Norwegian University
of Life Sciences

Master's Thesis 2023 30 ECTS
Faculty of Science and Technology

Wavelet Transforms for Rain and Snow Classification with Commercial Microwave Links: Evaluation Using Real-World Data

Rasmus Falkeid Hagland
Water and Environmental Technology

Men også regnet liker jeg.
Det væter byen
og klistrer den som et frimerke på landskapet.
Om natten kommer drømmen, den fattige postmannen,
og sender konvolutten til Rio i karnevaldstida.
Regnet trommer på taket.
Gamle uteseilere danser samba i søvne.

From *Gledespunkter*, 1985
Kolbein Falkeid (1933 - 2021)

Preface

This thesis marks the end of my 5-year integrated master's degree at the Norwegian University of Life Sciences (NMBU). When deciding what to write about in my thesis, working with hydrological data and learning more about data science were decisive factors in choosing the topic. Working on this project was challenging at times, and the learning curve was steep. Still, I've managed to see it through and acquired new knowledge outside my field of study.

The original plan was to work with a smaller data set of telecommunication links from Ås and record snow and rain events myself. Therefore, four wireless pluviometers and two climate stations were placed near the microwave links to use as reference data. Eventually, I was provided with a much larger data set of the telecommunication network from Norway, and the study was extended beyond Ås. As such, the wireless weather stations were not used in this study, but I will still like to thank *Marko Jakovljevic* for assisting me with all the electronics and setting up the LoraWAN network. The location of the pluviometers can be seen in Figure A.1 found in Appendix A and will hopefully be useful to someone else.

Although this is my master's thesis, this project could not have been completed without the help of my supervisors. I would express great gratitude to *Vegard Nilsen*, *Nils-Otto Kitterød*, and *Mareile Astrid Wolff* for providing me with good advice and feedback during this semester. I was also encouraged to go to the EGU 2023 conference in Vienna, where I was given the opportunity to present some of my findings from this project. I will also thank my dear friends *Thorstein* for the help with the proofreading and *Eivind* for assisting me in setting out the pluviometers. Additionally, I would like to express extra gratitude to my main supervisor *Erlend Øydvin*. Without you, this project would not be possible. It was always enjoyable working with you throughout this process. Thank you for helping me with the data and programming, for meaningful discussions, for letting me travel to Vienna with you, and for always challenging me to investigate and explore more into the topics.

Looking back at my five years here at NMBU, I can not help but to think about all the friends I made and the fun memories from Studentsamfunnet in Ås, the study trips, and the everyday campus life with my fellow students. I would also like to thank my family for always supporting me. Lastly, I will thank *Anna* for always being there for me and distracting me away from the writing process whenever needed. It is hard to even imagine my time at NMBU without you!

Ås, May, 2023

Rasmus Falkeid Hagland

Abstract

The need for improved precipitation estimations has prompted the exploration of opportunistic alternatives such as utilizing commercial microwave links (CML), particularly in areas with poor coverage of weather radars and rain gauges. It has been known that rainfall-induced attenuation in the microwave signal can be used to determine rainfall intensity accurately. However, detecting other types of precipitation, such as dry snow, remains a challenge. This study evaluates the feasibility of using wavelet transform combined with a random forest classifier to identify rain and snow events.

Real-world signal attenuation data from telecommunication operators and precipitation data from nearby disdrometers in Norway were used to develop the classification methods proposed in this study. The rain classifier was based on data from June 2022, while the snow classifier was evaluated using data from December 2021. The operating frequency of the CMLs used in this study was between 30-40 GHz. The algorithm for rain detection performed similarly to other wet-dry classification methods, with a mean Matthews correlation coefficient (MCC) of 36 % among 52 CMLs. The snow detection algorithm, however, showed no correlation between signal attenuation from 41 CMLs and dry snowfall.

In conclusion, the wavelet transforms effectively extract useful information from signal attenuation for rain classification but are unsuitable for detecting snow. Moreover, the study recommends testing commercial microwave links with higher operating frequencies than those used in this study, combined with temperature data, to improve the possibilities of dry snow detection.

Sammendrag

Behovet for forbedret nedbørestimering har ført til utforskning av opportunistiske alternativer som bruk av kommersielle radiolinjer (CMLer), spesielt i områder med dårlig dekning av værradarer og regnmålere. Det har vært kjent at nedbørrelatert demping i radiolinjene kan brukes til å bestemme nedbørintensiteten nøyaktig. Å oppdage andre typer nedbør, som tørr snø, er imidlertid fortsatt en utfordring. Denne studien ser på muligheten for å bruke wavelet-transformasjon kombinert med en random forest klassifiseringsalgoritme for å identifisere regn- og snøhendelser.

Måledata fra signaldemping fra telekommunikasjonsoperatører og nedbørsdata fra nærliggende disdrometre i Norge ble brukt til å utvikle klassifiseringsmetodene som ble foreslått i denne studien. Metodene for å klassifisere regn var basert på data samlet inn i juni 2022, mens metoden for å klassifisere snø ble utviklet ved hjelp av data fra desember 2021. Driftsfrekvensen til CML-ene som ble brukt i denne studien var mellom 30-40 GHz. Algoritmen for regndeteksjon viste tilsvarende resultater som andre regnklassifiseringsmetoder, med en gjennomsnittlig Matthews korrelasjonskoeffisient (MCC) på 36 % blant 52 CMLer. Algoritmen for å oppdag snø viste imidlertid ingen sammenheng mellom signaldempning fra 41 CMLer og snøhendelser.

Det kan konkluderes med at wavelet-transformasjon effektivt kan ekstrahere ut nyttig informasjon fra signaldempning for som kan brukes til regnklassifisering, men viserer seg å være uegnet for å oppdage snø. I tillegg anbefaler studien å teste radiolinjer med høyere driftsfrekvenser enn de som ble brukt i denne studien, kombinert med temperaturdata, for å forbedre mulighetene for å oppdage snøhendelser.

Table of Contents

List of Figures

List of Tables

List of Acronyms

API	Application programming interface
CML	Commercial microwave link
CNN	Convolutional neural network
CV	Cross-validation
CWT	Continues wavelet transform
DSD	Drop size distribution
DWT	Discreet wavelet transform
FN	False negative
FP	False positive
ITU	International Telecommunication Union
MCC	Matthews correlation coefficient
NVE	Norwegian Water Resources and Energy Directorate
RFC	Random Forest Classifier
RSD	Rolling standard deviation
STFT	Short-time Fourier transform
TN	True negative
TP	True positive
TRSL	Transmitted-received signal levels
WAA	Wet antenna attenuation

1. Introduction and background

In this chapter, the scope and goal of this study, as well as the motivation, are described. Then an introduction to the principles of how precipitation estimates with the use of signal attenuation in commercial microwave links are possible, followed by a summary of existing methods for wet-dry classification and a brief overview of previous research done on the detection of other precipitation types, such as snow, sleet, and hail. Then a short introduction to the mathematics behind wavelet transform is given, and lastly, the theory needed to understand how the random forest classifier can be used to detect precipitation.

1.1 Motivation

The global water cycle is a complex process that tracks water's movement, distribution, and storage. Water can be stored in many forms and can be found underground, in the oceans, at the surface, and in the atmosphere (Dingman, 2015). Accurate and precise precipitation measurements are therefore crucial as precipitation is the driving factor for understanding the water cycle. Traditionally precipitation is measured with point observation and weather radars. Although this gives a good estimation and precise single-point observations, it is limited by the rain gauge networks and radar's spatial distribution. Many sparsely populated regions are therefore affected by sampling errors (Sun et al., 2018). According to Kidd et al. (2017), if we let each rain gauge in the world represent the actual rainfall within an area of a 5.0 km radius, just 1 % of the Earth's surface area is covered.

According to estimates from Hirabayashi et al. (2008) and Singh (2001), snowfall only represents about 5-10 % of the world's precipitations. However, in mountainous regions and regions near the poles, snowfall is an essential contributor to the transportation of latent energy and moisture. In mainland Norway, about 40 % of the runoff comes from snow melting (Haddeland and Holmqvist, 2015), and snowfall makes up about one-third of the total precipitation (NVE, 2019). Furthermore, about one-sixth of the world's population relies on either snow melting or glaciers as their primary source of drinking water (Barnett et al., 2005).

Retrieving accurate snow measurements and estimating snowfall across larger areas is difficult due to several factors (Rasmussen et al., 2012). Firstly, snow is highly influenced by turbulence near the surface, making it challenging to measure, which leads to undercatch in windy conditions (Wolff et al., 2015). Secondly, the uneven distribution of snow, influenced by topography, further complicates the measurement process (Rasmussen et al., 2012). Therefore, correct snowfall measurements constitute some of the main challenges within the field of hydrology (Witze, 2016). Better snow estimates are vital to improve weather forecasting, predict river floods during spring, estimate the magazine level for hydropower plants, and gather more information and knowledge about how the climate changes (Kochendorfer et al., 2022).

During the last two decades, new and opportunistic methods of gathering precipitation measurements have been explored. This includes the use of satellite images from space (Kidd, 2001; Bellerby et al., 2000), using sensors in windshield wipers in cars (Rabiei et al., 2013), and image detection from on-board car cameras (Nashashibi et al., 2010). Another way to measure precipitation is to exploit signal attenuation in commercial microwave links (CMLs) intended for telecommunication. For rainfall measurements, this is already well proven by Messer et al. (2006) and Leijnse et al. (2007). The relation between the specific attenuation k [dB/km] in the microwave signal and the rainfall intensity R [mm/h] has shown to have a better correlation than for weather radars relativity Z [-] and R [mm/h] (Chwala and Kunstmann, 2019). The process of retrieving good precipitation estimation from CMLs consists of several steps. Some areas within this field still need to be investigated more as CMLs have almost exclusively been applied for rainfall measurements in temperate climate zone (Chwala and Kunstmann, 2019). New methods for detecting snowfall are always welcome in countries at higher latitudes, such as Norway, where precipitation frequently falls as solid particles.

This thesis aims to investigate whether discrete wavelet transformations, combined with machine learning, can be used to detect rain and snow in the signal attenuation levels of CMLs using real-world data. The goal is to provide a new and opportunistic way of measuring snow and explore new ways to improve the process of detecting precipitation in microwave signal attenuation. To answer this, two research questions have been asked:

- **Research question 1**: *Can wavelet transform and machine learning be used to detect rain with attenuation data from commercial microwave links?*

- **Research question 2**: *Is it possible to distinguish between periods with and without snowfall using the same methodology?*

1.2 Measuring precipitation with microwave links

Precipitation weakens the signals in wireless telecommunication networks. This can be used to estimate a path-average rainfall between two antenna towers, thus allowing rain measurement over larger areas with relatively good resolution in both time and space. (Chwala and Kunstmann, 2019). Figure 1.1 illustrates how the hydrometeors attenuated the microwave signal. CML networks consist of several antenna towers that transmit and receive microwave signals from antenna to antenna to form a network that covers an entire country. They are mainly used by telecommunication operators as part of the mobile telecommunication network called the backhaul network and are among the most critical forms of infrastructure in modern society. CML networks typically use frequencies between 5 GHz and 40 GHz, but with the 5G network under construction, frequencies up to 80 GHz would be used.CMLs with longer distances between the antenna towers tend to have lower operating frequencies, while the operating frequencies are higher where the distance is shorter. (Chwala and Kunstmann, 2019).

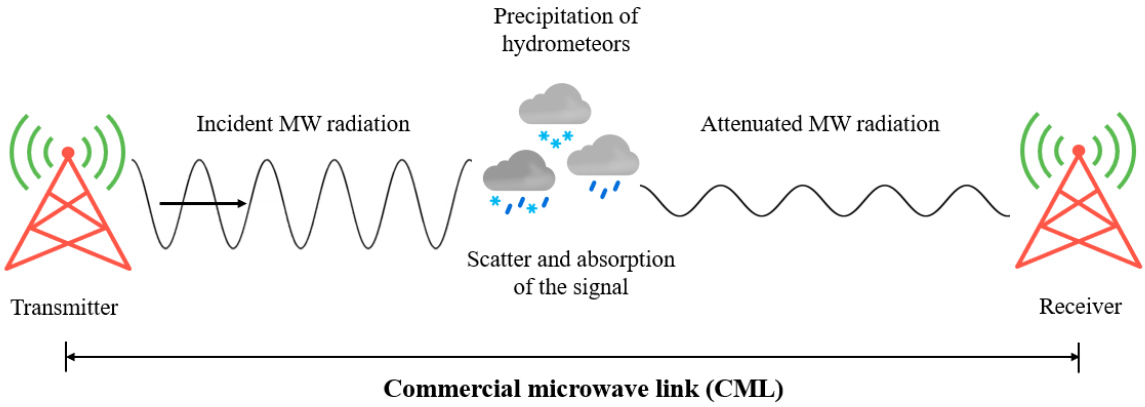

Figure 1.1: A illustration of a CML, showing two antenna towers transmitting and receiving the microwave (MW) signal, different types of precipitation that attenuate the signal. This figure was created with the inspiration of a similar figure from Chwala and Kunstmann (2019).

The fact that precipitation affects telecommunications signals has been known for a long time. Almost 80 years ago, Robertson and King (1946) showed that rainfall significantly impacted the attenuation of microwave signals. This is particularly true for signals with frequencies between 10 GHz and 50 GHz. This corresponds to wavelengths of around 10 mm and is of the same order of magnitude as the raindrops. The attenuation of the microwave signals is caused by the principle that precipitation has a different dielectric constant than the atmospheric gasses in the surroundings. Therefore, the signal is scattered (Leijnse et al., 2007) and induces attenuation. Since hydrometeors, such as rain and snow, often are irregularly shaped, the geometric diameter of the object is

defined as the equivalent diameter. When a raindrop has an equivalent diameter smaller than 1 mm, the shape is assumed to be spherical (Pu et al., 2020). However, when the equal diameter is greater than 6 mm, the droplet tends to have a more flattened shape at the bottom, making it asymmetrical across the horizontal axis (Magono, 1954); this gives the raindrops dielectric polarization that scatters the microwave signal.

The relationship between rainfall intensity R [mm/h] and the specific signal attenuation k [dB/km] due to rain can be described through a simple power function as Equation 1.1. Here, a [-] and b [-] are two constants in the equation (Equation 1.1) that will vary with the frequency and polarization of the signal, as well as the air temperature and drop size distribution (DSD) in a rainfall event. The International Telecommunication Union, (ITU, 2003), has done research on microwave signal attenuation for decades and developed a series of recommended values for the two constants a and b. The value of a will increase with frequency because high-frequency signals are more affected by rain. The constant b is shown to vary between 0.86 and 1.38, but for frequencies around 20 GHz to 40 GHz, it is observed to be close to 1. This simplifies Equation 1.2 to a linear relationship between signal attenuation and rain rate. The accumulated signal attenuation A [dB] can be expressed as an integral as seen in Equation 1.1.

$$k = aR^b \tag{1.1}$$

$$A_{rain} = \int_0^L k(l)dl = \int_0^L aR(l)^{b \approx 1} dl = a\bar{R}L \tag{1.2}$$

1.2.1 Other sources of attenuation

The CML signal is affected differently by melting hydrometeors, such as snow, graupel, and hail, than by liquid precipitation. This is mainly due to the different physical and dielectric properties the precipitation got under formation in the atmosphere (Alonso et al., 2021). The fact that the different precipitation types affect the CML signals differently indicates that it is possible to distinguish between different precipitation types. (Pu et al., 2020). Falling precipitation is not the only source that causes fluctuations in the transmitted signal between antenna towers. Accumulation of tiny water drops sticking to the transmitting and receiving antenna called *wet antenna attenuation* (WAA) also adds an amount of attenuation, which must be adjusted before the rain rate can be derived (Chwala and Kunstmann, 2019). Baseline attenuation refers to signal loss and fluctuations occurring in wireless communication systems due to certain atmospheric conditions and gases in the air. These conditions can cause attenuation during dry and rainy events, as explained in studies by Chwala and Kunstmann (2019) and Lian et al. (2022). The baseline attenuation is often calculated using the attenuation before and after rainfall. Although the baseline attenuation is set to a constant level, Ostrometzky

and Messer (2018) have developed a more advanced and dynamic method for estimating the baseline attenuation. In a commercial microwave link, the total attenuation level can be expressed as in Equation 1.3 and is the sum of the attenuation from the precipitation, the wet antenna effect, and the baseline.

$$A_{total} = A_{precipiatation} + A_{WAA} + A_{baseline} \tag{1.3}$$

1.3 The workflow for estimating rainfall

The workflow for using CMLs on larger scales and generating rainfall fields can be divided into five steps. Step one is to acquire the CML data. This includes getting the transmitted signal levels (TSL), received signal levels (RSL), polarization, location, and frequency. Then data quality control and some prepossessing are required to clean the data. Step two is to identify wet and dry events during the period. The process is further explained in Section 1.4. Step three is to determine the baseline during the wet events and do a correction for the WAA effect. After the baseline is determined, the four is to derive the rain rate for a single link by using the equation for the k-R relation (Equation 1.1). The fifth step is to repeat steps two to four on multiple CMLs covering a larger area and then generate rainfall fields using spatial interpolation techniques. (Chwala and Kunstmann, 2019). Figure 1.2 illustrates the workflow.

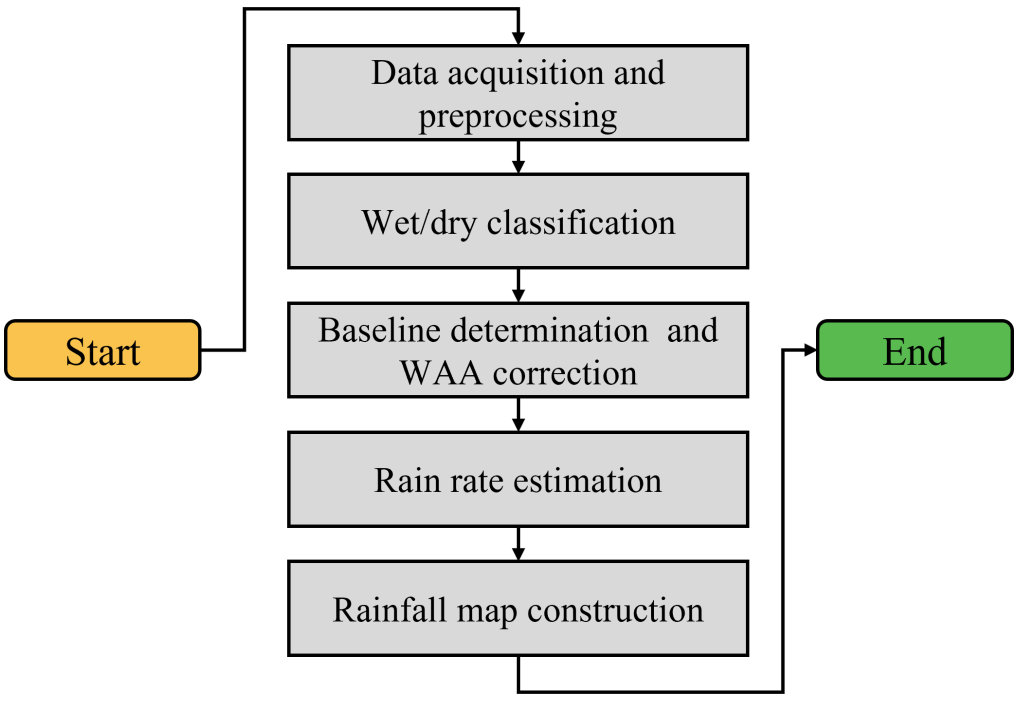

Figure 1.2: Schematic illustration of a typical workflow for rainfall estimated from commercial microwave links (CMLs). Inspired by Chwala and Kunstmann (2019).

