UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA EXPERIMENTAL YACHAY

Escuela de Ciencias Matemáticas y Computacionales

TÍTULO: Forecasting groundwater level recession patterns through ARIMA and Hidden Markov Model: A comparative study

Trabajo de integración curricular presentado como requisito para la obtención del título de Matemática

Autor:

Chaglla Aguagallo Diana Karina

Tutor:

Ph.D. Morocho Cayamcela Manuel Eugenio

Urcuquí, Julio 2021

Urcuquí, 25 de junio de 2021

SECRETARÍA GENERAL
(Vicerrectorado Académico/Cancillería)
ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES
CARRERA DE MATEMÁTICA
ACTA DE DEFENSA No. UITEY-ITE-2021-00019-AD

A los 25 días del mes de junio de 2021, a las 13:00 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

Presidente Tribunal de Defensa	Dr. GALLO FONSECA, RODOLFO , Ph.D.
Miembro No Tutor	Dr. ACOSTA ORELLANA, ANTONIO RAMON , Ph.D.
Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.

El(la) señor(ita) estudiante **CHAGLLA AGUAGALLO, DIANA KARINA**, con cédula de identidad No. **1803984283**, de la **ESCUELA DE CIENCIAS MATEMÁTICAS Y COMPUTACIONALES**, de la Carrera de **MATEMÁTICA**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-15-No.174-2015**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **FORECASTING THE GROUNDWATER LEVEL RECESSION PATTERNS THOUGH ARIMA AND HIDDEN MARKOV MODELS: A COMPARATIVE STUDY.**, previa a la obtención del título de **MATEMÁTICO/A**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Presidente Tribunal De Defensa	Dr. GALLO FONSECA, RODOLFO , Ph.D.	10,0
Tutor	Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.	9,5
Miembro Tribunal De Defensa	Dr. ACOSTA ORELLANA, ANTONIO RAMON , Ph.D.	10,0

Lo que da un promedio de: **9.8 (Nueve punto Ocho)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

CHAGLLA AGUAGALLO, DIANA KARINA
Estudiante

DIANA KARINA CHAGLLA AGUAGALLO
Firmado digitalmente por DIANA KARINA CHAGLLA AGUAGALLO
Fecha: 2021.07.08 20:17:21 -05'00'

Dr. GALLO FONSECA, RODOLFO , Ph.D.
Presidente Tribunal de Defensa

RODOLFO GALLO FONSECA
Firmado digitalmente por RODOLFO GALLO FONSECA
Fecha: 2021.06.29 11:05:58 -05'00'

Firmado electrónicamente por:
MANUEL EUGENIO
MOROCHO CAYAMCELA

Dr. MOROCHO CAYAMCELA, MANUEL EUGENIO , Ph.D.
Tutor

ANTONIO RAMON ACOSTA ORELLANA
Firmado digitalmente por ANTONIO RAMON ACOSTA ORELLANA
Fecha: 2021.06.28 21:54:07 -05'00'

Dr. ACOSTA ORELLANA, ANTONIO RAMON , Ph.D.
Miembro No Tutor

TATIANA BEATRIZ TORRES MONTALVAN
Firmado digitalmente por TATIANA BEATRIZ TORRES MONTALVAN
Fecha: 2021.06.28 21:53:25 -05'00'

TORRES MONTALVÁN, TATIANA BEATRIZ
Secretario Ad-hoc

Autoría

Yo, **Diana Karina Chaglla Aguagallo**, con cédula de identidad **1803984283**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el autor del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, Julio del 2021.

DIANA KARINA CHAGLLA AGUAGALLO

Firmado digitalmente por DIANA KARINA CHAGLLA AGUAGALLO
Fecha: 2021.07.08 20:17:51 -05'00'

Diana Karina Chaglla Aguagallo
CI: 1803984283

Autorización de publicación

Yo, **Diana Karina Chaglla Aguagallo**, con cédula de identidad **1803984283**, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Urcuquí, Julio del 2021.

DIANA KARINA CHAGLLA AGUAGALLO

Firmado digitalmente por DIANA KARINA CHAGLLA AGUAGALLO
Fecha: 2021.07.08 20:18:26 -05'00'

Diana Karina Chaglla Aguagallo
CI: 1803984283

Dedication

Este trabajo esta dedicado a toda mi familia que me apoya desde el inicio de todo. A mis padres Sonia y Ricardo, a mis hermanos Hilary y Dylan, a mi perrito Hobi por ser mi luz de todos los dias y a mi tía Janeth sin su ayuda nada de esto hubiese sido posible.

Acknowledgments

Quiero empezar agradeciendo a todos los profesores que formaron parte de mi vida en todos estos años, gracias por sus enseñanzas y por ser mi más grande ejemplo a seguir. En especial quiero agradecer al profesor Manuel Morocho que no negó su ayuda cuando más lo necesitabamos y a Diego Peluffo, por ser mi amigo, mi tutor y mi guía a lo largo de este camino, gracias por la confianza, por ayudarme a crecer tanto académica como personalmente y también gracias por abrirme las puertas de su grupo de investigación. Aprovecho para agradecer a SDAS Research Group por el apoyo y a FONAG por su apertura en este trabajo en colabaración, en particular a Paola Fuentes por solventar mis dudas a lo largo del proyecto.

Agradezco también a todos mis amigos, a Arita que está junto a mí desde que todo esto fue un sueño y ha sido mi compañera incondicional. A Juan González gracias por ser mi cotutor no oficial, por tu amistad y por toda la ayuda brindada en la realización de este trabajo. A Dome y Lai, siempre serán la mas bonita coincidencia, gracias por ser mis hermanas de otra madre y por hacer mi vida universitaria mucho menos sola, las llevaré en mi corazón siempre. A Alejandro y Celi, gracias por estar conmigo desde el inicio y quedarse hasta el final. A mis colegas, mi familia matemática MATE 4G, gracias por las risas y los buenos momentos, los recordaré siempre.

Finalmente, gracias a todos los artistas que hicieron las canciones que me acompañaron en las noches de desvelo principalmente Thank You, Encore, Epiphany y Pied Piper. A mi familia, no me va a alcanzar la vida para agradecerles todo lo que han hecho por mí. Y a todo aquel que ha formado parte de mi vida de una u otra manera y cree en mí y me apoya.

Abstract

The high Andean wetlands -considered as water conservation areas have suffered significant damage due to overgrazing and systematic drainage. In this regard, the Fideicomiso Mercantil Fondo para la Protección del Agua (FONAG) has developed a baseline to get information about their state of degradation by creating 18 wells that allow measuring its groundwater level, and hydraulic dynamics in rainy and dry seasons.

As a remedial action against the damage, artificial drains have been plugged to monitor the vegetation recovery process, soils and water dynamics of wetlands. Nonetheless, it has not been possible to identify whether the restoration action helps to mitigate the damage existing in the area; and until now, there is no enough amount of data to conduct further research. To overcome this challenge, we propose a custom time series forecasting model, which consists of three main stages. First, we acquire a high-quality data set to use it in our model implementation and analysis. Secondly, we select the parameters that better fit both models, in particular the auto-regressive integrated moving average and hidden Markov model. Finally, we present a comparative study between the aforementioned models. The proposed strategy is expected to measure the effectiveness of the chosen model, and serve as a baseline, enabling further work with the time-series collected data in other wells located in the wetland.

Keywords: Time series forecasting, groundwater level, wetland, ARIMA, Hidden Markov Model.

Resumen

Los humedales altoandinos -considerados como áreas de conservación de agua- han sufrido daños importantes debido al pastoreo excesivo y al drenaje sistemático. Por esta razón, el Fideicomiso Mercantil Fondo para la Protección del Agua (FONAG) ha desarrollado una línea de base con el fin de obtener información sobre su estado de degradación mediante la creación de 18 pozos que permiten medir sus niveles de agua subterránea y dinámica hidráulica en épocas lluviosas y secas.

Como acción de restauración contra el daño, se han taponado los drenajes artificiales con el propósito de monitorear el proceso de recuperación vegetal, de los suelos, y la dinámica del agua de los humedales. Sin embargo, no se ha podido identificar si la acción de restauración ayuda a mitigar el daño existente en el área; y hasta ahora no se dispone de la suficiente cantidad de datos para realizar una investigación adicional. Para solucionar este problema, proponemos un modelo personalizado de predicción de series de tiempo, que consta de tres etapas principales. En primer lugar, adquirimos un conjunto de datos de alta calidad para usarse en la implementación y análisis de los modelos. En segundo lugar, seleccionamos los parámetros que mejor se ajusten a los modelos a desarrollar, en particular el modelo autorregresivo integrado de media móvil y el model oculto de Markov. Finalmente, presentamos un estudio comparativo entre los modelos antes mencionados. Se espera que la estrategia propuesta mida la efectividad de los modelos y sirva como base para poder trabajar con datos recopilados en series de tiempo de los otros pozos ubicados en el humedal.

Palabras Clave: Predicción de series de tiempo, nivel de agua subterránea, humedal, ARIMA, Modelo Oculto de Markov.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Time series forecasting has several approaches in different areas such as: scientific, economic, climatic, industrial, etc. Its importance lies in the fact that good forecasting can be helpful to take preventive decisions in order to solve a problem even before it happens. In this particular case, being the groundwater a natural resource mainly used for domestic supply and irrigation, accurate forecasting of groundwater level will provide a measurement of resource quantity and allowable exploitation, as well as the sustainable utilization and the scientific management of the resource of groundwater [4].

Even though, this area has been often neglected because of the existing time component that makes the analysis a little bit more difficult to carry out, over the years many studies have been developed; for instance, the researchers reviewed this problem in [5]. The techniques applied to accomplish this matter range from the simplest to the more complex such as neural networks. This work proposes a comparative study of two forecasting methods utilized in groundwater level time series, namely autoregressive moving average (ARIMA) model that has been widely employed and hidden Markov model, an approach usually implemented for speech recognition problems, but has gained some popularity in prediction area.

The structure of this work is divided into 6 chapters. Specifically, introduction that contains the problem statement 1.1, contribution 1.2 and the objectives of the work 1.3. Overview and background that includes a short literature review of previous related works 2.1, brief context of the work 2.2, and a background about time series analysis and prediction 2.3, together with a detailed explanation of the models to be developed 2.4 and 2.5. Methodology where an outline 3.1 of the work is showed, in addition to a complete description of the steps to realize in order to accomplish the study, and the metrics to evaluate the performance of the models. Experimental setup that described the considered time series 4.1, as well as the parameters of the models 4.3. The outcomes of the developed experiments are presented in the results and discussion section by using plots and tables. Finally, Conclusion 6.1 giving the final remarks of the study and an analysis of whether the objectives were successfully fulfill and future works 6.2 that might be carried out after appropriate modifications to the proposed models.

1.1 Problem statement

Wetlands make an important contribution to the protection and enhancement of groundwater quality. Unfortunately, due to anthropogenic hazards, they have been affected leading to harm in the water regulation of the ecosystem.

Concretely, Pugllohuma wetland located in Antisana ecological reserve that is a valuable water resource of Quito city has been marred because of the overgrazing and systematic drainage. In this regard, FONAG since 2016 has started a restoration action and monitoring of the wetland dynamic. Nevertheless, they have not been able to obtain enough amount of data to make an extensive analysis of the groundwater level before and after the restoration, and verify if the restoration action helps to mitigate the damage in the wetland dynamic.

