Prediction of groundwater levels in a changing climate

Master Thesis

Rebeca Quintero Gonzalez
Supervisor: Jamal Jokar Arsanjani
MSc Geoinformatics - Spring 2021

AALBORG UNIVERSITY
COPENHAGEN

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Jamal Jokar Arsanjani for his insightful feedback and help all throughout this thesis project. Thanks to his support I could overcome many obstacles and I could bring my work to a new level.

Secondly, I would like to thank Mads Robenhagen Mølgaard and Magnus Marius Rohde for all their help in the first stages of this project, especially regarding access to data and all the help provided with the Jupiter dataset.

Finally, I would like to thank Sergio Garcia for all his support during this process.

SUMMARY

Shallow groundwater is defined as the uppermost water table, and is a key resource to human activities and ecosystems. As global mean temperatures continue to rise, regional patterns on precipitation and temperatures will be altered, affecting the groundwater recharge rates and water table elevation, causing negative socio-economic and environmental impacts, and increasing the need to predict the evolution of the water table. Machine Learning (ML) is being increasingly used in forecasting these environmental problems and has been proven to be quite efficient for modelling and predicting changes in the water table, offering better opportunities for management of water resources, and providing adaptation plans for possible events and risks caused by the change in the water table levels.

The aim of this project is to gain insights about the future water level changes based on different climate change scenarios using ML algorithms while addressing the following research questions: 1. How will the water table be affected by climate change in the future based on different Socio Economic Pathways (SSPs)?: 2. Do ML models perform well enough in predicting changes of the groundwater in Denmark? If so, which ML model outperforms for forecasting these changes?

To answer the aforementioned questions, three ML algorithms were used in R: Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Random Forest (RF). The ML models were trained with time-series data of groundwater levels taken at several wells in the Hovedstaden region, for the period 1990-2018. Several independent variables were used to train the models, including different soil parameters, topographical features and climatic variables for the time period and region selected.

Results show that the RF model outperformed the other two, resulting in a higher R-squared (0.50) and Mean Absolute Error (MAE) (1.01 m) and was therefore used for future predictions on the water table for three different climatic scenarios. The model showed sturdy outcomes resulting in successful metrics in regards to the criteria defined by the literature; however, these results should be considered carefully, taking into account the presence of possible signs of overfitting and the limitations in the available data.

The future prediction maps for the different scenarios show little variation in the water table. Nevertheless, predictions show that it will rise slightly, mostly in the order of 0-0.25m, especially during winter, increasing the areas where the water table will be less than 1m deep from the surface in Hovedstaden. These changes should be paid attention to since even slight fluctuations in the water table can have notorious repercussions, and ought to be accounted for and managed.

The work done in this project can be used to visualize areas where the water levels are expected to change, as well as to give an overview of how big the changes will be. This can provide better perspectives when planning and adapting both for climate change and for the impacts that changes in the water table will cause, allowing for decision makers to be ahead of situations where the risks might be high.

Additionally, the approaches and models developed with this thesis could be replicated and applied to different study areas, allowing for the possibility to extend this model to a national level, improving the prevention and adaptation plans in Denmark and providing a more global overview of future water level predictions to more efficiently handle future climate change scenarios.

ABBREVIATION LIST

ANN - Artificial Neural Network
ARMA - AutoRegressive Moving Average
ARIMA - AutoRegressive Integrated Moving Average
AS - Assessment Report
CIMP - Coupled Model Intercomparison Projects Phase
DT - Decision Tree
GIS - Geographic Information Sciences
IDE - Integrated Development Environment
IPCC - Intergovernmental Panel on Climate Change
MAE - Mean Absolute Error
ML - Machine Learning
NOVANA - National Monitoring Assessment Programme for the Aquatic and Terrestrial Environment
PCA - Principal Component Analysis
PC - Principal Component
RCP - Representative Concentration Pathways
RF - Random Forest
RMSE - Root Mean Squared Error
R2 - Coefficient of determination (R squared)
SARIMA - Seasonal ARIMA
SSP - Shared Socioeconomic Pathways
SVM - Support Vector Machine

Table of Contents

List of figures

List of tables

1. INTRODUCTION

Shallow groundwater is defined as the uppermost water table, and is a key resource to human activities and ecosystems (Gleeson et al., 2016). Groundwater is widely used as a source for drinking water (e.g. about 75% of EU inhabitants depend on groundwater for their water supply), it is an important resource for industry and agriculture, it directly affects terrestrial ecosystems by impacting the vegetation's access to water, and it represents an important link in the hydrological cycle, providing the base flow for surface water systems (European Commission, 2021).

Changes in groundwater levels can impact the state of terrestrial and aquatic ecosystems, human health and food provision, and even pose flooding hazards and cause severe droughts. For instance, several studies have proven that both low and high groundwater levels can negatively impact field crops, being the ideal water table height between 1 and 2 m below the surface (Kahlown et al., 2005; Zipper et al., 2015), and high groundwater levels can intensify the risk of flooding, which might be especially hazardous in urban areas (Jankowfsky et al., 2014).

As global mean temperatures continue to rise, regional patterns on precipitation will be altered (Collins et al., 2013) and extreme climatic events will occur with a higher intensity and frequency (Seneviratne et al., 2012). This means that the hydrological cycle will be affected, producing changes in precipitation and evapotranspiration patterns, and altering the periodicity and intensity of climatic events such as storms or droughts (Collins et al., 2013; Wuebbles et al., 2017), causing more frequent and severe droughts and flooding events, and affecting the groundwater recharge rates and table elevation (Bates et al. 2008).

However, these hydrological changes will not be regionally uniform. Different regions have different projections (Collins et al., 2013), and while some countries will suffer from high reductions in the water table levels and more frequent droughts, countries at higher latitudes, such as Denmark, are expected to potentially raise their water table due to the increase in precipitation and subsequent water recharge rates (Woldeamlak et al. 2007). For instance, among other changes, it is expected in Denmark that the water table will rise during the wet season due to the increased precipitation, while summers are expected to become drier due to the increase in temperatures and evapotranspiration (Danish Nature Agency, 2012).

Because of the impact and repercussions that these changes in the water level could have, methodologies that can predict the evolution of the water table are growing in importance (e.g. Lakshamanan et al., 2015; Rolnick et al., 2019). By being able to model and predict future climate change scenarios and their consequences and impact on both the environment and people's lives, more and better adaptation plans, as well as prevention measures, can be made.

Machine learning (ML) is a branch of artificial intelligence that focuses on providing systems with the ability of learning automatically from data, improving their decision-making and predictive accuracy over time without being explicitly programmed to do so (IBM Cloud Education, 2020). ML is increasing in popularity among all fields, and it has been a main component of spatial analyses in GIS, being widely used for classification of spatial components, modelling of spatial varying relationships and predicting changes over time (Singh, 2019).

This increase in popularity can mostly be attributed to the advantages of these data-driven models in mitigating the difficulties associated with physics-based models (Fahimi et al., 2017; Bowes et al., 2019); this is, physical relationships and parameters do not need to be defined. ML algorithms only need to process the data, and will find and approximate the relationships between model inputs and outputs through an iterative learning process (Solomatine & Ostfeld, 2008). Moreover, availability to data is improving by the day thanks to the internet, sensors, and improvements in data collection, making ML algorithms an up-and-coming and more reliable choice. On the other hand, although ML algorithms are not intended to replace physics-based models, in many cases they have been found to perform better (Mohanty et al., 2013), being a useful and resourceful tool for predicting future changes.

Because of the possibilities that ML brings in planning and preparing for future scenarios, these approaches are playing a crucial role in climate change prevention and adaptation. ML's relevance is thus rising, being increasingly used in modelling the impact of climate change in many fields and from different perspectives (e.g. Lakshamanan et al., 2015; Rolnick et al., 2019), including the impact that climate change will have on the shallow groundwater.

Therefore, the aim of this project is to investigate the future variation of water level under climate change and explore the possibilities of Machine learning algorithms to groundwater level prediction. Specifically, the goal is to study the potential changes on the shallow water table in Denmark, where current predictions state that groundwater levels will rise due to the increased precipitation that is expected in countries at such latitudes.

1.1. Problem statement

Current predictions expect Denmark to receive more precipitation in the future due to climate change. With higher precipitation and a rise of the temperatures, as well as an increase in the number of sporadic events of very heavy precipitation, the hydrological cycle is expected to be affected by climate change, and local events of flooding, as well as drier soils in the summer, are within the predictions made by the Danish Nature Agency for Denmark. Thus, the water table is expected to suffer changes due to climate change, with increased risks of both floods and droughts.

ML is increasing in popularity as more and more data is available and easily accessible, and this methodology is being increasingly used in forecasting environmental problems and changes such as those caused by climate change, including changes in the groundwater levels. Moreover, ML has been proven to be quite efficient for modelling and forecasting changes in hydrology settings, and predictions of changes in the water table will offer better opportunities for management water resources, planning and provide adaptation plans for possible events and risks caused by the change in the water table levels.

Thus, the aim of this project is to gain insights about the future water level changes based on different climate change scenarios using ML algorithms while addressing the following research questions:

- How will the water table be affected by climate change in the future based on different Socio Economic Pathways (SSPs)?

- Do ML models perform well enough in predicting changes of the groundwater in Denmark? If so, which ML model outperforms for forecasting these changes?

2. BACKGROUND

The following section includes background and description on the current trends on climate change scenarios and machine learning applied to hydrology, along with a short description of some projects in the same topic as state of the art.

2.1. Climate Change Scenarios

Global Climate Models simulate the physics, chemistry and biology of the Earth and are used to generate climate projections, being one of the primary means used to understand future climate changes. These models are constantly updated to improve their resolution, and approximately 100 of them are used by the Intergovernmental Panel on Climate Change (IPCC) for generating their assessment reports on climate change and for the resulting models (Hausfather, 2019).

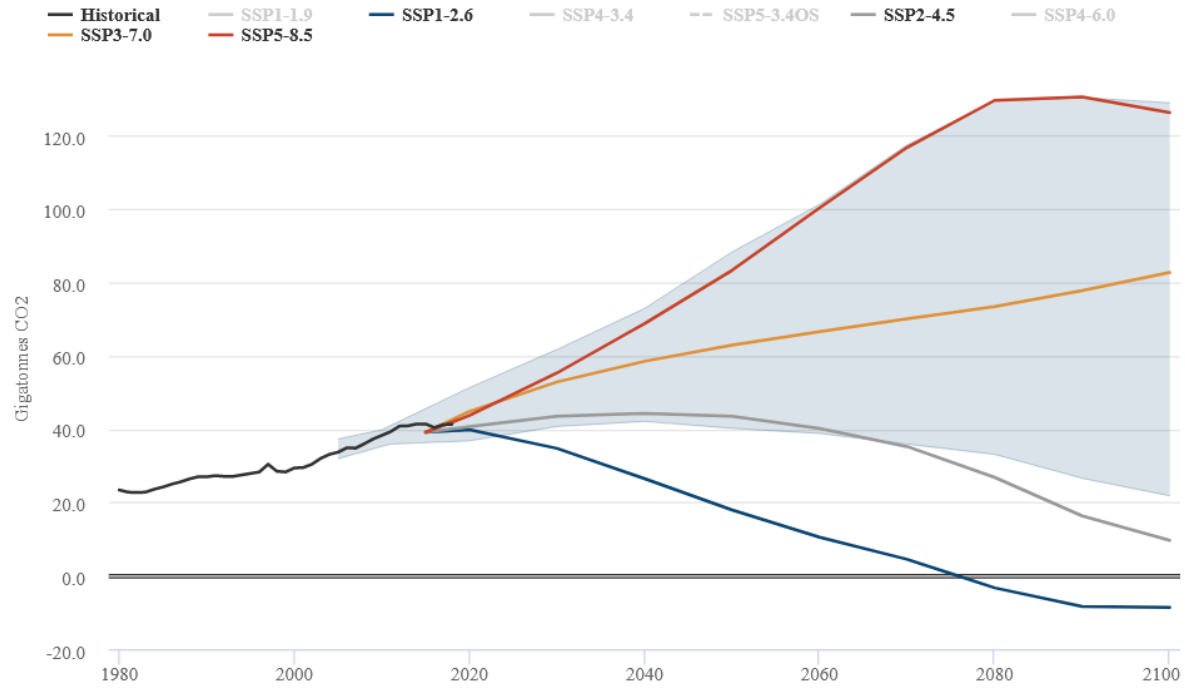


Figure 1. CO_2 emissions in the four available CMIP6 scenarios. Source: Hausfather (2019)

The latest future projections are part of the Coupled Model Intercomparison Projects Phase 6 (CMIP6), which is still being updated and will be used for the IPCC sixth assessment report (AR6). The release of these latest projections has also served to develop a set of emission scenarios, which are driven by different socioeconomic assumptions and greenhouse gas emissions, called the Shared Socioeconomic Pathways (SSPs). These scenarios provide a range of climate change outcomes up to 2100. These comprise eight scenarios in total, and four of them are available at WorldClim:

- SSP1-2.6 — low emission scenario, assumes that policies will be implemented to limit warming to below 2°C and shows a mean estimate of 2°C warming
- SSP2-4.5 — scenario where efforts are made to limit warming to 3°C by 2100, with a slow decline of CO_2 emissions.
- SSP3-7.0 — newly added CMIP6. It is seen as a "middle of the road" scenario showing 4.1oC of warming.
- SSP5-8.5 — high emission scenario with a mean warming of 5°C. It is often regarded as a "business as usual" scenario.

These four SSPs are available in WorldClim at three different spatial resolutions, 2.5, 5 and 10 minutes, which are accomplished by WorldClim by processes of downscaling and calibration (WorldClim, 2020). For this project, the best resolution available of 2.5 minutes (4.5km) was used. The resolution for these models is usually 30 seconds (1km), but due to ongoing testing of the CMIP6, the release date of the data has been pushed back (WorldClim, 2020).

Additionally, four temporal resolutions of the SPPs are available: 2021-2040, 2041-2060, 2061-2080 and 2081-2100. For this project, the temporal resolution 2060-2100 was selected along with the SSPs that follow best current EU and Danish plans and legislation regarding gas emissions. Thus, the selected SPPs were SSP2-4.5, SSP3-7.0 and SSP5-8.5.

2.2. Machine Learning for groundwater prediction

With the development of information science and technology, many modelling techniques based on physical principles have been developed and used to explore and understand groundwater dynamics, and to provide quantitative assessments of groundwater resources (Singh, 2014). These physical models have been widely developed and applied to simulating groundwater dynamics, improving the understanding of hydrologic and water resource systems (Markstrom et al., 2008).

The difficulties of physical models arise from the fact that it is necessary to develop and solve fluid mechanics and thermodynamics equations, applying detailed boundary conditions, and to describe the dynamics of the hydrological system in order to obtain the input-output relationship. However, solutions for physically based models often require simplifying assumptions, because physiographic and geomorphic characteristics of most hydrologic systems are complicated, and have a large degree of uncertainty in the boundary conditions (Brutsaert 2005). Additionally, these models also possess other limitations, such as the requirements on the accuracy of the data or the limitations on the computation resources. Physical models require a large quantity of accurate data, which can never be ascertained with absolute accuracy (e.g., the physical properties of aquifers) (Chen et al., 2020).

To overcome these limitations that physical models present, more data-driven models based on ML approaches are being studied and applied by researchers as an alternative to physical models (Solomatine & Shrestha, 2009), and more ML models are being developed for forecasting in hydrologic research fields (e.g. LeCunt et al., 2015; Guio Blanco et al., 2018). In the ML approach, physical relationships and parameters do not need to be defined. ML models are data-driven, meaning that ML algorithms only need to process the data, and will find and approximate the relationships between the macro-description of the behaviour of a system (model output) and the behaviour of the constituents of this system (model inputs) through an iterative learning process (Guergachi & Boskovic 2008; Solomatine & Ostfeld, 2008).

Overall, ML models have given very promising results for modelling hydrological systems and dynamics, and for forecasting groundwater levels, outperforming in many cases the results from physical models (Mohanty et al., 2013). Of the many algorithms researched, Random Forest is being widely used for groundwater modelling, giving very robust results (e.g. Gulo Blanco et al., 2017; Kenda et al., 2018; Koch et al., 2019). Different algorithms within Neural Networks are also widely used for groundwater table forecasting (e.g. Maier & Dandy, 2000; Bowes et al., 2019). Finally, SVMs are also a popular choice and are obtaining good results (Hussein et al., 2020), and usually outperforming other models (Chen et al., 2020).

For these reasons, this project studies the performance of three ML algorithms when trained with historical groundwater measurements and different geological, topographic, and climatic variables, in order to forecast changes on the depth to shallow groundwater. The selected algorithms are Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) based on the information obtained from the literature reviewed. Further descriptions of these algorithms are provided in section "4.3. Machine Learning Algorithms" of this report.

2.3. State of the art

Modelling of the shallow water table at high spatial resolution using random forest
(Koch et al., 2019)

This paper describes a study on Denmark focused on a large part of the Jutland Peninsula. It had four main objectives: 1. train a RF model to predict the depth to the shallow water table at a high spatial resolution, 2. Outline a simple method that unifies a physically based model and ML, 3. Conduct a sensitivity analysis to make better interpretations of the RF model prediction, and 4. Assess the uncertainty related to the RF model.

The main aim of the project was to model the depth to the shallow groundwater at a 50 m spatial resolution by training a RF model. The authors decided to disregard the temporal variability of

groundwater dynamics close to the surface, and decided to model an extreme winter event that characterizes a minimum depth, when the aquifer systems are replenished after several months of typically rainfall and low evapotranspiration.

For this purpose, the authors extracted groundwater head observations covering the entire model domain, from the national database, Jupiter. They filtered the observations to a maximum filter depth of 10m, keeping the measurements for the shallow groundwater table, for a 20-year period between 1998 and 2017. Several filters to exclude outliers were used.

Wells with more than five observations were grouped according to their hydrogeological setting, and their standard deviations were studied to define sinus curves amplitudes for each group. This was done because long time series of groundwater head measurements are scarce, and these sinus curves amplitudes could be applied to low-frequency sampled wells, so any given observation could be transformed to an expected high water table corresponding to an extreme wintertime event. The minimum and maximum of the sinus curves was set to February and August, respectively. With this method, at wells with more than five observations the recorded minimum depth was selected for training the model, while at wells with fewer observations, the predefined sinus model was applied to transform the measurements to an expected wintertime minimum. However, this model neglects seasonal and inter-annual variability.

RF was the selected algorithm due to its wide use for geophysical and environmental contexts, for its good performance in previous research, and because it undergoes a sensitivity analysis of the covariates, allowing to assess the importance of each variable to the model. An "out of the bag" validation method was used for the training, and the accuracy of the model was interpreted using the R2, RMSE (m) and MAE (m) metrics. The final metrics obtained were R2= 0.56, RMSE = 1.13 and MAE = 0.76, meaning that more than half of the variance contained in the training data was captured by the RF model, and the MAE accounted for 76 cm. Given these results, the authors deemed the model to have an overall good performance.

This RF model was then used to predict the depth to the shallow water table for an extreme wintertime event. The mean depth to the groundwater obtained with the model constituted 1.9m for the entire modelling area, and around 29% of the domain showed a depth to the groundwater of less than 1m, while approximately 14% of the area was characterized by a depth of less than 50cm to the groundwater. According to the authors, these results make apparent the severity of the risk of groundwater-induced floods.

On the other hand, this RF model was trained for a single event, disregarding the temporal dynamics of the shallow groundwater system. Since the model was designed as a simple screening tool, this can be considered an advantage, but it does not account for much of the complexity of the system, which is a clear shortcoming of the proposed method.

As a conclusion, the metric scores obtained for the model were overall very satisfying and in the range of what can be considered very acceptable in groundwater flow modelling (Henriksen et al., 2003). Thus, RF appears to be a good choice for modelling complex, non-linear variables, with an accuracy that would be difficult to obtain with physical based models. The model also showed a bias on wells located in clayley moraine sediments, which are very heterogeneous landscapes, but the current available national hydrogeological data does not require enough resolution to resolve these heterogeneities adequately. Nevertheless, the RF showed satisfactory results and it is deemed as a valuable option for modelling and forecasting the depth to the shallow water table.

HIP - Hydrological Information and Forecasting System

(Styrelsen for Dataforsyning og Effektivisering, 2021a)

HIP is a Danish project that exhibits free public hydrological data, model calculations and forecasts for the future to support the work on climate adaptation and water management and planning. It is meant to provide easy access to free public data on terrestrial hydrological conditions to support climate adaptation, water management, emergency assessments and other planning where water plays a role.

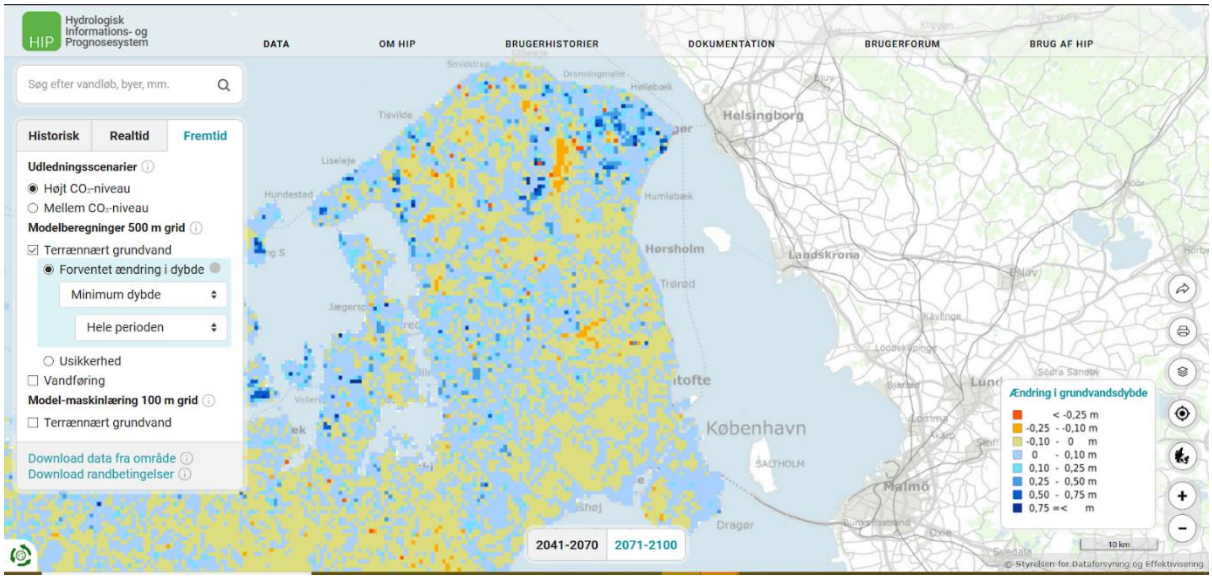


Figure 2. Example of the HIP web-based interface.

HIP is presented as a web-based interface (Styrelsen for Dataforsyning og Efektivisering, 2021b) showing data and model calculations on terrestrial hydrological conditions. It includes current and historical measurements of terrestrial groundwater, watercourses and sea levels from municipalities, regions, and the state. There is also available future forecasting on groundwater levels and water flow with a spatial resolution of 500m, which has been

downscaled to a spatial resolution of 100m and made available on the web-map as well for visualization. Forecasts are available for a near (2041-2070) and distant (2071-2100) future, and are based on two climate scenarios: a medium CO_2 emissions scenario (RCP 4.5) and a high CO_2 emissions scenario (RCP 8.5).

The model calculations to obtain the data, both historical and future predictions, have been made by GEUS using the National Water Resource Model (DK-model), which is a nationwide 3-dimensional physically based computer model of the freshwater cycle, and the downscaling of the data has been made with ML models.

The DK-model is used today to calculate water cycles and groundwater dynamics as well as climate-related changes. The model uses data on the entire hydrological cycle to calculate the depth to the groundwater, the soil water content in the root zone, water flow in watercourses, infiltration, and flow in groundwater. It was originally developed between 1996-2003 for the Ministry of the Environment. Nowadays, an updated version is used in connection with water plans for quantitative condition assessment, nitrate calculations in groundwater, monitoring programmes (e.g. sea load, mapping models and groundwater protection) as well as in a number of ongoing research projects. The main focus of the model has been the calculation of the models in a 500m grid, and model calibration and validation has focused on the deeper groundwater.

