SARIMA short to medium-term forecasting and stochastic simulation of streamflow, water levels and sediments time series from the HYDAT database

Adnane Stitou

Thesis Submitted under the supervision of

Dr. Ousmane Seidou

To the Faculty of Graduate and Postdoctoral Studies

In Partial Fulfilment of the Requirements for the Degree of

Master of Applied Science in Civil Engineering

Under the auspices of the Ottawa-Carleton Institute for Civil Engineering

uOttawa

Civil Engineering, Department of

University of Ottawa

Ottawa, Ontario, Canada

Abstract

This study aims to investigate short-to-medium forecasting and simulation of streamflow, water levels, and sediments in Canada using Seasonal Autoregressive Integrated Moving Average (SARIMA) time series models. The methodology can account for linear trends in the time series that may result from climate and environmental changes. A *Universal Canadian forecast Application* using python web interface was developed to generate short-term forecasts using SARIMA. The Akaike information criteria was used as performance criteria for generating efficient SARIMA models. The developed models were validated by analyzing the residuals. Several stations from the Canadian Hydrometric Database (HYDAT) displaying a linear upward or downward trend were identified to validate the methodology. Trends were detected using the Man-Kendall test.

The Nash-Sutcliffe efficiency coefficients (Nash ad Sutcliffe, 1970) of the developed models indicate that they are acceptable. The models can be used for short term (1 to 7 days) and medium-term (7 days to six months) forecasting of streamflow, water levels and sediments at all Canadian hydrometric stations. Such a forecast can be used for water resources management and help mitigate the effects of floods and droughts. The models can also be used to generate long time-series that can be used to test the performance of water resources systems.

Finally, we have automated the process of analysis, model-building and forecasting streamflow, water levels, and sediments by building a python-based application easily extendable and user-friendly. Therefore, automating the SARIMA calibration and forecasting process for all Canadian stations for the HYDAT database will prove to be a very useful tool for decision-makers and other entities in the field of hydrological study.

Acknowledgments

The realization of this thesis has been possible through the moral support and advice of:

- **My parents and sisters:** A big thank you to them for their encouragement, help, and continuous moral and financial support. Without their relentless spirit and dedication, none of this would be possible.

- **Thesis supervisor:** I would like to express my sincerest gratitude to Dr. Ousmane Seidou, Professor in the Department of Civil Engineering at the University of Ottawa, not only for his advice and support but also for believing in me as a candidate in the M.A.Sc program.

- **Thesis committee:** I would like to deeply thank my committee members, Dr. Ioan Nistor, Professor in the Department of Civil Engineering at the University of Ottawa and Dr. Abdolmajid Mohammadian, Professor in the Department of Civil Engineering at the University of Ottawa for their guidance, advice, and corrections.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

ACF	Autocorrelation Function
AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
ARMA	Autoregressive moving average
ARIMA	Autoregressive integrated moving average
^{o}C	Degree Celsius
DEM	Digital Elevation Model
GIS	Geographic Information System
HYDAT	Hydro climatological Data Retrieval Program
H_0	Null hypothesis
H_a	Alternative hypothesis
Km	Kilometer
Km^2	Kilometer square
MK	Mann-Kendall Trend Test
m/s	Meter per second
mm	Millimeter
mg/L	Milligrams per liter
m3/s	Cubic meters per second
l/s	Litres/second
PACF	Partial Autocorrelation Function
RMSE	Root mean square error
RHBN	Reference Hydrometric Basin Network
SARIMA	Seasonal Autoregressive integrated moving average
SQL	Structured Query Language
SWAT	Soil Water Assessment Tool
TDR	Tile drain rate
t/d	Tonnes per day
WHO	World Health Organisation
σ	Sigma
Φ	Phi
Θ	Theta
ε	Epsilon

Chapter 1 – Introduction

Water is a self-renewing and constant resource as it follows a never-ending cycle of precipitation, evaporation, and sublimation. It is one of the most abundant and consistent resources on the planet, but less than one percent of the earth's water is accessible freshwater (University of Wisconsin, 2013). Water is often a resource that is in high demand because of the growing populations in developing countries. In 2011, the world health organization (WHO) proclaimed that 'safe' water for human consumption is a basic human right as it is vital for health and hygiene.

Recent changes in the world's climate have made water accessibility and availability a critical obstacle towards human development and ecosystem preservation. According to Whitfield (2012), the warming climate would accelerate the hydrological cycle in future years. The warming of the earth is evident today from the various observations of NASA, the US EPA, or Environment Canada. These include an increase in global average air and sea level. These changes generate significant challenges to those in charge of assessing water resources systems and planning for future environmental impacts. Water bodies must be monitored, and their hydrological characteristics must be forecasted to prevent disasters. In other words, managing the quantity, quality, and accessibility to water resources is essential to the survival and well-being of humans, the environment, the economy, and political stability. This chapter introduces the hydrological system in the scope of Canadian water resources, the purpose, and context of the thesis, its objectives, and layout.

1.1 Purpose of the study

The Government of Canada has indicated decades ago that water is the "most undervalued and neglected natural resource" and "in no part of Canada is freshwater of sufficient quality and quantity that it can continue to be overused and abused" as it is "a scarce commodity with real value that must be managed accordingly" (The Federal Water Policy, 1993). Water management is a complex issue, especially in Canada, as it involves both provincial and federal jurisdictions. Indeed, provinces manage the water resource within their territorial boundaries, while the federal government has jurisdiction over international water agreements (Saunders, and Wenig, 2006).

Canadians often believe that water resources in Canada are well managed (Bakker, 2007), mislead by the abundance of fresh water available in the country (McFarlane and Nilsen, 2003). According to Statistics Canada (2017), Canadians are the second-highest users of freshwater per capita, after the United States. Indeed, Canada has one of the largest freshwater reserves in the world, holding 20% of the world's total freshwater located in lakes, rivers, streams, wetlands aquifers, mountain snow and ice pack (Environment Canada, 2012). Despite that fact, freshwater quality and availability are declining in some areas in Canada. It has been established that every Canadian uses about 250 liters of treated water per day when it comes to residential use (Environment Canada, 2014). About 90% of consumable water is processed from the surface (Statistics Canada, 2011).

Water abundance can be either a blessing or a problem depending on the way it's managed and monitored. Indeed, climate change intensifies the water cycle, which increases flooding (Milly, Wetherald, and Delworth, 2002; Hirabayashi et al., 2013). Flooding is a common natural hazard (Buttle et al., 2016) that can threaten human life and is mainly caused by change in

hydrometeorological conditions (e.g. excessive rain, snow melting) in Canada (Pietroniro et al., 2004). Preserving the effective control of flood represents a significant challenge for government and local authorities (Chiang, Willems, and Berlamont, 2010).

According to Environment and Climate Change Canada (2013), the impact of flood damage and repercussions on humans such as death can be minimized with effective warning systems and heavy rain forecasts. In Canada, several forecasting centers have been established to monitor conditions that cause flooding. When there is a probability of flooding, forecasters are responsible for contacting dam operators, municipal officials, emergency personnel, and media. Among flood protection methods is the construction of dikes, dams, and diversion channels, which can be expensive, time-consuming and do not always provide full safety (Environment and Climate Change Canada, 2013). The flooding events in Montreal, Ottawa, and elsewhere across the St-Laurent River reminded everyone of the importance to plan for such catastrophic events (Statistics Canada, 2017). Construction near rivers and other natural bodies will demand further investigation and a significant focus safety.

The behavior of Canadian water bodies is further exacerbated by changes in the climate. Climate change is a critical factor in our era. The scientific community agrees that "scientific evidence for warming of the climate system is unequivocal" (Intergovernmental Panel on Climate Change, 2001). Observations of increases in global average air and ocean temperatures, melting of snow and ice, and the rising global average sea level data (Figure 11) observed is undeniable. It is, therefore, critical to monitor water bodies movements in time and space. Such observations can be done either on a daily, weekly, or yearly basis. Observations taken as such are known as *time*

series. Environment Canada is the government agency responsible for monitoring and analyzing this data. Water level, flow, and other characteristics such as sediment transport are constantly monitored in North-America to provide useful historical data showing trends over long periods, and that can help forecast and estimate future changes.

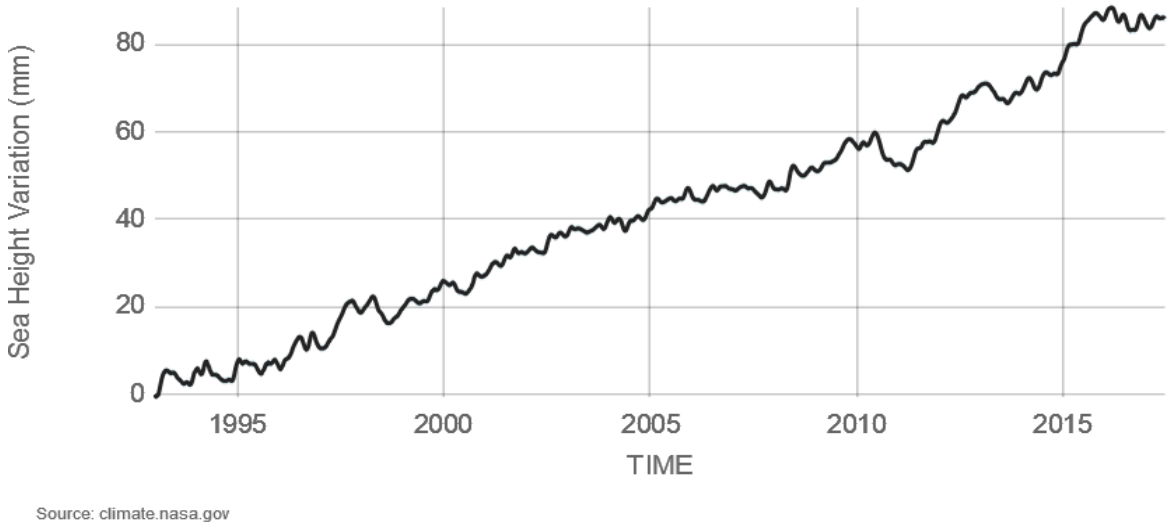


Figure 1.Satellite Data (1993-2017) For Sea Height Variation (NASA, 2017)

Considering the previous facts, it becomes evident that the study of water forecasting is much needed. Assuming that past meteorological conditions are representative of the future, mathematical models can be built to plan and simulate future conditions regarding water quality, quantity, and socio-economic data. As such, operational models are very useful in forecasting water flows, levels, quality, and other variables over time (Environment and Climate Change Canada, 2013). More specifically, a forecast is a planning tool used to predict the uncertain future using past and present data (Abraham, and Ledolter, 2005).

This study is one of the first Canadian studies to empirically investigate forecasts of streamflow, water levels, and sediments using SARIMA modeling over such a large of historical data. Indeed, no previous study has examined water forecasting in Canada to develop a python web interface

realized for all HYDAT stations across Canada: *Universal Canadian forecast Application* (Stitou, Adnane, 2018). Findings from our work will not only contribute to having a better view of all HYDAT stations across Canada with their respective information but also facilitate environmental policies for decision-makers and concerned entities (e.g. engineers, biologists, hydrologists…).

1.2 Objectives

The overall goal of this thesis is to develop Seasonal Autoregressive Integrated Moving Average (SARIMA) models that can generate short to medium-term forecasts and long-term simulations of streamflow, water levels and sediments for specific stations in Canada. This thesis has the following specific objectives:

- Demonstrate the usefulness of SARIMA models for forecasting and long-term simulations
- Extract appropriate stations from the HYDAT database: Naturally Regulated, with Flow data, Level data, and sediment data, with both Water Flow and Water Level data recorded for long periods
- Find the best parameters p, n and q for the Box-Jenkins method, and Generate SARIMA predictions for short and mid future periods and validate them against the observed data
- Create an application/interface to automate the process

1.3 Thesis Layout

The present thesis is divided into six chapters:

- **Chapter 1 - Introduction:** This chapter covers the hydrological system, Canadian water resources, the purpose and context of this thesis, the objectives, and the thesis layout.
- **Chapter 2 - Literature review:** This chapter offers background on time series analysis concepts and the SARIMA method and its properties

- **Chapter 3 - Methodology:** This chapter covers the approach used to generate statistical factors and forecast the historical environmental data
- **Chapter 4 - Study Cases:** This chapter covers the validation of the SARIMA approach by comparing the effectiveness of the results over the various data stations in Canada.
- **Chapter 5 – *Universal Canadian forecast Application*** (Stitou, Adnane, 2018©): Showcases the *SARIMA historical data forecast application (Universal Canadian forecast Application* (Stitou, Adnane, 2018©*)* and its possible use as a universal tool.
- **Chapter 6 - Conclusion:** Summary of the main conclusions generated by this thesis and outlook on future work

1.4 Scope/Limitations

This thesis objective is to produce reliable forecasts for short to medium-term periods. It does not take into account precipitation data. While alternative water levels and streamflow forecasting models are available (non-linear models, neural networks, etc.), only SARIMA models are considered in this thesis.

1.5 Novelty and knowledge contribution

Currently, there is no software able to forecast water parameters using all Canadian hydrometric stations available to the public. This thesis is one of the first studies contributing to the area of information technology combined with civil engineering applied to water forecast across all Canada. The findings from our work contribute not only to have better usage of the information of all HYDAT stations across Canada, but also facilitates decision-making for concerned entities (e.g. engineers, biologists, hydrologists…) regarding environmental policies.

Chapter 2 – Literature review

In this chapter, we present the Canadian water resources system in detail, along with the challenges associated with them. We define time series, present some of their characteristics such as trend, cyclical, seasonal and irregular components, and explain why they are relevant to statistical analysis. We also present the concepts of deterministic time series, stochastic process, stationarity, and linearity in time series. Moreover, a few models of time series are presented. Finally, we explain the application of time series to this thesis.

2.1 The hydrological cycle

A hydrological cycle represents the continuous water movement between earth and atmosphere, for instance, the evaporation of water from the surface, sublimation of ice or snow, or again plants transpiration (Figure 1). Clouds form when water vapor goes up in the air to cool and condenses (Statistic Canada, 2017). Precipitation can only happen when there is a sufficient abundance of water droplets heavy enough in the cloud to fall on the ground; in other words, allowing it to rain, snow, or hail. Water is provided to the ecosystems in several manners, such as in dew and fog (Statistic Canada, 2017). It can also be found in glaciers or snowpacks, not to mention that it can run into streams, rivers, and lakes, penetrate soils to create soil moisture or again reach the water table by traveling through the soil, to turn into groundwater (Statistic Canada, 2017). Groundwater flows underground before passing through several components (wells, springs, seepage into streams, rivers, lakes, and oceans) before getting discharged to the surface water ecosystem.

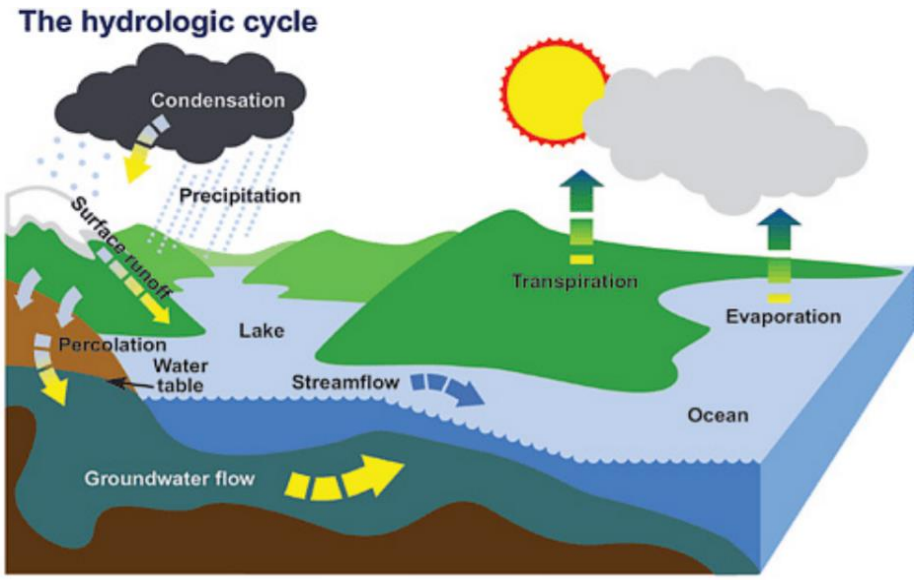


Figure 2. Hydrological cycle (Statistics Canada, 2017)

2.2 Canadian water resources

Canada has one of the largest reserves of freshwater in the world, holding 20% of the total world freshwater located in lakes, rivers, streams, wetlands aquifers, mountain snow, and ice pack and only 7% of the world's renewable freshwater resources (Environment Canada, 2012). The latter is the largest producer of hydroelectricity in the world providing more than 13% of total output (Environment and climate change Canada, 2013). Since 1984, Canada has added 39 rivers into its Canadian Heritage Rivers System, a cooperative program governed by all Canadian provinces and territories for their long-term management and conservation in terms of historical, natural and entertainment purposes (Environment and Climate Change Canada, 2013) (Figure 2). Canada holds over 8,500 rivers and more than 2 million lakes covering 9% of the country's area (Monk, Baird, Curry, Glozier, and Peters, 2010). These rivers and lakes hydrology have an important impact on the aquatic fauna and habitat, especially for plankton, plants, benthic macroinvertebrates and vertebrates including amphibians, fish, birds, and reptiles (Monk, Baird, Curry, Glozier, and Peters, 2010).

Despite their abundance, Canadian water resources systems are subject to intense pressure from several stressors, including global warming. Several rivers have experienced changes in regime, ranging from shifts in the mean to decreasing or increasing trends. Given that the hydrological regime affects the performance of several human activities as well as ecosystem health, it would be interesting to be able to forecast (or simulate) the hydrological regime of a river at various time horizons. The results of the forecast and simulations can be used to optimize water resource system operation (short to medium term forecast) or design (long term simulations).

Figure 3. Canadian Heritage Rivers System (Environment and Climate Change Canada, 2013)

Canadian freshwater is found in glaciers, ice, and snow, streams, lakes, rivers, wetlands, and groundwater and soil moisture (Statistic Canada, 2017). Renewable freshwater is defined as the water that can replenish lakes, rivers, and aquifers regularly. On the other hand, non-renewable water is the water kept in ice caps, glaciers and deep aquifers and that can be replenished at a very low rate compared to human needs (Statistic Canada, 2017). The water yield refers to the

estimation of renewable freshwater, and it is obtained through data collected from rivers and streams in the country (Statistic Canada, 2017). In Canada, there are 25 drainage regions (Figure 3) sectioned into five ocean drainage areas (Pacific Ocean, Arctic Ocean, Gulf of Mexico, Hudson Bay or the Atlantic Ocean) (Statistic Canada, 2017). Managing water resources is challenging due to temporal and regional differences in water demand and supply due to factors such as economic growth, resource development, climate change and extreme weather events (Statistics Canada, 2017).

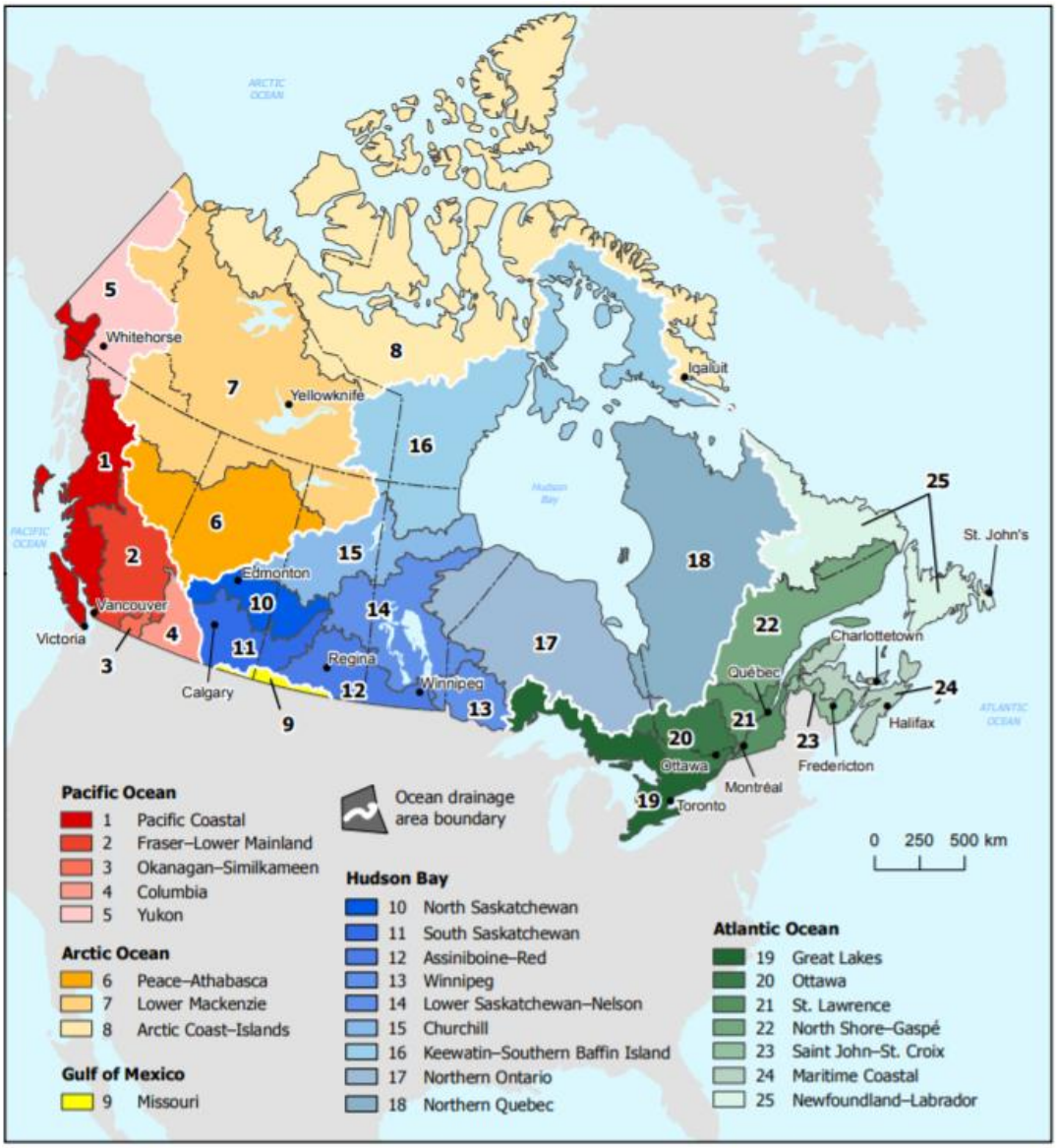


Figure 4. Drainage regions of Canada (Statistics Canada, 2009)

Water yield is defined as the estimates of freshwater runoff going into rivers and streams. In Canada, the average water yield between 1971 and 2013 is 3,478 km3 or 103,899 m3 per person (Table 1) (Statistics Canada, 2017). However, Canadian water yield distribution differs from one province to the other (Figure 4) (Statistics Canada, 2017). The highest yields are in the Pacific Coastal drainage region in British Columbia with an average annual water yield per unit area of

1.5 m3 /m2, while the lowest yields are in the Prairies (Missouri, Assiniboine– Red, South Saskatchewan, and North Saskatchewan) with a value of 0.05 m3 /m2 (Statistics Canada, 2017). In Canada, monthly water yield differs throughout the year. Usually, in late summer and fall, water yield decreases and the lowest yield is during winter, while during spring and early summer the most renewable freshwater is produced (Statistics Canada, 2017) (Figure 5).

Average annual water yield and selected statistics by drainage region, 1971 to 2013

	code	Total area[1]	Population, 2011	Average annual water yield, 1971 to 2013[2]	Water yield per area, 1971 to 2013[2]	Water yield per capita, 1971 to 2013	Water yield variability index, 1971 to 2013[3]	Average annual evapotranspiration, 1981 to 2010[4]
		km^2	persons	km^3	m^3/m^2	m^3/person	monthly CV	m^3/m^2
Canada	...	9,978,923	33,476,688	3,478.2	0.35	103,899	1.05	0.23
Pacific Coastal	1	334,455	1,505,007	510.2	1.53	339,002	0.50	0.26
Fraser–Lower Mainland	2	233,104	2,336,941	129.3	0.55	55,337	0.83	0.33
Okanagan–Similkameen	3	15,603	327,548	4.3	0.27	13,070	1.44	0.41
Columbia	4	87,323	160,896	67.9	0.78	422,042	1.04	0.41
Yukon	5	332,906	32,280	106.0	0.32	3,283,759	..	0.14
Peace–Athabasca	6	485,145	406,303	99.5	0.21	244,789	1.01	0.31
Lower Mackenzie	7	1,330,490	52,844	246.3	0.19	4,660,913	..	0.17
Arctic Coast–Islands	8	1,764,280	20,133	231.3	0.13	11,488,870	..	0.11
Missouri	9	27,096	8,439	0.5	0.02	62,660	2.14	0.33
North Saskatchewan	10	150,151	1,559,613	10.4	0.07	6,700	1.04	0.34
South Saskatchewan	11	177,623	2,168,447	10.3	0.06	4,732	1.10	0.34
Assiniboine–Red	12	190,704	1,464,936	8.4	0.04	5,702	2.49	0.39
Winnipeg	13	107,655	82,775	25.9	0.24	312,611	0.87	0.40
Lower Saskatchewan–Nelson	14	360,887	216,586	51.7	0.14	238,501	0.92	0.32
Churchill	15	313,568	94,292	51.0	0.16	541,004	0.75	0.24
Keewatin–Southern Baffin Island	16	939,569	13,968	192.0	0.20	13,745,664	..	0.13
Northern Ontario	17	691,809	134,355	199.2	0.29	1,482,638	..	0.31
Northern Quebec	18	940,193	109,239	516.3	0.55	4,726,324	..	0.18
Great Lakes	19	317,860	11,287,184	133.3	0.42	11,806	0.72	0.48
Ottawa	20	146,353	1,957,937	64.3	0.44	32,858	0.80	0.47
St. Lawrence	21	118,733	6,583,552	72.3	0.61	10,983	0.76	0.39
North Shore–Gaspé	22	369,095	508,869	290.4	0.79	570,600	0.91	0.27
Saint John–St. Croix	23	41,903	413,581	29.8	0.71	72,156	0.89	0.48
Maritime Coastal	24	122,057	1,515,262	103.6	0.85	68,343	1.03	0.47
Newfoundland–Labrador	25	380,361	515,698	324.2	0.85	628,662	..	0.23
Newfoundland (Island)[5]	...	111,186	487,808	125.6	1.13	257,404	0.62	0.34

Table 1.Average Annual water yield and selected statistics by drainage regions in Canada from 1971 to 2013 (Statistics Canada, 2017; Wang, Yang, Luo and Rivera, 2013)

1. The total area includes land and water.
2. The water yield estimates are 42-year annual averages, with the exception of estimates for drainage region 1(41 years of data); regions 5, 7, 17 and 18 and portions of 8, 16 and 25 (20 years of data); drainage region 8 and 16 (23- year of data) (Spence and Burke, 2008).
3. The variability index is measured using the coefficient of variation (CV) that allows the comparison of all months during 42 years.
4. The estimate of evapotranspiration at a 1 km resolution, excluding the Great Lakes.

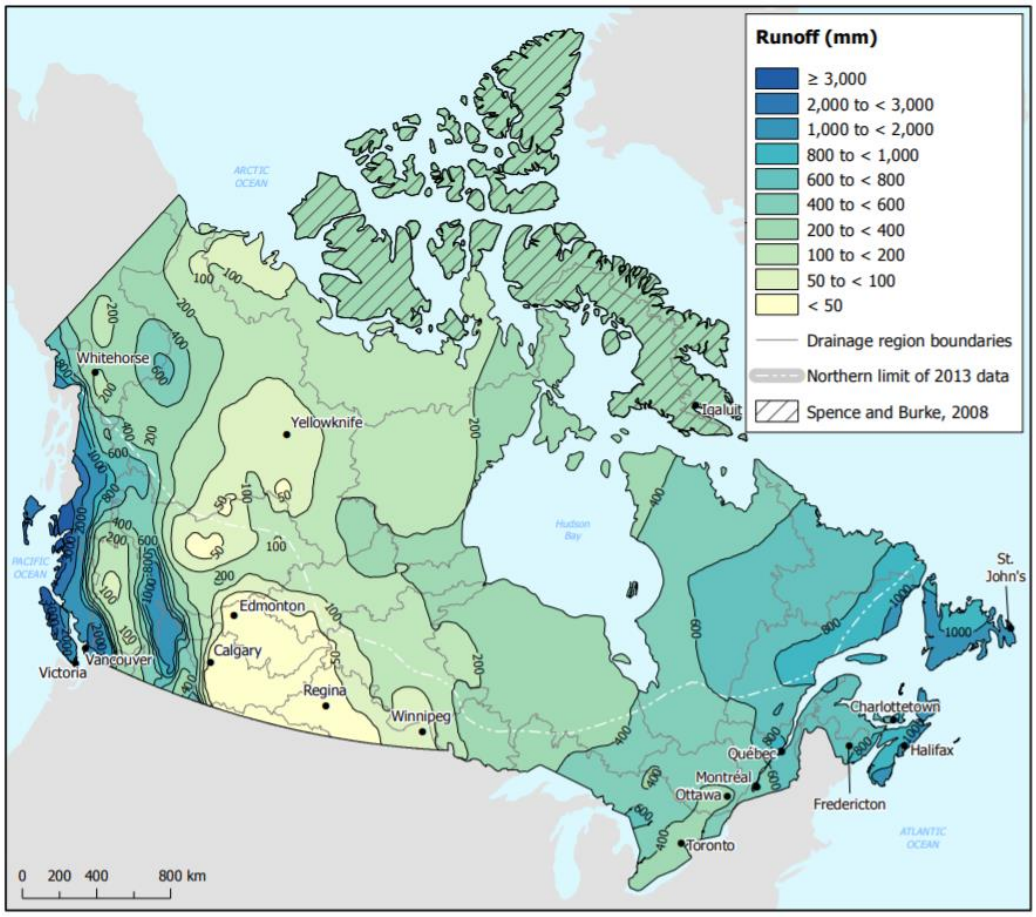


Figure 5. Canadian annual average runoff from 1971 to 2013 (Statistics Canada, 2017)

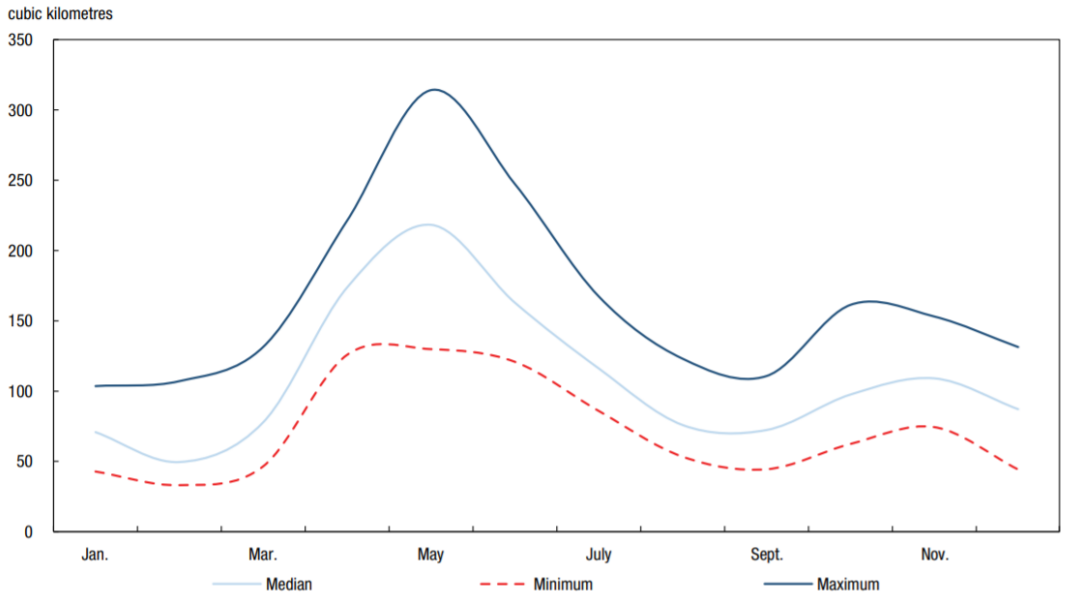


Figure 6. Monthly water yield in southern Canada between 1971 and 2013 (Statistics Canada, 2017)

2.2.1 Stream flows

The majority of Canadian rivers are showing an important seasonal flow variation (Burn and Whitfield, 2016). There is minimal annual flow during late summer when there are low precipitation and high evaporation rates. The flow is also minimum during late winter when precipitations are frozen in snow and ice. Minimum annual flows can impact negatively aquatic habitats as they influence water temperatures, the levels of dissolved oxygen, the minimum flow requirement for aquatic species, and the increase of summer temperatures (Schindler, 1997). According to a study realized on 172 sites, the annual flow has increased of 13% from 1970 to 2005, and most of these sites are located in northern Montane Cordillera, Taiga Plains, Taiga Shield, Boreal Cordillera and Arctic ecozones (Monk, Baird, Curry, Glozier, and Peters, 2010). The maximum annual flows happen in late spring and early summer due to melting snow and rainstorms. About 20 % of the 172 sites are showing a decrease in the maximum flow, while 5% of them showed an increase in maximum flow (Monk, Baird, Curry, Glozier, and Peters, 2010).

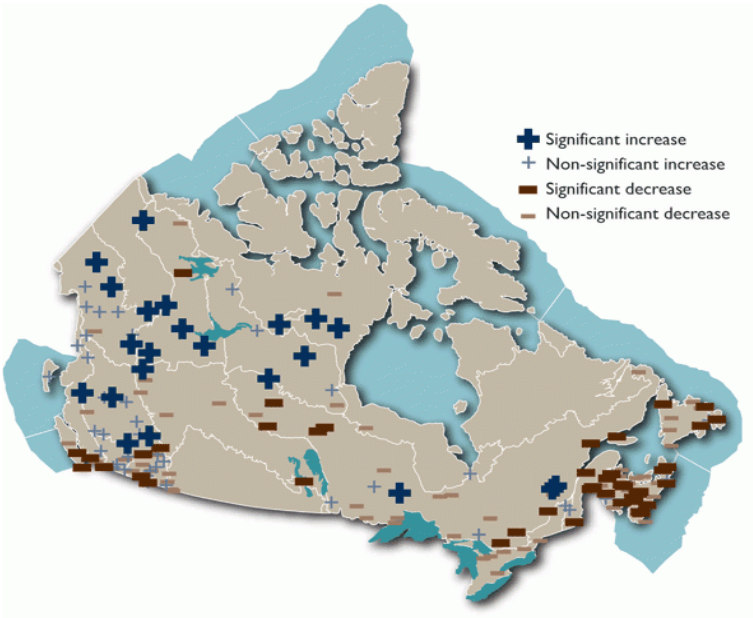


Figure 7. Trends in minimum river flow in Canadian natural rivers from 1970 to 2005 (Monk, Baird, Curry, Glozier, and Peters, 2010).

Figure 8. Trends in maximum river flow in Canadian natural rivers from 1970 to 2005 (Monk, Baird, Curry, Glozier, and Peters, 2010).

2.2.2 Canadian prairies water level in closed-basin lakes

In southeast Alberta and south Saskatchewan, dry climate and repeated glaciations resulted in several closed-plain lakes that drain internally. These lakes are vulnerable to changes in the variation of precipitation and evaporation. These variations affect water levels and salinity (Van der Kamp, Keir, and Evans, 2008). In 16 of these closed-basin lakes, water levels have decreased in a range of 4 to 10 meters between 1910 and 2006 (Van der Kamp, Keir, and Evans, 2008). This phenomenon is caused by several factors including ditches, dams, dugouts, wetland drainages, agricultural practices (Monk, Baird, Curry, Glozier, and Peters, 2010), decline of summer fallow (Javorek, and Grant, 2010) and an increase in spring temperatures which encourages high evaporation and low stream runoff (Zhang, Harvey, Hogg, and Yuzyk, 2001).

