POLITECNICO
MILANO 1863

SCHOOL OF CIVIL, ENVIRONMENTAL AND LAND MANAGEMENT ENGINEERING
MASTER OF SCIENCE IN ENVIRONMENTAL AND LAND PLANNING ENGINEERING

A MACHINE LEARNING FRAMEWORK TO DESIGN BASIN SPECIFIC DROUGHT INDEXES

Master Thesis by:
Sami Miaari
Matr. 943122

Advisor:
Prof. Andrea Castelletti

Co-Advisors:
Dr. Elena Matta
Dr. Matteo Sangiorgio

Academic Year 2021 – 2022

Acknowledgements

I would like to address my sincere gratitude to my supervisor, Professor Andrea Castelletti, for providing me the opportunity to meet my interest of conducting a research on such critical topic. A special thanks to my co-supervisor, Elena Matta, for continuously guiding my progress and providing insights on critical decisions in addition to background support. Thanks for Matteo Sangiorgio and Davide Cananzi for the technical support in the modeling procedure who helped me improve my technical skills.

I wish to share my immense gratitude for my family who always supported my academic career, in addition to my friends that are always available in hard times.

Abstract

Droughts can have a serious impact on health, agriculture, economy, energy and environment. Monitoring and forecasting such phenomena is an essential task for mitigation, adaptation, and reduction of future risks by policymakers, mainly for areas that are characterized by high water demand with respect to the available fresh water. The Nile River Basin is a region vulnerable to climate change facing high water stress, since almost half of the basin is an arid area, where the population relies on 93% of their fresh water from the Nile River. Considering that the Nile waters are fully exploited to fulfill the high water demands downstream, climate change will exacerbate the water-related challenges, leading to an increased risk of water scarcity and food insecurity. This raises the necessity of a reliable drought detection index to support water managers and decision makers in avoiding such risks. Thus, this thesis focuses on the design of a basin-specific drought index over the Nile River Basin, which consists in an interesting case study characterized by climate and topographical heterogeneity. Also, the Nile transboundary waters are cause of conflicts among the sharing countries and still generate international discussion.

The index design is performed through the application of a machine learning FRamework for Index based Drought Analysis: FRIDA *Zaniolo et al.* (2018). The proposed framework was applied on the sub-basin scale, due to the complexity and heterogeneity of the system. Thus, the Nile River Basin is divided into 10 sub-basins according to hydrology, climate, and land cover. The framework is based on 3 main steps: information retrievement, feature extraction, and index construction. The first is performed by the definition of a target variable and collection of prediction variables: the Normalized Drought Vegetation Index is chosen as the target variable, since the economy of the basin countries is highly dependent on agriculture, highlighting the ability of the target variable in detecting agricultural and meteorological droughts. The prediction variables are chosen based on literature, identifying the high correlation with drought events, long term effect, and case study characteristics. The second step is performed by using Wrapper for Quasi-Equally Informative Sub-

set Selection as the tool for feature extraction according to the relatively better performance of wrappers than filters. The third step—index construction—performed using two alternative regression models (i.e., linear and Artificial Neural Networks).

A common thread highlighted from the output results is the consistency of the study area basin subdivision criteria, however an additional criteria of considering the spatial resolution of the input variables can be an improvement, since the sub-basins have different surface areas, topographic, and climatic characteristics. Thus, accounting for noise filtering, avoiding the effect of extremes and outliers in a portion of the sub-basin. Furthermore, the results convey high performance of the designed basin specific drought indexes in reproducing the target variable, where the index predictive accuracy in 90% of the sub-basins is outstanding; while the low performance in some cases is due to multiple factors, mainly land cover and surface area. The study contributes a further proof of the expected feature extraction efficiency in meeting the objectives of the proposed framework, reproducing a drought index for the Nile River Basin highly accurate in most of the cases. The application of FRIDA can provide an appropriate drought monitoring index that can adapt to basins with different characteristics.

Sommario

La siccitá puó avere impatti molto seri sulla salute, l'agricoltura, l'economia, l'energia e l'ambiente. Monitorare e prevedere tale fenomeno é compito essenziale della politica per mitigare, adattare e ridurre rischi futuri, soprattutto per le aree caratterizzate da un'alta domanda di acqua rispetto alla disponibilitá di acqua dolce. Il bacino fluviale del Nilo é una regione vulnerabile ai cambiamenti climatici e ad un elevato stress idrico dal momento che circa metá del bacino é un'area arida dove la popolazione dipende per il 93% dall'acqua dolce del fiume Nilo. Considerando che attualmente l'acqua del Nilo é enormemente utilizzata al di sopra della sua disponibilitá e che il cambiamento climatico aumenta il rischio dell'insicurezza idrica e alimentare degli stati circostanti, si pone la necessitá di individuare un indice per identificare l'occorrenza degli eventi di magra, che sia efficiente ed in grado di supportare i diversi decisiori ed operatori locali nella gestione delle risorse idriche. Questo studio si concentra sulla progettazione di un indice di magra specifico per il bacino del fiume Nilo, caso studio caratterizzato da un'eterogeneitá climatica che varia da tropicale ad arido. Inoltre, il bacino del Nilo é frutto di discussioni internazionali per la critica gestione (finora non coordinata) delle risorse idriche tra i paesi limitrofi.

La progettazione dell'indice viene effettuata attraverso l'applicazione del Framework for Index based Drought Analysis (FRIDA) sviluppato da *Zaniolo et al.* (2018). Il framework si basa su 3 fasi principali: recupero delle informazioni, estrazione delle caratteristiche e costruzione dell'indice. La prima consiste nella definizione della variabile target e nella raccolta dei predittori di input; la variabile target é stata scelta come Normalized Drought Vegetation Index (NDVI), poiché l'economia dei paesi del bacino é fortemente dipendente dall'agricoltura, ció evidenzia la caratteristica della variabile target per rilevare le siccitá agricole oltre a quelle meteorologiche, mentre le variabili di previsione sono state scelte in base all'elevata correlazione con gli eventi di siccitá (i.e., precipitazioni e temperatura), oltre alla portata fluviale a causa del basso flusso del fiume Nilo e all'evapotraspirazione in quanto variabile critica causata

dall'elevata evaporazione soprattutto nel Lago di Nasser, mentre l'umiditá del suolo é stata considerata per tenere conto dell'effetto a lungo termine di una siccitá. La seconda fase é stata eseguita utilizzando Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS) come strumento per l'estrazione delle caratteristiche, in base alle prestazioni relativamente migliori dei wrapper rispetto ai filtri. La terza fase di costruzione del nuovo indice é stata eseguita attraverso due diversi modelli di regressione (lineare e a reti neurali artificiali), che permettono di confrontare le prestazioni dei due modelli scelti, in modo tale che i modelli siano caratterizzati uno per la semplicitá e l'altro per l'elevata accuratezza predittiva quando vengono applicati a sistemi complessi.

Dai risultati ottenuti si puó osservare l'importanza della suddivisione dell'area di studio del bacino in base alla risoluzione spaziale dei dati di input raccolti, in modo che la procedura garantisca il filtraggio del rumore e allo stesso tempo eviti un sistema complesso di grandi dimensioni. I risultati dello studio hanno mostrato un'ottima performance degli indici di magra specifici per il bacino nel riprodurre la variabile target, dove l'accuratezza dell'indice nel 90% dei sottobacini é risultata molto alta (nella maggior parte dei casi con un coefficiente di determinazione superiore al 85%). D'altra parte, lo studio ha fornito una conferma dell'efficienza di W-QEISS nel raggiungere l'obiettivo del framework proposto, dove si é costruito un indice di magra sulla base delle caratteristiche predittive selezionate piú accurato di quello che si sarebbe ottenuto utilizzando tutti i predittori di input.

Contents

List of Figures

List of Tables

1

Introduction

Climate change and population growth are becoming the key factors for limiting sustainable human resources development and natural systems conservation. Drought is one of the main effects of climate change worldwide, it is considered as one of the most disruptive natural hazards affecting living creatures and leading to serious challenges to the environment, economy and society *Ahmadalipour et al.* (2019); *Carrao et al.* (2016); *Mishra et al.* (2017). Monitoring this phenomenon becomes an essential procedure due to its substantial effects. A drought is not only defined by the lack of precipitation, but in addition to meteorological shortage it can also affect agriculture and socio-economic sectors.

Drought on river basins is currently a worldwide issue, specially principle rivers that are considered as a hydrological vein for the population highly relying on it as their only source of fresh water, mainly due to the river passage through arid areas. The Colorado River is one of the cases, recently the Colorado River Basin encountered a major loss in stream flow induced by high temperatures in the basin caused by climate change while previous droughts were mainly due to lack of precipitation. However, with the current climate change state, higher temperature will influence on more stream flow reduction in the Colorado River *Udall and Overpeck* (2017). Such cases highlight on the issue of future droughts and the impacts following it, and thus the need of developing monitoring and prediction models.

The development of data-driven models require an accurate, fine, continuous, long time series of data. Ground stations have been used in the past for monitoring meteorological and hydrological variables, however these tech-

niques are not widely used anymore due to the limited covered areas and spatial resolution (e.g., very scarce ground station data in Africa). Recently remote sensing observations make the best source of data since this technique provides fine and long time series without any limitations on the covered area. Moreover, reanalysis data are also being produced through the combination of model data with observations by using laws of physics and therefore obtaining fine, global datasets.

1.1 Remote sensing and reanalysis data for drought detection

The monitoring of a drought is traditionally performed through ground-based stations, where measurements of different variables are recorded e.g. precipitation, temperature, wind speed, humidity. The problem with this system of monitoring is the number of instruments and their location; For example, agricultural areas include few monitoring instruments *AghaKouchak* (2015). In addition, this system suffers from the limitation that observed value conflict with the same variable in two different monitoring stations *AghaKouchak and Nakhjiri* (2012).

Indices based on information acquired from Remote Sensing (RS) are recently being developed and employed in the agricultural sector *Monteleone et al.* (2020). Information collected by satellite are acquired using a sensor which operates without any physical contact with the investigated area *Wambua* (2019). Although it can be widely useful, yet it suffers from drawbacks , e.g. data continuity, uncertainty, changes in the sensor, and short time period *AghaKouchak* (2015).

The remote sensing data acquisition are highly used for indicators intended to be used in the agricultural sector since the collected images provide the ability to visualize the changes in the vegetation cover *Zaniolo* (2020). These changes in vegetation cover are considered as structural and physiological by minimizing the water stress where long droughts can lead to a permanent change in the structure *AghaKouchak* (2015).

Reanalysis data are also considered as a good source of input data in addition to remote sensing, it combines model data with observations from across the world into a globally complete and consistent dataset based on the laws of physics. Reanalysis produces data going many years back in time, providing an accurate past climate description. The obtained dataset is characterized by temporal and spatial resolutions that allows the dataset to be useful for all kind of land surface applications such as floods or droughts

1.2 Aim of the thesis

Droughts can have very harmful effects on the environmental, social and economical aspects *Robba* (2021). Dry lands in arid and semiarid areas such as northern regions of the Nile River Basin are under risk of land degradation and desertification caused by climate change and aridification *Huang et al.* (2017); *Park et al.* (2018). Fresh water is essential in the Nile River Basin for agriculture and hydropower generation, however water demand is constantly growing with the increase of population, while at the same time the total supplied water by the river is fully utilized, mainly by the downstream countries (Egypt and Sudan). This pressure on water resources, with the lack of a proper management of the transboundary waters, raises the need to have a reliable drought monitoring and forecasting model in order to support decision makers in natural resources management.

The objective of this study is to provide a basin specific drought index considering the Normalized Drought Vegetation Index (NDVI) as the target variable, that is able to identify the vegetation health and thus possible meteorological and agricultural droughts. Moreover, the most relevant input variables for the prediction of the designed index are intended to be evaluated for every sub-basin and on the whole Nile River Basin. Furthermore, the study allows to compare the predictive efficiency between the subset of input variables and that of using all the input predictors.

Zaniolo et al. developed FRamework for Index-based Drought Analysis (FRIDA) for data-driven design of regulated basins *Zaniolo et al.* (2018). The proposed framework was chosen owing to the applied advanced feature extraction method, that automatically simplifies the variable selection, facilitating and optimizing the index evaluation due to less dimensionality and higher accuracy. Moreover, the proven good performance of the framework when applied with a vegetation index as a target variable *Cananzi* (2021)(i.e., NDVI in this study) is supporting the decision of the application of the proposed framework in this study. FRIDA is made up of 3 main steps: information retrievement, feature extraction, and index construction. The study focuses on using Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS) for the feature extraction process *Karakaya et al.* (2015); *Taormina et al.* (2016), although wrappers are computationally more expensive but given better performance than filters *Galelli et al.* (2014). Moreover, the index construction is performed by applying 2 different regression models: linear regression, Artificial Neural Networks. The models were chosen the first due to simplicity, while the second based on the high predictive accuracy for complex systems disregarding

the time consumption.

1.3 Outline of the thesis

The thesis starts with an introduction about drought and indexes, followed by Chapter 2 that provides a literature review about the most widely used drought indexes, defining their properties and calculation methodology. Moreover, the proposed FRIDA framework is demonstrated in Chapter 3 through the brief description of the Input Variable Selection (IVS) methodology and the adopted W-QEISS. Chapter 4 defines the case study of the River Nile Basin and properties of the basin (i.e., hydrology, climate change, droughts, and transboundary conflicts) followed by a section for the characteristics of the collected data. Chapter 4 contains the illustration of the variables and the reason for the target variable selection in addition to results of the feature extraction process with the different regression models performance on the whole Nile River Basin and on the sub-basin scale which defines the motivation for the decision of the best performing model. In addition to a performance comparison of the newly constructed index using the selected input predictors and all the input variables.

2

State of the art

A precise definition of a drought is hard to be identified due to its complex nature and multiple impacts, confusing the definition of the main drought characteristics such as duration, intensity, severity, and the spatial extent. Another challenge that rises with droughts is the start and end recognition, all along with the impacts referred to it that are nonstructural and could be extended to large geographical areas. Moreover, human activities can contribute to droughts by multiple factors such as over farming, deforestation, or overexploiting the available water. The mentioned characteristics of a drought highlight on the difference from other natural hazards *Wilhite* (2000).

Droughts are usually defined by the lack of water availability in any of its supplied forms such as precipitation, runoff, or groundwater *Beran and Rodier* (1985). Droughts can be classified into four categories differentiated by hydrological, meteorological, agricultural, socio-economical *Dracup et al.* (1980); *Heim Jr* (2002). The meteorological drought that is defined as the lack of precipitation over a long period of time will result to a decrease in the streamflow, groundwater, or reservoir level, as known by the hydrological drought which exist for a certain period after the end of a meteorological drought. The socioeconomic is related to the supply and demand of economic products *Wilhite and Glantz* (1985), while the agricultural drought is recognized by a deficiency in water availability needed for crop production *Heim Jr* (2002).

2.1 Drought indicators

Throughout the aim of describing the physical characteristics of a drought, many indicators have been proposed to define the duration, severity, and spatial extent; these indicators are usually based on multiple parameters e.g. precipitation, temperature, streamflow *Steinemann et al.* (2005); *Hayes et al.* (2012). Drought indices are numerical representations of the severity or magnitude of a drought *Hao and Singh* (2015), they fully define multiple variables into one single value to support decision making *Hayes* (2002).

Many studies have been made to propose drought indices that provided the opportunity to identify indices for the different typologies of droughts *Sol'áková et al.* (2014). The first widely used index was the Palmer Drought Severity Index (PDSI) *Palmer* (1965) which was considered as a landmark in the development of drought indices *Vicente-Serrano et al.* (2010). Later, McKee et al. introduced the Standardized Precipitation Index (SPI) *McKee et al.* (1993) nowadays on the most widely used drought indicator for evaluating meteorological droughts, as the World Meteorological Organization (WMO) has recommended the application of this indicator *Sol'áková et al.* (2014); *Hayes et al.* (2011). Many other indicators were developed such as the Standardized Streamflow Index (SSI) *Nalbantis* (2008) and Standardized Runoff Index (SRI) *Shukla and Wood* (2008).

Droughts are affected by several variables for instance climatic factors described by high temperature and low relative humidity *Wilhite* (2005). As a result, a single drought indicator is not able to identify the drought condition and the impacts behind it. For example, a meteorological drought indicator which is defined as a deficit in precipitation, may not detect an agricultural drought that is mainly identified by the soil moisture. Therefore, the recognition of a drought from a multivariate perspective is needed to characterize all the drought variables, which can be performed by the combination or joint of some hydrological variables and drought indices *Hao and Singh* (2015). In order to perform a complete analysis of a drought, then the joint index should include indices related to the rainfall, water deficit, and soil moisture to explain the drought conditions for meteorological, hydrological, and agricultural droughts, respectively *Keyantash and Dracup* (2002). Numerous joint drought indices have been proposed such as the aggregate joint index that integrates the drought variables related to each of the its physical forms (i.e. meteorological, hydrological, and agricultural) *Keyantash and Dracup* (2004), another joint index that was developed by the implementation of a copula with the Kendal distribution to consider droughts from precipitation and streamflow *Kao and Govindaraju* (2010), a similar methodology was applied by Hao and AghaK-

ouchak who developed a copula-based joint index that was identified as the Multivariate Standardized Drought Index (MSDI) which considered meteorological and agricultural drought conditions by the probabilistic combination of the SPI and the Standardized Soil Index (SSI) *Hao and AghaKouchak* (2013). One of the most common combined drought indices is the Standardized Precipitation Evapotranspiration Index (SPEI) that is based on the precipitation expressed by the SPI and temperature that recognizes multi-scalar characters in addition to the ability to include the effects of temperature in the assessment of a drought *Vicente-Serrano et al.* (2010).

Some drought indicators are developed based on data from remote sensing, such as Normalized Drought Vegetation Index (NDVI) which is based on the soil cover *Xing et al.* (2020). The condition of the vegetation outlines how much light will be absorbed and reflected. Therefore, the health of the vegetation can be evaluated from the reflected wavelengths since satellite remote sensors can detect the amount of photo synthetically active radiation the vegetation can absorb.

2.1.1 Standardized Precipitation Index (SPI)

In the last century, the droughts were evaluated based on the Palmer Drought Severity Index (PDSI). However, this index suffered from some limitations so it was not considered as the best index to be used for the monitoring and supporting drought management decisions. Therefore, by the end of the 20th century the Standardized Precipitation Index (SPI) was developed by McKee et al, which provided better results than the Palmer indices regarding the abnormal wetness and dryness of lands *Guttman* (1999). The SPI is a probability based index having the ability to be applied for different timescales, this is important depending on the uses, for example agricultural interests require a short term analysis while water supply management requires the analysis of droughts over a long period of time spanning years.

The calculation procedure can be performed by a parametric or non-parametric approaches; however, the former is preferred over the latter because of the extrapolation problems which the latter suffers of. A study performed by Solakova et al. showed that higher differences can be identified between both approaches in terms of drought severity but less differences for durations and interarrival times that is more on drought entity, but less on drought identification *Sol'áková et al.* (2014).

The parametric approach is based on the collection of a precipitation dataset of 40-60 years as recommended by Guttman *Guttman* (1999), averaging the peri-

ods according to different timescales 3, 6, 9, 12, 24 and 48 months. Each dataset is fitted to a gamma function to define the relationship of probability to precipitation from historic records as recommended by most studies without testing the distribution function *Guenang and Kamga* (2014), thus the probability of any observed precipitation data can be obtained and used with an inverse normal (Gaussian) function to calculate the precipitation deviation for a normally distributed probability density with a mean zero and standard deviation unity. The final SPI value ranges from -2 to 2, where the boundaries of this interval can be assigned as the thresholds for extreme events.

One of the limitations of the SPI is the chosen distribution function. The index offers results that can be compared, but using different distributions provides different outcomes so for the same observed precipitation time series the results comparison might be confusing or misleading *Guttman* (1999). Furthermore, some users might have a weak background about the calculation procedure or not enough time to define the most suitable distribution. However, it was found by Guenang and Kamga that the distribution function depends on the location of the stations and length of the data time interval *Guenang and Kamga* (2014). Therefore, it is recommended to test multiple probability distributions for short length data records *Mishra and Singh* (2010), while for long time periods it is recommended to have a data record length of 40-60 years to achieve the stability of the parameters estimation in the central part of the distribution and 70-80 years for the stability in the tails of the precipitation distribution *Guttman* (1999). Moreover, the index is based only on precipitation data which make it unable to explain the effects on drought influenced by temperature changes *Guenang and Kamga* (2014); *Hao and AghaKouchak* (2013). Based on a study done by Pathak and Dodamani on the Ghataprabha river basin in India provided results that recommends the usage of SPI only in humid regions *Pathak and Dodamani* (2020). This suggestion is also provided by the analysis of a study made on 41 stations in Iran for data of more than 30 years *Jamshidi et al.* (2011). This could be considered as a drawback to this indicator as it limits the regions where it can be applied.

The SPI is highly recommended and widely used (e.g., *Vicente-Serrano et al.* (2010); *López-Moreno and Vicente-Serrano* (2008); *Mo and Schemm* (2008); *Bordi et al.* (2009); *Bothe et al.* (2011); *Santos et al.* (2010); *Zhu et al.* (2011); *Sienz et al.* (2012)) since it has the characteristics of a simple, spatially invariant, and probabilistic and can be used for various time periods according to the preferability of the user *Guttman* (1998, 1999). Also, it is able to capture historical drought conditions *Hao and AghaKouchak* (2013).

