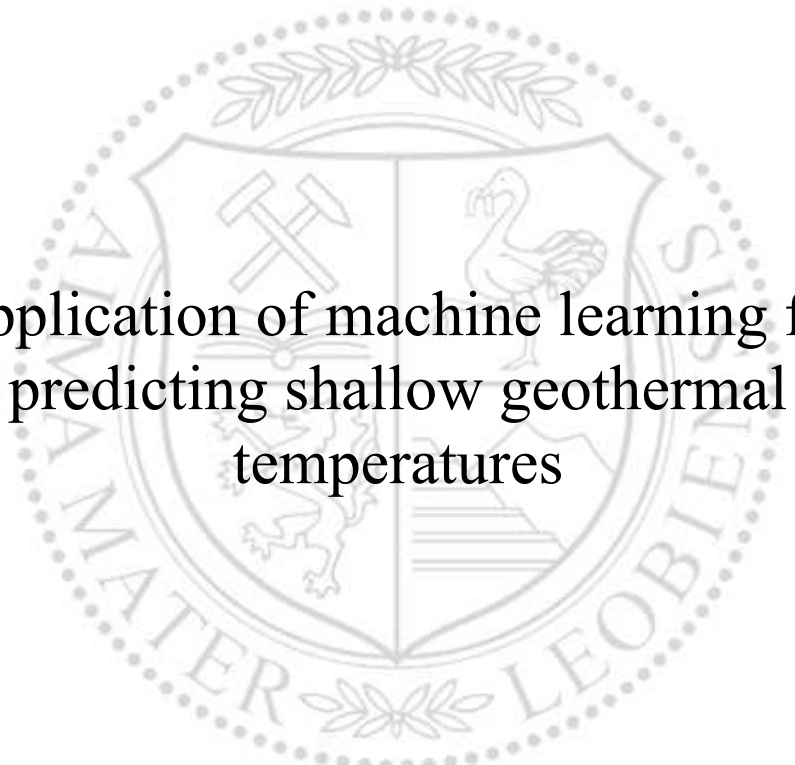




Chair of Geoenery Production Engineering

Master's Thesis



Application of machine learning for
predicting shallow geothermal
temperatures

Viktoriiia Skosareva

May 2024



AFFIDAVIT

I declare on oath that I wrote this thesis independently, did not use any sources and aids other than those specified, have fully and truthfully reported the use of generative methods and models of artificial intelligence, and did not otherwise use any other unauthorized aids.

I declare that I have read, understood and complied with the "Good Scientific Practice" of the Montanuniversität Leoben.

Furthermore, I declare that the electronic and printed versions of the submitted thesis are identical in form and content.

Date 16.05.2024

Signature Author
Viktoriia Skosareva

Viktoriia Skosareva
Master Thesis 2024
Petroleum Engineering

Application of Machine Learning for Predicting Shallow Geothermal Temperatures

Supervisor: Yoshioka, Keita; Univ.-Prof. PhD
Co-supervisor/Advisor: Sobhani, Ameneh;
Karsh. A.

Chair of Geoenery Production Engineering

Dedicated to my parents, my brother, and my grandparents

Abstract

Nowadays, the energy sector is experiencing the transition towards clean energy sources that reduce carbon emissions. Geothermal energy can cover the demand for energy grid stability and contribute to decarbonisation. The heat extracted from underground can be used as heat for space cooling and heating. The extraction of heat from shallow underground depths (up to 200 m) is the basis of this research.

To provide long-term sustainability of projects involving geothermal deployment and heat extraction, it is necessary to carefully monitor and manage the geothermal resources. Groundwater temperature (GWT) is a critical environmental parameter influencing the utilisation of geothermal systems. Accurate prediction of GWT is essential for assessing the efficiency and optimising the performance of geothermal installations. This thesis investigates the predictive modelling of GWT using machine learning (ML) techniques.

To accurately predict GWT with the help of ML algorithms, a comprehensive dataset with selected features was assembled, incorporating measurements from the eHYD database, along with weather data from Visual Crossing. The data from 1996 to 2016 of shallow geothermal wells in Vienna, district 22, was used for this purpose. Steps for dataset preprocessing are described, and ML algorithms are applied and evaluated to predict the temperature. Metric evaluations were used to check the accuracy of the predictions. The analysis revealed that, according to these evaluations, machine learning models are capable of providing high accuracy in predicting GWT, as well as showed the importance of accounting for climate factors when investigating geothermal applications on the city scale.

Zusammenfassung

Heutzutage erlebt der Energiesektor den Übergang zu sauberen Energiequellen, die die Kohlenstoffemissionen reduzieren. Geothermie kann den Bedarf an Energienetzstabilität decken und zur Dekarbonisierung beitragen. Die aus dem Untergrund entnommene Wärme kann als Wärme für die Raumkühlung und -heizung genutzt werden. Die Entnahme von Wärme aus geringen unterirdischen Tiefen (bis zu 200 m) ist die Grundlage dieser Forschung.

Um die langfristige Nachhaltigkeit von Projekten zu gewährleisten, die den Einsatz von Geothermie und die Wärmegewinnung beinhalten, ist es notwendig, die geothermischen Ressourcen sorgfältig zu überwachen und zu verwalten. Die Grundwassertemperatur (GWT) ist ein kritischer Umweltparameter, der die Nutzung von Geothermieanlagen beeinflusst. Eine genaue Vorhersage der GWT ist für die Bewertung der Effizienz und die Optimierung der Leistung von Geothermieanlagen unerlässlich. Diese Arbeit untersucht die Vorhersagemodellierung von GWT mit Techniken des maschinellen Lernens (ML).

Um die GWT mit Hilfe von ML-Algorithmen genau vorhersagen zu können, wurde ein umfassender Datensatz mit ausgewählten Merkmalen zusammengestellt, der Messungen aus der eHYD-Datenbank sowie Wetterdaten von Visual Crossing enthält. Dazu wurden die Daten von 1996 bis 2016 von flachen Geothermiebohrungen in Wien, Bezirk 22, verwendet. Die Schritte für die Datenvorverarbeitung werden beschrieben und ML-Algorithmen werden angewendet und ausgewertet, um die Temperatur vorherzusagen. Mit metrischen Auswertungen wurde die Genauigkeit der Vorhersagen überprüft. Die Analyse ergab, dass maschinelle Lernmodelle in der Lage sind, eine hohe Genauigkeit bei der Vorhersage von GWT zu liefern, und zeigte die Bedeutung der Berücksichtigung von Klimafaktoren bei der Untersuchung geothermischer Anwendungen auf der Stadtskala..

Table of Contents

Abstract.....	v
Zusammenfassung.....	vi
Table of Contents.....	vii
Chapter 1.....	9
1.1 Background and Context.....	9
1.2 Scope and Objectives.....	10
1.3 Achievements.....	11
1.4 Technical Issues.....	12
1.5 Overview of Dissertation.....	12
Chapter 2.....	15
2.1 Introduction to Geothermal Energy.....	15
2.2 Shallow Geothermal Energy: Mechanisms, Applications, and Environmental Benefits.....	16
2.3 Overview of Shallow Geothermal Applications All Over the World.....	18
2.4 Focus on Projects in Vienna: Shallow Geothermal Energy, Subsurface Urban Heat Effect, and Sustainable Solutions.....	20
2.5 Influence of Climatic Factors on Ground Water Temperature.....	22
2.6 Machine Learning Applications in Geothermal and Shallow Geothermal Energy..	23
Chapter 3.....	27
3.1 Data cleaning and preprocessing steps.....	27
3.2 Data interpretation.....	31
3.3 Identifying outliers.....	33
3.3.1 Graphical identification of outliers.....	34
3.3.2 Quantitative identification of outliers.....	40
3.4 Categorical or Binary Encoding.....	46
3.5 Normalisation Step.....	47
Chapter 4.....	49
4.1 K-Nearest Neighbors (KNN) Method.....	49
4.2 Deep Neural Networks (DNN).....	50
4.2.1 Convolutional Neural Networks (CNN).....	52
4.3 Random Forest.....	53
4.4 Machine Learning Models' Performance.....	54
4.4.1 K-Nearest Neighbors Method.....	54
4.4.2 Convolutional Neural Networks.....	55
4.4.3 Random Forest.....	56
4.5 Machine Learning Models Results.....	56
4.6 Feature Importance Analysis.....	57

4.7	Data Normalization Step Implementation.....	59
Chapter 5.....		61
5.1	Summary	61
5.2	Evaluation	62
5.3	Future Work.....	62
References		65
List of Figures		69
List of Tables		70
Abbreviations.....		71

Chapter 1

Introduction

1.1 Background and Context

To reduce CO₂ emissions, it is important to draw attention to the heating sector decarbonisation, as the energy supply for urban heating can cause up to almost a third of the total CO₂ emissions Tissen C. et al. (2021). Geothermal energy is a solution in this regard. According to statistics (IRENA and IGA, 2023), in 2021, geothermal energy had an installed capacity of 15,96 GW (electric). Also, the technical potential of geothermal energy is estimated at 200 to 5000 GW (thermal), which can meet up to 18% of global electricity demand.

The heat obtained from shallow depths (up to 200 m) is utilised mainly for space cooling and heating with the help of geothermal heat pump systems (Manzella, 2017). For instance, according to the European Geothermal Energy Council (2013), out of 5000 district heating systems in Europe, 237 were known as geothermal district heating (GDH) systems by 2014.

Geothermal energy is a prominent source that can be utilised to contribute to the decarbonisation strategy. Machine learning (ML) algorithms can be used during different stages to optimise geothermal deployment. This project aims to investigate the ML applications in predicting underground water temperatures in shallow geothermal wells. These temperatures are affected by multiple factors, namely geological conditions, weather patterns, and human activities, making their prediction complex but crucial for optimising system performance.

Therefore, if it becomes possible to accurately predict the temperature and create a temperature map of the area under investigation, then it will improve the project's performance in terms of energy, time, and money savings. Furthermore, in the case of urban aquifers, it is important to take into consideration additional influencing factors, features such as the urban effect, and subsurface infrastructures, including tunnels, subways and others (Wang et al., 2023). These

heat sources influence underground water warming and should be considered when creating temperature maps. Therefore, the project is aimed at searching for the most suitable ML model and acquiring more knowledge concerning the issue of sustainable operation of geothermal systems in urban areas.

1.2 Scope and Objectives

The scope of this thesis includes the study of shallow geothermal systems with a focus on the utilisation of machine learning to predict groundwater temperatures. The objectives are: first, to conduct a literature analysis of the factors influencing groundwater temperature fluctuation in urban areas and study cases where shallow geothermal systems are implemented, specifically in Vienna, and second, to assess and conclude on the performance of various machine learning models in predicting groundwater temperature. Machine Learning techniques are capable of capturing complex, nonlinear relationships between various variables in geothermal systems. A detailed analysis of variables influencing the temperature profile is studied, and the algorithms capable of describing their relationships are discussed.

The literature review is conducted to develop an understanding of geothermal energy applications and the role of ML techniques in GWT prediction. Subsequently, the steps of initial data handling are described, which involves collecting, cleaning, and preprocessing data. The clean dataset then serves as the basis for applying several machine learning algorithms to establish a predictive model that can be used by urban planners and developers to enhance the efficiency of geothermal installations. The workflow shown in Figure 1.1 below depicts the steps described in subsequent Chapters 3 and 4 and implemented starting with the data acquisition up to model training and optimisation.

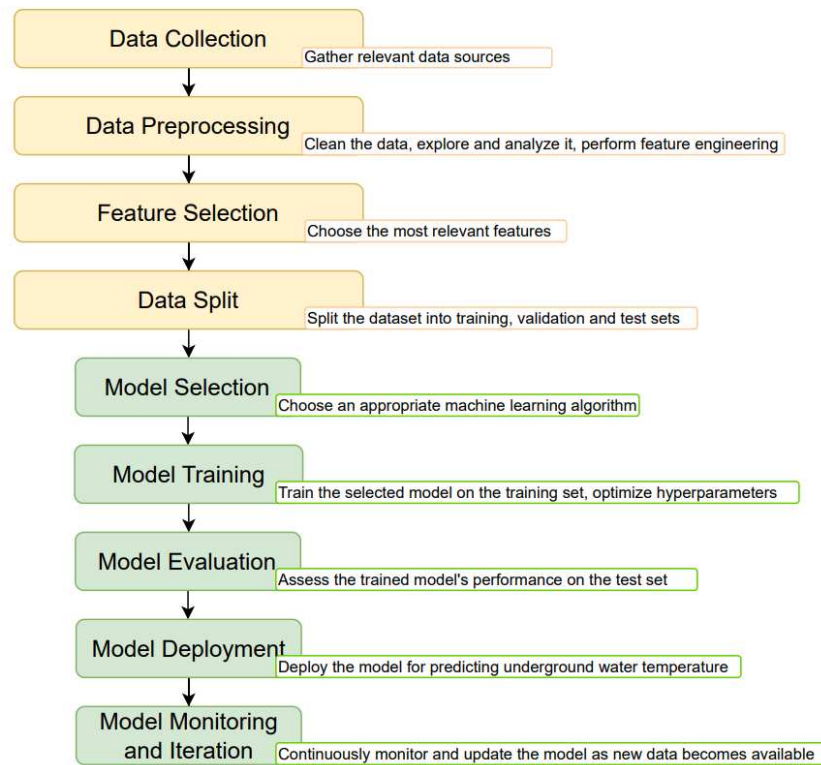


Figure 1.1 – Workflow for Underground Water Temperature Prediction Using Machine Learning

1.3 Achievements

As a result of this research, several practical outcomes can be highlighted:

- *Dataset Development:* A dataset for groundwater temperature prediction, which includes the most important features for analysis, such as various environmental and temporal factors, was compiled and prepared. This dataset can be used to study the performances of other machine-learning applications in geothermal energy in Vienna.
- *Evaluation of Machine Learning Models:* Different ML models, namely K-Nearest Neighbors (KNN), Deep Neural Networks (DNN), and Random Forest, were evaluated. The most suitable model for predicting groundwater temperatures was determined. The KNN model showed the best overall performance in terms of accuracy and computational efficiency.
- *Feature Importance Analysis:* Through the application of feature analysis techniques, the study identified key variables that most significantly affect groundwater temperatures. Understanding these factors helps in designing more efficient geothermal systems and improving model accuracy.
- *Improvements in Predictive Accuracy:* The models developed through this thesis showed relatively high levels of accuracy, enhancing the capability for precise temperature control within geothermal systems.

These outcomes extend academic understanding of machine learning implementation in geothermal energy systems and provide insights that can be applied to optimise and enhance these systems in urban sites.

1.4 Technical Issues

During the research, several technical challenges arose, primarily related to data management, which did not influence the modelling process but should be considered when processing raw data:

- *Diverse Data Sources:* The data for this study was collected from multiple sources, each with a different format and measuring frequency. Integrating these varied data sources into a coherent dataset suitable for analysis required attention to ensure consistency and compatibility.
- *Varying Time Scales:* The data collected featured different time scales and years. Some datasets provided data measured daily, while others offered monthly data. It was necessary to decide on the study period and apply averaging.
- *Insignificant Features:* The initial datasets included numerous features, not all of which were relevant or significant for predicting groundwater temperatures. These features were identified and removed.
- *Missing Data Handling:* Handling missing data was a major technical issue. The datasets contained gaps and incomplete records. The approach outlined in Chapter 3 was utilised to fill these gaps and to ensure the integrity and accuracy of the final model.

As data accuracy impacts the performance of the ML algorithms, addressing these technical issues was essential to provide accurate and reliable data.

1.5 Overview of Dissertation

The dissertation is organised into several chapters:

- Chapter 2: Literature Review – This chapter covers the fundamentals of geothermal energy, impact of environmental factors on groundwater temperatures, and provides examples of machine learning applications in this field worldwide and in Vienna.
- Chapter 3: Data Cleaning and Preprocessing – Outlines the methods used to prepare the dataset for machine learning analysis to ensure high data quality and relevance.
- Chapter 4: Machine Learning Methods and Algorithms – This chapter evaluates ML models for their effectiveness in predicting groundwater temperatures.

- Chapter 5: Conclusions – outlines the summary and results of the thesis and discusses potential steps for future research.

This thesis discusses insights for sustainable and efficient urban geothermal systems development.

Chapter 2

Literature Review

This section addresses the theoretical aspects. The review starts with an introduction to geothermal energy, and in this part of the chapter, definitions of both deep and shallow systems are given. Then, the review is followed by a detailed discussion of shallow geothermal energy, its mechanisms and applications. The following part covers the application of geothermal energy in Vienna, focusing on shallow implementation. The next part describes the relationship between climatic factors and groundwater temperature. Finally, the role of machine learning and its application for geothermal energy utilisation is discussed.

2.1 Introduction to Geothermal Energy

Decarbonisation is a worldwide concern that results from increasing energy consumption, population growth, and industrial development. To overcome climate change and minimise CO₂ emissions, it was stated in the Paris Agreement on Climate Change that the global average temperature increase is to be limited to 1.5°C above preindustrial levels (Secretariat, United Nations Framework Convention on Climate Change, 2015). Energy utilisation for cooling and heating purposes both in the industrial and residential sectors is considered to contribute significantly to CO₂ emissions. One of the methods to decarbonise this sector is the implementation of shallow geothermal energy (SGE) systems.

Geothermal energy is known to be energy accumulated as heat and derived from the core layer of the Earth, which serves as a reliable and eco-friendly alternative to fossil fuels. The term "geothermal" comes from the Greek language, where the word "geo" means "Earth" and "therme" - "heat" (Chettri and Sankarananth 2022). This heat, located in the fluids and rocks beneath the Earth's crust, can be utilised for heating, cooling, and electricity generation (Chettri and Sankarananth 2022).

Geothermal resources are located at different depths, which predetermines the heat utilisation. The exploration and utilisation of geothermal energy depends on the type of system, which can be categorised as deep and shallow geothermal systems. Deep geothermal systems, also referred to as geothermal reservoirs, are often located at depths up to 10 kilometres. These systems are typically associated with high temperatures (high-temperature geothermal resources are considered to be resources with a temperature higher than 150°C) and are suitable for generating electricity. The process of electricity production from deep geothermal sources involves drilling wells into geothermal reservoirs to access steam or hot water, which can then drive turbines connected to electricity generators (Vlahović, M., Stević, Z 2020).