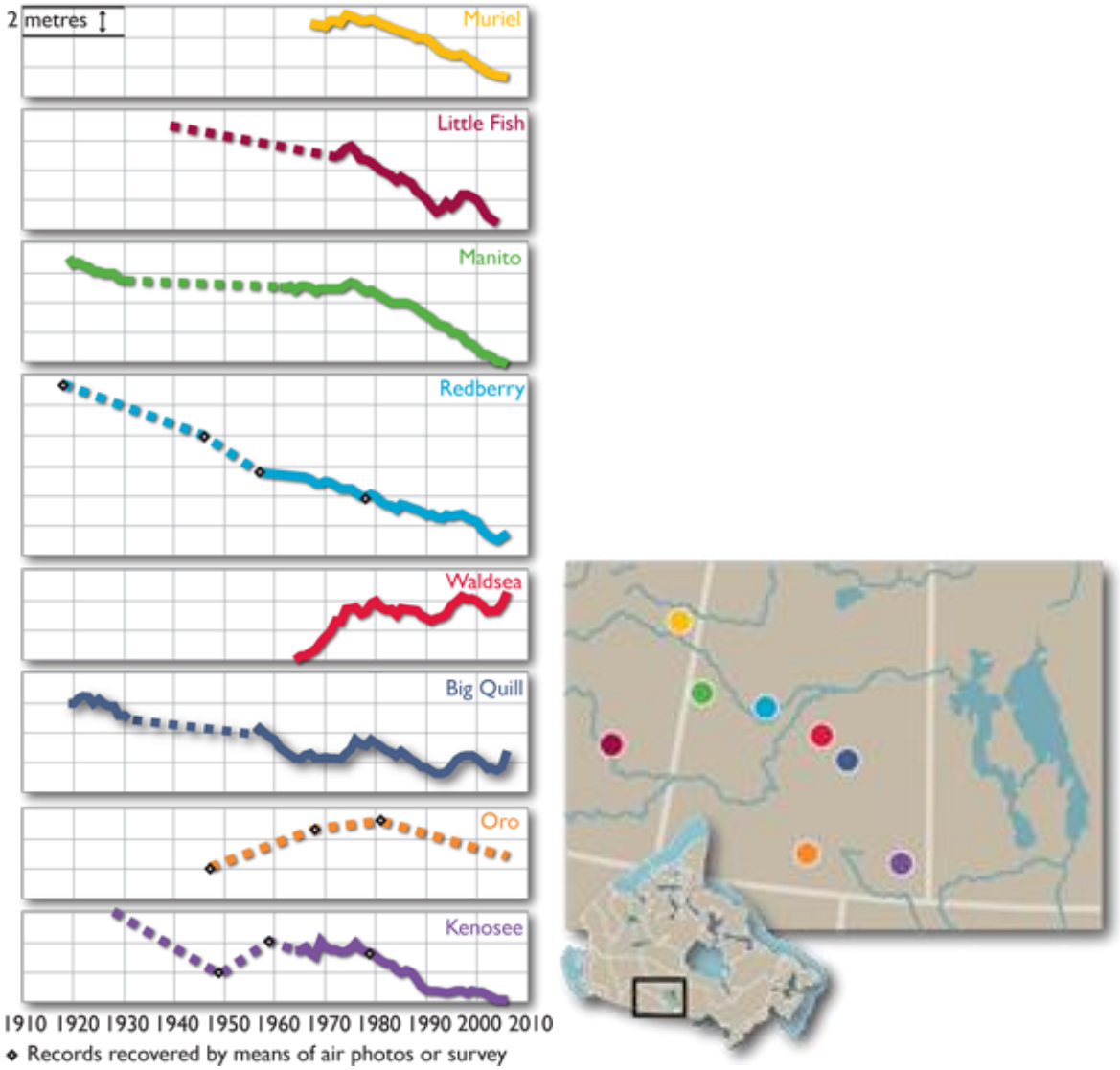


Figure 9. Water levels in the prairie (Southeast Alberta and south Saskatchewan) closed-basin lakes from 1910 to 2010 (Van der Kamp, Keir, and Evans, 2008).

2.2.3 Freshwater discharge fluctuations in the Arctic and North Atlantic

Canada is known for having plenty of freshwater resources. It has been established that Canada's renewable freshwater supply is the third biggest in the world and the second biggest when it comes to the amount per capita in developed countries (Statistic Canada, 2017). Nonetheless, from 1964 to 2003, there was a decrease of 10 % of Canadian freshwater discharge into the Arctic and North

Atlantic Oceans (Déry and wood, 2005; McClelland, Dery, Peterson, Holmes and Wood, 2006). It is important to mention that fluctuations in the number of freshwater discharges in northern seas influence ocean process and marine species (Peterson, Holmes, McClelland, Volosmarty and Lammers, 2002; Peterson, McClelland, Curry, Holmes, Walsh and Aargaard, 2006).

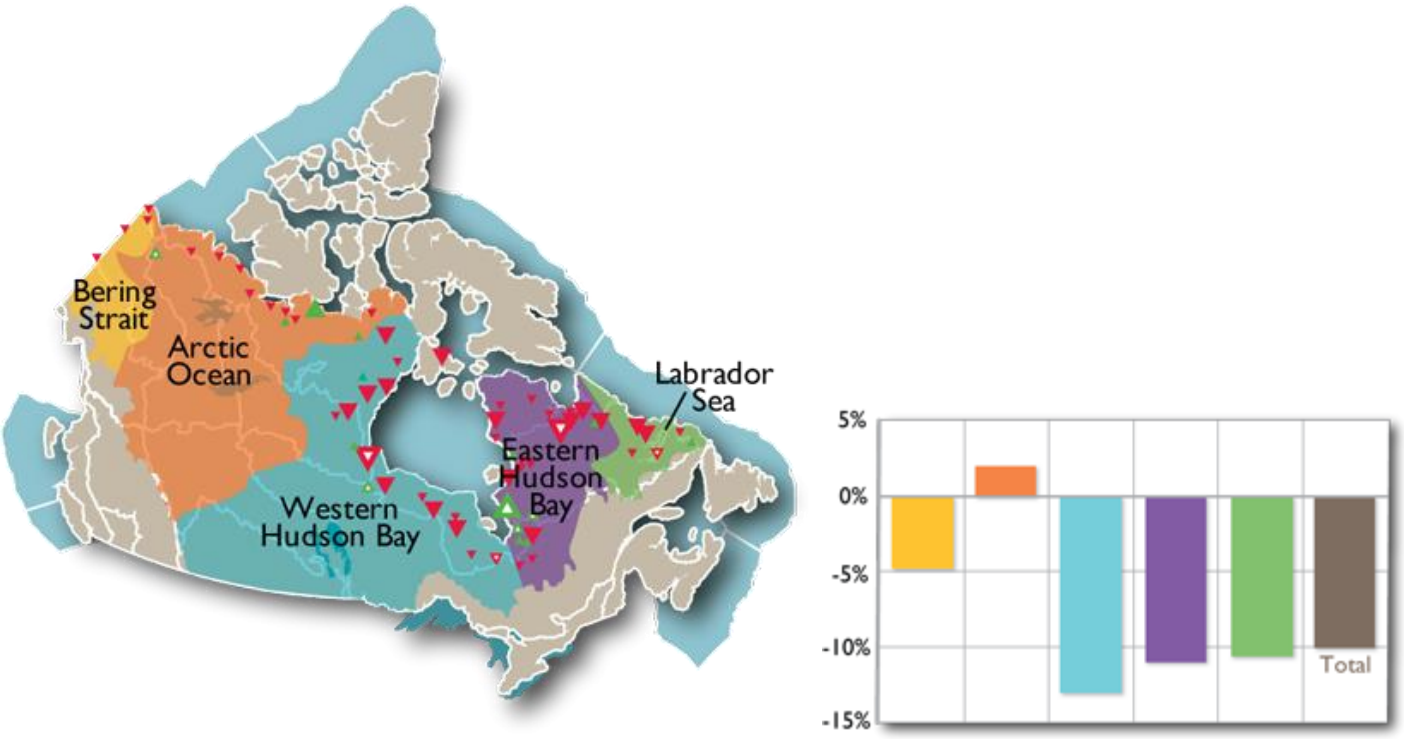


Figure 10. a) Change in the freshwater discharge into the Arctic and North Atlantic from 1964 to 2003 with triangles of several sizes indicating the magnitude of change for the decrease (red) or increase (green) inflow b) freshwater discharge Percent change from 1964 to 2003 (Déry and Wood, 2005)

From 1971 to 2013, the average Canadian annual water yield was of 3,478 km^3 or 0.349 m^3/m^2, which is equal to a 349mm depth in the area occupied by the country (Table 2) (Statistic Canada, 2017). It has been established that annual water yield in the southern Canadian areas varied between 1974 and 1987 going from high (1.544km^3) to low (1.165km^3). From 1972 to 1987, the water yield decreased and then had a

gradual recovery until 2012, which was then followed by a dip at the end of the '90s and early 2000s (Statistic Canada, 2017).

Renewable freshwater resources, water use and gross domestic product for selected countries

	Total area, 2014	Population, 2015	Total renewable freshwater resources[1]			Total water withdrawals[2]		Gross domestic product, 2014	
			Per year	Per capita	Per area	Per year	Per capita		
	km^2	thousand	km^3/year	m^3 per capita	m^3/m^2	km^3/year	m^3 per capita	million US dollars	US dollars per capita[3]
Algeria	2,381,740	39,667	12	294	0.005	8	225	213,518	5,383
Argentina	2,780,400	43,417	876	20,181	0.315	38	898	543,490	12,518
Australia	7,741,220	23,969	492	20,527	0.064	20	824	1,471,439	61,389
Brazil	8,515,770	207,848	8,647	41,603	1.015	75	370	2,346,523	11,290
Canada	9,978,923	35,852	3,478	103,899	0.349	38	1,078	1,785,390	49,799
China	9,600,000	1,407,306	2,840	2,018	0.296	554	411	10,430,590	7,412
Egypt	1,001,450	91,508	58	637	0.058	78	911	282,242	3,084
France	549,090	64,395	211	3,277	0.384	33	521	2,829,192	43,935
Germany	357,170	80,689	154	1,909	0.431	33	411	3,868,291	47,941
India	3,287,260	1,311,051	1,911	1,458	0.581	761	602	2,054,941	1,567
Mexico	1,964,380	127,017	462	3,637	0.235	80	658	1,294,695	10,193
Russian Federation	17,098,250	143,457	4,525	31,543	0.265	66	456	1,849,940	12,895
South Africa	1,219,090	54,490	51	936	0.042	13	270	349,819	6,420
United States	9,831,510	321,774	3,069	9,538	0.312	486	1,543	17,348,072	53,914

Table 2. Renewable freshwater resources, water use, and gross domestic product for selected countries (Statistic Canada, 2017).

1. Renewable water resource data are the long-term total renewable freshwater resources from 1971 to 2012 done by Statistics Canada.
2. Total water withdrawal volumes have been calculated using available data for 2013 (Australia and Canada, 2012 for Algeria, 2011 for Mexico and Argentina), 2010 (for France, Germany, Brazil, India, Egypt and the United States, 2005 for China, 2001 for the Russian Federation) and 2000 (South Africa).
3. Per capita gross domestic product was calculated using 2015 population data collected from United Nations Statistics Division, 2016, "GDP by Type of Expenditure at current prices - US dollars," National Accounts Estimates of Main Aggregates, http://data.un.org/Data.aspx?q=GDP+US+dollars&d=SNAAMA&f=grID:101;currID:USD;pcFlag:0&c=2,3,5,6&s=_crEngNameOrderBy:asc,yr:desc&v=1; Statistics Canada, CANSIM Tables 153-0116 and 051-0001, www5.statcan.gc.ca/cansim/home-accueil?&lang=eng&MM=as.

In 2017, the highest water yield per unit area was 1.5 m^3/m^2, and happens to be in British Colombia, more specifically in the Pacific coastal drainage region, whereas Missouri Assiniboine–Red, South Saskatchewan and North Saskatchewan's drainage areas in the prairies have shown the lowest yields with an average annual water yield per unit area of 0.05 m^3 /m^2 (Statistic Canada, 2017).

In spring and beginning of the summer, the water yield is at its highest flow as it varies all year long. In southern Canada, it was observed that the median monthly water yield peaked at 2018km^3 in May but soon decreased in August reaching 76km^3, followed by another low (50km^3) in February (Statistic Canada, 2017).

Spring flows are known to dominate Water yields in the Okanagan–Similkameen and the Assiniboine–Red drainage regions. The median water yields in those regions for spring/summer (April, May, June) represent 75% and 77% of annual flows between 1971 and 2013. To highlight the difference, later during the year, when the peak median water yields happened, and regularly declined in drainage areas in south and north of Saskatchewan (Statistic Canada, 2017). In 2013, Canada used water for electric power generation, transmission and distribution (68%), for manufacturing purposes (10%), households needs (9%), national agriculture (5%) and mining, oil and gas extractions (3%)(Statistic Canada, 2017) (Figure 10).

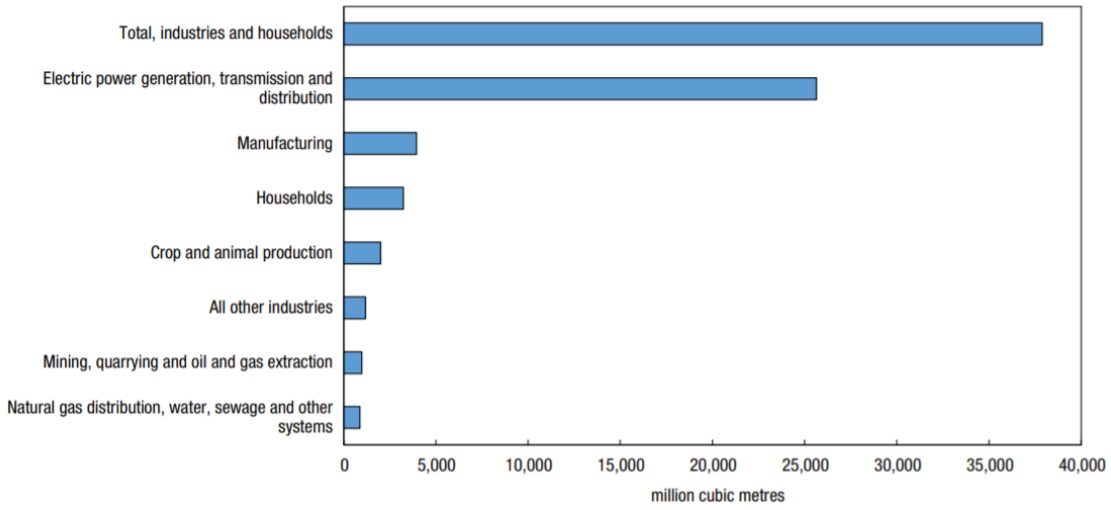


Figure 11. Water use in Canada by sectors in 2013 (Statistic Canada, 2017)

2.2.4 Sediments in Canadian water bodies

Water is essential to sculpt the Canadian landscape as it moves sediments which are an important volume of soil eroded from the landscape before moving toward river systems and dropped in lakes or seas (Environment and Climate Change Canada, 2013). Canada's hydrological system transports annually millions of tonnes of sediments. For instance, the Fraser River transports 20 million tonnes of sediments per year on average in the marine environment (Environment and Climate Change Canada, 2013).

The first step of the sediment cycle is the process of erosion when particles or fragments in rocks are weathered (Environment and Climate Change Canada, 2013). It has been established that erosion on the surface of the earth is affected by water, wind, glaciers, plant, and animal processes. Fluvial sediment movement occurs when water is the main cause of erosion.

Sediment transportation starts on the land surface with various erosions initiated by rain before water bodies (e.g., rivers, streams, gullies…) act as conduits and containers. At the end of the cycle, when movement energy is at the lowest, sediment deposit in channels and deltas, and on lakes and river beds. In Canada, sediments are considered in three categories (Environment and Climate Change Canada, 2013):

1) suspended load (suspended in water, mostly sand, silt, and clay)
2) bed load (rolling/bouncing near the bottom, stony material such as gravel and cobbles)
3) bed material (stationary on the bed)

Sediments have a major role in terms of toxic chemicals transportation, navigation, fisheries, aquatic habitat conditions, forestry, water supply, energy production, and agriculture (Environment and Climate Change Canada, 2013). For example, deposition of high quantities of sediments in watercourses can decrease water depth and consequently cause navigation problems. According to Chambers, Dale, Scrimgeour, and Bothwell (2000), agricultural practices (e.g., tillage) are greatly responsible for the increased sediment load into watercourses. The increased sediment load fills streambeds excessively which contributes to the frequency and depth of flooding (Libby and Boggess 1990). Also, a higher load of suspended sediment increases turbidity impacts negatively aquatic ecosystems and navigation (Hoyer and Jones, 1983). Environment Canada regularly collects sediment samples to study through time the physical and chemical

characteristics of Canadian watercourses sediments (Environment Canada 1994). During sampling, the quality and quantity of sediments are done as follow (Environment and Climate Change Canada, 2013):

1) Special samplers for suspended load collect data before being shipped to laboratories for further analysis. It is done mostly during high-flow levels when a high quantity of sediments is transported (spring, summer, fall rainstorms).

2) Special samplers are lowered to the watercourse bed to rest and trap the material transported along the bottom. It is done usually in the spring during high sediment discharges. After data collection, the bed-load samples are shipped to the laboratory.

3) Special samplers for bed-material sediments can scoop or extract a core from the bed. The sampling can also be done by hand from exposed stream/bar banks and is usually done during summer during low flow settings as it exposes some areas of the stream bed. The samples are sent to the laboratory.

The laboratory analyzes the concentration (ratio of dry sediments to the total water-sediment mixture in mg/L) and particle size (gravel, cobbles, clay, silt or sand) found in the samples. For example, the Fraser River in British Columbia has an average annual suspended sediment load of 20 000 000 tonnes compared to only 946 tonnes for Northeast Branch in Newfoundland (Environment and Climate Change Canada, 2013).

2.3 Definition of time series

A time series is an ordered sequence of data points measured over successive times (Akgun, 2003; Cochrane, 1997; Hornik, Stinchcombe &White, 1989; Hipel & McLeod, 1994). It reflects a set of vectors where the variables *x(t)* are arranged in chronological order. Examples of such data could be the daily recorded Water Level of a river, the weekly measured water flow, or the monthly concentration of sediments displaced. In time series, records of a single variable or more than one variable are respectively known as univariate and multivariate. The observations can be continuous when measured at every instance of time or can be discrete when recorded at various intervals. Finally, in time-series observations are plotted in a graph against corresponding time (e.g. week, month, year, seasons, etc).

2.4 Time series components

The main components affecting Times Series are Trend, Cyclical, Seasonal, and Irregular components (Akgun, 2003; Cochrane, 1997; Hornik, Stinchcombe &White, 1989; Hipel & McLeod, 1994). These patterns could be used to identify the type of data exhibited from the time series plots and thus choose appropriate forecasting procedures. First, a trend is the movement generating the increase, decrease or stagnation over a long period. Usually, the duration of the trend is more than a year. The trend can be linear, non-linear, exponential or quadratic. Second, the cyclical variation affecting time series is the change that repeats itself in the cycle due to circumstances. The variations differ from one cycle to another and occur in a non-regular fashion in a period of cycle that is greater than a year. Third, the seasonal variation is a short-term wavelike fluctuation regularly occurring over fixed intervals of time. Its duration is less than one year. Often, this variation corresponds to the seasons and occurs due to factors including the weather conditions. The last component is the irregular variation reflecting a random deviation of the

observations from the underlying component. This residual variation represents the variation remaining after considering the trend, cyclical and seasonal variations. The occurrence of rare events such as earthquakes, floods, and wars, can explain the irregular fluctuation.

These four components can be used to form additive or multiplicative models. In one hand, an additive model $Y(t) = T(t) + S(t) + C(t) + I(t)$[1], the seasonal, cyclical and irregular variations do not depend on the level of the trend variation. Moreover, in such a model, the variations are absolute values, and the components are independent of each other. On the other hand, in a multiplicative model $Y(t) = T(t) \times S(t) \times C(t) \times I(t)$, the seasonal, cyclical and irregular variations depend on the level of the trend variation. The more the trend variation is higher, the more the observed variation is intensive. In the multiplicative model, the four components are dependent on each other.

2.5 Relevance of time series analysis

Time series analysis is the procedure of fitting a time series to a model to extract meaningful statistics and characteristics of the data, to understand the nature of the series and to perform future simulations and forecasting (Hipel & McLeod, 1994). In time series forecasting, the prediction of future events is made through models capturing past observations. The main aims of time series Analysis are description, explanation, prediction, modeling, and/or control (Gottman, 1981). First, observations can be described by time plots reflecting the trends and periodicities, seasonal

[1] Y(t) is the observation at time t.
T(t) is the trend variation at time t.
S(t) is the seasonal variation at time t.
C(t) is the cyclical variation at time t.
I(t) is the irregular variation at time t .

behavior and other features. For example, histograms can help examine the symmetry of the amplitude of time series. Visualization of the time plots is relevant to the identification of the time series data patterns and appropriate forecasting methods. Second, explaining time series using observed variation allows understanding relationships between observations. For example, the difference in weather temperature can be explained by considering various two different seasons (winter vs. summer). Third, predicting future values of time series using observed variations might be suitable for analysis purposes and control of problems in various situations. For example, predicting the weather temperature at the end of the winter season might be useful in the management of flood risk. Forth, another objective of time series is to identify suitable statistical univariate or multivariate models that can be used. Multivariate models are based on past values of a given variable while multivariate models are based on both past and current values. Finally, the last objective of time series is to control the quality of the process of generating the predicted data. When aberrations are identified in the model, corrective measures are applied to optimize its settings and to ensure conformity to the target values.

2.6 Time series concepts

In this section, we explain the concepts of deterministic time series, stochastic process, stationarity, and linearity in time series.

2.6.1 Concept of Deterministic time series

Many time-dependent mathematical problems can be modeled and determined with precision based on various laws such as the physics law for example (Gyasi-Agyei, 2012). For example, it is possible to calculate the energy generated inside a reactor based on the rules of thermodynamics and the fact that in a closed and controlled environment, the results are not subject to probability.

These models are called deterministic and are expressed as functions such as $Y_t = f(t)$ where t =1,2,….n. Deterministic time series have no random or probabilistic aspects. In deterministic time series, past and future values are specific to values of a given time.

2.6.2 Concept of Stochastic Process

A Stochastic process is a sequence of random variables ordered in time following probabilistic laws. Time Series is non-deterministic since the observations occurring in the future can't be predicted with certainty. In theory, given a behavioral model for a system, we can predict future values of a time series measured from that system, based on past observations. However, in practice, physical systems are affected by many kinds of disturbances, so the predicted values always reflect the stochastic, or statistical, characteristic of a time series. The stochastic process serves as a model for the analysis of an observed time series. It is expressed as $Y_t = X(t)$, where X is a random variable (Cryer & Chan, 2008). For example, future values for Environmental data measured in an uncontrolled open environment, such as water flows, water levels, or suspended solids, can be established by a statistical model within specified probabilistic limits. Indeed, Unknown factors that are random such as defective equipment might affect the results.

2.5.2.1 Concept of Stationarity in Stochastic Process

Stationarity is defined as the assumption that the probability governing the variation in the observed data stays the same over time (Cryer & Chan, 2008). Indeed, the stationarity assumption implies that the probability distribution of observations remains the same for all times t and therefore has a constant mean. The fluctuation of observations does not follow an upward or

downward trend, and the Time Series has constant variance. It assumes that the process is in a state of statistical equilibrium. Otherwise, the time series is non-stationary.

There are two types of stationary processes. On the one hand, a process {Yt} is said to be strictly stationary if the marginal distribution of Y at time t[$p(Y_t)$] is the same as at any other point in time. Therefore $p(Y_t)\ is\ equal\ to\ p(Y_{t+k})$ and $p(Y_t, Y_{t+k})$ does not depend on t if t larger or equal to 1 and k is an integer value. A stationary time series implies therefore that the mean, the variance, and covariance of the series Y_t are time-invariant. Stationary data assumes a time series data displaying a zero mean and a constant variance over time.

On the other hand, a process {Yt} is said to be weakly stationary if the statistical moments of the process up to order k depend only on time differences. It is important to underline that the time of occurrences of the data being used to estimate the observations is not relevant. A weakly stationary process following normal distribution is considered to be strictly stationary (Cochrane, 1997).

In practice, classifying a process as stationary or not depends on the period of observation considering that series can be stable in a short period and non-stationary in a longer one (Cochrane, 1997). Time Series is said non-stationary when trends or seasonal patterns can be observed. In these cases, it is suitable to remove these patterns and achieve stationarity. Non-stationary characteristics are usually observed when the period of historical observations is greater. Stationarity in a Times Series data is required for an adequate mathematical model for future forecasting (Hipel & McLeod, 1994).

To detect stationarity or non-stationarity in a time series data, various tests can be used including Dickey and Fuller Test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test are two tests t

(Gyasi-Agyei, 2012; Cochrane, 1997). Dickey and Fuller's Test aims to confirm or infirm whether a time series is non-stationary and possesses a unit root. The hypotheses are H_0: X_t is non-stationary and H_1: X_t is stationary. H_1 can be tested in regression equation

$$\Delta X_t = \beta_0 + \alpha t + \beta_1 X_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta X_{t-1} + \varepsilon_t \quad ; \text{ If P} - \text{value} > 0.05, \text{H0 is to be accepted.}$$

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test is another used to identify stationarity. KPSS aims to confirm or infirm whether a time series is stationary around a mean or linear trend. The hypotheses are H_0: X_t is stationary and H_1: X_t is non-stationary. The KPSS test breaks up a series three parts including deterministic trend (βt), a random walk (r_t), and a stationary error (ε_t) and H_1 can be tested in regression equation $x_t = r_t + \beta t + \varepsilon_1$ If p-value > 0.05, H_0 is to be accepted. To remove the trend and seasonal patterns making the series non-stationary, differencing can be used to stabilize the mean of a time series.

2.6.3 Concept of Linearity

In time series, a model is linear if the current value of the series is a linear function of past observations. A stochastic process s (X_t, $t \in Z$) is said to be a linear process if for every $t \in Z$ we have $X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}$ where $a_j = 1$, (ε_t, $t \in Z$) is iid with $E\varepsilon t = 0$, $E\varepsilon^2_t < \infty$, and $\sum_{j=0}^{\infty} |a_j| < \infty$ (Box & Jenkins, 1970; Park, 1999; Parelli, 2001). In the literature, Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) are two linear time series models. These will be described in section 2.5.

A non-linear time series model is a model reflecting a non-linear function of the past observations. A Non-linear model is appropriate for predicting volatility changes in time series. In the literature,

Autoregressive Conditional Heteroskedasticity (ARCH) model, Generalized ARCH (GARCH), Exponential Generalized ARCH (EGARCH), Threshold Autoregressive (TAR), the Non-linear Autoregressive (NAR) model, and the Nonlinear Moving Average (NMA) model are examples of non-linear time series models. The most popular non-linear models are ARCH and NMA. ARCH represented as $y_t = \varepsilon_t + \alpha\sqrt{\varepsilon_t^2}$ is a non-linear model in mean but not in variance while NMA represented as $y_t = \varepsilon_t + \alpha\varepsilon_{t-1}^2$ is a non-linear model in the variance, but linear in the mean.

2.7 Linear time series Models

In this section, we present three models of time series, including ARMA, ARIMA, and SARIMA.

2.7.1 Autoregressive Moving Average (ARMA) Models

The Autoregressive Moving Average ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms. This model is appropriate for univariate time series modeling, and when the sequence of the observations is linear (Charles, 2011). ARMA is a weighted average of previous values $\varphi_1 y_{i-1}$ plus a fixed constant φ_0 and a random error term ε_i. The error terms are assumed to be independently distributed based on a Normal distribution with zero mean and a constant variance $N(0, \sigma^2)$. The error terms are independent of the y values. ARMA represented by equation as follows:

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_1 y_{i-1} + \sum_{j=1}^{p} \theta_j \varepsilon_{t-j}$$

2.7.2 Autoregressive Integrated Moving Average (ARIMA) Models

One of the most popular time series analysis tools is the Autoregressive Integrated Moving Average (ARIMA) method, introduced by Box and Jenkins (1992). This method is appropriate for discrete systems where the observations occur at equally spaced intervals of time (Box et al. 1994).

The ARIMA model can be effectively implemented on current and past recorded values to produce conservative estimates for the future.

ARIMA is an algebraic statement describing how observations on a time series are statistically related to past observations and residual terms from the same series. The ARIMA model is a statistical non-stationary Stochastic model combining an autoregressive component of order p and a Moving average component of order q written $w_t = \nabla^d Y_t = (1 - B)_t \ Y_t$.

ARIMA is the combination of an Autoregressive process over previous lags, and a Moving Average process over the lags error components. The integrated part is the differentiation of data to convert Non-Stationary data to Stationary. The ARIMA(p,d,q) model can be represented by the following equation:

$$w_t = \sum_{t=1}^{p} \propto_i w_{t-1} \ + \sum_{t=1}^{q} \theta_i e_{t-1} \ + \mu + e_t$$

2.7.3 Seasonal Autoregressive Integrated Moving Average (SARIMA) Models

SARIMA is the seasonal ARIMA where non-stationarity is removed using seasonal differencing of order k which is the difference between an observation and the corresponding observation from the previous year that is calculated as $z_t = y_t - y_{t-s}$ (Kihoro, Otieno & Wafula, 2004). The SARIMA $(p,d,q)x(P,D,Q)^s$ model where z_t is the seasonally differenced series can be represented by the following equation:

$$\Phi_p \ (L^s)_{\Phi_p} \ (L)(1 - L)^d (1 - L^S)^D y_t = \ \Theta_Q (L^s)_{\Phi_p} \ (L)_{\varepsilon_t} ,$$

$$i.e. \Phi_p \ (L^s)_{\Phi_p} \ (L)_{z_t} = \Theta_Q (L^s) \theta_p \ (L) \varepsilon_t ,$$

2.8 Box-Jenkins Methodology for selecting an appropriate time series model

The Box-Jenkins Model is a methodology used to select a times-series model that can generate an accurate forecast (Box & Jenkins, 1970). It is a systematic method leading to identify, fit, check, and use integrated autoregressive, moving average (ARIMA) time series models that is appropriate for time series of at least 50 observations. The Box-Jenkins methodology uses three iterative phases for forecasting future values of the time series.

First, the phase of identification consists of preparing the data and selecting a model. The objective of this step is to select values of d and then p and q in the ARIMA (p,d,q) model. During this phase, the data is transformed to stabilize the variance to reflect stationary variables, to detect seasonality in the dependent series, and to use plots of autocorrelation functions (ACF & PACF) to decide if an autoregressive or moving average component should be used in the model. Differencing could be used to achieve stationarity.

Second, the phase of estimating and testing uses algorithms (e.g., maximum likelihood, non-linear-squares estimation) to identify coefficients that best fit the model. Indeed, Akaike Information Criterion (AIC) $AIC(p) = n\ln(\hat{\sigma}_\epsilon^2/n) + 2p$ or Bayesian Information Criterion (BIC) $BIC(p) = n\ln(\hat{\sigma}_\epsilon^2/n) + p + p\ln(n)$ where n is the number of effective observations and $\hat{\sigma}_\epsilon^2$ is the sum of squared residuals could be used (Faraway & Chatfield, 1998; Kihoro, Otieno & Wafula, 2004). The number of model parameters minimizing either AIC or BIC defines the optional model.

Finally, in the third phase, the time series model is used to test if the model is conforming to the specifications of a stationary process. If the model is inadequate, the iterative process is to be repeated to build a better model.

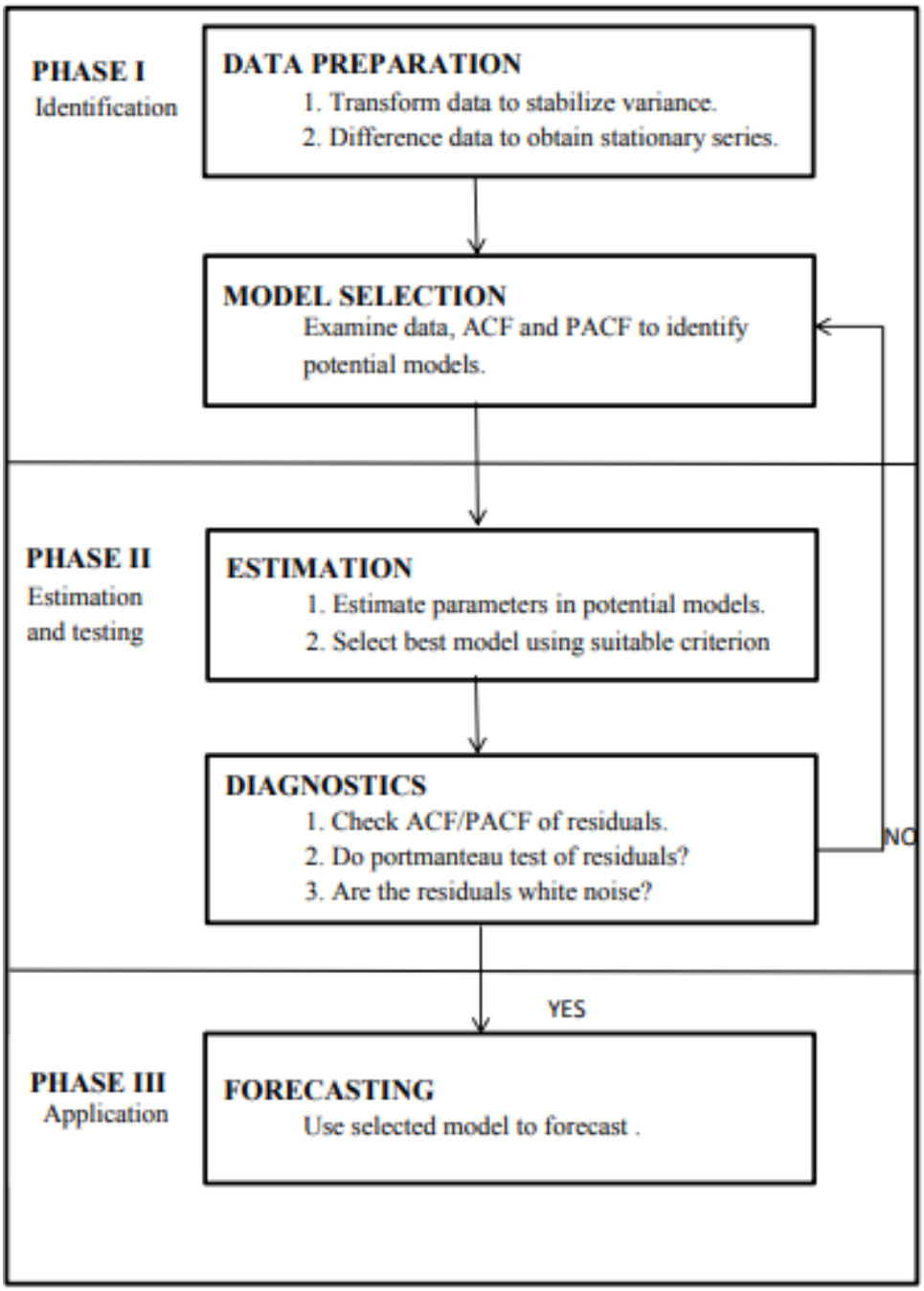


Figure 12.The Box-Jenkins methodology (Spyros et al., 1998)

2.9 Forecasting

The forecasting Principles project started in 1997 and assessed relevant information about knowledge in forecasting (Armstrong and Pagell, 2003). According to forecasting principles, it is advised to combine forecasts from two or more forecasting methods if the uncertainty is high (Armstrong and Pagell, 2003). A joint probability model was used by Wang and Robertson (2011), to study forecasting flows at several sites in temporary streams, by defining zero flow occurrences as censored data. The adaptation of these methodologies when it comes to the spline compartment model is possibly accurate for forecasting intermittent flow on several sites.

Various water demand models used approaches for long-term (future decades) and short-term (up to six months) forecasting (Kame'enui, 2003). Valid forecasts of stream flows are highly reliable to water resources planning and management if done months or seasons before (Chiew et al., 2003; Nigam, 2016). In terms of seasonality, according to Dyck, Cool, Rodriguez and Sadiq (2014), in terms of geography, Canada has the most seasonal variability when it comes to weather, in the world. Seasonal forecasting of streamflow consists of running climate models to generate forecasts of several weather variables (e.g. rainfall) to be used into hydrological models to forecast streamflow (Chiew et al., 2003). According to Pokhrel, Robertson, and Wang (2013), predicting streamflow using a hydrological model is useful for several purposes including flood forecasting at short- and long-term scales of water resource assessment.

However, models used for predictions can be inaccurate due to errors in several components such as model structures, input data, and parameters. One way to solve these issues is to calibrate the model before its application which will reduce uncertainty in the generated prediction (Pokhrel,

Robertson, and Wang, 2013). Important efforts are continuously provided to develop hydrological, statistical, and satellite-driven methods, to increase the lead time of forecasting while omitting the use of complex methods (Armstrong and Pagell, 2003).

In terms of river flood forecasting, to provide an accurate prediction, it is necessary to monitor and control the measurement and notifications of certain components such as water levels, precipitation, and velocity (Merkuryeva and Kornevs, 2013). River flood forecasting uses mining historical data and specific domain knowledge as their primary reference to accurately determine flood forecasting (Merkuryeva and Kornevs, 2013). It is necessary to use space and ground-observed data from satellites and terrestrial stations such as climatological stations, automatic rain gauge, and meteorological stations when it comes to effective flood monitoring and control (Merkuryeva and Kornevs, 2013).

In short terms, the forecast of river flow takes a few hours to a few days to complete. In long terms, the forecast of river flow can reach up to nine months (Georgakakos, and Krzysztofowicz, 2001). According to Webster and Hoyos (2004), five to ten days is considered to be an appropriate forecasting lead time for an increase in flood response and vigilance in areas at risk of flood. All components must be involved to generate a perfect or almost perfect model of a flood forecasting system, which is uncommon in a real scenario (Lorenz, 1963). Also, it has been assessed that to diminish social tragedies, decrease financial damage and improve public safety, an effective flood alarm system (on short term flow forecasting) would be accurate mitigation.

According to Boland (1985), extrapolative methods are explained by the change in water use through time. Therefore, future water utilization is related to water use in the past, and those are the only two variables considered (Dekay, 1985). These methods consider the least of data, therefore making them easier for users. Extrapolation can be achieved in several ways that can either be simple or more complex. For instance, extrapolation can be done using graphical or mathematical methods, whereas change can be described by linear, exponential, logistic or as conforming to other relationships (Prasifka, 1988). An extrapolative method that is often used is the time series analysis, where the Box-Jenkins or ARIMA (AutoRegressive Integrated Moving Average) model is frequently applied (Chen, 1988; Quevedo et al., 1988).

Moreover, it is risky to assume that past trends will remain the same in the future. Therefore, time-series models might not give reliable predictions beyond a few years (Boland, 1985; McDonald and Kay, 1988). According to George (1985), there are two severe disadvantages when it comes to time-series models. The first one is that they are not accurate when great alterations in determining variables happen in the future, and they're not suitable to be used as policy tools. The second disadvantage is that their starting values carry the highest weight in the forecast. Therefore, they can be overly sensitive to the first data points (George, 1985).

Several variables are important for the water forecasting models, including the rainfall and runoff information, evaporation, interception, snowmelt, and catchment physical characteristics, depending on the context (Beven, 2001). For example, in Canada, snowmelt may be considered as the most important variable in terms of the annual maximum water discharge that increase the risk of flooding (Beven, 2001). Finally, the methods available use historical data from the past and the

present to forecast future changes. Forecasts are very useful and provide decision-makers with time to plan, execute, control, mitigate against possible disasters, and optimize solutions for future natural changes.