Five climate models are used for emission scenario RCP 4.5 and 17 climate models for emission scenario RCP 8.5; thus the latter allows for more robust results. Projected sea level rises have been implemented based on the Danish Meteorological Institute Climate Atlas (Climate Atlas, 2020) for the two climate scenarios RCP 8.5 and RCP4.5 for the period 2041-2070 and 2071-2100.

Daily values have been extracted and have been used to calculate statistically processed products for the depth to groundwater close to the ground, water flow in watercourses and the water content of the soil for three 30-year periods, respectively: reference period 1990-2019, near future 2041-2070 and distant future 2071-2100. By basing the statistics on 30-year periods, the variations that can occur over time in especially extreme values are taken into account. This provides a robust basis for calculating e.g. mean, minimum and maximum values, T-year events (T = 2-100 years), etc. Statistics have been extracted for 48,653 watercourse calculation points (Q-points) and for all grid cells in 500 meters.

The model calculations are available as the median change in depth to terrestrial groundwater and the earth's water content, as well as the median climate factor for water flow in watercourses. The median value here describes the predominant direction of change, while the standard deviation describes the typical spread for the ensemble of climate models.

As a summary, the model calculations provide a nationwide overview of the expected climate-projected median change in depth to the terrestrial groundwater in the periods 2041-2070 and 2071-2100 in a resolution of 500x500 meters (downscaled to 100x100m). The model calculations can be used to screen how large a change in depth is expected for terrestrial groundwater in the future. The climate-projected changes in depth to terrestrial groundwater are available as the median change for selected statistically processed products.

Soil water status and water table depth modelling using electromagnetic surveys for precision irrigation scheduling

(Hedley et al., 2013)

This paper is a study on a sand plain field with a high fluctuating water table, both spatially and temporally. The aim of the study was to develop a spatio-temporal method to predict soil water status and water table depth, and to describe spatio-temporal variation of soil moisture and water table depth by using hourly updated soil moisture in the first 50cm of soil.

The authors used wireless sensors to obtain hourly measurements of soil moisture and water table depth at several data loggers. The time series obtained with this methodology were then used to build a spatio-temporal model using ML learning in order to predict future soil water status and water table depth. The selected models were Multiple Linear Regression (MLM) and Random Forest (RF).

Several topographic and geological variables were used for training the models, such as the Saga Wetness Index (SWI), Topographic Wetness Index (TWI) or a DEM. Additionally, rainfall data for the period covered by the time-series was also used in the training of the models, accounting for any rain events that could affect the groundwater levels and soil moisture. Finally, and due to the small size of the dataset, the authors decided to not split the data into training and validation sets.

After training both models, it was found that RF gives improved predictions compared to the MLM model, obtaining a very good performance for both the water table depth (R^2 = 0.91; RMSE = 7.17 cm) and the soil moisture (R^2 = 0.94; RMSE = 0.03 m^3m^{-3}). The authors conclude that the predictions made by the RF model seem to follow rather well the natural trends and patterns of the soil moisture and the water table in the selected study area. They also argue that spatial patterns are inconsistent as the soils dry out, but are temporally more stable on dry soils, which also affects the predictions. Therefore, model predictions improve for drier soils, while more samples are required in wetter conditions to develop prediction models.

Finally, the authors conclude that the RF approach was found to be more accurate than a multiple linear regression modelling (MLM) approach for spatio-temporal modelling of water table depth and soil moisture, most likely because RF implements a more thorough interrogation of the data, and gives very robust results. Further research should focus on improving the models and test their robustness over longer periods of time.

3. METHODOLOGICAL FRAMEWORK

In the following section, both the study area and the data utilized for this project are described and explained, including sources and the methodology used for pre-processing the data for further use in the thesis.

3.1. Study area

Denmark is a small country with 43.000 km^2. Basically all of the landscape in Denmark can be labelled as cultural, with barely no pristine nature areas left, since most of the land has been altered. Specifically, the study area selected, Hovedstaden, comprises 2,568 km², of which approximately 43% is agricultural, 39% is urban and less than 20% is natural (forest, wetlands, etc.) (Statistics Denmark, 2021).

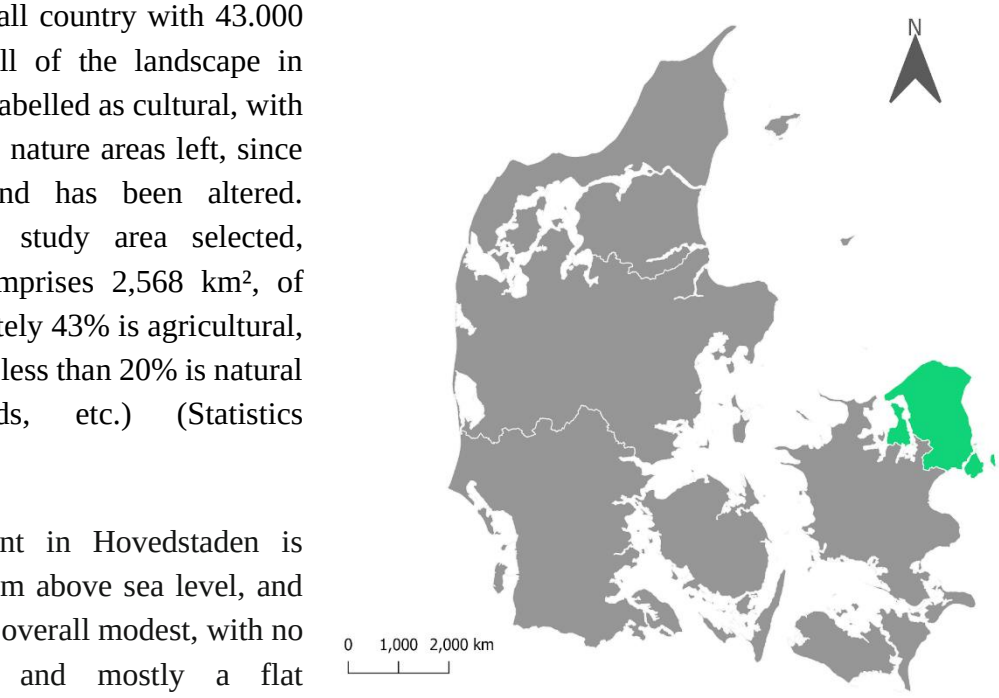


Figure 3. Map of Denmark with the study area resalted in green.

The highest point in Hovedstaden is approximately 96m above sea level, and the topography is overall modest, with no high elevations and mostly a flat landscape that has been highly modified by glaciers from the Quaternary. Thus, the uppermost sediments are predominantly sandy and clayey tills and sandy or gravelly meltwater deposits that allow groundwater recharge in most areas (Jorgensen & Stockmarr, 2009).

Streams are mainly groundwater-fed and relatively small when compared to other European rivers. The climate in the region is coastal temperate, with average annual precipitation of approximately 700mm, and a mean annual temperature of approximately 7°C (Worldbank, 2021). Due to climate change, the mean annual temperature has increased almost 1.5°C during the last century, and annual precipitation has increased 15% (100mm) since 1874, when records began. On the other hand, there are also fewer days of snow cover, with longer warmer seasons and higher rates of evapotranspiration (OECD, 2013).

According to the OECD (2013), the projected impacts of climate change in the region show a rise in annual temperatures of 3-5°C depending on the emissions scenario, leading to fewer days with frost and snow cover. Precipitation will increase 10-40% in winter, and will be

reduced in the summer in the order of 10% to 25%, with a clear tendency towards more episodes of extreme precipitation that will yield 20-30% more water than today. These changes in precipitation and temperature averages will lead to a reduced formation of groundwater in summer and an increased formation the rest of the year, which will affect the use of groundwater for drinking water or irrigation, and an increased risk of pollution. Danish water supply relies almost entirely on unpolluted groundwater, so the aforementioned repercussions might lead to limitations on water extraction (Danish Nature Agency, 2012).

It is a constated fact that shallow groundwater systems are and will be challenged by climate change. Previous studies have predicted a rise of groundwater levels by up to 1.5m for a 100-year event relative to present average conditions (Kidmose et al., 2013), and climate change will bring changes of at least 0.5m in the water table for 26% of Denmark (Henriksen et al., 2012). As stated before, changes in the groundwater table levels affect crops and ecosystems, might limit water extraction and might increase the risk of both local floods and pollution infiltrations to the groundwater (Danish Ministry of the Environment, 2021). And even though Denmark has abundant groundwater resources, some regions are experiencing pressure on groundwater due to rising temperatures and evapotranspiration (Statistics Denmark, 2017), and the need for irrigation systems might increase in the future. Therefore, the main concern of these changes on the groundwater table on the region are on the impacts it will cause on water supply, from drinking water supply to irrigation issues, while flooding risks are seen as very local hazards that will be mainly caused by heavy precipitation and cloudbursts along with coastal floods from sea level rise (Jebens et al., 2016).

All these problems call for comprehensive modelling tools that can support environmental decision making aiming at tackling current and future challenges related to shallow groundwater.

The region of Hovedstaden was selected due to computational constraints, since the databases were too large for a broader analysis.

3.2. Dependent variable: depth to shallow water table

The depth to water table measurements that were used for this project as the dependent variable are part of a dataset provided by GEUS, called “Jupiter”. Jupiter is a public national well database with data on groundwater, drinking water, raw materials, environmental and geotechnical data (GEUS, 2021a). In Denmark, water supply data are reported to the Jupiter database, so the database contains data collected from different companies and organisations in Denmark (Miljøstyrelsen, 2020).

The database contains information about more than 280,000 wells, including:

- technical structure of the well
- geographical location
- administrative information
- geological description
- water level measurements
- groundwater chemical tests and analyses

Additionally, the database also contains information about more than 35,000 water abstraction plants (waterworks, irrigation systems, etc.), including:

- geographical location
- administrative information
- drinking water chemical tests and analyses
- abstracted water volumes
- permits for water abstraction

This database can be accessed and downloaded as a Postgres database (among other formats) containing more than 100 tables on the different measurements and administrative data and metadata on all the wells (more info at: GEUS, 2021b).

For this project, the dataset for Hovedstaden was used (available at: GEUS 2021c). It was downloaded as a Postgres database (backup Postgres file) and processed using PgAdmin. The measurements of the depth to the water table are available in the "watlevel" table of the database, in a column named "watlevgrsu", and it shows the calculation of the depth to the groundwater, measured as metres under the surface. This column, along with the "timeofmeas", which accounts for the date in which the measurements were collected, were selected from this table.

On the other hand, the coordinates for the location of the wells ("xutm" and "yutm") and the depth of the well ("drilldepth") were selected from the table "borehole". Then, all the selected columns were merged in a new table by using the id of each of the wells ("boreholeid"). Finally, the resulting table containing all the necessary data was connected to QGIS for further processing.

The database was connected to QGIS, where the points were clipped using a shapefile with a polygon for the outline of the Hovedstaden region, excluding points outside. This was mainly due to the high number of points to process, as it would take a lot of computational effort and time. After this, the dataset was exported as a shapefile and further pre-processed in R.

Once loaded in R, the measurements were given a proper time format and filtered to a time period between 1990 and 2018. Wells with a filter depth deeper than 10m were excluded, keeping only measurements for the shallow water table. Outliers were also removed by

dropping out measurements deeper than the filter depth, and a maximum of 5m over the surface was set too. Finally, for wells with several observations in the same month, these daily observations were summarized and transformed into monthly observations. Once this pre-processing was done, close to 1200 wells remained, covering more than 10000 measurements of ground water levels in total, and were then used for training and testing the different ML models.

It is important to note that the number of observations varies greatly from one well to another, some wells have very complete time series while others have fewer measurements of the depth to the water table. Additionally, the available time series are very irregular, as the time frames between measurements vary greatly, from monthly to annual measurements depending on the well, and even changing abruptly in the same well (fx. from monthly measurements to measurements every other month).

3.3. Independent variables

In total, 27 variables were acquired for this project, although only 20 of them were used to train the final models (see table 1).

The data used for this project were chosen by following the recommendations from the literature (Hedley et al., 2013; Hengl et al., 2018; Meyer, 2018; Koch et al., 2019). Some of the variables could be obtained from different sources, while others had to be calculated.

The clay content was developed by Adhikari et al. (2013) and the dataset is composed of four tif files, each showing a different depth level of the soil. Like this, clay 1-4 represents the content of clay as a percentage in the first 0-30, 30-60, 60-100 and 100-200 cm of soil, respectively.

The depth to clay was calculated in R by using available soil types and their depth, provided by Geo. Thus, the depth to the clay layers of soil was used to obtain a single layer with the total depth to the first layer of clay, in metres.

The soil drainage class was developed by Møller et al. (2017). It consists of a tif raster file with categorical values from 1 to 5 depending on the drainage class of the soil, being 1 very well-drained soils, while 5 is very poorly drained soils.

The soil type was obtained from GEUS, and it is a shapefile containing the different types of soils as polygons. This shapefile was transformed to raster in R. More information about the different soil categories can be read at the source (GEUS, 2021d).

The horizontal and vertical distance to water bodies were both calculated in SAGA by using both the Digital Elevation Model (DEM) and the water bodies files. Additionally, the topographic wetness index (TWI), flow accumulation, slope and incoming solar radiation for each month of the year were also calculated in SAGA by using the DEM.

Table 1. Data collected for the project. In grey the ones that were not used for training/testing the final ML models.

Type/group	**Variable**	**Type**	**Resolution**	**Source**
Geology	Clay content (1-4)	Continuous	30 m	Adhikari et al. (2013)
	Depth to clay	Continuous	30m	-
	Soil drainage class	Categorical	30m	Møller et al. (2017)
	Soil type	Categorical	N/A	GEUS
Topography	DEM	Continuous	25m	Copernicus
	Topographic wetness index	Continuous	25m	-
	Flow accumulation	Continuous	25m	-
	Slope	Continuous	25m	-
	Incoming solar radiation	Continuous	25m	-
Water	Horizontal distance to nearest waterbody	Continuous	25m	-
	Vertical distance to nearest water body	Continuous	25m	-
	Water body (lakes, streams, coastline)	Categorical		Koch et al. (2019)
	Sea level	Continuous	N/A	NOAA
Land cover	Corine	Categorical	100m	Copernicus
	Imperviousness	Continuous	20m	Copernicus
Bioclimatic variables (monthly historical data)	Precipitation	Continuous	4.5km	WorldClim
	Minimum temperature	Continuous	4.5km	
	Maximum temperature	Continuous	4.5km	
	Average temperature	Continuous	4.5km	
Coordinates	xutm	Continuous	25m	-
	yutm	Continuous	25m	-
Bioclimatic variables - Future projections	Precipitation	Continuous	4.5km	WorldClim
	Average temperature	Continuous	4.5km	

Finally, the DEM, imperviousness and Corine land cover were all obtained from Copernicus as raster files, while all the climatic variables, both historical (Harris et al., 2014) and future predictions (Fick & Hijmans, 2017), were obtained from WorldClim. The reason to choose this historical climate data was the available temporal resolution, as no other datasets were found that covered the period selected for this project in monthly measurements. Similarly, the Corine land cover and the imperviousness from Copernicus also have a rather suitable temporal scale, covering several years during the selected period, and thus changes on both land cover and imperviousness could be accounted for.

Further pre-processing of the data was all performed in RStudio. All the data was reprojected to a Danish projection system (EPSG:25832) and resampled to a resolution of 25m by the nearest neighbour method. This method ensures that no new values are created when resampling the data, which is especially important for categorical variables like the land cover, as new values do not account for any of the categories established for this dataset. Finally, the data was all clipped to the extent of the study area and exported as tif files that were later on used for training the models and for the predictions.

4. METHODS

The following section comprises an explanation of all the tools and methodologies explored and utilized for the development of this thesis, adding some background for their better understanding, and explaining the choices and steps followed for its implementation.

4.1. Utilized tools

4.1.1. PostgreSQL

PostgreSQL is an open-source object-relational database system that uses the Structured Query Language (SQL) combined with many features that allow to store and scale many types of data workloads. It is part of the Postgres project at the University of California at Berkeley and has more than 30 years of active development on the core platform (PostgreSQL, 2021).

For this project, pgAdmin, which is an Open Source administration and development platform for PostgreSQL, was used to load the Jupiter database, select the necessary data such as coordinates or water level measurements, and export it to QGIS where further processing was performed.

4.1.2. QGIS

QGIS is an open source and commonly used Geographic Information System (GIS) desktop mapping software. It provides many tools for processing, analysing, and visualizing geographical data. In this project, QGIS was used to connect to the PostgreSQL database and for minor pre-processing of the wells dataset, as well as for visualization purposes. The Layout tool in QGIS was also used to produce the resulting maps showcased in this thesis.

4.1.3. RStudio

RStudio is an open-source Integrated Development Environment (IDE) for R. It is mainly used for data science and provides a graphic interface to R, making it more user friendly and easy to use, adding many features, and equipping the user with tools for plotting, history, debugging and workspace management (RStudio, 2021).

R is a scripting language developed by Ross Ihaka and Robert Gentleman in 1995, first developed to teach intro statistics (Data Carpentry, 2019). It is a widely used language for statistical computing and graphics, since it allows to work with very large datasets, while providing powerful scripting capabilities. There are also many software packages available in different fields of science, although it is mainly used by data scientists for developing statistical software and data analysis.

In this project, RStudio was used for most of the data processing, and it was the main tool used for building the ML models and for evaluating them, and for obtaining the predictions on future scenarios. Among the several packages utilized, the main packages used were raster, caret, dplyr and factoextra.

Raster package

The raster package (Hijmans, 2020) has various tools for manipulating spatial data, mainly in raster format. For this project, the raster package was used to handle all the variables in raster format, resampling and resizing them to the extent of the study area, with functions such as *extent*, *resample* and *crop*. It was also used to prepare the data for the ML models by extracting the cell values of the rasters to each of the groundwater measurement points, which is performed with the *extract* function. Finally, the raster package also allows to make predictions with the ML fitted models obtained, by using the *predict* function.

Caret package

The caret package (Classification and Regression Training) (Kuhn, 2008) provides tools for pre-processing, developing and evaluating predictive models, with access to many different ML algorithms and several options for model training and tuning, where one can choose between a myriad of modelling techniques (Kuhn, 2019). Thus, caret provides the tools to build a predictive model from the beginning stages to model training, and prediction and validation of the results.

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

Figure 4. Explanation of the caret's train function. Source: Kuhn (2019).

In this project, caret was used for training the models with the three ML algorithms selected, including the evaluation of the models, which is done with the *train* function in the package. With the *train* function, caret can evaluate, with resampling, the effect of model tuning

parameters on performance, choose the "optimal" model across these parameters, and estimate model performance from a training set. Three types of resampling methods can currently be specified in caret: *k*-fold cross-validation (once or repeated), leave-one-out cross-validation and bootstrap.

After resampling, the process produces a profile of performance metrics to guide the user as to which tuning parameter values should be chosen. By default, the function automatically chooses the tuning parameters associated with the best value, but these can also be specified by the user (Khun, 2019).

As a summary (Figure 4), the train function fits a model for each set of resampled data with a different tuning parameter combination each time, and each model is used to predict the corresponding output samples. Then, the resampling performance is estimated by aggregating all the output sample sets. Using these performance estimates, the function evaluates the right combination of tuning parameters. Finally, the final model is fit with the entire training dataset.

Dplyr

Dplyr (Wickham et al., 2021) is a grammar of data manipulation. It provides a consistent set of verbs that help in solving the most common data manipulation challenges, such as:

- mutate: adds new variables that are functions of existing variables
- select: picks variables based on their names.
- filter: picks cases based on their values.
- summarise: reduces multiple values down to a single summary.
- arrange: changes the ordering of the rows.

These all can be combined with *group_by()* which allows to perform any operation "by group", and was heavily used in this project for pre-processing the wells dataset.

Factoextra

Factoextra (Kassambra & Mundt, 2020) is an R package that makes it easier to extract and visualize the output of exploratory multivariate data analyses, such as Principal Component Analysis, which was its main use during this project.

4.1.4. Saga GIS

SAGA (System for Automated Geoscientific Analyses) (Conrad et al., 2015) is a free open-source GIS software designed for an easy and effective implementation of spatial algorithms. It offers a wide set of geoscientific methods within an easily approachable user interface. The SAGA development was started by a small team of researchers from the Dept. of Physical Geography in Göttingen and is under constant development. It can be used as an independent software, but it has also been implemented into QGIS and it is available as a package in R (RSAGA), so the analyses and tools can also be accessed through QGIS and R.

SAGA's main objective is to give users an effective but easily learnable platform for the implementation of geoscientific methods, providing a fast-growing set of geoscientific methods ready to be used in numerous applications (Conrad et al., 2015). Most of the modules available in SAGA focus on Digital Elevation Models and Terrain Analysis, like analytical hill shading, local geomorphometry and geomorphographic classifications, terrain parameters related to hydrology, channel network and watershed basin extraction, and the creation of profiles and cross section diagrams.

For this project, SAGA was used to calculate the slope, flow accumulation, both vertical and horizontal distance to water bodies, and the monthly incoming solar radiation for the study area.

4.2. Correlation and PCA

4.2.1. Correlation

In statistical terms, correlation is used to denote association between two quantitative variables. In general, it is useful to determine if a dataset contains highly correlated variables because these might add more complexity to the models due to the presence of redundant information (Kuhn & Johnson, 2013). Using highly correlated variables might result in unstable models and might negatively affect their predictive performance.

When a pair of variables are significantly correlated, they are considered to have collinearity, or multicollinearity if this occurs between multiple predictors. Usually, the degree of association between variables is measured by the correlation coefficient r, also called Pearson's correlation coefficient. This correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1 (Kent State University, 2021). Complete correlation or a perfect collinearity happens when predictors are "exact linear functions of each other" (Dormann et al., 2013), and the correlation coefficient is expressed by either + 1 or -1. When this occurs, one of the variables should be omitted from the models. On the other hand, complete absence of correlation is represented by 0. A positive correlation shows that when one variable increases, the other

increases too, while a negative correlation shows that when one variable decreases, the other increases.

A way to easily visualize correlation between variables is by computing a correlation matrix, which can be obtained by contrasting the relationships of the different predictors with each other, in the form of a table or matrix. This allows to easily visualize which pairs have the highest correlation. The diagonal of the matrix is always a set of ones, since the correlation between a variable and itself is always 1.

For this project, a correlation matrix was computed to quickly visualize the degree of correlation between the variables, and a recommended threshold of ±0.7 (Dormann et al., 2013) was used to exclude highly correlated variables.

4.2.2. PCA

Principal Component Analysis (PCA) is a widely spread technique used in exploratory data analysis and for making predictive models. It is based on the Pearson's correlation coefficient, and it helps in reducing the dimensionality of a dataset while preserving as much variability as possible, helping to avoid the issue of overfitting the ML models (Jolliffe & Cadima, 2016). Basically, a lower number of independent variables reduces noise and eliminates redundancy in datasets that may already experience high correlation with one another (Goonewardana, 2019).

By preserving as much variability as possible, the PCA finds new variables that are linear functions of those in the original dataset, and are plotted against orthogonal axes that represent a variable. As more axes are added, these successively maximize the variance while being uncorrelated with each other. After the examination of each variable's magnitude (Eigenvalue) and direction (Eigenvector), a covariance matrix displays the results where each variable is listed within a Principal Component (PC). Depending on the percentage of variance threshold that is set, the number of PCs can be narrowed down to those that exhibit higher variance and are thus more significant. PCs 1 and 2 usually contain the biggest variance, and variables contributing to these two first PCs are overall the most important for the model, as they contribute for most of the variance (STHDA, 2017).

However, since variance is a concept that is only applied to variables that are continuous and numeric, PCA is not suitable to variables that are categorical or discrete (Linting et al., 2007).

The PCA is a built-in function in R that can be applied with the function *prcomp()*. In this project, PCA was used to explore the variance and importance of the dataset to reduce dimensionality by leaving out the least important variables.

