STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic Year 2021-2022

Save Some For a Rainy Day:

Assessing the Economic Risk of Extreme Rainfall Events

Emma Hellström (23908)

Abstract

Extreme rainfalls cause large economic damage, they are also expected to become more frequent due to climate change. To design efficient climate adaptation policy and insurance pricing, it is necessary to assess which areas and properties are the most at risk. This study utilizes detailed insurance claim data and weather observations from 795 weather stations in Sweden. Many previous studies of extreme rainfall and the damage it brings are event-based. This thesis contributes to the current literature by describing extreme rainfall risk and its sources, based on events from every Swedish municipality between 2011-2020. The main findings are that the number of extreme rainfalls during a year and the daily amount of rainfall, leads to more damages but not larger damages. Furthermore, the result suggests that properties with a simpler type of water connection or no connection seem to have smaller damages. Another finding is that properties connected to a municipality-managed system are more frequently damaged.

Keywords: Extreme Precipitation, Pluvial Flooding, Insurance Data, Damage Assessment
JEL: Q51, Q54

Supervisor: Marion Leroutier
Date submitted: December 5, 2021
Date examined: December 15, 2021
Discussant: Denise Shen
Examiner: Kelly Ragan

Acknowledgements

First and foremost, I would like to thank my supervisor Marion Leroutier for her valuable advice and support throughout the process. I would also like to express my gratitude towards Laura Ni and Kathrin Larsson from If, for providing me with data and sharing their expert knowledge with great patience and clarity. Lastly, I would like to thank my friends and family for their great support. And a special mention to Max Alteg and Ingrid Löfman who went over and beyond by reading drafts of the thesis. Thank you!

Contents

1. Introduction

According to IPCC's Sixth Assessment Report, the number of heavy precipitation events has increased on all continents since 1950. But only in Northern Europe can this also be attributed to anthropogenic climate change with high certainty. If global warming reaches 2 degrees Celsius, IPCC projects an intensification of heavy precipitation events with high confidence compared to 1995-2014 (Masson-Delmotte et al., 2021). Global warming will have many different effects on the earth's climate. This thesis focuses on the anticipated increase of extreme precipitation events in Sweden. It is difficult to attribute single events to a trend like climate change. However, during August 2021 two extreme rainfalls led to flooding (i.e. pluvial flooding) in Sweden. This resulted in more than 250 million SEK in damage for one of the municipality-owned water companies (SVT, 2021a). In addition, the total amount paid out by insurance companies is estimated to be almost 500 million SEK (Svensk Försäkring, 2021c). Fortunately, only material property was damaged during these events, and only economic damage will be considered in this paper. But these events clearly demonstrate that extreme rainfalls can lead to pluvial flooding and cause great destruction. It is therefore central for policymakers and insurance companies to understand when heavy rainfalls cause damage and why. This thesis aims at improving this understanding by investigating which areas and types of properties are the most at risk.

In line with previous research, damage frequency and size are proxied with claim frequency and severity. A claim is made when an insurance policyholder has incurred damage on insured property. The claim severity equals the monetary amount the policyholder receives. I have decomposed total economic risk from extreme rainfalls into these two components. This leads me to my two research questions: (1) What impact does extreme rainfall have on claim frequency? And (2) what impact does extreme rainfall have on claim severity? In addition, sources of extreme rainfall risk connected to the building's characteristics are also investigated. Extreme rainfall risk is analyzed through a framework by Bouwer (2013). This framework identifies three components of extreme weather risk, namely: Exposure, vulnerability and hazard probability.

To better assess the sources of extreme rainfall risk, data provided by the insurance company If is carefully matched with data on the daily amount of rainfall from SMHI. SMHI stands for the Swedish Meteorological and Hydrological Institute and is an expert agency reporting to the Ministry of the Environment. By using Generalized Linear Models (GLM), claim severity and claim frequency is estimated. GLM is commonly used within the insurance sector due to its flexibility. However, only one previous study has estimated extreme precipitation risk with a GLM. I intend to contribute to the current research field by demonstrating the appropriateness of GLMs in this setting. Another contribution is the identification of new important risk sources.

I hypothesize that an additional extreme rainfall over a local area and year will both increase the claim severity and claim frequency. My main results are threefold. Firstly, a statistically significant, robust and positive relationship between the number of claims and the number of extreme rainfalls was found. Which aligns with previous research. This relationship increases in magnitude for stricter definitions of an extreme rainfall event. Secondly, no overall statistically-significant relationship could be found between the number of extreme rainfalls and claim severity. Only the strictest definition (90mm/24 hours) yielded significant and robust results. Unexpectedly, the effect was negative. Which probably can be attributed to several flaws in the model specification. This result does not align with previous research. Thirdly, two sources of vulnerability to incur rainfall-caused damage are identified. Properties with a simpler type of water connection or no connection seem to on average incur smaller damages. While properties connected to a municipality-managed system are estimated to have a higher claim frequency. Unexpectedly, the size of the floor area has neither an effect on the claim frequency nor on the claim severity. Finally, maps of the average and weighted extreme rainfall risks are presented.

The remainder of the paper is structured as follows. Section 2, summarises previous literature and provides some background on the Swedish insurance industry. Section 3, lists the research questions together with the hypotheses and sub-hypotheses. Section 4, provides an overview of the data and how the data sources were matched. Section 5, motivates the choice of the econometric model and the model specifications. The results are presented in section 6, followed by a set of robustness checks. A visual presentation of risk heterogeneity between Swedish municipalities is also included. Section 7, discusses the results and their limitations. Finally, section 8 concludes the thesis.

2. Literature Review and Background

2.1 Climate Change and Extreme Weather Events

Many studies of trends in extreme weather event losses attribute it to climate change (McNamara and Jackson, 2019). But evidence of increased hazard frequency and/or severity due to climate change is often faulty or even lacking (Bouwer, 2013). Therefore, studies proving that hazards have become more severe over time by linking them to larger economic damage are sometimes misleading. Bouwer (2013) argues that an increased natural hazard severity might not stem from climate change, but rather from increased exposure (e.g., higher population density and asset value). With more valuable assets located in a certain area, damages will seem larger even if the frequency and severity of extreme weather events remain constant. In other words, omitting exposure from the analysis can lead to a false upward trend in economic damage caused by natural hazards. Another problem with studies only including meteorological variables when estimating the economic cost of hazards is that changes in vulnerability are not captured. Vulnerability can e.g., be assessed based on the building's age, type of drainage system, if the property has a basement, or its distance to water (Gradeci et al., 2019).

In contrast to exposure, omitting vulnerability from the analysis can both lead to an under- and overestimation of historical costs depending on the geographic area studied (Mechler and Bouwer, 2015).

Extreme precipitation is a particular form of extreme weather. IPCC argues that extreme precipitation already has become more frequent in all land regions (Masson-Delmotte et al., 2021). They also describe pluvial flooding, which is caused by extreme precipitation, as one of the big sources of climate-related risk in northern Europe. However, these statements are questioned by some of the previously mentioned researchers (Mechler et al., 2020). Whether or not extreme precipitation has historically increased due to anthropogenic climate change is clearly controversial. However, most researchers and IPCC both believe that climate change will make extreme weather and extreme precipitation more frequent in the future (Bouwer, 2013). I will therefore not attempt to attribute my findings to climate change since my period of study is too short (10 years). But since many papers equate historical trends in extreme precipitation with climate change, a clarification was deemed necessary, and a risk-assessment useful for future climate adaptation policy.

2.2 Costs of Extreme Weather Events

Losses due to extreme weather events are usually grouped into three categories: direct, indirect, and macroeconomic (Mechler and Bouwer, 2015). Direct losses are often physical and immediate, such as damage to property and loss of life. Indirect losses can be found in the aftermaths of an extreme weather event, such as damage to the local economy e.g. production loss or years of education lost. The final type of damage, macroeconomic, is similar to indirect losses but affects the economy on a higher scale (e.g., country-level). To gain a holistic understanding of damage made by extreme weather events and specifically pluvial flooding all three aspects are needed. However, due to the inherent complexity of analyzing indirect effects, only direct damages are often analyzed. Direct damage can in turn both be economic and non-economic. A research review of loss and damage due to climate change notes that it is more common to discuss both economic and non-economic costs than just one of the types (McNamara and Jackson, 2019). Narrowing down the scope to home insurances and extreme weather events, the most researched type of hazard is flooding (i.e., both fluvial and pluvial flooding). The great majority of this research is made on events in either the US or in Europe (Lucas et al., 2021), which could partly stem from a funding bias. Another probable contributing factor is data availability. Insurance industries in the US and Europe have collected long time series of claims data, which is a popular type of data to use when estimating the economic costs of extreme weather events.

2.3 Pluvial Flooding

Pluvial flooding can be defined as "... flooding that results from rainfall-generated overland flow and ponding before the runoff enters any watercourse, drainage systems or sewer, or cannot enter it because the network is full to capacity." (Falconer et al., 2009). Hence, whether heavy precipitation causes pluvial flooding depends not only on the duration or amount of rain, but also on the vulnerability of the area. Since data on precipitation is more accessible than detailed data on abnormal water-levels due to pluvial flooding, extreme precipitation is often used as a proxy for pluvial flooding (Frame et al., 2020).

The definition of extreme precipitation differs between authors. However, most agree that the relationship between damage (e.g. number of claims) and weather variables (e.g. precipitation) is nonlinear. Most use various forms of tipping point measurements when defining weather-conditions leading to damage (Lyubchich and Gel, 2017). Pastor-Paz et al. (2020), define a day with extreme precipitation as when the daily accumulated rain is within the 95th, 98th or 99th percentile of the daily historical rainfall distribution in that area. By constructing local and yearly thresholds extreme precipitation is relativized, and the emphasis is put on 'extreme'. This definition implies that damages are caused by unusually much rain, and thus areas who often experience heavy rainfall are less vulnerable than other dryer ones. Other authors define absolute conditions for extreme precipitation. These conditions are often combinations of mm of rain and time. It is common to separate between short and intensive rainfalls e.g., (>= 25 mm) during three hours, and longer but less intense rainfalls e.g., (>67mm) during three days (Grahn and Nyberg (2017), Botzen et al. (2010)). As described by the definition of pluvial flooding provided above, the threshold for extreme precipitation should capture scenarios when the drainage systems are insufficient and the redundant water causes damage. This absolute definition either implicitly assumes that all areas have similar drainage capacity or requires good covariates capturing variation in vulnerability.

Extreme precipitation can also cause different types of damages depending on the building's construction, drainage systems and the duration of the rainfall. Shorter (1-10h) and intense rainfalls tend to cause damage on property close to the main sewers in urban areas. Rainfalls that last for several days cause more evenly spread out claims, without a connection to the area's topography (Sörensen and Mobini, 2017). In a study by Spekkers et al. (2015), the type of damage caused by intense precipitation is investigated. They found that the most common types of precipitation-caused damage on residential property are roof- and wall leakage, followed by blocked roof gutters, melting snow and sewer flooding. Sewer flooding caused the most expensive claims. Roof leakage is unlikely to result from pluvial flooding but rather as a result of the rain itself, illustrating that extreme precipitation and pluvial flooding are not synonymous but rather that the former is a prerequisite of the latter. This thesis will both include damages directly caused by extreme precipitation and indirectly

through pluvial flooding. Another important confounding factor when estimating damages, is the amount of rain the day(s) before the extreme precipitation event. Even if a rainfall is extreme it might not cause any damage if the soil and drainage systems were dry the day before (Torgersen et al., 2015).

Bouwer (2013) presents a framework for quantifying economic damage from extreme weather events. The framework consists of three components; exposure, vulnerability and hazard probability. The former two concepts does not only contribute to an accurate estimation of the relationship between hazard severity and climate change as discussed above. They can also in combination with hazard probability identify the magnitude of the damage and which location is damaged the most. Following Bouwer (2013) exposure is defined as "...the presence of people and assets in areas subject to the occurrence of natural hazards". And vulnerability is defined as "...the susceptibility to loss and damage, including the capacity to anticipate, cope with, resist, and respond to impacts". Lastly, hazard probability is the likelihood that a location will be subject to an extreme weather event.

$$Risk = Exposure \times Vulnerability \times Hazard\,probability \tag{1}$$

By decomposing extreme weather risk, information about risk heterogeneity between demographic, geographic and social groups can be obtained. Identifying heterogeneity in extreme weather risk, is central in designing efficient climate adaptation policy. Since the largest marginal benefit can often be achieved by mitigating the risk for the most vulnerable ones (Hsiang et al., 2019).