Indeed, the forecast model $\check{z}_t(l)$ over a future lead period l benefits decision-makers with a huge advantage in economic planning, production planning, production control, and optimization of the system concerned. The forecast function $\check{z}_t(l)$ estimates future periods based on current and previous periods available. The forecast function is optimized by selecting factors that minimize the mean square of the deviations $(z_{t+1} - \check{z}_t(l))$ between each lead time l and the actual value at time t. The prediction accuracy must also be calculated be setting probability limits.

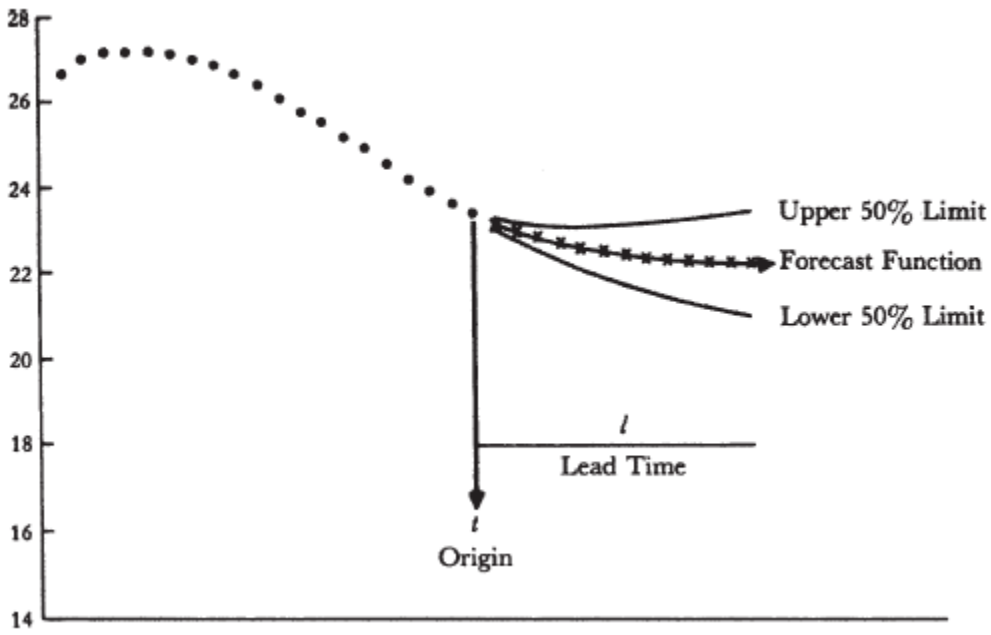


Figure 13.Values of a time series with forecast function and 50% probability limits

2.10 Application to the thesis

Discrete time-series are usually measured at fixed intervals to allow for proper analysis. The HYDAT database presents discrete time series of daily averages measured at constant daily intervals (Figure 13). Contrary to *deterministic* time series, future values of hydrological data (Water level, Water flow, Sediment concentration), such as in HYDAT, cannot be calculated by some exact mathematical function. These *Statistical time series* evolve in time according to probabilistic laws and are deemed to be *stochastic processes*. In these processes, observations are a realization of that random variable following a probability density function.

Stochastic Models comprise a very important Stationary time series that display constant statistical properties over time. Most statistical forecasting methods assume the data properties such as mean, variance, or autocorrelation can be rendered Stationary using appropriate mathematical transformations. By assuming a time series to be stationary due to the cyclic nature of the variable, the expectation is that its statistical properties will remain constant in the future. This is true for seasonal dependant observations such as water flows and water levels for example. Therefore, every year during the summer period, the flow will be higher than the yearly mean, and every year during winter we can reasonably expect the flow to be slower to the yearly mean.

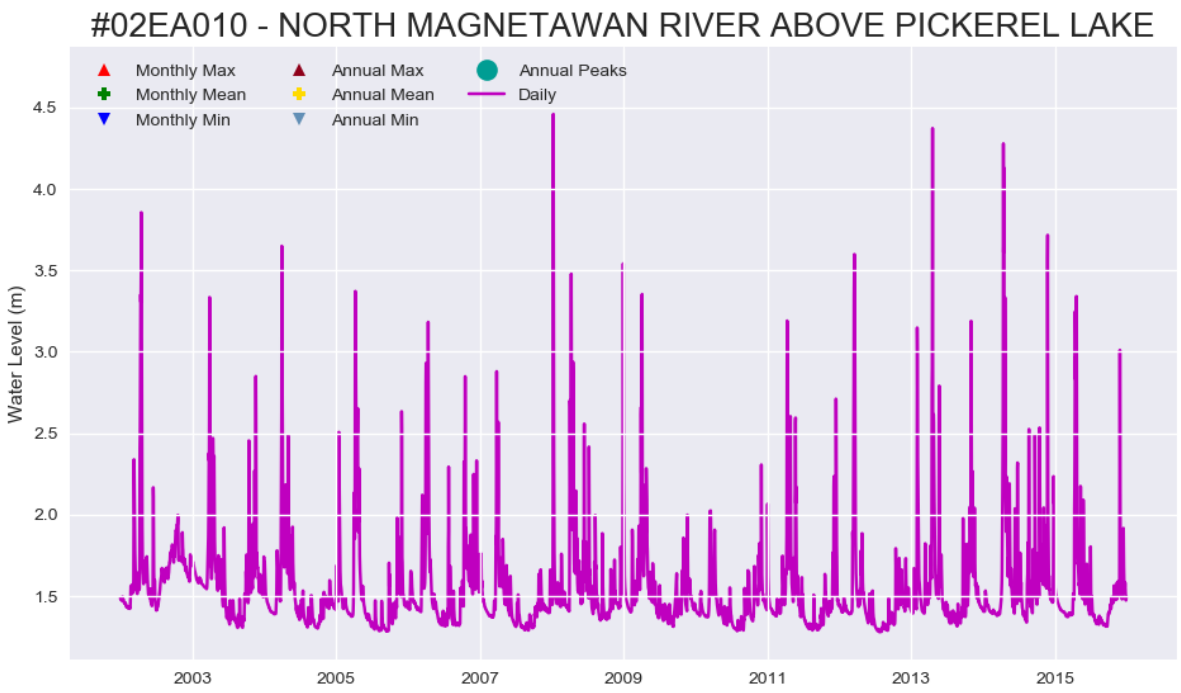


Figure 14.Daily Water Level measurements for Station #02EA010

This thesis will cover the forecasting of historical Environmental data around Canada. We will focus on identifying SARIMA models, estimating its parameters (p,d,q) through various methods such as Autocorrelation Functions (ACF) and Partial autocorrelation functions (PACF), determining the parameters' values based on. We assume that time series in the HYDAT database time series follows a seasonal pattern because of the nature of such data (Water Level, Water Flow, Sediment transported).

$$\varphi(y)(1-y)^d(Z_t-\mu)=\theta(y)a_t \tag{2.1}$$

where,

Z_t: *the observed series*

μ: *the series mean*

φ: *autoregressive polynomial of order p*

$\theta: moving\ average\ polynomial\ of\ order\ q$

$\varepsilon: random\ error\ term\ N(0, \sigma^2)$

The SARIMA model is the following:

$$y_i = \varphi_0 + \varphi_1 y_{i-1} + \varphi_0 y_{i-2} + \cdots + \varphi_p y_{i-p} + \varepsilon_i + \theta_1 \varepsilon_{i-1} + \cdots + \theta_q \varepsilon_{i-q} \quad (2.1)$$

Chapter 3 - Methodology

3.1 Data collection

Water Survey of Canada from *Environment Canada* has eight regional offices collecting hydrometric data across the country and stored in two centrally-managed databases (HYDEX and HYDAT). HYDEX contains information on streamflow's, water levels and sediment stations (active and discontinued stations) across Canada (Government of Canada, 2018c). It contains information about the stations (location, equipment, and types of data). HYDAT contains the actual data for all stations listed in HYDEX for various variables (daily and monthly means of flow, water levels and sediment concentrations) (Government of Canada, 2018c).

Water Survey of Canada collects and compiles three types of data for water currents around the country: water level, water flow, and water sediment concentration. Data are collected from gauging stations and recorded in the relational database HYDAT (Hydro Climatological Data Retrieval Program). This historical database is regularly updated and has measurements that go back as early as 1892 (late 19th Century). The time series recorded provides us with an unprecedented overview of the trends natural water bodies follow. The HYDAT data is collected for most water bodies daily when collection is possible. Therefore, we can have access to three time-series (Level vs. Time; Flow vs. Time; Sediment vs. Time) for each of the 7791 water bodies monitored. These time series can be analyzed to predict future values with controlled reliability.

For this study, all 7791 stations across Canada were selected from the Reference Hydrometric Basin Network (RHBN) (Water Survey of Canada, 2012) (Figure 15). Each station has a name and number, the period of record, province location, type of data available (e.g. flow, water level,

sediments loads…), latitude, longitude, and gross drainages area (km2) (Figure 16 and Figure 17) (Government of Canada, 2018). Time series data from all stations are showing graphically four types of data (as lines or points) on daily, monthly, annual and during peaks basis (Environment Canada, 2016).

1) **Flow**: Flow rate for a water body in m3/s,
2) **Level**: Water level in meters above station Datum,
3) **Suscon:** Suspended material concentration in mg/l,
4) **Load**: Sediment load in tons.

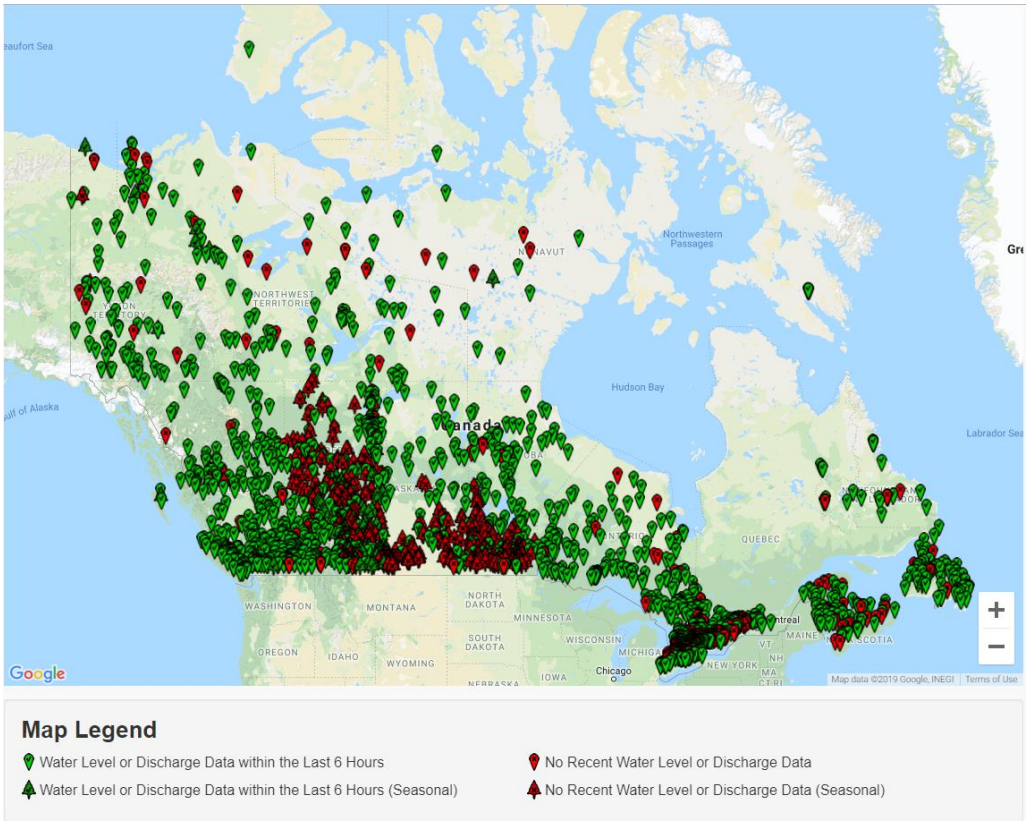


Figure 15.Real-time Hydrometric Data Search Results Location of 7791 HYDAT Stations in Canada (Water Survey of Canada, 2018)

Check All	Station Name	Years	Province	Station Number	Data Availability	Latitude	Longitude	Gross Drainage Area (km^2)
	ABENAQUIS (RIVIERE DES)	1972-1978	QC	02PJ035	Flow	46°13'24" N	70°31'17" W	152
	ABERCROMBY (RIVIERE)	1975-1986	QC	02OB036	Flow	45°51'57" N	73°54'02" W	58.5
	ABERDEEN LAKE AT THE OUTLET	1968-1986	BC	08LC043	Level	50°06'17" N	119°04'23" W	
	ABITAU RIVER ABOVE CUMING LAKE	1988-2017	SK	07QC005	Flow and Level	59°59'58" N	108°46'26" W	3,780
	ABITIBI (LAC) À MANCEBOURG	1946-2013	QC	04MA004	Level	48°45'00" N	79°13'40" W	1,840

Figure 16.Historical Hydrometric Data Search Results. Example of 5 HYDAT stations with their respective water data (level and flow) (Government of Canada, 2018)

Check All	Station Name	Years	Province	Station Number	Data Availability	Latitude	Longitude	Gross Drainage Area (km^2)
	ABBOTT RIVER ABOVE MIDSHIPMAN BAY	1975-1975	NU	10VC011	Concentration, Instantaneous	75°14'00" N	95°41'00" W	191
	ABITIBI RIVER AT ONAKAWANA	1978-1994	ON	04ME003	Loads, Concentration, Instantaneous	50°36'10" N	81°24'52" W	27,500
	AISHIHIK RIVER NEAR WHITEHORSE	1972-1972	YT	08AA001	Loads, Concentration, Instantaneous	60°51'40" N	137°03'40" W	4,300
	ALBANY RIVER ABOVE NOTTIK ISLAND	1983-1983	ON	04GD001	Loads, Concentration, Instantaneous	51°38'29" N	86°23'27" W	32,400
	ALBANY RIVER NEAR HAT ISLAND	1975-1994	ON	04HA001	Loads, Concentration, Instantaneous	51°19'50" N	83°50'00" W	118,000

Figure 17.Historical Hydrometric Data Search Results. Example of 5 HYDAT stations with their respective sediment data (Government of Canada, 2018b)

3.2 Data variables

3.2.1 Streamflow

Streamflow is defined as the water quantity flowing past a point in a unit of time in a river. The units are either liters/second (l/s) or cubic meters per second (m^3/s) (Water Survey of Canada, 2012). Measurements are done periodically to generate time series. These measurements can be

done from a bridge, by boat, by stream wading, or by using a cableway strung (Water Survey of Canada, 2012).

3.2.2 Water levels

Water levels are recorded continuously at each station using a mechanical or electronic recorder (Water Survey of Canada, 2012). Several measurements of water depth and velocity are required across the section of a given water body to measure its flow rate; the measurements are used to get the average discharge (Water Survey of Canada, 2012).

3.2.3 Sediments

Water bodies' sediment load is displayed in tonnes per day (t/d), while the sediment concentration is measured in milligrams per liter (mg/L) (Environment Canada, 2001)

3.3 Model building

3.3.1 Stationarity checking

The Mann-Kendall test is useful in determining whether or not a linear monotonic trend is present in time series. This reliable test is used generally in climatological, hydrological and environmental time series. According to Pohlert (2018), this test is a non-parametric trend similar to Kendall's correlation coefficient concept. The null hypothesis (H_0) indicated a no monotonic trend which is tested against three alternative hypotheses (H_a):

1) Presence of a monotonic upward trend
2) Presence of a monotonic downward trend
3) Presence of either a monotonic upward trend or a monotonic downward trend

According to Mann (1945), Kendall (1975) and Gilbert (1987), there are several assumptions regarding the data when using the Mann-Kendall test:

1) Data is distributed independently and identically when no trend is present;
2) Measurements are the observables true states at specific times of measurements;
3) Methods used when collecting samples, measuring, and handling data are unbiased.

There are several advantages when using the Mann-Kendall test:

1) No assumptions about the distribution of the data. As such, there is no requirement for the data to be normally distributed.
2) Missing data do not affect the test. However, when the number of of sample points is reduced, the statistical significance is negatively affected;
3) Time series length and irregular spacing of measurement's time points do not affect the test.

However, there are a few limitations regarding the Mann-Kendall test:

1) Data with periodicities cannot be handled by the test. All periodic effects in the data must be removed before doing the Mann-Kendall test.
2) Shorter datasets tend to get more negative results. It is recommended to have a longer time series for a more effective trend detection computation.

3.3.1.1 Application of the Mann-Kendall test

Step 1: Computing the indicator function ($sgn(x_i - x_j)$) using a time series of length n as follow:

$$sgn(x_i - x_j) = \begin{cases} 1 & x_i - x_j > \varepsilon \\ 0 & |x_i - x_j| \leq \varepsilon \\ -1 & x_i - x_j < -\varepsilon \end{cases}$$

This indicator function indicates if the measurements at time *i* and *j* are negative, positive, or equal to zero.

Step 2: Computing the mean and variance of the quantity calculated in step 1. The mean (*E[S]*) is given by equation 1, and the variance (*VAR(S))* is given by equation 2 where *p* is the data total number of tie groups, and *qk* is the total number of data point the kth tie group. As an example, a time series with the following measurements: {16, 58, 23, 16, 67, 45, 58, 58, 10}, has two tie groups for the measurements {16} and {58} that is p=2, with the number of data points being q1=2 for {16} and q2=3 for {58}.

$$E[S] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_i - x_j)$$

Equation 3.1

$$VAR\ (S) = \frac{1}{18}(n(n-1)(2n+5) - \sum_{k=1}^{p} qk(qk-1)(2qk+5))$$

Equation 3.2

The Man-Kendall test statistic is computed using the E[S] and VAR(S) to ensure for samples with larger datasets a normal or almost normal distribution of the test statistic Z_{MK}, as follow:

$$Z_{MK} = \begin{cases} \dfrac{E[S]-1}{\sqrt{VAR(S)}}, & E[S] > 0 \\ 0, & E[S] = 0 \\ \dfrac{E[S]+1}{\sqrt{VAR(S)}}, & E[S] < 0 \end{cases}$$

Equation 3.3

3.3.1.2 Hypothesis testing:

The Type I error rate (significance level α of Man-Kendall test) allows the user to determine whether to reject or accept the alternative hypothesis (H_a) for all variants of (H_a) independently as follow:

1) Ha= An monotonic upward trend exists:

   In that case, if $Z_{MK} \geq Z_{1-\alpha}$, Ha is accepted, where the $Z_{1-\alpha}$ represents the $100(1-\alpha)^{th}$ percentile of the standard normal distribution.

2) Ha= A monotonic downward trend exists:

   In that case, if $Z_{MK} \leq -Z_{1-\alpha}$, Ha is accepted.

3) Ha = An upward or downward monotonic trend exists:

   In that case, if $|Z_{MK}| \geq Z_{1-\alpha/2}$, Ha is accepted, where the notation $|\cdot|$ is the absolute value function.

3.3.2 Using SARIMA models

One of the most common methods used in time series forecasting is known as the ARIMA model, which stands for Autoregressive Integrated Moving Average. ARIMA is a model that can be fitted to time series data to better understand or predict future points in the series. There are three distinct integers (p, d, q) that are used to parametrize ARIMA models. Because of that, ARIMA models are denoted with the notation ARIMA (p, d, q). Together these three parameters account for seasonality, trend, and noise in datasets:

- p is the autoregressive part of the model. It allows us to incorporate the effect of past values into our model. Intuitively, this would be similar to stating that it is likely to be warm tomorrow if it has been warming the past three days.
- d is the integrated part of the model. This includes terms in the model that incorporates the amount of differencing (i.e. the number of past time points to subtract from the current value) to apply to the time series. Intuitively, this would be similar to stating that it is likely to be same temperature tomorrow if the difference in temperature in the last three days has been very small.
- q is the moving average part of the model. This allows us to set the error of our model as a linear combination of the error values observed at previous time points in the past.

When dealing with seasonal effects, we make use of the seasonal ARIMA, which is denoted as ARIMA (p,d,q)(P,D,Q)s. Here, (p, d, q) are the non-seasonal parameters described above, while (P, D, Q) follow the same definition but are applied to the seasonal component of the time series. The term s is the periodicity of the time series (4 for quarterly periods, 12 for yearly periods, etc.).

The seasonal ARIMA method can become daunting because of the multiple tuning parameters involved. We will describe the procedure used to automate the process of identifying the optimal set of parameters for the seasonal ARIMA time series model.

The general ARIMA model can be expressed by Equation-1 (Vandaele 1983)

$$\varphi(B)x_t = \theta(B)a_t$$

Where:

$\varphi(B) = 1 - \varphi_1 B^1 - \varphi_2 B^2$ (Non-seasonal autoregressive polynomial)

B (Backward shirt operator in the equation)

φ_p(Model's autoregressive parameters)

P(Order of autoregressive polynomial)

x_t(Stationary series after defencing)

$X_{t=1-\varphi_1 B^1-\varphi_2 B^2}$ (Stationary series after defencing)

X_t(Dependent variable)

$\theta(B) = 1 - \theta_1 B^1 - \theta_q B^2$ (Non-seasonal moving average polynomial)

θq (Moving average of the model)

q(Order of moving average polynomial)

at (White noise process)

The number of parameters p, n and q determine the order of the model ARMA (p, n, q)x(P, N, Q)s. The SARIMA model, on the other hand, considers seasonality trends in the data. Since water flow, water level, and sediment content in water bodies is seasonal dependent, SARIMA can be very useful in improving the model. This model can be either multiplicative or non-multiplicative. The general form of SARIMA model in its multiplicative form is described by equation-x:

$$\varphi(B)\Phi(B^s)x_t = \theta(B)\Theta(\mathrm{B}^s)\mathrm{a_t}$$

Where,

$$x_t = (1 - B)^d (1 - B^s)^D X_t$$

$$D\ is\ the\ number\ of\ seasonal\ differencing$$

$$\Phi(\mathrm{B}^s) = 1 - \Phi_1 B^s - \Phi_1 B^s - \cdots$$

$$- -\Phi_\mathrm{p} B^{Ps} is\ the\ seasonal\ autoregressive\ polynomial$$

Streamflow forecasting is an important indicator used in water resources management. It can be often critical for planning purposes. Medium forecasting at weekly, or monthly scales or long-term forecasting at seasonal or even annual time scales can be particularly useful in guiding decision-makers towards informed and efficient strategies in the operation of reservoirs, irrigation management, or the legal aspects of water resources management and planning. For these purposes, a large number of forecasting methods have been developed over the years to enhance the accuracy and reliability of results. Statistical Methods such as time series models, regression models, artificial neural network models, fuzzy logic and nearest-neighbor model have been developed and tested in various fields of study and data.

Time series analysis has been the most popular in the last decades due to its systematic model building process. Box & Jenkins (1976) defined a three-step methodology for building time series models: identification of the model, estimation of parameters, and diagnostic check to correct or improve the model. This iterative process has proved often to be the most efficient and reliable model in time series forecasting.

In time series analysis, there are two categories: univariate and multivariate models. Univariate models are those dealing with a single time series. The Autoregressive Integrated Moving Average

model, known as ARIMA and its various derivatives such as the Seasonal ARIMA (SARIMA), deseasonalized ARIMA, and periodic ARIMA has been used extensively in modeling and forecasting monthly streamflow data (Noakes et al. 1985, Abrahart, 2000).

Time series models used in streamflow forecasting are mostly linear models, they assume a normally distributed process, although most streamflow processes are nonlinear (Wang 2006). Therefore, it is of interest to generate Seasonal ARIMA models for a large number of stream flows to assess the efficiency and reliability of this time series analysis model.

3.3.3 Generating the best factors (overfitting or not)

Akaike information criterion (AIC) by Akaike (1973)

The Akaike information criterion (AIC) is very useful for model selection (Akaike, 1973). Seasonal ARIMA model selection is frequently made by using AIC to select a model offering a good estimate while having a minimum of parameters. Adding additional parameters leads to model's penalization by the AIC. When comparing a set of models for a given data, a criterion is estimated by the AIC which is the reference model minimum value. This value is calculated using the maximized value of the likelihood function as follow (Posada, 2004):

$$AIC = -2l + 2k$$

where k is the model's number of parameters, and l is the log-likelihood estimate of the selected model ($l = \log(L)$) with L as the likelihood function. Thus, the best model is selected based on the AIC minimal value obtained. For models having almost identical values, the following formula can be used for further comparison:

$$\Delta AIC = \Delta_i = - AIC_{min} + AIC_i$$

where AIC_{min} is the model's lowest AIC value, and AIC_i is the i^{th} model value (model of interest). Significant evidence of the model interest is shown when $\Delta_i < 2$ (Burnham and Anderson, 2002). However, the model of interest has less support when three $\leq \Delta_i \leq 7$, while $\Delta_I > 10$ show important support against the model of interest (Burnham and Anderson, 2002).

In other words, the model with the lowest AIC value is to be selected as it describes best the data analyzed. Adding more parameters increases the uncertainty of parameter estimation. Thus, the number of parameters and the enhanced fit should balance one another to minimize the AIC value.

For this study, to select the best fit seasonal ARIMA, the Akaike information criterion (AIC) will be performed. For example, between two seasonal ARIMA models (e.g., Model A with $\Delta_I = 9.5$ and Model B with $\Delta_I = 1.5$), the model to select would be Model B with $\Delta_I = 1.5$.

3.3.4 Fitting the data

The minimum number of factors is to be used, and residuals must follow a normal distribution. Otherwise, it would mean that the model is overfitting the data and will produce unreliable forecasts. The general trend of the time series must be captured instead of the exact data trend.

3.3.5 Forecast over a known period and verification of the result

For any statistical model, the bare minimum data required for fitting depends on 1) the number of parameters to estimate and 2) the number of random variations in the data. Statistical perspective, it is important to have a higher number of observations than parameters (Hanke, Reitsch and Wichern, 1998). It is important to mention that using this minimum number of data required does not guarantee adequate estimates of seasonality.

When forecasting using seasonal models, the number of data required depends on the model type and the number of randomness in the data (Hanke, Reitsch and Wichern, 1998). There are no strict rules about the sample size to use, and the minimum required size is only used when the number of random variations in the data is small. In practice, data usually contains a high number of randomness causing the size of sample requirements to increase consequently (Hanke, Reitsch and Wichern, 1998).

According to Makridakis, Wheelwright, and Hyndman (1998), seasonal ARIMA models use p+q+P+Q parameters, and d+mD, if differencing is required. As an example, Box, Jenkins and Reinsel (1994) detailed a monthly model ("airline" model) ARIMA (0,1,1)(0,1,1)12 which contains 15 parameters (0+1+0+1+1+12). Consequently, 16 is the required minimum number of observations as p+q+P+Q+d+mD+1 are required to estimate seasonal ARIMA models.

Some researchers have pointed out some minimal data length required for several forecasting models (Hanke, Reitsch and Wichern, 1998), while others oppose the idea assuming that these values are ignoring the underlying data variability. Also, short time series makes reliable

forecasting challenging for several reasons; an example of such reasons is the difficulty of choosing starting values (Hanke, Reitsch and Wichern, 1998).

One approach is to verify if there are sufficient data to accurately estimate the model. Then, it is important to test the model and verify if it performs well on data excluded from the sample. We verify the results by various means such as Mean Square (MSE), Root Mean Squared Error (RMSE), and model efficiency E as defined by Nash and Sutcliffe (1970).

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})\,(F_i - \bar{F})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(F_i - \bar{F})^2}\right]^2$$

The Root Mean Squared Error (RMSE) is a good indication of the average discrepancy between the forecast values and the actual data recorded. The RMSE is calculated based on Equation-X:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2}$$

We will calculate the model efficiency as defined by Nash and Sutcliffe (1970).

$$E = 1 - \frac{\sum(Y_i - F_i)^2}{\sum(Y_i - \bar{Y})^2}$$

It has been established that to provide satisfactory performance, and the model efficiency must be 90% or above. If the percentage ranges between 80 to 90%, the performance is rather good, and if it is under 80%, it is rather questionable and not satisfactory (Nash and Sutcliffe, 1970).

Chapter 4 – Results and discussion

The main results of this analysis are the development of a tool (called universal forecast application and presented in Appendix 1) that's able to generate short to medium-term forecasts of three variables from the HYDAT database. The tool can analyze all 7791 stations in the database. In the next sections, we are going to present the analysis flowchart for particular stations displaying stationary, and non-stationary behavior with examples of results that are generated, then assess the quality of the forecasts at 7791 these stations

4.1 Analysis flowchart

4.1.1 Data visualization

Time series provide the opportunity to forecast future values. Based on previous values, time series can be used to forecast trends in economics, weather, and capacity planning, to name a few. The specific properties of time-series data mean that specialized statistical methods are usually required.

We will aim to produce reliable forecasts of time series in the HYDAT database. We will begin by analyzing and discussing the concepts of autocorrelation, stationarity, and seasonality for the chosen station data (Rideau river, station #01AD003), and proceed to apply one of the most commonly used methods for time-series forecasting: the seasonal ARIMA method.

One of the methods available in Python to model and predict future points of a time series is known as SARIMAX, which stands for Seasonal Autoregressive Integrated Moving Averages with exogenous regressors. Here, we will primarily focus on the ARIMA component, which is used to fit time-series data to better understand and forecast future points in the time series.

We will study the historical data of the Rideau River of Ottawa (Station #02LA004).

- Water Level average monthly data in meters
- Water Flow average monthly data in m^3/s
- Sediment flow average monthly data in

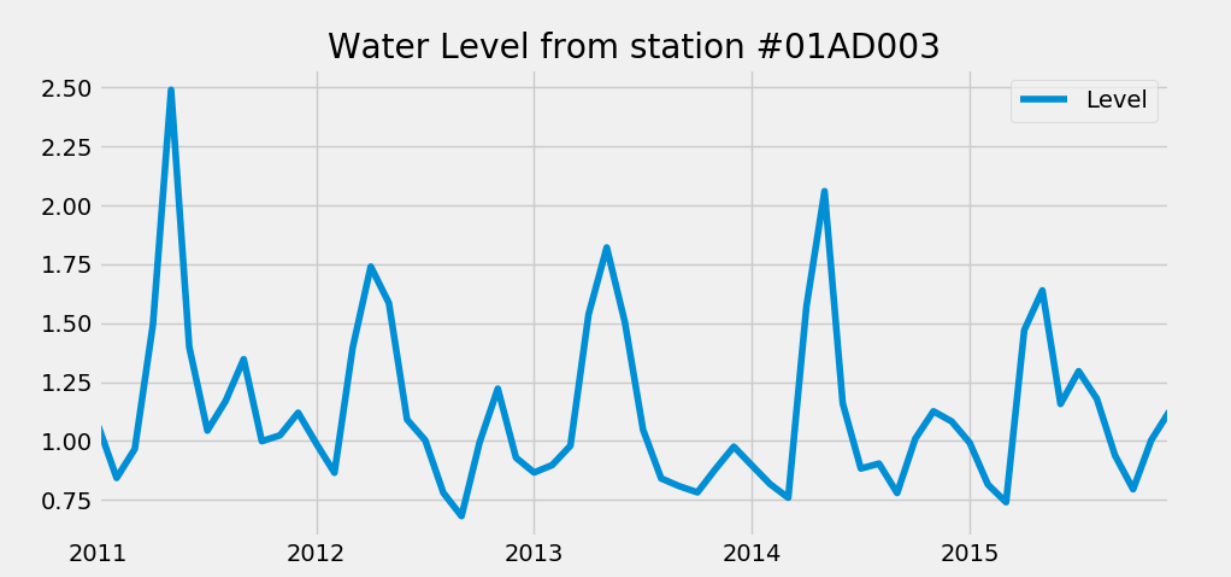


Figure 18.Water Level data in m for Station #01AD003 – Rideau river

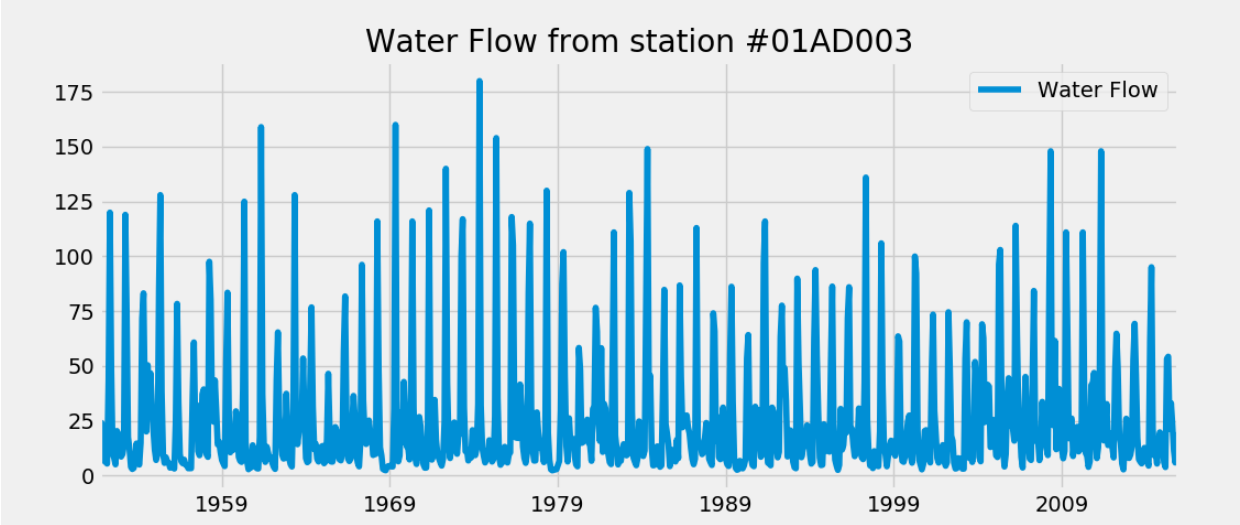


Figure 19.Flow monthly data in m3/s for Station #01AD003 – Rideau river

4.1.2 Preliminary analysis

We can better visualize our time series data by breaking it down into Trend, seasonal, and residual components (fig-3.3). We observe that we have almost stationary data from the Trend component, and the residuals seem to be normally distributed across the recorded data.

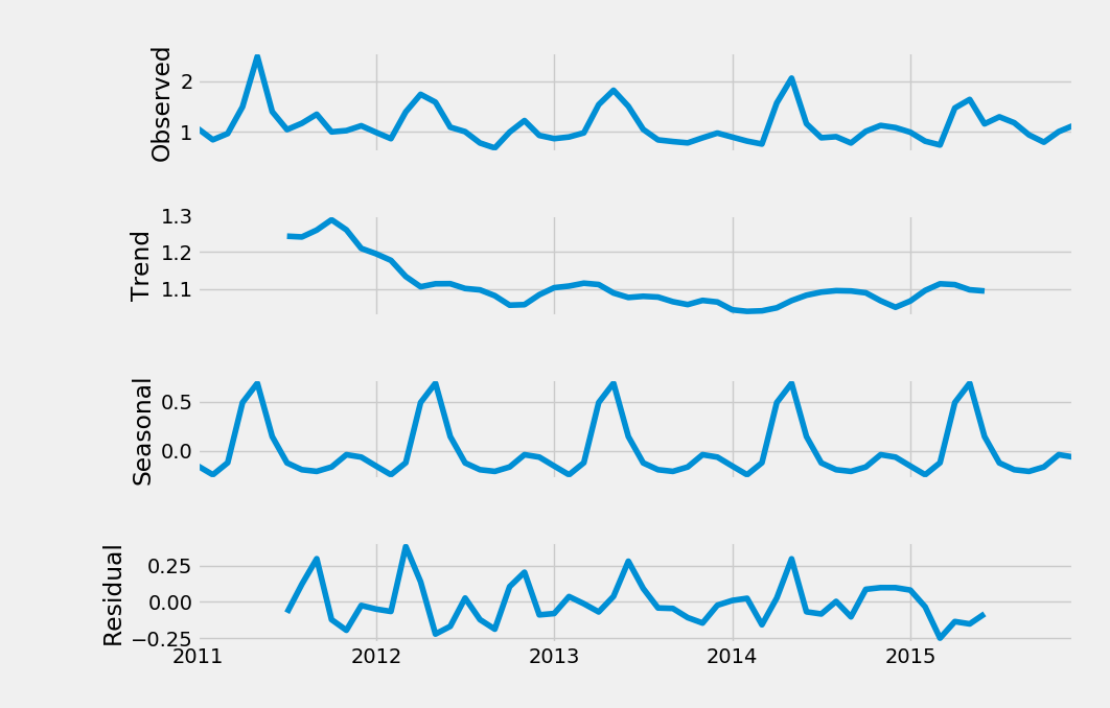


Figure 20. Observed time series data are broken down into Trend, Seasonal and Residual components

4.1.3 Parameters selection/fine-tuning

When looking to fit time series data with a seasonal ARIMA model, our first goal is to find the values of ARIMA(p,d,q)(P,D,Q)s that optimize a metric of interest. There are many guidelines and best practices to achieve this goal, yet the correct parametrization of ARIMA models can be a painstaking manual process that requires domain expertise and time. Other statistical programming languages such as R provide automated ways to solve this issue, but those have yet to be ported

over to Python. In this section, we will resolve this issue by writing Python code to programmatically select the optimal parameter values for our ARIMA(p,d,q)(P,D,Q)s time series model.

We will use a "grid search" to iteratively explore different combinations of parameters. For each combination of parameters, we fit a new seasonal ARIMA model with the SARIMAX () function from the statsmodels python module and assess its overall quality. Once we have explored the entire landscape of parameters, our optimal set of parameters will be the one that yields the best performance for our criteria of interest. Let's begin by generating the various combination of parameters that we wish to assess:

We can now use the triplets of parameters defined above to automate the process of training and evaluating ARIMA models on different combinations (Table-3.1). In Statistics and Machine Learning, this process is known as grid search optimization for model selection.