1.2 Contribution

With the aim of providing a solution to tackle existing issues, this work proposes a comparative study of two forecasting methods. Subsequently, we performed a selection of one of them intending to get a suitable prediction of groundwater level, in order to know in advance the future dynamic, and strengthen the restoration action if its needed.

1.3 Objectives

1.3.1 General Objective

To develop a methodology and an experimental setup to compare representative time-series-analysis techniques hidden Markov model and autoregressive moving average (ARIMA) on the evaluation of groundwater level in terms of accuracy.

1.3.2 Specific Objectives

- To propose a data preparation stage through suitable time-series pre-processing approaches for accomplishing proper, clean data for subsequent analysis.
- To implement and tune selected representative time-series techniques ARIMA and Hidden Markov model for forecasting future groundwater level values.
- To design a comparative study between ARIMA and Hidden Markov model by the use of metrics in the predicted groundwater level values including a methodological and experimental approach.

Chapter 2

Overview and Background

2.1 Related Works

In this section, some related works about time series forecasting will be reviewed.

Patle, Singh, Sarangi, Rai, Khanna and Sahoo (2015)

The study was carried out using the data of pre and post-monsoon from 1974 to 2010 in Karnal district of Haryana-India. The goal was to identify the trends using the Mann-Kendall test and Sen's slope estimator, and ARIMA for time series modeling of groundwater levels for forecasting. According to [6], the use of ARIMA model was supported by the previous accurate forecasting studies developed in [7], with the monthly reference crop evapotranspiration for the Jordan Valley, in [8] with streamflow for Salt River basin in Arizona, in [9] with precipitation, monthly average temperature and relative humidity, in [10] with monthly river flow Selangor riverand and Bernam rivers of Selangor state in Malaysia, and in [11] using SARIMA to get a temporal behavior of groundwater tables.

The time series was divided with a 70:30 ratio for training and testing, respectively. The train data set is from 1974 to 1999 and the test data set from 2000 to 2010. After the selection of the model with Bayesian information criteria (BIC), its accuracy was evaluated by comparing the test data set with the forecasted using the following metrics: root mean square error , mean absolute percentage error, mean absolute error, MaxAPE and MaxAE. Forecasting results indicated that pre and post-monsoon groundwater levels would decline by 12.97 m/yr and 12.0 m/yr over the observed groundwater levels in 2010. Average rates of decline of pre and post monsoon groundwater levels in Karnal district during 1974-2010 were 0.23 m/yr and 0.27 m/yr, respectively, which would increase to 0.32 m/yr and 0.30 m/yr for the period of 2011 to 2050.

Yan and Ma (2016)

Yan and Ma used a combined model of ARIMA and radial basis function network (RBFN) for the prediction of monthly groundwater level fluctuations for two observation wells in the city of Xi'an, China. The data is divided from the year 1998 to 2008 used for training,

and the data from the year 2009 to 2010 for testing the proposed hybrid model. As is mentioned in [4], ARIMA could have some forecasting failures when the problem to be treated is not linear, for this reason this study use the ARIMA model to estimate the linear part and RBFN to the nonlinear residuals of the groundwater level time series, with the objective of getting better forecasting results.

Next, the metrics employed in this paper, in order to select the best structure and parameters of ARIMA and RBFN models are root mean square error (RMSE), the mean absolute error (MAE) and the correlation coefficient (R^2). The results showed that ARIMA, RBFN, and the hybrid model have high fitting accuracy in the training sets. However, this does not necessarily mean a higher forecasting accuracy in the testing sets. After making the comparison between the two pure models and the hybrid model the results show that the last one is more competent in forecasting groundwater level.

Khadr (2015)

Khadr [12] executes several homogeneous hidden Markov models (HMMs) to forecast droughts using the Standardized Precipitation Index in daily precipitation data set, collected from January 1960 to December 2007 from 22 meteorological stations in the upper Blue Nile basin. The measures of goodness used to evaluate the forecast performance of HMM models include mean absolute deviations (MAD), the coefficient of determination (R^2), root mean square error (RMSE) and correlation coefficient (C_r). Moreover, to investigate whether there is a significant difference between the mean from the observed and predicted data a Z-test for the means was employed in the analysis.

The hidden Markov model was used after the computation of the SPI index, defining seven states: State 1 – Extremely wet, State 2 – Very wet, State 3 – Moderate wet, State 4 – normal state, State 5 – Moderate drought, State 6 – Severely drought, State 7 – Extremely drought. The data set from 1994 to 2007 was used to validate the forecast applying the metrics mentioned above. The results show that hidden Markov models could be used to forecast SPI time series of multiple timescales for more than one month ahead. Furthermore, since HMM outcome is potentially skillful, it can be used as a essential tool to improve drought management and for medium-short term planning in water resources management.

Chen, Shin and Kim (2016)

In [13] a new probabilistic scheme to forecast droughts using a discrete-time finite state-space hidden Markov model (HMM) aggregated with the representative concentration pathway is proposed. The study begins citing [14] as one of the first application of HMM to stochastic hydrology, although recently has been used as a forecasting tool in prediction groundwater level fluctuations [15], in earthquake probability [16], and in sea surface elevation [17]. A standard precipitation index (SPI) times series is employed to carry out the training and forecasting validation, the data was obtained after calculated a continuous record of monthly precipitation data from 5 stations located in specific regions of South Korea. The data from 1974 to 2002 is the input for HMM as a training data set while the data from 2003 to 2021 is assigned to testing.

Once the training data is modeled, the forecasted mean value are examined using point forecast skill score (SS), in addition to a probabilistic verification using ranked probability score (RPSS) and the relative operating characteristic (ROC), as a tool for evaluating the quality of the probabilistic forecasts. In the first evaluation stage, is conducted a comparative analysis of the performance of HMM with RCP information, HMM without RCP information, ARMA with the parameters obtained from correlation and partial correlation plot, and a three-layered feed-forward artificial neural network (ANN). The results indicated that the HMM-RCP forecast mean values displayed a significant improvement in forecasting skill score over the other models. In the second evaluation stage, the probabilistic forecast performance of the HMM-RCP and HMM without RCP is measured by comparing the RPS and RPSS for different lead times, in this case the results show that HMM-RCP is a useful tool to forecast drought, capturing the number and duration of drought events occurred during the validation period. The study ends making the 3 following conclusions about HMM-RCP: is able to forecast future SPI considering uncertainties, can provide an accurate long lead time probabilistic forecast, and precisely reflect the statistical properties of future droughts (occurrence, duration, severity).

Adamowski and Chan (2011)

Adamokski and Chan proposed in [18] to make a comparative study of coupled wavelet neural network, autoregressive moving average (ARIMA), and artificial neural networks (ANN) models. Hidrological and meteorological data are used for ANN and WA-ANN, specifically monthly time series of precipitation, average temperature, and average groundwater level. The precipitation and temperature data sets were collected from the website of Environment Canada, while the groundwater level time series were obtained from two wells located in the cities of Mercier and St-Remi in Canada, all of them goes from November 2002 to October 2009.

While in the ARIMA model the training data set corresponds to the period from November 2002 to February 2009, and the testing data set from March 2009 to October 2009, in ANN and WA-ANN data series were divided into a training set from November 2002 to June 2008, validation set from July 2008 to February 2009, and testing set from March 2009 to October 2009. The models were calibrated using the time series destined to training/validation, then the performance was evaluated using the coefficient of determination (R^2), Nash–Sutcliffe model efficiency coefficient (E), and root mean squared error (RMSE) comparing the testing data set to the forecasted values. After the evaluations were done this research concludes that WA-ANN demonstrated a better accuracy in the forecasting of groundwater level time series of the two cities, and conclude that making the hypothesis that the WA–ANN models are more accurate because wavelet transforms provide useful decompositions of the original time series, improving the performance of ANN forecasting.

Summary

To summarize the previously mentioned related works, Table 2.1 is presented.

Name of work	Authors	Year	Proposed models
"Time series analysis of groundwater levels and projection of future trend"	Patle, Singh, Sarangi, Rai, Khanna, and Sahoo	2015	ARIMA
"Application of integrated arima and rbf network for groundwater level forecasting"	Yan, and Ma	2016	ARIMA + RBFN
"Forecasting of meteorological drought using hidden markov model (case study: The upper blue Nile river basin, Ethiopia)"	Khadr	2015	Hidden Markov Model + Standardized Precipitation Index
"Probabilistic forecasting of drought: a hidden markov model aggregated with the rcp 8.5 precipitation projection"	Chen, Shin, and Kim	2016	Hidden Markov Model + RCP
"A wavelet neural network conjunction model for groundwater level forecasting"	Adamowski, and Chan	2011	WA-ANN, ANN and ARIMA

Table 2.1: Related works summary.

2.2 Context

According to [19], the purpose of time series analysis is generally twofold: to understand or model the stochastic mechanisms that give rise to an observed series, and to predict or forecast the future values of a series based on the history of it. Along the years several techniques of time series forecasting have been developed, broadly speaking it is possible to divide the techniques into three main groups judgmental forecasted bases in subjective judgment, and any other relevant information, univariate method where the forecast depends only on the present and past data of a single time-series and multivariate methods where forecast of a given variable depends on one or more additional time series variables [20].

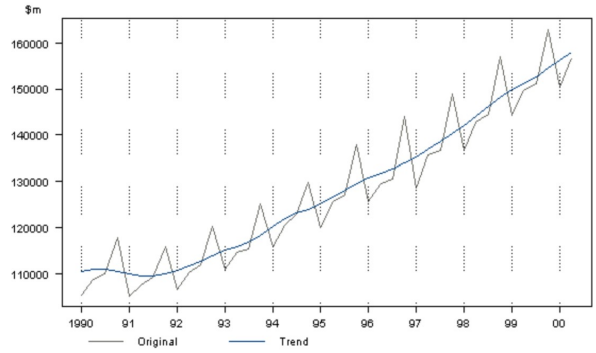
In this work, two univariate methods are addressed namely autoregressive moving average (ARIMA) and hidden Markov model (HMM). ARIMA model has been widely used in many

areas of time series forecasting due to its statistical properties and easy implementation. Although it might have some difficulties when the problem to be dealt with is non-linear, the researches where it was employed has shown an acceptable accuracy prediction. On the other hand, hidden Markov model has several applications, such as speech recognition, gene finding, gesture tracking, etc. However, recently has been gaining popularity thanks to its applications in the forecasting of meteorological drought, groundwater prediction, and market sales forecasting. Moreover, it has a strong statistical foundation, and owing to the fact that it is an efficient learning algorithm that can take place directly from the raw sequence data, it might be helpful when the quality of the data is not good enough.

