FITTING SARIMA MODELS TO RHÔNE TEMPERATURE DATA

NATHAN GAVILLET[1] & LUC KUNZ[2]

May 21, 2024

CONTENTS

ABSTRACT

This paper aims to predict the monthly average temperature of the Rhône at the Porte-du-Scex station using SARIMA (Seasonal Autoregressive Integrated Moving Average) models. Utilizing data spanning from 1968 to the present (i.e. April 2024), we explore different SARIMA configurations and employ model selection criteria such as AICc (Akaike Information Criterion corrected) to identify the most suitable model. Our methodology involves analyzing the trend and seasonality of the time series through linear regression and differencing techniques, followed by modeling the resulting stationary time series using an ARMA (Autoregressive Moving Average) approach. Subsequently, we conduct a thorough analysis of the residuals to ensure the validity of the selected model. Our findings reveal that SARIMA models are suitable for the data at hand, as well as a pronounced seasonal pattern, with summer months exhibiting an average temperature approximately 5°C warmer than winter months. Overall, this study provides valuable insights into the predictive modeling of river temperature dynamics and highlights the importance of rigorous model selection techniques in time series analysis.

Keywords: SARIMA models, temperature prediction, time series analysis, model selection, Rhône temperature.

[1] *Statistics Student, University of Neuchâtel, Switzerland. Email: nathan.gavillet@unine.ch*
[2] *Mathematics Student, University of Neuchâtel, Switzerland. Email: luc.kunz@unine.ch*

1 INTRODUCTION

The Rhône, a vital waterway coursing through Europe, serves as a crucial ecosystem and resource for numerous communities. Understanding its environmental dynamics, particularly the temperature variations over time, holds significance for various scientific, ecological, and societal endeavors ([Khalanski et al., 2009]). In this study, we undertake a comprehensive analysis of time series data supplied by the Federal Office for the Environment (FOEN), focusing on the temperature evolution of the Rhône.

The analysis of temperature variations in the Rhône holds paramount importance across various domains. For instance, temperature serves as a fundamental determinant influencing several interconnected variables, including water conductivity and oxygen concentration ([Hardenbicker et al., 2017], [Ducharne, 2008]). Our primary objectives encompass a multifaceted approach towards analyzing the time series data of the Rhône: we aim to employ advanced time series modeling techniques (i.e. ARMA models) to gain deeper insights into the temporal patterns and fluctuations exhibited by the river's temperature. We also seek to provide meaningful interpretations and forecasts of temperature movements, thereby elucidating the impacts of climatic changes and seasonal variations on the river's thermal regime. By traversing the entire spectrum of time series analysis – from descriptive statistics to predictive modeling – we aim to contribute to a comprehensive understanding of the Rhône's temperature dynamics.

We anticipate observing an increasing trend in the temperature of the Rhône over the years, reflecting the broader phenomenon of climate change. Additionally, we expect to identify clear seasonal patterns, with temperatures peaking during the summer months. Furthermore, our preliminary analyses suggest the presence of temporal dependencies within the data, indicating that past temperature values can serve as valuable predictors for future observations. The data is expected to exhibit a high degree of reliability and accuracy ([Denzler, 2019]). Temperature measurements, in particular, are considered relatively straightforward and reliable, given the advancements in sensor technology and data collection methodologies ([Bouffard et al., 2019]).

Section 2 describes the data and the context in which it has been gathered, section 3 exhibits the modeling methodology used to conduct the analysis. Section 4 is dedicated to the presentation and discussion of the results and section 5 concludes the paper.

2 DESCRIPTION

The data under study originates from the Porte-du-Scex monitoring station situated along the Rhône. It encompasses daily (resp. monthly) average temperature measurements spanning from January 1, 1968 to April 10, 2024, comprising a total of 20,555 data points (resp. 676). The dataset exhibits full integrity, with no missing values observed throughout its entirety. However, it is worth noting that recent data entries (from 2021 onwards) still necessitate validation from the FOEN.

Table 1 provides summary statistics of the data. It reveals for instance that daily temperature averages consistently remain above zero degrees throughout the observation period.