1.4 Existing wet-dry classification methods

Rain

Numerous approaches have been suggested for detecting rain events in the CML attenuation. The research field has two primary methodologies depending on the available type of data (Polz et al., 2020). One way to retrieve wet and dry events in CML networks is to compare links close to each other and classify a period as wet if all the links experience signals attenuation in the same period. This method is explained in more detail in work by Overeem et al., 2016. The second methodology, which is also used in this study, focuses on attenuating from individual links. The most widely used and simplest single-link method is the rolling standard deviation (RSD) method developed by Schleiss and Berne (2010), which applies a sliding window over the signal loss time series. A period is then wet if the standard deviation is above a given threshold value, typically 0.8. Other single-link methods include the short-time Fourier transform (STFT) of the signal attenuation levels time series, which classify wet and dry events in the frequency domain by Chwala et al. (2012), a convolutional neural network (CNN) method from Polz et al. (2020) and a method using a Markov switching the based model from Wang et al. (2012).

Common to all methods used for classifying wet events is that they come with a trade-off. Either being more restrictive and only detecting the more significant rainfall events or having a greater likelihood of classifying the fluctuations in dry periods as wet (Chwala and Kunstmann, 2019). Although several methods have been developed for detecting wet events, most studies have only been used with CML data up to 50 links (Polz et al., 2020). Furthermore, there are differences in how the signal attenuation data is sampled. Some telecommunication operators measure the lowest and highest attenuation levels in 15 minutes period of time, and other measures the attenuation level every 10 seconds. Therefore, variations in sampling techniques can make comparing methods challenging. Method comparison will hopefully become more accessible in the future as Andersson et al. (2022) have published some open access CML data for benchmark testing.

Snow

There have been relatively fewer suggested methods for detecting other types of precipitation. One way of detecting graupel and wet snow was proposed by Pu et al. (2020). This method utilizes a network of multi-frequency microwave links where the signal attenuation has been synthetically generated based on signal scattering physics. The drop size distribution data (DSD) from disdrometers was used as ground truth to then classify the precipitation types. The study concluded that the accuracy increased proportionally

with the operating frequency of the CMLs, significantly above 50 GHz, and that dry snow was difficult to detect. Another method proposed by Ostrometzky et al. (2015) made a classification algorithm based on a kernel Fisher discriminant analysis (KDA) combined with a decision tree. The results showed that the algorithm could distinguish between rain and sleet (wet snow). Due to the relatively low temporal resolution of the measured signal attenuation of 15 minutes, an entire period of dry snow was never measured. Therefore it was not clear if the snow would be possible to detect.

1.5 The wavelet transform

Wavelet transform is an advanced form of signal analysis in the field of mathematics that deals with signal processing and multi-frequency analysis (non-stationary signals). The key concept of the wavelet transform is that it decomposes a non-stationary signal into different components in both time and frequency. The term *wavelet transform* refers to the transformation of an integrable signal *f(t)* to the domains of both time and frequency, meaning that we can obtain the localization of frequencies. The idea of a wavelet transformation started in the early 1900s (Vistnes, 2016), but was not fully used until the 1980s and was mainly utilized in the field of seismology to analyze seismic waves inside the earth. Today wavelet transform is used in multiple fields of science, including image processing, biomedical engineering, and robotic control systems.

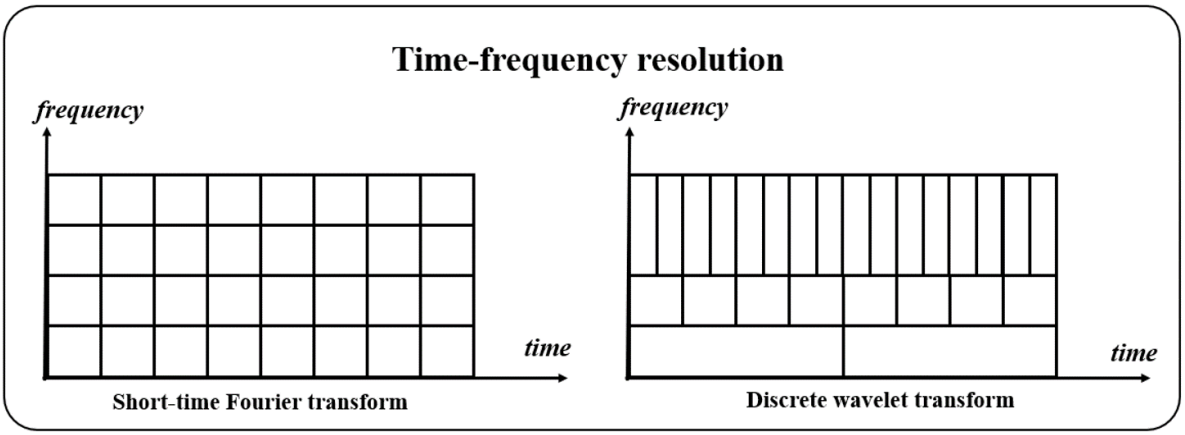

Figure 1.3: Illustration of the time-frequency resolution for the STFT versus the wavelet transform. It can be seen how the wavelet time-frequency resolution varies at different frequencies, while the STFT has a constant time-frequency resolution. Recreation of a figure by Yoo et al. (2022).

A fundamental principle of signal processing is the *uncertainty principle,* which states that there is an inevitable trade-off between the precision of time localization and the precision of frequency localization when analyzing signals (Debnath, 2001). Compared to short-time Fourier Transform (STFT), where a standard Fourier transform is applied inside a sliding window with fixed width in time. The wavelet transform will have

different time-frequency resolutions (Figure 1.3) that change for the different frequency ranges. The time-frequency relationship in wavelets is designed such that the *time resolution* is good where the frequency is high, and the *frequency resolution* is good where the frequency is low. This is due to the unique property of wavelets that allows them to adapt to the varying frequency content of a signal over time. (Debnath and Shah, 2015). How wavelet transform handles the uncertainty principle between time resolution and frequency resolution makes it a flexible tool that is computationally efficient when analyzing non-stationary signals.

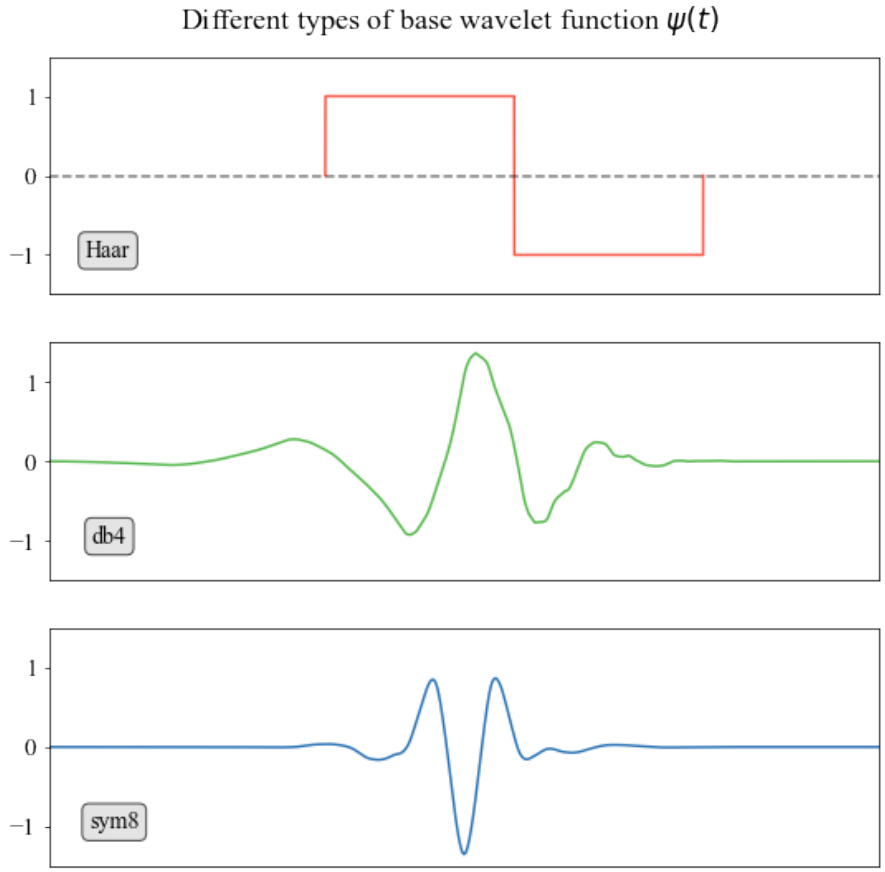

Figure 1.4: Examples of three different wavelet basis functions. In red is the Haar wavelet, also known as the Daubechies level 1 wavelet (db1). In green is the db4 wavelet, and in blue is the symlet level 8 wavelet (modified version of Daubechies wavelets). The area under the curves for all the wavelets always equals zero. Please note that the x-axis (time) for these three wavelets is not to scale, as this figure only tries to show the differences in their shapes.

The wavelet transform is computed by convolving the input signal *f(t)* with $\psi_{s,\tau}$, which represents a basis function often called the *mother wavelet* or just *wavelets*. The computation process is illustrated in Figure 1.5, where one can see how the wavelet transform with the use of the Haar wavelet. The wavelet function is described in a general form in Equation (1.4) and some examples of different wavelets is seen in Figure 1.4. The wavelet acts as a filter that passes over the whole signal while the base wavelet is stretched and

compressed. It is done by changing τ for the time-shift (*translation*) and changing the scaling factor s, which is a *dilation* of the wavelet. This means that the wavelet is stretched or compressed to match all the frequencies in the signal. In the areas where the translated and dilated wavelet is similar with *f(t)*, the wavelet transform returns greater values called wavelet coefficients. The result is a set of decomposed wavelet coefficients of the original signal that provide information about the time localization of the frequencies in *f(t)*. In other words, the wavelet transform act as a mathematical microscope and a useful tool in signal processing for detecting sharp edges, compressing images, or classifying signals.

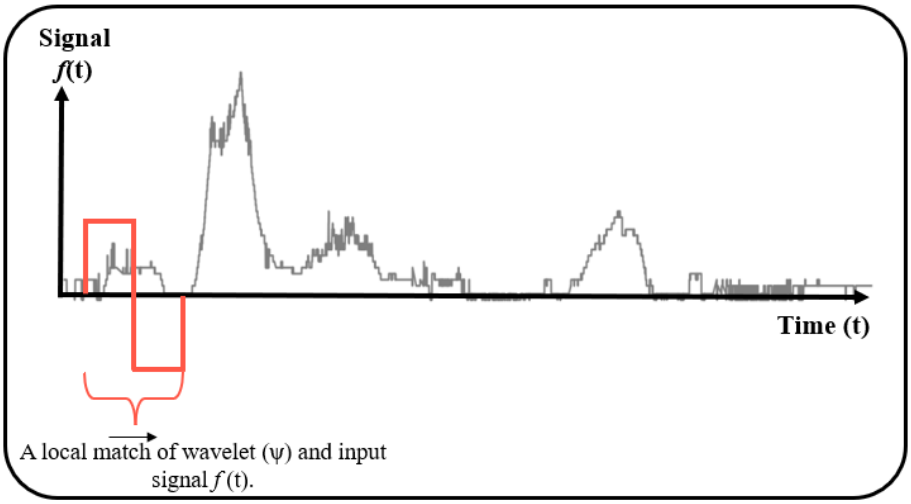

Figure 1.5: Convolution of a signal *f(t)* with a wavelet function $\psi_{s,\tau}$. In this example, the Haar wavelet is used to compute the wavelet transform from a signal attenuation time series. The better the match between the wavelet and signal, the greater the output coefficients. This figure is created with inspiration from Pinello (2012) to illustrate the convolution of a signal.

$$\psi_{s,\tau} = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-\tau}{s}\right), \qquad s, \tau \in \mathbb{R}, a \neq 0 \tag{1.4}$$

The wavelet transform can be divided into two implementation methods. There is the continuous wavelet transform (CWT) found in Equation 1.5 and the discreet wavelet transform (DWT) found in Equation 1.6. The a and b in the DWT represent the τ and s from the CWT, respectively, and are now integers instead of continues values (Srivastava, 2018). N represents the length of the signal.

$$CWT(s,\tau) = \frac{1}{\sqrt{|s|}}\int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{1.5}$$

$$DWT[a,b] = \frac{1}{\sqrt{b}}\sum_{m=0}^{N} f[t_m]\psi\left[\frac{t_m - a}{b}\right] \tag{1.6}$$

The scaling of the wavelets (1.4) is the key factor that separates the continuous and discreet wavelet transform. In CWT, the wavelet is time-shifted and compressed continuously until the frequencies in *f(t)* are covered for the chosen scaling level. However, this may cause the need for high computational complexity and return many very similar values that don't offer much more information. In DWT, the scaling happens at certain intervals called scaling levels. The τ and s from CWT is now replaced by a and b and are based on powers of 2 (dyadic) instead of changing continuously. Thus, the transform is much faster to compute and only includes the most valuable values. The coefficients a and b is described in Equation 1.7 and Equation 1.8 where the parameters j and k control how the wavelet transform decomposes the signal.

$$a = k2^{-j} \tag{1.7}$$

$$b = 2^{-k} \tag{1.8}$$

The resulting values from DWT produce a pair of detail coefficients (cD) and a set of approximation coefficients (cA). The detail coefficients represent the higher frequencies and can be interpreted as a high-pass filter bank. In contrast, the approximation coefficients are the smoothed residual values and can be construed as a low-pass filter. For each level of j, a new pair of cD and cA are computed by taking the DWT of the previous cA. This can be seen in Figure (1.6). This process can be repeated until the length of the signal becomes one and is called multiresolution analysis (MRA) (Srivastava, 2018).

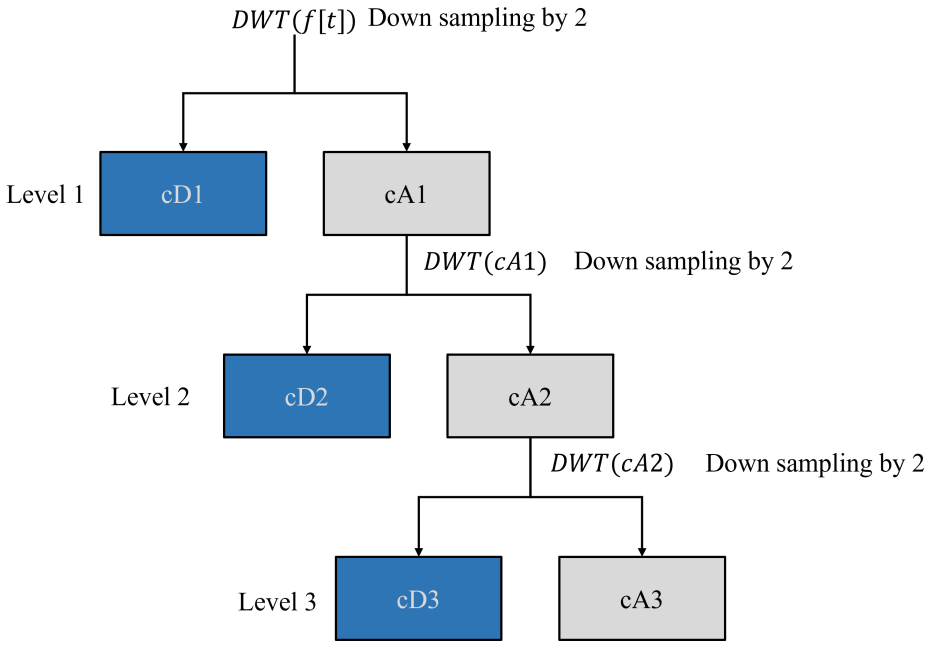

Figure 1.6: DWT for three levels of decomposition, filtering the signal into detail coefficients (cD) and approximation coefficients (cA). For each level of decomposition, the output length is reduced by a factor of two.

1.6 Machine learning

Machine learning uses statistical methods and algorithms to develop models that can make decisions or predictions without human instructions and is in the subfield of artificial intelligence (AI). The machine learning model is learning because it uses its own set of constraints and rules to make a prediction. Although machines don't have real humanlike cognitive properties, their ability to recognize patterns in large data sets is meant to imitate how the human brain works. Machine learning models learn by giving them a subset of the data with the correct labels or values (Tidemann, 2023). This process is called *supervised learning* and is the method used in this study. There are numerous machine learning models available for various purposes. This study utilizes the Random Forest Classifier, a more advanced version of Decision Trees.

1.6.1 Random forest classification

To understand the functioning of the Random Forest Classifier (RFC) algorithm, it is essential to briefly explain decision trees. A simple decision tree asks multiple yes-no questions until it has split the data into categories (Breiman, 2001b). The algorithm starts with a root node and splits the data based on the most informative features into branches. The splitting of the data ends when a certain criterion is met. At the end of each branch, we find the leaf nodes. Leaf nodes represent the final prediction made by the decision tree. A weakness of decision trees is that they are susceptible to overfitting, meaning they may fit the training data too closely and ends up paying attention to the variance in the training data set. This may cause the model to perform poorly with new data.(Criminis et al., 2011).

The RFC algorithm can be used to address the overfitting issue with decision trees. RFC utilizes several decision trees at the same time, and the predictions of each tree are combined to determine the most popular result through a major-voting process. Random forest is an ensemble learning method comprising several classifiers at once. A simple illustration of the RFC can be seen in Figure 1.7, showing how it combines several decision trees. The algorithm uses an ensemble method called bootstrap aggregation to build all the subtrees in the random forest. This involves generating multiple randomized subsets from the original data set and building decision trees from each subset. (Breiman, 2001a). One of the advantages of RFC is that it can find relationships in non-linear systems. On the other hand, if the RFC gets too complicated, meaning it consists of many decision trees, it becomes a *black box* model. Meaning there may not be any physicality to the model. Although RFC is less exposed to overfitting than decision trees, overfitting may still be an issue.

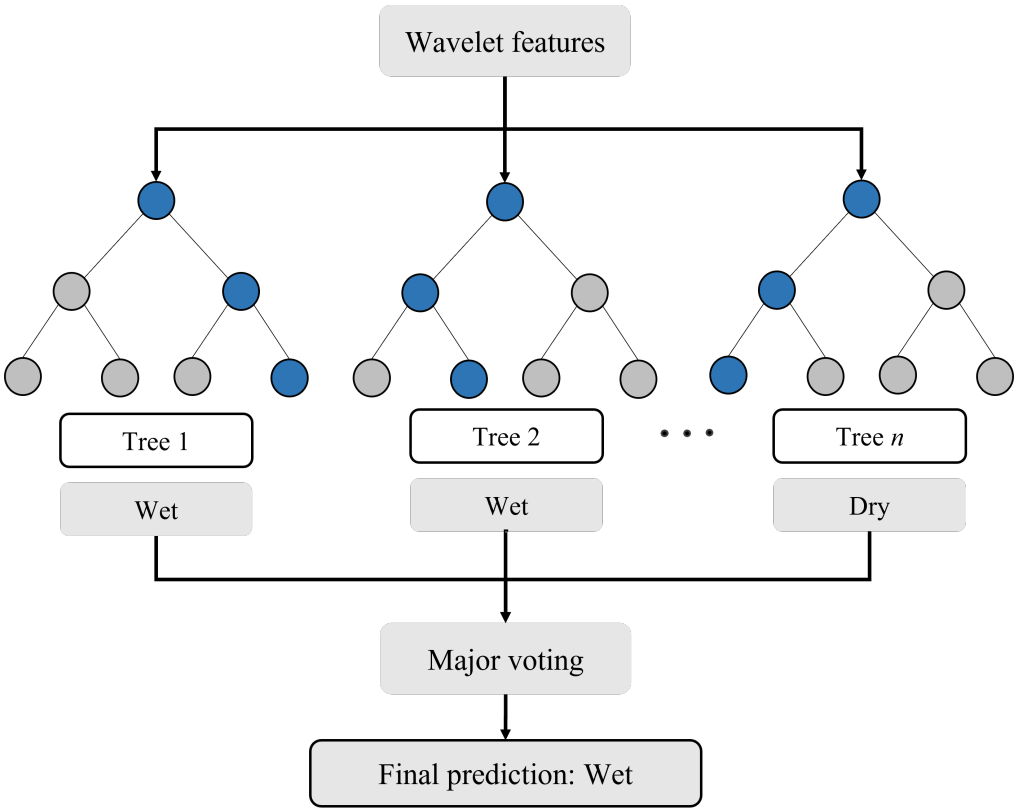

Figure 1.7: An illustration of how a Random Forest Classifier is built consisting of multiple decision trees. In this case, the wavelet coefficients are used as features to predict if there is a rainy or non-rainy event. Figure recreated from Khan et al. (2021), but modifications for the data in this study.