Conversely, shallow geothermal energy refers to the heat within the upper 200 meters of the Earth's surface. This form of geothermal energy is predominantly used for direct heating and cooling applications. Shallow geothermal systems can have temperatures below 20 °C and utilise ground-source heat pumps (GSHPs) to transfer heat to or from the ground. This offers an efficient method for heating buildings during winter and cooling them during summer. If heat is utilised for electricity production, the GSHPs system performs with a 3-5 times higher coefficient of performance compared to traditional combustion-based heaters (Ahmed et al. 2022), (Manzella 2019).

The environmental benefits of both deep and shallow geothermal energy are substantial, providing a low-carbon alternative to conventional energy sources. Shallow geothermal energy, in particular, has seen increased application in urban areas, thus helping to reduce CO₂ gas emissions and reduce urban heat island effects (Shao et al. 2023). As the global efforts move towards more sustainable energy solutions implementation, geothermal energy utilisation will be essential in meeting energy needs while preserving the environment.

2.2 Shallow Geothermal Energy: Mechanisms, Applications, and Environmental Benefits

Shallow geothermal energy (SGE) systems are a promising and sustainable solution for heating and cooling applications, which utilise the heat that is held within the Earth's crust up to depths of 200 meters and temperatures below 25°C (Shao et al. 2023). Ground-source heat pumps (GSHPs) serve as the key technology for transferring heat between the ground and the built environment. GSHPs are used for heat extraction from the ground at relatively low temperatures, utilising a heat pump to increase this temperature for space heating or hot water use. GSHP systems include a borehole heat exchanger, followed by heat pump system, and,

subsequently, heat distribution system, as schematically shown below in Figure 2.1 (Ahmed et al. 2022):

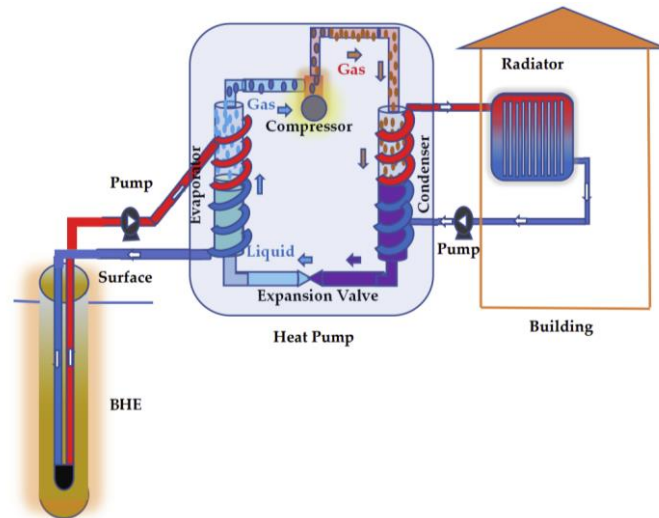


Figure 2.1 – Schematic flow diagram of GSHP systems for heating

The primary mechanism for heat extraction in shallow geothermal systems involves circulating a fluid (water or an antifreeze solution) through a closed-loop pipe system installed in the ground. The fluid absorbs ground heat and is then returned back to the surface, where the heat pump unit increases the temperature further for space heating or hot water use. In cooling mode, the system operates in reverse, removing heat from the building and disposing it into the ground. This dual capability makes shallow geothermal energy systems versatile and efficient for year-round climate control (Manzella 2019).

SGE systems fall into two primary categories: open and closed systems. The choice between these two systems depends on the available land, geological conditions, and specific heating and cooling needs. Open systems involve direct utilisation of groundwater as a heat carrier, extracted and re-injected into aquifers through groundwater wells. This system is suitable for sites with appropriate hydrogeological conditions, sufficient permeability and proper chemical quality of groundwater. Closed systems, on the other hand, employ underground heat exchangers in various configurations with a circulating heat carrier medium (Sanner 2003). Carrier fluid in such systems does not have any direct contact with the rock. Horizontal systems are often less costly but require significant land area, making them suitable for rural or suburban settings (Manzella 2019). Vertical systems, while more expensive, have a smaller footprint and are thus more applicable in urban environments where space is limited (Ninikas et al. 2017).

Shallow geothermal energy, characterised by minimal environmental impact, varies according to the scale of resource exploitation. It is crucial to mitigate potential risks like thermal pollution, ecosystem disruption due to drilling, and chemical pollution from geothermal fluids.

Practices such as re-injection of spent fluids are recommended to sustain resource integrity and prevent land subsidence (Manzella 2019).

SGE systems operate with minimal electrical consumption, primarily for operating the heat pump and circulation pumps. Thus, shallow geothermal systems contribute significantly to the reduction of carbon emissions. By displacing conventional fossil fuel-based heating and cooling solutions, SGE systems also play a crucial role in city-scale applications, helping to mitigate the urban heat island effect (Wan et al. 2022).

Research and application of shallow geothermal energy systems, particularly in regions with significant heating and cooling demands, have demonstrated their potential to contribute to sustainable urban development. Providing an energy-saving solution for building climate control, these systems mitigate the urban heat island effect, further enhancing their environmental benefits (Shao et al. 2023). The next part of this chapter covers more cases of shallow geothermal systems applications worldwide.

2.3 Overview of Shallow Geothermal Applications All Over the World

Utilising the constant temperatures near the Earth's surface, shallow geothermal energy systems have been increasingly implemented worldwide. Their implementation proves the efficiency of this renewable energy source in a variety of climatic and geographical settings. In this part of the chapter, implementation of SGE systems is discussed.

When discussing shallow aquifers located in urban areas, it is important to mention the term “subsurface urban heat islands (SSUHI)”, which explains the nature of groundwater temperature increase as a result of anthropogenic activities. The integration of shallow geothermal energy systems within urban environments has gained significant attention for its potential to mitigate the SSUHI effect and contribute to sustainable city development. The study conducted by Böttcher and Zosseder (2022) on Munich's aquifer provides an analysis of how human activity and natural factors influence groundwater temperature distribution in the city.

The effect of SSUHI is widely discussed nowadays, and this phenomenon drives the development of projects that aim to optimise the efficiency of thermal energy utilisation. On the other hand, this phenomenon, characterised by elevated groundwater temperatures, can adversely affect groundwater quality, microbial ecosystems, and the efficiency of geothermal heat pumps. If analysing depths up to 10 m, the groundwater temperature (GWT) here highly depends on seasonal air temperature variations. Nevertheless, heat loss from buildings or subsurface constructions also influences GWT significantly, making surface sealing and the

density of buildings the features to consider. Surface sealing emerges as the strongest anthropogenic warming influence. Thus, the study showed that the areas with sealed ground surfaces contribute to GWT warming in Munich. Nevertheless, when conducting the analyses at deeper depths, natural factors greatly influence the study, which is why, when studying dependencies of GWT on a city scale, it is essential not to neglect hydrogeological factors.

Hydrogeological factors such as aquifer thickness and depth-to-water exhibit significant cooling effects; therefore, it is important to maintain permeable surfaces and green areas within cities to combat SSUHI effects. Moreover, the depth-dependent analysis reveals that the influence of certain factors, such as aquifer thickness and Darcy velocity, changes with depth, highlighting the complexity of subsurface temperature dynamics. The study concludes that for designing and implementing shallow geothermal systems, it is crucial to consider both anthropogenic and natural factors (Böttcher and Zosseder 2022).

Talking about more projects implemented worldwide, it can be mentioned that in US, the Oregon Institute of Technology operates entirely on geothermal energy, including both deep and shallow systems, serving as a model for campuses and communities seeking to transition to renewable energy sources (Oregon USA Oregon Institute of Technology.). Similarly, in Iceland, a country with vast geothermal resources, shallow geothermal systems complement the deep geothermal power plants, providing heating for swimming pools, greenhouses, and residential spaces (Shelare et al. 2023).

One more example is the study of Ninikas et al. (2017), which shows the successful utilisation of a shallow heat source via the operation of a water source heat pump installed in Glasgow, UK, at a subway station. Water entering the tunnels, with a mean temperature of 14.2°C, poses challenges due to its volume and quality, affecting subway operations. This water is collected in chambers and then directed to the sewerage network. By redirecting this water through the heat pump system, it became possible to cover heating and hot water needs for the station, substituting the previously utilised electric-fired system for heating. This project helped reduce energy consumption by up to 60% and is an excellent example of shallow geothermal systems implementation to reduce CO₂ footprint (Ninikas et al. 2017).

Other examples of implementation of geothermal heat pumps (GHP) can be found in Turkey. The first GHP system was installed in Istanbul in 1998 for residential heating and cooling. Since then, the application of GHP systems has expanded and nowadays is implemented in large commercial buildings, shopping malls, office buildings, educational institutions, and even innovative projects like the maintenance building at Sabiha Gökçen Airport, one of the largest energy pile GHP. The use of groundwater, which has a temperature of 17°C in one of the shopping malls in Antalya, allowed consuming 50% less electricity (Cetin and Paksoy 2013).

Implementing shallow geothermal systems has numerous environmental benefits, including reducing greenhouse gas emissions, lowering energy consumption, and having minimal ecological impact. Additionally, these systems contribute to energy security and can drastically lower operational costs, making them an attractive option for sustainable development (Manzella 2019).

The future of shallow geothermal energy looks promising as cities endeavour to reduce carbon footprint. Shallow geothermal energy is perceived as a solution that is both reliable and sustainable. Next, the emphasis will be drawn to the application of this system in Vienna, Austria.

2.4 Focus on Projects in Vienna: Shallow Geothermal Energy, Subsurface Urban Heat Effect, and Sustainable Solutions

In Austria, 30% of the energy is consumed for district heating and hot water supply. To achieve climate neutrality and heating decarbonisation by 2040, according to Energy Innovation Austria (2021), it is necessary to draw attention to renewable sources, where geothermal energy can play a crucial role. It is assumed that by 2040, it will be possible to generate up to 15 TWh of heat from near-surface geothermal energy sources. Geothermal energy can find its application in air conditioning, seasonal heat storage, cooling purposes, heating networks and energy storage systems. Thus, near-surface geothermal heat utilisation is discussed in the Manage GeoCity project, taking place in Graz. The simulation was made to analyse the influence of heat extraction and storage activities on the GWT. The analysis revealed the possibility of a reduction of up to 85% in greenhouse gas emissions. Utilising heat pumps for heating purposes can positively result in mitigating the SSUHI effect. Also, the Geological Survey of Austria (GBA) offers maps that provide information about shallow geothermal energy potential and can be utilised for planning.

The study of V. Ostermann (2011) examines the potential of geothermal energy in Austria. The study addresses shallow geothermal energy capacity at depths of 10-15 m up to 50 m, where temperatures vary between 10 and 15°C, and focuses on its application for building heating. The analysis is performed via modelling of the country's territory and calculation of geothermal energy potential. When comparing the available geothermal energy potential with thermal energy demand, the author concludes that 99.5% of heating energy demand can be covered by shallow geothermal energy utilisation across most inhabited regions. This provides a great opportunity for Austria to minimise carbon dioxide emissions (Ostermann 2011).

Vienna, the capital of Austria, has demonstrated a progressive approach to utilising shallow geothermal energy, integrating it into the city's strategy for sustainable development and energy efficiency. The city's geothermal initiatives address both heating and cooling needs while contributing to environmental goals.

In Vienna, 1070 closed-loop and 762 open-loop systems are already in operation. Vienna has an area of 415 km² and is divided into 23 districts. The city has two parts: eastern and western, split by the Danube River. Vienna has four hydrogeological zones, with the Holocene having the highest permeability and average thickness of 7-14 m. It is estimated that 30% of the city is suitable for shallow geothermal energy utilisation, with the 21st and 22nd districts posing a high potential for SGE projects. Tissen C. et al. (2021) provides a solution to identify shallow geothermal locations in Vienna. The suggested GeoEnPy tool, evaluating anthropogenic heat input and shallow geothermal potential, can calculate the technical geothermal potential and heat supply rate, thus identifying the most feasible sites. As for anthropogenic heat flux analysis, the author concludes that the highest mean heat flux originates from underground car parks and DH (district heating) systems, while the negative can be found in tunnels. The analysis showed that 68% of Vienna's heating demand could be supplied by borehole heat exchanger (BHE), preferably with its installation in the aquifer east of the Danube River (districts 21 and 22), which is the most productive aquifer with the highest yield. The eastern and southern districts have the highest heat supply rates and sustainable potential, making them ideal for geothermal systems.

As discussed previously, the subsurface urban heat island (SUHI) effect is the phenomenon of urban ground and groundwater warming. Among the sources that lead to increased groundwater temperatures (GWT) in urban aquifers are heat input from buildings and increased ground surface temperature (GST). Also, subsurface infrastructures, namely district heating (DH) pipes and sewers, contribute to SUHI. In addition to this, even air conditioning can contribute to summertime overheating, as described in (Österreicher and Sattler 2018). Therefore, a lot of projects are aimed not only at seeking solutions for heating geothermal utilisation but also at mitigating the SUHI effect. Thus, Vienna's geothermal projects contribute to minimising this effect by utilising the ground as a heat sink in summer months, thereby reducing overall urban temperatures and improving living conditions. An example of heat sink implementation is a project in Baden bei Wien where a hydrogeological model for BHE installation was created for waste heat geothermal storage at a depth of 10-20 m (Haslinger et al. 2022).

The city's approach to geothermal energy, particularly in addressing the subsurface urban heat effect, illustrates the potential of shallow geothermal systems not only as a source of renewable energy but also as a tool for urban climate adaptation.

2.5 Influence of Climatic Factors on Ground Water Temperature

The efficiency of shallow geothermal systems, particularly very shallow installations, is closely connected to the dynamics of groundwater temperature (GWT). Different climatic factors, such as air temperature, precipitation, solar radiation, and others, influence GWT. Understanding these relationships is critical to optimising the performance of geothermal systems.

Air temperature plays a significant role in influencing GWT, as seasonal variations in ambient temperature affect the thermal regime of the subsurface. During colder months, the ground acts as a thermal reservoir: it absorbs and stores heat, and in warmer months, the ground can serve as a cooling source due to its relatively lower temperature compared to the air. Studies have shown that air temperature fluctuations can significantly impact GWT, highlighting the importance of considering local climate conditions in the design and operation of geothermal systems (Hare et al. 2023). Ambient temperature directly impacts GWT through conductive and convective heat transfer mechanisms. Higher air temperatures can increase GWT, especially in shallow subsurface regions.

Precipitation contributes to groundwater recharge and can influence GWT by affecting soil moisture levels and the thermal properties of the subsurface. Heavy rainfall normally increases groundwater levels, potentially leading to a temporary cooling effect on GWT due to the thermal inertia of water. Conversely, prolonged periods of drought can lead to lower groundwater levels, which may result in increased GWT due to reduced heat dissipation (Rushlow et al. 2020).

Snow acting as an insulating layer can result in heat loss reduction from the ground to the atmosphere. Its presence can significantly change GWT, especially in regions with seasonal snow cover.

Solar radiation directly impacts the Earth's surface temperature and, consequently, GWT. Solar heating of the ground surface can increase GWT, especially in areas with high solar insolation and minimal cloud cover. The extent of this influence is dependent on the surface reflecting power, vegetation cover, and the presence of built-up areas, which can absorb and retain heat.

Solar Energy represents the total solar energy received over a period, affecting the thermal regime of the ground. Higher cumulative solar energy can increase the energy stored in the ground, raising the GWT. *Sunlight duration* impacts the amount of solar energy received by the Earth's surface, directly influencing ground temperature through radiation absorption. Longer daylight hours increase the ground's exposure to solar radiation, thereby increasing GWT.

The UV index indirectly indicates the potential for solar radiation to affect surface and subsurface temperatures, although its direct impact on GWT might be less significant than solar radiation and energy metrics.

Clouds can act as a thermal blanket, trapping longwave radiation from the Earth's surface. High *cloud cover* can lead to warmer night-time temperatures, affecting the diurnal variation of GWT.

The dew point temperature directly influences humidity and moisture levels in the air. This, in turn, can influence the rate of evapotranspiration and, eventually, the ground temperature. With higher dew point, air contains more moisture, potentially leading to higher rates of evapotranspiration and cooling effects on the ground surface under certain conditions.

Wind can boost the convective heat transfer between the ground surface and the atmosphere. Higher *wind speeds* increase the heat removal rate from the surface, potentially lowering the GWT in exposed areas.

High *humidity levels* can result in a reduction in the effectiveness of evaporative cooling from the ground surface, leading to higher ground temperatures. Humidity also affects the thermal conductivity of air, influencing heat transfer rates.

The closeness of groundwater to the surface can significantly influence GWT due to water's high specific heat capacity. Areas with shallow *groundwater levels (GWL)* might exhibit more stable temperatures compared to regions with deeper groundwater levels.

The parameters mentioned above and their influence on GWT underscore the necessity for a comprehensive approach to geothermal energy planning and management. The practical chapter will outline the conducted analysis of each feature's importance to identify the factors that significantly influence the prediction of Ground Water Temperature (GWT). Understanding these dependencies is crucial for optimising geothermal energy extraction.

2.6 Machine Learning Applications in Geothermal and Shallow Geothermal Energy

Machine learning (ML) is considered as a subset of artificial intelligence (AI) (Vesselinov et al. 2022). It focuses on the algorithms and different statistical models development with their subsequent optimisation, thus allowing computers to perform tasks without explicit instructions. Machine learning algorithms have a distinctive feature: over time, they can enhance their performance by continuously learning from data, thus allowing them to identify patterns, make predictions, and adapt to new inputs without being explicitly programmed for

each task. Smart data analysis will become more and more necessary for technological progress as the amount of data increases (Smola and Vishwanathan, 2008).

Shallow geothermal energy systems harness the thermal energy which is stored in the upper layers of the crust for heating and cooling purposes. In the context of geothermal energy extraction (Vesselinov et al. 2022), advanced technologies are needed to reduce geothermal exploration and development risks and costs. These technologies can be used to assess geothermal prospectivity, energy affordability, and resource development.