2.1.2 Multivariate Standardized Drought Index (MSDI)

The detection and monitoring of a drought is very hard to be perfectly spotted by one single drought indicator; thus, the development of a multi-variate index becomes essential to join the properties of more than one indicator. The perfect index should consider all the categories of a drought: meteorological, agricultural, hydrological, and socio-economical *Heim Jr* (2002). The MSDI was proposed to characterize drought conditions at different timescales. This index is based on a joint distribution function of two indicators, SPI and SSI which considers the precipitation data and soil moisture data, respectively *Hao and AghaKouchak* (2013). These two indices were considered for the MSDI development as the precipitation provides the ability to early detect a drought while the soil moisture provides a delayed detection. Moreover, soil moisture takes several more months to show the end of an extreme event *Hao and AghaKouchak* (2013). Therefore, by considering the precipitation and soil moisture data, this index has the ability to characterize a meteorological and agricultural drought.

The calculation methodology of this index requires the use of a copula to determine the joint distribution of the two variables, disregarding their marginal distribution. It starts by assuming the precipitation and soil moisture as random variables X and Y, respectively. In such way, the copula C allows to express the joint distribution and the cumulative joint probability p *Nelsen* (2007); *Sklar* (1959):

$$P(X \leq x, Y \leq y) = C[F(X), G(Y)] = p \tag{2.1}$$

Where C is the copula, F(X) and G(Y) are the marginal cumulative distribution functions of the random variables X and Y, respectively. Multiple copulas have been developed, in the MSDI the frank copula has been applied which has a symmetric dependence structure, as expressed in the following equation *Nelsen* (2007); *Salvadori et al.* (2007).

$$C(u, v) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right] \tag{2.2}$$

Where u and v are the marginal cumulative probabilities of precipitation and soil moisture, respectively. The parameter θ can be estimated from the Kendall,Äôs rank correlation τ.

Finally, the MSDI can be determined as the inverse of φ, the standard normal distribution function *Hao and AghaKouchak* (2013).

$$MSDI = \varphi^{-1}(p) \tag{2.3}$$

This equation can transform the joint probability p to the MSDI which allows cross-comparison of different drought indices *Hao and AghaKouchak* (2013).

During the calculation procedure of the index, copula functions are required to derive the joint probability distribution of precipitation and soil moisture. However, these copula functions give different distribution results that should be tested for their goodness-of-fit by applying the Cramér-von Mises or the Kolmogorov-Smirnov statistic to evaluate the performance of copulas in modeling the relation between precipitation and soil moisture *Genest et al.* (2006); *Genest and Favre* (2007). The study made by Hao and AghaKouchak for the development of the MSDI included the test of three different copulas, i.e., Clayton, Frank, and Gumbel. The results from the performed goodness-of-fit tests showed that the Frank and Gumbel copulas can model most of the months, while Clayton rarely fits *Hao and AghaKouchak* (2013). Since the MSDI is developed as an extension to the SPI proposed by McKee et al. *McKee et al.* (1993), so it has a very similar characteristics to the SPI such as the ability to monitor droughts with different time scales (e.g. 1-,3-,6-month). The performed study on the MSDI resulted that it can demonstrate the drought start and drought persistence, which are similar properties of SPI and SSI, respectively *Hao and AghaKouchak* (2013). Therefore, the MSDI has the advantage of detecting the drought as early as the SPI, along with describing the drought process relying on the state of two variables: precipitation and soil moisture. Consequently, it reaches the main goal behind the implementation of a multi-variate index, avoiding the affection of a temporary change in one variable on the index. It is widely known the ability of the SPI to identify drought severity, the newly developed MSDI had shown the same property. This index has the ability to identify the drought severity with a lower intensity than that identified by SPI affected by the drawback of SSI in the recognition of drought severity. However, in the case of a severe drought recognized by both SPI and SSI, then the MSDI resembles a more severe drought than both indicators. Furthermore, MSDI is able to detect drought before SPI and SSI, thus improving the earlier detection better than each single index *Hao and AghaKouchak* (2013). Finally, the newly developed MSDI can be applied for risk analysis since it allows to obtain a probability of occurrence similar to that of SPI and SSI.

The usage of multivariate parametric copula functions to obtain the joint distribution requires parameter estimation and testing the goodness-of-fit *Hao and AghaKouchak* (2014). Therefore, the choice of the copula function should be verified, in some cases this limitation is avoided by using empirical joint probability formulas *Hao et al.* (2014).

2.1.3 Standardized Precipitation Evapotranspiration Index (SPEI)

Many empirical studies have been performed to evaluate the effect of temperature on drought severity. For instance, a general circulation model was used by Abramopoulos et al. to show that evaporation and transpiration consume up to 80% of rainfall *Abramopoulos et al.* (1988); moreover, they found that temperature anomalies cause drying of about the same efficiency of that of rainfall shortage. Furthermore, the greatest damage to cultivation and natural system, with the increase of evapotranspiration and water stress in the devastating central European drought in the summer of 2003 were mainly caused by the extremely high temperature over Europe in June and July *Rebetez et al.* (2006).

Temperature is expected to increase during the twenty-first century due to global warming *Solomon et al.* (2007) which will effect on the water demand as a result of the increased evapotranspiration *Sheffield and Wood* (2008). It has been proved that indicators based only on meteorological conditions are not able to clarify the droughts caused by increase of temperature *Dubrovsky et al.* (2009). Therefore, the use of drought indices which consider temperature is mandatory *Vicente-Serrano et al.* (2010). The SPEI proposed by Vicente-Serrano et al. has a role to combine the sensitivity to changes in evaporation demand, that is caused by temperature fluctuations, with the multi-temporal SPI. This index was meant to detect, monitor, and explore the effects of global warming on the droughts *Vicente-Serrano et al.* (2010).

The SPEI calculation methodology relies on the same procedure for SPI using monthly or weekly precipitation data. The methodology is based on the monthly difference between the precipitation and the potential evapotranspiration (PET). The PET is estimated from the Penman-Monteith method, but it requires a lot of data for solar radiation, temperature, wind speed, and relative humidity. Many other empirical equations were developed to calculate the PET in the case of scarce data *Allen et al.* (1998). A simple method can be used for the estimation of the PET as recommended by Vicente-Serrano, which is provided by Thornthwaite in 1948 *Thornthwaite* (1948). However, the method used for the calculation of the PET is not very critical, since the evaluation of this variable is performed for the reason to attain a relative temporal estimation *Vicente-Serrano et al.* (2010), this hypothesis is proved by Mavromatis, who demonstrated that the usage of a simple or a complex method for the PET calculation provides similar results on the evaluated drought index *Mavromatis* (2007). Finally, the deficit D_i can be obtained by the difference between the precipitation P_i and the

PET_i for each month i *Vicente-Serrano et al.* (2010).

$$D_i = P_i - PET_i \tag{2.4}$$

The D_i values are aggregated at different time scales as $D^k_{(i,j)}$ where i stands for the month, j for the year, and k for the time scale. Three parameter distribution is used to calculate the SPEI as x can have a value in the range $\gamma < x > \infty$, where γ is the parameter of origin of the distribution so x can have negative values. In the study done by Vicente-Serrano, the L-moment ratio diagrams were used.

In order to standardize the D series, the L-moment ratios are adjusted by different candidate distributions (Pearson III, log-normal, general extreme value, log-logistic). The Kolmogorov-Smirnoff test was applied to choose the best candidate distribution. As for the study done by Vicente-Serrano, the log-logistic three parameter distribution was used that has the parameters of scale, shape and origin which are calculated from the L-moment procedure that is a robust and easy approach *Ahmad et al.* (1988). The cumulative probability distribution function F(x) from the log-logistic is used to obtain the value of the SPEI index as the standardized values of F(x) by using the relation proposed by Abramowitz and Stegun *Abramowitz et al.* (1988).

$$SPEI = W - \frac{C_0 + C_1 W + C_2 W^2}{1 + d_1 W + d_2 W^2 + d_3 W^3} \tag{2.5}$$

Where $C_0, C_1, C_2, d_1, d_2, d_3$ are constants and $w = \sqrt{-2\ln(P)}$ for $P \leq 0.5$

Such that P is the probability of exceeding an obtained D value, $P = 1 - F(x)$. While $P = F(x)$ for $P > 0.5$ and the sign of SPEI is reversed.

The value of the index is standardized so it can be compared with other SPEI values over time and space.

The evapotranspiration parameter has a major role in the explanation of drought variability when considered by drought indices that are based on soil water balances, where this parameter becomes comparable to the importance of precipitation under some circumstances *Hu and Willson* (2000). This concept is verified by Narasimhan and Srinivasan who concluded that the usage of only evapotranspiration data in the calculation of a drought index whose role is to monitor an agricultural drought, has shown better results than the indices based on precipitation *Narasimhan and Srinivasan* (2005). Therefore, the consideration of evapotranspiration in the calculation procedure of drought indices is highly recommended even though there exists a complexity in the determination of its influence on drought conditions *Vicente-Serrano et al.* (2010).

The use of an index, that takes temperature into account, is becoming essential as many studies have shown an increase in future drought severity due to an increase in temperature *Beniston et al.* (2007); *Sheffield and Wood* (2008). As the severity of the droughts will be proportionate to the variability in temperature, this phenomenon will be well weighed by the SPEI. In addition, the SPEI has an advantage over other indices which include the temperature in the calculation procedure due to its simplicity, less data requirement, and the multi-scalar character *Vicente-Serrano et al.* (2010). As a result, the use of SPEI is preferred for the identification, analysis, and monitoring of droughts in all regions of the world *Vicente-Serrano et al.* (2010).

In addition to the characteristics of SPEI which give it the preferability over other indicators, it also has the same ability as other indices for the measurement of drought severity based on its duration and intensity, in addition to the identification of the onset and end of drought episodes *Vicente-Serrano et al.* (2010). Therefore, it allows for time and space comparison of drought severity. Most widely used indices rely on the consideration of PET effect on the severity of a drought, the SPEI is able to identify different types of droughts and their effects in the context of global warming which give it the main advantage over the other indices. For more detailed explanation of the procedure, the reader can refer to the research done by Vicente-Serrano *Vicente-Serrano et al.* (2010).

2.1.4 Normalized Difference Vegetation Index (NDVI)

NDVI is identified as a remote sensing indicator of the vegetation cover status. The assessment of the vegetation cover is quantified according to the electromagnetic spectrum *AghaKouchak et al.* (2015). The red and infrared bands can identify the absorption of a green vegetation; Therefore, NDVI is made up based on the red and infrared bands *Sruthi and Aslam* (2015).

$$NDVI = \frac{\lambda_{NIR} - \lambda_{VISIBLE}}{\lambda_{NIR} + \lambda_{VISIBLE}} \tag{2.6}$$

This indicator has values of results ranging from -1 to +1 according to the presence of vegetation *Pettorelli et al.* (2005). Therefore, it can be used for defining the vegetation condition.

The computation and usage of this index depends on the availability of remote sensing observations for the required time period. However, the non-linearity and a delayed response of the vegetation to a deficit in rainfall can be considered as a drawback. Moreover, this indicator implies the hypothesis that the vegetation cover has not changed from the past and will not change in the

future, which is a strong assumption that could lead to non-reliable results in the future.

2.2 Comparison among different drought indicators

It can be recognized that the decision on indicators to be used for the detection and monitoring of droughts is dependent on the scope and reason behind the application process. Thus their classification is based on the relevant drought category. Moreover, these indicators are evaluated based on data that highlights their relation to one of the mentioned categories. For example, the SPI is determined based on precipitation data, which implies that a drought evaluated by this index is a meteorological drought as it is based on meteorological data *McKee et al.* (1993). However, some indices are determined by joining two indicators or using the data of two categories, such as the SPEI *Vicente-Serrano et al.* (2010). This index is determined by using the SPI (i.e. precipitation data) and temperature. As a result, the SPEI is considered as an index that can be used for monitoring meteorological and agricultural droughts *Vicente-Serrano et al.* (2010). The SPI is known for its capability in identifying the intensity of drought severity from a reduction in precipitation; However, the evapotranspiration can improve the detection of drought severity based on information from temperature *Vicente-Serrano et al.* (2010) while this parameter is not considered in the SPI since it accounts for only precipitation data in the calculation procedure. Therefore, the use of SPEI over SPI is highly recommended due to its ability to account for temperature variability and extremes in the context of global warming. Another interesting newly developed index is the MSDI, that is based on the joint copula of SPI and SSI. Hence it can be noticed that the MSDI can detect a meteorological and an agricultural drought *Hao and AghaKouchak* (2013). However, the NDVI is widely used due to it is reliability on remote sensing data and high accuracy in detecting vegetation cover. This drought indicator has the advantage of simplicity and ease of computation from spectral bands, as it does not require the application of a distribution function but it is a ratio-based index. In addition, it provides global coverage, since it is based on the land cover and data from remote sensing, which can be acquired for any point on the surface of Earth. Nevertheless, it also suffers from some limitations as the lagged response and non-linearity.

The comparison between the indicators is mainly based on the type of input data (e.g., precipitation, temperature) and evaluation methodology, as it is preferred to follow a direct and simple calculation procedure rather than a complex one, mainly depending on the different types of the indexes as shown in

Table 2.1: *Characteristics of the most common drought indexes (X indicates the presence of the characteristic outlined)*

Characteristic	SPI	MSDI	SPEI	NDVI
Timescale aggregation	X	X	X	
Meteorological drought	X	X	X	X
Agricultural drought		X	X	X
Type	2-parameter distribution	Joint copula	3-parameter distribution	Ratio-based
Restricted areas	X			
Deferred response				X

Table 2.1. The SPI is one of the most widely used drought indicators, since it requires only precipitation data and has a simple application methodology which is based on a two-parameter distribution. However, this advantage comes with the cost of finding the best suitable distribution function, since many functions give results but not all provide the best output *Guttman* (1999).

Even though the SPI is widely used, it has the drawback of a limited regions preventing this indicator from providing reliable results compared to the other indices. Some studies were done and concluded that the SPI is recommended to be used only in humid areas *Pathak and Dodamani* (2020); *Jamshidi et al.* (2011).

Furthermore, based on previous studies SPI and SPEI showed similar results *Omar* (2020). Pathak and dodamani in 2020 compared the SPI, SDI, and SPEI where their results showed that no harmonization exists between SPI and SPEI at any timescale of the period under study *Omar* (2020); *Pathak and Dodamani* (2020).

The SPEI has a similar methodology to the SPI, but uses a three-parameter distribution which gives it an advantage, as it allows to avoid the minor problems faced in the results of SPI, which uses two-parameter gamma distribution for short time scales in low precipitation values *Vicente-Serrano et al.* (2010); *Wu et al.* (2007). However, it requires more data as it is also evaluated using temperature data in addition to the precipitation. The inclusion of temperature makes the index more complex, as it requires the calculation of PET, evaluated either by the Penman-Monteith method that requires many data, or using complex empirical methods.

By comparing these two indices with the MSDI, it can be found that the later has the most complicated methodology, compared to the previous two. This is because MSDI was based on the joint function of two indicators using the copula function. Furthermore, it suffers from the limitation of choice of the best copula function by applying a goodness-of-fit test *Hao and AghaKouchak* (2013).

NDVI can be identified with the least complex methodology between the

mentioned indexes since it is a ration based index as can be noticed from Table 2.1. Although it suffers from a deferred response, yet it still has the properties to be used as a meteorological and agricultural drought index. Therefore, providing NDVI an advantage over the other considered indicators, in addition to the characteristic of this variable as a remote sensing index, having the advantage of collecting information over all the study area without restrictions on resolution.

3

Methods and Tools

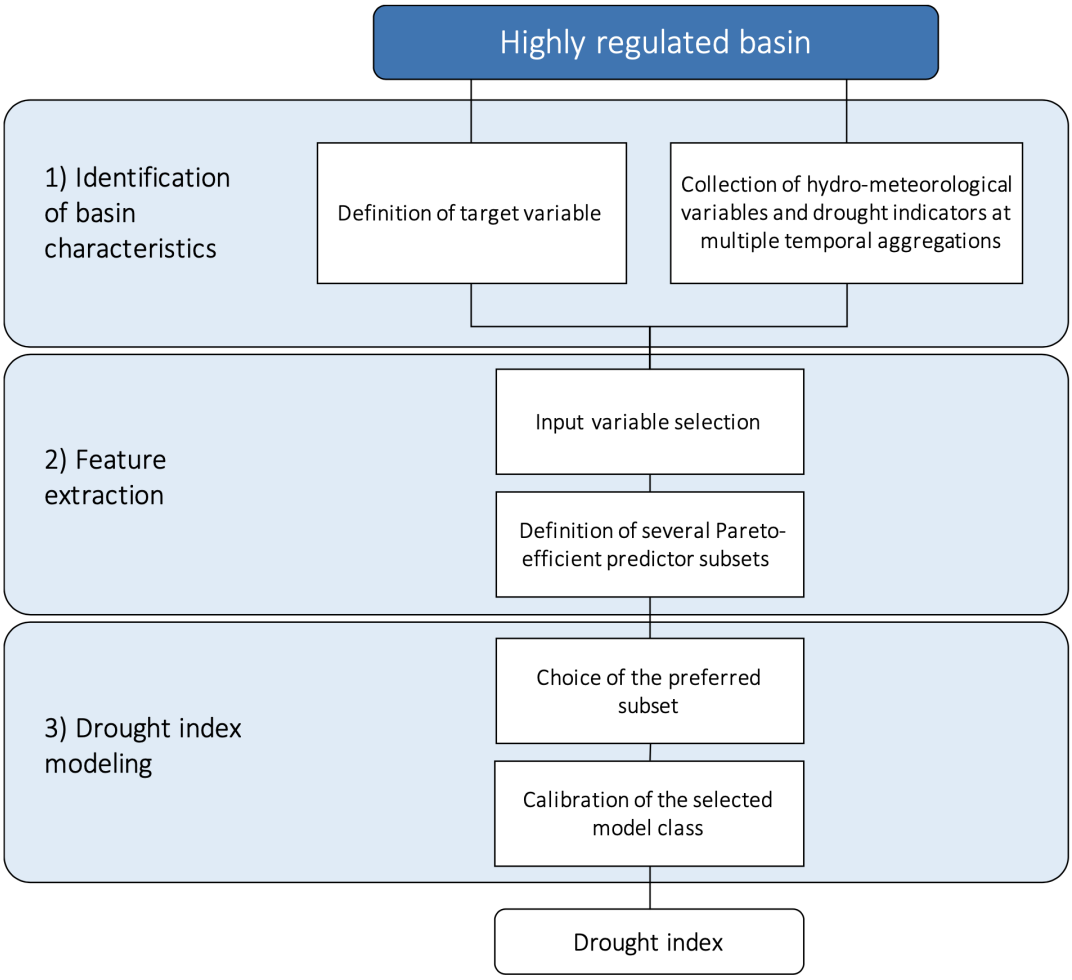
Ad-hoc index formulations are usually implemented according to empirical combinations made up of several hydro-meteorological variables. Although these customized formulations are considered effective in the design basin, yet they lack the ability to be generalized and applied in different contexts.

Zaniolo et al. (2018) developed a novel framework for the automatic design of basin-customized drought index known as "FRamework for Index based Drought Analysis" (FRIDA). When compared to ad-hoc empirical approaches, FRIDA is fully-automated, generalizable, and can be applied across different basins.

3.1 FRamework for Index based Drought Analysis (FRIDA)

FRamework for Index based Drought Analysis is developed by Zaniolo for data-driven design of regulated basins *Zaniolo et al.* (2018). The result is an index that represents all the effects of a drought in the system by considering all the relevant information to determine it. The special property of this framework is using an advanced feature extraction method, that simplifies the variable selection automatically. This property facilitates and optimizes the process of evaluating the index, due to the less dimensionality and higher accuracy. The framework is made up of three steps as shown in Figure 3.1. The first step is the identification of basin characteristics that is an initial empirical process for the selection of a target variable and candidate predictors. The target variable is defined according to the impacts of a drought on the basin such as soil mois-

Figure 3.1: *FRIDA framework proposed by Zaniolo et al. (2018)*



ture deficit or water supply deficit, the predictors are candidate features that allow to mimic the target variable, usually consisting of hydro-meteorological observations and drought indicators over multiple spatio-temporal scales.

The second step is the feature extraction which is the core of the framework. It is based on the selection of the most relevant variable subsets that best describe the target variable by applying a feature extraction algorithm. A genetic algorithm can iteratively explore and find Pareto optimum subsets from the input space of the candidate predictors. The objectives of the Pareto results are (1) accuracy, (2) cardinality of the subset, (3) relevance for defining highly informative subsets and, (4) redundancy to ensure low intra-subset similarity. The accuracy is evaluated by the calibration of a predefined model (e.g., Extreme Learning Machine (ELM)) until reaching the stopping criteria *Huang et al.* (2006); *Cananzi* (2021). While the relevance and redundancy improve the search for a more diversified and comprehensive set of solutions. *Zaniolo* (2020)

applied an algorithm known as Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS) to obtain efficient subsets in a selection matrix *Karakaya et al.* (2015).

The third and last step of the framework is the drought index modeling where the preferred subset is selected by the user considering additional objectives neglected in the input variable selection search. The last step implies the fit of a regressor to the data of the Pareto results and the target variable. The choice of the model is depending on the application, usually a highly non-linear machine learning such as Artificial Neural Network (ANN) could be a good option considering accuracy and flexibility. However such tools have a weak strength considering several stakeholders *Estrela and Vargas* (2012). Therefore, a simpler model (e.g., a linear model) could be of better benefit as it provides an immediate physical understanding, with less approximation skills.

3.1.1 Data-driven models

Data-driven models are the most widely used type of models for exploring and defining the complex relation between several variables in an environmental system due to the complex and dynamic relation between these variables *Limburg et al.* (2002); *Cananzi* (2021). This type of models is empirical and thus based on mathematical equations not defined from physical processes but from the analysis of a time series of data *Solomatine et al.* (2009). Artificial Neural Networks (ANNs) represent the most widely used family of data-driven models *Maier and Dandy* (2000), they are powerful models that can partially understand environmental systems.