4.3. Machine Learning algorithms

In this section, the ML algorithms used for this project are described. The selected algorithms for the project were RF, SVM and ANN because of their popularity in hydrological studies (see section 2.3.), but methods like SARIMA were also investigated during the first phases of development, and will be also briefly described in this section, along with the reasons why they were finally not used for the project.

In this project, regression ML was used. Regression analyses consist of a set of machine learning methods that allow us to predict a continuous variable (y) based on the value of one or multiple predictor variables (x) (STHDA, 2021). The goal of this methodology is to build a mathematical equation which defines y as a function of the x variables, and then use the equation to predict the outcome (y) based on new values of the predictor variables (x).

4.3.1. Random Forest

Random forest (RF) is an ensemble of decision tree (DT) algorithms, first developed by Breiman (2001). Decision Tree is a supervised learning algorithm used for both classification and regression problems. The DT consists of nodes and branches (Figure 5), forming a "tree". The number of branches depends on the number of criteria used to split the data into. The data is split in a binary manner following the criteria of each branch until achieving a threshold unit. These criteria are inferred from the training data. Each split of the data creates a node. The nodes in the DT are referred to as "parent" and "child" depending on their directional connection. The first node is the "root" and consists of the entire dataset, prior splitting, and the final nodes, with no further connections, are denoted as "leaf".

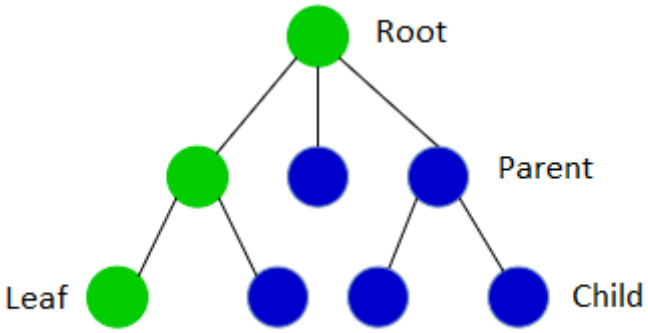


Figure 5. Example of a DT. Source: Analytics Vidhya (2020)

The RF algorithm is an extension of bootstrap aggregation of DTs (Figure 6) which can also be used for both classification and regression problems. The RF algorithm constructs multiple DTs that are trained in parallel with random bootstrapped samples of the training dataset, using different subsets of available features (Breiman, 2001). This method ensures that each DT is

unique, since each of them is fit on a slightly different training dataset, showing a slightly different performance, and reducing the variance of the RF.

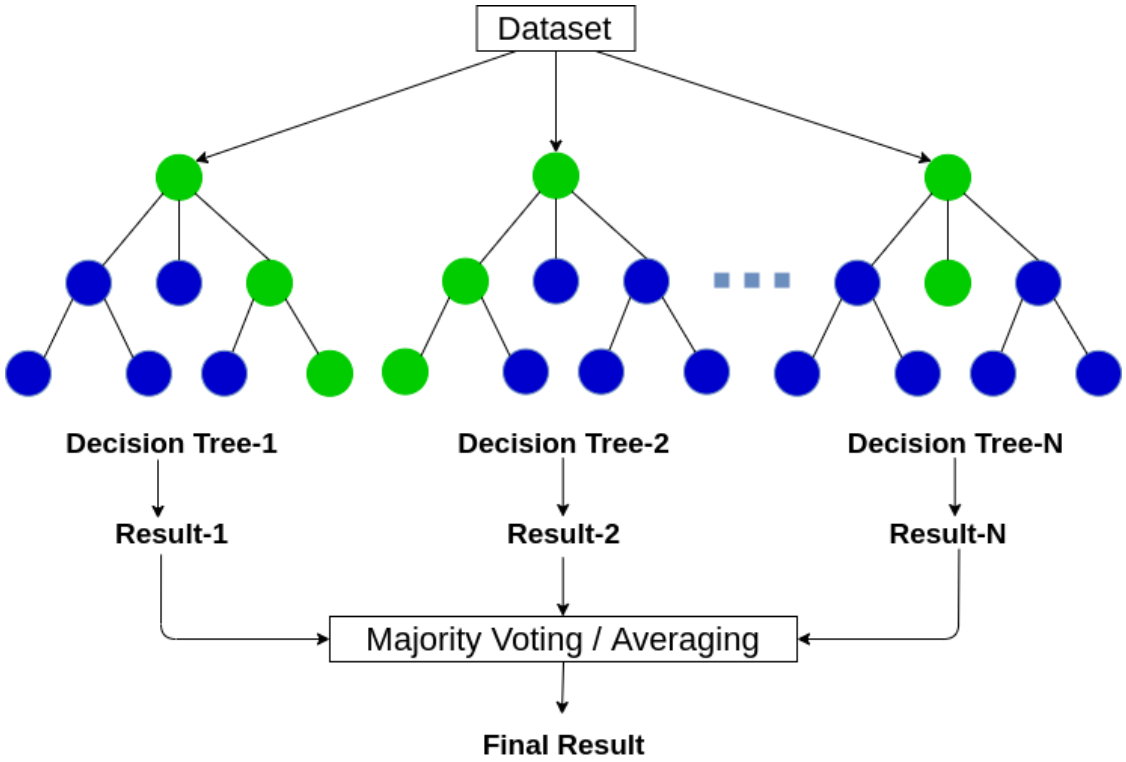


Figure 6. Example of a RF model with DTs. Source: Analytics Vidhya (2020)

For the final prediction, all the predictions from the individual trees are aggregated and averaged, resulting in better performance than any single tree in the model (Brownlee, 2020). A prediction on a regression problem is the average of the prediction across the trees in the ensemble, while a prediction on a classification problem is the majority vote class label predicted across the decision trees.

Thus, RF constructs a large number of DTs from bootstrap samples from the training dataset. At each split point in the construction of the trees, the RF algorithm also randomly selects a subset of input features, forcing each DT to be more different and reducing the correlation between prediction errors of each DT. These predictions are then averaged to make a final prediction, giving the final model (Brownlee, 2020).

This method also allows the RF to predict the importance of the variables by looking at how much of the error of prediction increases when one of the variables is left out (out of the bag, oob) while the rest are left fixed (Catani et al., 2013).

For tuning RF, the hyperparameters to take into account are the number of trees (ntree) and the number of variables randomly sampled as candidates at each split (mtry).

4.3.2. Artificial Neural Networks

ANNs are pieces of a computing system designed to simulate the way the human brain analyses and processes information. They are built like the human brain, with neuron nodes that are interconnected like a web. ANNs are built up with hundreds or even thousands of artificial neurons which are called processing units and are interconnected by nodes. These units are made up of both input and output units. Input units receive information and the neural network attempts to learn about this information to produce one output report, or prediction (Frankenfield, 2020). These nodes can be weighted to communicate each one's strength and affect the final model outputs. The weights of these nodes are combined before being passed through an activation function that ultimately translates the input into output with a value range of 0-1, in a process called feed-forward (Zhou, 2019).

ANNs also use a set of learning rules called backpropagation or backward propagation of error. ANNs go through a training phase where they learn to recognize patterns in data, and where the network compares the output with the actual measurements or what it was "supposed to be". The difference between both outcomes is then adjusted using backpropagation, meaning that the network goes backwards from the input to the output units to adjust the weight of the connections between the units, until the difference between outcomes produces the lowest error possible (Frankenfield, 2020).

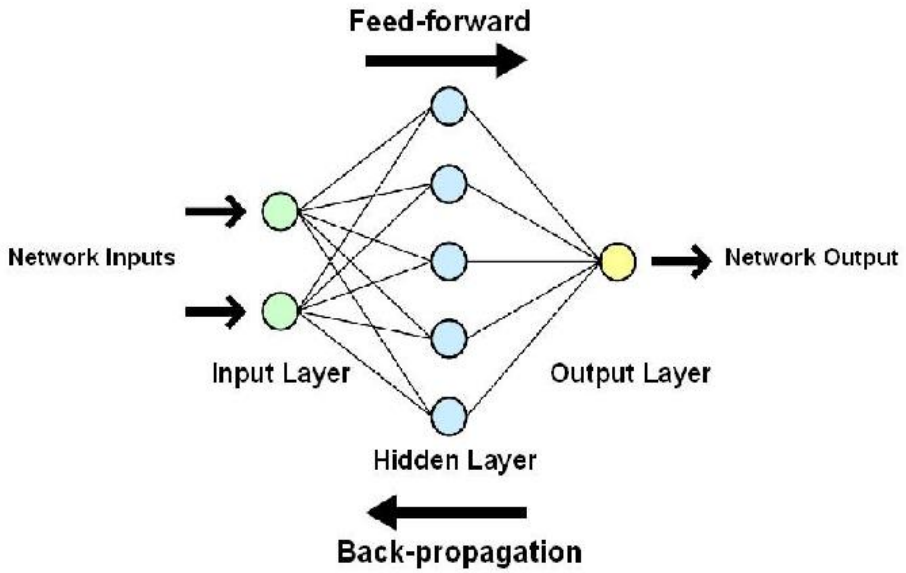


Figure 7. Example of an ANN network. Source: Güllü & Yilmaz (2011).

Any number of neurons can be embedded in the network as hidden layers, which are located between input and output and where the functions to weight the inputs are applied. These layers go through an activation function, which "fires" if the value of the input is larger than a

threshold. Should the combined weight of a neuron ensure its activation, it is passed through on to the next layer (Chauhan, 2019).

The hyperparameters that are necessary to tune for ANNs in the *caret* package are the decay, which is the regularization parameter to avoid overfitting, and the size, which serves to adjust the number of units in the hidden layers.

4.3.3. Support Vector Machines

Support Vector Machines are a set of supervised ML methods used for classification, regression and outlier detection. SVMs use a binary or multi-class approach to data segregation, finding a hyperplane that maximizes the margin between the classes of the data. The vectors that define the hyperplane are the support vectors. In the case of regression problems, a margin of tolerance is approximated, but the idea is the same: to minimize the error by individualizing the hyperplane that maximizes the margin.

SVMs are effective in high dimensional spaces, even in cases where the number of dimensions is greater than the number of samples. SVMs are also versatile, as they can be used for both linear and non-linear data by applying different Kernel functions to the decision function of the support vectors. However, SVMs are negatively influenced by large, noisy datasets, proving less suitable for these types of datasets than other ML algorithms (Sayad, 2021).

When tuning a SVM model, the hyperparameters used are c (cost), which controls training errors and margins, and sigma, which determines the reach of a single training instance.

4.3.4. Univariate models: ARMA, ARIMA and SARIMA

Some time series models such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) were investigated at the start of this project as plausible options for predicting change within wells. ARIMA is a widely used forecasting method for univariate time series data forecasting, and while it can handle trends on the data, it cannot handle data with seasonal components. SARIMA is an extension to ARIMA that supports the direct modelling of the seasonal component of the data, which makes it more suitable for groundwater level forecasting (Brownlee, 2019). These models have been used in several studies to investigate patterns between the input and the output of groundwater data to make future predictions (e.g. Hussein et al., 2020).

These models are considered linear fitting models and have the advantage of accounting for the correlations that arise between data points, and have a great performance when used for forecasting on univariate datasets. However, linear fitting is not ideal in describing the nonlinear behaviour of groundwater, and univariate approaches are not best suited for this type of data

(Hussein et al., 2020). Most of the time series characteristics of hydrology and water resources systems are nonstationary, and it is necessary to use methods that can model these behaviours to optimize water systems. Thus, using classic statistical methods such as SARIMA, that imply that the data are stationary, are not a good choice, as its linear approach can lead to poor performance when testing the model with unseen data (Ticlavica & Torres., 2012).

Additionally, these kinds of models do not support long-term forecasting, such as the climate scenarios used in this project, since the long term forecast asymptotically approaches the mean value of the time series data (Shumway & Stoffer, 2011). For long-term forecasting of water systems such as groundwater modelling, ML models have been proven to be more suitable than univariate models (Nourani et al., 2009). Hence, these models were dropped as ML algorithms are more suitable for both this kind of data and the time-period for the forecasting.

4.4. Model validation and performance

4.4.1. K-fold cross-validation

Cross-validation is a resampling procedure used to evaluate ML models. For this procedure, the input dataset (the dependent variable) is resampled into training and testing sets or folds. The training data is used to train the ML model, while the testing data is used to evaluate the model's performance by predicting on unseen data, which reduces the possibility of overfitting (Scikit-learn, 2020a).

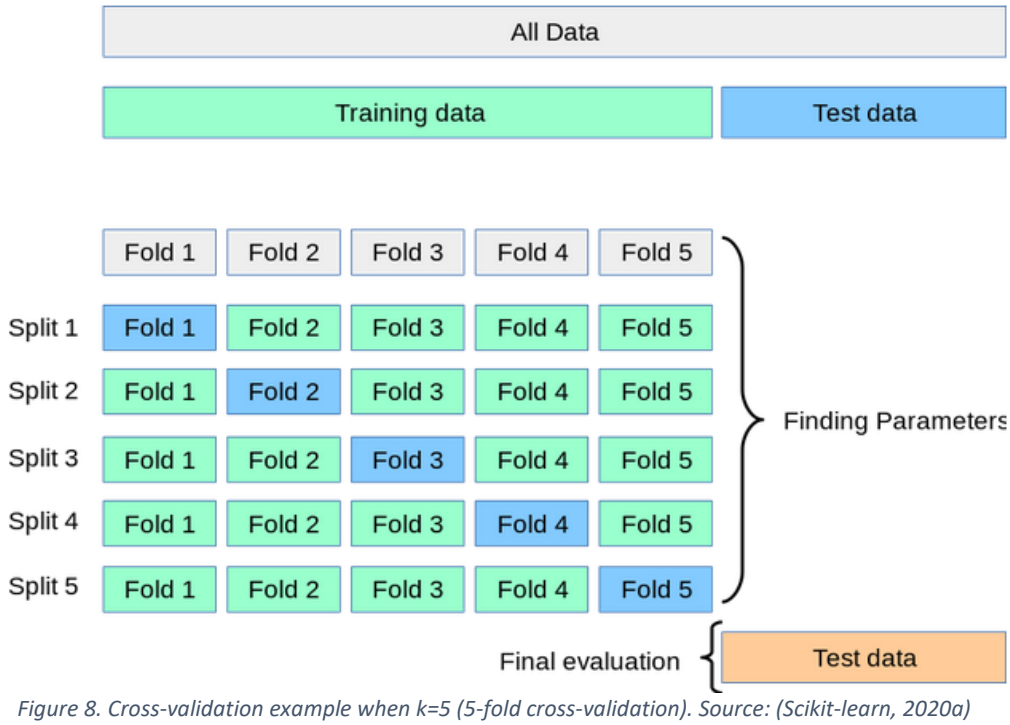


Figure 8. Cross-validation example when k=5 (5-fold cross-validation). Source: (Scikit-learn, 2020a)

In k-fold cross-validation, the input dataset is split into a number k of groups or folds. Common values of k are 3, 5, 7 and 10 (Brownlee, 2014). When deciding the value of k to use for the cross-validation, the most important factor is the size of the input data, so both the training and testing folds are large enough to be statistically representative of the broader dataset (Brownlee, 2018).

Once the k parameter has been set, each fold will be used as training data k-1 times, and will be used for testing once. Hence, for each iteration, the model will be trained with k-1 folds and one fold will be used to test the model. In a 5-fold cross-validation (k=5), five models will be obtained through this process, and each of the models will be trained with 4 folds and tested in one fold, which will be different for each model (Figure 8). Once the process is over, these models are discarded, and their skills scores are collected and summarized for use (Brownlee, 2018).

4.4.2. Metrics for model performance

Coefficient of Determination - R^2

The coefficient of determination or R^2 is a statistical measure of how close the data are to a fitted regression line. It is thus the percentage of the dependent variable variation that is explained by a linear model (Rieuf, 2017).

In a regression ML model, the R^2 represents the proportion of variance of the dependent variable that has been explained by the independent variables in the model. It provides an indication of goodness of fit, providing a measure of how well unseen samples are likely to be predicted by the model through the proportion of explained variance (Scikit-learn, 2020b). How close the data are to the predicted data, showing how much of the variance contained in the training data is captured by the ML model.

If $\hat{y}$ is the predicted value of the i-th sample and y_i is the corresponding true value for total n samples, the estimated R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

The R^2 ranges from 0 to 1. A score of 0 would indicate that the model explains none of the variability of the response data, thus indicating that the predictive ability of the model is low. On the other hand, a score of 1 would indicate that the model explains all the variability of the response data.

Mean Absolute Error - MAE

The Mean Absolute Error (MAE) is a measurement of the average magnitude of the errors in a set of forecasts, being a measure of accuracy for continuous variables. The MAE is expressed as the average of the absolute values of the differences between the predicted and the actual values, and a form of verification of regression ML models. Since the MAE is a linear score, the individual differences are all weighted equally in the average (Scikit-learn, 2020b). Additionally, one of the advantages of the MAE is that it is more robust to outliers and does not penalize the errors as extremely as other metrics, such as the RMSE.

The MAE can be defined as:

$$MAE = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i|$$

Where $\hat{y}$ is the predicted value of the *i*-th sample, and y_i is the corresponding true value, and *n* is the total number of samples or errors.

Root Mean Squared Error - RMSE

The Root Mean Squared Error (RMSE) is a measurement of the model's prediction error. Similar to the MAE, the RMSE corresponds to the average difference between the observed values and the predicted values by the model (STHDA, 2021).

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. It expresses the difference between forecast and corresponding observed values, where each is squared and then averaged over the sample, and finally, the square root of the average is taken. The lower the RMSE, the better the model. Additionally, since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors.

The RMSE is computed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Where $\hat{y}$ is the predicted value of the *i*-th sample, and y_i is the corresponding true value, and *n* is the total number of samples or errors.

The RMSE will always be larger or equal to the MAE. The greater the difference between them, the greater the variance in the individual errors in the sample.

4.5. Implementation

First of all, the groundwater data was pre-processed as explained in section 3.2., while the independent variables were all loaded into R and pre-processed to have the same extent and resolution, as explained in section 3.3.

Once all the data was pre-processed in R, the *extract* function from the raster package was used to get the background data for each of the groundwater observations at each of the observations' location, and according to the date when they were collected (an example of the code can be seen in Figure 9). Like this, monthly precipitation and temperature were linked to each observation by date, and imperviousness and landcover were also extracted according to the year. The result of this process was a dataframe containing the groundwater measurements for each well along with the corresponding background data for each observation.

```
#loop to extract precipitation data for each month from 1990 to 1999
for (i in seq_along(prec1990list)) {
  sub<- subset(pointsfinal, date==datelist$date[[i]]) #subset gw observations by month
  sub$ID <- seq.int(nrow(sub)) #give the observations a unique ID
  data<-raster::extract(p1990[[i]], sub, df=T) #extract precipitation for each observation
  pder <- merge(sub, data, by.x="ID", by.y="ID") #merge extracted precipitation with gw observations
  names(pder)[9] <- "precip"
  data1990<-rbind(data1990, pder) #merge dataframe with all months/years
  print(paste0("Extracted ", i, " out of ", length(prec1990list)))
}
```

Figure 9. Example code showing the loop used to extract the climatic variables by date for all the groundwater observations.

After the background data was extracted, the next step was to perform a PCA and a correlation test to better understand the importance of the different variables, excluding those that are highly correlated or explain little of the variation of the data, in order to reduce noise and avoid overfitting of the ML models. First, the correlation test was performed to detect collinearity between the variables by using the *cor* function built in R, and *corrplot* was used to see the correlation matrix (Figure 10). The correlation matrix shows the pairwise correlation between variables, both positive (blue) and negative (red), showing a stronger correlation with darker colours and bigger circles (Figure 11).

```
c<-cor(cor.factor[,c(6:30)]) #check correlation
corrplot(c, is.corr = T)
```

Figure 10. Function to perform the correlation test and matrix.

By looking at the correlation matrix (Figure 11), it is easy to detect a strong positive correlation between the temperature variables, and some correlation as well between these and the incoming solar radiation (ins), especially with the maximum temperature (tmax). The Pearson correlation coefficient was also calculated separately and a threshold of ±0.70 (Dormann et al., 2013) was applied for the correlation coefficient, excluding variables that were highly correlated. It was decided to keep the average temperature (tmean) in order to have a bigger

overview on the variance between minimums and maximums, and also because the maximum temperature showed the highest correlation to the incoming solar radiation.

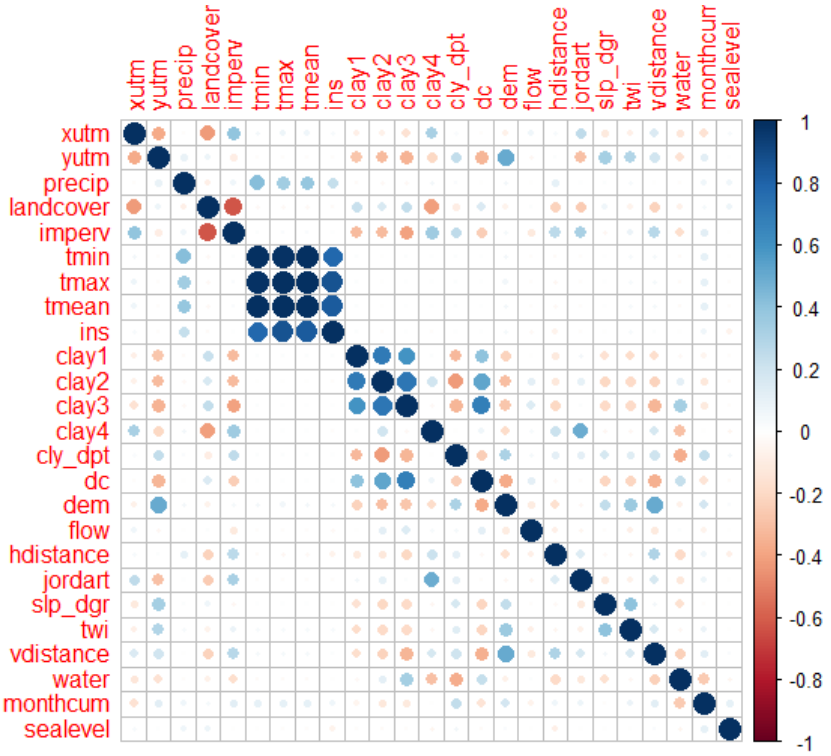


Figure 11. Correlation matrix for assessing the correlation between variables, from 1 to -1.

Then, a PCA was also performed to understand how much of the variation of the data was explained by each of the independent variables, thus finding their importance in order to exclude the least important ones. Since PCA cannot handle categorical variables well, this process was performed only with the continuous data, leaving land cover, drainage class and the type of soil out of the analysis. The PCA was performed with the function *prcomp* which is built-in in R, and *factoextra* was used to visualize the results (Figure 12). In the final graph obtained (Figure 13) it can be seen the contribution of each variable to the first and second PCs, which are the ones explaining most of the variation of the data.

```
#PCA
pcadat[,c("landcover", "dc", "jordart")]<-NULL #exclude categorical variables

data.pca<- prcomp(pcadat[,c(2:20)], center = T, scale. = T) #perform pca

#plot the contribution to pc1 and pc2
fviz_pca_var(data.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
)
```

Figure 12. Code snippet of the PCA analysis and the visualization method used.