1.6.2 Training and validation of a model

When applying a machine learning model to solve problems, splitting the data into training and testing sets is common when the available data quantity is low. The training data includes data with the correct labels, while the testing set only contains the features and is used to evaluate the model's performance. The challenge of training a model in this manner is overfitting. This happens when the model has learned the training data too well and has essentially memorized the patterns in the data rather than learning general patterns that can be applied to new data.

A cross-validation method is typically applied to avoid or minimize the risk of overfitting when training. Cross-validation means to split the data set into K several folds. $K-1$ folds are used as the training set, and then the Kth fold is used as the testing set for validation. Then another fold is chosen as the testing set. This process continues till all the folds have been used as testing sets. A simple illustration is shown in Figure 1.8 with five folds.

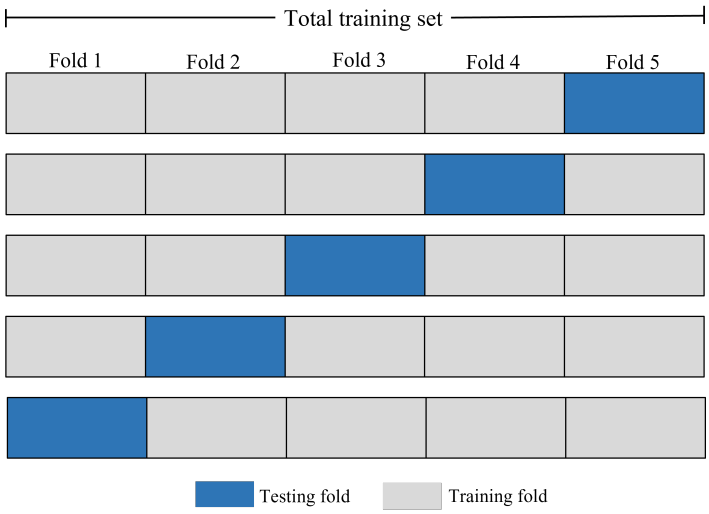

Figure 1.8: Example of how the training data is split into folds when performing a Cross-validation. In this example, there are five folds. The blue represents the data used for testing, while the training data is colored gray.

1.6.3 Hyperparameter tuning

Hyperparameter tuning is an essential step in building machine-learning models, and the goal is to set the best values for the parameters that control the learning process. Every machine-learning algorithm possesses its unique set of tunable hyperparameters that specify the behavior and functionality of the algorithm. An example of hyperparameters in the random forest classification method is how many leaf nodes the tree can split into or how deep the random forest can be. When hyperparameter tuning is done correctly, it can improve the overall performance of the machine-learning model.(Koehrsen, 2018). An explanation of the hyperparameters in the random forest classifier can be found below (Hoffman, 2020).

- **Number of trees:** The total number of individual decision trees in the forest. Often more trees will increase the performance, but it may also lead to overfitting, increased training time, and memory in the computer.
- **Maximum features:** The Maximum number of coefficients (wavelet coefficients) allowed to be considered when looking for the most informative split.
- **Maximum depth:** Determines how deep every tree can be.
- **Minimum samples for split:** Minimum number of coefficients required to split a node further, controlling the complexity within the random forest.
- **Minimum samples for leaves:** The minimum number of coefficients required in a leaf node.

2. Methods

The method used in this study can be divided into three steps to address the goal and research questions proposed in Section 1.1. First, the data used in this study will be presented and described. The second part explores how to implement the discrete wavelet transform (DWT) with machine learning and how this can be used to make algorithms for detecting rain and snow using data from three locations in Norway. Then the third step will explain how the algorithms were applied to a larger data set and compared to existing methods.

For rain detection, three other methods were used to compare the classification method developed in this study, but for snow, there are no available methods to use for comparison. Data from June 2022 was used to detect rain events, and data from December 2021 was used for trying to detect snow events. Disdrometer data was used as a ground truth reference.

2.1 Study areas and data acquisition

Three areas were selected in Norway (Figure 2.1). The primary location was Ås, just south of Oslo. The two other sites were chosen due to their proximity to CMLs with nearby disdrometers and because the probability of snow was higher in Trondheim (50 m.a.s.l) and in the high mountains at Dovre (950 m.a.s.l). In Ås, two CMLs were used for exploring the wavelet's potential in wet-dry classification. Several rain gauges were available, including more advanced meteorological climate stations and disdrometers (about 90 m.a.s.). In Trondheim and at Dovre, only one microwave link was used at each location, together with weather stations.

In Ås, the annual precipitation rate was 1082 mm (Wolff et al., 2021), with a continental type of climate (Dannevig and Harstveit, 2022). Trondheim's climate is characterized by copious amounts of precipitation and moderately mild winters, known as a maritime climate (Dannevig, 2020). The average temperature is approximately 0°C in January, and the annual precipitation rate in Trondheim ranges from 750 mm up to as much as 1500 mm depending on where you are in Trondheim (seNorge.no, 2023). In the mountain

ranges at Dovre, the climate is characterized by an open environment with few trees, strong winds, and a polar climate (Dannevig and Harstveit, 2022). The annual rainfall rate barely exceeds 500 mm per year, and the mean temperatures range from +5°C in the summer and -10°C in the winter months (SeNorge.no, 2023).

Figure 2.1: Map of Norway's three main study areas: Ås, Trondheim, and Dovre. More details of the study areas can be seen in Appendix A.

The CML data used in this study was supplied by Ericsson and Telia by a personal agreement and should only be used for research purposes. Data from two time periods, June 2022 and December 2021, was provided. The data from the network providers consist of six measurements per minute of received signal levels (RSL) and transmitted signal levels (TSL), each with a resolution of 0.1 dB. The lowest RSL (min RSL)

Table 2.1: The four CMLs from the study areas were used when developing the algorithms. The polarization *V* means a vertical polarization.

Link name	Length [km]	Polarization	Frequency [GHz]	Location
AKH007	1.47	V	32	Ås
AKHa01	4.48	V	21	Ås
STR437	1.16	V	39	Trondheim
OPL080	12.8	V	11	Dovre

value and the highest TSL (max TSL) for each minute were extracted for further use. By subtracting the min RSL values from the max TSL, the transmitted-received-signal level (TRSL) was found and is used as the main input data for the classification models. The CML data also included the distance of the link, operating frequency, location, and polarization. The data set was further organized into two channels. Channel 1 from location A to B and channel 2 from location B to A. Despite originating from the same link but in opposite directions, the TRSL values exhibited some differences between the two channels that could be useful for wet-dry detection. Information on the four CMLs used for this part can be seen in Table 2.1. Disdrometers were used to label the precipitation types. Like the traditional pluviometers (rain gauges), the disdrometer is a device used for point observation of precipitation. Compared to measuring the mass of the precipitation like pluviometers that fall into a funnel, the disdrometer typically uses an optical sensor to detect the DSD and velocity of precipitation to classify the precipitation type. Thus the disdrometers are not vulnerable to clogging due to drifting debris or freezing issues. (Islam et al., 2012). Disdrometers were chosen as they can detect rain and snow and are available sources of ground truth in Norway. The data from the disdrometers used in this study were provided by the Norwegian Meteorological Institute (MET) and are owned by the Norwegian Public Roads Administration (SVV). The data was received through the Frost application programming interface (API, `https://frost.met.no/index.html`) operated by MET. The main reason for choosing disdrometers by SVV was to ensure that all the disdrometers were of the same type (or almost the same) and that they were placed in proximity to roads so that they covered most parts of Norway. The main disdrometers type used by SVV, and thus the main disdrometer used in this study, is the OTT Parsivel with a time resolution of 10 minutes. Table 2.2 shows simplified classification codes for the different precipitation types. A period is classified as wet if the disdrometers give a value of 512, and periods with snow result in a value of 2048. The location of the disdrometers relative to CMLs can be seen in the Appendix A.

Table 2.2: Simplifies disdrometers codes for the different precipitation types with the OTT Parsivel.(Tangen, 2020). This study used the disdrometers as a ground truth reference for rain and snow when classifying what caused attenuation in the microwave signal.

Code	Precipitation type
0	No Precipitation
256	Drizzle
512	Rain
1024	Hail
2048	Snow
4096	Unspecified

2.2 The discrete wavelet scalogram

A multilevel decomposition of the discreet wavelet transform (DWT) was performed directly on data from the CMLs from three locations to try to understand how to utilize its properties in rain and snow detection. The discreet wavelet transform (DWT) was chosen over the continuous wavelet transform (CWT) to reduce computational complexity. The Haar wavelet (Figure 1.4) was selected for the mother wavelet as it is the most straightforward wavelet to work with. The Haar-wavelets scaling is exactly 2^J for each level J of decomposition, meaning that the length of the signal is divided by a factor of two for every scaling step. This process is also known to as a dyadic wavelet transform. Thus the total input signal length must be a factor of 2^J when using a dyadic down sampling. In Figure 2.2, a DWT was applied to the TRSL values from the link AKHa01-AKH016 from Ås in June 2022. The DWT was computed through the PyWavelets library for Python with the multilevel decomposition function called *wavedec*. The figures below 2.2 show nine decomposition levels, resulting in the detail coefficients (cD). Figure 2.2 also show the TRSL values in black and the detected rain events in blue. With the length of the input signal being 40960 minutes, the first decomposition (cD1) yielded an array of the finest frequencies filtered out with a length of $40960/2^1$. The following detail coefficients were computed all the way down to cD9 with a length of $40960/2^9$ for the last level. After the DWT were computed, all the detail coefficient from cD1 to CD9 were re-sampled up so that they all had the same length and a rectangular scalogram with dimensions of 9 by 40960. The time-frequency resolutions of the scalogram are seen in Figure 1.3.

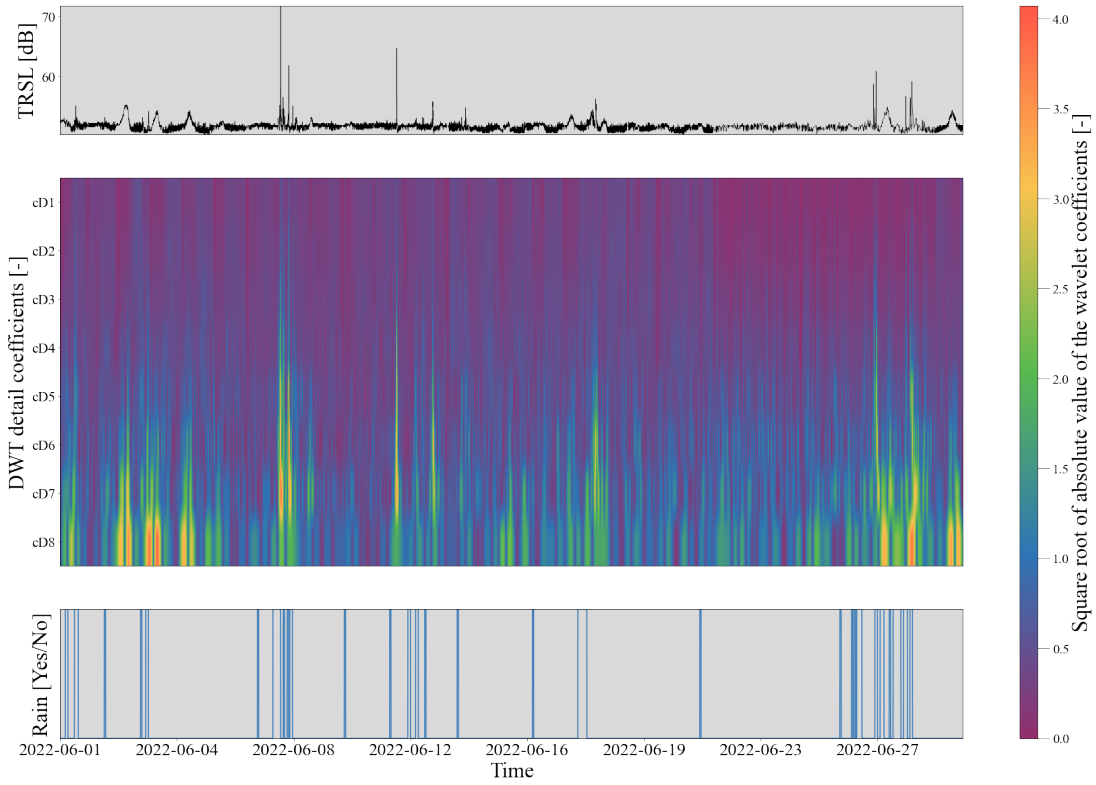

Figure 2.2: Discreet wavelet transform of CML during June 2022 in Ås, Norway. The wavelet coefficients are set to absolute values and then taken to the square root of them for better visualization.

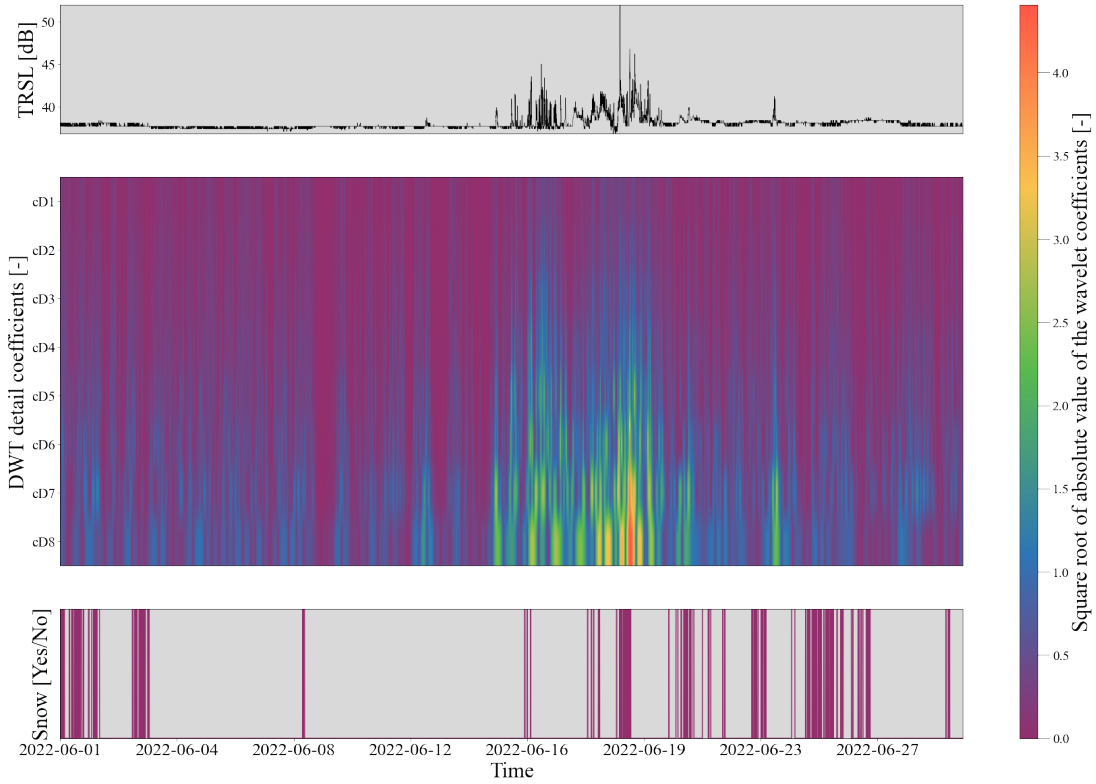

Figure 2.3: Discreet wavelet transform of CML during December 2021 in Trondheim, Norway. The wavelet coefficients are set to absolute values and then taken to the square root of them for better visualization.

As seen in the scalogram (Figure 2.2), the attenuation caused by precipitation is clearly visible, especially at the 4th (cD4) to 8th (cD8) level of decomposition. Greater amplitudes of wavelet coefficients on the lower decomposition levels (cD1 - cD3) capture the finest fluctuations (higher frequencies), possibly due to noises and background fluctuations. In the higher decomposition levels (cD4 - cD9), the frequencies with more extended periods are captured (lower frequencies), and the fluctuation that was induced by rain might be found here. If the signal were further decomposed, it would result in a less informative coefficient and capture only longer trends of frequencies which do not give any information about rain. A wavelet transform of the link STR437-STR139 from Trondheim in December 2021 was also computed (Figure 2.3) to determine if the snowfall was visible in the scalogram. However, results are not as clear as with rainfall.

2.3 Developing the algorithms for classification

As the input length of the attenuation timer series is the main parameter for both DWT and CWT, using the scalogram itself was impossible when trying to develop an algorithm to be versatile for the different lengths of TRSL data. For example, an input length of 40960 minutes at decomposition level 6 would have a length of 640 minutes, while an input signal with a length of 20480 minutes (about two weeks of TRSL data) has a length of 320 minutes at 6 levels of decomposition. Furthermore, when computing the higher levels (e.g., cD8 and cD9) of DWT coefficients, the time resolution was drastically decreased due to the time-frequency trade-off. This means that if the algorithm would need, for example, eight decomposition levels, the input signal would need to be of such length that it was divisible with 2^8 (256). This length issue can be solved through a more sophisticated method that considers the input length and finds the correct scaling level s from Equation 1.5. Then the wet periods can be found by setting the suitable threshold to the scalogram, but using the CWT may need more computational work.

Another way to solve this issue is by applying a sliding window technique that slides over the TRSL values and performs a DWT inside the window frame as a feature extraction method that could be combined with various machine-learning classification techniques. This technique was inspired by Asman et al. (2022) (transient overvoltage in and power quality controllers), Bennet et al. (2014) (classification of heartbeat sounds), and Mei et al. (2021) (DNA microarray technology) where they did a wavelet transform inside a sliding window in combination with machine learning. The sliding window approach lets the algorithm be independent of the input length of the TRSL time series. This study selected the sliding window DWT approach, which had the most promising potential

in the early testing period. This method will be called the Sliding Window Wavelet Transform (SWWT) method. It will be used for rain and snow classification, with small modifications for each precipitation type, meaning that two algorithms were developed.

Therefore, the sliding window approach was investigated and developed further to address the challenges described above. With a short window frame, R consisting of k time steps that slide along the TRSL time series (Equation 2.1), one could ensure a DWT of a signal with constant length L [minutes] independently of the total length of the TRSL time series. We obtain Equation 2.2 by taking the DWT inside the sliding window, and the resulting wavelet coefficients become comparable for every CML. The a and b are described in Section 1.5 and describe how the base wavelet $\psi_{a,b}$ shall behave. Utilizing the information in the wavelet coefficients, an RFC model could be used to classify whether the signal scalogram inside the window was due to a precipitation event (rain or snow) or a non-precipitation event. The RFC would create multiple decision trees to find suitable thresholds to distinguish between precipitation and no precipitation.

$$R(t) = \left\{ R_k | k \in \left\{ t - \frac{L}{2}, ..., t + \frac{L}{2} \right\} \right\} \tag{2.1}$$

$$DWT[a,b] = \frac{1}{\sqrt{b}} \sum_{m=0}^{k-1} R(t_m) * \psi \left[\frac{t_m - a}{b} \right] \tag{2.2}$$

By testing with different sizes for the window length, wavelet decomposition levels, and which wavelet coefficients to include, a window length of 256 minutes was shown to perform well with the wavelet coefficients from levels 4 to 8 (cD4 - cD8) for rain detection. For snow detection, a shorter window of length 32 minutes with wavelet coefficients from levels 1 to 4 (cD1 - cD4). This was done to try to look for snow-induced attenuation in the finer frequencies as it was clear that it was not to be found in similar frequencies as rain as seen in the scalograms from Figure 2.2 and Figure 2.3. The extracted wavelet coefficients w were placed vertically in a 1 by 256-minute vector for each minute in the total TRSL series from a design matrix X. Corresponding correct labelled precipitation types by disdrometer was used as the target vector y. The X matrix is then given to a random forest classifier and evaluated against the correct labels in vector y. Figure 2.5 illustrates the input data.