The selection of the proper model inputs is vital as it effects on the performance *Solomatine et al.* (2009). However, for ANN models the effect of the inputs becomes less important resulting in less attention given to the input selection process *Maier and Dandy* (2000).

Although this approach is possible, but it has some drawbacks such as the curse of dimensionality, irrelevant inputs negatively affect on the learning process, and local minima can influence on the accuracy of the results *Cananzi* (2021). It can be concluded that implementing an approach for appropriate input selection provides a higher level of optimality for the model *Bowden et al.* (2005), such approaches are called input variable selection techniques.

3.1.2 Input Variable Selection

The selection of appropriate relevant inputs and features becomes necessary for improving the performance of the model, since a large dataset of inputs can

lead to overfitting due to irrelevant features *Karakaya et al.* (2015). The technique is generally a preprocessing algorithm to obtain a more compact and relevant subset for the calibration of a model *Bowden et al.* (2005).

Previous modeling of hydrological systems by ANNs included a proper input selection step, however it was performed by some traditional techniques such as trial and error or cross-correlation *Galelli et al.* (2014). Nevertheless, these traditional techniques suffer from limitations and drawbacks which requires the need to apply robust IVS methods that are able to capture interdependencies, redundancies, and non-linearities of the hydrological system *Snieder et al.* (2020).

The IVS algorithm can be used for two purposes: classification and regression. In the classification, the algorithm is meant to identify the relevant and non-redundant variables to provide a categorical output of the model. In regression, the algorithm is meant to combine the selected features to generate the output which is expected to be numerical and able to reproduce the chosen target variable *Karakaya et al.* (2015); *Taormina et al.* (2016).

It is important that the IVS algorithm is able to identify relevant and non-redundant input variables to have a reliable, efficient, and fast model. Irrelevant input variables increase the complexity and are uninformative; on the other side, redundant and relevant input variables can increase the dimensionality of the model without any additional information. Therefore, the exclusion of some relevant input variables can decrease the dimensionality on the expense of decreasing the accuracy of the model. Hence, the result of the IVS process should select the lowest number of input variables that best describe the system with the minimum redundancy *Guyon and Elisseeff* (2003).

3.2 Feature extraction by Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS)

The feature extraction process is important for obtaining a compact and informative subset *Cunningham* (2008). The IVS technique is one of the feature extraction algorithms, and it is applied for the identification of relative predictors for the model calibration *Bowden et al.* (2005). Two main classes of IVS are the most used, Filters and Wrappers. Filters methodology is based on evaluating the relevance of each variable separately, computing an error metric on the features *Yang and Pedersen* (1997); *Sharma* (2000); *Galelli and Castelletti* (2013). However, wrappers evaluate the relevance of the ensemble of a variable, by assessing the prediction performance of a learning machine calibrated on the input

dataset, and therefore considering the interactions and dependencies between variables *Guyon and Elisseeff* (2003). Although wrappers are computationally more expensive but give better performance than filters *Galelli et al.* (2014).

One of the wrapper algorithms is the W-QEISS *Karakaya et al.* (2015); *Taormina et al.* (2016). Typically wrappers can solve a two-objectives optimization problem and give the results as Pareto efficient solutions. The two objectives are the minimization of the model complexity and the maximization of its accuracy. According to *Karakaya et al.* (2015), two subsets can be quasi equally informative if they have almost the same classification performance with respect to a given learning algorithm.

3.2.1 Methodology

Let $\overline{f}(S_i)$ be a predictive accuracy from 0 to 1, where 0 represents no predictive skills and 1 represents perfect predictive accuracy. For S_j as a subset of predictors and another subset S_i. If the two subsets have quasi equal predictive accuracies with respect to a given model class so they are δ-quasi equally informative.

$$\overline{f}(S_i) \geq (1-\delta)\overline{f}(S_i) \; for \; 0 \leq \delta \leq 1 \tag{3.1}$$

Some of the advantages of W-QEISS that come with its feature of exploring equally informative subsets are:

- Capable of determining the importance of each predictor by studying the frequency in the most informative subsets.
- Ability to better understand synergistic relationships between variables.
- It gives the possibility to identify variables providing the same information to avoid redundancy.
- It Allows the option to decide between different combinations of predictors with different characteristics depending on the predictive accuracy and reliability.
- For missing data, the predictor can be replaced with an alternative one.

3.2.2 Objective functions

The innovative feature of W-QEISS is based on a four-objective optimization problem consisting of accuracy, complexity, relevance, and redundancy. The accuracy guarantees an exact reproduction of the data while complexity is important for simplifying the model. Relevance and redundancy are an asset for

a better subset search to ensure high informative content with minimum similarity between subsets.

Symmetric uncertainty (SU), which measures the dependence and similarity between variables, is used by three out of the four objective functions. SU is a function of Shannon entropy *Shannon* (1948) that measures the uncertainty of a random variable defined by:

$$H(\cdot) = -\sum p(\cdot) \log p(\cdot) \tag{3.2}$$

SU ranges from 0 to 1, such as 0 means independent variables and 1 means complete dependence between the variables. The SU is computed for two features A and B as below:

$$SU(A,B) = \left[\frac{2 \cdot (H(A) + H(B) - H(A,B))}{H(A) + H(B)}\right] \tag{3.3}$$

The objective functions are defined by X as the pool of candidate predictors, y as the output variable and S as a subset of X:

1. For a metric of relevance defined as the first objective function as f_1 (S) to be maximized is defined as:

$$f_1(S) = \sum_{x_i \in S} SU(x_i, y) \tag{3.4}$$

   Where $SU(x_i, y)$ represents the symmetric uncertainty between x_i (feature) and y (output). Therefore, the relevance measures the explanatory power with respect to the output.

2. Redundancy is defined as a metric $f_2(S)$ to be minimized.

$$f_2(S) = \sum_{x_i, x_j \in S, i<j} SU(x_i, x_j) \tag{3.5}$$

   The symmetric uncertainty in this function is used between two predictors x_i and x_j. It is evaluated by minimizing the redundancy so that output features are mutually dissimilar.

3. $f_3(S)$ defined as the cardinality that is the number of selected predictors which is to be minimized to ensure the minimum complexity

$$f_3(S) = |S| \tag{3.6}$$

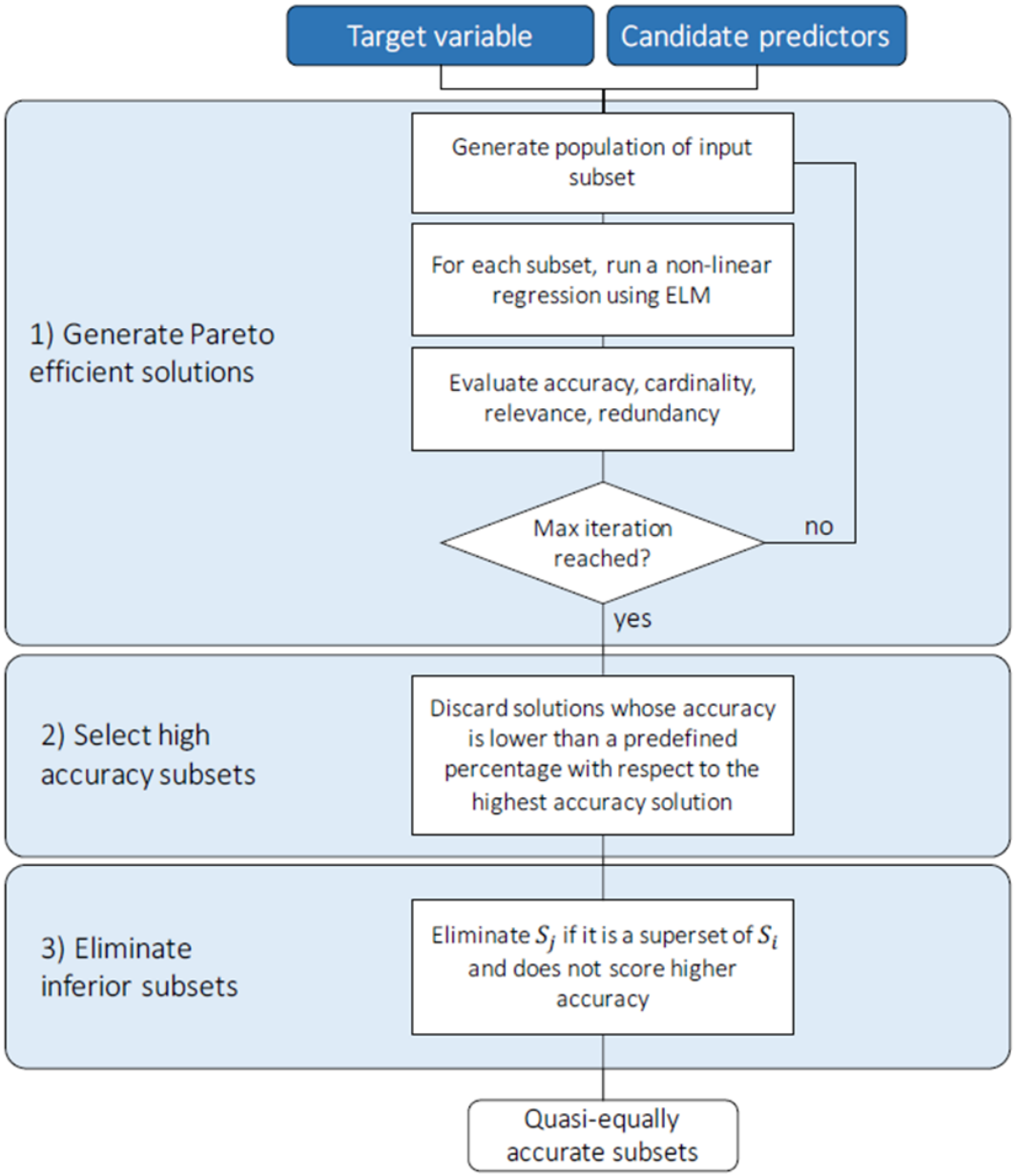
4. The predictive accuracy is identified as $f_4(S)$ and should be maximized. This objective function is also evaluated by the SU between the observed

and predicted outputs, y and y(S) respectively.

$$f_4(S) = SU\left(y, \hat{y}(S)\right) \tag{3.7}$$

The algorithm for W-QEISS is run following three steps as identified in Figure 3.2 *Zaniolo et al.* (2018):

Figure 3.2: *Wrapper for Quasi-Equally Informative Subset Selection (W-QEISS) Zaniolo et al. (2018)*



1. Alternate subsets of predictors are found by applying a global multi-objective optimization algorithm. Results in a set A of Pareto-efficient solutions

based on the objective functions.

2. Identify the best accuracy value of the subsets in A based on the predictive accuracy. In addition to the δ-quasi equally informative subsets based on the same objective function and a pre-selected value of δ in order to eliminate from A subsets, whose accuracy is lower than $(1-\delta)f_4^*$.

3. The removal of subsets in A_δ that do not improve the accuracy of the model based on the comparison of the subsets. For example, the subset S_j is considered inferior to S_i, if it contains the same features as S_i and does not improve the accuracy of the model. Therefore, the result would be the δ-quasi equally informative subset A_δ^*.

3.2.3 Implementation

W-QEISS should be designed according to the following properties:

- Find a large number of equally informative subsets of predictors.
- Flexible modeling, dealing with linear and non-linear relationships.
- Computationally efficient.

These features can be achieved by the combination of the Borg MOEA for an efficient global search, and the Extreme Learning Machine for a flexible and efficient model architecture *Cananzi* (2021).

3.2.4 Borg MOEA

Borg Multi-Objective Evolutionary Algorithm ensures that the selection, crossover and mutation operators are selected based on their capability to generate efficient solutions. Borg MOEA is known for featuring an e-box dominance archive. The idea behind an e-box dominance is based on the subdivision of the objective space into hyper-boxes of e as the side length. The selection of Pareto-efficient solutions from each e-box guarantee convergence and diversity of the search process. Moreover, Borg MOEA utilizes time continuation through the introduction of mutated e-box dominance archived solutions for the diversification of the population, for maintaining the diversity and avoiding local minima *Taormina et al.* (2016). These properties guarantee for Borg MOEA to outperform other types of MOEA by the number of solutions, scalability to objective functions, ease-of-use and overall consistency across multiple problems *Reed et al.* (2013).

3.2.5 Extreme Learning Machines

Extreme Learning Machines (ELMs) are applied as a substitute learning algorithm to single-hidden layer feedforward neural networks *Huang et al.* (2006). These algorithms are used for problems of multiple classifications, clustering, regression, and feature engineering problems *Cananzi* (2021). It is based on an input layer, one or more hidden layers and the output layer *Ahmad et al.* (2018). ELM randomly chooses the weight of the inputs and analytically evaluates the weight of the outputs instead of performing a calibration of the parameters of the underlying functions as in traditional ANN, which hunt for the best combination of parameter values *Huang et al.* (2004); *Maier et al.* (2010). This characteristic fasten the learning process without loss in accuracy *Cananzi* (2021). These models can learn faster and have higher generalize capability than other feedforward network models *Huang et al.* (2006); *Ahmad et al.* (2018). Therefore, ELM model is used in the input variable selection process since computation time is necessary for the feature selection, while other models characterized by higher accuracies can be used for the regression model.

4

Case Study: The Nile River Basin

The Nile River is considered one of the longest and main hydrological veins in the world. It spreads on the 35 degree latitude covering around 6850 km with flow direction from south to north *Allan* (1995). The catchment basin of the Nile spreads over a large area of almost 10% the African continent passing through eleven countries. The basin is constituted by a severely stressed ecosystem shared by about 400 million people depending with their economy on agriculture *DESA* (2013). Studies on the projected flow shows that half of the population of the basin will be living below water scarcity level by the year 2030 *Baecher et al.* (2000); *Food and of the United Nations (FAO)* (2000). The Nile River Basin spreads over almost 3.25 Million Km^2 *Revenga et al.* (1998) which makes it a vast basin; thus it is split into 10 sub-basins by the Nile Basin Initiative in Figure 4.1. This study has been performed on the sub-basin level due to the large area of the basin where some of the input variables could be non-relevant in a sub-basin while be the opposite in another due to the heterogeneity of the River Nile Basin.

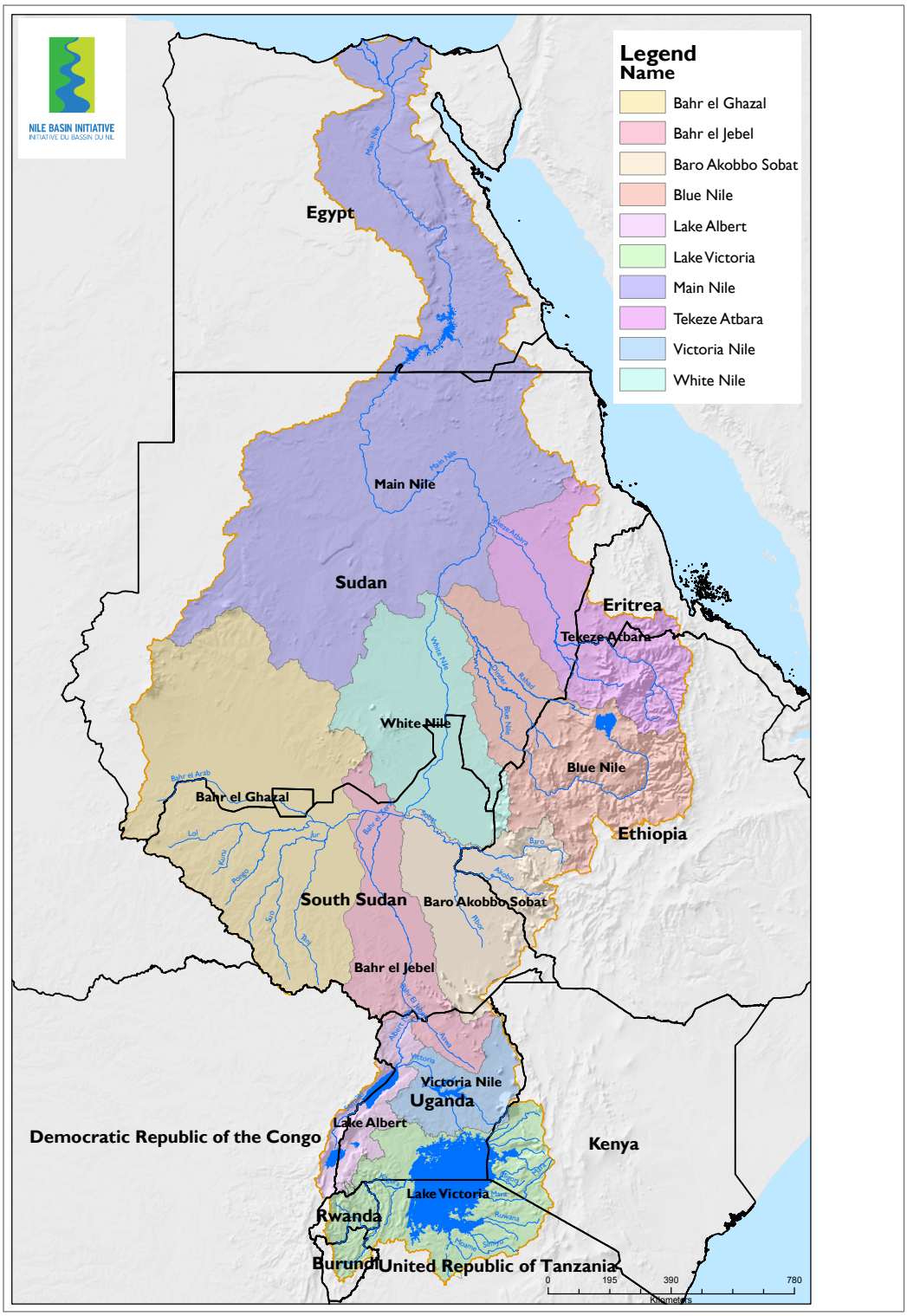


Figure 4.1: *River Nile major sub-basins NBI (2016)*

4.1 Basin characteristics and challenges

The substantial size of the basin makes it characterized by five climate zones of tropical, subtropical, arid, semi-arid and Mediterranean *Digna et al.* (2018) and thus it encompasses a unique wildlife and broad biodiversity. Two mountainous plateaus characterize the River Nile basin, the first known as the Equatorial or Lake Plateau on the southern part of the basin, where the second known as the Ethiopian or Abyssinian Plateau on the eastern part *Afan et al.* (2020).

4.1.1 Hydrology and climate

The Nile River is mainly formed by the Blue Nile and the White Nile as the major tributaries. The Blue Nile originates at Lake Tana in Ethiopia while the White Nile starts from Lake Victoria, the first is characterized by a highly seasonal flowrate contributing to almost 80% of the total river flow while the second has a relatively steady flow supplying around 10% to 20% of the streamflow *Afan et al.* (2020); *Omar* (2020)*Hilhorst* (2011).

The full utilization of a river basin requires a balance of infrastructure and economic development, in addition to stability by forcing environmental policies *Baecher et al.* (2000). The main driver for such balance is caused by the low streamflow volume compared to the water demand needed by the supplied population, since the Nile River Basin is considered a middle-range basin of about 2% of the Amazon water mass and 20% that of Mekong *Menniken* (2010). The low streamflow volume is explained by the small upstream portion of the basin providing most of the fresh water, hence the river flows for more than 50% of its route in arid and hyper-arid dry areas *Hilhorst* (2011). Therefore, the management of such a river basin should consider the full range of impacts on the whole boundary.

Climate change is considered as one of the most crucial challenges encountering natural ecosystems and human habitats *Solomon et al.* (2007); *IPCC* (2007, 2013). Intense and long droughts have been observed during the twentieth century mainly derived by an increase in temperature and decrease in precipitation levels *IPCC* (2007). Previous studies regarding this issue have highlighted that rainfall fluctuations highly affect on climate variability of the Lake Victoria basin and the Ethiopian Highlands *Conway* (2005). Future climate change would be the main cause of influencing higher stress on the already vulnerable water resources in the Nile River Basin, thus raising the need of an effective drought mitigation, adaptation and eventually reduce the future risks by policymakers.

4.1.2 Drought risk

Dry lands in arid and semiarid areas such as northern regions of the Nile River Basin are considered in risk caused by climate change and aridification which intensify the land degradation and desertification *Huang et al.* (2017); *Park et al.* (2018). A study performed by haile et al. concerning long term drought projections on East Africa, where the majority of the Nile River Basin is located, shows that the increase of temperature is mainly contributing to future climate system in the region, which projects an increase in droughts due to the fast warming in the studied area with respect to the global mean *Haile et al.* (2020).

The magnitude of projected droughts and impacts on environment and society are still largely unexplored *Haile et al.* (2020), which rises a necessity of investigating future droughts severity levels (moderate, severe, and extreme), and characteristics (region, duration, frequency, and intensity). Furthermore, since the river passes through arid countries such as Egypt which relies on the Nile river for 93% of its conventional water resources *Omar* (2020), in addition to high contribution of the Blue Nile in the river streamflow; Therefore, a decline in the precipitation over the Blue Nile basin would subsequently cause a hydrological drought in the river *Di Baldassarre et al.* (2011); *Trombetta* (2020).