Machine learning has emerged as a powerful tool for optimising and predicting various aspects of geothermal heat extraction. ML methods have been applied to different stages of geothermal heat extraction, including site selection, reservoir characterisation, drilling and completion optimisation, and power plant operation. Thus, PCA (principal component analysis) was used to identify promising regions for development. Also, unsupervised ML techniques combined with SMEs (subject-matter experts) were used to acquire knowledge of the fault location and propagation behaviour, which helped further exploration. NMFk (non-negative matrix factorisation) method allowed based on the gravity and seismic data to construct temperature and heat flow maps in Utah, which provided more insides into drilling locations. ML was also applied to study the Great Basin region for hidden systems locations. Another example is the application of clustering algorithms (image data was used to assess the geothermal energy potential, and as a result, it was possible to analyse the direction and pattern of Brady's field). Supervised ML methods were applied to image geothermal reservoir properties (and monitor reservoir evolution) and forecast seismic events (Vesselinov et al. 2022).

Three main groups of activities for ML application are as follows:

- Resource assessment: ML algorithms can help predict subsurface geology and identify areas with high geothermal potential. Machine learning models can generate 3D maps of subsurface reservoirs and estimate their energy capacity by analysing geological data, such as temperature, pressure, and resistivity.
- Drilling optimisation: ML can also be used to optimise the drilling process and reduce costs. By analysing drilling data and real-time sensor data, machine learning algorithms can predict drilling performance, identify potential problems, and adjust drilling parameters in real time.
- Plant operation optimisation: ML can help improve the efficiency of geothermal power plants by predicting system performance and identifying areas for optimisation. By analysing data from sensors and control systems, machine learning models can identify patterns and anomalies that can help

plant operators optimise equipment performance and reduce maintenance costs.

Overall, machine learning algorithms can find various applications in different stages of geothermal deployment and can be used in:

- the judgment process of a drilling site selection;
- temperature prediction;
- geothermal systems productivity prediction;
- the prediction of negative environmental effects;
- thermo-economical optimisation;
- operation optimisation of a geothermal power plant and heating systems (Wang et al. 2023).

Machine Learning techniques are capable of capturing complex, nonlinear relationships between various variables (dependent and independent) in geothermal systems. *Data* is crucial to developing an ML model. Although the other two types of data exist (lab-scale and synthetic data), data used in this work can be classified as *distributed field-scale data*, meaning that various remote sensing and geophysical methods were used to collect data over a region. This data can have a potential downside as they are generally noisy and may need extensive preprocessing. The ML methods can be categorised as the following (Wang et al. 2023):

- The data-driven methods learn dependencies from the provided data.
- The physics (or science)-informed methods, together with provided data, use also existing knowledge (constitutive relationships, conservation laws, mathematical models) about the solved problem.

Also, ML methods can be split into three following categories:

- *Supervised ML methods* – algorithms that map inputs (i.e. permeability or temperature) to outputs (i.e. thermal power produced or geothermal potential). The training of such algorithms requires massive datasets.
- *Unsupervised (self-supervised) ML* – algorithms able to identify patterns from data, and then associate these patterns with physical meaning. The method can extract hidden signals from data.
- *Semi-supervised ML* – algorithms that can learn by combining labelled data with unlabeled data samples, which allows learning accuracy improvement (Vesselinov et al. 2022).

When talking about machine learning applications for geothermal system optimisation, the process requires sophisticated analysis of various parameters, including GWT, system design,

and environmental conditions. Machine learning models are capable of handling complex, multivariate datasets, making them ideal for identifying optimal configurations for geothermal installations. Thus, deep learning algorithms are applied to predict geothermal reservoir performance, enabling more effective management of heat extraction processes and minimising the risk of thermal breakthroughs (Gudmundsdottir, H. and Horne, R. 2020).

Enhanced Geothermal Systems (EGS) are another mechanism for deep geothermal energy exploitation, but heat, in this case, is extracted from hot dry rocks through the injection of fluids. Machine learning models (i.e. Recurrent Neural Network) are being used to simulate the complex hydrothermal dynamics within EGS, guiding the development of more effective and sustainable extraction techniques. These models help the assessment of potential sites, optimisation of fluid injection strategies, and prediction of long-term system performance (Xue Z. 2023)

This study investigates one of the most important machine learning applications, namely **temperature prediction**. Furthermore, in the case of urban aquifers, it is important to take into consideration additional influencing factors, features such as the heat input from buildings, and subsurface infrastructures, such as district heating pipes, tunnels, subways and others, as these heat sources influence groundwater warming. Artificial Neural Network is one of the most popular methods for temperature prediction (Wang et al. 2023). Random Forest was investigated for subsurface temperature prediction and showed the highest accuracy (Shahdi et al. 2021).

Integrating machine learning techniques in geothermal energy projects represents a significant advancement in the field, offering precise and robust data analysis, system optimisation, and predictive modelling. These ML applications improve the operational efficiency of geothermal installations and contribute to sustainable energy source utilisation.

Overall, the principles of geothermal energy, particularly shallow geothermal systems used in urban areas like Vienna were reviewed. The mechanism of these systems was discussed, along with their environmental benefits and growing importance in providing sustainable energy solutions. Also, the literature review aimed to gain knowledge about projects implemented worldwide, and particularly in Vienna. The review also touched on how weather and seasonal changes affect groundwater temperatures, which are crucial for these geothermal systems. The chapter concluded by pointing out the potential of using machine learning to improve predictions related to these temperatures. The next chapter provides information on handling real-world data before its utilisation in machine learning algorithms. It will cover how to clean and prepare the data for analysis, ensuring that the information used to feed into machine learning models is accurate and ready for effective use.

Chapter 3

Data cleaning and preprocessing

3.1 Data cleaning and preprocessing steps

Data quality is considered to be a term uniting factors like accuracy, completeness, consistency, timeliness, believability and interpretability (Chakrabarti et al. 2008). Elements such as accuracy, completeness and consistency define data quality.

Initially, the real-world data may come together with inaccuracy, inconsistency or missing values. This will significantly influence the model's accuracy and can lead to unreliable results. To ensure the reliability and integrity of the further analysis of the data, as well as its correctness, cleaning and preprocessing are required. A description of the process workflow is provided in this chapter. Most of the steps for data manipulation mentioned below can be conducted using Pandas Python Library, as it has the most relevant features for data merging, cleaning, filtering and sorting.

The following workflow was applied in this work (Data Science Horizons, 2023):

- Data Collection – the process of collecting the initial data from various available sources. The most significant data sets in this work, namely the information about the groundwater temperatures and levels, were downloaded from the eHYD (Bundesministerium für Landwirtschaft, Regionen und Tourismus, eHYD.), and the datasets related to the weather information were available on visualcrossing.com (Visual Crossing Corporation. Weather Data Services). One additional step that was applied in the beginning is data alignment, as all data sets should fall within the same period.

- Data Reduction – this process may include feature selection or clustering to focus on the most relevant variables, helping reduce the data size (Chakrabarti et al. 2008), while allowing the production of the same analytical results.
- Data Cleaning – the data obtained from the sources should be checked for missing values and noise, and further steps, such as error identification and inconsistencies correction, should be conducted. If duplicates are identified, they should be subject to removal. The other step within this process is missing values handling.
- Data Integration – as the data is obtained from multiple different sources, it is also needed to deal with integration, which involves columns aligning and proper naming, tables and datasets merging.
- Data Transformation – for data to be usable for analysis, encoding of categorical variables and normalising numerical features may be needed, where data are scaled to fall within a smaller range, like from 0 to 1.

Some of the forms of data preprocessing are depicted in Figure 3.1 below:

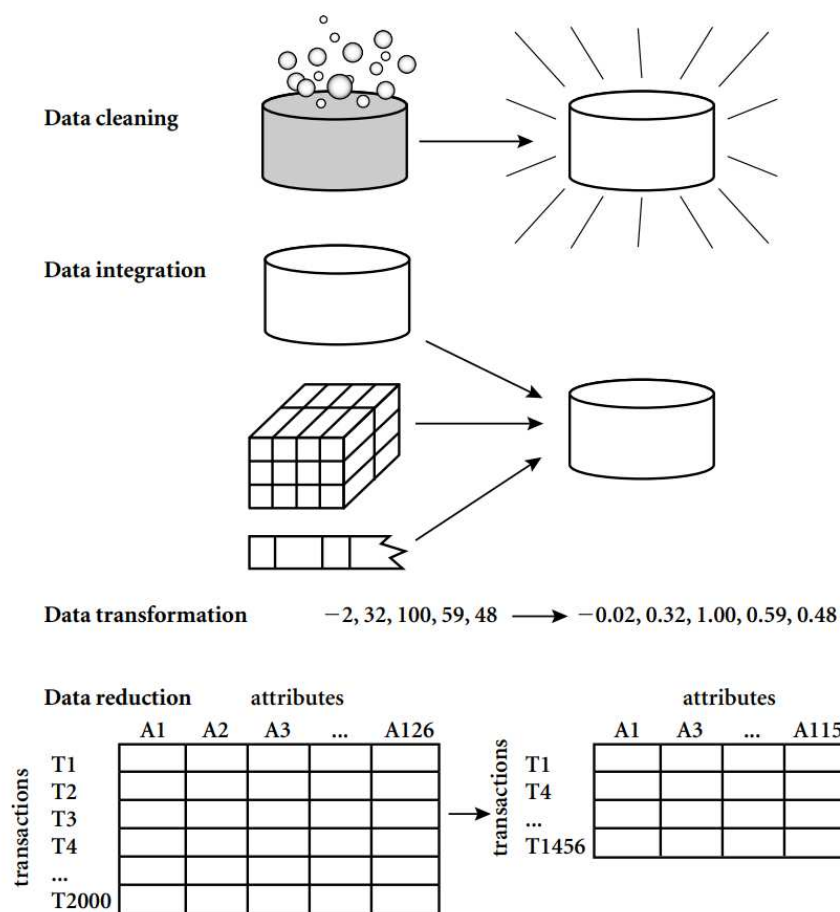


Figure 3.1 – Forms of data processing (Chakrabarti et al. 2008)

Low-quality data is considered to be data with errors, missing values, and inconsistencies, and data quality assessment is a must-do step before any analysis or preprocessing. The following steps can allow proper low-quality handling to ensure the elimination of any inconsistencies in the data set.

Missing values can occur for a wide variety of reasons. Among them, human error appearing during data entry can be mentioned. Also, this can be because of issues with the process of data collection or if certain data fields are considered not applicable. This is the most common challenge when ensuring data quality.

Missing values may cause skewed analyses as well as can result in bias occurring in the models. Appropriate handling of these values is an important step that should be performed to maintain the analysis integrity. But before handling them, it is necessary to identify them, which may be done using the command presented below (Figure 3.2), which prints the count of missing values in each column:

```
# Importing necessary library
import pandas as pd

# Loading your dataset
df = pd.read_csv('your_file.csv') # Replace 'your_file.csv' with your filename

# Checking for missing values in each column
missing_values = df.isnull().sum()
print(missing_values)
```

Figure 3.2 – Identification of missing values command (Data Science Horizons, 2023)

Missing data handling is a crucial step before conducting the analysis because it can introduce bias or inaccuracies. There are several techniques to handle it, with data imputation among them. This is the process which includes missing data replacement with substituted values. This can be the mean imputation (applied for continuous data), median (used for ordinal data), or mode (for categorical data). Statistical models or ML algorithms can be utilised to predict missing values based on other data, which is predictive imputation. Also, one of the other methods is to use a constant to fill the missing value (Chakrabarti et al. 2008).

Inconsistent data is one of the obstacles that can occur in diverse forms. An example of this can be numeric data stored as text, casing in string data, or date formats. To ensure data integrity, the following command shown in Figure 3.3 may be needed that converts the values in numeric_column to a numeric format, converting non-numeric values to NaN:

```
# Example: Converting a column with numeric values stored as strings to numeric format
df['numeric_column'] = pd.to_numeric(df['numeric_column'], errors='coerce')
```

Figure 3.3 – Command allowing to convert to numeric format (Data Science Horizons, 2023)

Outliers are considered data points that significantly differ from other observations in the dataset. They can appear for reasons such as measurement or data entry errors. Although they may be valid, they are considered as extreme observations. Outliers can impact the data analysis results, and also they can negatively affect predictive modelling performance. It is, therefore, critical to identify and handle them in the appropriate way. The process of outlier detection is described in more detail in Chapter 3.3. Among statistical methods, the Z-score is the method where a data point with a Z-score greater than 3 or less than -3 is considered an outlier, and the IQR method states that the data points that fall below the first quartile or above the third quartile are considered to be an outlier. Visualisation or graphical methods suggest box plots and scatter plots utilisation for outliers detection.

There are several strategies to handle outliers, and one of them is deletion. This approach is the most straightforward and simplest but can be applied if the origin of it is known or obvious (applicable in case of incorrectly measured data) and can lead to a data point loss. Another way is transformation, such as log or square root, which can reduce the extreme values.

Data normalisation and scaling help to standardise the range of independent variables or features of data. This critical data preprocessing step aims to standardise or rescale the values of different features in a dataset and should also be applied before running the machine learning algorithms. The objective is to ensure that all the features are brought to a common scale or range, which helps achieve several goals. First, by scaling features to the same range, we can exclude the scenario where a feature dominates the others during modelling, making it easier to compare their contributions. Second, the normalisation process ensures the better performance of machine learning algorithms when features are on a similar scale. Normalised data can also lead to improved interpretability of model coefficients. It allows to make more meaningful comparisons between feature weights.

Several common normalisation techniques include Min-Max scaling, Z-score (Standardization) scaling, and Robust scaling. For instance, Min-Max scaling can be applied to rescale data to a predefined range (for example, [0, 1]), while Z-score scaling transforms data to have a mean of 0 and a standard deviation of 1. Robust scaling performs similarly to min-max scaling but uses the interquartile range rather than the min-max, making it robust to outliers (Chakrabarti et al. 2008).

Encoding categorical variables is a technique applicable to data that may be divided into groups. The one-hot encoding technique allows assigning a binary value of 0 or 1 based on a threshold.

Data integration approach should be applied when dealing with data from different sources, and involves data merging, joining or concatenating. When merging, the combination of two or more data sets is done based on common columns. Joining is the process of combining columns of different data frames into a single result data frame. Concatenating is the process of changing a data set by adding rows or columns.

3.2 Data interpretation

The data preparation, cleaning and preprocessing process included the following steps:

1) Ground Water Level (GWL) and Ground Water Temperature (GWT) data were taken as monthly data for the exact location 341438 (the closest data set, District 22, Vienna) in the years range of 1996-2016 from eHYD (Bundesministerium für Landwirtschaft, Regionen und Tourismus. (n.d.). eHYD.). The measurements of GWL were conducted at different time (from 00:00 to 12:50) and in different months (with most measurements taken in July), as shown in Figure 3.4:

Month	05 - May				06 - June				07 - July						08 - August			09 - September									
Day	16	21	27	3	7	13	23	25	2	13	14	16	24	25	26	27	31	4	10	22	6	10	total				
Time	00:00	2019																						00:00	1		
	05:40																2014								05:40	1	
	06:30						2016																		06:30	1	
	06:50									2017															06:50	1	
	06:55								2015														2018		06:55	2	
	07:01		1996																						07:01	1	
	07:15					2013																			07:15	1	
	08:30													2001											08:30	1	
	08:35																2003								08:35	1	
	08:40		1999																						08:40	1	
	08:55															2012									08:55	1	
	09:10																		2000						09:10	1	
	09:25																					2002			09:25	1	
	09:50											2005													09:50	1	
	10:25				2004																				10:25	2	
	10:35																								10:35	1	
	11:00									2009				2008												11:00	1
	11:11												1998	1997											11:11	2	
	12:50					2010													2011					2007	12:50	3	

Figure 3.4 – Ground Water Level (GWL) Measurements

As for GWT measurements, recordings were taken monthly (on the first date) at 00:00, and more information about the measurements is shown in Figure 3.5 below:

File (Measurement point Nr)	Exact location (Grad. Min. Sec.)		Exact depth of measurement	Number of measurements	Year Start	Year End
341438	16 28 33	48 11 47	6,05	348	1991	2019

Figure 3.5 – Ground Water Temperature (GWT) Measurements

2) Weather data was taken from the weather website visualcrossing.com (Visual Crossing Corporation. Weather Data Services) for District 22 as daily data. To conduct further analysis of the dependency of groundwater temperature on the weather temperature, the following

features were identified as the most influencing: weather temperature, dew, humidity, precipitation, snow data, wind speed, cloud cover, solar radiation, solar energy, UV-index and sunlight time. The data range was also chosen between 1996 and 2016 years. The sunlight duration time was calculated as the difference between the sunset and sunrise time.