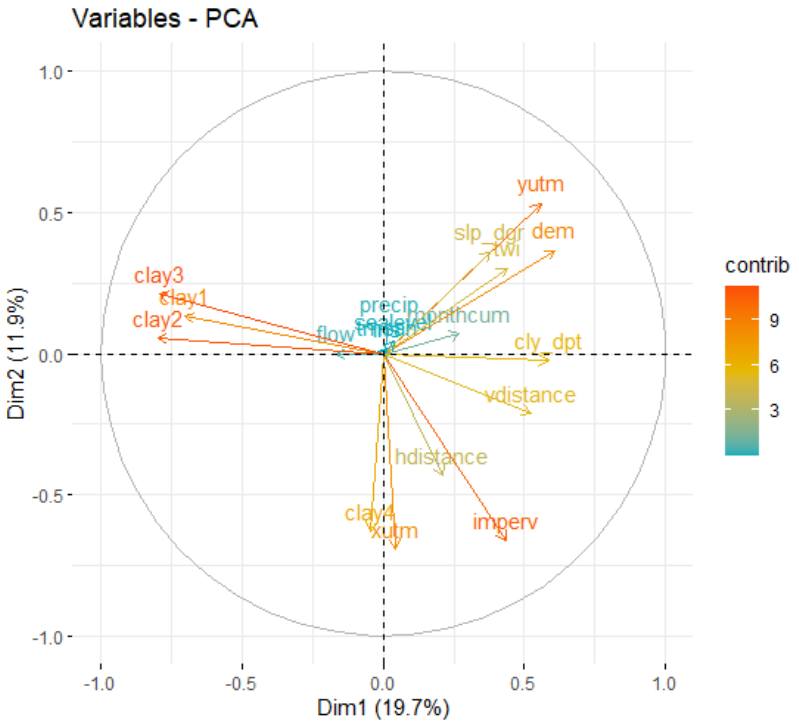


Figure 13. Results from the PCA. Contribution of each variable to the first and second PCs.

After performing the PCA, the flow accumulation and sea level variables were also excluded, as these added little to the main PCs. The PCA also showed that both precipitation and temperature were quite low in the contribution to the first and second PCs, although their contribution was higher in following PCs. Additionally, these climatic variables are the ones with the highest variation as they are historical monthly data, while the other variables are mostly static, so these two were kept in the analyses. The PCA does not really look into the relationship between the independent variables and the values of the dependent variable, while RF and ANN do look into that when analysing the importance of the variables. Therefore, these were left in the analyses for the models to pick up their importance, as they are supposed to be relevant for changes in the water level and thus, for forecasting future scenarios where these variables will suffer big changes.

After dropping high correlated variables and those that explained little of the variation of the data, 20 independent variables were used for the models (see section 3.3.). The wells dataset was then divided into training and testing data in an approx. 70:30 ratio, but maintaining the temporality, so data from 1990 to 2016 was used for training, and the period for 2017-2018 was used for testing. This was done following the recommendations on the literature, since partitioning time series randomly could cause a look-ahead bias (Dixon et al., 2019).

For hyperparameter tuning, the caret package can optimize the hyperparameters for each model by using a random search, selecting the optimal values for each of the hyperparameters. This method allows to easily find the best hyperparameters for each model, while making it possible

to build all the models in a similar fashion so they can be easily compared. To train the models with caret, it is necessary to input the independent variables as a dataframe and the dependent variable as a vector separately, and some parameters need to be set:

- Method: the ML algorithm to be used
- trainControl: defines the type and number of resampling, as well as the search method. For this project, a 5-fold cross-validation was used.
- metric: determines how the final model is defined, by selecting the tuning
- parameters with the highest value of the objective function. Since these are regression models, it was set to "Rsquared".
- tuneLength: sets the size of the default grid of the tuning parameters. It was set to 10 for all the three models.

```
set.seed(10)

rfmodel <- caret::train(train[,predictors],train$groundwater,
                        method="rf",
                        metric="Rsquared",
                        ntree = 1000,
                        tuneLength=10,
                        importance=TRUE,
                        trControl=trainControl(method="cv", number=5, search="random"))
```

Figure 14. Code snippet of the train function of the caret model with the hyperparameter tuning selected for the RF model as an example.

The parameters were selected according to the literature and by "trial an error" when setting the models. For reproducibility, a seed number was set before each model. After setting all the parameters, the models were run, obtaining the metrics for each model (R^2, MAE and RMSE).

The testing data on 2017-2018 was then used to evaluate the models with the *predict* function from the raster package, and R^2 and MAE were calculated by comparing the actual with the predicted values obtained by the models. This dataset was then exported as a shapefile to further check on QGIS, where the location of the wells and the different errors were checked and visualized, and finally made into a map showing the location of the largest errors.

Since the RF model was the one with the best outcome, another three RF models were also trained based on three different land cover types. Thus, 3 models were trained by dividing the data into three categories: nature, agricultural and urban. This was done to check whether the location of the wells affected the model, or if the placement of the wells was affecting the observations.

Because static variables can cause overfitting in these types of models (Meyer et al., 2018), both a RF and ANN models were trained excluding all but the climatic variables to see if the

static variables were in fact overfitting the models. However, results on this were not good and further analyses on this matter were dropped.

For the future predictions, data for the different climate scenarios for both winter and summer was used. However, not all the data was available for the future climatic scenarios. For instance, climatic data was available for all the three different SSPs selected, but future predictions on land cover and imperviousness were not available. Therefore, the latest data available for these variables was used, which accounts for the period 2018-2020.

The variables for the different SSPs were pre-processed following the same steps as for the predictors used to train the models. This data was added to R as a raster stack and used with the *predict* function (Figure 15) from the raster package along with the first trained RF model, as this was the one with the best results. Two predictions were made for each of the SSPs selected (SSP2-45, SPP3-70, SPP5-85), one for the winter and one for the summer.

```
predict245f<-stack(predictors2,monthf,insf,fprec245,ftemp245) #all predictors winter SSP245
prediction245f<-predict(predict245f, pcarfmodel) #ssp245 winter
```

Figure 15. Example of the code used for predicting one of the future SPP scenarios selected.

The resulting maps obtained from the predictions were exported as tif files and further processed in QGIS for better visualization and for making the final maps.

5. RESULTS

The following section is used to show and describe the results obtained from the ML algorithms, starting with the direct metrics obtained from the ML models, and finishing with the prediction maps developed with the selected ML model.

Comparison of the models

As explained in the implementation section, after performing the PCA and the correlation test, four variables were excluded for being too highly correlated or for explaining too little of the variation of the model, and further training of the models was performed without them.

Table 2. Result scores obtained from the training of the three different models.

ML Model	R2	RMSE	MAE
RF	0.75	0.98	0.61
ANN	0.63	1.19	0.85
SVM	0.65	1.15	0.75

After training the three different models and obtaining their metrics, it can be seen (Table 2) that the RF model was the one with the better scores, obtaining the highest R^2 (0.75) and the lowest MAE (0.61m) and RMSE (0.98m) of the three.

Table 3. Result scores from the evaluation of the three ML models.

ML Model	R2	MAE
RF	0.50	1.01
ANN	0.41	1.17
SVM	0.25	1.31

After evaluating the models on the test data for the period of 2017-2018, the scores dropped slightly for all three models (Table 3). Small differences in performance are expected when comparing training and testing, but it might also indicate that there could be some overfitting in the models. Still, RF performance seemed to be best with a R^2 of 0.5 and MAE of 1m, which according to the literature (Henriksen et al., 2003) can be considered as a good result, and was selected for further predictions.

Looking at the results from the evaluation performed with the RF model on the test dataset, it can also be seen that the largest errors (of approximately 5-3m) are all located in urban areas (Figure 16), mostly in Copenhagen (with the exception of three observations located in rural areas). These errors account for 6% of the number of errors, and almost 20% of the total error, being an important cause of the increase in the MAE and the reduction of the R^2 score.

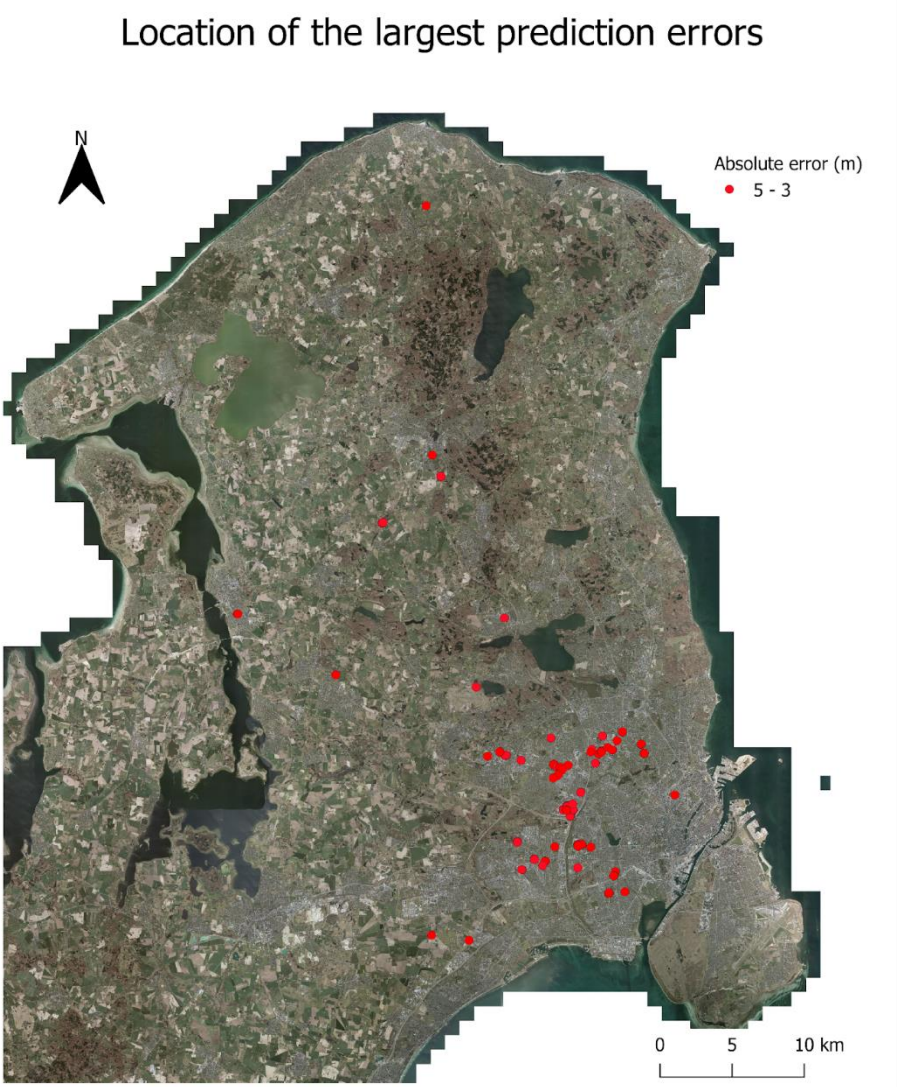


Figure 16. Map with the location of the observations that accounted for the largest prediction errors with RF the model.

When looking at the importance of each variable based on the RF and ANN models (Figure 17), static variables like the elevation (dem), horizontal and vertical distance to water bodies (hdistance and vdistance, respectively) or the depth to clay (cly_dpt) have relatively high importance in both models, being the predominant factors explaining the variation within the groundwater data. Imperviousness (imperv) also has a rather high importance in both models, while precipitation and temperature are surprisingly quite low, especially in the ANN model.

These results also go along with the PCA performed, where the climatic variables had a rather low importance in the first two principal components. No results are shown from SVM, as this model does not have a built-in function to account for variable importance.

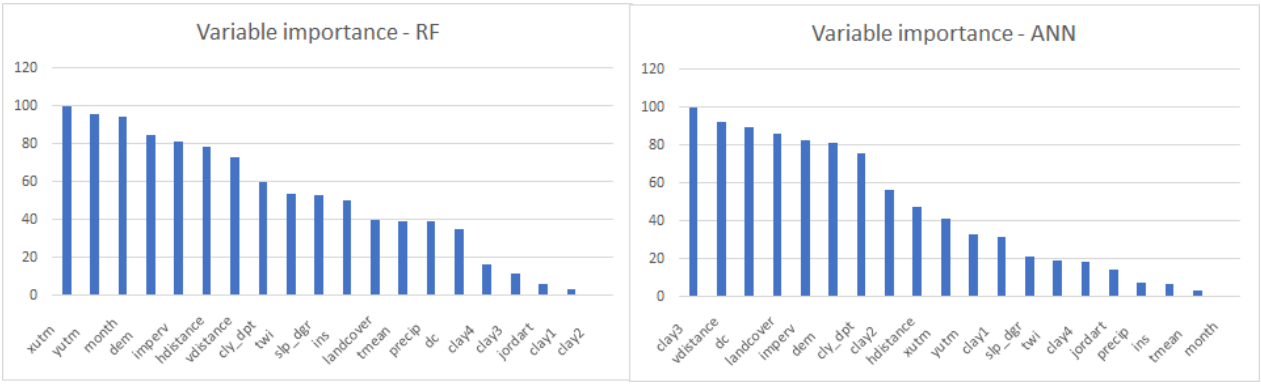


Figure 17. Importance of the variables based on the RF (left) and the ANN (right) models.

Three more RF models were also trained with wells located in different land cover types. When dividing the data by land cover, three land cover types were used: urban, agricultural and nature, to check if the location of the wells affected the predictive efficiency of the model.

Table 4. Scores for the RF models trained with data from different locations based on land cover.

Land cover type	R2	MAE
Urban	0.70	0.63
Agricultural	0.65	0.69
Nature	0.86	0.38

After training the three models with RF, it can be seen that the model seems to perform best on natural areas, while the worst seem to be within agricultural areas, which might indicate that there are other factors apart from the variables used in the model affecting some of the observations. However, it is also important to note that the number of observations on natural areas is lower than those on either urban or agricultural land cover types.

Another reason for having least accuracy in farms would be that farms' moisture level is manipulated through irrigation and planting high water consumption crops. Moreover, they are not planted every few years for maintenance purposes and hence the water level variation changes naturally. This means that farmlands are sometimes actual agricultural crops and

sometimes are left as grass/barren land depending on the region (Levin, 2006; Danish Agriculture and Food Council, 2019).

Then, a RF model without all the static variables was also trained to account for a possible overfit in the model caused by these. As explained before, land cover and imperviousness were also treated as static because they changed little during the period selected for this project. Therefore, only precipitation and temperature were used for this model.

However, the results were unsatisfactory, with low metrics after training (R^2 of 0.39 and MAE of 1.31) and even worse results after evaluating the model on the test dataset (R^2 of 0.12 and MAE of almost 2m), showing a very low predictive power. Thus, it seems that although the static variables can overfit a model, they are also explaining the variation of the data to some extent.

Future predictions

Two prediction maps for each of the climatic scenarios were obtained from the predictions made with the RF model, one for the winter season and one for the summer season. Additionally, two maps were also obtained for the current situation, one for winter and one for summer, respectively. The final maps for each season and each climatic scenario can be seen in figures 18 and 19.

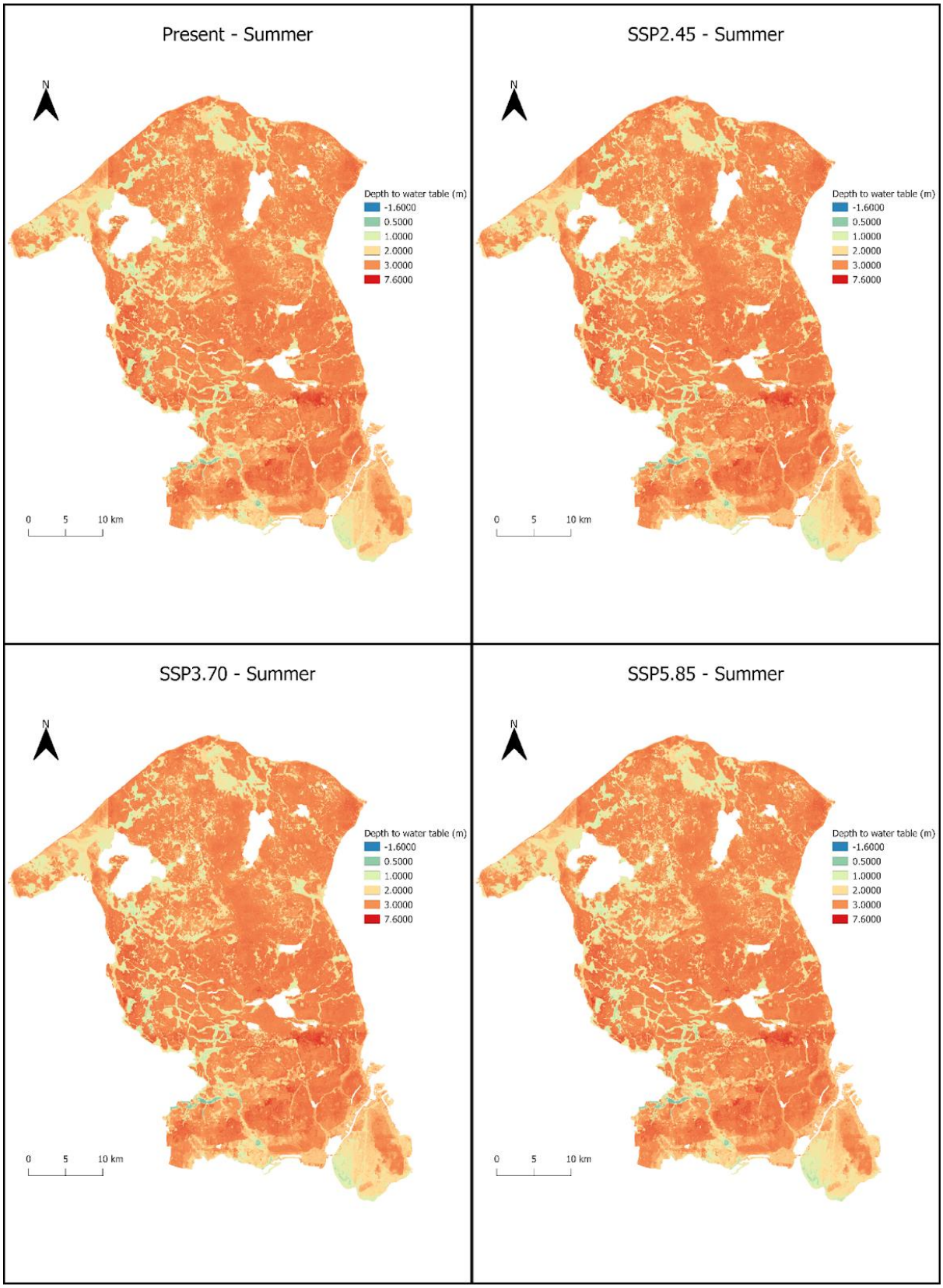


Figure 18. Comparison of the resulting maps based on the predictions made with the RF model for the different climate change scenarios for the summer season.

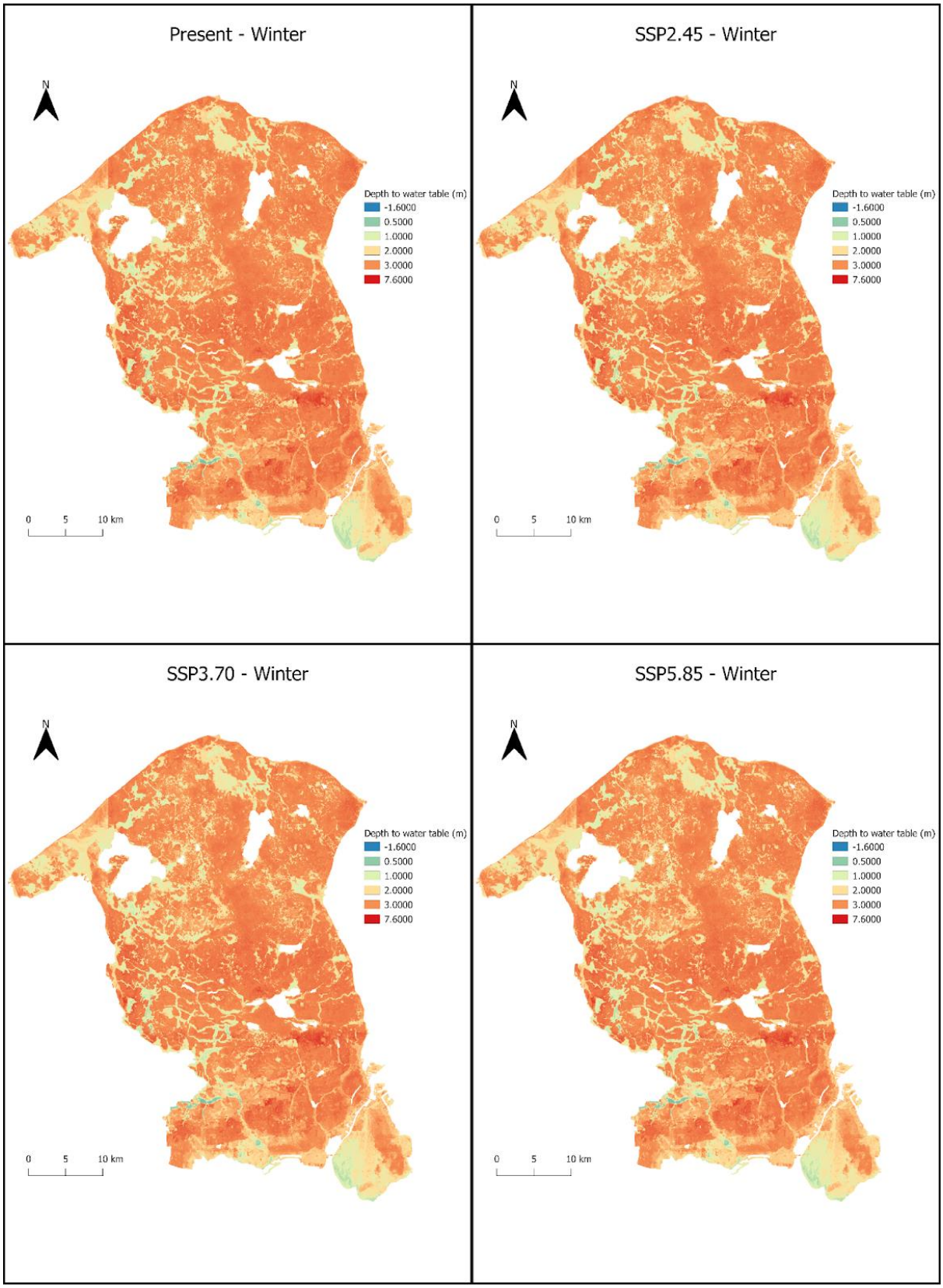


Figure 19. Comparison of the resulting maps based on the predictions made with the RF model for the different climate change scenarios for the winter season.

Visually, there are no noticeable changes between climatic scenarios, although differences between the present and future water levels can be noticed, especially in some specific areas like to the east of Copenhagen (Figure 20). The maps show that overall, the future water table will be higher, both in winter and summer, being the SSP2.4-5 an exception with similar levels to the current situation. Additionally, it shows that summers will be drier than winters, in line with the expectations from the literature (Koch et al., 2019).