In Figure (2.4), the process for the method is schematically illustrated from input to output. In Figure D.1 found in Appendix D, the Python code for the **Sliding window** part and **Wavelet feature extraction** part can be seen. The code shows the rain detection version, but the method for detecting snow was done similarly.

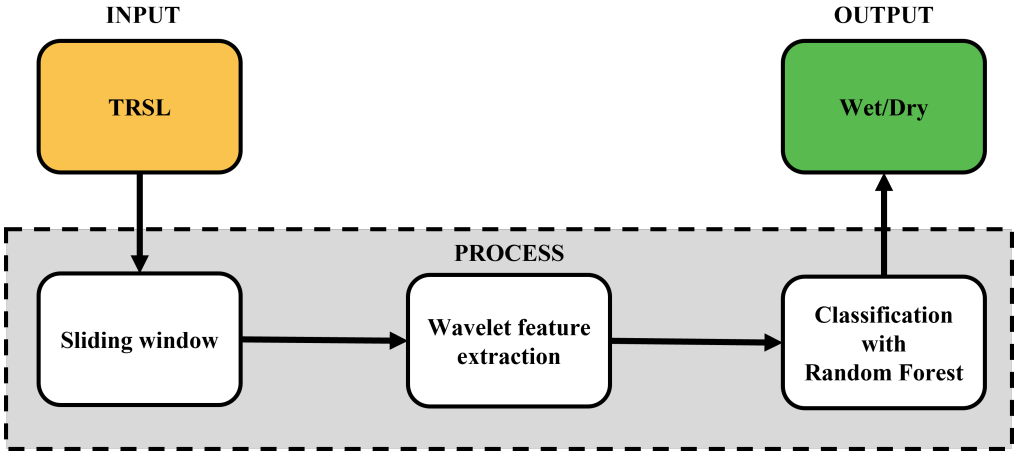

Figure 2.4: A schematic illustration of the classification process. The TRSL (Transmitted-received signal levels) are given as inputs, and the sliding window takes a portion of the signal and applies a discrete wavelet transform. A selected set of wavelet coefficients is then given to a random forest classification model for binary prediction. The illustration was inspired by a figure from Cherkassky et al. (2014).

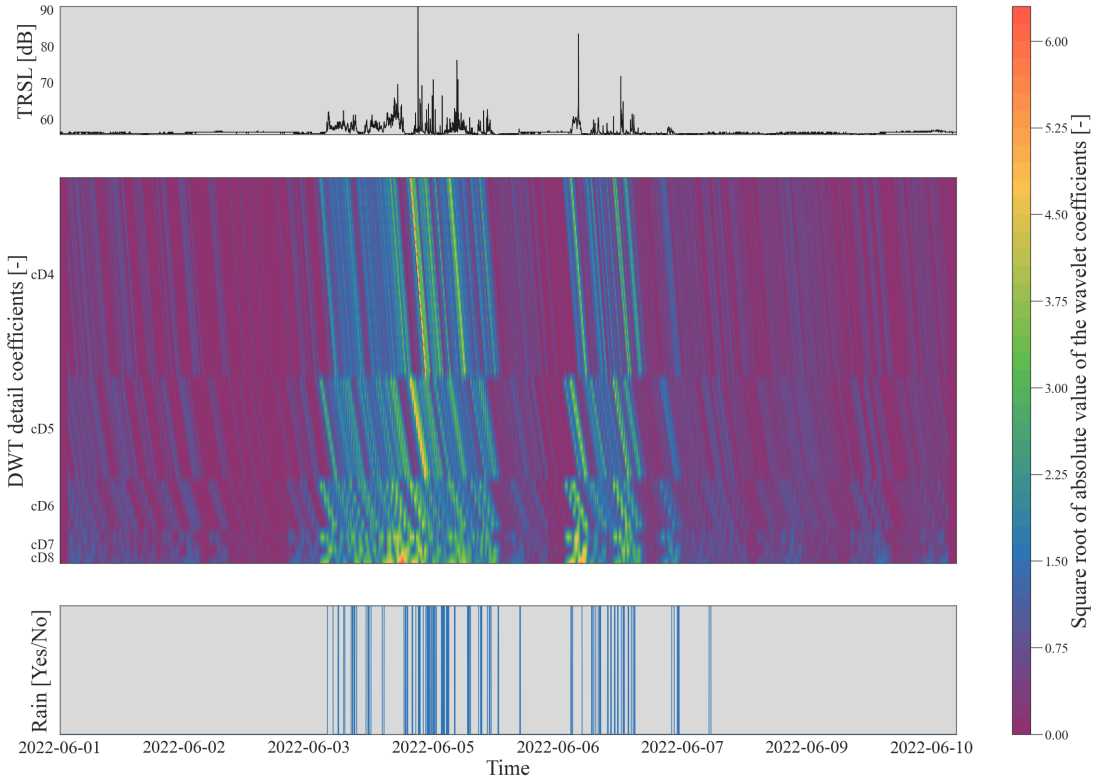

Figure 2.5: This figure shows how to extracted wavelet coefficients from the SWWT algorithm could look like. As the length of the coefficients gets divided by two for every decomposition, there are more cD4 values than cD5 values, and so on. For cD8, there is only one coefficient. Due to the sliding window, the coefficients drift along the window. The coefficients are transformed for better visualization.

2.4 Evaluation of the performance

For evaluating the performance of the classification of rain or snow, the Matthews Correlation Coefficient (MCC) was used. Matthews Correlation Coefficient (MCC) is a statistical measure that ranges from -1.0 to 1.0 and is suitable for binary classification problems. A perfect prediction would have an MCC score of 1.0, whereas a random prediction would have a score close to zero. An MCC score of -1 indicates a perfect negative classification. The MCC-score is high when the numerator in Equation 2.3 has a greater product between the true positives (TP) and the true negatives (TN) than the product of the false positives (FP) and false negatives (FN). The denominator in the MCC is used to scale the MCC score between -1.0 to 1.0. The reason for choosing the MCC-scoring method for evaluation is that it is more robust than the accuracy score (Equation 2.4) when the data is imbalanced (Chicco and Jurman, 2020). In the case of wet-dry classification, the data is often heavily imbalanced against non-precipitation events and is used in other studies for classification methods of wet-dry detection.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{2.3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

2.5 Model training and validation

The smaller data set from Ås, Trondheim, and Dovre served mainly as a testing and developing data set for the methods used in this study. The purpose was to see if there was any potential for using disdrometers and wavelet transform. A larger data set would be needed to evaluate the SWWT method's performance on a more extensive scale. As shown in Figure 2.2 and Figure 2.3, rain affected the CMLs much more than snow. Therefore the two different methods of the SWWT were developed. For rainfall detection, the SWWT was modified to use wavelet coefficients from cD4 to cD8, and for snow, the SWWT was modified to use cD1 to cD4. For the rain classification method, the goal was to make it perform well on several different CMLs. Since there are no methods for detecting dry snow from TRSL fluctuations yet, the goal was to determine if there was any correlation between all the data sets. Both data sets for rain and snow detection were collected from various areas in Norway. The CMLs are given three-letter code names to indicate their location in Norway. The explanations for the CML-code names can be found in Table (2.3).

Table 2.3: The explanations for the code names for the CMLs in the data sets

Link name code	County (Old County name)
AAD	Agder (Aust-Agder)
AKH	Viken (Akershus)
BSK	Viken (Buskerud)
FIM	Troms og Finnmark (Finnmark)
HRD	Vestland (Hordaland)
MOR	Mør og Romasdal (Møre og Romasdal)
NOR	Nordland (Nordland)
NTR	Trøndelag (Nord-Trøndelag)
OPL	Innlandet (Oppland)
OSL	Oslo (Oslo)
SOF	Vestland (Sogn og Fjordane)
STR	Trøndelag (Sør-Trøndelag)
TLM	Vestfold og Telemark (Telemark)
TRM	Troms og Finnmark (Troms)
VAD	Agder (Vest-Agder)

Rain classification

A data set of 165 CMLs from the Norwegian CML network was provided by Ericsson and Telia for June 2022. Then some of the insufficient disdrometer data was removed by eliminating disdrometers with a large amount of *Unspecified* data with the code 4096 from Table 2.2. CMLs with larger baseline shifts resulted in sharper edges in the signal attenuation levels were also removed. A total of total, 28 CMLs were removed. The rest of the CMLs were filtered so that only CMLs within a 4.0 km range of a disdrometer were left, resulting in 58 CMLs for the evaluation part. In CMLs where it could be found missing data, the nearest neighbor interpolation technique was used. A map of the location for the CMLs used for training and validation for the wavelet-based rain model can be found in Appendix B in Figure B.1.

The CML data was split into two parts. One part is for training and testing (training set), and one part is for an independent validation test (validation set). The training set consisted of six CMLs that were selected by links having a good correlation between the rolling standard deviation of the TRSL and the rain events classified. This selection method ensured the model's training on *good* CML data with less noise and prevented the model from learning from noisy CML data or disdrometer data. The six CMLs used in the training set were removed from the whole CML data set to form the validation subset of 52 CMLs. The selected CMLs from the training set can be seen in Table 2.4.

To further ensure that the validation set was independent of the training and testing data, the first ten days of June 2022 (from June 1st to June 10th) were used for the training data, and the time interval from June 12th to June 29th for the validation data.

The reason for this was that several CMLs from the same region that was to be found in both the training set and validation set had similar precipitation patterns and TRSL patterns. The random forest classifier could learn from the patterns in the noise rather than from learning what wavelet coefficients correspond to rain or not rain. To avoid this, the described separation of the data was applied.

Table 2.4: CMLs included in the training of the SWWT classifier for rain. It shows the CMLs name, location, and gigahertz (GHz) operating frequency. The last digit in the link name indicates if the signal is horizontal (1) or vertically (2) polarized.

Link name	Location	Frequency [GHz]
`AKH806_1/1/2.AKH210_1/1/2`	Viken (Akershus)	32
`HRD892_1/1/1.HRD107_1/1/1`	Vestland (Hordaland)	38
`AAD107_1/1/1.AAD057_1/1/1`	Agder (Aust-Agder)	39
`AAD107_1/1/2.AAD057_1/1/2`	Agder (Aust-Agder)	39
`NOR181_1812.NOR101_1012`	Nordland (Nordland)	38
`NOR073_0732.NOR007_0072`	Nordland (Nordland)	37

After the training subset and validation subset were separated, the training set was given to a hyperparameter tuning- and cross-validation algorithm from **Scikit-learn**. A method, called grid search, was done with the GridSearchCV function (Pedregosa et al., 2011). The GridSearchCV were given a set of different hyperparameters, and all possible combination was tested. For each combination, a 5-fold CV (cross-validation) was performed with a custom MCC function with 10-minutes time aggregation for evaluation instead of the default accuracy score. The time aggregation technique will be described later in Section 2.6. The tested and optimal values can be seen in Table 2.5. Figure 2.6 shows how the model's performance reaches a maximum performance level after about 30 trees in the random forest model. This is called the optimum level of model complexity. If the number of trees is further increased, the model will start to overfit and struggle with new validation data (Olteanu, 2018).

Table 2.5: Hyperparameters included in the tuning of the model. It shows the range of the values tested and the result of the best parameter combination. The explanation of the hyperparameters is found in Section 1.6.3

Hyperparameter	Values tested	Optimal
Number of trees	5, 10, 20, 30, 80, 120	30
Maximum features	log2, auto, sqrt	auto
Maximum depth	10, 30, 60, None	30
Minimum samples for split	2, 4, 8	2
Minimum samples for leafs	1, 2, 4	1

Once the model had been tuned and fitted with the training data, it was applied to the 52 CMLs in the validation data. It was applied to both channels (both signal directions),

and a logical *or*-gate was used to classify wet events. The MCC score for 1, 10, 20, 30, 40, 50, 60, and 120 minutes of time aggregation was computed. The location of the CMLs used for the rain classification is found in Figure B.1 in the Appendix B.

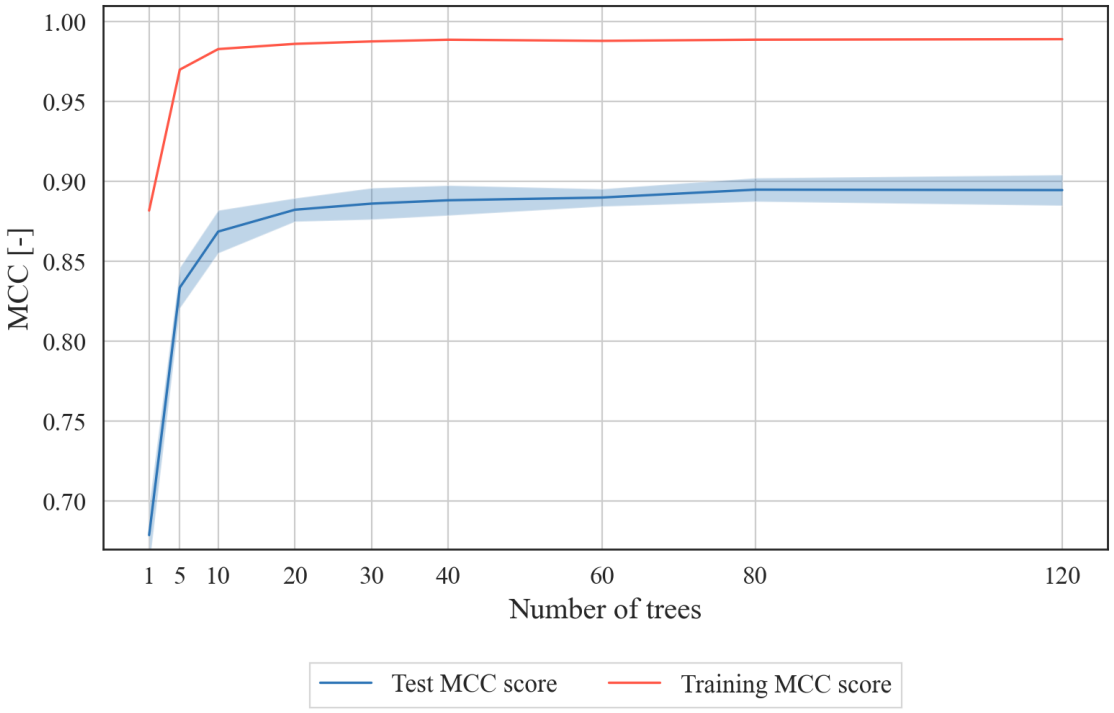

Figure 2.6: A learning curve for the SWWT random forest classifier (RFC) for rain detection. The figure shows how the model performs as the complexity of the model increases (number of trees). The blue line shows the mean MCC score after a 5-fold cross-validation (CV) with test data. The shaded area shows the standard deviation. In red, one can see how the model performs with training data.

Snow classification

The procedure utilized to eliminate bad CML and disdrometer data from the June 2022 data set for the rain detection method was also done for the snow detection methods. The December data set consisted of 128 CMLs, but after removing the bad data and restricting the distance to 4 km between CML and the disdrometer, 41 CMLs were left.

The challenge was to find good CMLs for the training of the SWWT model for snow, the data set from December 2021 was divided into two parts. Part one started from December 1st to December 15th, and part two started from December 16th to December 30th. The first part was used for training, and the second one for validation. This was done to find at least some CMLs that could show some correlation between signal attenuation and snowfall. The reason for this was to see which CMLs were more suited for capturing snow-induced attenuation and that these CMls could help train the model better. Three models (A, B, and C) based on the SWWT were used. The three models had different

RFC hyperparameters and are found in Table 2.6 to test different model complexities. At the same time, the MCC scores of the results were aggregated to amplify to MCC score, so it should be easier to pick out the links with better performance.

Based on the results from the training and testing with snow data from December 2021 seen in Figure 3.4 found in the results in Section 3.2. Eight CMLs with a greater MCC score than 0.1 was selected as the new training data for retraining the SWWT model for snow classification. The training CMLs are shown in Table 2.7. The hyperparameter set from model B (Table 2.6) was shown to be the most promising with respect to overfitting. As with the rain classification, both signal directions were used with a logical *or*-gate was used to classify wet events. Then the mean MCC scores were computed. The location of the CMLs used for the snow classification is found in Figure B.2 in the Appendix (B).

Table 2.6: The table shows the different settings for the random forest classifier used. Resulting in 3 model combinations (A, B, and C). Then each model was aggregated in time for 1, 10, and 60 minutes

Model	Number of trees	Max depth
A	5	5
B	10	25
C	50	None

Table 2.7: CMLs included in the training of SWWT for snow detection. It shows the CMLs name, location, and gigahertz (GHz) operating frequency. The last digit in the link name indicates if the signal is horizontal (1) or vertically (2) polarized.

Link name	Location	Frequency [GHz]
`AKH806_1/1/2.AKH210_1/1/2`	Viken (Akershus)	32
`AAD107_1/1/2.AAD057_1/1/2`	Agder (Aust-Agder)	39
`MOR356_3562.MOR193_1932`	Møre og Romsdal (Møre og Romsdal)	26
`STR420_1/1/1.STR168_1/2/1`	Nordland (Nordland)	29
`VAD152_1/1/2.VAD002_1/1/2`	Agder (Vest-Agder)	39
`VAD229_1/1/2.VAD002_1/3/2`	Agder (Vest-Agder)	39
`NOR073_0732.NOR007_0072`	Nordland (Nordland)	37
`MOR394_3942.MOR013_0132`	Møre og Romsdal (Møre og Romsdal)	32

2.6 Time aggregation

Time aggregation of the disdrometer and TRSL time series was done because a small discrepancy in time between the measured precipitation by the disdrometers and the rain-induced TRSL fluctuation in the CMLs was discovered. The discrepancy was mainly caused by the distance from the disdrometer to the CMLs, resulting in a spatial-temporal bias that would lower the actual correlation. A correction method was used on both

the disdrometer and predicted precipitation for rain and snow events to ensure the overlapping events that most likely should happen simultaneously. Additionally, the time aggregation helps to amplify the MCC score for the methods and it becomes easier to differentiate between various methods. The goal was not to achieve a higher correlation but rather to correct the distance between the links and CMLs. Due to the temporal resolution of the disdrometer data of 10 minutes, a time aggregation of 10 minutes was found most suitable when comparing and validating the result. The correction becomes unrealistic if the time aggregation is extended beyond hundreds of minutes. The aggregation technique is illustrated in Figure 2.7.

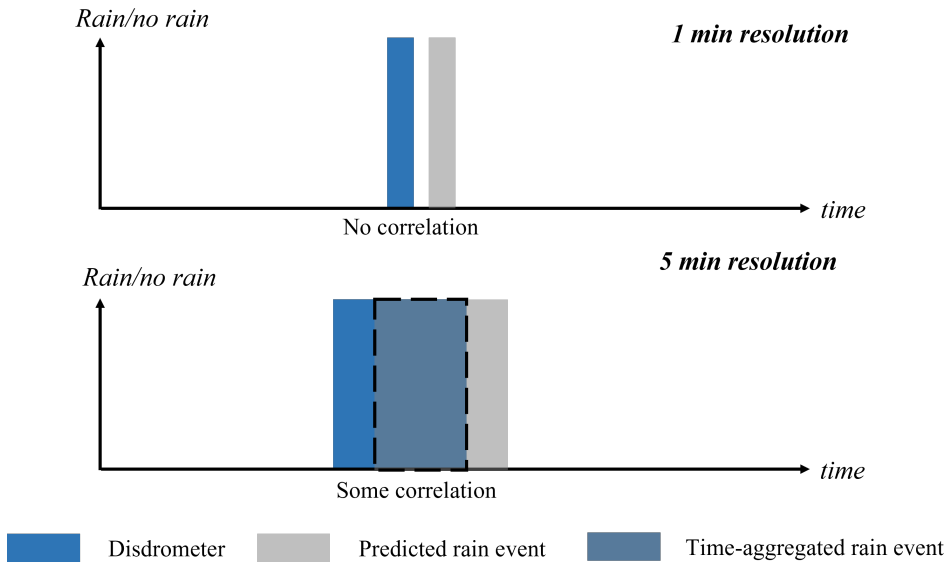

Figure 2.7: Illustration of the temporal aggregation technique. A 1-min resolution is aggregated up to a 5-min resolution to ensure some overlap and yields a more realistic MCC score.