2.4 Pluvial Flooding in Scandinavia

The still very limited selection of research conducted on pluvial flooding in Scandinavia has primarily been made by researchers from the Centre for Climate and Safety at Karlstad University. In two studies written by researchers from Karlstad University, insurance claim data on residential property from Länsförsäkringar is utilized. These studies argue that private insurance payouts due to flood damage on residential property have steadily increased over the past 25 years in Sweden, but with large yearly variation (Grahn and Nyberg (2017), Grahn and Olsson (2019)). Another finding made is that the majority of flood-related claims were reported during the summer. Similar findings were also found in the Norwegian city Fredrikstad, where 79 percent of flooding compensation during 2006 - 2012 were connected to events that occurred between July and September (Torgersen et al., 2015). One explanation of the skewed distribution of claims is snow. It is snowing in large parts of Scandinavia during the winter months, and this type of precipitation does not translate into pluvial flooding. It is therefore common to only investigate the connection between extreme precipitation and damage during the summer months. Grahn and Olsson (2019) estimated a 4-16 percentage increase in yearly damages caused by pluvial flooding to residential property, given one additional extreme rainfall event per year. However, since they utilized aggregated insurance data,

only 304 observations were used in the estimation. They used >6mm/15min as the threshold for extreme rainfall. In an earlier paper by Grahn and Nyberg (2017) data at the residential building level was used. They defined extreme rainfall based on three criteria, a daily, one over 9 respectively 3 hours. They estimated 42 percent higher claim severity on movable goods if the damage were made by extreme rainfall. Damage to the property's structure was estimated 65 percent larger if it was caused by extreme rainfall. The latest study on pluvial flooding in Sweden was conducted on a single extreme rainfall over the Swedish city Malmö in 2014. Their main finding was that the type of sewer system (separate or combined) is an important determinant of which property is damaged by an extreme precipitation event. But they could not detect any difference in average claim size (Mobini et al., 2021). In contrast to other studies of extreme precipitation in Sweden, could this study (only considering Malmö) not detect any trend in the number of claims related to heavy precipitation.

Residential Property Insurance in Sweden

Residential property insurances often have at least two components. One part insuring the property itself *[Swedish: byggnadsförsäkring]* and the other part insures movable goods *[Swedish: lösöresförsäkring]*. It is very common to purchase both components from the same insurance company, and the policyholder seldom switches insurance provider. When comparing the number of detached houses in 2017 with the number of detached property insurance policies in 2017, the insurance take-up rate is almost 85 percent. However, the definitions of a detached house or a "villa" may differ between the data sources (SCB (2018), Svensk Försäkring (2021b)). In addition, Grahn and Nyberg (2017) assumes that the insurance take-up rate is almost 100 percent since it is a requirement for using the property as collateral in a house mortgage. As of today, residential property insurances in Sweden always cover damage caused by extreme precipitation events equal to or larger than 1mm/minute or 50mm/24 hours. The water must have either entered the building on the ground floor level through e.g. a window or a door or through the sewer system (Konsumenternas (2021)). However, the industry organization Svensk Försäkring believes that Swedish insurance companies will refuse to offer insurance coverage to property in certain areas in the future (SVT, 2021b). This development has already begun in the Netherlands, where approximately 40 percent of the insurance companies no longer offer residential property insurance with severe local precipitation coverage (Botzen, 2010). Since the fundamental rule of insurance is that the damage must be unanticipated, more frequent extreme precipitation events make property insurance increasingly difficult and expensive. Another important consideration is that many properties are connected to the municipality-owned sewer- and water system. Hence, damage inflicted to property due to under-sized sewer systems should be compensated by the municipalities and not the private insurance companies. The economic risk of increasingly severe and frequent extreme precipitation events is shared by insurance companies, property owners and municipalities.

2.5 My Contribution

In the following section, I will explain my contribution to the existing literature on loss and damage from extreme weather events. More specifically, how I am contributing to the literature on flood risk assessment as outlined by Gradeci et al. (2019). In their review of this interdisciplinary research field, they pinpointed access to reliable insurance claim data as the main challenge. They argue that the lack of high-quality data makes it impossible to match damage and hazard with sufficient accuracy. In addition, the lack of a consistent classification of damages makes it difficult to identify which damages have been caused by pluvial flooding versus fluvial flooding. Thus, one contribution of this thesis is the usage of a detailed insurance claim data set. In contrast to similar papers, I have been able to match the insurance data to the weather data rather accurately. Adding on, the claims included in my data set exclusively stem from damages caused by heavy rainfall. These two advantages will allow me to contribute with a more accurately identified relationship between extreme rainfall and economic damage.

In a research review on home insurances against extreme weather events by Lucas et al. (2021), it is stated that approximately a fourth of the published papers stems from Environmental Science. The second most frequent research field is Social Sciences, followed by the category Economics, Econometrics and Finance. Since this is a thesis within economics, the contribution will be focused on climate adaptation policy guidance. In contrast to similar papers only considering meteorological variables (e.g, Haug et al. (2011), I am including covariates capturing heterogeneity in vulnerability, exposure and hazard probability. By identifying sources of vulnerability connected to certain building characteristics or geographical areas, I aim to contribute with valuable insights for future policy work.

My thesis contributes to current literature in three more aspects. Firstly, to my knowledge only one previous paper has analyzed the economic costs of extreme weather using a Generalized Linear Model (GLM), which is very suitable for modeling insurance data. Instead, a common choice of method when modeling claim severity is OLS (Pastor-Paz et al. (2020), Grahn and Nyberg (2017) and Grahn and Nyberg (2014)). By assuming Gamma distributed standard errors which is common practice within the insurance industry, I intend to contribute with an alternative econometric model that better describes the true distribution of the dependent variable (see section 5.1.1.). Secondly, I am including and comparing a set of both absolute thresholds and station-specific thresholds. Previous research has only either used one definition or a set of either absolute or relative thresholds. In my opinion, a comparison between the two types of thresholds will yield valuable insight on which type is preferred and how they differ. Lastly, many previous studies have only investigated damages from extreme rainfall based on a single event. Thus, analyzing properties from every Swedish municipality and controlling for e.g, type of water connection will provide novel insights.

3. Hypotheses

This paper aims to identify sources of extreme rainfall risk and analyze the relationship between hazard and damage. This is achieved by investigating differences across geographical areas and building characteristics. The overall risk has two components, namely how often the damage occurs and how large the damage is when it occurs. Damage frequency is proxied with claim frequency and damage severity is proxied with claim severity. Two research questions are designed to answer how these two components are affected by extreme rainfall events. The sub-hypotheses aim at investigating heterogeneity in extreme rainfall risk. They are based on what previous literature has identified as sources of vulnerability, exposure and hazard probability. The motivation for dividing total risk into these two components is that the frequency and severity can be affected oppositely by the same factor. A house with a low standard is more likely to incur damage if subjected to a hazard. But the size of the damage may be smaller since the building and the content within the building are less expensive. Thus, it is not obvious if the overall riskiness is larger or smaller for a house with low standard versus a house with high standard. Ohlsson and Johansson (2010) also suggests a decomposition of total risk, since claim frequency can often be predicted with larger certainty. Claim severity is usually much harder to predict.

Research question 1: What impact does extreme rainfall have on claim frequency?

$[H1]$: Claim frequency increases with the number of extreme rainfalls.
To accurately test and estimate this relationship is not as trivial as it might seem at a first glance. As concluded by Gradeci et al. (2019), matching damages with the correct weather conditions has been a main challenge in previous literature. I expect to find a statistically significant and positive relationship between these two variables. Furthermore, I also expect this relationship to remain after controlling for heterogeneity in extreme rainfall risk e.g. geographical area, and building characteristics. In previous literature, claim frequency is estimated to increase with 4-16 percent (Grahn and Olsson, 2019) and 24-57 percent (Pastor-Paz et al., 2020) with an additional extreme rainfall. These results are not directly comparable since they both consider different geographical areas and do not utilize the same definition of extreme rainfall. Thus an important robustness check is to test this hypothesis with a set of different extreme rainfall thresholds.

$[H1a]$: Claim frequency increases with the average floor area.
$[H1b]$: Claim frequency is higher if the property has no or a low standard water system.
$[H1c]$: Claim frequency is higher if the property is connected to a municipality-managed water system than if it is connected to a private water system.
$[H1d]$: Claim frequency is lower for a residential property than for holiday homes.

These four hypotheses are designed to disentangle heterogeneity connected to the property's vulnerability. Hypothesis $H1a$ is based on evidence from previous literature. Studies find that damages caused by extreme rainfall can be found on the roof, the ground floor walls and in the basement (Spekkers et al. (2015) Gradeci et al. (2019)). But to claim compensation from a Swedish insurance company, damages must be made by water entering from either the ground floor or the sewer system. Hence, I conclude that the average floor area is the best proxy for capturing vulnerability since it is often proportional to the size of the basement, ground floor wall area and the number of windows and doors.

Sub-hypothesis $H1c$ and $H1d$ build on the assumption that a property has the same type of water system and sewer system. Several studies identify the type of sewer system and sewer capacity as important determinants of property damage (Mobini et al. (2021), Sörensen and Mobini (2017), Spekkers et al. (2015)). Systems with lower capacity or both collect sewers from households and stormwater are found to have higher vulnerability. I hypothesize that properties connected to municipality-managed systems are more vulnerable than properties connected to private systems. An argument is that municipality-managed systems are usually connected to many properties and hence lower sewage capacity is plausible. Another argument is that municipality-managed systems can be combined while this is rarely the case for private systems. Another assumption is that the type of water connection holds information about the property's overall standard. If the property is not connected to any water system or only has access to water during the summer, I assume that they have a lower standard. A lower standard is in turn connected to higher vulnerability. However, if being connected to a water system is a large risk factor then the opposite relationship will be found. I believe that the overall lower standard will overpower a potential risk increase from being connected to a high-quality system. The effect of not having a high standard water system is probably larger for residential homes than holiday homes. Hence, I am both running separate regressions for residential- and holiday homes, and interactions in the main model. Lastly, $H1d$ builds on the same reasoning as $H1b$. I expect that residential property is built more robust than holiday homes and thus has a lower vulnerability leading to a lower claim frequency.

Research question 2: What impact does extreme rainfall have on claim severity?

$[H2]$: Claim severity increases with the number of extreme rainfalls.

$[H2a]$: Claim severity increases with the average floor area.
$[H2b]$: Claim severity is lower if the property has no or a low standard water system.
$[H2c]$: Claim severity is larger for a residential property than for holiday homes.

Given that a property is damaged, the size of the damage depends on the property's vulnerabil-

ity. Previous literature has established that there is a positive and significant relationship between claim size and extreme weather events. Grahn and Nyberg (2017) estimate that damage on movables is on average 42 percent higher if the damage was caused by extreme rainfall, and 92 percent higher if the extreme rainfall damaged the property's structure. Pastor-Paz et al. (2020) also detect a positive relationship between claim size and the number of extreme weather events for all three types of daily thresholds (95th, 98th and 99th). I form the hypothesis that the number of extreme rainfalls during a year increases claim severity. Hence, that the damage function is non-linear and that the marginal damage increases with the number of extreme rainfalls. It seems likely that the vulnerability to incur a water damage increases after an extreme rainfall since it often removes natural obstacles such as sand, gravel and vegetation. This makes it easier for the rainwater to reach the property. There is also evidence that heavy rainfalls saturate the ground which makes an additional extreme rainfall more potent (Sörensen and Mobini (2017), Grahn and Nyberg (2014)). However, claim severity could also decrease with the number of extreme rainfalls. This could be the case if property-owners or municipalities take measures in preventing damage from an additional hazard, as a response to the first one. For example, by digging better ditches or investing in new sewer systems. This would break the assumption of independent observations, which is further discussed in section 5.1.3. These three sub-hypothesis are very similar to the sub-hypothesis in research question 1, and they build on the same literature. In contrast to claim frequency, a property with an overall lower standard is expected to have a lower claim severity. And holiday homes are also expected to have lower claim severity than residential property since it is intuitive that the property's standard and content are less expensive.