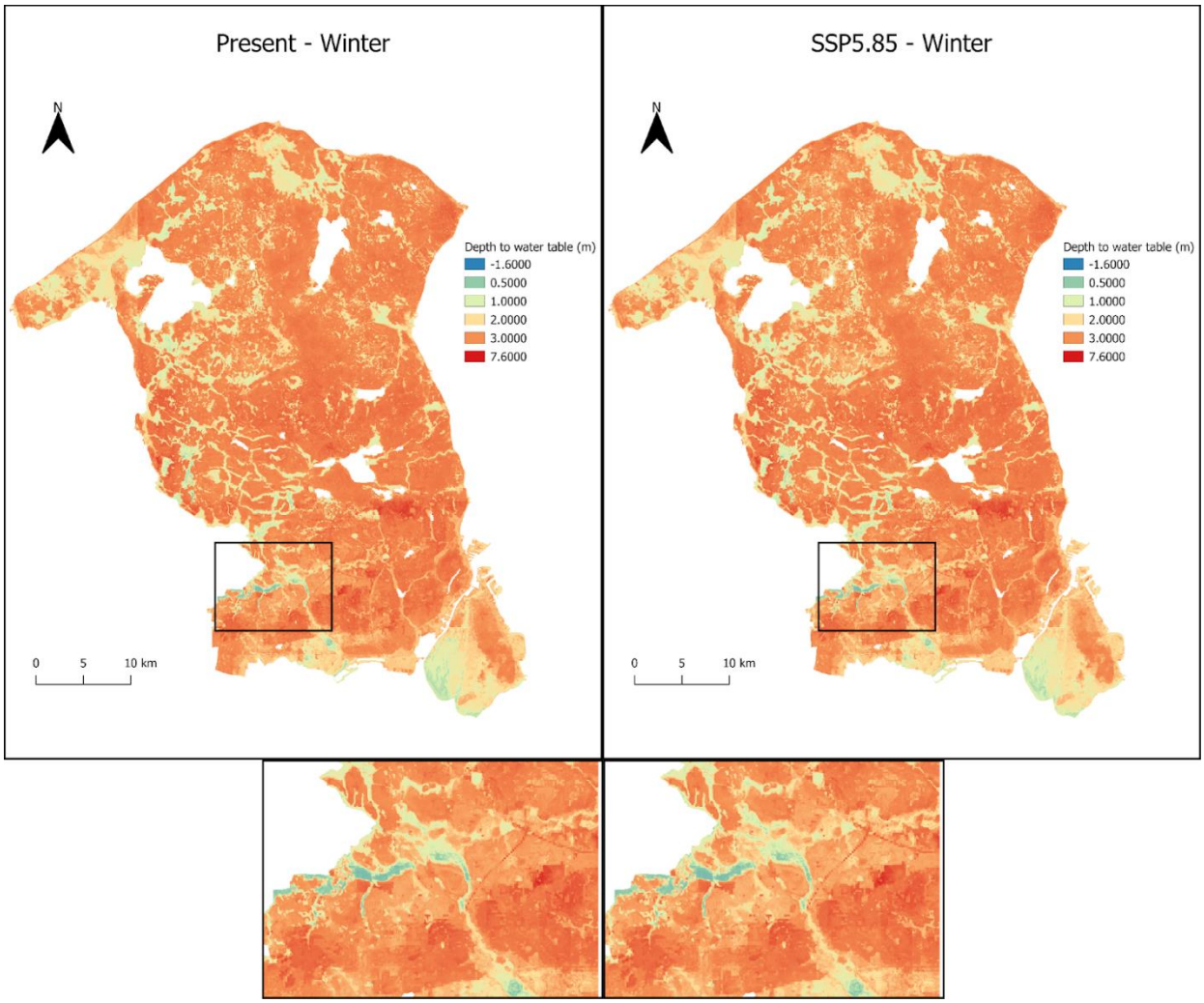


Figure 20. Comparison with one of the SSPs in a zoomed in area where there is noticeable change in the water table depth.

Comparing a specific area in present and future scenarios (Figure 20), it can be appreciated that the orange and red shades are lighter in the future scenarios, along with more blue shades, meaning that the groundwater will be closer to the surface in a larger area (Table 5).

The average groundwater levels of the different scenarios oscillated between 2.95 and 3m, being rather stable over the study area regardless of the month and scenario. However, when considering how the groundwater levels are distributed, we can see that in future scenarios there

is an increase in the areas where the water table will be closer to the surface, even in the summer periods. Since the first meter from the surface is the most important, as water table at this depth highly affects the surface, 1m was the focus for showing changes on the water table depth.

Table 5. Comparison of the % of groundwater in the first meter from the surface.

Scenario	Winter (1m (%))	Summer (1m (%))
2018	1.40	1.26
245	1.40	1.25
370	1.43	1.29
585	1.43	1.41

On the other hand, the differential maps obtained between present and future scenarios (Figures 21 and 22) show that overall, water levels will rise during the winter, while in the summer months most of the area will remain the same (yellow shades) or suffer a slight decrease in the water level (orange shades), although several areas will see a raise in the water table too (green/blue shades). In both cases changes will still be limited, and most of the area will suffer changes within 0-0.25m. On the other hand, while it could be seen before that the first meter of soil will overall be impacted by a rising water level (Table 5), it seems that the water at deeper levels will fall in summer, although not much (between 0 and 0.25m).

Figure 21. Differential maps showing changes in the water level between the present and each of the future SSPs selected for the summer season. Positive numbers indicate a rise in the water table, while negative numbers indicate a fall.

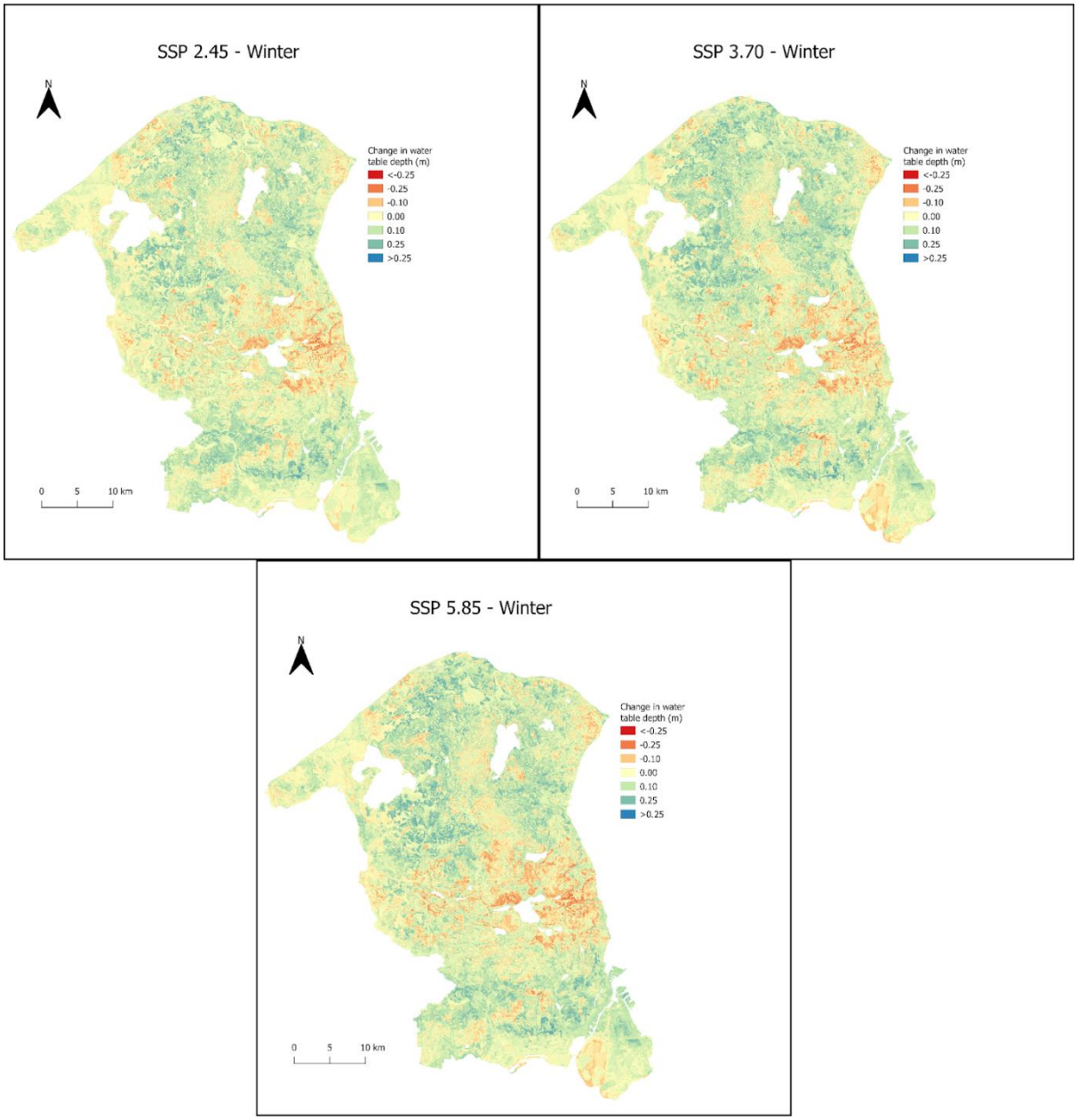


Figure 22. Differential maps showing changes in the water level between the present and each of the future SSPs selected for the winter season. Positive numbers indicate a rise in the water table, while negative numbers indicate a fall.

When further studying these comparisons between the different future scenarios and the present, we can see that the fluctuation in the maximum rise values for the water levels are larger than the maximum fall values (Table 6), and that the increase in the maximum rise values is especially noticeable in the scenarios with higher emissions (especially during the winter).

Table 6. Maximum change in the water level (rises and falls) for each of the scenarios compared to the present levels for both winter and summer.

	Winter			Summer		
Scenario	SSP 2.4-5	SSP 3.7-0	SSP 5.8-5	SSP 2.4-5	SSP 3.7-0	SSP 5.8-5
Max. rise (m)	+0.67	+0.83	+0.82	+0.70	+0.70	+0.71
Max. fall (m)	-0.52	-0.47	-0.49	-0.52	-0.49	-0.52

When looking at these results, it should be noted that, since the model is based on water levels from wells with no observations on either streams or water bodies, the model seems to predict drier scenarios than what should be expected, and the groundwater in streams is not represented over the surface. These limitations, together with the aforementioned possibility of overfitting, should be taken into consideration when inferring conclusions based on the model predictions.

6. DISCUSSION

Comparison of the models

RF had the best metrics on both the training and testing sets, and was, for this reason, deemed best for further analyses and for this project. As reviewed from the literature, RF is a model widely used in hydrology settings such as groundwater level forecasting, and has proven to provide very robust results and accurate predictions in several studies (e.g. Hedley et al., 2013; Gulo Blanco et al., 2017; Koch et al., 2018).

On the other hand, although ANN and SVM are also popular choices in this field, they were outperformed by the RF model in this project. Even though it is a possibility that the tuning of these two models was insufficient because it was more focused on allowing comparison between the three models, it is not uncommon for RF to outperform models such as ANN (Robbach, 2018). Thus, with the data available for this project and with the tuning possibilities researched for each model, it can be concluded that the best model was RF, followed by ANN and finally SVM.

As it was mentioned before, the accuracy drop of the RF model when evaluating with the testing data, might indicate that it is slightly overfit. As a measure to improve the model, this possible overfit was accounted for and different steps were followed to attempt to counterfeit it. As seen in the literature, it is possible in models that use spatio-temporal data to get overfit when static variables are used in the model (Meyer et al., 2018). However, removing those static variables from the model did not improve the results, which might mean that perhaps the climatic variables alone are insufficient to predict changes in the groundwater level of the study area.

Additionally, it was also tried to divide the groundwater data by season, as perhaps having a single model with data for both wet and dry seasons could be too much and confusing for the model. But this division of the data into wet and dry seasons did not improve the results either, giving similar metrics and predictions to the ones already obtained with the original RF model, and therefore this approach was disregarded and, due to resources and time constraints, this was not further investigated.

However, this should not take from the value of the results from the RF model, since related literature shows that a slight accuracy drop in prediction is to be expected, and the metrics of the model are within the expectations for a successful model (Henriksen et al., 2003).

Future predictions

As seen from the results, even though there are slight differences in the water levels amongst the different present and future scenarios, there seems to be less fluctuation than expected (Kidmose et al., 2013) in these values according to the selected RF model, and with the available data. There are different reasons that could explain this limited change.

On one hand, artificially drained areas are not expected to have a rise in the water table depth, but rather the increase in precipitation will cause an increase in the drainage water runoff (GEUS, 2017). The selected study area is highly anthropomorphic, being composed of mostly agricultural and urban areas, and it is likely that the artificial drainage in the region is well extended over the area, although information on this specific factor could not be found for this project. Therefore, it is possible that the water table will not suffer major changes in the region.

On the other hand, both precipitation and temperature are expected to increase in the future, which would translate into a higher water level due to higher recharge rates as a result of precipitation, but also into a lower water level due to increased evapotranspiration caused by the rise in temperatures (Danish Nature Agency, 2012). Because of these two opposite effects taking place, it is possible that average water levels for long periods such as dry and wet seasons, as compared in this project, will not change much in the future.

Additionally, the model picked some static variables as being important in explaining the variation of the water level. Since these variables will not change for each well for the different future scenarios, it is reasonable that there would just be small changes between their groundwater levels.

Regardless of the small changes in the average groundwater levels, it can be seen that in future scenarios (especially for the scenarios with higher emissions), the water table will be closer to the surface in some areas, albeit apparently just a bit. Still, even small changes on the water table on the first meter of the terrain can greatly impact the ecosystems and human activities on the surface (Zipper et al., 2015), and should be paid close attention to. This is especially relevant when considering that a water table rising closer to the surface would translate into a higher risk of flooding, especially in case of heavy precipitation events (Koch et al., 2019). It is highly possible that the risk of flooding will be very local, and specific to certain events such as heavy storms events (Jebens et al., 2016).

It is expected that sporadic precipitation events will be more common and increase in intensity in the future in Denmark (Danish Nature Agency, 2012). These sporadic events are not accounted for in the average precipitation data used in the model. The model presented in this thesis is not trained to pick up on these extreme scenarios, but it simply draws an overall trend based on how the groundwater levels have changed month by month, and how they will change

based on the climatic changes in the future. For this reason, any raise, even if small, should be paid close attention to, since any raise in the average groundwater level of an area estimated by the model, could actually mean much more significant rises during these extreme scenarios, translating into possible flooding or pollution infiltration scenarios.

Current concerns regarding the groundwater in Denmark are mostly focused on possible pollution infiltration to the groundwater, as this is a valuable resource for water supply (OECD, 2013; Danish Ministry of the Environment, 2021). Changes in the water table levels can increase the risks of pollution of the groundwater, as pollutants infiltrate through the soil and come in contact with the water (GEUS, 2017). In addition to this, there are other concerns regarding the effect that a higher water level will have on artificial infrastructures such as sewers, water pipes or buildings. If the groundwater level rises, it will result in an increased pressure on the foundations of buildings that are below the groundwater level. Moreover, the groundwater can infiltrate through leaks in sewer and water supply pipes, and pollutants in the soil can infiltrate into the pipes with the water (Miljø Metropolen, 2011).

Limitations of the model

As mentioned before, there are multiple factors that should be taken into consideration when inferring conclusions based on the model predictions.

The fact that the studied region is so affected by human activities could impact the ML model predictions. These areas are susceptible to water table modifications due to artificial drainage and irrigation, which are factors that the current model is not taking into account, and that might be affecting the measurements of the wells, thus affecting the predictions of the model as well. If the changes in a well are not reflected on the precipitation or the temperature of that month because there are other factors affecting it, the model will not learn properly about the trends and variation of those wells and its predictions will not be accurate. This might also be implied in the fact that the climatic variables used to train the models had overall a rather low importance, meaning that they explained little of the variability of the data. This was also hinted at by the models trained with data located in different land cover types, as the model trained with observations in natural areas gave higher metrics than the other models, which could mean that other factors are in fact affecting the other areas.

Regardless, based on the metrics obtained from the RF model, it seems that this is a good method for forecasting future water table levels, albeit the conclusions must be drawn carefully. Based on the model, little change will be seen on average in the future scenarios, which could be caused partly by the artificial drainage in the regions, and partly because changes on the water table will be affected more by sporadic events. However, the model showed some overfitting so these results should be taken with caution. Additionally, adjustments should be

made to further improve the results, such as adding measurements on streams and lakes to add examples of flooded areas to the model. Finally, it is possible that the model is slightly biased due to the high number of observations in Copenhagen, and observations from other land cover types could further improve the model and its predictions.

Limitations of the data

There were several limitations in the implementation of this project in relation to the available data that impacted the results from the ML models and thus the predictions made based on this model. These limitations are mostly caused by issues regarding the data's quality and resolution.

First of all, the observations from the wells are monthly measurements of the groundwater level taken at a seemingly random day within the month, which might not be an ideal temporal resolution. For instance, a measurement of groundwater level could be taken at a very dry period of the month, which might not be reflected by the average monthly precipitation if it started raining just after this measurement was taken. This issue negatively affects the ML process making less accurate the relationship of both monthly temperature and monthly precipitation with the measurement of groundwater level.

These irregularities in the measurements could also be compromising the results since these monthly measurements are also being taken at different times for each of the different wells, not only making each time-series very irregular, but also creating irregularities among wells. For instance, the water level of two wells in the same area might have been taken at completely different days in the same month, with very different climatic conditions. However, because the climatic variables used in this project are monthly averages, the model will learn that the water levels from these two wells were taken under the same climatic circumstances, when in fact that might not have been the case. Therefore, any variance between the wells that could be caused by climatic factors might be attributed to some other factor by the model.

Moreover, it seems that some land cover types and areas of the region selected might be underrepresented. Most of the observations of the groundwater levels are located in either urban or agricultural areas, and from these more than 50% were taken in Copenhagen. Meanwhile, natural areas have very few observations, and areas like lakes and streams, or humid biomes like peatbogs, etc. have no observations at all. This can cause the ML model to predict incorrectly in some of the unknown or underrepresented areas, which could explain the possible overfit mentioned in the results, and could be limiting the predicting possibilities of the model, since it does not have the necessary data to learn what is happening in these undersampled locations.

Additionally, there are some limitations due to the unavailability of future predictions for a few of the independent variables. Some independent variables, such as the imperviousness, showed high importance in the RF model, but no future forecasts of these variables were available, so current data was used for the future predictions. Making these variables static for current and future scenarios, even though it was a necessity with the limitation of available data, is unrealistic and adds a degree of uncertainty to any predictions made by the model for the future SSPs selected.

Implications to society and decision makers

As previously explained, the fluctuations in the water level according to the selected model will be limited, but an overall trend seems to indicate that the water table will rise, especially during the winter. As mentioned, this can bring problems to human access to groundwater and cause damages to artificial structures such as buildings or pipes.

The drinking water supply in Denmark is entirely based on groundwater (Danish Nature Agency, 2012). With the high percentage of land used for agriculture, the Danish Government has determined that the groundwater in the region is vulnerable to nitrate pollution, a risk that is further increased with the rise of the water table (Danish Ministry of the Environment, 2021). The closer the water table is to the surface, the higher the risk of infiltrated pollutants in the soil getting in contact with the water, affecting the drinking water supply. Currently, there is a monitoring program to periodically check the quality of the groundwater at a national level, called NOVANA (National Monitoring Assessment Programme for the Aquatic and Terrestrial Environment) (Danish Ministry of the Environment, 2021), but further actions might be required in the future if the risk of pollution of the groundwater increases, such as further filtering of extracted water or restrictions to fertilizers and chemicals used in farms.

On the other hand, many infrastructures might be affected by a rise in the water table. As a start, supply and wastewater pipes and sewers located in the groundwater are subject to infiltration as a function of depth below the water table (Fung & Babcock, 2020). The water infiltrates through defects in the structures and issues from external pipe deterioration, resulting in pollutants in the soil infiltrating into the pipes with the water (Miljø Metropolen, 2011), and in a higher maintenance and treatment costs of the infrastructures (Fung & Babcock, 2020). Similarly, the foundation of buildings can also be affected by a rise in the water table, which will also require a higher maintenance and repair costs, but which will also increase the risks of deteriorating foundations and subsequent hazards if the damages are not prevented or treated (Miljø Metropolen, 2011).

Additionally, several studies have analysed the best groundwater levels for crops and vegetation, finding that the best suitable depth for the water table is between 1-2m (Kahlown et al., 2005;

Zipper et al., 2015). Changes in the water table depth, both higher and lower, will affect crops and farms, as well as natural ecosystems. Furthermore, a higher water table will be more likely to surpass this optimal depth in case of heavy precipitation events, and it would increase the risk of flooding (Koch et al., 2019). This could not only negatively impact ecosystems on the surface, but it could even translate into losses for the agricultural sector and possible risks to human's safety (Danish Nature Agency, 2012).

For these reasons, preventive measures should be taken regarding possible water level rise in the region. Drainage methodologies might be required in agricultural areas to maintain an optimal water level in places where the water table will be too close to the surface. Artificial infrastructures might also need to be monitored more regularly, and predictions on the location of submerged infrastructures could also be used for adaptation plans. Finally, as it is expected that precipitations will get stronger and more sporadic, it might be necessary to focus on prevention plans for big sporadic changes in the water table, including more accurate predictions based on storm events (see section 7. Future work).

The results obtained from the model can be used in these planning and adaptation processes, as they show both areas where the water levels will change, and how much. Knowing where the largest fluctuations on the water table will occur can give decision makers an idea of which areas will require more attention, and where to put more focus and resources to prevent damages and risks caused by changes in the water levels. Moreover, it can also serve to decide which methods and plans to put into action depending on where these changes will occur.

7. FUTURE WORK

Based on the results obtained from the models and the aforementioned limitations due to the quality and accessibility of the data, some measures could be followed to improve the models in future works. For instance, obtaining new data with a better spatial distribution of the groundwater observations, with more measurements on different types of land cover, could bring new information to the model and improve the predictions. This would also include measurements at streams, lakes, and other flooded areas, so the model can have better examples of water bodies and predict better in those cases.

On the same line, the temporal resolution of the groundwater observations might not be sharp enough to account for some specific changes in the water table depth. Since future climate predictions show that heavy storm events will be more frequent, and that these will affect the water table levels and increase the risk of flooding, monthly measurements of groundwater levels and climate variables are not enough. For a better visualization of the effect of these storm events, it would be best to use samples collected at shorter times, such as daily groundwater measurements, along with local measurements of precipitation and other climatic factors. In this way, a model could be trained to catch such local changes on the groundwater level and make more accurate predictions for possible sporadic precipitation events.

On the other hand, the study area is highly affected by human activities that can impact the levels of the groundwater. Having data on factors such as irrigation or artificial drainage that are affecting the observations would also give the model the information needed to account for the effects of these and could help in making more accurate predictions in areas highly affected by human activities, such as agricultural fields or cities.

8. CONCLUSION

For this project, and with the data available, RF outperformed the SVM and ANN models, giving the most robust predictions, providing sturdy outcomes and resulting in successful metrics in regards to the criteria defined by the literature.

On the other hand, the model showed possible signs of overfitting, which together with the fact that the study area is highly anthropogenic, the unevenness of the data, and considering that the model does not account for sporadic precipitation events, could be biasing its results and predictions. Because of these reasons, conclusions based on the model should be made cautiously, and considering the discussed circumstances and context.

The analysis of the results from the RF models shows that the water levels will not change that much in future climate scenarios. However, predictions show that it will rise slightly, mostly in the order of 0-0.25m, especially during winter, increasing the areas where the water table will be less than 1m deep from the surface in Hovedstaden.

This is especially relevant since even slight fluctuations in the water table can have notorious repercussions, and should be accounted for and managed. Higher water levels can affect the drinking water supply and put pressure on artificial infrastructures such as pipes or buildings, they can cause economic losses by damaging crops, and they can increase the risk of flooding with the subsequent hazards posed to humans.

The work done in this project can be used to visualize areas where the water levels are expected to change, as well as to give an overview of how big the changes will be. This can provide better perspectives when planning and adapting for both climate change and for the impacts that changes in the water table will cause, allowing for decision makers to be ahead of situations where the risks might be high.

Future work should focus on obtaining a more spatially and temporally spread representation of samples, accounting for human modifications and impacts if possible, and ideally improve the measurement collection on the wells, including local measurements of climatic variables.

Most limitations to this project were due to the irregularities in the data, as well as for variables with no available data. With future work improving the sampling of the data and the availability of independent variables, the model used in this project could be replicated and adapted to this new, better quality data, making improvements on the model and its results, providing even more accurate and reliable predictions.

Additionally, the approaches and models developed with this thesis could be replicated and applied to different study areas, allowing for the possibility to extend this model to a national level, improving the prevention and adaptation plans in Denmark and providing a more global overview of future water level predictions to more efficiently handle future climate change scenarios.

BIBLIOGRAPHY

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B. and Greve, M. H. (2013) *High-Resolution 3-D Mapping of Soil Texture in Denmark*. Soil Sci. Soc. Am. J. **9:** e105519 https://doi.org/10.2136/sssaj2012.0275

Analytics Vidhya (2020) *Decision Tree vs. Random Forest - Which algorithm should you use?* Retrieved May 2021 from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Bates, B., Kundzewicz, Z., Wu, S. & Palutikof, J. 2008: Climate change and water. Intergovernmental Panel on Climate Change, Technical Paper 6, 200 pp. Geneva: IPCC.

Bowes, B. D., Sadler, J. M., Morsy, M. M. Behl, M. and Goodall, J. L. (2019) Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks. *Water* **11**(1098).