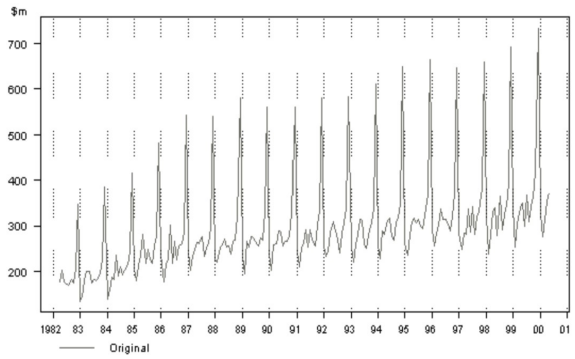
2.3 Background

2.3.1 Time series

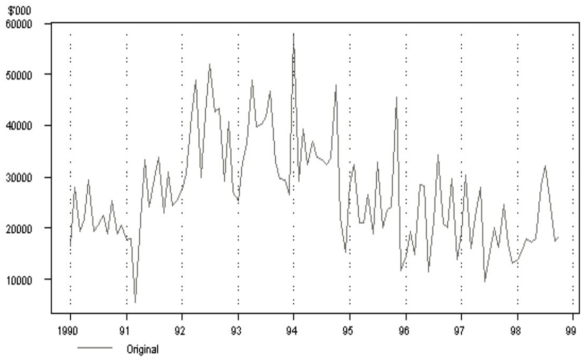
Time series is the data obtained from observations collected sequentially over time [19], are frequently used in any domain of applied science that involves temporal measurements, for instance: statistics, pattern recognition, econometric, mathematical finance, weather forecasting, earthquake prediction, etc. An observed time series can be decomposed into three components: the trend which represents the long-term direction, the systematic calendar related movements named seasonal component and the residual part of the series [1]. To illustrated the time series components the Fig. 2.1 of monthly retail sales in New South Wales retail department stores is presented.



(a) Trend



(b) Seasonal



(c) Residual

Figure 2.1: Time series components of monthly retail sales.
Source: [1]

According to [21], time series can be classified in stationary and non-stationary. Stationarity means that the time series if both the mean and the variance are time-invariant [22], while non-stationary time series do not fulfill this condition. In [2], Fig. 2.2 is presented as an example of the transformation process from a non-stationary time series to a stationary, applied in monthly US net electricity generation data.

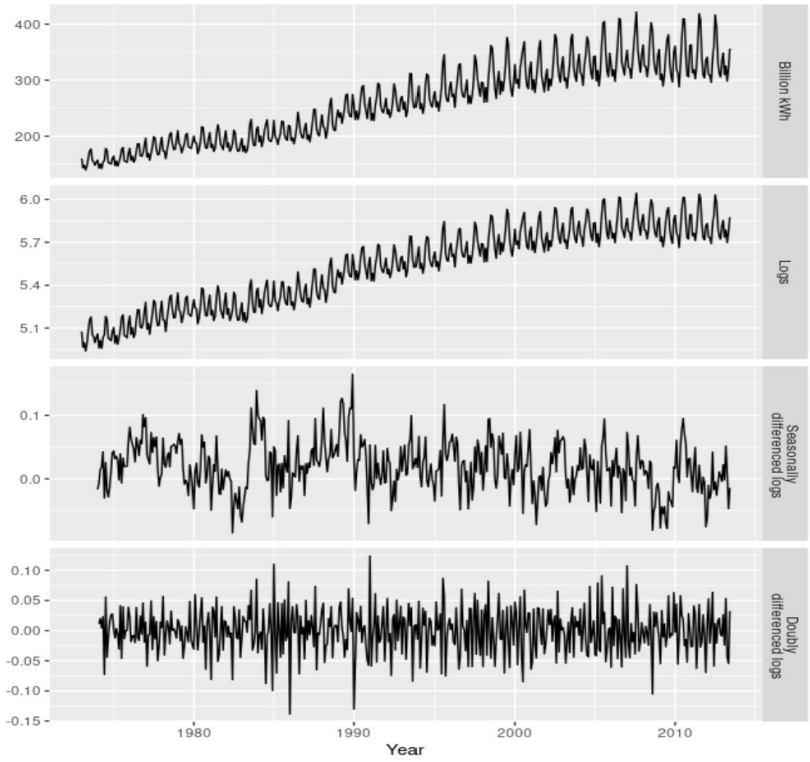


Figure 2.2: Process from non stationary to stationary time series of monthly US net electricity generation.
Source: [2]

2.3.2 Time Series Forecasting

In this section, a brief description of some time series forecasting methods is given.

Simple Moving Average (SMA)

A simple moving average (SMA) is one of the easiest methods for forecasting. It is an average of a subset of periods in a time series [23]. In order to do the forecasting, SMA uses the total average of all past data of time series [19], applying the Eq. 2.1.

$$\hat{Y}_{t+1} = \frac{\sum_{t=1}^{p} Y_t}{p} \tag{2.1}$$

where Y_t is the actual value at time t, p the number of terms in moving average and $\hat{Y}_{t+1}$ the forecasted value for the next period [24]. The accuracy of this method depends on whether the time series has a constant mean or not.

Naive Model

Naive approach is based solely on the most recent information available assuming that recent periods are the best predictors of the future [24], as is illustrated in Eq. 2.2

$$\hat{Y}_{t+1} = Y_t \tag{2.2}$$

The main problem with this model is that it discards all observations except for the most recent one, for this reason this is not a reliable forecasting method.

Simple Exponential Smoothing (SES)

Simple Exponential Smoothing (SES) uses a weighted moving average of past data as the basis for a forecast [25]. The Eq. 2.3 calculated the forecasted values.

$$F_{t+1} = \alpha y_t + (1 - \alpha) F_t \tag{2.3}$$

where y_t is the actual values at the time t, F_t is the forecast value of the variable Y at the time t, F_{t+1} is the forecast value at the time $t + 1$ and α is the smoothing constant this is a value between 0 and 1. SES uses the complete time series, incrementing exponentially the weight when the data is more recent.

Artificial Neural Networks (ANN)

Artificial neural networks are an information processing system that roughly replicates the behavior of a human brain by emulating the operations and connectivity of biological neurons [26]. ANN is usually employed when the data has non linear components, due to its accuracy in those kinds of problems. In [2], explained that a neural network can be thought of as a network of "neurons" which are organized in layers. The predictors (or inputs) form the bottom layer, and the forecasts (or outputs) form the top layer. There may also be intermediate layers containing "hidden neurons" as is showed in Fig. 2.3.

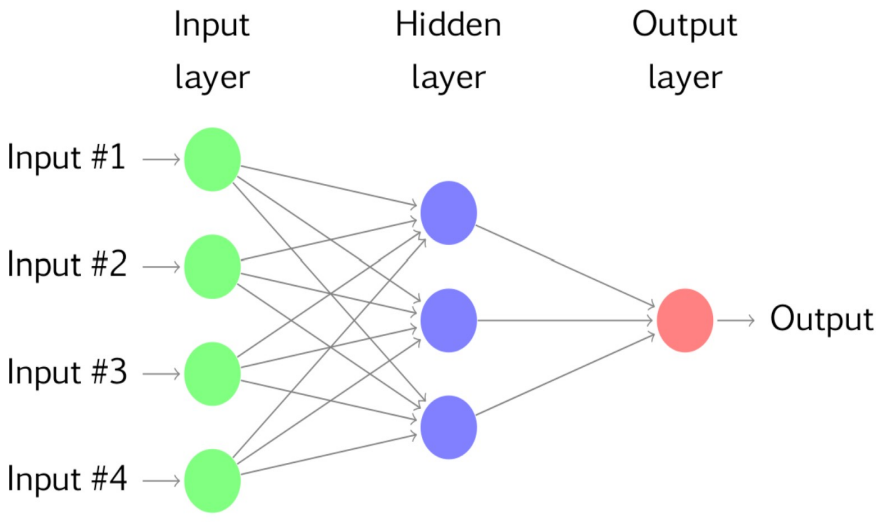


Figure 2.3: Example of artificial neural network with hidden neurons.
Source: [2]

Making a summary of the ANN procedure, the output is a non-linear modification of the previous result obtained from the weighted linear combination input in the last layer, each layer of nodes receives as inputs the outputs of the previous layer. To get a broad perspective a systematic review of ANN applied in time series forecasting can be seen in [27].

Long Short-Term Memory network (LSTM)

The long short-term memory network is a type of recurrent neural network used in deep learning. According to [28], the input units are fully connected to a hidden layer consisting of memory blocks with 1 cell each. The cell outputs are fully connected to the cell inputs, to all gates, and to the output units. Moreover, the gates, the cell and the outputs are biased. Briefly speaking LSTM is organized in cells that include several operations, has an internal state variable, which is passed from one cell to another and modified by operation gates [3]. Figure 2.4 is presented to illustrated the architecture of LSTM.

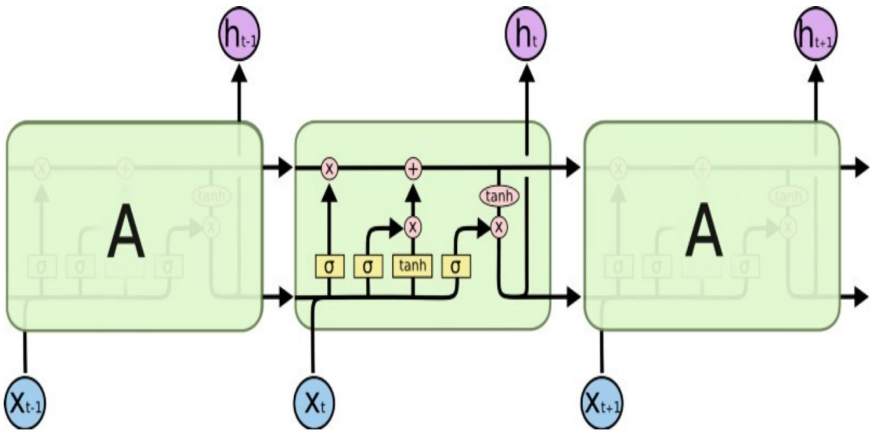


Figure 2.4: Architecture of long short-term memory network (LSTM).
Source: [3]

Support Vector Machine (SVM)

Support vector machine (SVM) is a type of neural network. In [29], is stated that SVM uses a linear model to implement nonlinear class boundaries through some nonlinear mapping the input vectors into the high-dimensional feature space, building a hyperplane or sets of hyperplanes in it. SVM has been mainly used in classification and regression problems, but there exists literature that endorses its functionality it time series forecasting. Namely, the following researches [30], [31], [32], [33], where a wide explanation of SVM forecasting performance can be found.

2.4 ARIMA

Autoregressive integrated moving average (ARIMA) is popular a stochastic linear model for time series forecasting, were originally developed by Box and Jenkins in 1970. The model uses the variation and regression of the data to forecast future values by identifying the trend, in that way the future estimations are supposed to be a linear combination relying upon the past data and past errors, expressed as follows:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}, \tag{2.4}$$

where y_t is the actual value, ε_t is the random error at time t, ϕ_i and θ_j are the coefficients, the last terms p and q are the parameters referred as an autoregressive and moving average, respectively [34].

ARIMA model consists of three parts, the AR part (p) shows that the time series is regressed on its past data, MA part (q) indicates that the forecast error is a linear combination of past respective errors and I part illustrate that the data values have been replaced with differenced values of in order to obtain stationary data, which is the requirement of the ARIMA model approach [35]. The p and q parameters together with d, which represents the number of non-seasonal differences necessary in order to get stationarity, build the called ARIMA(p, d, q) model.

The Box-Jenkins method of time series modeling has four steps as is mentioned in [6].

1. Model identification: the time series is analyzed to find stationarity, a required condition of ARIMA. Subsequently, the autocorrelation function (ACF) and partial autocorrelation (PACF) are also examined.

2. Parameter estimation: the initial values of the parameters p, q are determined by the ACF and PACF, respectively.

3. Diagnostic: Once the parameters are given ARIMA estimates the coefficients ϕ and θ of Eq. (2.4) through the Maximum Likelihood Estimation method.