Table 3.Seasonal ARIMA parameter combinations

ARIMA(p,d,q) combinations considered	Seasonal parameters (P, D, Q, S) considered
(0, 0, 0)	(0, 0, 0, 12)
(0, 0, 1)	(0, 0, 1, 12)
(0, 1, 0)	(0, 1, 0, 12)
(0, 1, 1)	(0, 1, 1, 12)
(1, 0, 0)	(1, 0, 0, 12)
…	…

When evaluating and comparing statistical models fitted with different parameters, each can be ranked against one another based on how well it fits the data or its ability to accurately predict future data points. We will use the AIC (Akaike Information Criterion) value, which is conveniently returned with ARIMA models fitted using statsmodels python module. The AIC measures how well a model fits the data while taking into account the overall complexity of the model. A model that fits the data very well while using lots of features will be assigned a larger AIC score than a model that uses fewer features to achieve the same goodness-of-fit. Therefore, we are interested in finding a model that yields the lowest AIC value.

Our procedure consists of iterating through combinations of parameters and uses the SARIMAX function from statsmodels to fit the corresponding Seasonal ARIMA model. Here, the order argument specifies the (p, d, q) parameters, while the seasonal order argument specifies the (P, D, Q, S) seasonal component of the Seasonal ARIMA model. After fitting each SARIMAX ()model, the AIC score is generated, and the best combination of ARIMA and Seasonal parameters is selected based on the lowest value returned.

Table 4.Akaike Information Criterion (AIC) values for each seasonal ARIMA parameters combination

Seasonal ARIMA parameters	AIC value
ARIMA(0, 0, 0)x(0, 0, 1, 12)12	+6787.3
ARIMA(0, 0, 0)x(0, 1, 1, 12)12	+1596.7
ARIMA(0, 0, 0)x(1, 0, 0, 12)12	+ 1058.9
…	…
ARIMA(1,1,1)x(1, 1, 1, 12)12	+277.8

…	…

The output of our code suggests that SARIMAX (1, 1, 1)x(1, 1, 1, 12) yields the lowest AIC value of 277.78. We should, therefore, consider this to be optimal option out of all the models we have considered.

4.1.4 Fitting the seasonal ARIMA model

Using a grid search, we have identified the set of parameters that produces the best fitting model for our time series data. We can proceed to analyze this particular model in more depth.

We'll start by plugging the optimal parameter values into a new SARIMAX model and generating the summary of the parameters in Fig-3.4:

Table 5.Seasonal ARIMA parameters summary output

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3182	0.092	3.441	0.001	0.137	0.499
ma.L1	-0.6254	0.077	-8.162	0.000	-0.776	-0.475
ar.S.L12	0.0010	0.001	1.732	0.083	-0.000	0.002
ma.S.L12	-0.8769	0.026	-33.812	0.000	-0.928	-0.826
sigma2	0.0972	0.004	22.633	0.000	0.089	0.106

The summary attribute that results from the output of SARIMAX returns a significant amount of information, but we'll focus our attention on the table of coefficients. The coefficient column shows the weight (i.e., importance) of each feature and how each one impacts the time series. The P>|z| column informs us of the significance of each feature weight. Here, each weight has a p-value lower or close to 0.05, so it is reasonable to retain all of them in our model.

When fitting seasonal ARIMA models, or any other models for that matter, it is important to run model diagnostics to ensure that none of the assumptions made by the model have been violated.

We can generate the diagnostics plots allows us to quickly generate model diagnostics and investigate for any unusual behavior (fig-4.4)

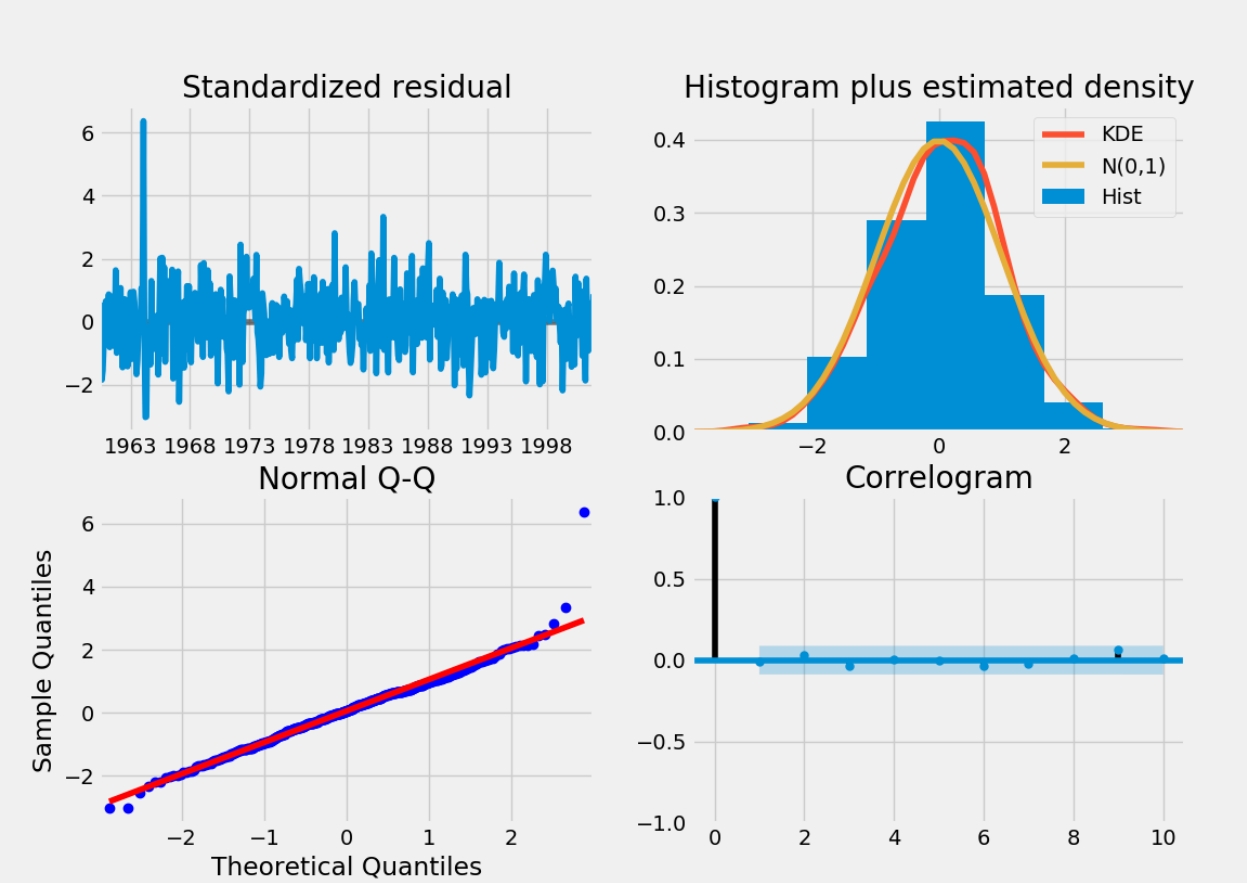


Figure 21.Seasonal ARIMA parameters summary output

Our primary concern is to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved.

In this case, our model diagnostics suggest that the model residuals are normally distributed based on the following:

- In the top-right plot, we see that the red KDE line follows closely with the N(0,1) line (where N(0,1)) is the standard notation for a normal distribution with mean 0 and standard deviation of 1). This is a good indication that the residuals are normally distributed.
- The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1). Again, this is a strong indication that the residuals are normally distributed.
- The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e., correlogram) plot on the bottom right, which shows that the time series residuals have a low correlation with lagged versions of itself.

Those observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values.

Although we have a satisfactory fit, some parameters of our seasonal ARIMA model could be changed to improve our model fit. For example, our grid search only considered a restricted set of parameter combinations, so we may find better models if we widened the grid search.

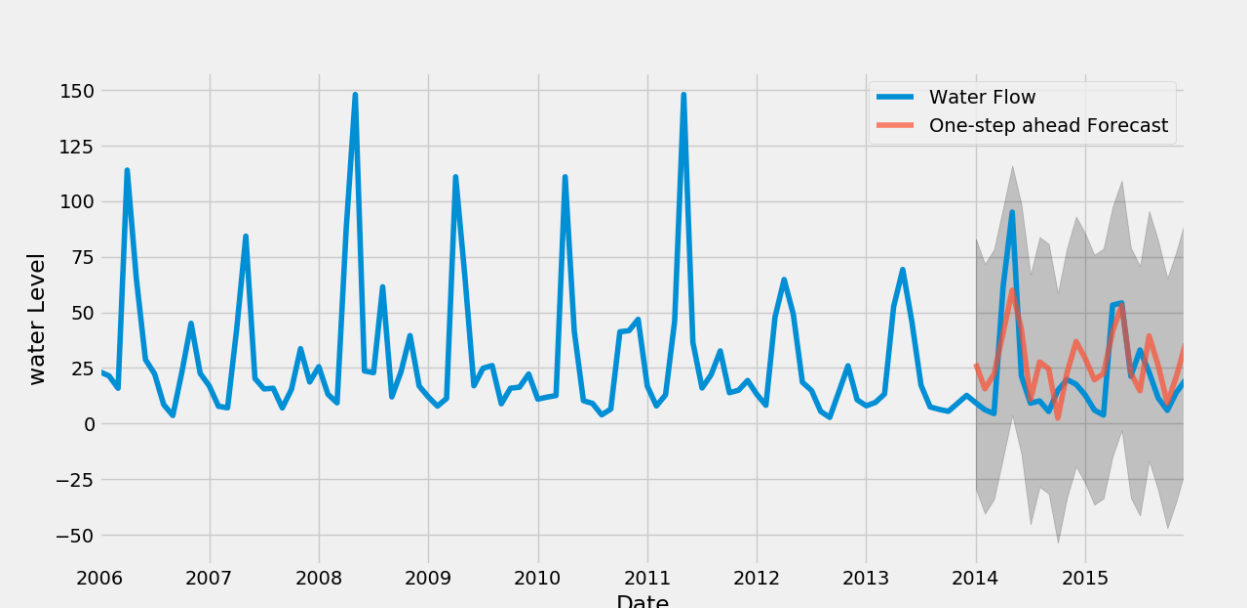


Figure 22.Seasonal SARIMAX (1,1,1,4)x(1,1,1) model with 4 months period

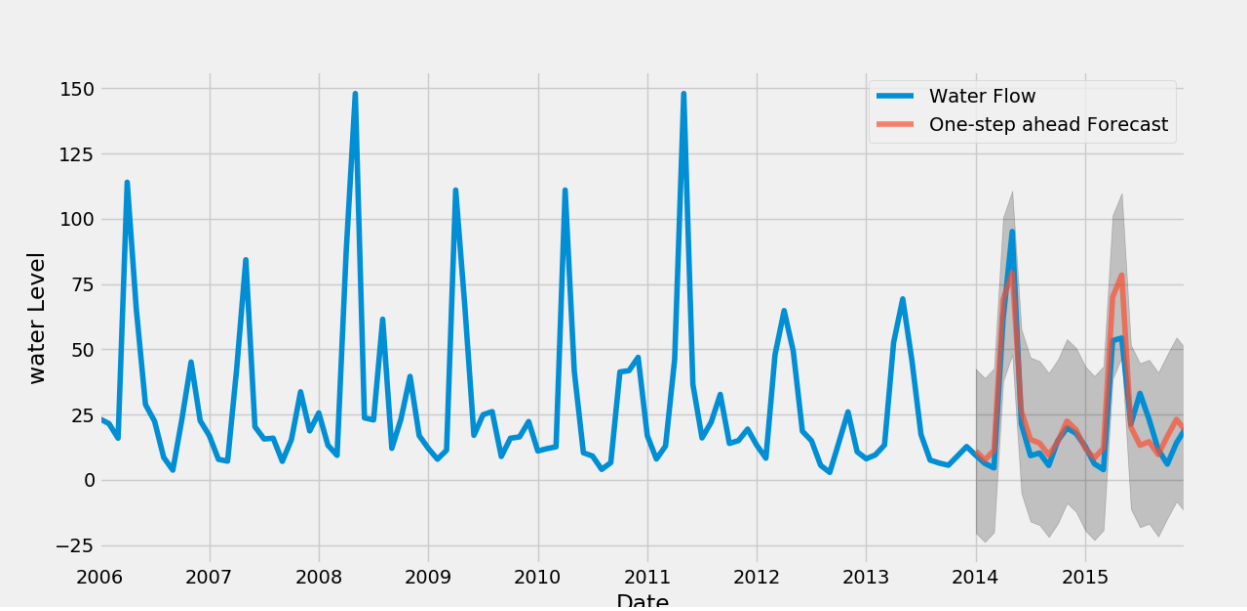


Figure 23.Seasonal SARIMAX (1,1,1,12)x(1,1,1) model with 12 months period

4.2 Forecast validation at particular stations

We have obtained a model for our time series that can now be used to produce forecasts. We start by comparing predicted values to real values of the time series, which will help us understand the

accuracy of our forecasts. The get_prediction() and conf_int() attributes allow us to obtain the values and associated confidence intervals for forecasts of the time series. We were able to test our fitting model over 20 different stations (Figure 27 to 46). The model fitting efficiency range between 0.89 and 0.92, which indicates a strong likelihood to generate a good forecast.