2.7 Comparison of methods for rain detecting

As Section 1.4 mentions, multiple methods exist for detecting rain events. Therefore it was natural to compare the SWWT with some of the existing methods. The classification methods used to compare the SWWT rain method was the simple yet efficient method from Schleiss and Berne, 2010 that computes the rolling standard deviation (RSD) and sets a threshold for detecting parts of the TRSL signal with great fluctuations. The short-time Fourier transform method (STFT) from Chwala et al., 2012, and a convolutional neural network (CNN) developed by Polz et al., 2020 in Germany.

When selecting a method for detecting wet and dry events for a larger set of CMLs, it is advantageous that the method doesn't have too many parameters to optimize because of the differences in the operating frequencies of the CMLs, length, polarization, and

much more. For the RSD method, a standard window size was set to 50 minutes, and the threshold was set to 0.8. The STFT method has four parameters to adjust for each CML: The window size, the f_{divide}, a dry mean, and the threshold limit (σ). The f_{divide} and σ depend on the individual length and frequency of the CMLs. In the study from Chwala et al. (2012), a window size of 256 minutes was recommended, and the optimum values for f_{divide} and σ were found for four CMLs. The average link length for the data set used was about 1.5 km, therefor f_{divide} was set to 0.0031 Hz and σ to 1.0. The dry mean could be calculated by manually defining a dry period's start and end or by defining an interval and automatically letting the algorithm find the dry period. The STFT method was not fully optimized in the comparison and will probably give better results if it was. The CNN doesn't need any parameters applied directly to the Norwegian data set. A challenge for the CNN method is that it was trained on hourly precipitation data, therefore may have some difficulties when evaluated against 10-min precipitation data.

Both the STFT and CNN method was imported through the Python-based Pycomlink library version **0.3.4** developed by Chwala (2017).

3. Results

This chapter presents the results from the rain and snow classification. The results will be presented in two sections. Section one will focus on the wavelet-based rain detection method and how the SWWT method developed in this study compares against three other classification methods. The results from the wavelet-based snow detection method will be presented in the second section of this chapter. Additional results are found in Appendix C.

3.1 Rain detection

The results from the rain event detection are shown in Figure 3.1. The four methods for detecting rain events were applied to the validation data set of 82 CMLs. The mean MCC score was computed for time aggregation from 1 minute (no time aggregation) to 120 minutes. The SWWT and CNN methods show similar results at 1-minute aggregation, but the CNN method quickly outperforms the SWWT after 30 minutes of time aggregation. The STFT method performed somewhat well for 1-minute time aggregation but did not improve much with time aggregation above 30 minutes. Out of the four methods, the RSD method did score the lowest.

In Figure 3.2, the MCC scores were plotted against the distance between the link and the disdrometers to see if the MCC would decrease with the distance. From the figures (Figure 3.2), it could be seen that up to a distance of 4 km, there was no clear sign of decreasing performance with distance in predicting wet events, and none of the methods stood out.

To see how the different methods predict wet events, they were compared over a short period shown in Figure 3.3. For a week, from June 14th to June 21st, a CML in the Trøndelag region was selected to further investigate. The CML from Trøndelag was selected because it had rain, snow, and mixed events. The rain events and all the predictions from the models were time aggregated up to 10 minutes. The true positive events are indicated in green, where the methods predict a wet event when the disdrometers measure the rain. The false negative events are in red and indicate

missed rain events from the models, while the disdrometer recorded a rain event. The false positive events are in yellow and indicted the *false alarms*, where the models would predict rain in a non-rain event. From Figure 3.3, it can be seen that both the CNN method and the STFT method do predict many of the rain events but also cause the most *false alarms* among the models, especially the STFT. The CNN got an MCC score of 42 %, and the STFT received a score of 31 % in the selected period. The SWWT method misses some wet events but did not label too many dry events as wet according to the disdrometer reference. The SWWT method got an MCC score of 36 %. As expected, the RSD method seems to be the worst of the methods. With several missed rain events and with an MCC score of 23 %, it is shown in Figure 3.1.

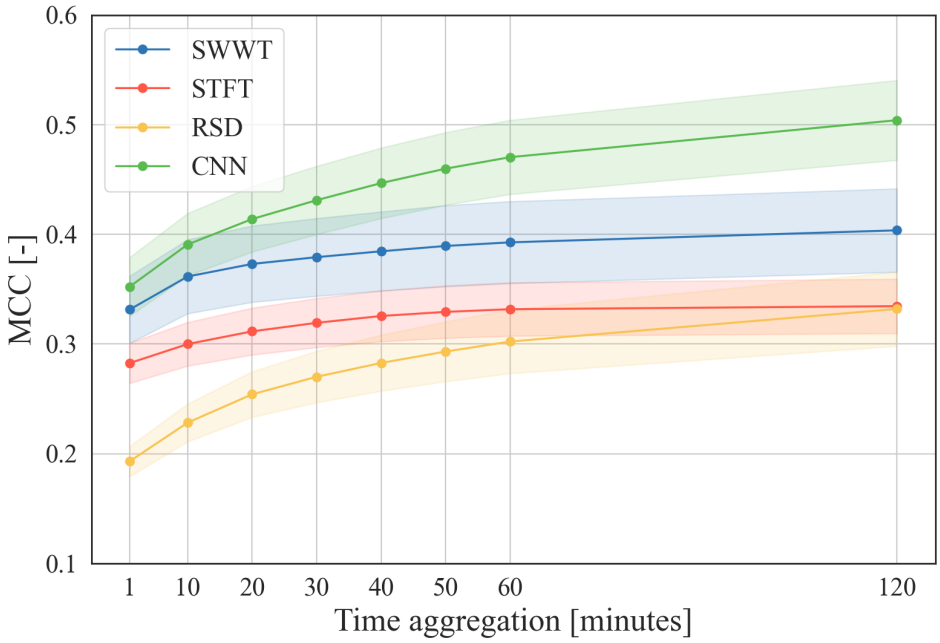

Figure 3.1: The average MCC score (line) and variance (shaded area) across the 52 links. As the time aggregation increased, the MCC scores also rose but slowly flattened out when aggregated. The CNN (convolutional neural network) network) colored in green has shown to have the best performance. The SWWT (sliding window wavelet transform) in blue was right behind the CNN method for shorter time aggregations (>30 min). The STFT (short-time Fourier transform) in red achieved medium results. The RSD (rolling standard deviation) had the lowest MCC score.

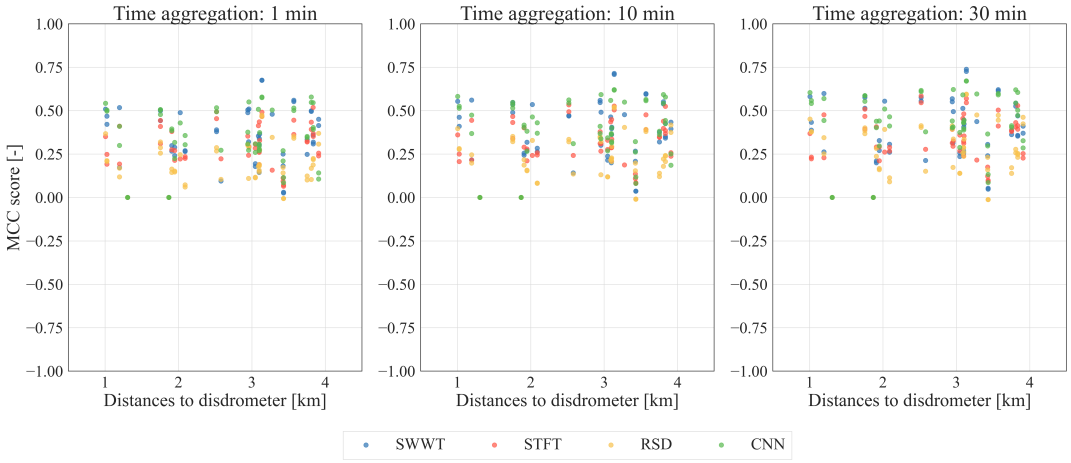

Figure 3.2: MCC-score as a function of distance to disdrometer for three different time aggregations, 1 minute, 10 minutes, and 60 minutes, and with the four different methods. The SWWT (sliding window wavelet transform) method in blue, the STFT (short-time Fourier transform) method in red, the RSD (rolling standard deviation) method is in yellow, and the CNN (convolutional neural network) method in green. There was no sign of a decrease in the MCC score as a function of distance from the disdrometer.

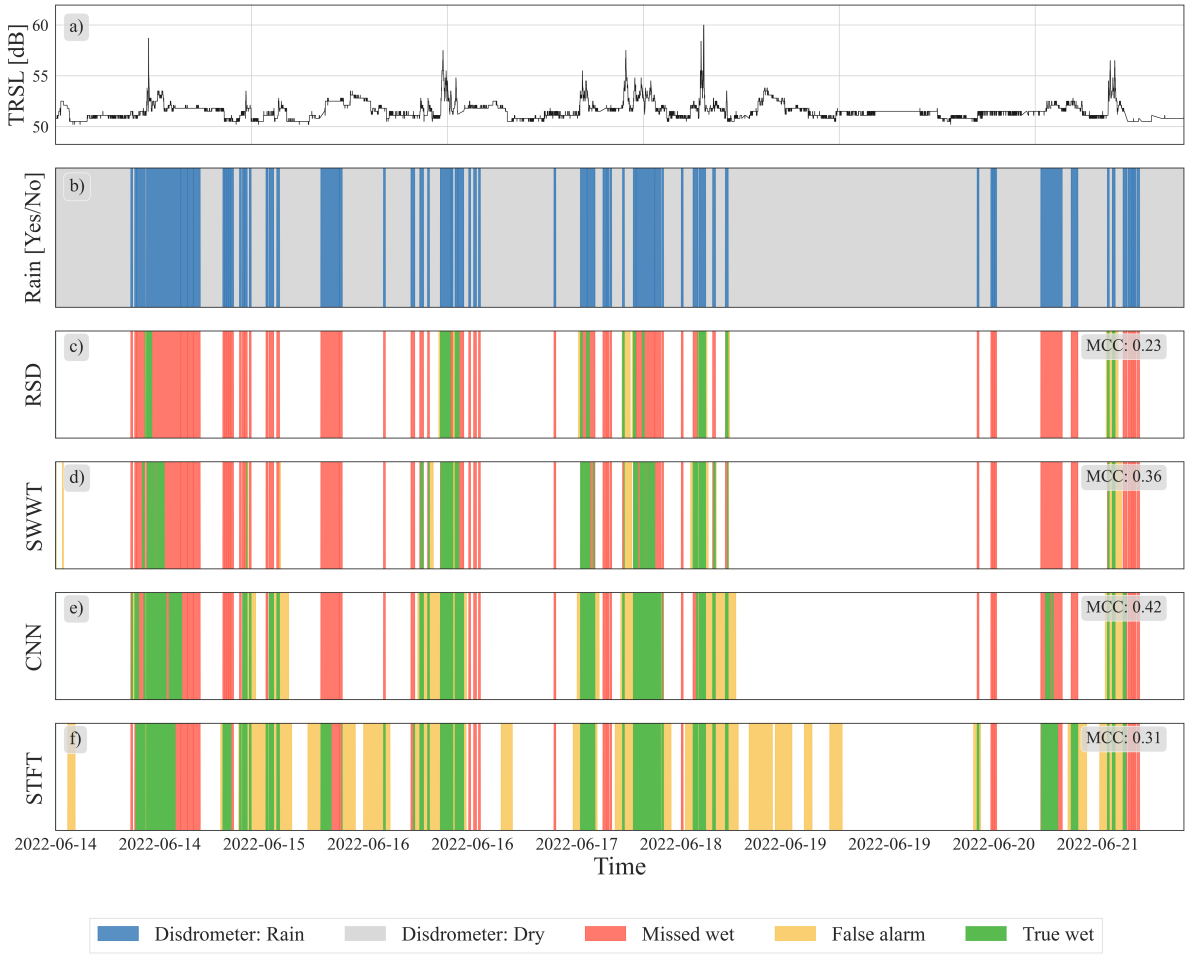

Figure 3.3: An example of how the different methods performed compared to each other. The TRSL levels are shown in (a), the ground truth from the disdrometer is shown in (b), the RSD method from Schleiss and Berne (2010) is shown in (c), the SWWT method is shown in (d), the CNN method from Polz et al. (2020) is shown in (e), and lastly, the STFT by Chwala et al. (2012) is shown in (f). The figure inspired by Polz et al. (2020).

3.2 Snow detection

The results from the method for finding CMLs to use for training data of the SWWT for snow classification are shown in Figure 3.4. This method is described in Section 2.5. Figure 3.4 shows all the combinations of the hyperparameters (Table 2.6) and time aggregation. The three models combined with time aggregations for 1, 10, and 60 minutes resulted in nine plots. The model complexity is increasing from the top row plots to the bottom, and the time aggregation is increasing from the left column of plots to the right. The eight best (MCC>0.1) CMLs from the data set were further used to retrain the SWWT model for snow classification. The purple line shows the MCC score of the validation set (second half of December 2021). The gray line showed the MCC score of the testing data (first half of December 2021).

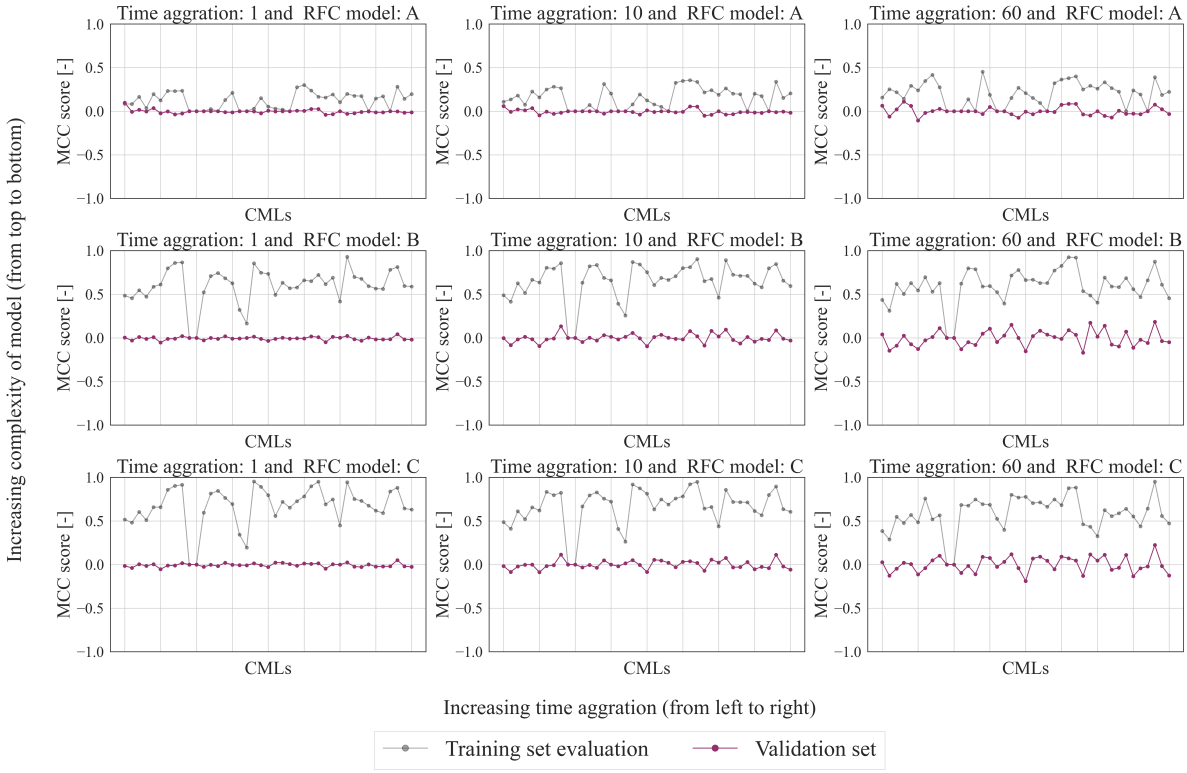

Figure 3.4: Training and testing for all of the 41 of CMLs from December 2021. In gray is the training data set evaluation, and in purple is the test data evaluation. The plots from top to bottom have increasing maximum depth and number of trees of the random forest model, and the plots from left to right have an increasing time aggregation. This figure was used to find CMLs with the best correlation between snow and signal attenuation. CMLs with an MCC score greater than 10 % were selected for training data

The results of the retrained SWWT method trained with the best-performing CMLs for snow classification are shown in Figure 3.5. The results show a very poor MCC score that stays close to 0 % for all the time aggregations from 1 to 120 minutes. The first and second standard deviation is also shown as shaded areas.

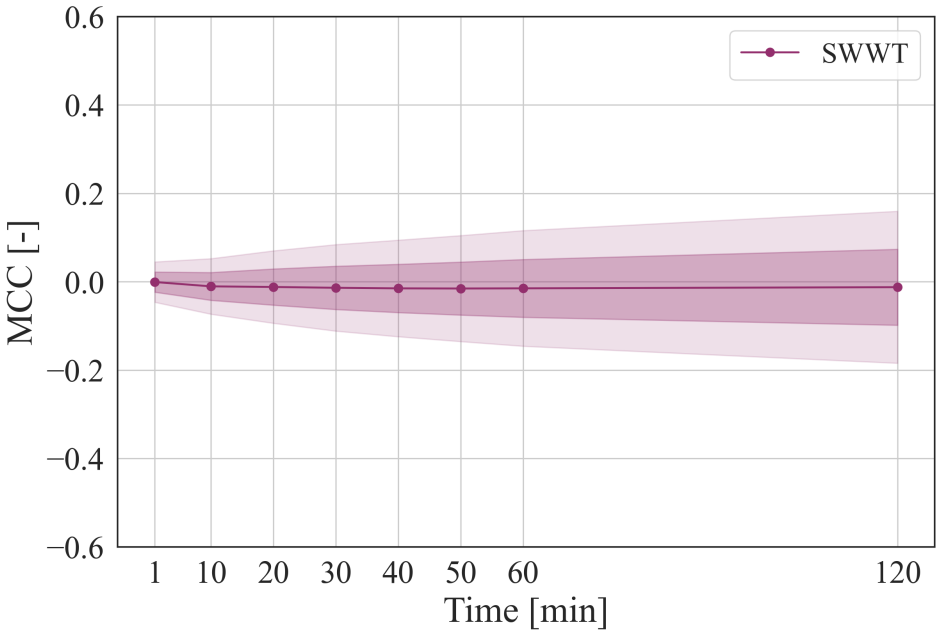

Figure 3.5: The resulting MCC scores from the SWWT model trained with best-performing CMLs from Figure 3.4. The shaded areas show the first two standard deviations. From this figure, no correlation between snow and TRSL can be seen.

As with the rain detection method, a figure of a selected period for a CML in Trondheim can be seen in Figure 3.6 with the predicted outcomes from the SWWT method (e). The figure shows one of the CMLs from the validation set plotted with rain (b), snow (c), and temperature (d) from the nearby disdrometer and weather stations. Again the predicted values for snow are indicated in green, yellow, and red for true positives, false positives, and false negatives, respectively. The figure shows a high amount of missed snow events. Most correct and false predictions occur when the disdrometers show rain events or higher air temperatures, indicating that the model did not learn how to detect snow events. Showing the difficulties of detecting dry snow events.

