasonality.

The condition phase then distributes the months of interest across the years of seasonality based on the mean value of the yearly variability of the dataset as follows: The overall variability for each year of seasonality (Eq. (8)) is utilized to measure the degree to which each year of seasonality differs from the average of all years which is used to identify changes in seasonal patterns over time, then by relating the overall variability for a given year of seasonality to the average value for the months of interest over the years (Eq. (9)), it is possible to allocate the overall variability for a given year of seasonality to the months of interest, allowing for potential seasonal trends to be implemented.

$$S_{yR} = \sigma^2(t_y \mid S_y) - \mu \sum_{t_y=1}^{n} \sigma^2{}_{(t_y)} \tag{8}$$

$$MES_y = min\left\{S_{yR} \;\middle|\; \mu\left(\sum_{t_y=1}^{n} \sigma^2{}_{(t_{Moi})}\right)\right\} \tag{9}$$

where S_{yR} is the overall variability for each year of seasonality, and MES_y is the monthly effect based on the years of seasonality.

The conditional Hidden Seasonal Markov component consists of state and condition phases, this component allocates the months of interest effect value from the state phase (Eq. (6)) to the time series based on the monthly effect of the years of seasonality from the condition phase (Eq. (9)). This component is a process that allows the SMSAR model to accurately capture the seasonal yearly fluctuations in the time series and allocate the effect value accordingly, it writes:

$$HSM = (Ev_t \mid MES_y) \tag{10}$$

where HSM is the conditional Hidden Seasonal Markov component value.

The SMSAR model (Eq. (11)) is an improvement of the MSAR model (Eq. (3)) by integrating a conditional Hidden Seasonal Markov component (Eq. (10)). Subsequently, the simulation is enhanced by considering an additional value based on the effect values of the months of interest (Eq. (6)) and allocates this value to the time series concerning the years of seasonality (Eq. (9)) by integrating a seasonal transition matrix (Eq. (12)), allowing for potential seasonal trends to be implemented. This method is proposed to simulate severe values, i.e., high river flow values exceeding the outliers (Eq. (14)), writes:

$$Y_t \mid Y_{t-1} = \sum_{k=1}^{M} P(\mathcal{S}_t)\left(\sum_{i=1}^{p} \varphi_{i|k} Y_{t-i}\right) + \left(HSM \mid (\mathcal{S}_m)\right) \tag{11}$$

$$\mathcal{S}_m = \begin{bmatrix} t_{Moi}{}^{\langle \Psi,\Omega \rangle} & \cdots & t_{Moi}{}^{\langle \Psi,\Omega \rangle} \\ \vdots & \ddots & \vdots \\ t_{Moi}{}^{\langle \Psi,\Omega \rangle} & \cdots & t_{Moi}{}^{\langle \Psi,\Omega \rangle} \end{bmatrix}; \quad \left(t_{Moi}{}^{\langle \Psi,\Omega \rangle} \mid t\right) \epsilon \{1, \dots, 12\} \tag{12}$$

where $\mathcal{S}_m$ is the seasonal transition probability matrix that is estimated as the probabilistic transition between the months of interest effect value of the HSM states, Ψ represents the current state, and Ω

represents the transition state within the months of interest. In this particular application, the seasonal transition probability matrix (Eq.(13)), writes:

$$\mathcal{S}_m = \begin{bmatrix} {t_{Moi}}^{\langle 1,1\rangle} & {t_{Moi}}^{\langle 1,2\rangle} & {t_{Moi}}^{\langle 1,12\rangle} \\ {t_{Moi}}^{\langle 2,1\rangle} & {t_{Moi}}^{\langle 2,2\rangle} & {t_{Moi}}^{\langle 2,12\rangle} \\ {t_{Moi}}^{\langle 12,1\rangle} & {t_{Moi}}^{\langle 12,2\rangle} & {t_{Moi}}^{\langle 12,12\rangle} \end{bmatrix} \tag{13}$$

2.3. Severe values

The severe values detection method is proposed to detect the outliers of the monthly distributed dataset over their historical period. The method is based on the interquartile range of the monthly dataset [45] and considers the months of interest to discriminate the limit of the severe values, it writes:

$$E = \frac{\sum_i^{t_{Moi}} \left[Q_i + 1.5[Q_i - Z_i] \right]}{n_{t_{Moi}}} \tag{14}$$

where E is the limit to detect severe values, Q_i is the upper quartile of the dataset, Z_i is the lower quartile of the dataset, and $n_{t_{Moi}}$ is the number of months of interest counted from the t_{Moi} in Eq. (6).

3. Dataset Description

Thames River is considered to illustrate the proposed model within the framework of extreme events simulation. The Thames is considered a vital testbed since it is the longest river in England and has one of the longest flow records associated with any United Kingdom flow gauging station from 1883. Several studies on flood risk management indicated that the Thames is a flood-prone river and is considered a vital testbed [46–49]. The historical dataset comprises river flow measurements from the Thames local station number 3400TH, referenced as 39001 in Kingston. This dataset is extracted from the United Kingdom national river flow archive and spans from 1883 to 2020 [50].

In Fig. 3, the historical monthly dataset and yearly variance of the Thames River demonstrate the dynamic nature and fluctuations of the river's flow values. The fluctuations within the monthly river flow values and yearly variance values ensure that the Thames is a vital testbed for extreme events research and analysis.

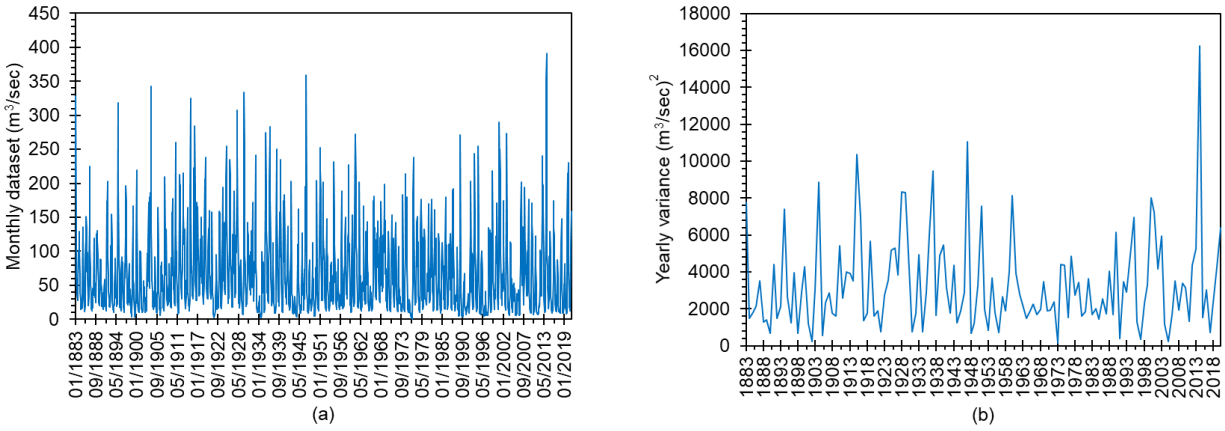

Fig. 3. Dataset description. (a) Monthly dataset, (b) Yearly variance.

4. RESULTS AND DISCUSSION

This section estimates the accuracy of the proposed SMSAR model, its ability to simulate extreme events, and forecast river flow. All results will be compared with those provided by the MSAR model.

4.1. Data Processing

This section aims to assess the quality of the proposed SMSAR model and its ability to detect severe values, i.e., maximum river flow values. The analysis sections are as follows: Statistical analysis, Probability distribution, Accuracy analysis, and Extreme events analysis.

4.1.1. Statistical analysis

This section describes the relation between the simulations and the dataset. In addition, we analyze the relative frequency and the intensity of upcrossings of predefined intervals of the models, this analysis displays the number of times a simulation passes upwards or downwards through an interval. Since it is important to correctly interpret the statistical analysis, the results of the simulations are then compared with those of the dataset to discuss their variations and assess their quality.