4. Forecasting: the selected model is used to fit and forecasted values.

2.5 Hidden Markov Model

A Hidden Markov model (HMM) is a statistical model that can be used to describe the evolution of observable events that depend on internal factors, which are not directly observable [36]. It consists of two stochastic processes, the visible process of observable events and the invisible process that is a Markov chain going from one state to another.

HMM is characterized by the following components:

- Number of observed events or symbols.
- Number of hidden factors underlying the observation called states.
- A transition probability matrix, representing the probability of moving from one state to another.
- An emission probability matrix, expressing the probability of an observation being generated from a state.
- An initial probability distribution, standing for the probability of being in a certain state at the beginning.

All the components mentioned above joined together to build the model in the following way: for the initial state we denote the initial probability $\pi = \{\pi_i\}$ which represents the probability to be in state i at time $t = 1$. Then, one state transits to another at time t to time $t+1$ with a determined probability. These probabilities from entering the state j to the current state i forms the transition probability matrix $F = \{f_{ij}\}$, which dimension depends on the number of states S. Finally, given the sequence of observations $O = \{O_1, O_2, ..., O_N\}$ the emission probability matrix $E = \{e_j(O_t)\}$ represents the probability of observing O_t at state j [37].

The goal of HMM is to capture the hidden information from the observable estimating the most likely regime, including the associated time-varying means and volatilities, in order to create a reliable predictive model. Therefore, by analyzing the pattern in the past it can forecast the future outcomes of the time series.

Chapter 3

Methodology

3.1 Outline

The methodology to implement and tune the forecasting methods ARIMA and hidden Markov model are summarized in Fig. 3.1 is based in Crisp-DM methodology and consists of the following stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

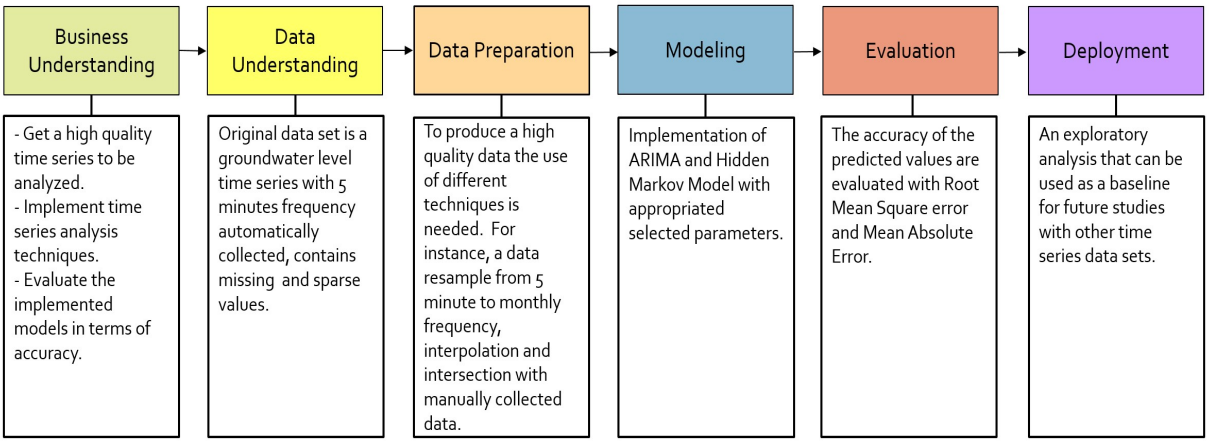


Figure 3.1: Methodology for forecasting the groundwater level.

3.2 Data understanding

The data was obtained from Pugllohuma wetland located in Antisana ecological reserve, it was provided by The Water and Paramo Scientific Station (ECAP) and FONAG.

The original data set consists of 2 time series with 5 minutes frequency, the first one is collected using a sensor located in 1 of the 18 piezometers in the wetland that verifies the groundwater level measured in cm below 0, which represents the water-free space in the one meter long piezometer, and the second one is the precipitation accumulated in that

lapse. Both of them were collected from November 2016 to September 2020.

Due to some instrumental errors, the groundwater level time series include periods of missing data and sparse values that do not match to the groundwater level values manually taken biweekly or monthly as is shown in Fig. 3.2

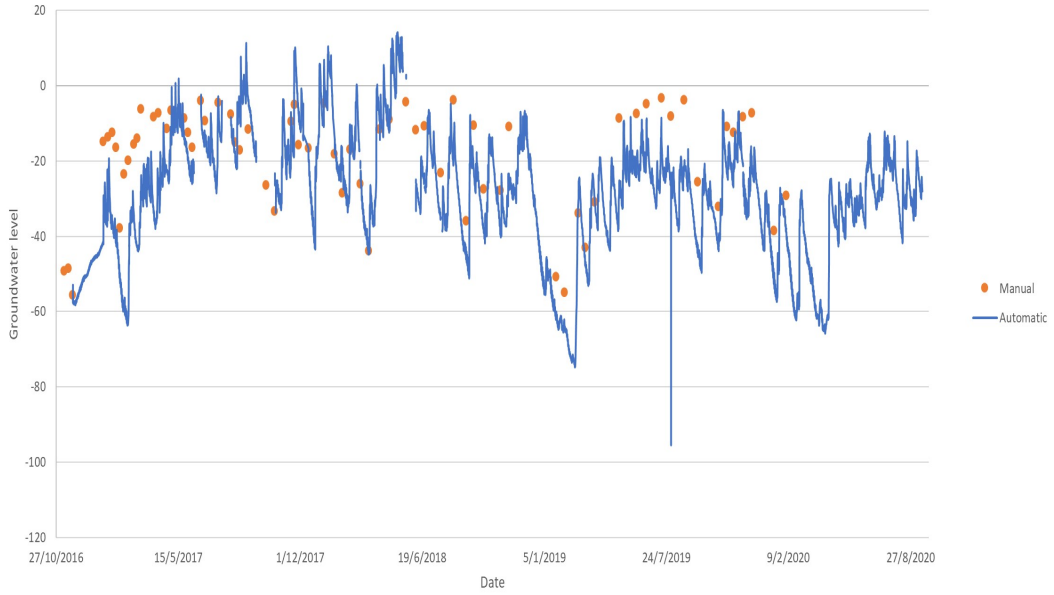


Figure 3.2: Groundwater level time series without pre-processing.

Moreover, a boxplot showing the outliers, morphology and symmetry of the unprocessed data is presented in Fig. 3.3 and a complete description in Table 3.1

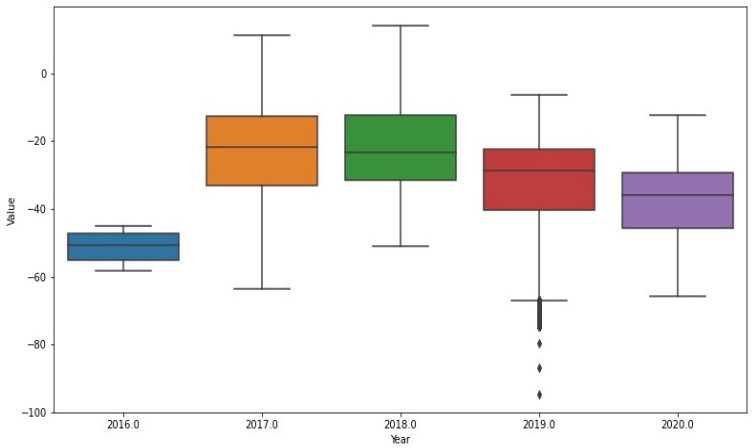


Figure 3.3: Boxplot of groundwater level time series without pre-processing.

Description	
number of samples	307709
mean	-27.58411
std	15.86747
min	-94.65000
25%	-36.57000
50%	-26.28000
75%	-17.62000
max	14.19000

Table 3.1: Statistical description of unprocessed time series.

By the beginning of 2020 the data series was collected with a new sensor, in particular INW, that aims to recollect error-free data closer to the actual values. On the other hand, the precipitation time series does not have inconsistencies and it was previously validated by FONAG technicians.

3.3 Data preparation

The beginning of the period from November 2016 to April 2017 is excluded from the study due to the high amount of inconsistencies with the real groundwater level values that have been manually taken. Thereafter, to get a high-quality data to be analyzed later, the following steps are needed.

3.3.1 Resample

After removing the significant outliers, with the aim of reducing the dimension of the data the time series is averaged in the case of groundwater level and added in precipitation, so in that way, a resample is obtained changing the frequency from 5 minutes to monthly.

3.3.2 Interpolation

To the treatment of gaps with missing values, the data is interpolated in Python, using the last value before each gap and the first one after as a reference to fill every empty period.

3.3.3 Quality Control and Data Selection

To get better approximations to the actual values of groundwater level, the precipitation time series and emptying time constant (a previous study made by FONAG) is used. Afterward, the values of the remaining time series are fitted taking the emptying time constant as a baseline for the dry days and the amount of accumulated precipitation for

the rainy days.

Making a brief data analysis the following boxplot for groundwater level processed time series per year is presented in Fig. 3.4. The boxplot shows the absence of outliers in all years except 2017, this atypical value belongs to a pre-intervention dry period, so the value will be preserved.

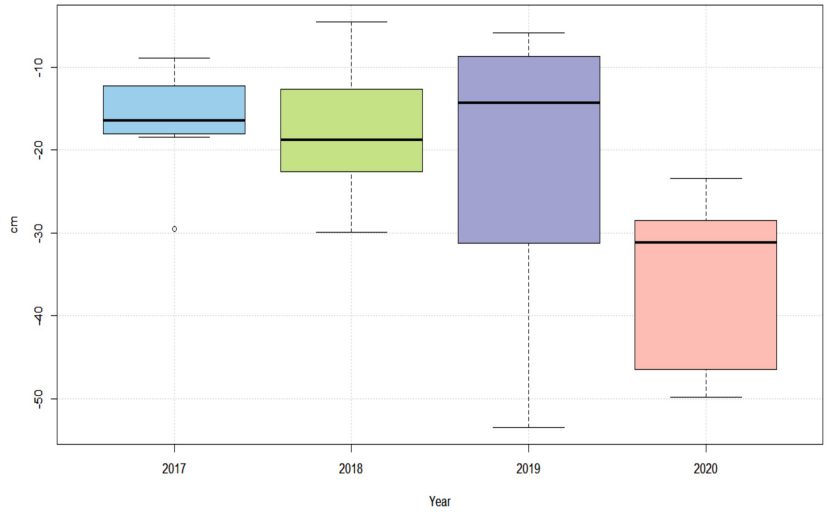


Figure 3.4: Processed time series boxplot.

Finally, with the purpose of produce the predictive models, the groundwater level time series is split into two differentiated data sets with a proportion of $70\% - 30\%$ for training and testing the models, respectively. The train data set is used to generate the models, and the test data set to validate the accuracy of the models. In this particular case, the train time series goes from May 2017 to August 2019, and the test time series goes from September 2019 to September 2020, as is presented in Fig. 3.5

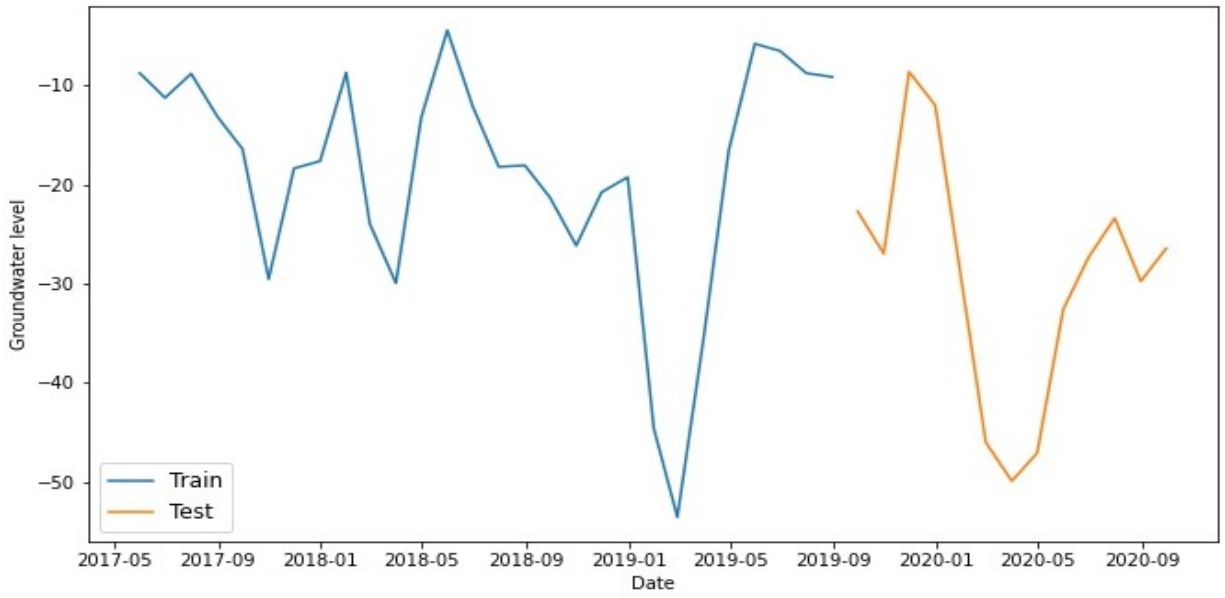


Figure 3.5: Groundwater Level time series splitted in train and test data sets.

3.4 Modeling

3.4.1 ARIMA

Following the ARIMA steps mentioned in Section 2.4 a decomposition, the autocorrelation, and partial autocorrelation should be done to check the time series characteristics, as is presented in Fig. 3.6 and Fig. 3.7.

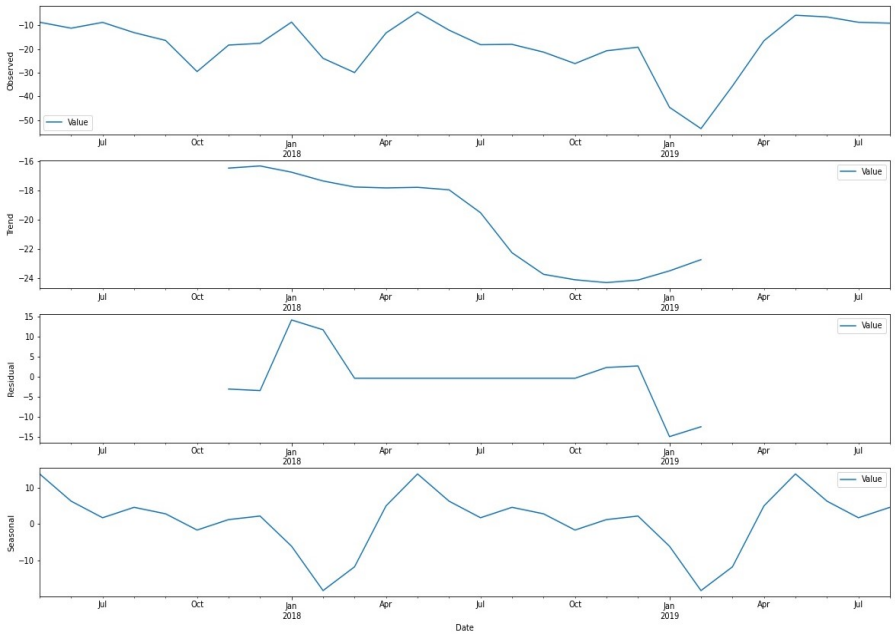
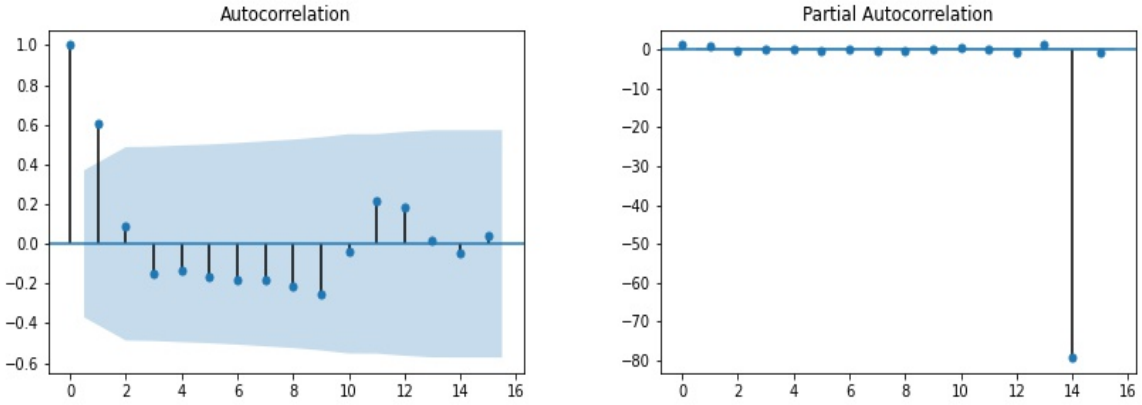


Figure 3.6: Groundwater level time series decomposition.



(a) Autocorrelation (b) Partial Autocorrelation

Figure 3.7: Autocorrelation and partial autocorrelation of groundwater level time series.

Since the condition of stationarity fails the data is transformed by carrying out a differentiation. Once a stationary time series is obtained the next step is to find the p, q, and d parameters. Even though these parameters can be set from the plots in Fig. 3.7 this method might be limiting and tricky.

Therefore, the automatic process `auto_arima` from Python `pmdarima` library is used [38]. This process gets information from the training data and picks up the parameters that best fit the model by analyzing the Akaike information criterion (AIC) and getting the ones that minimize the value, then the model is prepared on the training data by calling the `fit()` function, and finally the predictions are made by calling `predict()` function and specifying the time to be predicted [39].

3.4.2 Hidden Markov Model

HMM attempts to perform the following steps with the components discussed in Section 2.5:

1. Compute the probability of the occurrence of the observation sequence.
2. Determinate the parameters of the model, that best explains the observations.
3. Define the most probable state sequence from a given observation sequence.

The Forward-Backward, Baum-Welch, and Viterbi are algorithms used to perform the aforementioned steps, respectively. These algorithms are explained in more detail in [40].

Therefore, the `GaussianHMM` process from `hmmlearn.hmm` Python library is employed [41], which executes the required algorithms to implement HMM, with the following fixed parameters: *n_components* being the number of hidden states, *covariance_type* indicating the type of covariance matrix that the model will take, and *n_iter* specifying the number of iterations to run of the Baum-Welch algorithm, and *init_params* that controls which

element are initialized prior to training, is replaced among 's' for startprob, 't' for transmat, 'm' for means and 'c' for covars. The goal is to find the better number of hidden states, until find the better one to make the implementation. Once the training time series is fitted then the forecasted process is carried out as is stated in Section 2.5.

3.5 Evaluation

The performance of ARIMA and hidden Markov model is measured with the following metrics applied in the testing data set. The computational cost can be considered to evaluate the quality of the model. Nevertheless, is this case due to the limited number of samples this approach is discarded.

3.5.1 Root Mean Squared Error

Mean squared error (RMSE) measures the square root of the average squared difference between the forecasted values and the actual values. Is one of the most common metric to evaluate a regression model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{3.1}$$

In Eq. 3.1 the following notation is considered:

- x_i = the actual values.
- $\hat{x}_i$ = the predicted values.
- n = is the number of elements.

3.5.2 Mean Absolute Error

Mean absolute error (MAE) is the mean of the absolute value between the predicted and observed values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{3.2}$$

In Eq. 3.2 the same notation of Eq. 3.1 is employed.

3.6 Deployment

After carrying out the models and its evaluation, an strategy to deploy the project is needed. First of all the model with better accuracy is selected to apply to the other piezometers located in the wetland, thereby is expected to be able to take preventive measures if these are imperative in order to mitigate the drainage of the wetland.

Chapter 4

Experimental Setup

4.1 Datasets

For the experiments, the considered data sets are a monthly time series split into training data and test data as is detailed in Section 3.3.3 and showed in Fig. 3.5. However, it is important to emphasize that ARIMA model requires an additional condition of stationarity to be executed.

4.2 Experiment description

The mechanism of the proposed approach is explained in Section 3.4. Additionally, using the parameters defined in Table 4.3 and Table 4.4 the following experiments are set in motion.

- Experiment 1: It is done fixing the parameters from the beginning.
- Experiment 2: It is done replacing the values of the parameters until finding the better ones.

4.3 Parameter settings and methods

4.3.1 Experiment 1 parameters

The parameters to implement ARIMA using p and q obtained from ACF and PACF plots, see Fig. 3.7, are stated in Table 4.1

ARIMA	
Parameter	Value
p	1
d	0
q	14
freq	'M'

Table 4.1: ARIMA parameters for experiment 1.

For HMM the parameters are detailed in Table 4.2

HMM	
Parameter	Value
n_components	2
covariance_type	"full"
n_iter	100
init_params	'stmc'

Table 4.2: Hidden Markov model parameters for experiment 1.

The last parameter 'stmc' is a default setting, meaning that all the components explained in Section 3.4.2 are initialized prior to training.

4.3.2 Experiment 2 parameters

The parameters to carry out ARIMA are indicated in Table 4.3.

ARIMA	
Parameter	Value
start_p	1
start_q	1
max_p	10
max_q	10
m	12
d	None

Table 4.3: ARIMA parameters for experiment 2.

Where $start\_p$, $start\_q$, max_p and max_q correspond the the initial and final values of the parameters mentioned in Section 2.4. Moreover, $d = None$ will automatically take the value, and $m = 12$ indicates that the data is monthly seasonal differentiated.

Likewise, the parameters for hidden Markov model are showed in Table 4.4.

HMM	
Parameter	Value
n_components	2
covariance_type	"full"
n_iter	100
init_params	'm'

Table 4.4: Hidden Markov model parameters for experiment 2.

4.4 Performance measure

To evaluate the performance of experiment 1 and experiment 2 the metrics stated in Section 3.5 are used.

Chapter 5

Results and Discussion

The results of each experiment mentioned in Section 4.2 are presented in figures and tables in order to get better visualization.

5.1 Experiments

The study consists of two experiments discussed in the following subsections.

5.1.1 Experiment 1

The experiment is developed with determined parameters from the begging of the implementation in the case of ARIMA model the parameters are chosen according to the Figure 3.7. On the other hand, for HMM the parameter *init_params* is set to default, with 2 hidden states.