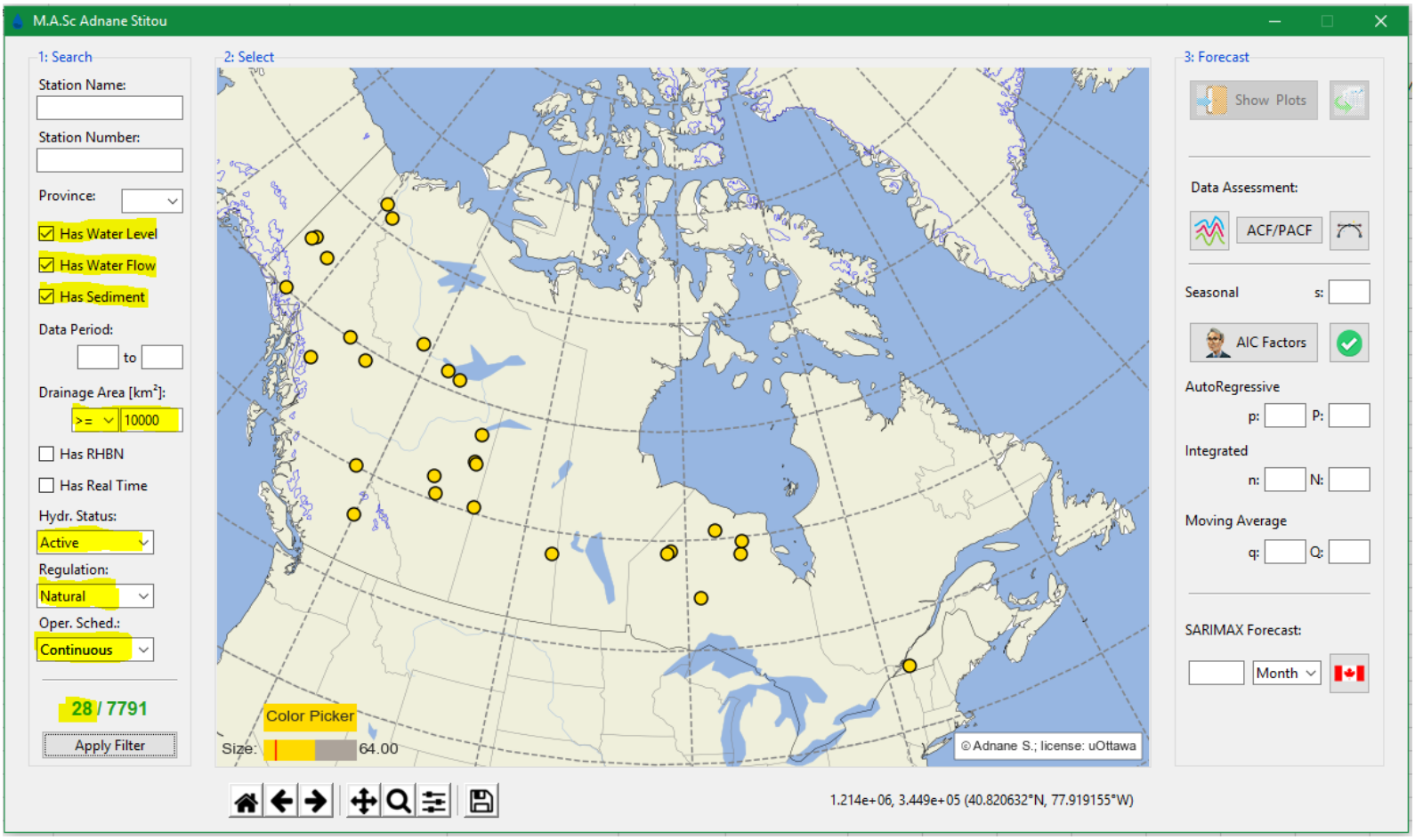


Figure 24.Canada 2

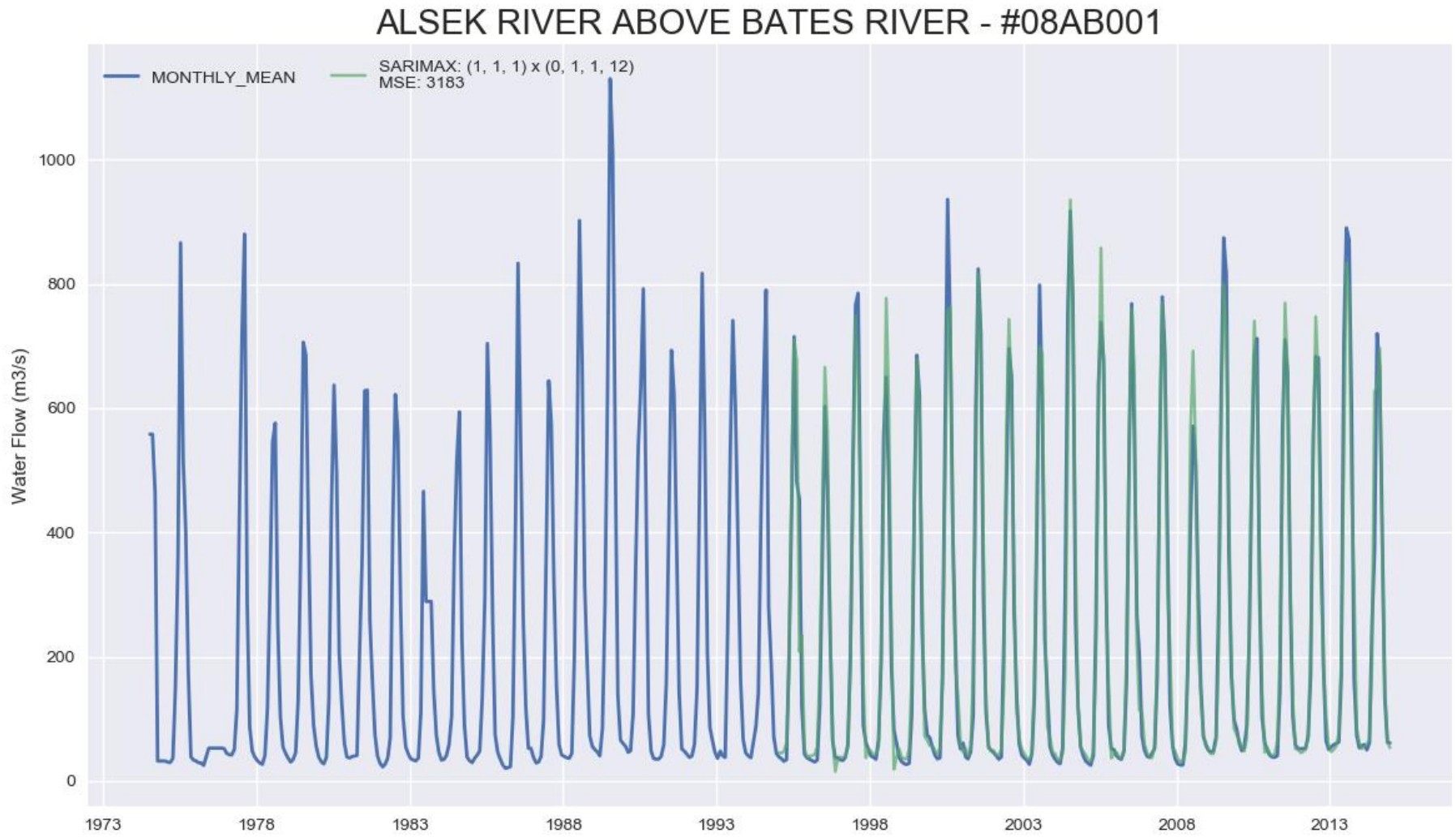


Figure 25. Alsek River above bates river

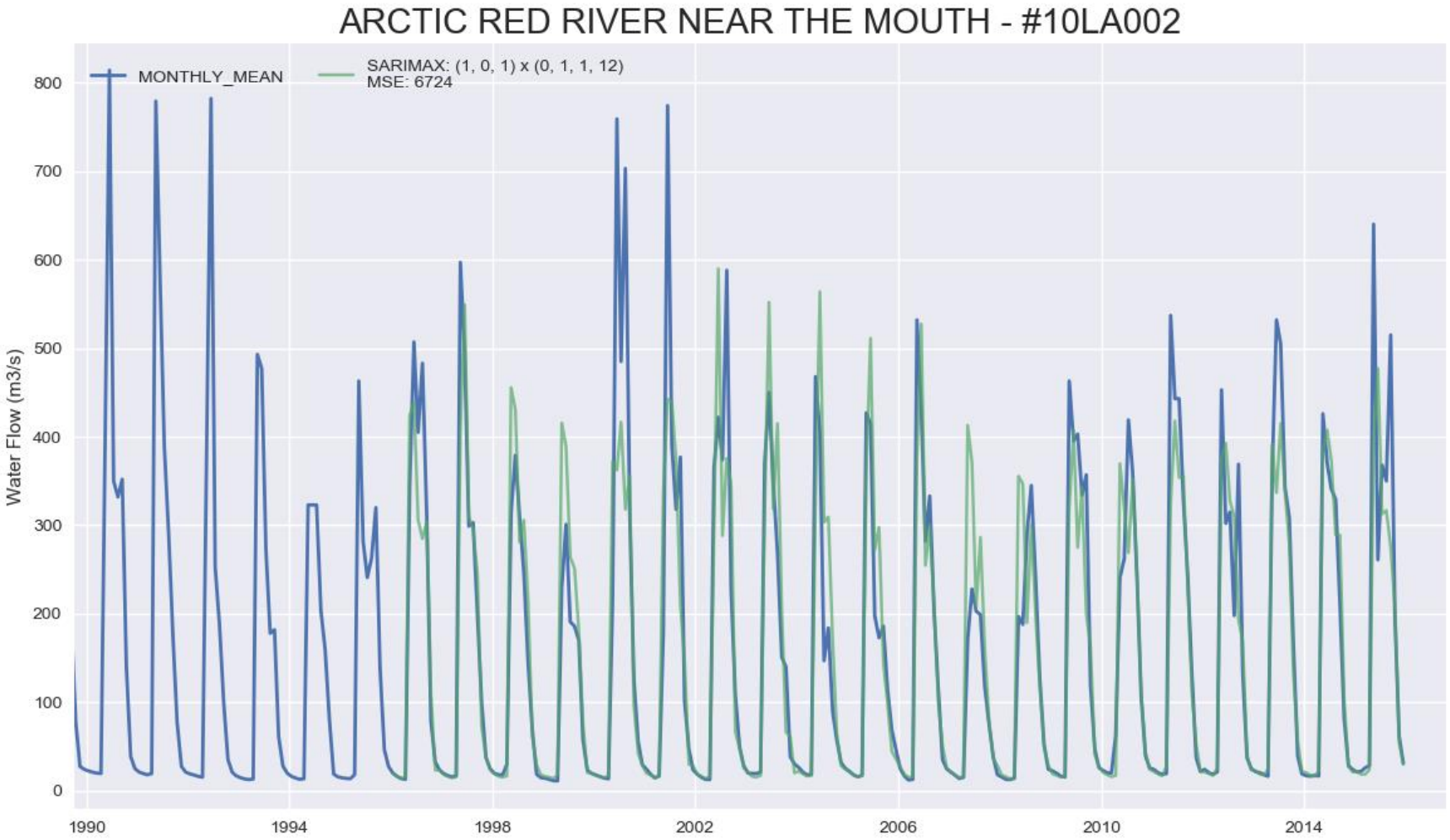


Figure 26.Arctic red river near the mouth

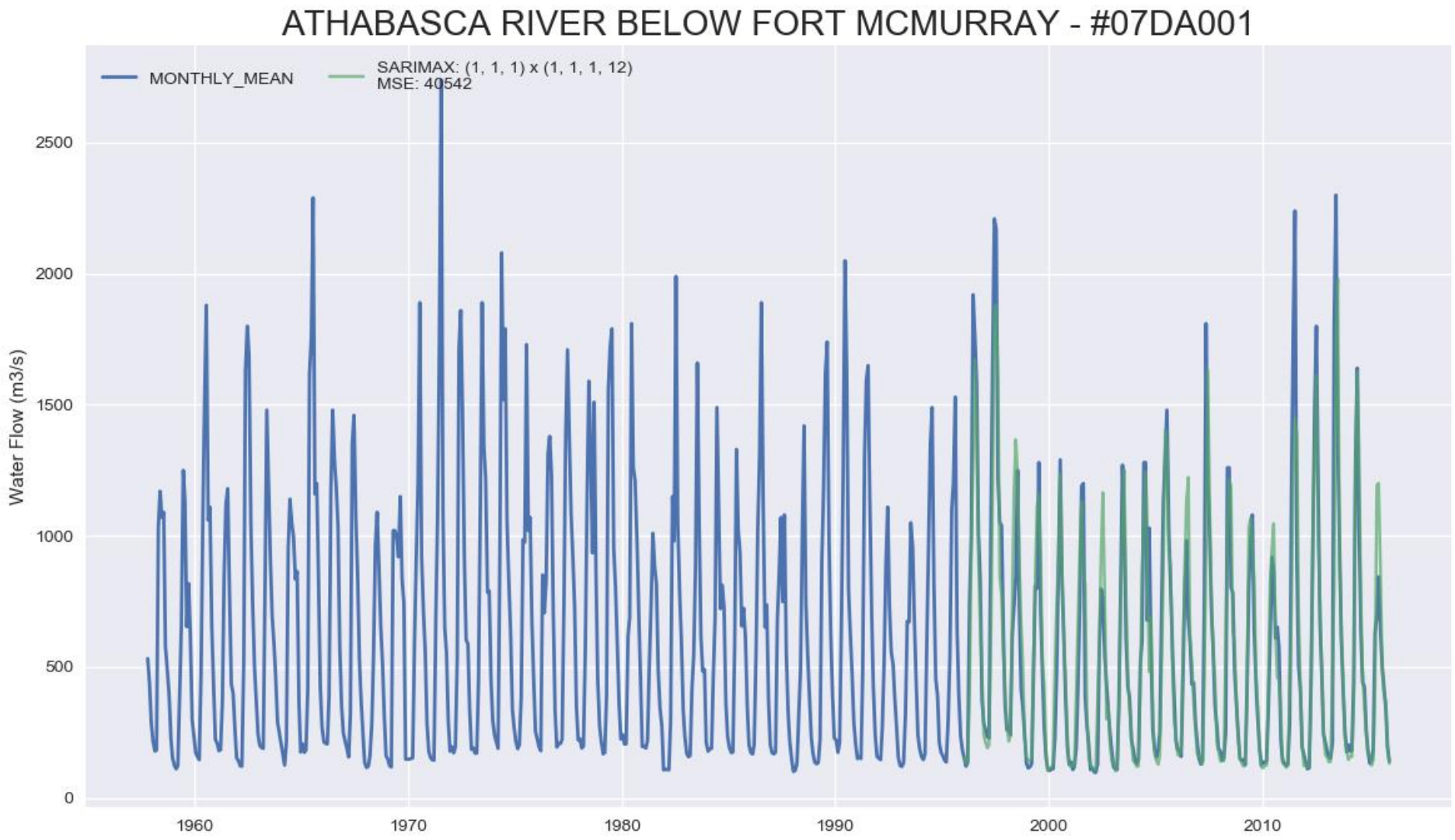


Figure 27.Athabasca river below fort McMurray

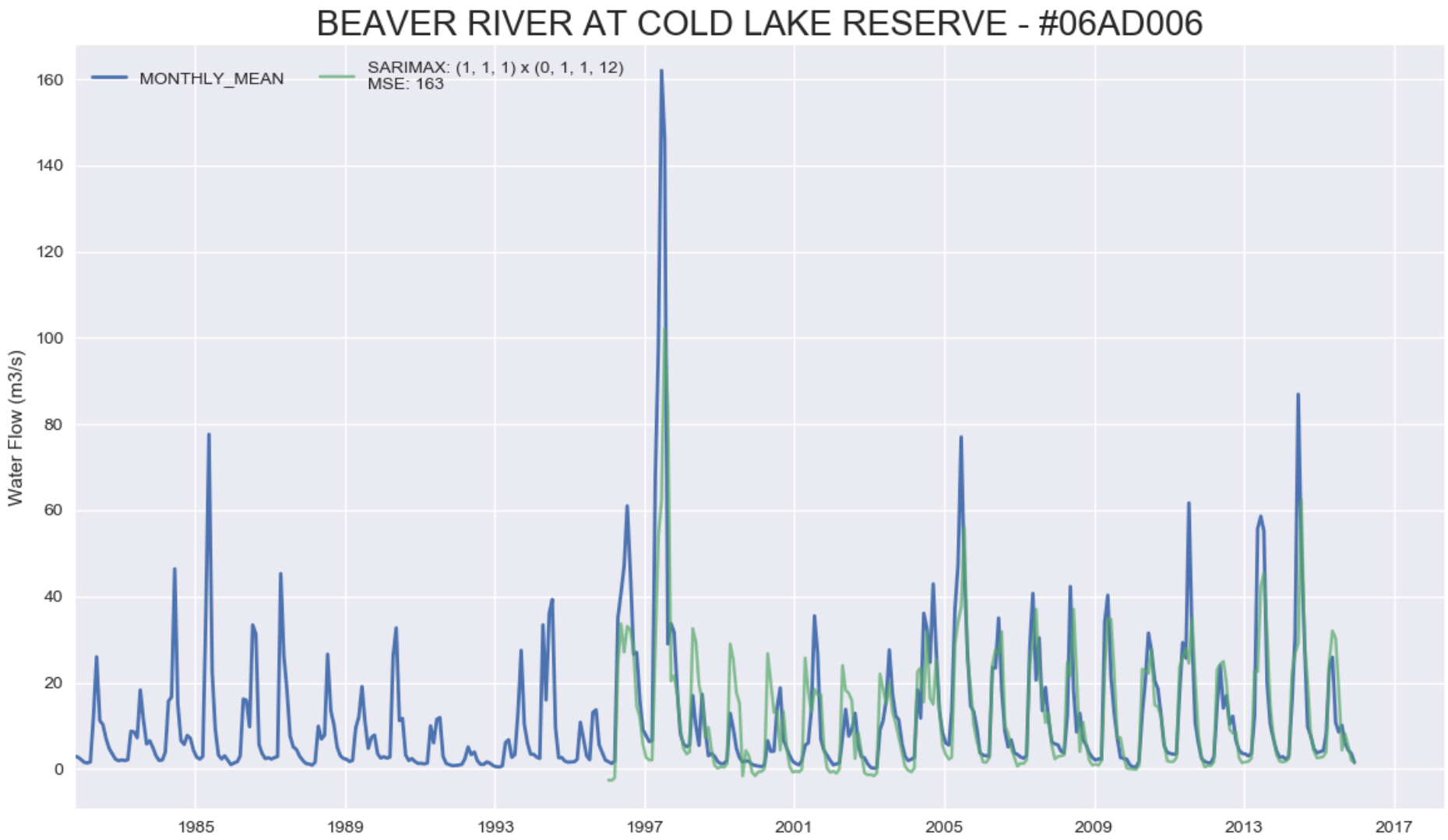


Figure 28.Beaver river at cold lake reserve

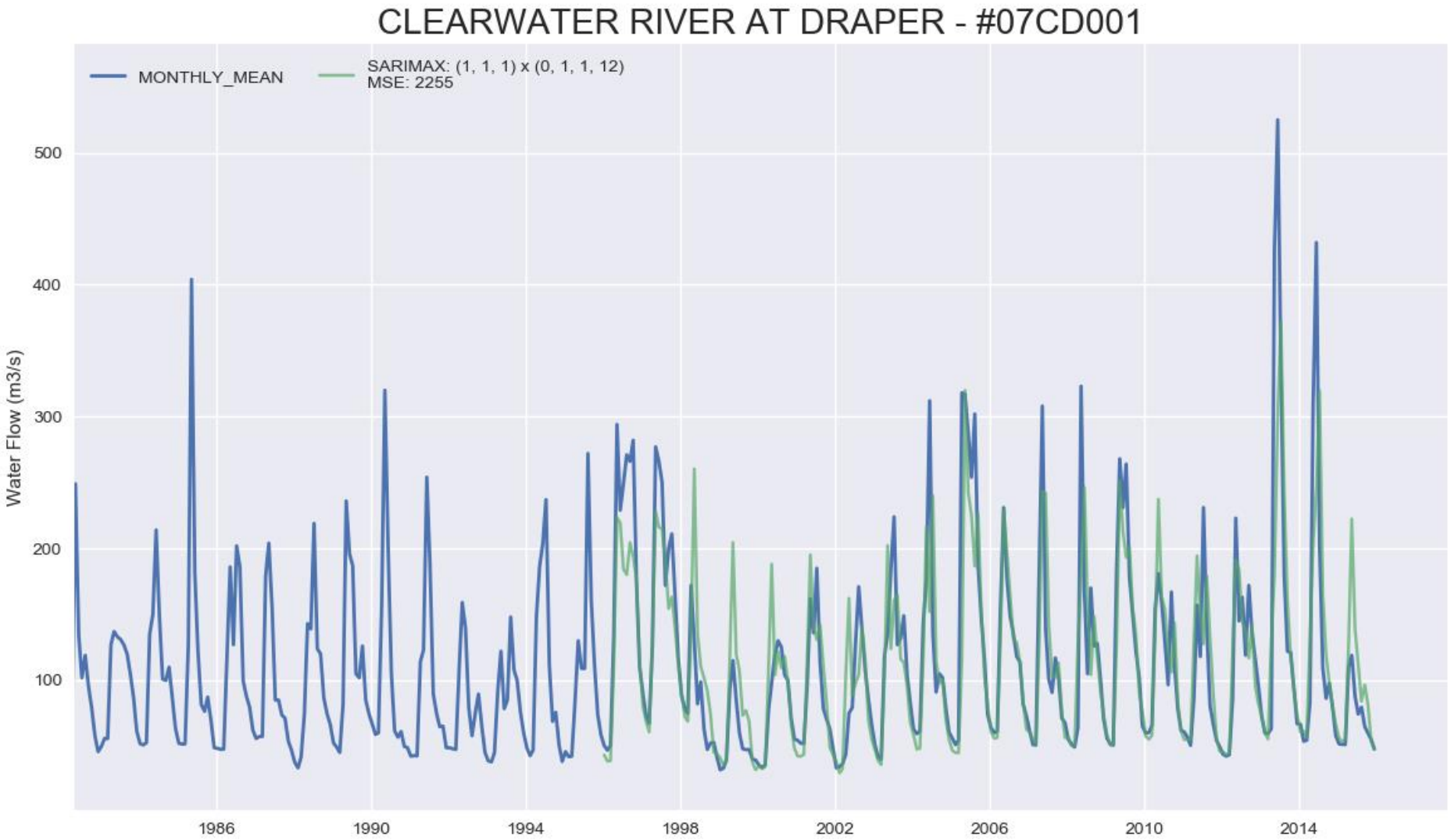


Figure 29.Clearwater river at draper

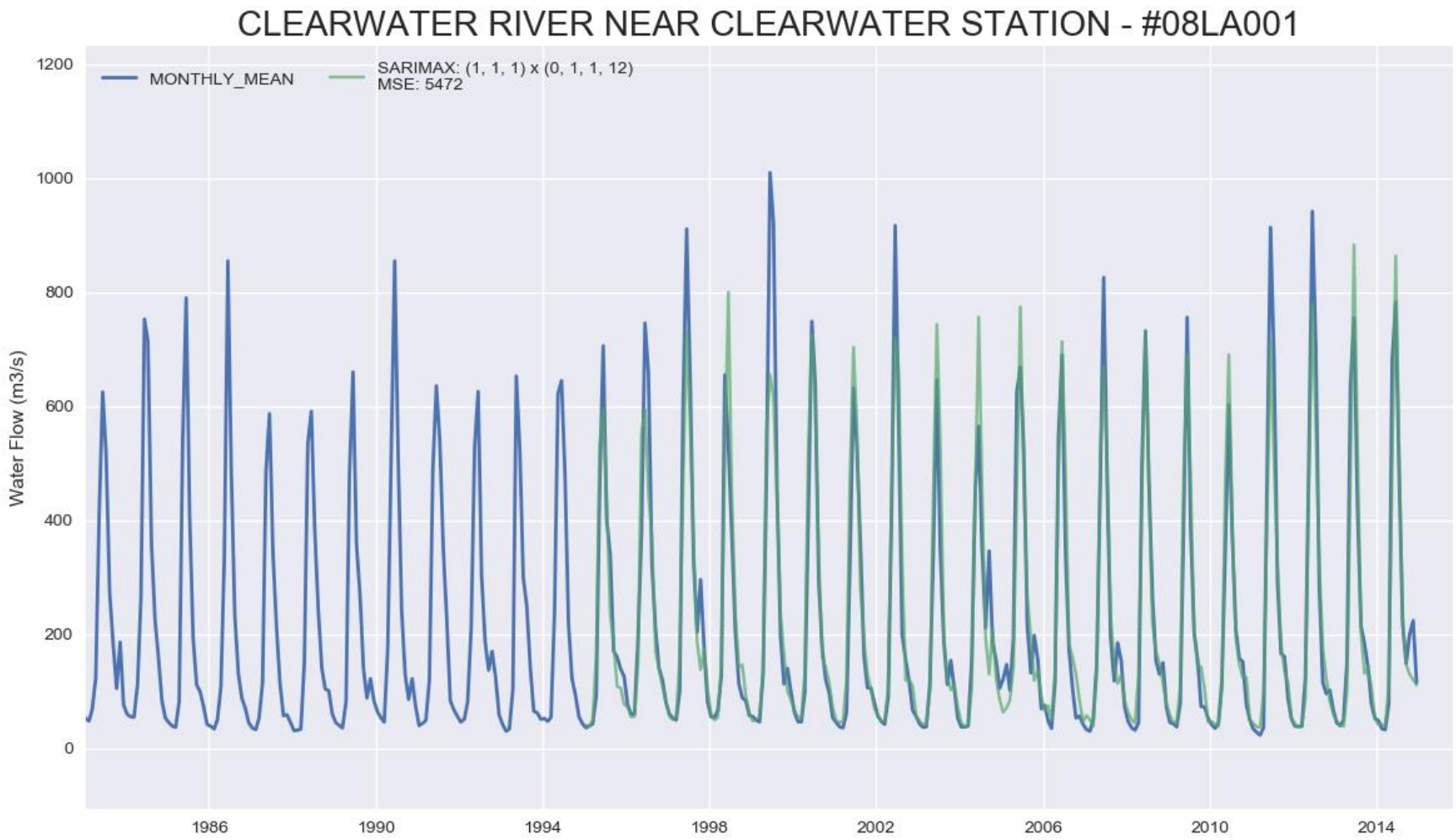


Figure 30.Clearwater river near clearwater station

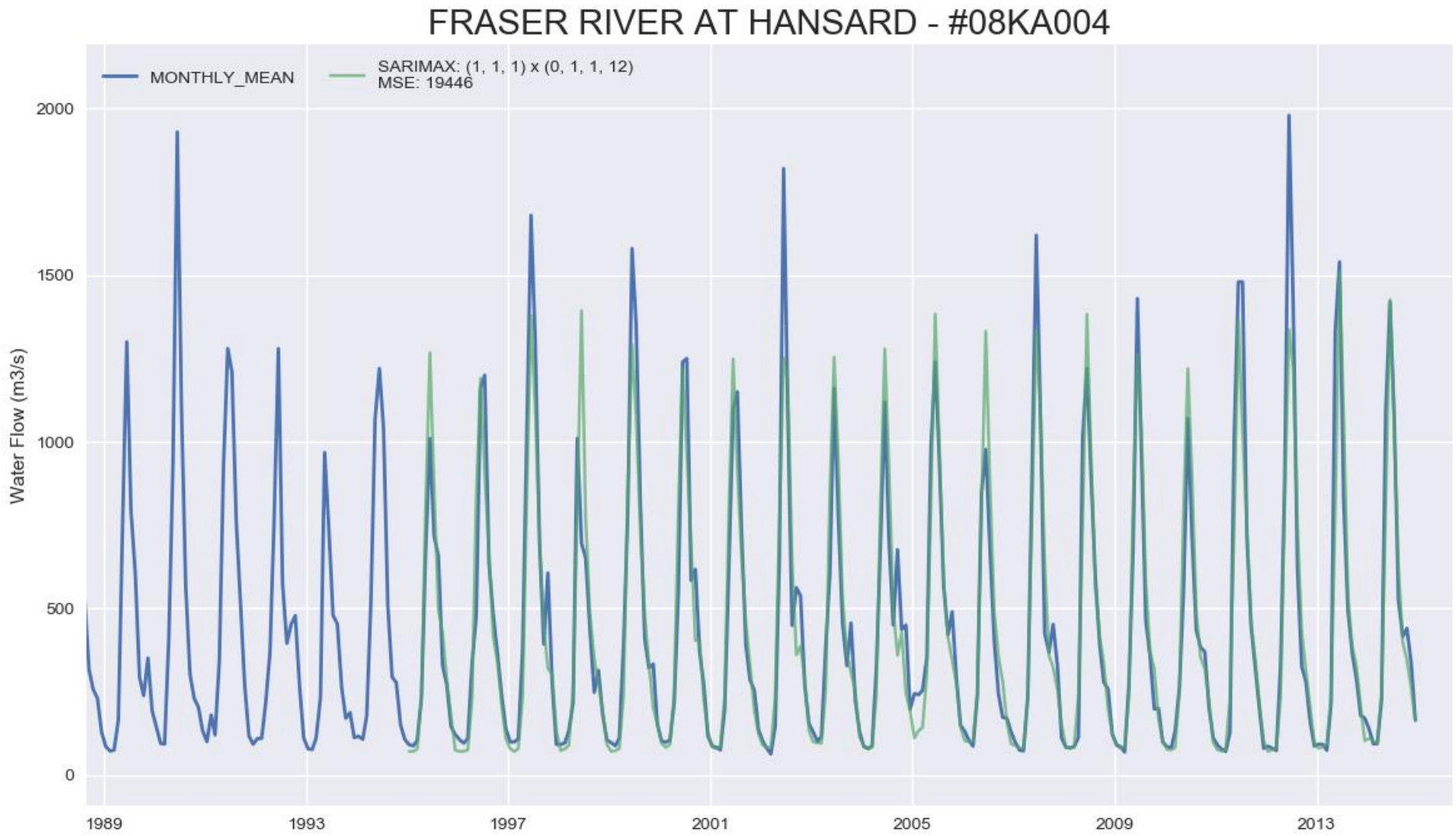


Figure 31.Fraser river at Hansard

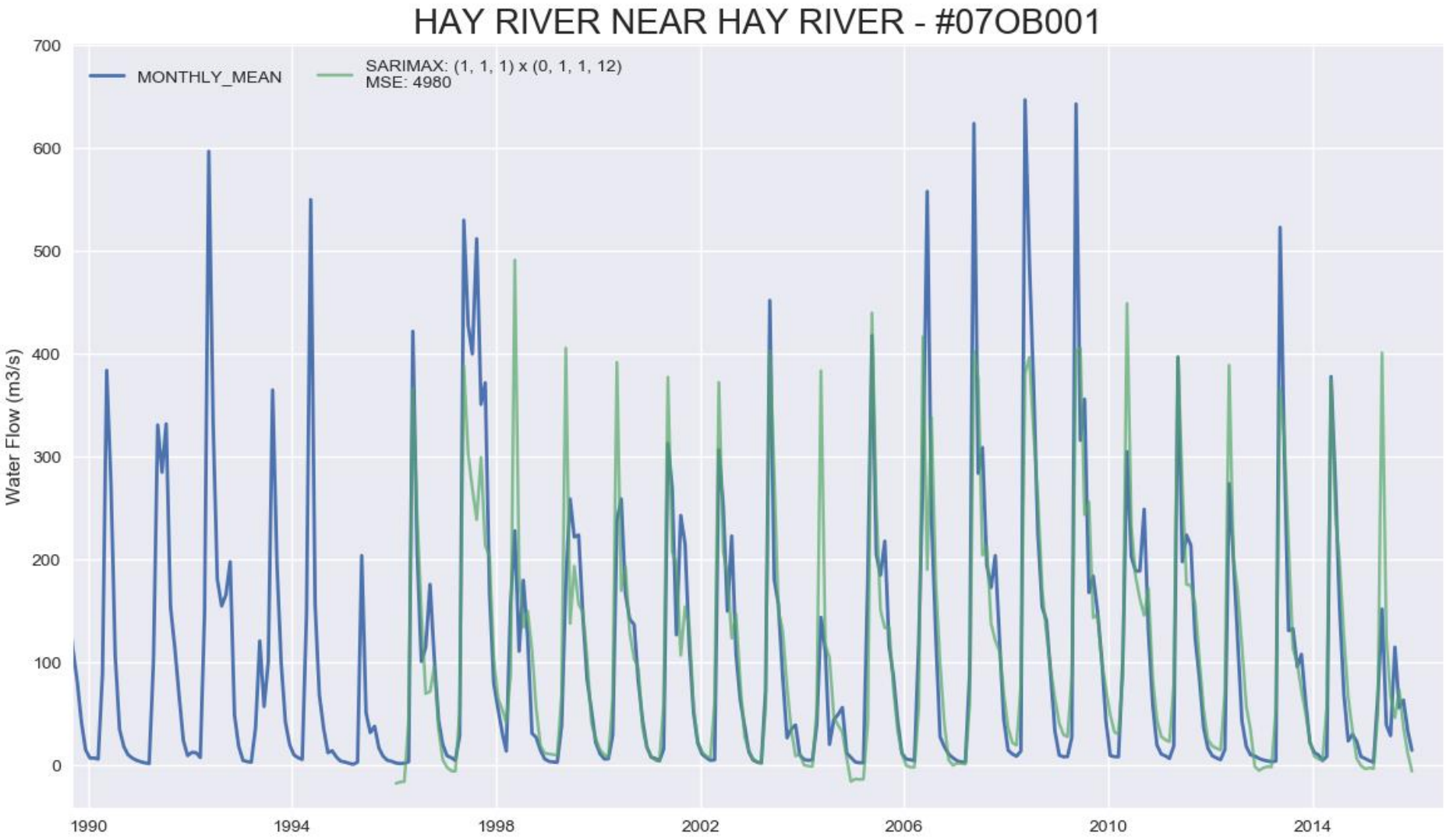


Figure 32.Hay river near hay river

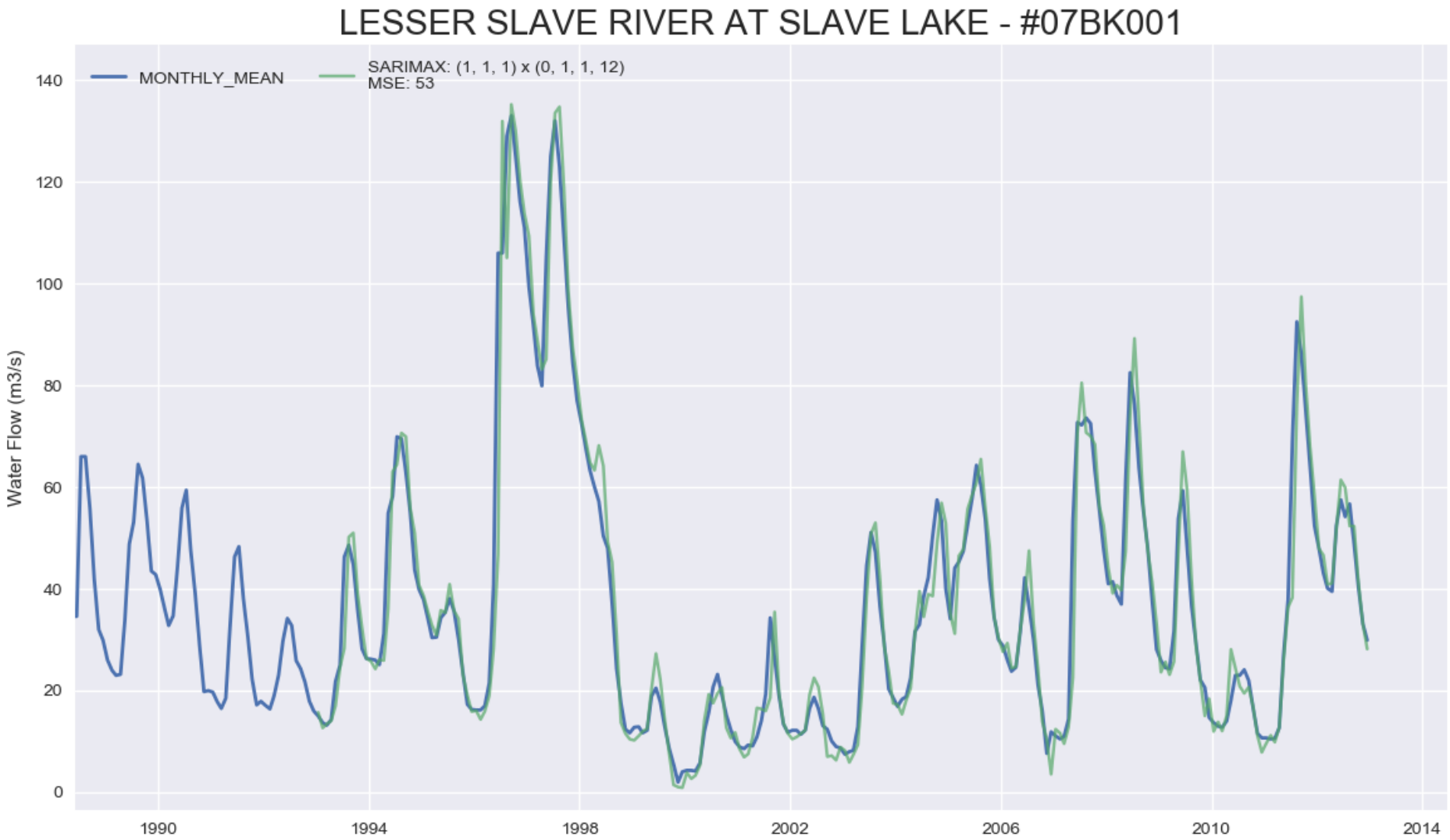


Figure 33.Lesser slave river at slave lake

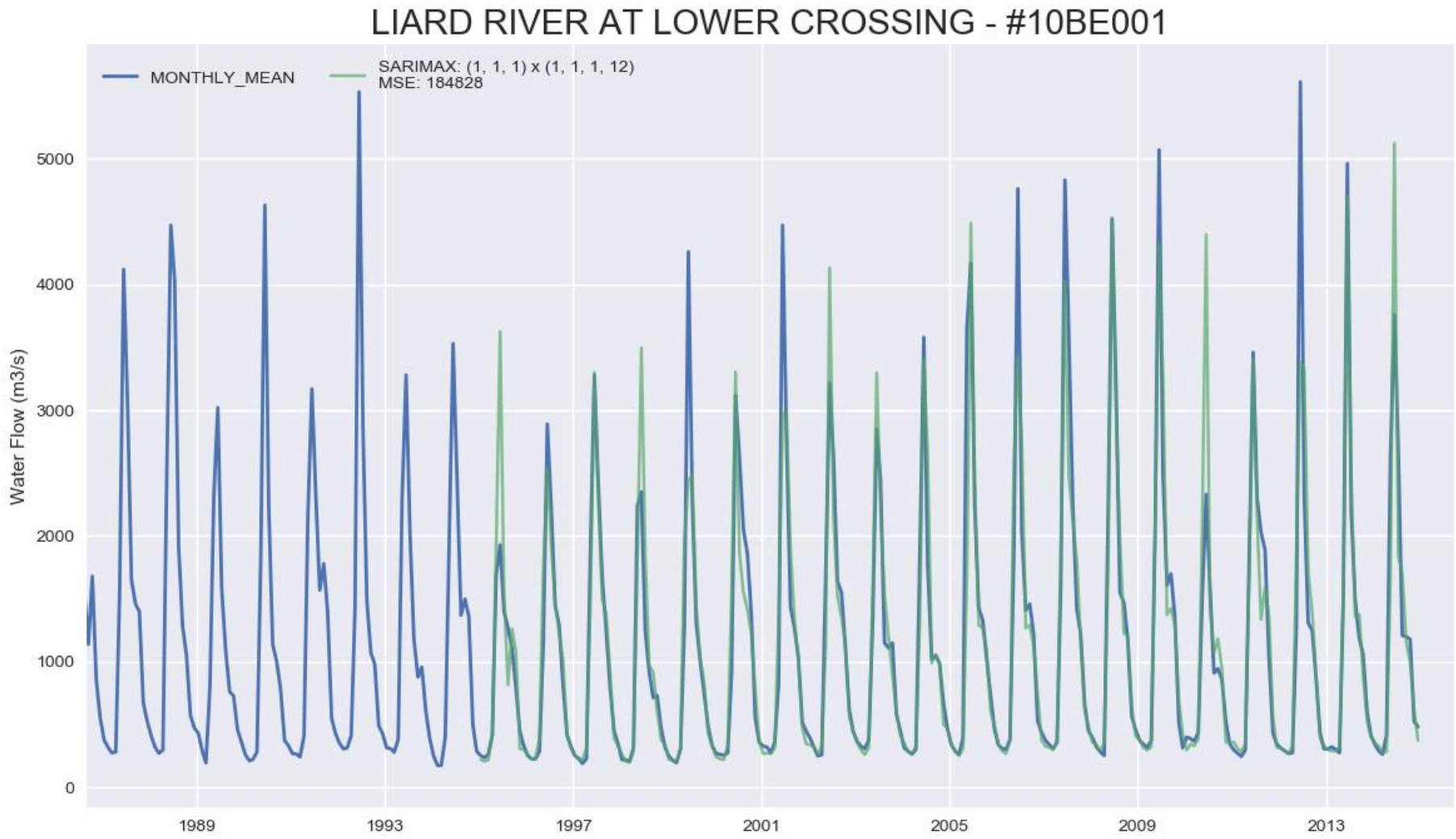


Figure 34.Liard river at the lower crossing

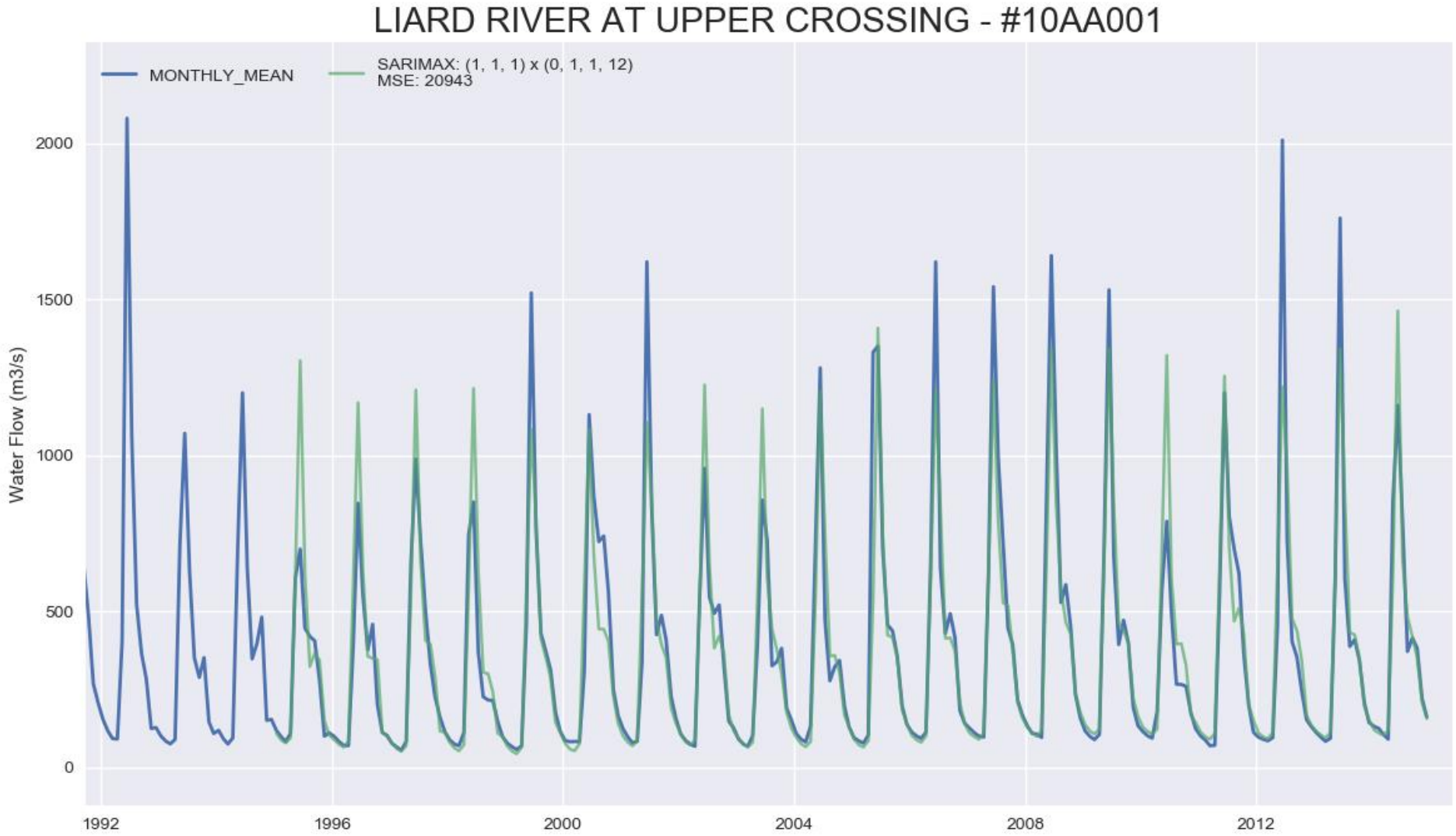


Figure 35.Liard River at the upper crossing

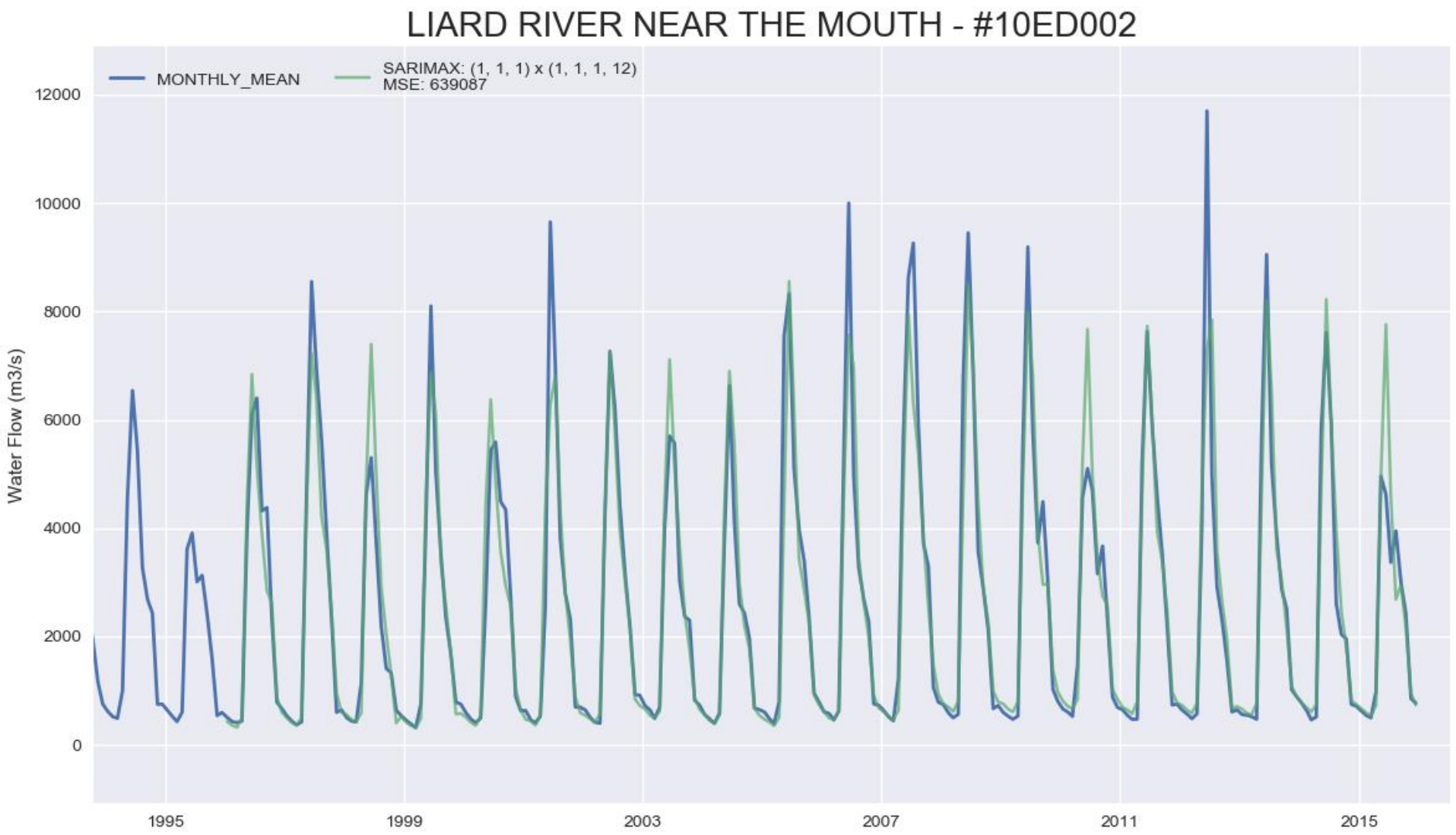


Figure 36. Liard river near the mouth

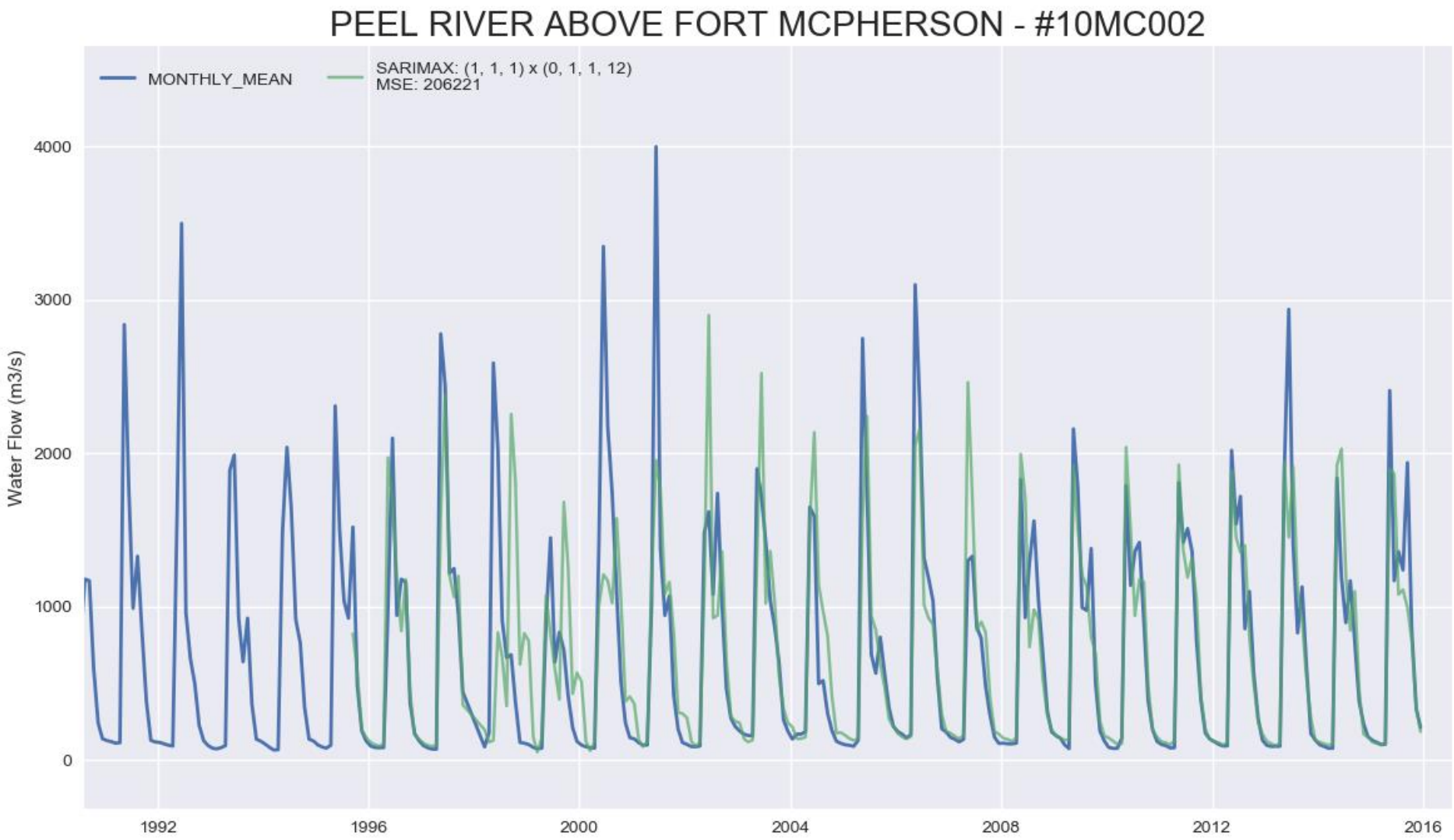


Figure 37.Peel river above fort Mcpherson

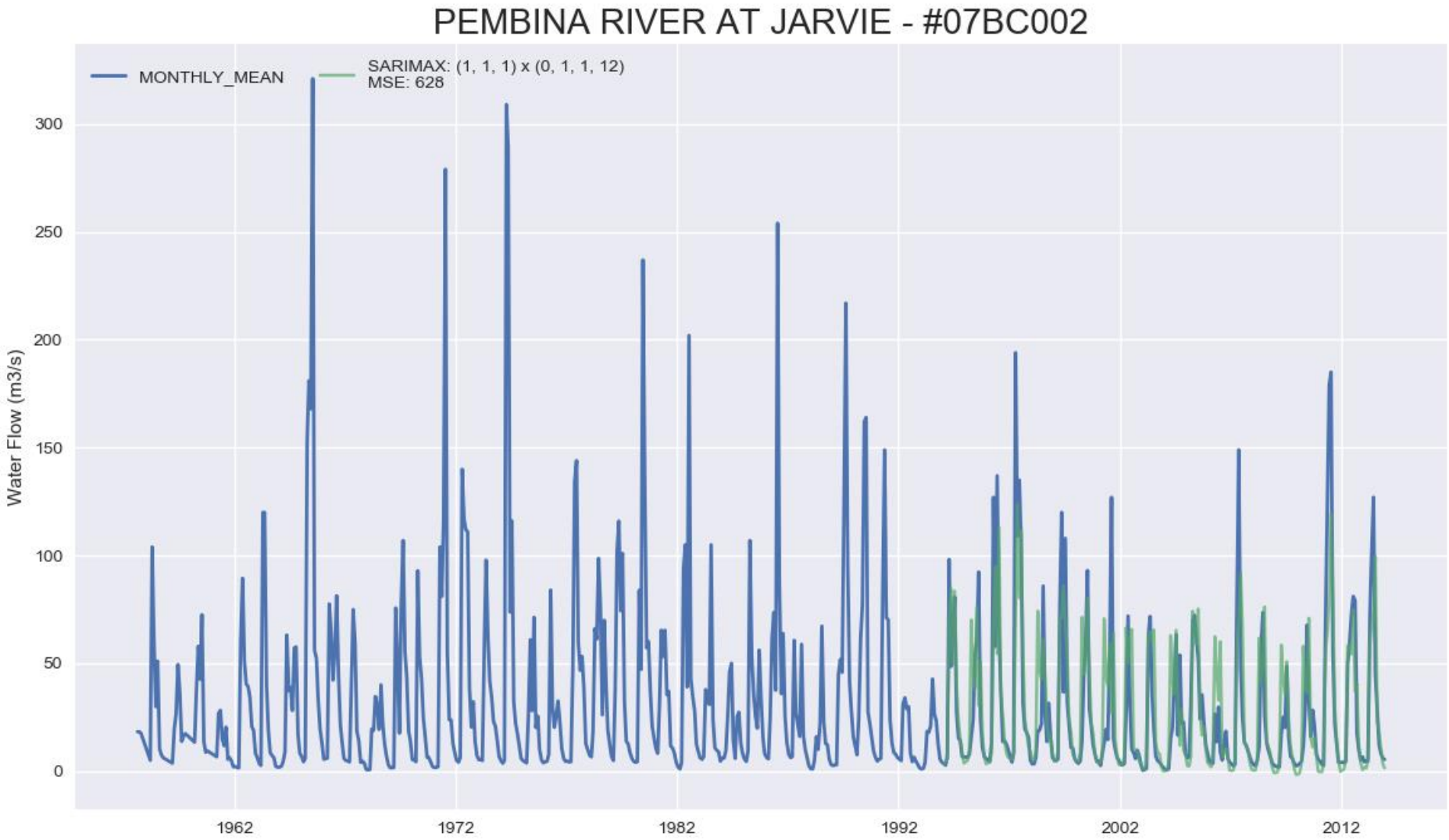


Figure 38.Pembina river at Jarvie

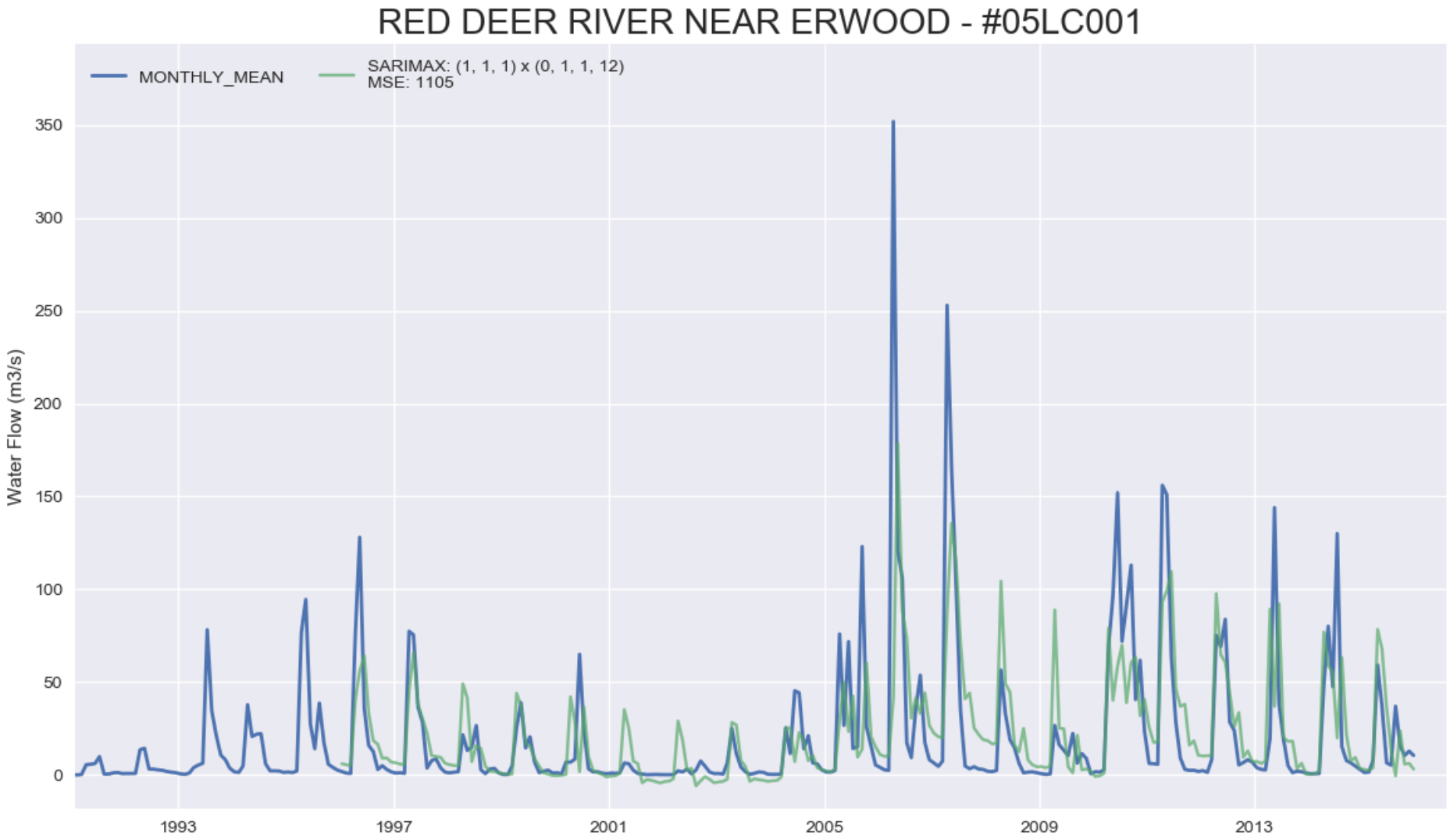


Figure 39.Red deer river near Sherwood

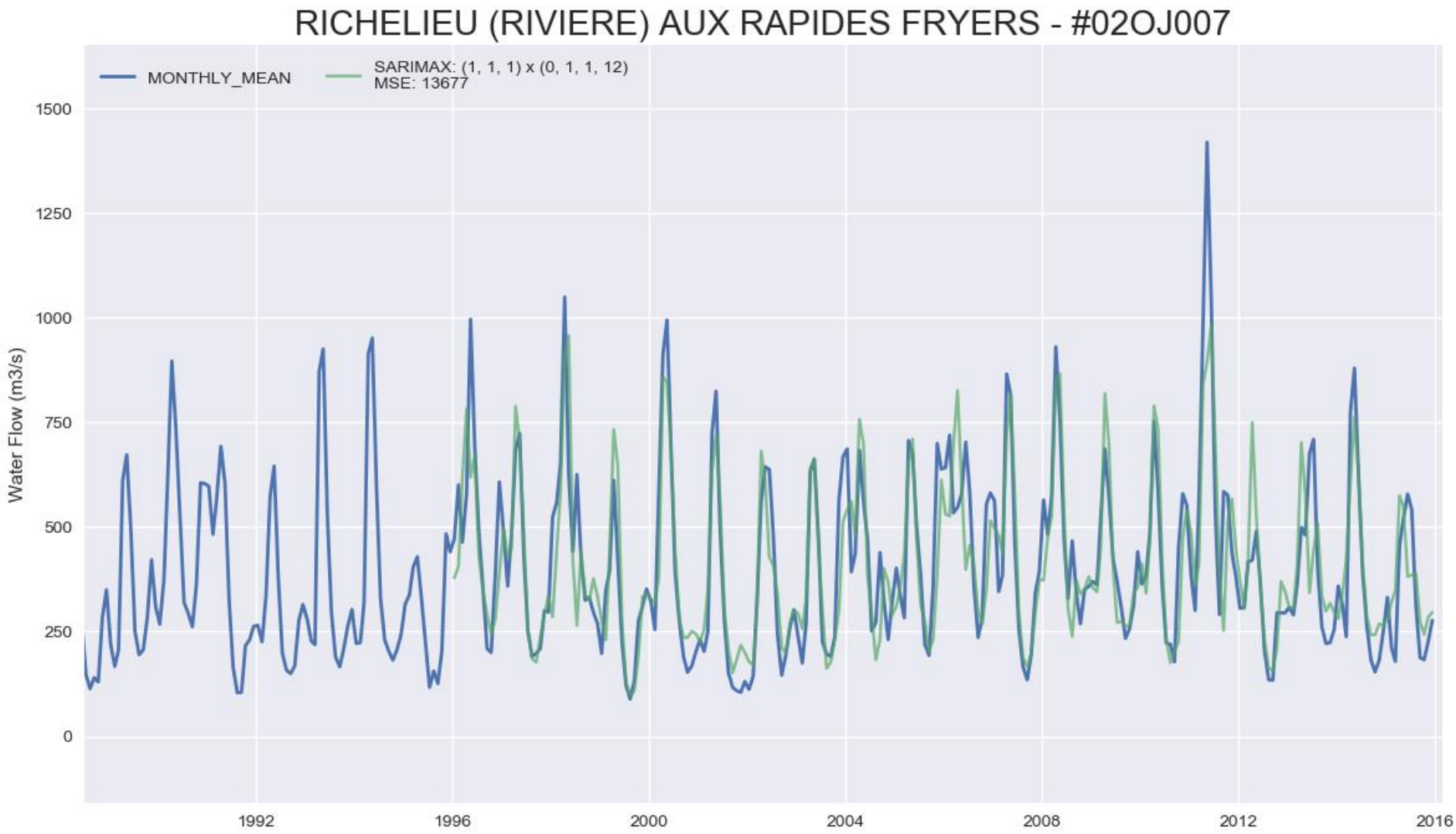


Figure 40. Richelieu (riviere) aux rapides fryers

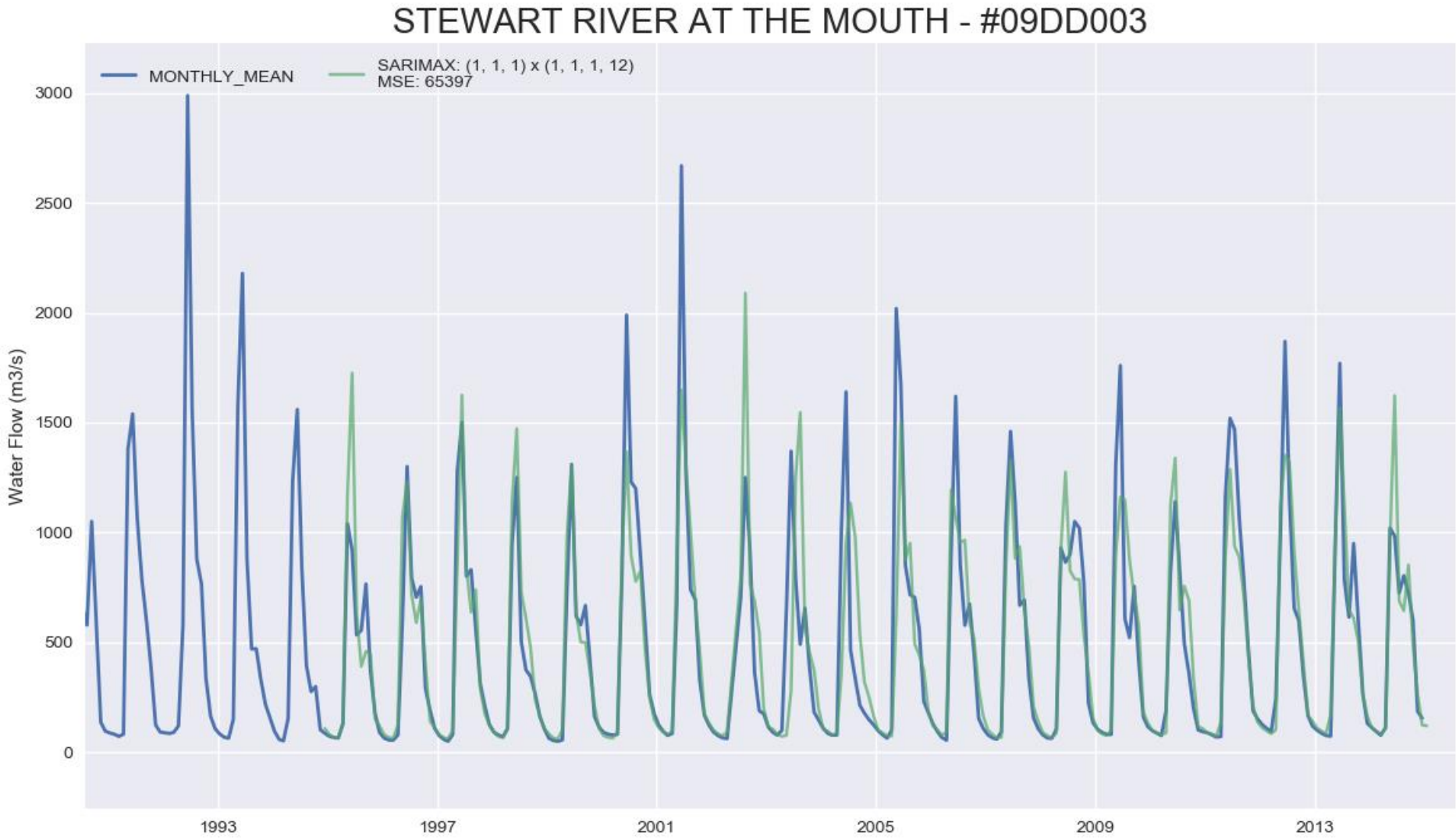


Figure 41.Stewart river at the mouth

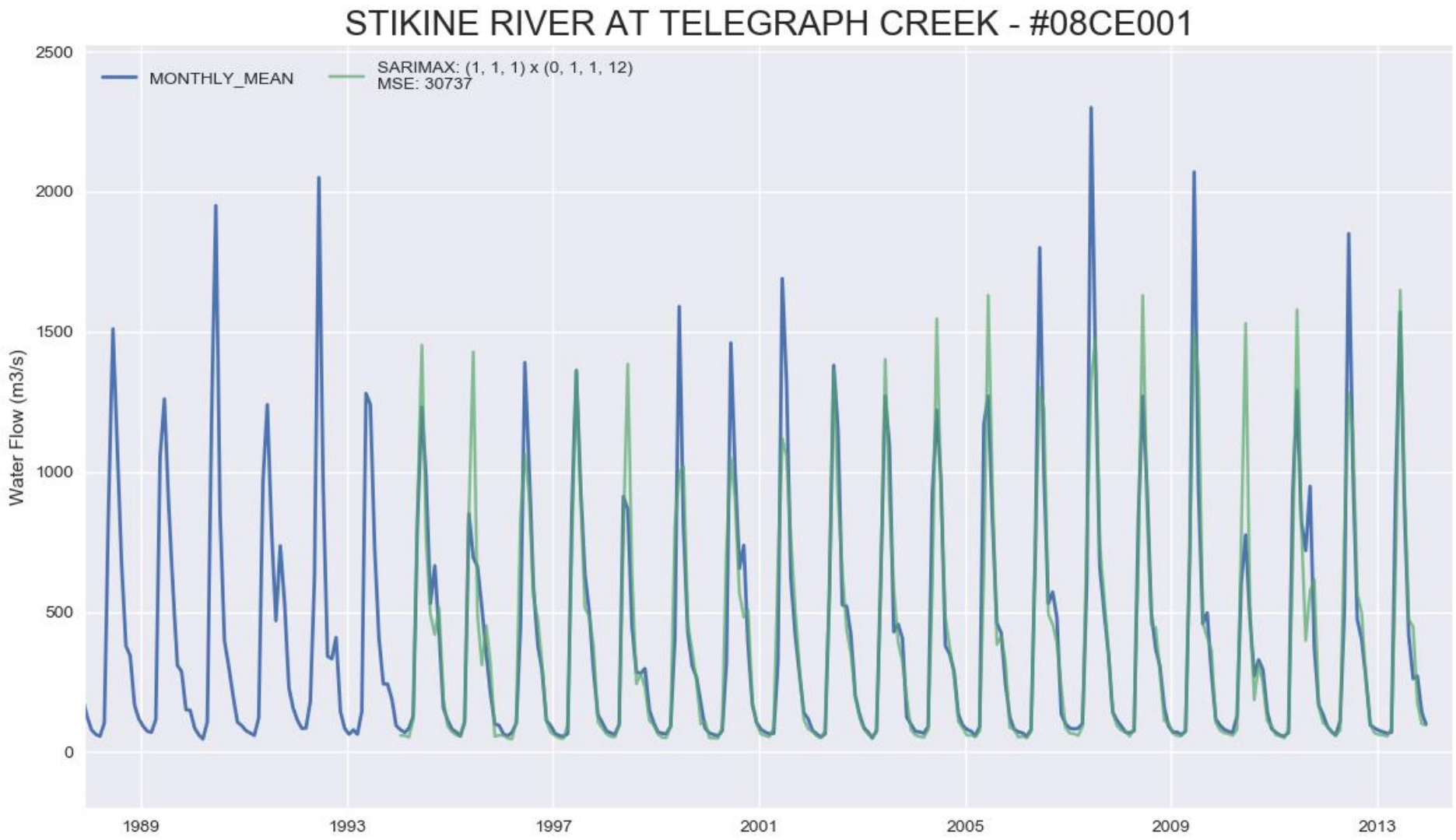


Figure 42.Stikine river at telegraph creek

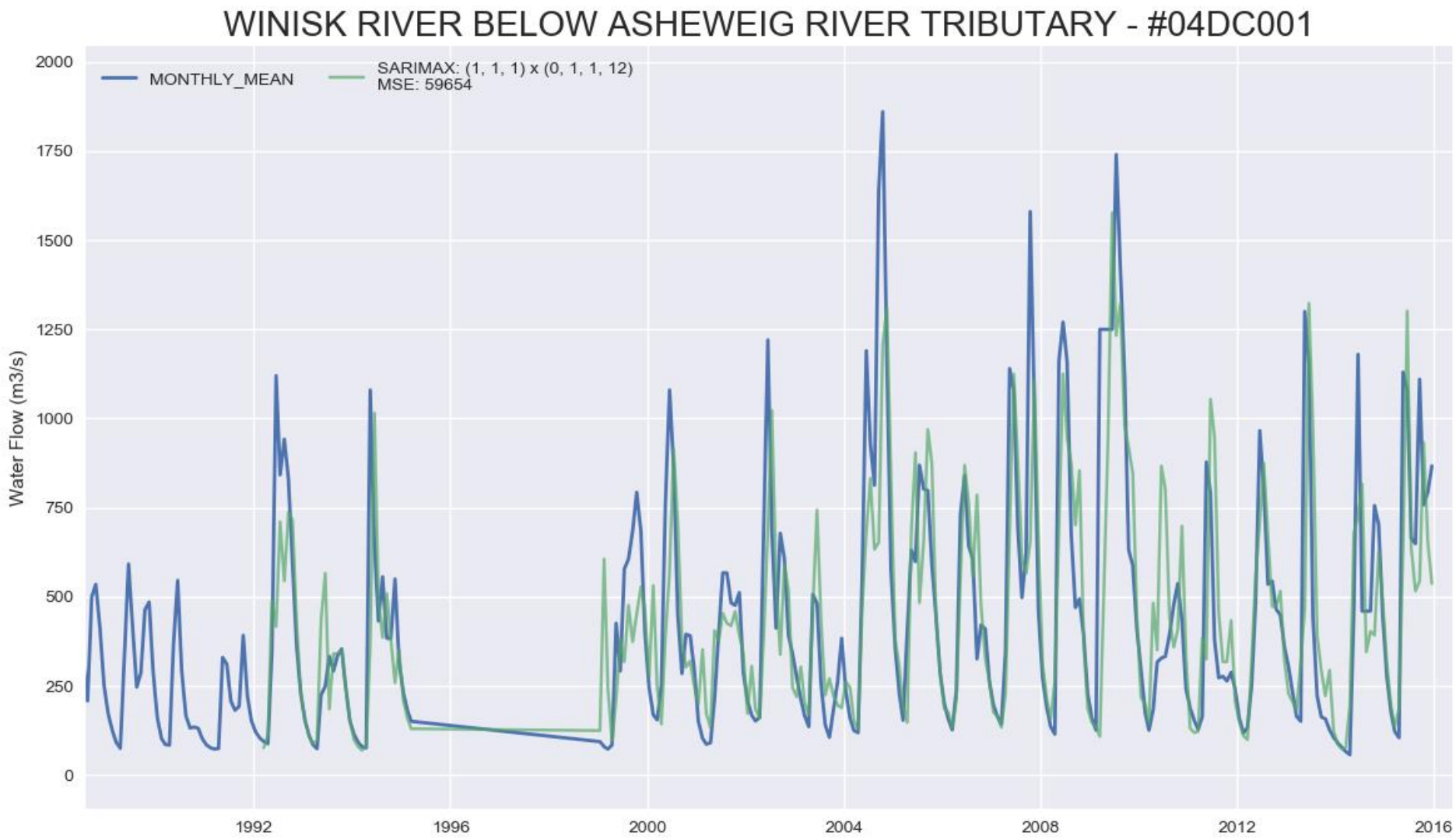


Figure 43.Winisk River below asheweig river tributary

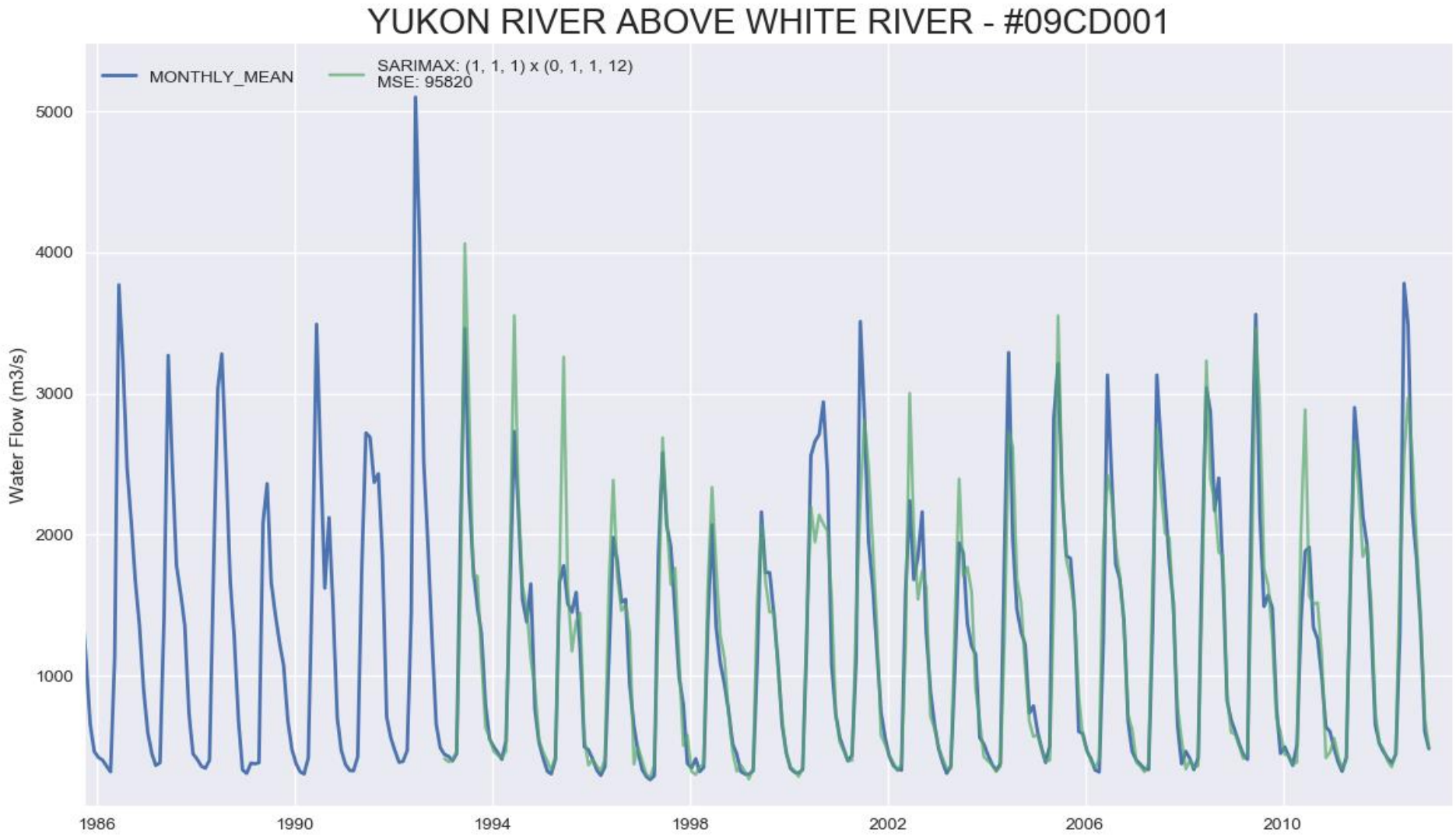


Figure 44.Yukon river above the white river

Table 6. Stations detailed information

STATION_NUMBER	STATION_NAME	DRAREA	LATITUDE	LONGITUDE	SARIMAX	MSE_SARIMX
08AB001	ALSEK RIVER ABOVE BATES RIVER	16200	60.118309	-137.977753	(1,1,1 x 0,1,1)12	3183
10LA002	ARCTIC RED RIVER NEAR THE MOUTH	18750	66.787437	-133.089645	(1,0,1 x 0,1,1)12	6724
07DA001	ATHABASCA RIVER BELOW FORT MCMURRAY	132588	56.78035	-111.402191	(1,1,1 x 1,1,1)12	40542
06AD006	BEAVER RIVER AT COLD LAKE RESERVE	14504	54.35516	-110.217278	(1,1,1 x 0,1,1)12	163
07CD001	CLEARWATER RIVER AT DRAPER	30800	56.68528	-111.255417	(1,1,1 x 0,1,1)12	2255
08LA001	CLEARWATER RIVER NEAR CLEARWATER STATION	10300	51.649471	-120.066566	(1,1,1 x 0,1,1)12	5472
08KA004	FRASER RIVER AT HANSARD	18000	54.078671	-121.850357	(1,1,1 x 0,1,1)12	19446
07OB001	HAY RIVER NEAR HAY RIVER	51700	60.743	-115.859642	(1,1,1 x 0,1,1)12	4980
07BK001	LESSER SLAVE RIVER AT SLAVE LAKE	13567	55.304871	-114.75621	(1,1,1 x 0,1,1)12	53
10BE001	LIARD RIVER AT LOWER CROSSING	104000	59.412498	-126.097221	(1,1,1 x 1,1,1)12	184828
10AA001	LIARD RIVER AT UPPER CROSSING	32600	60.050831	-128.906937	(1,1,1 x 0,1,1)12	20943
10ED002	LIARD RIVER NEAR THE MOUTH	275000	61.742722	-121.227966	(1,1,1 x 1,1,1)12	639087
10MC002	PEEL RIVER ABOVE FORT MCPHERSON	70600	67.258888	-134.888809	(1,1,1 x 0,1,1)12	206221
07BC002	PEMBINA RIVER AT JARVIE	13103	54.450291	-113.993317	(1,1,1 x 0,1,1)12	628
05LC001	RED DEER RIVER NEAR ERWOOD	11000	52.85947	-102.195061	(1,1,1 x 0,1,1)12	1105
02OJ007	RICHELIEU (RIVIERE) AUX RAPIDES FRYERS	22000	45.398472	-73.258438	(1,1,1 x 0,1,1)12	13677
09DD003	STEWART RIVER AT THE MOUTH	51000	63.282219	-139.25444	(1,1,1 x 1,1,1)12	65397
08CE001	STIKINE RIVER AT TELEGRAPH CREEK	29000	57.900269	-131.159714	(1,1,1 x 0,1,1)12	30737
04DC001	WINISK RIVER BELOW ASHEWEIG RIVER TRIBUTARY	50000	54.499611	-87.227692	(1,1,1 x 0,1,1)12	59654
09CD001	YUKON RIVER ABOVE WHITE RIVER	149000	63.0825	-139.496933	(1,1,1 x 0,1,1)12	95820

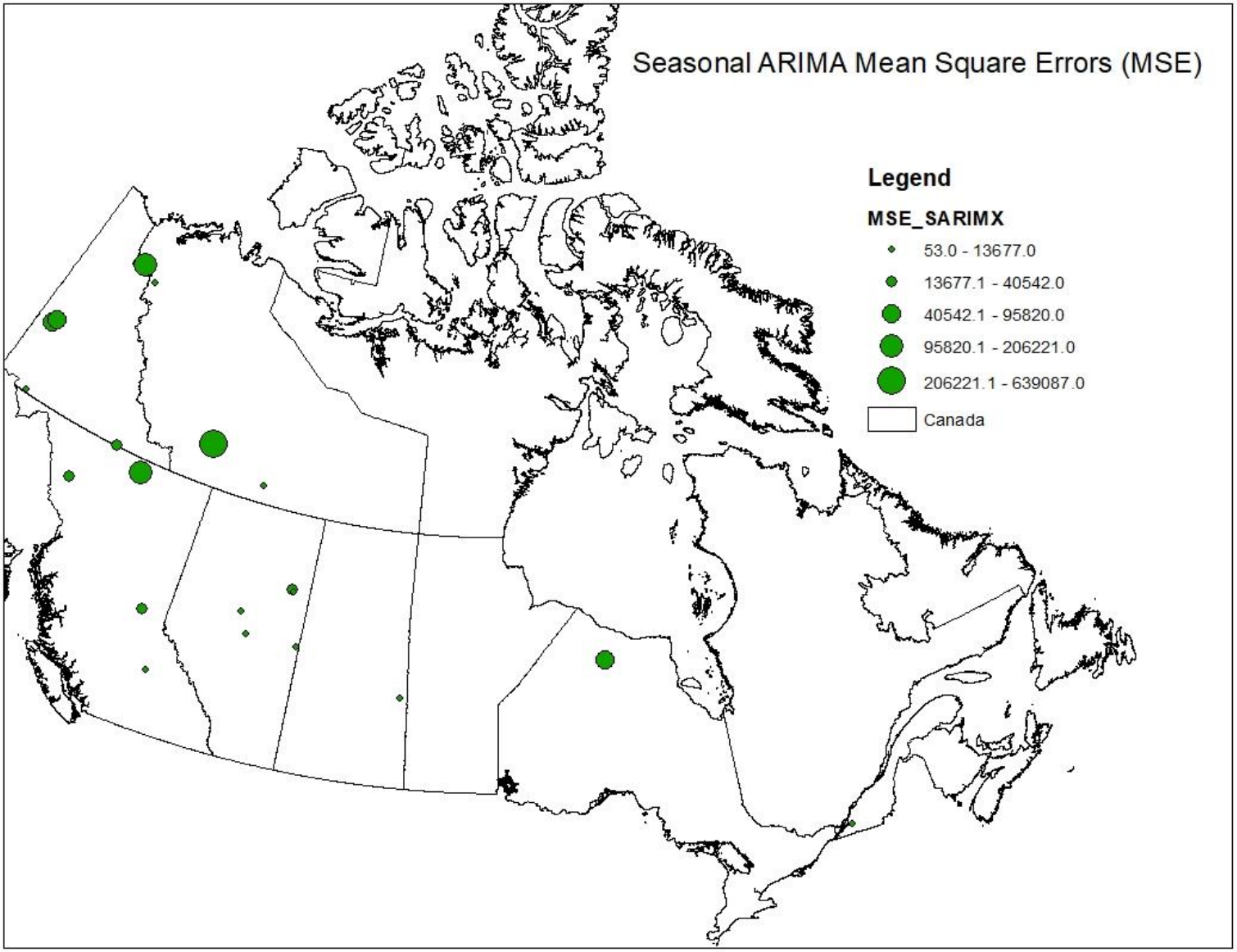


Figure 45. SARIMA mean square errors (MSE)

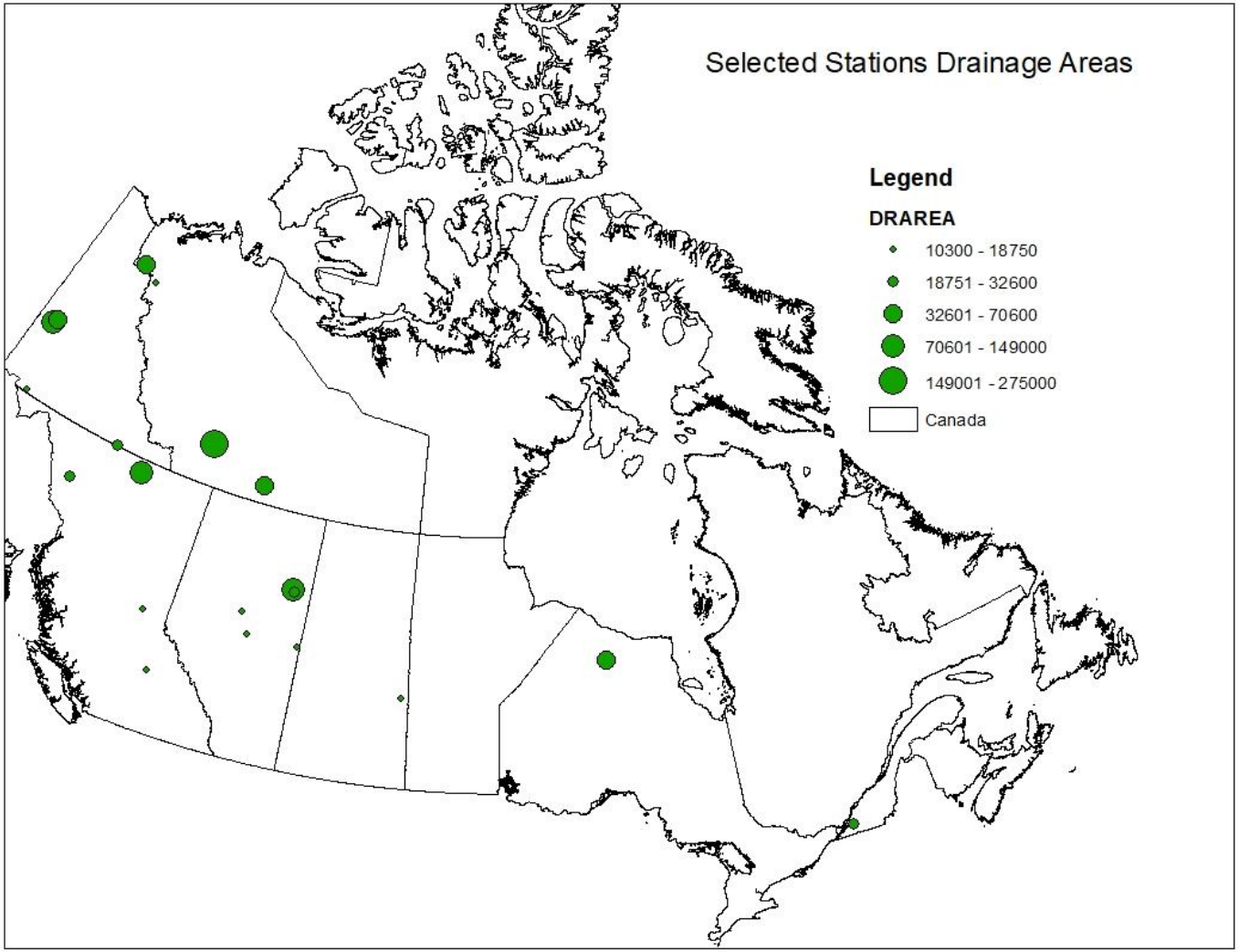


Figure 46. Selected Stations drainage areas

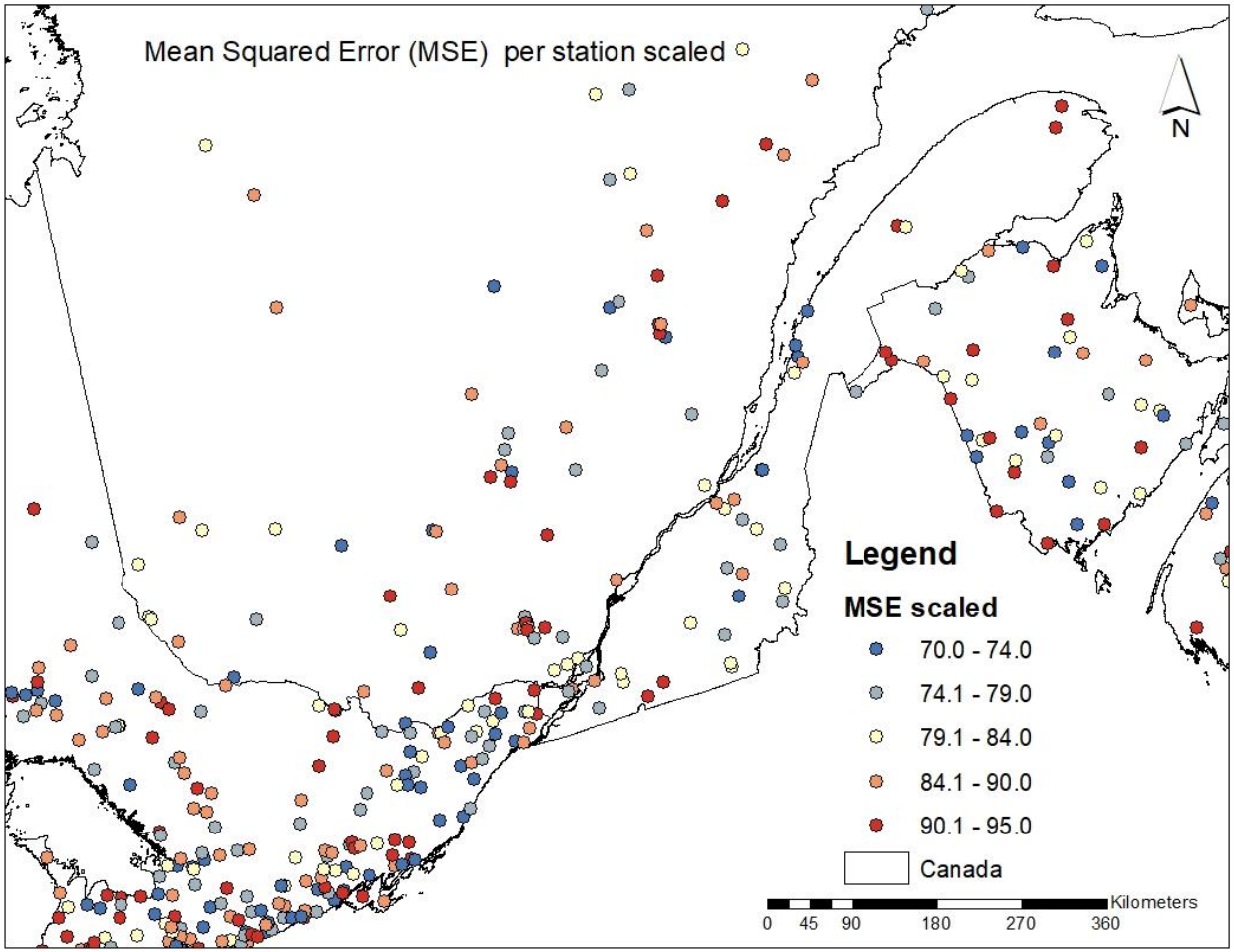


Figure 47.Canad

4.3 Case study of stations with Upward and downward trends

4.6.1 Study Case #2 – Time series with downward Trend

An interesting station to study is #06AC007, that of Redberry Lake near Krydor. This station is located in Alberta and is well known for its shallow and fairly clear water (Mitchell and Prepas, 1990). The data recorded displays a clear downward trend.

This station displays a clear trend and will be used to test the accuracy of our methodology. The Redberry Lake station display a trend across a span time from 1981 to 2013 and is a typical case study for our model. We have generated the Man-Kendall trend (figure 48). The water level of this station follows a negative slope of – 0.0116. Using our python application, we have run all possible SARIMA combinations and generated Akaike factors for all SARIMA factors combinations. It was determined that a SARIMA model of (1,1,1) x (1,0,1,12) would be the most representative model of the trend and the least likely to overfit the data. To analyze the efficiency of the model, all factors have been generated, and the residual's distribution has been studied (Figure zz).

The residuals are overall equally distributed in the positive and negative sides of the first graphic, and they seem to be normally distributed (top right graphic). The QQ-plot for the residuals shows that most values are closely following the normal distribution along the red line with only a few discrepancies at the extremes of the line. The correlogram also supports the idea of our model not being an overfit of the data. Indeed, most points do not display a correlation with the current points.

These facts are good indicators and allow us to plot a short to medium-term forecast of six months for Redberry Lake. The generated results can be used assess the efficiency of the model (Table xxx). This station has a mean square error of 0.019, a RMSE of 0.14, and a model efficiency e to be 0.92, which is above 0.9 and is indicator of a reliable SARIMA model according to Nash ad

Sutcliffe (1970). Following this test case, we can confirm that our methodology generated a reliable forecast stations showing extreme downward trends. Now, we must verify the efficiency of our methodology for stations showing extreme upward trends.

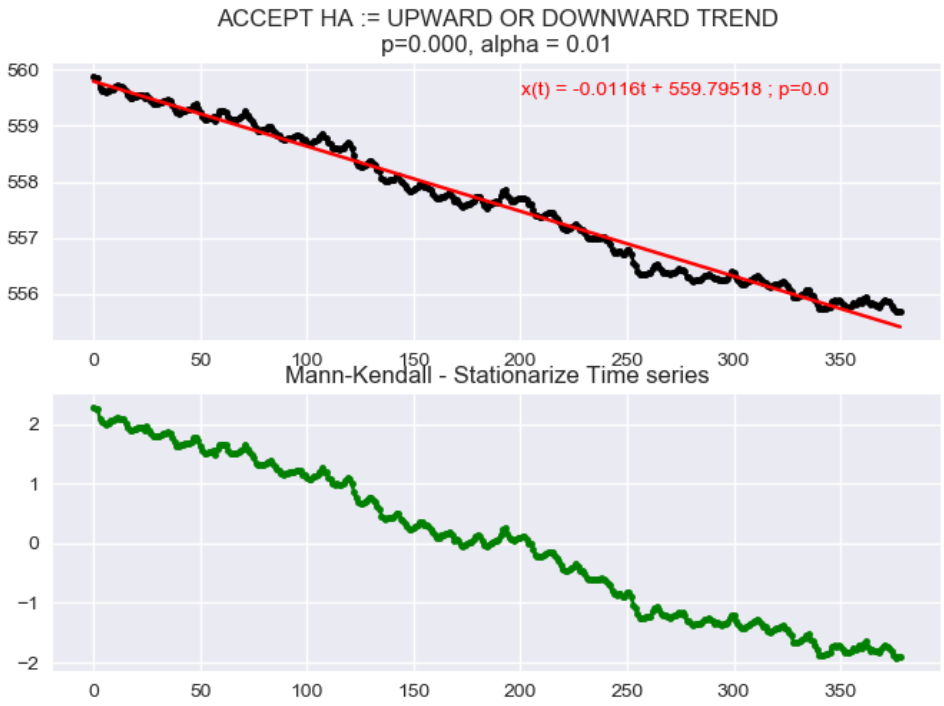


Figure 48.Mann-Kendall

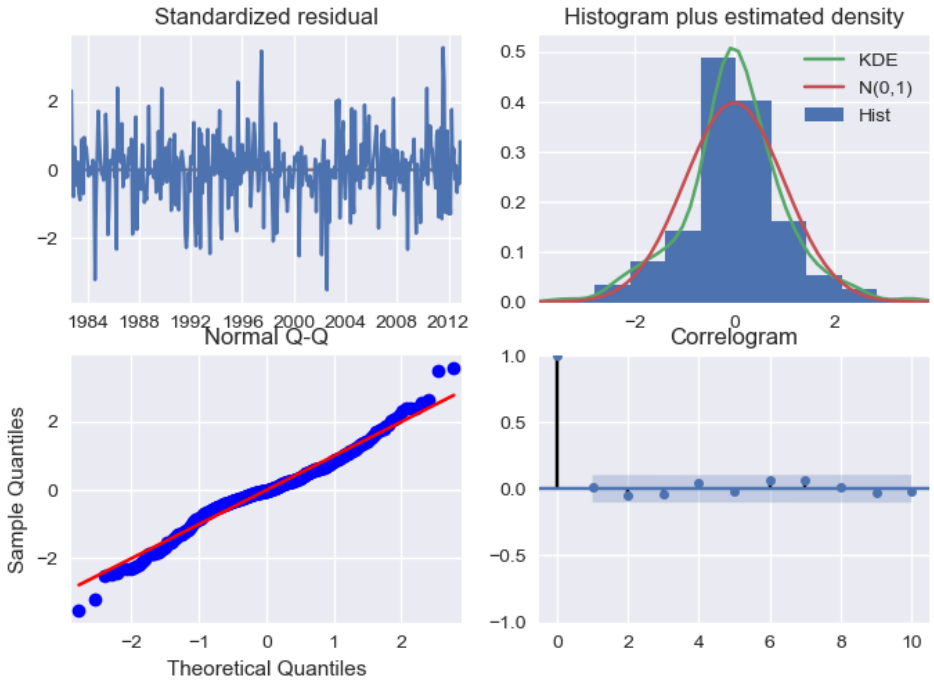


Figure 49.QQ-Plot, Histogram, correlogram, and standardized residuals

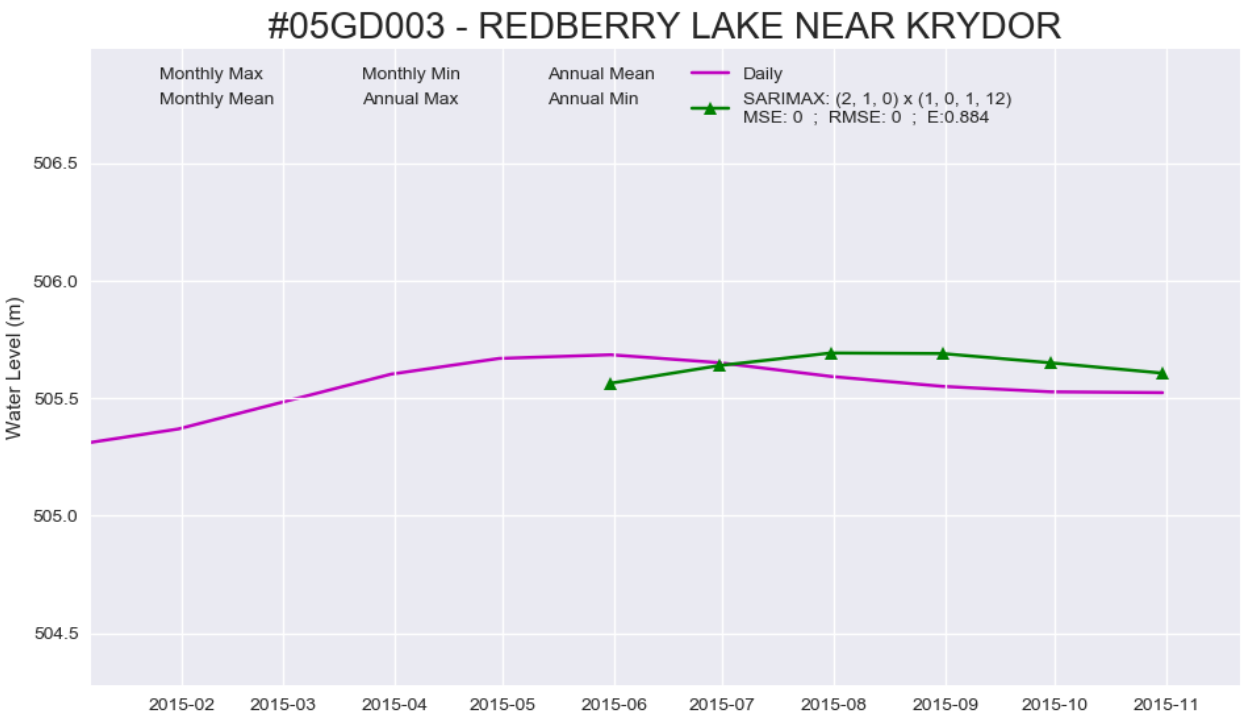


Figure 50.Redberry lake Kryder

4.6.2 Study Case #3 – Time series with upward Trend

An interesting station to study is #05KA010, that of Waldsea Lake near Humboldt. This station is located in Saskatchewan and is well known for its partially saline characteristics with important concentrations of Chlorine, Sodium, Magnesium, Sulfate, and Chlorine. The data recorded displays a clear upward trend.

To test the efficiency of our methodology, we must choose critical stations that display a clear trend. We will analyze a positively trended station and negatively trended station. The Waldsea Lake near Humboldt station display an upward trend across a span since 1976. This makes it a typical study case to test our model. We have generated the Man-Kendall trend (figure 51). We can see that the water level for the Waldsea River follows a positive slop of 0.00647. Using our python application, we have run all possible SARIMA combinations. We have generated Akaike factors for all SARIMA Factors combinations. It was determined that a SARIMA model of (1,1,1)

x (1,0,1,12) would be the most representative model of the trend and the least likely to overfit the data. To check the efficiency of the model, we have generated the factors, and we have analyzed the distribution of the residuals (figure 52). As we can see, the residuals are generally equally distributed in the positive and negative sides of the first graphic, and they seem to be normally distributed (top right graphic). The qq-plot for the residuals shows that most values are closely following the normal distribution along the red line with only a few discrepancies at the extremes of the line. The correlogram also supports the idea of our model not being an overfit of the data. Indeed, we observe that most points do not display a correlation with the current points.

Knowing these facts, we feel confident to plot a short to medium-term forecast of six months for the Waldsea. We have calculated a mean square error of 0.006, a RMSE error of 0.08, and we measured the model efficiency E to be 0.913, which is above 0.9 and is indicator of a reliable SARIMA model according to Nash ad Sutcliffe (1970). Following this test case, we can confirm that our methodology generated a reliable forecast stations showing extreme upward trends.

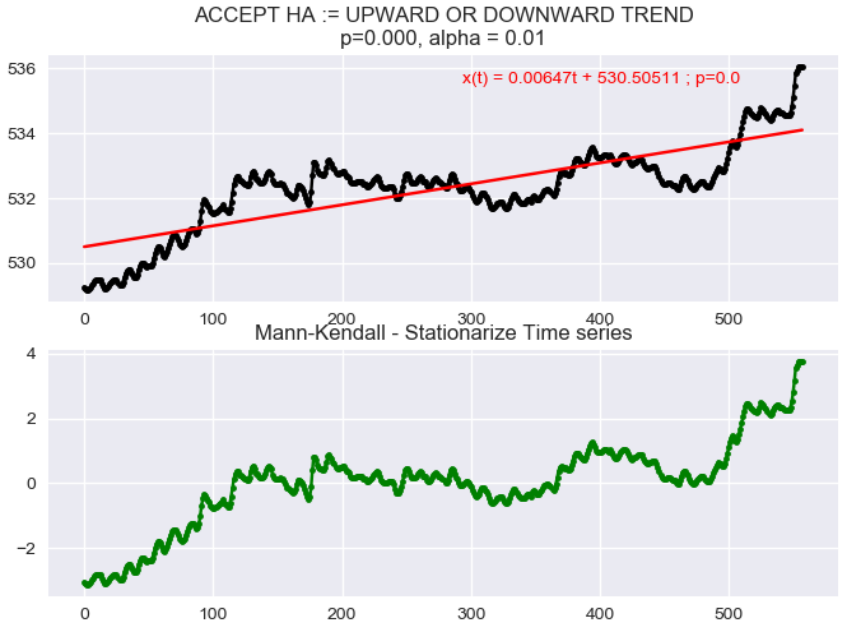


Figure 51.Mann-Kendall

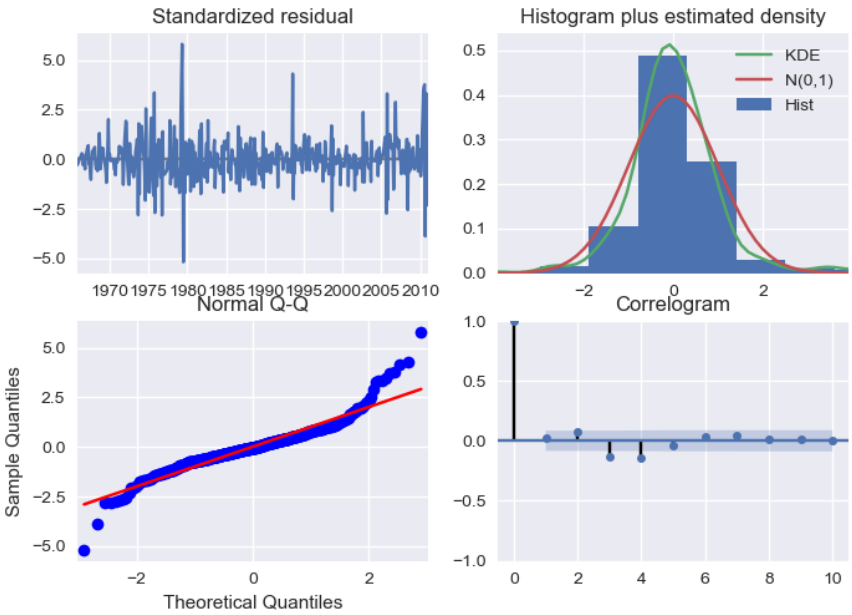


Figure 52.QQ-Plot, Histogram, correlogram, and standardized residuals

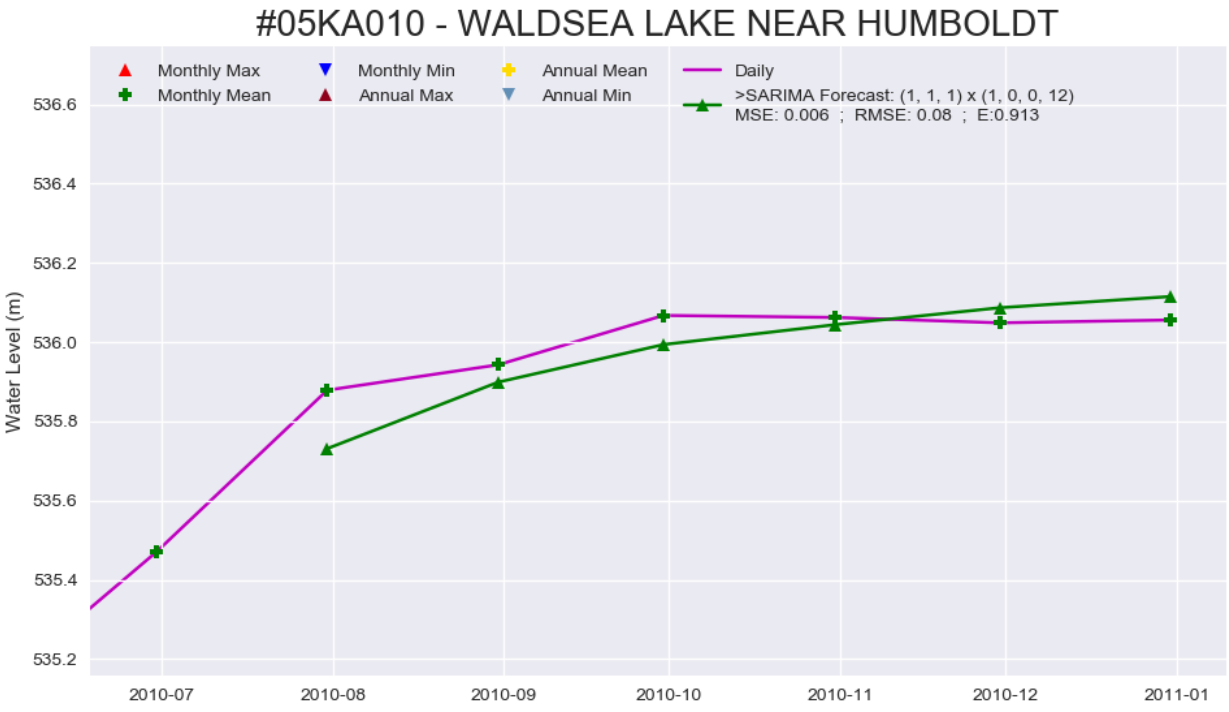


Figure 53.Waldsea Lake near Humboldt Forecast (6 months)

Chapter 5 - Conclusion

In this thesis, we have been able to highlight the needs for forecasting water flow, water level and sediment concentrations useful for several entities including hydroelectric industries, municipalities (water consumptions regulation), meteorological previsions, construction companies (building near water bodies), civil engineers (e.g. building bridges) and biologists (e.g. water conditions affecting aquatic life). We have then developed an efficient methodology to analyze hydrological time series data in which we identified stationarity of the series using Man-Kendall. The Akaike information criteria was used as it is a solid basis for generating efficient SARIMA models. We developed means by which to validate our models by analyzing the residuals. Then, we generated a SARIMA model for downward trend of (1,1,1)x(1,0,1,12) with a slope of x(t)=-0.0116t+559.795, and another for an upward trend equal to (1,1,1)x(1,0,1,12) with a slope of x(t)= 0.00647t+ 530.505. Their respective efficiency factors are 0.88 and 0.913, which is an indication that those are acceptable models according to Nash ad Sutcliffe (1970). Using the HYDAT database, we have been able to identify stations displaying a stationary, upward and downward trend to validate our model.

Finally, we have automated the process of analysis, model-building, and forecast (streamflow, water levels, and sediments) by building a python-based application easily extendable and user-friendly. Therefore, automating the SARIMA process overall Canadians station for the HYDAT database will prove to be a very useful tool for decision-makers and many other entities in the field of hydrological study.

5.1. Main contributions

The main contributions to the literature as well as a summary of findings, the strengths, limitations, and the areas for future research are presented below.

5.1.1 Main findings and contributions

The main finding the SARIMA method makes a solid forecast for short- and medium-term periods but may be less efficient for stations with unique or extreme conditions that can trigger sudden peaks or drops. The use of exogenous factors might be useful in improving the model in such cases.

This research also contributes to the area of water forecasting through the development of a python-based application, the *Universal Canadian forecast Application* (Stitou, Adnane, 2018©). First, it documents all HYDAT stations across Canada in terms of water flow, water level, and sediment concentrations.

5.1.2 Strengths of the thesis

This study has several strengths, including:

- **Novelty and knowledge contribution**

To the knowledge of the author, there is no software tool that can forecast water parameters throughout all Canadian hydrometric stations. This thesis is one of the first studies contributing to the area of information technology combined with civil engineering applied to water forecast across all Canada. Findings from our work contribute not only to have a better view of all HYDAT

stations across Canada with their respective information but also facilitate to environmental policies for decision-makers and concerned entities (e.g. engineers, biologists, hydrologists…).

- **Development of a new web interface application for water forecasting in Canada**

This is the first web interface realized for all HYDAT stations across Canada. Web-based systems provide entry to climate change associated information for a wide variety of end-users. The particular complexity associated with climate change information in the actual development of user-friendly applications remains a challenge (Liu, 2003). Thus, an end-user requires appropriate web-based interface for easy access to climate change data. A major challenge for climate data web interface revolves around accessibility and ease of use the application by users. The *Universal Canadian forecast Application* (Stitou, Adnane, 2018©) solves these issues by allowing users to search the database based on multiple criteria and easily visualize the data on the interface:

- Station Name
- Station Number
- Data Types available: Level, Flow, Sedimentation

5.1.3 Limitations of the thesis

This study has some limitations, including:

- **Limited data available for some stations**

Due to the lack of data for some Canadian Stations, the results of this thesis might not be as precise as desired. Therefore, more data sampling is necessary to get results closer to perfection.

- **Lack of comparisons of our results with other studies**

Because of the limited number of studies on water foresting in Canada using software application, we discussed our results mainly in the light of the general literature on Canadian forecasting. No comparisons with the results of previous studies were possible to confirm or contradicts our findings.

5.1.4 Future Area of research

Future areas of studies could include adding exogenous factors to improve the forecasts. Indeed, it is possible to improve forecasts by taking into account extreme events recorded in the past that do not follow the main trend of the time series.

Besides, we could implement the US stations from the US Hydrological database into the app with the long-term objective of enlarging the monitoring of water bodies to other countries in Europe and Asia. Finally, it would be interesting to compare the statistical SARIMA method with empirical methods such as Artificial Neural Networks and other Machine learning methods that are being developed

References

Abraham, B. and Ledolter, J. (2005). *Statistical Methods for Forecasting*, Wiley & Sons, 1983 (Second edition, 2005)

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov, and F. Cs´aki, eds., 2nd International Symposium on Information Theory, Akad´emia Kiad´o, Budapest, pp. 267–281.

Akgün, B. (2003). Identification of Periodic Autoregressive Moving Average Models, M.Sc, Dept. of Statistics, Middle East Technical University, Turky.

Ankit Kumar Nigam. (2016). "Analysis of Water Demand and Forecasting Water Demand for Year 2048 Jabalpur City", SSRG International Journal of Civil Engineering (SSRG - IJCE), V3(7),41-46 July 2016. ISSN:2348 – 8352. www.internationaljournalssrg.org/IJCE/index.html. Published by: Seventh Sense Research Group.

Armstrong, J. S., & Pagell, R. (2003). The Ombudsman: Reaping Benefits from Management Research: Lessons from the Forecasting Principles Project. Retrieved from https://repository.upenn.edu/marketing_papers/102

Bakker, Karen. 2007. Eau Canada: The Future of Canada's Water. Edited by Karen Bakker. UBC Press, University of British Columbia, Vancouver BC.

Beven. K. 2001. Rainfall-Runoff Modeling: A Primer. John Wiley and Sons Ltd.

Boland, P., Forecasting Water Use: A Tutorial, IN: Torno, H.C. (ed.), Computer Application in Water Resources, p907-916, 1985.

Box, G.E.P., Jenkins, G. (1970). Time Series Analysis, Forecasting, and Control,", Holden-Day, San Francisco, CA.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). Time Series Analysis: Forecasting and Control (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: a practical information-theoretic approach, 2nd edition. Springer-Verlag, New York.

Burn, Donald H and Whitfield, Paul H.(2016). Changes in floods and flood regimes in Canada, Canadian Water Resources Journal/Revue Canadienne des Ressources hydriques, 41(1-2), 2016, 139 – 150.

Buttle JM, Allen DM, Caissie D, Davison B, Hayashi M, Peters DL, Pomeroy JW, Simonovic SP, St.-Hilaire A, Whitfield PH. 2016. Flood processes in Canada: regional and special aspects. Canadian Water Resources Journal 41. DOI:10.1080/07011784.2015.1131629.

Chambers, P.A., A. M. Anderson, C. Bernard, L.J. Gregorich, B. McConkey, P.H. Milburn, J. Painchaud, N.K. Patni, R.R. Simard, and L.J.P van Vliet. 2000a. Surface Water Quality. In: Coote, D.R., and L.J. Gregorich (Eds.). The Health of our Water: Toward sustainable agriculture in Canada. Research Branch, Agriculture and Agri-Food Canada, Publication 2020/E.

Chambers, P.A., Dale, A.R., Scrimgeour, G.J., and M.L. Bothwell. 2000b. Nutrient enrichment of northern rivers in response to pulp mill and municipal discharges. Journal of Aquatic Ecosystem Stress Recovery 8: 53-66.

Charles, Akinmutmi. (2011) , A seasonal Arima Modelling of Residential Coal Consumption Series in Nigeria, Department of Mathematics, Ahmadu Bello University, Nigeria.

Chatfield, C. (1996). The Analysis of time series, 5th ed., Chapman & Hall, New York.

Chen Yuechun, Application of Time Series Analysis to Water Demand Prediction, IN: Coulbeck, B. and C.H. Orr (ed.), Computer Applications in Water Supply: Voll, p289-296, 1988.

Chiang, P.-K. P. Willems, J. Berlamont, "A conceptual river model to support real-time flood control," in Demer,' River Flow 2010 - Dittrich, Koll, Aberle & Geisenhainer, 2010, pp. 1407-1414.

Chiew, F. H. S., S. L. Zhou, and T. A. McMahon (2003), Use of seasonal streamflow forecasts in water resources management, J. Hydrol., 270(1–2), 135–144, doi:10.1016/S0022-1694(02)00292-5.

Cochrane, John. (1997). Time Series for Macroeconomics and Finance. Graduate School of Business, University of Chicago, and spring

DeKay, C. F., The Evolution of Water Demand Forecasting, Management And Operations, 77(10):54-61, 1985.

Déry, S.J. and Wood, E.F. 2005. Decreasing river discharge in northern Canada. Geophysical Research Letters 32:1-4. doi:10.1029/2005GL022845, 2005.

Dyck, R., Cool, G., Rodriguez, M., & Sadiq, R. (2014). Treatment, residual chlorine and season as factors affecting variability of trihalomethanes in small drinking water systems. Frontiers of Environmental Science & Engineering, 1–9. http://doi.org/10.1007/s11783-014-0750-1

Environment and Climate Change Canada. (2013).Water – How we manage it. Retrieved from http://www.ec.gc.ca/eau-water/default.asp?lang=En&n=3DC41CC0-1

Environment Canada. (2012). Environment. Retrieved from https://www150.statcan.gc.ca/n1/pub/11-402-x/2011000/chap/env/env-eng.htm

Environment Canada. (1994). Guidance document on collection and preparation of sediments for physiochemical characterization and biological testing. Ottawa (ON): Environment Canada. Environmental Protection Series Report EPS 1/RM/29.

Environment Canada. (2001). HYDAT CD-ROM User's manual. Retrieved from https://www.mcgill.ca/library/files/library/HydatManual_PDF.pdf

Environment and Climate Change Canada. (2013). Water – The Transporter. Retrieved from http://www.ec.gc.ca/eau-water/default.asp?lang=en&n=ADB791B6-1

Environment Canada. (2014). Canadian Environmental Sustainability Indicators: Data Sources and Methods for the Residential Water Use in Canada Indicator. Retrieved from https://www.ec.gc.ca/indicateurs-indicators/default.asp?lang=en&n=D43360E1-1

Environment Canada. (2016). Water survey: Version 2.1. Retrieved from http://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/ECDataExplorer_EN.pdf

Faraway, J., Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data. Applied Statistics 47, pages: 231–250.

George, S.S., Energy Forecasting Techniques: an Overview, IN: Morlan (ed.), Energy Forecasting: Proceedings of the Energy Division Session of the ASCE Conference in Detroit, New York, p12-30,1985.

Georgakakos, K.P., R. Krzysztofowicz, "Probabilistic and Ensemble Forecasting (Editorial)," Journal of Hydrology, 249(1), 2001, pp.1-4

Government of Canada. (2018).Historical Hydrometric Data Search Results. Retrieved from https://wateroffice.ec.gc.ca/search/historical_results_e.html?search_type=station_name&station_name=&start_year=1850&end_year=2019&minimum_years=&gross_drainage_operator=%3E&gross_drainage_area=&effective_drainage_operator=%3E&effective_drainage_area=

Government of Canada. (2018b). Sediment Data Search Results. Retrieved from https://wateroffice.ec.gc.ca/search/sediment_results_e.html?search_type=station_name&station_name=&start_year=1850&end_year=2003&minimum_years=&gross_drainage_operator=%3E&gross_drainage_area=&effective_drainage_operator=%3E&effective_drainage_area=

Government of Canada (2018c).National Water Data Archive: HYDAT. Retrieved from https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html

Gottman, M John. (1981). Time Series Analysis, A Comprehensive Introduction for Social Scientists. Cambridge University Press, British

Gyasi-agyei, Kwame. (2012), Analysis and modeling of prevalence of measles in the ashnti region of chana, Kwame Nkrumah University of Science and Technology, West Africa.

Gilbert, R.O. (1987). Statistical Methods for Environmental Pollution Monitoring, Wiley, NY.

Hanke, J. E., Reitsch, A. G. and Wichern, D. (1998). Business Forecasting (6th ed.), Englewood Cliffs, NJ: Prentice-Hall.

Heaton, D. 2003. Effects of water level fluctuations. In State of the Great Lakes 2003. Edited by Environment Canada and U.S. Environmental Protection Agency. Governments of Canada and the United States. pp. 86-87.

Hipel, K.W., McLeod. A.I. (1994). Time Series Modelling of Water Resources and Environmental Systems. Amsterdam, Elsevier.

Hirabayashi Y, Mahendran R, Koirala SR, Konoshima L, Yamazaki D, Watanabe S, MKim H, Kanae S. 2013. Global flood risk under climate change. Nature Climate Change 3: 816–821.

Hornik, K., Stinchcombe, M., White, H. (1989). Multilayer feed-forward networks are universal approximators", Neural Networks 2, pages: 359–366.

Hoyer, M.V. and J.R. Jones. 1983. Factors affecting the relation between phosphorus and chlorophyll-a in Midwestern reservoirs. Canadian Journal of Fisheries and Aquatic Sciences 40: 192-199.

Hyndman, R.J. & Kostenko, A.V., 2007. "Minimum Sample Size requirements for Seasonal Forecasting Models," Foresight: The International Journal of Applied Forecasting, International Institute of Forecasters, issue 6, pages 12-15, Spring.

Javorek, S.K., and Grant, M.C. 2010. Trends in wildlife habitat capacity on agricultural land in Canada, 1986-2006. Canadian Biodiversity: Ecosystem Status and Trends 2010, Technical Thematic Report Series No. 14. Canadian Councils of Resource Ministers. Ottawa, ON. In press.

Kame'enui, A. (2003). "Water demand forecasting in the Puget Sound region: Short and long-term models."M.S. thesis, Dept. of Civil and Environmental Engineering, Univ. of Washington, Seattle, WA.

Kendall, M.G. (1975). Rank Correlation Methods, 4th edition, Charles Griffin, London.

Kihoro, J.M., Otieno, R.O., Wafula, C. (2004). Seasonal Time Series Forecasting: A Comparative Study of ARIMA and ANN Models. *African Journal of Science and Technology* (AJST), 5(2), pages: 41-49.

Kundzewicz, Z.W. et al., 2013, "Flood risk and climate change: global and regional perspectives," Hydrological Sciences Journal, Vol. 59, no. 1, p. 1–28.

Last, W.M. and Ginn, F.M. 2005. Saline systems of the Great Plains of Western Canada: an overview of the limnogeology and paleolimnology. Saline Systems 1:10.

Libby, L.W. and W.G. Boggess. 1990. Agriculture and water quality: where are we and why? In: Agriculture and Water Quality: International Perspectives. J.B. Braden and S.B. Lovejoy (Eds.). Lynne Rienner Publishers, Boulder Colorado.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. Link, Google Scholar

Makridakis, S. G., Wheelwright, S. C. & Hyndman, R. J. (1998). Forecasting: methods and applications (3rd ed.), New York : John Wiley & Sons.

Mann, H.B. (1945). Non-parametric tests against trend, Econometrica 13:163-171.

McClelland, J.W., Dery, S.J., Peterson, B.J., Holmes, R.M. and Wood, E.F. 2006. A Pan-Arctic evaluation of changes in river discharge during the latter half of the 20th century. Geophysical Research Letters 33:1-4.

McDonald, A.T., and D. Kay, Water Resources: Issues and Strategies, Jone Wiley & Sons, New York, 1988.

McFarlane, S., and E. Nilsen. (2003). Canada West Foundation Water Report: On Tap, Urban Water Issues in Canada Discussion Paper. Canada West Foundation, Calgary, Alberta.

Merkuryeva Galina V. & Kornevs Maksims, 2013. "Water Flow Forecasting and River Simulation for Flood Risk Analysis," Information Technology and Management Science, Sciendo, vol. 16(1), pages 42-46.

Milly PCD, Wetherald RT, Delworth TL. (2002). Increasing risk of great floods in a changing climate. Nature 415: 514–517.

Mitchell, P.A. and E.E. Prepas (Eds). 1990. The Atlas of Alberta Lakes. The University of Alberta Press, Edmonton. 675 pp.

Monk, W.A., Baird, D.J., Curry, R.A., Glozier, N. and Peters, D.L. 2010. Ecosystem status and trends report: biodiversity in Canadian lakes and rivers. Canadian Biodiversity: Ecosystem Status and Trends 2010, Technical Thematic Report Series No. 20. Canadian Councils of Resource Ministers. Ottawa, ON. In press.

Nash, J. E., and Sutcliffe, J. V. 1970. River flow forecasting through conceptual models: 1. A discussion of principles. Journal of Hydrology, 10(3), 282-290.

Park, H.(1999). Forecasting Three-Month Treasury Bills Using ARIMA and GARCH Models", Econ 930, Department of Economics, Kansas State University.

Parrelli. R. (2001). Introduction to ARCH & GARCH models", Optional TA Handouts, Econ 472 Department of Economics, University of Illinois.

Peterson, B.J., Holmes, R.M., McClelland, J.W., Volosmarty, C.J. and Lammers, R.B. 2002. Increasing river discharge to the Arctic Ocean. Science 298:2171-2173.

Peterson, B.J., McClelland, J., Curry, R., Holmes, R.M., Walsh, J.E. and Aargaard, K. 2006. Trajectory shifts in the Arctic and Subarctic freshwater cycle. Science 313:1061-1066.

Pohlert, T. (2018). "Non-Parametric Trend Tests and Change-Point Detection". R-package `trend`. Accessed on: 19 April, 2018. Retrieved from https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

Pokhrel, P., Robertson, D. E., and Wang, Q. J.: A Bayesian joint probability post-processor for reducing errors and quantifying uncertainty in monthly streamflow predictions, Hydrol. Earth Syst. Sci., 17, 795-804, https://doi.org/10.5194/hess-17-795-2013, 2013.

Postel, S.L. 1992. Last oasis: facing water scarcity. Worldwatch Institute. New York, NY. 240 p.

Prasifka, D.W., Current Trends in Water Supply Planning: Issues, Concepts and Risks, Van Nostrand Reinhold Company, New York, 1988.

Quevedo, J., G. Cembrano, A. Valls, and J. Serra, Time Series Modelling of Water Demand—A Study on Short-Term and Long-Term Predictions, IN: Coulbeck, B. and C.H. Orr (ed.), Computer Application in Water Supply, p268-287, 1988.

Ricciardi, A. and Rasmussen, J.B. 1999. Extinction rates of North American freshwater fauna. Conservation Biology 13:1220-1222.

Saunders, J. O., and Wenig, M. 2006. Whose water? Canadian water management and the challenges of jurisdictional fragmentation. In Eau Canada: The Future of Canada's Water, ed. K. Bakker, 119-141. Vancouver: University of British Columbia Press.

Schindler, D.W. and Donahue, W.F. 2006. An impending water crisis in Canada's western Prairie Provinces. Proceedings of the National Academy of Sciences of the United States of America 103:7210-7216.

Schindler, D.W. 1997. Widespread effects of climate warming on freshwater ecosystems in North America. Hydrological Processes 11:1043-1067.

Skinner, J.A., Lewis, K.A., Bardon, K.S., Tucker, P., Carr, J.A. and B.J. Chambers. 1997. An Overview of the Environmental Impact of Agriculture in the UK. Environmental Management 50: 111-128.

Spence, C. and Burke, A. (2008). Estimates of Canadian Arctic Archipelago runoff from observed hydrometric data," Journal of Hydrology, Vol. 362, pp. 247–259.

Statistics Canada. (2009). Environment, Energy and Transportation Statistics Division, special tabulation from Pearse, P.H., F.Bertrand and J.W. MacLaren, 1985, Currents of Change: Final Report of the Inquiry on Federal Water Policy, Environment Canada, Ottawa.

Statistics Canada. (2011). Survey of Drinking Water Plants. Retrieved from www.statcan.gc.ca/pub/16.../16-403-x2013001-eng.pdf

Statistics Canada. (2017). Section 2: Freshwater supply and demand. Retrieved from https://www150.statcan.gc.ca/n1/pub/16-201-x/2017000/sec-2-eng.htm

Statistic Canada. (2017). Human Activity and the Environment: Freshwater in Canada. Retrieved from https://www150.statcan.gc.ca/n1/en/pub/16-201-x/16-201-x2017000-eng.pdf?st=ASCddoOo

The Federal Water Policy. (1993). MUNICIPAL WATER ISSUES IN CANADA. http://publications.gc.ca/Collection-R/LoPBdP/BP/bp333-e.htm#(40)txt

University of Wisconsin (2013). Great Lakes and Wisconsin Water Facts. Great Lakes and Fresh Water. Retrieved from http://www.seagrant.wisc.edu/Home/AboutUsSection/PressRoom/Details.aspx?PostID=796

Van der Kamp, G., Keir, D. and Evans, M.S. 2008. Long-term water level changes in closed-basin lakes of the Canadian prairies. Canadian Water Resources Journal 33:23-38.

Wang, Q. J., and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, Water Resour. Res., 47, W02546, doi:10.1029/2010wr009333, 2011.

Wang, S., Y. Yang, Y. Luo and A. Rivera, 2013, "Spatial and seasonal variations in evapotranspiration over Canada's landmass," Hydrology and Earth System Sciences, Vol. 17, no. 9, pp. 3561–3575, doi: 10.5194/ hess-17-3561-2013.

Water Survey of Canada. (2018). Real-Time Hydrometric Data Map Search. Retrieved from https://wateroffice.ec.gc.ca/google_map/google_map_e.html?map_type=real_time&search_type=province&province=all

Water Survey of Canada. (2012).The Hydrometric Network. Retrieved http://www.ec.gc.ca/rhc-wsc/default.asp?lang%20=En&n=E228B6E8-1

Webster, P. J., and C. Hoyos, 2004: Prediction of monsoon rainfall and river discharge on 15–30-day time scales. *Bull. Amer. Meteor. Soc.*, 85, 1745–1765, https://doi.org/10.1175/BAMS-85-11-1745. Link, Google Scholar

Whitfield PH. 2012. Floods in future climates: a review. Journal of Flood Risk Management 5: 336–365.

WHO (World Health Organization). (2011). Guidelines for Drinking Water Quality (Vol. 4). Geneva.

Wilcox, D.A., Thompson, T.A., Booth, R.K. and Nicholas, J.R. 2007. Lake-level variability and water availability in the Great Lakes. U.S. Geological Survey Circular No. 1311. 25 p.

Zhang, X., Brown, R., Vincent, L., Skinner, W., Feng, Y. and Mekis, E. 2010. Canadian climate trends 1950-2007. Canadian Biodiversity: Ecosystem Status and Trends 2010, Technical Thematic Report Series No. 5. Canadian Councils of Resource Ministers. Ottawa, ON. In press.

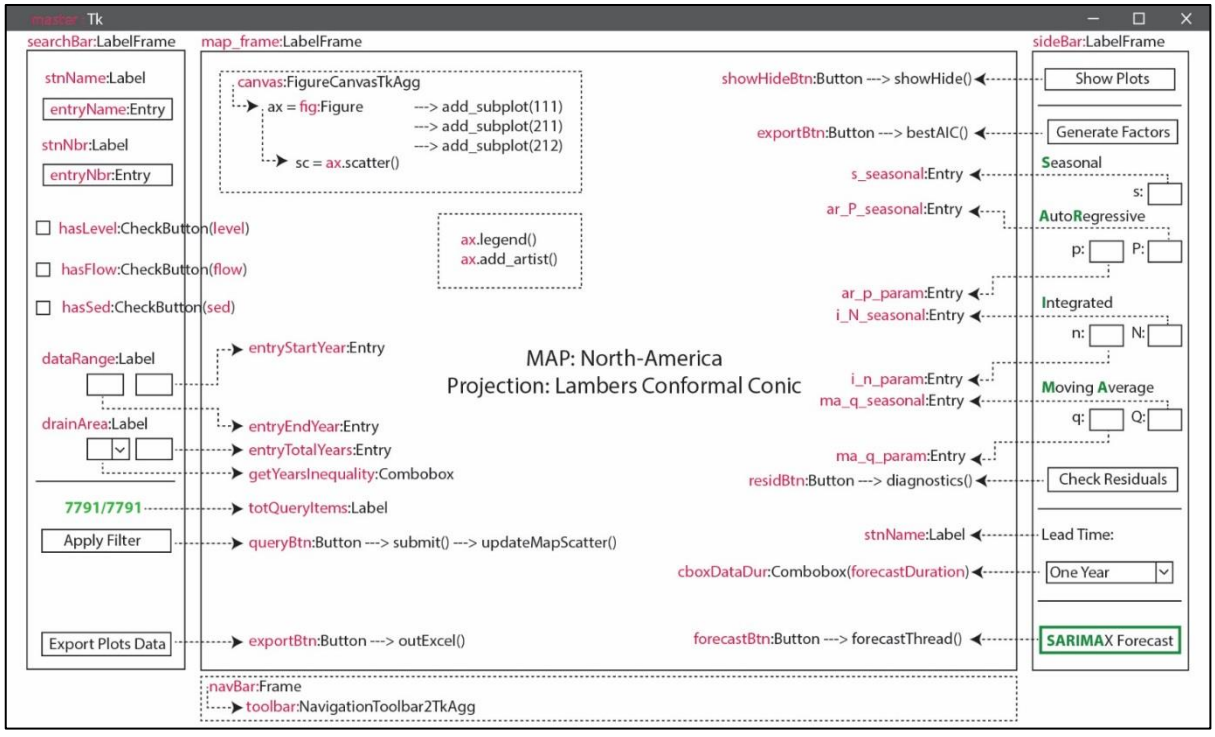
Zhang, X.B., Harvey, K.D., Hogg, W.D. and Yuzyk, T.R. 2001. Trends in Canadian streamflow. Water Resources Research 37:987-998.

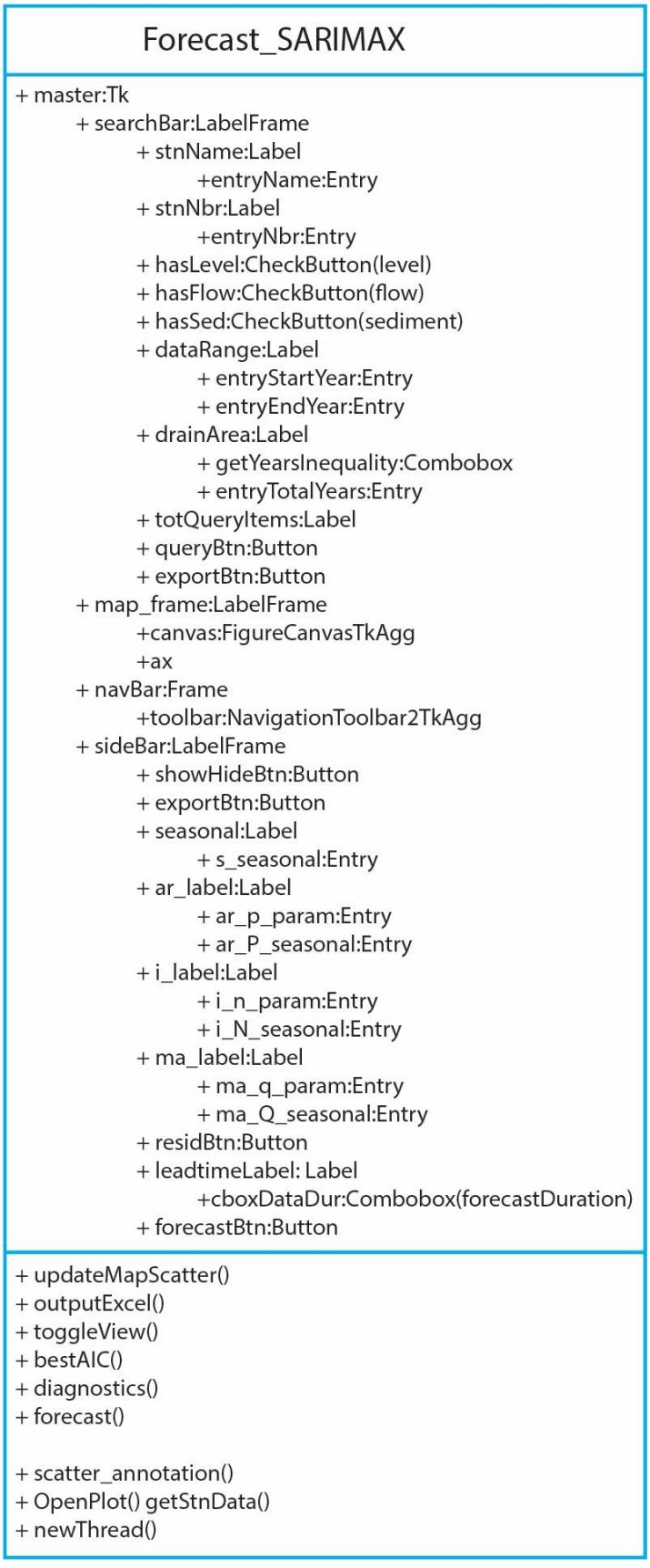
Appendix I- Universal Canadian forecast Application (Stitou, Adnane, 2018©)

A.1 Building the application and Data screening

An application was developed to search the database and assess and forecast the data. This python based application is useful in querying the database and visualizing data such as generating forecasts using the seasonal ARIMA model. The selection process of appropriate stations was done as follows: 1) selecting only stations with complete data, 2) filling missing values with reasonable interpolations.

We chose to create a user-friendly application in the python language to help users generate their initial course SARIMA model forecasts. The fact that the application interface is python based allows for multiplatform compatibility. Indeed, our Python-based application is accessible through all under all mainstream operating systems (Windows, Linux, Mac, Web, Mobile) and generates live forecasts in a user-friendly interface. This can prove to be a key element for developers to extend the potential of this application.




Forecast_SARIMAX
+ master:Tk
+ searchBar:LabelFrame
+ stnName:Label
+entryName:Entry
+ stnNbr:Label
+entryNbr:Entry
+ hasLevel:CheckButton(level)
+ hasFlow:CheckButton(flow)
+ hasSed:CheckButton(sediment)
+ dataRange:Label
+ entryStartYear:Entry
+ entryEndYear:Entry
+ drainArea:Label
+ getYearsInequality:Combobox
+ entryTotalYears:Entry
+ totQueryItems:Label
+ queryBtn:Button
+ exportBtn:Button
+ map_frame:LabelFrame
+canvas:FigureCanvasTkAgg
+ax
+ navBar:Frame
+toolbar:NavigationToolbar2TkAgg
+ sideBar:LabelFrame
+ showHideBtn:Button
+ exportBtn:Button
+ seasonal:Label
+ s_seasonal:Entry
+ ar_label:Label
+ ar_p_param:Entry
+ ar_P_seasonal:Entry
+ i_label:Label
+ i_n_param:Entry
+ i_N_seasonal:Entry
+ ma_label:Label
+ ma_q_param:Entry
+ ma_Q_seasonal:Entry
+ residBtn:Button
+ leadtimeLabel: Label
+cboxDataDur:Combobox(forecastDuration)
+ forecastBtn:Button
+ updateMapScatter()
+ outputExcel()
+ toggleView()
+ bestAIC()
+ diagnostics()
+ forecast()
+ scatter_annotation()
+ OpenPlot() getStnData()
+ newThread()

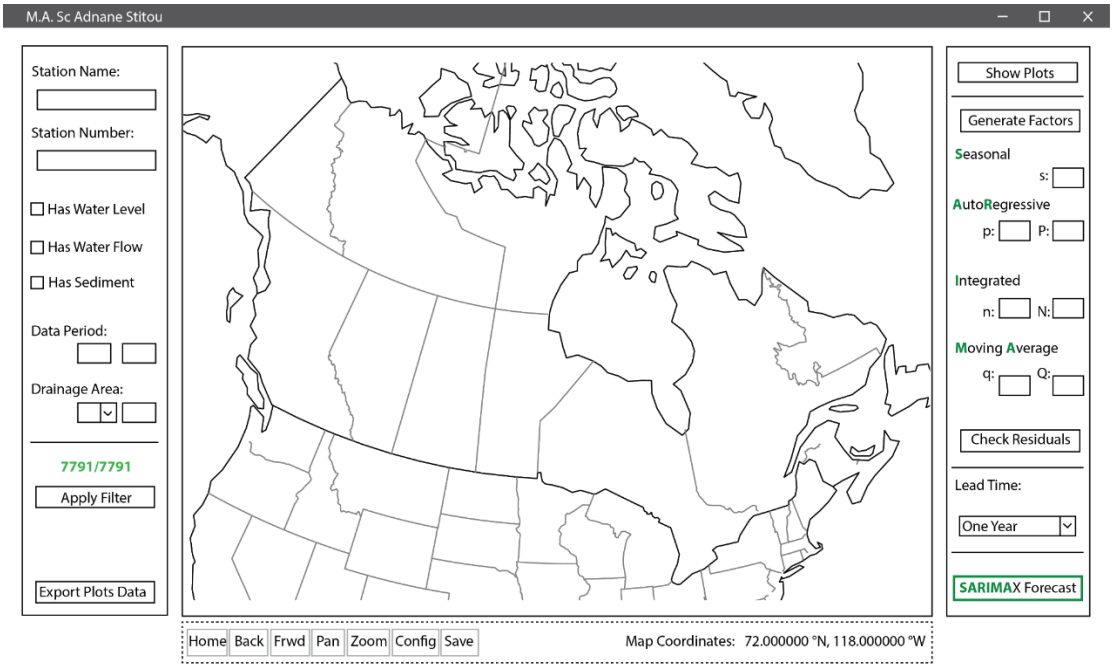


Figure 54. Canada map

A.1.1 Description of the interface

The interface presents a search area where it is possible to search by station name, station number, data types available (water level, water flow, sediment concentration), years of data available, data recording methods, and conditions affecting the stations (natural…). Based on these criteria, an SQL query statement is generated, and the results are automatically used to update the Canada map with all stations respecting the conditions selected. Users can hover on top of the station to visualize its name and location. They can zoom in and out and click on the station of interest.

The interface is divided into three components (Figure 54). The first is the Search section:

- Station Name: search by station name
- Station Number: search by station number as defined by Environment Canada in HYDAT
- Has Water Level: the station has water level data in m available
- Has Water Flow: the station has flow data in m^3/s available

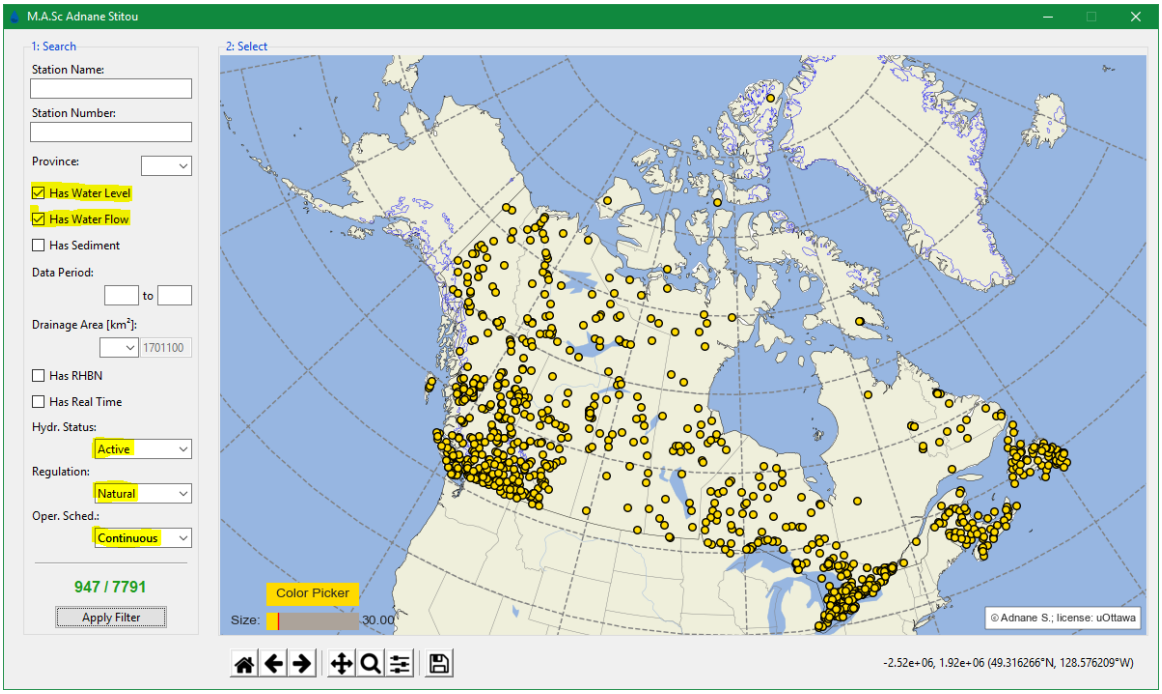


Figure 55.Hydrological Stations (7791) available in the HYDAT database

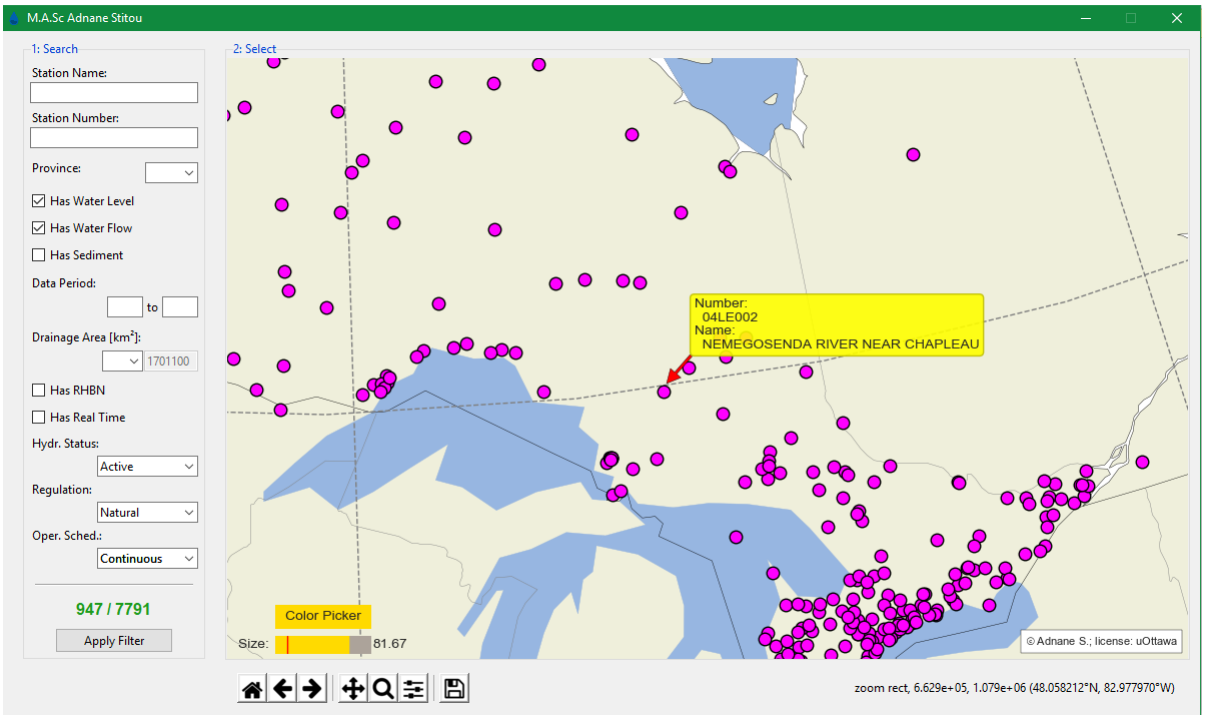


Figure 56.Search by station name and data available

Once we open a station, we get all data available which might include water level (m), water flow (m3/s), and sediment concentration (ml per liter) and sediments transported. For each data type, we have plots of time series for daily, monthly, annual and annual peaks. On the left side, we see all tools necessary to generate a SARIMA forecast following the methodology described previously for this thesis. First, we can select the time series type we would like to analyze, we can:

- Generate a seasonal decomposition of the time series
- Autocorrelation function (ACF) and Partial autocorrecting function (PACF) plots
- Generate man-Kendall plot to determine whether the time series displays a trend

Once we visualize our data, we are ready to generate the best SARIMA model based on the Akaike information criteria. After a few seconds or minutes, we get the best factors for seasonality of 12, 4 or 3 depending on the user choice. We can check the distribution of the residuals by generating QQ-plots, correlograms, histograms, standardized residual plots. We can still improve our model in case the residuals are not normally distributed. It is possible to do a second-degree search to improve the model further but may take more time computationally. Once we are satisfied with our model, we are ready to generate a fit that will give us an indication of the precision of the model using the MSE, RMSE, and efficiency factor E. These factors are an indication of the precision of the model.

The user can choose to do a second-degree search, meaning 0, 1 and 2 will be used to generate factors combinations for the Akaike information criterion. This will take longer computational time to generate but might give us more complex model that will capture our data in a more precise way (Figure 57).

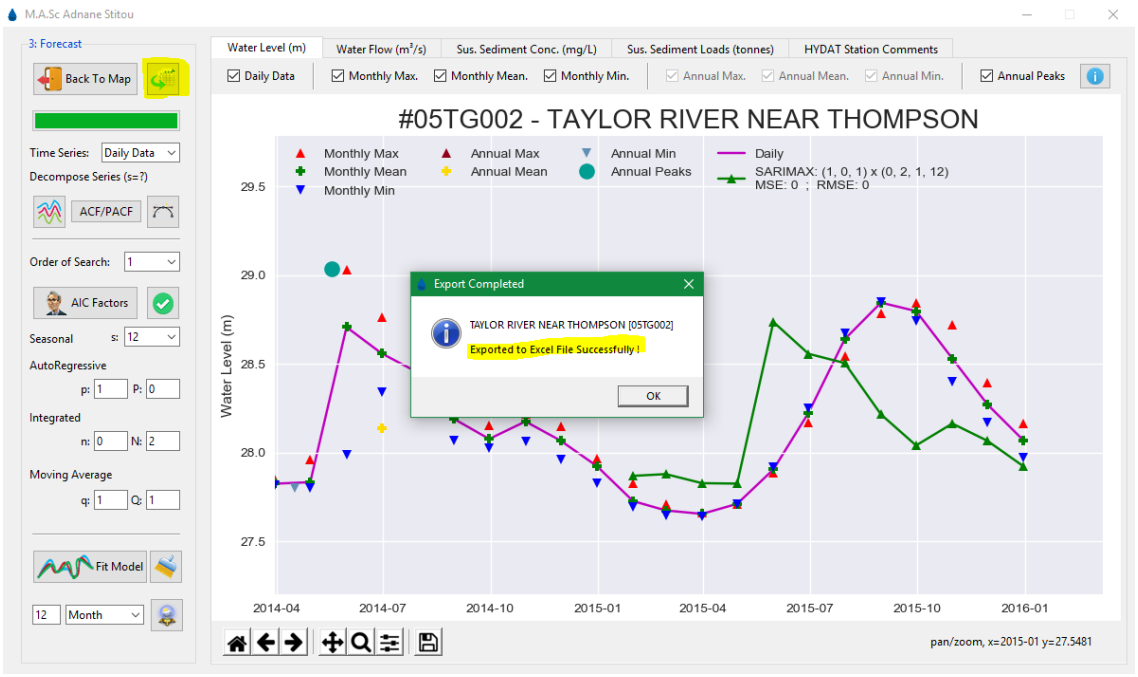


Figure 57.Visualizing the data (Water Level, Water Flow, Sediment) of the Station

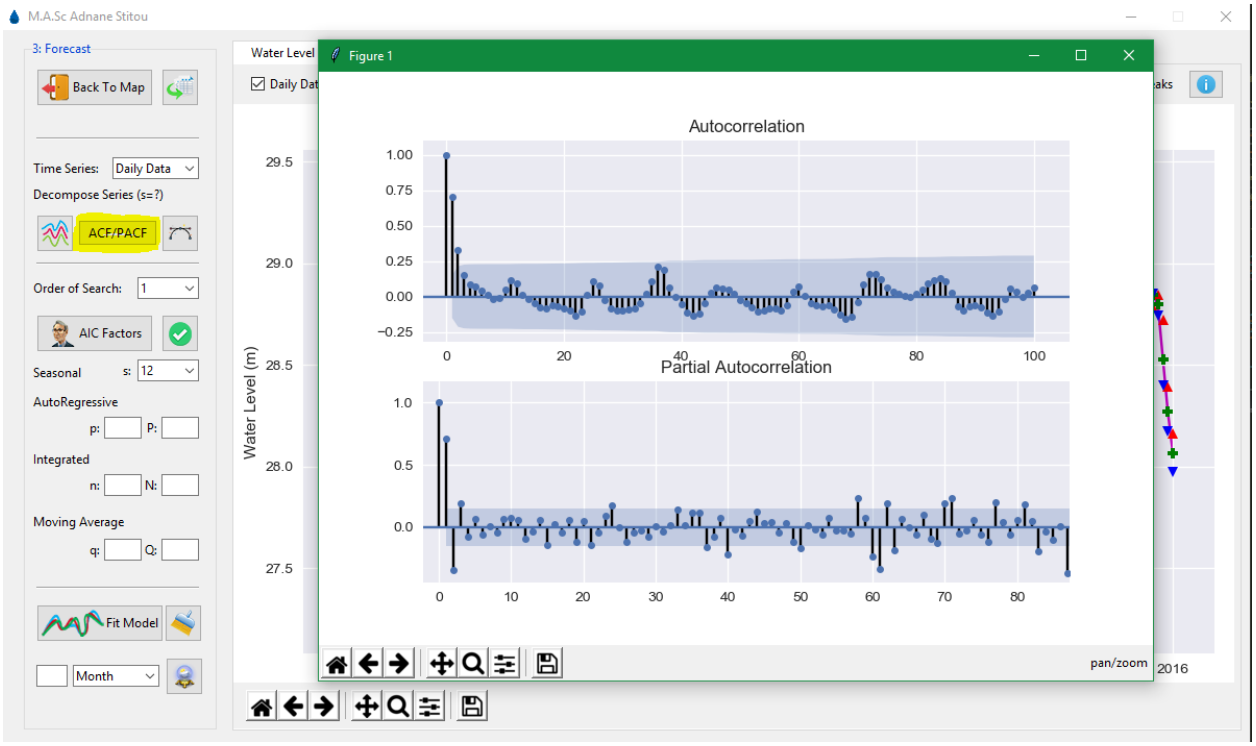


Figure 58.Visualizing the data (Autocorrelation & Partial Autocorrelation)

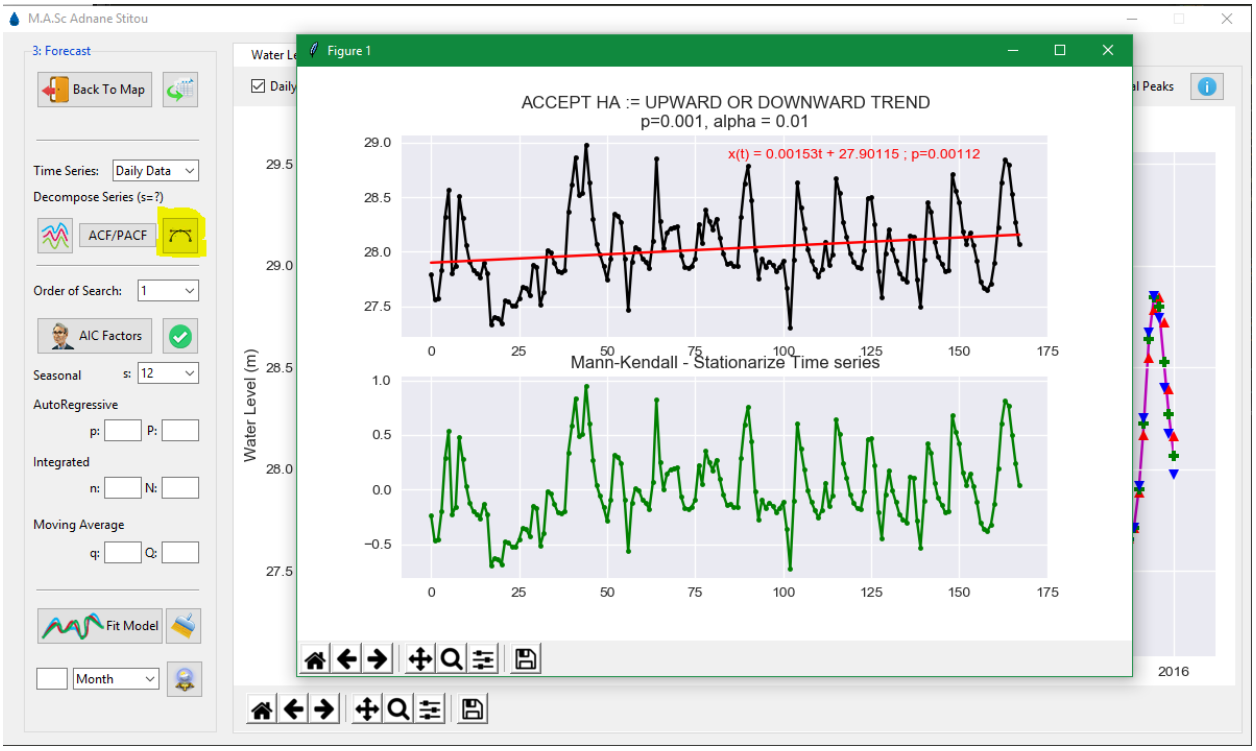


Figure 59.Visualizing the data – Trend

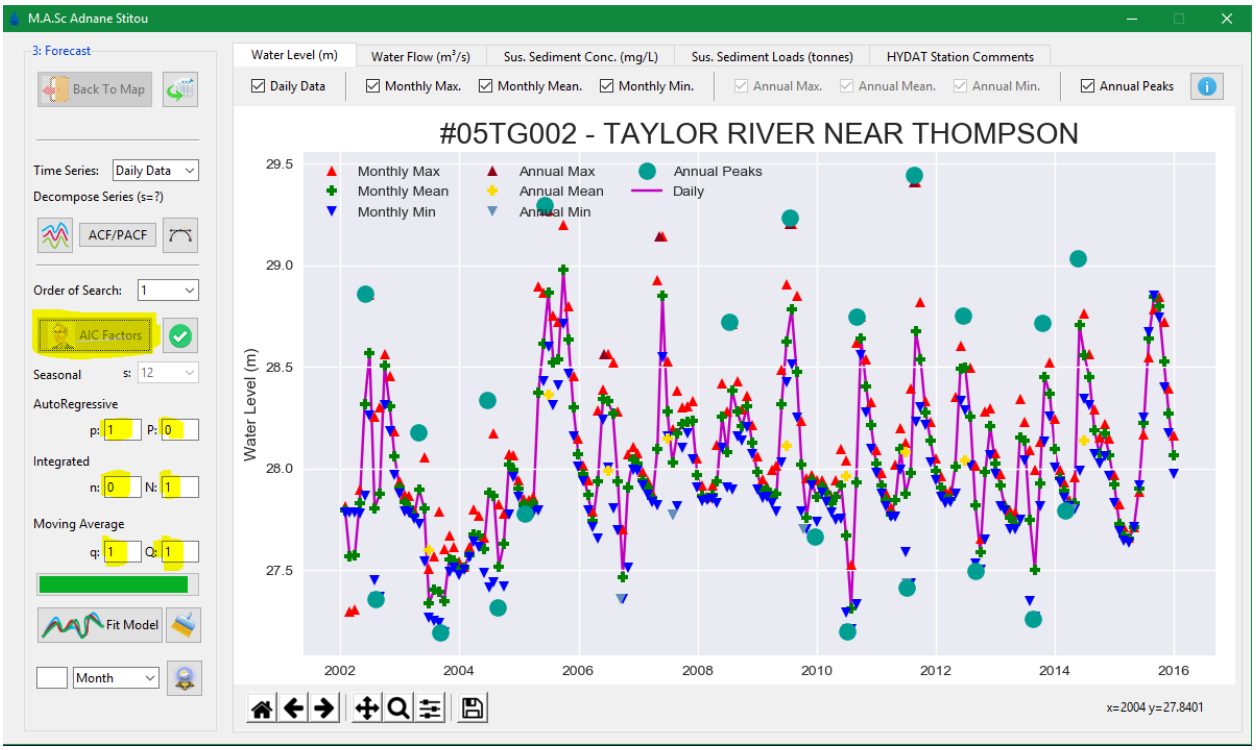
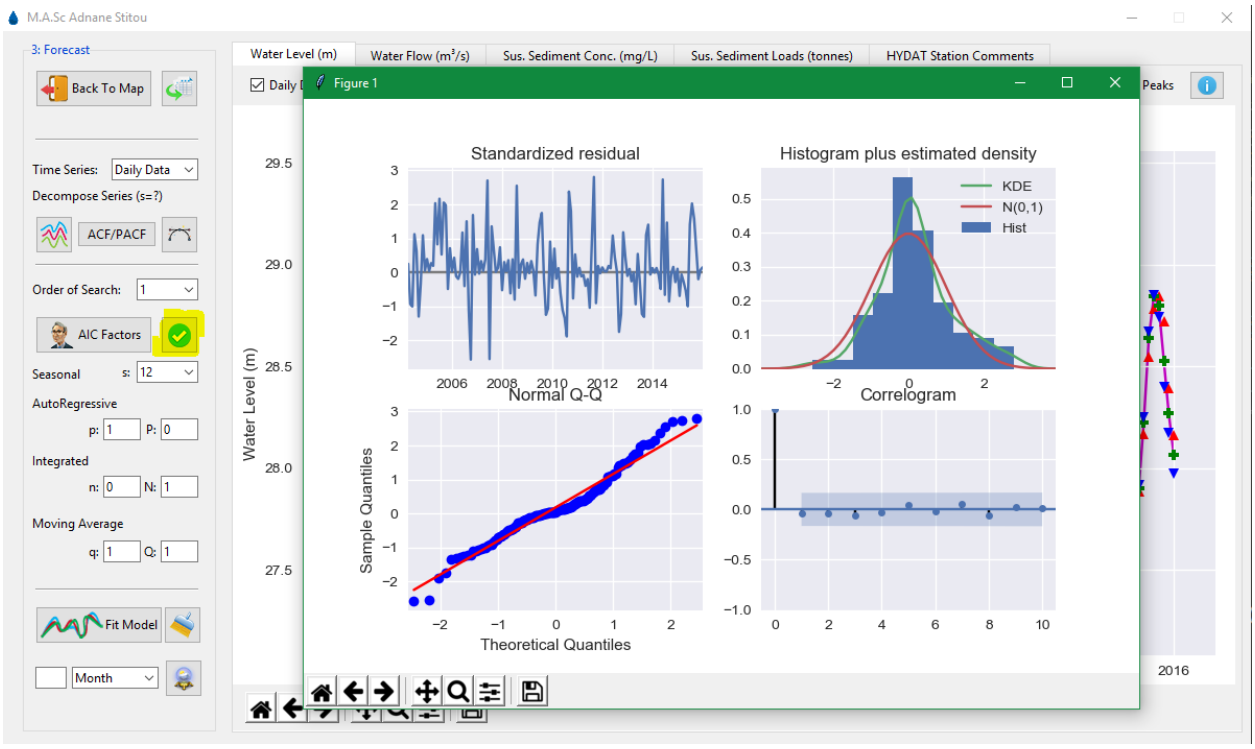


Figure 60.Generating AIC Factors



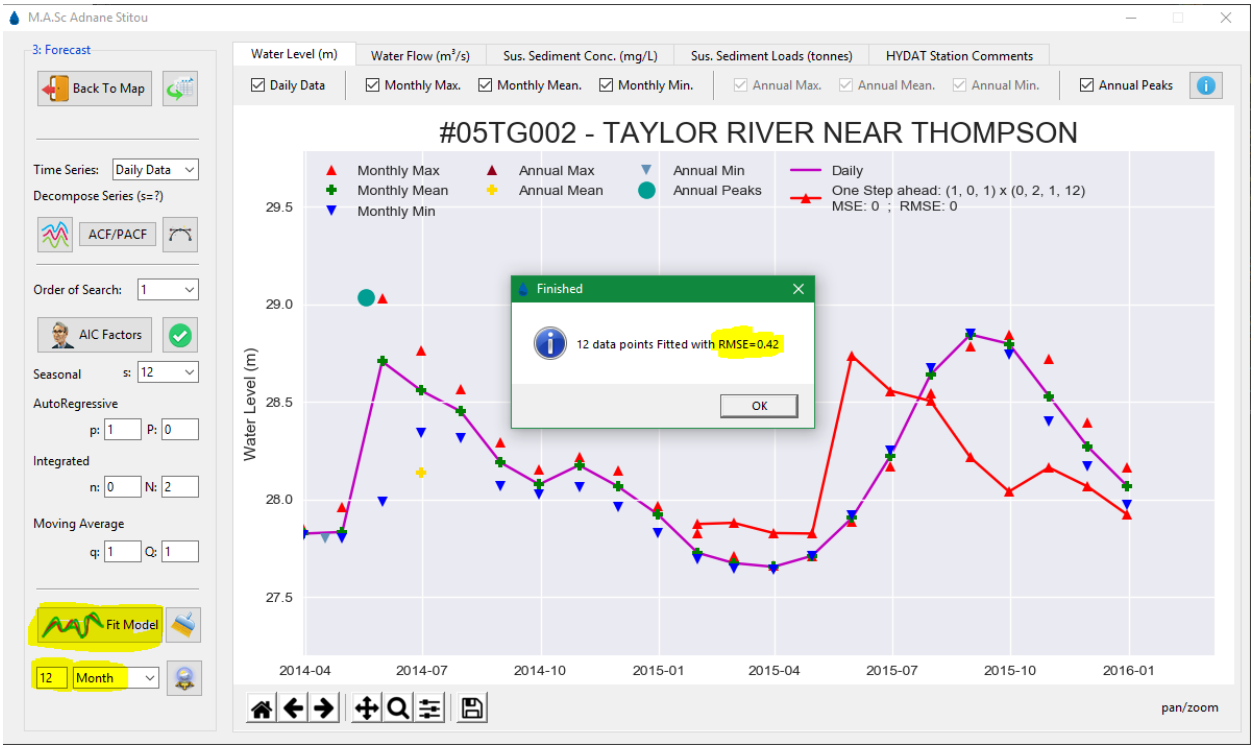


Figure 61. Generating rmse error of initial fit

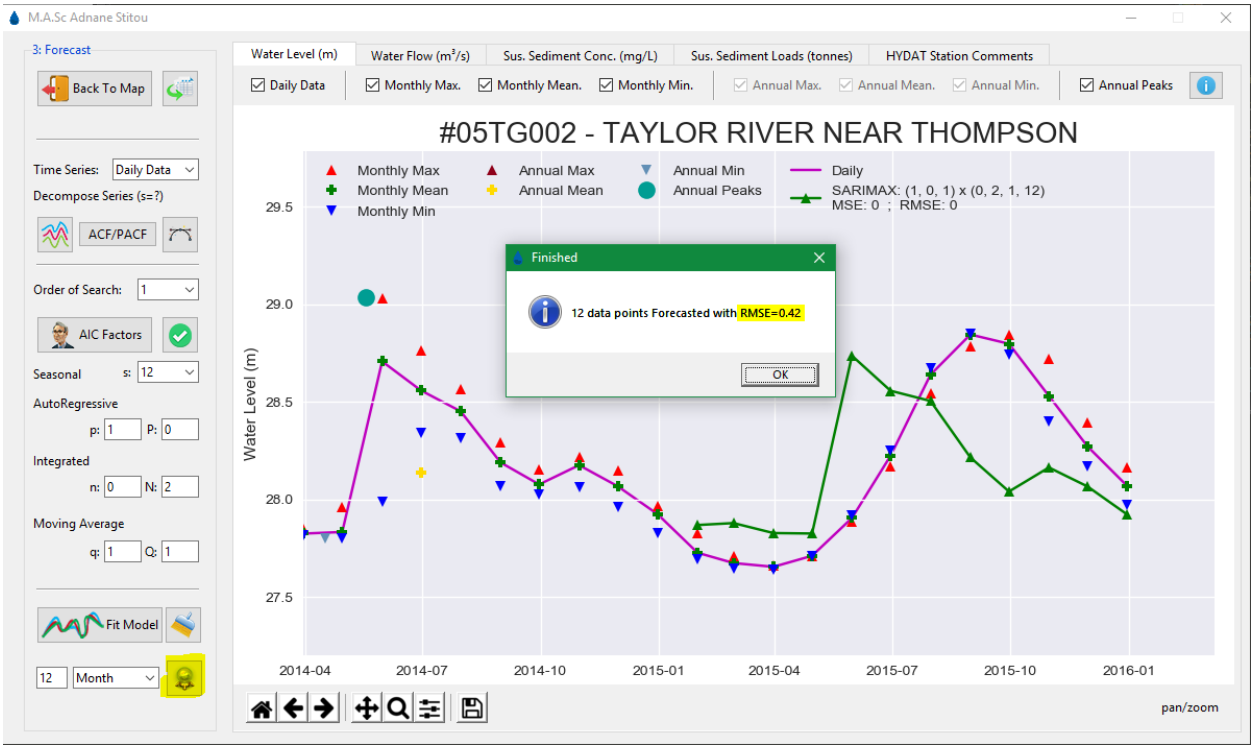


Figure 62.Generating Forecast

Finally, we can compare the theoretical model fit with an actual forecast by eliminating known data from the time series and generating a new plot with new factors MSE, RMSE, and efficiency factors to check how much the model captures the general trend of the time series, and is not just overfitting the existing data.

A.1.2 Extendibility of the app and future area of research

This user-friendly tool will make life easier for decision-makers to forecast water body's hydrological characteristics around Canada by generating reasonable and reliable efficient forecasts. This is a keys aspect for many fields such as the hydroelectric industries, municipalities (water consumption regulation), construction companies (building near water bodies), civil engineers (e.g. building bridges) which rely greatly on water flow forecasts to plan for future

operations. A possible improvement of the interface could be the inclusion of exogenous factors that take into account specifies unique to each station.

Also, the extendibility of this application should allow us to include USA hydrological stations to the map and maybe someday extend it all over the continent and the rest of the world in a universal monitoring program. This could allow us to understand better the dynamic trends and interactions displayed by water bodies such as rivers, lakes, and oceans.

Another way to compare and maybe improve our forecast results is to compare the efficiency of statistical methods such as SARIMA with newer machine learning methods such as ANN (Artificial Neural Networks). These machine warning methods have shown recently a great deal of potential.

Appendix II- Universal Canadian forecast Application (Stitou, Adnane, 2018©)

B.1 Universal Canadian forecast Application Python Source Code

```
#!/usr/bin/env python3.7
import sys
print(sys.version)

from tkinter import *
from tkinter import ttk
from tkinter import messagebox
from tkinter import filedialog
from tkcolorpicker import askcolor

import sqlite3
import pandas as pd

import numpy as np
import threading
from multiprocessing import Process

import textwrap
import xlsxwriter

import itertools
import warnings

from matplotlib.figure import Figure
from matplotlib.backends.backend_tkagg import FigureCanvasTkAgg
from matplotlib.backends.backend_tkagg import NavigationToolbar2Tk

import cartopy.crs as ccrs
import cartopy.feature as cfeature
from matplotlib.widgets import Slider
from matplotlib.widgets import Button

from matplotlib.offsetbox import AnchoredText

from matplotlib import style
style.use("seaborn")

import statsmodels.api as sm
import multiprocessing
from multiprocessing import Pool, Process, Queue, Manager
import time
import os
import warnings
warnings.filterwarnings("ignore")


def getAIC(tasks): ...


class Forecast: ...
    # 3. LEFT Graphic components
        def activateDataPeriod(self): ...

        def activateDrainageArea(self): ...

        def getCanadaMap(self): ...

    # 4.Search functions
        def printSearch(self): ...

        def getSQLStatement(self, toFind): ...

        def runSQL(self, mysql): ...

        def updateMapScatter(self, df): ...

        def setMapWidgets(self): ...

        def search(self): ...

    # 5. MIDDLE Graphic components ---> MODIFY like hoverClick.py
        def hover(self, event): ...

    # 6. RIGHT Graphic components
        def setSideBarWidgets(self, forMethod='SARIMAX'): ...

        def createNotebook(self): ...

        def getStnDfs(self, stnNbr): ...

        def updateNbPlots(self): ...

        def showHidePlots(self, msg): ...

        def showChkBxInfo(self): ...

        def toggle_Nb(self, event=None): ...

        def exportToExcel(self): ...

        def addThread(self, func): ...

        def runInParallel(self, *fns): ...

        def sarimaxForecast(self, request): ...


def main():
    root = Tk()
    app = Forecast(root)
    type(app)
    root.mainloop()


if __name__ == "__main__":
    main()
```