In Fig. 4, the outputs of the simulations (SMSAR and MSAR) and the dataset values are presented. In addition, deterministic values are presented to illustrate the proportion of variance, i.e., the coefficient of determination (R^2) and the strength of the linear association, i.e., Pearson correlation coefficient (R). It is observed that the SMSAR model provides a better fit to the dataset. This is evidenced by the improved results of the deterministic values of the coefficient of determination (R^2) and the Pearson correlation coefficient (R). In a sense, the SMSAR model less deviates from the dataset and presents a better linear relation with the dataset.

In Fig. 5, the residuals are fitted to their simulations distribution, this aims to introduce the variations within the range limit values of the simulations and compare them to the range limit of the dataset, i.e., maximum value. The range limit value of the dataset is 391 m^3/sec, this value is surpassed in one state by the SMSAR model with a value of 426 m^3/sec. On the other hand, the MSAR model has a lower range limit in comparison to the dataset with a range limit value of 254 m^3/sec.

The results demonstrate that, first, the river flow variations are better presented by the SMSAR model. Second, low and high river flow values are well presented by the embedded 2 regime-switching within the MSAR and SMSAR models which is explained by the positive Pearson correlation coefficient value. It is interesting to note that, the SMSAR model improves this value.

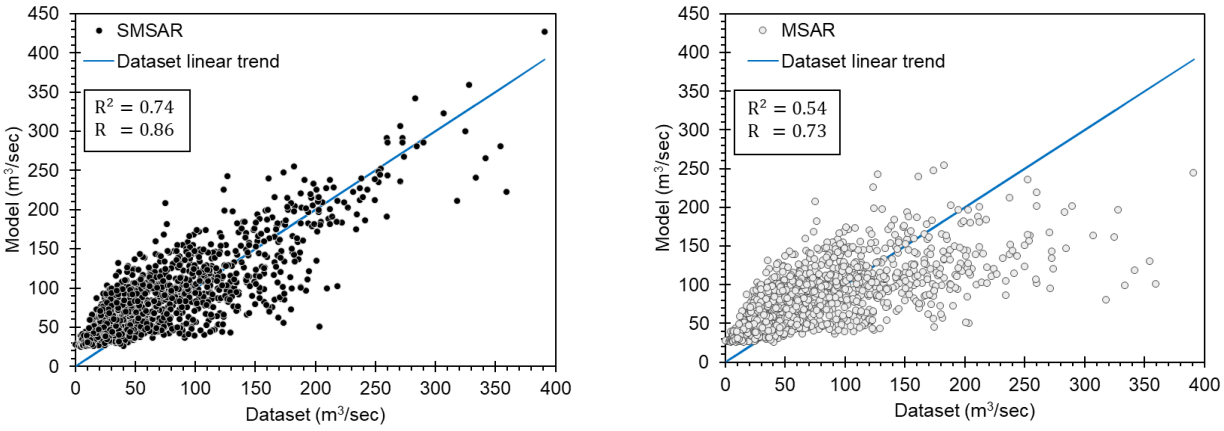

Fig. 4. Comparison of the river flow dataset with the SMSAR and MSAR simulations.

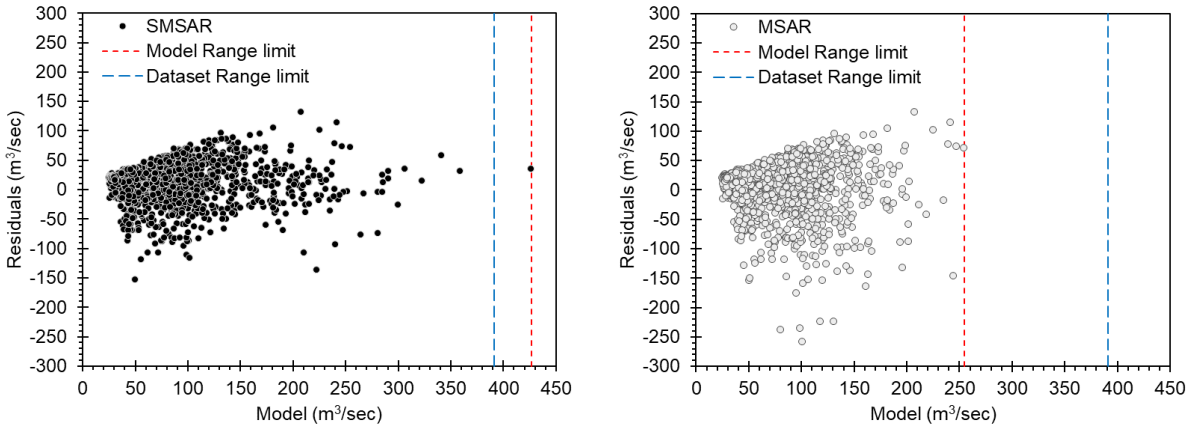

Fig. 5. Analysis of the residuals for each model.

In line with the statistical analysis, the relative frequency, and the intensity of upcrossings illustrate the occurrence of specific events, i.e., the percentage of occurrence, and the shifting between time series levels, respectively. It is important to note that the following findings are based on an interval sequential distribution of 10 m^3/sec. In a sense, the simulations are based on a sequential distribution of the river flow starting from 0 to the maximum river flow value with a range value of 10 m^3/sec for each increment. The interval sequential value is determined to compromise between having intervals with small but also large enough values that most of the observations neither fall in a single range interval nor end up with empty range intervals.

In Fig. 6, the observed relative frequency distribution emphasizes that when exceeding 170 m^3/sec, better results are achieved by the SMSAR model. Moreover, the SMSAR model seems to improve the findings when exceeding 250 m^3/sec since the MSAR model did not present values above this limit as shown in Fig. 5. However, both models show larger differences when they are below 70 m^3/sec.

The results demonstrate two statements. First, two regimes-switching within the MSAR model were not efficient to explain extremely low and high values. Second, this analysis found evidence that the SMSAR model suffers from the same limitations associated with the MSAR model for low values. However, the SMSAR model seems to improve findings for high values since the SMSAR model improves the pre-initiated MSAR simulation by embedding the conditional Hidden Seasonal Markov component (Eq. (10)) that allocates the months of interest effect value to the time series based on the monthly effect of the years of seasonality.

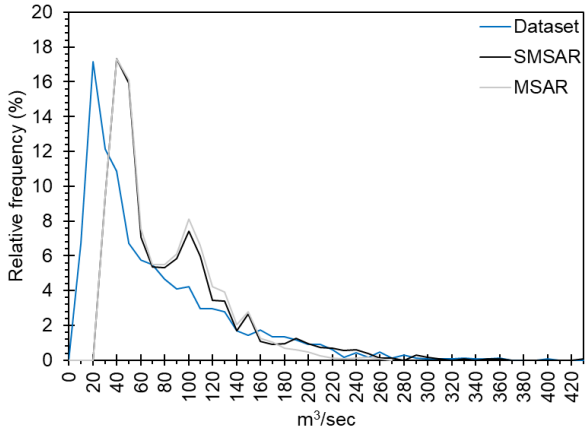

Fig. 6. Statistical distribution study.

In Fig. 7, the statistical intensity study indicates that both models underestimate the shifting behavior when below 40 m^3/sec. However, better results are seen for the SMSAR model above 150 m^3/sec. The results of the study support the notion that the Hidden Seasonal Markov component within the SMSAR model reallocates the statistical properties of the pre-initiated MSAR simulation.