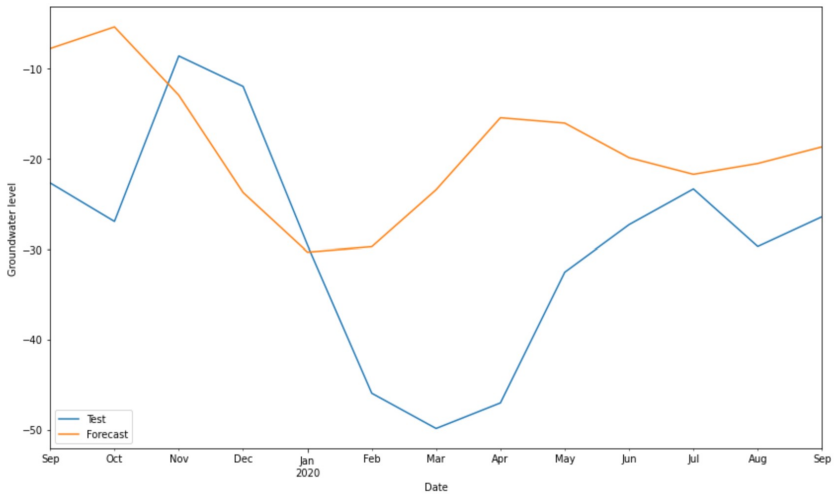


Figure 5.1: Plot of ARIMA(1,0,14) test vs. forecast data sets.

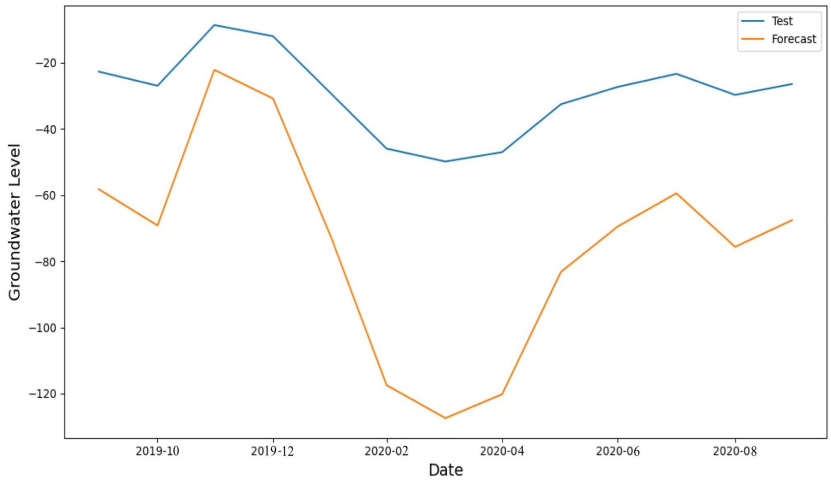


Figure 5.2: Plot of HMM with default settings test vs. forecast data sets.

The results in the selected metrics are detailed in Table 5.1.

Errors \ Model	ARIMA	HMM
MAE	13.072672	45.539267
RMSE	15.868396	49,160797

Table 5.1: MAE and RMSE results of experiment 1.

As is observed in the metrics results in Table 5.1, the forecasted accuracy is unsatisfactory this might happen because of the parameters enunciated in Table 4.1 and Table 4.2 are not appropriate to produce good results with the type of time series to be forecasted, for this reason the AIC evaluation is required to fine-tune the models until finding the best value parameters.

In ARIMA model predictions, see Fig. 5.1, the peaks in October 2019, March, and July 2020 are not well approximated. Moreover, the forecasted time series presents a short groundwater level recession from October 2019 to January 2020 which clearly differs from the recession period from December 2019 to March 2020 of the observed time series. Therefore, the model with these parameters can be discarded for future applications because of the inconsistencies when predicting the groundwater level recession.

On the other hand, HMM is able to predict the months when the peaks occur, but not entirely accurate. As shown in Fig. 5.2 the minimum and maximum values of these pikes do not coincide. However, besides these failures the recession forecasted period concurs with the observed except the lapse from March to April 2020, in which the real values keep decreasing but in the predicted increase.

5.1.2 Experiment 2

The experiment is developed following the methodology explained in Section 3.4, and setting the parameters specified in Table 4.3 and Table 4.4.

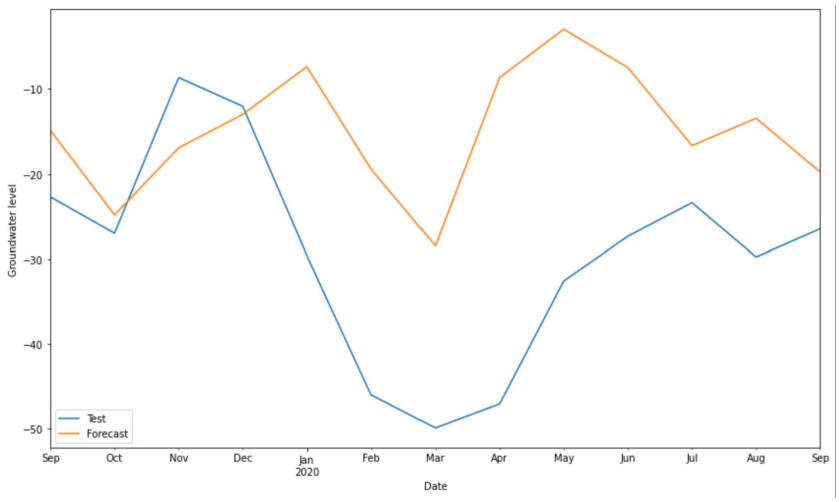


Figure 5.3: Plot of ARIMA(1,0,2) test vs forecast data sets.

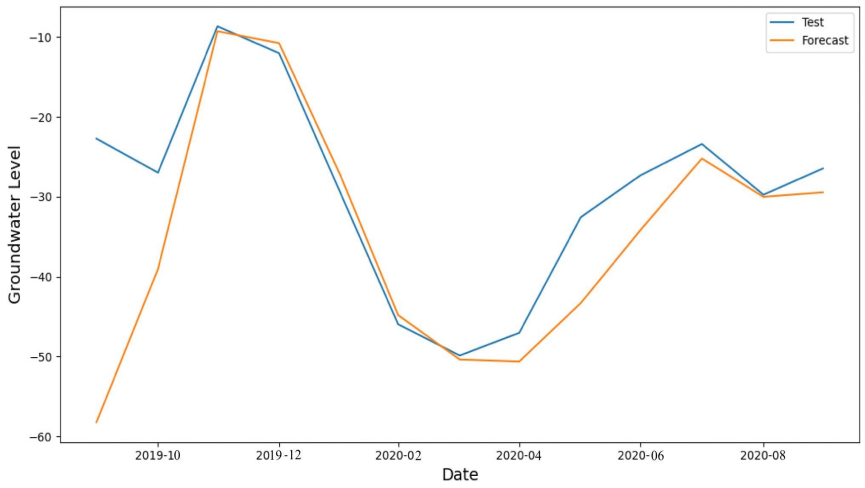


Figure 5.4: Plot of HMM test vs forecast data sets.

The results in the selected metrics are detailed in Table 5.2.

Errors \ Model	ARIMA	HMM
MAE	15.926025	6.310349
RMSE	19.406509	11.133436

Table 5.2: MAE and RMSE results of experiment 2.

An evident difference between the error results in the execution of both models is seen in Table 5.2. Meanwhile, in ARIMA model the errors between the observed and forecasted time series remain high, in HMM are reduced significantly. This outcome was obtained by letting HMM initialize its development with the mean parameter prior to training, and with 2 hidden states representing whether in that month, a considerable amount of precipitation is detected or not.

ARIMA attempts to predict the behavior of the real time series values, but failed principally due to the characteristics and the quality of the data. For this reason, the recession period according to the predicted values starts one month later in comparison with the actual decline of groundwater level and does not reach the minimum value that the series met, see Fig. 5.3. On the opposite, HMM get the result illustrated in Fig. 5.4 since it is insensitive with raw data. Therefore, there is no doubt that this model fits better to the type of this time series in specific, and can be used to successfully forecast future values that will serve as a baseline for future analysis.

Chapter 6

Conclusions and future work

6.1 Conclusion

Time-series forecasting of groundwater level can be used as a practical tool to enhance the management of water resources, in which groundwater plays a fundamental role. In this work, after a deep prior analysis of related works, we present two time series forecasting methods with the aim to conduct a comparative study between them, namely ARIMA, and hidden Markov model.

The raw data obtained from FONAG was previously treated before the implementation of both models, in that way a closer approximation of the series through the monitoring time in Pugllohuma wetland was acquired. After that, a selection of the best model parameters was done to develop the models to carry out the comparative study using evaluation metrics, particularly mean absolute error and root mean squared error.

Owing to the poor quality of the initial data set, and taking into account the instrumental errors in the device that capture the groundwater level information through time, the results in ARIMA model were severely affected, leading to substandard forecasting. Inversely, since HMM is insensitive to the raw data characteristics, the performance was not really affected when the appropriate parameters were selected. Finally, after analyzing all the results hidden Markov model is selected as the more suitable for this kind of time series.

Nevertheless, two other forecasting techniques were evaluated as alternative methods to accomplish the proposed goal. As shown in A.1 and A.2, better forecasting was obtained without the need for resampling the daily time-series into monthly. Therefore it will interesting to discover if, with this kind of time series, long-short term memory and artificial neural network models improve the performance of the system when predicting future values. On the other hand, it is expected that with a better data quality more enhanced results are obtained, as is shown in A.3 and A.4 where the experiments were carried out with the air quality London monthly average of Mean Roadside: Nitrogen Dioxide data which was acquired from [42], and more accurate forecast were achieved especially with HMM with default settings.

6.2 Future work

The application area of time series forecasting is extensive, and as has been demonstrated there exist a lot of methods that can also be used in the hydrology field. Hereunder, some different approaches that have not been explored in this study are stated.

- Tackle the weaknesses of ARIMA model presented in this study, and get better results with a combination of another model that can cope with these issues, making in that way a new proposed hybrid model for time series forecasting
- Generalize the study with different kinds of time series, luckily a time series with a more appropriate initial quality.
- Apply the improved generalized model to the data of other piezometers located in Pugllohuma wetland.
- Implement multivariate time series analysis and forecasting, to understand and evaluate the complete dynamic of the wetland, taking into account all the variables influencing it.

Bibliography

[1] "Time series analysis: The basics," https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics#:~:text=An20observed20time20series20can,unsystematic2C20short20term20fluctuations)., accessed: 2021-03-18.

[2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[3] "A guide for time series prediction using recurrent neural networks (lstms)," https://blog.statsbot.co/time-series-prediction-using-recurrent-neural-networks-lstms-807fa6ca7f, accessed: 2021-03-20.

[4] Q. Yan and C. Ma, "Application of integrated arima and rbf network for groundwater level forecasting," *Environmental Earth Sciences*, vol. 75, no. 5, p. 396, 2016.

[5] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006, twenty five years of forecasting. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207006000021

[6] G. Patle, D. Singh, A. Sarangi, A. Rai, M. Khanna, and R. Sahoo, "Time series analysis of groundwater levels and projection of future trend," *Journal of the Geological Society of India*, vol. 85, no. 2, pp. 232–242, 2015.

[7] M. R. Hamdi, A. N. Bdour, and Z. S. Tarawneh, "Developing reference crop evapotranspiration time series simulation model using class a pan: a case study for the jordan valley/jordan," *Jordan. Journal of Earth and Environmental Sciences*, vol. 1, no. 1, pp. 33–44, 2008.

[8] M. Karamouz and B. Zahraie, "Seasonal streamflow forecasting using snow budget and el niño-southern oscillation climate signals: Application to the salt river basin in arizona," *Journal of Hydrologic Engineering*, vol. 9, no. 6, pp. 523–533, 2004.

[9] S. A. Shamsnia, N. Shahidi, A. Liaghat, A. Sarraf, and S. F. Vahdat, "Modeling of weather parameters using stochastic methods (arima model)(case study: Abadeh region, iran)," in *International conference on environment and industrial innovation*, vol. 12, no. 1, 2011, pp. 282–285.

[10] R. Samsudin, P. Saad, and A. Shabri, "River flow time series using least squares support vector machines," *Hydrology and Earth System Sciences*, vol. 15, no. 6, pp. 1835–1852, 2011.

[11] D. K. Panda and A. Kumar, "Evaluation of an over-used coastal aquifer (orissa, india) using statistical approaches," *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, vol. 56, no. 3, pp. 486–497, 2011.

[12] M. Khadr, "Forecasting of meteorological drought using hidden markov model (case study: The upper blue nile river basin, ethiopia)," *Ain Shams Engineering Journal*, vol. 7, no. 1, pp. 47–56, 2016.

[13] S. Chen, J. Y. Shin, and T.-W. Kim, "Probabilistic forecasting of drought: a hidden markov model aggregated with the rcp 8.5 precipitation projection," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 5, pp. 1061–1076, 2017.

[14] W. Zucchini and P. Guttorp, "A hidden markov model for space-time precipitation," *Water Resources Research*, vol. 27, no. 8, pp. 1917–1923, 1991.

[15] Y. Chebud and A. Melesse, "Operational prediction of groundwater fluctuation in south florida using sequence based markovian stochastic model," *Water resources management*, vol. 25, no. 9, pp. 2279–2294, 2011.

[16] D. W. Chambers, J. A. Baglivo, J. E. Ebel, and A. L. Kafka, "Earthquake forecasting using hidden markov models," *Pure and applied geophysics*, vol. 169, no. 4, pp. 625–639, 2012.

[17] T. Hokimoto and K. Shimizu, "A non-homogeneous hidden markov model for predicting the distribution of sea surface elevation," *Journal of applied statistics*, vol. 41, no. 2, pp. 294–319, 2014.

[18] J. Adamowski and H. F. Chan, "A wavelet neural network conjunction model for groundwater level forecasting," *Journal of Hydrology*, vol. 407, no. 1-4, pp. 28–40, 2011.

[19] J. D. Cryer and K.-S. Chan, *Time series analysis: with applications in R*. Springer Science & Business Media, 2008.

[20] C. Chatfield, *Time-series forecasting*. CRC press, 2000.

[21] W. W. Wei, "Time series analysis," in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2006.

[22] R. S. Tsay, *Analysis of financial time series*. John wiley & sons, 2005, vol. 543.

[23] "An overview of time series forecasting models part 1: Classical time series forecasting models," https://medium.com/@shaileydash/an-overview-of-time-series-forecasting-models-part-1-classical-time-series-forecasting/models-2d877de76e0f, accessed: 2021-03-19.

[24] J. E. Hanke and D. W. Wichern, "Business forecasting, eight ed," 2005.

[25] E. Ostertagova and O. Ostertag, "Forecasting using simple exponential smoothing method," *Acta Electrotechnica et Informatica*, vol. 12, no. 3, p. 62, 2012.

[26] S. Palani, S.-Y. Liong, and P. Tkalich, "An ann application for water quality forecasting," *Marine pollution bulletin*, vol. 56, no. 9, pp. 1586–1597, 2008.

[27] A. Tealab, "Time series forecasting using artificial neural networks methodologies: A systematic review," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 334–340, 2018.

[28] F. A. Gers, D. Eck, and J. Schmidhuber, "Applying lstm to time series predictable through time-window approaches," in *Neural Nets WIRN Vietri-01*. Springer, 2002, pp. 193–200.

[29] K.-j. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307–319, 2003.

[30] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & operations research*, vol. 32, no. 10, pp. 2513–2522, 2005.

[31] C. Zhao, H. Zhang, X. Zhang, M. Liu, Z. Hu, and B. Fan, "Application of support vector machine (svm) for prediction toxic activity of different data sets," *Toxicology*, vol. 217, no. 2-3, pp. 105–119, 2006.

[32] U. Thissen, R. Van Brakel, A. De Weijer, W. Melssen, and L. Buydens, "Using support vector machines for time series prediction," *Chemometrics and intelligent laboratory systems*, vol. 69, no. 1-2, pp. 35–49, 2003.

[33] T. B. Trafalis and H. Ince, "Support vector machine for regression and applications to financial forecasting," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 6. IEEE, 2000, pp. 348–353.

[34] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[35] V. Kotu and B. Deshpande, "Chapter 12 - time series forecasting," in *Data Science (Second Edition)*, second edition ed., V. Kotu and B. Deshpande, Eds. Morgan Kaufmann, 2019, pp. 395–445. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128147610000125

[36] B.-J. Yoon, "Hidden markov models and their applications in biological sequence analysis," *Current genomics*, vol. 10, no. 6, pp. 402–415, 2009.

[37] M. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: a new approach," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. IEEE, 2005, pp. 192–196.

[38] T. G. Smith *et al.*, "pmdarima: Arima estimators for Python," 2017–, [Online; accessed ¡today¿]. [Online]. Available: http://www.alkaline-ml.com/pmdarima

[39] "Forecasting with arima using python," https://levelup.gitconnected.com/simple-forecasting-with-auto-arima-python-a3f651271965, accessed: 2021-02-04.

[40] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[41] R. W. S. D. *et al.*, "hmmlearn," https://github.com/hmmlearn/hmmlearn, 2018.

[42] "London average air quality levels," https://data.london.gov.uk/dataset/london-average-air-quality-levels, accessed: 2021-04-20.

Appendices

Appendix A

Alternative methods

A.1 LSTM

A.1.1 Architecture of the proposed ANN

LSTM Summary		
Layer	Input	Output
Input	(None, None)	(None, None)
Lambda	(None, None)	(None, None, 1)
Bidirectional(LSTM)	(None, None, 1)	(None, None, 64)
Bidirectional_1 (LSTM)	(None, None, 64)	(None, 64)
Dense	(None, 64)	(None, 1)
Lambda_1	(None, 1)	(None, 1)

Table A.1: Architecture of the alternative method LSTM.

A.1.2 Results

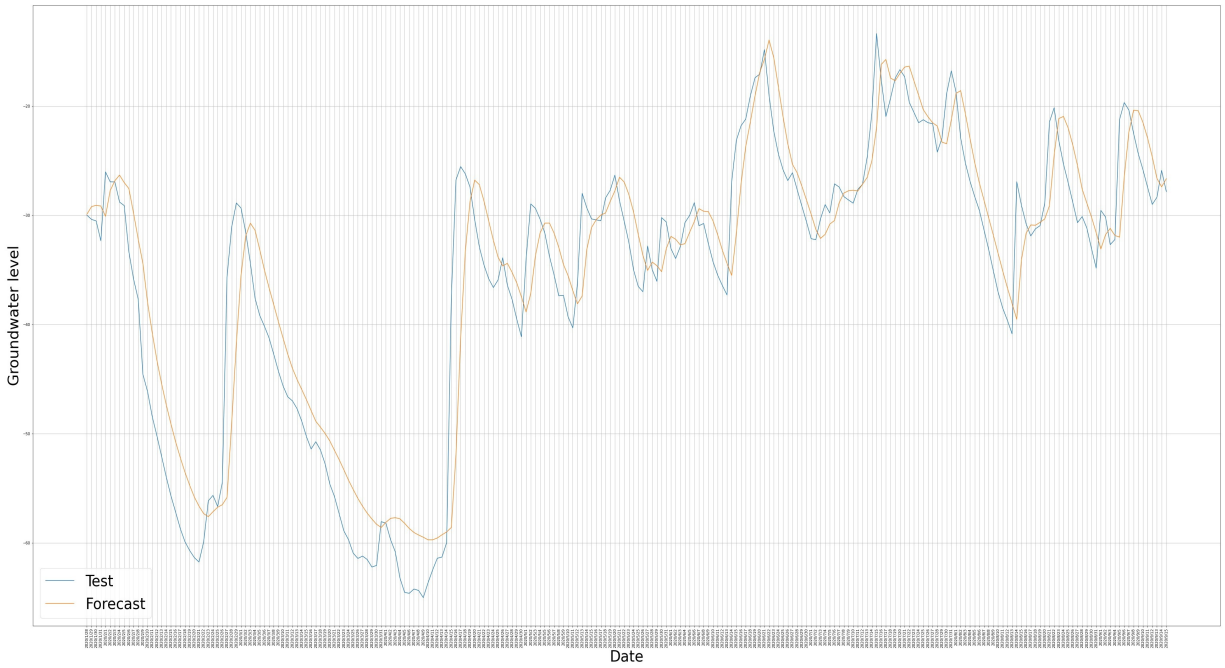
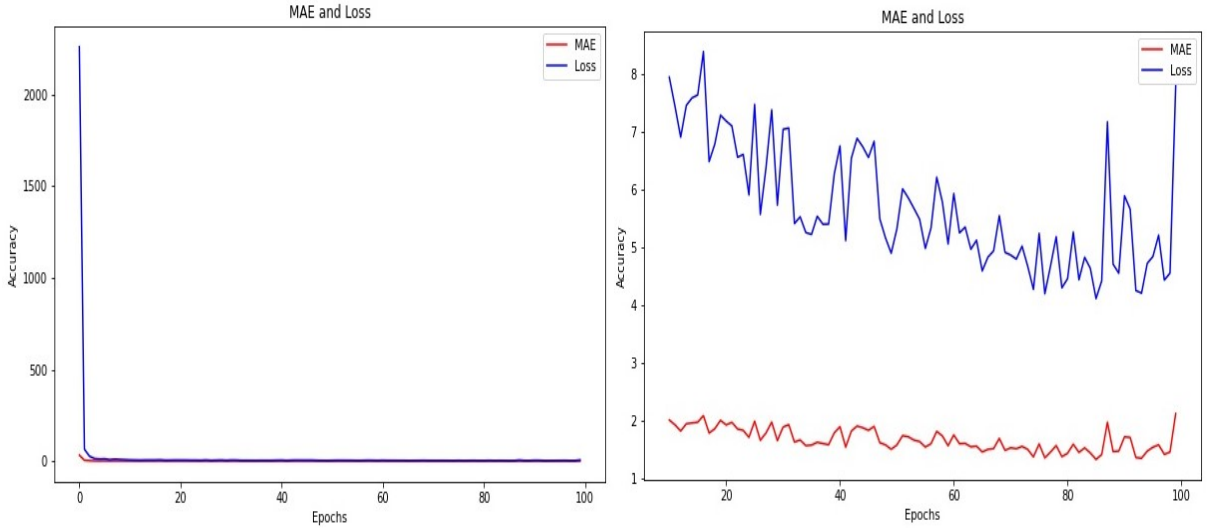


Figure A.1: Plot of LSTM test vs forecast data sets.



(a) Plot MAE vs Loss

(b) Zoomed plot MAE vs Loss

Figure A.2: Mean absolute error vs loss in LSTM method.

Errors \ Model	LSTM
MAE	3.750401
RMSE	5.048310

Table A.2: MAE and RMSE results of LSTM.

A.2 ANN

A.2.1 Architecture of the proposed ANN

ANN Summary		
Layer	Input	Output
Input	(None, 30)	(None, 30)
Dense_3	(None, 30)	(None, 10)
Dense_4	(None, 10)	(None, 10)
Dense_5	(None, 10)	(None, 1)

Table A.3: Architecture of the alternative method ANN.