- 3) In the above datasets, the missing data was filled in as the average of the same dates from the same months of different years.
- 4) The weather data was transformed from daily to monthly data via averaging.
- 5) The weather data was merged with GWT and GWL, and the resulting file was subjected to better representation (via changing commas to dots, all data representation with two decimals after dots, and presenting UV-index as a whole number and sunlight time as time representation)

After the data merge process, the following result file was created, as shown in Figure 3.6:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	datetime	temp	dew	humidity	precip	snow	snowdepth	windspeed	cloudcover	solarradiation	solarenergy	uvindex	Sun light time	GWL	GWT
2	1996-01	-2,78	-4,46	88,44	0,72	0,48	11,91	24,64	83,48	56	4,82	3	08:50:30	153,25	10,9
3	1996-02	-2,71	-6,41	76,66	0,78	0,43	13,92	30,87	66,49	88,08	7,6	4	10:13:01	153,23	10,9
4	1996-03	2,17	-1,94	75,83	0,48	0,15	4,49	25,72	69,94	157,64	13,58	5	11:56:17	153,26	10,1
5	1996-04	10,07	3,73	68,49	1,56	0,01	1,71	25,23	57,4	217,38	18,78	7	13:42:13	153,34	8,3
6	1996-05	15,64	10,79	74,88	1,25	0	0,02	27,14	65,57	246,73	21,3	7	15:13:30	153,33	9,3
7	1996-06	18,99	12,83	69,39	0,42	0	0	23,09	48,8	278,38	24,04	8	16:00:38	153,24	10,2
8	1996-07	18,54	12,3	69,11	0,33	0	0	26,39	56,7	278,59	24,07	8	15:36:22	153,16	11,9
9	1996-08	19,14	13,78	72,97	1,21	0	0,01	22,93	57,99	240,95	20,82	7	14:15:21	153,03	12,6
...
244	2016-03	6,74	1,22	69,34	0,44	0	0	19,85	74,89	155,07	15,37	3	11:36:49	153,52	9,7
245	2016-04	11,23	4,11	63,87	1,31	0	0	18,61	64,29	217,32	18,75	7	13:42:44	153,4	9,7
246	2016-05	15,65	9,02	66,68	3,5	0	0	20,56	66,2	245,98	21,23	7	15:13:52	153,46	10,8
247	2016-06	20,18	13,38	67,11	3	0	0	17,37	59,64	262,26	22,62	7	16:00:40	153,47	12,1
248	2016-07	22,24	14,73	64,43	3,58	0	0	18,53	59,55	280,66	24,24	8	15:36:01	153,46	14,1
249	2016-08	20,34	13,41	66,49	1,95	0	0	17,21	47,96	241,63	20,9	7	14:14:50	153,43	15,9
250	2016-09	18,81	12,03	67,23	0,81	0	0	13,86	47,12	195,1	16,83	6	12:33:00	153,39	16,8
251	2016-10	9,97	6,67	80,73	2,19	0	0	16,85	79,33	105,28	9,11	4	10:47:46	153,36	16,7
252	2016-11	5,18	1,85	80,04	1,23	0	0	18,2	70,75	60,7	5,22	3	09:14:08	153,26	15,9
253	2016-12	1,48	-1,96	79,14	0,61	0,13	0,15	19,47	68,13	62,14	5,36	3	08:24:51	153,24	13,7
254															

Figure 3.6 – Intermediate data set after the initial preprocessing step

The features and dimensions of the values in each column are mentioned below:

- 1) datetime: This column represents the date in the format "YYYY-MM" (Year and Month). It is used as a timestamp for the data points.
- 2) temp: This column displays the temperature in degrees Celsius (°C). It represents the average monthly temperature.
- 3) dew: The "dew" column represents the dew point temperature. The dew point is the temperature at which air becomes saturated, and moisture in the air begins to condense into water droplets. It is measured in degrees Celsius (°C).

- 4) humidity: This column represents the relative humidity percentage (%). Relative humidity is a measure of the amount of moisture in the air relative to the maximum amount the air could hold at that temperature.
- 5) precip: This column represents the amount of precipitation in millimetres (mm). Precipitation includes any form of water, such as rain, snow, freezing rain and ice.
- 6) snow: This column represents the amount of snowfall, in centimetres (cm), for that month.
- 7) snowdepth: The "snowdepth" column represents the depth or thickness of the snow on the ground. It is also measured in centimetres (cm). Snow is the amount of snow that fell during the exact time period. Snow depth should be interpreted as the average amount of snow being on the ground at a specific period. Snow depth will build up with snowfall and decrease when melting.
- 8) windspeed: This column represents the wind speed in kph units (kilometres per hour, km/h) during that month.
- 9) cloudcover: The amount of sky covered by clouds during the month is indicated in the "cloudcover" column in per cent (%).
- 10) solarradiation: Solar radiation is known as the amount of energy received from the sun. The measurement unit is watts per square meter (W/m^2). It indicates the amount of solar energy available during that month.
- 11) solarenergy: This is another measure of solar energy in megajoules per square meter (MJ/m^2).
- 12) uvindex: The UV (Ultraviolet) index measures the strength of ultraviolet radiation from the sun. The risk of skin damage from sun exposure is often assessed by it (no units).
- 13) Sun light time: This column represents the amount of daylight time for each month, typically in hours and minutes (HH:MM:SS).
- 14) GWL: The "GWL" values represent groundwater level and are measured in meters. This term defines the depth where water is found below the ground's surface.
- 15) GWT: The "GWT" values represent groundwater temperature and are measured in degrees Celsius ($^{\circ}\text{C}$). It represents the temperature of the groundwater.

3.3 Identifying outliers

It is possible that features that differ significantly from most of the data can occur in the real data set. These anomalies or inconsistencies in the data can result in problems in further analysis; for instance, the quality of the decision system can be adversely affected by them. The data set can be damaged due to several factors, among which human errors can be mentioned together with the inherent variability of the field. Also, rounding and transcription errors, as

well as tool glitches, may occur (Nassreddine et al. 2023). Outlier identification is one of the key steps in data preprocessing for predictive modelling in various fields, including geothermal water production.

Therefore, it is important to detect these data anomalies through data examination. In data analysis terminology, this process is known as “outlier detection”. According to Hawkins (1980), an outlier is defined as *an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. Barnett and Lewis (1994), as well as Hodge, V.J. and Austin, J. (2004), indicate that an outlier is *one that appears to deviate markedly from other members of the sample in which it occurs*. Similarly, Johnson (2002) defines an outlier as *an observation in a data set that appears to be inconsistent with the remainder of that data set*.

Several techniques are used to find outliers. The techniques applied to the given dataset are described below. These are graphical and quantitative identification methods.

3.3.1 Graphical identification of outliers

The first technique to identify outliers involves graphing data. This includes histograms, or graphs such as box plots; also, scatter plots can be used. Graphical methods offer a visual way to start the outlier detection process. They can give insights into the distribution and highlight potential outliers in the given dataset. The steps for identifying a potential outlier start with using a single construct technique visual tool. This is followed by using a multiple construct technique to run an additional check. The final procedure is error handling, which requires error correction or removal (Aguinis et al., 2013).

3.3.1.1 Single construct technique – histograms

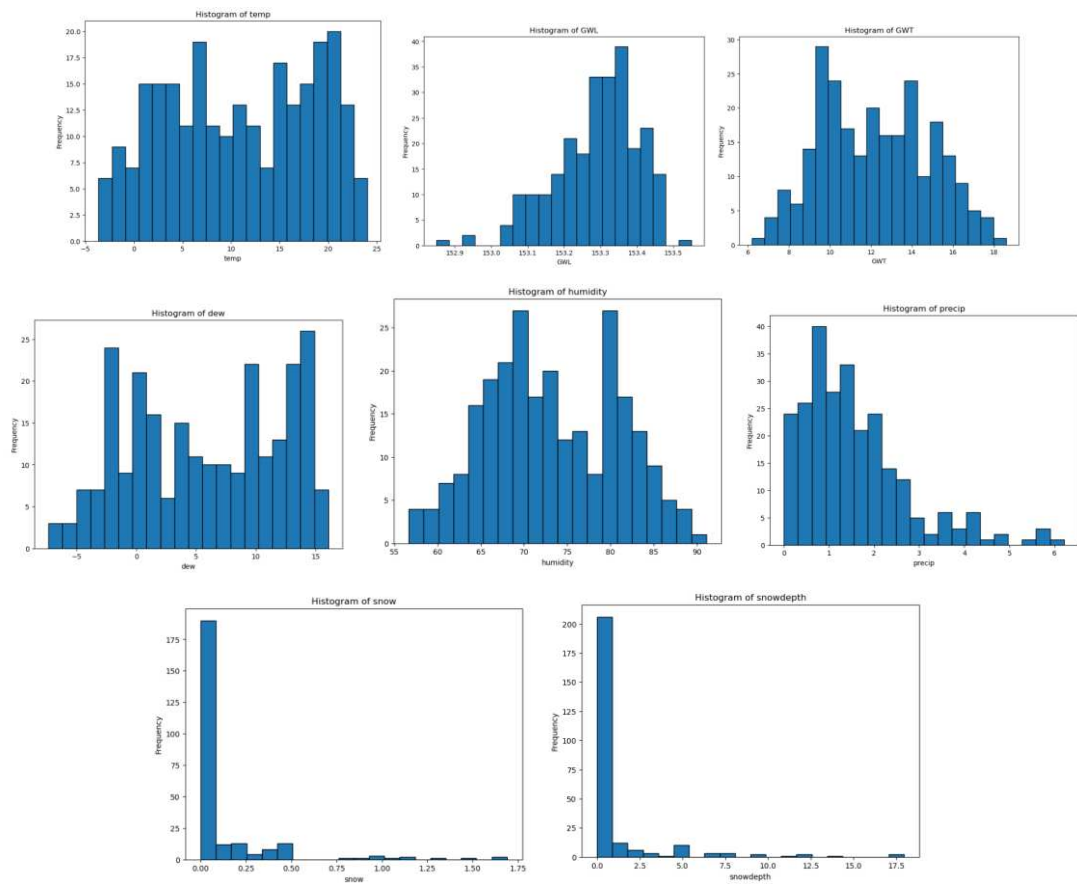
As a *single construct technique*, it was decided to use *histograms*, which allow visualisation and analysis of the distribution of one variable within a given time. A histogram is considered to be a graphical representation of the distribution of a dataset. To create a histogram, it is needed to split the range of values in the dataset into equally sized intervals or bins (in the dataset we work on, the data is provided with the monthly frequency); therefore, each bin represents a specific range of values, namely – one month. The height of each bar in the histogram corresponds to the number of data points that fall within that bin. After the histogram is created, it is essential to understand the shape: it can reveal important information about the data distribution. Common shapes include normal (also referred to as bell-shaped), skewed (that can be positively or negatively), bimodal (this means having two peaks), and uniform (also termed flat). Outliers in a histogram are the data points that are located far from the central part

of the distribution. Outliers may appear as isolated bars in the tails or as individual points; most importantly, these points or bars will show a notable difference from the rest of the data.

In our case, we can apply histograms to each numeric column (features) in our dataset. For each numeric column (e.g., "temp," "dew," "humidity," etc.), a separate histogram was created.

The number of bins was chosen: as there are 191 months in the data set, using 191 bins might not provide a clear visualisation, so it was important to select the proper number of bins that best represent the data distribution. Too few bins could oversimplify the data, while too many bins could obscure important patterns. Using the Square Root Rule (the rule is to use the square root of the number of data points as the number of bins) suggests that for 191 data points, that would be around 14 bins. The data output was also checked with different bin numbers, such as 10 to 20, and the result with 20 bins allowed us to see patterns and potential outliers better.

The histograms for each column were made using the plotting library Matplotlib in Python. The results are represented below in Figure 3.7:



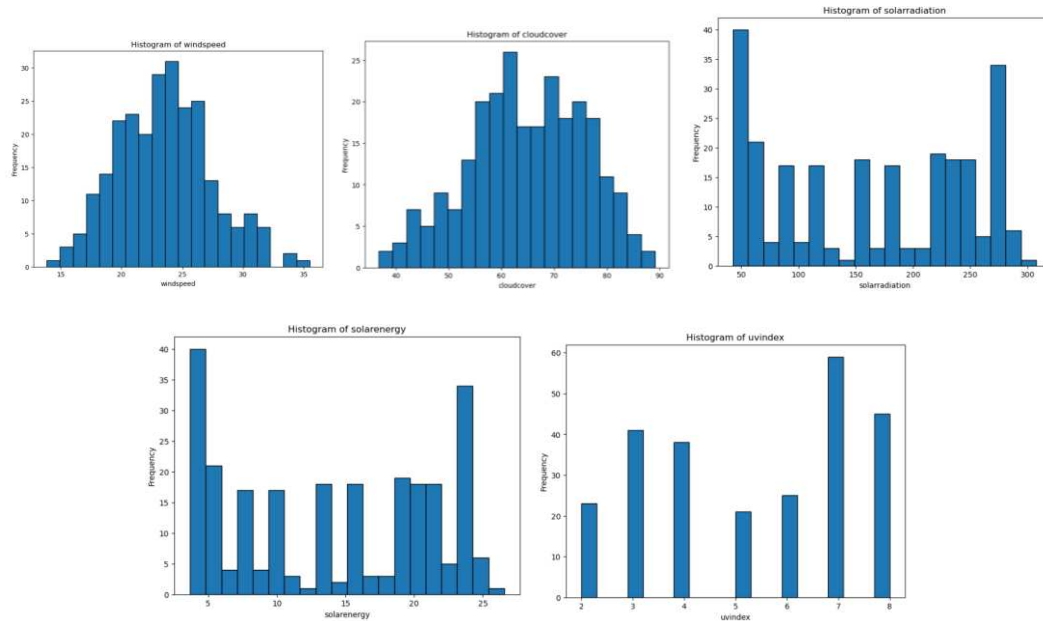


Figure 3.7 – Visual outliers identification via histogram plotting

The visual inspection step included the examination of the shape of each histogram. The bars that are significantly shorter than the majority may mean that the frequency of the value occurrence was minimal; this can represent a potential outlier. Data points (or bars) that are located far from the central part of the distribution may also be potential outliers. First, if we analyse the y-axis, it can be noted that there are some graphs with the y-axis range of 0-20, which means that most bins have relatively few data points (not more than 20 data points). This suggests a relatively evenly distributed dataset with no extreme values. A range of 0-40 indicates that some bins have a moderate count of data points, but there are no bins with extremely high counts. The histograms with snow and snow depth require more careful examination: the wider range, like 0-175 or 0-200, indicates that some bins have a significant number of data points, and there may be some bins with a relatively high count. This could suggest a dataset with some extreme values or a strong concentration of data in specific intervals. Therefore, theoretically, other values besides these extreme values can be treated as outliers.

Symmetry and skewness analysis:

- 1) Precipitations and GWT can be considered a positively skewed distribution: here, the right tail of the histogram is longer or stretched out to the right, meaning that most data point concentration is located on the left side. On the right side, only a few extreme values can be observed.
- 2) Unlike previously mentioned, GWL and cloud cover represent the negatively skewed distribution with more data points concentrated on the right side.

3) The histogram of windspeed has a bell-shaped curve and can be considered a symmetric distribution. This means the data is evenly distributed around the central value, and there are an equal number of data points on both sides of the mean.

4) Solar energy, solar radiation and UV index histogram can be interpreted as a multimodal distribution, as there are more than two distinct peaks in a histogram. This means the data may have several different modes or groups.

Looking at the GWL histogram, it can be noticed that there are some tails (152.9 and 153.5). Theoretically, these tails can represent the extreme values of the distribution, but practically, this is an insignificant variation of the depth measurement. With a scale of approximately 153 meters, this fluctuation is neglectable.

The snow values require a detailed analysis. To continue the procedure, it can be necessary to encode snow depth as a *categorical or binary variable*. For example, a threshold can be established, with 1 indicating the presence of snow (above the threshold) and 0 indicating no snow. This can help maintain the information about snow, even without treating it as having outliers.

3.3.1.2 Multiple construct technique – scatter plots

Single construct technique (histograms) provided insights into the distribution characteristics of parameters. We identified skewness, symmetry, and multimodality in the data, providing a foundation for further exploration. While the GWL histogram revealed minor fluctuations within the scale, snow depth emerged as a point of interest requiring more detailed treatment.

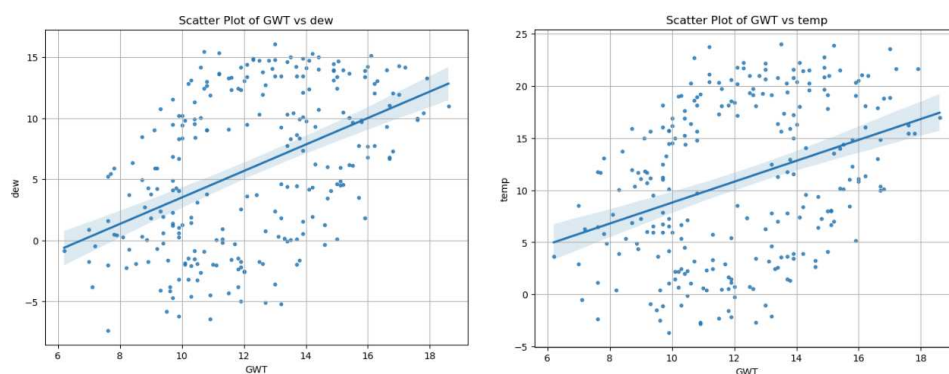
The procedure was followed by *the multiple construct technique*, namely *scatter plot*. This plot contains two main components – points and a trend line. Each data point is represented as a dot or marker on the plot, with one variable on the x-axis (usually the independent variable) and another on the y-axis (usually the dependent variable). A trend line can be additionally drawn for the general trend representation or to highlight the correlation between the variables. A data point located far away from the centroid of the data can be treated as a potential outlier (Aguinis et al. 2013).

In this project, the dependency of groundwater temperature on various weather parameters is studied. The following pairs of variables that are likely to have an impact on groundwater temperature were selected.

- 1) Groundwater Temperature (GWT) vs Weather Temperature (temp): This is a fundamental relationship to explore since weather temperature directly affects groundwater temperature.

- 2) Groundwater Temperature (GWT) vs. Dew Point Temperature (dew): Dew point temperature can affect the condensation and evaporation of moisture in the atmosphere, which may influence groundwater temperature.
- 3) Groundwater Temperature (GWT) vs. Humidity: Humidity levels can influence temperature by affecting heat transfer and energy exchange.
- 4) Groundwater Temperature (GWT) vs. Precipitation: It is important to investigate the effect of precipitation on groundwater temperature, especially if there are significant rainfall events.
- 5) Groundwater Temperature (GWT) vs. Solar Radiation or Solar Energy: These variables can influence the heating or cooling of the ground surface, potentially impacting groundwater temperature.
- 6) Groundwater Temperature (GWT) vs. UV Index: This relationship is also of interest, as UV radiation can affect heat absorption.
- 7) Groundwater Temperature (GWT) vs. Wind Speed: Wind can influence the rate of heat exchange taking place between the ground and the atmosphere, which may impact groundwater temperature.
- 8) Groundwater Temperature (GWT) vs. Cloud Cover: cloud cover can affect solar radiation and heat exchange.
- 9) Groundwater Temperature (GWT) vs. Sunlight Time: Sunlight duration can directly affect the heating of the ground surface and, consequently, groundwater temperature.