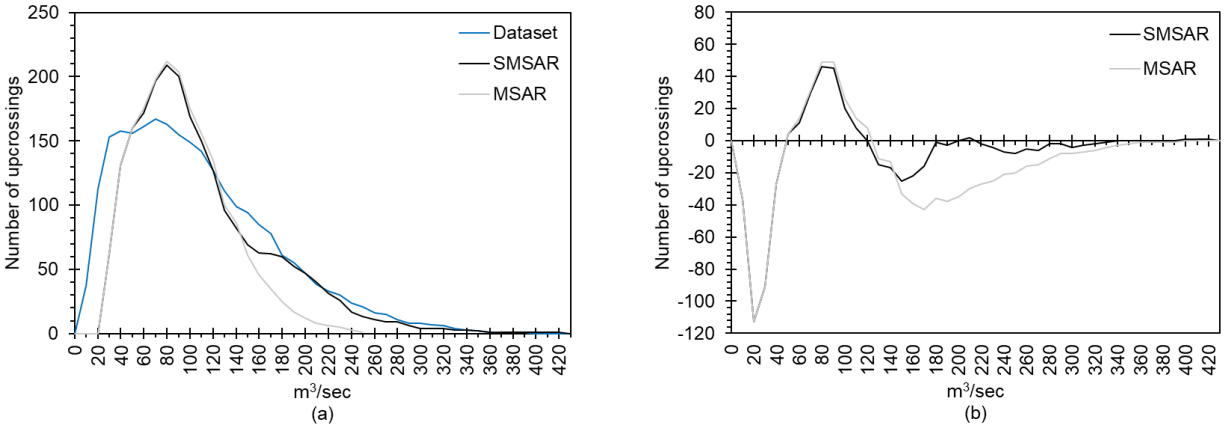

Fig. 7. Statistical intensity study. (a) The number of upcrossings, (b) Difference in the number of upcrossings.

4.1.2. Probability distribution

This section presents the probability distribution and parametric classifications of the distribution, i.e., spread and shape. The results of the simulations are compared with the values obtained from the dataset to discuss their variations and assess their quality in terms of a probability distribution.

In Fig. 8, the non-parametric kernel probability distribution is used to represent the statistical classification, i.e., asymmetry and tail distribution of the river flow. The results indicate that the dataset, SMSAR, and MSAR models are classified by shape distribution as asymmetrical right-skewed distributions with values of 1.59, 1.68, and 1.06, respectively. In a sense, the dataset and the SMSAR model values of the distributions are concentrated around low values and the right tail presents a low percentage of occurrence for higher values. However, the MSAR model distribution is classified by shape as a slightly asymmetrical right-skewed distribution with no outliers or extreme values since the skew value is nearly equal to 1.

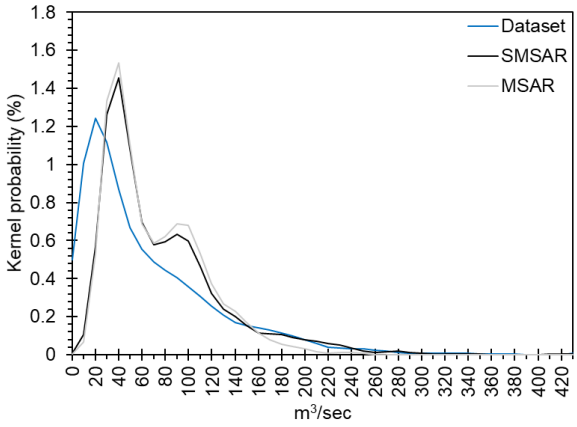

Fig. 8. Probabilistic analysis.

A graphical analysis (Q-Q plot) is carried out between the probability distributions by plotting their quantiles against the theoretical quantile, i.e., normal distribution. The applicability of this analysis is to test whether the distributions are normally distributed.

In Fig. 9, it is noted that the samples' quantiles do not align with the theoretical quantiles. This indicates that the distributions are not normally distributed. In addition, the shape of distributions is asymmetrical right-skewed since the upper end of the Q-Q plot deviates from the linear trend. The lower end of the Q-Q plot is horizontal, which appears to be the case of repetitive values and presents a high concentration around lower values within the distributions. The results also indicate that the SMSAR model improves the representation of the distribution with respect to the dataset in comparison to the MSAR model.

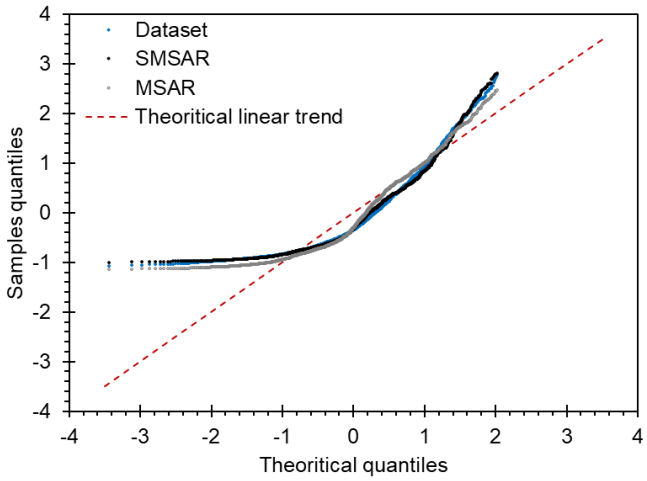

Fig. 9. Q-Q plot.

4.1.3. Accuracy analysis

This section evaluates the ability of the models to fit the dataset by comparing the results of their residuals and error indicators. In addition, this section demonstrates whether there is missing statistical information that should be accounted for in the models since the Markov-Switching regime's approach suffers from the limitation that volatile changes in the time series neglect the time series assumption of independent errors.

In Fig. 10, the MSAR model residuals are widely distributed, reaching a value of -258 m^3/sec. On the other hand, the SMSAR model residuals are less scattered, with a variance that is 57.11 % of the MSAR model's.

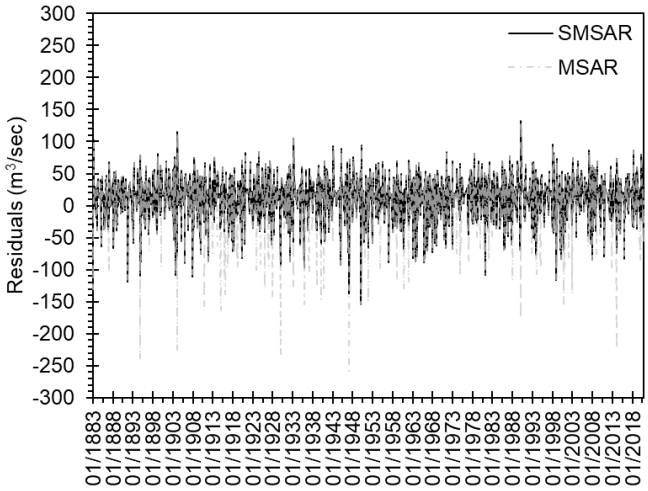

Fig. 10. Models' residuals distribution.

The distribution of the residuals is then analyzed by presenting their relative frequency to demonstrate to which degree the residuals are distributed. The relative frequency is based on an interval sequential distribution of 10 m^3/sec. In a sense, the simulations are based on a sequential distribution of the river flow starting from 0 to the maximum river flow value with a range value of 10 m^3/sec for each increment.

In Fig. 11, the distributions of the residuals indicate that slightly higher frequencies are achieved with the SMSAR model in comparison to the MSAR model within the distribution range between 0 m^3/sec to 40 m^3/sec. Moreover, lower frequencies for the SMSAR model are observed within the distribution below -50 m^3/sec.

This is evidence for the underestimation considered within the MSAR model due to the model's two regimes dynamic. However, this is improved by considering the Hidden Seasonal Markov component within the SMSAR model. Together, the present results of the models' residuals distribution and histogram confirm that the SMSAR model seems to provide a better fit to the dataset because of the fewer deviations.