Breiman, L. (2001) Random forests. *Mach. Learn.* **45**: 5-32.

Brownlee, J. (2014) *How To Estimate Model Accuracy in R Using The Caret Package*. Retrieved April 2021 from Machine Learning Mastery: https://machinelearningmastery.com/how-to-estimate-model-accuracy-in-r-using-the-caret-package/

Brownlee, J. (2018) *A gentle introduction to k-fold cross-validation*. Retrieved April 2021 from Machine Learning Mastery: https://machinelearningmastery.com/k-fold-cross-validation/

Brownlee, J. (2019) *Comparing classical and machine learning algorithms for time series forecasting*. Retrieved May 2021 from Machine Learning Mastery: https://machinelearningmastery.com/findings-comparing-classical-and-machine-learning-methods-for-time-series-forecasting/

Brownlee, J. (2020) *Random forest for time series forecasting.* Retrieved May 2021 from Machine Learning Mastery: https://machinelearningmastery.com/random-forest-for-time-series-forecasting/

Brutsaert, W. (2005) *Hydrology, an introduction.* Cambridge University Press, NY.

Catani, F., Lagomarsino, D., Segoni, S. and Tofani, V. (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat. Hazards Earth Syst. Sci.* **13**: 2815-2831.

Chauhan, N. S. (2019) *Introduction to Artificial Neural Networks (ANN)*. Retrieved May 2021, from Towards Data Science: https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9

Chen, C., He, W., Zhou, H., Xue, Y. and Zhu, M. (2020) A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Scientific Reports* **10**(3904): 1-13.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J. L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi,C., Weaver, A.J. and Wehner, M. (2013) Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Bïhner, J. (2015) System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev. 8, 1991-2007. doi:10.5194/gmd-8-1991-2015.

Danish Agriculture & Food Council (2019) Denmark - a food and farming country. *Landbrug Fødevarer*. Danish Agriculture & Food Council.

Danish Ministry of the Environment (2021): Groundwater Monitoring in Denmark. *The Danish action plan for promotion of eco-efficient technologies.* Retrieved April, 2021 from: Geus, Danish

Ministry of the Environment. https://eng.ecoinnovation.dk/media/mst/8051419/CASE_Groundwatermonitoring_print.pdf
Danish Nature Agency (2012) Mapping climate change - Barriers and opportunities for action. *Task Force on Climate Change Adaptation*.
Data Carpentry (2019). Intro to R and RStudio for Genomics: Introducing R and RStudio IDE Retrieved May 2021 from Data Carpentry: https://datacarpentry.org/genomics-rintro/01-introduction/index.html
Dixon, M. F., Polson, N. G. and Sokolov, V. O. (2019) Deep learning for spatio-temporal modeling: Dynamic traffic flows and high frequency trading. *Applied Stochastic Models in Business and Industry* **35**: 788-807.
Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitao, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. and Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**(1): 27-46.
Fahimi, F., Yaseen, Z. M., El-Shafie, A. (2017) Application of soft computing based hybrid models in hydrological variables modeling: A comprehensive review. *Theor. Appl. Climatol.* **128**: 875-903.
Fick, S. E. and Hijmans, R. J. (2017) WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**(12): 4302-4315.
Frankenfield, J. (2020) *Artificial Neural Network (ANN).* Retrieved May 2021 from Investopedia: https://www.investopedia.com/terms/a/artificial-neural-networks-ann.asp
Fung, A. and Babcock, R. (2020) A flow-calibrated method to project groundwater infiltration into coastal sewers affected by sea level rise. *Water* **12**(1934): 1-13.
GEUS (2017) Groundwater monitoring 1989-2017 - Summary. Danish Ministry of Energy, Utilities and Climate.
GEUS (2021a) *National boringsdatabase (Jupiter).* Retrieved from: De Nationale Geologiske Undersøgelser for Danmark og Grønland: https://www.geus.dk/produkter-ydelser-og-faciliteter/data-og-kort/national-boringsdatabase-jupiter
GEUS (2021b) *Dokumentation af PCJupiterXL tabeller og koder.* Retrieved from: De Nationale Geologiske Undersøgelser for Danmark og Grønland: https://data.geus.dk/tabellerkoder/index.html?tablename=WATLEVEL
GEUS (2021c) *Download PCJupiter.* Retrieved April 2021 from De Nationale Geologiske Undersøgelser for Danmark og Grønland: https://data.geus.dk/JupiterWWW/downloadpcjupiter.jsp?xl=1
GEUS (2021d) Download jordartskort. Retrieved April 2021 from De Nationale Geologiske Undersøgelser for Danmark og Grønland: https://www.geus.dk/produkter-ydelser-og-faciliteter/data-og-kort/danske-kort/download-jordartskort
Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., and Cardenas, M. B. (2016) The global volume and distribution of modern groundwater. *Nat. Geosci.* **9**: 161-167. https://doi.org/10.1038/ngeo2590.
Goonewardana, H. (2019) *PCA: Application in Machine Learning*. Retrieved May 2021, from Medium Apprentice Journal: https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db
Guergachi, A. and Boskovic, G. (2008) System models or learning machines? *Appl Math Comp* **204**:553–567.
Guio Blanco, C. M., Brito Gomez, V. M., Crespo, P. and Ließ, M. (2018) Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. *Geoderma* **316**: 100–114. https://doi.org/10.1016/j.geoderma.2017.12.002
Güllü, M., & Yilmaz, I. (2011). Application of Back Propagation Artificial Neural Network for Modelling Local GPS/Levelling Geoid Undulations: A Comparative Study.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. (2014) Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *International Journal of Climatology* **34**: 623-642. doi:10.1002/joc.3711

Hausfather, Z. (2019) *CMIP6: the next generation of climate models explained.* Retrieved April 2021 from CarbonBrief: https://www.carbonbrief.org/cmip6-the-next-generation-of-climate-models-explained

Hedley, C.B., Roudier, P. , Yule, I.J., Ekanayake, J. and Bradbury, S. (2013) Soil water status and water table depth modelling using electromagnetic surveys for precision irrigation scheduling. *Geoderma* **199**: 22-29. https://doi.org/10.1016/j.geoderma.2012.07.018

Hengl, T., Nussbaum, M. and Wright, M. N. (2018) Random Forest for spatial data. *GeoMLA.*

Henriksen, H. J., Troldborg, L., Nyegaard, P., Sonnenborg, T. O., Refsgaard, J. C. and Madsen, B. (2003) Methodology for construction, calibration and validation of a national hydrological model for Denmark. *J. Hydrol.* **280**: 52-71. https://doi.org/10.1016/S0022-1694(03)00186-0

Henriksen, H. J., Højberg, A. . L., Olsen, M., Seaby, L. P., van der Keur, P., Stisen, S., Troldborg, L., Sonnenborg, T. O., and Refsgaard, J. C. (2012) Klimaeffekter på hydrologi og grundvand (Klimagrundvandskort), Geological Survey of Denmark and Greenland. Copenhagen, Denmark: 1-116.

Hijmans, R. J. (2020). *Package 'raster'*. Retrieved May, 2021, from The Comprehensive R Archive Network: https://cran.rproject.org/web/packages/raster/raster.pdf

Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A. and Vaccari, M. (2020) Groundwater prediction using machine-learning tools. *Algorithms* **13**(300). https://doi.org/10.3390/a13110300

IBM Cloud Education (2020) *Machine Learning*. IBM. Accessible: https://www.ibm.com/cloud/learn/machine-learning

Jankowfsky, S., Branger, F., Braud, I., Rodriguez, F., Debionne, S., and Viallet, P. (2014) Assessing anthropogenic influence on the hydrology of small peri-urban catchments: Development of the object-oriented PUMMA model by integrating urban and rural hydrological models. *J. Hydrol.* **517**: 1056-1071. https://doi.org/10.1016/j.jhydrol.2014.06.034

Jebens, M., Sørensen, C. S. and Piontkowitz, T. (2016). Danish risk management plans of the EU Floods Directive. E3S Web of Conferences, 7.

Jorgensen, L. F. and Stockmarr, J. (2008) Groundwater monitoring in Denmark: characteristics, perspectives and comparison with other countries. *Hydrogeology Journal* **17**: 827-842.

Jolliffe, I. T. and Cadima, J. (2016) Principal component analysis: a review and recent developments. *Phil.Trans.R.Soc.A* **374**: 20150202. http://dx.doi.org/10.1098/rsta.2015.0202

Kahlown, M. A., Ashraf, M. and Zia-Ul-Haq (2005) Effect of shallow groundwater table on crop water requirements and crop yields. *Agr. Water Manage*. **76**: 24-35. https://doi.org/10.1016/j.agwat.2005.01.005,.

Kassambra, A. and Mundt, F. (2020) *factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* Retrieved May, 2021 from: CRAN - R project: https://cran.r-project.org/web/packages/factoextra/index.html

Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P. and Mladenić, D. (2018) Groundwater modeling with machine learning techniques: Ljubljana polje Aquifer. *Proceedings* **2**(11): 697. https://doi.org/10.3390/proceedings2110697

Kent State University (2021) *Pearson Correlation*. Retrieved May 2021 from Kent State University: https://libguides.library.kent.edu/SPSS/PearsonCorr

Kidmose, J., Refsgaard, J. C., Troldborg, L., Seaby, L. P., and Escrivà, M. M.: Climate change impact on groundwater levels: ensemble modelling of extreme values, Hydrol. Earth Syst. Sci., 17, 1619–1634, https://doi.org/10.5194/hess-17-1619-2013, 2013.

Koch, J., Berger, H., Henriksen, H. J. and Sonnenborg, T. O. (2019) Modelling of the shallow water table at high spatial resolution using random forests. *Hydrol. Earth Syst. Sci.* **23**: 4603-4619.

Kuhn, M. (2008). Building Predictive Models in R Using the Caret Package. J*ournal of Statistical Software* **28**(5).

Kuhn, M. (2019) *The caret package*. Retrieved May, 2021 from: https://topepo.github.io/caret/index.html

Kuhn, M. and Johnson, K. (2013) Applied Predictive Modeling. New York, USA: Springer.
Lakshamanan, V., Gilleland, E., McGovern, A. and Tingley, M. (2015) Machine Learning and data mining approaches to climate science. Springer, NY.
LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. Nature. **521**: 436–444. https://doi.org/10.1038/nature14539
Levin, G. (2006) Dynamics of Danish agricultural landscape and role of organic farming. *National Environmental Research Institute - Danish Ministry of the Environment.*
Linting, M., Meulman, J., Groenen, P. and Van der Kooij, A. (2007). Nonlinear Principal Components Analysis: Introduction and Application. In: Psychological Methods. American Psychological Association.
Maier, H. R. and Dandy, G. C. (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. https://doi.org/10.1016/S1364-8152(99)00007-9
Markstrom, S. L., Niswonger, R. G., Regan, R. S., Prudic, D. E. and Barlow, P. M. (2008) *GSFLOW - Coupled Ground-Water and Surface-Water Flow Model Based on the Integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005)*. Report No. 6-D1, 240.
Meyer, H. (2018) Introduction to Cast. *R-Project*.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M. and Nauss, T. (2018) Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* **101**: 1-9.
Miljø Metropolen (2011) Copenhagen climate adaptation plan. *Copenhagen Carbon Neutral by 2025.* Miljø Metropolen.
Miljøstyrelsen (2020) *Indberetning og godkendelse af vandforsyningsdata (Jupitervejledningen).* Retrieved from: Miljøministeriet: https://mst.dk/service/nyheder/nyhedsarkiv/2020/maj/indberetning-og-godkendelse-af-vandforsyningsdata-jupitervejledningen/
Mohanty, S., Jha, M.K., Kumar, A. and Panda, D.K. (2013) Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi–Surua Inter-basin of Odisha, India. *J. Hydrol*. **495**: 38-51.
Møller, A. B., Iversen, B. V., Beucher, A., and Greve, M. H. (2017) Prediction of soil drainage classes in Denmark by means of decision tree classification. *Geoderma* **352**: 314-329. https://doi.org/10.1016/j.geoderma.2017.10.015
Nourani, V., Mehdi, K. and Akira, M. (2009) A multivariate ANN-wavelet approach for rainfall–runoff modeling. *Water Resource Management* **23**: 2877-2894.
OECD (2013), *Water and Climate Change Adaptation: Policies to Navigate Uncharted Waters*, OECD Studies on Water, OECD Publishing, Paris.
PostgreSQL (2021) *About*. Retrieved from: The PostgreSQL Global Development Group: https://www.postgresql.org/about/
Rieuf, E. (2017) *How to interpret R-squared and Goodness of fit in regression analysis.* Retrieved May 2021 from Data Science Central: https://www.datasciencecentral.com/profiles/blogs/regression-analysis-how-do-i-interpret-r-squared-and-assess-the
Robbach, P. (2018) *Neural Networks vs. Random Forest - Does it always have to be deep learning?* Retrieved May 2021 from Frankfurt School of Finance and Management: https://blog.frankfurt-school.de/neural-networks-vs-random-forests-does-it-always-have-to-be-deep-learning/
Rolnick, D., Donti, P., Kaack, L., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E., Mukkavilli, K., Kording, K., Gomes, C., Ng, A., Hassabis, D., Platt, J. and Bengio, Y. (2019). Tackling Climate Change with Machine Learning. https://www.researchgate.net/publication/333773164_Tackling_Climate_Change_with_Machine_Learning

RStudio (2021). *RStudio IDE Features*. Retrieved May 2021, from RStudio: https://rstudio.com/products/rstudio/features/
Sayad, S. (2021) *An introduction to data science*. Retrieved May 2021 from Saedsayad: https://www.saedsayad.com/data_mining_map.htm
Scikit-learn (2020a) *Cross-validation: evaluating estimator performance*. Retrieved April 2021 from Scikit-learn: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation
Scikit-learn (2020b) *Metrics and scoring: quantifying the quality of predictions*. Retrieved May 2021 from Scikit-learn: https://scikit-learn.org/stable/modules/model_evaluation.html
Seneviratne, S.I., N. Nicholls, D. Easterling, C.M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K. McInnes, M. Rahimi,M. Reichstein, A. Sorteberg, C. Vera, and X. Zhang, (2012) Changes in climate extremes and their impacts on the naturalphysical environment. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen,M. Tignor, and P.M. Midgley (eds.)]. A Special Report of Working Groups I and II of the Intergovernmental Panel on ClimateChange (IPCC). Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 109-23.
Shumway, R. H. and Stoffer, D. S. (2011) Time series analysis and its applications. Third edition. Springer. USA.
Singh, A. (2014) Groundwater resources management through the applications of simulation modeling: A review. *ScTEn* **499**: 414-423. https://doi.org/10.1016/j.scitotenv.2014.05.048
Singh, R. (2019) *Where deep learning meets GIS*. ESRI. Available: https://www.esri.com/about/newsroom/arcwatch/where-deep-learning-meets-gis/
Solomatine, D.P. and Ostfeld, A. (2008) Data-driven modelling: Some past experiences and new approaches. *J. Hydroinform* **10**: 3-22.
Solomatine, D. P. and Shrestha, D. L. (2009) A novel method to estimate model uncertainty using machine learning techniques. *Water Resour Res* **45**.
Statistics Denmark (2017), Geography, environment and energy: Statistical Yearbook 2017.
Statistics Denmark (2021) *AREALDK: Land by land cover, region and unit.* Retrieved from StatBank Denmark: Geography, environment and energy: https://www.statbank.dk/statbank5a/SelectVarVal/saveselections.asp
STHDA (2017) *Principal Component Methods in R: Practical Guide*. Retrieved May 2021 from: Statistical tools for high-throughput data analysis: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/#access-to-the-pca-results
STHDA (2021) *Regression Analysis Essentials For Machine Learning.* Retrieved May 2021 from Statistical tools for high-throughput data analysis: http://www.sthda.com/english/wiki/regression-analysis-essentials-for-machine-learning
Styrelsen for Dataforsyning og Effektivisering (2021a) *Dokumentation.* Retrieved from: HIP - Hydrologisk Informations- og Prognosystem: https://hip.dataforsyningen.dk/docs#future/7/720396.8/6180507.2/0/b02/1121795191264/day/
Styrelsen for Dataforsyning og Effektivisering (2021b) *HIP*. Retrieved from: HIP - Hydrologisk Informations- og Prognosystem: https://hip.dataforsyningen.dk/#realtime/2/600000/6225000/0/b01/1125477661289/day/
Ticlavica, A. M. and Torres, A. (2012) Data driven models and Machine Learning approach in water resources systems. *Utah Water Research Laboratory*.
Woldeamlak, S., Batelaan, O. & de Smedt, F. (2007) Effects of climate change on the groundwater system in the Grote-Nete catchment, Belgium. *Hydrogeology Journal* **15**: 891-901.
WorldClim (2020). WorldClim. Retrieved April 2021, from Downscaling future and past climate data from GCMs: https://worldclim.org/data/downscaling.html
Wuebbles, D. J., Fahey, D. W., Hibbard, K. A., Dokken, D. J., Stewart, B. C. and Maycock, T. K. (2017) Climate Science Special Report: Fourth National Climate Assessment, Volume I. Global Change Research Program: Washington: DC, USA.

Zipper, S. C., Soylu, M. E., Booth, E. G. and Loheide, S. P. (2015) Untangling the effects of shallow groundwater and soil texture as drivers of subfield-scale yield variability, *Water Resour. Res.* **51**: 6338-6358. https://doi.org/10.1002/2015WR017522

Wickham, H., François, R., Henry, L. and Müller, K. (2021) dplyr: A Grammar of Data Manipulation. Retrieved May 2021 from: CRAN - R project: https://CRAN.R-project.org/package=dplyr

Worldbank (2021) *Climate data: Historical*. Retrieved from Worldbank: https://climateknowledgeportal.worldbank.org/country/denmark/climate-data-historical

Zhou, V. (2019) *Machine Learning for Beginners: An Introduction to Neural Networks.* Retrieved May 2021 from Towards Data Science: https://towardsdatascience.com/machine-learning-for-beginners-an-introductionto-Neural-networks-d49f22d238f9

Annex 1. Github repository for the project

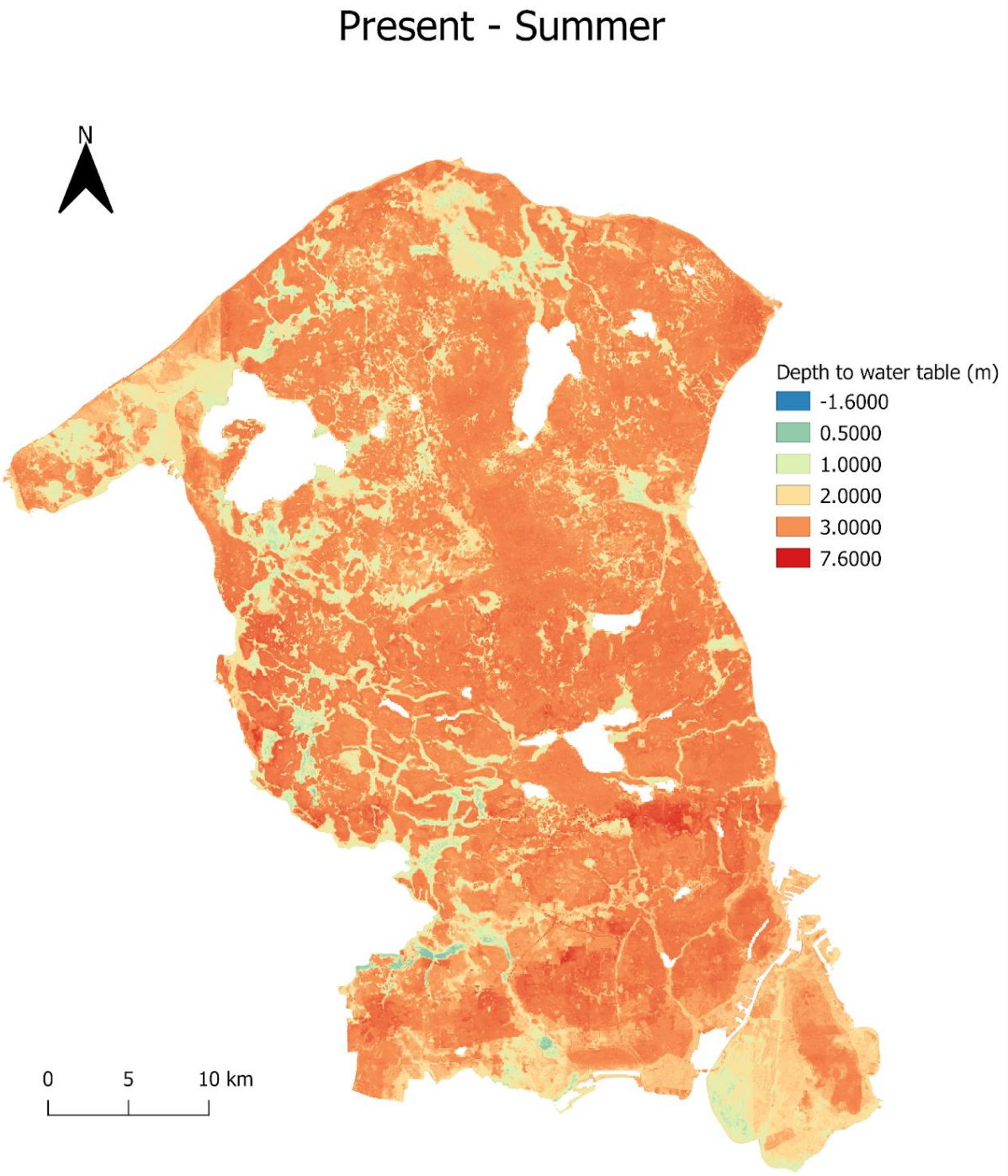
The R scripts utilized for both pre-processing the data and running the ML algorithms can be found in the following repository in github:

https://github.com/RebeQuiGon/Master_Thesis_Groundwater

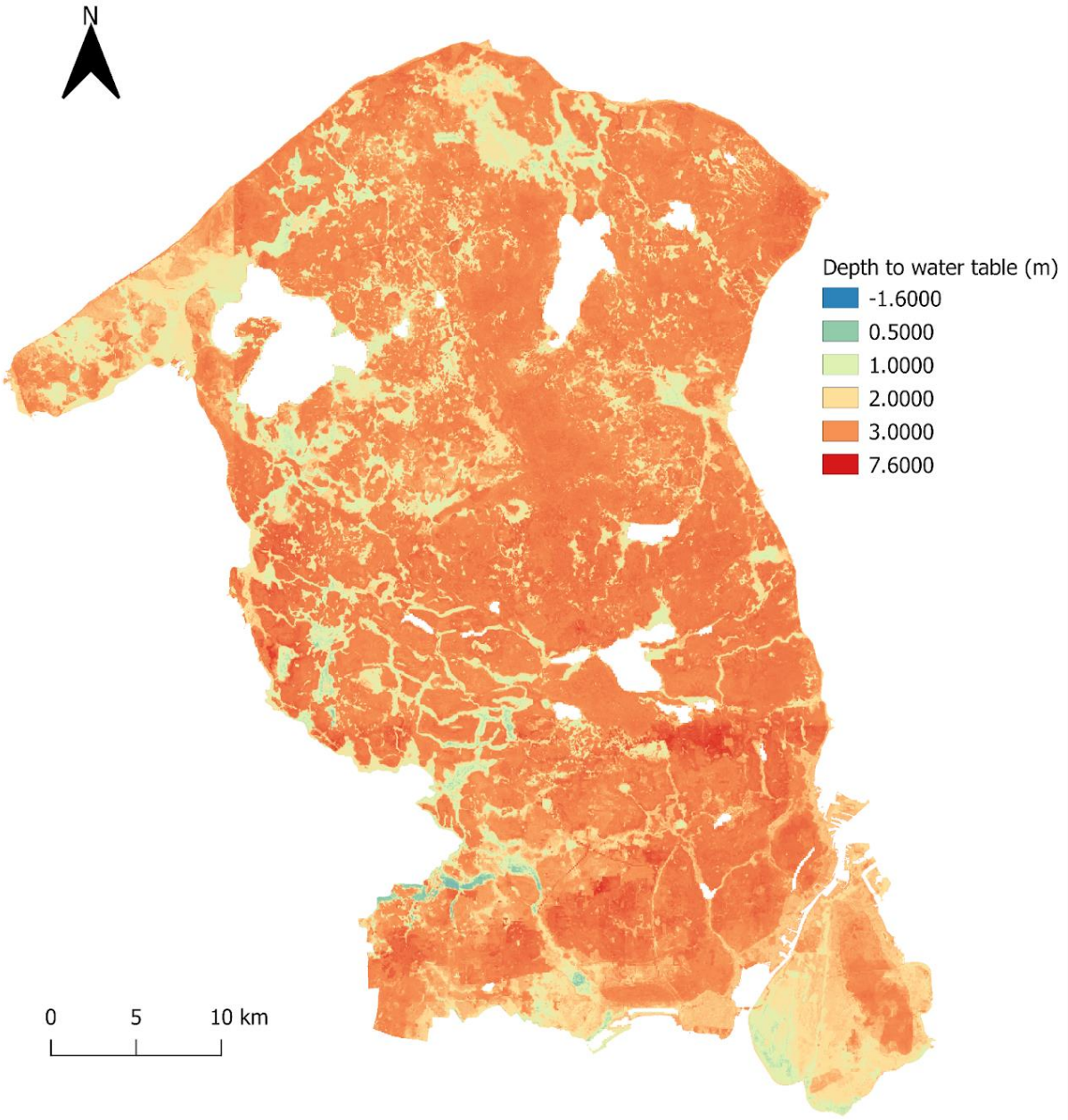
The repository contains two scripts, one showing the process for pre-processing all the data and another one for the ML procedures and the predictions made with the RF model.