4. Data

4.1 Insurance Data

The data has been provided by the insurance company If. It contains information on Swedish structure- and content insurance policies held by private policyholders between 2011 and 2020. There are observations of insured objects located in every Swedish municipality. It contains information on both the type of property insured and claim history. As previously explained, this type of insurance is often divided into two parts. One part insuring the property itself (structure), and one part insuring the movable goods inside (content). In an effort to narrow down the thesis' scope three important limitations have been imposed. (1) Only properties with data on both content- and structure insurance policies will be included. In other words, if the policyholder only has either a content- or structure insurance policy at If, they will be excluded. This limitation only has a marginal effect on the sample size and is done to maintain consistency. This gives each policyholder the same weight which will prove important in the method section. (2) Another limitation imposed is to exclude apartments from the data. Apartments are often located in urban areas where local infrastructure is central in understanding when and how heavy precipitation causes damage (Mobini et al., 2021). Additional support for excluding apartments is that rainfall-caused damage manifests itself very differently. Even

if only the ground floor is damaged in an apartment all residents are affected. Hence the complexity increases in terms of insurance payout and responsibility. Especially when considering rented apartments. The types of properties included in the sample are therefore; single-family detached houses, multi-family detached houses, townhouses and cottages. The properties can both function as residential- and holiday homes. Limitations 1 and 2 together restrict the sample to only owner-owned properties. This is advantageous since if the policyholder also owns the property and the movables inside he or she will claim when the damage is perceived to be of significance, both in relation to the policyholder's living standard and the deductible. It is less straightforward when the policyholder of the structural policy is not living in the property themselves, since they might be either unaware of the damage or less willing to claim since they are not personally affected. By excluding apartments and policyholders with only either a content or a structure insurance policy, I have tried to ensure that all policyholders have homogeneous incentives to claim. (3) The third limitation is that only claims based on rainfall-caused damage will be included. The claim categorization has been made by the insurance company. This will rule out damages made by other types of precipitation such as snow.

4.1.1 Insurance Policy Characteristics

The insurance data can broadly be divided into insurance policy characteristics and building characteristics. The data set provided by the insurance company includes information about how much compensation (in SEK) was paid out following a claim. I inflation-adjusted claim cost to the 2010 price level by using Statistic Sweden's customer price index (SCB, 2021b). I also truncated the claim cost data at the 99.5th percentile. This is common practice when working with insurance data since few but large outlier values are common. The share of policyholders that have claimed compensation for damage caused by extreme rainfalls is very low. The claim frequency per insurance policy and year is less than 0.1 percent. The probability slightly differs between municipalities but the regional variation is largely driven by a few extreme events that caused a lot of pluvial flooding e.g., the flooding of Malmö in 2014. It is also extremely uncommon for a policyholder to claim compensation for several different rain-related damages.

4.1.2 Building Characteristics

Building characteristics are included in the analysis to describe heterogeneity in vulnerability and exposure as suggested by among others Mechler and Bouwer (2015). The average floor area is constructed by dividing the total living area by the number of floors. The data also includes information about the type of water system the property is connected to and if it is a holiday home or a residential building. Lastly, a variable constructed by the insurance company is included. This variable indicates if the property has a certain characteristic that makes it more probable to incur a water damage. I am referring to this variable as the "vulnerability characteristic". I decided to include this factor despite no more information about it can be presented. The main motivation is that it greatly improves the

model's fit. Buildings with a total living area larger than 500 square meters were excluded since those few outlier properties would have skewed the estimation.

The unique identifier in this data set is the policy number. A new policy number is created if a new customer purchases an insurance and vice versa. The property insurances included in this data are often kept for many years. However, when a customer moves or if they switch insurance company the policy number disappears. A customer could also come back to the insurance company which increases the complexity further. Systematic attrition could therefore be a problem. Attrition is systematic when the reason for dropping out is related to the response variable (Wooldridge, 2010). Being damaged by an extreme rainfall could make the policyholder more willing to move away if he/she thinks it is dangerous to stay or difficult in any other way. The policyholder might also switch insurer if they are dissatisfied with the compensation. But it may also be the case that the policyholder is more inclined to stay since he or she must repair the damage, or are very happy with the insurance money they received. In an effort to assess if the data suffers from systematic attrition, I estimated the probability of staying until 2020 conditional on whether the policyholder has claimed or not. I also included station area controls. I found that policyholders that have claimed were a lot more likely to stay until 2020. In addition, when visually comparing the attrition rates in the (1) full data set, (2) station areas where someone has claimed and (3) claiming policyholders, no large differences were found. The great majority were still customers in 2020. Hence, systematic attrition does not seem to be an issue here.

4.2 Meteorological Data

Data on daily accumulated precipitation between 2011 and 2020 was gathered from SMHI's open API (SMHI, 2021). 842 weather stations were active during the period of interest. However, only 462 stations were active during the entire period (2011-2020). The location of SMHI's weather stations can be seen in figure 1 below.

Figure 1: Active Weather Stations (2011-2020)

Since an extreme precipitation event neither affects the entire population nor only one property at a time, the unit of study is based on the area surrounding each weather station (henceforth station

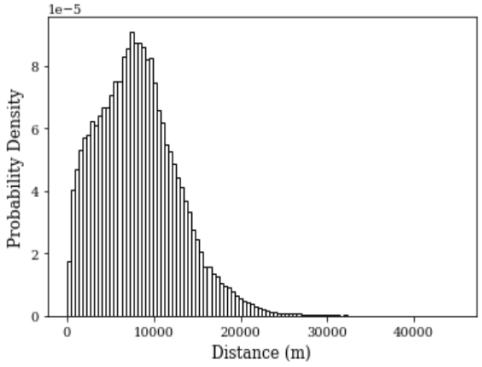
area). I assume that all properties with station x as its closest station are subject to extreme rain when the station is. Determining each property's shortest Haversine distance[1] to a weather station is done with a ball tree nearest neighbor method. The matching is done with SWEREF99 coordinates and performed by the Python library scikit-learn (Pedregosa et al., 2011). The ball tree nearest neighbor method was recommended by Tenkanen and Heikinheimo (2020), due to its memory efficiency. The SWEREF99 coordinate system is harmonized with the European Terrestrial Reference System (Lantmäteriet, 2021). I am matching along two dimensions, north and east. The closest and second closest stations were mapped to each property. The mean and median distances to the closest station were 7.9 km and 7.5 km, respectively. And the mean and median distances to the second closest station were 14.1 km and 13.7 km, respectively. As previously mentioned, almost half of the active weather stations started or stopped reporting weather data during 2011 and 2020. As a consequence, if a station had a missing daily observation during a year, all daily observations during that year were dropped and replaced with data from the second closest station. Only 3 percent of the properties had missing values during the same year in both their closest and second closest stations. These observations were omitted from the sample together with 6 properties where the closest active station was more than 50 km away. After these adjustments, data from 795 stations are utilized. In figure 2a, the distribution of the distance to the nearest active station per property and year is displayed.

During the winter large parts of Sweden are covered with snow. Hence, precipitation during these months is not rain but snow, which does not result in pluvial flooding. Since Sweden is relatively long, winters in northern Sweden are very different from winters in the south. In contrast to previous studies of extreme precipitation in Scandinavia (e.g Grahn and Nyberg (2017), Grahn and Olsson (2019)), I am not limiting the period of interest to the summer months. I find this necessary since the claims in my data set are spread out over the year while Grahn and Nyberg (2017) among others found the vast majority of their claims during the summer. This could stem from different claim classifications, or potentially a larger share of policyholders located in southern Sweden in my data set. To single out which days it was snowing and not raining, additional data on the type of precipitation from SMHI's open API was sourced. This data also spans from 2011 to 2020, but due to poor data quality, it was aggregated per month and on the county level (25 counties). If it snows the majority of days with precipitation in a month and county, then all days with precipitation are set to zero. Hence, the remaining days with precipitation which are later used in the estimations only represent days with rain. For more information about the types of precipitation and which county-month combinations were set to zero, see Appendix. When only considering the daily rainfalls within the 95th percentile per station, it is clear that large daily rainfalls are more frequent during the summer months. But there are also heavy rainfalls (e.g. larger than 50mm) in the year's last months. If I limited my study to the summer months, these heavy rainfalls would not be taken into account while they still in fact are both extreme and rainfalls.

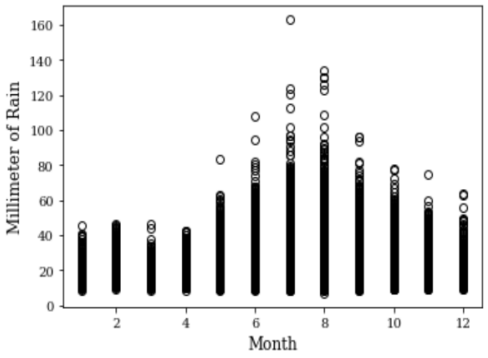
[1]The Haversine formula calculates the distance between two points on a sphere, based on their longitude and latitude values.

Figure 2: Closest Active Stations (2011-2020)

(a) Distance to the Closest Active Station



(b) Extreme rainfalls per Month



4.2.1 Defining Extreme Rainfall Events

As described earlier, previous studies have defined extreme precipitation in many different ways. Since this thesis investigates extreme precipitation from an economic angle, it is important that the definition describes an amount of precipitation capable of causing damage and thus costs. The literature review found that both absolute and relative definitions of extreme precipitation events have been used in previous literature. Since this definition most likely will heavily affect what results and conclusions I can draw from this analysis, a set of both relative and absolute thresholds will be used. Adding on, only daily precipitation thresholds will be used since they are most commonly used in the literature and used by practitioners.

Two absolute thresholds will be employed, 90mm and 50mm of precipitation over 24 hours. The former threshold of 90mm, is defined by SMHI who states that at this level of precipitation damage to vulnerable areas is likely (SMHI, 2012). The latter threshold of 50mm, is the definition of a heavy rainfall *[Swedish: skyfall]* used by the insurance company who provided the data. In addition to the absolute thresholds three relative definitions of an extreme precipitation event are included. Inspired by Pastor-Paz et al. (2020) area-specific precipitation threshold values based on the 95th, 98th and 99th percentiles were calculated. In similarity to their study, my thresholds are also only based on days that had precipitation. But in contrast, my thresholds are not yearly but based on precipitation during the entire period (2011-2020). The reason is that extreme rain does not occur in every area and year, hence some yearly thresholds would be very low in certain areas and not represent rainfalls that cause material damage. See table 1 below for average thresholds and the number of extreme rainfalls per weather station (795 stations) between 2011 and 2020 in Sweden.

Table 1: Average Number of Extreme Precipitation Events per Station

	90mm	**50mm**	**99th**	**98th**	**95th**
Average Threshold (mm)	90.0	50.0	24.7	19.9	13.9
Max Threshold (mm)	90.0	50.0	37.2	30.5	23.0
Min Threshold (mm)	90.0	50.0	14.6	11.8	8.4
# of extreme rainfalls (2011-2020)	17	523	10 525	21 106	51 354

4.2.2 Linking the Independent Variable of Interest to the Dependent Variables

In order to estimate the causal relationship between an extreme weather event and economic damage, the two variables must be linked together over two dimensions, spatial and temporal. As mentioned above, the matching of stations and property was made with a ball-tree nearest neighbor approach. Hence, I am implicitly assuming that the weather station with the closest radius distance to a certain property is that property's closest match. The median distance between the closest active station and property is 7.8km. One assumption is therefore that an extreme precipitation event affects an area with a radius of at least 7.8km. Another assumption is that the shape of the area affected by heavy precipitation is circular, which clearly is a strong assumption. In reality is the size and shape of the affected area are determined by a set of meteorological variables such as temperature, wind speed and the area's topography (SMHI, 2017). Analyzing these factors is outside the scope of this economics thesis. Therefore, I am assuming that the distance is sufficiently short and a ball-tree approach sufficiently accurate to enable a causal analysis of extreme precipitation and economic damage.

The second dimension is time. As seen in the summary statistics table above, the data is aggregated on a yearly level. Therefore it is hard to assess if an extreme precipitation event and a claim that occurred during the same year are related or not. A yearly aggregation was needed since it would not have been practical to perform regressions on the disaggregated data set since it is too large. The insurance data both contain information on when the claim was made and when the policyholder believes the damage was made. In table 2 below the percentage of claims with at least one extreme weather event occurring in the same station area within 1, 5 or 10 days before or after is presented. The three safety margins of 1, 5, and 10 days are used to take reporting errors into account.