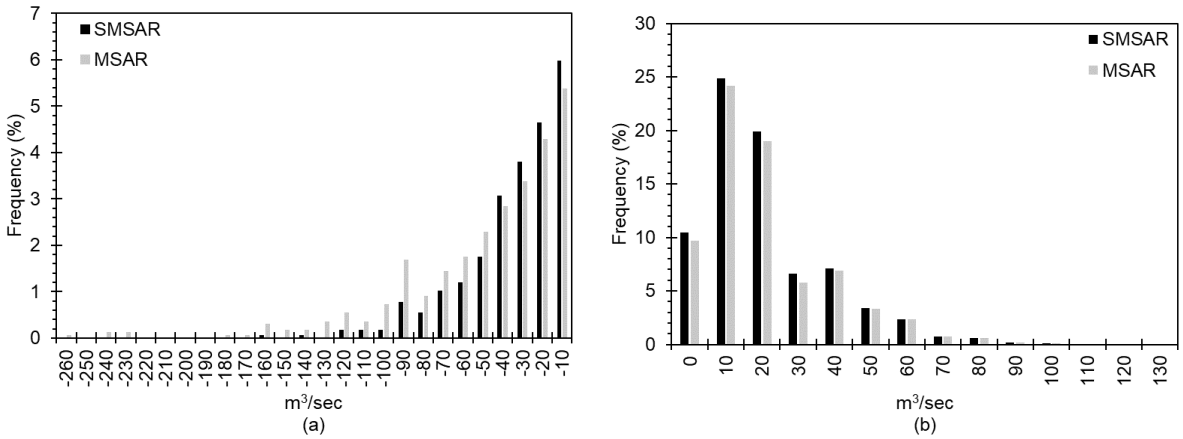

Fig. 11. Models' residuals histogram. (a) Negative residuals histogram, (b) Positive residuals histogram.

It is important to accurately interpret the residuals to assess the accuracy of the models compared to the dataset, therefore, in Table 1, the residuals are presented in terms of descriptive values to provide deterministic meaning. The error indicators for the SMSAR model present an increase in precision.

Table 1. Error indicators.

Model	RMSE*	MAE**	MAPE***	SME****
SMSAR	33.28	26.65	89.99	0.75
MSAR	41.30	30.73	91.84	1.00

* Square root of the average square errors. (m^3/sec)
** Mean absolute error. (m^3/sec)
*** Mean absolute percentage error. (%)
**** Standard mean error. (m^3/sec)

The autocorrelation of the residuals is computed to indicate whether there is still missing statistical information that should be accounted for within the SMSAR model. This aspect is important since the model includes a Markov-Switching process which may arise autocorrelation in the residuals, which in turn violates the assumption of independent errors. In a sense, exceeding the confidence interval indicates missing statistical information that should be considered in the model.

In Fig. 12, it is noted that the autocorrelation of the SMSAR model residuals is lower than that obtained from the MSAR model. However, there is some missing statistical information not considered by the SMSAR model for non-seasonal lags. Indeed, non-seasonal lags are the lags that exclude the seasonal monthly lags, i.e., 12, and 24.

The results demonstrate that the proposed conditional Hidden Seasonal Markov component within the SMSAR model significantly improves the results for seasonal lags, which implies that an improvement in the seasonal patterns is associated with the SMSAR model. Also, the SMSAR model still slightly suffers from the same limitations associated with the MSAR model analysis for the non-seasonal lags, i.e., errors are not completely random, but instead are following a pattern. Despite that, the SMSAR model also improves the results for the non-seasonal lags.

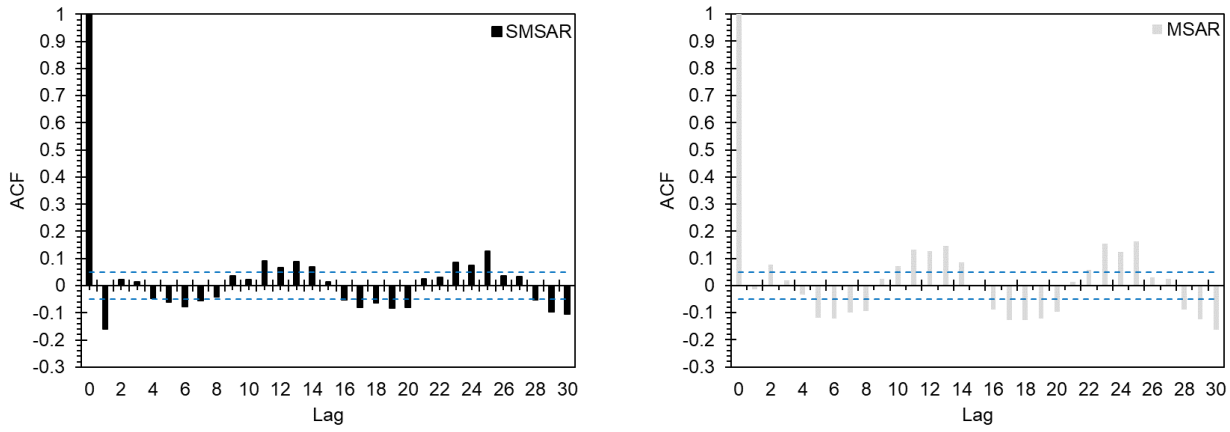

Fig. 12. Autocorrelation function (ACF) of the residuals.

4.2. Extreme events analysis

One major limitation of recent studies in time series analysis is the ability to simulate extreme values. In this section, the ability of the SMSAR and MSAR models to simulate extreme events is tested.

The proposed extreme events limit is considered based on the interquartile range of the pre-determined months of interest of the monthly distributed dataset; this method is mentioned in Eq. (14) (See section 2.3). Subsequently, the river flow value of 260 m^3/sec is considered to be the limit of extreme events.

In Fig. 13, the dataset and the SMSAR model present 19 and 15 extreme events, respectively. On the other hand, the MSAR model failed to simulate any extreme event. It is important to mention that

despite the impressive performance achieved by the SMASR model within the framework of extreme events detection, the SMSAR model underestimated five extreme events on 11/1894, 12/1910, 12/1929, 03/1947, and 02/1990, and overestimated one extreme event on 02/1915. The results ensure that the SMSAR model is capable of simulating extreme events since the SMSAR model detected 15 extreme events out of 19 extreme events in a monthly time series of length 1656 starting from 1883 to 2022.

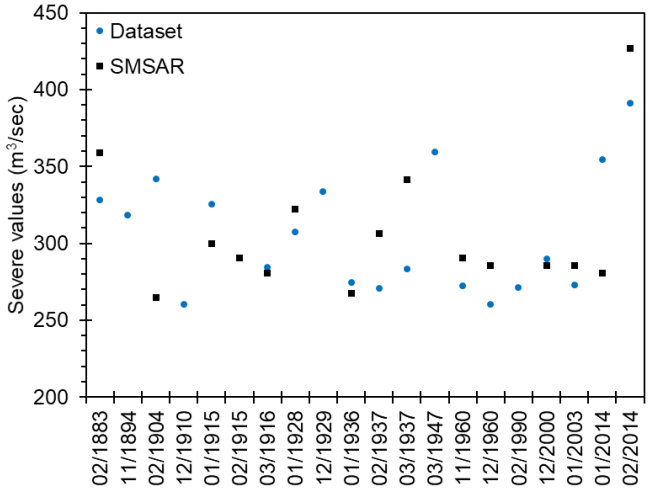

Fig. 13. Extreme events detection.

In Fig. 14, the difference in extreme events values between the SMSAR simulation and the dataset shows that the deviations of the SMSAR model are concentrated within the range of -10 m^3/sec to 20 m^3/sec with a mean absolute error value of 30 m^3/sec.