Table 1: Temperature summary statistics

	Temperature (°C)					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Daily average	0.740	5.110	7.700	7.296	9.300	13.300
Monthly average	2.587	5.183	7.687	7.285	9.331	10.941

Figure 1 displays monthly average temperature variations of the Rhône spanning the past 56 years.

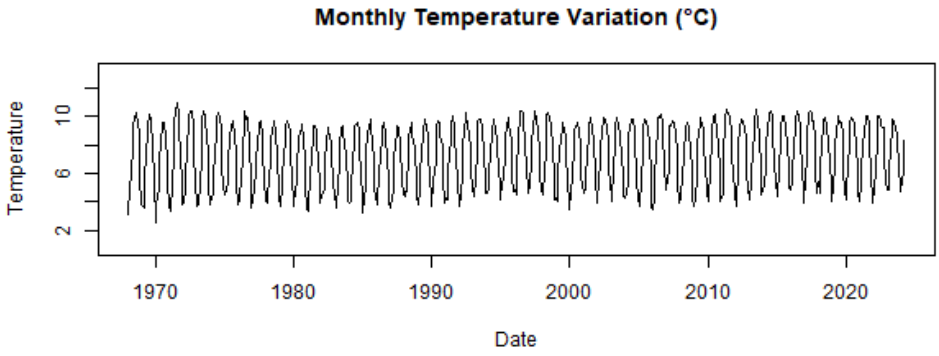

Figure 1: Monthly temperature variations of the Rhône over the last 56 years

A discernible pattern emerges, revealing a modest upward linear trend alongside distinct seasonal fluctuations. The summer months seem to consistently exhibit higher temperatures, aligning with expectations of warmer weather during this period. This observation is further corroborated by the analysis presented in Figure 2.

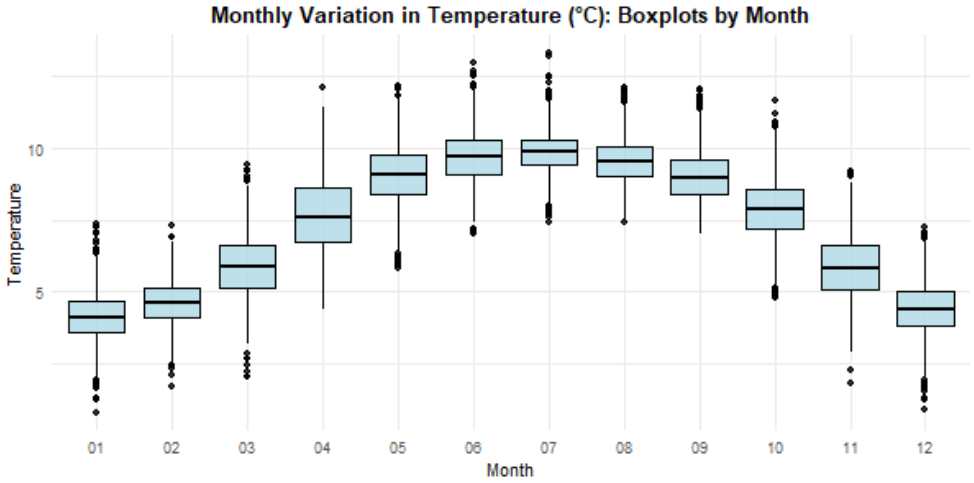

Figure 2: Temperature of the Rhône. Boxplots by month

Based on the boxplots shown above, certain observations appear to be potential outliers. Nonetheless, as depicted in Figure 1, these values fall within an acceptable range.

Regarding the trend, an insightful comparison of monthly averages for three different years is depicted in Figure 3. Notably, it showcases a noteworthy tendency towards higher temperatures in more recent years. This observation will be elabo-

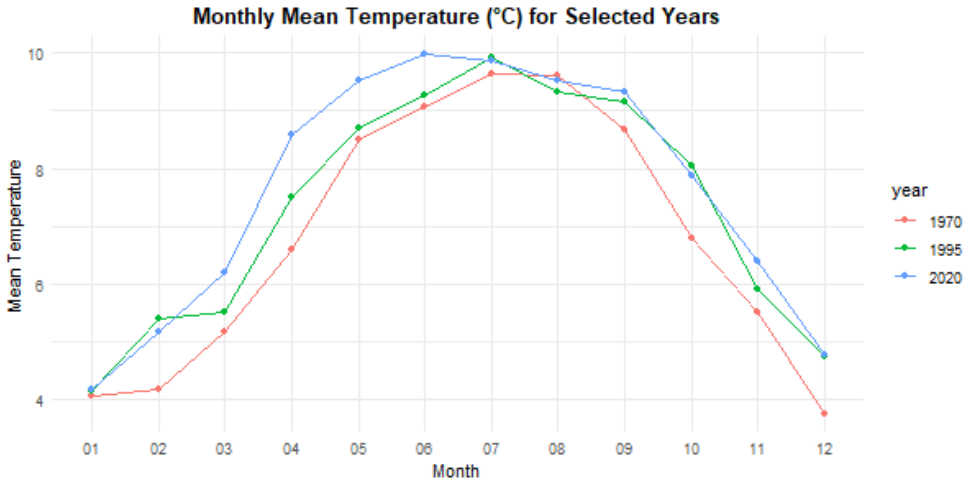

Figure 3: Monthly temperature average for three different years

rated upon in Sections 3 and 4, where a comprehensive analysis of the temperature trend and seasonality will be conducted using linear regression.

3 MODELING

This section is dedicated to the comprehensive description of the models utilized in the analysis. In the following, let $\{X_t\}$ denote the time series representing the average temperature of the Rhône measured on a monthly basis, and let n be the number of data points, i.e. 676.

3.1 Trend and seasonality

We first model the trend and seasonality using a linear regression model as follows:

$$X_t = m_t + s_t + Y_t, \quad t = 1, \dots, n \tag{1}$$

where:

- m_t is a polynomial trend of degree k, i.e.: $m_t = a_0 + a_1 t + ... + a_k t^k$,
- s_t is a centered periodic function, i.e.:
  - $s_{t+d} = s_t$, where d is a known period length (in our case: $d = 12$),
  - The sum over one period is null, i.e. $\sum_{j=1}^{d} s_{t+j} = 0$.
- Y_t is the error term.

We model the seasonal component using binary variables with January as the reference month and end up with the following linear model:

$$X_t = a_0 + a_1 t + ... + a_k t^k + b_2 \mathbb{1}_{\{\text{Feb}\}} + b_3 \mathbb{1}_{\{\text{Mar}\}} + ... + b_{12} \mathbb{1}_{\{\text{Dec}\}} + Y_t, \quad t = 1, ..., n$$

Afterwards, we assess the stationarity of the residual time series ($\{\hat{Y}_t\}$) by employing diagnostic tools such as autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Upon confirming its stationarity, we proceed to model it using an autoregressive moving average (ARMA) approach.

3.2 Autoregressive Moving Average (ARMA) Model

The ARMA model is a fundamental time series modeling approach that combines autoregressive (AR) and moving average (MA) components to capture temporal dependencies within the data.

Let $\{Y_t\}$ denote a stationary time series. The ARMA(p, q) model is formulated as follows:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \tag{2}$$

where:

- Z_t is a white noise error term with mean zero and constant variance σ^2 (i.e. $Z_t \sim WN(0, \sigma^2)$).
- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients representing the influence of past observations on the current value,
- $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients representing the influence of past white noise errors on the current value,

In addition, it is imperative that the polynomials $(1 - \phi_1 z - \ldots - \phi_p z^p)$ and $(1 + \theta_1 z + \ldots + \theta_1 z^q)$ share no common factors, as any common factor would allow for simplification of the model. The parameters p and q determine the orders of the autoregressive and moving average components, respectively.

In the forthcoming sections, it is important to emphasize the significance of causality within time series analysis. A process is considered causal if its output is solely determined by past and present inputs, without any influence from future inputs. In the context of time series analysis, a causal ARMA model implies that the current value of the series is influenced only by past observations and past stochastic shocks, with no dependency on future values. Causal models are essential for making reliable predictions and interpreting the underlying dynamics of the time series data.