By analysing these relationships using scatter plots and trend lines, we can gain insights into the dependencies of groundwater temperature on various weather parameters and identify potential correlations or trends, as well as find outliers. The scatter plots are shown below in Figure 3.8:



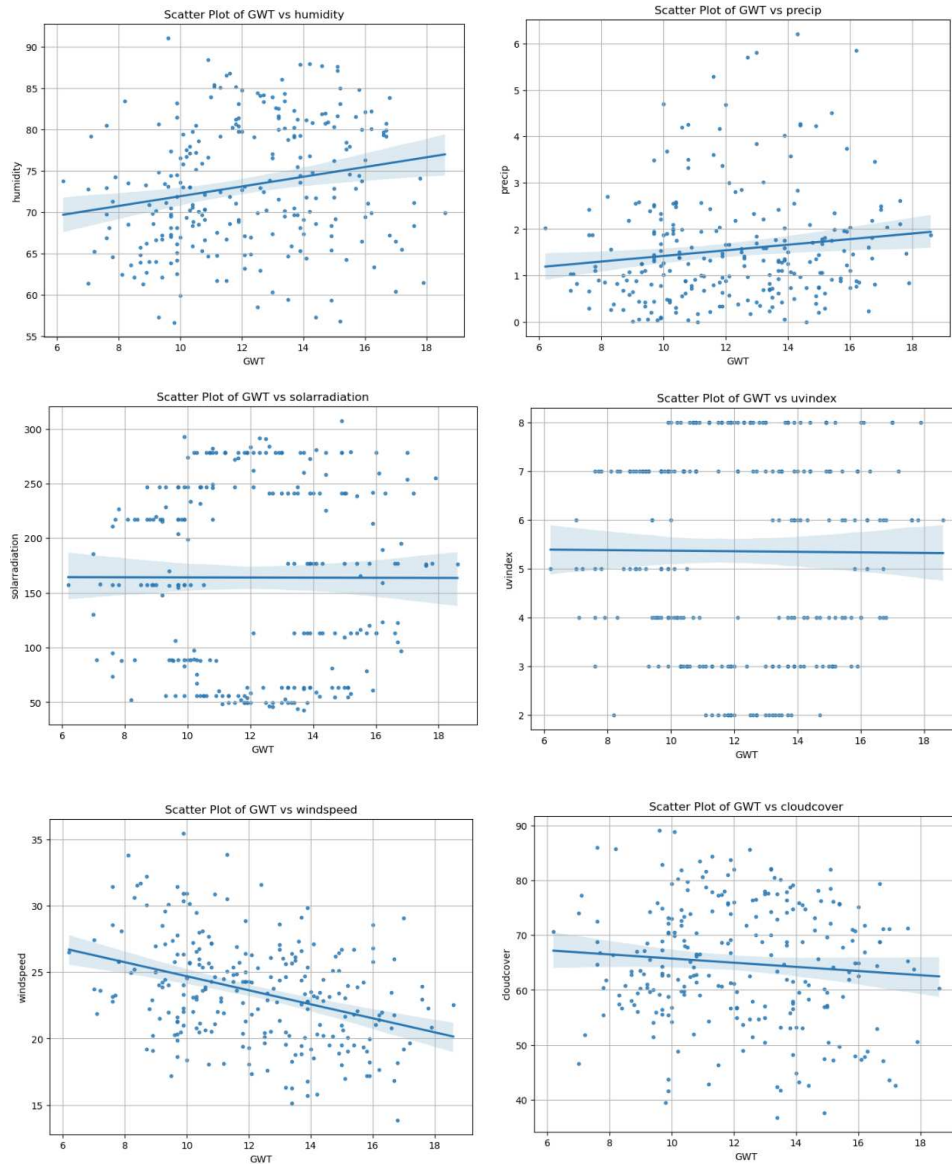


Figure 3.8 – Visual outliers identification via multiple construct technique

In terms of outlier detection, the visual inspection strategy can allow for the noticing of data points with significant deviations from the trend line. Unusual patterns can also represent outliers. However, interpreting and understanding these deviations in the context of our data is important. For example, if we observe data points for precipitation, we notice that most of the points cluster around low values, such as 0, 1, 2, or 3 mm. There are some points that show significantly higher values, like 6 mm, which are far from the trend line, but they should not be considered data errors, as it is natural. Similarly, in most cases, the data points shown above appear valid and meaningful, but for data analysis, they should be eliminated for higher modelling accuracy.

Nevertheless, the scatter plots represent the dependencies, which can be interpreted as follows:

- 1) **Trend Line Direction:** plots of the ground-water temperature with dew point, weather temperature, humidity and precipitations show positive steepness of the trend line. The upward direction of the trend can be explained naturally: the higher each value is, the higher the ground-water temperature. Conversely, the wind speed and the cloud cover show a negative steepness, and this inverse relationship can be explained in the following way: cloud cover can negatively affect the heat exchange.
- 2) **Strength of Correlation:** When analysing how closely the data points clusters are located around the trend line, it is possible to identify the strongest dependencies between the variables. The strong correlations are indicated by closely packed points. When comparing precipitations and humidity, it can be clearly seen that the feature “precipitations” has a stronger dependency and should be considered to have a greater impact than the “humidity” feature.

In summary, the analysis of scatter plots with trend lines involves a combination of visual inspection together with a contextual understanding of the data. This analysis can help identify potential outliers and assess the strength and direction of relationships between variables. Nevertheless, this is a qualitative approach. Therefore, further quantitative methods are needed to get more precise insights into outlier identification.

3.3.2 Quantitative identification of outliers

Statistical methods are used to detect outliers precisely based on statistical tests or procedures. These methods are essential for making data-driven decisions and identifying unusual observations that may impact the reliability of analyses or models. In this work, two commonly used techniques for identifying outliers are used: the Z-score methodology and the Interquartile Range (IQR).

3.3.2.1 Z-score methodology

The Z-score method, also known as the standard score, is a system that helps representing abnormal behaviour occurrences in terms of their association with the standard deviation and mean of a collection of arguments. It is a statistical measure that quantifies how many standard deviations a data point is away from the mean of the dataset. Estimating the Z-score is plotting the items into a scattering diagram, where the standard deviation is indicated as 0 and the mean as 1. Estimating Z-scores, we can meet several objectives: to eliminate the properties of the position and scale of the data points, thus permitting dissimilar datasets to be associated exactly. To recognise anomaly, the Z-score technique utilises the following technique: after plotting the data items in the scattering diagram, it considers as anomalies the items that are far from the value zero (Anusha et al., 2019).

The standard Z-Score includes the calculation of the distance between data points and the mean divided by the standard deviation. The Z-Score can be positive or negative. If the value stays above the mean, it is positive; otherwise, it is negative. For each data point in a specific column (e.g., GWT), Z-score can be calculated using the formula (Jamshidi, E. J):

$$z = \frac{x - \mu}{\sigma}, \quad (3.1)$$

where z – stands for the Z-score, x is considered as the data point, μ represents the mean of the column, and σ is the standard deviation of the column.

Then, a threshold for the Z-score above which data points are considered outliers needs to be determined. Common threshold values include 2 or 3 standard deviations from the mean (in this work, 3 was chosen).

The code was applied to all columns, and this is the result:

- 1) *Sun light time* (as the data was in time representation it was converted to seconds), *GWT*, *temp*, *dew*, *humidity*, *cloudcover*, *solarradiation*, *solarenergy*, *uvindex* showed no outliers.
- 2) *GWL* – analysis identified two outliers: lines 22 and 23, with the values of 152.85 and 152.94, as shown in Figure 3.9:

	cloudcover	solarradiation	solarenergy	uvindex	Sun light time	GWL
22	73.85	63.78	5.49	3	09:15:08	152.85
23	84.39	49.85	4.29	2	08:25:00	152.94

Figure 3.9 – Outliers defined for column representing *GWL*

- 3) *Precip* – five outliers were determined: lines 79, 127, 140, 161, and 224 with the highest values of precipitations (more than 5.25 millimetres), as shown in Figure 3.10:

Z-Score Outliers:						
	datetime	temp	dew	humidity	precip	
79	2002-08	20.64	16.08	76.55	5.80	
127	2006-08	18.07	12.69	72.48	5.70	
140	2007-09	14.06	9.33	74.83	6.21	
161	2009-06	18.09	12.32	71.92	5.29	
224	2014-09	16.07	12.85	82.22	5.85	

Figure 3.10 – Outliers defined for column representing precipitations

- 4) *Snow* – several outliers were identified, as shown in Figure 3.11:

```

Z-Score Outliers:
  datetime temp  dew  humidity  precip  snow
84  2003-01 -0.13 -3.25   80.48   1.39  0.95
96  2004-01 -1.05 -4.57   77.92   2.01  1.03
97  2004-02  3.62 -1.60   71.19   2.39  0.99
98  2004-03  4.88  0.07   73.25   2.19  1.16
108 2005-01  2.24 -1.92   75.14   1.50  1.12
109 2005-02 -1.07 -4.67   77.35   1.95  1.64
119 2005-12  0.58 -2.38   81.22   2.04  0.98
179 2010-12 -2.66 -5.08   84.16   0.97  1.45
204 2013-01  0.42 -2.22   83.46   2.70  1.69
205 2013-02  1.15 -2.00   80.50   2.42  1.32

```

Figure 3.11 – Outliers defined for column representing the amount of snowfall

- 5) *Snowdepth* – several outliers (notably, which differ from *snow*), as shown in Figure 3.12:

```

  datetime temp  dew  humidity  precip  snow  snowdepth
0  1996-01 -2.78 -4.46   88.44   0.72  0.48   11.91
1  1996-02 -2.71 -6.41   76.66   0.78  0.43   13.92
12 1997-01 -2.50 -3.82   91.09   0.81  0.48   17.94
48 2000-01 -0.93 -4.10   80.67   1.25  0.48   17.33
109 2005-02 -1.07 -4.67   77.35   1.95  1.64   11.98
110 2005-03  4.36 -1.72   66.23   0.59  0.26    9.13
120 2006-01 -3.65 -6.19   83.18   1.21  0.03    9.18
179 2010-12 -2.66 -5.08   84.16   0.97  1.45   11.16

```

Figure 3.12 – Outliers defined for column representing the snow depth

- 6) *Windspeed* – one outlier (line 37) was identified, as shown in Figure 3.13:

```

Z-Score Outliers:
  datetime temp  dew  humidity  precip  snow  snowdepth  windspeed
37 1999-02  1.05 -3.18   74.53   0.08  0.45    7.58    35.48

```

Figure 3.13 – Outliers defined for column representing wind speed

The most outliers were identified in the columns “snow” and “snow depth”. After running the analysis and learning about the physical meaning of the column “snow”, it was decided to exclude this feature, and to continue only with the “snowdepth”. Talking about “snowdepth”, – it was noticed that winter months with snow presence and heavy snow were treated as the outlier, which requires further data processing, namely binary coding. Precipitation and windspeed – these can be the highest values, representing months with stronger winds and more precipitation.

3.3.2.2 Interquartile range (IQR)

Another technique which can be used is Interquartile range or IQR. Interquartile range is similar to range but is applied when working with data sets with extreme outliers. The IQR is the value of the middle half of the data set. IQ is used to determine "fences" around data and then define outliers, which will be represented by values outside the fences (Figure 3.14). The interquartile range can be found by subtracting the Q_1 value (this is the value below which 25 per cent of the

distribution lies) from the Q_3 value (which is the value below which 75 per cent of the distribution lies).

The upper fence value can be calculated as $Q_3 + (1.5 * IQR)$, and the lower fence $Q_1 - (1.5 * IQR)$ (Nassreddine et al. 2023).

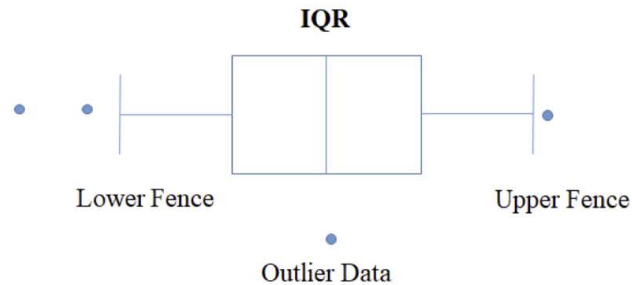


Figure 3.14 – IQR technique (Nassreddine et al. 2023)

In other words, the Inter Quartile Range tells us the variation in the data. Any value which lies outside the range of (25th Percentile–1.5x Inter Quartile Range) to (75th Percentile + 1.5x Inter Quartile Range) is detected as an outlier where IQR is defined as 75th Percentile – 25th Percentile.

Thus, Nassreddine G. (2023) presents an example using the Python programming language. In this example, a temperature measurement in Iraq was presented using: $T = [25, 30, 35, 4, 75, 18, 45, 35, 25, 70, 12, 15]$. The temperature is normal, between 0 and 40 degrees. Figure 3.15 shows the code in Python for illustrating IQR and the results of detecting outliers' data using the interquartile range (IQR). According to this picture, the extremes are: [75, 45, 70].

```
[12]
T = [25,30,35,4,75,18,45,35,25,70,12,15]
Q1 = np.percentile(T, 10, interpolation = 'midpoint')
Q2 = np.percentile(T, 20, interpolation = 'midpoint')
Q3 = np.percentile(T, 40, interpolation = 'midpoint')

IQR = Q3 - Q1
print('Interquartile range is', IQR)
low_lim = Q1 - 1.5 * IQR
up_lim = Q3 + 1.5 * IQR
print('low_limit is', low_lim)
print('up_limit is', up_lim)
outlier = []
for x in T:
    if ((x > up_lim) or (x < low_lim)):
        outlier.append(x)
print(' outlier in the dataset is', outlier)

Interquartile range is 11.5
low_limit is -3.75
up_limit is 42.25
 outlier in the dataset is [75, 45, 70]
```

Figure 3.15 – Code in Python for interquartile range (Nassreddine et al. 2023)

The IQR performs less sensitively to extreme values than the Z-score and is especially useful for datasets with non-normally distributed data. This is how the IQR methodology was applied to identify outliers:

- 1) IQR was calculated as: $IQR = Q_3 - Q_1$, where Q_3 is the third quartile, and Q_1 is the first quartile.
- 2) Lower and upper thresholds for identifying outliers were determined. Data points outside of these thresholds were considered outliers. Common thresholds include 1.5 times the IQR below Q_1 and above Q_3 . In this work, 1.5 is used.

The code was applied to all columns, and this is the result:

- 1) *Sun light time, GWT, temp, dew, humidity, cloudcover, solarradiation, solarenergy, uvindex* showed no outliers.
- 2) *Precipitation* – several outliers were identified, as shown in Figure 3.16, and the boxplots before (on the left) and after (on the right) outlier handling are shown in Figure 3.17:

IQR Outliers:

	datetime	temp	dew	humidity	precip
66	2001-07	20.87	13.91	67.15	4.22
68	2001-09	13.97	10.01	78.48	4.50
79	2002-08	20.64	16.08	76.55	5.80
101	2004-06	18.16	12.43	71.11	4.17
127	2006-08	18.07	12.69	72.48	5.70
140	2007-09	14.06	9.33	74.83	6.21
149	2008-06	20.28	14.15	70.12	4.19
161	2009-06	18.09	12.32	71.92	5.29
172	2010-05	14.53	10.23	76.56	4.70
175	2010-08	19.57	13.99	71.70	4.24
198	2012-07	21.60	14.88	68.09	4.02
209	2013-06	18.44	13.05	72.68	4.69
220	2014-05	14.81	9.02	70.32	4.26
222	2014-07	21.54	15.03	68.58	4.27
224	2014-09	16.07	12.85	82.22	5.85

Figure 3.16 – Outliers defined for column representing precipitations

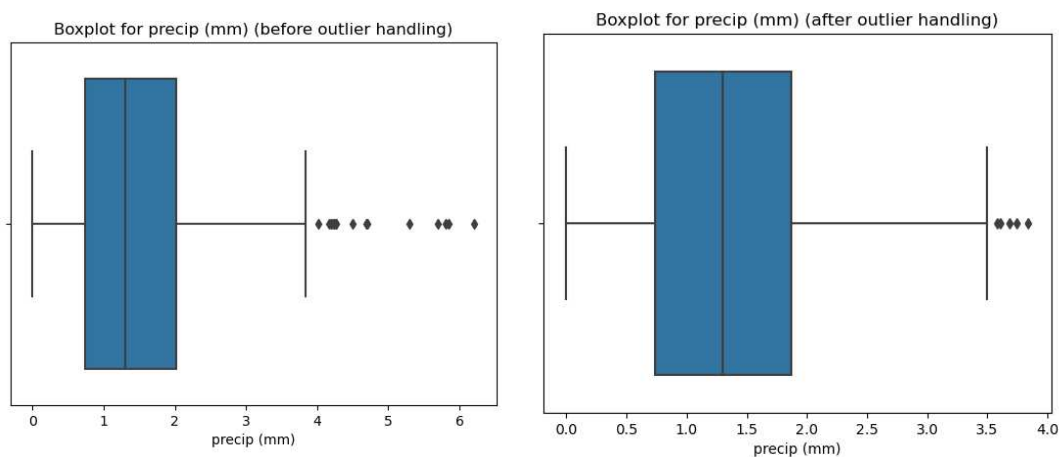


Figure 3.17 – Boxplots before (on the left) and after (on the right) outlier handling