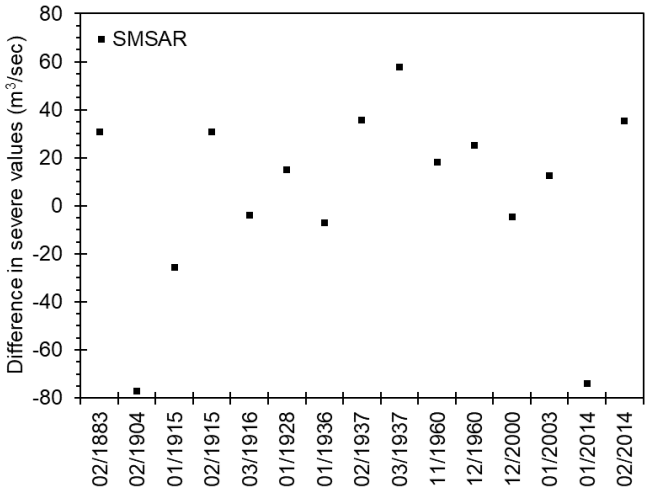

Fig. 14. Difference in severe values between the simulation and the dataset.

Together, the present results of the extreme events analysis go beyond previous reports, showing that the Markov-Switching approach is capable of simulating extreme events. This is achieved by embedding a conditional Hidden Seasonal Markov component within the SMSAR model. However, it is important to highlight the fact that the conditional component has an additional effect on the simulation. Therefore, overestimated states from the MSAR model won't be improved by the SMSAR model. For example, the overestimated extreme event on 02/1915. Overall, improved results are seen for the SMASR model within the framework of extreme events detection.

4.3. Models Forecasting

To date, it remains unclear whether predicting directly the dataset or a simulation fitted to the dataset yields better results. In addition, the impact of the training period on pattern recognition has not been clearly investigated, which raises concerns about the precision of the forecasts. Therefore, to test whether the previous statement is verifiable, the deep machine learning, i.e., Recurrent NN method [51] is considered for forecasting the dataset, the MSAR, and the SMSAR models. This method comprises the effect of monthly seasonality and the historical period length effect of 20 and 40 years on the period from 01/1931 to 12/1933. This period is chosen to include the highest influence of the Conditional Hidden Seasonal Markov component within the proposed SMSAR model of the training periods (earlier than 01/1931). Subsequently, by comparing the forecasts from the Neural Networks method of the dataset (Dataset NN), the SMSAR model (SMSAR NN), and the MSAR model (MSAR NN) with the available dataset of the mentioned period (Dataset), it is likely to investigate whether predicting the dataset or the alternatively based simulations provide better results.

In Fig. 15a, the observed Neural Networks forecasts are based on a 20-year Historical Period Length, the results are compared with the realistic dataset of this period. Overall, the result indicates that the Dataset NN, the MSAR NN, and the SMSAR NN forecasts can recognize the dataset patterns.

In Fig. 15b, the observed Neural Networks forecasts are based on a 40-year Historical Period Length. Overall, the result indicates that the Dataset NN, the SMSAR NN, and the MSAR NN forecasts can recognize the dataset patterns. In addition, forecasts based on a 40-year Historical Period Length present a better fit than the forecasts based on a 20-year Historical Period Length.

The results of the Neural Networks forecasts demonstrate that extending the training period within the Neural Networks forecasting method enhances its capacity for accurately recognizing historical patterns.

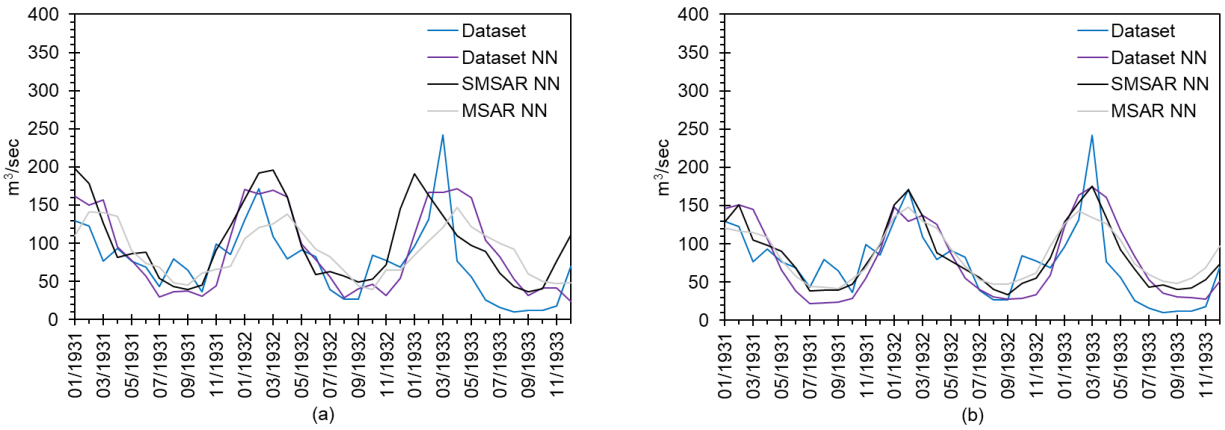

Fig. 15. Neural Network simulation forecasts. (a) 20-year and, (b) 40-year Historical Period Length.

The precision of the forecasts is presented in terms of descriptive values to determine the degree of accuracy and error. This provides an insight into which degree of inference whether predicting directly the dataset or alternatively a fitted simulation to the dataset is attributed to providing better results. Furthermore, it highlights the impact of the training period on pattern recognition.

In Table 2, the error indicators provide the accuracy of the Neural Networks simulations with respect to the dataset on the forecasted period from 01/1931 to 12/1933. The results of the simulations of the

40-year Historical Period Length yield better results than those of the 20-year Historical Period Length in terms of the degree of accuracy and error.

Table 2. Neural Network simulations error indicators.

	20-year Historical Period Length			40-year Historical Period Length		
Models	MAE*	RMSE**	R^2***	MAE*	RMSE**	R^2***
Dataset NN	35.6	45.4	0.42	30.0	36.8	0.52
SMSAR NN	38.2	46.1	0.48	22.5	27.2	0.71
MSAR NN	37.2	45.6	0.20	27.1	33.7	0.55

* Mean absolute error. (m^3/sec)
** Square root of the average square errors. (m^3/sec)
*** Coefficient of determination.

The results of the forecasting using the NN method indicate three key findings:

- The disproportion between the results of the fitted simulations, i.e., SMSAR NN and MSAR NN supports the notion that simulating a time series model requires access to experts who understand the relevant cause and effect of the dataset distribution, in which experts must be able to provide a recognizable pattern relevant to their priorities within the dataset. In this case, the proposed SMSAR model recognizes well the pattern and presents a significantly better fit than the MSAR model.

- Predicting a simulation fitted to the dataset is attributed to providing better results. In a sense, a simulation can integrate missing signals in the dataset and provide a recognizable pattern. This is indicated within the results of the SMSAR NN and MSAR NN compared to the Dataset NN.

- The training period impacts the accuracy of the forecasts. In this case, forecasts of a 40-year Historical Period Length yield a better fit than those of a 20-year Historical Period Length, since the length of the training process affects the ability of the Neural Networks to detect the patterns.

5. CONCLUSIONS AND PERSPECTIVES

This paper proposes a stochastic Seasonal Markov-Switching Autoregressive (SMSAR) model to recognize the dynamic pattern of the river flow and to simulate the hydrological fact of rivers, i.e., high river flow values for a short time interval are followed by low river flow values for a long-time interval. In addition, it aims at improving the capacity of simulating river flow extreme events. This paper also investigates the precision of the forecasts when considering the training period, i.e., Historical Period Length effect on the pattern recognition, and whether the predictive analysis of the dataset or equivalent simulations yields better results. The proposed framework provides valuable insights into simulating extreme events and offers a promising approach for enhancing our understanding of hydrological processes. The main conclusions are summarized as follows.

- Simulating extreme events of a hydrological time series poses a significant challenge for policymakers in achieving better management of water resources and assessing flood risks through a comprehensive understanding of the underlying dynamics of the hydrological dataset.