3.3 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

In this section, we provide descriptions for both the ARIMA and SARIMA models. While the first one captures the general behavior of a time series, the second model extends it by incorporating seasonal components to address periodic variations in the data.

3.3.1 Autoregressive Integrated Moving Average (ARIMA) Model

The ARIMA model is a generalization of the autoregressive moving average (ARMA) model that includes differencing to make the time series stationary.

Let $\{X_t\}$ be a time series. $\{X_t\}$ is an ARIMA(p, d, q) process if the time series $\{Y_t\} := \nabla^d X_t$ is a causal ARMA(p, q) process satisfying:

$$\phi(B) Y_t = \theta(B) Z_t \tag{3}$$

where:

- ∇ is the differencing operator: $\nabla X_t = X_t - X_{t-1} = (1 - B) X_t$,
- B is the backshift operator: $B X_t = X_{t-1}$,
- $Z_t \sim WN(0, \sigma^2)$

- $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ is the autoregressive polynomial of order p,
- $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ is the moving average polynomial of order q.

The parameters p, d, and q determine the orders of the autoregressive, differencing, and moving average components, respectively.

3.3.2 Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The SARIMA model extends the ARIMA model by incorporating seasonal components to account for periodic variations in the time series data.

Let $\{X_t\}$ be a time series. $\{X_t\}$ is a Seasonal ARIMA(p, d, q)x(P, D, Q)$_s$ process with period s if the time series $\{Y_t\} := \nabla^d \nabla_s^D X_t$ is a causal ARMA process satisfying:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t \tag{4}$$

where:

- ∇_s is the lag-s differencing operator: $\nabla_s X_t = X_t - X_{t-s} = (1 - B^s)X_t$,
- s is the seasonal period,
- B is the backshift operator,
- $Z_t \sim WN(0, \sigma^2)$
- $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$ is the non-seasonal autoregressive polynomial of order p,
- $\Phi(z) = 1 - \Phi_1 z - ... - \Phi_P z^P$ is the seasonal autoregressive polynomial of order P,
- $\theta(z) = 1 + \theta_1 z - ... - \theta_q z^q$ is the non-seasonal moving average polynomial of order q,
- $\Theta(z) = 1 + \Theta_1 z - ... - \Theta_Q z^Q$ is the seasonal moving average polynomial of order Q.

The parameters p, d, q, P, D, and Q determine the orders of the autoregressive, differencing, and moving average components, both seasonal and non-seasonal.

3.4 Order selection, residuals diagnostic, and predictions

The selection of model parameters is guided by visually analyzing the ACF and PACF plots as well as by a likelihood-based approach aimed at minimizing a criterion such as the Akaike information criterion (AIC). Subsequently, an analysis of the residuals is conducted to assess the adequacy of the selected model. Forecasts for the Rhône temperature over the next h months (e.g., h = 36) are then generated using the fitted models with optimal parameters and best linear prediction. This entails estimating the coefficients of a linear combination of past observations that minimizes the mean square prediction error. Furthermore, confidence intervals are computed to quantify the uncertainty surrounding temperature predictions. For more comprehensive insights, readers are referred to Chapters 5 and 6 of [Brockwell and Davis, 2016].

4 RESULTS AND DISCUSSION

In this section, we delve into the analysis and interpretation of the results obtained from the modeling and forecasting of the monthly temperature time series data.

4.1 Model 0 - Trend and Seasonality

Table 2 displays the results of the linear model described in section 3.1. The coefficients in the table represent the average deviations in monthly mean temperature relative to the reference month, January. The robust statistical significance, indicated by notably low p-values for all coefficients, underscores meaningful deviations from the reference month. For instance, August exhibits an average temperature approximately 5.40°C higher on average than that of January.