To see the MCC scores of each CML from both the SWWT model for rain detection and the SWWT for snow detection, they can be found in Figure C.2 and Figure C.1 in Appendix C. The CMLs are ordered alphabetically by their name, which indicates the region where to find the CMLs.

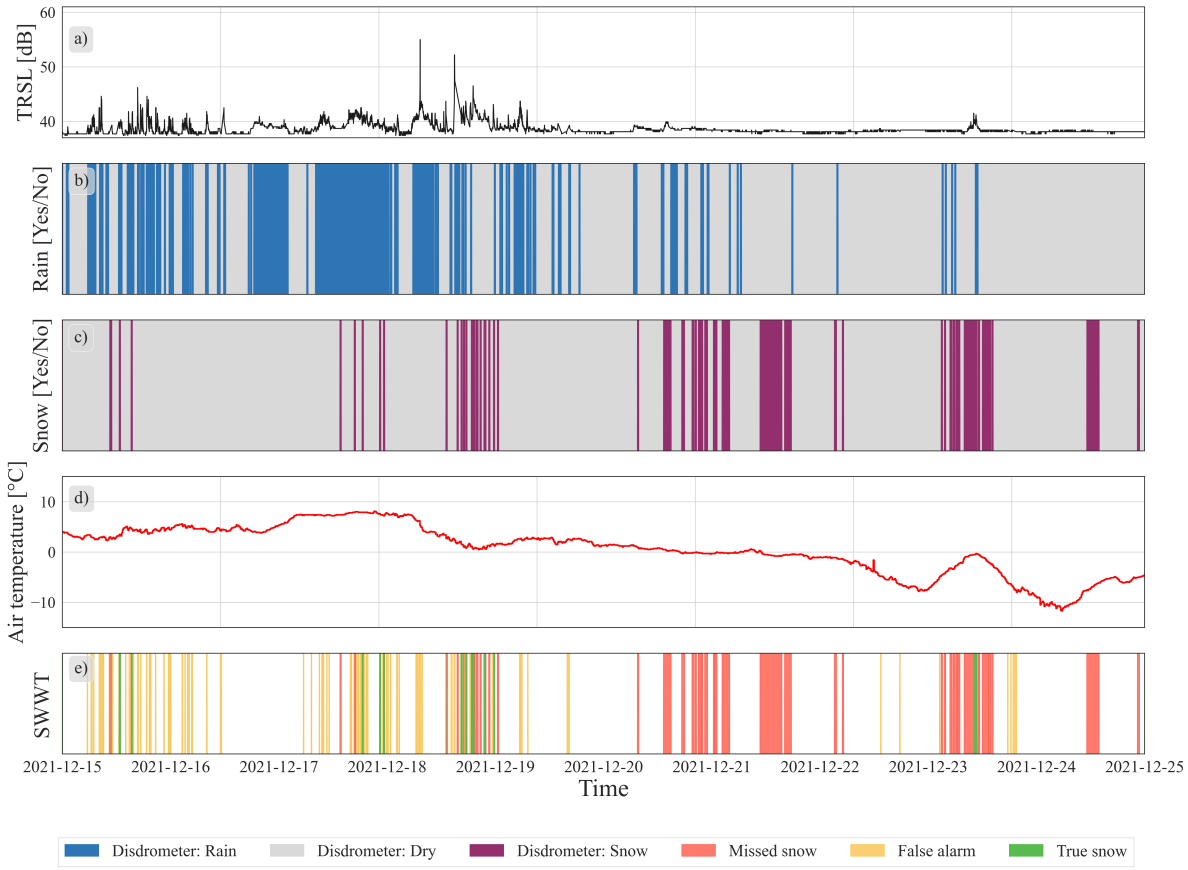

Figure 3.6: An example of the snow classification method on a single CML. The TRSL levels are shown in (a), in (b), the rainfall from the disdrometer is shown, and snow events are shown in (c). The temperature is shown in (d), and the predictions from the SWWT are shown in (e). Time aggregation of 10 minutes. The figure is inspired by Polz et al. (2020).

4. Discussion

This chapter will discuss the results of rain and snow detection to address the thesis's primary goals on how wavelet transform can be used to classify rain and snow and how CMLs can supplement snow measurements. The proposed research questions will be discussed consecutively, starting with **Research question 1** (*Can wavelet transform and machine learning be used to detect rain with attenuation data from commercial microwave links?*) followed by **Research question 2** (*Is it possible to distinguish between periods with and with- out snowfall using the same methodology?*) found in Section 1.1. As there are some similarities between the rain and snow algorithms, common uncertainties will be discussed more generally at the end of this chapter.

4.1 Wavelet transform as feature extraction

The results from Figure 3.1 showed that a discrete wavelet transform (DWT) as a method for feature extraction coupled with a random forest classifier could detect rain. This means that extracting useful information from the attenuation time series with wavelet transform could be a viable option. CMLs have almost exclusively been applied for rainfall-only situations, and as previously mentioned in this study, several rainfall detection methods already exist. Testing a wavelet approach for rain detection was still interesting, as it would serve as a feasibility check and give an idea of how it could be used for other types of precipitation, such as snow.

Comparing the SWWT method to other methods showed that it did score better than the STFT and RSD methods. Compared to the CNN method, the SWWT was not far away for time aggregations up to 20 minutes. The differences between the SWWT and CNN were apparent when aggregated further. As the temporal resolution of the disdrometers was 10 minutes, the models were mainly compared at a 10-minutes time aggregation. Therefore, the SWWT performed well compared with existing methods. On the other hand, it is important to remember that the STFT method needed to be fully optimized for large-scale usage, and it is unknown whether the STFT could outperform the SWWT method.

Usually, a classification algorithm's MCC score between 0.2 and 0.5 would be considered poor. In the papers by Chwala et al. (2012) and Polz et al. (2020), where they proposed the STFT and CNN methods, respectively, both methods showed better results than in this study. As seen in Figure (3.3), there are rain events that are captured by neither method. Better data pre-processing routines could be done for the CML and disdrometer data to improve the correlation scores. The result from the SWWT method compared with the other three methods gave enough evidence that wavelet transform could pick up important information that could be used for wet-dry classification.

Another important finding from the rainfall classification methods was that the MCC score showed no apparent decrease with the increasing distance between the CML and its closest disdrometer. This can be seen in Figure 3.2, where the MCC score for the four wet-dry methods is plotted against the distance between CMLs and the disdrometers. This may indicate that using CMLs could be useful when rain events cover larger areas in space.

Looking closer into the differences between the methods, as exemplified in Figure 3.3, it can be seen how the four methods work in a given case. The STFT and the CNN methods tend to predict wetness for extended periods once they first find a wet event. This is seen as there is a lot more yellow to be seen. This type of error is usually not that bad, as the events colored yellow by the STFT and CNN often occur when there is low attenuation anyway. On the other hand, if there is measured high attenuation, this may lead to an overestimation in the rain intensity calculations.

For the SWWT method, it can be seen that it misses some rain events and that the predictions made are sometimes delayed compared to ground truth observations. The SWWT method also predicts shorter rain events compared with CNN and STFT methods. This could be an important feature that can further improve baseline estimation. By being more exact in the prediction, it could be possible to model a more dynamic baseline during a rain event compared to the current methods, where the baseline often is set to a given attenuation level during the whole period.

4.2 The challenges with detection of dry snow

The results seen in Figure 3.5 from the SWWT model for snow detection showed no correlation between TRSL and snowfall. Although the results from the example shown in Figure 3.6 show that dry snow is difficult to detect, wet snow (a mixture of snow and rain) may be possible to classify. Still, almost all the dry snow alone was undetectable. During the few instances where the model predicts snow correctly, rainfall was also present in the disdrometer data before or after. Although few studies have tried to

classify dry snow in TRSL data previously, similar results found in this were found by Pu et al. (2020). They concluded that dry snow had too little effect on attenuation when using synthetic data.

It is likely that dry snow was not detected because it cannot be differentiated from the surrounding air at the frequencies used in the CMLs found in the data. Most likely due to the dielectric properties of dry snow particles, and when dry snow is compared with liquid precipitation such as rain or wet snow, dry snow has a much lower dielectric permittivity (Ori et al., 2014). This can also be seen below in Figure 4.1, where a histogram from a single CML from Trøndelag shows the number of times a given level of signal attenuation was measured and colored with respect to the precipitation type classified by the nearby disdrometer.

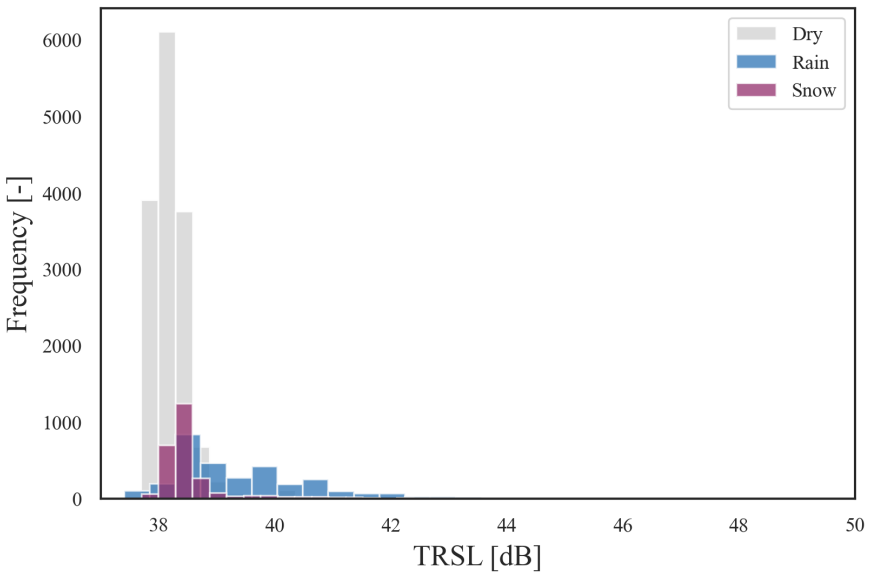

Figure 4.1: The figures shows that snow (purple) and non-precipitation (gray) events occur at the same attenuation levels most of the time. While higher levels of attenuation are mostly caused by rain (blue). A nearby disdrometer made the precipitation classification.

The operating frequencies for the available CMLs used in this study were between 30 GHz and 40 GHz. In a study by Amaya et al. (2014), where they conducted an experiment on snow attenuation in Canada at a 20.2 GHz terrestrial microwave signal, they found that snow-induced attenuation was somewhat modest with a 1-2 dB of attenuation. According to a study by Renaud and Federici (2019), where they theoretically investigated high-frequency attenuation using the Mie theory approach for electromagnetic radiation, it was concluded that dry snow, rain, and air would be distinguishable from around 100 GHz and above. Still, this may lead to challenges as detecting snow with high-frequency (>80GHz) satellite microwave observations increase the noise levels in the data (Liu and Seo, 2013).

It is possible that the dry snow was not detected due to the snow-induced attenuation having a pattern of attenuation that was not accurately captured in the low-resolution sampling of the CML signal attenuation data. The CMLs used for this study use a sampling rate of six instantaneous measurements per minute with an attenuation resolution of 0.3 dB. If the snow had a very small but recognizable pattern, it could have been seen if the sampling resolution of the RSL and TSL were higher.

As with the rain detection methods, some interesting observations can be seen when looking closely into the results from a given period, as seen in Figure 3.6. There was a time interval without rainfall observation right before the 19th of December. At the same time, there was a significant level of TRSL and fluctuations and snow measurements from the disdrometer. The cause of this phenomenon is not exactly clear. It would likely be due to a transition phase between rain and snow, ice accumulation on the antenna, or simply missing rain data from the disdrometer.

In an effort to simplify the snow classification model and, let it only rely on the TRSL time series, the model did not use air temperature data. However, air temperature is an important physical parameter when distinguishing between rain and snow. Figure 4.2 shows two plots. At the top, the hourly mean air temperature from a nearby weather station of a CML and disdrometer in Trondheim was plotted against a set of wavelet coefficients representing the fluctuations in the TRSL time series. Each point in that figure represents a one-minute observation that is colored based on the prediction made by the SWWT for snow classification. The bottom plot of Figure 4.2 shows the correct precipitation types labeled by the disdrometer.

Figure 4.2 shows that very few of the snow events were correctly classified. Between -7.5°C and -5°C, most of the measured snow can likely be considered dry snow, making it difficult to detect. Between -2.5°C and +2.5°C, some snow-induced attenuation could be seen, but this was most likely due to wet snow in a transition phase from cold to warm weather. Therefore, it might have been possible to improve the detection of wet snow if the air temperature had been included in the model. The SWWT for snow detection did not mislabel many rain events.

For comparison, the SWWT model trained for rain was tested on the same December 2021 data from Trondheim. This is seen in Figure (4.3) and shows that the SWWT model mislabels many dry events as rain where the temperature is between -2.5°C to +2.5°C. This indicated that CMLs for precipitation estimation should only be considered if the temperature is above +3°C. No time aggregation was used in both Figure 4.2 and Figure 4.3. Therefore, the prediction from Figure 4.2 is somewhat different from what is seen in Figure 3.6, where a 10-minute time aggregation was used.

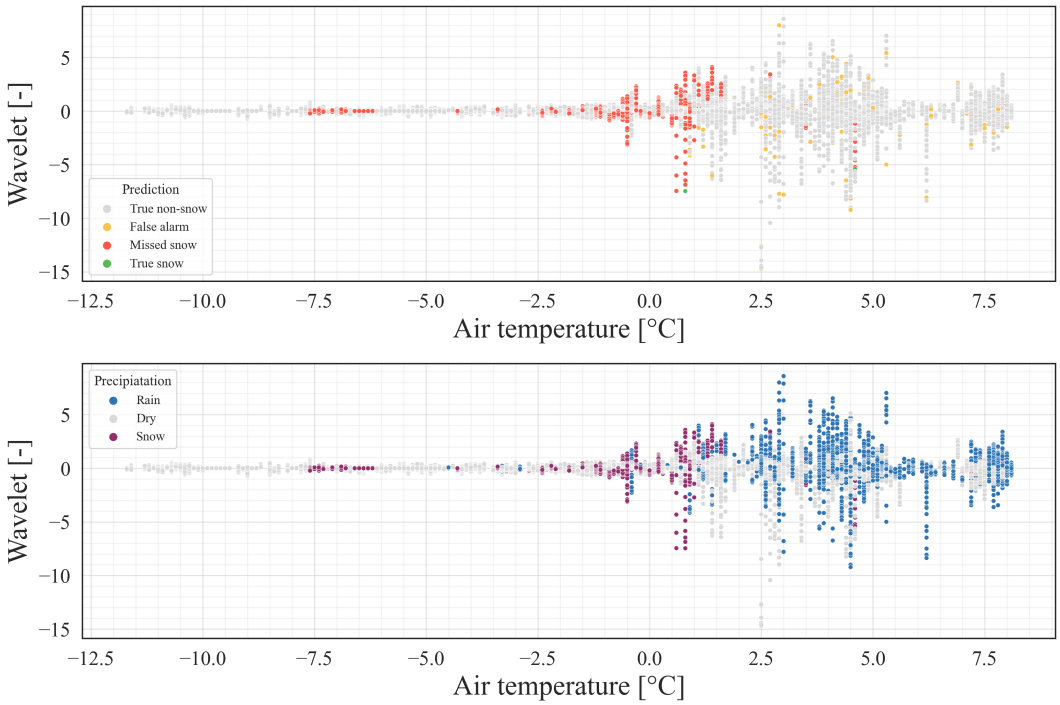

Figure 4.2: The figure shows two plots. At the top, the air temperature is plotted against one of the wavelet coefficients, and the observations are colored by the prediction made by the SWWT for snow classification. The bottom plot shows the correct precipitation types labeled by the disdrometer.

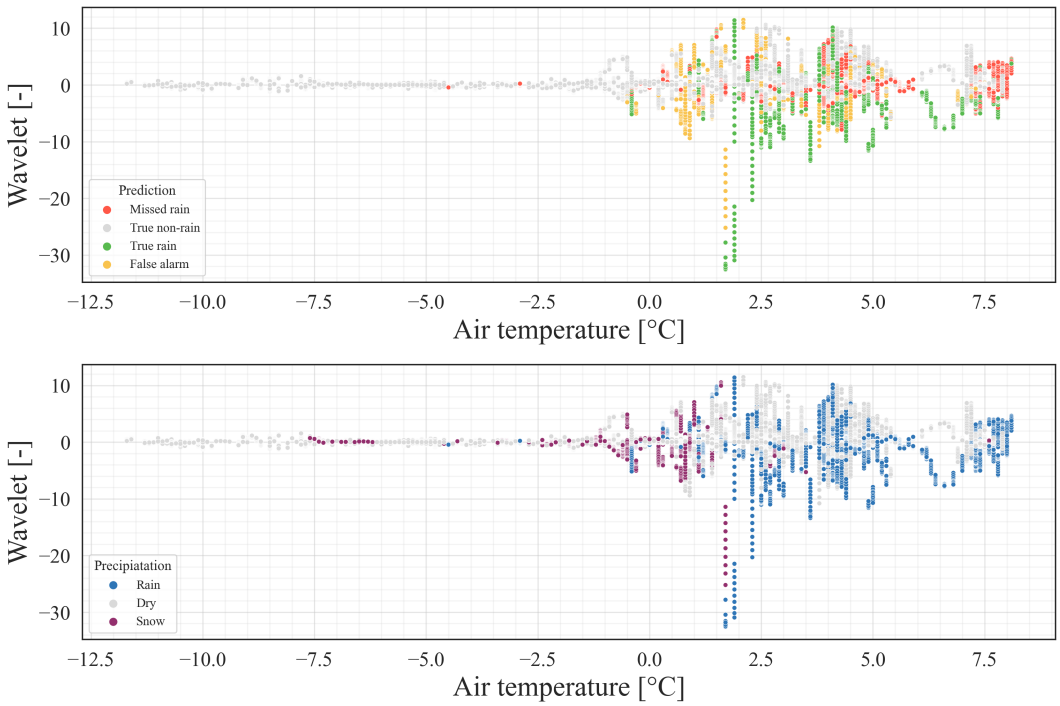

Figure 4.3: The figure shows two plots. At the top, the air temperature is plotted against one of the wavelet coefficients, and the observations are colored by the prediction made by the SWWT for rain classification. The bottom plot shows the correct precipitation types labeled by the disdrometer.

Lastly, Norway's two largest telecommunications network providers, Telenor and Telia, are building and planning to expand their 5G coverage in Norway (Telia Norge AS, 2023; Telenor Norge, 2023). Although optical fibers are more common when expanding telecommunication networks today, antenna towers with wireless communication are currently being built. This may lead to more CMLs with higher operating frequencies than those used in this study and could improve snow detection in the future.

4.3 Additional discussion

The rain and snow detection methods are based on the same algorithm, so they share some common aspects that could be improved. Even though the main objective of this thesis was to classify precipitation events in CML time series, the SWWT methods for rain and snow need to be optimized with respect to the choice of a base wavelet function, window size, and which wavelet coefficients to include. There are many base wavelet functions to choose from. Wavelet transform offers an advantage over short-time Fourier transform as it is not restricted to sine and cosine functions. It enables the selection of various wavelet functions that can be customized according to the specific requirements of the task. For this study, no testing with different wavelets other than the then Haar wavelet was done.

When using the Haar wavelet, the window size was restricted to powers of two (Figure 1.4). There was a direct connection between how many wavelet coefficients are produced and the selected decomposition levels. If, for instance, the method had included all the decomposition levels from level 1 to level 10, a window size of 1024 minutes would be needed, and there would be 1023 detail coefficients. This would have provided more information but also miscellaneous details. Optimizing the window size and wavelet function may not necessarily lead to a significant improvement in snow detection.