    - The Markov-Switching Autoregressive model offers a flexible framework for capturing regime shifts and changes in the time series dynamics, enabling the model to simulate the hydrological fact of rivers. However, simulating extreme values requires

considering the associated limitations of modeling extreme values by implementing additional methodological advancements to improve the efficacy.

- In this study, the proposed SMSAR model provides valuable insights into simulating the extreme values and offers a promising approach for enhancing our understanding of the hydrological processes by implementing an additional conditional Hidden Seasonal Markov component, allocating effect value of the months of interest to the time series based on the monthly effect of the years of seasonality.

- The Markov-Switching process holds promise as a valuable framework for simulating extreme events, particularly by effectively addressing the challenges posed by switching regimes. Consequently, it's essential to investigate the performance of the proposed SMSAR model in comparison to the traditional MSAR model to provide insights into the SMSAR model's effectiveness and its potential advantages.
  - An investigation of the statistical analysis is essential to interpret the differences between the proposed SMSAR and MSAR models to identify the advantages and limitations of each model and to simulate extreme events. This investigation concluded the following:
    - The SMSAR model deviates less from the dataset and presents a better linear relation with the dataset, providing a better fit to the dataset. This is evidenced by the improved results of the deterministic values of the coefficient of determination (R^2) and the Pearson correlation coefficient (R).
    - The MSAR model encountered challenges in accurately simulating both extremely low and high values. Subsequently, this SMSAR model suffers from the same limitations associated with the MSAR model for low values below 70 m^3/sec. However, the SMSAR model seems to enhance the simulation of high values exceeding 170 m^3/sec.
    - The Hidden Seasonal Markov component within the SMSAR model reallocates the statistical properties of the pre-initiated MSAR simulation, providing a better-shifting behavior above 150 m^3/sec.
  - An investigation of the probability distribution is essential to interpret the parametric classifications of the distribution, i.e., spread and shape, identifying the advantages and limitations of each model. This investigation concluded the following:
    - The SMSAR model and the dataset are classified by shape distribution as asymmetrical right-skewed distributions, characterized by concentration around low values, while the occurrence of higher values is minimal in the right tail. On the other hand, the MSAR model distribution is a slightly asymmetrical right-skewed distribution with no extreme values.
  - An investigation of the accuracy analysis is essential to evaluate the performance and to assess the uncertainties of the models. This investigation concluded the following:
    - The residuals of the SMSAR model residuals are less scattered compared to those of the MSAR model with a variance representing 57.11 % of that observed

in the MSAR model. In this context, the SMSAR model indicates a higher precision in the error indicators compared to the MSAR model.

- The SMSAR model significantly improves the missing statistical information within the MSAR model for seasonal lags. In addition, the SMSAR model also improves the results for the non-seasonal lags.

- An investigation of the extreme events analysis assesses the models' ability to simulate extreme events, which further can be integrated into a framework to provide valuable insights for decision-making. The findings of this investigation are summarized as follows:

  - The SMSAR model significantly improves the ability of the regime-switching technique to simulate extreme events. In this study, the SMSAR model succeeded in simulating 15 extreme events out of the 19 extreme events of the dataset. On the other hand, the regime-switching MSAR model failed to simulate any extreme event.

  - It is important to highlight the fact that the SMSAR model suffers from the fitting limitation associated with the MSAR model, in which the overestimation of the MSAR simulation cannot be adjusted if found.

- Forecasting river flow values is essential for informed decision-making in flood mitigation plans, as it addresses critical challenges to the resilience of the infrastructure systems against the adverse impacts of extreme hydrological events.

  - An investigation of River flow forecasting assesses whether predicting directly the dataset or a simulation fitted to the dataset yields better results. Additionally, it explores the impact of the training period on pattern recognition within the SMSAR model.

    - Predicting the fitted SMSAR simulation yields better results than predicting the dataset directly. In a sense, a simulation can integrate missing signals in the dataset and provide a recognizable pattern.

    - The training period significantly influences the accuracy of the forecasts as evidenced by the improvements adjusted by the Recurrent NN forecasts of the 40-year Historical Period Length compared to the 20-year Historical Period Length, with superior results observed for the SMSAR forecasts.

- Future research should be devoted to exploring several innovative advancements, as follows.

  - Developing the proposed SMSAR model since the model underestimates some extreme events and shows some missing statistical information within the simulation. In this context, the authors suggest the following:

    - Considering the yearly seasonal effect as a Hidden Seasonal Markov component within the model to increase the quality of the model.

    - Adjusting the effect value within the proposed model to influence all the months. This will improve the ability of simulating low river flow values.

    - The proposed methodology could be also adapted to simulate other extreme events such as floods, fires, extreme winds, etc. Specific databases and analysis

should be carried out to test the ability of the proposed methodology to deal with these extreme events.

- Integrating a framework to predict the impact of a changing climate on the river flow values. In this context, the authors suggest implementing a multivariate model to relate the climatic variables with the simulated river discharge from the reference period, followed by relating climatic variables under a changing climate to conduct future river discharge values.

6. ACKNOWLEDGEMENTS

This paper is carried out in the framework of the Strengthening the Territory's Resilience to Risks of Natural, Climate and Human Origin (SIRMA) project, which is financed by the European Regional Development Fund (ERDF) through INTERREG Atlantic Area Program with application code: EAPA_826/2018. This publication does not necessarily reflect the opinion of the European Union. Neither the INTERREG Europe program authorities nor the authors are responsible for any use that may be made of the information contained therein.

Credit author statement

Bassel Habeeb: conceptualization, methodology, software, formal analysis, data curation, writing – original draft, visualization. **Emilio Bastidas-Arteaga**: conceptualization, methodology, formal analysis, writing – review & editing, supervision, project administration, funding acquisition. **Mauricio Sánchez-Silva**: conceptualization, writing – review & editing, supervision. **You Dong**: conceptualization, writing – review & editing, supervision.

7. REFERENCES

[1] H. Bai, Y. Aoues, J.-M. Cherfils, D. Lemosse, Design of an Active Damping System for Vibration Control of Wind Turbine Towers, Infrastructures 6 (2021) 162. https://doi.org/10.3390/infrastructures6110162.

[2] E. Bastidas-Arteaga, C.-P.E. Soueidy, O. Amiri, P.T. Nguyen, Polynomial chaos expansion for lifetime assessment and sensitivity analysis of reinforced concrete structures subjected to chloride ingress and climate change, Struct. Concr. (2020) 1–12. https://doi.org/10.1002/suco.201900398.

[3] T. Eljufout, H. Toutanji, M. Al-Qaralleh, Fatigue Stress-Life Model of RC Beams Based on an Accelerated Fatigue Method, Infrastructures 4 (2019) 16. https://doi.org/10.3390/infrastructures4020016.

[4] M. Sánchez-Silva, G.-A. Klutke, Reliability and Life-Cycle Analysis of Deteriorating Systems, Springer International Publishing, Cham, 2016. https://doi.org/10.1007/978-3-319-20946-3.

[5] H. Guo, Y. Dong, E. Bastidas-Arteaga, X.-L. Gu, Probabilistic failure analysis, performance assessment, and sensitivity analysis of corroded reinforced concrete structures, Eng. Fail. Anal. 124 (2021) 105328. https://doi.org/10.1016/j.engfailanal.2021.105328.

[6] Q.C. Truong, C.-P. El Soueidy, Y. Li, E. Bastidas-Arteaga, Probability-based maintenance modeling and planning for reinforced concrete assets subjected to chloride ingress, J. Build. Eng. 54 (2022) 104675. https://doi.org/10.1016/j.jobe.2022.104675.

[7] E. Zacchei, E. Bastidas-Arteaga, Multifactorial Chloride Ingress Model for Reinforced Concrete Structures Subjected to Unsaturated Conditions, Buildings 12 (2022) 107. https://doi.org/10.3390/buildings12020107.