Table 2: Regression results - Model 0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.7556	0.0713	52.64	0.0000***
February	0.4587	0.0902	5.09	0.0000***
March	1.7215	0.0902	19.09	0.0000***
April	3.4974	0.0902	38.78	0.0000***
May	4.9309	0.0906	54.43	0.0000***
June	5.5128	0.0906	60.85	0.0000***
July	5.7230	0.0906	63.17	0.0000***
August	5.3994	0.0906	59.60	0.0000***
September	4.8680	0.0906	53.73	0.0000***
October	3.6845	0.0906	40.67	0.0000***
November	1.6799	0.0906	18.54	0.0000***
December	0.2246	0.0906	2.48	0.0134*
Trend	0.0012	0.0001	12.38	0.0000***

Additionally, the significant positive trend coefficient suggests a linear increase in temperature over time. Overall, these results strengthen the observations discussed in section 2, contributing to our understanding of the temperature dynamics in the studied region. Figure 4 displays the ACF of the residual time series of Model 0 and of the seasonally-differenced time series (i.e. $\nabla_{12} X_t$).

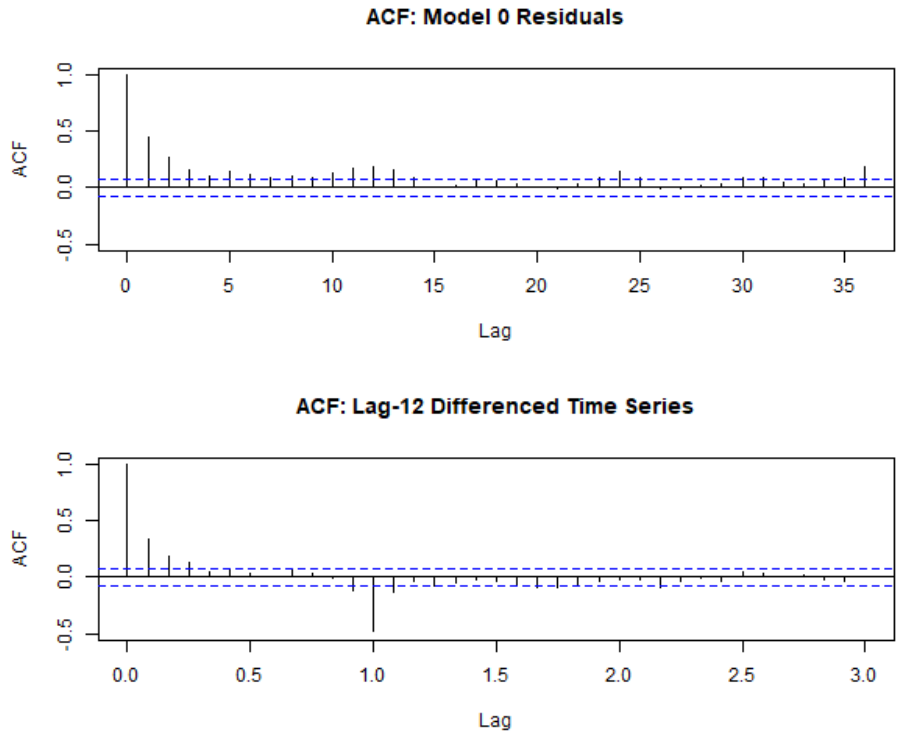

Figure 4: ACF: Model 0 versus Lag-12 differenced monthly temperature

The ACF plot of the residuals of Model 0 does not exhibit the expected patterns. Specifically, there is no exponential decrease or significant spike at a specific lag. Moreover, many lags, up to lag number 60, are statistically significant. As a result, using this method to model the temperature may not be appropriate. Conversely, the ACF plot of the seasonally-differenced time series reveals an interesting pattern: an exponential decrease for the initial lags followed by a pronounced spike at lag 12. Subsequent analysis of the partial autocorrelation function (PACF) of this series is presented in Figure 5.

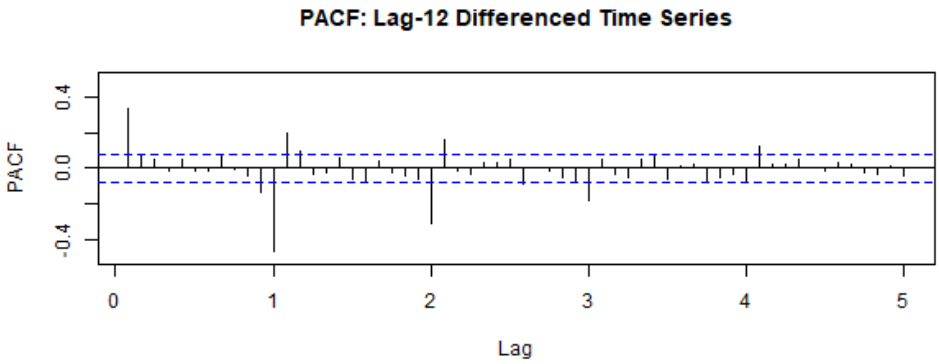

Figure 5: PACF: Lag-12 differenced monthly temperature

The above Figure exhibits an exponential decrease of the seasonal component and a spike at lag 1 when looking at the first few lags. Overall, based on the previous analysis, it appears that a $SARIMA(1,0,0)(0,1,1)_{12}$ model is suitable for the data.

4.2 SARIMA Models

Table 3 presents the outcomes of three SARIMA models. The first one is the model suggested by the analysis of the previous section. Models 2 and 3 are derived using the `auto.arima()` function in R. The seasonal period is set to 12 as expected. Model 2 is optimized to minimize the AICc criterion, while model 3 focuses on the BIC.

Table 3: Comparison of SARIMA(p, d, q)(P, D, Q) models

	Model 1	**Model 2**	**Model 3**
Parameters	$(1,0,0)(0,1,1)$	$(2,0,2)(0,1,1)$	$(1,0,1)(0,1,1)$
ar1	0.4161	1.5397	0.6826
	(0.0359)	(0.0988)	(0.0679)
ar2	—	−0.5457	—
	(—)	(0.0958)	(—)
ma1	—	−1.1787	−0.3249
	(—)	(0.1113)	(0.0890)
ma2	—	0.2091	—
	(—)	(0.1002)	(—)
sma1	−0.8467	−0.8649	−0.8349
	(0.0281)	(0.0254)	(0.0253)
$\hat{\sigma}^2$	0.1846	0.1811	0.1829
Log Likelihood	−387.42	−380.35	−383.93
AIC	782.83	772.7	775.86
AICc	782.89	**772.83**	775.92
BIC	800.83	799.69	**793.86**

The models agree on the seasonal component, with all of them having a seasonal MA coefficient between -0.8 and -0.9. This implies that the time series embeds serial seasonal autocorrelation. Furthermore, the parameters seem to exhibit way larger absolute values than their standard errors across all models, suggesting statistical significance.

4.3 Residual analysis

The analysis of residuals constitutes a pivotal aspect in the process of model selection and validation within statistical frameworks. Figure 6 displays the diagnostic of the residuals of Model 2 obtained using the `tsdiag()` function in R.

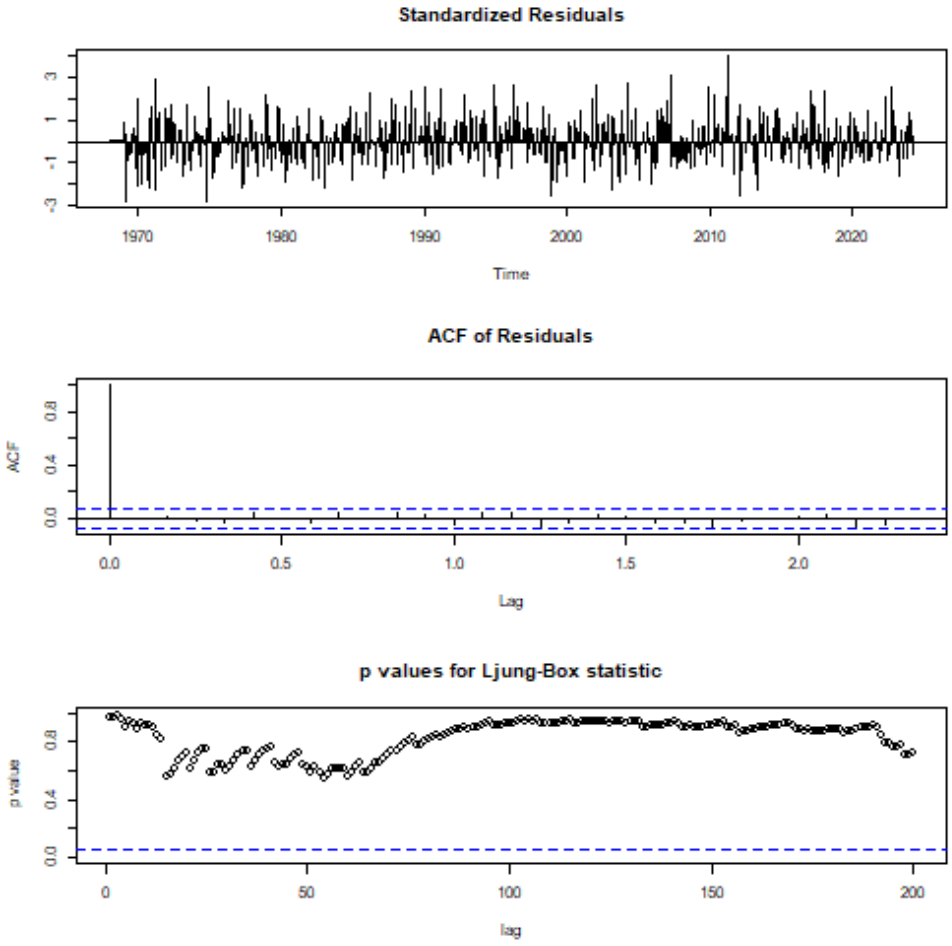

Figure 6: Residual analysis of Model 2

In our context, the adherence of residuals to the characteristics of a white noise (i.e., $WN(0, \hat{\sigma}^2)$) signifies the suitability of a model for representing the underlying data.

The insights drawn from Figure 6 validate the adequacy of Model 2: the residuals time series does not integrate autocorrelation or any specific pattern. In addition, it should be mentioned that a Shapiro-Wilk test rejects the assumption of normality within residuals. Similar conclusions can be drawn for Model 3. However, differences in p-values of the Ljung-Box statistic are observed, with smaller ones apparent at lower lags compared to Model 2. Despite this, they remain above the critical threshold. The Shapiro-Wilk test once again rejects the assumption of normality

within the residuals in the case of Model 3. Ultimately, Model 1 exhibits autocorrelation within the residuals, indicating inadequacy and suggesting the necessity of a more complex model, such as Model 2 or Model 3.

As only valid models warrant further consideration, Models 2 and 3 become the focal points for subsequent analysis. The selection process post-residuals analysis hinges on comparing AICc/BIC values and the number of parameters, with preference accorded to parsimony in instances of near-equal performance. However, the dichotomy between AICc and BIC values necessitates a nuanced evaluation, balancing model fit against complexity. Considering the need for constructing a predictive model, prioritizing AICc minimization favors Model 2 over Model 3, although the latter boasts fewer parameters.

4.4 Forecasts

While the computed models fail the Shapiro-Wilk test for normality, the residuals conform to white noise characteristics per the Ljung-Box test, enabling the computation of confidence intervals. Nonetheless, the non-normality of residuals produces confidence intervals deviating from the norm.

Figure 7 displays monthly average temperature forecasts for the next 36 months using model 2.

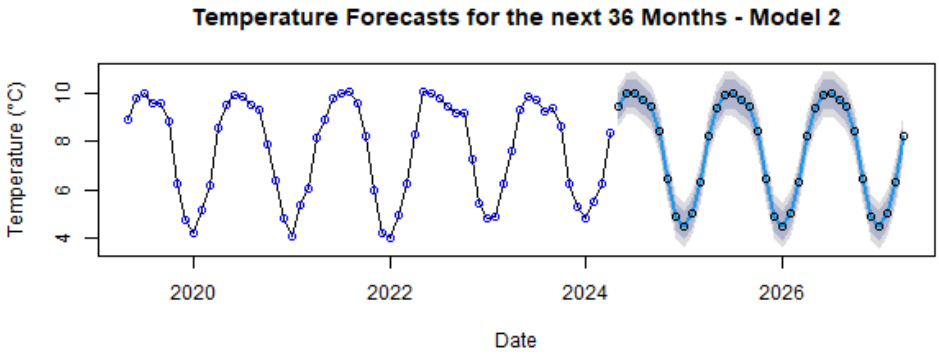

Figure 7: Monthly average temperature forecasts for the next 36 months using Model 2.

The faithful replication of the water temperature's seasonal cycle underscores its natural perpetuation. Consequently, the predicted values are anticipated to remain non-zero. Furthermore, extended forecasting periods, such as 100 years, result in slightly widening confidence intervals, reflecting heightened uncertainty with temporal distance from the present.

4.5 Limitations and improvements

> All models are wrong, but some are useful.
>
> George E. P. Box (1919 – 2013)

The first drawback observed in our modeling endeavors pertains to the non-normality exhibited by the residuals, as evidenced by the Shapiro-Wilk tests, consequently resulting in confidence intervals deviating from the conventional normal 95% bounds.

Our analysis relies solely on SARIMA models for predicting monthly temperature averages. While SARIMA models are powerful tools for time series forecasting,

they may not capture all the underlying complexities of the data: our study focuses on analyzing patterns within the temperature data itself. We do not consider external factors (i.e. exogenous variables) that may influence monthly temperature variations, such as atmospheric conditions, land use changes, or anthropogenic activities. Quantifying those variables may provide relevant insights to the analysis. Finally, the study confirms what is already known about temperature variations in the study area. While validation of existing knowledge is important, it does not provide any substantial insight into the underlying processes behind temperature changes.

5 CONCLUSION

The study focuses on temperature variations of the Rhône, leveraging advanced time series analysis techniques and data provided by the Federal Office for the Environment (FOEN). Through our comprehensive analysis, we have observed significant temporal patterns and fluctuations in the river's temperature, reflecting the complex interplay of climatic factors, such as seasonal variations.

Our findings highlight the increasing trend in the temperature of the Rhône over the years, which may be indicative of broader climate change trends. We also have identified clear seasonal patterns, with temperatures peaking during the summer months, which aligns with expectations based on known climatic patterns. Furthermore, our analyses have revealed the presence of temporal dependencies within the data, suggesting the potential for utilizing past temperature observations to inform future predictions.

It is worth noting that our study is not without limitations: it only confirms existing knowledge. We also acknowledge the inherent uncertainties associated with time series analysis and modeling, as well as the challenges of predicting complex environmental systems. Moreover, our analysis solely focuses on SARIMA modeling and may benefit from the inclusion of exogenous variables to enhance predictive accuracy and provide substantial insights on the dynamics underlying temperature variations.

Overall, our study contributes to the collective knowledge of the Rhône's temperature dynamics and underscores the importance of continued monitoring and research efforts to support effective environmental management and conservation initiatives.

REFERENCES

[Bouffard et al., 2019] Bouffard, D., Dami, J., and Schmid, M. (2019). Swiss lake temperature monitoring program. Report, Federal Office for the Environment (FOEN), Hydrology Division, CH-3003 Bern.

[Brockwell and Davis, 2016] Brockwell, P. and Davis, R. (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing.

[Denzler, 2019] Denzler, L. (2019). Magazine «l'environnement» 1/2019 - un réseau de vie pour la suisse.

[Ducharne, 2008] Ducharne, A. (2008). Importance of stream temperature to climate change impact on water quality. *Hydrology and Earth System Sciences*, 12(3):797–810.

[Hardenbicker et al., 2017] Hardenbicker, P., Viergutz, C., Becker, A., Kirchesch, V., Nilson, E., and Fischer, H. (2017). Water temperature increases in the river rhine in response to climate change. *Regional Environmental Change*, 17:299–308.

[Khalanski et al., 2009] Khalanski, M., Carrel, G., Desaint, B., Fruget, J.-F., Olivier, J.-M., Poirel, A., and Souchon, Y. (2009). Étude thermique globale du rhône. impacts hydrobiologiques des échauffements cumulés. *Hydroécologie Appliquée*, 16.