Table 2: Mapping Claims to Extreme Precipitation Events

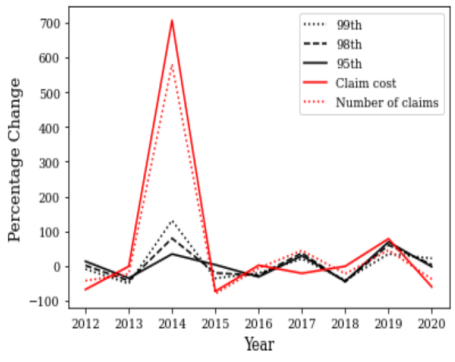
% of claims with at least one extreme precipitation event within...	**90mm**	**50mm**	**99th**	**98th**	**95th**
1 day	0.5	25	49	51	61
5 days	0.5	26	49	61	73
10 days	0.5	27	56	65	80
# of ext. events (2011-2020)	17	523	10 525	21 106	51 354

The range of selected thresholds is very broad, from 90mm/24 hours which was only recorded 17 times by the selected weather stations, to the 95th station-specific threshold, which defined 51 354 station/days as extreme between 2011-2020. The percentage of claims with an extreme weather within 1-10 days also differs a lot between thresholds. Three possible explanations to why the rate is not one hundred percent are: (1) Smaller precipitation events than the defined threshold values cause damage. Hence, the water damage does not stem from an extreme precipitation event, but rather a regular rainfall or a very long period of rain. (2) The closest station does not accurately represent the property's weather. (3) The reporting error is larger than 10 days which seems unlikely since the share of claims with an extreme event within 1 day is not drastically smaller than the same measurement within 10 days.

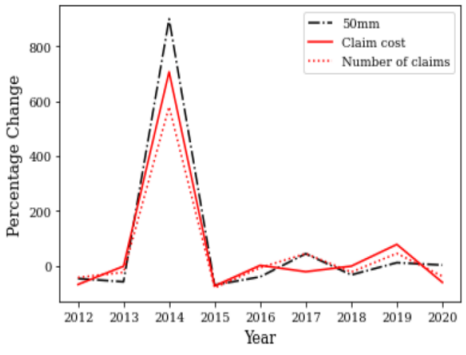
The following two figures illustrate the percentage change in the variable of interest[2] and the dependent variables. The black lines are based on the yearly number of extreme weather events recorded by the active weather stations. The red lines are based on the total number of extreme weather events or total claim cost. The peak in claim cost and the number of claims in 2014 is largely driven by the flooding of Malmö which is the event studied by Sörensen and Mobini (2017). A third of the claims during this year are made in Skåne, while the other two-thirds are geographically spread out. The fact that the number of extreme rainfalls and claim data follow each other so well is not obvious or causal. If an extreme rainfall affects a densely populated area more claims can be expected than if the same hazard affects a more sparsely populated area. However, these plots indicate that there is a positive relationship. The absolute threshold of 50mm seems to best follow the peak in claims and claim cost during 2014, but this is necessarily not causal

Figure 3: Yearly Variation in Variables of Interest

(a) Relative Thresholds



(b) Absolute Threshold



[2]The highest threshold of 90mm of rain per 24 hours is not included since there are no events that intense during some of the years. Hence a percentage change plot is very uninformative.

4.2.3 Summary Statistics

In the following summary statistics table, data per station area and year is presented. The data set spans 10 years and there are on average approximately 644 weather station areas per year.

Table 3: Summary Statistics

Variables	Mean	St.Dev.	Min	Max	N
Dimensions:					
Year			2011	2020	6 443
Station area			102180	99270	6 443
Average Values:					
Share Residential	0.5	0.3	0	1	6 443
Share w. Vuln_ch	0.3	0.1	0	0.7	6 443
Share w. Quality_con	0.8	0.2	0	1	6 443
Average Floor Area	92.1	16.6	15.0	144.3	6 443
Claim Frequency	0.0	0.0	0	0.2	6 443
Average Claim Cost* (truncated)	32.2	484.3	0	30 446	6 443
# Extreme Rainfalls (90mm)	0.0	0.1	0	1	6 443
# Extreme Rainfalls (50mm)	0.0	0.3	0	3	6 443
# Extreme Rainfalls (99th)	1.6	1.3	0	9	6 443
# Extreme Rainfalls (98th)	3.3	1.9	0	11	6 443
# Extreme Rainfalls (95th)	10.0	3.2	0	22	6 443
Average Water System Type			1	3	6 443
Summed Values:					
Number of Insured Properties	464.1	1123.8	1	21 125	6 443

* Average Claim Cost per Property Insured

4.3 External Validity

A key question is if the insurance company's customers can be seen as a representative selection of the population since they are not randomly selected. If they are systematically different from the population this will limit the generalizability of the thesis' findings. As mentioned in the combined background and literature review, almost all Swedish households both have a content- and structure insurance. Hence, there is no strong adverse selection problem connected to purchasing this type of insurance. However, there might be a selection problem connected to who purchases their insurance from this specific company. The insurance data is provided by If. If's market share within private property insurance was in 2019-2020 between 16 and 18 percent (Svensk Försäkring, 2021a). It is therefore important to investigate if the properties within the sample are systematically different from the others in a way that impacts the frequency and/or severity of the damages caused by extreme rainfall.

An argument in favor of strong external validity is the fact the data is geographically diverse. There are several observations from every municipality in Sweden which in part accounts for the most im-

portant factor affecting the number of extreme rainfalls, namely geographic location. One way to at least partially test for this is to compare the median household living area at the municipality-level, between the sample and the population. If the surface that can be damaged is approximately the same size, then the estimated effect of an extreme weather event can be assumed to hold some external validity. Data on the population's living area was collected from SCB (Statistics Sweden). Only observations from households living in one- or two-dwelling buildings which were either tenant-owned [*Swedish: bostadsrätt*] or owner-occupied [*Swedish: ägarrätt*] in 2018 were extracted to match the sample (SCB, 2021b). Only observations on policies insuring residential houses from 2018 were included from the sample to make a comparison possible. The municipality average living area in the sample is 15-60 percent higher than the population's. This implies that the properties in the sample have a higher exposure than the population average in terms of square meters. An on average larger living area could imply that the sample's policyholders are richer than the municipality average. But it could also indicate that they live in less densely populated areas. If they are richer that would imply a lower vulnerability since the house's standard is likely to be higher. But a wealthy policyholder probably also has more expensive content inside the house, which increases its exposure. Hence, it is unclear if this would increase or decrease total risk. If they instead live in less populated areas where house prices in general are lower, then a larger living area must not imply a higher standard but only an increased vulnerability from having a larger roof, floor area etc. It is therefore not entirely clear how this difference affects the external validity. But since the sample also includes a large set of smaller properties and the model controls for average floor area, external validity to the Swedish population can still somewhat be achieved.

5. Method

5.1 Choice of Econometric Model

5.1.1 OLS vs GLM

The number of claims per policyholder and year is often Poisson distributed, which is also the case in this data set. Almost all policyholders will never need to claim compensation for a damage, hence almost all observations will have zero claims. Very few will claim once, and even fewer will claim several times. Out of the observations with claims, the claim size is also not normally distributed. A small share of the claims is much larger than the median claim. Despite that the data is truncated at the 99.5th percentile, it follows a Gamma distribution. It is therefore not appropriate to assume normally distributed errors, which is an assumption that must hold in order to construct accurate confidence intervals and perform hypothesis testing with Ordinary least squares (OLS) (Wooldridge, 2010). Another advantage of using a GLM with a log-link function when estimating the probability of a rare event occurring (i.e, claim frequency) is that the multiplicative model is easier to interpret than an additive model. In addition, it is possible to use the log-link function despite that the depen-

dent variable often is zero, since the log-link function transforms the mean of the dependent variable, which is a linear function of the independent variables (Ohlsson and Johansson, 2010).

There are several examples of similar papers using OLS regressions to analyze claim severity (Pastor-Paz et al. (2020),Grahn and Nyberg (2014) and Grahn and Nyberg (2017)). As previously argued, assuming a positively skewed distribution of the standard errors will likely yield more accurate standard errors, than assuming normality. Hence, in contrast to many other similar studies I will not employ an OLS when estimating claim severity, but instead employ a generalized linear model for both the claim severity- and claim frequency models.

5.1.2 GLM

Generalized linear models (GLM) is a generalization of the ordinary linear model. The response variable (Y_i) follows an exponential dispersion model (EDM), which enables me to assume another distribution of the standard error than the normal distribution. The GLM model is fitted with a maximum likelihood estimation. Broadly, this procedure assumes the optimal values for the parameters by maximizing the probability of conforming with the observed data under the model's restrictions (Hastie et al., 2009). I have consciously chosen to not include any polynomials since overfitting the model would heavily reduce the appropriateness of extrapolating my findings to the population. The models are therefore rather parsimonious. The data is aggregated into cells since this reduces the number of observations and hence improves the calculation speed considerably. The data is grouped per year (t), station area (s) and building characteristics (b). The unique combination of these traits is defined as a cell (i). There are 60 888 cells in the data. In table 4 below a clarification of what a cell contains and consists of is made.

Table 4: The Building Blocks of a Cell (i)

Dimensions	**Cell Averages**	**Cell Sums**
Year	Number of Extreme Rainfalls	Claim Cost
Station Area	Average Floor Area	Number of Claims
Building Type		Number of Policyholders
Type of Water Connection		
Quality Connection		
Vulnerability Characteristic		

These cells are weighted in the estimation with the exposure variable (w_i). In equation 1, the probability distribution of an EDM is presented. The natural parameter θ depends on i, while the dispersion parameter ϕ is constant, both parameters are determined by the assumed distribution of the response variable. The distribution must belong to the exponential family of distributions (e.g,

Normal-, Poisson- or Gamma distribution).

$$f_{Y_i}(y_i;\theta_i,\phi) = exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i,\phi,w_i)\Big\} \tag{2}$$

and

$$\mu_i \doteq E(Y_i) = b'(\theta_i) \tag{3}$$

$$Var(Y_i) = b''(b'^{-1}(\mu_i))\phi/w_i \tag{4}$$

Claim Frequency

Let X_i be the number of claims in cell i with weight w_i, and further assume that $X_i \sim Poisson(\lambda)$. This is a common assumption for claim frequency since most policyholders never claim. When estimating claim frequency, w_i represents the number of insured properties in cell i. Within the insurance industry, w_i is set to sum of shares of a year that the insurance policies have been active. Since the focus of this thesis is not to perform a risk analysis of this specific insurance company's customer portfolio, but rather assess the economic risk of extreme rain for the population, w_i is equal to 1 per property and year regardless of the number of portfolio coverages (see section 4.1 for more information about what type of insurance policies are included in the data set). Since I am more interested in estimating the claim frequency on the policy-year level than on the cell level, the average number of claims per cell is calculated $Z_i = X_i/w_i$. Z_i is a transformation of the original Poisson distributed X_i, therefore Ohlsson and Johansson (2010) refers to Z_i as *relatively Poisson distributed*. The EDM is adapted to a Poisson distribution in equation 2, together with its mean and variance.

$$f_{Z_i}(z_i;\theta_i) = exp\big\{w_i(z_i\theta_i - e^{\theta_i}) + c(z_i,w_i)\big\} \tag{5}$$

and

$$E(Z_i) = e^{\theta_i} \tag{6}$$

$$Var(Z_i) = e^{\theta_i}/w_i \tag{7}$$

Where c(z_i,w_i) = $w_i z_i$log(w_i) - log($w_i z_i$!), $\theta_i = log(\mu_i)$, $\phi = 1$ and b(θ_i) = e^{θ_i}. In the Poisson distribution the dispersion parameter is equal to 1 and therefore is the variance equal to the weighted mean. A common problem with Poisson models is overdispersion. As shown above, one assumption of the Poisson is that the variance is equal to the weighted mean. This assumption is violated if there is variation within each cell i. If each cell is not homogeneous, it is possible that the within-cell variation is higher than the variance of the Poisson distribution i.e. the mean. Overdispersion does not bias the estimates but downward bias the confidence intervals. If the model's deviance is considerably larger than the residual degrees of freedom, the model suffers from overdispersion. And if the opposite is true, the model suffer from underdisperion (Ohlsson and Johansson (2010), Olsson (2002)). I am testing if my Poisson models have this problem in the Method section below.