4.1.3 Transboundary issues

Fresh water supplied by the Nile River plays a significant role of the socio-economic development of the countries in the basin. Water availability is essential for agricultural productivity, which makes it the dominant economic sector in most of the Nile riparians, providing employment and improving the living standards, in addition to the utilization of the Nile fresh water for hydropower generation. The Nile basin is afflicted with land degradation, in addition to decreasing water quality and quantity *Tekuya* (2020). However, it has great potential of improving the social and economic life through a collaborative utilization of water resources between the basin countries.

Population growth along the Nile River Basin is expected to increase significantly with no effective policies applied yet considering the extraordinary pressure on natural resources since the water supplied by the Nile River is essential for providing food security and rural development. As a consequence of the limited streamflow, the water supplied by the river is currently fully utilized by for industrial, domestic, and agriculture mainly by Egypt and Sudan resulting in less than 10 km^3 of the river water in a year flowing into the Mediterranean sea which is the minimum requirement for the environmental contribution *FAO* (2011).

Most of the upstream riparians are characterized by an abundant but variable rainfall which causes adverse impact on rainfed agriculture. Some of the upstream countries, notably Ethiopia, widely invested in hydraulic structures for mitigating the effects of weather uncertainties. Multiple dams have been built through the Blue Nile to control the variability of the flow and fully utilize the water resources of the basin as illustrated in Figure 4.2.

Nevertheless, the development of hydropower dams could cause environmental, social and economic costs if not correctly planned, specially dam sizing *Bertoni et al.* (2017) and location *Jozaghi et al.* (2018); *Schmitt et al.* (2018). The full utilization of the dam requires usually a long period of reservoir filling, withholding a large fraction of the river flow from downstream users. The speed of the reservoir filling process is the direct reason for potential conflicts between downstream and upstream countries *Zaniolo* (2020). However, variable hydroclimatic regimes makes harder the correct decision of impoundment strategies since the resulting filled reservoir could widely vary between a wet and a dry period for the same filling policy *Zaniolo* (2020).

The filling procedure of large dams has been the main reason for international conflicts in the past. A transboundary tension that met global resonance is the filling of The Grand Ethiopian Renaissance Dam (GERD) over the Blue Nile, with no specific agreement yet on water sharing or reservoir operation. This conflict was mainly raised due to the high reliability of Egypt and Sudan on the river water, such that these countries have nearly the full consumption of the river water, where the main purpose of the GERD construction for Ethiopia is the power generation rather than consumptive use, however the GERD operation will significantly effect on downstream flow patterns. Therefore, the main concern for Egypt and Sudan is regarding the initial filling period management in addition to the long-term operations. Many studies have been performed about this concern, such as the one done by Wheeler et al. which uses synthetic flows incorporating future changes projected by climate models. The simulation of three periods (i.e., reservoir filling, new normal period, and severe multi-year drought) resulted in considerable drop of levels in the Aswan High Dam (AHD), while the new normal period will benefit Ethiopia and Sudan without significant effects on Egypt, however the multi-year period includes high risks of harmful impacts that requires an effective management of the system *Wheeler et al.* (2020).

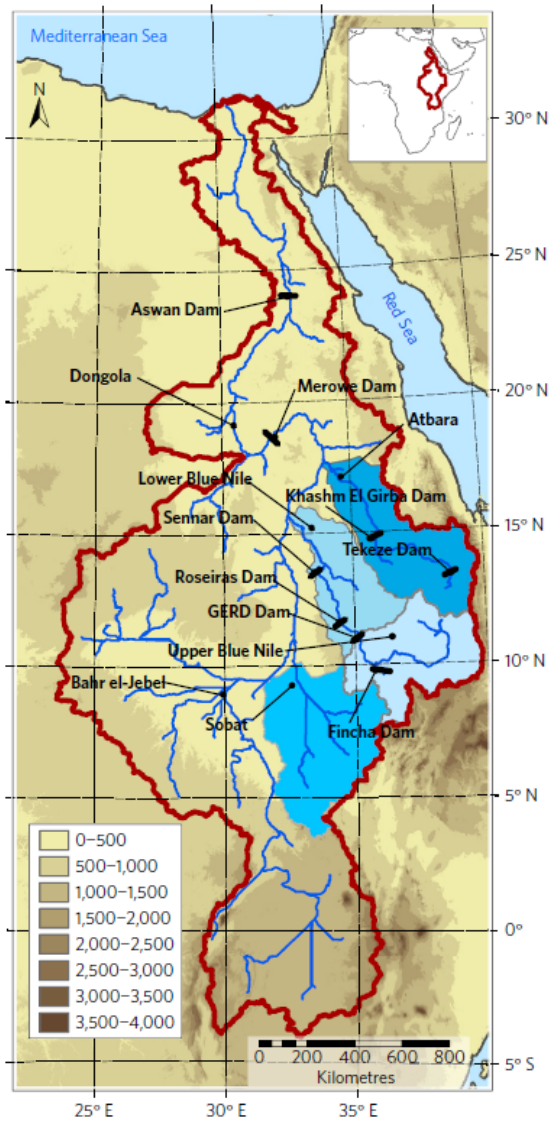


Figure 4.2: *Topographic map of the Nile River Basin showing the Blue Nile, Bako Akobbo-Sobat, Tekeze Atbara, and Bahr El Jebel sub-basins with the dams in these basins Siam and Eltahir (2017).*

4.2 Data collection

The collected datasets used to model drought over the Nile River sub-basins are time series of remote sensing data of precipitation in addition to maximum and minimum temperature (Tmax and Tmin), since studies have proven the correlation between drought and the chosen meteorological variables, moreover remote sensing data would provide fine spatial resolution over the whole study area *Sousa et al.* (2020); *Funk et al.* (2019). Furthermore, reanalysis data of soil moisture *Data and DISC)* (2015), river discharge *Harrigan et al.* (2019), and evapotranspiration (ET) *Muñoz Sabater et al.* (2019) were added to the whole col-

Table 4.1: *Properties of the collected variables*

Variable	Source	Spatial resolution	Temporal resolution
Precipitation	CHIRPS	0.05°x0.05°	Daily
Tmin	CHIRTS	0.05°x0.05°	Daily
Tmax	CHIRTS	0.05°x0.05°	Daily
ET	ERA5	0.1°x0.1°	Hourly
River discharge	Copernicus	0.1°x0.1°	Daily
Soil moisture	MERRA-2	0.5°x0.625°	Hourly
NDVI	NOAA STAR	4kmx4km	Weekly

lected datasets. River discharge and ET were chosen based on the case study where these two variables are critical (considerably low streamflow of the river and high ET specially in Lake Nasser at the AHD *Wheeler et al.* (2020), while soil moisture would provide a good representation of the long term effect of drought *Hao and AghaKouchak* (2013).

The time horizon spreads from 1984 to 2016 which is the common time horizon between all the available datasets, however excluding the year 2004 was necessary due to the missing data for the target remote sensing variable NDVI in this specific year. It can be noticed from Table 4.1 that the resolution of the datasets differ between the different resources. In order to overcome this issue, the mean value over the area under study (sub-basin scale) was calculated for every variable so that eventually one value represents the whole area per time step with respect to every variable. Furthermore, the temporal resolution was adjusted as a weekly time step the same as that of the target variable.

4.2.1 Remote sensing

Satellite or Remote Sensing observations are described by the procedure of data collection by analyzing imaginary data from a sensor without any physical contact with the target under study *Wambua* (2019). Remote sensing has been evolving through the years to become a reliable resource for developing remote sensing indices for drought monitoring, some of which are vegetation indicators *Monteleone et al.* (2020). Although drought monitoring through remote sensing data has been highly beneficial, however it suffers from multiple challenges such as data continuity, unquantified uncertainty, sensor changes, community acceptability, and the short length of available record *AghaKouchak et al.* (2015).

Normalized Difference Vegetation Index (NDVI)

The computation of the NDVI relies on the difference between the red spectral region, which defines the absorption of light by the chlorophyll, and the Near Infrared spectral region (NIR), where the leaves reflection occurs. The Center for Satellite Applications And Research (STAR) provides a Blended Vegetation Health Product (Blended-VHP) which is a re-processed Vegetation Health data set determined from an Advanced Very High Resolution Radiometer (AVHRR) and later from a Visible Infrared Imaging Radiometer Suite (VIIRS) with a Global Area Coverage data of a 4 km resolution and a weekly time step.

Precipitation

Estimating rainfall variations in space and time is an important factor for drought early warning and monitoring environmental variables. However, satellite data are areal averages affected by complex terrains that result in underestimating the intensity of extreme rainfall events. On the other hand, station data provide precipitation grids which suffer in rural areas where rain gauge stations could be limited. A collaboration between the USGS Earth Resources Observation and Science (EROS) Center created Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) as a complete, reliable, up-to-date datasets for trend analysis and seasonal drought monitoring *Sousa et al.* (2020). The CHIRPS datasets has been improved by removing systematic-bias through the application of a high resolution (0.05°x0.05°) gridded precipitation climatologies with a daily temporal resolution.

Temperature

High resolution mean weather variables are necessary to assess the impact on local and non-linear sectors such as health and agriculture. The variations in the absolute value of the weather variable are highly correlated to impacts on human health and crops. Nowadays there is a scarcity in the accuracy of temperature data monitoring, although this information is important for the evaluation of extreme temperatures which can wilt crops or decimate livestock herds leading to a famine.

However the Climate Hazards Center (CHC) developed a quasi-global high resolution (0.05°x0.05°) data set of daily maximum and minimum temperatures known as CHIRTS-daily *Funk et al.* (2019).

4.2.2 Reanalysis data

Evapotranspiration

The total amount of water evaporated from the surface of the Earth including transpiration from vegetation is provided from ERA5-Land reanalysis dataset *Muñoz Sabater et al.* (2019). Reanalysis combines model data with observations using the laws of physics into a complete and consistent dataset. ERA5 atmospheric variables are used to control the simulated land fields known as atmospheric forcing. This control is made to avoid the deviation of the model estimates from reality. Hence, the observations contribute in an indirect influence on the production of the ERA5-Land through the atmospheric forcing in the process of running the simulation. Moreover, the parameters of the atmospheric forcing (air temperature, air humidity and pressure) follow a lapse rate correction where the variables are corrected to account for the difference in altitude between the grid of the forcing and the higher resolution grid of ERA5-Land of (0.1°x0.1°) provided by an hourly temporal resolution *Muñoz Sabater et al.* (2019).

Soil moisture

The water content of the soil defined as the soil moisture is provided by the Modern-Era Retrospective analysis for Research and Applications the second version (MERRA-2) *Data and DISC)* (2015). The available dataset is a global reanalysis data used to assimilate space-based observations of aerosols and represent their interactions with other physical processes in climate system. This dataset collection consists of land surface diagnostics with a spatial resolution of (0.5°x0.625°) and an hourly temporal resolution.

River discharge

The river discharge is known as the amount of water passing through a section of area in a given time. The dataset is a global modeled daily data from the Global Flood Awareness System (GloFAS) supported by the Copernicus Emergency Management Service (CEMS) *Harrigan et al.* (2019) characterized by (0.1°x0.1°) spatial and daily temporal resolutions.

4.2.3 Sub-basins

The study was performed on the sub-basin scale according to the defined sub-basins in Figure 4.1 *NBI* (2016). The HydroSHEDS database provides freely

available vector layers (shapefiles) of hydrographic data products such as catchment boundaries, river networks, and lakes *Lehner and Grill* (2013). All the collected datasets were cropped to the sub-basin scale based on the layers provided by HydroSHEDS. Due to the special case of the Main Nile sub-basin of considering only the irrigated area for evaluating the NDVI in this sub-basin, the irrigated area was collected from MICRA2000 with a resolution of 10kmx10km *Portmann et al.* (2010).

5

Results and Discussion

This Chapter includes the results obtained from the application of FRIDA on the case study of the Nile River Basin. The Chapter starts with a brief description of the target variable, followed by the output of the input variable selection process known as the feature extraction over the Nile River Basin. In addition to a demonstration of the chosen regression model. Furthermore, a representation of the FRIDA results applied on the sub-basin scale illustrating the most common outcome in addition to specific cases. The results obtained by applying FRIDA on the Nile River Basin are compared with those considering all features (i.e., disregarding the major first two steps of FRIDA). The different outputs are finally discussed.

5.1 Target variable

The decision of a target variable in a model setup can be challenging, since it should be representative based on the study area characteristics and the scope of the research. Chapter 4 represents the need to have a reliable drought monitoring and forecasting index for the Nile River sub-basins. Studies performed using NDVI as the target variable highlighted the reliability of implementing a vegetation index for drought monitoring.

The economy of the Nile River Basin countries mainly depend on agriculture *DESA* (2013), which defines that the majority of the basin is utilized for agriculture. Consequently, the decision of a remote sensed vegetation health indicator can consistently describe the drought conditions and severity. The

NDVI is one of the most widely known vegetation health indexes, it has been applied in different studies with different basin conditions where it has proven the ability to describe the vegetation condition *Cananzi* (2021); *Zaniolo* (2020). Through the analysis of NDVI over time it is possible to realize the stress level of the vegetation in addition to different characteristics related to natural disturbances. Therefore, the choice of NDVI as the target variable aligns with the characteristics of the study area and the scope of the research, turning it to be the best target variable alternative for feature extraction in this case study.

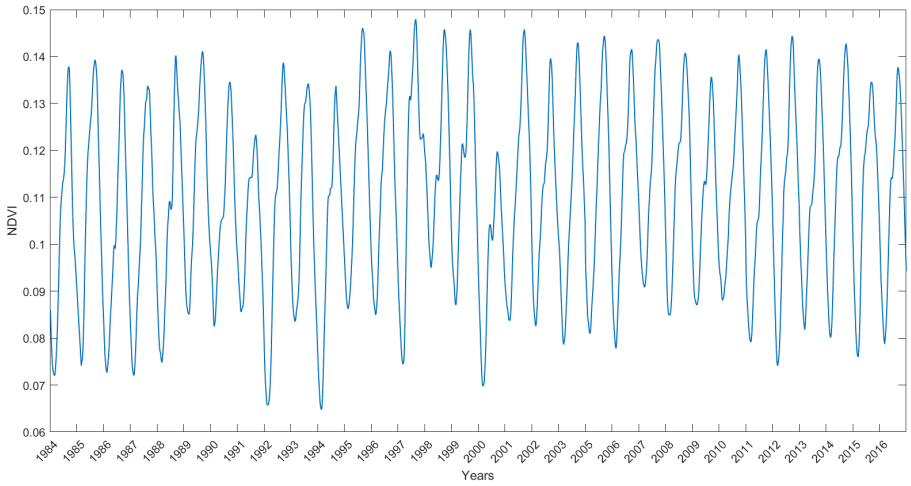
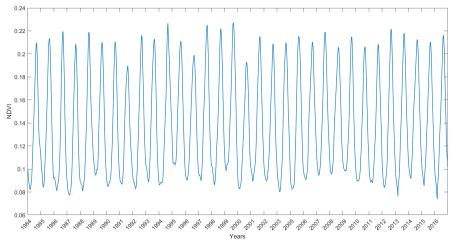
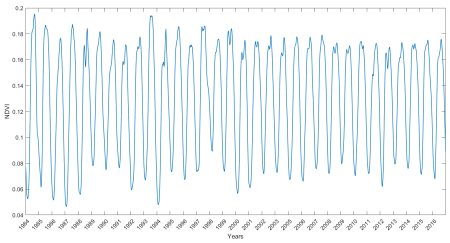


Figure 5.1: *Mean NDVI values for the River Nile Basin (STAR)*

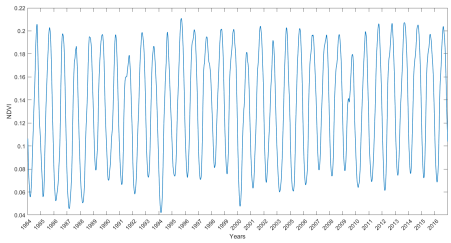
The values of NDVI usually range from 0 to 0.1 for barren rock, sand or snow. While this range becomes higher for green vegetation to reach 0.8, however for dense and healthy vegetation the range changes for a minimum of 0.6 and a maximum of 0.9. The mean NDVI values for Nile River Basin are presented in Figure 5.1. These values are evaluated as the average over the whole basin area excluding water bodies (i.e. Lake Albert and Lake Victoria) and desert (in the Main Nile sub-basin), which is affected by the arid lands mainly in the Main Nile sub-basin that is characterized by paucity of vegetation cover. Therefore, while lakes Albert and Victoria were excluded from the input datasets of the related sub-basins since water bodies represent zero vegetation cover, only the irrigated area was considered for extracting the NDVI values in those sub-basins, as well as for the Main Nile sub-basin *Portmann et al.* (2010). Figure 5.2 shows the mean NDVI values for all the sub-basins and it can be noticed the difference of range of values between the sub-basins based on its vegetation cover and characteristics (e.g., precipitation, humidity, and soil moisture).



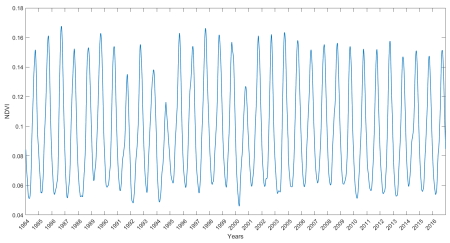
(a) Bahr El Ghazal



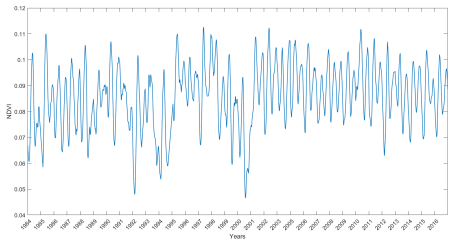
(b) Bahr El Jebel



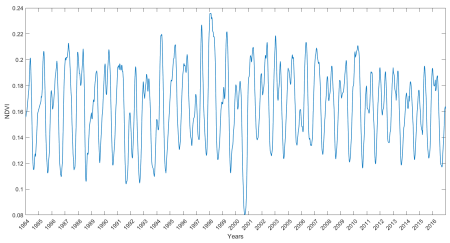
(c) Bako Akobbo-Sobat



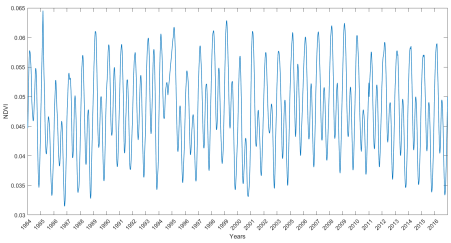
(d) Blue Nile



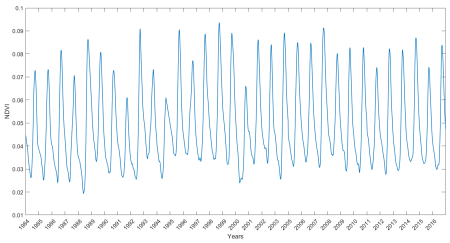
(e) Lake Albert



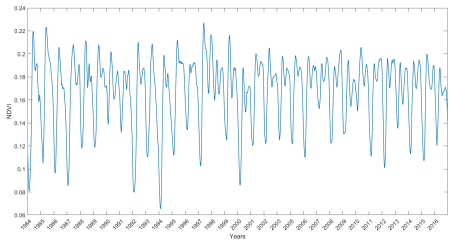
(f) Lake Victoria



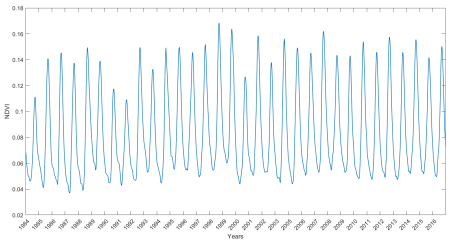
(d) Main Nile



(e) Tekeze Atbara



(f) Victoria Nile



(g) White Nile

Figure 5.2: *Observed NDVI values for all the sub-basins*

Table 5.1: *Candidate predictors time aggregation*

Feature Type	Feature Name	Time Aggregation (in weeks)
Time information	Year	Not aggregated
	Week	Not aggregated
Candidate predictors	Precipitation	1,2,4
	Minimum temperature	1,2
	Maximum temperature	1,2
	Mean temperature	1,2,4
	Evapotranspiration	1,3,6,16
	River discharge	1,2,4,16
	Soil moisture	1,16,52
Traditional drought indicator	SPI	1,3,6,16,52

5.2 Feature selection

The Input Variable Selection process is applied using the defined W-QEISS for feature extraction on the sub-basin scale, with candidate predictors characterized by a lag time aggregated as the same temporal resolution as the target variable, a weekly time step. The candidate variables are the year, week of the year, hydro-meteorological variables (i.e. precipitation, temperature, evapotranspiration, river discharge, and soil moisture), and a traditional drought indicator (i.e. SPI).

The choice of input variables was based on the capability of meteorological variables (i.e., precipitation and temperature) in identifying droughts, while the main reasons for choosing evapotranspiration and river discharge are the relatively small streamflow of the Nile River and high evaporation in the basin which are two critical points in this case study, moreover soil moisture is used as a predictor to provide a representation of the long term effect of a drought. The input variables have been also aggregated over different time periods as shown in Table 5.1 based on their physical processes (e.g. minimum and maximum temperature are not able to define the trajectory for a period longer than 2 weeks) and multiple trials on the Input Variable Selection process (e.g. evapotranspiration aggregated over 2 and 4 weeks were asynchronously selected, therefore 3 weeks average made the best time aggregation). The length of the dataset is 1664 data points, corresponding to weekly values for the period between 1984 to 2016 excluding 2004 and the number of candidate predictors of 28.

The settings of the W-QEISS were adjusted according to the guidelines proposed by Huang et al. and Karakaya et al. *Huang et al.* (2006); *Karakaya et al.* (2015) supported by a trial and error procedure to define the best parameters for the case study. Hence the maximum cardinality was chosen as 12 since the

Nile River Basin is a complex system and a low number of predictors could be insufficient to represent the whole system, the number of hidden neurons in the ELM is set to 28, based on the trade-off between a high number of neurons and a high computation time, while the number of folds for the k-fold cross validation set to 32 same as the number of years. However, for Borg MOEA, the number of function evaluations was set to 200,000 and the epsilon to 0.001 according to trial and errors in addition to experience gathered from previous work *Zaniolo* (2020). The W-QEISS was run for every sub-basin on the same settings, the result of the W-QEISS is a plot of a frequency matrix for alternative subset of predictors with a SU within a range of δ with respect to the best performing subset, the δ value varies for each sub-basin from 1% to 20% according to the SU of the lower performing subsets.

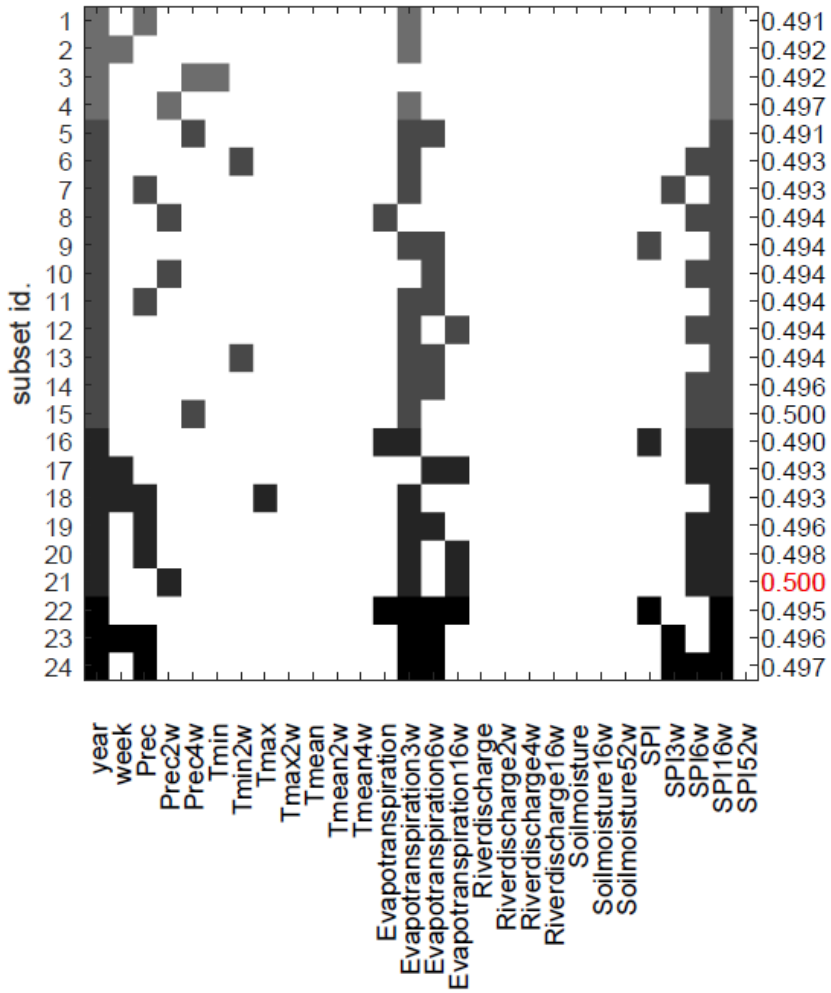


Figure 5.3: *Feature selection results for the Nile River Basin*

Figure 5.3 illustrates the frequency matrix of the results obtained for the Nile River Basin that shows a good performance of the feature selection with a sub-

set characterized by the highest SU of 0.5, where the right axis represents the SU from 0 to 1 so that the value 1 stands for perfect performance through the 4 objective functions (i.e. relevance, redundancy, cardinality, and predictive accuracy). The alternative subsets are sorted in ascending order of cardinality (from top to bottom), and accuracy (within each cardinality level). In order to read the matrix, the rows stand for the subset and columns for the candidate predictors, a rectangular marker represents the selected predictors for the subset. The marker color varies with the corresponding cardinality, the darker the color the higher the cardinality. The plot provides the possibility to identify the relevance of the predictors by the observation of vertical bars through joining markers over multiple rows. In the case of feature selection for the whole Nile River Basin, it can be noticed that the drought index SPI averaged over 16 weeks and the periodicity represented by the year are highly relevant variables as their absence would affect the model performance, where SPI provides the ability to detect the extremes as discussed in Chapter 2 and year for the stationary trajectory of the NDVI values shown in Figure 5.4.

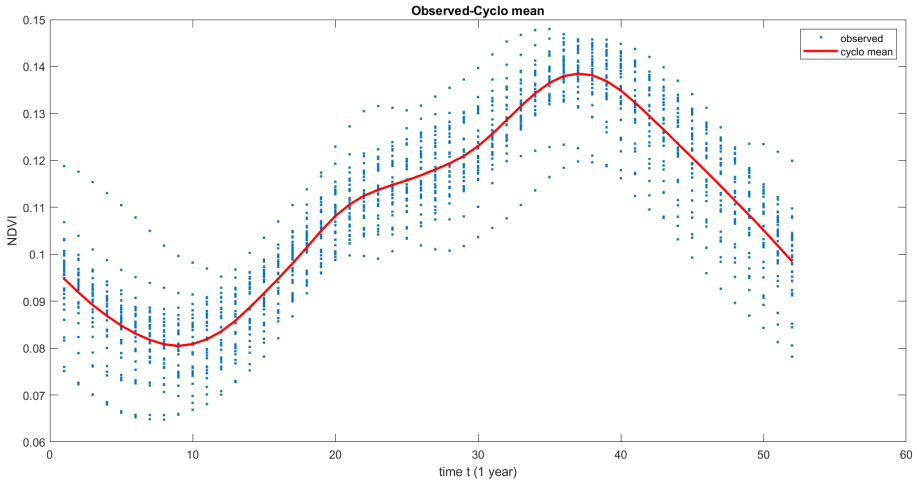


Figure 5.4: *NDVI observed values and cyclo-stationary mean in the Nile River Basin.*

Furthermore, evapotranspiration averaged over 3 weeks time period seems to be another highly relevant predictor, as it can be noticed the inclusion of this variable in 83% of the selected subsets, which provides the ability of the model to explain the changes in evaporation demands caused by temperature fluctuations as described in Chapter 2 by *Vicente-Serrano et al.* (2010). In this case, since the highest performing subsets have a close range of SU, it was preferred to select a subset with lower cardinality such as the one defined by subset id 4 with a negligible loss in SU from the highest performing one. The chosen subset has

a cardinality of 4 predictors, the most 3 relevant variables (i.e. year, evapotranspiration 3 weeks aggregation and SPI 16 weeks aggregation) in addition to the precipitation averaged over 2 weeks.

Although the feature extraction process is able to identify the relevant predictors, it also provides the possibility to detect the irrelevant variables. The results of the W-QEISS shows zero correlation between the irrelevant predictors, river discharge and soil moisture in this case, and NDVI values over the Nile River Basin which can be explained for the first by the different streamflow characteristics through the Nile River (e.g. the Blue Nile has a highly stochastic monthly flow while the White Nile has a stable one as discussed in Chapter 4), while the second is excluded due to the different characteristics of the considered geographical area where the Nile River Basin includes relatively humid areas in the south while arid areas represent the majority of the northern part of the basin which can highly affect the ability of soil moisture in describing the vegetation conditions.

5.3 Drought Index Modeling at the Nile River Basin scale

The model class decision is depending on different criteria that varies over case studies and scopes. An Extreme Learning Machine was used for the feature selection process in the W-QEISS due to the advantage of high speed of computation and a relatively high accuracy with respect to alternative non-linear models. However, modeling the drought index requires a high accuracy with low necessity of a fast regression model, hence a slower but a more accurate traditional back propagation one layer feedforward Artificial Neural Network makes the better alternative to be applied for the drought index modeling. Moreover, the performance of the ANN model is also compared with an interpretable linear model providing a good compromise between accuracy and transparency. Although a linear model is characterized by the simplicity, yet it is known to struggle with high seasonality features in contrary to non-linear models which can effortlessly handle such variations. Therefore, in order to provide a fair comparison between the applied models, the seasonality is removed by depurating the predictors of their cyclo-stationary mean.

Regarding the settings of the non-linear model, the number of hidden neurons was set to 3 as a common value between the sub-basins with 200 training iterations of the ANN model. Furthermore, a k-fold cross validation with 32 number of folds was applied for the calibration and validation of both models.

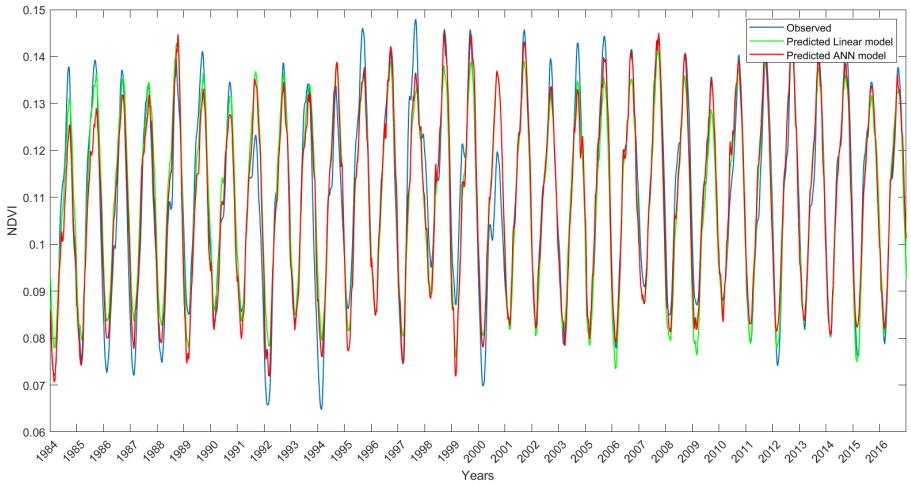


Figure 5.5: *Observed and predicted NDVI values for the Nile River Basin using linear and ANN models*

The calibrated linear model representing the predicted NDVI values is reported in Figure 5.5 providing a highly satisfying performance with an accuracy measured by the coefficient of determination in cross validation of $R^2_{linear} = 0.887$. In a further analysis, the ANN model calibration and cross validation presented in Figure 5.5 scores a slightly higher accuracy of $R^2_{ANN} = 0.9057$. Therefore, it shows that the FRIDA indexes are highly able to reproduce the target NDVI trajectory with a slightly preferable performance to non-linear models allowing ANN to be the better option for the Nile River Basin. However, more detailed analysis on the sub-basin scale was performed.

5.4 Drought index modeling at the Nile sub-basin scale

The River Nile Basin constitutes a large geographical area so that the prediction of a target variable could differ from a sub area to another caused by the different characteristics and vegetation cover between the sub basins given the heterogeneity of the Nile River Basin. Therefore, it was decided to run the same analysis on the 10 sub-basins of the Nile River Basin shown in Figure 4.1. The results of feature selection process and index modeling are presented for the sub-basins Blue Nile, Main Nile, and Lake Albert for the sake of synthesis and order for the thesis. The presented sub-basins are chosen such that Blue Nile is a good example of the high performing sub-basins, Main Nile is considered a special case since the majority of the area has a low index of vegetation cover, and Lake Albert as the lowest performing sub-basin. However, the results of the 7 remaining sub-basins are reported in the appendix for completeness. It can be observed that their performance is similar to that of the Blue Nile.

5.4.1 Blue Nile

The feature selection results for the Blue Nile are illustrated in Figure 5.6 where it can be noticed that the selected subset with the highest SU has an index of 0.633 with the lowest cardinality of 3 predictors. The chosen predictors provide a physical representation of NDVI trajectory such that the week describes the periodicity while river discharge, aggregated over different time periods, is included in all the subsets since it provides the ability to identify the stochastic flow of the Blue Nile River in the considered sub-basin. Moreover, it can be noticed that SPI aggregated over 16 weeks makes one of the major predictors since it was selected in 100% of the best performing subsets.

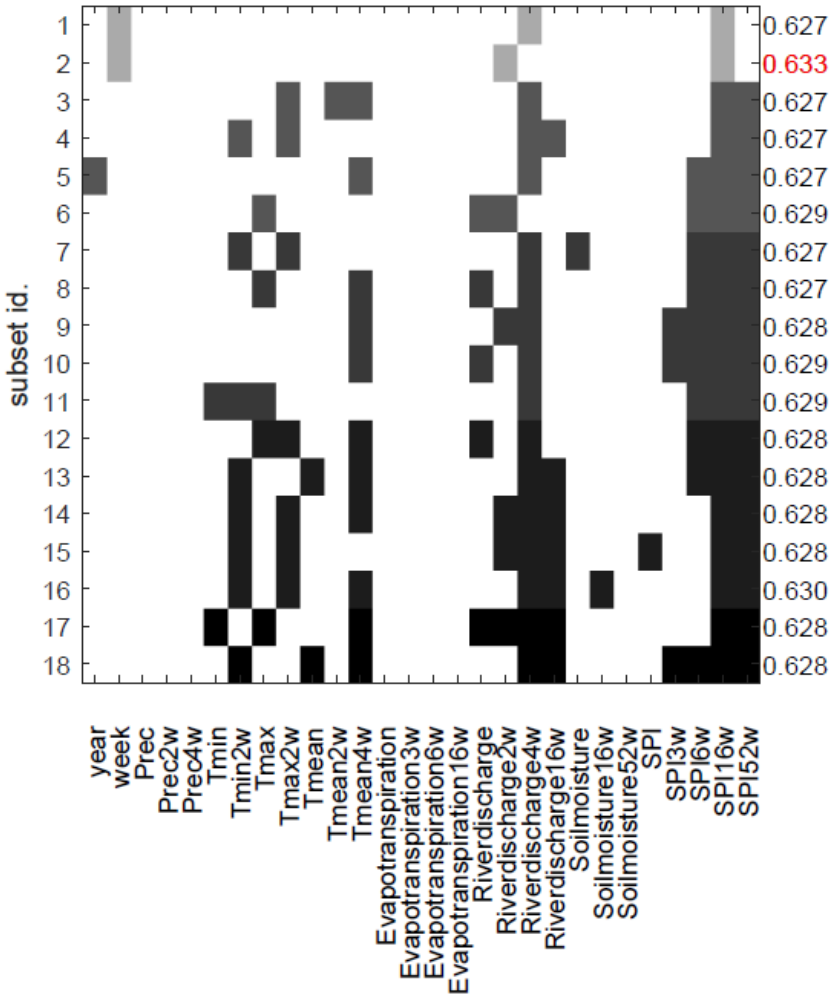


Figure 5.6: *Feature selection results for the Blue Nile*

Modeling of the drought index for the Blue Nile sub-basin resulted in a highly reliable and accurate index illustrated in Figure 5.7 with a coefficient of determination of $R^2_{linear} = 0.928$ and $R^2_{ANN} = 0.9288$ for the considered linear

and non-linear models, respectively.

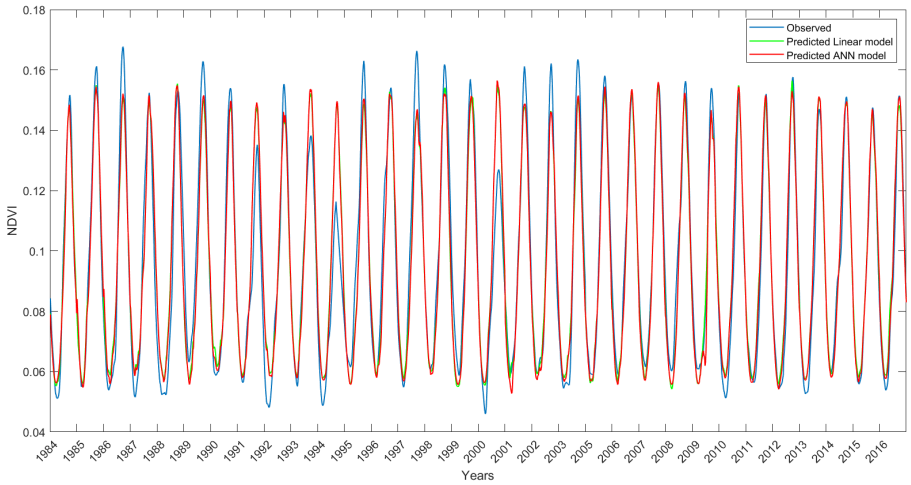


Figure 5.7: *Observed and predicted NDVI values for the Blue Nile sub-basin using linear and ANN models*

5.4.2 Main Nile

The Main Nile sub-basin, defining the northern part of the Nile River Basin as shown in Figure 4.1, was considered a more complex case since the majority of the sub-basin is composed of arid non-vegetated area. In order to avoid the filtering of the NDVI trajectory over the large area of unvarying values, the irrigated districts in Egypt were considered for the extraction of NDVI values for the studied sub-basin. Although only the section of area with a vegetation cover was considered for the Main Nile, yet the feature selection was incapable of finding a subset able to reproduce the NDVI trajectory with a high accuracy. Figure 5.8 illustrates the results of the input variable selection process where the best performing subset constitutes only the week as an input variable describing the periodicity of NDVI without considering any other predictor. Despite the fact that periodicity of the NDVI values can be a major predictor, though using only this variable can make the model vulnerable and unable to detect extremes or natural hazards (e.g. heat waves or droughts). Therefore, a more reasonable decision is to select the second best performing subset with a SU of 0.371 composed of precipitation, temperature, river discharge and SPI as predictors. By analyzing the alternative subsets, it is obvious the necessity of including meteorological and hydrological variables in addition to the widely used drought index SPI. The meteorological variables in addition to SPI provide the ability to detect the extremes, while river discharge is necessary to identify the vegetation health since the irrigated area in the Main Nile sub-basin

is mainly supported by fresh water from the river due to the low precipitation levels in this region as can be noticed from Figure 5.9.

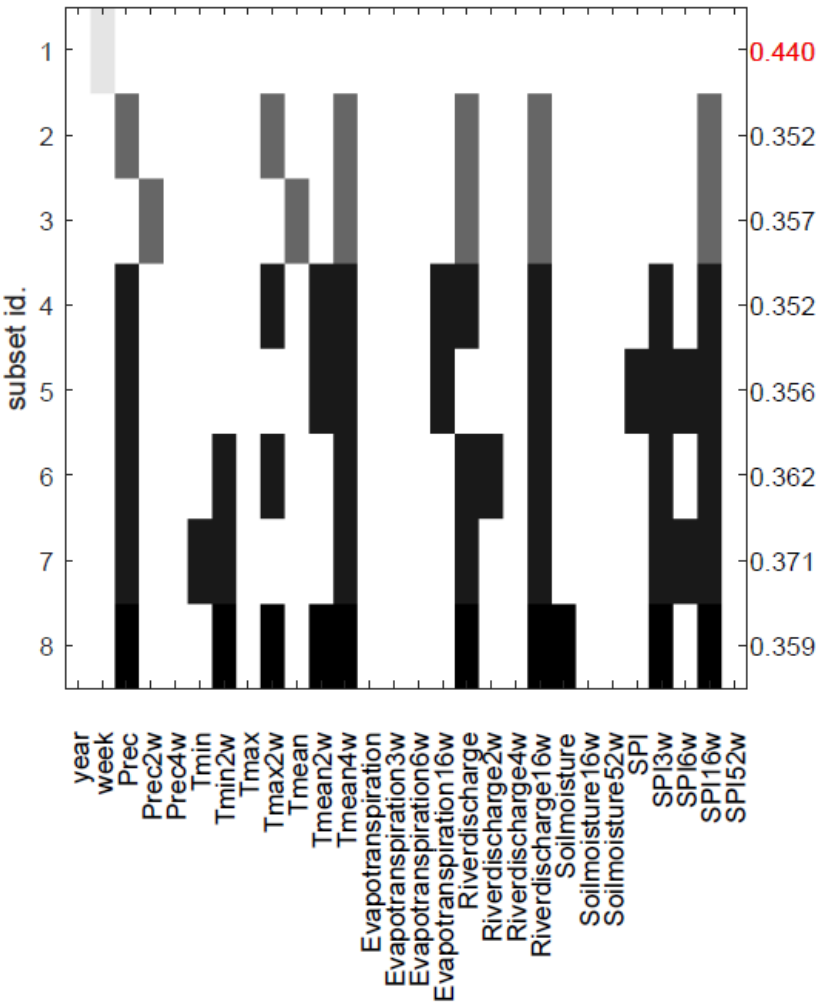


Figure 5.8: *Feature selection results for the Main Nile sub-basin.*

The drought index modeling was expected to have a lower performance than the other sub-basins as a result of the low SU in the feature selection process. The linear and ANN models results are represented in Figure 5.10 where it can be noticed the lower accuracy in terms of coefficient of determination with $R^2_{linear} = 0.7844$ for the linear model while the coefficient is slightly higher for the ANN model with $R^2_{ANN} = 0.7875$. However, considering the complexity of the study case, such behavior can be assumed as an acceptable performance to be applied for monitoring and forecasting droughts on the Main Nile sub-basin, where precipitation occurs very seldom throughout the year, it has high interannual variability and hard predictability *Siam and Eltahir* (2017).

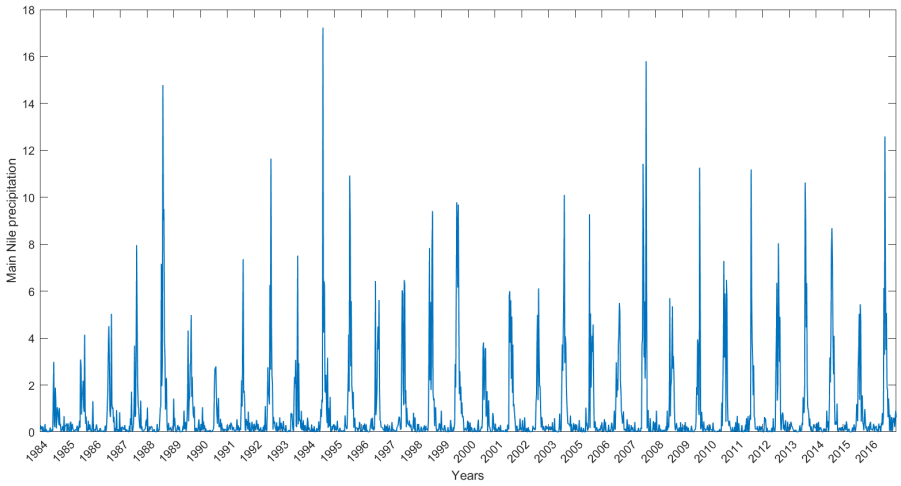


Figure 5.9: *Precipitation over the Main Nile sub-basin.*

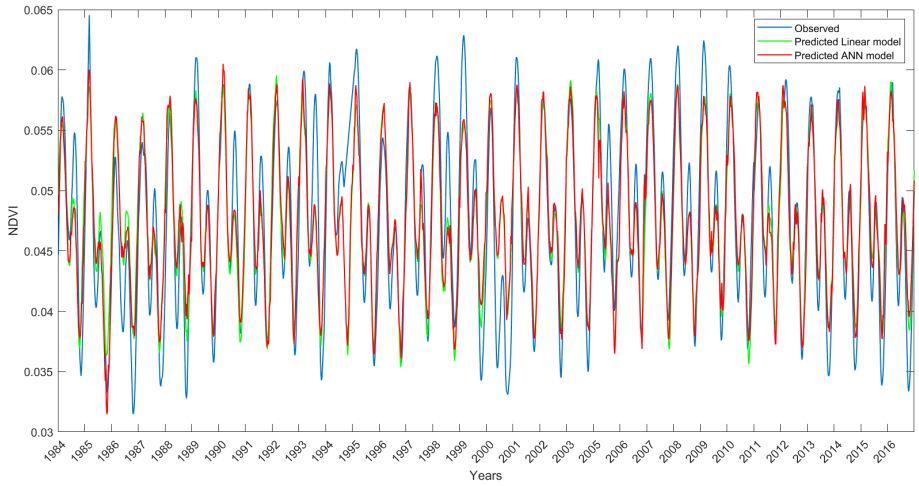


Figure 5.10: *Observed and predicted NDVI for the Main Nile sub-basin using linear and ANN models.*

5.4.3 Lake Albert

The application of a vegetation index on a region mainly covered by water bodies, such as Lake Albert sub-basin, is incapable of fully representing the considered area due to the scarce vegetation cover. Based on the fact that both sub-basins Lake Albert and Lake Victoria are mainly constituted by water bodies, then a low performance of reproducing the NDVI trajectory can be expected. Figure 5.11 illustrates the results of input variable selection for Lake Albert where the low SU can be observed with a maximum value reaching 0.204.

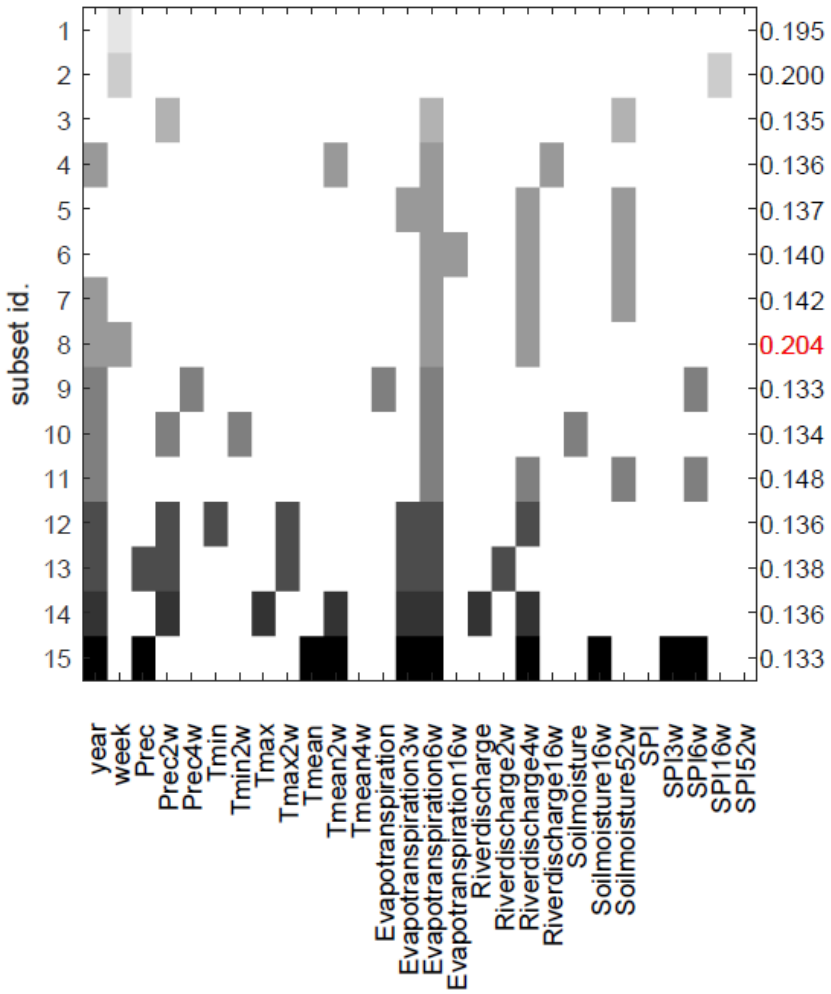


Figure 5.11: *Feature selection results for Lake Albert sub-basin.*

The obtained SU for Lake Albert is the lowest between all the other sub-basins; however, this is explained by the exclusion of water bodies, where Lake Albert sub-basin is mainly covered by a water body (i.e., lake Albert), in addition to the highly irregular trajectory of NDVI in Lake Albert sub-basin as shown in Figure 5.2. This irregularity is mainly explained by the relatively small area with respect to the other considered watersheds (Lake Albert has the smallest area of 44432 km^2 out of all the considered sub-basins, 306989 km^2 for Blue Nile), the increase of a surface area applies better noise filtering for input variables, hence a small surface area with a relatively coarse spatial resolution of input predictors would deteriorate the prediction performance. The selected features concerning the best performing subset in terms of SU result to be the year and week representing the variability of NDVI over the years and weeks of the year, in addition to evapotranspiration and river discharge. The selection of river discharge is considered necessary in sub-basins where water

bodies are a major component, such as Lake Albert. While evapotranspiration is a major predictor in this case since it was selected in 85% of the subsets, due to the large water surface which contributes to high evaporation and therefore making the evapotranspiration variable capable of explaining the climatic situation and hazards.

By applying the selected subset for the drought index modeling, a low accuracy was obtained represented by $R^2_{linear} = 0.4151$ and $R^2_{ANN} = 0.5498$ for both linear and ANN models as shown in Figure 5.12, respectively. The weak performance of the models is expected in correlation with the low SU of the selected subset due to the incapability of reproducing the NDVI values using ELM (feature selection) or linear and ANN (index modeling). However, it can be noticed the linear model highlighting its better ability to detect model irregularities.

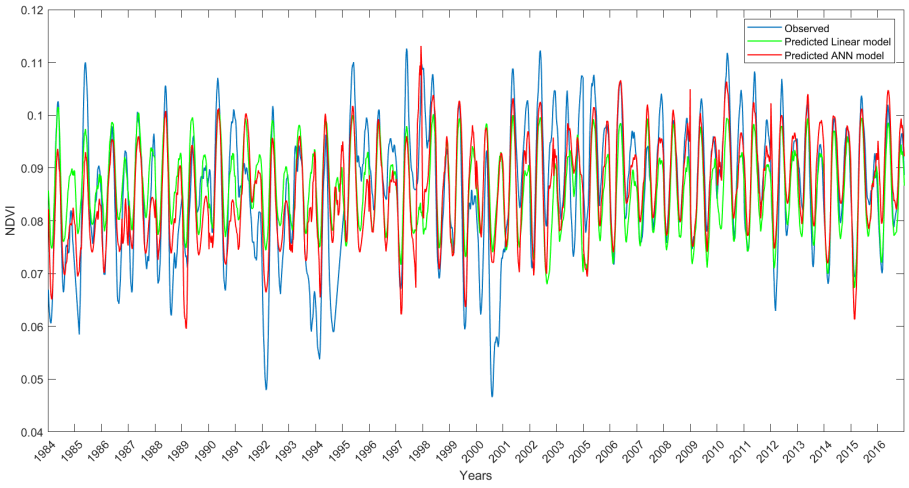


Figure 5.12: *Observed and predicted NDVI for the Lake Albert sub-basin using linear and ANN models.*

5.4.4 All Nile sub-basins

The output subsets of predictors obtained from the input variable selection for all the 10 Nile sub-basins are presented in Table 5.2, along with the relevant SU and the drought index models performance (i.e., coefficient of determination R^2). From the feature selection results, although some variables are common between most of the sub-basins (e.g. SPI), chosen predictors diversity can be observed between sub-basins. The difference of selected features highlights the heterogeneity of the Nile River Basin such that relative and informative input variables differ between sub-basins based on the geographical region, climate, land cover. However, SPI is the most selected predictor specially aggregated over 16 weeks which proves the ability of SPI in representing a drought and its

Table 5.2: *Feature selection outputs subsets of predictors for all 10 Nile sub-basin with the relevant SU and drought index models performance.*

Sub-basin	Year	Week	Precipitation	Tmin	Tmax	Tmean	ET	River discharge	Soil moisture	SPI	SU	R^2_{linear}	R^2_{ANN}
Bahr El Ghazal		x	4w							16w	0.675	0.9496	0.9538
Bahr El Jebel		x								16w	0.627	0.9379	0.9459
Bako Akobbo-Sobat	x	x	2w				1w		1w	6w, 16w	0.606	0.9243	0.9345
Blue Nile		x						2w		16w	0.633	0.928	0.9288
Lake Albert	x	x					6w	4w			0.204	0.4151	0.5498
Lake Victoria	x						6w	4w	52w		0.342	0.7841	0.8565
Main Nile			1w	1w, 2w		4w		1w, 16w		3w, 6w, 16w	0.371	0.7844	0.7875
Tekeze Atbara		x	2w							3w, 16w	0.495	0.8587	0.8775
Victoria Nile	x	x				4w	6w				0.248	0.7207	0.7843
White Nile	x								1w	6w, 16w	0.603	0.9234	0.937

high correlation with vegetation health. On the other hand, it is interesting to notice that temperature was one of the least selected predictors, and maximum temperature was not chosen in any of the sub-basins. This can be explained by the availability of other meteorological input variables that can provide the same information as temperature (e.g., ET).

Finally, After performing the study on many different cases, eventually by comparing the different accuracies between the chosen prediction models (i.e. linear and ANN) shown in Table 5.2, it shows that ANN is the better alternative due to the higher performance than the linear one, specially for Lake Albert and Lake Victoria sub-basins. Therefore, tolerating the use of a more complex and time consuming model for a better prediction of the target variable.

5.5 Feature selection VS All features

In Chapter 3, it was described how the first step of FRIDA is based on the subset selection of predictors known as the Input Variable Selection process, that was applied using W-QEISS in this study. In order to prove the importance of this step and to ensure that the outputs meets the required objectives, the drought index model was performed using all input predictors to compare different accuracies of the obtained index with that of using the selected features. The final comparison results are shown in Figure 5.13 such that the performance of the model is shown in vertical bars using the subset identified in Table 5.2 for every sub-basin.

By observing the models performance in Figure 5.13, it can be noticed that the models performance using inputs selected by W-QEISS is relatively similar, some cases better, to that of implementing all the input variables disregarding the selected subsets. Although in some cases (e.g. Lake Victoria) using all features shows slightly better performance than that of using the selected features, however in other cases (e.g. Main Nile) decreasing the complexity and modeling the index using selected features as inputs provides higher accuracy

than that of implementing all features. Therefore, since the difference of performance can be considered negligible in all cases between the prediction with all features or the selected subset, then it can be verified that the W-QEISS is meeting the required objective of decreasing the complexity and cardinality of the system without influencing on the prediction accuracy of the drought index model.

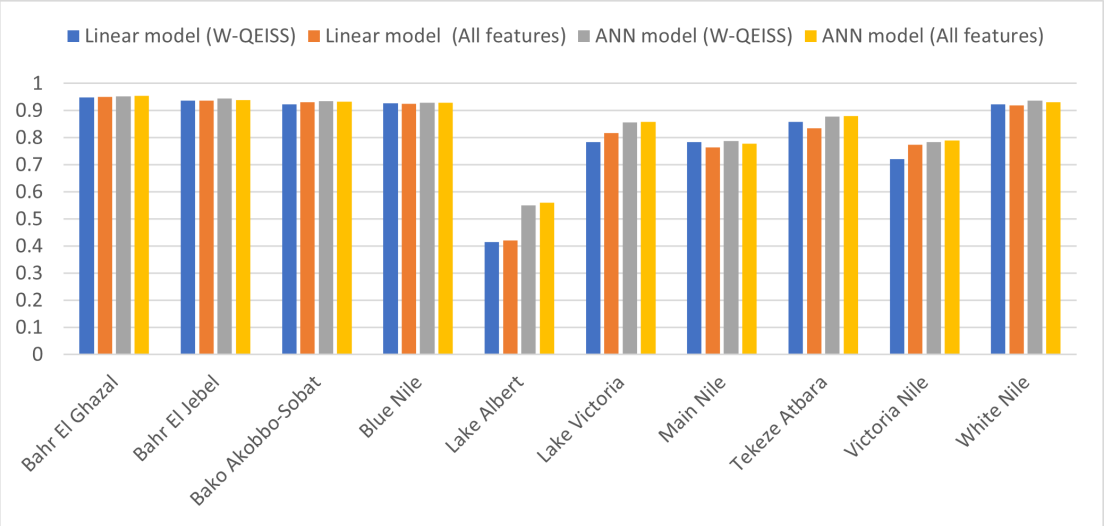


Figure 5.13: *Drought index models performance using all features and predictors selected by W-QEISS*

6

Conclusions and Future research

This research is aimed at providing a drought monitoring and forecasting index through the application of a novel framework (i.e., FRamework for Index based Drought Analysis (FRIDA)), based on machine learning tools and applied on the Nile River Basin. The Nile River Basin was chosen as a case study due to the challenges raised by:

- Climate change, where the analysis of climate trends illustrates a projected warming and decreasing precipitation, hence increasing drought frequency and severity.
- Transboundary water management, identified by the political conflicts and lack of coordination among the upstream and downstream countries.

The proposal of a customized basin-specific drought index is a necessity for supporting water management in an area challenged by climate change, water stress, and transboundary waters shared among different countries. Hence, motivating research in providing further insights and tools for better drought monitoring and, consequently, improve water management. Thus, this study is focused on the development of an index able to identify drought conditions (e.g., frequency, severity, intensity), using different prediction variables according to the study area.

The presented framework is supposed to be a portable and easy-to-use methodology that is based on the application of the Wrapper for Quasi Equally Informative Subset Selection (W-QEISS) as a feature extraction technique. This algorithm is supposed to find the subset with the highest predictive accuracy, in

addition to other multiple quasi equally informative subsets, providing awareness of the predictors relevance and a better understanding of the physical processes. The W-QEISS employs a deep learning machine for the predictors selection by developing a model able to reproduce the target variable, i.e. Normalized Drought Vegetation Index (NDVI). The choice of target variable offers a fully data-driven approach based on the direct consideration of the drought state proxy and its impact on the crop biomass. Furthermore, NDVI makes the best alternative for this case study, since the Nile River Basin countries highly depend on agriculture for their economy that highlights on the application of an indicator able to detect agricultural droughts. However, the chosen target variable is affected by the land cover in the study area, due to the filtering effect explained by the inclusion of almost zero values referring to non-vegetated land. In order to avoid this issue, a subdivision of the main study area into smaller sub-basins based on land cover, vegetation health, and climate; consequently, providing the possibility to identify the differences and similarities between the sub-basins.

In this work, the employment of FRIDA on the Nile River Basin to design an accurate drought index shows success in most of the Nile sub-basins. The outcomes highlight on the importance to apply a consistent basin subdivision. In fact, the sub-basins have different surface areas, topographic, and climatic characteristics. Thus, in future work, a subdivision accounting also for the spatial resolution of input variables is suggested to obtain the best performance of FRIDA; therefore ensuring noise filtering, where the mean value over the sub-basin can be fully representative, not affected by extremes or outliers in a portion of the sub-basin; hence, avoiding a behavior such as the case of Lake Albert. On the other hand, this will relevantly increase the computation time and costs.

Moreover, in this work there is the assumption that land cover was constant through the past and will remain in the future, which can affect on the NDVI efficiency as a target variable. Furthermore, the performance of FRIDA is highly dependent on predictors diversity, numerosity, and length of the time series. Therefore, introducing variables such as ground water level and air humidity may improve the behavior of future designed indexes.

Future studies can address the development of a classification instead of a regression model for the design of drought indexes which are expected to perform better in extremes detection and identification of drought conditions.

Bibliography

Abramopoulos, F., C. Rosenzweig, and B. Choudhury (1988), Improved ground hydrology calculations for global climate models (gcms): Soil water movement and evapotranspiration, *Journal of Climate*, pp. 921–941.

Abramowitz, M., I. A. Stegun, and R. H. Romer (1988), Handbook of mathematical functions with formulas, graphs, and mathematical tables.

Afan, H. A., et al. (2020), Input attributes optimization using the feasibility of genetic nature inspired algorithm: Application of river flow forecasting, *Scientific reports*, *10*(1), 1–15.

AghaKouchak, A. (2015), A multivariate approach for persistence-based drought prediction: Application to the 2010–2011 east africa drought, *Journal of Hydrology*, *526*, 127–135.

AghaKouchak, A., and N. Nakhjiri (2012), A near real-time satellite-based global drought climate data record, *Environmental Research Letters*, *7*(4), 044,037.

AghaKouchak, A., A. Farahmand, F. Melton, J. Teixeira, M. Anderson, B. D. Wardlow, and C. Hain (2015), Remote sensing of drought: Progress, challenges and opportunities, *Reviews of Geophysics*, *53*(2), 452–480.

Ahmad, I., M. Basheri, M. J. Iqbal, and A. Rahim (2018), Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection, *IEEE access*, *6*, 33,789–33,795.

Ahmad, M., C. Sinclair, and A. Werritty (1988), Log-logistic flood frequency analysis, *Journal of Hydrology*, *98*(3-4), 205–224.

Ahmadalipour, A., H. Moradkhani, A. Castelletti, and N. Magliocca (2019), Future drought risk in africa: Integrating vulnerability, climate change, and population growth, *Science of the Total Environment*, *662*, 672–686.

Allan, J. A. (1995), *The Geographical Journal*, *161*(1), 90–91.

Allen, R. G., L. S. Pereira, D. Raes, M. Smith, et al. (1998), Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56, *Fao, Rome*, *300*(9), D05,109.

Baecher, G., R. Anderson, B. Britton, K. Brooks, and J. Gaudet (2000), The nile basin: Environmental transboundary opportunities and constraints analysis, *International Resources Group, for USAID, Washington DC*.

Beniston, M., et al. (2007), Future extreme events in european climate: an exploration of regional climate model projections, *Climatic change*, *81*(1), 71–95.

Beran, M. A., and J. A. Rodier (1985), *Hydrological aspects of drought*, Unesco.

Bertoni, F., M. Giuliani, and A. Castelletti (2017), Scenario-based fitted q-iteration for adaptive control of water reservoir systems under uncertainty, *IFAC-PapersOnLine*, *50*(1), 3183–3188.

Bordi, I., K. Fraedrich, and A. Sutera (2009), Observed drought and wetness trends in europe: an update, *Hydrology and Earth System Sciences*, *13*(8), 1519–1530, doi: 10.5194/hess-13-1519-2009.

Bothe, O., K. Fraedrich, and X. Zhu (2011), Large-scale circulations and tibetan plateau summer drought and wetness in a high-resolution climate model, *International Journal of Climatology*, *31*(6), 832–846.

Bowden, G. J., G. C. Dandy, and H. R. Maier (2005), Input determination for neural network models in water resources applications. part 1,Äîbackground and methodology, *Journal of Hydrology*, *301*(1-4), 75–92.

Cananzi, D. (2021), Improving drought monitoring via machine learning: a new impact-based drought index.

Carrao, H., G. Naumann, and P. Barbosa (2016), Mapping global patterns of drought risk: An empirical framework based on sub-national estimates of hazard, exposure and vulnerability, *Global Environmental Change*, *39*, 108–124, doi: https://doi.org/10.1016/j.gloenvcha.2016.04.012.

Conway, D. (2005), From headwater tributaries to international river: Observing and adapting to climate variability and change in the nile basin, *Global Environmental Change*, *15*(2), 99–114.

Cunningham, P. (2008), Dimension reduction, in *Machine learning techniques for multimedia*, pp. 91–112, Springer.

Data, G. E. S., and I. S. C. G. DISC) (2015), Global modeling and assimilation office (gmao)(2015), merra-2 tavg1_2d_slv_nx: 2d, 1-hourly, time-averaged, single-level, assimilation, single-level diagnostics v5. 12.4.

DESA, U. (2013), Population division (2015) world population prospects: the 2015 revision, key findings and advance tables, *United Nations, New York*, *53*.

Di Baldassarre, G., et al. (2011), Future hydrology and climate in the river nile basin: a review, *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, *56*(2), 199–211.

Digna, R. F., M. E. Castro-Gama, P. Van der Zaag, Y. A. Mohamed, G. Corzo, and S. Uhlenbrook (2018), Optimal operation of the eastern nile system using genetic algorithm, and benefits distribution of water resources development, *Water*, *10*(7), 921.

Dracup, J. A., K. S. Lee, and E. G. Paulson Jr (1980), On the definition of droughts, *Water resources research*, *16*(2), 297–302.

Dubrovsky, M., M. D. Svoboda, M. Trnka, M. J. Hayes, D. A. Wilhite, Z. Zalud, and P. Hlavinka (2009), Application of relative drought indices in assessing climate-change impacts on drought conditions in czechia, *Theoretical and Applied Climatology*, *96*(1), 155–171.

Estrela, T., and E. Vargas (2012), Drought management plans in the european union. the case of spain, *Water resources management*, *26*(6), 1537–1553.

FAO (2011), Synthesis report: Fao-nile basin project, gcp/int/945/ita, 2004-2009.

Food, and A. O. of the United Nations (FAO) (2000), *New Dimensions in Water Security: Water, society and ecosystem services in the 21st century*, FAO.

Funk, C., et al. (2019), A high-resolution 1983–2016 t max climate data record based on infrared temperatures and stations by the climate hazard center, *Journal of Climate*, *32*(17), 5639–5658.

Galelli, S., and A. Castelletti (2013), Tree-based iterative input variable selection for hydrological modeling, *Water Resources Research*, *49*(7), 4295–4310.

Galelli, S., G. B. Humphrey, H. R. Maier, A. Castelletti, G. C. Dandy, and M. S. Gibbs (2014), An evaluation framework for input variable selection algorithms for environmental data-driven models, *Environmental Modelling & Software*, *62*, 33–51.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *Journal of hydrologic engineering*, *12*(4), 347–368.

Genest, C., J.-F. Quessy, and B. Rémillard (2006), Goodness-of-fit procedures for copula models based on the probability integral transformation, *Scandinavian Journal of Statistics*, *33*(2), 337–366.

Guenang, G. M., and F. M. Kamga (2014), Computation of the standardized precipitation index (spi) and its use to assess drought occurrences in cameroon over recent decades, *Journal of Applied Meteorology and Climatology*, *53*(10), 2310–2324.

Guttman, N. B. (1998), Comparing the palmer drought index and the standardized precipitation index 1, *JAWRA Journal of the American Water Resources Association*, *34*(1), 113–121.

Guttman, N. B. (1999), Accepting the standardized precipitation index: a calculation algorithm 1, *JAWRA Journal of the American Water Resources Association*, *35*(2), 311–322.

Guyon, I., and A. Elisseeff (2003), An introduction to variable and feature selection, *Journal of machine learning research*, *3*(Mar), 1157–1182.

Haile, G. G., Q. Tang, S.-M. Hosseini-Moghari, X. Liu, T. Gebremicael, G. Leng, A. Kebede, X. Xu, and X. Yun (2020), Projected impacts of climate change on drought patterns over east africa, *Earth's Future*, *8*(7), e2020EF001,502.

Hao, Z., and A. AghaKouchak (2013), Multivariate standardized drought index: A parametric multi-index model, *Advances in Water Resources*, *57*, 12–18, doi: https://doi.org/10.1016/j.advwatres.2013.03.009.

Hao, Z., and A. AghaKouchak (2014), A nonparametric multivariate multi-index drought monitoring framework, *Journal of Hydrometeorology*, *15*(1), 89–101.

Hao, Z., and V. P. Singh (2015), Drought characterization from a multivariate perspective: A review, *Journal of Hydrology*, *527*, 668–678.

Hao, Z., A. AghaKouchak, N. Nakhjiri, and A. Farahmand (2014), Global integrated drought monitoring and prediction system, *Scientific data*, *1*(1), 1–10.

Harrigan, S., E. Zsoter, C. Barnard, F. Wetterhall, P. Salamon, and C. Prudhomme (2019), River discharge and related historical data from the global flood awareness system, v2. 1, copernicus climate change service (c3s) climate data store (cds).

Hayes, M., M. Svoboda, N. Wall, and M. Widhalm (2011), The lincoln declaration on drought indices: universal meteorological drought index recommended, *Bulletin of the American Meteorological Society*, *92*(4), 485–488.

Hayes, M. J. (2002), *Drought indices*, National drought mitigation center, University of Nebraska.

Hayes, M. J., M. D. Svoboda, B. D. Wardlow, M. C. Anderson, and F. Kogan (2012), Drought monitoring: Historical and current perspectives.

Heim Jr, R. R. (2002), A review of twentieth-century drought indices used in the united states, *Bulletin of the American Meteorological Society*, *83*(8), 1149–1166.

Hilhorst, B. (2011), Information products for nile basin water resources management. synthesis report.

Hu, Q., and G. D. Willson (2000), Effects of temperature anomalies on the palmer drought severity index in the central united states, *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *20*(15), 1899–1911.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2004), Extreme learning machine: a new learning scheme of feedforward neural networks, in *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*, vol. 2, pp. 985–990, Ieee.

Huang, G.-B., Q.-Y. Zhu, and C.-K. Siew (2006), Extreme learning machine: theory and applications, *Neurocomputing*, *70*(1-3), 489–501.

Huang, J., et al. (2017), Dryland climate change: Recent progress and challenges, *Reviews of Geophysics*, *55*(3), 719–778, doi: https://doi.org/10.1002/2016RG000550.

IPCC (2007), Climate change 2007: The physical science basis.

IPCC (2013), The physical science basis, the working group i contribution to the un ipcc,Äôs fifth assessment report (wg1 ar5).

Jamshidi, H., D. Khalili, M. R. Zadeh, and E. Z. Hosseinipour (2011), Assessment and comparison of spi and rdi meteorological drought indices in selected synoptic stations of iran, in *World Environmental and Water Resources Congress 2011: Bearing Knowledge for Sustainability*, pp. 1161–1173.

Jozaghi, A., B. Alizadeh, M. Hatami, I. Flood, M. Khorrami, N. Khodaei, and E. Ghasemi Tousi (2018), A comparative study of the ahp and topsis techniques for dam site selection using gis: A case study of sistan and baluchestan province, iran, *Geosciences*, *8*(12), 494.

Kao, S.-C., and R. S. Govindaraju (2010), A copula-based joint deficit index for droughts, *Journal of Hydrology*, *380*(1), 121–134, doi: https://doi.org/10.1016/j.jhydrol.2009.10.029.

Karakaya, G., S. Galelli, S. D. Ahipaşaoğlu, and R. Taormina (2015), Identifying (quasi) equally informative subsets in feature selection problems for classification: a max-relevance min-redundancy approach, *IEEE transactions on cybernetics*, *46*(6), 1424–1437.

Keyantash, J., and J. A. Dracup (2002), The quantification of drought: an evaluation of drought indices, *Bulletin of the American Meteorological Society*, *83*(8), 1167–1180.

Keyantash, J. A., and J. A. Dracup (2004), An aggregate drought index: Assessing drought severity based on fluctuations in the hydrologic cycle and surface water storage, *Water Resources Research*, *40*(9).

Lehner, B., and G. Grill (2013), Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, *Hydrological Processes*, *27*(15), 2171–2186, doi: https://doi.org/10.1002/hyp.9740.

Limburg, K. E., R. V. O'Neill, R. Costanza, and S. Farber (2002), Complex systems and valuation, *Ecological economics*, *41*(3), 409–420.

López-Moreno, J. I., and S. M. Vicente-Serrano (2008), Positive and negative phases of the wintertime north atlantic oscillation and drought occurrence over europe: a multitemporal-scale approach, *Journal of Climate*, *21*(6), 1220–1243.

Maier, H. R., and G. C. Dandy (2000), Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, *Environmental modelling & software*, *15*(1), 101–124.

Maier, H. R., A. Jain, G. C. Dandy, and K. P. Sudheer (2010), Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, *Environmental modelling & software*, *25*(8), 891–909.

Mavromatis, T. (2007), Drought index evaluation for assessing future wheat production in greece, *International Journal of Climatology*, *27*(7), 911–924, doi: https://doi.org/10.1002/joc.1444.

McKee, T. B., N. J. Doesken, J. Kleist, et al. (1993), The relationship of drought frequency and duration to time scales, in *Proceedings of the 8th Conference on Applied Climatology*, vol. 17, pp. 179–183, California.

Menniken, T. (2010), *Hydrological regionalism in the Mekong and the Nile Basin: International politics along transboundary watercourses*, Dr. Kovač.

Mishra, A., T. Vu, A. V. Veettil, and D. Entekhabi (2017), Drought monitoring with soil moisture active passive (smap) measurements, *Journal of Hydrology*, *552*, 620–632, doi: https://doi.org/10.1016/j.jhydrol.2017.07.033.

Mishra, A. K., and V. P. Singh (2010), A review of drought concepts, *Journal of hydrology*, *391*(1-2), 202–216.

Mo, K. C., and J. E. Schemm (2008), Relationships between enso and drought over the southeastern united states, *Geophysical Research Letters*, *35*(15).

Monteleone, B., B. Bonaccorso, and M. Martina (2020), A joint probabilistic index for objective drought identification: the case study of haiti, *Natural Hazards and Earth System Sciences*, *20*(2), 471–487.

Muñoz Sabater, J., et al. (2019), Era5-land hourly data from 1981 to present, *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, *10*.

Nalbantis, I. (2008), Evaluation of a hydrological drought index, *European Water*, *23*(24), 67–77.

Narasimhan, B., and R. Srinivasan (2005), Development and evaluation of soil moisture deficit index (smdi) and evapotranspiration deficit index (etdi) for agricultural drought monitoring, *Agricultural and forest meteorology*, *133*(1-4), 69–88.

NBI (2016), *Guidelines for Wetlands Ecosystems Valuation in the Nile Basin*.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Omar, M. (2020), Impact of meteorological drought in upper blue nile basin on the hydrological drought of nile river in egypt, doi: 10.35940/ijeat.F1213.089620.

Palmer, W. C. (1965), *Meteorological drought*, vol. 30, US Department of Commerce, Weather Bureau.

Park, C.-E., et al. (2018), Keeping global warming within 1.5 c constrains emergence of aridification, *Nature Climate Change*, *8*(1), 70–74.

Pathak, A. A., and B. Dodamani (2020), Comparison of meteorological drought indices for different climatic regions of an indian river basin, *Asia-Pacific Journal of Atmospheric Sciences*, *56*(4), 563–576.

Pettorelli, N., J. O. Vik, A. Mysterud, J.-M. Gaillard, C. J. Tucker, and N. C. Stenseth (2005), Using the satellite-derived ndvi to assess ecological responses to environmental change, *Trends in ecology & evolution*, *20*(9), 503–510.

Portmann, F. T., S. Siebert, and P. Doll (2010), Mirca2000-global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling, *Global Biogeochemical Cycles*, *24*(1), doi: https://doi.org/10.1029/2008GB003435.

Rebetez, M., H. Mayer, O. Dupont, D. Schindler, K. Gartner, J. P. Kropp, and A. Menzel (2006), Heat and drought 2003 in europe: a climate synthesis, *Annals of Forest Science*, *63*(6), 569–577.

Reed, P. M., D. Hadka, J. D. Herman, J. R. Kasprzyk, and J. B. Kollat (2013), Evolutionary multiobjective optimization in water resources: The past, present, and future, *Advances in water resources*, *51*, 438–456.

Revenga, C., S. Murray, J. Abramovitz, and A. Hammond (1998), Watersheds of the world: Ecological value and vulnerability-world resour, *Inst. and World Watch Inst, Washington, DC*, p. 164.

Robba, F. (2021), Assessing the potential of seasonal drought forecasts for triggering drought management strategies.

Salvadori, G., C. De Michele, N. T. Kottegoda, and R. Rosso (2007), *Extremes in nature: an approach using copulas*, vol. 56, Springer Science & Business Media.

Santos, J. F., I. Pulido-Calvo, and M. M. Portela (2010), Spatial and temporal variability of droughts in portugal, *Water Resources Research*, *46*(3).

Schmitt, R. J., S. Bizzi, A. Castelletti, and G. Kondolf (2018), Improved trade-offs of hydropower and sand connectivity by strategic dam planning in the mekong, *Nature Sustainability*, *1*(2), 96–104.

Shannon, C. E. (1948), A mathematical theory of communication, *The Bell system technical journal*, *27*(3), 379–423.

Sharma, A. (2000), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1,Äîa strategy for system predictor identification, *Journal of Hydrology*, *239*(1-4), 232–239.

Sheffield, J., and E. F. Wood (2008), Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, ipcc ar4 simulations, *Climate dynamics*, *31*(1), 79–105.

Shukla, S., and A. W. Wood (2008), Use of a standardized runoff index for characterizing hydrologic drought, *Geophysical research letters*, *35*(2).

Siam, M. S., and E. A. Eltahir (2017), Climate change enhances interannual variability of the nile river flow, *Nature Climate Change*, *7*(5), 350–354.

Sienz, F., O. Bothe, and K. Fraedrich (2012), Monitoring and quantifying future climate projections of dryness and wetness extremes: Spi bias, *Hydrology and Earth System Sciences*, *16*(7), 2143–2157, doi: 10.5194/hess-16-2143-2012.

Sklar, M. (1959), Fonctions de repartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris*, *8*, 229–231.

Snieder, E., R. Shakir, and U. Khan (2020), A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models, *Journal of Hydrology*, *583*, 124,299.

Sol'áková, T., C. De Michele, and R. Vezzoli (2014), Comparison between parametric and nonparametric approaches for the calculation of two drought indices: Spi and ssi, *Journal of Hydrologic Engineering*, *19*(9), 04014,010.

Solomatine, D., L. M. See, and R. Abrahart (2009), Data-driven modelling: concepts, approaches and experiences, *Practical hydroinformatics*, pp. 17–30.

Solomon, S., M. Manning, M. Marquis, D. Qin, et al. (2007), *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, vol. 4, Cambridge university press.

Sousa, K. d., A. H. Sparks, W. Ashmall, J. v. Etten, and S. Ø. Solberg (2020), chirps: Api client for the chirps precipitation data in r, *Journal of Open Source Software*.

Sruthi, S., and M. M. Aslam (2015), Agricultural drought analysis using the ndvi and land surface temperature data; a case study of raichur district, *Aquatic Procedia*, *4*, 1258–1264.

Steinemann, A. C., M. J. Hayes, and L. Cavalcanti (2005), Drought indicators and triggers, *Drought and water crises: Science, technology, and management issues*, pp. 71–92.

Taormina, R., S. Galelli, G. Karakaya, and S. Ahipasaoglu (2016), An information theoretic approach to select alternate subsets of predictors for data-driven hydrological models, *Journal of Hydrology*, *542*, 18–34, doi: https://doi.org/10.1016/j.jhydrol.2016.07.045.

Tekuya, M. (2020), The egyptian hydro-hegemony in the nile basin: the quest for changing the status quo, *The Journal of Water Law*, *26*, 2.

Thornthwaite, C. W. (1948), An approach toward a rational classification of climate, *Geographical review*, *38*(1), 55–94.

Trombetta, G. (2020), From individualism to full cooperation: optimal operation of the nile river basin storages under varying levels of cooperation.

Udall, B., and J. Overpeck (2017), The twenty-first century colorado river hot drought and implications for the future, *Water Resources Research*, *53*(3), 2404–2418, doi: https://doi.org/10.1002/2016WR019638.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno (2010), A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, *Journal of climate*, *23*(7), 1696–1718.

Wambua, R. M. (2019), Spatio-temporal characterization of agricultural drought using soil moisture deficit index (smdi) in the upper tana river basin, kenya, *International Journal of Engineering Research andAdvanced Technolog*, *5*(2), 93–106.

Wheeler, K. G., M. Jeuland, J. W. Hall, E. Zagona, and D. Whittington (2020), Understanding and managing new risks on the nile with the grand ethiopian renaissance dam, *Nature communications*, *11*(1), 1–9.

Wilhite, D. A. (2000), Drought as a natural hazard: concepts and definitions.

Wilhite, D. A. (2005), *Drought and water crises: science, technology, and management issues*, Crc Press.

Wilhite, D. A., and M. H. Glantz (1985), Understanding: the drought phenomenon: the role of definitions, *Water international*, *10*(3), 111–120.

Wu, H., M. D. Svoboda, M. J. Hayes, D. A. Wilhite, and F. Wen (2007), Appropriate application of the standardized precipitation index in arid locations and dry seasons, *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *27*(1), 65–79.

Xing, X., C. Yan, Y. Jia, H. Jia, J. Lu, and G. Luo (2020), An effective high spatiotemporal resolution ndvi fusion model based on histogram clustering, *Remote Sensing*, *12*(22), 3774.

Yang, Y., and J. O. Pedersen (1997), A comparative study on feature selection in text categorization, in *Icml*, vol. 97, p. 35, Nashville, TN, USA.

Zaniolo, M. (2020), Feature representation learning in complex water decision making problems.

Zaniolo, M., M. Giuliani, A. F. Castelletti, and M. Pulido-Velazquez (2018), Automatic design of basin-specific drought indexes for highly regulated water systems, *Hydrology and Earth System Sciences*, *22*(4), 2409–2424.

Zhu, X., O. Bothe, and K. Fraedrich (2011), Summer atmospheric bridging between europe and east asia: Influences on drought and wetness on the tibetan plateau, *Quaternary International*, *236*(1), 151–157, doi: https://doi.org/10.1016/j.quaint.2010.06.015, quaternary Paleoenvironmental Change and Landscape Development in Tibet and the Bordering Mountains.

A

Appendix

A.1 Bahr El Ghazal

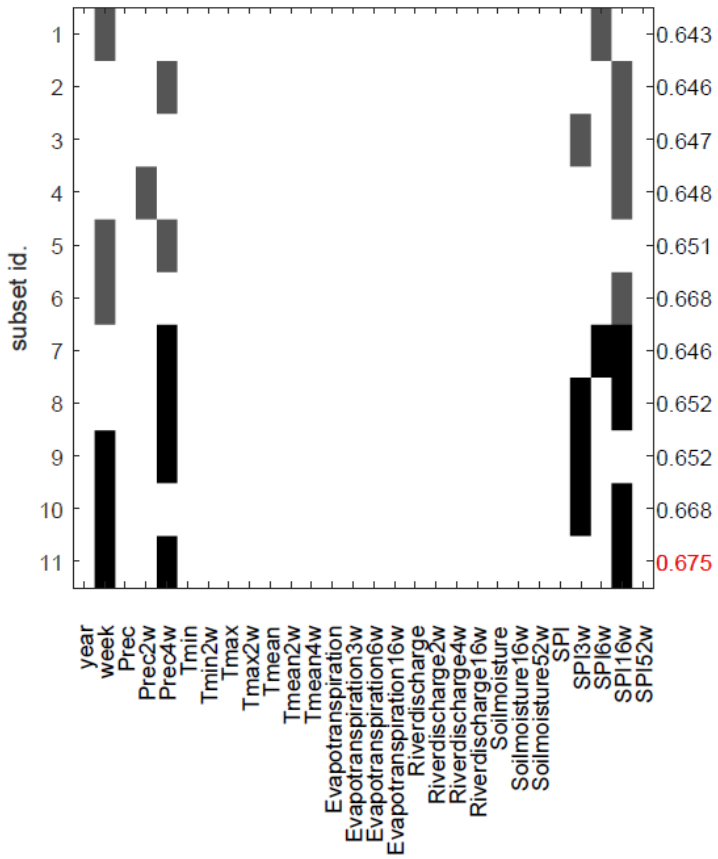


Figure A.1: *Feature selection results for Bahr El Ghazal sub-basin.*

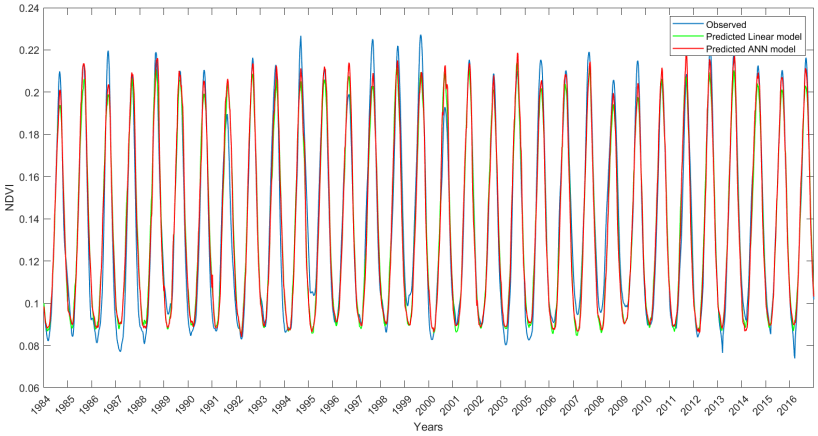


Figure A.2: *Observed and predicted NDVI for the Bahr El Ghazal sub-basin using linear and ANN models.*

A.2 Bahr El Jebel

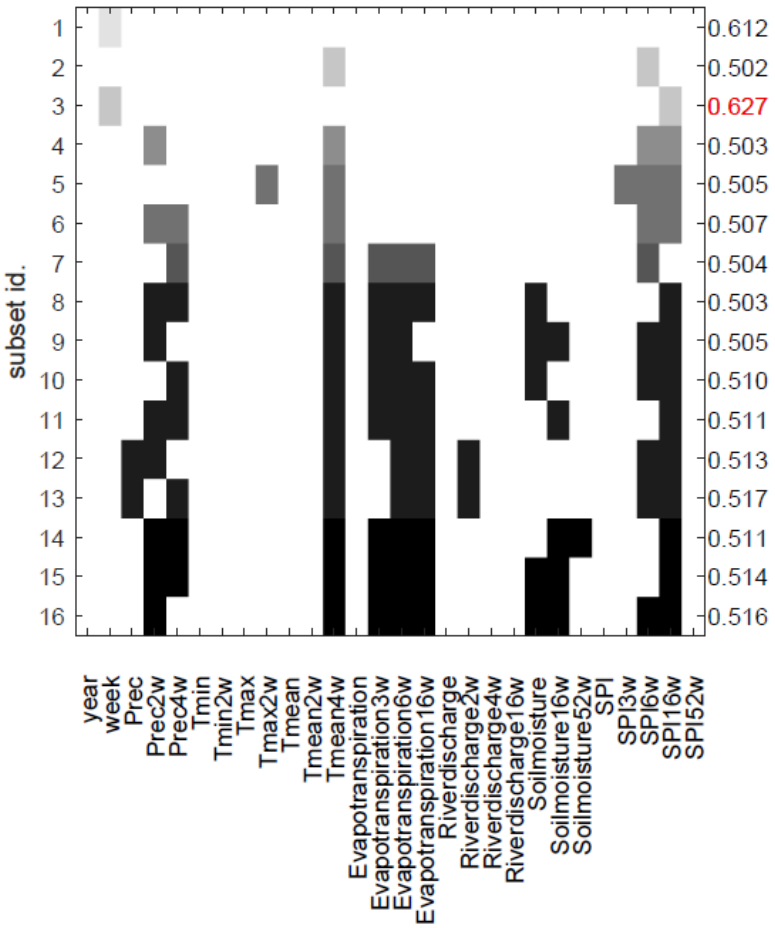


Figure A.3: *Feature selection results for Bahr El Jebel sub-basin.*

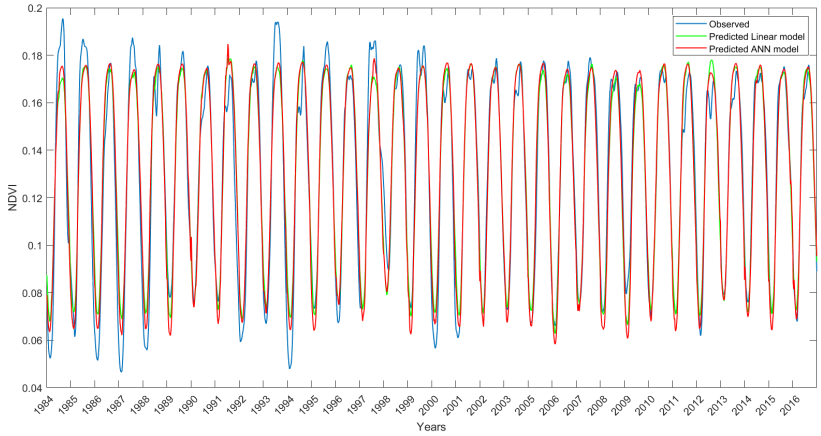


Figure A.4: *Observed and predicted NDVI for the Bahr El Jebel sub-basin using linear and ANN models.*

A.3 Bako Akobbo-Sobat

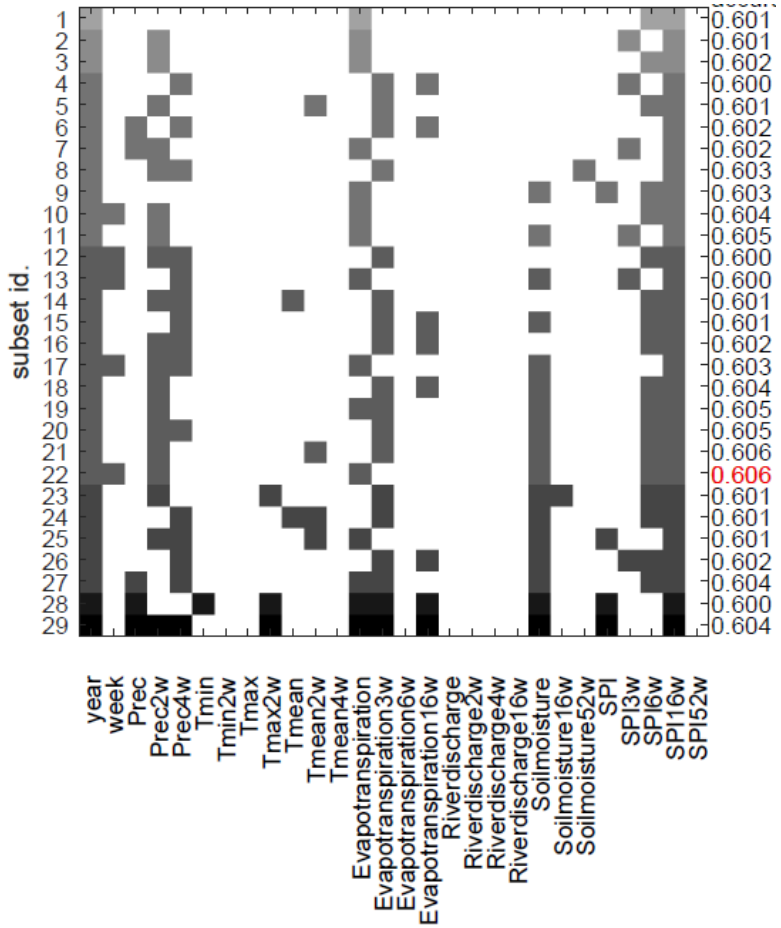


Figure A.5: *Feature selection results for Bako Akobbo-Sobat sub-basin.*

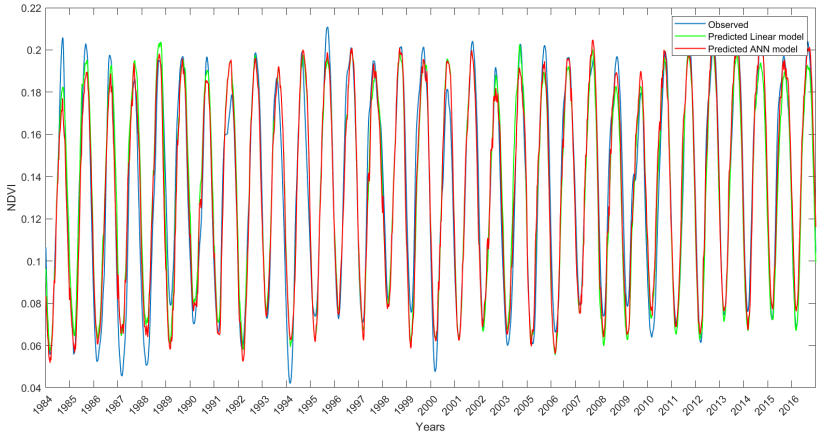


Figure A.6: *Observed and predicted NDVI for the Bako Akobbo-Sobat sub-basin using linear and ANN models.*

A.4 Lake Victoria

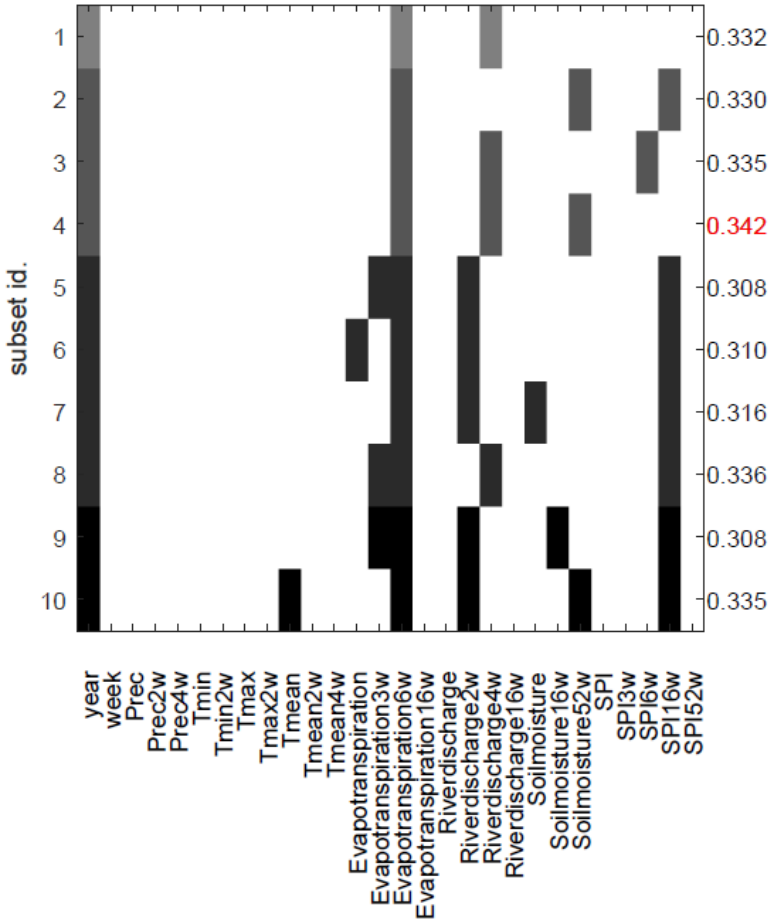


Figure A.7: *Feature selection results for Lake Victoria sub-basin.*

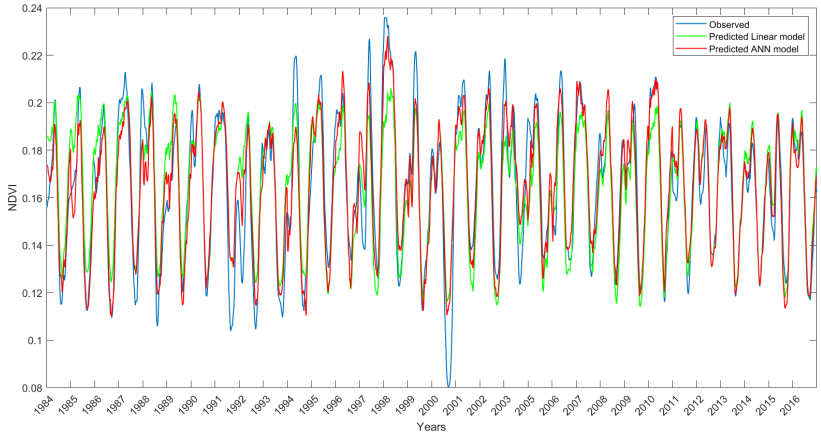


Figure A.8: *Observed and predicted NDVI for the Lake Victoria sub-basin using linear and ANN models.*

A.5 Tekeze Atbara

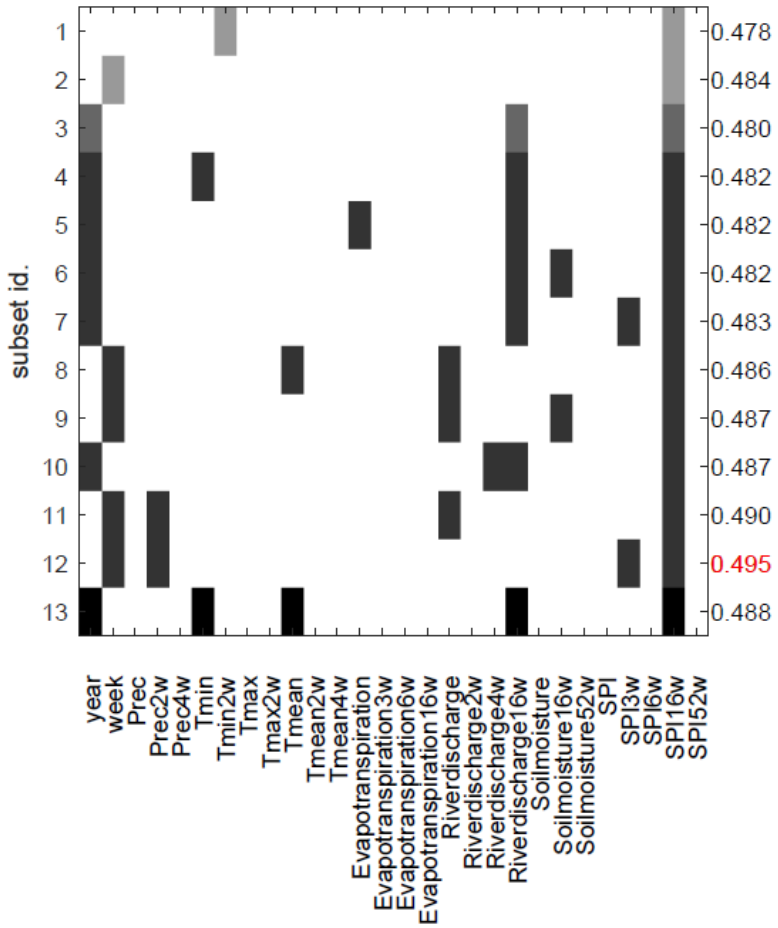


Figure A.9: *Feature selection results for Tekeze Atbara sub-basin.*

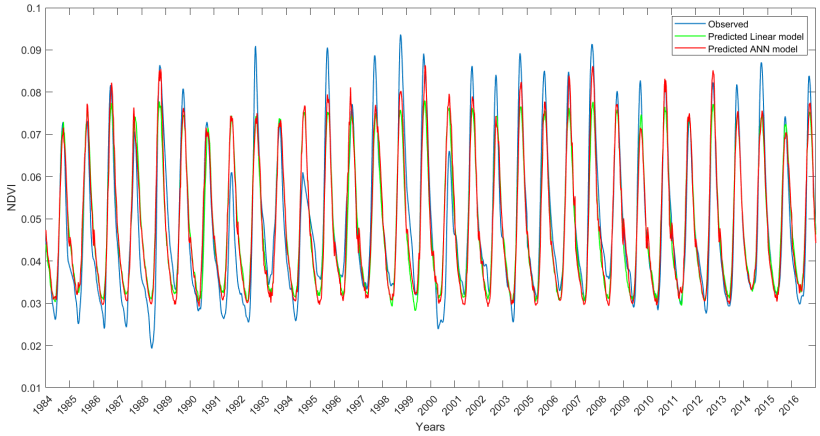


Figure A.10: *Observed and predicted NDVI for the Tekeze Atbara sub-basin using linear and ANN models.*

A.6 Victoria Nile

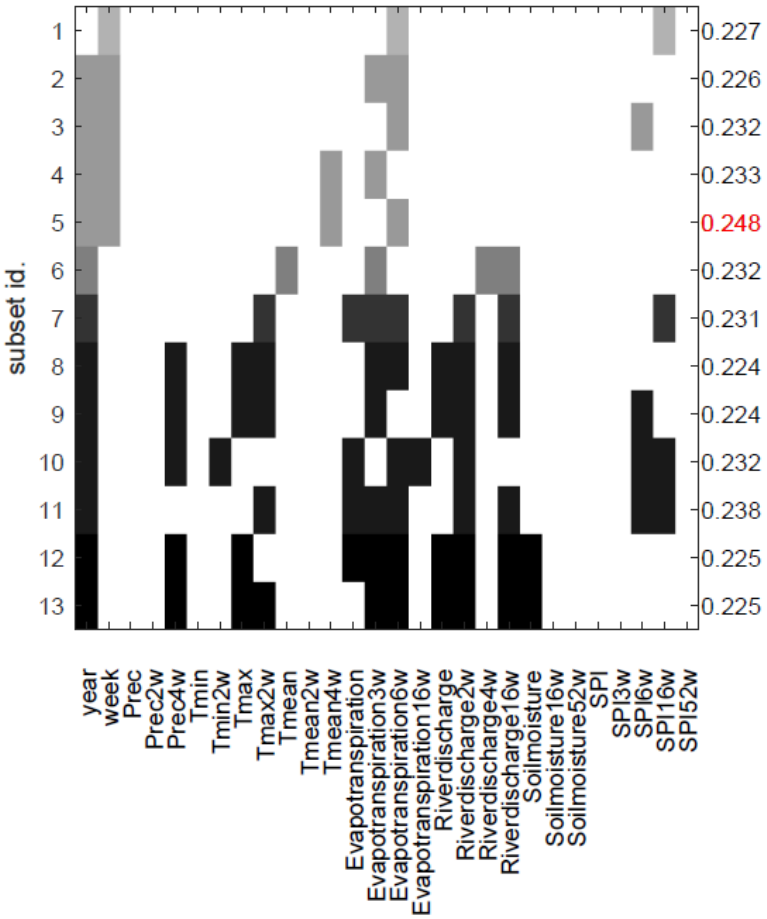


Figure A.11: *Feature selection results for Victoria Nile sub-basin.*

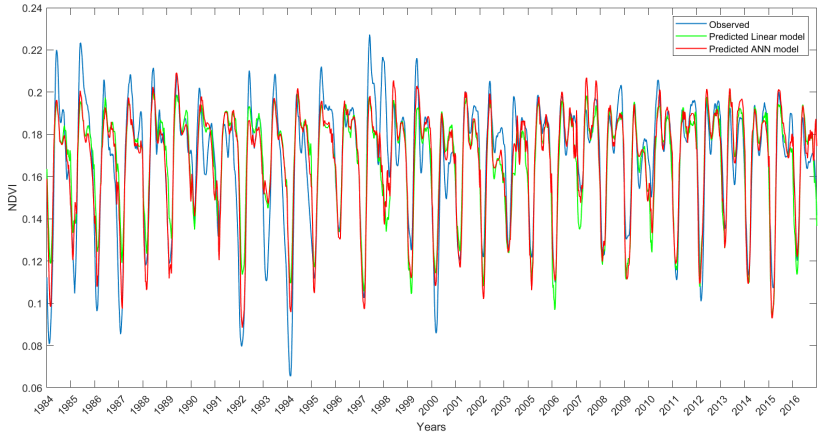


Figure A.12: *Observed and predicted NDVI for the Victoria Nile sub-basin using linear and ANN models.*

A.7 White Nile

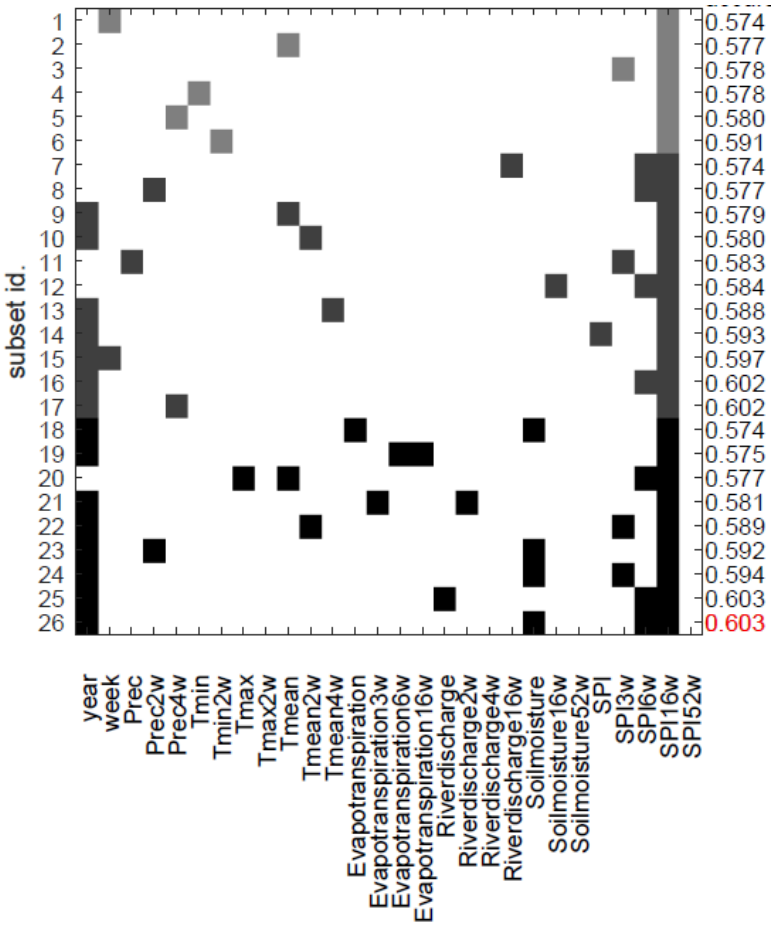


Figure A.13: *Feature selection results for White Nile sub-basin.*

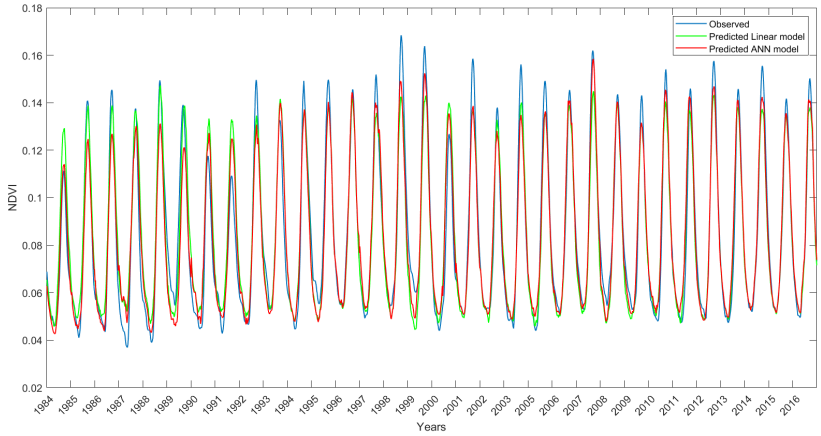


Figure A.14: *Observed and predicted NDVI for the White Nile sub-basin using linear and ANN models.*