The data is not included in the repository because permission was required for some of the layers. Moreover, the size of some of the files exceeds the maximum set by github. Nevertheless, the script provides for a replicable process, and the same results can be obtained after acquiring the data described in this project.

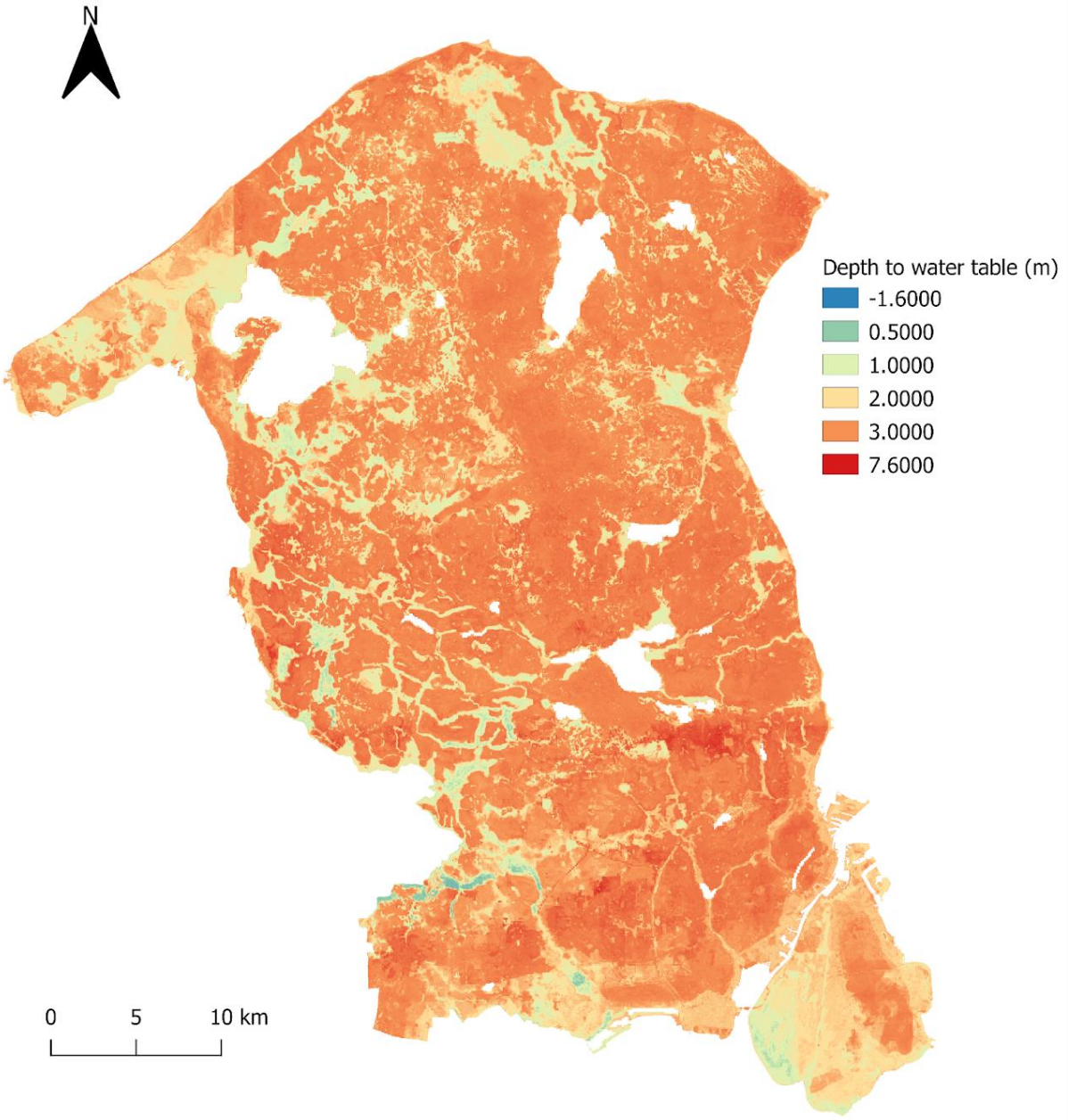
Annex 2. Prediction maps



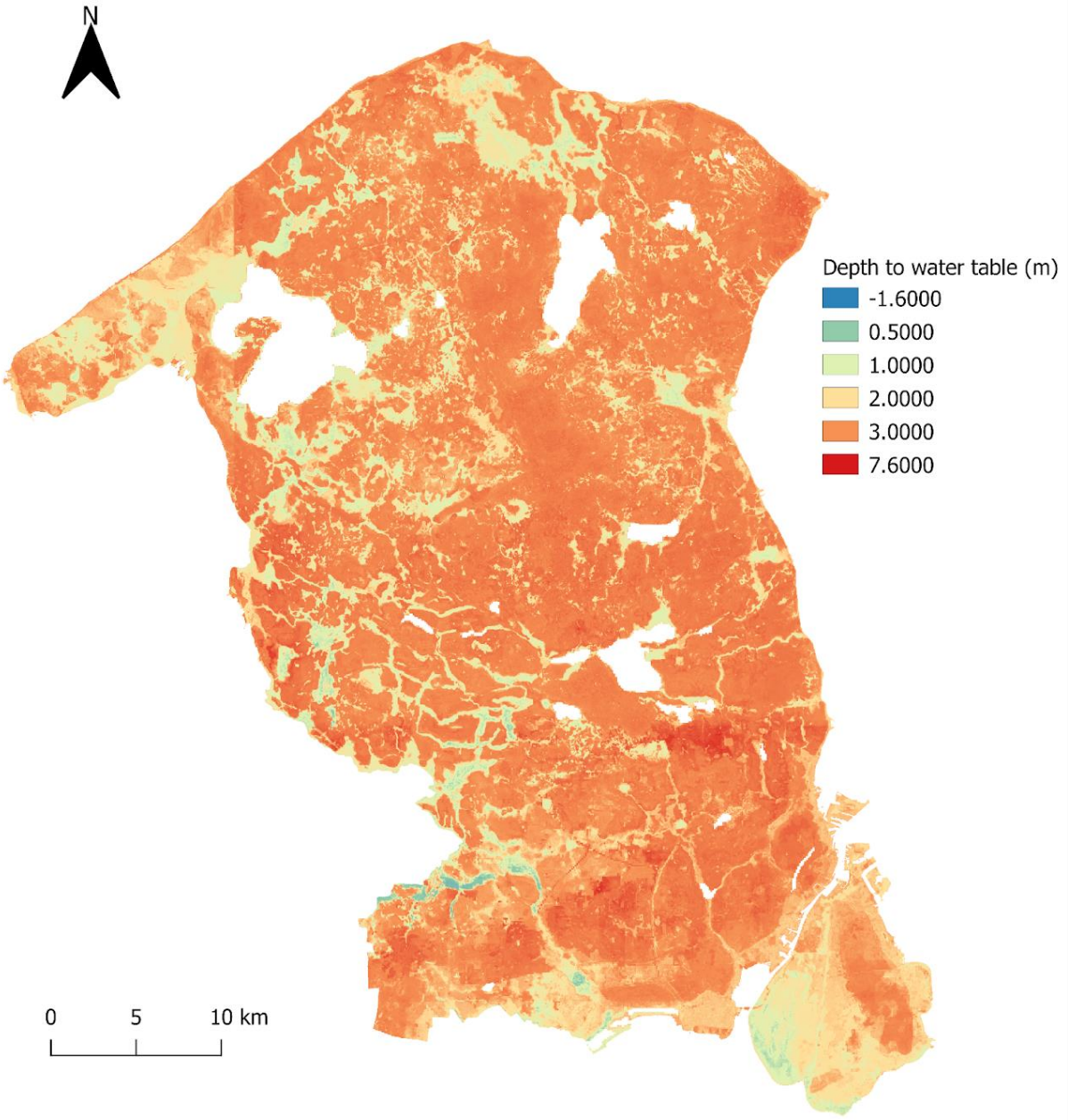
SSP2.45 - Summer



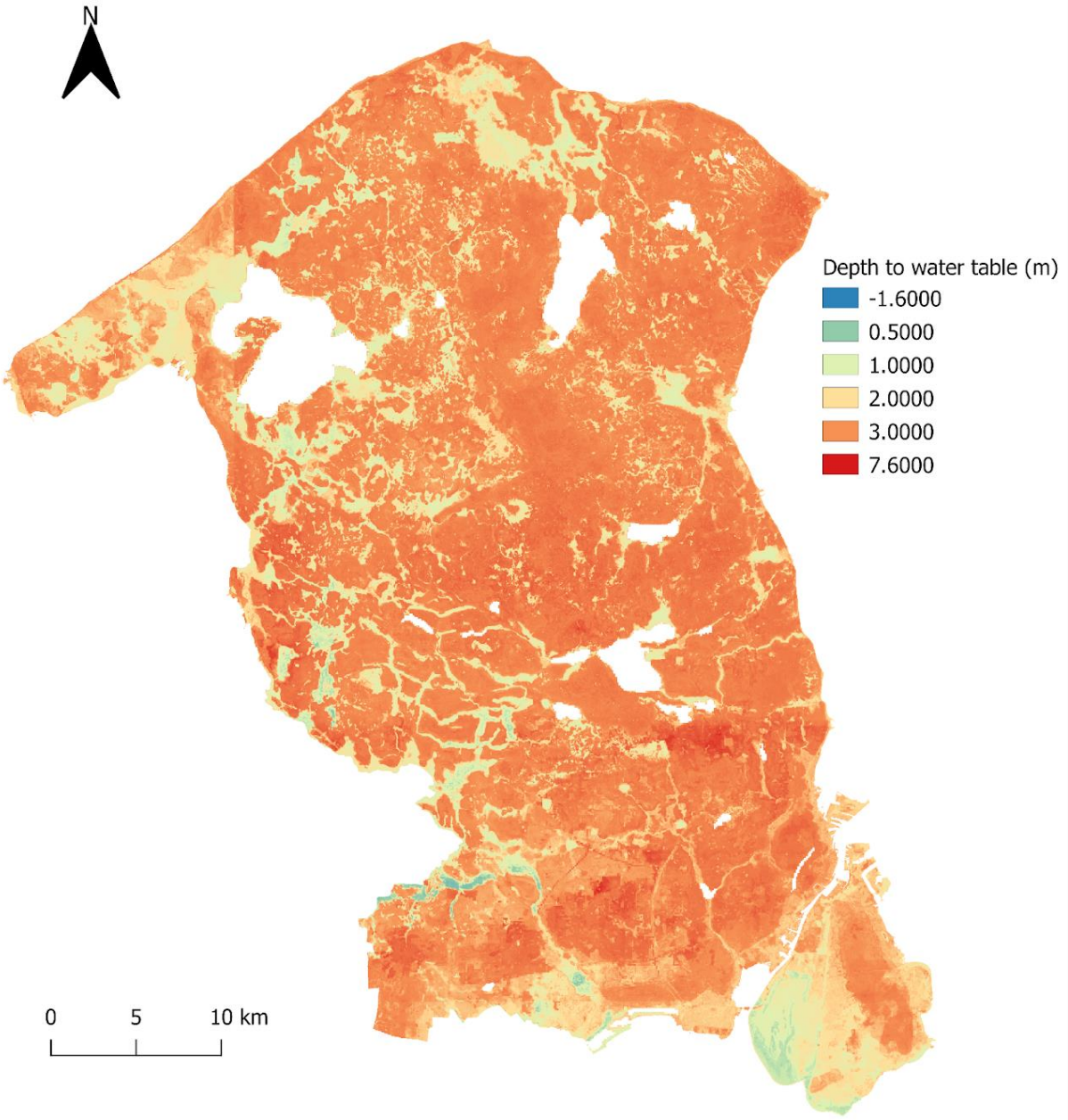
SSP3.70 - Summer



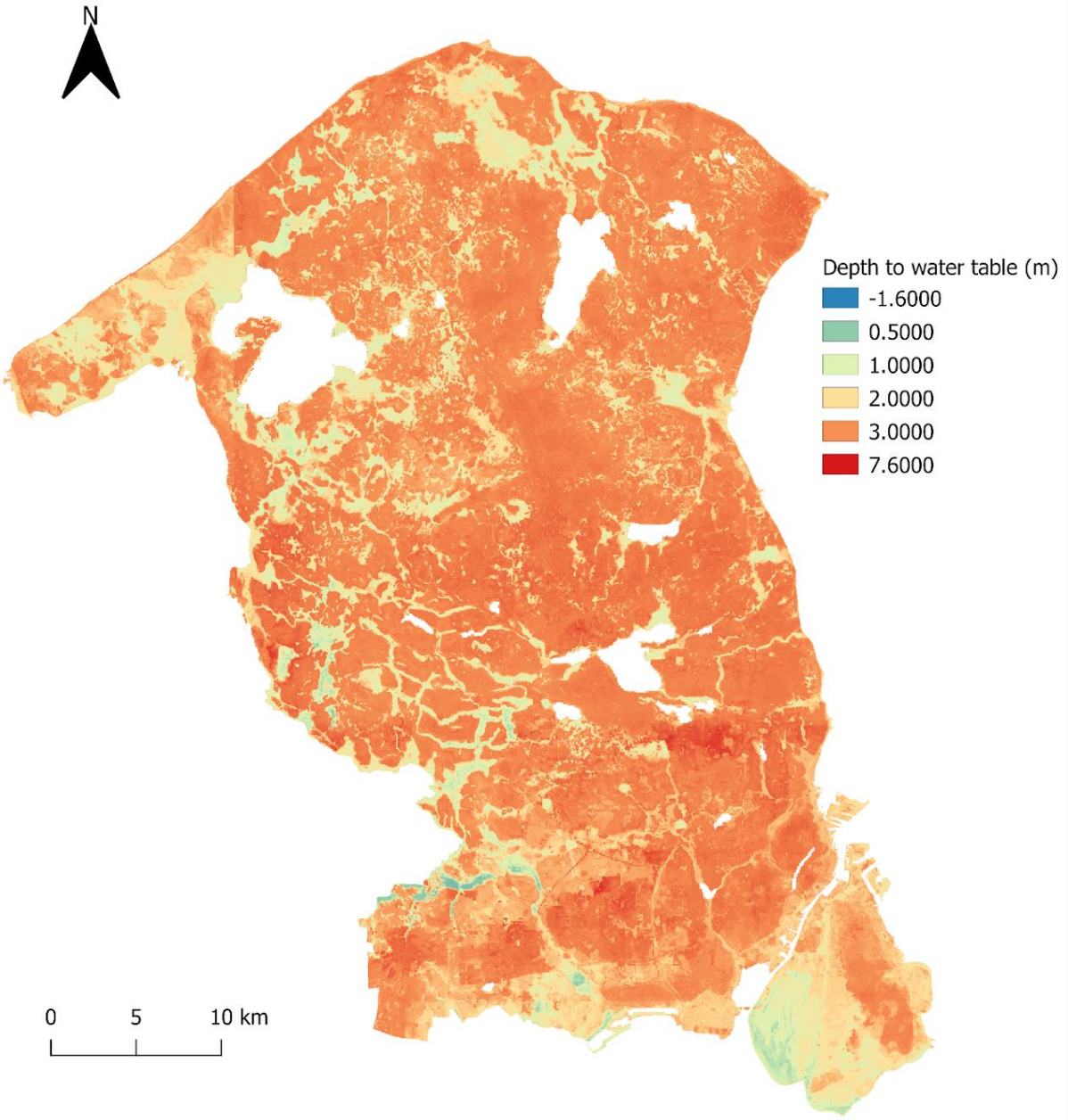
SSP5.85 - Summer



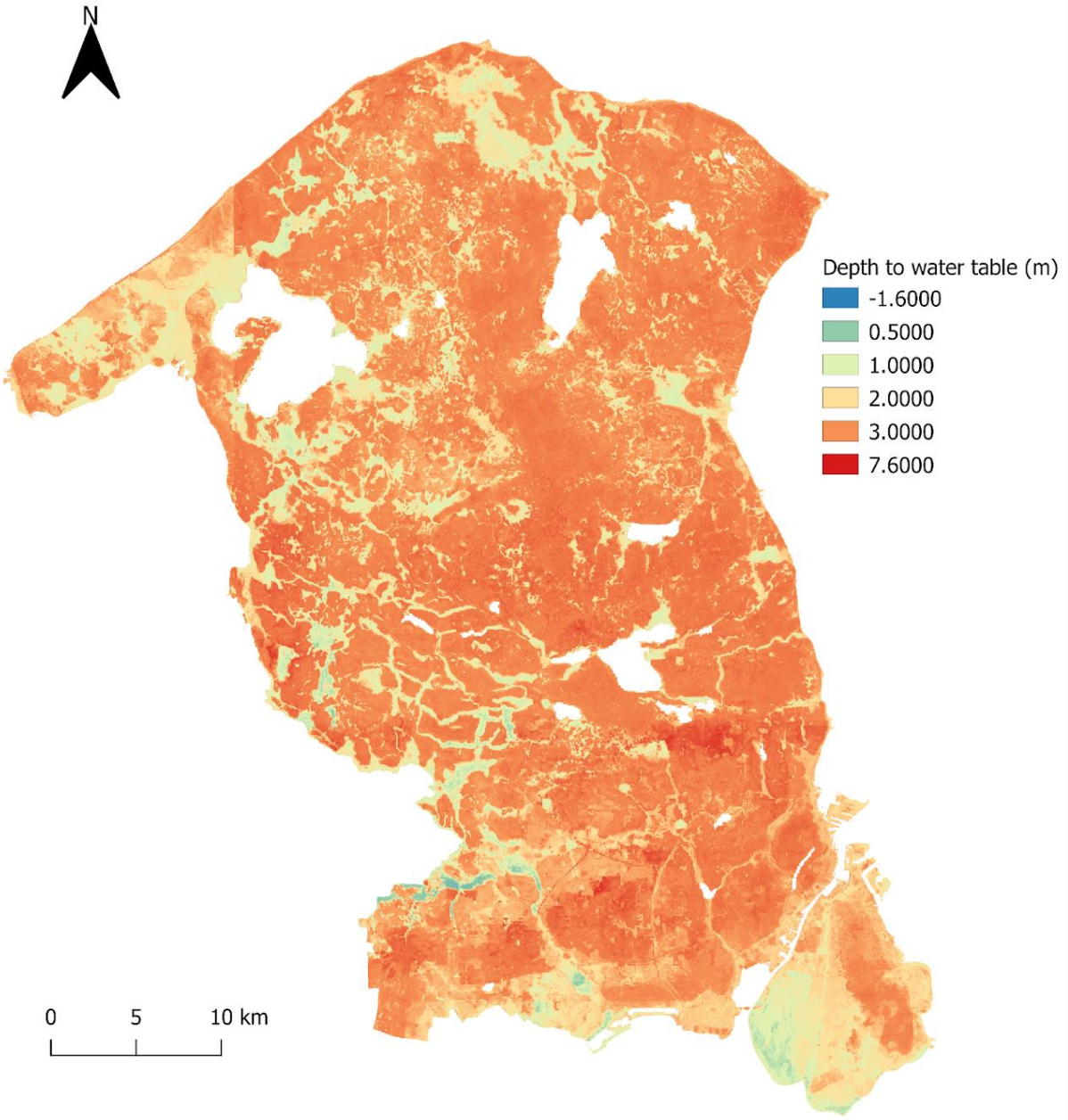
Present - Winter



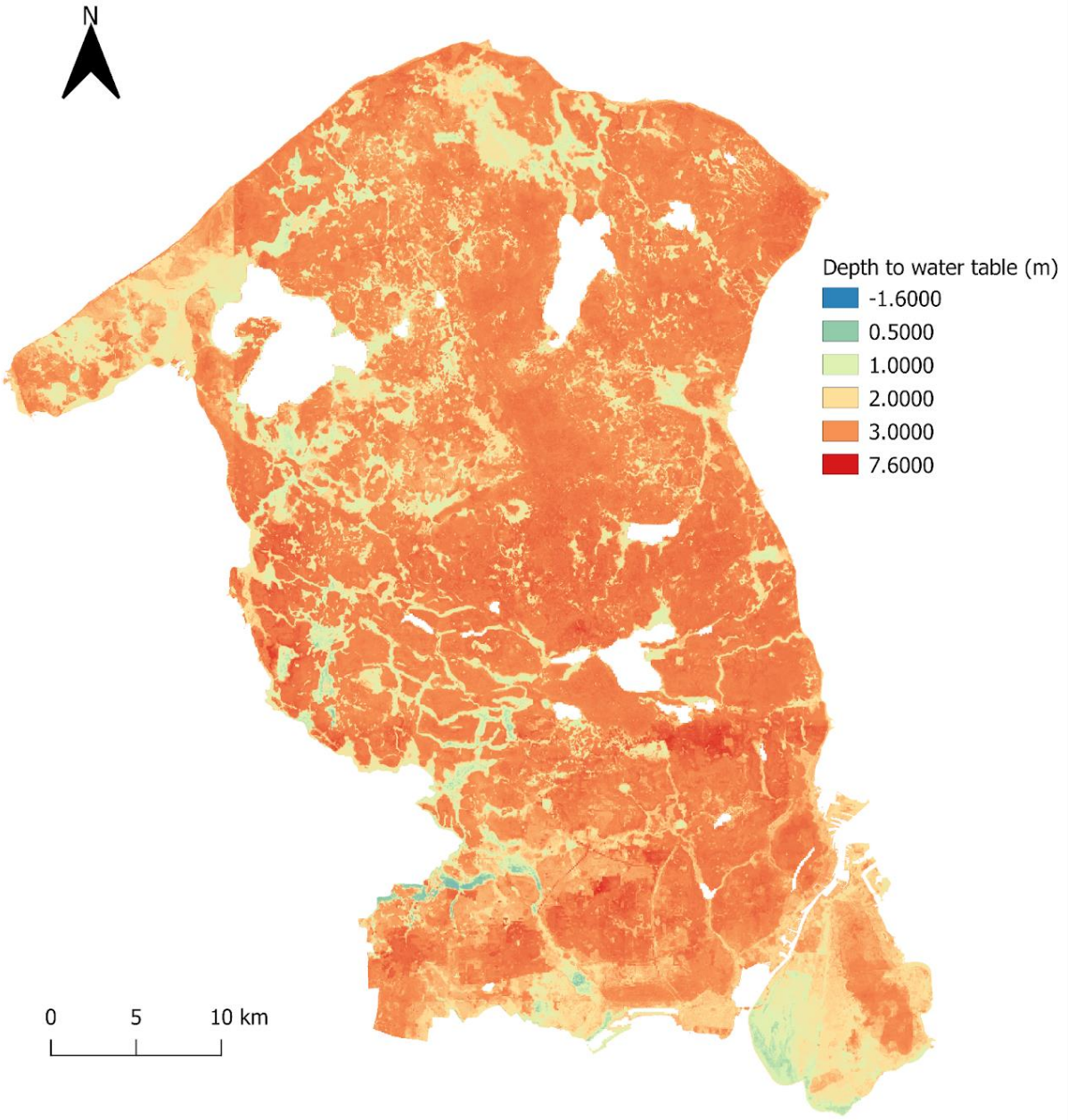
SSP2.45 - Winter



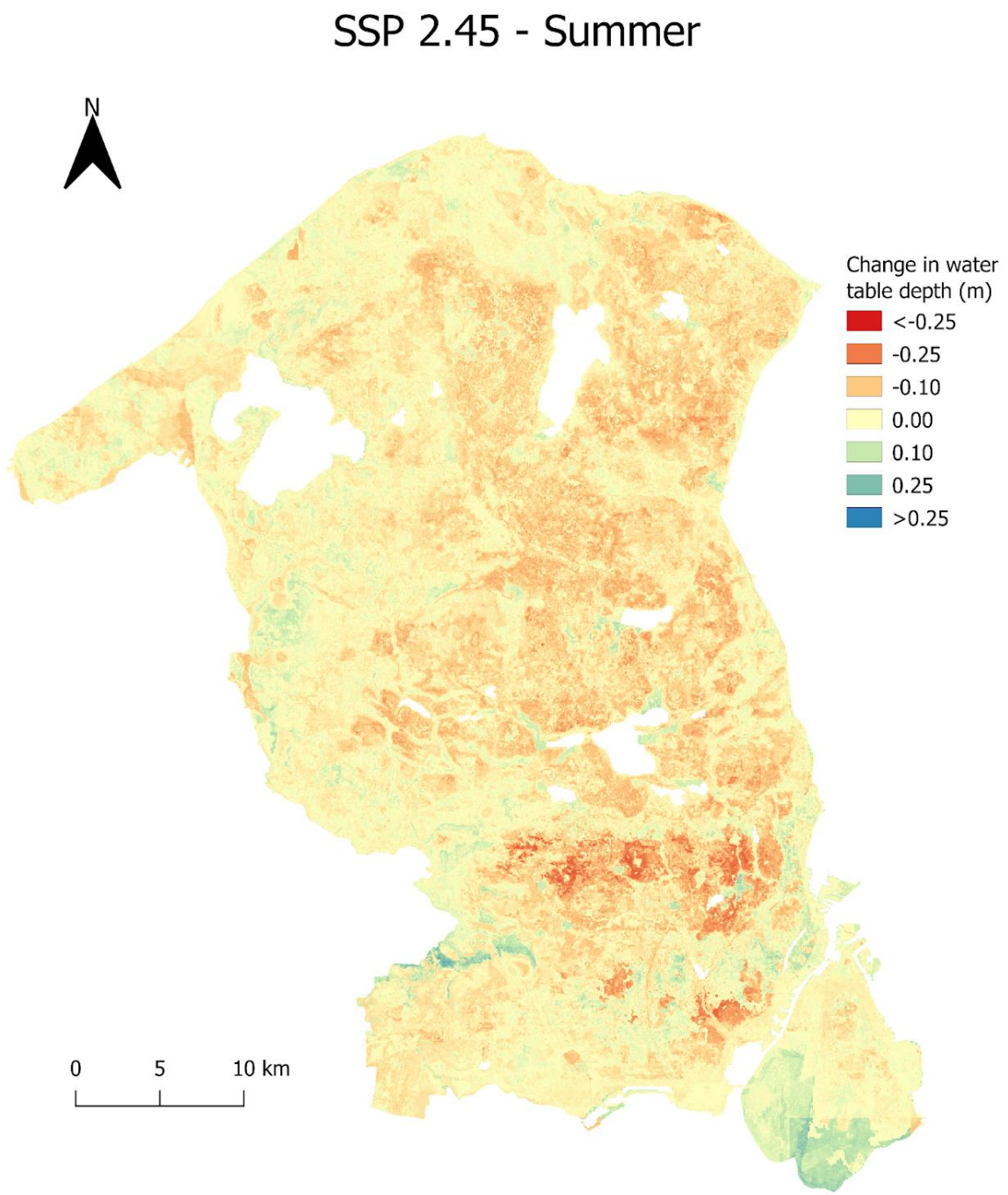
SSP3.70 - Winter