A.2.2 Results

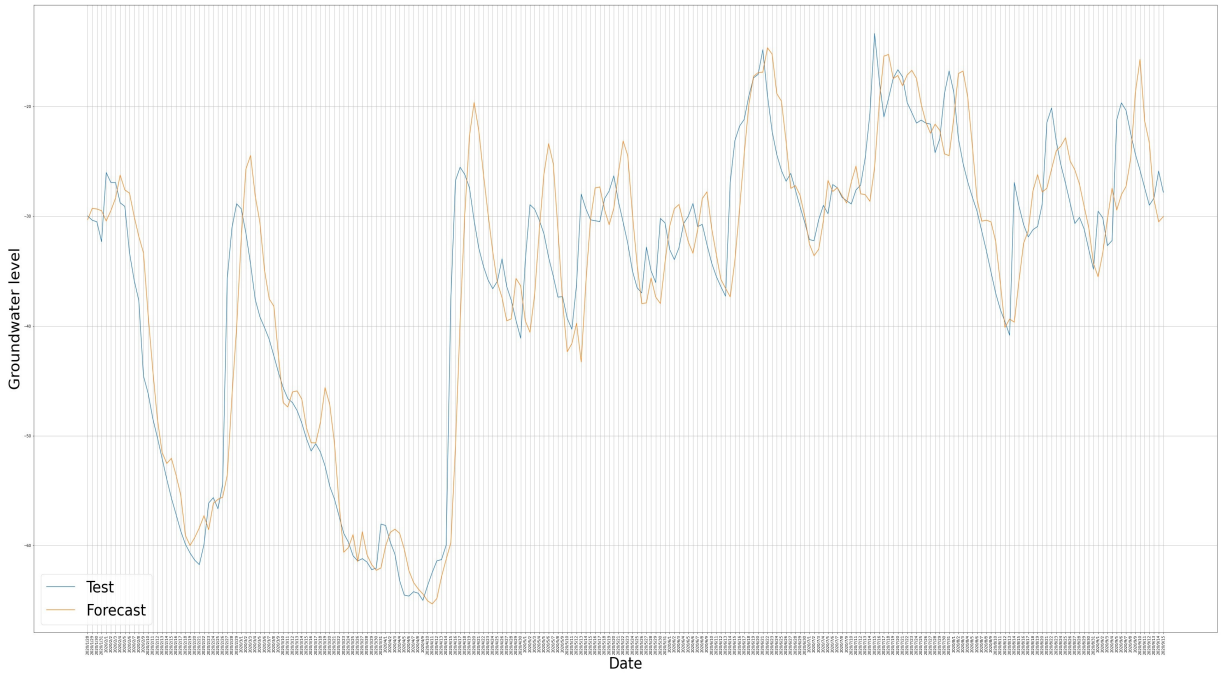
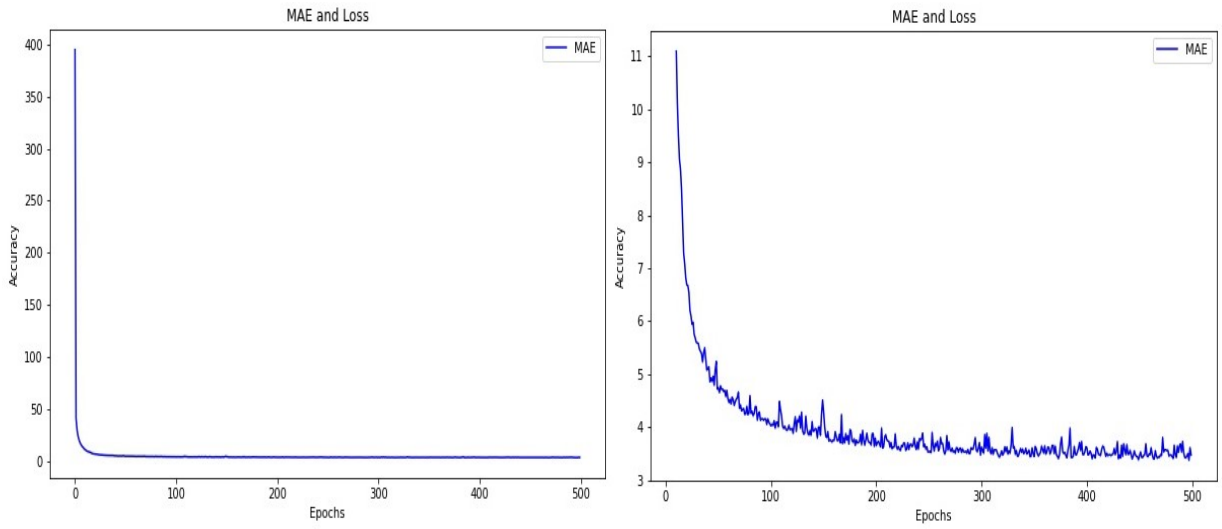


Figure A.3: Plot of ANN test vs forecast data sets.



(a) Plot MAE vs Loss (b) Zoomed plot MAE vs Loss

Figure A.4: Mean absolute error vs loss in ANN method.

Errors \ Model	ANN
MAE	3.855601
RMSE	5.377836

Table A.4: MAE and RMSE results of ANN.

A.3 Experiment 1 with London average air quality levels data

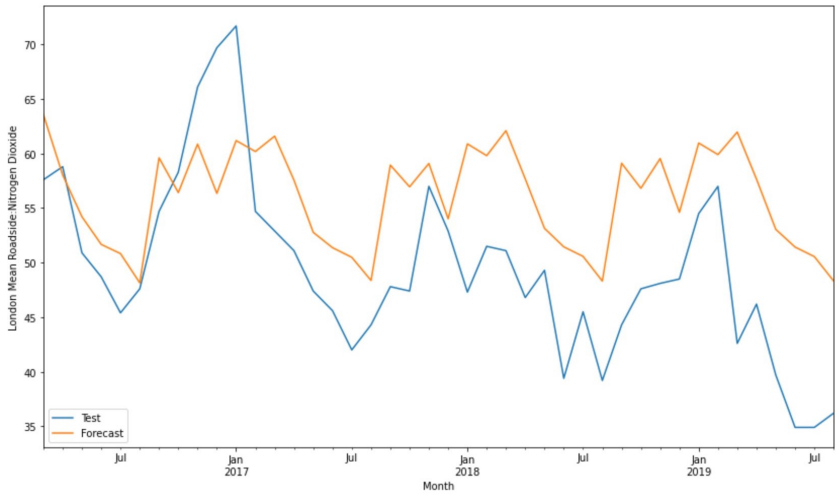


Figure A.5: Plot of ARIMA(0,0,0) test vs. forecast data sets.

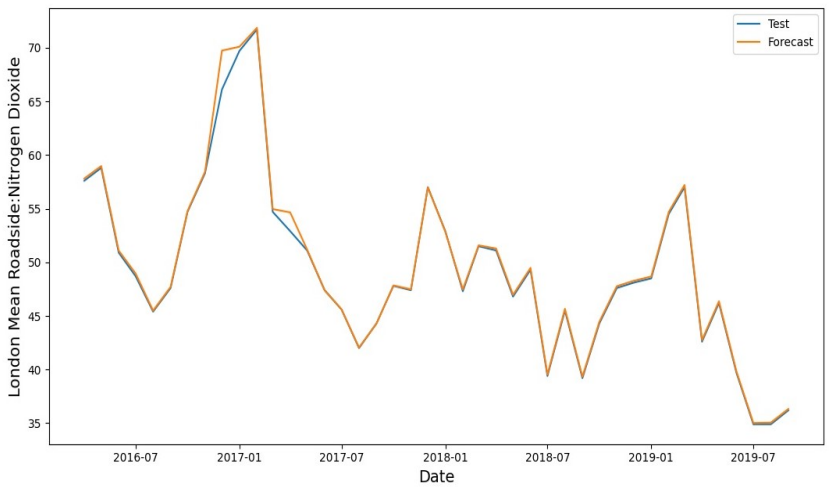


Figure A.6: Plot of HMM with default settings test vs. forecast data sets.

Errors \ Model	ARIMA	HMM
MAE	8.011653	0.259872
RMSE	9.238388	0.639629

Table A.5: MAE and RMSE results of experiment 1 with London average air quality levels data.

A.4 Experiment 2 with London average air quality levels data

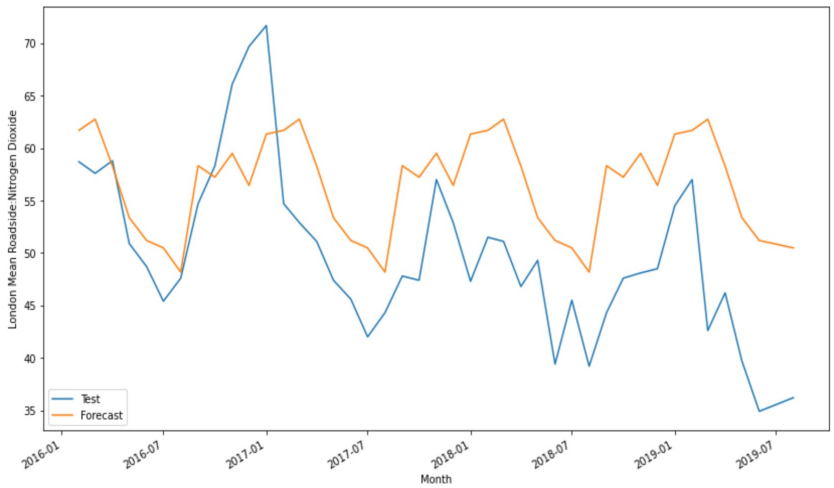


Figure A.7: Plot of AUTOARIMA test vs. forecast data sets.

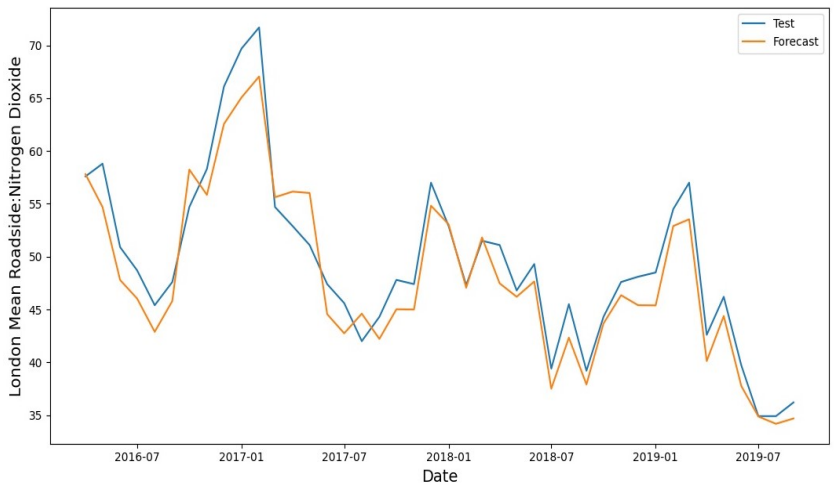


Figure A.8: Plot of HMM test vs. forecast data sets.

Errors \ Model	ARIMA	HMM
MAE	8.020230	2.685155
RMSE	9.238034	3.433010

Table A.6: MAE and RMSE results of experiment 2 with London average air quality levels data.