Claim Severity

A different dispersion parameter is used in the severity estimation, and only policies with claims are included since the response variable must be strictly positive. The weighting variable w_i is also different. Now, w_i represents the number of claims in cell i. The total claim cost per cell i is G_i. I am further assuming that $G_i \sim Gamma(w_i\alpha, \beta)$ since the distribution is positively skewed and the variance increases exponentially with the mean (see eq. 10). In similarity to the frequency estimation above, I am not investing the aggregated response i.e. the number of claims per cell G_i. Rather I am interested in $S_i = G_i/w_i$, which is the average claim cost per claim made in cell i. The density function for claim severity is:

$$f_{S_i}(s_i;\theta_i,\phi) = exp\left\{\frac{(s_i\theta_i + log(-\theta_i)}{\phi/w_i} + c(s_i,\phi,w_i)\right\} \tag{8}$$

and

$$E(S_i) = \alpha/\beta \tag{9}$$

$$Var(S_i) = (\alpha/\beta)^2\phi/w_i \tag{10}$$

Where $c(s_i,\phi,w_i) = \log(w_i s_i/\phi)w_i/\phi$-$\log(s_i)$-$\log\Gamma(w_i/\phi)$, $\theta_i = -1/\mu_i = -1/\frac{\alpha}{\beta}$, $1/\alpha = \phi > 0$ and $b(\theta_i) = -log(-\theta_i)$. The index parameter α and the scale parameter β determine the shape of the Gamma distribution. I am modelling with the Python package Statsmodels (Seabold and Perktold, 2010). In this package, the optimal values of α and β are selected automatically. As guidance in selecting the candidate model for claim severity and claim frequency, AIC (Akaike's Information Criteria) statistics are calculated. This measurement of in-sample prediction error is based on the model's log-likelihood. The log-likelihood represents the logged and summed probability that each independent outcome fits the data. Hence higher probabilities result in a less negative likelihood of the actual data given the model's constraints. To avoid overfitting the model, the AIC penalizes the log-likelihood proportionally to the number of parameters used to estimate the model. The model with the lowest AIC has the highest fit given the number of parameters used in the estimation (Hastie et al., 2009).

5.1.3 GLM Assumptions

Policy independence: The responses i.e. the number of claims or claim costs must be independent of each other. Even though this is one of the basic assumptions of the GLM model it is often violated in practice. This assumption is also violated in this setting since a single extreme rainfall affects several insured properties. Ohlsson and Johansson (2010) argues that when analyzing the insurance risk of an extreme and widespread catastrophe, other types of models are superior to the GLM. But since the extreme precipitation events analyzed in this thesis are not geographically widespread but rather very local, a GLM model is still preferable.

Time independence: The responses i.e. the number of claims or claim cost in year t, must be independent of claims or claim cost in previous years. This assumption would be violated if the policyholder improved the drainage system after the property incurred a rain-related damage. This scenario is not unlikely, however extreme precipitation events are still very rare and thus the willingness to invest in better drainage- or sewer systems could be limited since the risk of another event has historically been small.

Homogeneity: An important feature of the GLM is exposure, which gives every unique cell a certain weight. Hence, every observation in the model represents a unique combination of the policy's characteristics which impacts its riskiness. The homogeneity assumption holds if all policies in the same unique cell have the same level of risk i.e. have the same probability of claiming or claiming equally large. This assumption would be violated if there is an omitted variable describing risk heterogeneity within a unique cell. In a literature review of Gradeci et al. (2019), is the property's material presented as a possibly important explanatory variable. Since I can't access data on material type and it could impact the probability of claiming and claim size, this assumption would be violated. However, this type of problem is hard to avoid since data access is always limited. With the current selection of explanatory variables, I believe that this assumption holds, but these issues should be kept in mind.

5.2 Models

I have constructed three specifications to answer my two research questions. Both claim frequency and claim severity are estimated with these three sets of covariates. Henceforth I am shifting from the cell annotation to a more recognizable format. As previously explained, each cell represents a unique combination of year (t), station area (s) and building characteristics (b). Each unique combination is thus weighted with w_{bst}. In the claim frequency estimation the dependent variable, number of claims made by policies insuring property with characteristics b, in year t and station area s, is represented by N_claims_{bst}. Note that each cell is weighted by the number of insured properties. While when estimating claim severity, the dependent variable is the total claim cost per property

with characteristics b, in year t and station area s. It is represented by Claim_cost$_{bst}$. Note that each cell is weighted with the number of claims made. As previously explained and motivated, I am estimating these GLM models with a log-linkage function. Hence the dependent variables are related to the covariates through the following functions; μ_{bst} = E(N_claims$_{bst}$) and η_{bst} = E(Claim_cost$_{bst}$).

Specification 1

$$ln(\mu_{bst}) = \beta_0 + \beta_1 N\_ext_{bst} + \beta_2 Vuln\_ch_{bst} + \beta_3 Avg\_floorarea_{bst} + \gamma_s + \alpha_t + \epsilon_{bst} \tag{11}$$

$$ln(\eta_{it}) = \beta_0 + \beta_1 N\_ext_{bst} + \beta_2 Vuln\_ch_{bst} + \beta_3 Avg\_floorarea_{bst} + \alpha_t + \epsilon_{bst} \tag{12}$$

N_ext_{bst} represents the number of extreme rainfalls, and $Vuln\_ch_{bst}$ is an indicator variable equal to 1 if the building has a certain characteristic that makes it more vulnerable to water damage. $Avg\_floorarea_{bst}$ represents the property's average floor area. Lastly, γ_s and α_t represents station area and year dummy variables. The number of extreme rainfalls is random both at a local geographical level and over time. It is difficult to predict the exact geographical location of an extreme rainfall and when it will happen. Hence there are no obvious factors impacting both the number of extreme weather events and the number of claims/claim costs within a region and year. However, as described earlier the likelihood of an extreme rainfall depends on the local topography, wind speed and temperature. These meteorological factors likely influence how robust the property is built, since it needs to endure the local weather conditions. Another possible problem could be if locations with more extreme rainfalls also have more rain in general which makes it a less attractive settlement area, leading to less valuable assets located in that region. This is not improbable and hence station area controls are implemented to account for location-specific heterogeneity. However, since the claim severity estimation by design only includes observations with a positive claim cost, and many station areas only claim once station area fixed effects can't be implemented. Thus it is possible that these coefficients are biased, which is further discussed in section 7. The error term is represented by ϵ_{bst}.

Specification 2

The second specification also includes $Quality\_con_{bst}$. Which is an indicator variable equal to one if the property's water connection is of high quality. The main motivation for not including $Quality\_con_{bst}$ in the first specification is that I want to compare the specifications' AIC values to see if the additional control contributes to the overall model. Another reason is that $Quality\_con_{bst}$ and $Avg\_floorarea_{bst}$ are correlated and hence it is informative to only include the former first and then both together.

$$\begin{aligned} ln(\mu_{bst}) =& \beta_0 + \beta_1 N\_ext_{bst} + \beta_2 Vuln\_ch_{bst} + \beta_3 Avg\_floorarea_{bst} + \\ & \beta_4 Quality\_con_{bst} + \gamma_s + \alpha_t + \epsilon_{bst} \end{aligned} \tag{13}$$

$$ln(\eta_{bst}) = \beta_0 + \beta_1 N\_ext_{bst} + \beta_2 Vuln\_ch_{bst} + \beta_3 Avg\_floorarea_{bst} + \beta_4 Quality\_con_{bst} + \alpha_t + \epsilon_{bst} \tag{14}$$

Specification 3

Specification 3 is an additional extension of specification 1. Two additional control variables are included and interacted. $Type\_con_{bts}$ represents what type of water connection the property has. The property can be connected to a municipality-managed water system which is the reference category, or a private water system or have "other" water system. Properties not connected or connected to a very simple system are classified as "other". $Build\_type_{bst}$ is an indicator variable equal to 1 if the building is a residential property, it is otherwise a holiday home. As explained in the Hypothesis section. The different building types in combination with water connection types are likely differently vulnerable. I e.g, expect that the difference in vulnerability between properties with "other" versus municipality-managed system to be larger for residential properties than holiday homes. It is therefore interesting to interact these two variables with each other.

$$ln(\mu_{bst}) = \beta_0 + \beta_1 N\_ext_{bst} + \beta_2 Vuln\_ch_{bst} + \beta_3 Avg\_floorarea_{bst} + \beta_4 Type\_con_{bst} + \beta_5 Build\_type_{bst} + \beta_6 (Type\_con_{bst} \times Build\_type_{bst}) + \gamma_s + \alpha_t + \epsilon_{bst} \tag{15}$$

$$ln(\eta_{bts}) = \beta_0 + \beta_1 N\_ext_{bts} + \beta_2 Vuln\_ch_{bts} + \beta_3 Avg\_floorarea_{bts} + \beta_4 Type\_con_{bts} + \beta_5 Build\_type_{bts} + \beta_6 (Type\_con_{bts} \times Build\_type_{bts}) + \alpha_t + \epsilon_{bts} \tag{16}$$

Standard Errors

The standard errors are clustered at the level of treatment i.e. at the station area level. I deem it necessary to cluster the standard errors since it is likely that the residuals within each station area are correlated. E.g, if one policyholder claims and receives compensation this might have spillover effects on neighboring policyholders.

6. Results

The models are estimated using a log linkage function. I have therefore taken the exponential of the coefficients and adjusted the standard errors accordingly. The coefficients in the following tables should be interpreted as factors. E.g, in table 5 is the claim frequency estimated to increase approximately 3.14 times with an additional extreme rainfall, all else equal. I have decided to use 50mm of rain during 24 hours as my main definition of an extreme rainfall. However, all six models have been estimated using the other four thresholds. The regression results for the other control variables were very similar in magnitude, sign and significance level. I therefore deemed it sufficient to only compare the threshold coefficients, which is done in section 6.3. Furthermore, I will mainly comment on whether or not the two main hypotheses H1 and H2 can be supported. A brief analysis of the sub-hypothesis will also be included.

6.1 Claim Frequency

I am in the following section testing my hypothesis and sub-hypothesis connected to research question 1; *What impact does extreme rainfall have on claim frequency?*. The results in table 5 indicate that the occurrence of an additional extreme rainfall increases the claim frequency by approximately 3.14 times. The estimate is significant at the 1 percent level. This supports my hypothesis and aligns with findings of Pastor-Paz et al. (2020) and Grahn and Olsson (2019). A closer comparison will be conducted below when the same thresholds are used. Shifting the focus to the building characteristics. The effect of average floor area on claim frequency is statistically significant (<0.05) in the first two specifications. However, the magnitude of the effect is in practice zero. The coefficient also loses statistical power when controlling for building type. This is somewhat expected since the size of the property is likely to be correlated with the type of building. Residential property is often larger than holiday homes. Hence, there is no support for H1a. The vulnerability characteristic is significant at the 1 percent level across all specifications. This characteristic is likely to be positively related to

Table 5: Claim Frequency Main Results

	1	2	3
No. of Extreme Rainfalls (50mm) [H1]	4.14***	4.14***	4.13***
	(0.91)	(0.91)	(0.91)
Vulnerability Characteristic	5.45***	5.40***	5.24***
	(0.45)	(0.44)	(0.39)
Average Floor Area [H1a]	1.01***	1.00**	1.00
	(0.00)	(0.00)	(0.00)
Quality Connection [H1b]		1.43**	
		(0.25)	
Private Connection [H1c]			0.79***
			(0.09)
Other Connection			0.52*
			(0.20)
Holiday Home [H1d]			0.46***
			(0.11)
Holiday Home × Private Connection			1.51*
			(0.35)
Holiday Home × Other Connection			1.56
			(0.73)
Station Area Control:	Yes	Yes	Yes
Year Control:	Yes	Yes	Yes
N	60 688	60 688	60 688
AIC	9 719	9 717	9 698

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

the average floor area. Hence, some of the explanatory power within the average floor area might be controlled for with the vulnerability characteristic dummy. However, I expect the average floor area to explain the dependent variable through other channels not controlled for by the other independent

variables. Furthermore, the coefficient for high-quality connection is statistically significant at the 5 percent level, indicating that properties with a higher standard also have higher claim frequency. Which is the opposite to the expected relationship outlined in the sub-hypotheses. In turn, this indicates that the variable may not be an adequate proxy of the building's overall standard. Instead, this indicates that being connected to a high-quality water system increases the risk of being damaged. The third specification in table 5 includes controls for the type of connection, type of property, and the two variables interacted. The base category in "Holiday Home" is residential property. The base category for private water connection and "other" water connection is municipality-managed (MM) connection. The estimated claim frequency for residential property with a private water connection is 79 percent of the claim frequency for residential property connected to a MM network. Or differently put, properties with a MM connection are estimated to have $(1/0.79) \approx 27$ percent higher claim frequency than properties with a private connection. Which confirms my hypothesis. Residential properties with a simpler type of water system or no water systems (coded as "other"), are estimated to also have lower claim frequency than the base category. However, this finding is only indicative, since the estimate is only significant at the 10 percent level. Claim frequency for residential property with a MM connection is estimated to be 117 percent (1/0.46) larger than claim frequency for holiday homes with the same connection type. For an in-depth analysis of these unexpected results, see the Discussion section below. No estimates with a sufficiently high statistical level were found for the two interactions. Specification 3 has the best fit when comparing AIC values. As previously disused, overdispersion might be a problem when assuming a Poisson distribution. When comparing the deviance with the degrees of freedom in each model, I unexpectedly find that my models are underdispersed (deviance/df ≈ 0.1). Olsson (2002) states that underdispersion is very uncommon in practice, but preferable to overdispersion since the standard errors now are overestimated.