[8] B. Imam, Chapter Six - Climate Change Impact for Bridges Subjected to Scour and Corrosion, in: E. Bastidas-Arteaga, M.G. Stewar (Eds.), Clim. Adapt. Eng., Butterworth-Heinemann, 2019: pp. 165–206. https://doi.org/10.1016/B978-0-12-816782-3.00006-1.

[9] A. Malekjafarian, L.J. Prendergast, Experimental Demonstration of a Mode Shape-Based Scour-Monitoring Method for Multispan Bridges with Shallow Foundations, J Bridge Eng 25 (2020) 13. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001586.

[10] L.J. Prendergast, K. Gavin, A review of bridge scour monitoring techniques, J. Rock Mech. Geotech. Eng. 6 (2014) 138–149. https://doi.org/10.1016/j.jrmge.2014.01.007.

[11] P.F. Lagasse, Case studies and lessons learned from recent scour-related bridge failures in the United States, (1999). https://doi.org/10.5169/SEALS-60752.

[12] E. Bastidas-Arteaga, Towards Climate Change Adaptation of Existing and New Deteriorating Infrastructure, in: J.C. Matos, P.B. Lourenço, D.V. Oliveira, J. Branco, D. Proske, R.A. Silva, H.S. Sousa (Eds.), 18th Int. Probabilistic Workshop, Springer International Publishing, Cham, 2021: pp. 39–51. https://doi.org/10.1007/978-3-030-73616-3_3.

[13] B. Van Den Bout, V.G. Jetten, C.J. Van Westen, L. Lombardo, A breakthrough in fast flood simulation, Environ. Model. Softw. (2023) 105787. https://doi.org/10.1016/j.envsoft.2023.105787.

[14] S. Rasiya Koya, N.V. Giron, M. Rojas, R. Mantilla, K. Harvey, D. Ceynar, F. Quintero, W.F. Krajewski, T. Roy, Applicability of a flood forecasting system for Nebraska watersheds, Environ. Model. Softw. 164 (2023) 105693. https://doi.org/10.1016/j.envsoft.2023.105693.

[15] A.S. Leon, Y. Tang, L. Qin, D. Chen, A MATLAB framework for forecasting optimal flow releases in a multi-storage system for flood control, Environ. Model. Softw. 125 (2020) 104618. https://doi.org/10.1016/j.envsoft.2019.104618.

[16] C. Aguilar, M.J. Polo, Assessing minimum environmental flows in nonpermanent rivers: The choice of thresholds, Environ. Model. Softw. 79 (2016) 120–134. https://doi.org/10.1016/j.envsoft.2016.02.003.

[17] C.H.F. Toro, S. Gómez Meire, J.F. Gálvez, F. Fdez-Riverola, A hybrid artificial intelligence model for river flow forecasting, Appl. Soft Comput. 13 (2013) 3449–3458. https://doi.org/10.1016/j.asoc.2013.04.014.

[18] D.J. Pedregal, R. Rivas, V. Feliu, L. Sánchez, A. Linares, A non-linear forecasting system for the Ebro River at Zaragoza, Spain, Environ. Model. Softw. 24 (2009) 502–509. https://doi.org/10.1016/j.envsoft.2008.09.010.

[19] P. Elek, L. Márkus, A long range dependent model with nonlinear innovations for simulating daily river flows, Nat. Hazards Earth Syst. Sci. 4 (2004) 277–283. https://doi.org/10.5194/nhess-4-277-2004.

[20] H. Aksoy, Markov chain-based modeling techniques for stochastic generation of daily intermittent streamflows, Adv. Water Resour. 26 (2003) 663–671. https://doi.org/10.1016/S0309-1708(03)00031-9.

[21] J. Szilagyi, G. Balint, A. Csik, Hybrid, Markov Chain-Based Model for Daily Streamflow Generation at Multiple Catchment Sites, J. Hydrol. Eng. 11 (2006) 245–256. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:3(245).

[22] Z.-Q. Lu, L.M. Berliner, Markov switching time series models with application to a daily runoff series, Water Resour. Res. 35 (1999) 523–534. https://doi.org/10.1029/98WR02686.

[23] L.E. Baum, T. Petrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, Ann. Math. Stat. 37 (1966) 1554–1563. https://doi.org/10.1214/aoms/1177699147.

[24] J.D. Hamilton, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica 57 (1989) 357. https://doi.org/10.2307/1912559.

[25] V. Monbet, P. Ailliot, M. Prevosto, Survey of stochastic models for wind and sea state time series, Probabilistic Eng. Mech. 22 (2007) 113–126. https://doi.org/10.1016/j.probengmech.2006.08.003.

[26] C. Engel, C.S. Hakkio, The Distribution of Exchange Rates in the EMS, Int. J. Finance Econ. 1 (1996) 55–67. https://doi.org/10.1002/(SICI)1099-1158(199601)1:1<55::AID-IJFE5>3.0.CO;2-C.

[27] V. Monbet, P. Ailliot, M. Prevosto, Survey of stochastic models for wind and sea state time series, Probabilistic Eng. Mech. 22 (2007) 113–126. https://doi.org/10.1016/j.probengmech.2006.08.003.

[28] I. Cárdenas-Gallo, R. Akhavan-Tabatabaei, M. Sánchez-Silva, E. Bastidas-Arteaga, A Markov regime-switching framework to forecast El Niño Southern Oscillation patterns, Nat. Hazards 81 (2016) 829–843. https://doi.org/10.1007/s11069-015-2106-y.

[29] A.J. Lawrance, N.T. Kottegoda, Stochastic Modelling of Riverflow Time Series, J. R. Stat. Soc. Ser. Gen. 140 (1977) 1. https://doi.org/10.2307/2344516.

[30] M. Campolo, P. Andreussi, A. Soldati, River flood forecasting with a neural network model, Water Resour. Res. 35 (1999) 1191–1197. https://doi.org/10.1029/1998WR900086.

[31] J.L. Uc-Castillo, A.E. Marín-Celestino, D.A. Martínez-Cruz, J. Tuxpan-Vargas, J.A. Ramos-Leal, A systematic review and meta-analysis of groundwater level forecasting with machine learning techniques: Current status and future directions, Environ. Model. Softw. 168 (2023) 105788. https://doi.org/10.1016/j.envsoft.2023.105788.

[32] R.R. Trippi, E. Turban, eds., Neural networks in finance and investing: using artificial intelligence to improve real-world performance, Probus Pub. Co, Chicago, Ill, 1993.

[33] K.E. Adikari, S. Shrestha, D.T. Ratnayake, A. Budhathoki, S. Mohanasundaram, M.N. Dailey, Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions, Environ. Model. Softw. 144 (2021) 105136. https://doi.org/10.1016/j.envsoft.2021.105136.

[34] Z. Jiang, S. Yang, Z. Liu, Y. Xu, Y. Xiong, S. Qi, Q. Pang, J. Xu, F. Liu, T. Xu, Coupling machine learning and weather forecast to predict farmland flood disaster: A case study in Yangtze River basin, Environ. Model. Softw. 155 (2022) 105436. https://doi.org/10.1016/j.envsoft.2022.105436.

[35] M. Castangia, L.M.M. Grajales, A. Aliberti, C. Rossi, A. Macii, E. Macii, E. Patti, Transformer neural networks for interpretable flood forecasting, Environ. Model. Softw. 160 (2022) 105581. https://doi.org/10.1016/j.envsoft.2022.105581.

[36] M. Bildirici, Ö. Ersin, Modeling Markov Switching ARMA-GARCH Neural Networks Models and an Application to Forecasting Stock Returns, Sci. World J. 2014 (2014) 1–21. https://doi.org/10.1155/2014/497941.

[37] I.-F. Kao, Y. Zhou, L.-C. Chang, F.-J. Chang, Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting, J. Hydrol. 583 (2020) 124631. https://doi.org/10.1016/j.jhydrol.2020.124631.

[38] M. Valerie, Non-Homogeneous Markov Switching Autoregressive Models, (2022). https://cran.r-project.org/web/packages/NHMSAR/NHMSAR.pdf.

[39] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 1994.

[40] R. Shibata, Selection of the order of an autoregressive model by Akaike's information criterion, Biometrika 63 (1976) 117–126. https://doi.org/10.1093/biomet/63.1.117.

[41] J.D. Hamilton, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica 57 (1989) 357. https://doi.org/10.2307/1912559.

[42] S.N. Durlauf, L. Blume, eds., Macroeconometrics and time series analysis, Palgrave Macmillan, Basingstoke, 2010.

[43] C.-J. Kim, Dynamic linear models with Markov-switching, J. Econom. 60 (1994) 1–22. https://doi.org/10.1016/0304-4076(94)90036-1.

[44] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inf. Theory 48 (2002) 1518–1569. https://doi.org/10.1109/TIT.2002.1003838.

[45] S. Arciniega-Esparza, A. Hernández-Espriú, J.A. Breña-Naranjo, M.H. Young, A. Pedrozo-Acuña, An outlier detection approach for water footprint assessments in shale formations: case Eagle Ford play (Texas), Environ. Earth Sci. 79 (2020) 454. https://doi.org/10.1007/s12665-020-09197-8.

[46] L. Alfieri, P. Salamon, A. Bianchi, J. Neal, P. Bates, L. Feyen, Advances in pan-European flood hazard mapping: ADVANCES IN PAN-EUROPEAN FLOOD HAZARD MAPPING, Hydrol. Process. 28 (2014) 4067–4077. https://doi.org/10.1002/hyp.9947.

[47] H.-I. Eum, D. Sredojevic, S.P. Simonovic, Engineering Procedure for the Climate Change Flood Risk Assessment in the Upper Thames River Basin, J. Hydrol. Eng. 16 (2011) 608–612. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000346.

[48] S. Lavery, B. Donovan, Flood risk management in the Thames Estuary looking ahead 100 years, Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 363 (2005) 1455–1474. https://doi.org/10.1098/rsta.2005.1579.

[49] V.A. Bell, A.L. Kay, S.J. Cole, R.G. Jones, R.J. Moore, N.S. Reynard, How might climate change affect river flows across the Thames Basin? An area-wide analysis using the UKCP09 Regional Climate Model ensemble, J. Hydrol. 442–443 (2012) 89–104. https://doi.org/10.1016/j.jhydrol.2012.04.001.

[50] National River Flow Archive, (n.d.). https://nrfa.ceh.ac.uk/data/search.

[51] R. Hyndman, Forecasting Functions for Time Series and Linear Models, (2022). https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Appendix. 1. Pseudo-algorithm of the simulation methodology

```
BEGIN
READ Dataset

# Define the monthly time series
INPUT (Dataset, Start = Date, End = Date, Frequency =12)
        OUTPUT Time series

# Stationary check:
COMPUTE KPSS test (Time series)
        OUTPUT P-value && Statistic value
        IF (P-value > p_Criteria && Statistic value < Intercept_Critical) THEN
                PRINT (Time series is stationary)
        ELSE
                PRINT (Time series is non-stationary)
                SET Differencing order to zero
                        CALCULATE Integration (Time series)
                        OUTPUT Differencing order && Integrated Stationary Time series
        END IF

# Choosing a model to fit the time series dataset:
COMPUTE Autocorrelation function (Time series)
        OUTPUT Decay
        IF (Decay = Slowly) THEN
                PRINT (Long-term models are required, i.e., MSAR or ARFIMA)
                COMPUTE Structural break test (Time series)
                        OUTPUT P-value
                        IF (P-value < 0.05) THEN
                                PRINT (MSAR model is mandatory)
                        ELSE
                                PRINT (ARFIMA model is a preferable choice)
                        END IF
        ELSE
                PRINT (Short-term models are required, i.e., ARIMA)
        END IF

# Simulate the MSAR model:
OBTAIN Array (Time series)
        OUTPUT Time series array
INIT Akaike Info Criterion to zero
INIT j to 2
FOR i = 1 to 4
        COMPUTE MSAR (Time series array, φ[i], M[j]) # Eq.(3)
                OUTPUT Fitted MSAR model
                COMPUTE Akaike info criterion
                STORE Fitted model && Akaike info criterion in Result [MSAR]
END FOR
READ the lowest value of Akaike info criterion in Result [MSAR][i]
        OUTPUT Fitted MSAR model
```

```
# Calculate the hidden seasonal Markov component stating phase:
COMPUTE Months of interest increase rate (Moi_t): # Eq. (5)
COMPUTE Months of interest effect value (Ev_t): # Eq. (6)

# Calculate the hidden seasonal Markov component conditioning phase:
COMPUTE Yearly seasonality (S_y): # Eq. (7)
COMPUTE Monthly effect value with respect to yearly seasonality (MES_y): # Eq. (9)

# Calculate the Hidden Seasonal Markov component value:
COMPUTE Conditional Hidden Seasonal Markov component value (HSM): # Eq. (10)

# Simulate the SMSAR model:
COMPUTE SMSAR (Fitted MSAR model, HSM): # Eq. (11)
OUTPUT SMSAR fitted model

# Extreme events:
CALCULATE E # Eq. (14): Severe values limit
SET n to length (Time series)
SET p to 1
WHILE p < n
        IF (Time series[p, 1], MSAR fitted model[p, 2] && SMSAR fitted model[p, 3]) >= E)
        THEN
                OUTPUT Extreme events values
                STORE Extreme events values in Result [Extreme events] [p, q]
        ELSE
                INCREMENT p
        END IF
END WHILE

# Neural Network forecasting
READ Time series, MSAR fitted model && SMSAR fitted model
STORE Time series, MSAR fitted model && SMSAR fitted model in Result [Simulations] [x, y]
SET y to 1
FOR x = 1 to n
        Neural network (Simulations [x, y], non-seasonal lags = 11, seasonal lags= (1,12))
        OUTPUT NN fitted simulations
        STORE NN fitted simulations in Result [Neural Network] [x, y]
        INCREMENT y
END FOR
GENERATE Forecast ([Neural Network] [1: n, 1: 3], periods of forecasting = 36)
            OUTPUT Forecasted values
            STORE Forecasted values in Result [Forecasts] [1: 36, 1: 3]
```

The simulation methodology in Appendix. 1 provides a detailed illustration of the methodology and consists of five main phases, i.e., Dataset preparation, model identification, model fitting, extreme events detection, and forecasting the dataset and the simulations (MSAR and SMSAR models) using Recurrent Neural Network system. The input of the time series analysis is the historical river flow values of the Thames River. These phases are demonstrated as follows:

In the first phase of this process, the time series dataset is presented with a monthly frequency and tested for non-stationary behavior. This is performed to determine whether the dataset's mean and variance change over time. However, the stationary behavior of the time series will not be accurately computed when the time series has long-term behavior. Subsequently, the long-term behavior is examined in the model identification phase, and the type of the model is specified. In this case of long-term behavior, a structural break test is performed to indicate whether a Markov-Switching Autoregressive model is suitable for simulating the dataset.

In the third phase, two regimes within the Markov-Switching process are presented to describe the latent regimes equivalent to wet and dry periods. In addition, the autoregressive order is estimated using the Akaike Info Criterion (AIC) method. Additionally, the simulation is approached with a conditional Hidden Seasonal Markov component, this component consists of state and condition phases. The state phase defines the months of interest and provides their effect value. In parallel, the condition phase determines the years of seasonality and allocates the months of interest effect value to the years of seasonality.

The fourth phase detects the extreme events based on the interquartile range of the monthly dataset and considers the months of interest to discriminate the limit of the severe values. Finally, the forecasting phase uses the NN system, i.e., Recurrent NN to provide forecasts. This forecasting method comprises the effect of monthly seasonality.