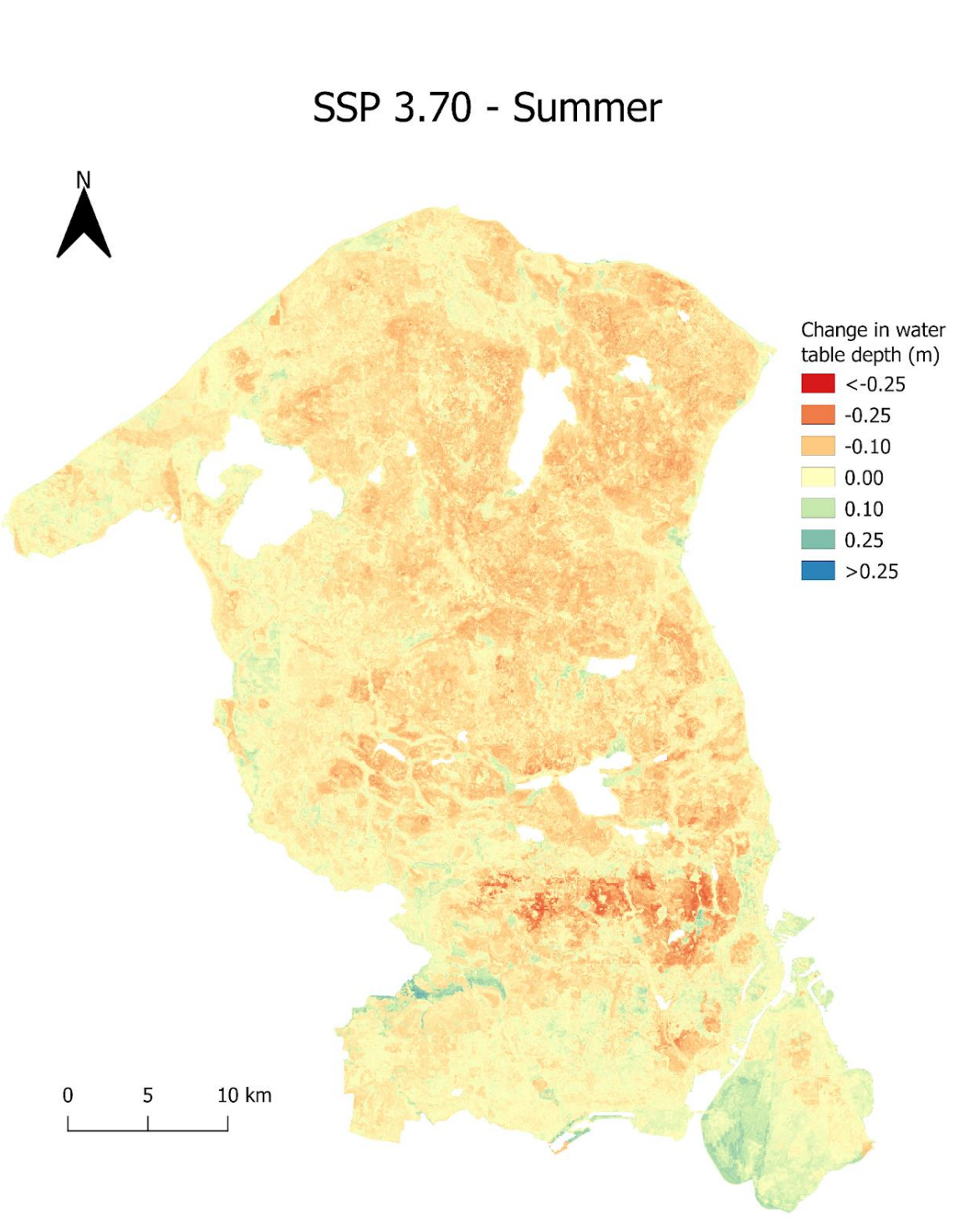


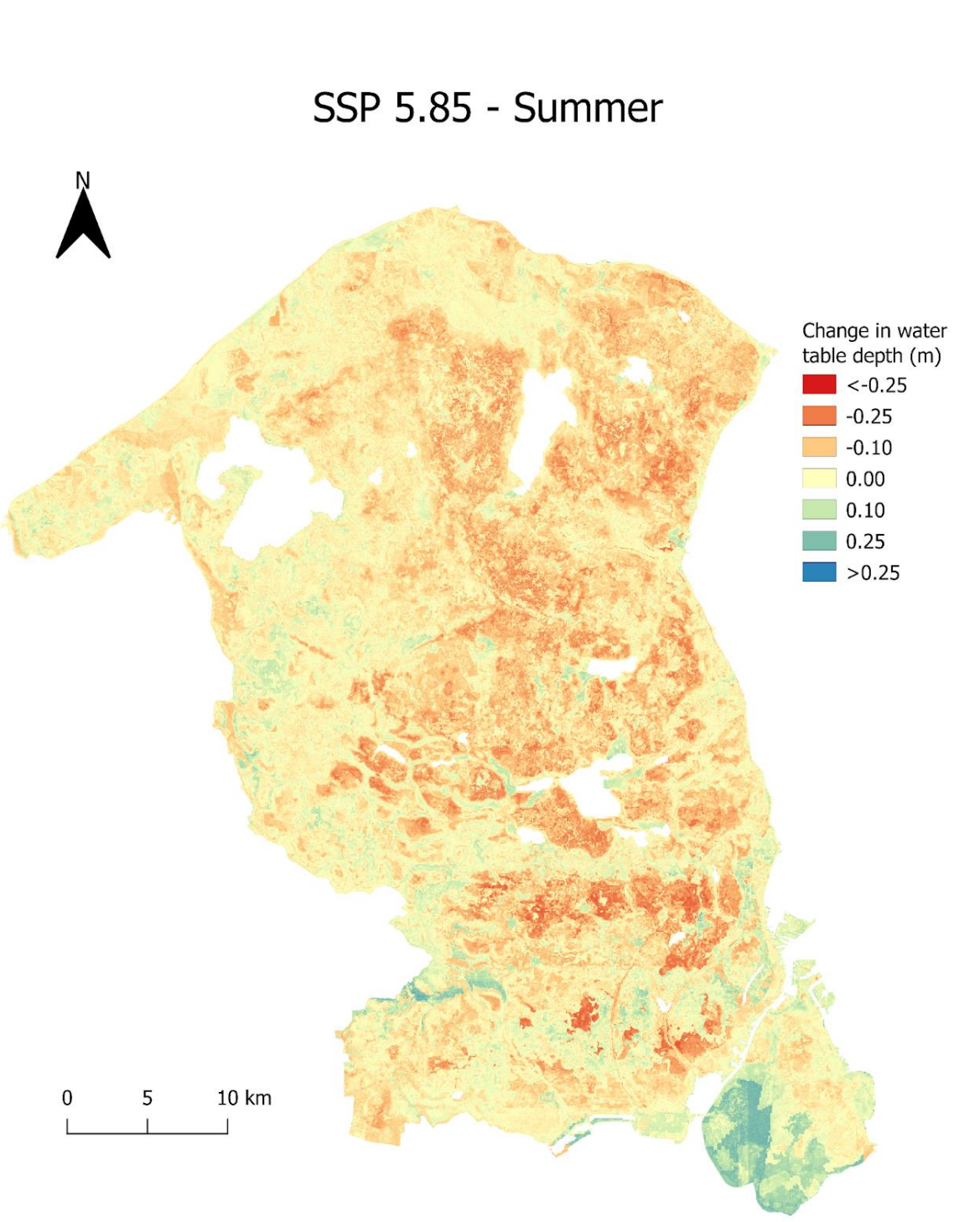
SSP5.85 - Winter

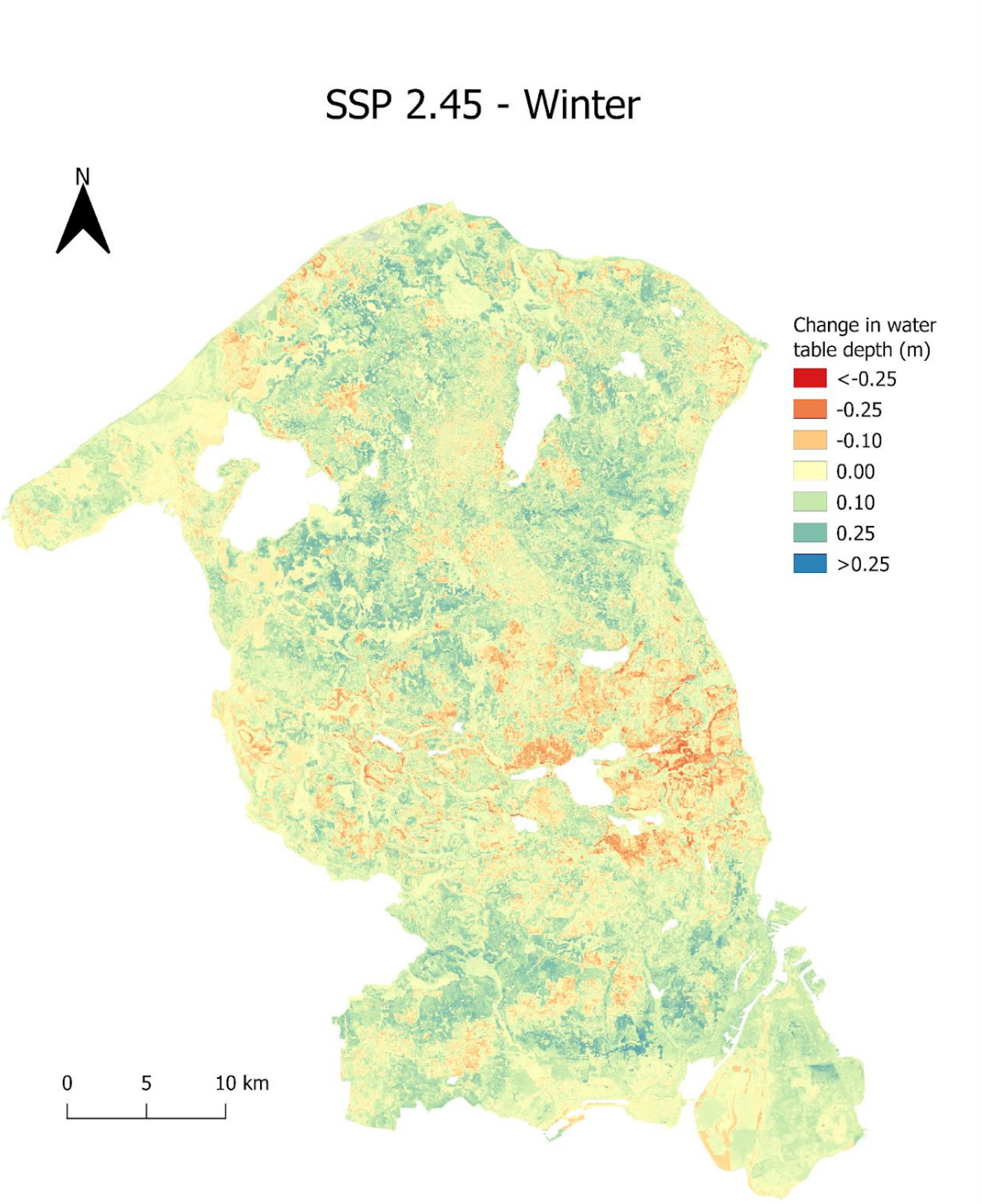


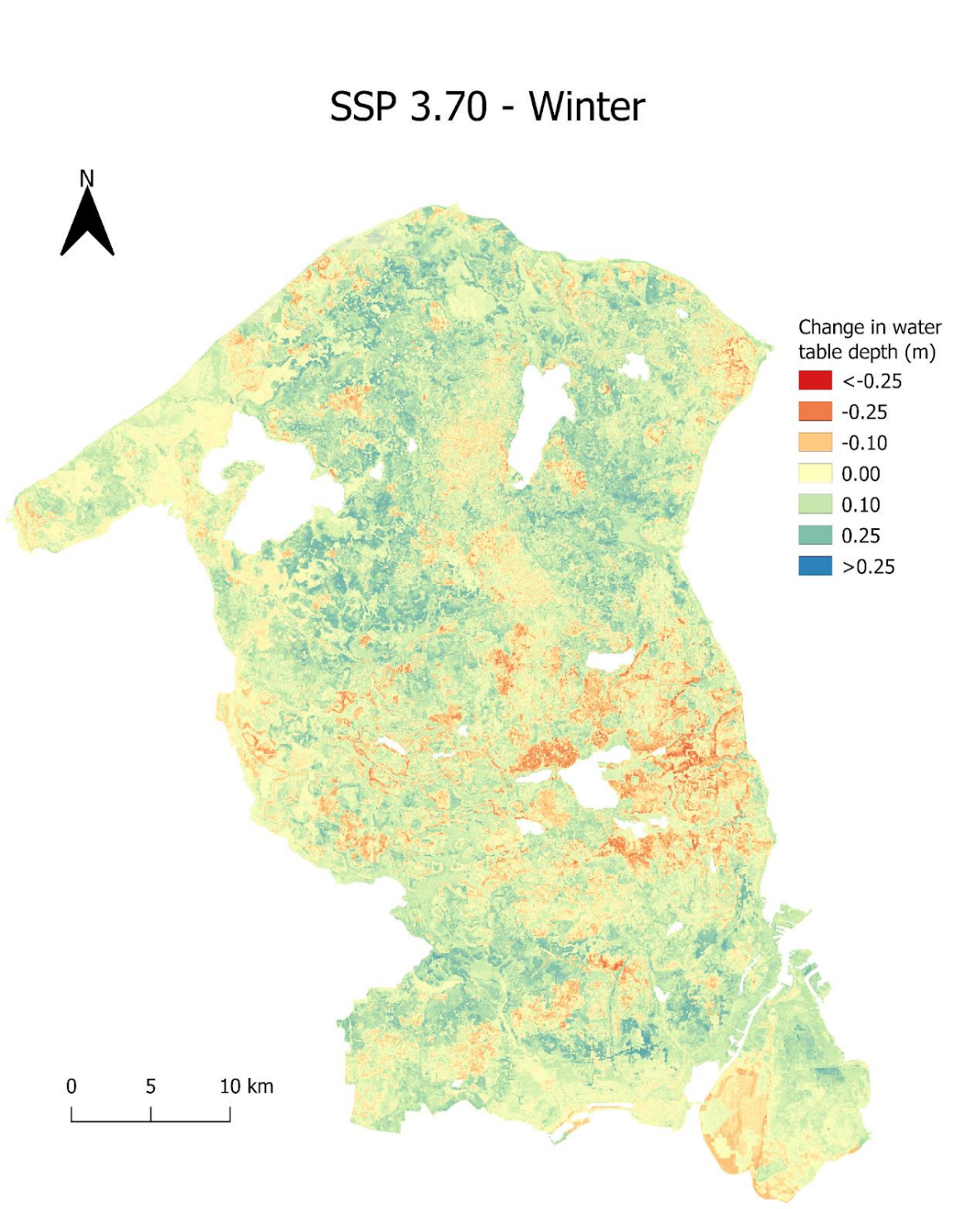
Annex 3. Differential maps

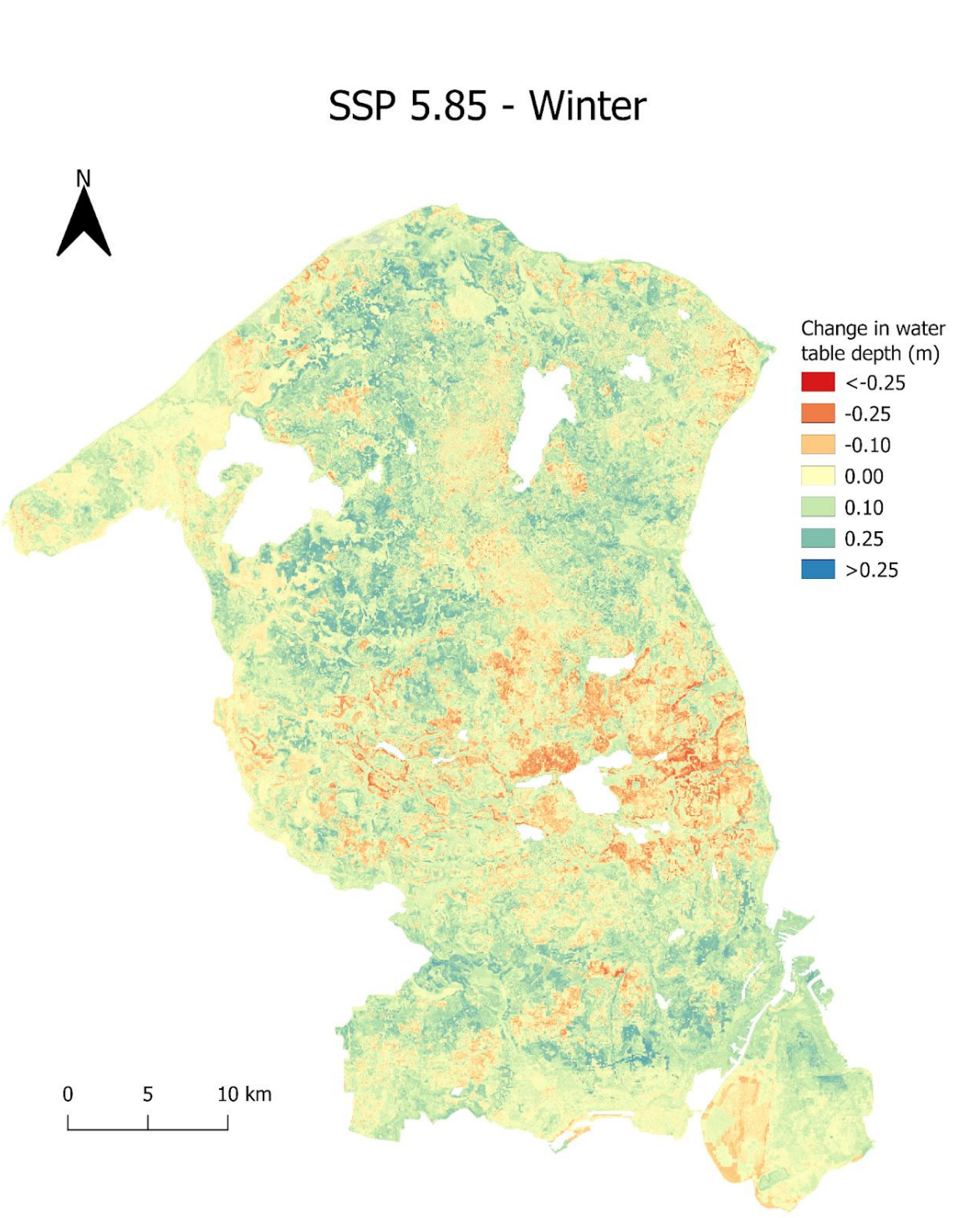



SSP 3.70 - Summer
N
Change in water table depth (m)
<-0.25
-0.25
-0.10
0.00
0.10
0.25
>0.25
0
5
10 km


SSP 5.85 - Summer
N
Change in water table depth (m)
<-0.25
-0.25
-0.10
0.00
0.10
0.25
>0.25
0
5
10 km


SSP 2.45 - Winter
N
Change in water table depth (m)
<-0.25
-0.25
-0.10
0.00
0.10
0.25
>0.25
0
5
10 km


SSP 3.70 - Winter
N
Change in water table depth (m)
<-0.25
-0.25
-0.10
0.00
0.10
0.25
>0.25
0
5
10 km


SSP 5.85 - Winter
N
Change in water table depth (m)
<-0.25
-0.25
-0.10
0.00
0.10
0.25
>0.25
0
5
10 km