- 7) *Snow* - several outliers were identified, as shown in Figure 3.18:

```

IQR Outliers:
  datetime  temp  dew  humidity  precip  snow
0  1996-01 -2.78 -4.46  88.44  0.72  0.48
1  1996-02 -2.71 -6.41  76.66  0.78  0.43
11 1996-12 -2.26 -4.31  86.59  0.22  0.25
12 1997-01 -2.50 -3.82  91.09  0.81  0.48
13 1997-02 3.90 -0.71  73.52  0.26  0.45
23 1997-12 2.48 0.13  85.85  0.67  0.25
24 1998-01 2.00 -1.51  79.79  0.31  0.48
25 1998-02 5.95 -0.25  66.76  0.33  0.45
35 1998-12 -1.05 -3.57  83.90  0.42  0.25
36 1999-01 0.41 -2.01  85.16  0.00  0.48
37 1999-02 1.05 -3.18  74.53  0.00  0.45
47 1999-12 1.15 -2.19  79.77  1.48  0.25
48 2000-01 -0.93 -4.10  80.67  1.25  0.48
49 2000-02 4.67 0.41  74.30  0.90  0.43
59 2000-12 2.43 0.17  86.05  1.60  0.25
61 2001-02 3.17 -0.89  75.92  0.62  0.35
71 2001-12 -2.10 -5.22  80.04  1.00  0.29
83 2002-12 -0.23 -2.60  84.75  1.67  0.25
84 2003-01 -0.13 -3.25  80.48  1.39  0.95
93 2003-10 8.37 4.35  76.83  1.05  0.19
96 2004-01 -1.05 -4.57  77.92  2.01  1.03
97 2004-02 3.62 -1.60  71.19  2.39  0.99
98 2004-03 4.88 0.07  73.25  2.19  1.16
108 2005-01 2.24 -1.92  75.14  1.50  1.12
109 2005-02 -1.07 -4.67  77.35  1.95  1.64
110 2005-03 4.36 -1.72  66.23  0.59  0.26
119 2005-12 0.58 -2.38  81.22  2.04  0.98
121 2006-02 -0.50 -3.80  79.22  1.03  0.41
122 2006-03 3.67 -0.87  73.75  2.02  0.48
142 2007-11 3.94 0.53  79.24  1.33  0.43
156 2009-01 -1.54 -3.83  85.20  0.85  0.40
157 2009-02 1.47 -1.86  79.47  1.91  0.92
167 2009-12 1.37 -1.61  81.24  1.28  0.39
168 2010-01 -2.16 -4.99  81.42  1.41  0.78
169 2010-02 0.67 -3.17  76.69  0.47  0.35
178 2010-11 7.39 4.19  81.19  1.07  0.42
179 2010-12 -2.66 -5.08  84.16  0.97  1.45
180 2011-01 0.59 -1.96  83.95  0.91  0.35
192 2012-01 2.45 -2.09  73.06  2.09  0.31
193 2012-02 -2.34 -7.36  69.79  0.66  0.23
204 2013-01 0.42 -2.22  83.46  2.70  1.69
205 2013-02 1.15 -2.00  80.50  2.42  1.32
206 2013-03 2.95 -1.75  72.81  1.03  0.31
207 2013-04 11.75 5.24  66.90  0.29  0.19
227 2014-12 3.64 0.88  79.30  1.33  0.19
228 2015-01 2.92 0.29  80.28  1.96  0.19
229 2015-02 2.19 -1.47  77.52  1.10  0.44
240 2016-01 0.78 -2.52  79.72  0.89  0.37
    
```

Figure 3.18 – Outliers defined for column representing the amount of snowfall

8) *Snowdepth* – several outliers were detected, as shown in Figure 3.19:

```

IQR Outliers:
  datetime  temp  dew  humidity  precip  snow  snowdepth
0  1996-01 -2.78 -4.46  88.44  0.72  0.48  11.91
1  1996-02 -2.71 -6.41  76.66  0.78  0.43  13.92
2  1996-03 2.17 -1.04  75.83  0.48  0.15  4.49
3  1996-04 10.07 3.73  68.49  1.56  0.81  1.71
11 1996-12 -2.26 -4.31  86.59  0.22  0.25  4.81
12 1997-01 -2.50 -3.82  91.09  0.81  0.48  17.84
13 1997-02 3.90 -0.71  73.52  0.26  0.45  4.91
14 1997-03 5.81 0.50  71.31  1.20  0.15  2.60
23 1997-12 2.48 0.13  85.85  0.67  0.25  1.66
24 1998-01 2.00 -1.51  79.79  0.31  0.48  4.97
25 1998-02 5.95 -0.25  66.76  0.33  0.45  4.85
26 1998-03 5.33 -1.90  82.17  0.83  0.15  1.52
35 1998-12 -1.05 -3.57  83.90  0.42  0.25  3.20
36 1999-01 0.41 -2.01  85.16  0.00  0.48  4.65
37 1999-02 1.05 -3.18  74.53  0.00  0.45  7.58
38 1999-03 7.24 1.05  70.93  0.80  0.15  1.43
45 1999-10 11.09 6.80  76.32  0.73  0.81  2.50
46 1999-11 3.45 1.53  87.90  1.75  0.07  0.88
47 1999-12 1.15 -2.10  79.77  1.48  0.25  3.63
48 2000-01 -0.93 -4.10  80.67  1.25  0.48  17.33
49 2000-02 4.67 0.41  74.30  0.90  0.43  3.45
50 2000-03 6.46 1.62  72.97  1.87  0.15  1.45
56 2000-09 15.42 10.42  74.09  1.47  0.00  2.68
59 2000-12 2.43 0.17  86.05  1.60  0.25  1.69
83 2002-12 -0.23 -2.60  84.75  1.67  0.25  1.98
84 2003-01 -0.13 -3.25  80.48  1.39  0.95  4.86
96 2004-01 -1.05 -4.57  77.92  2.01  1.03  6.66
97 2004-02 3.62 -1.60  71.19  2.39  0.99  1.50
98 2004-03 4.88 0.07  73.25  2.19  1.16  6.31
108 2005-01 2.24 -1.92  75.14  1.50  1.12  3.75
109 2005-02 -1.07 -4.67  77.35  1.95  1.64  11.98
110 2005-03 4.36 -1.72  66.23  0.59  0.26  9.13
119 2005-12 0.58 -2.38  81.22  2.04  0.98  2.55
120 2006-01 -3.65 -6.19  83.18  1.21  0.03  9.18
121 2006-02 -0.50 -3.80  79.22  1.03  0.41  5.17
122 2006-03 3.67 -0.87  73.75  2.02  0.48  2.22
142 2007-11 3.94 0.53  79.24  1.33  0.43  1.36
156 2009-01 -1.54 -3.83  85.20  0.85  0.40  1.13
157 2009-02 1.47 -1.86  79.47  1.91  0.92  4.64
167 2009-12 1.37 -1.61  81.24  1.28  0.39  1.11
168 2010-01 -2.16 -4.99  81.42  1.41  0.78  8.07
169 2010-02 0.67 -3.17  76.69  0.47  0.35  7.65
179 2010-12 -2.66 -5.08  84.16  0.97  1.45  11.16
180 2011-01 0.59 -1.96  83.95  0.91  0.35  0.92
192 2012-02 -2.34 -7.36  69.79  0.66  0.23  1.63
204 2013-01 0.42 -2.22  83.46  2.70  1.69  6.93
205 2013-02 1.15 -2.00  80.50  2.42  1.32  5.06
    
```

Figure 3.19 – Outliers defined for column representing the snow depth

3) *Windspeed* – the same one outlier (line 37) was identified, as shown in Figure 3.20, and the boxplots before (on the left) and after (on the right) outlier handling are shown in Figure 3.21:

```

IQR Outliers:
  datetime  temp  dew  humidity  precip  snow  snowdepth  windspeed
37 1999-02 1.05 -3.18  74.53  0.08  0.45  7.58  35.48
    
```

Figure 3.20 – Outliers defined for column representing wind speed

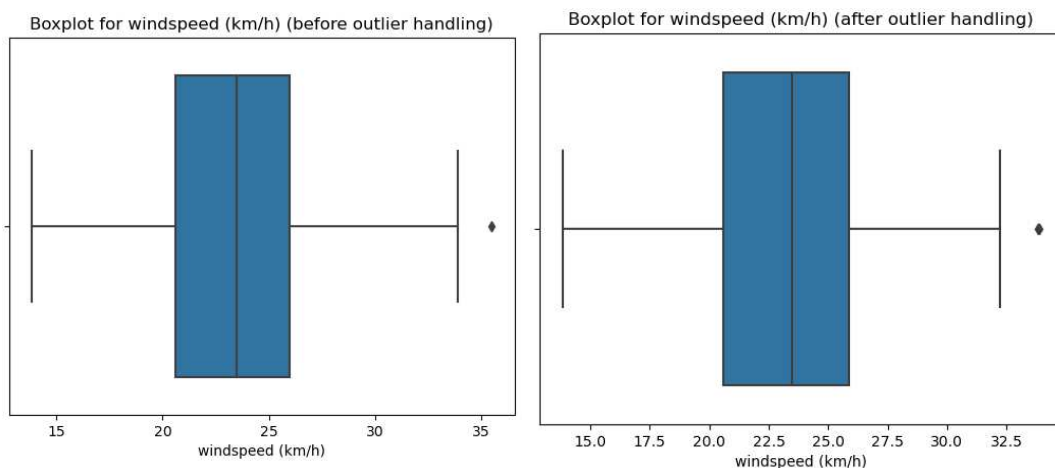


Figure 3.21 – Boxplots before (on the left) and after (on the right) outlier handling

As for the windspeed, this method identified the same outlier as it was identified by the z-score method. As for precipitation, the IQR method could identify even more outliers, which confirms the fact that this method is a more sensitive one (compared to z-score). As for the feature “snow”, the algorithm identified all the values above 0.19 as an outlier, which is not the right interpretation, as values above 0.19 represent the presence of snow and must be considered for the analysis. Similarly, when the algorithm analysed the feature “snowdepth”, most of the values representing snow were treated as outliers.

3.4 Categorical or Binary Encoding

The previous steps identified the necessity of an additional study of the features “snow” and “snowdepth”. Comparing the two features, it can be noted that the “snow” column represents only the amount of snowfall (of new snow that has fallen in the period) and does not provide information representing the depth of the snow. The latter is of importance. Therefore, it can be reasonable to continue working only with the feature “snowdepth”, which represents the depth or thickness of the snow on the ground.

During the previous step, most of the values representing the presence of snow were identified as outliers. To maintain the information about snow events without treating them as outliers is possible if we encode snow depth as a categorical or binary variable.

Binary encoding is a data transformation technique commonly employed in data preprocessing and feature engineering. It serves the purpose of converting continuous or categorical data into a binary format, where 1 typically represents the presence of a specific condition or attribute, while 0 denotes its absence. This approach is particularly useful for simplifying and categorising complex data into a format that is more amenable to analysis, modelling, or classification. In the context of environmental data, as mentioned earlier, binary encoding can

be applied to represent the presence or absence of specific weather-related events, such as snowfall above a certain threshold. This simplifies the interpretation and integration of such events into further analysis, facilitating insights and patterns discovery.

One-hot encoding is a technique used to convert categorical data into a format that can be provided to ML algorithms to improve prediction accuracy. This process involves creating a new binary column for each category of the variable, which can enhance the model's ability to recognise and respond to different snow conditions.

The minimum amount of the “snowdepth” was 0, and the maximum was 17.94. Based on the observed snow depth ranges, it was decided to categorise the snow depth into four distinct groups (represented in the column “Snow_Presence”): no snow (0-2 cm), light snow (3-7 cm), moderate snow (8-15 cm), and heavy snow (more than 15 cm). Each category is represented by a unique integer in a new column named "snow presence," as shown in Figure 3.22, where:

- "0" indicates no snow (0-2 cm),
- "1" represents light snow (3-7 cm),
- "2" shows moderate snow (8-15 cm), and
- "3" denotes heavy snow (more than 15 cm).

By categorising the snow depth into these defined groups, the data becomes more manageable and interpretable for both statistical analyses and machine learning models.

datetime	temp (°C)	dew (°C)	humidity (%)	precip (mm)	windspeed (km/h)	cloudcover (%)	solarradiation (W/m ²)	solarenergy (MJ/m ²)	uvindex	Sun light time (HH:MM:SS)	GWL (m)	GWT (°C)	Snow_Presence
1996-01	-2,78	-4,46	88,44	0,72	24,64	83,48	56	4,82	3	08:50:30	153	10,9	2
1996-02	-2,71	-6,41	76,66	0,78	30,87	66,49	88,08	7,6	4	10:13:01	153	10,9	2
1996-03	2,17	-1,94	75,83	0,48	25,72	69,94	157,64	13,58	5	11:56:17	153	10,1	1
1996-04	10,07	3,73	68,49	1,56	25,23	57,4	217,38	18,78	7	13:42:13	153	8,3	0
1996-05	15,64	10,79	74,88	1,25	27,14	65,57	246,73	21,3	7	15:13:30	153	9,3	0
1996-06	18,99	12,83	69,39	0,42	23,09	48,8	278,38	24,04	8	16:00:38	153	10,2	0
1996-07	18,54	12,3	69,11	0,33	26,39	56,7	278,59	24,07	8	15:36:22	153	11,9	0
1996-08	19,14	13,78	72,97	1,21	22,93	57,99	240,95	20,82	7	14:15:21	153	12,6	0
1996-09	12,5	9,28	81,61	1,33	25,61	82,06	176,66	15,26	6	12:33:33	153	13,2	0
1996-10	11,32	7,77	80,15	0,7	25,68	63,47	113,44	9,79	4	10:48:17	153	13,4	0
1996-11	7,35	3,71	79,11	0,38	31,59	71,81	63,78	5,49	3	09:14:30	153	12,4	0
1996-12	-2,26	-4,31	86,59	0,22	25,38	77,18	49,85	4,29	2	08:24:53	153	11,5	1
1997-01	-2,5	-3,82	91,09	0,81	20,26	89,11	56	4,82	3	08:52:03	153	9,6	3

Figure 3.22 – Result after one-hot encoding of the “snowdepth” column into “Snow_Presence”

3.5 Normalisation Step

In the dataset studied, various features with different units and scales are present, such as temperature in Celsius, precipitation in millimeters, windspeed in kilometers per hour, and solar radiation in watts per square meter, among others. These features naturally span different numerical ranges, which can lead to issues when training machine learning models.

Firstly, the disparity in the scales of features can cause certain features to dominate the learning process, leading the model to give more weight to features with larger numerical values. As a result, the model may become biased towards those features and overlook the contributions of others, potentially leading to suboptimal performance.

Secondly, some machine learning algorithms are sensitive to the scale of the input features. For instance, algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) compute distances between data points, and having features with different scales can distort these distances, leading to inaccurate predictions.

To address these issues, data normalisation and scaling were applied as a crucial preprocessing step. By normalising the data, all features were brought to a common scale or range, effectively mitigating the problems associated with disparate scales. This ensures that no individual feature dominates the others during model training, making it easier for the algorithm to learn meaningful patterns from the data.

In Chapter 4, a comparison of the performance of machine learning algorithms is given before and after the normalisation step implementation. Normalising the data improved model performance and interpretability.

To conclude, this chapter provided a detailed walkthrough of the essential steps needed to prepare data for analysis, which includes collecting and cleaning the data and, generally, preprocessing it for modelling. This involves removing or correcting data that may be incorrect (known as outliers), filling in missing values, and standardising the data to ensure that the models treat all features equally. These steps are crucial because good data quality is key to obtaining reliable and meaningful results from machine learning models. In this chapter, the data set features used are described, and the features selected are interpreted. Also, theoretical steps of data preprocessing are implemented practically on the given dataset, making it ready for further machine learning model training. Next, the focus will shift to applying various machine learning models to this prepared data. The discussion will include exploring different algorithms to see which one performs best at predicting groundwater temperatures based on the processed data. The technical body of the dissertation consists of a number of chapters (just one here, but there will usually be more). Follow a logical structure in how you present your work. This will usually be the phases of the software development cycle, the modules of your system, etc.

Chapter 4

Machine Learning Methods and Algorithms

There are several methods used for geothermal applications: deep neural networks, convolutional neural networks, support neural machines, and others. Among the methods, KNN, CNN and Random Forest models were selected to be used for problem investigation. Therefore, the project objective is to analyse the features influencing groundwater temperature fluctuation to make further temperature map creation more accurate using predictive ML methods (meaning that they can predict outcomes based on input data). The goal of the proposed models is to serve as a tool for geothermal heat supply system optimisation and more accurate performance monitoring.

4.1 K-Nearest Neighbors (KNN) Method

Instance-based learning methods such as nearest neighbour are straightforward approaches to approximating real- or discrete-valued target functions. Instance-based learning methods store the training examples. Generalising beyond these examples is postponed until a new instance (or example) is classified. Each time a new query instance occurs, its relationship to the previously stored examples is analysed to assign a target function value for this new instance. Instance-based methods are sometimes referred to as "lazy" learning methods because they delay processing until a new instance is classified. A key advantage of this kind of lazy learning is that instead of estimating the target function once for the entire instance space, these methods can estimate it locally and differently for each new instance to be classified (Chakrabarti et al. 2008). K-Nearest Neighbors (KNN) is an instance-based learning method.

The abbreviation KNN stands for "K-Nearest Neighbour". It is a supervised machine-learning algorithm, meaning it uses labelled input data in order to learn a function that can generate an appropriate output for new, unlabeled data. The algorithm can be used to solve both classification (discrete values) and regression problems (which had real numbers) statements.

The symbol 'K' denotes the number of nearest neighbours considered when predicting or classifying a new, unknown variable. KNN calculates the distance from all points in the proximity (also referred to as distance, similarity, or closeness) of the unknown data and filters out the ones with the shortest distances to it. Therefore, it can often be called a distance-based algorithm. The straight-line distance (also called the Euclidean distance) is a popular and familiar choice to calculate the distance.

The KNN Algorithm implementation includes the following steps. After data set loading, K should be initialised (which represents the chosen number of neighbours to consider for the algorithm). For each example in the data, the distance between the query and the current example from the dataset should be calculated; also, this distance and the index of the example should be then added to an ordered collection. Once all distances are calculated, this collection should be sorted in ascending order based on the distances. Then, we should pick the first K entries from the sorted collection and get the labels of these entries. If the task is regression, the mean of the K labels must be returned. If the task is classification, return the mode (Machine Learning Basics with the K-Nearest Neighbors Algorithm, Towards Data Science).

In order to determine the optimum value for K for a given dataset, it is necessary to run the algorithm multiple times using different values of K. This value for K should be selected, that minimises the number of errors encountered while also ensuring that the algorithm remains effective at making accurate predictions for new data.

In the paper (Shahdi et al. 2021), for constructing the subsurface temperature prediction map KNN regression method was chosen. In the paper (Khankishiyev, O., Salehi, S., Karami, H., & Mammadzada, V. 2024) KNN method was chosen, and the study showed excellent results and minimal computational power requirements, making it a practical choice for real-world applications.

4.2 Deep Neural Networks (DNN)

Over recent years, deep learning (DL) has been successfully utilised in various applications, namely for complex pattern recognition. It is a subset of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Deep learning models have the capability to automatically extract the necessary features for detecting or classifying data from raw input. This allows us to achieve a high level of accuracy that may surpass human-level performance in multiple tasks like natural language processing, image and speech recognition, and autonomous driving.

An artificial neural network (ANN) is a computing system designed to recognise patterns and solve problems by learning from data. They typically consist of three layers - input, hidden, and output layers, as shown in Figure 4.1. These layers contain nodes or neurons, and each connection can transfer a signal from one neuron to another. The receiving neuron processes the signal and passes it on to other neurons connected to it. The model can learn and generalise complex, multivariate non-linear relationships between the input and output, requiring low computational cost (thanks to the one-time calibration).

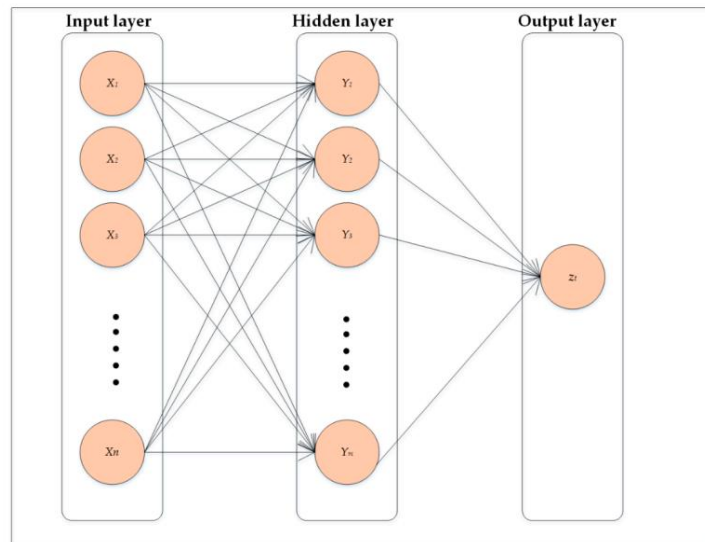


Figure 4.1 – Structure of the ANN model (Taheri et al. 2023)

ANNs are the most widely used AI models used to approximate environmental and hydrological components, particularly soil temperatures (Kisi et al. 2015).

Traditional neural network models have limitations in processing large amounts of data and have limited generalisation ability and scalability. Additionally, their training process is rather slow, and they can easily fall into local optima. To overcome these weaknesses, deep learning (DL) models were developed. DL models utilise a multiple hidden layer structure to process large multi-feature data.

DL models, also known as large-sized neural networks, have powerful computing capabilities compared to traditional neural networks. High-level features are automatically extracted from low-level input data by these models through the use of intermediate layers, allowing complex nonlinear functions to be learned (as illustrated in Figure 4.2). DL models are commonly used in various fields, including search engines and image recognition. Given the highly complex and nonlinear structural characteristics of soil, DL models prove to be more effective in analysing soil particle size and texture compared to traditional models.

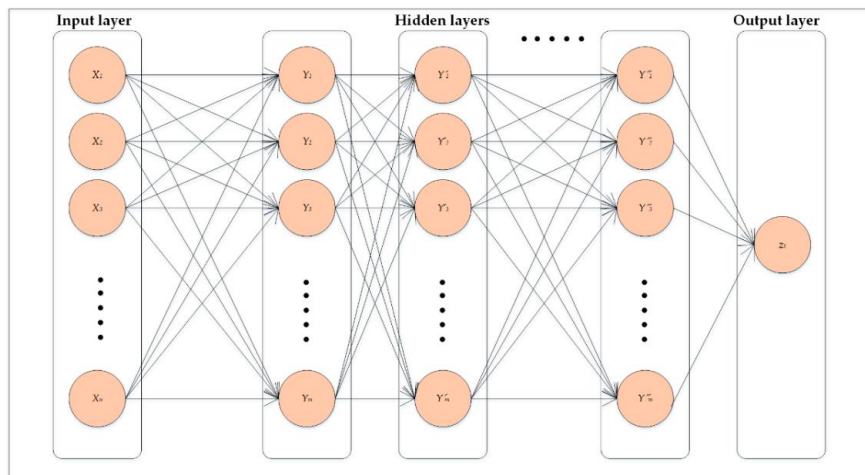


Figure 4.2 – Structure of the Deep Learning Model (Taheri et al. 2023)

Convolutional Neural Networks (CNN), a class of deep NN, are designed for processing data that comes in the form of arrays (like images). A CNN learns to recognise patterns and features through a process called convolution, which filters inputs for useful information as they pass through convolutional layers.

4.2.1 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a specialised type of neural network designed for processing structured grid data, such as images or time series data. CNNs utilise convolutional layers to automatically learn spatial hierarchies of features from the input data.

CNNs consist of convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply a set of learnable filters (kernels) to the input data, extracting features through convolution operations. While convolutional layers can be followed by additional convolutional layers or pooling layers, the fully-connected layer is the final layer. With each layer, the CNN increases in its complexity. Fully connected layers aggregate the features learned from convolutional and pooling layers to produce the final prediction, as shown in Figure 4.3.

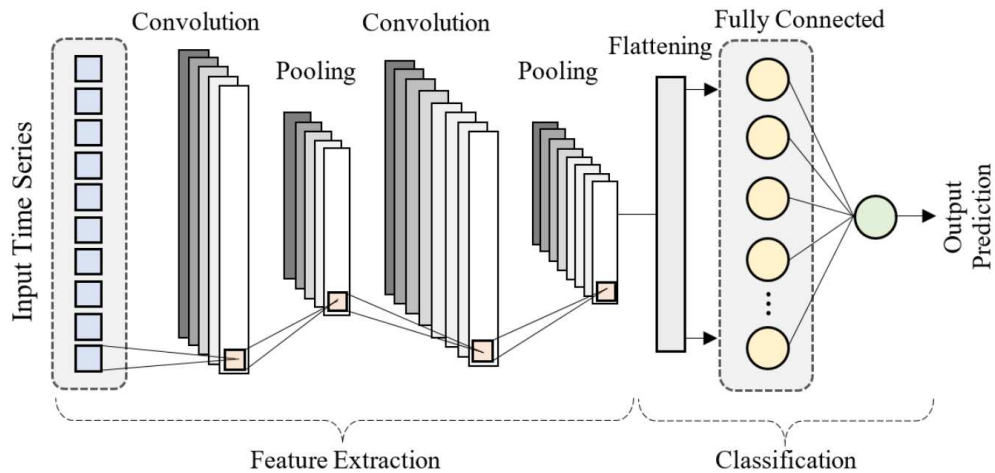


Figure 4.3 – Schematic diagram of a CNN model for time series data forecasting

CNNs are well-suited for tasks involving spatial or temporal dependencies, making them ideal for image processing and time series analysis. They automatically learn hierarchical representations of features, reducing the need for manual feature extraction. CNNs can process spatial and temporal data from a wide variety of sources. Thus, satellite imagery, or geological surveys, as well as temperature records can be utilised to detect anomalies in GWT.

4.3 Random Forest

Random Forest is a learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. It employs multiple decision trees to achieve a more accurate prediction than any individual tree could provide on its own, as shown in Figure 4.4.

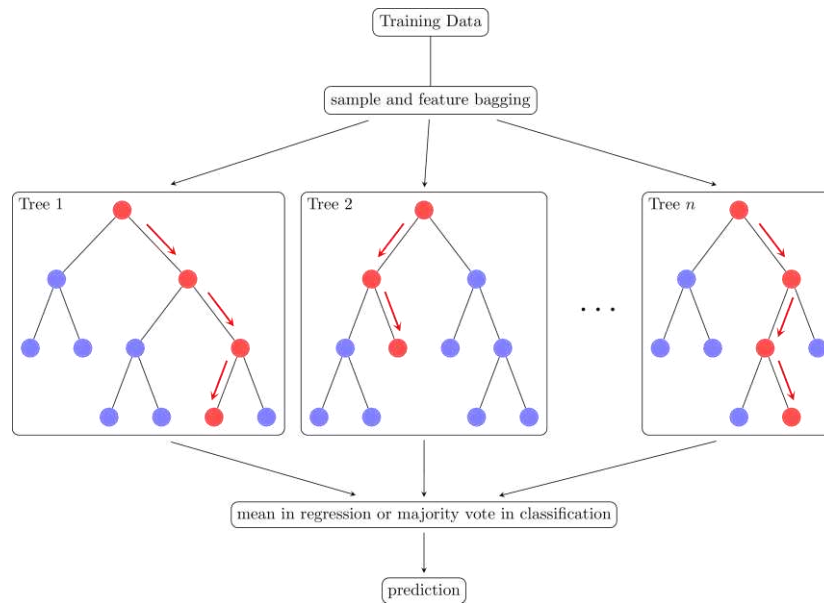


Figure 4.4 – Random Forest model illustration (Random Forest Diagram)

Applying Random Forest to predict GWT involves analysing spatial and temporal data to capture the underlying patterns affecting subsurface temperatures. This model excels in handling the multifaceted nature of environmental data, characterised by complex interactions between seasonal variations.

4.4 Machine Learning Models' Performance

4.4.1 K-Nearest Neighbors Method

The code for K-Nearest Neighbors Model was written in Python. The target variable was groundwater temperature, which is why column “GWT” was excluded from normalisation. Mean Squared Error, Root Mean Squared Error, Mean Absolute Error and R^2 Score were calculated, as shown in Table 4.1.

Table 4.1 – KNN Model performance on the given dataset

Mean Squared Error	0.1430
Root Mean Squared Error	0.3782
Mean Absolute Error	0.2776
R-squared (R2) Score	0.8675

Converting the result into °C, the Mean Squared Error (MSE) of the K-Nearest Neighbors (KNN) model on the test data is approximately 1.0089 °C. This value represents the average

squared difference between the actual groundwater temperature and the predicted groundwater temperature.

The lower the MSE value, the better the model's performance because with a lower MSE, the predicted temperatures are closer to the actual temperatures. In our case, an MSE of approximately 1.0089°C suggests that, on average, the squared difference between the predicted groundwater temperatures and the actual temperatures is quite low, indicating a high level of accuracy in the model's predictions.

RMSE is the square root of the mean squared error and provides a measure of the average magnitude of the error. It is in the same unit as the predicted variable ($^{\circ}\text{C}$), making it more interpretable. An RMSE of approximately 1.0044 indicates that the model's predictions are, on average, about 1.0044 units away from the actual groundwater temperature values. This suggests that the model has a good predictive performance, with relatively small errors in predicting groundwater temperature. Compared to RMSE, Mean Absolute Error is less sensitive to outliers since it doesn't square the errors. The value of 0.7373°C suggests the model is quite accurate, with minimal average error in its predictions. The R^2 score (also referred to as the coefficient of determination) shows the proportion of the variance in the dependent variable that can be predicted from the independent variables. An R^2 score of approximately 0.8675 suggests that about 86.75% of the variability in groundwater temperature can be explained by the model's inputs, which, in this case, are the weather data variables. This is considered a strong score, especially in environmental data prediction, indicating that the model effectively captures the relationship between weather conditions and groundwater temperature.

4.4.2 Convolutional Neural Networks

In the code for the CNN Model, the dataset was split into training and testing sets, with 80% of data used for training and 20% for testing. The first layer is a 1D convolutional layer with 64 filters and a kernel size of 2. This layer automatically learns 64 different filters that capture various features from the input sequence. The second is the output layer with 1 neuron (as it is a regression problem) and does not use an activation function, implying a linear output. The model is trained for 100 epochs with a batch size of 32. An epoch is a complete pass over the entire training dataset, and the batch size is the number of samples processed before the model is updated.

The MSE of this regression model on our test data is approximately 2.006°C , and RMSE had a value of 1.416°C . Here, the realistic difference (MAE) is approximately 1.192°C . Normalised (or scaled) errors are shown in Table 4.2. These values are worse compared to the KNN Model performance. The R-squared score is 0.7265.

Table 4.2 – CNN Model performance on the given dataset

Mean Squared Error (Scaled)	0.2844
Root Mean Squared Error (Scaled)	0.5333
Mean Absolute Error (Scaled)	0.4486
R-squared (R2) Score (Scaled)	0.7365

4.4.3 Random Forest

Similarly, Table 4.3 represents the performance of the Random Forest model. The Mean Squared Error (MSE) is approximately 1.2313 °C, and MAE had a value of 0.8753°C. These values are worse if compared to the performance of the KNN model.

Table 4.3 – Random Forest Model performance on the given dataset

Mean Squared Error (Scaled)	0.1746
Root Mean Squared Error (Scaled)	0.4178
Mean Absolute Error (Scaled)	0.3296
R-squared (R2) Score (Scaled)	0.8383

4.5 Machine Learning Models Results

Upon implementing various machine learning models on the dataset aimed at predicting groundwater temperature (GWT) based on other features, notable differences in performance were observed between Convolutional Neural Networks (CNN), K-Nearest Neighbors (KNN), and Random Forest Regressor, the results are summarised in Table 4.1 below:

Table 4.4 – Machine Learning Models Results

Machine Learning Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R-squared (R2) Score
CNN	0.2844	0.5333	0.4486	0.7365
KNN	0.1430	0.3782	0.2776	0.8675
Random Forest	0.1746	0.4178	0.3296	0.8383

The KNN model exhibited competitive performance, showcasing lower MSE, RMSE, and MAE values compared to the CNN and Random Forest models. Notably, the KNN model

achieved a relatively high R-squared score of 0.8675, indicating its ability to explain approximately 86.8% of the variance in GWT based on the selected features. This suggests that the KNN model may provide a more reliable estimation of GWT compared to the CNN or Random Forest models.

The Fandom Forest Model, while not outperforming KNN in error metrics, demonstrated a competitive R² score (0.8383), suggesting it has a decent capability to explain the variability in the dataset. This could indicate that the random forest model may require further tuning of its parameters or that the dataset characteristics do not favour this model as much as others.

Convolutional Neural Network did not perform as well in this context, recording the highest MSE and RMSE and the lowest R² score. However, its slightly higher error values highlight potential areas for improvement, possibly through deeper network architectures or more advanced feature extraction techniques.

Taking into account that the target variable (GWT) values in the data set varied from 6.2 to 18.6°C, the MAE can also be represented in percentage; thus, models KNN, RF and CNN have the following percentage of error, respectively: 6.04%, 7.17% and 9.77%.

4.6 Feature Importance Analysis

In this study, an analysis of feature importance to identify the factors that significantly influence the prediction of Ground Water Temperature (GWT) was conducted. Feature importance analysis is crucial as it provides insights into the relative contribution of each feature towards the predictive performance of the model. Since KNN does not provide feature importances like tree-based models such as Random Forest, it was decided to analyse the feature importances only using the Random Forest model.

The results obtained from the analysis showed the following feature importance, as also shown in Figure 4.5:

- dew (°C): 0.2777
- Sun light time (HH:MM:SS): 0.2515
- humidity (%): 0.0929
- temp (°C): 0.0895
- windspeed (km/h): 0.0651
- solarradiation (W/m²): 0.0450
- solarenergy (MJ/m²): 0.0429
- GWL (m): 0.0408
- cloudcover (%): 0.0338

- precip (mm): 0.0271
- uvindex: 0.0244
- Snow_Presence: 0.0099

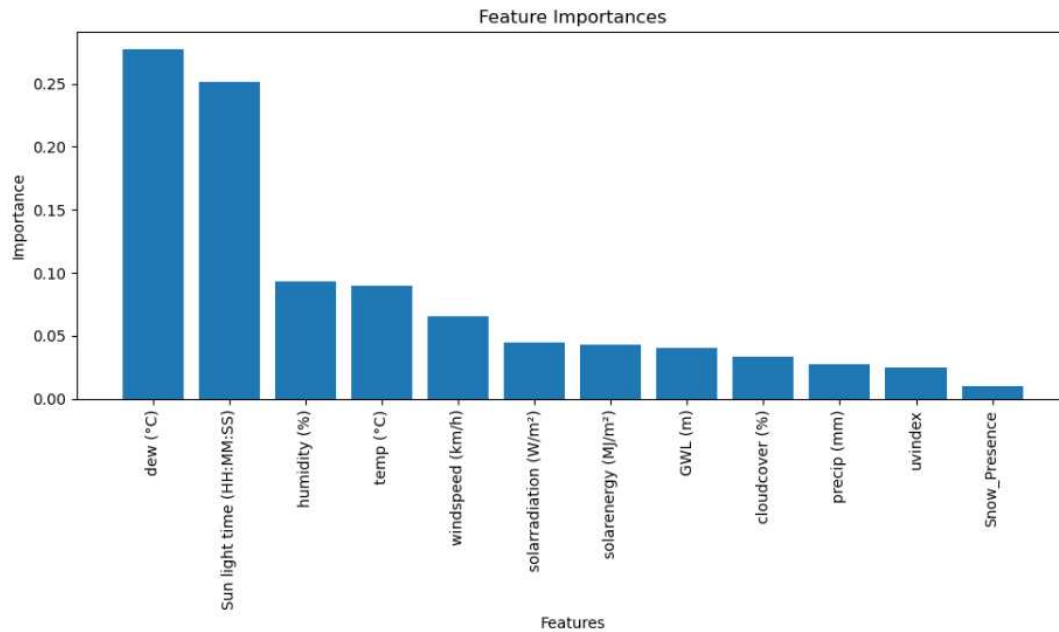


Figure 4.5 – Graphical representation of features and their importance in a Random Forest Regressor model

From the above results, it is evident that dew, followed by sun light time, are the most influential features in predicting GWT. These findings suggest that environmental factors such as dew and sunlight duration have a significant impact on the groundwater temperature, which aligns with our domain knowledge. Additionally, meteorological factors like windspeed, humidity, temperature, and cloud cover also play crucial roles in predicting GWT, albeit to a lesser extent compared to dew and sunlight duration.

Furthermore, the relatively lower importance of features like UV index, snow presence, and precipitation indicates that these variables have less influence on GWT prediction in the studied context.

In conclusion, understanding the relative importance of various features enables better interpretation of model predictions and aids in identifying key drivers of groundwater temperature fluctuations. This knowledge can be valuable for water resource management and decision-making processes, helping develop more effective strategies for groundwater utilisation.

4.7 Data Normalization Step Implementation

In our case, before normalisation, the "Sun light time (HH:MM:SS)" column was transformed into seconds to facilitate uniform scaling alongside other numerical features. This transformation maintains the temporal information while ensuring compatibility with the normalisation process.

During the analysis of the dataset and when applying various machine learning models, significant insights were gained regarding **the necessity and impact of normalisation** as a preprocessing step. Prior to normalisation, the dataset exhibited disparate scales across its features, which posed challenges for effective model training and performance. Subsequently, without normalisation, machine learning models such as Convolutional Neural Networks (CNN), Random Forest Regressor, and K-Nearest Neighbors (KNN) yielded suboptimal results in terms of Mean Squared Error (MSE). The values in °C of the Mean Squared Error before normalisation are shown in Table 4.2 below:

Table 4.5 – Values of Mean Squared Error (MSE) before the normalisation step

Machine Learning Model	Mean Squared Error
CNN	5.4162
KNN	3.4069
Random Forest	2.9287

For instance, without normalisation, the CNN model exhibited a comparatively high Mean Squared Error (MSE) of 5.416, suggesting considerable discrepancies between predicted and actual values. Similarly, the Random Forest Regressor and KNN models also demonstrated elevated MSE values of 2.929 and 3.407, respectively. These findings underscored the challenges posed by disparate feature scales, as the models struggled to effectively learn from and generalise to the dataset.

However, upon implementing normalisation, marked improvements were observed across all models, indicative of the significant impact of this preprocessing step. For the KNN model, as depicted in Table 4.1 above (Chapter 4.4.1 K-Nearest Neighbors Method), normalisation led to a drastic reduction in MSE to 1.009°C, along with lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values, enhancing predictive accuracy and precision. Similarly, the CNN and Random Forest Regressor models showcased substantial MSE reductions, accompanied by notable improvements in MAE, RMSE, and R-squared (R²) scores.

As a result of normalising the data, we observed improvements in model performance and interpretability. Models trained on normalised data typically generalise better to unseen data, resulting in more accurate predictions. These results underscore the critical role of normalisation in mitigating the challenges associated with disparate feature scales. Bringing all features to a normalised scale allowed us to achieve more effective model training, enabling algorithms to distinguish meaningful patterns and relationships within the dataset.

Overall, by applying data normalisation as part of the preprocessing step, it is possible to achieve better performance, stability as well as interpretability of ML models on the given dataset.

To conclude Chapter 4, the thesis examined several machine learning models, including K-Nearest Neighbors (KNN), Deep Neural Networks (DNN), and Random Forest, to find out which model predicts groundwater temperatures most effectively. Each model was evaluated based on its ability to predict GWT. For evaluation, metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) were used. The analysis showed that the KNN model could achieve the best results. After this, the feature analysis was described to find the most influencing features in the dataset. This helps understand the importance of feature handling and features' influence on ML models, providing ideas for further improvement of the models. Also, the importance of the data normalisation step is highlighted.

Chapter 5

Conclusion

5.1 Summary

This thesis focused on enhancing the prediction of groundwater temperatures in shallow geothermal systems in Vienna using machine learning techniques. The research addressed several objectives: identifying the key variables affecting groundwater temperatures, evaluating machine learning models for their prediction accuracy, and understanding the influence of these variables on temperature predictions.

The objectives set at the beginning of the work were achieved, namely:

- Literature review: The theoretical chapter covered different aspects relevant to the topic. After introducing the general terms, shallow geothermal energy mechanisms and environmental benefits were discussed. To analyse the prospects of this system implementation, several cases mentioned in different literature sources were examined. Then, attention was drawn to the area of interest – Vienna. Cases presented in the literature allowed us to conclude on the possibility of utilising the potential of shallow geothermal systems not only as a source of renewable energy but also as a tool for urban climate adaptation. To gain a better understanding of the problem, different climatic factors were studied in terms of their influence on groundwater temperature fluctuation. This step allowed us to minimise the number of features needed for further analysis. Subsequent literary research was aimed at studying the machine learning applications in geothermal and shallow geothermal energy. This allowed conclusions on the most promising algorithms to be analysed.
- Data Preprocessing: The next step of the thesis was focused on data analysis and preparation. Data was interpreted, collected and integrated, and the number of features was reduced to the selected. Initially, the data contained inconsistencies and missing

values; therefore, it was cleaned. Then, the dataset was screened for outliers, and several strategies were applied to handle them. Also, encoding was applied together with normalising. These steps were needed to make the given dataset ready for further machine learning model training.

- Machine Learning Model Development and Evaluation: Different models were theoretically described and practically analysed, including K-Nearest Neighbors (KNN), Deep Neural Networks (DNN), and Random Forest. KNN could provide the most accurate predictions for this specific application.
- Feature Importance Analysis: Additionally, the analysis was run to conclude on the features, that could influence the machine learning models' performance greatly. Dew point and sunlight time were found to be the most influential features in predicting GWT.

5.2 Evaluation

The objectives set at the beginning of this research were successfully met. Firstly, the thesis achieved its primary goal of theoretical selection and practical development of a predictive model for groundwater temperatures using machine learning. Python codes within the Jupyter Notebook environment were also developed to manage data preprocessing and to implement various machine learning models. These scripts represent a reusable resource for further studies, as they can be adapted for future projects. The systematic approach to data cleaning and preprocessing ensured the high quality and reliability of the data used for modelling.

5.3 Future Work

In terms of further development, it should be mentioned that while K-Nearest Neighbors (KNN) demonstrated the highest accuracy in predicting groundwater temperatures, the Convolutional Neural Network (CNN) also showed promising results. Therefore, CNN, being a type of Deep Neural Network (DNN), has the potential to be explored more extensively in future work. Namely, deeper and more complex CNN architectures could be developed. Future research could concentrate on increasing the complexity of CNN models by adding more layers and experimenting with different types of layers and hyperparameters.

During the literature review, it was noted that urban structures, such as buildings, roads, and parks, significantly influence the Surface Urban Heat Island (SUHI) effect, which in turn affects groundwater temperatures. Future studies should focus on refining the models to better account for these urban factors. Incorporating more data on urban infrastructure and land use could improve the models' ability to predict temperature variations influenced by human activities

and urban development. By predicting groundwater temperatures more accurately, the suggested models can help optimise the energy efficiency of buildings and contribute to more sustainable urban development practices.

References

Aguinis, Herman; Gottfredson, Ryan K.; Joo, Harry (2013): Best-practice recommendations for defining, identifying, and handling outliers. In *Organizational Research Methods* 16 (2), pp. 270–301.

Ahmed, Abdelazim Abbas; Assadi, Mohsen; Kalantar, Adib; Sliwa, Tomasz; Sapińska-Śliwa, Aneta (2022): A critical review on the use of shallow geothermal energy systems for heating and cooling purposes. In *Energies* 15 (12), p. 4281.

Barnett, Vic; Lewis, Toby (1994): *Outliers in statistical data*: Wiley New York (1).

Böttcher, Fabian; Zosseder, Kai (2022): Thermal influences on groundwater in urban environments—A multivariate statistical analysis of the subsurface heat island effect in Munich. In *Science of The Total Environment* 810, p. 152193.

Bundesministerium für Landwirtschaft, Regionen und Tourismus. (n.d.). eHYD. Available online at <https://ehyd.gv.at/>.

Cetin, Aysegul; Paksoy, Halime (2013): Shallow geothermal applications in Turkey. In *EGC2013*, Pisa, Italy.

Chakrabarti, Soumen; Neapolitan, Richard E.; Pyle, Dorian; Refaat, Mamdouh; Schneider, Markus; Teorey, Toby J. et al. (2008): *Data mining: know it all*: Morgan Kaufmann.

Chettri, Nimesh; Sankarananth, S. (2022): *Geothermal Energy: Definition and Its Applications*. Technoarete Transactions on Renewable Energy, Green Energy and Sustainability.

Data Science Horizons. *Data Cleaning and Preprocessing for Data Science Beginners*. (2023). Available online at <https://www.datasciencehorizons.com>.

Energy Innovation Austria (2021): *Heat from the depths: Geothermal energy as an energy technology of the future in Austria*. Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology. Available online at <https://www.energy-innovation-austria.at>.

European Geothermal Energy Council. *Market report 2013/2014* (2013).

Gudmundsdottir, H. and Horne, R. (2020): Prediction modeling for geothermal reservoirs using deep learning.

Hare, Danielle K.; Benz, Susanne A.; Kurylyk, Barret L.; Johnson, Zachary C.; Terry, Neil C.; Helton, Ashley M. (2023): Paired Air and Stream Temperature Analysis (PASTA) to evaluate groundwater influence on streams. In *Water Resources Research* 59 (4), e2022WR033912.

Haslinger, Edith; Turewicz, Veronika; Hammer, Andreas; Götzl, Gregor (2022): Assessment of Deep and Shallow Geothermal Resources and Measurement of Waste Heat Potentials from Industrial Processes for Supplying Renewable Heat for Industry and Urban Quarters. In *Processes* 10 (6), p. 1125.

Hawkins, Douglas M. (1980): *Identification of outliers*: Springer.

Hodge, Victoria; Austin, Jim (2004): A survey of outlier detection methodologies. In *Artificial intelligence review* 22, pp. 85–126.

IRENA and IGA (2023): *Global geothermal market and technology assessment*, International Renewable Energy Agency, Abu Dhabi; International Geothermal Association, The Hague.

Johnson, Richard Arnold; Wichern, Dean W. (2002): *Applied multivariate statistical analysis*.

Khankishiyev, O., Salehi, S., Karami, H., & Mammadzada, V. (Ed.) (2024): *Identification of Undesirable Events in Geothermal Fluid/Steam Production using Machine Learning*. Proceedings of the 49th Workshop on Geothermal Reservoir Engineering Stanford University, Stanford, CA, USA.

Kisi, Ozgur; Tombul, Mustafa; Kermani, Mohammad Zounemat (2015): Modeling soil temperatures at different depths by using three different neural computing techniques. In *Theoretical and applied climatology* 121, pp. 377–387.

Machine Learning Basics with the K-Nearest Neighbors Algorithm. (n.d.). Towards Data Science. Available online at <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>.

Manzella, A. (2017): *Geothermal energy*.

Manzella, A. (2019): General introduction to geothermal energy. In *Geothermal Energy and Society*, pp. 1–18.

Nassreddine, Ghalia; Younis, Joumana; Falahi, Thaer (2023): Detecting Data Outliers with Machine Learning. In *Al-Salam Journal for Engineering and Technology* 2 (2), pp. 152–164.

Ninikas, Konstantinos; Hytiris, Nicholas; Emmanuel, Rohinton; Aaen, Bjorn (2017): Heat energy from a shallow geothermal system in Glasgow, UK: Performance evaluation design. In *Environmental Geotechnics* 7 (4), pp. 274–281.

Oregon USA Oregon Institute of Technology.: *Oregon Institute of Technology Geothermal Uses and Projects - Past, Present, Future*. Available online at https://chptap.ornl.gov/profile/174/OIT_ORC-Project_Profile.pdf.

Ostermann, Viktoria (2011): GEO-Pot: Geothermal Energy Potential in Austria. In : Geotechnical Engineering: New Horizons: IOS Press, pp. 333–338.

Österreicher, Doris; Sattler, Stefan (2018): Maintaining comfortable summertime indoor temperatures by means of passive design measures to mitigate the urban heat island effect—A sensitivity analysis for residential buildings in the City of Vienna. In *Urban Science* 2 (3), p. 66.

Random Forest Diagram. Available online at <https://tikz.net/random-forest/>.

Rushlow, Caitlin R.; Sawyer, Audrey H.; Voss, Clifford I.; Godsey, Sarah E. (2020): The influence of snow cover, air temperature, and groundwater flow on the active-layer thermal regime of Arctic hillslopes drained by water tracks. In *Hydrogeology Journal* 28 (6), pp. 2057–2069.

Sanner, Burkhard (2003): Shallow geothermal energy. In *GHC bulletin In Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ)*. International Geothermal Association. Available online at https://www.geothermal-energy.org/pdf/IGAstandard/ISS/2003Germany/II/4_1_san.pdf.

Secretariat, United Nations Framework Convention on Climate Change. (2015): Report of the Conference of the Parties on Its Twenty-First; United Nations Framework Convention on Climate Change. Bonn, Germany.

Shahdi, Arya; Lee, Seho; Karpatne, Anuj; Nojabaei, Bahareh (2021): Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. In *Geothermal Energy* 9, pp. 1–22.

Shao, Qi; Zhang, Chengwei; Sun, Mingyuan; Chen, Tao; Yu, Xueyou; Yin, Huiyong; Wang, Bingfeng (2023): Research and Application of Shallow Geothermal Energy Development and Utilization in Shandong Province. In *ACS omega* 8 (26), pp. 23545–23553.

Shelare, S.; Kumar, R.; Gajbhiye, T.; Kanchan, S. (2023): Role of Geothermal Energy in Sustainable Water Desalination—A Review on Current Status, Parameters, and Challenges. *Energies* 2023, 16, 2901. <https://doi.org/10.3390/en16062901> Academic Editor: Mohammadali Ahmadi Received 10.

Smola, Alex; Vishwanathan, S. V.N. (2008): Introduction to machine learning. In Cambridge University, UK 32 (34), p. 2008.

Taheri, Mercedeh; Schreiner, Helene Katherine; Mohammadian, Abdolmajid; Shirkhani, Hamidreza; Payeur, Pierre; Imanian, Hanifeh; Cobo, Juan Hiedra (2023): A review of machine learning approaches to soil temperature estimation. In *Sustainability* 15 (9), p. 7677.

Tissen, Carolin; Menberg, Kathrin; Benz, Susanne A.; Bayer, Peter; Steiner, Cornelia; Götzl, Gregor; Blum, Philipp (2021): Identifying key locations for shallow geothermal use in Vienna. In *Renewable Energy* 167, pp. 1–19.

Vesselinov, V.; Mudunuru, M.; Ahmmed, B.; Karra, S.; O'Malley, D. (2022): Machine Learning to Discover, Characterize, and Produce Geothermal Energy. In : *Machine Learning Applications in Subsurface Energy Resource Management*: CRC Press, pp. 45–70.

Visual Crossing Corporation. Weather Data Services: Visual Crossing Corporation. Available online at <https://www.visualcrossing.com>.

Vlahović, M., Stević, Z (Ed.) (2020): Utilizing renewable resources—converting geothermal energy to electricity. *Proceedings-8th International Conference on Renewable Electrical Power Sources/Zbornik radova-8. Međunarodna konferencija o obnovljivim izvorima električne energije*: Beograd: Savez mašinskih i elektrotehničkih inženjera i tehničara Srbije-SMEITS.

Wan, Xiao-fan; Zhang, Hao; Shen, Chuan-bo (2022): Visualization analysis on the current status and development trend of geothermal research: Insights into the database of Web of science. In *Frontiers in Energy Research* 10, p. 853439.

Wang, Long; Yu, Ziwang; Zhang, Yanjun; Yao, Peiyi (2023): Review of machine learning methods applied to enhanced geothermal systems. In *Environmental Earth Sciences* 82 (3), p. 69.

Xue Z., Chen Z. (Ed.) (2023): Deep learning based production prediction for an enhanced geothermal system (EGS). *SPE Canadian Energy Technology Conference*: SPE.

List of Figures

Figure 1.1 – Workflow for Underground Water Temperature Prediction Using Machine Learning.....	11
Figure 2.1 – Schematic flow diagram of GSHP systems for heating.....	17
Figure 3.1 – Forms of data processing (Chakrabarti et al. 2008).....	28
Figure 3.2 – Identification of missing values command (Data Science Horizons, 2023).....	29
Figure 3.3 – Command allowing to convert to numeric format (Data Science Horizons, 2023)	30
Figure 3.4 – Ground Water Level (GWL) Measurements	31
Figure 3.5 – Ground Water Temperature (GWT) Measurements.....	31
Figure 3.6 – Intermediate data set after the initial preprocessing step.....	32
Figure 3.7 – Visual outliers identification via histogram plotting	36
Figure 3.8 – Visual outliers identification via multiple construct technique	39
Figure 3.9 – Outliers defined for column representing GWL.....	41
Figure 3.10 – Outliers defined for column representing precipitations	41
Figure 3.11 – Outliers defined for column representing the amount of snowfall	42
Figure 3.12 – Outliers defined for column representing the snow depth.....	42
Figure 3.13 – Outliers defined for column representing wind speed.....	42
Figure 3.14 – IQR technique (Nassreddine et al. 2023).....	43
Figure 3.15 – Code in Python for interquartile range (Nassreddine et al. 2023)	43
Figure 3.16 – Outliers defined for column representing precipitations	44
Figure 3.17 – Boxplots before (on the left) and after (on the right) outlier handling	44
Figure 3.18 – Outliers defined for column representing the amount of snowfall	45
Figure 3.19 – Outliers defined for column representing the snow depth.....	45
Figure 3.20 – Outliers defined for column representing wind speed.....	45
Figure 3.21 – Boxplots before (on the left) and after (on the right) outlier handling	46
Figure 3.22 – Result after one-hot encoding of the “snowdepth” column into “Snow_Presence”	47
Figure 4.1 – Structure of the ANN model (Taheri et al. 2023).....	51
Figure 4.2 – Structure of the Deep Learning Model (Taheri et al. 2023)	52
Figure 4.3 – Schematic diagram of a CNN model for time series data forecasting.....	53
Figure 4.4 – Random Forest model illustration (Random Forest Diagram)	54
Figure 4.5 – Graphical representation of features and their importance in a Random Forest Regressor model	58

List of Tables

Table 4.1 – KNN Model performance on the given dataset	54
Table 4.2 – CNN Model performance on the given dataset.....	56
Table 4.3 – Random Forest Model performance on the given dataset.....	56
Table 4.4 – Machine Learning Models Results	56
Table 4.5 – Values of Mean Squared Error (MSE) before the normalisation step	59

Abbreviations

ML	Machine Learning
KNN	K - Nearest Neighbors
DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
GWL	Ground Water Level
GWT	Ground Water Temperature
SGE	Shallow Geothermal Energy
GSHP	Ground-Source Heat Pumps
BHE	Borehole Heat Exchanger
DH	District Heating
GHP	Geothermal Heat Pump
SSUHI	Subsurface Urban Heat Islands
PCA	Principal Component Analysis
SME	Subject-Matter Expert
NMFk	Non-negative Matrix Factorization
EGS	Enhanced Geothermal Systems
RNN	Recurrent Neural Network
SVM	Support Vector Machines
IQR	Interquartile Range
UV	Ultraviolet
SUHI	Surface Urban Heat Island
MAE	Mean Absolute Error

MSE

Mean Squared Error

RMSE

Root Mean Squared Error