There also exist uncertainties in the data used in this study. In a study by Haugen (2015), the OTT Parsivel disdrometer, used by SVV, scored the lowest compared with other disdrometer models. The same study also shows that all the disdrometer models have poor accuracy when detecting frozen precipitation, such as snow. The wind is the main driver of the uncertainties in disdrometers and may increase the falling velocity of the snow so that it is categorized as a heavier hydrometeor. Regarding the CML data, there are some sources of uncertainties. Although an attempt was made to remove the CMLs with the most noise and missing data, some CMLs in the data behaved abnormally, had blackouts, or were generally more sensitive to atmospheric fluctuations. Sometimes high rainfall intensities may cause blackouts in the CML attenuation time series, leaving an empty gap. Even though more sophisticated methods for detecting and filling the missing gaps exist (Polz et al., 2023), this was not done in this study. As a result, a couple of high-attenuation events have probably been missed.

5. Conclusions

This study developed a new method for wet-dry detection in commercial microwave links and tested the model for the rain and snow classification method using real-world data. A frequency analysis technique was developed utilizing the discrete wavelet transform to extract informative features from the signal attenuation time series in the microwave links. Nearby disdrometers were used as ground truth reference as they could detect both snow and rain and were available across Norway.

The method for detecting rain events was shown to perform at similar levels to other wet-dry algorithms when tested on the data used in this study. Therefore **research question 1** is answered, concluding that wavelet transform combined with simple machine learning techniques can be used to distinguish between rain and no rain events. Still, it does not perform better than advanced deep-learning techniques.

To answer **research question 2**, the methods from this study were tested on data from a cold Norwegian winter period of December 2021. The findings suggest the correlation between signal attenuation in microwave links with lower operating frequencies and snowfall is low and that other parameters, such as air temperature, should be considered.

5.1 Further work

It remains to be seen whether CMLs could be used to detect snow at high frequencies (>80GHz) and supplement traditional snow measurements. Therefore it is suggested that high-frequency CMLs are looked into individually to see if any snow-induced attenuation can be found. Temperature data can also be included, as temperature models are available on a larger spatial scale and could help indicate the precipitation type. Furthermore, the absence of attenuation during snow events can be combined with other remote-sensing techniques, such as weather radars. As the disdrometers are not an accurate source of precipitation classification, especially for snow, further work should also include better ground truth measurements.

References

Alonso, R., Pozo, J. M. G. d., Buisán, S. T., and Álvarez, J. A. (Feb. 2021). Analysis of the Snow Water Equivalent at the AEMet-Formigal Field Laboratory (Spanish Pyrenees) During the 2019/2020 Winter Season Using a Stepped-Frequency Continuous Wave Radar (SFCW). *Remote Sensing* 13 (4): 616. DOI: 10.3390/rs13040616.

Amaya, C., García-Rubia, J.-M., Bouchard, P., and Nguyen, T. (Oct. 2014). Experimental assessment of snow-induced attenuation on an Earth-space link operating at Ka-band. *Radio Science* 49 (10): 933–944. DOI: 10.1002/2014RS005547.

Andersson, J. C. M., Olsson, J., van de Beek, R. (Z., and Hansryd, J. (Dec. 2022). OpenMRG: Open data from Microwave links, Radar, and Gauges for rainfall quantification in Gothenburg, Sweden. *Earth System Science Data* 14 (12): 5411–5426. DOI: 10.5194/essd-14-5411-2022.

Asman, S. H., Abidin, A. F., Yusoh, M. A. T. M., and Subiyanto, S. (Mar. 2022). Identification of transient overvoltage using discrete wavelet transform with minimised border distortion effect and support vector machine. *Results in Engineering* 13: 100311. DOI: 10.1016/j.rineng.2021.100311.

Barnett, T., Adam, J., and Lettenmaier, D. (Nov. 2005). Potential impacts of a warming climate onwater availability in snow-dominatedregions. *Nature* 438: 303.

Bellerby, T., Todd, M., Kniveton, D., and Kidd, C. (Dec. 2000). Rainfall Estimation from a Combination of TRMM Precipitation Radar and GOES Multispectral Satellite Imagery through the Use of an Artificial Neural Network. *Journal of Applied Meteorology* 39 (12): 2115–2128. DOI: 10.1175/1520-0450(2001)040<2115:REFACO>2.0.CO;2.

Bennet, J., Arul Ganaprakasam, C., and Arputharaj, K. (2014). A Discrete Wavelet Based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis. *The Scientific World Journal* 2014: 1–9. DOI: 10.1155/2014/195470.

Breiman, L. (2001a). Detection of Ventricular Fibrillation Using Random Forest Classifier. *Machine Learning* 45 (1): 5–32. DOI: 10.1023/A:1010933404324.

Breiman, L. (Jan. 2001b). *Random Forests*. Tech. rep. California: Statistics Department University of California.

Cherkassky, D., Ostrometzky, J., and Messer, H. (2014). Precipitation classification using measurements from commercial microwave links. *IEEE Transactions on Geoscience and Remote Sensing* 52 (5): 2350–2356. DOI: 10.1109/TGRS.2013.2259832.

Chicco, D. and Jurman, G. (Dec. 2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (1): 6. DOI: 10.1186/s12864-019-6413-7.

Chwala, C., Gmeiner, A., Qiu, W., Hipp, S., Nienaber, D., Siart, U., Eibert, T., Pohl, M., Seltmann, J., Fritz, J., and Kunstmann, H. (Mar. 2012). Precipitation observation using

microwave backhaul links in the alpine and pre-alpine region of Southern Germany. *Hydrology and Earth System Sciences Discussions* 9: 741–776. DOI: 10.5194/hessd-9-741-2012.

Chwala, C. (Mar. 2017). *Pycomlink*.

Chwala, C. and Kunstmann, H. (Mar. 2019). *Commercial microwave link networks for rainfall observation: Assessment of the current status and future challenges*. DOI: 10.1002/WAT2.1337.

Criminis, A., Shotton, J., and Konukoglu E (2011). Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Microsoft Research technical report TR-2011-114*: 20–25.

Dannevig, P. (Mar. 2020). *Trøndelag - klima*. URL: https://snl.no/Trondelag_-_klima.

Dannevig, P. and Harstveit, K. (Oct. 2022). *Klima i Norge*. URL: https://snl.no/klima_i_Norge.

Debnath, L. (2001). *Wavelet Transforms and Time-Frequency Signal Analysis Applied and Numerical Harmonic Analysis*. Boston: Birkhauser: 220–224. URL: https://books.google.no/books?id=DUdd3iIvmDUC.

Debnath, L. and Shah, F. A. (2015). *Wavelet Transforms and Their Applications*. 2nd ed. Boston: Birkhauser. DOI: 10.1007/978-0-8176-8418-1.

Dingman, S. L. (2015). *Physical hydrology*. 3rd ed. Long Grove, Illinois: WAVELAND PRESS, INC: 47–70.

Haddeland, I. and Holmqvist, E. (Apr. 2015). Norway's historical and projected water balance in TWh. *Geophysical Research Abstracts, EGU General Assembly 2015*.

Haugen, M. (2015). Vurdering av meteorologiske Present Weather Sensors fra Haukeliseter testfelt.

Hirabayashi, Y., Kanae, S., Motoya, K., Masuda, K., and Döll, P. (2008). A 59-year (1948-2006) global meteorological forcing data set for land surface models. Part II: Global snowfall estimation. *Hydrological Research Letters* 2: 65–69. DOI: 10.3178/hrl.2.65.

Hoffman, K. (Nov. 2020). *Random Forest Hyperparameters Explained*. URL: https://kenhoffman.medium.com/random-forest-hyperparameters-explained-8081a93ce23d.

Islam, T., Rico-Ramirez, M. A., Han, D., and Srivastava, P. K. (Apr. 2012). A Joss-Waldvogel disdrometer derived rainfall estimation study by collocated tipping bucket and rapid response rain gauges. *Atmospheric Science Letters* 13 (2): 139–150. DOI: 10.1002/asl.376.

ITU (2003). *ITU-R P.838-2: Specific Attenuation Model for Rain for Use in Prediction Methods, International Telecommunication Union*. Tech. rep.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., and Raazi, S. M. K.-R. (Sept. 2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity* 2021: 1–18. DOI: 10.1155/2021/2553199.

Kidd, C. (July 2001). Satellite rainfall climatology: a review. *International Journal of Climatology* 21 (9): 1041–1066. DOI: 10.1002/joc.635.

Kidd, C., Becker, A., Huffman, G. J., Muller, C. L., Joe, P., Skofronick-Jackson, G., and Kirschbaum, D. B. (Jan. 2017). So, How Much of the Earth's Surface Is Covered by Rain Gauges? *Bulletin of the American Meteorological Society* 98 (1): 69–78. DOI: 10.1175/BAMS-D-14-00283.1.

Kochendorfer, J., Earle, M., Rasmussen, R., Smith, C., Yang, D., Morin, S., Mekis, E., Buisan, S., Roulet, Y.-A., Landolt, S., Wolff, M., Hoover, J., Thériault, J. M., Lee, G., Baker, B.,

Nitu, R., Lanza, L., Colli, M., and Meyers, T. (Feb. 2022). How Well Are We Measuring Snow Post-SPICE? *Bulletin of the American Meteorological Society* 103 (2): E370–E388. DOI: 10.1175/BAMS-D-20-0228.1.

Koehrsen, W. (Jan. 2018). Hyperparameter tuning the random forest in python. URL: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74.

Leijnse, H., Uijlenhoet, R., and Stricker, J. N. M. (Apr. 2007). Hydrometeorological application of a microwave link: 2. Precipitation. *Water Resources Research* 43 (4). DOI: 10.1029/2006wr004989.

Lian, B., Wei, Z., Sun, X., Li, Z., and Zhao, J. (June 2022). A Review on Rainfall Measurement Based on Commercial Microwave Links in Wireless Cellular Networks. *Sensors* 22 (12): 4395. DOI: 10.3390/s22124395.

Liu, G. and Seo, E.-K. (Feb. 2013). Detecting snowfall over land by satellite high-frequency microwave observations: The lack of scattering signature and a statistical approach. *Journal of Geophysical Research: Atmospheres* 118 (3): 1376–1387. DOI: 10.1002/jgrd.50172.

Magono, C. (Feb. 1954). On The Shape Of Water Drops Falling In Stagnant Air. *Journal of Meteorology* 11 (1): 77–79. DOI: 10.1175/1520-0469(1954)011<0077:OTSOWD>2.0.CO;2.

Mei, N., Wang, H., Zhang, Y., Liu, F., Jiang, X., and Wei, S. (Oct. 2021). Classification of heart sounds based on quality assessment and wavelet scattering transform. *Computers in Biology and Medicine* 137: 104814. DOI: 10.1016/j.compbiomed.2021.104814.

Messer, H., Zinevich, A., and Alpert, P. (May 2006). Environmental Monitoring by Wireless Communication Networks. *Science* 312 (5774): 713–713. DOI: 10.1126/science.1120034.

Nashashibi, F., de Charrette, R., and Lia, A. (Dec. 2010). Detection of unfocused raindrops on a windscreen using low level image processing. In: *2010 11th International Conference on Control Automation Robotics & Vision*. IEEE: 1410–1415. DOI: 10.1109/ICARCV.2010.5707398.

NVE (Jan. 2019). *Snø*. URL: https://www.nve.no/vann-og-vassdrag/vannets-kretsloep/snoe/.

Olteanu, A. (Jan. 2018). Learning Curves for Machine Learning in Python. *Dataquest*. URL: https://www.dataquest.io/blog/learning-curves-machine-learning/.

Ori, D., Maestri, T., Rizzi, R., Cimini, D., Montopoli, M., and Marzano, F. S. (Aug. 2014). Scattering properties of modeled complex snowflakes and mixed-phase particles at microwave and millimeter frequencies. *Journal of Geophysical Research: Atmospheres* 119 (16): 9931–9947. DOI: 10.1002/2014JD021616.

Ostrometzky, J., Cherkassky, D., and Messer, H. (2015). Accumulated Mixed Precipitation Estimation Using Measurements from Multiple Microwave Links. *Advances in Meteorology* 2015: 1–9. DOI: 10.1155/2015/707646.

Ostrometzky, J. and Messer, H. (Jan. 2018). Dynamic Determination of the Baseline Level in Microwave Links for Rain Monitoring From Minimum Attenuation Values. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (1): 24–33. DOI: 10.1109/JSTARS.2017.2752902.

Overeem, A., Leijnse, H., and Uijlenhoet, R. (June 2016). Retrieval algorithm for rainfall mapping from microwave links in a cellular communication network. *Atmospheric Measurement Techniques* 9 (5): 2425–2444. DOI: 10.5194/amt-9-2425-2016.

Øydvin, E., Hagland, R. F., Nilsen, V., Wolff, M. A., and Kitterød, N.-O. (Apr. 2023). *Classification of snow and rainfall using commercial microwave links*. Vienna, Austria. DOI: 10.5194/egusphere-egu23-12061.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*: 2825–2830.

Pinello, L. (2012). Multi Layer Analysis. PhD thesis. Palermo: Università degli Studi di Palermo: 8.

Polz, J., Chwala, C., Graf, M., and Kunstmann, H. (July 2020). Rain event detection in commercial microwave link attenuation data using convolutional neural networks. *Atmospheric Measurement Techniques* 13 (7): 3835–3853. DOI: 10.5194/amt-13-3835-2020.

Polz, J., Graf, M., and Chwala, C. (Feb. 2023). Missing Rainfall Extremes in Commercial Microwave Link Data Due To Complete Loss of Signal. *Earth and Space Science* 10 (2). DOI: 10.1029/2022EA002456.

Pu, K., Liu, X., Hu, S., and Gao, T. (July 2020). Hydrometeor identification using multiple-frequency microwave links: A numerical simulation. *Remote Sensing* 12 (13). DOI: 10.3390/rs12132158.

Rabiei, E., Haberlandt, U., Sester, M., and Fitzner, D. (Nov. 2013). Rainfall estimation using moving cars as rain gauges – laboratory experiments. *Hydrology and Earth System Sciences* 17 (11): 4701–4712. DOI: 10.5194/hess-17-4701-2013.

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K., and Gutmann, E. (June 2012). How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed. *Bulletin of the American Meteorological Society* 93 (6): 811–829. DOI: 10.1175/BAMS-D-11-00052.1.

Renaud, D. L. and Federici, J. F. (Aug. 2019). Terahertz Attenuation in Snow and Sleet. *Journal of Infrared, Millimeter, and Terahertz Waves* 40 (8): 868–877. DOI: 10.1007/s10762-019-00607-y.

Robertson, S. and King, A. (Apr. 1946). The Effect of Rain upon the Propagation of Waves in the 1- and 3-Centimeter Regions. *Proceedings of the IRE* 34 (4): 178p–180p. DOI: 10.1109/JRPROC.1946.234239.

Schleiss, M. and Berne, A. (July 2010). Identification of Dry and Rainy Periods Using Telecommunication Microwave Links. *IEEE Geoscience and Remote Sensing Letters* 7 (3): 611–615. DOI: 10.1109/LGRS.2010.2043052.

SeNorge.no (Feb. 2023). *Monthly mean temperature*. URL: https://www.senorge.no/.

seNorge.no (Feb. 2023). *Precipitation 1991-2020*. URL: https://www.senorge.no/.

Singh, P. S. (2001). Snow and Glacier Hydrology. *Springer* 37 (Water Science and Technology Library): 1–45.

Srivastava, M. (Aug. 2018). Improving Signal Resolution And Reducing Experiment Time In Electron Spin Resonance Spectroscopy Via Data Processing Methods. PhD thesis. Ithaca: Cornell University Library.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., and Hsu, K.-L. (Mar. 2018). A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Reviews of Geophysics* 56 (1): 79–107. DOI: 10.1002/2017RG000574.

Tangen, M. (2020). Comparing the performance of machine learning-based methods to traditional temperature threshold methods for precipitation phase partitioning across Norway. *Department of Physics Faculty of mathematics and natural sciences*.

Telenor Norge (Apr. 2023). *5G er fremtidens mobilnett*. URL: https://www.telenor.no/5g/.

Telia Norge AS (Apr. 2023). *Dette er 5G*. URL: https://www.telia.no/nett/5g/.

Tidemann, A. (Jan. 2023). *Kunstig intelligens*. URL: https://snl.no/kunstig_intelligens.

Vistnes, A. I. (2016). Wavelettransformasjon. In: *Svingninger og bolgers fysikk (Norwegian Edition)*. Versjon 21.04.2016. Oslo: UiO. Chap. 13.

Wang, Z., Schleiss, M., Jaffrain, J., Berne, A., and Rieckermann, J. (July 2012). Using Markov switching models to infer dry and rainy periods from telecommunication microwave link signals. *Atmospheric Measurement Techniques* 5 (7): 1847–1859. DOI: 10.5194/amt-5-1847-2012.

Witze, A. (Apr. 2016). Snow-sensing fleet to unlock water's icy secrets. *Nature* 532 (7597): 17–17. DOI: 10.1038/532017a.

Wolff, M. A., Isaksen, K., Petersen-Øverleir, A., Ødemark, K., Reitan, T., and Brækkan, R. (Feb. 2015). Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study. *Hydrology and Earth System Sciences* 19 (2): 951–967. DOI: 10.5194/hess-19-951-2015.

Wolff, M. A., Thue-Hansen, V., and Grimenes, A. A. (Feb. 2021). *METEOROLOGISKE DATA FOR ÅS 2020*. Tech. rep. Ås: Norges miljø- og biovitenskapelige universitet.

Yoo, J., Lee, C.-H., Jea, H.-M., Lee, S.-K., Yoon, Y., Lee, J., Yum, K., and Hwang, S.-U. (Sept. 2022). Classification of Road Surfaces Based on CNN Architecture and Tire Acoustical Signals. *Applied Sciences* 12 (19): 9521. DOI: 10.3390/app12199521.

Appendix A. Maps of study the areas

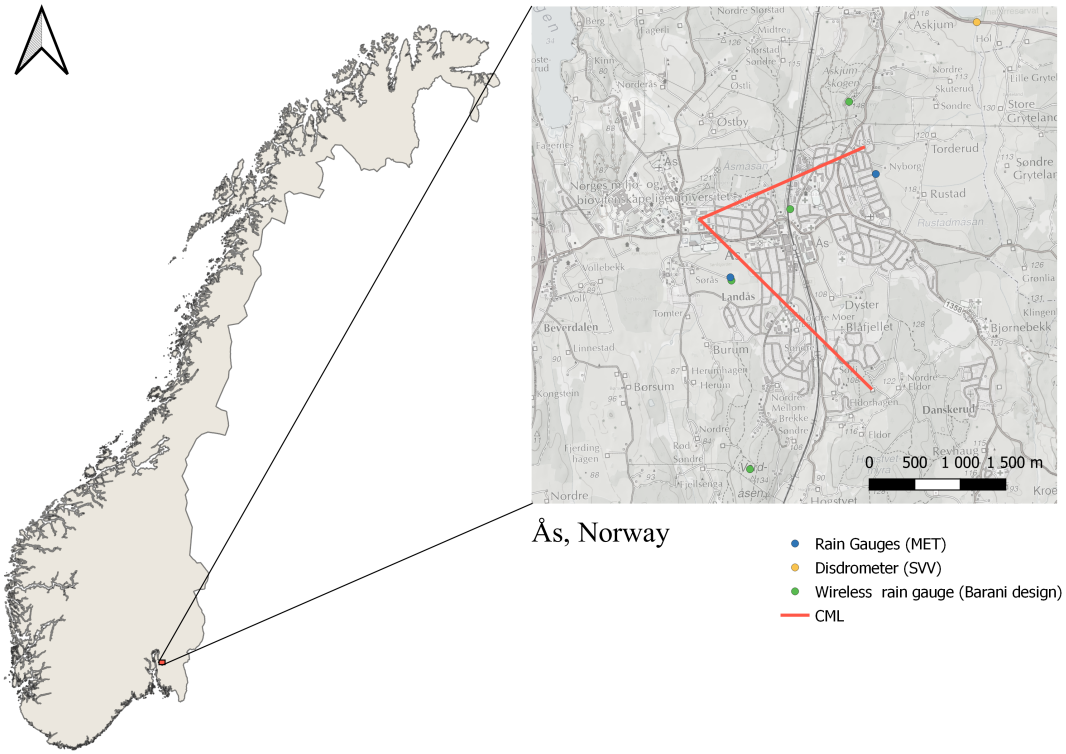

Figure A.1: The map shows the ground instrumentation and CMLs from Ås. In blue, one can see the rain gauges; in yellow, one can see disdrometers from FROST API https://frost.met.no/index.html. In green, one can see four wireless rain gauges from Barani design https://www.baranidesign.com/. The true CMLs (red) coordinates have been fuzzed to hide their real locations.