Holiday homes and residential properties may differ in several important aspects related to rainfall risk. E.g. where they are located within a station area, its standard, and type of policyholder. To further investigate these differences, the same three specifications are fitted on two subsets of the data based on building type. In table 6 it is evident that the number of extreme rainfalls and claim frequency is positively related for both building types. The estimates are highly significant across the three specifications (<0.01). But the estimated effect is larger in magnitude for residential property (≈ 4.5) versus holiday homes (≈ 1.8). This indicates that holiday homes are more robust against extreme rainfalls than residential property. Another interesting result is that the expected claim frequency is only different per connection type for residential property. Thus the results in main table 5, were largely driven by the residential property data.

Table 6: Claim Frequency Results by Building Type

	Residential Property			**Holiday Homes**		
	1r	2r	3r	1h	2h	3h
No. of Extreme Rainfalls (50mm) [H1]	4.51***	4.50***	4.50***	1.89***	1.89***	1.75**
	(0.96)	(0.96)	(0.95)	(0.42)	(0.42)	(0.49)
Vulnerability Characteristic	5.46***	5.42***	5.35***	3.72***	3.78***	3.18***
	(0.46)	(0.45)	(0.44)	(0.87)	(0.90)	(0.86)
Average Floor Area [H1a]	0.99	0.99	0.99*	1.01	1.00	1.01
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)
Quality Connection [H1b]		1.91*			1.26	
		0.71			(0.32)	
Private Connection [H1c]			0.77**			1.27
			(0.09)			(0.34)
Other Connection			0.46**			0.97
			(0.18)			(0.37)
Station Area Controls:	Yes	Yes	Yes	Yes	Yes	Yes
Year Controls:	Yes	Yes	Yes	Yes	Yes	Yes
N	28 412	28 412	28 412	32 276	32 276	32 276
AIC	7 980	7 980	7 973	1 705	1 705	2 843

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Models denoted "r" are based on residential property data, and models denoted "h" are only based on data from holiday homes.

6.2 Claim Severity

As mentioned in the methods section, these models are only run on a subset of the data including only policy years with one or more claims. Moreover, in compliance with standard practice within the insurance industry is the data truncated based on claim size. The top 99.5th largest claims are removed since they are outlier values. In similarity to the frequency regression above, a log-link is used in the Gamma regressions, and the coefficients and the standard errors in table 7 have been adjusted in the same fashion as in table 5 and 6. In this section, I aim to answer the second research question: *What impact does extreme rainfall have on claim severity?* Main hypothesis 2 is tested. The null hypothesis that the number of extreme rainfalls does not affect claim severity cannot be rejected. Further, I cannot reject the null hypotheses for any of the sub-hypotheses. However, a statistically significant difference between a residential property with a simpler type of water connection versus a municipality-managed system was found. The same relationship could not be found for holiday homes.

Table 7: Claim Severity Main Results

	1	2	3
No. of Extreme Rainfalls (50mm) [H2]	0.88	0.88	0.87
	(0.10)	(0.10)	(0.10)
Vulnerability Characteristic	1.10	1.10	1.14
	(0.14)	(0.14)	(0.15)
Average Floor Area [H2a]	1.00	1.00	1.00*
	(0.00)	(0.00)	(0.00)
Quality Connection [H2b]		1.00	
		(0.26)	
Private Connection			0.80
			(0.11)
Other Connection			0.23***
			(0.11)
Holiday Home [H1c]			1.06
			(0.34)
Holiday Home × Private Connection			1.49
			(0.65)
Holiday Home × Other Connection			4.74***
			(2.83)
Station Area Control:	No	No	No
Year Control:	Yes	Yes	Yes
N	924	924	924
AIC	23 162	23 164	23 158

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

As explained earlier, it is unfortunately not possible to control for station area fixed effects since there is no time dimension for most of the observations in this subset of the data. This could pose

some threats to the identification, such as differences in the local environment which both impact the number of extreme rainfalls and claim cost. The AIC values are very similar indicating that the model's fit did not improve considerably when controlling for connection type. The model with the lowest AIC value is specification 3.

Using the same motivation as in the claim frequency section, the data is divided by building type. The results are presented in table 8 and are rather mixed. The number of extreme weather events does not have a statistically significant effect (at the 5 percent level) on claim severity (i.e., average claim size) for residential property. For holiday homes however, the result indicates a negative relationship. This contradicts findings of Pastor-Paz et al. (2020) and does not support hypothesis 2. A more elaborate discussion can be found in section 7.

Table 8: Claim Severity Results by Building Type

	Residential Property			**Holiday Homes**		
	1r	2r	3r	1h	2h	3h
No. of Extreme Rainfalls (50mm) [H2]	0.97	0.95	1.44*	0.60**	0.58**	0.34
	(0.13)	(0.13)	(0.29)	(0.13)	(0.13)	(0.75)
Vulnerability Characteristic	1.15	1.15	1.26	1.44	1.54	2.41
	(0.16)	(0.16)	(0.22)	(0.42)	(0.46)	(4.15)
Average Floor Area [H2a]	1.00*	1.00**	1.00	1.00	1.00	1.01
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)
Quality Connection [H2b]		4.02***			0.73	
		(1.87)			(0.17)	
Private Connection			0.78			0.22
			(0.17)			(0.83)
Other Connection			0.30			0.77
			(0.25)			(2.24)
Station Area Controls:	No	No	No	No	No	No
Year Controls:	Yes	Yes	Yes	Yes	Yes	Yes
N	782	782	782	142	142	142
AIC	19 693	19 689	19 604	3 467	3 467	3 506

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Models denoted "r" are based on residential property data, and models denoted "h" are only based on data from holiday homes.

6.3 Robustness Checks

The main focus of the robustness checks is on the independent variable of interest. I deem it likely that the definition of an extreme rainfall will heavily impact the effect of an additional extreme rainfall on claim severity and claim frequency. In addition to this comparison, two model diagnostics tests are performed. The first one assess the models' robustness toward influential observations and the second one analyze the residual deviance.

As seen in table 9, a statistically significant relationship at the 1 percent level, between the number of extreme rainfalls and claim frequency is estimated for all thresholds values. Therefore I can conclude that the positive and statistical significant relationship is robust to the definition of an extreme rainfall. The magnitude of the relationship increases with the strictness of the definition, as expected. This is an additional indicator that the causal relationship between extreme rainfalls and economic damage (claims) is captured. Grahn and Olsson (2019) estimated the marginal effect from an additional extreme rainfall to be between 4 and 16 percent. This study also took place in Sweden, but they utilized another definition of an extreme rainfall event, namely 6mm/15 min. Which indicates that daily thresholds capture more destructive extreme rainfall events. Area-specific relative thresholds were used in Pastor-Paz et al. (2020). However, they investigated extreme rainfalls in New Zealand where the rainfalls are much intenser than in Sweden. In addition, they based the thresholds on 1 year of data while I based them on 10 years of data. Despite these central differences, are the magnitudes of the estimated marginal effects very similar. Pastor-Paz et al. (2020) found claim frequency to increase with 24 (95th), 41 (98th) and 57 (99th) percent with an additional extreme rainfall. While the results in table 9 show marginal increases of 18 (95th), 36 (98th) and 65 (99th) percent. Suggesting that the yearly top percentile of daily rainfalls in New Zealand and the top 10-year percentile of daily rainfalls in Sweden have a similar impact on claim frequency. The lowest AIC score and hence the best fit is achieved with model 3, using the 50mm threshold. This model is selected as the claim frequency candidate model. The same threshold comparisons per building type can be found in Appendix, table 13. The effect is larger in magnitude for residential property than for holiday homes regardless of threshold value used.

Table 9: Threshold Comparison - Frequency

	1	2	3
No. of Extreme Rainfalls (90mm)	10.07***	10.10***	10.17***
	(7.26)	(7.28)	(7.37)
AIC	*10 515*	*10 512*	*10 491*
No. of Extreme Rainfalls (50mm)	4.14***	4.14***	4.13***
	(0.91)	(0.91)	(0.91)
AIC	*9 719*	*9 717*	*9 698*
No. of Extreme Rainfalls (99th)	1.65***	1.65***	1.64***
	(0.12)	(0.12)	(0.12)
AIC	*9 768*	*9 765*	*9 746*
No. of Extreme Rainfalls (98th)	1.36***	1.36***	1.35***
	(0.08)	(0.08)	(0.08)
AIC	*10 043*	*10 040*	*10 022*
No. of Extreme Rainfalls (95th)	1.18***	1.18***	1.18***
	(0.04)	(0.04)	(0.04)
AIC	*10 161*	*10 158*	*10 138*

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.
Every coefficient in this table stems from a separate regression.

In contrast to the frequency models, no unanimous strong evidence of a connection between the number of claims and claim cost could be found. Although, the strictest threshold of 90mm and surprisingly the 98th percentile threshold yield statistically significant estimates. Since there is a maximum of one extreme rainfall within an area and year, should the coefficient of an extreme rainfall, defined as 90mm, be interpreted as a dummy variable. Namely, how much larger the estimated average claim size is if there has been an extreme rainfall event during the year. The estimated effect is negative and statistically significant (<0.05), which implies that the average claim cost is approximately (1/0.4) 1.5 times larger if there have been no extreme rainfalls in the area and year. All of these results contradict hypothesis 2, and contrast to findings by Grahn and Nyberg (2017). They found that extreme rainfalls caused approximately 42 percent larger damages than less intense rainfalls. The lowest AIC value was found for specification 3, using the 98th percentile threshold value. This is the claim severity candidate model. The same threshold comparisons per building type can be found in Appendix, table 14. The results are mixed in terms of significance. Statistical significant and slightly negative effects were found for residential property using the two lowest thresholds. A negative effect was also found with the 90mm threshold. However, no significant estimate was found for either building type when controlling for the type of water connection. Hence, the results' robustness can be questioned.

Table 10: Threshold Comparison - Severity

	1	2	3
No. of Extreme Rainfalls (90mm)	0.43**	0.43**	0.39***
	(0.16)	(0.16)	(0.13)
AIC	*23 160*	*23 163*	*23 156*
No. of Extreme Rainfalls (50mm)	0.88	0.88	0.87
	(0.10)	(0.10)	(0.10)
AIC	*23 162*	*23 164*	*23 158*
No. of Extreme Rainfalls (99th)	0.96	0.96	0.97
	(0.04)	(0.04)	(0.04)
AIC	*23 163*	*23 166*	*23 161*
No. of Extreme Rainfalls (98th)	0.92***	0.92***	0.93***
	(0.02)	(0.02)	(0.02)
AIC	*23 151*	*23 153*	*23 149*
No. of Extreme Rainfalls (95th)	0.97*	0.97*	0.97*
	(0.01)	(0.01)	(0.01)
AIC	*23 158*	*23 160*	*23 1534*

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.
Every coefficient in this table stems from a separate regression.

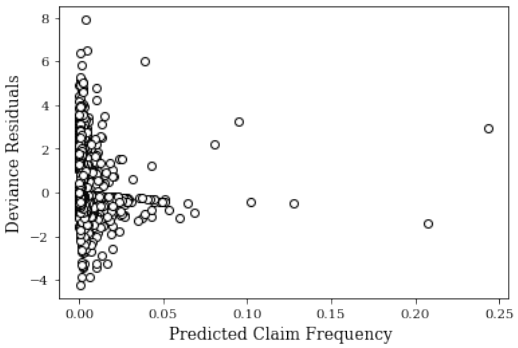
Two additional tests were conducted on the candidate models. A potential threat to the robustness of the findings is influential observations (Olsson, 2002). If an influential observation is deleted or changes slightly it will affect the estimates. Cook's Distance[3] is calculated for the candidate models. A rule of thumb is that observations with a value higher than 1 or a lot higher than the other values are influential observations. Since there are large extreme rainfalls e.g. in Malmö 2014 this could pose a threat to the models' robustness. In the frequency model, only one observation had a distinctively large value. However, its Cook's Distance equaled only 0.004. There were no abnormally large values in the claim severity estimation at all. Hence, the estimations seem robust to influential observations. Another way of assessing the finding's robustness is to plot the deviance residuals against the predicted values (Olsson, 2002). As seen in figure 5a, the deviance in the claim frequency model is almost symmetric which is very positive. An optimal fit would yield small and normally distributed deviances around 0. The model's largest flaw is that it predicts claim frequencies close to 0, while claim frequency estimated by the saturated model is either larger or smaller. It also overestimates the claim frequency slightly when predicting higher claim frequencies. Plausible explanations are that it lacks important covariates, the assumption that claim frequency is Poisson distributed is faulty and/or that the linkage function is unsuitable. I believe that the first reason is the most important one, which is a limitation of the model. However, the model still fits the data well, the median deviance residual is very close to 0. In figure 5b, is the density of observations with negative deviance slightly higher. Indicating that the proposed model overestimates claim severity

[3]Cooks's Distance equals change in the estimated coefficients when removing one observation. The model is thus refitted without one observation. If the change is very large in comparison to the other observations, and/or the full model's variance it is considered influential (Olsson, 2002).

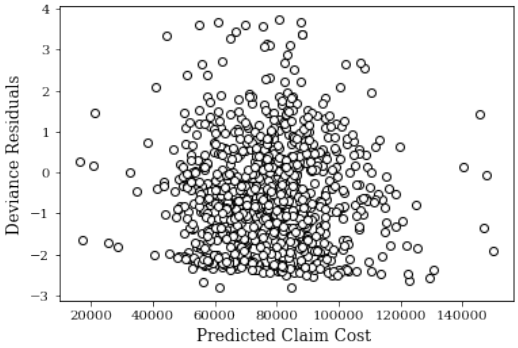
in comparison to the saturated model. The median deviance residual equals -0.73. Apart from that, the Gamma model has a rather good fit.

Figure 4: Deviance Residual Plots

(a) Claim Frequency (50mm)



(b) Claim Severity (98th)



6.4 Extreme Rainfall Risk

By multiplying the fitted values from the claim frequency and claim severity candidate models, the risk factor is constructed. As previously explained, this is the risk of economic damage from extreme rainfall events as outlined by Bouwer (2013). But since the candidate models utilize different thresholds, I have decided to use the 50mm threshold for both the claim severity and claim frequency models. My motivation is that claim frequency is expected to better describe the total risk than claim severity. The hazard probability, vulnerability and exposure vary across Sweden. It is therefore interesting to see how the relative risk differs geographically. I grouped the observations per municipality and calculated the average risk. All values are normalized between 0 and 1. Figure *(a)* displays regional differences in the average extreme rainfall risk. Figure *(b)* displays the relative number of detached houses per municipality (SCB, 2021a). Lastly, figure *(c)* illustrates the weighted risk. The weighted risk is calculated by multiplying the number of detached houses with the average risk level. The highest estimated average risk can be found in the municipalities Grästorp, Vara and Malmö. These results should however be interpreted with caution due to limitations discussed in section 7.1. The general pattern is that municipalities close to the coast have a higher average risk. Especially along the southeast coast. This pattern is strengthened in the weighted risk figure. When comparing figure *(a)* and *(b)*, it seems like the number of properties might be positively related to the average risk. However, it is also possible that the causal channel is not running through "property density" but rather the distance to water. This is further discussed in next section.

Figure 5: Municipality Differences

(a) Average Risk (b) No. of Detached Houses (c) Weighted Risk

7. Discussion

The results broadly support main hypothesis 1 and not main hypothesis 2. The number of extreme rainfalls seems to be positively related to claim frequency. All thresholds were sufficiently high to capture rainfalls so extreme that they caused economic damage. Employing the framework presented by Bouwer (2013), hazard probability is captured by the number of extreme rainfalls. Exposure is described by the weight variable, i.e. the number of policyholders. The relationship between claim frequency and the number of extreme events is therefore not biased by changes in exposure. Which has been a problem in previous research according to e.g. Bouwer (2013). The third factor, vulnerability, is captured by building characteristics such as connection type, building type, and average floor area. Several sources of vulnerability were discovered. The results suggest that residential properties (RP) with a private connection or a simpler type of water system have lower claim frequency than RP with a municipality-managed (MM) water system. The risk of being connected to a MM system is likely to be higher in municipalities with under-sized sewer systems. However, it is not unlikely that properties connected to a MM water system are different in other aspects too.

It is possible that these properties also are more often located in densely populated areas. Hence, a contributing factor to higher vulnerability could be the level of urbanization. But since station area controls are used, and the median diameter is only approximately 16km the level of urbanization is already largely taken into account. Another finding is that holiday homes have a lower claim frequency than RP, despite having a lower standard. This result is unexpected but could in part be explained by the fact that holiday homes are often located in more sparsely populated areas than RP. While this is certainly not the only explanation it is still a plausible one. An improvement of the model would hence be to control for the level of urbanization at the property's location. Many properties are also located along the coasts, as seen in figure 6b. It is possible that RPs are more

often located along the coasts within a station area than holiday homes. However, the opposite could also be true. Regardless, it would be very advantageous to control for these two factors despite that station area controls are already implemented. Another potential explanation for the relatively higher vulnerability of RP is that water damages may be harder to detect if the policyholder is not at the property during the event. Assuming this is the case, damages to holiday homes may sometimes go unnoticed since the water has dried up before the policyholder visits.

The result does not support that there is a connection between an additional extreme rainfall and claim cost. Admittedly, the estimated coefficient using the 98th percentile was statistically significant at the 1 percent level. However, since neither a significant effect was found using the 95th nor the 99th threshold, the result is not very robust. Although the Cook's distance plot did not indicate a problem with influential observations, the significance is likely to disappear if another year of data is included. A reason why no significant results were found is that only extreme rainfalls before the damage contribute to claim severity. Since the number of extreme rainfalls includes hazards both before and after damage the effect is probably diluted. Another improvement to better capture the intended effect would have been to only include a dummy variable for the years with at least one extreme rainfall in the local area. This would not be informative for the lower thresholds, but still valuable if a sufficiently strict definition is used. Despite lacking robustness, the results from the model with a 98th percentile threshold, suggest that the average claim cost slightly decreases with the number of rainfalls. This indicates that there is a learning effect present. Property owners and/or the municipality are thus more prepared and less vulnerable if hit again by an extreme rainfall in the same year. However, these results are not robust. The same set of relative thresholds were used in Pastor-Paz et al. (2020). But their study investigated extreme rainfalls in New Zealand. They found a statistically significant and positive effect between the number of extreme rainfalls and claim cost. However, both the building's robustness and the size of extreme rainfalls are likely to differ a lot between Sweden and New Zealand. The yearly average top percentile values for the 95th, 98th and 99th percentile were 36, 52 and 66 respectively. In comparison, the same thresholds based on 10 years of data from Sweden were only 14, 20 and 25. Disregarding other differences, this indicates that the hazards defined by the percentile thresholds are too low to distinguish rainfalls so extreme that an additional one causes more extreme damage.

Extreme rainfalls defined as 90mm, seem however sufficiently large. These events are so rare that the number of extreme rainfalls can be interpreted as a dummy variable. The results suggest that the average claim cost in an area subject to an extreme rainfall is lower. The negative effect is unexpected. Since only 17 days with rainfalls larger or equal to 90mm were recorded during this 10 year period by 795 stations it is possible that this result is statistical by chance. But if I give this result the benefit of the doubt and interpret it causally, other aspects of rainfall than mm of rain during 24 hours seem to cause damage. A large share of the claims is made despite no extreme rainfalls in the

station area within 5 days as demonstrated in table 2. This indicates that e.g, the rainfall's duration may be important as suggested in the literature review.

Another flaw of this model is the inability to control for station area-specific characteristics. Hence, it is plausible that areas affected by these extreme rainfalls are not entirely random and thus they are better prepared. This would explain why the estimated effect is negative. The only source of heterogeneity in claim cost vulnerability was found for connection type. Properties with a simpler or no water connection are also expected to have a lower average claim cost. This relationship can be explained in two ways. Firstly, properties with "other" connection types have an overall lower standard. However, no support was found for this hypothesis as seen in table 7. Secondly, the type of damage caused by malfunctioning sewer systems is a lot more expensive than other types of damages, which is probable. Mobini et al. (2021) found that the number of damages were higher for certain types of municipality-managed sewer systems, while no difference was found for the average damage size. In similarity, I found differences in claim frequency but not in claim severity between MM and private sewer systems. Which indicates that the type of damage is similar if it occurs.

7.1 Limitations

As briefly addressed throughout the thesis there are several limitations. Limitations connected to the data are twofold. Firstly, the data does not include claims that were made but did not receive any compensation. If this is common, then the treatment effect on claim frequency might be underestimated. While the treatment effect on claim severity is overestimated. However, I assume that the insurance company has correctly assessed which properties have been damaged by rainfall. A second limitation of the data is that damages caused by malfunctioning sewer systems managed by municipalities should be compensated by the municipalities. Hence, the insurance company claims compensation in turn from the municipality. If they receive money from the municipality the claim is removed from the data set. But since properties connected to a municipality-managed system are still estimated to have a higher claim frequency, this risk transfer is not perfect. Although, the actual risk of being connected to a municipality-managed system is likely underestimated in the models.

I concluded in the external validity section that the results could be generalized to the Swedish population. This was partly due to an argument based on the average house size. However, the results show that the average floor area does not impact extreme rainfall risk. Thus, this limits the external validity of the findings since I cannot assess if there are any meaningful differences between the sample and the population. Hence, figure 6 should be interpreted with caution. Lastly, the Poisson models suffer from underdispersion which is very rare and therefore suspect. The models struggle to fit observations with low-frequency levels, which is a large share of the observations and naturally limits the result's credibility. Nonetheless, the median deviance residual is close to zero.

8. Conclusion

Extreme rainfall frequency and severity are expected to increase as a result of climate change. But the exact location and timing of an extreme rainfall are hard to predict. Hence, it is of interest for insurance companies and policy-makers to identify sources of extreme rainfall risk. By identifying these sources, risk mitigation policies and insurance pricing models can be designed more efficiently. Total extreme rainfall risk is divided into a frequency and a severity component. These components are estimated using GLMs. Since the GLM framework does not require normally distributed errors it is more preferable to the commonly used OLS. I thereby test if the number of extreme rainfalls and building characteristics affect claim severity and/or claim frequency. One of the main findings is a positive and highly significant relationship between the number of claims and the number of extreme rainfalls. The magnitude of this relationship is heavily dependent on the definition of an extreme rainfall. Hence, I can conclude that not only the number of events but also the amount of rain during 24 hours affect claim frequency.

To my knowledge, no prior studies have investigated differences in claim frequency between properties connected to municipality-managed and private sewer systems. I am also unaware of papers investigating this difference for claim severity. Despite that, the type of sewer system is identified as a main risk factor in previous event-studies (Sörensen and Mobini (2017), Mobini et al. (2021), Spekkers et al. (2015)). Ultimately, this thesis provides evidence that being connected to a municipality-managed sewer system increases claim frequency by almost 27 percent, assuming that the type of sewer- and water system are the same. However, I was not able to rule out the possibility that other omitted factors such as property density or distance to water bias this estimate. By including better controls for the local environment e.g. within a station area, more persuasive findings would have been attained. This is therefore an important avenue for future research. Despite these concerns, station area controls are imposed and a positive effect makes empirical sense. I therefore deem it likely that this relationship exists. Hence, a proposal for future policy-making is to invest in municipality-managed sewer systems to mitigate this type of risk.

The second main finding is that no robust relationship between an additional extreme rainfall and claim severity was detected. However, very extreme rainfalls (90mm) are estimated to have a negative effect on claim size. This result contradicts previous research. But since it is not possible to control for station area-specific characteristics, a possible explanation is that areas affected by these events are more prepared and hence less vulnerable. Another possible conclusion is that the size of the damage is determined by other factors than the number of extreme rainfalls or mm of rain during 24 hours. Lastly, a lower average claim size is found for properties with only a simple- or no sewer connection. Confirming that the risk of damage and the damage size depends on the type of sewer. Whilst other intuitive building characteristics such as the floor area do not have an effect.

References

Botzen, W. J., van den Bergh, J. C., and Bouwer, L. M. (2010). Climate change and increased risk for the insurance sector: A global perspective and an assessment for the Netherlands. *Natural Hazards*, 52(3):577–598.

Bouwer, L. M. (2013). Projections of Future Extreme Weather Losses Under Changes in Climate and Exposure. *Risk Analysis*, 33(5):915–930.

Falconer, R., Cobby, D., Smyth, P., Astle, G., Dent, J., and Golding, B. (2009). Pluvial flooding: New approaches in flood warning, mapping and risk management. 2(3):198–208.

Frame, D. J., Rosier, S. M., Noy, I., Harrington, L. J., Carey-Smith, T., Sparrow, S. N., Stone, D. A., and Dean, S. M. (2020). Climate change attribution and the economic costs of extreme weather events: a study on damages from extreme rainfall and drought. *Climatic Change*, 162(2):781–797.

Gradeci, K., Labonnote, N., Sivertsen, E., and Time, B. (2019). The use of insurance data in the analysis of Surface Water Flood events – A systematic review. *Journal of Hydrology*, 568(October 2018):194–206.

Grahn, T. and Nyberg, L. (2017). Assessment of pluvial flood exposure and vulnerability of residential areas. *International Journal of Disaster Risk Reduction*, 21(December 2016):367–375.

Grahn, T. and Nyberg, R. (2014). Damage assessment of lake floods: Insured damage to private property during two lake floods in Sweden 2000/2001. *International Journal of Disaster Risk Reduction*, 10(PA):305–314.

Grahn, T. and Olsson, J. (2019). Insured flood damage in Sweden, 1987–2013. *Journal of Flood Risk Management*, 12(3):1–10.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*.

Haug, O., Dimakos, X., Vårdal, J., Aldrin, M., and Meze-Hausken, E. (2011). Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, pages 1 – 20.

Hsiang, S., Oliva, P., and Walker, R. (2019). The Distribution of Environmental Damages. *Review of Environmental Economics and Policy*, 13(1):83–103.

Konsumenternas (2021). Drabbad av översvämning – anmäl din skada och så gäller din försäkring. Retrieved 28-11-2021 from, `https://www.konsumenternas.se/arkiv---nyheter-bloggar-och-poddar/nyheter/2021/augusti/drabbad-av-oversvamning--sa-galler-din-forsakring/`.

Lantmäteriet (2021). Sweref 99. Retrieved 24-11-2021 from, `https://www.lantmateriet.se/sv/Kartor-och-geografisk-information/gps-geodesi-och-swepos/Referenssystem/Tredimensionella-system/SWEREF-99/`.

Lucas, C. H., Booth, K. I., and Garcia, C. (2021). Insuring homes against extreme weather events: a systematic review of the research. *Climatic Change*, 165(3-4).

Lyubchich, V. and Gel, Y. R. (2017). Can we weather proof our insurance? *Environmetrics*, 28(2):1–10.

Masson-Delmotte, V., P., Z., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and (eds.), Z. B. (2021). Ipcc, 2021: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change.

McNamara, K. E. and Jackson, G. (2019). Loss and damage: A review of the literature and directions for future research. *Wiley Interdisciplinary Reviews: Climate Change*, 10(2):1–16.

Mechler, R. and Bouwer, L. M. (2015). Understanding trends and projections of disaster losses and climate change: is vulnerability the missing link? *Climatic Change*, 133(1):23–35.

Mechler, R., Singh, C., and Ebi, K. (2020). Loss and damage and limits to adaptation: recent ipcc insights and implications for climate science and policy. *Sustain Sci*, 15:1245–1251.

Mobini, S., Nilsson, E., Persson, A., Becker, P., and Larsson, R. (2021). Analysis of pluvial flood damage costs in residential buildings – A case study in Malmö. *International Journal of Disaster Risk Reduction*, 62(April).

Ohlsson, E. and Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models.*

Olsson, U. (2002). *Generalized Linear Models An Applied Approach.*

Pastor-Paz, J., Noy, I., Sin, I., Sood, A., Fleming-Munoz, D., and Owen, S. (2020). Projecting the effect of climate change on residential property damages caused by extreme weather events. *Journal of Environmental Management*, 276(September):111012.

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*, 12:2825–2830.

SCB (2018). Drygt 4,8 miljoner bostäder i sverige. Retrieved 28-11-2021 from, `https://www.scb.se/hitta-statistik/statistik-efter-amne/boende-byggande-och-bebyggelse/bostadsbyggande-och-ombyggnad/bostadsbestand/pong/statistiknyhet/bostadsbestandet-2017-12-31/`.

SCB (2021a). Antal och andel hushåll efter region, boendeform och hushållets storlek. År 2012 - 2020. Retrieved 11-11-2021 from, `https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__HE__HE0111/HushallT26/`.

SCB (2021b). Consumer price index (cpi). Retrieved 25-10-2021 from, https://www.scb.se/en/finding-statistics/statistics-by-subject-area/prices-and-consumption/consumer-price-index/consumer-price-index-cpi/.

SCB (2021c). Counties and municipalities in numerical order. Retrieved 11-11-2021 from, https://www.scb.se/en/finding-statistics/regional-statistics/regional-divisions/counties-and-municipalities/counties-and-municipalities-in-numerical-order/.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

SMHI (2012). Extrem nederbörd. Retrieved 25-10-2021 from, https://www.smhi.se/kunskapsbanken/meteorologi/regn/extrem-nederbord-1.23060.

SMHI (2017). Nederbörd. Retrieved 25-10-2021 from, https://www.smhi.se/kunskapsbanken/meteorologi/nederbord-1.361.

SMHI (2021). Se.acmf meteorologiska observationer - nederbörd, summa 1 dygn. Retrieved 25-10-2021 from, https://www.smhi.se/data/utforskaren-oppna-data/se-acmf-meteorologiska-observationer-nederbord-summa-1-dygn.

Sörensen, J. and Mobini, S. (2017). Pluvial, urban flood mechanisms and characteristics – Assessment based on insurance claims. *Journal of Hydrology*, 555:51–67.

Spekkers, M. H., Clemens, F. H., and Ten Veldhuis, J. A. (2015). On the occurrence of rainstorm damage based on home insurance and weather data. *Natural Hazards and Earth System Sciences*, 15(2):261–272.

Svensk Försäkring (2021a). Försäkringsmarknaden kvartal 3, 2021. Retrieved 24-11-2021 from, https://www.svenskforsakring.se/statistik/marknadsstatistik/forsakringsmarknaden/.

Svensk Försäkring (2021b). Kvartal statistik: Bestånd. Retrieved 28-11-2021 from, https://www.svenskforsakring.se/statistik/statistikdatabas/.

Svensk Försäkring (2021c). Nästan en halv miljard i beräknat skadebelopp i gävleborg och dalarna. Retrieved 02-12-2021 from, https://www.svenskforsakring.se/aktuellt/nyheter/2021/nastan-en-halv-miljard-i-beraknat-skadebelopp-i-gavleborg-och-dalarna/.

SVT (2021a). En kvarts miljard kronor – gävle kommuns kostnad för översvämningarna. Retrieved 02-12-2021 from, https://www.svt.se/nyheter/lokalt/gavleborg/en-kvarts-miljard-kommer-oversvamningarna-kosta-gavle-kommun.

SVT (2021b). Risk att du blir utan försäkringsskydd om du köper fel hus. Retrieved 28-11-2021 from, https://www.svt.se/nyheter/inrikes/har-skulle-jag-definitivt-inte-kopa-hus-skatteintakterna-styr-risk-att-du-blir-utan-forsakringsskydd-om-du-koper-fel-hus.

Tenkanen, H. and Heikinheimo, V. (2020). *Automating GIS Pro*, chapter Nearest neighbor analysis with large datasets. Department of Geosciences and Geography, University of Helsinki. Retrieved 24-11-2021 from, `https://autogis-site.readthedocs.io/en/2019/notebooks/L3/nearest-neighbor-faster.html`.

Torgersen, G., Bjerkholt, J. T., Kvaal, K., and Lindholm, O. G. (2015). Correlation between extreme rainfall and insurance claims due to urban flooding – Case study fredrikstad, Norway. *Journal of Urban and Environmental Engineering*, 9(2):127–138.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, volume 1 of *MIT Press Books*. The MIT Press, 2 edition.

Appendix

Table 11: Type of Precipitation

Coded as..	**Types of Precipitation Found in SMHI's Data Set**	**English Translation**
Rain	Snöblandat regn	Snow and rain mixed
	Byar av snöblandat regn	Flurries of snow and rain mixed
	Regn	Rain
	Duggregn	Drizzle
	Regnskurar	Rainfalls
	Småhagel	Small hail
Snow	Snöhagel	Snow hail
	Kornsnö	Snow grains
	Snowfall	Snowfall
	Snöbyar	Flurries with snowfall
	Isnålar	Type of ice hail
	Ishagel	Ice hail
	Underkyld nederbörd	Freezing Rain

Table 12: Months and Counties With Snow

Months When it Snowed the Majority of days with Precipitation	**Counties**
No months with snow	7,8,9,10,12,13,14
February	6
January, February	1,3,4,5,17,18,19
January, February, March and December	20,21,22
January, February, March, November and December	23,24
January, February, March, April, November and December	25

The county code used can be found on SCB's website (SCB, 2021c). Unsurprisingly, counties in southern Sweden have fewer months when it is snowing the majority of days with precipitation. Additionally, counties located in the west also tend to have more months coded as "snow".

Regressions on holiday homes with the 90mm threshold are omitted in table 13 and 14. Only one claim was made in the same year and station area as an 90mm rainfall, hence the claim severity and claim frequency estimates reveal customer-specific information and can therefore not be included.

Table 13: Threshold Comparison Frequency

	Residential Property			**Holiday Homes**		
	1r	2r	3r	1h	2h	3h
No. of Extreme Rainfalls (90mm)	9.93***	9.93***	9.97***	–	–	–
	(7.31)	(7.31)	(7.34)	(-)	(-)	(-)
AIC	*8 792*	*8 791*	*8 783*	-	-	-
No. of Extreme Rainfalls (50mm)	4.51***	4.50***	4.50***	1.89***	1.89***	1.75***
	(0.96)	(0.96)	(0.95)	(0.42)	(0.42)	(0.49)
AIC	*7 980*	*7 980*	*7 973*	*1 705*	*1 705*	*2 843*
No. of Extreme Rainfalls (99th)	1.68***	1.68***	1.68***	1.36***	1.35***	1.43***
	(0.13)	(0.13)	(0.13)	(0.09)	(0.09)	(0.12)
AIC	*8 055*	*8 054*	*8 047*	*1 689*	*1 690*	*2 822*
No. of Extreme Rainfalls (98th)	1.38***	1.38***	1.38***	1.17***	1.17***	1.18***
	(0.09)	(0.09)	(0.09)	(0.06)	(0.06)	(0.06)
AIC	*8 317*	*8 316*	*8 310*	*1 700*	*1 701*	*2 837*
No. of Extreme Rainfalls (95th)	1.19***	1.19***	1.19***	1.07**	1.07**	1.07**
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)
AIC	*8 437*	*8 436*	*8 429*	*1 707*	*1 708*	*2 845*

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 14: Threshold Comparison Severity

	Residential Property			**Holiday Homes**		
	1r	2r	3r	1h	2h	3h
No. of Extreme Rainfalls (90mm)	0.31***	0.31***	0.47	–	–	–
	(0.11)	(0.10)	(1.68)	(-)	(-)	(-)
AIC	*19 688*	*19 684*	*19 610*	-	-	-
No. of Extreme Rainfalls (50mm)	0.97	0.95	1.44*	0.60**	0.58**	0.34
	(0.13)	(0.13)	(0.29)	(0.13)	(0.13)	(0.75)
AIC	*19 693*	*19 689*	*19 604*	*3 467*	*3 467*	*3 506*
No. of Extreme Rainfalls (99th)	0.98	0.98	1.01	0.97	0.97	0.84
	(0.04)	(0.04)	(0.07)	(0.08)	(0.08)	(0.25)
AIC	*19 693*	*19 689*	*19 610*	*3 472*	*3 474*	*3 511*
No. of Extreme Rainfalls (98th)	0.94**	0.94**	1.00	0.90	0.91	0.79
	(0.02)	(0.02)	(0.05)	(0.07)	(0.07)	(0.55)
AIC	*19 679*	*19 676*	*19 610*	*3 471*	*3 473*	*3 528*
No. of Extreme Rainfalls (95th)	0.97**	0.97**	1.02	1.02	1.03	1.00
	((0.01))	((0.01))	((0.03)	(0.05)	(0.06)	(0.34)
AIC	*19 679*	*1 9675*	*19 609*	*3 473*	*3 474*	*3 530*

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.