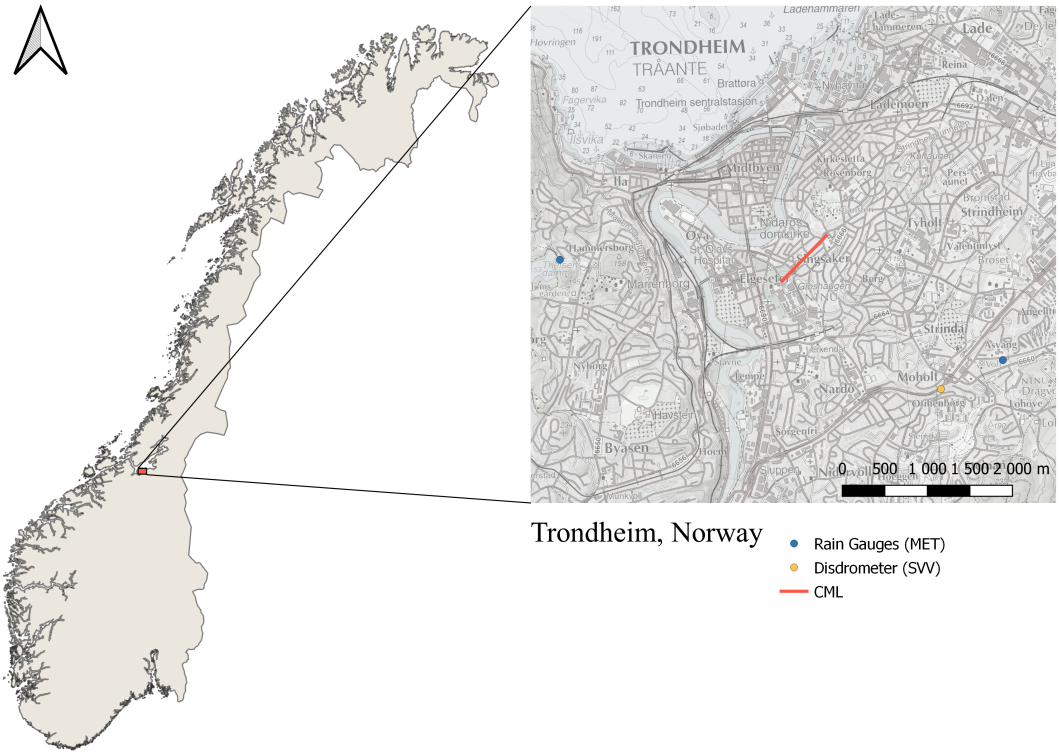

Figure A.2: The map shows the ground instrumentation and CMLs from Trondheim. In blue, one can see the rain gauges; in yellow, one can see disdrometers from FROST API https://frost.met.no/index.html. The true CMLs (red) coordinates have been fuzzed to hide their real locations.

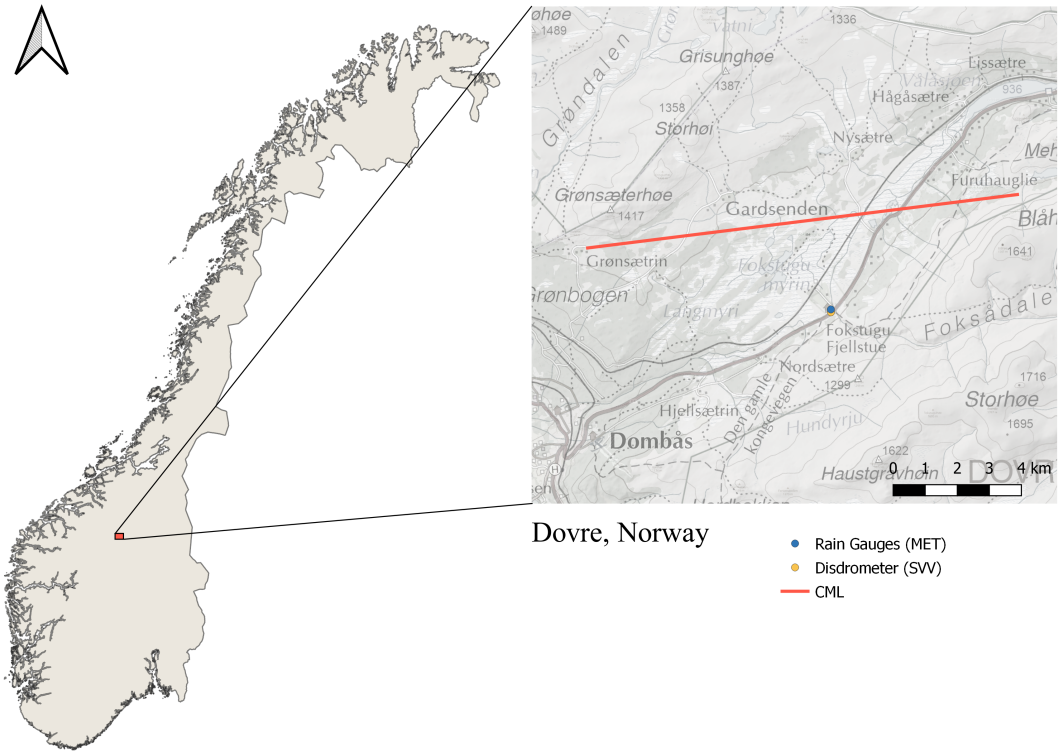

Figure A.3: The map shows the ground instrumentation and CMLs from Dovre. In blue, one can see the rain gauges; in yellow, one can see disdrometers from FROST API `https://frost.met.no/index.html`. The true CMLs (red) coordinates have been fuzzed to hide their real locations.

Appendix B. Spatial distribution of CMLs

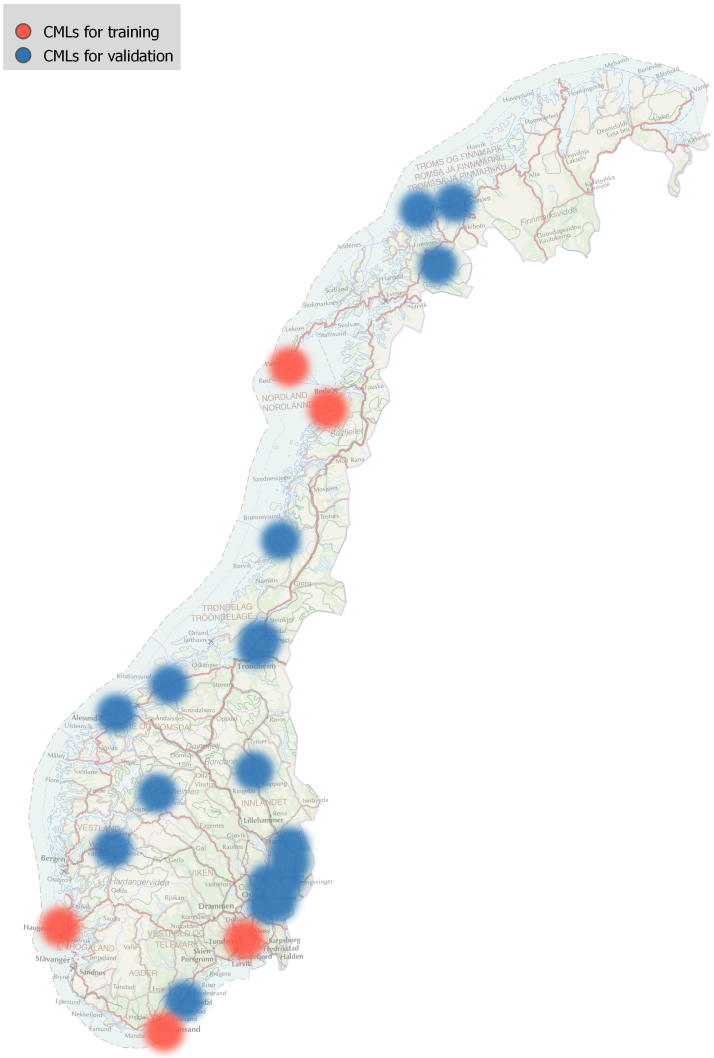

Figure B.1: A map illustrating the spatial distribution of the training/testing CMLs (red) and validation CMLs (blue) for the rain detection method. The maps do not show the true location of the CMLs.

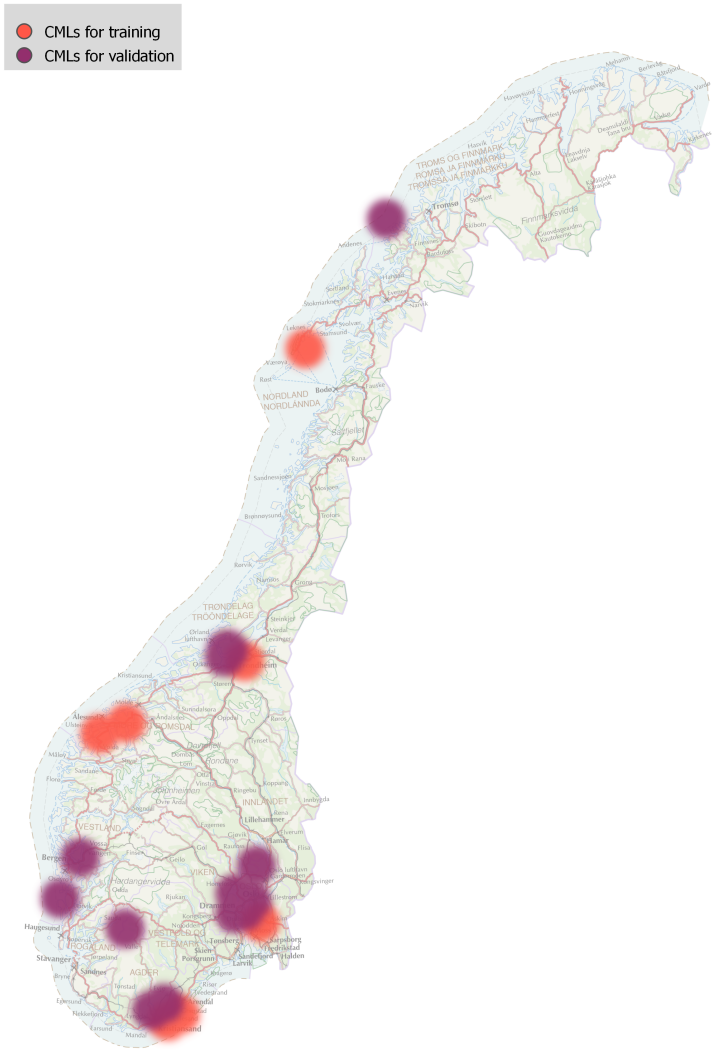

Figure B.2: A map illustrating the spatial distribution of the training/testing CMLs (red) and validation CMLs (purple) for the snow detection method. The maps do not show the true location of the CMLs.

Appendix C. Validation of each CML

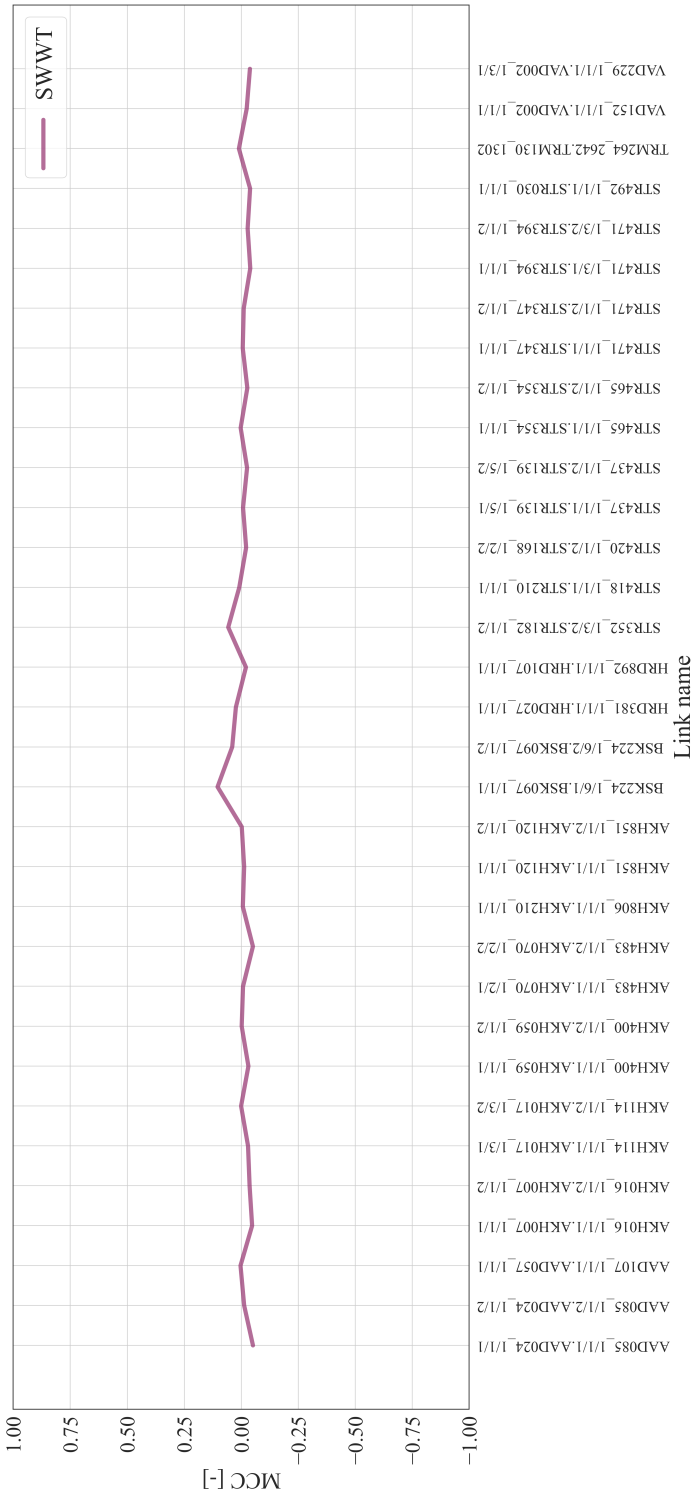

Figure C.1: Results from the SWWT for all CMLs for snow classification

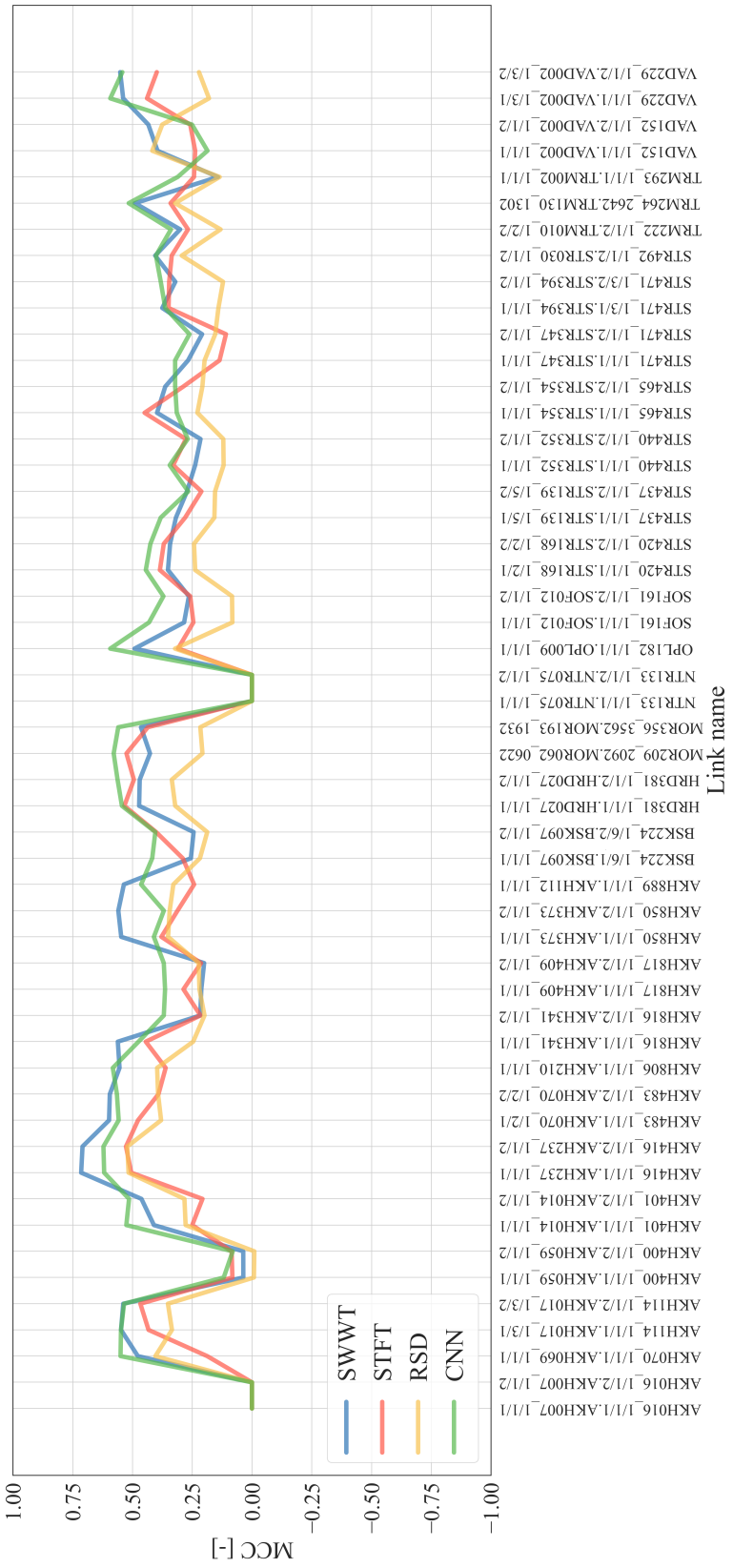

Figure C.2: Results from the four methods for all CMLs for rain classification

Appendix D. Python code: SWWT

```
        def DWT(window_frame):
            coeffs = pywt.wavedec(window_frame,'haar',level=8)
            # level=4 for the snow detection version
            cA8,cD8,cD7,cD6,cD5,cD4,cD3,cD2,cD1=coeffs
            # cD1,cD2,cD3,cD4 used in snow method
            return np.concatenate((cD4,cD5,cD6,cD7,cD8))

        def sliding_window_DWT(signal):
            window_size=2**8 # Power of 8 because of 8-level
            ↪ decomposition
            nWindows = len(signal) # Number of windows
            # Empty array for the results
            result = np.empty((nWindows,2**5-1)) # 31 detail coefficients
            x=np.empty(window_size//2,dtype=object) # Center the window
            ↪ frame
            vec=np.concatenate((x,signal,x))
            # Sliding the window over the signal.
            for i in range(nWindows):
                startIdx = i
                endIdx = startIdx + window_size
                tempIdx = slice(startIdx, endIdx) # indexing the window
                result[i] = DWT(vec[tempIdx]) # Store the result in an
                ↪ array
            # Pad the start and ending to keep the same size
            result[0:window_size//2]=np.nan
            result[-window_size//2:]=np.nan
            return result
```

Figure D.1: The two Python functions that were used when extracting wavelet coefficients for the SWWT method. The wavelet coefficients are then given to a Random forest classifier.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway