

Natural Catastrophe Models for Weather Related Events and Insurance Applications

By

Kroustallis I. Efstathios

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of PhD in Statistics

Athens, Greece

July 2023



Μοντέλα Φυσικών Καταστροφών Μετεωρολογικών Συμβάντων και Εφαρμογές στον Ασφαλιστικό κλάδο.

Κρουστάλλης Ι. Ευστάθιος

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα

Ιούλιος 2023

DEDICATION

To my professor Nikolaos Frangos.

ACKNOWLEDGEMENTS

Words cannot express my gratitude to my professor Dr. Nikolaos Frangos for his invaluable patience, inspiration, support and feedback. I also could not have undertaken this journey without the valuable input and support of the professors Dr. Dimitrios Karlis and Dr. George Tzougas, who generously provided knowledge and expertise.

Furthermore, I would also like to thank the National Observatory of Athens researchers and especially the Research Director Dr. Vassiliki Kotroni and the Senior Research Scientist Dr. Katerina Papagiannaki, for their valuable input and collaboration. Moreover, I would like to thank the vice-president of the Hellenic Actuarial Society, Ioannis Totos, the senior actuary specialized in NAT-CAT models, Ioanna Messini and my company colleagues for providing business insights and their total support during my research.

Lastly, I would be remiss in not mentioning my family, especially my parents, my brother, my wife Georgia, and my children Konstantina and Yiannis. Their belief in me has kept my spirits and motivation high during this process.

ABSTRACT

The aim of this analysis is to approach Solvency Capital Requirement (SCR) calculations in context with Natural Catastrophic events (NAT CAT) regarding the climate change effect on insurance, under the Solvency II regime. Due to climate change, the frequency and severity of natural catastrophes are expected to increase. Improved climate projections provide evidence that future climate change will increase climate-related extremes (e.g. heat waves, forest fires, heavy precipitation, droughts, top wind speeds and storm surges) in many European regions (European Environment Agency, 2017 and 2020). Through an analytical research on the insurance sector portfolio particularities in Greece, the severe weather-related events recorded for the last thirty years and the frequency and severity characteristics and distributions, we include the weather-related impact in the calibration of the "SCR" models and capture major risk assessment issues on the Greek insurance industry, highlighting weaknesses and proposing solutions to risk identification and recognition difficulties, products pricing implications and the capital requirements calculation models approach. One of the most representative supervised learning approaches, the decision tree based ensemble method called gradient boosting, is implemented for classifying the number of insurance claims caused by storms in Greece. Quantifying the meteorological event-related risks has become increasingly important in general insurance because extreme climate events may trigger excess claims that may harm the insurer's portfolio. On the other hand, it is challenging to model the relation between climate events and claim frequencies since climate events are often not fully recorded. Motivated by the above issues, we introduce a new class of compound frequency models for joint modelling of storm and storm-triggered claim frequencies observed by Greek insurance companies. The proposed models demonstrate the joint distribution of the actual storm and claim processes even though the observed data are incomplete. Additionally, geospatial covariates are included to assess their impacts on the storm and claim frequencies.

ΠΕΡΙΛΗΨΗ

Ο στόχος αυτής της ανάλυσης είναι να προσεγγίσει τους υπολογισμούς της κεφαλαιακής απαίτησης φερεγγυότητας (SCR) στο πλαίσιο των φυσικών καταστροφών (NAT CAT) σχετικά με την επίδραση της κλιματικής αλλαγής στις γενικές ασφαλίσεις, στο πλαίσιο του πλαισίου «Solvency II». Λόγω της κλιματικής αλλαγής, η συχνότητα και η σοβαρότητα των φυσικών καταστροφών αναμένεται να αυξηθούν. Οι πρόσφατες κλιματικές προβλέψεις φανερώνουν ότι η επερχόμενη κλιματική αλλαγή θα αυξήσει τις ακραίες συνθήκες που σχετίζονται με το κλίμα (π.χ. Υψηλές θερμοκρασίες, δασικές πυρκαγιές, ισχυρές βροχοπτώσεις, ξηρασία, θυελλώδεις άνεμοι, καταιγίδες) Μέσα από μια αναλυτική έρευνα για τα ιδιαίτερα χαρακτηριστικά του ασφαλιστικού κλάδου στην Ελλάδα, τα έντονα καιρικά φαινόμενα που καταγράφηκαν τα τελευταία τριάντα χρόνια και τα χαρακτηριστικά των κατανομών συχνότητας και σφοδρότητας, συμπεριλαμβάνουμε τις επιπτώσεις που σχετίζονται με τον μετεωρολογικό κίνδυνο στη βαθμονόμηση μοντέλων υπολογισμού της κεφαλαιακής επάρκειας (SCR) και αποτυπώνονται σημαντικά ζητήματα αξιολόγησης του ρίσκου στον ελληνικό ασφαλιστικό κλάδο, αναδεικνύοντας αδυναμίες και προτείνοντας λύσεις σχετικά με τις δυσκολίες εντοπισμού και αναγνώρισης κινδύνου, τις επιπτώσεις στην τιμολόγηση των προϊόντων και την προσέγγιση των μοντέλων υπολογισμού των κεφαλαιακών απαιτήσεων. Μία από τις πιο αντιπροσωπευτικές προσεγγίσεις μοντέλου «εποπτευόμενης εκμάθησης», βασισμένη σε δενδροδιάγραμματα αποφάσεων μέθοδο, η «gradient boosting», εφαρμόζεται για την ταξινόμηση του πλήθους των ασφαλιστικών συμβάντων που προκαλούνται από καταιγίδες στην Ελλάδα. Η ποσοτικοποίηση των κινδύνων που σχετίζονται με μετεωρολογικά συμβάντα έχει γίνει ολοένα και πιο σημαντική στις γενικές ασφαλίσεις, επειδή τα ακραία καιρικά φαινόμενα μπορεί να προκαλέσουν πολύ μεγάλες κεφαλαιακές απαιτήσεις που μπορεί να βλάψουν το χαρτοφυλάκιο της ασφαλιστικής εταιρείας. Από την άλλη πλευρά, είναι δύσκολο να μοντελοποιηθεί η σχέση μεταξύ μετεωρολογικών συμβάντων και συχνότητας των ασφαλιστικών ζημιών, καθώς τα μετεωρολογικά γεγονότα

συχνά δεν καταγράφονται πλήρως. Με κίνητρο τα παραπάνω ζητήματα, εισάγουμε μια νέα κατηγορία μοντέλων σύνθετων κατανομών συχνότητας για την από κοινού μοντελοποίηση της συχνότητας των ασφαλιστικών αποζημιώσεων και του πλήθους των ακραίων μετεωρολογικών φαινομένων που παρατηρούνται σε χαρτοφυλάκια ελληνικών ασφαλιστικών εταιρειών. Τα προτεινόμενα μοντέλα καταδεικνύουν την από κοινού κατανομή του πλήθους των μετεωρολογικών συμβάντων και των ασφαλιστικών αποζημιώσεων περιλαμβάνοντας γεωχωρικές μεταβλητές για την αξιολόγηση των επιπτώσεών τους.

Contents

1 Introduction **16**

2 Solvency II Regulation and NAT-CAT risk **20**

2.1 NAT-CAT risk - Solvency II - Standard Formula 20

2.1.1 Standard Formula Framework 20

2.1.2 Definition of NAT CAT Parameters used in the SF - EIOPA . . . 29

3 Climate Risk in the European Insurance Sector **31**

3.1 Including Climate Change in the NAT-CAT SCR Calibration. 36

3.2 Weather Related Perils in Greece 37

3.2.1 Temperature-Related Risks 39

3.2.2 Wind-Related Risk 47

3.2.3 Water-Related Risk 54

4 Insurance Data Characteristics - Descriptive Statistics **72**

4.1 Main Distributions Fit for Frequency and Severity. 87

4.1.1 Large Claims Identification 87

4.1.2 Distributions Fit - Frequency 93

4.1.3 Distribution Fit - Severity 101

5 Climate Related Risks Classification 120

5.1 Machine Learning Preliminaries 121

5.1.1 "Supervised" and "Unsupervised" Learning 121

5.1.2 Algorithms and Data-sets 122

5.1.3 Ensemble Learning 122

5.1.4 Decision Trees 123

5.1.5 Boosting 126

5.1.6 Multiclass Classification Case. 128

5.2 Model 130

5.3 Data-set Details and Manipulation 132

5.4 LGBM Model Results 134

6 Weather Related Risk - Frequency and Severity Models 138

6.1 Attritional Claims Modeling 138

6.1.1 Probability of Event Occurrence Estimation 138

6.1.2 Number of Claims Estimation 141

6.1.3 Severity Estimation 142

6.2 Large Claims Modelling 144

6.2.1 Frequency of Large Claims Estimation 144

6.2.2 Severity of Large Claims Estimation 150

6.3 Risk Distribution and Capitals Calculation 153

7 Multivariate Regression Models on Weather Related Events 157

7.1 Regression Models - Data-set description 159

7.2 Methodology 165

7.3 Models 166

7.3.1 Poisson-Poisson model 166

7.3.2 Negative binomial-Poisson model 169

7.3.3 Negative binomial - Negative binomial model 170

7.3.4 Finite mixture models 172

7.4 Inference 174

7.5 Analysis of fitting results 175

8 Concluding Remarks 185

9 Annex I 187

9.1 Main Distributions for Frequency Modeling 187

9.1.1 Negative Binomial Distribution 187

9.1.2 Poisson Distribution 188

9.2 Main Distributions for Claims Severity Modeling 189

9.2.1 Gamma Distribution 189

9.2.2 Pareto Distribution 191

9.2.3 Weibull Distribution 192

9.2.4 Lognormal Distribution 194

9.2.5 Burr Distribution 196

10 Annex II **197**

10.1 Hill's Tail Index Estimator 197

10.2 Akaike Information Criterion - AIC 197

10.3 Baysian Information Criterion - BIC or SBC 198

10.4 Kolmogorov–Smirnov Test 198

10.5 Cramér–von Mises Criterion 199

10.6 Anderson–Darling Test 200

List of Tables

1 Lines of Business- Solvency II table 24

2 Definition of perils and coverage in the SF 27

3 List of countries and perils that are currently included in the SF 28

4 Classification of climate-related hazards. 39

5 Ianos - Property Insurance Losses 52

6 Ianos Insurance Cost - Frequency - Severity 53

7 Ianos - Other Lines of Business 53

8 Claims Count per peril and cause 2019 - Greek insurance companies . . . 67

9 Claim Cost in millions per peril and cause 2019 - Greek insurance companies 68

10 Average Claim Cost per peril and cause 2019 - Greek insurance companies 68

11 Total NAT - CAT Claims estimation for the years 1993-2018 - Hellenic Assosiation of Insurance Companies 69

12 Weather related NAT - CAT events for the years 1993-2018 - Hellenic Assosiation of Insurance Companies 69

13 Athens Frequency and Severity table 73

14 Frequency & Severity of “Other Regions” of Greece 74

15 Claims severity distribution - Athens 75

16 Claims Severity distribution - Other Regions 75

17 Frequency Athens 76

18 Frequency - Other Regions 76

19 Variables Description 77

20 Other variables description 78

21 Meteorological and Geospatial Indices 78

22 Total Premium distribution 80

23 Greece Claims Distribution Percentiles 87

24 Frequency Distributions Gamlss 94

25 Severity Distribution Gamlss 102

26 Severity Athens Table 102

27 Severity Other Table 102

28 Risk Distribution - Simulations Table 154

29 Risk Distribution - Percentiles 155

30 Frequency table of (M, N) for the observations under the "Motors" branch. 161

31 Summary statistics for the explanatory variables 163

32 Summary of goodness-of-fit for various model specifications 175

33 Estimated parameters (and standard errors) for the selected models. . . . 178

34 Fitted frequency table of (M, N) under 5PP model. 182

35 Fitted frequency table of (M, N) under 5NP model. 182

36 Several correlation measures between the observed storm and claim frequencies (M, N) under various fitted models and the empirical data. 183

37 Several correlation measures between the true storm frequency and per-storm claim count $(\tilde{M}, \tilde{N}_t)$ under various fitted models and evaluation methods. 184

List of Figures

1 Contributing factors of the three main components of weather-related risks (Swiss Re, 2020) 26

2 Overview of the PESETA IV methodology. 31

3 Classification of climate-change effects and their relevance for the re/insurance industry - Swiss Re Institute. 33

4 Key observed and projected climate change and impacts for the main biogeographical regions in Europe (EEA, 2017). 35

5 Relative percentage of historical insured losses for geophysical, meteorological, climatological and hydrological events per country. 38

6 Land Cover - Usage - Mati 2018 41

7 Meteorological Conditions - Mati 2018 42

8 Wild Fire 2007 Peloponnese - Flood and mass movements impact. 44

9 Mean Maximum and Minimum Temperature Values for the Time Period 1946–2006 and the Year 2007 for the Summer Months in Greece. 45

10 Projected changes in summer soil moisture. 46

11 European average temperatures between 1850 and 2015 over land area relative to the pre-industrial period. 47

12 Path of Medicane Ianos, from 0000 UTC 15 Sep up to 0000 UTC 21 Sep 2020. Shading denotes the SST (at 1-K intervals), and the black dots the daily SST anomaly higher than 2 K on 15 Sep 2020. Symbols for floods, landslides, and damaging winds are placed over the most affected areas. . 48

13 (a) Total accumulated rainfall over Greece for the period spanning from 1200 UTC 17 Sep to 1200 UTC 19 Sep 2020 (48 h) and (b) zoom over central Greece, for the same period. 51

14 Projected changes in heavy precipitation in winter and summer. 55

15 (a) Frequency distribution (histogram) of R24, at the PC spatial analysis. (b) Distribution(boxplot) of R24 over the binary damage occurrence variable, DO. 57

16 Results for the logistic regression analysis of the probability of damage occurrence (DO) across the PCs of the AMA depending on R24 and controlled for the population. 58

17 (a) Fitted values of DO predicted probability as a function of R24 controlled for the population. (b) ROC curve, highlighting the TPR/FPR coordinates related to the AMA optimal R24 threshold given that the difference between TPR and FPR is maximized. 59

18 Optimal R24 thresholds (mm) above which flood damage is likely to occur, level of flood damage risk (3-color palette), and level of confidence in R24's discriminating performance, at the municipality level. 60

19 (a) Rainfall accumulation map of the storm event over the two catchments, along with basin average time series of rainfall (intensity and accumulation) of (b) Soures and (c) Agia Aikaterini. Note that the start/end time showing on the graphs (b,c) correspond to Nov. 14, 12:00 to Nov. 15, 15:00 UTC. The triangle denotes Elefsina meteorological station. 64

20 Percentage of lightning cases with hail 5cm for a 1971–2000 and b RCP8.5 2071–2100. 65

21 MSG SEVIRI infrared (IR10.8 channel) and visible (VIS) satellite retrieval. 66

22 NAT-CAT Weather related average claims cost. 70

23 NAT-CAT Weather related events occurrence during a year period (observations 1993-2018). 71

24 Construction Materials of buildings insured 81

25 Number of floors of the insured buildings 82

26 Construction Year distribution of the buildings insured 83

27 Item Type distribution 84

28 Sub-Item Type distribution 85

29 Ianos – Claims Distribution Greece 88

30 Large Claims Threshold selection 90

31 Athens Large Claims plot 91

32 Athens Hill plot 91

33 Other Regions Large Claims plot 92

34 Other Regions Hill plot 92

35 Frequency Distributions Fit - Athens 93

36 Frequency Distributions Fit - Other Regions 94

37 Poisson GoF - Athens 95

38 Negative Binomial GoF - Athens 96

39 Poisson Inverse Gaussian GoF - Athens 97

40 Poisson GoF - Other Regions 98

41 Negative Binomial GoF - Other Regions 99

42 Poisson Inverse Gaussian GoF - Other Regions 100

43 Severity Distributions Fit - Athens 101

44 Severity Gamma Distribution Fit - Athens 103

45 Severity Generalized Gamma Distribution Fit - Athens 104

46 Severity Generalized Pareto Distribution Fit - Athens 105

47 Severity Weibull Distribution Fit - Athens 106

48 Severity Gamma Distribution Fit - Other Regions 107

49 Severity Generalized Gamma Distribution Fit - Other Regions 108

50 Severity Generalized Pareto Distribution Fit - Other Regions 109

51 Severity Weibull Distribution Fit - Other Regions 110

52 Severity Distribution Fit - Athens 111

53 Severity Distribution Fit2 - Athens 112

54 Severity Distribution Fit - Other Regions 113

55 Severity Distribution Fit2 - Other Regions 114

56 Flood Vulnerability Map of Greece 115

57 Frequency Map of weather related claims 116

58 Severity Map of weather related claims 117

59 Risk Premium Map of weather related claims 118

60 Basic diagram to illustrate the process of "gradient boosting" process. . . 132

61 Number of Claims Distribution. 133

62 Metric during training. 135

63 The AUC plots for the three classes. 136

64 The most important features and their contribution to the classifying each example to a given class. 137

65 ROC-AUC plot for the above logistic regression model. 140

66 ROC-Threshold selection for the above logistic regression model. 140

67 Flood Large Claims observed 2012-2019. 145

68 Cullen and Frey graph for the observed number of claims. 146

69 Simulated Large Claims Counts for a sixty year period. 147

70 Negative Binomial Distribution Fit for Large Claims related to flood events. 148

71 Histogram of Pareto Distribution Fit for Large Claims severity related to flood events. 150

72 Cumulative Probability of Pareto Distribution Fit for Large Claims severity related to flood events. 151

73 Pareto Distribution Fit for Large Claims severity related to flood events. 153

74 Insurance weather related risk spatial distribution in Greece. 156

75 Distribution of observed postcodes per each explanatory variable. 163

76 Distribution of exposure units per each explanatory variable (red: motorcycle, green: car, blue: truck) 164

77 Partial dependence plots for the selected models (Green: 5PP, Red: 5NP). Left panel: $\hat{\mathbb{E}}[\tilde{M}|\mathbf{x}_{\cdot p} = x]$; Middle panel: $\hat{\mathbb{E}}[\tilde{N}|\mathbf{x}_{\cdot p} = x]$; Right panel: $\hat{\mathbb{E}}[N|\mathbf{x}_{\cdot p} = x]$. 181

Introduction

Weather-related phenomena such as floods, windstorms, hailstorms and wildfires can cause extensive financial losses. Insurance companies can find it challenging to distinguish between the impacts of the underlying meteorological perils for each given event and resulting claims due to various factors including complexity and the business cost involved. The individual impacts of floods, strong winds and hail are often combined and classified into a single cause for an insurance claim as as "storm event". To better price and reserve for future perils, it could be beneficial for insurers to distinguish between these underlying components even when within the same insurance claim. Providing a business solution to tackling this problem is important because by having a better model and understanding of the underlying perils for a claim, an insurer can achieve better pricing and reserving whilst policyholders can pay a fairer price for the policy that they buy. Insurers are expanding traditional techniques of analysis, for example, by incorporating data science techniques, to hopefully improve models, leverage larger sets of data, "unlock" hidden data patterns and free resources within the team, where currently a lot of the data manipulation, data cleansing and experience analysis is conducted by manpower. Given this and the potential large amounts of data involved, using machine learning approaches seems optimal to analyse such large sets of insurance, geographical and meteorological data.

Furthermore, in the case of non-life underwriting risk, climate change may have an impact on the sub-module premium risk. Climate impacts already observed may be priced in the premiums because non-life premiums are generally adjusted on an annual basis. Data used by EIOPA[1] for the calibration of the premium risk standard deviation can therefore be assumed to provide a current view of climate change. The non-life catastrophe risk sub-module is one of the central modules to be impacted by climate change. This sub-module consists of three separate and independent sub-modules dealing with natural catastrophe risk, man-made catastrophe risk and other catastrophe events. The following analysis focuses on the natural catastrophe (NAT CAT) module as climate

[1] European Insurance and Occupational Pensions Authority

change could lead to more frequent and severe events that may lead to higher insured losses for non-life insurers. Also, in the context of the "SCR"[2] calculation module, under the Solvency II regime, Greece is excluded for the flood peril. In the analysis below, insurance portfolio and data characteristics will be revealed, weather related risk assessment conducted and a methodological approach for the insurance weather-related risk estimation proposed.

In recent years, various non-life insurance studies related to climate risks have been carried out. Starting from [64], the focus has been on the increasing frequency and severity of flood risk due to climate change which could potentially increase property insurance claims. With this in mind, they built a "random forest" model with two steps: i. damage level classification and claim number classification; and ii. sub-sampling strategies, using a combination of spatial, topography, land and meteorological data. In [37] one can find a thorough review of recent advancements in weather-related risk modelling and assessment for agricultural and home insurance using various statistical and machine learning methodologies. For instance, [2] uses statistical copulas to identify how, by introducing agricultural revenue insurance contracts in Spain, this would influence insurance premiums compared to yield insurance schemes. Moreover, [9] builds a multiple classification and auto-regressive error model whilst they also perform a model-based cluster analysis to identify what signals would trigger a drought-insurance payout. Also [57] is concerned with the effects of climate tele-connections, such as the El Niño southern oscillation on US government-set premium rate pricing for cotton, using a moment-based maximum entropy modelling approach indicating that private insurers would have the potential to decrease claims paid by 10–15 percent on average. The following article, [51], accounts for the dependence between various weather events such as rainfall or snow-melt, and the count of water-related property insurance claims, aiming for modelling large claims counts within a mixture and extremal mixture modelling statistical framework based on a discretised generalised Pareto distribution. Moreover, they develop a temporal clustering algorithm for the derivation of new explanatory variables in an effort to better comprehend how claims and weather-related events relate. [38], highlights the

[2]Solvency Capital Requirement

importance of using machine learning methods to estimate flood hazards across large geographical territories in a computationally cost effective way, where they build a "random forest" model for classification in order to predict flood probability across the Texas Gulf Coast region using a large National Flood Insurance Program (NFIP) insurance claims geospatial dataset. In another article, [35], the authors give a summary of the methods and challenges faced by weather-based insurance indices, i.e. rainfall, water stress, drought, and remotely sensed vegetation indices suggesting that focusing on past data does not allow for accurately accounting for complex non-stationarities in space and time associated with events of high temperature or rainfall. [32] develops a "random forest" to predict parcel-level and tract-level flood insurance claims within New York State based on a US hydrologic and social demographic dataset for flood exposure mapping. In addition to weather risks affecting spatial and temporal correlation of insurance indices and their covariates, climatic change can also increase the volatility of weather variables, generating non-stationary loss distributions, which are challenging to estimate reliably, adding uncertainty to actuarial rate-making [39]. Furthermore, [11] develops Bayesian belief networks for helping insurers to assess weather index insurance viability in developing countries from a climate viewpoint. Finally, [58] examines how vulnerable EU river line flood insurance systems are to global reinsurance market conditions, i.e. hard and soft markets, and climate change by applying a proprietary flood insurance model.

Extreme weather events affect homes, commercial businesses, crops in agriculture, vehicles, as well as public facilities and infrastructure, often with significant economic consequences. The social and economic impacts of weather-related natural disasters are gaining particular importance as it has been recognized in recent years by various meteorological organizations worldwide that climate change is taking place. As a result, the frequency, intensity and duration of extreme weather events have increased and are expected to intensify even more in the future. Additionally, strong demographic changes are observed, such as increases in populations, infrastructure, goods, but also significant changes in land use. Also of particular concern is the fact that the urbanized population percentage is increasing annually and is expected to increase significantly

in the coming years. The effects of climate and social trends on the risk of severe weather events are evident globally, as average economic losses due to extreme weather conditions have shown a significant increase in recent years to exceed 100 billion in insurance claims worldwide in the year 2021 (Swiss Re Group). In Europe alone, the total economic losses due to extreme weather events exceeded 500 billion Euro over the period 1980-2020 (European Environment Agency). Greece often experiences extreme weather events with quite serious negative effects, due to its geomorphological features, social and economic characteristics, inadequate infrastructure, unregulated construction, and mismanagement of the country's natural resources. The most common phenomena that cause damage in Greece are floods, storms, and strong winds. With regards to insurance financial losses, indicatively, for the year 2020, 28.6% of the loss notices in the "Property" sector are due to extreme weather phenomena, while the amount of compensation amounts to 53.2% of the total compensation of the specific sector (Hellenic Association of Insurance Companies). Due to climate change, it is evident that weather-related risks affect the insurance portfolio to a greater extent recently compared to previous years and thus it becomes increasingly important to understand the relationship between climate events and insurance claims.

This study is based on weather related data from a major insurance company in Greece. The final data-set selected consists of two databases, the exposure database and the claims database. The selection was driven by the quality, accuracy and completeness of the information included.

Solvency II Regulation and NAT-CAT risk

NAT-CAT risk - Solvency II - Standard Formula

Standard Formula Framework

The Solvency II Standard Formula is part of a regulatory framework referred to as Solvency II. One part of the Solvency II framework requires that each insurer calculates its Solvency Capital Requirement (SCR) using a Standard Formula, an internal model, or some combination of the two. The SCR, whether calculated from the Standard Formula or otherwise, is the capital level corresponding to the Value-at-Risk (VaR) of the basic own funds of an insurance or reinsurance undertaking subject to a confidence level of 99.5% over a one-year period. The capital requirements for each peril are aggregated across "CRESTA" zones within each country using a prescribed correlation matrix intended to reflect tail dependency for 1-in-200 year events.

The Solvency II catastrophe underwriting risk charge is determined both net and gross of reinsurance. The charge net of reinsurance is a component of the total underwriting risk charge. The difference between the gross and net risk charges is referred to as the catastrophe risk mitigation effect and is a measure of reinsurance credit risk used in the counter-party default risk calculation.

Catastrophe risk charges consider both natural catastrophes and man-made catastrophes. Natural catastrophes consist of windstorm, flood, earthquake, hail, and subsidence. Storm surge is included with the windstorm peril. Man-made catastrophes include motor, fire, aviation, marine, liability, credit and surety-ship, and terrorism. The risk charges are based on a combination of scenarios and factor methods. Scenario methods (Method 1) are for use in calculating capital requirements for natural catastrophes in the European Economic Area (EEA). Factor methods (Method 2) are used when application of Method 1 may not be appropriate, e.g., for natural catastrophe exposures outside of the European Economic Area (EEA), miscellaneous business, and non-proportional

reinsurance business. For the factor method, the Standard Formula provides selected gross loss ratios by line of business. With either Method 1 or Method 2, insurers calculate their gross catastrophe losses based on either the prescribed scenarios (Method 1) or the prescribed gross loss damage ratios and gross written premiums (Method 2). Insurers then net down the calculated gross loss damages using their own reinsurance programs including any excess covers, any proportional covers, excess retention and reinstatement premiums as well as recoveries from any national pools.

The aggregate capital requirement for non-life (NL) natural catastrophes is determined as:

$$NLCAT = \sqrt{(Method1)^2 + (Method2)^2} \tag{1}$$

Thus, it is assumed that Method 1 and Method 2 apply to independent catastrophe exposures.

The Standard Formula provides tables showing the gross loss damage ratio ($Q_{country}$) for 1-in-200 year catastrophe events, by peril, within each CRESTA zone separately by country. The term "CRESTA" refers to the Catastrophe Risk Evaluating and Standardizing Target Accumulation organization, founded in 1977, which is an independent body established for the technical management of the natural peril coverage. CRESTA determines country-specific zones for the uniform and detailed reporting of exposure data relating to natural perils. The capital requirement for each CRESTA zone and each peril gross of reinsurance is $Q_{country}$ times the aggregated value of geographically weighted total insured value by peril for each country, where the weights are the zone relativity factors for each country provided by the Standard Formula. The capital requirements for each peril are aggregated across CRESTA zones within each country using a prescribed correlation matrix intended to reflect tail dependency for 1-in-200 year events. The correlation coefficients strongly depend on proximity and, for flood, the shape of river networks. The aggregation by-country-by-peril allows application of national pool arrangements.

Then the capital charges are aggregated across countries, by peril and finally across perils using prescribed covariance matrices between countries and between perils.

Insurers must allow for multiple events for wind, flood, and hail as follows. Each insurer evaluates the effect of two scenarios, A (one large event plus a second smaller event, with initial loss A1 and subsequent loss A2) and B (two moderate events, with initial loss B1 and subsequent and loss B2), on a gross basis and then net down for reinsurance. For each peril, the insurer calculates a capital requirement as the maximum of A (net) and B (net). Specifically, the capital charge for each peril is the maximum value of capital charges for scenario A and B net of reinsurance; where the capital charge for scenario A is equivalent to A1%, say 80%, of the capital charge for scenario A1 plus A2%, say 40%, of capital charge for scenario A2 (80% + 40% is greater than 100% reflecting the chance of more than one event); and where the capital charge for scenario B is equivalent to B1%, say 100%, of the capital charge for scenario B2 plus B2%, say 20%, of capital charge for scenario B2. The standard formula provides the 1-in-200 "$Q_{country}$" values for A1%, A2%, B1%, and B2% for each peril.

The Solvency Capital Requirement (SCR) for the flood risk, in the context with the Standard formula approach is calculated as follows:

$$SCR_{flood} = \sqrt{\sum_{r,s} corrFL_{(r,s)} * SCR_{(flood,r)} * SCR_{(flood,s)} + SCR^2_{(floodother)}} \qquad (2)$$

where $SCR_{flood,i} = max(scenario1, scenario2)$, scenario 1 includes an instantaneous gross loss of 65% of the specified flood loss in region r and a gross loss of 45% of the specified flood loss in region r and scenario 2 includes a gross loss of 100% of the specified flood loss in region r and a gross loss of 10% of the specified flood loss in region r.

For all regions the specified flood loss in a particular region r shall be equal to

the following amount:

$$L_{(flood,r)} = Q_{(flood,r)} * \sqrt{\sum_{i,j} corr_{(flood,r,i,j)} * WSI_{(flood,r,i)} * WSI_{(flood,r,j)}} \quad (3)$$

where the $Q_{(flood,r)}$ denotes the flood risk factor for region r, the sum includes all possible combinations of risk zones (i, j), $corr_{(flood,r,i,j)}$ is the correlation coefficient for flood risk in flood zones i and j of region r and finally the WSI is the weighted sums insured for flood risk in zones i and j of region r. For all regions and all risk zones of those regions the weighted sum insured for flood risk in a particular flood zone i of a particular region r shall be equal to the following:

$$WSI_{(flood,r,i)} = W_{(flood,r,i)} * SI_{(flood,r,i)} \quad (4)$$

where $W_{(flood,r,i)}$ denotes the risk weight for flood risk in risk zone i of region r and $SI_{(flood,r,i)}$ denotes the sum insured for flood risk in flood zone i of region r.

For all regions and all risk zones of those regions, the sum insured for a particular flood zone i of a particular region r shall be equal to the following:

$$SI_{(Flood,r,i)} = SI_{(property,r,i)} + SI_{(onshore-property,r,i)} + 1.5 * SI_{(motor)} \quad (5)$$

where $SI_{(property,r,i)}$ denotes the sum insured by the insurance or reinsurance undertaking for lines of business 7 and 19 in relation to contracts that cover flood risk, where the risk is situated in risk zone i of region r, $SI_{(onshore-property,r,i)}$ denotes the sum insured by the insurance or reinsurance undertaking for lines of business 6 and 18 in relation to contracts that cover onshore property damage by flood and where the risk is situated in risk zone i of region r, $SI_{(motor,r,i)}$ denotes the sum insured by the insurance or reinsurance undertaking for lines of business 5 and 17 as set out in Annex I in relation to contracts that cover flood risk, where the risk is situated in risk zone i of region r.

The amount of the instantaneous loss, without deduction of the amounts recoverable

from reinsurance contracts and special purpose vehicles is equal to the following amount:

$$L_{(Flood,Other)} = 1.1 * (0.5 * DIV_{flood} + 0.5) * P_{flood} \tag{6}$$

where P_{flood} is an estimate of the premiums to be earned by the insurance or reinsurance undertaking for each contract that covers the obligations during the following 12 months (for this purpose, premiums are gross, without deduction of premiums for reinsurance contracts) and DIV_{flood} is calculated from the below factor for geographical diversification of a particular segment s:

$$DIV_s = \frac{\sum_r (V_{(prem,r,s)} + V_{res,r,s})^2)}{(\sum_r (V_{(premium,r,s)} + V_{reserve,r,s}))^2} \tag{7}$$

where each of the sums cover all the geographical regions, $V_{(prem,r,s)}$ denotes the volume measure for premium risk of the segment s and the region r and $V_{(res,r,s)}$ denotes volume measure for reserve risk of the segment s and the region r.

Solvency II Lines of business mentioned above are presented in the table below:

Table 1: Lines of Business- Solvency II table

LoB code	Solvency II - Line of Business
1	Medical expense insurance
2	Income protection insurance
3	Workers' compensation insurance
4	Motor vehicle liability insurance
5	Other motor insurance
6	Marine, aviation and transport insurance
7	Fire and other damage to property insurance
8	General liability insurance
9	Credit and suretyship insurance
10	Legal expenses insurance
11	Assistance
12	Miscellaneous financial loss

Under Solvency II, capital requirements are determined on the basis of a 99.5% value-at-risk measure over one year. For the re-calibration process, it is important to note that it can take more than two years between parameters re-calibration and when undertakings will actually use these parameters. In addition, the fact that the SF parameters are not re-calibrated annually needs to be considered. It is key to ensure that the parameters are adequate for more than one year as the same parameters will be used during multiple years until a re-calibration will be done. It might therefore be important to introduce a forward-looking approach when performing a Nat Cat SF parameters re-calibration to ensure that the parameters are valid over the next 5-10 years. Let us assume the following example, a re-calibration is done in 2025, the next re-calibration takes place 5 years later in 2030. However, the industry will implement the parameters from the 2030 calibration only in 2032. This means that the parameters from the 2025 re-calibration need to be valid to be used until 2032 (and this is the under the condition that the new re-calibration took place 5 years later).

Solvency II is a risk-based approach and it is important to consider all aspects of the risks. For Nat Cat calculation, exposure, vulnerability and hazard should be taken into account. Moreover, in order to integrate climate change aspects in Solvency II, it is not sufficient to just consider changes in hazard (for example higher precipitation rates) but it is also necessary to consider the exposure and its corresponding vulnerability. If for example flood risk increases, it might well be possible that adaptation measures are taken and flood defences are installed so that the risk does not increase even though the hazard does.

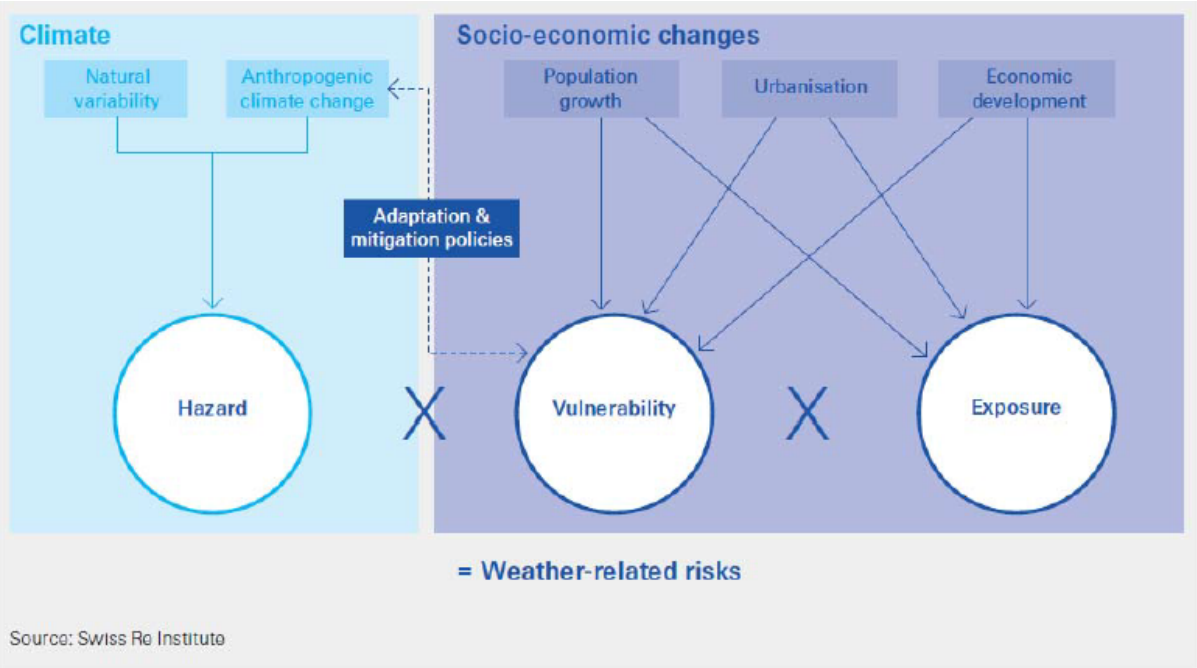


Figure 1. Contributing factors of the three main components of weather-related risks (Swiss Re, 2020)

If a weather related NAT-CAT model is available for a specific country, the factors of the model reflect the Nat Cat risk for the next 12 months. The impact of climate change is mostly not explicitly reflected in the current Nat Cat models. Any current climate change will be implicitly included in the recent data (available historical data) and scientific assumptions used to create the Nat Cat models. Also, Nat Cat models are not necessarily updated annually as it takes a lot of effort and resources. The models used in the calibration would therefore typically be a couple of years old. Up to today, the impact of the climate change is partially depicted through data of the most recent years. Standard-formula models do not take into account estimations about the development of basic meteorological indices for the upcoming decades. The climate change impact is implicitly estimated as an accounting view of the claims occurred because of extreme weather events which affect the exposed portfolio of an insurance company.

In the table below, perils included in the Standard Formula (SF) calculations are

described.

Table 2: Definition of perils and coverage in the SF

A/A	SF Peril name	Type of disaster	EM-DAT definition	SF
1	Earthquake	Geophysical	Sudden movement of a block of the Earth's crust - Ground shaking.	Ground movement. Tsunami & fire following the event are not included.
2	Flood	Hydrological	Overflow of water from a stream. Flood because of heavy rain (flash floods).	Fluvial, pluvial and flash flood events. Coastal, ice-jam flood are not included.
3	Windstorm	Meteorological	Cyclonic storms, Tropical cyclons, Convective storms.	Cyclonic storms, coastal flood, windstorm.
4	Hail	Meteorological	Sub-category of convective storms.	The SF also includes losses from tornadoes, lightning etc.
5	Subsidence	Geophysical	Sinking of the ground.	Refers to a swelling or shrinking of clay soils.

The "EM-DAT" is the "Emergency Events Database" of the Centre for Research on the Epidemiology of Disasters – CRED Université catholique de Louvain. Main weather related perils are flood, windstorm, and hail. Subsidence even it is considered as a geophysical peril, sometimes can be caused by high precipitations or high rain rates and floods.

Although climate change effect is considered to have significant impact on insurance obligations and claims the upcoming decades, Greece is not included in the Solvency II Standard Formula calculations. The following table provides an overview of the perils that are covered in the SF due to their materiality in the relevant countries.

Table 3: List of countries and perils that are currently included in the SF

Country	Windstorm	Earthquake	Flood	Hail	Subsidence
AT (Austria)	X	X	X	X	
BE (Belgium)	X	X	X	X	
BG (Bulgaria)		X	X		
CY (Cyprus)		X			
CZ (Czech Republic)	X	X	X	X	
DE (Germany)	X	X	X	X	
DK (Denmark)	X				
ES (Spain)	X				
FI (Finland)	X				
FR (France)	X	X	X	X	X
GR (Greece)		X			
HR (Croatia)		X			
HU (Hungary)	X	X	X		
IE (Ireland)	X				
IT (Italy)		X	X	X	
IS (Iceland)	X				
LI (Liechtenstein)	X	X	X	X	
LU (Luxembourg)	X			X	
MT (Malta)		X			
NL (Netherlands)	X			X	
NO (Norway)	X				
PL (Poland)	X		X		
PT (Portugal)		X			
RO (Romania)		X	X		
SE (Sweden)	X				
SI (Slovenia)	X	X	X	X	
SK (Slovakia)		X	X		

Climate change affects the natural perils in different ways and has an impact on the frequency, severity and regional - seasonal distribution of windstorms, floods, hail and subsidence. Unfortunately, a lot of countries, like Greece, are not included in the SF calculation and cannot estimate a possible financial impact of weather related CAT-risk. Moreover, the unavailability of appropriate databases in order for monitoring that kind

of perils is a major issue for any model calibration and estimation of the hazard.

Definition of NAT CAT Parameters used in the SF - EIOPA

The country factors used for the SF calculation model represent the per-occurrence 99.5% loss for a specific peril in the country which is under consideration, as a ratio of the total sums insured. This can be represented as the one in two hundred years per occurrence PML (Probable Maximum Loss) percentage. For each peril best estimates of each country's 1/200 year per occurrence PML were provided. Expert judgment was used to identify outliers and obtain consensus on the outcome. It was assumed that a peril is not significant for a given country, if its estimated country factor was less than 1/15 of the largest peril-specific factor for that country. In other words, the materiality threshold used to decide to include or not a specific peril or region in the SF was if its 200 year loss exceeds approximately the 1/15 of the highest 200 year peril loss.

The matrices for the correlations between perils and between countries were derived using an iterative discussion process using expert judgment. The same process was used in order to estimate the country factors. Estimates of the ratio of the 1 in 200 year annual aggregate loss to the 1 in 200 year per occurrence loss for each peril were provided. A consensus on how to distribute the 1 in 200 year aggregate loss between two occurrences for each peril was based on expert judgment.

Finally, The zonal relativities are proportional to the 1 in 200-year loss of each zone and the aggregation matrices reflect the correlation between zones at the 1 in 200 year loss level. The calculation and calibration of these relativities were derived using several underlying, stochastic event-based catastrophe risk models and an assumption about the relative distribution between the zones of the total sums insured within the country. While the methodology was consistent, not all countries and perils benefited from the same level of detailed model treatment. It should be noted that the zonal relativities and correlations only become relevant to the extent that the geographic

distribution of an undertaking's exposures deviate from the industry average distribution assumed in the calibration.

Climate Risk in the European Insurance Sector

The European Environment Agency (EEA) reports on climate change, impacts and vulnerability in Europe (EEA, 2017 and 2020) show that climate change is already having wide-ranging consequences for human health, the environment and economies across Europe. In addition, the JRC PESETA IV study stresses the urgent need to mitigate greenhouse gas emissions or adapt to climate change. Limiting global warming to well below 2°C would considerably reduce climate change impacts in Europe. Adaptation to climate change would further minimize unavoidable impacts in a cost effective manner (JRC, 2020)[17].

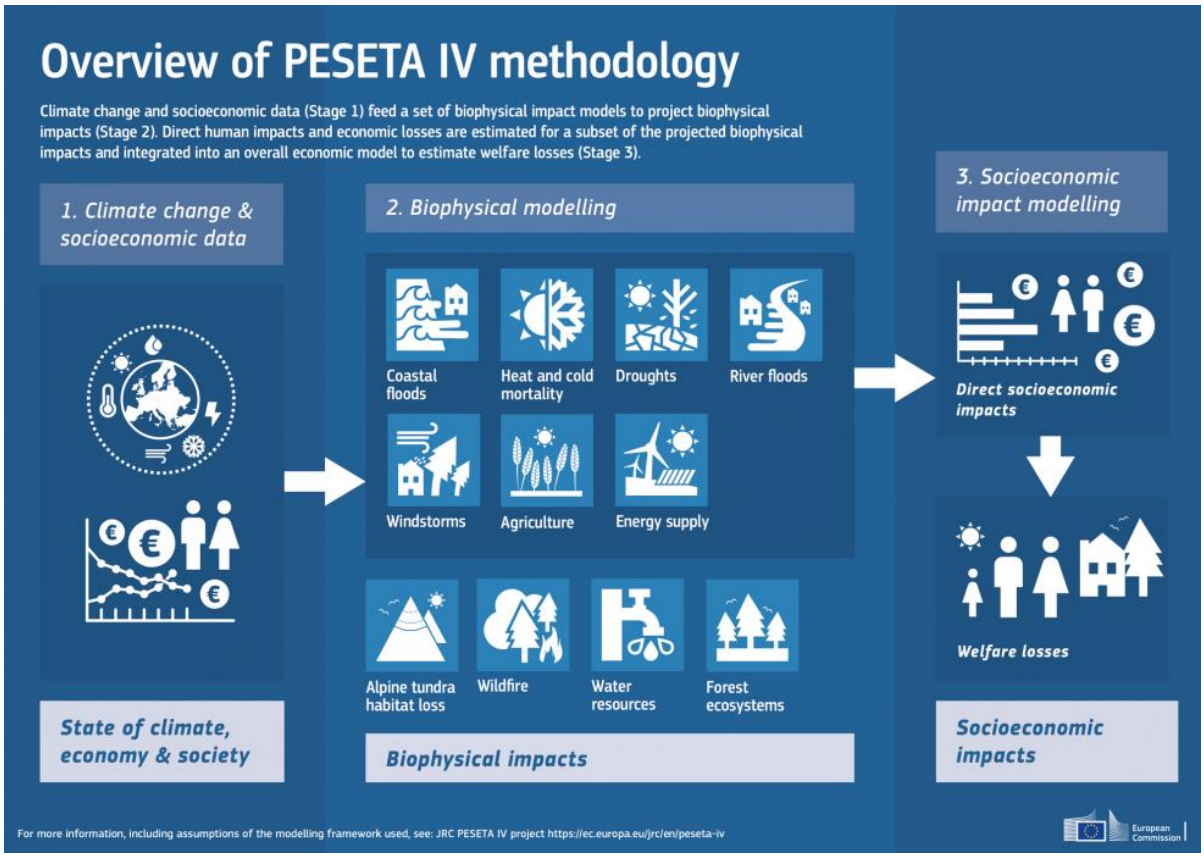


Figure 2. Overview of the PESETA IV methodology.

In order to understand better the impact of climate change in Europe, EIOPA focused on the hazards with general acceptance by the scientific community as the most significant the upcoming years. Swiss Re has assessed the risk in terms of levels of confidence as to expected outcomes across different weather and environmental variables based on available studies and general acceptance of the scientific community. A reduced level of confidence does not imply no impact of climate change but rather that less available data or scientific analysis are presently available (Swiss Re, 2020).

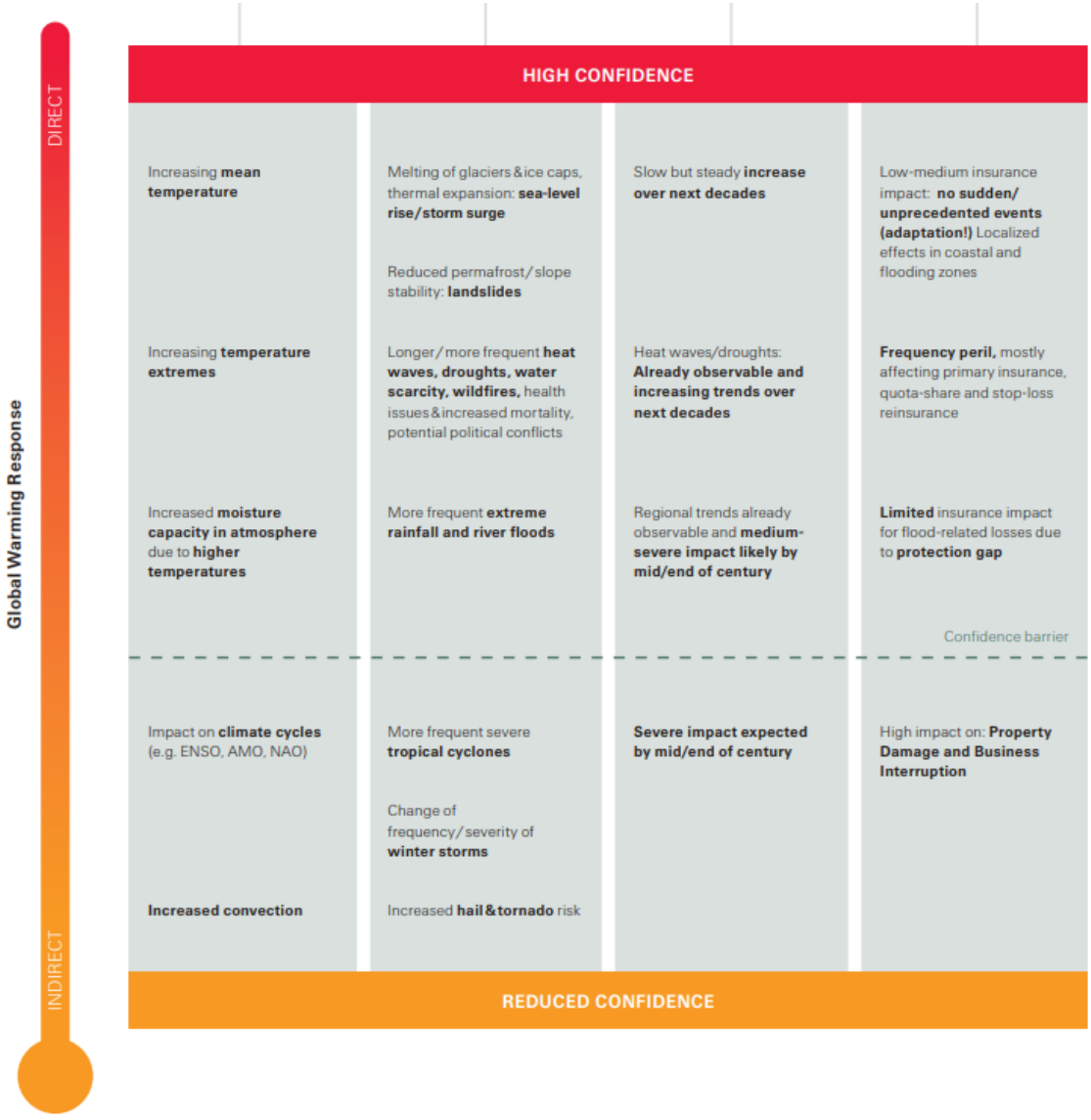


Figure 3. Classification of climate-change effects and their relevance for the re/insurance industry - Swiss Re Institute.

From the above graph it seems that confidence about observed and future trends is higher related to the increase in global temperatures and temperature extremes, but it is not taken into account the great impact of the severe convective storms which are getting worse in terms of severity and frequency.

In Europe, different regions and sectors will be affected differently by climate change. The rise in sea level has increased flood risks and contributed to erosion along European coasts. The observed increase in heat waves has had significant effects on human health, in particular in cities. Heat waves are also increasing the risk of electricity blackouts and forest wild fires. Climate change is affecting most of regions in Europe, but the impacts are not uniform and countries are facing different types of extreme weather conditions. South eastern and southern Europe are projected to be hot spot regions, having the highest numbers of severely affected domains. Coastal areas and floodplains in the western parts of Europe are also areas of high risk. The Alps and the Iberian Peninsula are additional hot spots for ecosystems and their services. Ecosystems and human activities in the Arctic will be strongly affected owing to the particularly fast increase in air and sea temperatures. This will cause fast melting of ice on land, sea ice, and thawing of permafrost both in the Arctic Circle and at high-elevation mountain sites outside the Arctic. Economic costs can potentially be high, even for modest levels of climate change, and these costs rise significantly for scenarios of greater levels of warming. It is very important to mention that the projected damage costs from climate change seem to be higher in southern Europe. However,the projected estimated economic impact of climate change considers only some sectors and show considerable uncertainty. The magnitude of future climate change and its actual impact from the middle of the century onwards depend on the effectiveness of global climate mitigation efforts. This can be substantially reduced by keeping the increase trend in global average temperature below a specific threshold.

In the map below main perils, results of the climate change, are depicted per region.

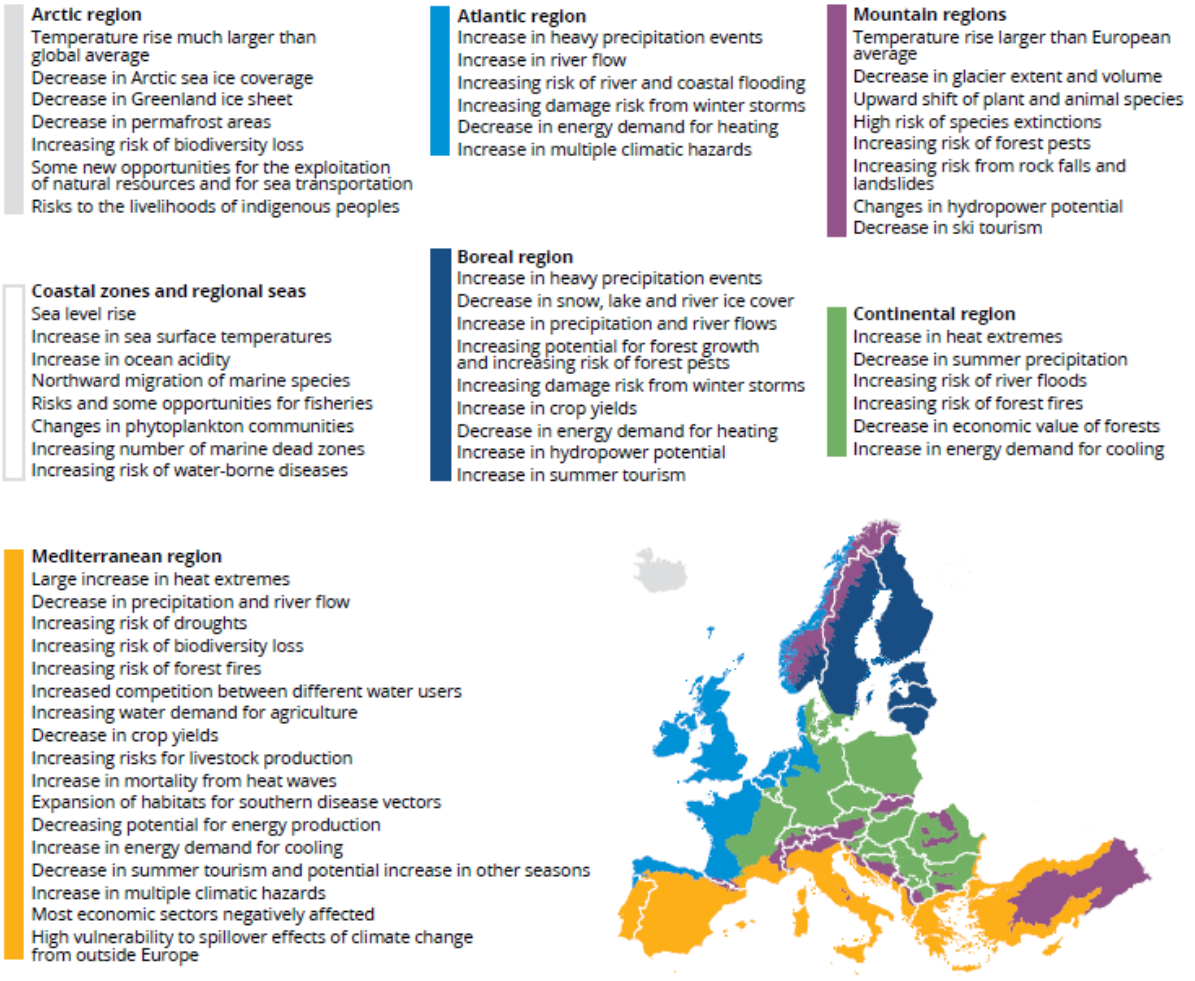


Figure 4. Key observed and projected climate change and impacts for the main biogeographical regions in Europe (EEA, 2017).

Adaptation measures can reduce the financial impact of these perils by taking appropriate actions and preventing or minimizing the incurred losses. For example, adapting building codes to future climate conditions and extreme weather events, building flood defences and raising the levels of dykes, developing drought-tolerant crops, choosing tree species and forestry practices less vulnerable to storms and fires can minimize the hazard and the vulnerability components of the weather-related risk.

According to JRC PESETA IV study, climate change adaptation can reduce unavoidable impacts of climate change in the EU in a cost-efficient way. For example, in case of unmitigated climate change, reducing flood peaks by installing retention reservoirs would reduce annual river flood damage at the end of the century by nearly 40 €billion per year and around 400,000 fewer people would be exposed each year to flooding in the EU and the UK. The annual investment from now until 2100 to install and maintain the reservoirs would be 3.3 €billion/year. There are additional benefits of nature-based storage areas, such as restoring the natural functioning of floodplain areas and improving ecosystem quality (JRC, 2020)[17].

Including Climate Change in the NAT-CAT SCR Calibration.

The current parameters in the NAT CAT Standard Formula do not explicitly consider climate change. It could be assumed that to a large extent the current calibration approach captures appropriately climate in the one-year time horizon specified under Solvency II. These parameters are also assumed to sufficiently capture forward looking trends to be used for a couple more years until a (re)calibration will be performed. Moreover, the current Standard Formula calibration uses NAT CAT models which are calibrated to reflect implicitly the recent climate change. There is a difficulty to quantify future climate change related impacts to catastrophe and extreme weather. The amalgamation of climate change with other variables with even more dominance, like urbanization, increased coastal settlements and population growth makes more complex any estimation.

A re-calibration could allow to update the parameters of the Standard Formula but maybe without specific consideration of the climate change impact. It is difficult to ensure that the parameters properly reflect the risk for the time they will be used by the undertakings until a new re-calibration will take place. The 2010 NAT CAT parameters were for example used by undertakings to calculate the NAT CAT SCR until 2020. So in a such a long time period many parameters could change and may not anticipate the correct profile of the risk.

The retrospective approach regarding the weather related CAT risk estimation

may be in a way anticipated if we believe that the observations period is adequate enough in order to describe trends which are assumed to reflect the short-term climate change time horizon. However, for certain perils and regions strongly impacted by climate change it might be important to consider a more forward-looking approach and not base the risk estimation only on historical data or scientific data which do not consider a projection approach of the hazard. Climate change means that using the assumption that past losses are a reliable way to estimate future losses may no longer be an appropriate approach. In this situation, it becomes even more important to adopt modelling based scientific principles to assess the risks using also meteorological and climatology indices.

In light of climate change, there is a clear need to ensure that model vendors and insurers collaborate with academic and scientific communities to develop a better understanding of the uncertainties involved in climate change and how these impacts can be quantified.

Weather Related Perils in Greece

Although weather related risk is not considered as the most significant NAT CAT risk in Greece there is an upward trend in recent years that should be taken into account. In the graph below it is presented that the relative percentage of historical insured losses for geophysical risk (earthquake) is greater than other perils (55% vs. 45%).

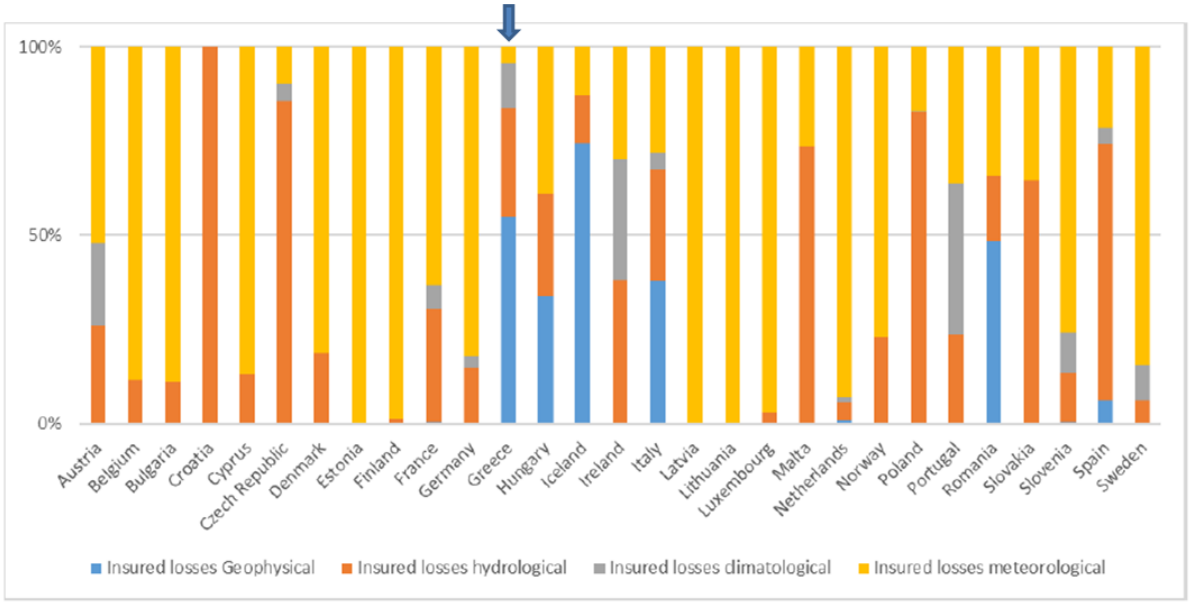


Figure 5. Relative percentage of historical insured losses for geophysical, meteorological, climatological and hydrological events per country.

Due to climate change, the frequency and severity of weather related natural catastrophes is expected to increase. Improved climate projections provide evidence that future climate change will increase climate-related extremes (e.g. heat waves, heavy precipitation, droughts, top wind speeds and storm surges) in many European regions. EIOPA is using the final report of the EU Technical Expert Group on Sustainable Finance (TEG, 2020) as they created a classification specific for climate related hazards and separate between chronic and acute climate-related hazards. The classification comprises four major hazard groups, with hazards related to water, temperature, wind, and mass-movements. All groups include acute (extreme) and chronic (slow-onset) hazards. EIOPA focuses on acute hazards which could be very risky regarding the stability and the solvency of insurance companies. In the table below there is the categorization of these perils.

Table 4: Classification of climate-related hazards.

Perils	Temperature-related	Wind-related	Water-related	Solid mass-related
Chronic	Temperature change	Wind patterns change	Precipitation patterns change	Coastal erosion
	Heat stress		Precipitation variability	Soil degradation
	Temperature variability		Ocean acidification	Soil erosion
	Permafrost thawing		Saline intrusion	Solifluction
			Sea level rise	
			Water stress	
			Drought	
Acute	Heat Wave	Tropical cyclone	Drought	Avalanche
	Cold wave/frost	Windstorm	Heavy precipitation (rain, hail, snow/ice)	Landslide
	Wildfire	Tornado	Flood (coastal, fluvial, pluvial)	Subsidence
			Glacial lake outburst	

Temperature-Related Risks

Large forest fires in recent years have affected various regions in Europe in which fires were not prevalent in the past. The unprecedented forest fires in several European countries in 2017 and 2018 coincided with record droughts and heatwaves in these years. As temperatures in the summer period gradually increase fire risk follows the same pattern. The large majority of wildfires in Europe are usually ignited by humans, either accidentally or intentionally. However, climatic factors and the availability of fuel determine the conditions under which fires occur and spread, once ignition has occurred. The extreme fire episodes and devastating fire seasons of recent years in Europe were, in most cases, driven by severe fire weather conditions. Thus, climate change is expected to have a strong impact on forest fire regimes in Europe.

Most significant wild fires the last decades in Greece were the forest fires in Peloponnese in 2007 and the wild fires of “Mati” in 2018. There were two very extreme

incidents with a lot of victims and very expanded losses in the property line of insurance business. Other significant forest fire incidents also occurred during the summer mainly in south and central Greece.

On 23 July 2018, Attica, Greece, was impacted by a major wildfire that took place in a wildland–urban interface area and exhibited extreme fire behavior, characterized by a very high rate of spread. One-hundred civilian fatalities were registered, establishing this wildfire as the second-deadliest weather-related natural disaster in Greece, following the heat wave of July 1987, while it is considered the second deadliest of the 21st century worldwide after the fires in Australia on February 7, 2009 which had killed 173 people. On the day of the deadly wildfire, a very strong westerly flow was blowing for more than 10 h over Attica. Wind gusts up to 30–34 m/s occurred over the mountainous areas of Attica, with 20–25 m/s in the city of Athens and surrounding suburban areas. This strong westerly flow interacted with the local topography and acted as downslope flow over the eastern part of Attica, with temperatures rising up to 39°C and relative humidity dropping to 19% prior to the onset of the wildfire. These weather elements are widely acknowledged as the major contributing factors to extreme fire behavior.[34]

The insurance losses occurred were estimated approximately €36 mil. (34.5 mil. for the Property line of business and 1.5 mil for Motor)[3] It is very important to mention that the specific region is actually part of the East suburbs of Athens and is an area with significant insurance exposure. In the map below it is obvious the land cover-use, which is mainly mixed urban-vegetated.

[3]http://www1.eaee.gr/deltia-typoy/xepernoyn-ta-eu-36-ekat-oi-provlepseis-ton-apozimioseon-tis-asfalistikis-agoras-gia-tis-katastrofikes-pyrkagies-tis-23is-7-2018-stin-attiki

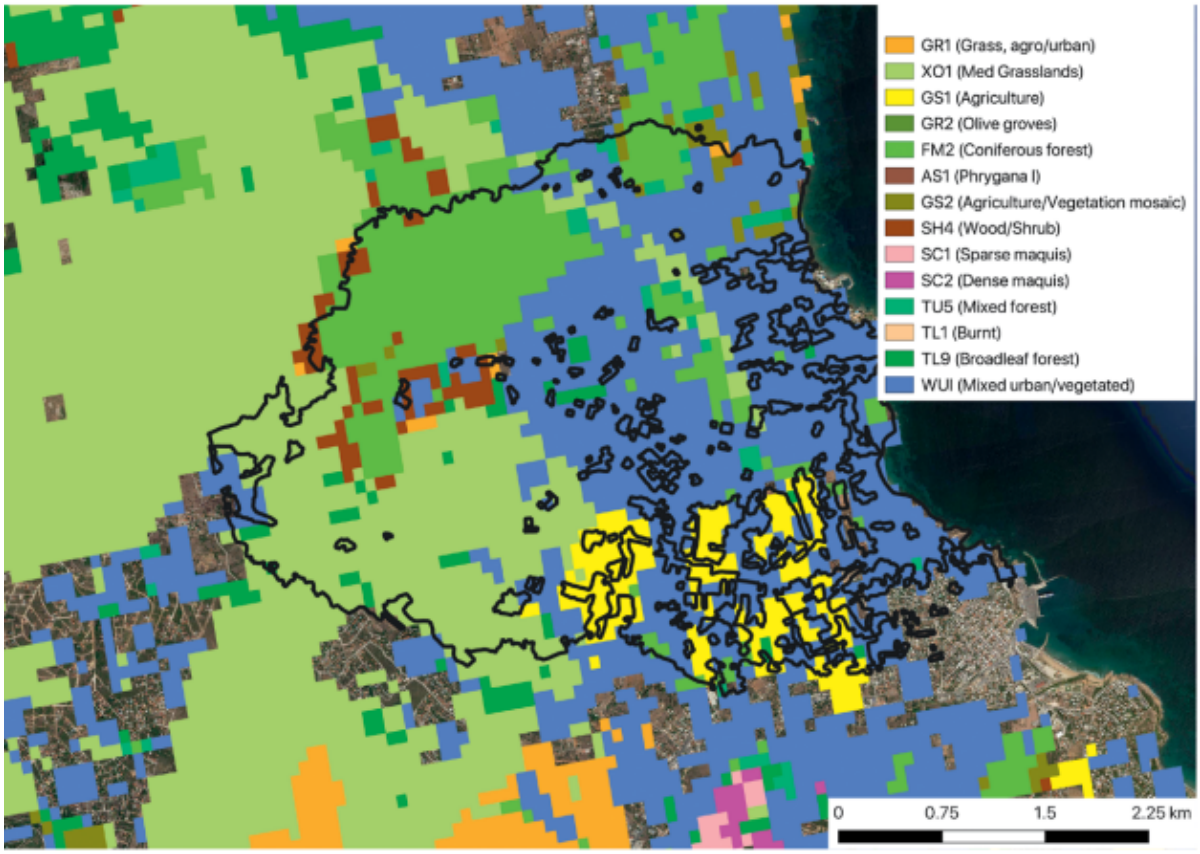


Figure 6. Land Cover - Usage - Mati 2018 .

The affected area was very close to the city of Athens and was densely inhabited. That is the main reason of the expanded insurance claims occurred in contrary with the relative narrow destruction path of the fire.

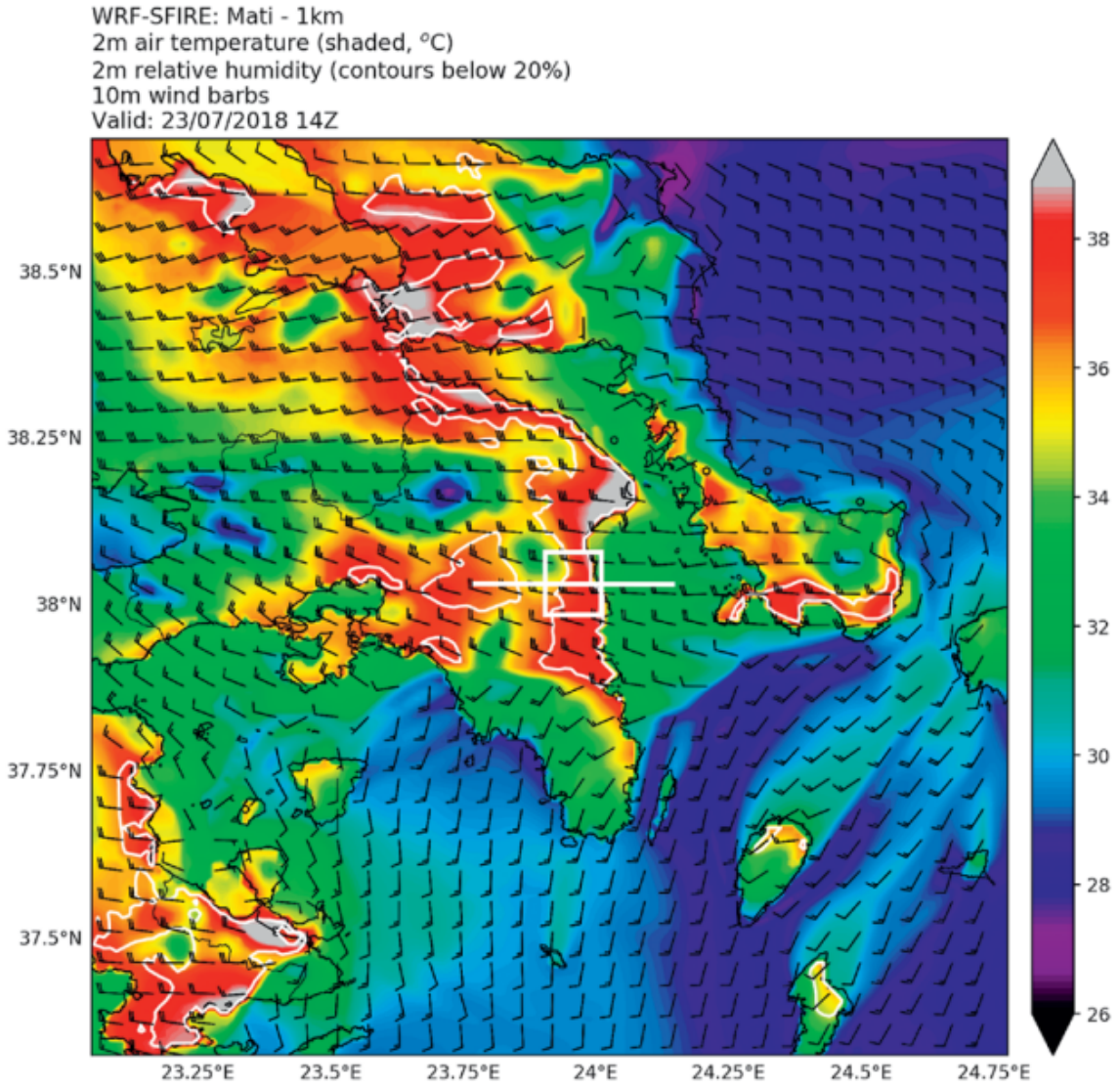


Figure 7. Meteorological Conditions - Mati 2018 .

The maximum daily temperature recorded at Rafina AWS during 23 July 2018 was 39°C (at 1230 UTC) and this was the highest temperature recorded in Attica not only during that day, but also during all summer months of 2018. The narrow band along the eastern coasts of Attica where temperature exceeded 38°C was also very dry with relative humidity (RH) less than 20% and very strong winds.

On the other hand, during the wild fires in Peloponnese in 2007 although the magnitude of this event is considered to be an extreme event regarding the past data and circumstances of the specific region, the insurance impact was not such extended. The forest fire of 2007 in Peloponnese in Greece corresponds to the largest fire in the country's record that burnt 1773 Km^2, causing 78 fatalities and very significant damages in property and infrastructure. Moreover, the distraction path was so expanded that triggered also other NAT CAT events like floods and mass movement events. Observational evidence based on several data sources collected during the period 1989–2016 show that the 2007 fire has contributed to an increase of average flood and mass movement events frequency by approximately 3.3 and 5.6 times respectively. Fire affected areas record a substantial increase in the occurrence of both phenomena, presenting a noticeably stronger increase compared to neighbouring areas that have not been affected. Examination of the monthly occurrence of events showed an increase even in months of the year were rainfall intensity presented decreasing trends. Although no major land use changes has been identified and chlorophyll is shown to recover 2 years after the fire incident, differences on the type of vegetation as tall forest has been substituted with lower vegetation are considered significant drivers for the observed increase in flood and mass movement frequency in the fire affected areas.[13]

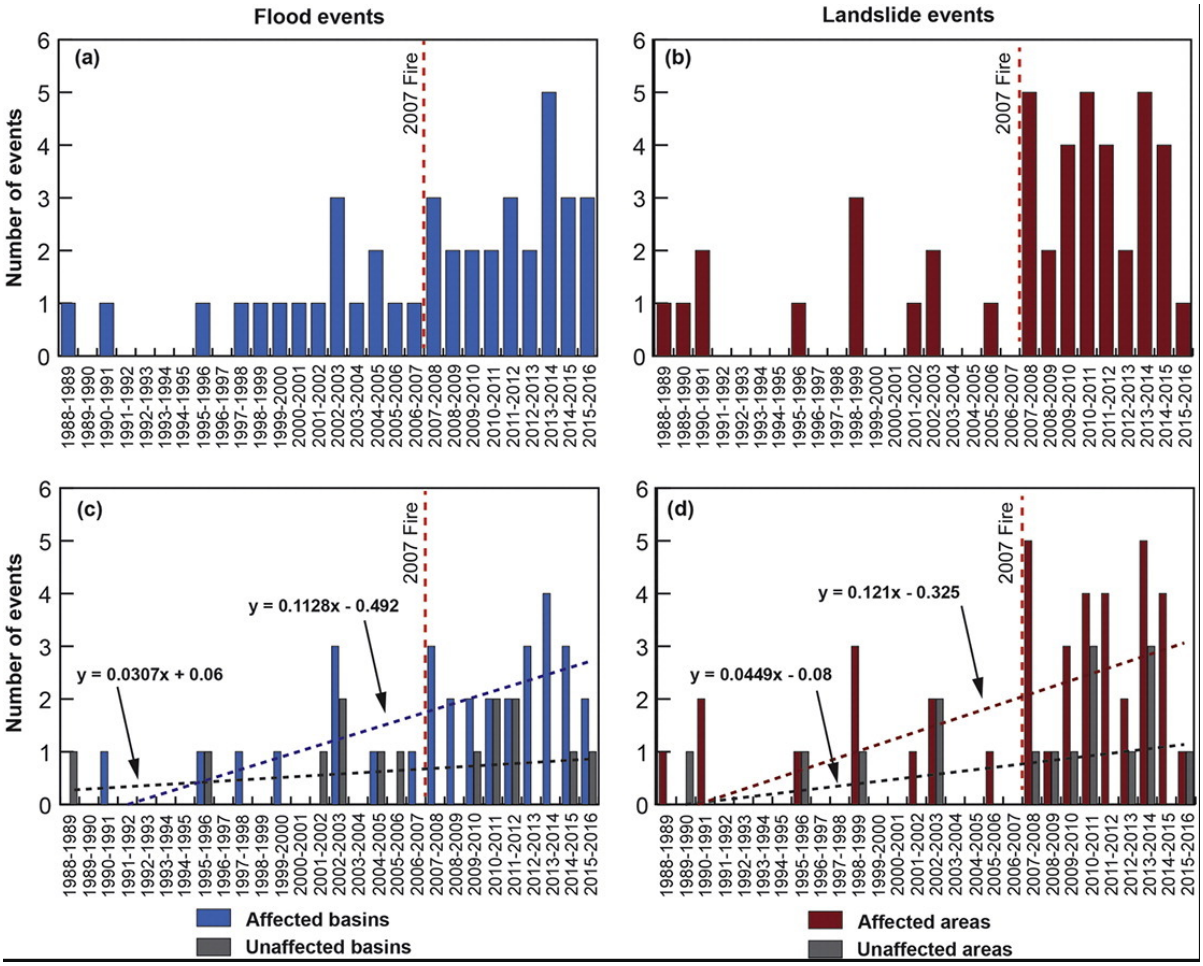


Figure 8. Wild Fire 2007 Peloponnese - Flood and mass movements impact.

From the above graph it is considered that the final impact of the wild fires is sometimes complex to estimate as beyond the immediate impact of the event the indirect financial insurance losses could not been foreseen. For example the flood events or subsidence triggered as presented in the above graph as a consequence of the forest fire maybe sometimes be more devastating and expanded from the initial event.

According to the response of the Insurance Europe Institution[4] to the EIOPA's position paper for the climate change, wildfire risk exists in Greece and constitutes an

[4]https://www.insuranceeurope.eu/

increasing concern since the Mati incident in 2018. But as that event led to an increase of the levels of state readiness towards the phenomenon to mitigate the exposure, it is mentioned that the specific peril should not be included to the standard formula calculations.

Climate change projections suggest substantial warming and increases in the number of heat waves, droughts and dry spells across most of the Mediterranean area and more generally in southern Europe, which would increase the length and severity of the fire season. For example, the wild fires of 2007 followed a remarkably warm period. Three major heat waves were detected, one in each summer month, with maximum and minimum temperature values breaking many records since late 1940s.[60]

	June	July	August	JJA
Tmax (1946–2006)	29.1°C	31.4°C	31.2°C	30.6°C
Highest Tmax value	(39.0°C)	(42.6°C)	(41.5°C)	
Tmax (2007)	31.6°C	34.7°C	32.7°C	33.0°C
Highest Tmax value	(41.6°C)	(43.6°C)	(39.6°C)	
Tmin (1946–2006)	18.3°C	20.7°C	20.5°C	19.8°C
(Highest Tmin value)	(25.0°C)	(26.9°C)	(26.4°C)	
Tmin (2007)	20.2°C	21.9°C	21.1°C	21.1°C
(Highest Tmin value)	(25.7°C)	(24.8°C)	(23.5°C)	

Figure 9. Mean Maximum and Minimum Temperature Values for the Time Period 1946–2006 and the Year 2007 for the Summer Months in Greece.

The projected changes in soil moisture anomaly (Palmer Drought Severity Index) show a strong latitudinal gradient, from pronounced drier conditions in southern Europe to wetter conditions in northern European regions in all seasons. The largest changes in the soil moisture index between 2021–2050 and the baseline period (1961–1990) are projected for the summer period in the Mediterranean, especially in north-eastern Spain and south-eastern Europe. Projections for the end of the 21st century show significant

decreases in summer soil moisture content in the Mediterranean region and central Europe, and increases in the north-eastern part of Europe.[1]

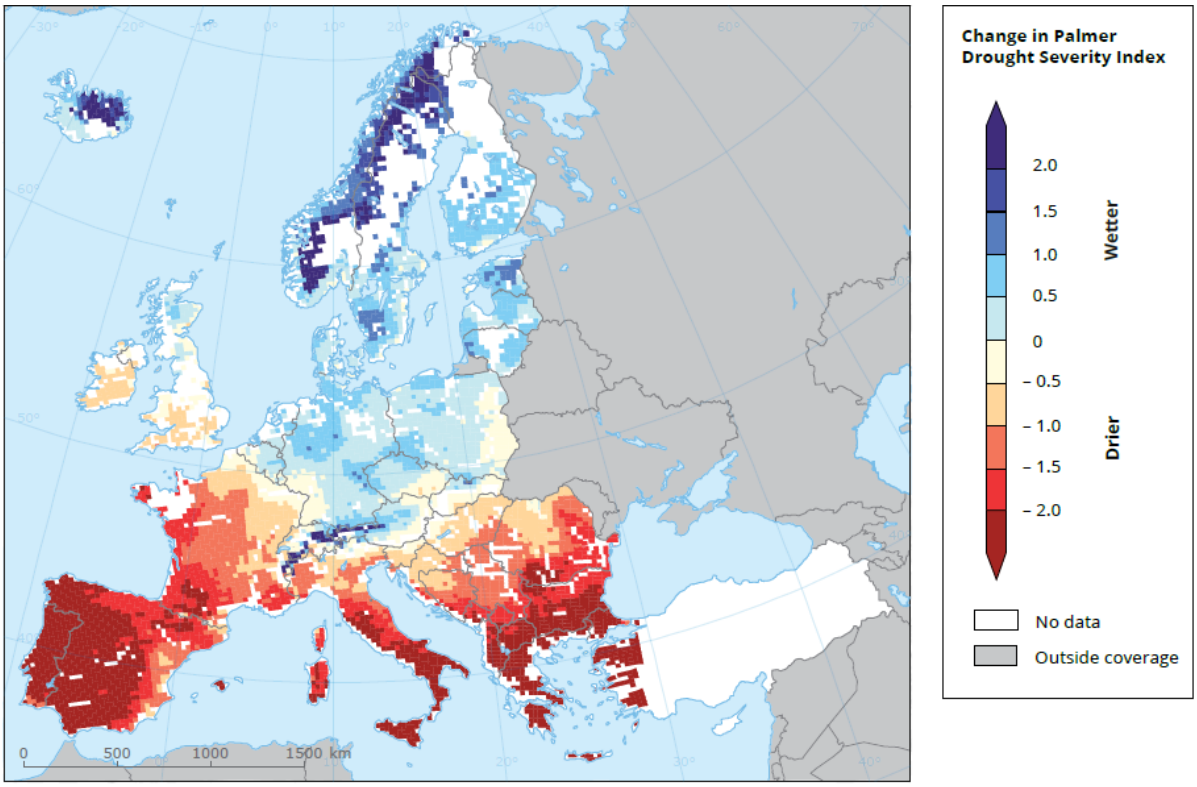


Figure 10. Projected changes in summer soil moisture.

Moreover the constant raising trend of the temperature may lead to more severe phenomena caused by high temperatures. All of Europe has warmed significantly since the 1960s The strongest warming has been observed over the Iberian Peninsula, particularly in summer and across central and north-eastern Europe. Winter warming has been strongest over Scandinavia. Mediterranean countries are affected by the temperature increase specially during summer. The years 2014 and 2015 were jointly the warmest calendar years in Europe since instrumental records began and the global average temperature is expected to continue to increase throughout this century as a result of projected further increases in greenhouse gas concentrations. The projected increase in fire risk in southern Europe are robust across different modelling approaches whereas the projections for

northern Europe are more uncertain.

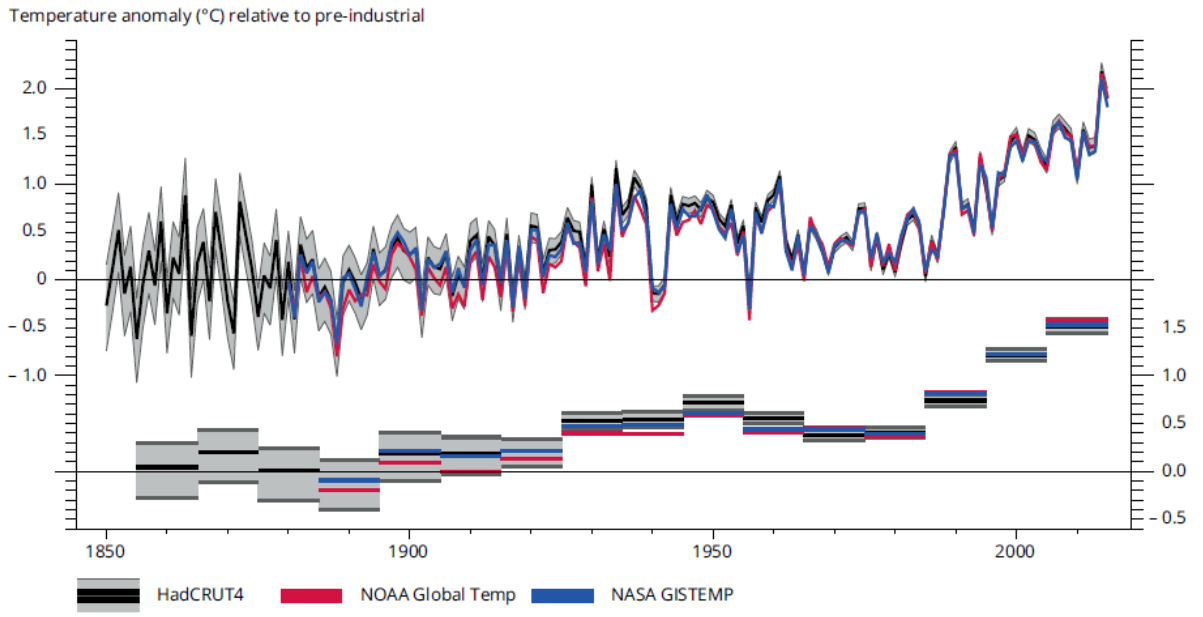


Figure 11. European average temperatures between 1850 and 2015 over land area relative to the pre-industrial period.

Wind-Related Risk

Windstorms are amongst the most damaging natural hazards in Europe, with approximately €5 billion of estimated annual losses in the EU. The number of reported windstorms has increased significantly over the last decades, yet there is no consensus about a climate-induced trend in windstorms over Europe.

In Greece extreme windstorms are not very often phenomena. Although, in case of specific meteorological conditions extended insurance claims may occur. Extreme wind storms may happen in case of several types of convective systems either in case of cyclonic conditions. During 15–21 September 2020, an intense "medicane"[5], named

[5]Hurricane of the Mediterranean sea

Ianos, formed over the warm Mediterranean Sea. Following a path of approximately 1,900 km, Medicane Ianos affected Greece resulting in four casualties and devastating damage in the western and central parts of Greece. Persistent gale force 1-min winds up to 44 m/s and wind gusts up to 54 m/s were recorded in Cephalonia Island (Ionian Sea), while record-breaking amounts of accumulated rainfall have been recorded in several Ionian islands, as well as in parts of central Greece. The Mediterranean Sea occasionally hosts intense cyclones with characteristics similar to tropical cyclones, such as a frontless and axisymmetric structure, deep convection close to their centers with a windless "eye" and sustained wind speeds close to hurricane strength.[33] Such cyclones are often named "medicanes". Although most of their lifetime is spent over the sea, when they reach the coasts of Mediterranean countries they can become very destructive.

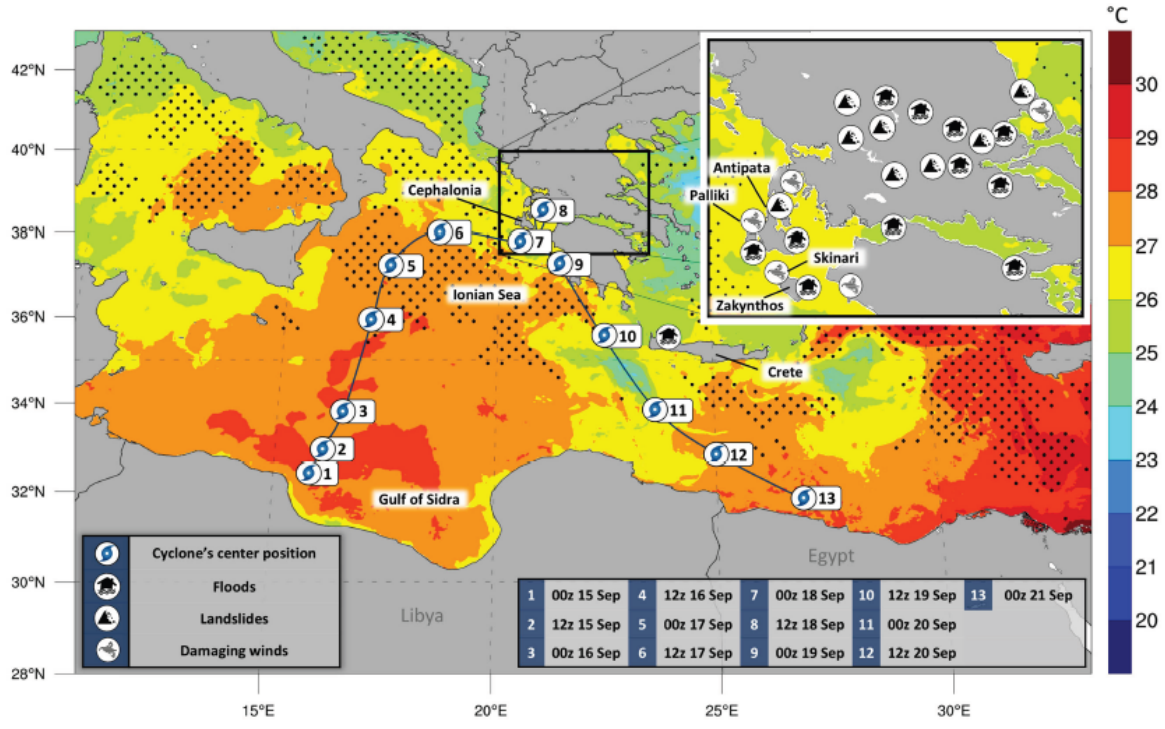


Figure 12. Path of Medicane Ianos, from 0000 UTC 15 Sep up to 0000 UTC 21 Sep 2020. Shading denotes the SST (at 1-K intervals), and the black dots the daily SST anomaly higher than 2 K on 15 Sep 2020. Symbols for floods, landslides, and damaging winds are placed over the most affected areas.

Ianos initially emerged as a surface cyclone close to a thunderstorm cluster in the Gulf of Sidra on 15 September 2020, off the coasts of Libya. Global and regional model guidance showed a highly possible transition of the surface cyclone to a powerful medicane, a fact that led the METEO unit at the National Observatory of Athens (NOA) to name the medicane "Ianos" in the morning of 16 September, following NOA's practice adopted since 2017 to name weather events that are expected to produce high socioeconomic impacts.[6] Based on satellite imagery, Ianos was evolved into a powerful medicane on 17 September 2020 when it started gradually to affect Greece with storm surges accompanied by high coastal waves, high rainfall amounts, and devastating winds, which resulted in four fatalities and extended infrastructure damages and landslides in the Ionian Islands and in central Greece. The torrential rainfall resulted in a large number of landslides and debris flows and more specifically in central Greece the were mapped more than 1,400 landslides in 2 days. Hurricane-force winds, resulted in extended damage in the Ionian Islands fortunately with no loss of human lives. At the same time the maximum wind gust recorded was 54.2 m/s and a record-breaking daily accumulated rainfall of 644.7 mm at Cephalonia which almost corresponds to the annual accumulated precipitation of Athens.

The 48-h (17 and 18 September 2020) accumulated rainfall in the Ionian Islands reached 769 and 250 mm in Cephalonia and Zakynthos Islands, respectively. During the same period, a very strong easterly flow was established over central Greece, associated with the low-level medicane flow in the Ionian Sea, which resulted in local strong low-level convergence zones and long-lasting rainfall with extreme rain rates. On 18 September 2020 the daily accumulated rainfall exceeded 250 mm in central Greece. The high amounts of precipitation in central Greece resulted in widespread flooding, extensive damage in rural and urban areas, and unfortunately four fatalities. The figure below shows the 48-h accumulated rainfall at the stations of Pertouli (P), Karditsa (K), and Mouzaki (M) in central Greece. The highest value was recorded in Pertouli (P), at the station located at the highest elevation among the three stations (1,170 m), with a total amount of 317 mm. Mouzaki (M) and Karditsa stations (K) recorded 268 and 212 mm

[6] www.meteo.gr/namedstorms.cfm

of rain within the same time period.

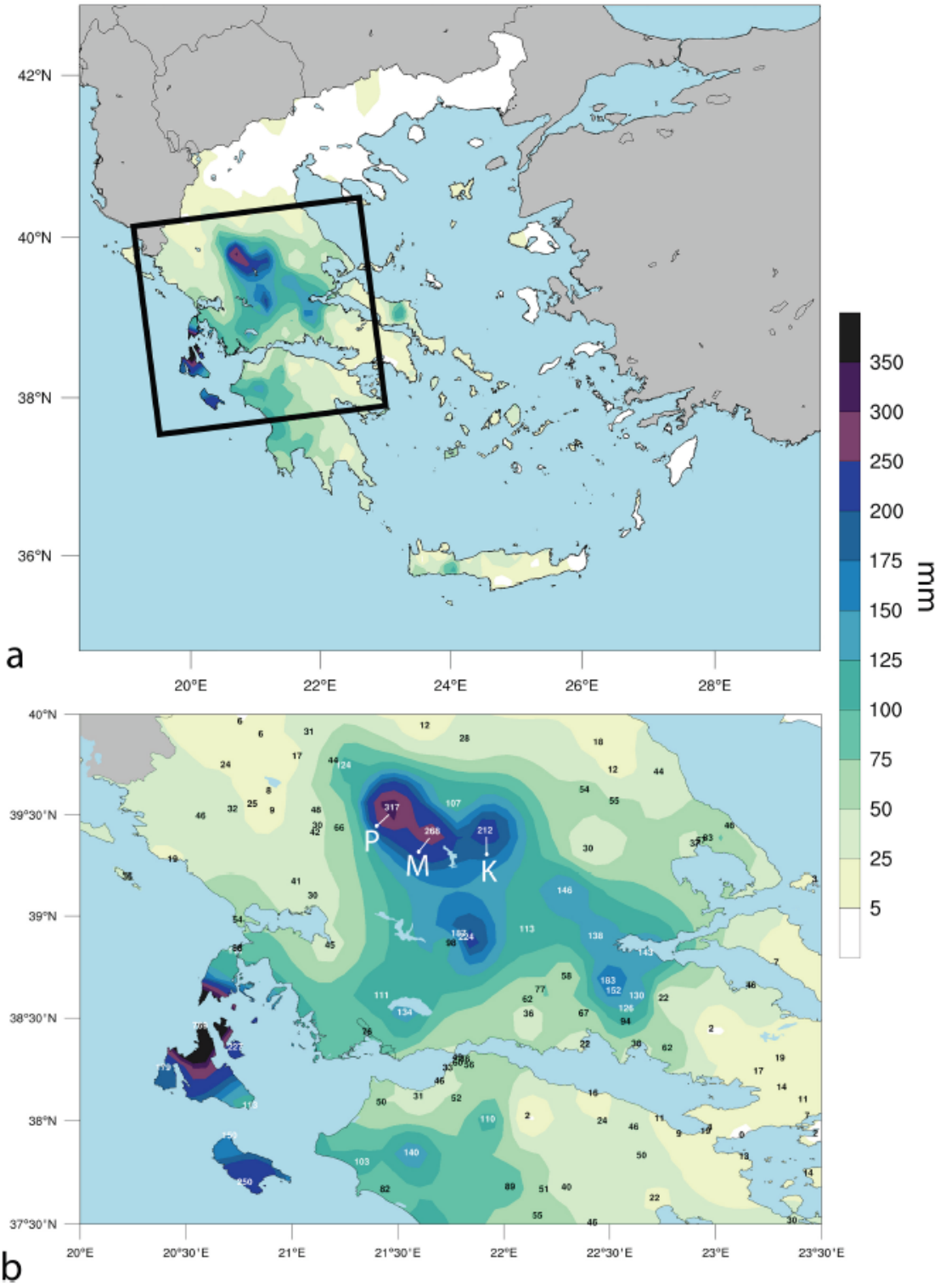


Figure 13. (a) Total accumulated rainfall over Greece for the period spanning from 1200 UTC 17 Sep to 1200 UTC 19 Sep 2020 (48 h) and (b) zoom over central Greece, for the same period.

The financial impact regarding the insurance losses was very significant although the bigger urban centers of Greece were not actually affected. For the total insurance industry of Greece, 2453 number of claims were reported with a first estimated claims cost of €30.9 mil. Of these, 1400 claims were related to the Property line of business, 1027 to the Motor business and 26 to the marine line.

According to the Hellenic Association of Insurance Companies[7] the outstanding claims reported after the event for the Property line of business, regarding all Greek insurance companies, are presented below per insured object. In the last column it is calculated the ratio of the outstanding claims divided by the sum insured per insured object type.

Table 5: Ianos - Property Insurance Losses

Insured Object Type	Claims Number		OS Claims		Sum Insured		OS/SI ratio
Industrial	41	2.9%	3026426	11.5%	505141234	18.7%	0.60%
Commercial	776	55.4%	19018930	71.9%	1158833751	42.9%	1.64%
Residential	533	38.1%	2468258	9.3%	89773301	3.3%	2.75%
Engineering	10	0.7%	323575	1.2%	938423673	34.8%	0.03%
Photovoltaic systems	40	2.9%	1616175	6.1%	9251793	0.3%	17.47%
Total	1400	100%	26453363	100%	2701423752	100%	0.98%

From the table above it is considered that Residential and Photovoltaic panels insured objects seem to be significantly affected in terms of claims cost and capital insured. In case of photovoltaic systems, the estimated claims cost is approximately the 14.5% of the capital insured for this object. This seems valid as the specific line of business is considered very vulnerable to extreme weather conditions and especially to strong winds and hail. Commercial insured object, mainly shops, seems also susceptible to this type of weather conditions.

Frequency seems higher for the Engineering part and Commercial and Industrial

[7]HAIC - http://www1.eaee.gr/

objects follows. Frequency is defined the ratio between the number of claims divided by the number of policies in force per insured object.

Table 6: Ianos Insurance Cost - Frequency - Severity

Insured Object Type	Claims Count	Policies Count	Frequency	OS Claims	Avg Cost
Industrial	41	293	14%	3026426	73815
Commercial	776	4627	17%	19018930	24509
Residential	533	12711	4%	2468258	4631
Engineering	10	20	50%	323575	32358
Photovoltaic systems	40	497	8%	1616175	40404
Total	1400	18148	8%	26453363	18895

In terms of the severity part of the claim, industrial and photovoltaic systems seem to be more susceptible with average costs exceeding €40000.

Moreover, other lines of business beyond the Property part are affected like the Motor and the Marine lob as presented in the table below.

Table 7: Ianos - Other Lines of Business

Line of Business	Claims Number	Claims Cost	Avg Cost
Motor	1027	3293860	3207
Marine	26	1191584	45830

Between 2016 and 2020, four strong medicanes have developed in the Ionian Sea and have caused damage in Greece, with the most recent and most destructive case being the medicane Ianos between 15 and 20 September 2020. In October 2016, medicane Trixie battered the south coasts of Greece with excessive rainfall up to 180 mm in less than 12 h in Crete and wind gusts stronger than 30 m/s. One year later, in November 2017, medicane Numa hit the west coasts of Greece resulting in severe flooding, landslides, and excessive precipitation. Medicane Zorbas in September 2018 was also one of the costliest weather-related disaster in Greece in the past decade, bringing hurricane-force

wind speeds in south Greece and causing severe flooding in central Greece (up to 506 mm of rainfall in 48 h)[33]. Based on the available data, Ianos was among the strongest medicanes observed in the Mediterranean in the last at least 25 years.

Water-Related Risk

Changes in the frequency and severity of heavy precipitation events can have considerable impacts on society, including on agriculture, industry and ecosystem services. An assessment of past trends and future projections of heavy precipitation is therefore essential for advising policy decisions on mitigation and adaptation to climate change. Flooding events including pluvial floods and flash floods are directly linked to heavy precipitation hazards. The risks posed by heavy precipitation hazards are also influenced by non-climatic factors, such as population density, floodplain development, defences and land-use changes. Hence, estimates of future changes in such risks need to consider changes in both climatic and non-climatic factors. Flooding caused by heavy rainfall is a weather-related hydrological hazard associated with significant economic losses and human fatalities. The societal impact of flash floods varies and depends on the hazard intensity and the vulnerability of the affected area as measured by its topographic and demographic features, the quantity and quality of the materials exposed, the anthropogenic interventions and the effectiveness of the prevention measures taken by the local authorities.[61]. Flood-producing rainfall amounts have a significant cumulative economic impact. Despite the advance in flood risk mitigation measures, the cost of rehabilitation and compensation of citizens by the state and insurance companies is increasing worldwide.

Studies show that particularly urban areas are vulnerable to flash floods that even less important rainfall amounts can cause. Despite the progress made towards protective infrastructure, risk communication, and understanding a wide pallet of vulnerability features, rainfall events cause repetitive and eventually severe financial losses for citizens, the state, and insurance companies. Part of the exposure of elements to the rainfall hazard can be addressed by appropriate and timely reactions, starting with identifying

the potential risk occurrence related to an upcoming hazardous event. Athens Metropolitan Area has experienced severe flash floods in the past and it is considered as an area particularly prone to floods. Rainfall thresholds are identified over which the probability of a flood event occurrence is very high.[41]

The projected changes in heavy precipitation (in %) in winter and summer from 1971-2000 to 2071–2100 for the RCP8.5 scenario based on the ensemble mean of different regional climate models (RCMs) nested in different general circulation models (GCMs), according to the EEA[8] analysis, are presented in the map below.[1]

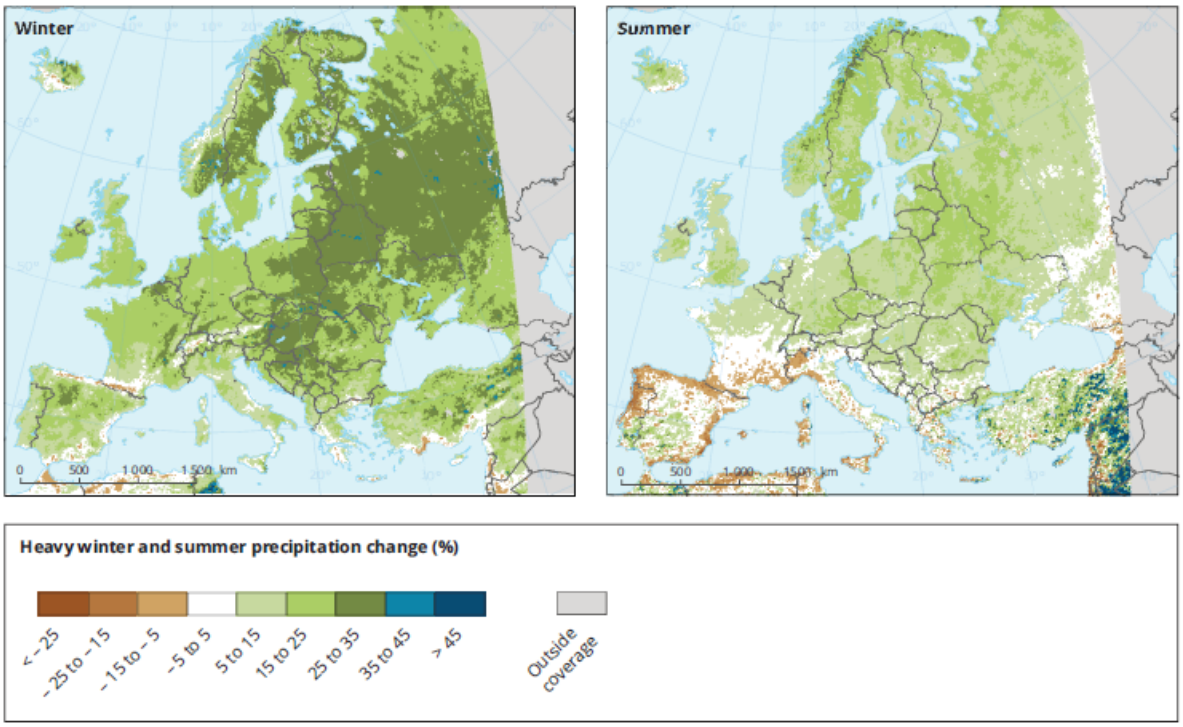


Figure 14. Projected changes in heavy precipitation in winter and summer.

Global warming is projected to lead to higher intensity of precipitation as well as longer dry periods in Europe. Modelling studies show that globally a warming

[8]European Environment Agency

atmosphere has an intensifying effect, with dry regions getting drier and wet regions getting wetter, and extremes of precipitation increasing in both the wettest and driest regions. Modelled projections of extreme precipitation events indicate an increase in the frequency, intensity and/or amount under future climate in Europe, and events currently considered extreme are expected to occur more frequently in the future. Globally, a 1-in-20-year annual maximum daily precipitation amount is likely to become a 1-in-5- to 1-in-15-year event by the end of the 21st century. Projections show an increase in heavy daily precipitation (here defined as the intensity of the heavy precipitation events defined as the 95th percentile of daily precipitation) in most parts of Europe in winter, by up to 35% during the 21st century. In summer the increase is also projected in most parts of Europe but decreases are projected for some regions in southern and south-western Europe. Similar patterns were found for other heavy precipitation indices.[1]

The Athens Metropolitan Area (AMA) of the Attica prefecture is the most populated region of Greece (3.8 Mio inhabitants), as it includes the city of Athens, the capital of the country. The climate is temperate, typical Mediterranean, and the average annual precipitation is approximately 450 mm, with the highest peaks recorded in late autumn and early winter. Regarding weather-related societal impacts, the AMA is the most affected region in Greece. According to the high-impact weather events (HIWE) database developed by the METEO[9] unit of the National Observatory of Athens (NOA), the study area suffers especially from rainfall-induced flash floods.[42]

The determination of appropriate rainfall thresholds that are likely to trigger flood damages contribute to flooding risk assessment and early warning systems. In that way meteorological peril can be estimated and capture a probable risk pattern for specific areas. During an observation period of 2012–2019, 228 rainfall events occurred in the Athens Metropolitan area, half of which (115) caused flood damages in parts of this region. Each event spatially affected from 1 up to 243 (out of the 284 in total) postal codes (M(SD) = 38.3 (56.3)), leading to 8,726 R24–DO pairs at the postal code (PC)

[9]https://www.meteo.gr/index-en.cfm

spatial analysis. The R24[10] ranged from 20.0 mm, the chosen minimum threshold, to 179.6 mm, with a mean value of 38.2 mm. The frequency distribution (histogram) of R24, at the postal code level, indicating that frequency decreases as R24 (above 20 mm) increases. In the graph below it is presented the statistical distribution (boxplot) of R24 over the binary damage occurrence variable, DO[11]. A one-way ANOVA revealed that there was a statistically significant difference in R24 between the cases (i.e., the R24–DO pairs) with and without damage occurrence ($F(1, 8724) = 387.38$, $p < 0.001$). The mean R24 was 36.8 mm among the cases without flood damages and 48.5 mm for those that caused flood damages, indicating a positive relationship between R24 and DO.[41]

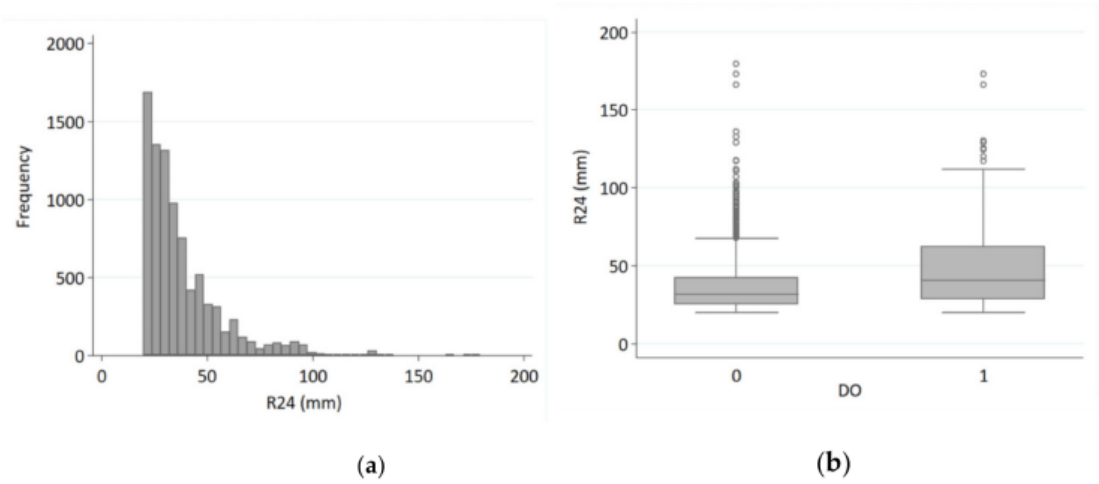


Figure 15. (a) Frequency distribution (histogram) of R24, at the PC spatial analysis. (b) Distribution(boxplot) of R24 over the binary damage occurrence variable, DO.

During the study period, the number of R24–DO pairs per PC, ranged from 2 to 66, with a mean value of 31 events. The number of damage occurrences per postal code ranged from 1 to 20, with a mean value of 4. In 12% (33) of the PCs, there was no damage claim recorded during the analysed period, while there was only one postal code in which all rainfall events caused damages. Logistic regression analysis was used

[10]Precipitation in mm observed in 24 hours
[11]DO: Damage Occurrence

to examine the effect of R24 on DO across the PCs of the AMA. The model is significant at the 5% level.

Variable	b	SE	*p* Value	95% Conf. Interval	
R24 [1]	3.03	0.18	0.000	2.68	3.39
Population [1]	1.51	0.11	0.000	1.29	1.72
Intercept	−12.87	0.54	0.000	−13.93	−11.80
N = 8726					
LR $chi^2(5)$ = 498.85					
Prob > chi^2 = 0.000					
Pseudo R^2 = 0.08					

[1] Variables were log-transformed.

Figure 16. Results for the logistic regression analysis of the probability of damage occurrence (DO) across the PCs of the AMA depending on R24 and controlled for the population.

The effect of R24 on DO was found positive and statistically significant (b = 3.03, p < 0.001). Given a specific population, an increase in R24 is associated with an increased likelihood of causing flood damages. The population was also found to affect DO positively (b = 1.51, p < 0.001) given a specific R24, confirming our initial hypothesis. In the graph below it is depicted the predicted probability (fitted values) of DO as a function of R24, controlled for the population. Results indicate that the DO probability in the AMA increases monotonically with increasing amounts of R24.

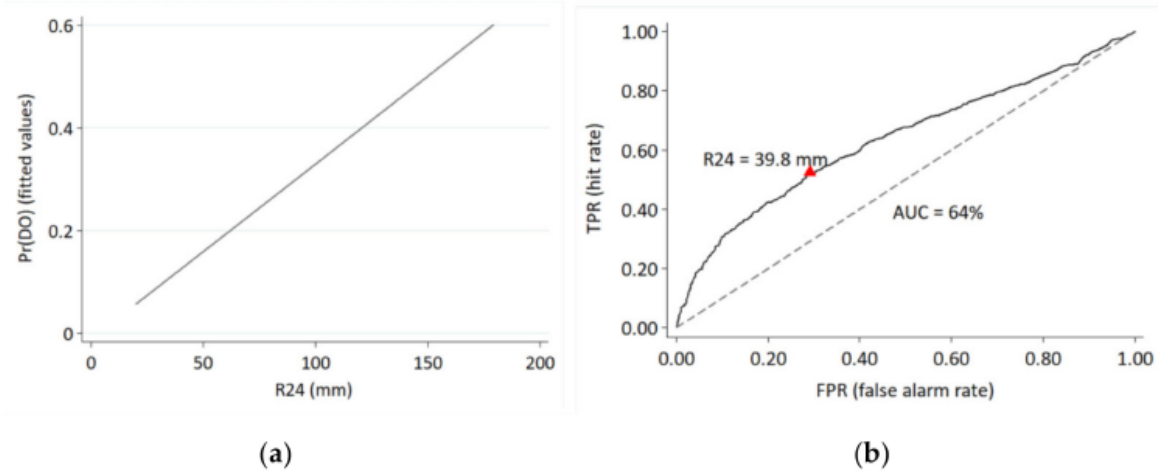


Figure 17. (a) Fitted values of DO predicted probability as a function of R24 controlled for the population. (b) ROC curve, highlighting the TPR/FPR coordinates related to the AMA optimal R24 threshold given that the difference between TPR and FPR is maximized.

Logistic regression was then applied separately for each municipality, except for the one in which there were no damage occurrences in order to examine thresholds for each case separately and indirectly to estimate the vulnerability of each region to the meteorological peril. Out of the 58 logistic models (corresponding to the 58 municipalities), only 19 (33%) were found to be statistically significant, i.e., the p-value was less than 0.05. Specifically, the sample size was found to have a statistically significant and negative effect on the R24 p-value, which indicates that a small statistical sample may be responsible for the statistical insignificance of the logistic models of some of the examined municipalities. Therefore, to overcome the limitation likely posed by the sample size to determine rainfall thresholds for as many municipalities as possible, we proceeded to merge neighboring municipalities for which the individualized logistic models were not found to be statistically significant (the p-value was greater than 0.05). The statistical results were then re-evaluated and found to be statistically significant for nine groups of merged municipalities. Overall, we defined optimal R24 thresholds for 78% of the AMA's municipalities. The optimal R24 thresholds were further classified

to estimate the local risk of flood related damage, considering that threshold rates are inversely proportional to the risk. Namely, the lower the R24 threshold associated with a high probability of damage, the higher the risk in the area since there is a higher probability of lower R24 occurrences, as shown by the frequency distribution of R24 in the Athens Municipality Area. Percentiles on the distribution of the optimal rainfall thresholds at the municipality level were used to define risk levels.

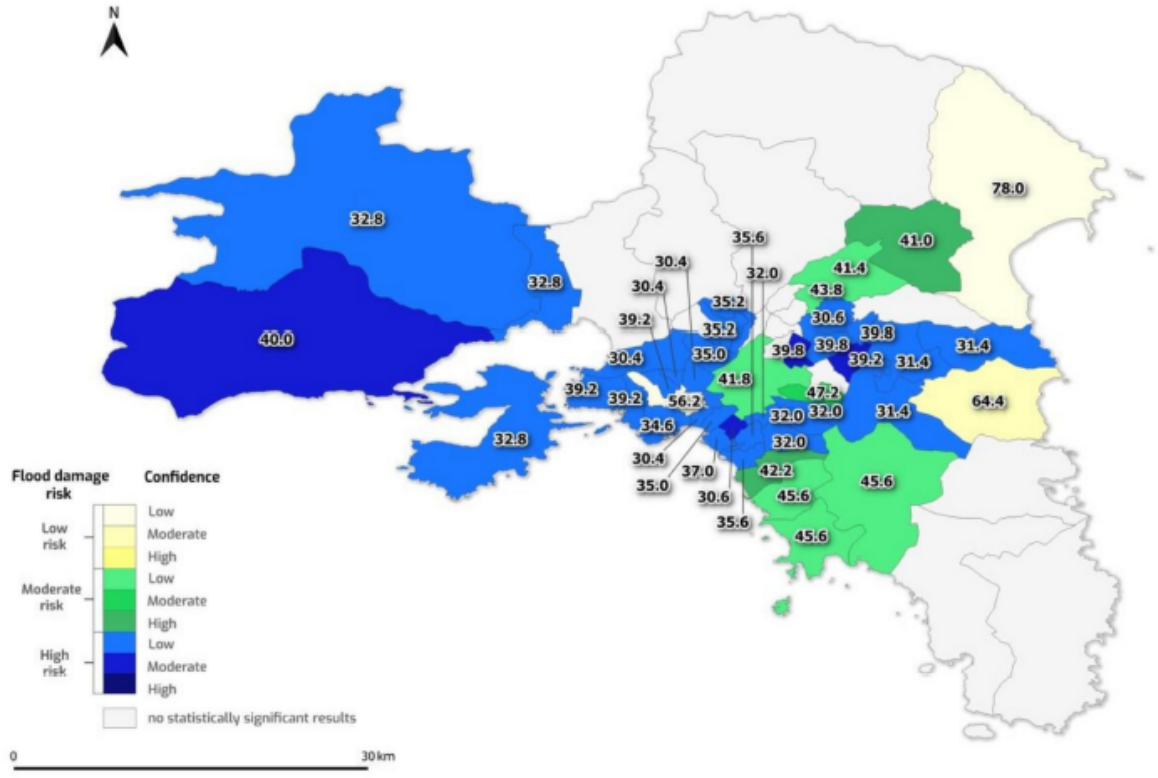


Figure 18. Optimal R24 thresholds (mm) above which flood damage is likely to occur, level of flood damage risk (3-color palette), and level of confidence in R24's discriminating performance, at the municipality level.

The optimal rainfall threshold estimated for the Athens Metropolitan Area as a whole is 39.8 mm. Papagiannaki et al.[43], who studied the relationship between flash floods and the rainfall hazard in the Attica basin using data from fire service operations

and meteorological observations, found that 50% of damaging events were associated with R24 between 30 mm and 60 mm. A relevant study using requests related to insurance claims received at meteorological services in Barcelona suggested an R24 of 40 mm as the rainfall threshold associated with societal impact in densely populated areas. Both studies highlighted the critical role for the degree of urbanization on how urban areas respond to rainfall and how susceptible can become to weather related risks. In order to expand the analysis to the whole country there are limitations that have to do with the density of the network of meteorological stations. For areas with statistical insignificance due to the lack of data, assumptions can be made although in such cases the insurance exposure is usually very limited.

Severe weather related to mid-latitude convective storms, such as floods, hail, tornadoes, and damaging wind gusts, constitute a sizable and increasing fraction of all weather-related economic losses. In convective storms, large amounts of buoyant air quickly move upward, leading to rapid condensation of the contained water vapor, releasing latent heat in the storm's updraft. Environmental vertical wind shear can further enhance the intensity and longevity of individual thunderstorms by promoting their organization, for example, into rotating supercells or into bow-echoes, coined in reference to their representation on weather radar.Such well-organized storms are the most prolific producers of convective severe weather. Anthropogenic climate change is associated with a multitude of changes in the troposphere. Such changes will affect the severity and frequency of convective storms. For instance, higher low level water vapor content amplifies convective instability, and hence, updraft intensity, while lower mid-troposphere humidity would suppress storm initiation, even in the presence of adequate instability.[48]

On November 15 of 2017, a high intensity convective storm, reaching 300mm in 13h in the core zone of the event, hit the western part of the region of Attica in Greece, causing a catastrophic flash flood in the town of Mandra and a tragic loss of 24 people, making it the most deadly flood in the country, in a period of 40 years.

The Mandra's plain is host to large industrial areas and small towns with dense urban fabric at locations and is crossed by two motorways built approximately parallel to the coastline. The average annual rainfall derived from the historical records of Hellenic National Meteorological Service of the nearby meteorological station of Elefsina is 373 mm with 59 mm occurring on average in November. The region of Attica is the part of the country with the highest frequency of floods with a rich record of floods. The area of Mandra has a history of flooding with several known events in the last decades. Existing flood databases and studies describe at least 5 flood events since the 1960s: November 1963, February 1978, January 1996 (causing two fatalities), December 1999 and February 2015.[12]

The storm event that triggered the catastrophic flash flood occurred between 14th November 2017 at 23:00 UTC and 15th November 2017 at 12:00 UTC with the majority of rainfall falling within 6h (01:00 to 07:00 UTC), indicating a particularly short duration and highly localized intense precipitation. The space-time pattern of the storm was captured by the NOA (National Observatory of Athens) X-band polarimetric radar (XPOL), which was located about 40km to the east of the area and recorded rainfall at high spatial (gate length of 150m and 0.6° angular resolution) and temporal (2min) resolution. According to the XPOL Doppler measurements (not shown here) the wind speed and direction near the surface during the core of the event was moderate to high (around 15m/s) with a general direction from the south and an eastern flow to the east of the flood area, which converged in the area of Pateras Mountain. Based on the radar-rainfall maps, the highest intensity core of the event was anchored over the catchments of Ag. Aikaterini and Soures with a southwest-northeast orientation and was approximately 18 km long and 4 km wide. This localized character of the triggering rainfall emphasizes further the great challenge that this type of storms poses on the conventional in-situ observational networks. Note that the Elefsina station mentioned previously, recorded only 27.6 mm for the same storm event highlighting the great benefit and necessity for utilizing radar rainfall observations for capturing such events. The basin-average rainfall, according to XPOL estimates, was 153 and 194 mm total accumulation and 41 and 57 mm/h maximum intensity (at 10 min temporal scale) for Ag. Aikaterini and Soures

catchment respectively[12]

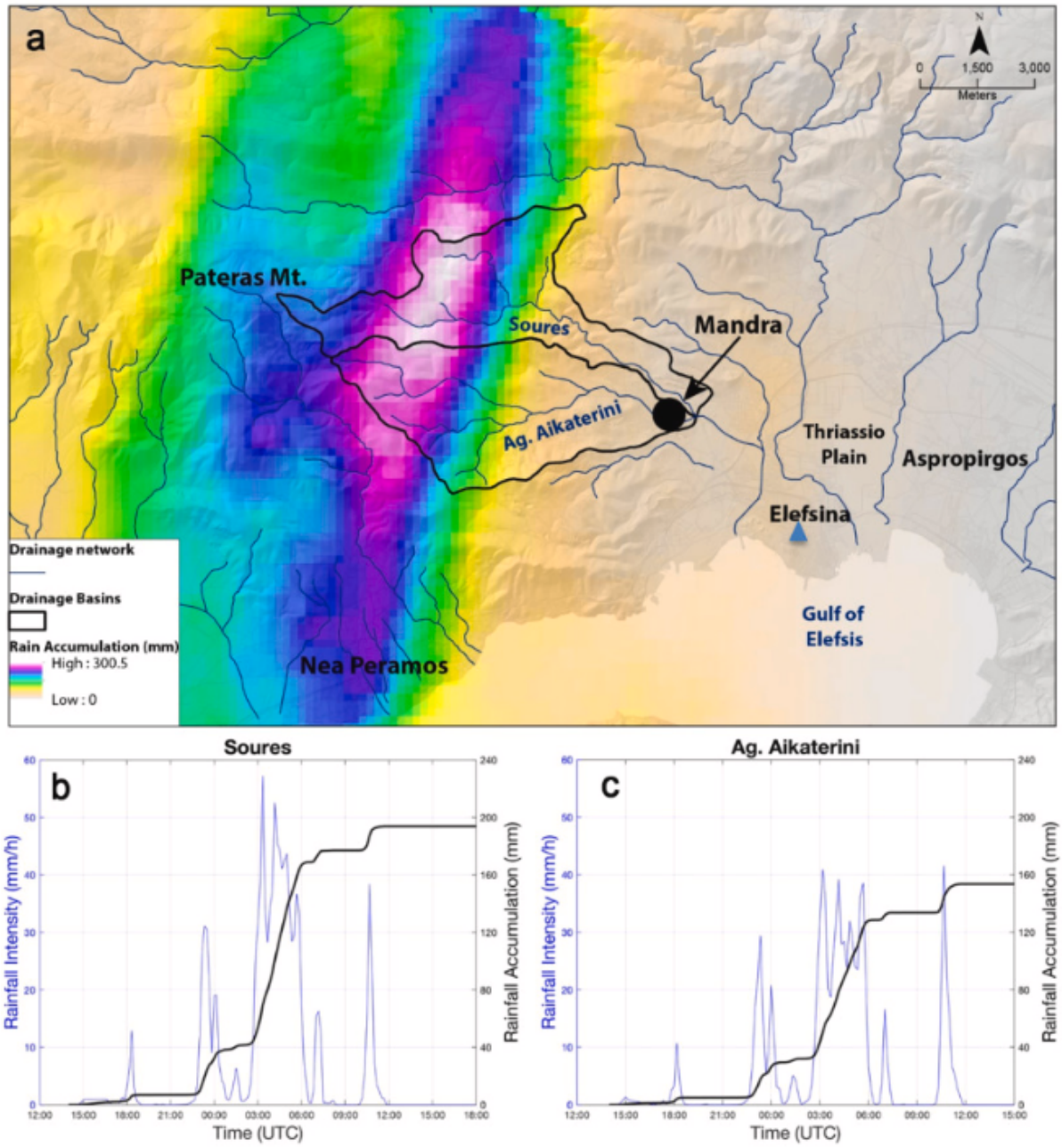


Figure 19. (a) Rainfall accumulation map of the storm event over the two catchments, along with basin average time series of rainfall (intensity and accumulation) of (b) Soures and (c) Agia Aikaterini. Note that the start/end time showing on the graphs (b,c) correspond to Nov. 14, 12:00 to Nov. 15, 15:00 UTC. The triangle denotes Elefsina meteorological station.

Hail events are among the costliest weather-related extreme events in several European regions, causing substantial damage to crop, vehicles, buildings and other infrastructure usually causing very significant insurance losses. According to the below graph climate models projections indicate more severe events in the future especially for Peloponnese area and central Greece.

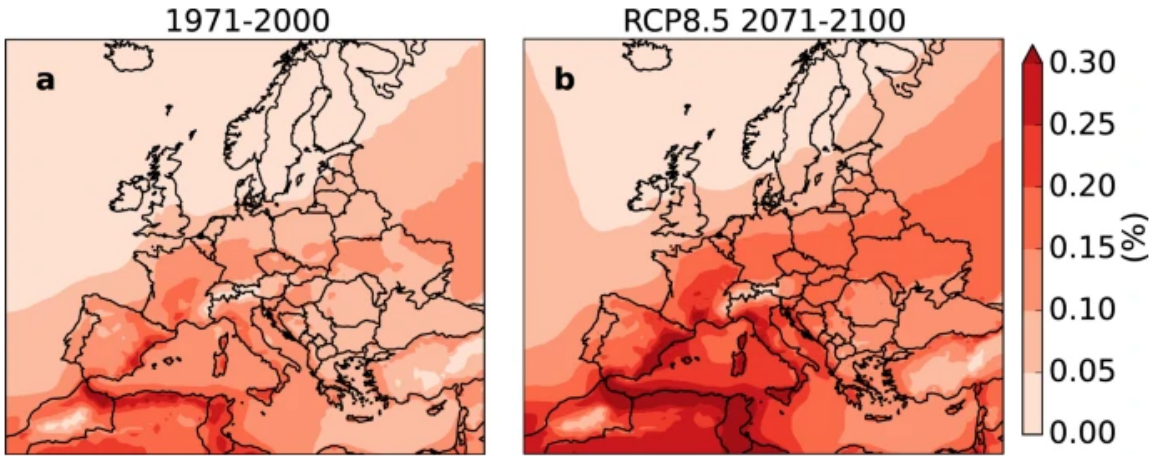


Figure 20. Percentage of lightning cases with hail 5cm for a 1971–2000 and b RCP8.5 2071–2100.

On 4 October 2019, giant hailstones of 11 cm (largest ever reported in Greece) were reported in northern parts of Attica in southern Greece. During the same day, multiple large hail reports of hailstones larger than 3 cm and 5 tornadoes were reported in the European Severe Weather Database along the track of a long lived severe thunderstorm that formed over northeastern Peloponnese and moved northeastwards to Attica and Euboea. This severe thunderstorm that produced large to giant hailstones revealed significant damages to cars (e.g. broken windshields, large dents), solar panels, solar water heaters, brick rooftops, greenhouses and tree crops. The majority of the significant damage reports were centred in the area of Oropos in northern Attica and Euboea island.[44]

In the satellite frame below colour shading denotes brightness temperatures below

− 40 ◦C. White crosses denote ZEUS[12] lightning strikes detected within 15 min of the satellite retrievals. Green triangles denote large hail reports of hailstones larger than 2 cm and the red triangles denote tornado reports from the European Severe Weather Database (ESWD).[44]

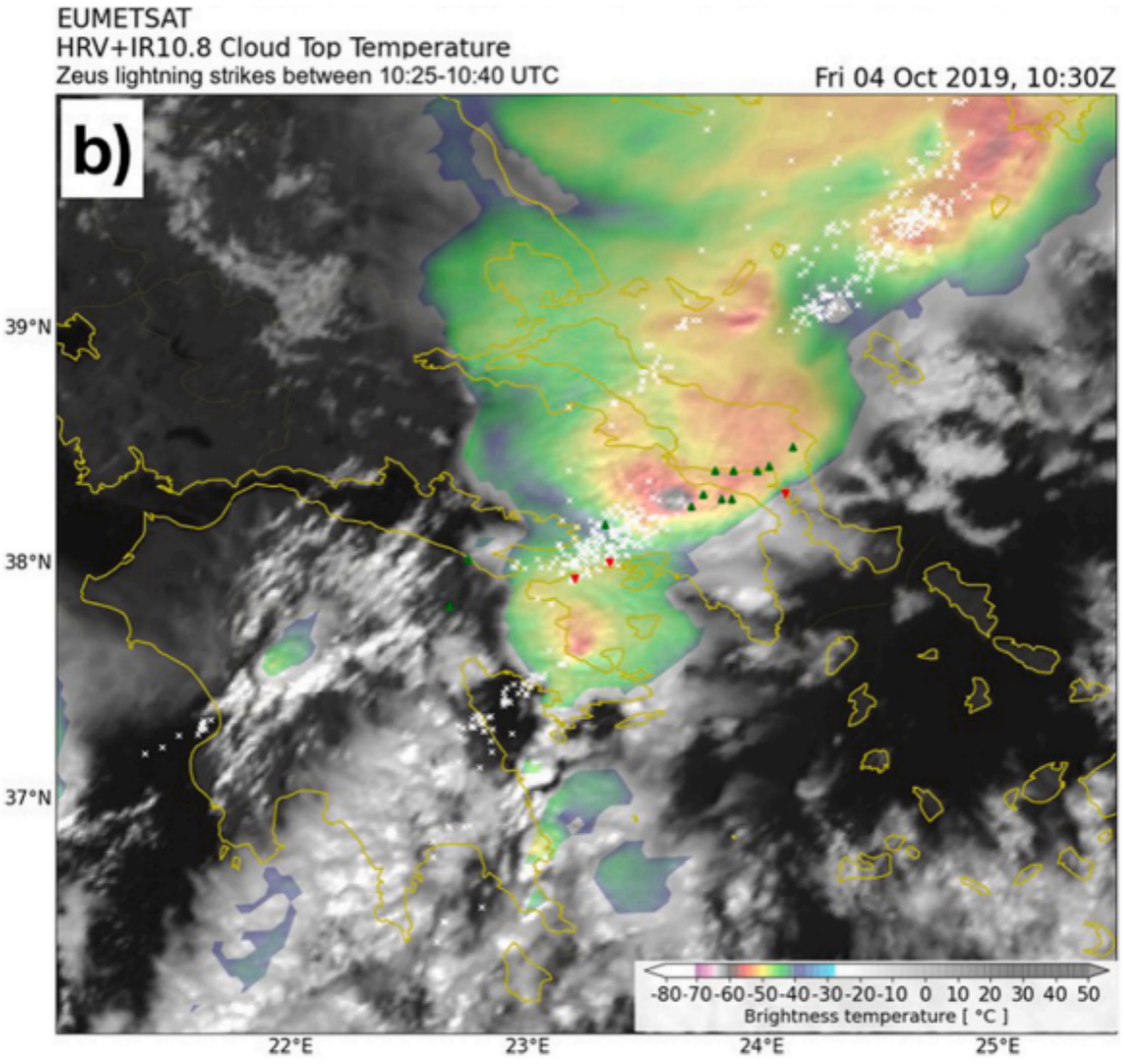


Figure 21. MSG SEVIRI infrared (IR10.8 channel) and visible (VIS) satellite retrieval.

[12]ZEUS is a long-range lightning detection network with receivers located at six sites over Europe https://www.meteo.gr/talos/en/zeus-info-en.cfm

Global warming will progressively increase storms frequency and severity in most of Europe. At the same time, the projected social and economic growth will further increase exposure to flood events. Future climate change is projected to increase the occurrence and frequency of once-in-a-century river floods in most regions of Europe, with the exception of parts of northern Europe, southern Spain and Turkey. Pluvial floods and flash floods, which are triggered by intense local precipitation events, are likely to become more frequent throughout Europe.[1]

In the tables below, claims characteristics of 22 insurance companies in Greece are depicted. Data are collected from the annual reports of HAIC.[13] In order to understand better the distribution of claims per peril and cause we select the year 2019 and at the same time compare with 2018, years with significant weather related events.

Table 8: Claims Count per peril and cause 2019 - Greek insurance companies

Peril Insured	Claims Count	Fire	EQ	Theft	Weather	Terrorism	Pipes	Plate Glass	Short	Other
Residential	12285	623	3016	1364	2868	28	2144	310	710	1222
Residential other	5	0	0	0	3	0	0	1	1	0
Hotels	768	38	8	17	449	5	54	49	92	56
Cargo	71	10	2	12	24	0	1	4	12	6
Photovoltaic systems	570	13	1	11	84	1	0	4	404	52
Super markets	422	16	7	95	46	36	6	55	37	124
Other	7743	425	357	1502	1193	540	420	1567	657	1082
Total	21864	1125	3391	3001	4667	610	2625	1990	1913	2542
2019 Distribution (%)	100%	5.2%	15.5%	13.7%	21.4%	2.8%	12.0%	9.1%	8.7%	11.6%
2018 Distribution (%)	100%	15.5%	1.1%	16.5%	21.0%	2.0%	13.4%	9.1%	6.7%	14.6%

In 2019, the 21.4% of the number of claims is referred to weather related perils which is the maximum beyond the other categories. Similar pattern is observed in 2018 with a percentage of 21% for the weather peril where at the same time "Fire" category experienced a significant increase. This increase has to do with the wild fires on "Mati" mentioned above where the fire peril is activated for the claims occurred. From business perspective it can be seen that the "Residential" insured object is mainly affected.

[13] Hellenic Assosiation of Insurance Companies

In terms of the claims cost part, it is observed a similar distribution with the weather related claims count, accounting for the 31.8% of the total claims cost. In the 2018 claims cost distribution, a major impact on the "Fire" line of business is observed due to devastating wild forest fires as described above.

Table 9: Claim Cost in millions per peril and cause 2019 - Greek insurance companies

Peril Insured	Claims Cost	Fire	EQ	Theft	Weather	Terrorism	Pipes	Plate Glass	Short	Other
Residential	24.56	3.70	6.68	2.72	5.46	0.15	3.09	0.15	0.59	2.04
Residential other	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Hotels	10.82	1.00	0.03	0.03	8.54	0.01	0.11	0.03	0.24	0.84
Cargo	0.30	0.08	0.00	0.15	0.05	0.00	0.00	0.00	0.02	0.01
Photovoltaic systems	1.12	0.06	0.01	0.10	0.23	0.00	0.00	0.00	0.60	0.13
Super markets	0.70	0.09	0.04	0.18	0.19	0.02	0.00	0.03	0.06	0.09
Other	33.50	10.40	1.33	4.97	8.12	0.82	0.87	0.97	1.02	5.00
Total	71.02	15.32	8.08	8.14	22.60	1.01	4.07	1.18	2.52	8.10
2019 Distribution	100%	21.6%	11.4%	11.5%	31.8%	1.4%	5.7%	1.7%	3.5%	11.4%
2018 Distribution	100%	56.8%	1.5%	5.4%	11.4%	0.6%	2.8%	0.7%	1.7%	19.2%

When calculating the average cost of the claims it is obvious that "Hotels", "Cargo" and then "Residential" categories are mostly affected in terms of more severe claims. Significant severity is experienced in case of "Fire" and "Weather" related perils.

Table 10: Average Claim Cost per peril and cause 2019 - Greek insurance companies

Peril Insured	Claims Cost	Fire	EQ	Theft	Weather	Terrorism	Pipes	Plate Glass	Short	Other
Residential	1999	5931	2214	1993	1903	5429	1440	468	825	1668
Residential other	3400	0	0	0	5667	0	0	0	0	0
Hotels	14091	26184	3125	1588	19027	1600	2037	653	2641	14982
Cargo	4225	8000	500	12083	1875	0	2000	750	1250	1333
Photovoltaic systems	1970	4308	7000	8727	2714	4000	0	750	1485	2481
Super markets	1666	5625	6000	1905	4152	667	500	564	1486	702
Other	4327	24471	3731	3312	6804	1522	2069	619	1554	4620
Total	3248	13614	2384	2713	4842	1656	1551	595	1317	3186

From the HAIC's[14] analysis of NAT-CAT events in Greece for the period 1993-2018 the most significant peril, beyond earthquake, is the flood related risk (45% of

[14]Hellenic Assosiation of Insurance Companies

the total claims count and 41.4% of the total claims cost for the above mentioned time period).

Table 11: Total NAT - CAT Claims estimation for the years 1993-2018 - Hellenic Assosiation of Insurance Companies

Claim Cause	Claims Count	Claims Count Distribution (%)	Estimated Cost	Claims Cost Distribution (%)	Avg Cost
Snow	646	2.8%	2400000	0.8%	3715
Heavy Rain	10412	45.4%	128600000	41.4%	12351
Forest Fire	1542	6.7%	45800000	14.8%	29702
Earthquake	10313	45.0%	133500000	43.0%	12945
Total	22913	100.0%	310300000	100.0%	13543

Table 12: Weather related NAT - CAT events for the years 1993-2018 - Hellenic Assosiation of Insurance Companies

Month	Year	Region	Claims Count	Claims Cost	Cause	Avg Cost
November	1993	Attica	358	4554600	Heavy Rain	12722
May	1994	Attica	138	1577523	Heavy Rain	11431
October	1994	Attica-Greece	1159	26077872	Heavy Rain	22500
October	1996	Attica	48	653353	Heavy Rain	13612
November	1996	Thrace	40	876478	Heavy Rain	21912
January	1997	Attica-Greece	311	6909486	Heavy Rain	22217
March	1998	Attica-Greece	1704	6023074	Heavy Rain	3535
July	2002	Attica-Greece	1350	28621262	Heavy Rain	21201
February	2004	Attica-Greece	646	2409739	Snow	3730
September	2004	Thessaloniki	387	2841395	Heavy Rain	7342
November	2004	Attica-Greece	209	1930133	Heavy Rain	9235
August	2007	Attica-Greece	232	5656344	Forest Fire	24381
August	2009	Attica	227	4095910	Forest Fire	18044
February	2013	Attica	765	5286423	Heavy Rain	6910
November	2013	Rhodes	93	3441998	Heavy Rain	37011
October	2014	Attica	1170	4296297	Heavy Rain	3672
October	2015	Attica-Greece	1065	10933233	Heavy Rain	10266
October	2017	Crete	141	5240193	Heavy Rain	37164
November	2017	West Attica	326	12303663	Heavy Rain	37741
July	2018	Attica	1083	36009746	Forest Fire	33250
September	2018	Attica-Greece	1148	7042149	Heavy Rain-Tornado	6134

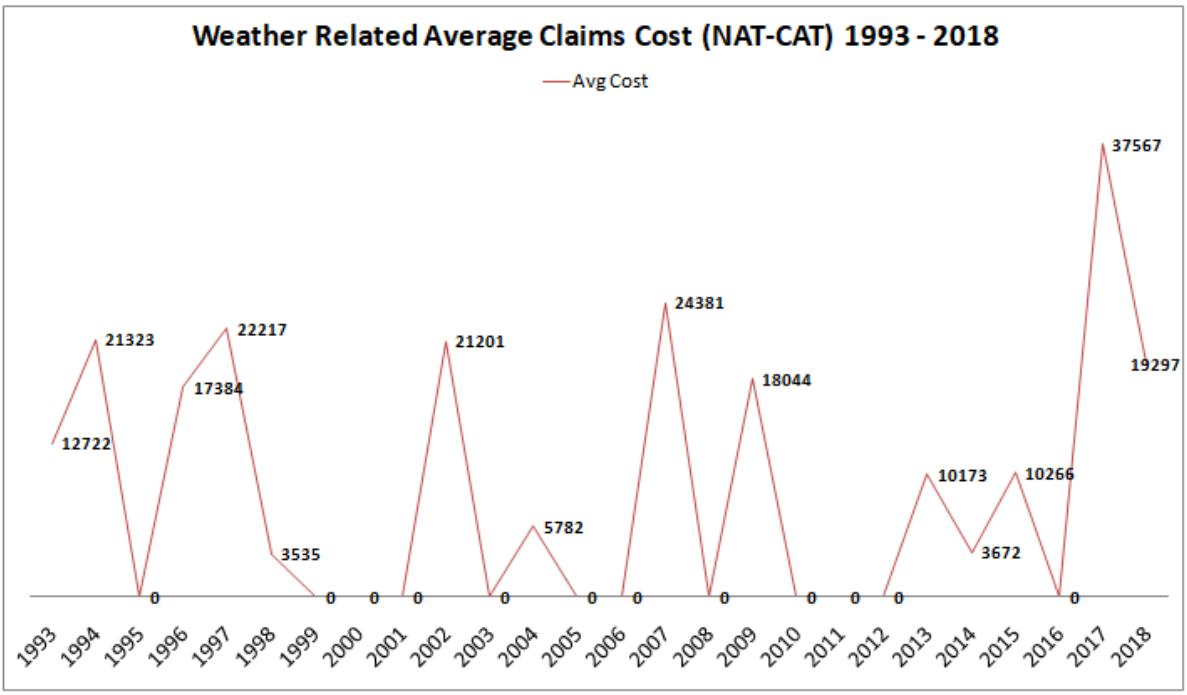


Figure 22. NAT-CAT Weather related average claims cost.

Weather related NAT-CAT events happen almost every year as mentioned in the previous table and graph with probability of occurrence of approximately 54%. The average claims cost ranges between €3535 to €37741. The average claims cost equals to the total cost of the insurance claims divided by the counts of claims for a year period of time. Extreme weather phenomena may affect all over Greece but with significant financial losses at great urban areas and industrial zones (etc. Attica region).

Regarding the 2017 maximum value of the average claims cost distribution, this has to do with "Mandra's" flood event which devastated west Attica region and affected both residential and industrial zones.

Most frequent and severe events are mainly heavy rain events which can occur at any time during the year but with high probability of occurrence during the Autumn period which denotes a significant seasonality trend.

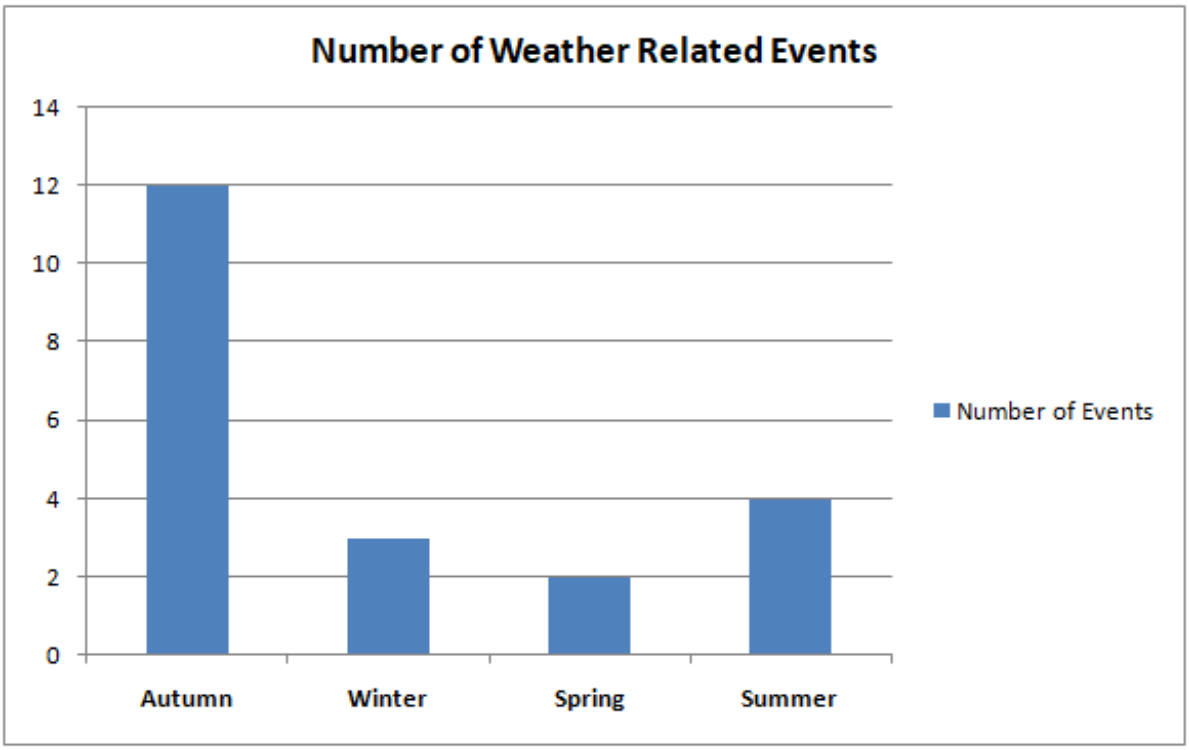


Figure 23. NAT-CAT Weather related events occurrence during a year period (observations 1993-2018).

In the period 1993-2018 Greece experienced 20 severe weather related events which had significant impact on insurance sector with total estimated incurred cost 177 mil. approximately, corresponding to 12600 claims. The events included in the HAIC's[15] database includes claims for disasters of great magnitude and intensity characterized as "NAT-CAT" events.

[15]Hellenic Association of Insurance Companies

Insurance Data Characteristics - Descriptive Statistics

Data used for the analysis derive from leading insurance companies in Greece in Property and Casualty insurance. Portfolio characteristics and particularities are revealed from the exposure data analysis and several parameters used, for example, "Construction Year" or the "Insured Object", "Location" etc. Furthermore, claims details are combined in order to make all the appropriate estimations.

The exposure database consists of policies mainly from the "Property" line of business which contains covers regarding weather related risks for the underwriting years 2012 - 2022. For each policy (with specific policy number) per year, cover, items and sub-items, all the characteristics of the insured objects are available. The Item type describes five groups of insurance contracts which are industrial, commercial, residential, property and common area insurance. The Sub-item variable describes the type of the object insured like content, building, cargo, machinery and agro insurance. The Premium variable concerns only weather related covers. The Claims database consists of claim identification codes and includes the incurred claims related to weather phenomena which are described per policy number, cover and postal code for accident years 2012-2022. Claims have been developed for accident years prior to 2022 and are well developed for 2022, as weather related claims are usually settled within a short period of time and incurred claims are anticipated to be close to the ultimate claims cost. The "Exposure" and "Claims" databases are merged together by policy ID, cover, sub-item code and postal code in order to produce a clear view of the final undertaken risk. The risk database is enriched with geospatial and meteorological information which is available per postal code. The final data-set is divided into two sub-groups, one for the Attica region and another for the rest of Greece. Covers that are usually activated by a weather event usually correspond to several types of phenomena like thunderstorms, strong winds and hail which usually coexist in a severe storm. That's why insurance companies often mix these covers together as it is very difficult for adjusters to locate the proximate cause of the experienced loss. The frequency and severity of these types of losses for Attica and other regions of Greece (mentioned below as "Other Regions") are described below,

where frequency is defined by the number of claims observed divided by the exposure variable.

Table 13: Athens Frequency and Severity table

Cover Name	Frequency	Severity
All risks	0.32%	438
Flood	0.40%	2,217
Flood-Thunderstorm-Hail-Windstorm	0.19%	433
Flood-Thunderstorm-Wind-Snow-Hail	0.66%	834
General Liability weather-pipes	0.22%	508
Hail	0.00%	0
Ice	0.01%	2,002
Snow	0.10%	3,672
Snow-Hail	0.00%	0
Thunderstorm-Windstorm	0.14%	1,251
Weather phenomena	0.54%	484
Totals	0.44%	952

From the above table it is clear that for AMA[16] the most important cover in terms of "Frequency" is the mixed cover of "Flood-Thunderstorm-Wind-Snow-Hail" which includes most weather phenomena and especially all types of precipitation and strong winds. Flood cover is also significant with an estimated frequency of 0.4%.

Severity is estimated by dividing the total incurred cost per cover by the number of claims reported. We can observe that for the "Flood" case, apart from the significant exposure ratio, the estimated "Severity" is also significant.

"Weather phenomena" is also a significant cover in terms of frequency and corresponds to all weather related risks and is usually a characteristic of Motor insurance contracts.

[16] Athens Metropolitan Area - Attica region

For the province areas of Greece the claims frequency distribution between the covers follows a similar pattern.

Table 14: Frequency & Severity of "Other Regions" of Greece

Cover Name	Frequency	Severity
All risks	0.00%	0
Flood	0.58%	13,983
Flood-Thunderstorm-Hail-Windstorm	0.28%	696
Flood-Thunderstorm-Wind-Snow-Hail	0.72%	2,402
General Liability weather-pipes	0.06%	1,095
Hail	0.00%	0
Ice	0.01%	7,688
Snow	0.02%	9,863
Snow-Hail	0.00%	0
Thunderstorm-Windstorm	0.24%	5,304
Weather phenomena	0.59%	615
Totals	0.52%	4,291

Severity impact in the "Flood" cover case seems very substantial. This is mainly driven by large claims that occur during weather events.

The most usual covers activated as a result of severe weather events are the "Flood", "Flood-Thunderstorm-Hail-Windstorm" and "Weather phenomena" covers for both the Athens and suburbs region and the rest of Greece with a total frequency of 0.52% for province areas and 0.44% for the Attica region. Severity is also significant for almost the same covers.

Claims are categorised into three severity groups, "C1", "C2" and "C3" according to the ultimate cost estimation per "Claim ID" [17]. For severity category 1 (C1), regarding claims less than €5000, for severity category 2 (C2) for claims between €5000 and €10000 and severity category 3 (C3) for claims greater than €10000, the same covers seem to have larger claims as presented in the table below.

[17] Claim Identification Code - unique for each claim

Table 15: Claims severity distribution - Athens

Claims Count Severity Category	C1	C2	C3	Totals
Cover Name				
All risks	100%	0%	0%	100%
Flood	93%	3%	4%	100%
Flood-Thunderstorm-Hail-Windstorm	98%	2%	0%	100%
Flood-Thunderstorm-Wind-Snow-Hail	98%	2%	1%	100%
General Liability weather-pipes	100%	0%	0%	100%
Hail	100%	0%	0%	100%
Ice	100%	0%	0%	100%
Snow	63%	38%	0%	100%
Thunderstorm-Windstorm	88%	12%	0%	100%
Weather phenomena	100%	0%	0%	100%
Totals	97%	2%	1%	100%

For claims incurred in the Attica region the cover which is activated and affected by large claims is the "Flood" cover. 4% of the number of claims is in the Category 3 (C3) which means that 4% of the claims occurred are greater than €10000.

A similar pattern of claims can be seen in the province areas of Greece.

Table 16: Claims Severity distribution - Other Regions

Claims Count Severity Category –>	C1	C2	C3	Totals
Cover Name				
All risks	100%	0%	0%	100%
Flood	69%	12%	19%	100%
Flood-Thunderstorm-Hail-Windstorm	100%	0%	0%	100%
Flood-Thunderstorm-Wind-Snow-Hail	89%	6%	6%	100%
General Liability weather-pipes	100%	0%	0%	100%
Hail	100%	0%	0%	100%
Ice	0%	100%	0%	100%
Snow	50%	0%	50%	100%
Thunderstorm-Windstorm	78%	5%	17%	100%
Weather phenomena	94%	6%	0%	100%
Totals	86%	6%	8%	100%

From the above tables we observe that claims tend to be more severe for province areas rather than the Athens region. In Attica, 97% of the total number of claims belong to severity category 1 (C1) in contrast with 86% of the number of claims of the province areas.

Regarding the frequency observed when taking into account both "Sub-item type" and "item type" we consider the "Residential - Building" combination as the most significant for the whole portfolio.

Table 17: Frequency Athens

Sub-Item_Type_Description	Building	Cargo	Content	Machinery	Totals
Item_Type_Description					
Commercial	0.42%	0.18%	0.28%	0.00%	0.34%
Commercial-Industrial	0.29%	0.29%	0.14%	0.19%	0.23%
Common area	0.15%	0.00%	0.00%	0.00%	0.15%
Residential	0.93%	0.00%	0.13%	0.00%	0.52%
Totals	0.70%	0.21%	0.18%	0.10%	0.44%

For the province areas we observe a significant frequency ratio in the case of the "Machinery-Commercial" combination as a lot of warehouses are located in province regions near Athens.

Table 18: Frequency - Other Regions

Sub-Item_Type_Description	Building	Cargo	Content	Machinery	Totals
Item_Type_Description					
Commercial	0.64%	0.38%	0.20%	0.43%	0.50%
Commercial-Industrial	0.63%	0.29%	0.38%	0.21%	0.45%
Common area	0.00%	0.00%	0.00%	0.00%	0.00%
Residential	0.95%	0.15%	0.00%	0.00%	0.55%
Totals	0.79%	0.26%	0.26%	0.31%	0.52%

When a claim occurs it is very difficult for the adjusters to find out what the main

phenomenon that triggered the claim was. Consequently, a mixed cover is frequently activated. In order to capture the weather risk as a whole we can aggregate covers in order to capture the total climate risk which finally affects the weather related NAT-CAT calculations.

In the table below there is a short description of the categorical variables included in an insurance database.

Table 19: Variables Description

Variables	Short Description	Categories	Categories Description
SubItemTypeCd / SubItemType Description	Sub-item insured.	C1	Building
		C2	Content
		C3	Cargo
		C4	Machinery
constructionBodycode	Manufacturing methods and materials used.	C1	Concrete
		C2	Steel and Other
FloorsType_C	Description of the type of floors insured per policy.	C1	Policies with at least insured ground floor, basement or roof
		C2	Other
ItemTypeCd / ItemType Description	Type of insured item.	C1	Residential
		C2	Commercial-Industrial
		C3	Commercial
		C4	Common area
Location	Location of insured object.	Athens	Athens and Attica region
		Other	Rest of Greece
PCause	Proximate cause of the claim.	NA	No claim
		Flood	Flood claim
		Storm	Storm related combined claim

Apart from the above categorical variables, the rest variables are described below:

Table 20: Other variables description

Variables	Short Description
PolicyNo_Flag	The policy number ID
Year	Underwriting year of the policy
postalCode	The postal code of the region
builtYear	The year of the of the building permit
nrOfFloors	Number of floors of the building
Floors_No	Number of floors insured per policy
TCapital_Insured	The sum insured per policy
TExposure	Total exposure for all weather covers
TPremiums	Total Premium written for all weather covers
TClaimsNo	Number of claims incurred for all weather covers
TIncurred	Total Cost of Incurred Claims for all weather covers

Table 21: Meteorological and Geospatial Indices

Variable	Description
floodvulnerability_indx	Indicators of flood vulnerability due to rainfall.
	Categorical variable of five levels (L1, L2,.., L5).
wind_severity_indx	Meteorological indicator based on wind intensity (eg. Knots/h).
	Categorical variable of three levels (L1,.., L3).
postcode_area	Area per postcode (m2)
postcode_perimeter	Perimeter per postcode (m)
postcode_coast_len	Length of the postcode coast (m)
postcode_flg_coast	Flag per postcode for coastal areas. Categorical variable (0-1)
postcode_alt_mean	Average altitude per postcode (m)
postcode_alt_min	Minimum altitude per postcode (m)
postcode_alt_max	Maximum altitude per postcode (m)
postcode_slo_mean	Average slope per postcode (degrees or percentage)
postcode_slo_min	Minimum slope per postcode (degrees or percentage)
postcode_slo_max	Maximun slope per postcode (degrees or percentage)
postcode_rgh_mean	Average surface roughness per postcode
postcode_rgh_min	Minimum surface roughness per postcode
postcode_rgh_max	Maximum surface roughness per postcode
postcode_corine_char	Corine data characteristics regarding the usage of the land.

Corine data characteristics are very helpful to identify the usage of the land. Visual interpretation is a method of recognizing, identifying and assessing objects recorded in aerial or satellite images. This method was applied creating the CLC database (Corine Land Cover database). It is based on the analysis and interpretation of elements of the recorded landscape objects. [18]

[18] https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/html

In order to capture the Premium distribution we observe that in the Athens area the majority of the objects insured are apartments or flats of the "Residential" line of business in contrast with province areas where most residential objects insured are in the first category "C1". This view is presented in the table below:

Table 22: Total Premium distribution

	FloorsType_C - Total Premium	C1	C2	Totals
Location	Item_Type_Description			
Athens	Commercial	88.80%	11.20%	100.00%
	Commercial-Industrial	96.70%	3.30%	100.00%
	Common area	98.90%	1.10%	100.00%
	Residential	28.20%	71.80%	100.00%
Other	Commercial	95.70%	4.30%	100.00%
	Commercial-Industrial	99.50%	0.50%	100.00%
	Common area	100.00%	0.00%	100.00%
	Residential	63.00%	37.00%	100.00%
	Totals	73.60%	26.40%	100.00%

The portfolio consists of buildings which are located all over Greece. Most of the buildings are constructed with concrete and the rest mainly from steel, stone and other materials.

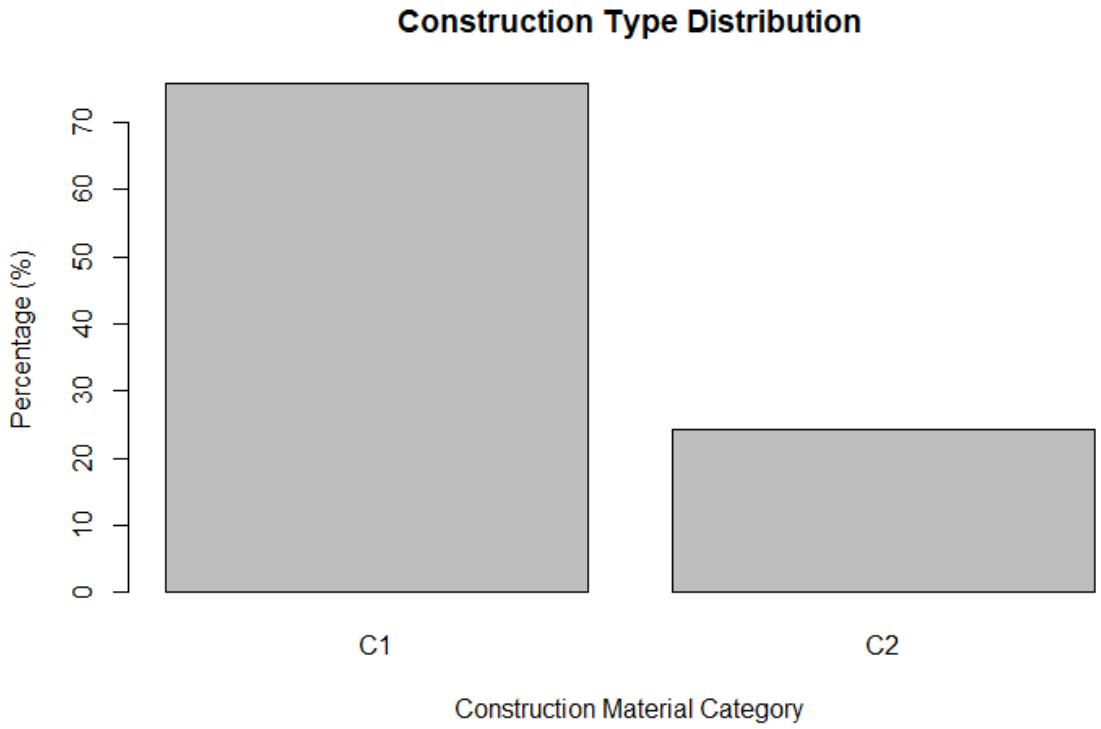


Figure 24. Construction Materials of buildings insured

Most buildings have only a ground floor as they are mainly stores but there are also residential areas which have buildings with several other floors.

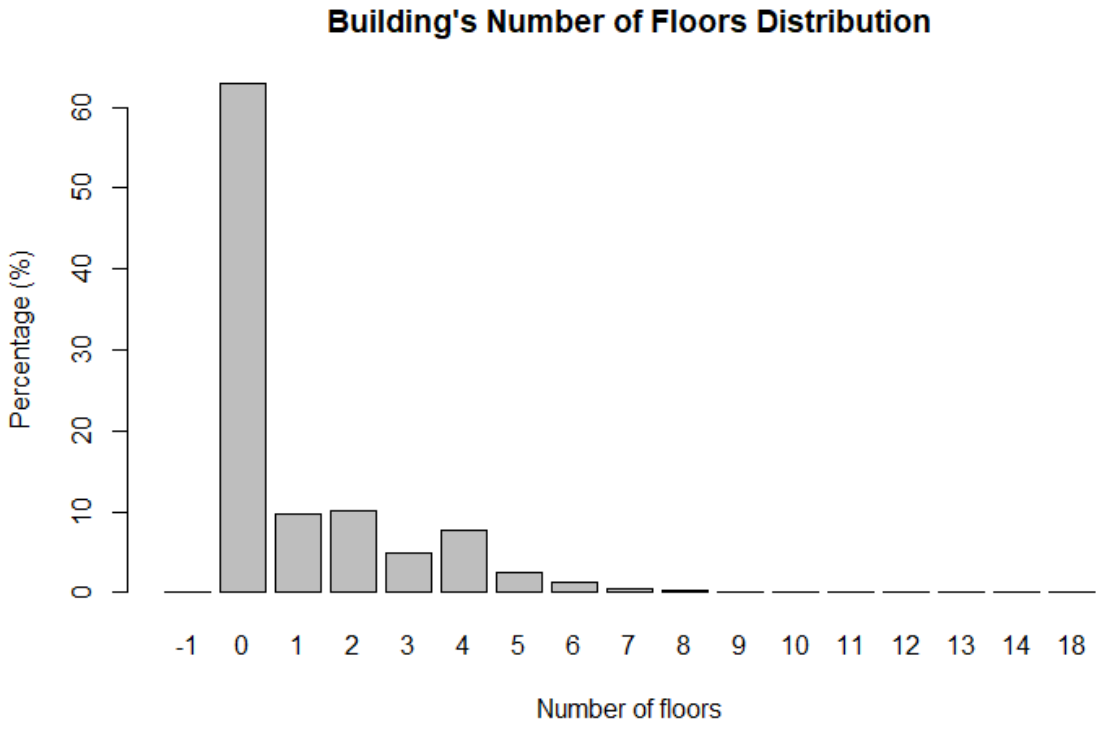


Figure 25. Number of floors of the insured buildings

In order to afford the price of a contract for a building with several floors, policy holders sometimes assume that the most vulnerable part of the building is the ground floor and the basement. As a result, they choose to insure only these specific parts in order to achieve a lower price for the contract. This approach is partially based on logic when we consider the experience of the portfolio, but in many cases serious claims may occur also in an apartment on a specific floor due to other weather phenomena such as hail, strong winds, thunderstorms or ice.

Most buildings in the portfolio were built during the decades of the sixties and the seventies when older urban regulations were in place.

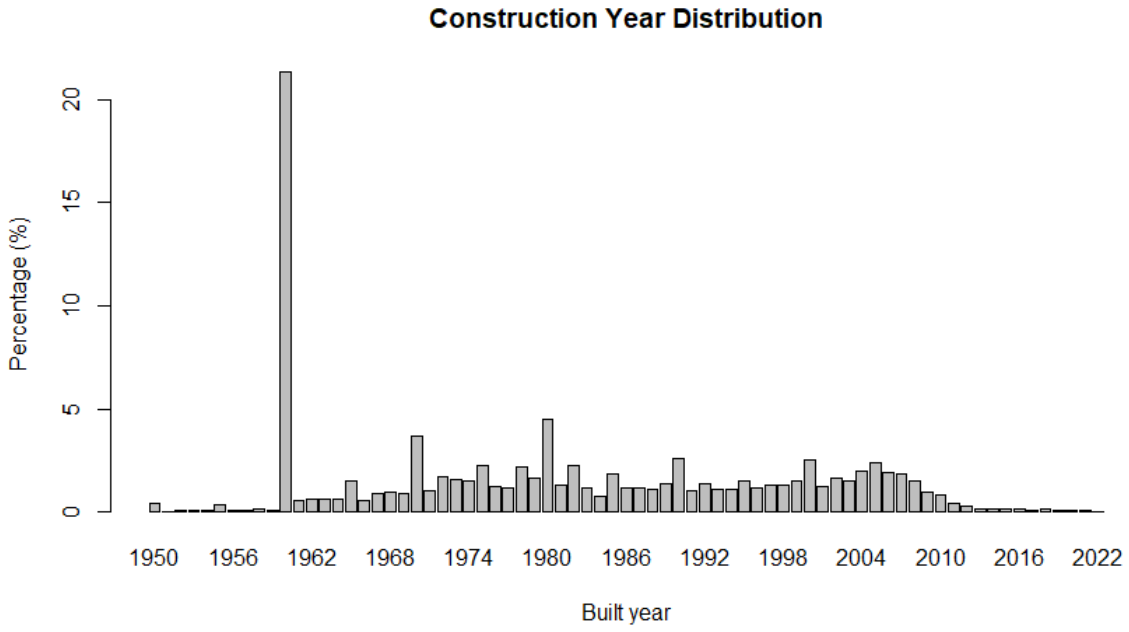


Figure 26. Construction Year distribution of the buildings insured

From the above graph we can observe the distribution of the buildings built in recent years peaks in 1960. This is related to the treatment of cases where the construction year is not clearly reported in the insurance company. When a building was constructed during the 1950's, but the issue year of the construction licence is not given to the company, the year "1960" is used as an assumption.

The portfolio consists mainly of residential property, but the commercial part has also a significant number of policies.

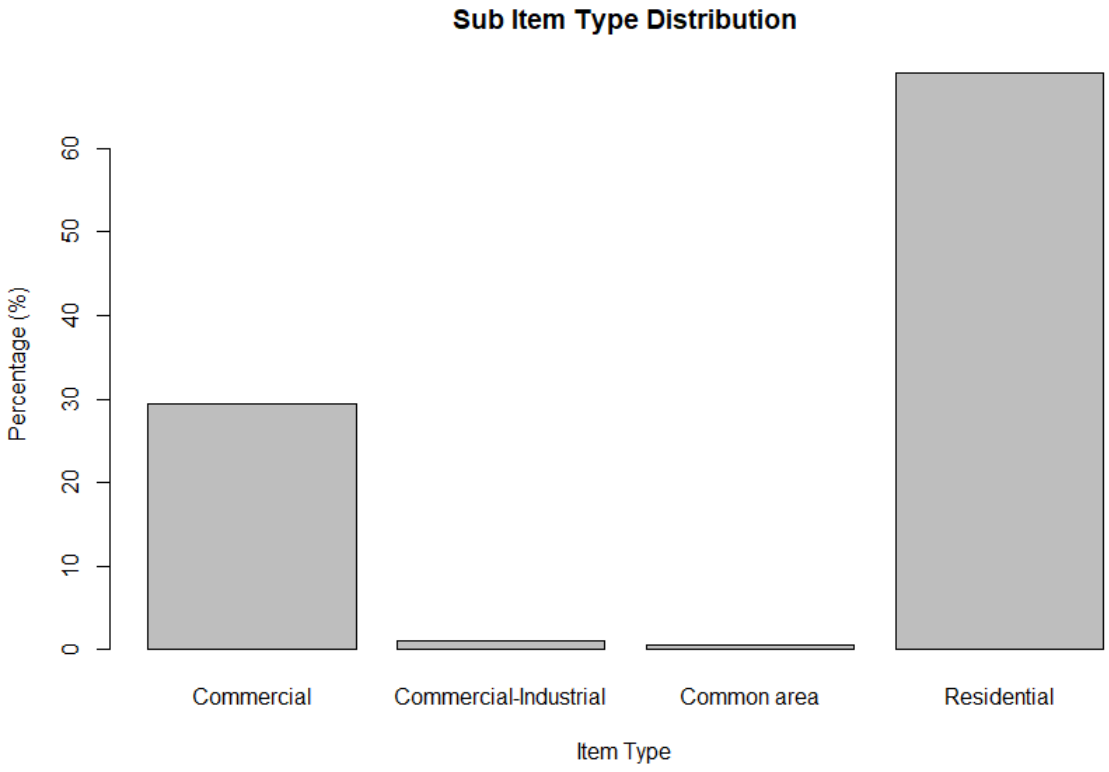


Figure 27. Item Type distribution

Commercial insurance protects businesses from losses due to unexpected events during normal business operations, such as lawsuits, natural disasters, or accidents. There are many types of commercial insurance for businesses, including coverage for property damage, legal liability, and employee-related risks, among others. Companies evaluate their commercial insurance needs based on potential risks, which can vary depending on the type of business and its environment. Residential insurance covers losses and damage to an individual's residence, along with furnishings and other assets in the home but also damage to the building itself. Industrial insurance covers include protection from weather related risks as well. All the above item types of insurance are

considered to be very vulnerable to extreme weather conditions.

The two basic parts insured in terms of "Property" weather related risks are the building and the contents protection.

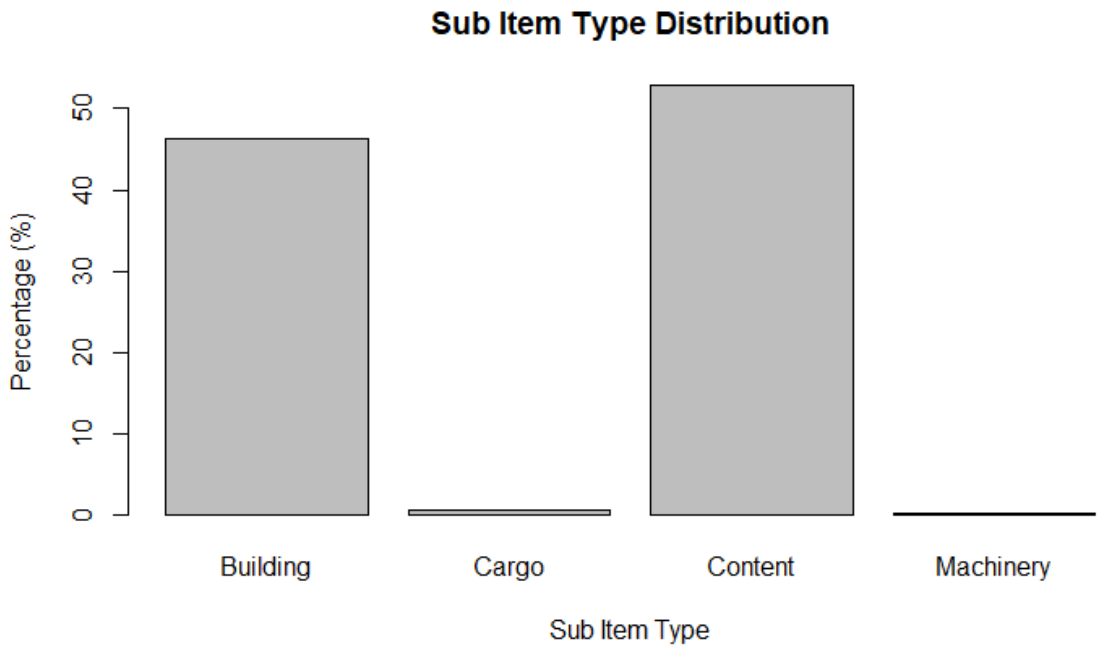


Figure 28. Sub-Item Type distribution

In terms of flood risk, content is very often affected. Usually the parts of the building affected are the ground floor and the basement. In the case of hail, extensive damage may occur also to the roof. Moreover, mixed claims can also occur in case of a lightning strike where we have extensive damage to the electrical installation of the building but also to the content as the electrical equipment of the house is very susceptible to this type of risk.

Cargo insurance safeguards shipments from loss, damage, or theft while they are in transit, to ensure goods are shipped safe and sound. Cargo insurance against extreme

weather conditions can be very beneficial as cargo is extremely vulnerable to weather phenomena. Machines insurance is generally connected with industrial risk where a weather phenomenon (e.g. lightning) can cause serious financial losses when essential equipment of the factory is affected (machine damage, business interruption, etc.).

Main Distributions Fit for Frequency and Severity.

Large Claims Identification

In order to fit specific distributions for frequency and severity and recognize possible specific patterns we separate the data-set into two groups, one for the Athens portfolio and another for the rest of Greece. In order to have homogeneity in terms of the claims distribution, we examine the data-sets for outliers and select an optimum threshold.

Table 23: Greece Claims Distribution Percentiles

Quantiles	Athens Claims Cost	Other Claims Cost
0%	30	30
5%	91	150
10%	92	197
15%	92	244
20%	144	294
25%	170	361
30%	199	422
35%	240	500
40%	272	602
45%	312	733
50%	356	850
55%	413	1,011
60%	492	1,251
65%	583	1,518
70%	673	1,930
75%	850	2,477
80%	1,000	3,310
85%	1,329	4,887
90%	2,000	8,599
95%	3,365	15,560
100%	84,126	694,352
Average:	1,025	4,648
Variance:	10,529,932	600,347,768
perc(99.5%)	13,860	83,527

From the above table it can be seen that we have larger claims in the province areas. This is partially explained by the occurrence of one of the most important meteorological events of the current decade and the catastrophe zone that was mainly affected. Specifically, the "Ianos" event, which was considered a Mediterranean cyclonic system, affected mainly the province areas of Greece and Athens to a lesser extent.

Below there is a "Hellenic's Association of Insurance Companies" map which depicts the areas affected by that event.

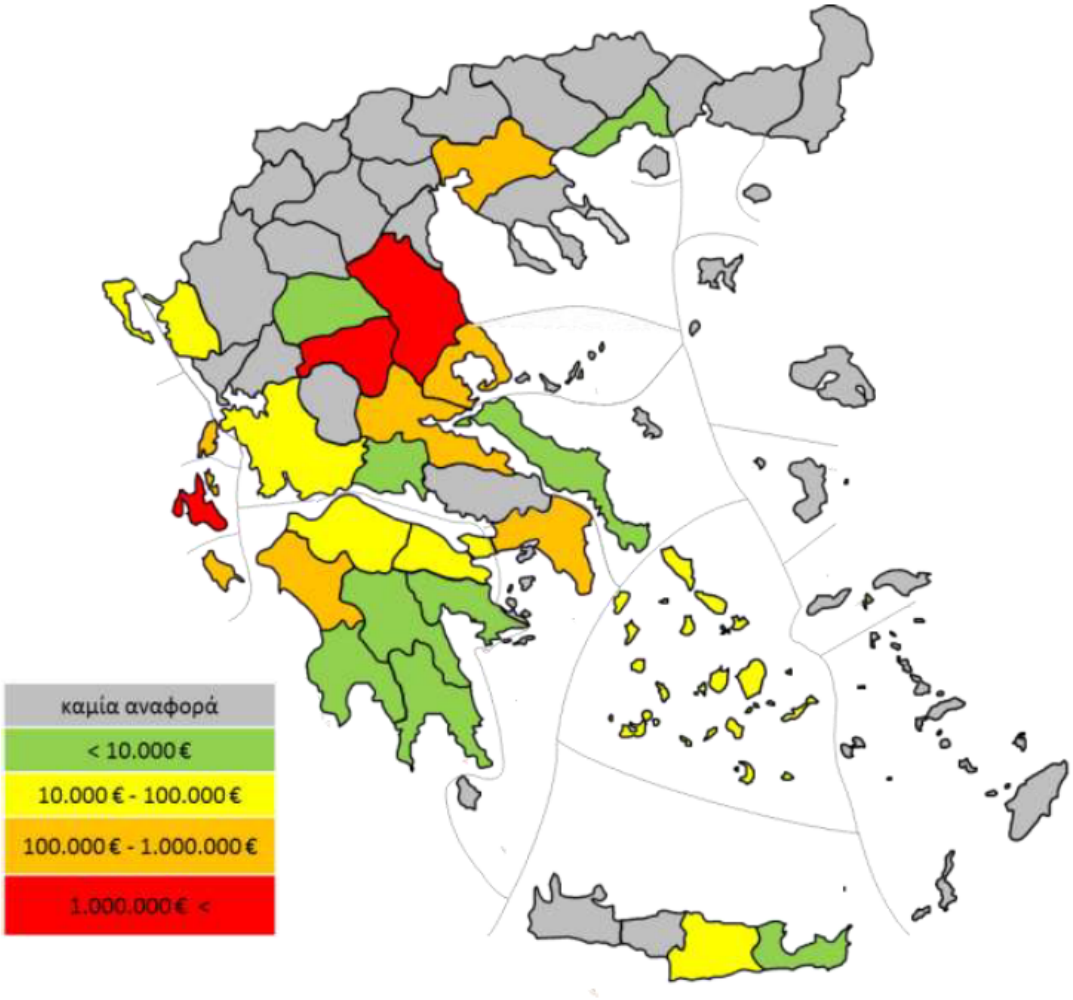


Figure 29. Ianos – Claims Distribution Greece

In the case of Athens claims, we see that an option for the threshold could be the amount of €10000. For consistency between the two data-sets the same threshold can be used for the province areas as well. The distribution of a random variable X with distribution function F has a heavy (right) tail if the moment generating function of X, $M_x(t)$, is infinite for all t > 0. In this case we have:

$$\int_{-\infty}^{\infty} e^{tx}\, dF(x) = \infty \quad \text{for all } t > 0. \tag{8}$$

Using a tail distribution function expression this equals to $\overline{F}(x) \equiv \Pr[X > x]$ as

$$\lim_{x\to\infty} e^{tx}\overline{F}(x) = \infty \quad \text{for all } t > 0.$$

With $X_1, X_2, ..., X_n$ and $X_i > X_{i+1}$ and calculating the $\ln\left(\frac{x_i}{x_{i+1}}\right)$ we have the below graph:

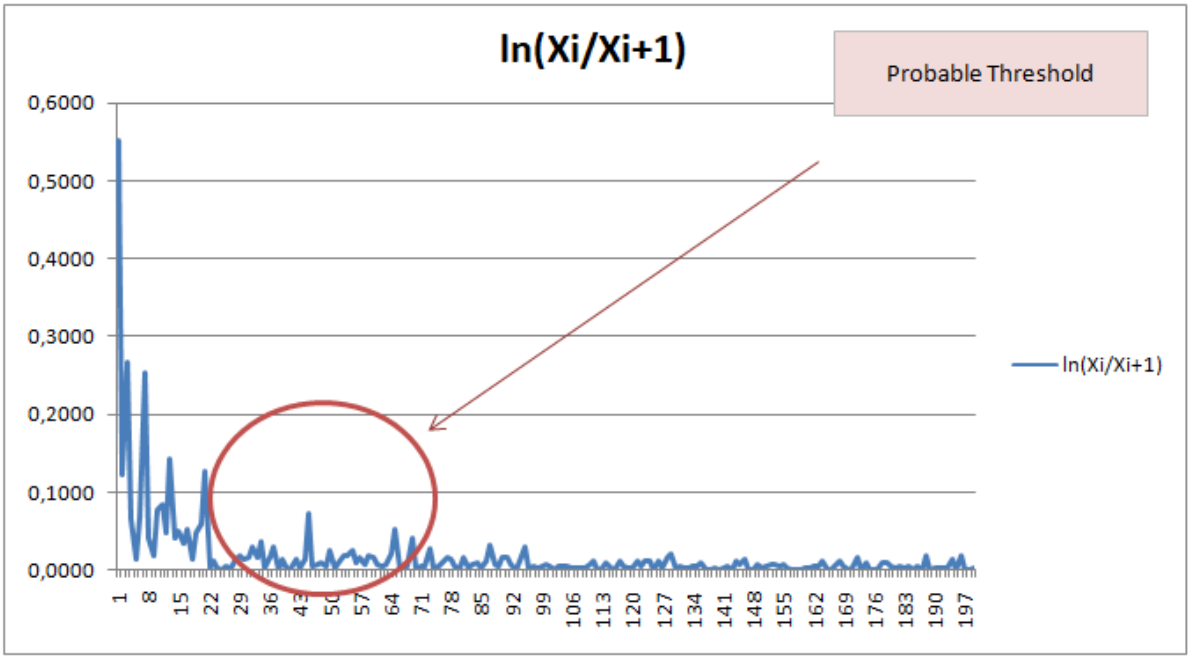


Figure 30. Large Claims Threshold selection

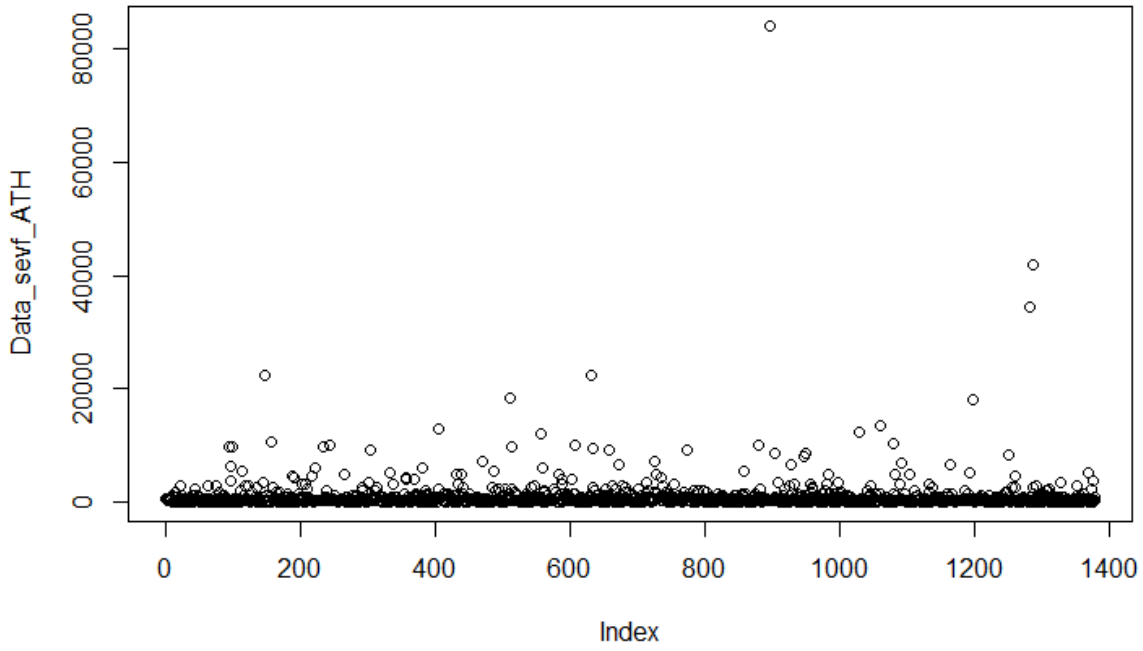


Figure 31. Athens Large Claims plot

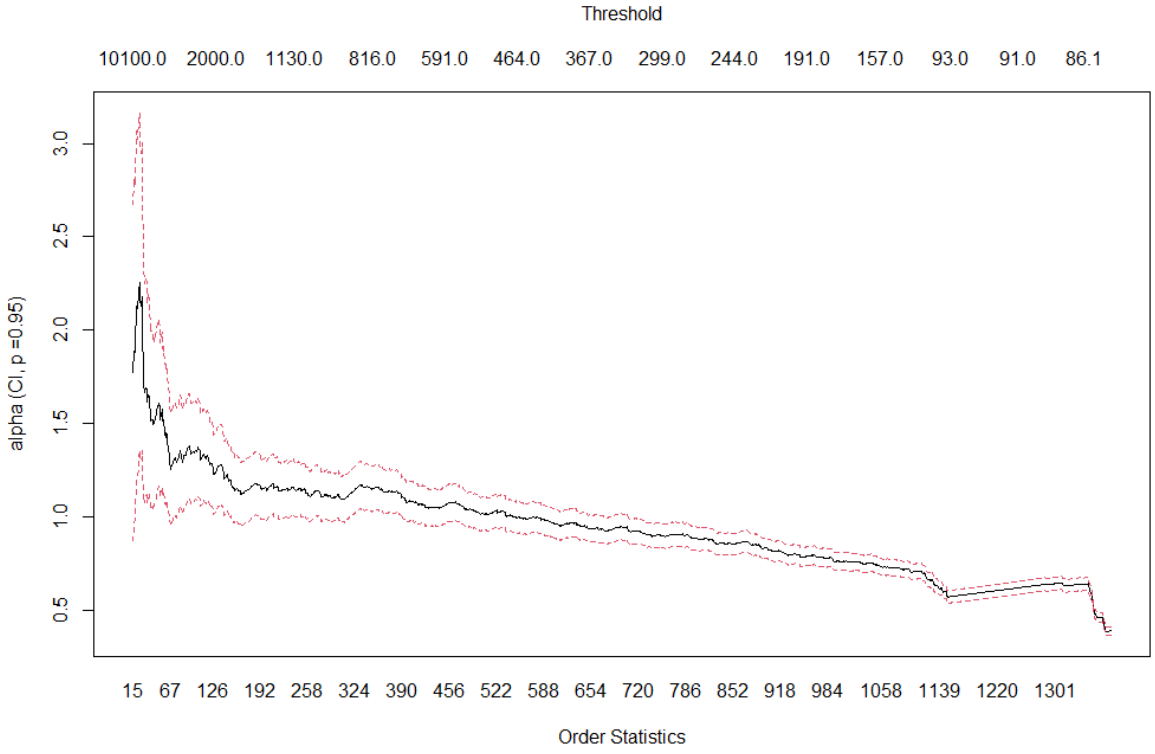


Figure 32. Athens Hill plot

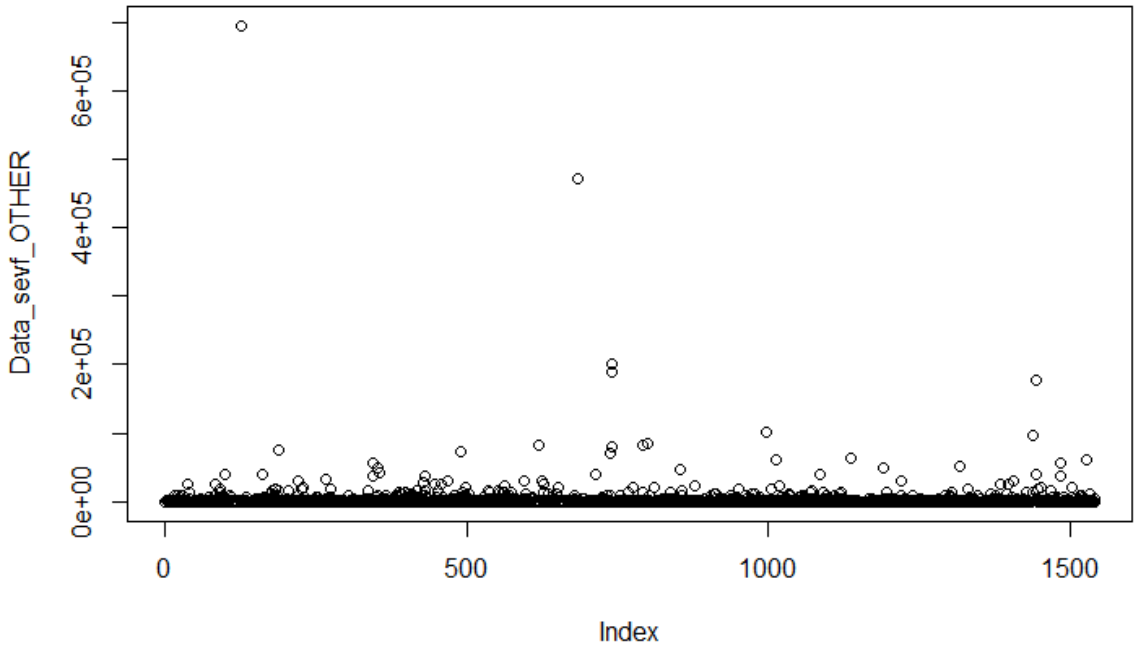


Figure 33. Other Regions Large Claims plot

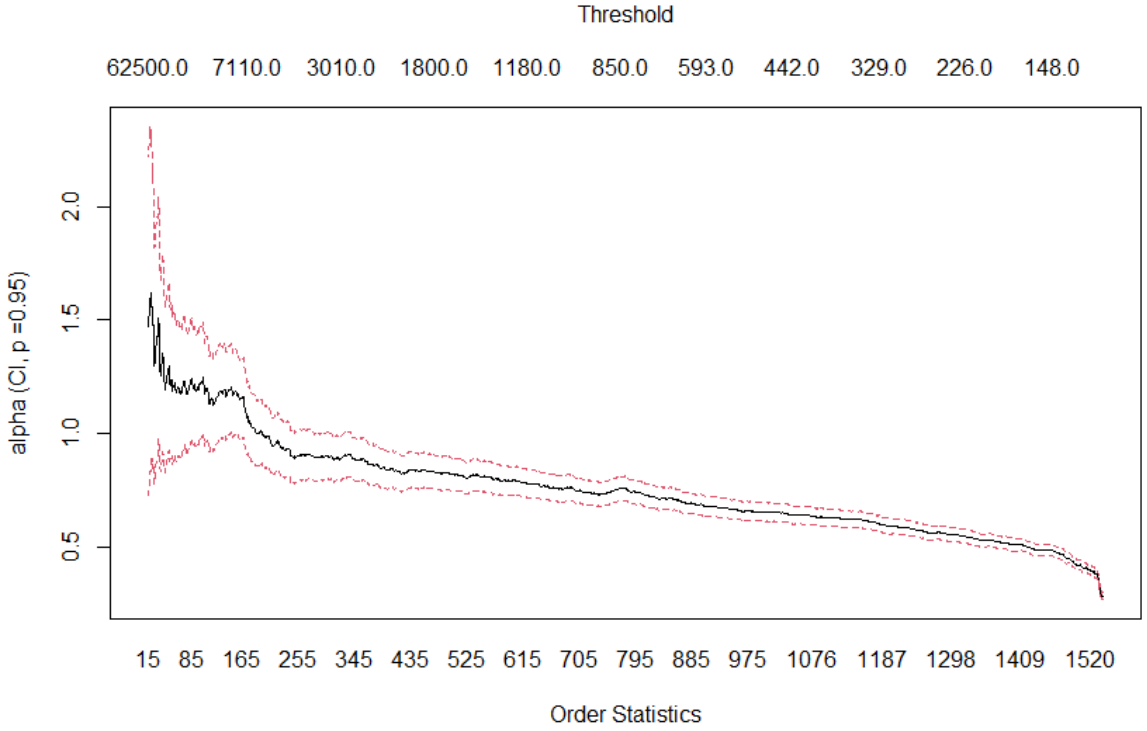


Figure 34. Other Regions Hill plot

Distributions Fit - Frequency

For both data-sets (Athens and Other Regions) we fit Poisson, Negative Binomial and Poisson Inverse Gaussian distributions using the package "Gamlss" in "R".

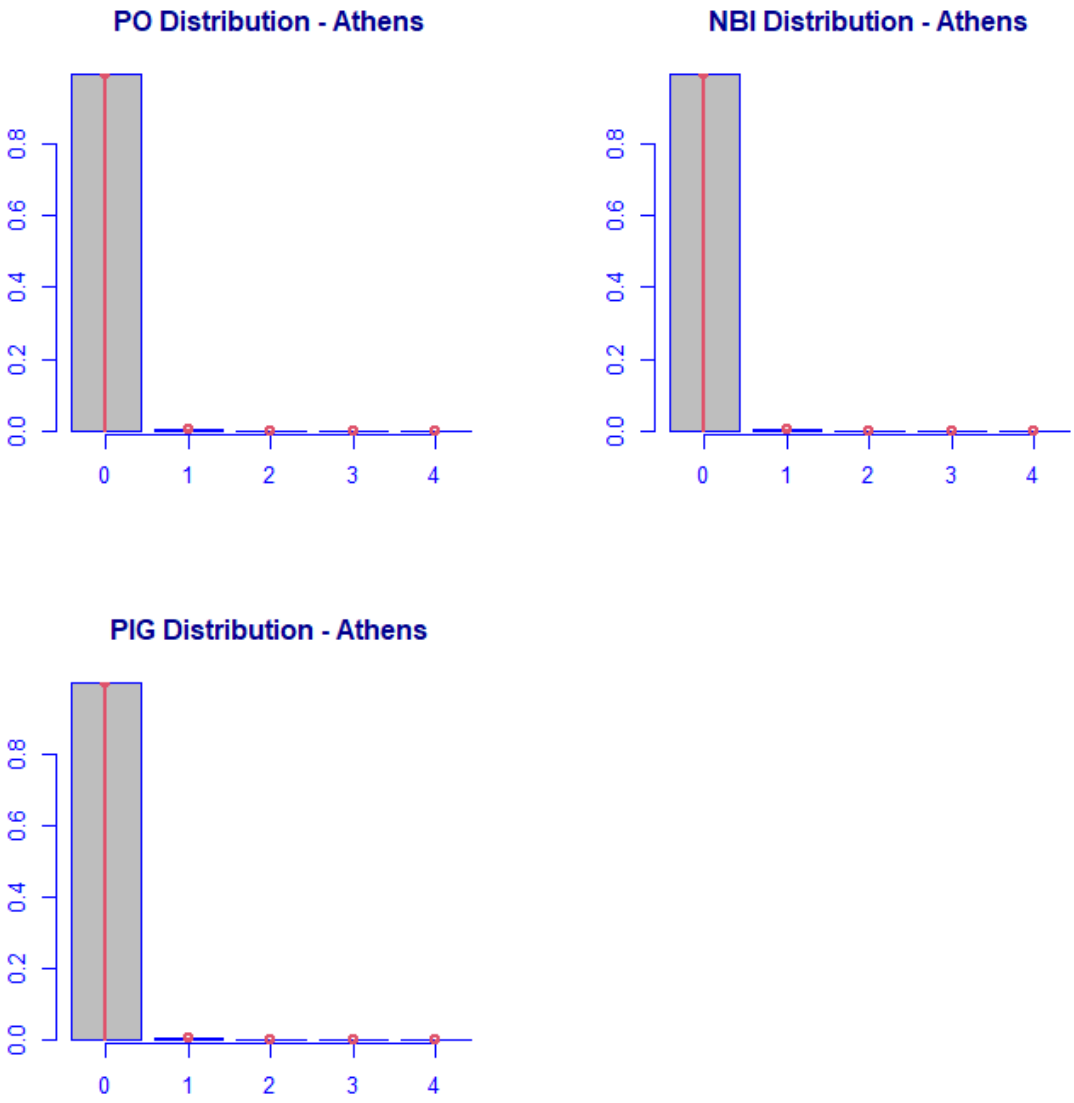


Figure 35. Frequency Distributions Fit - Athens

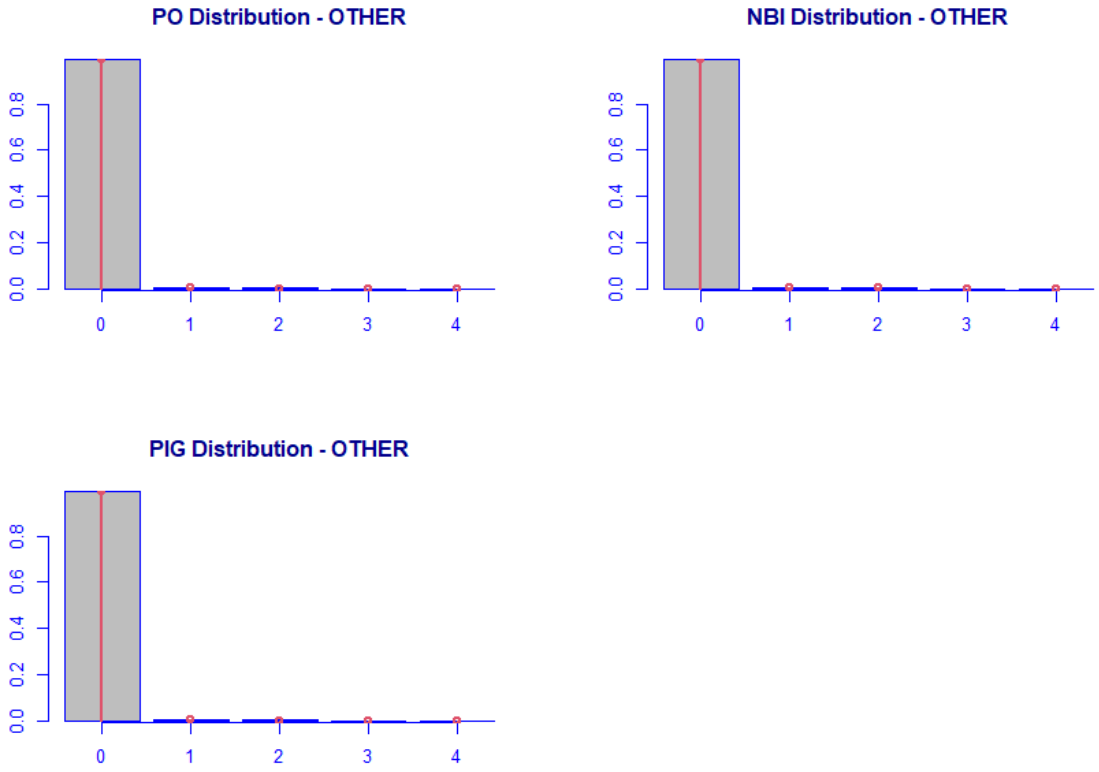


Figure 36. Frequency Distributions Fit - Other Regions

Table 24: Frequency Distributions Gamlss

Gamlss Frequency Distributions	Location	Global Deviance	AIC	SBC
Poisson	Athens	19624.8	19626.8	19637.6
	Other	21473.4	21475.4	21486.2
Negative Binomial	Athens	19041.0	19045.0	19066.7
	Other	20732.9	20736.9	20758.4
Poisson Inverse Gaussian	Athens	19046.5	19050.5	19072.2
	Other	20735.9	20739.9	20761.4

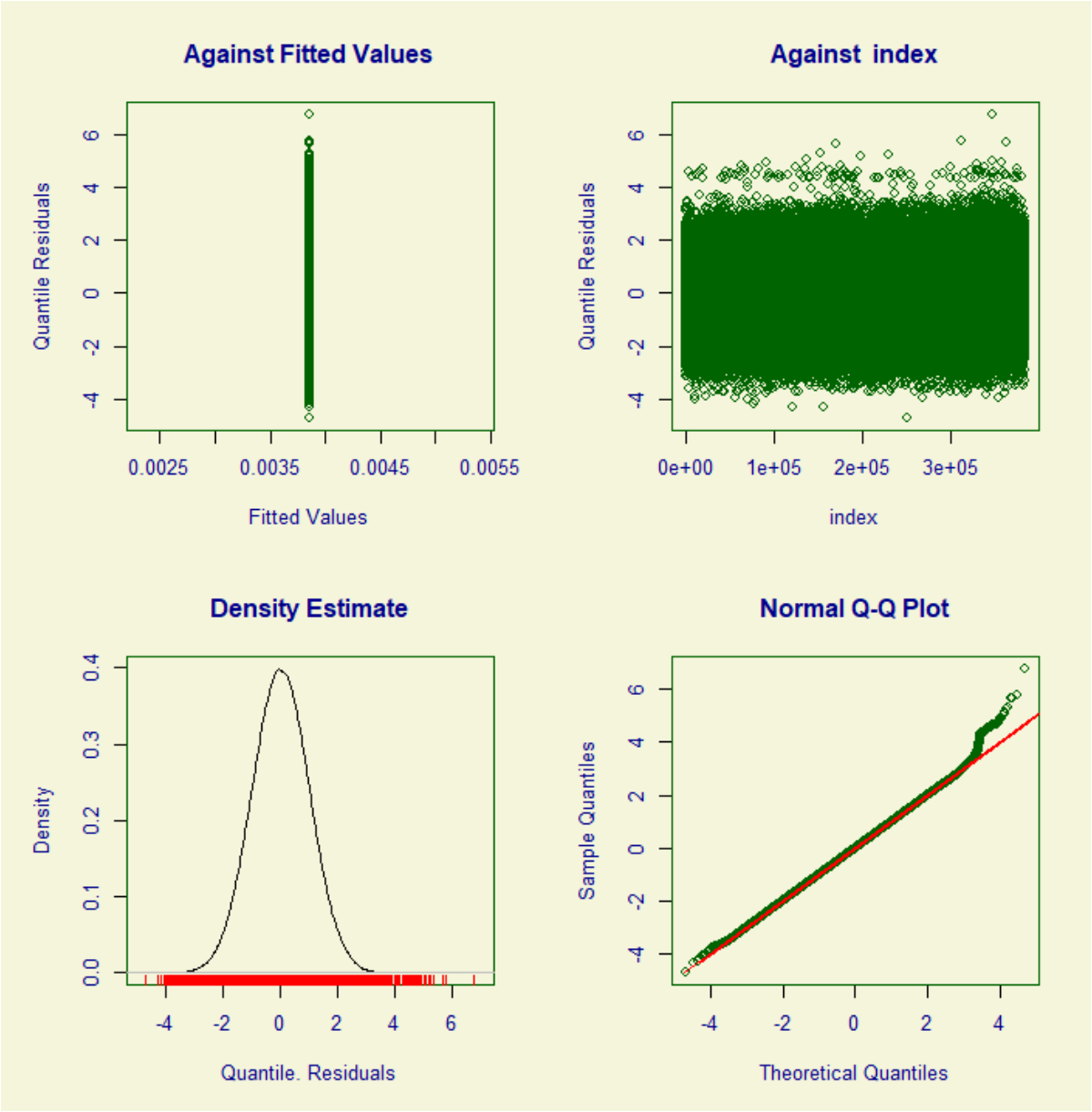


Figure 37. Poisson GoF - Athens

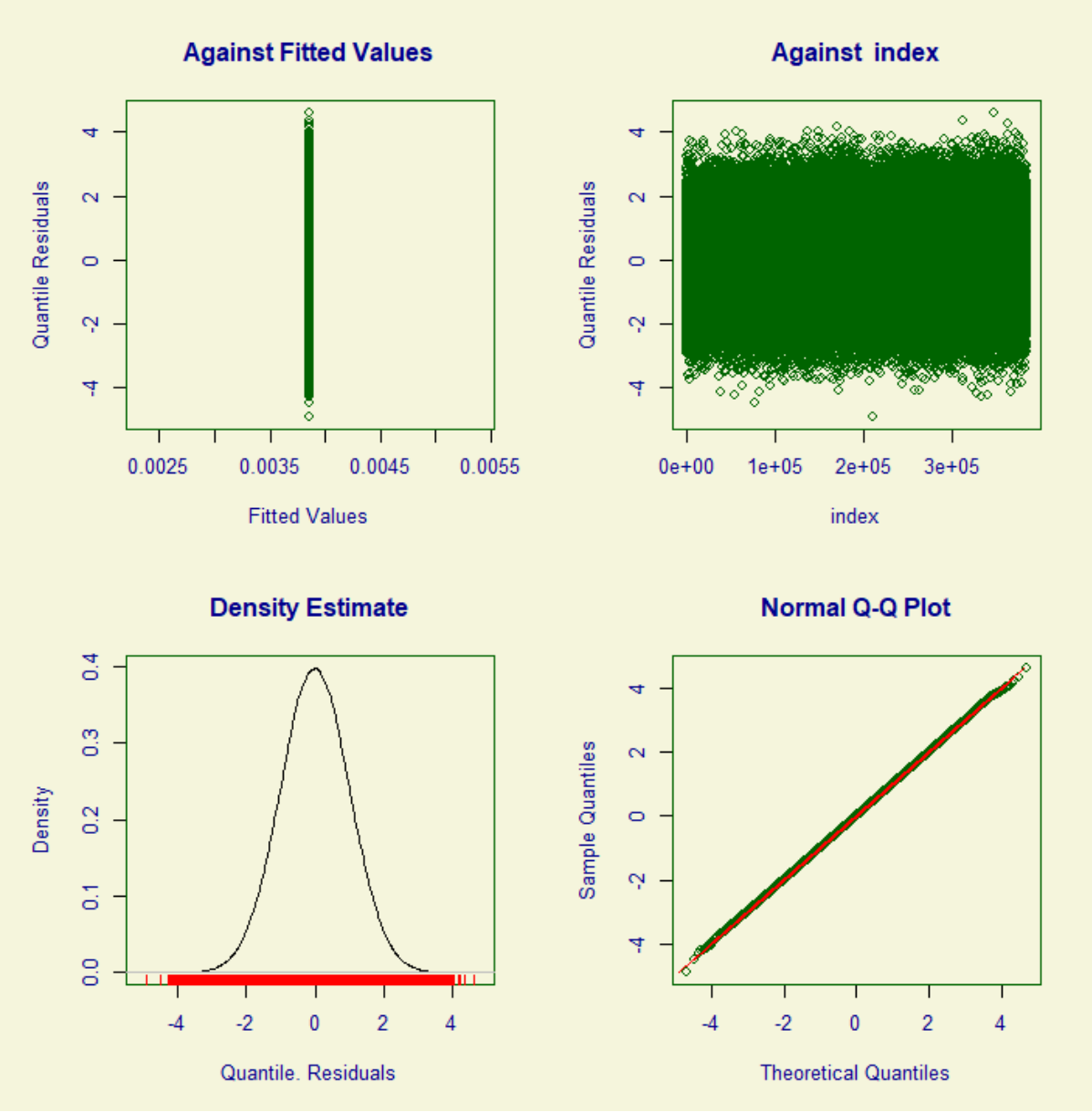


Figure 38. Negative Binomial GoF - Athens

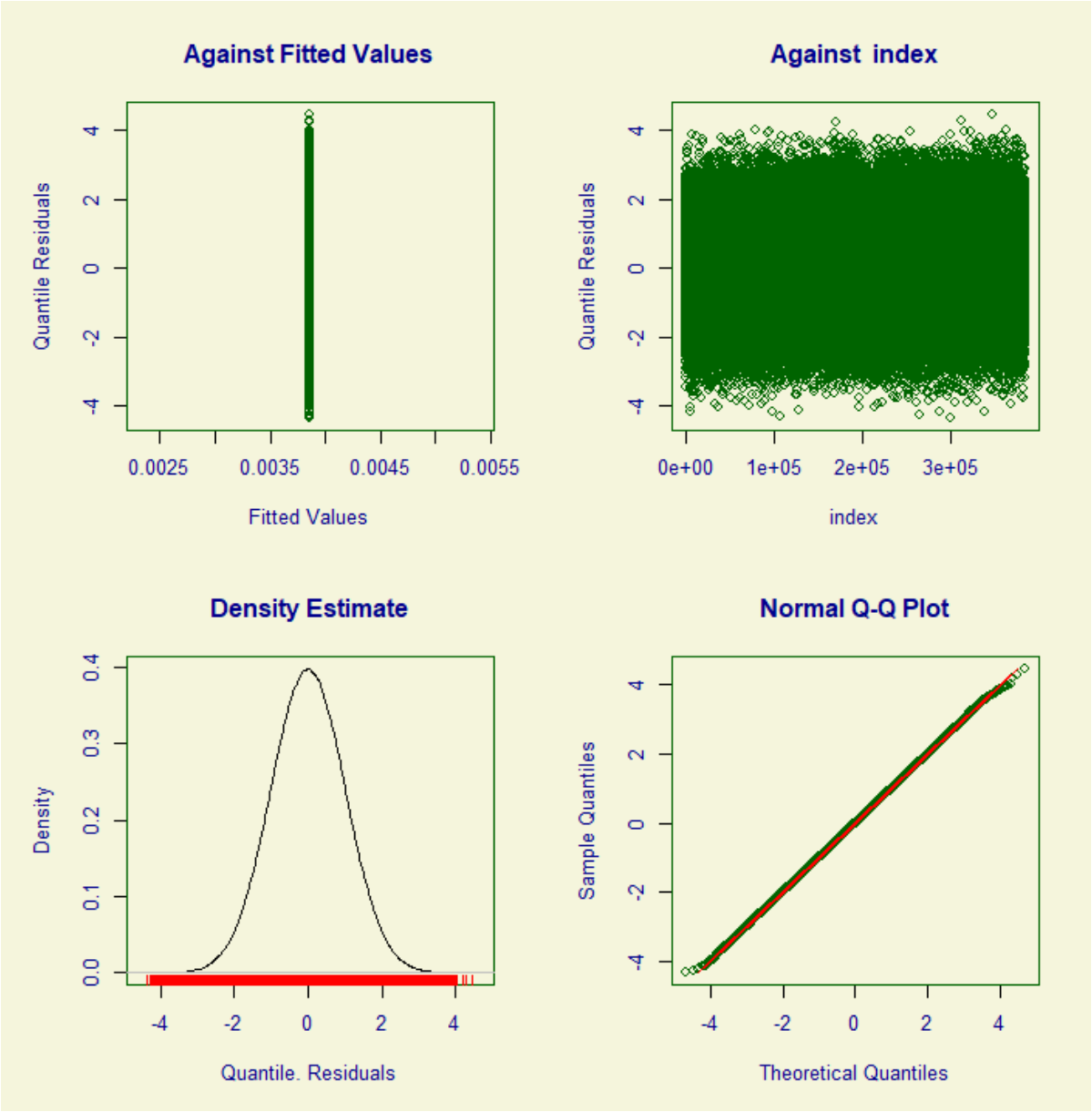


Figure 39. Poisson Inverse Gaussian GoF - Athens

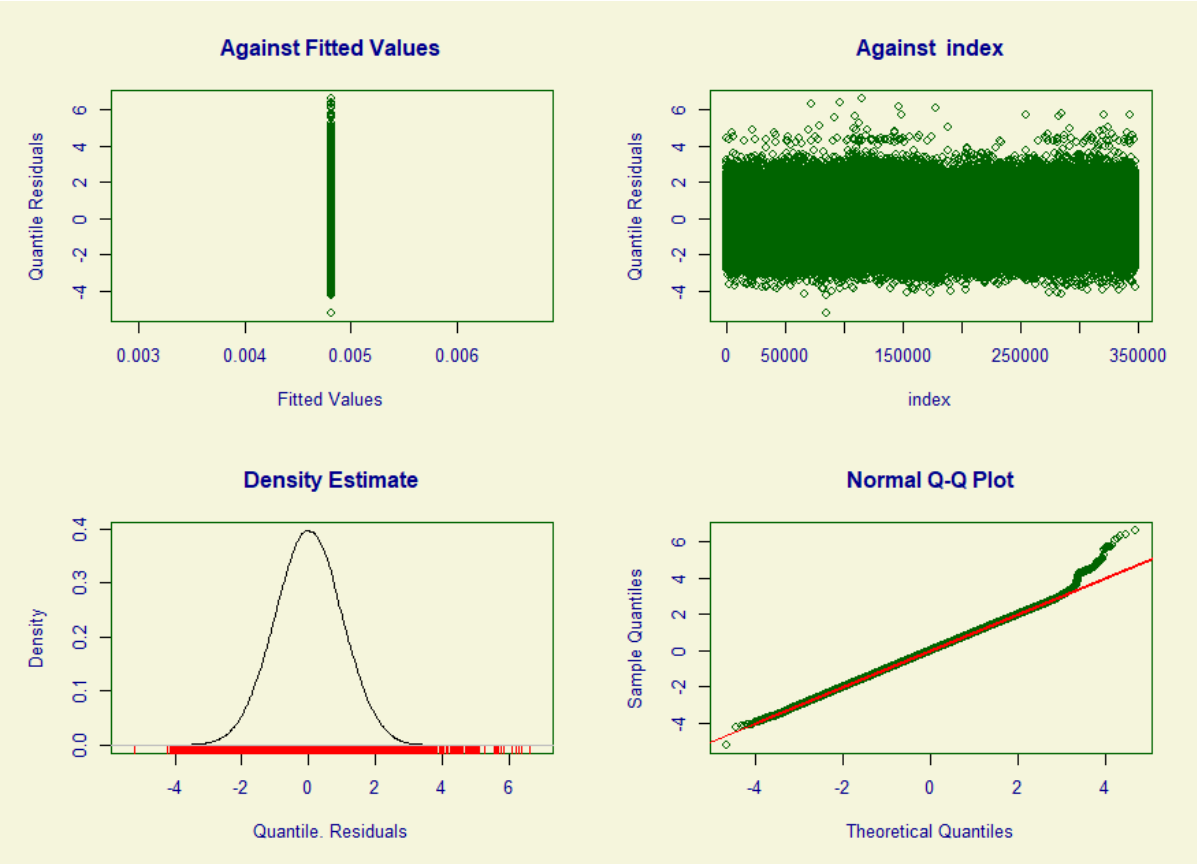


Figure 40. Poisson GoF - Other Regions

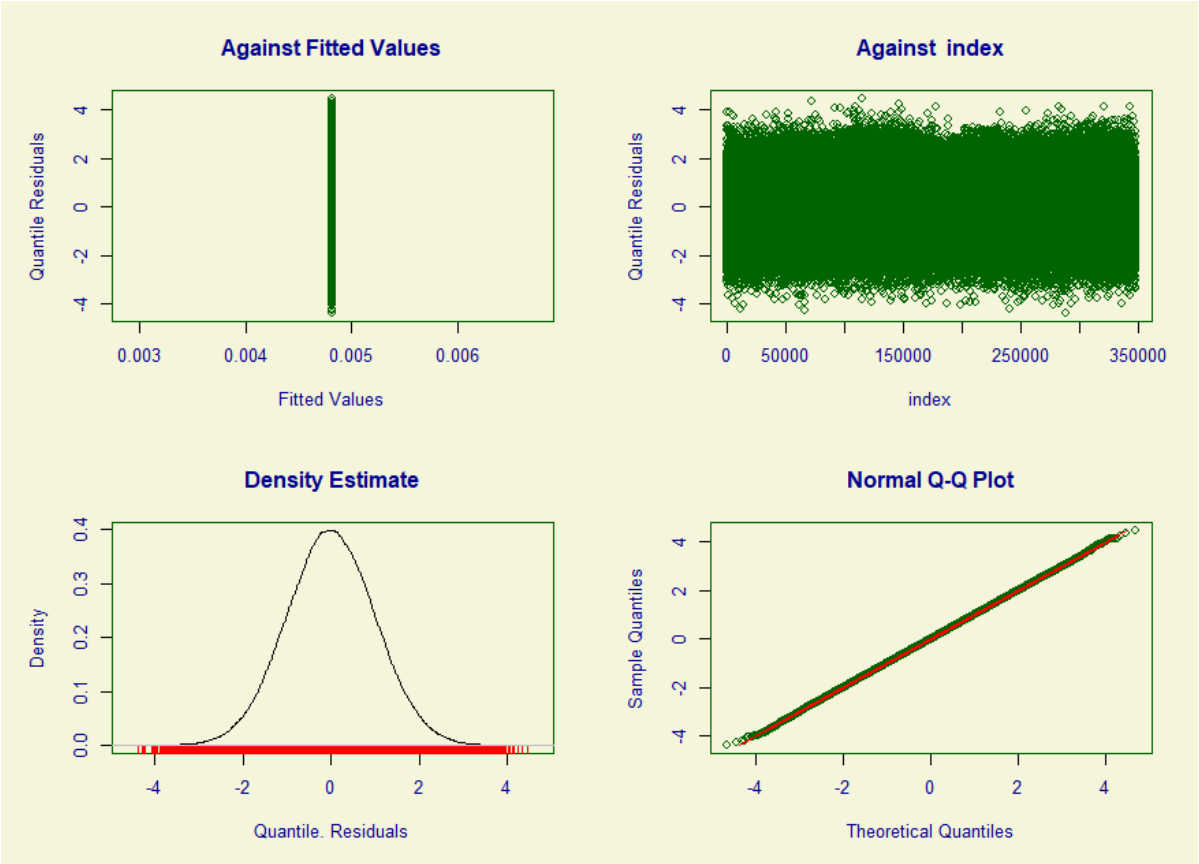


Figure 41. Negative Binomial GoF - Other Regions

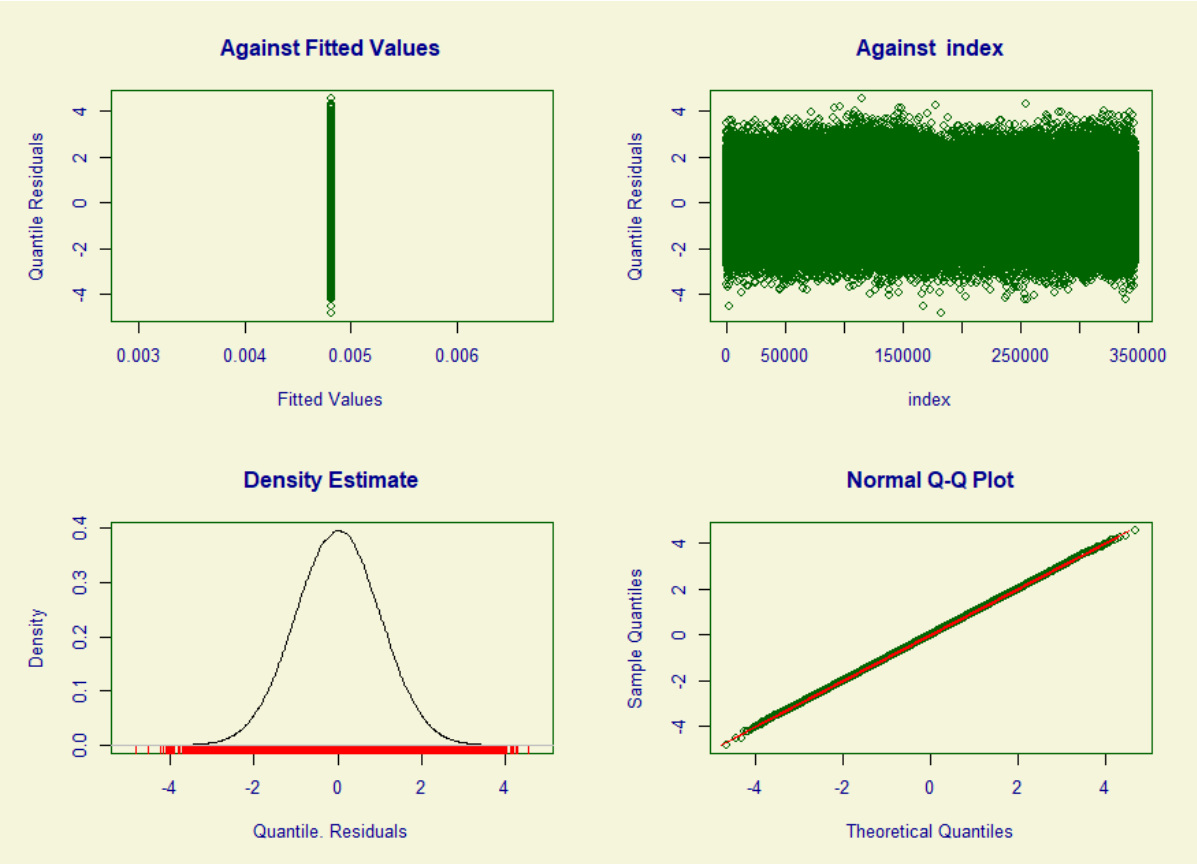


Figure 42. Poisson Inverse Gaussian GoF - Other Regions

From the above we consider that Negative Binomial and Poisson Inverse Gaussian distributions fits better.

Distribution Fit - Severity

For the severity of claims we examine Gamma, Weibull, Pareto, Log-normal, Burr, Generalized Gamma and Generalized Pareto distributions.

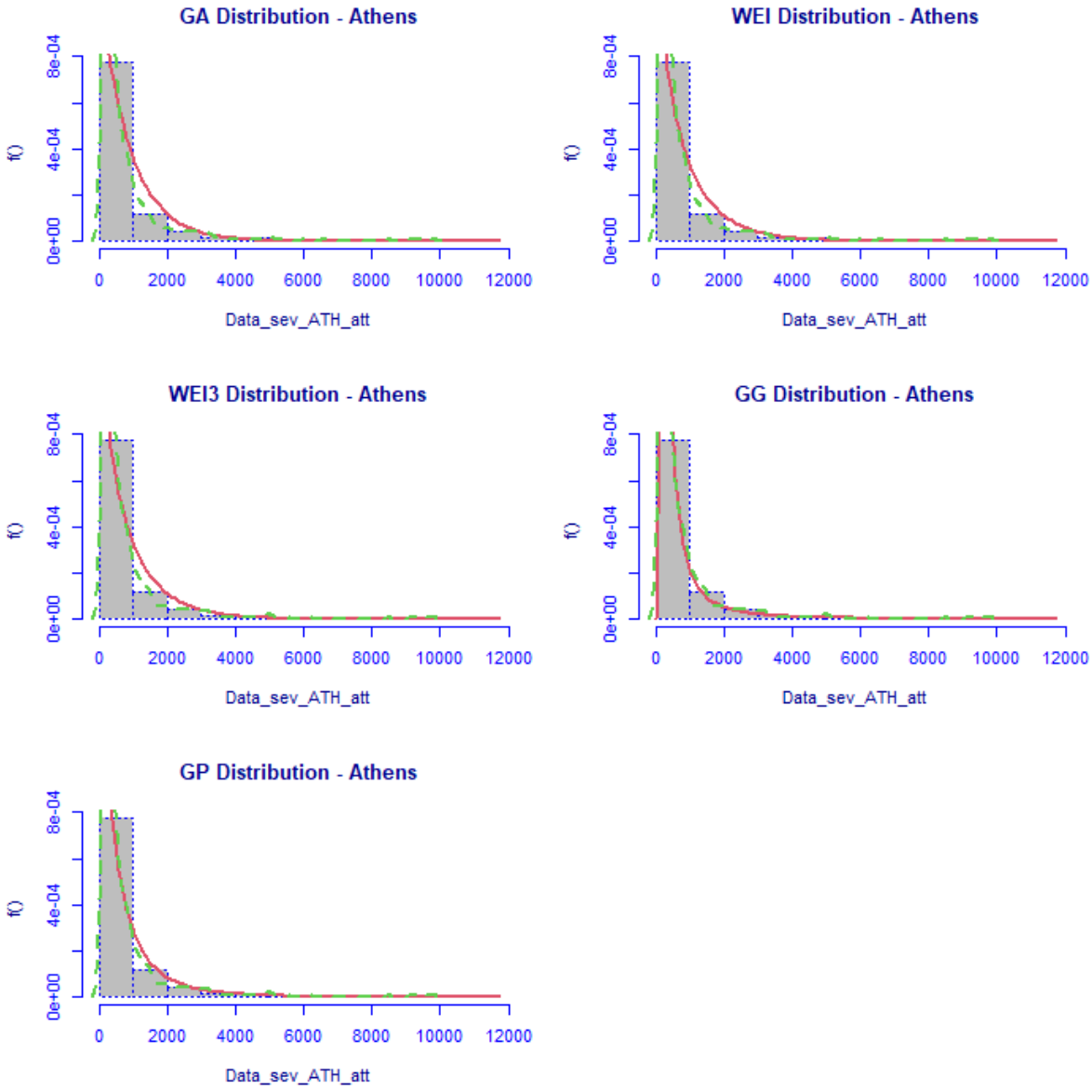


Figure 43. Severity Distributions Fit - Athens

Table 25: Severity Distribution Gamlss

Gamlss Severity Distributions	Location	Global Deviance	AIC	SBC
Gamma	Athens	17556.5	17560.5	17570.5
	Other	22743.8	22747.8	22758.2
Weibull	Athens	17536.8	17540.8	17550.8
	Other	22723.6	22727.6	22738.0
Weibull type 3	Athens	17536.8	17540.8	17550.8
	Other	22723.6	22727.6	22738.0
Generalized Gamma	Athens	16978.5	16984.5	16999.6
	Other	22427.5	22433.5	22449.1
Generalized Pareto	Athens	17414.6	17418.6	17428.6
	Other	22655.2	22659.2	22669.6

Table 26: Severity Athens Table

Severity - Athens				
Goodness-of-fit statistics	lnorm	llogis	Pareto	Burr
Kolmogorov-Smirnov statistic	0.07744	0.07038	0.16287	0.05893
Cramer-von Mises statistic	2.03009	1.13994	4.73654	1.01377
Anderson-Darling statistic	13.62382	10.83466	34.57627	5.67912
Goodness-of-fit criteria	lnorm	llogis	Pareto	Burr
Akaike's Information Criterion	17150.36	17163.51	17418.58	17032.14
Bayesian Information Criterion	17160.42	17173.56	17428.63	17047.22

Table 27: Severity Other Table

Severity - OTHER Regions				
Goodness-of-fit statistics	lnorm	llogis	Pareto	Burr
Kolmogorov-Smirnov statistic	0.05821	0.05160	0.09811	0.04899
Cramer-von Mises statistic	1.07450	1.05722	1.84820	1.00154
Anderson-Darling statistic	7.48477	8.65174	16.78687	7.42052
Goodness-of-fit criteria	lnorm	llogis	Pareto	Burr
Akaike's Information Criterion	22468.58	22548.72	22659.19	22531.50
Bayesian Information Criterion	22479.02	22559.16	22669.63	22547.17

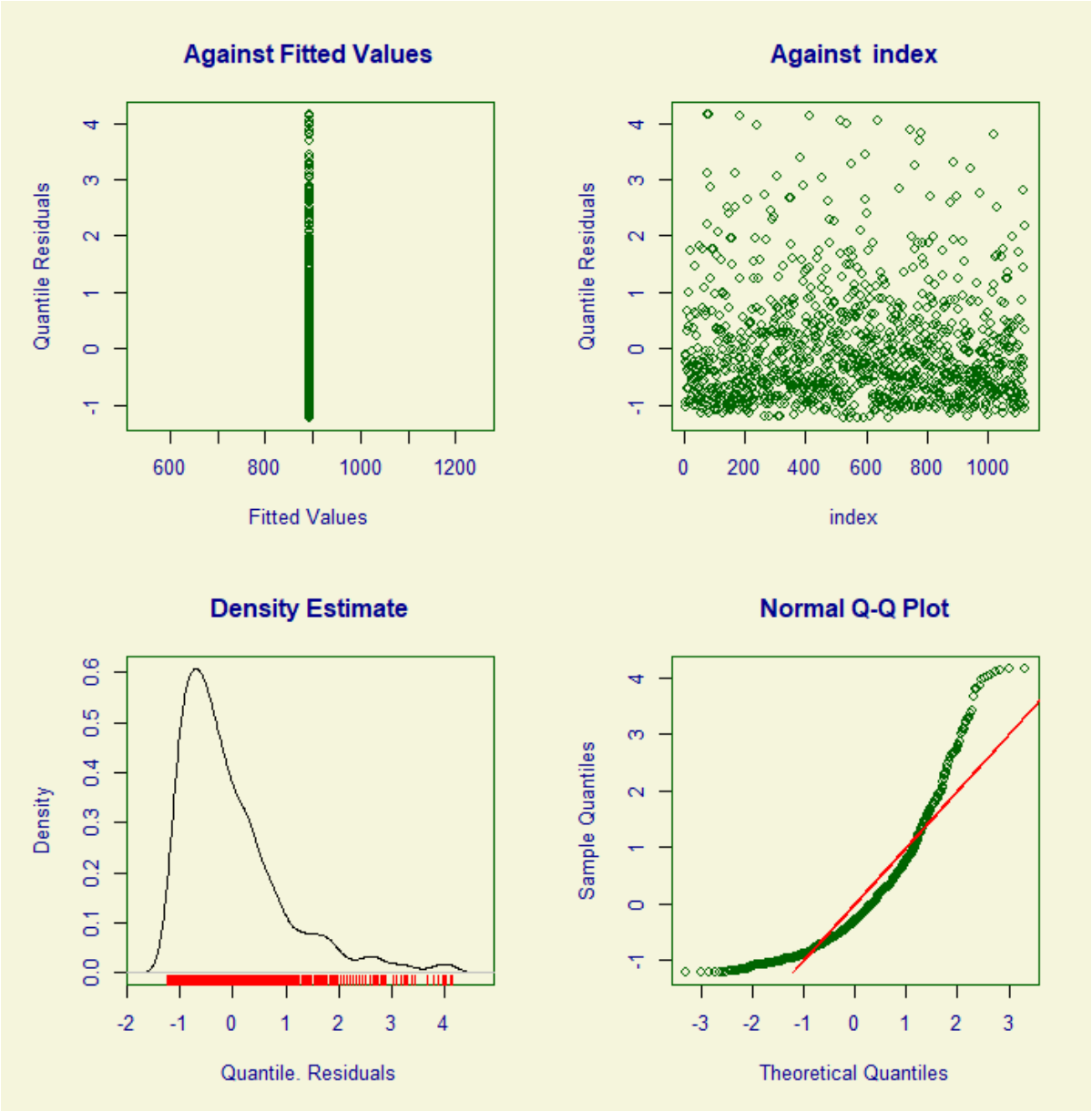


Figure 44. Severity Gamma Distribution Fit - Athens

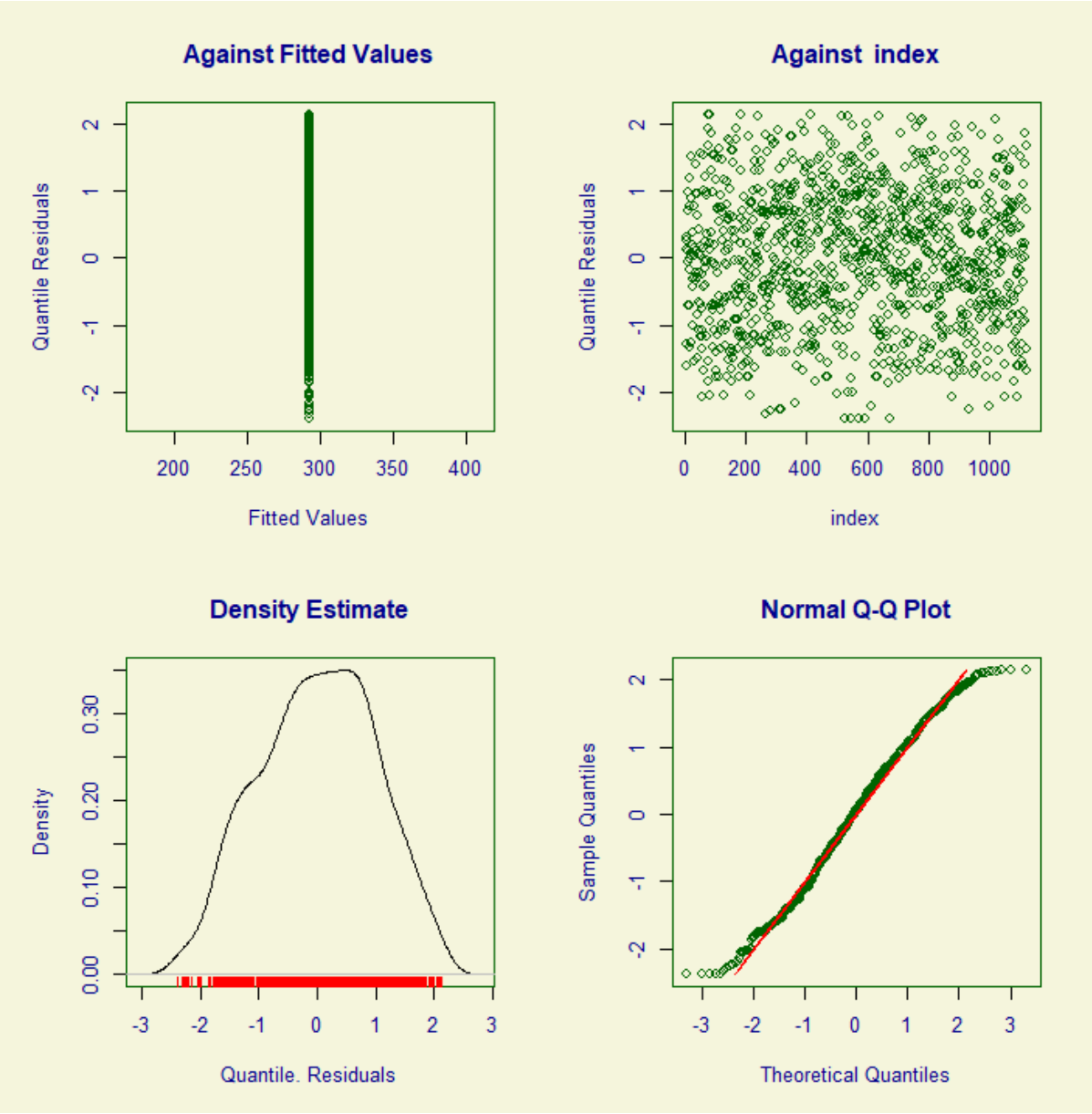


Figure 45. Severity Generalized Gamma Distribution Fit - Athens

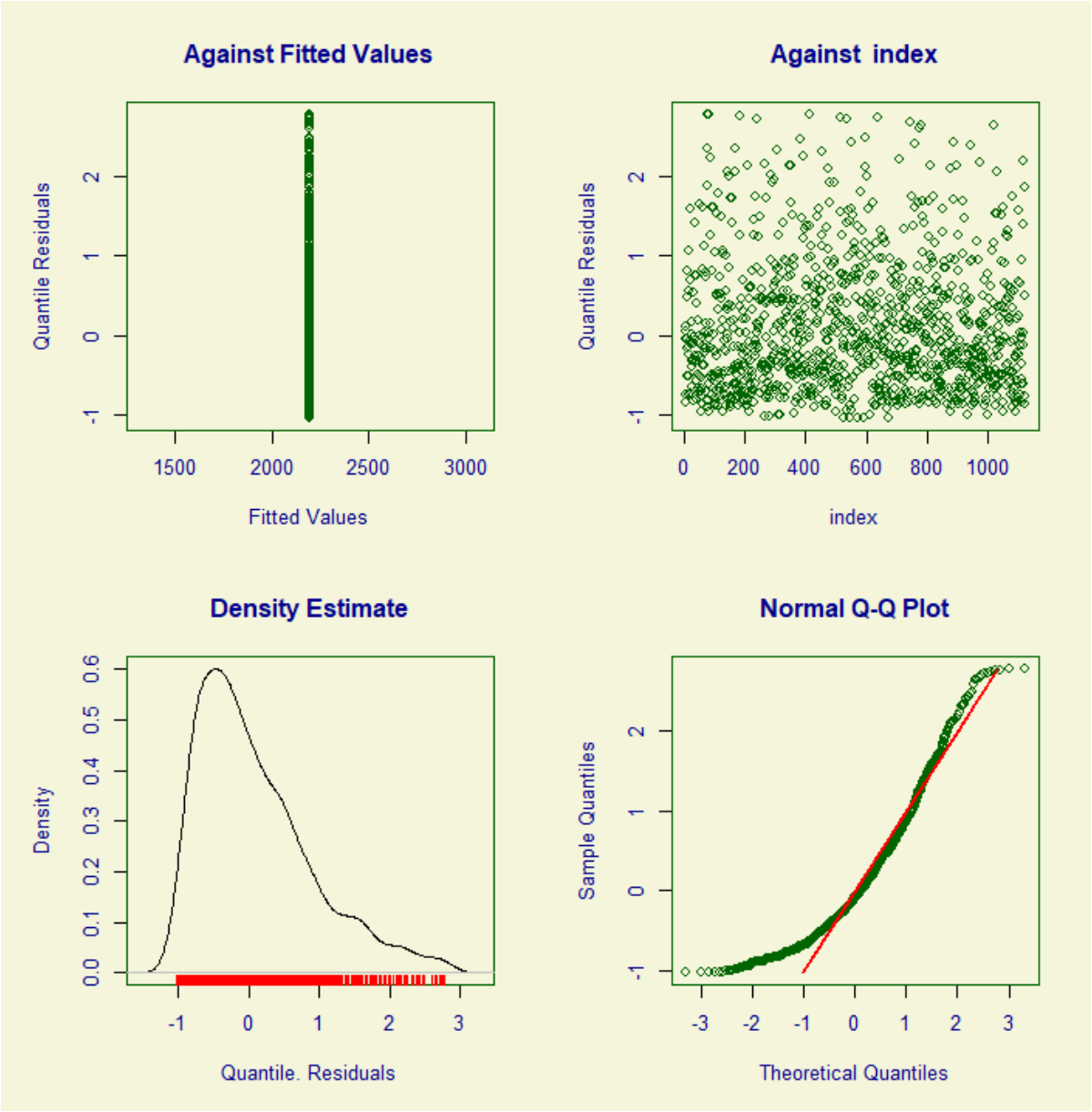


Figure 46. Severity Generalized Pareto Distribution Fit - Athens

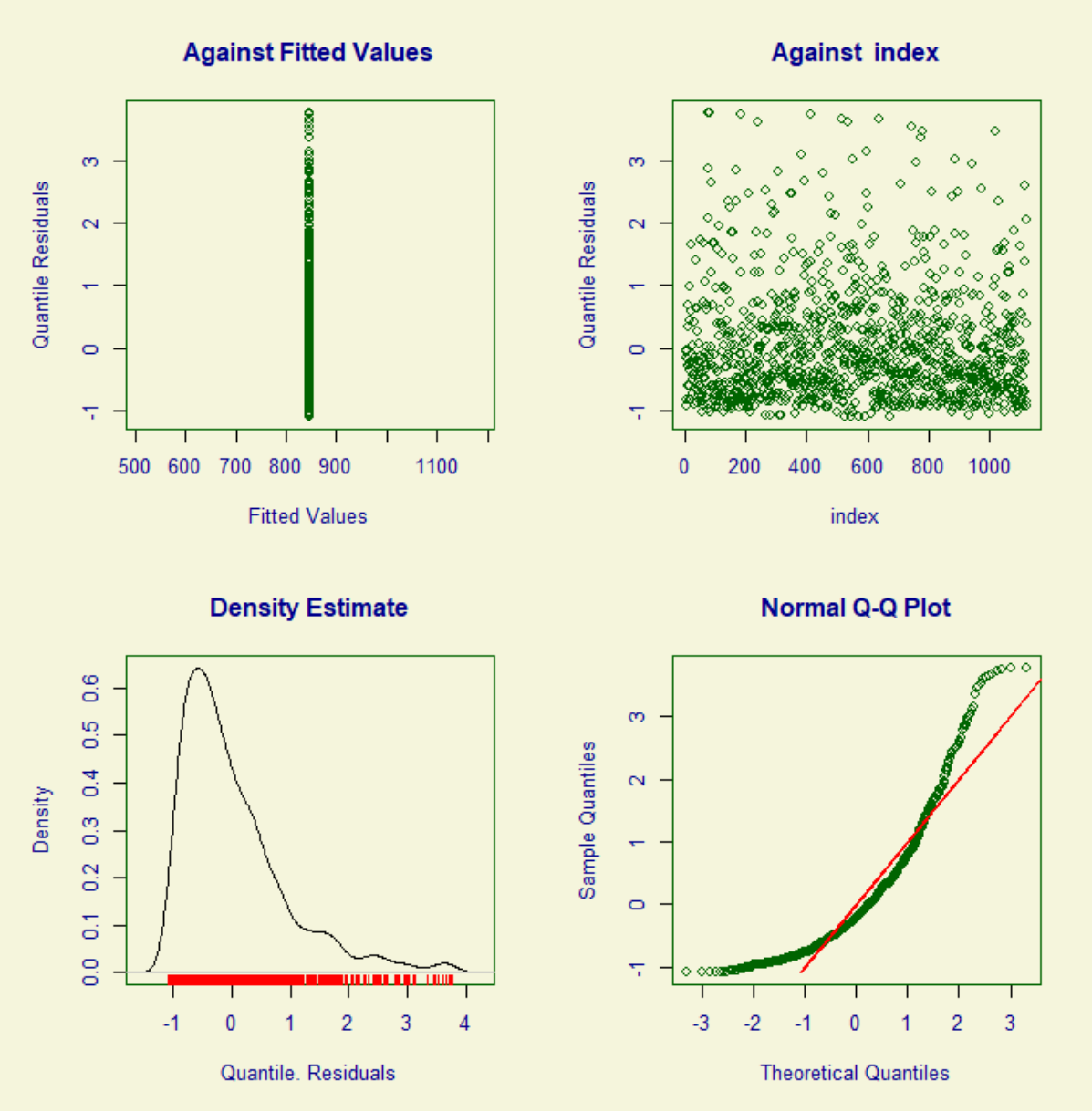


Figure 47. Severity Weibull Distribution Fit - Athens

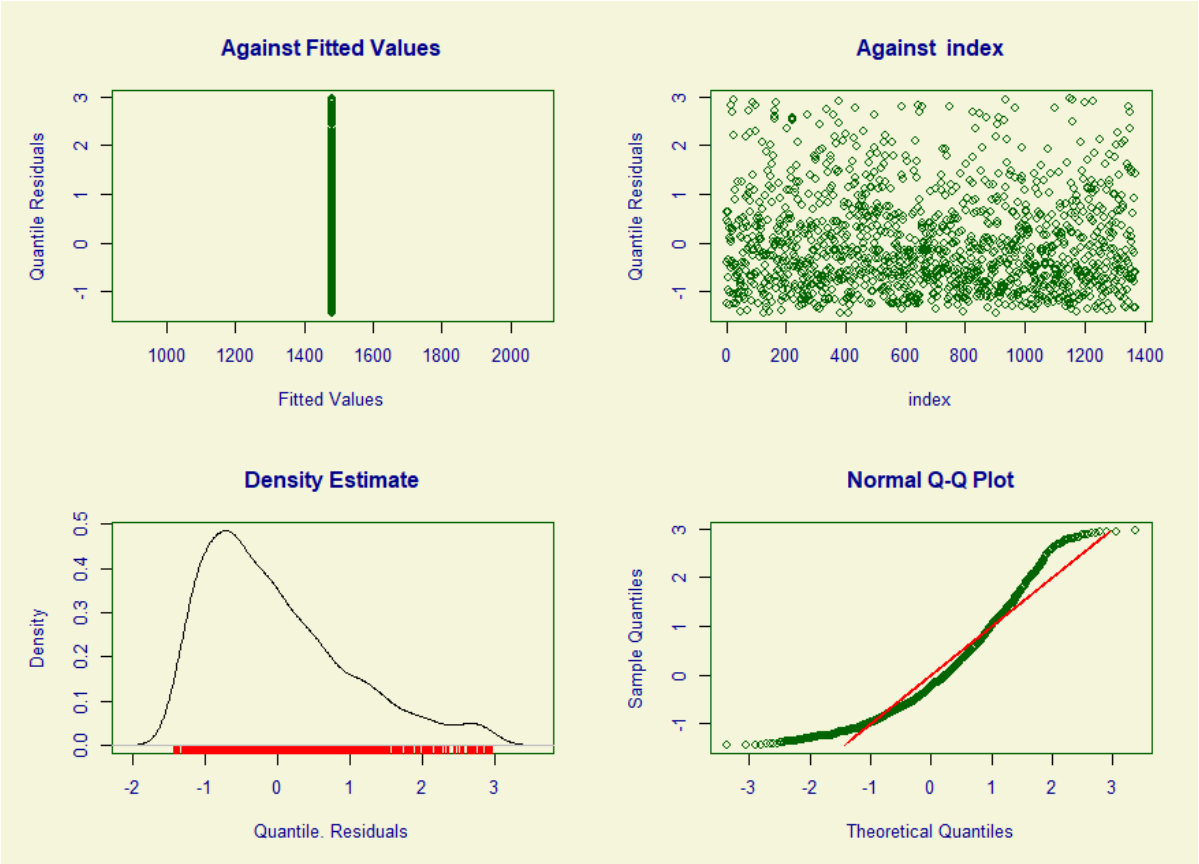


Figure 48. Severity Gamma Distribution Fit - Other Regions

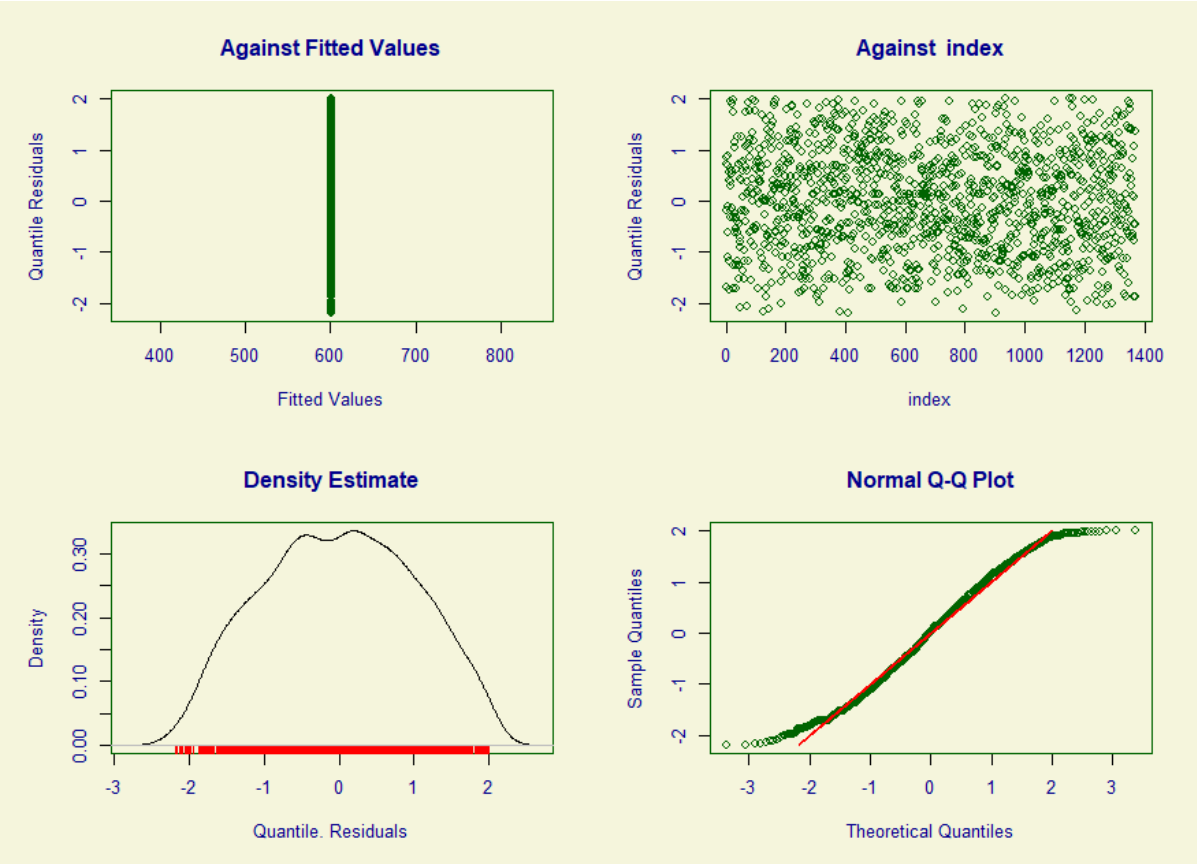


Figure 49. Severity Generalized Gamma Distribution Fit - Other Regions

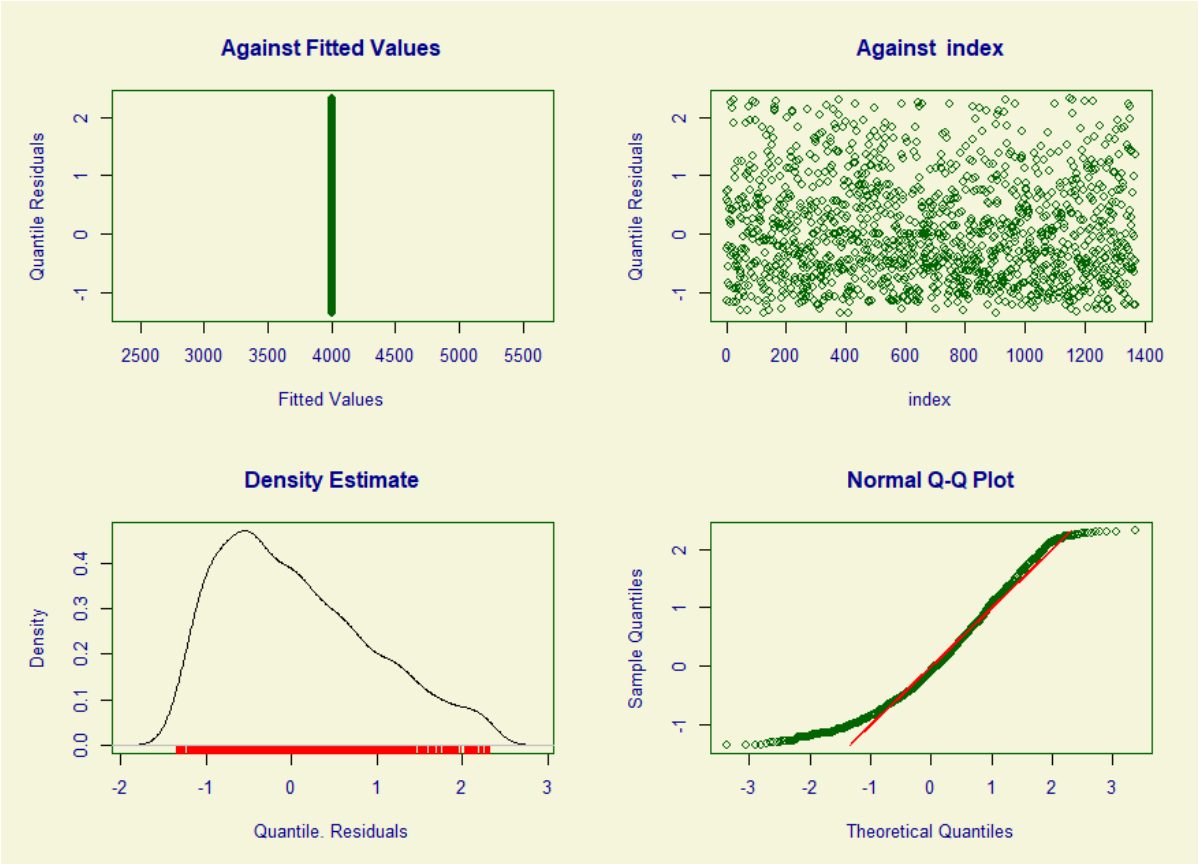


Figure 50. Severity Generalized Pareto Distribution Fit - Other Regions

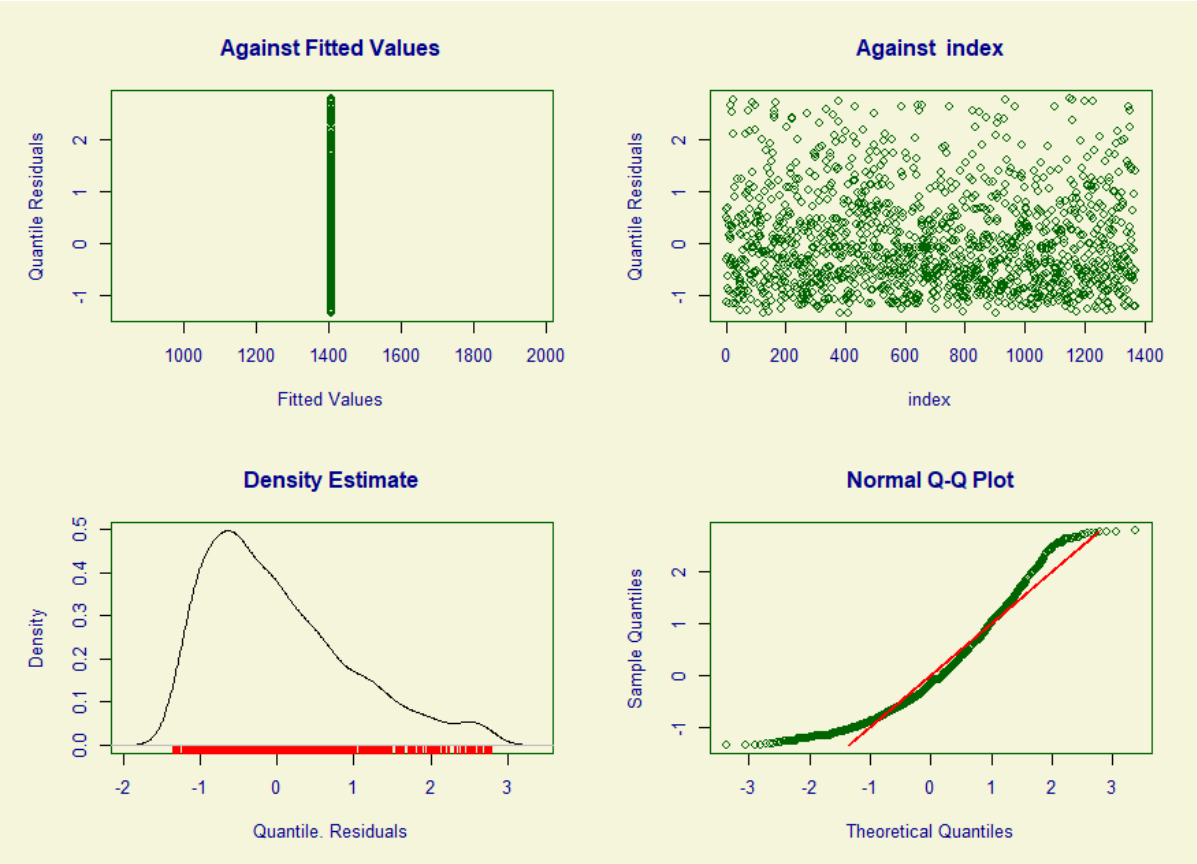


Figure 51. Severity Weibull Distribution Fit - Other Regions

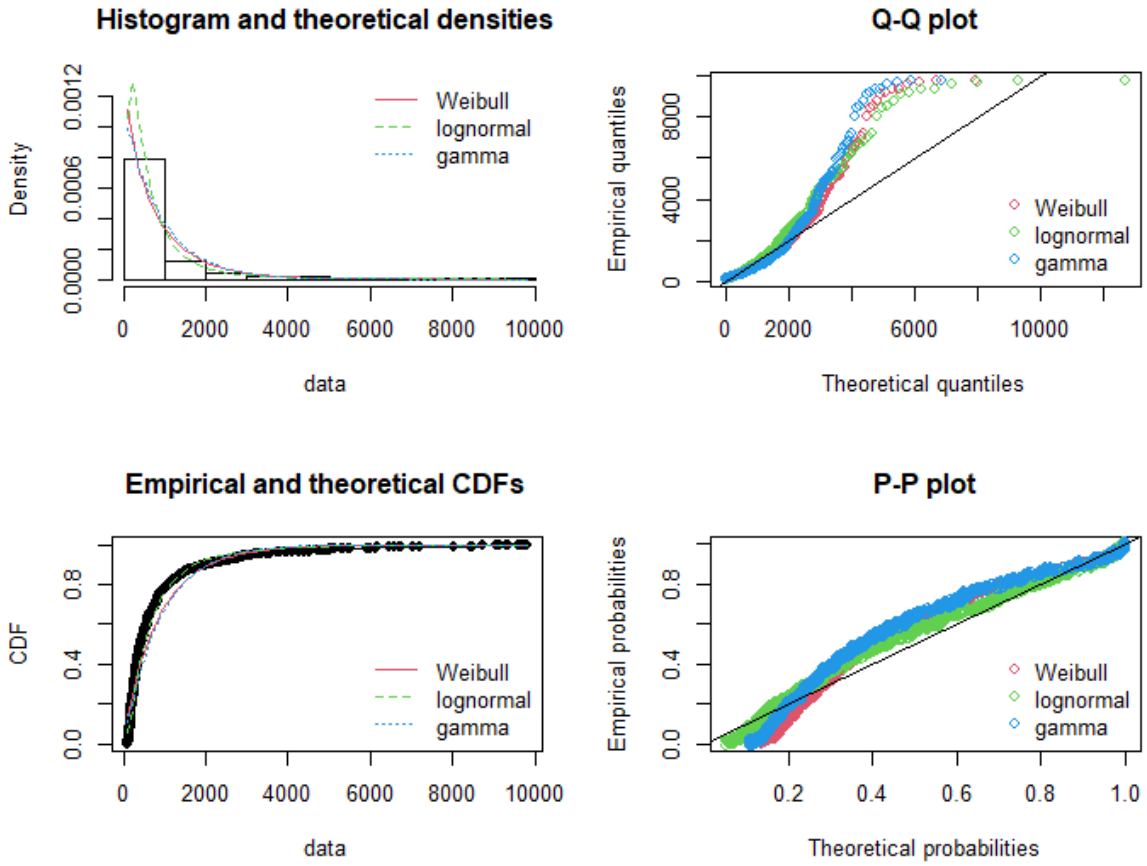


Figure 52. Severity Distribution Fit - Athens

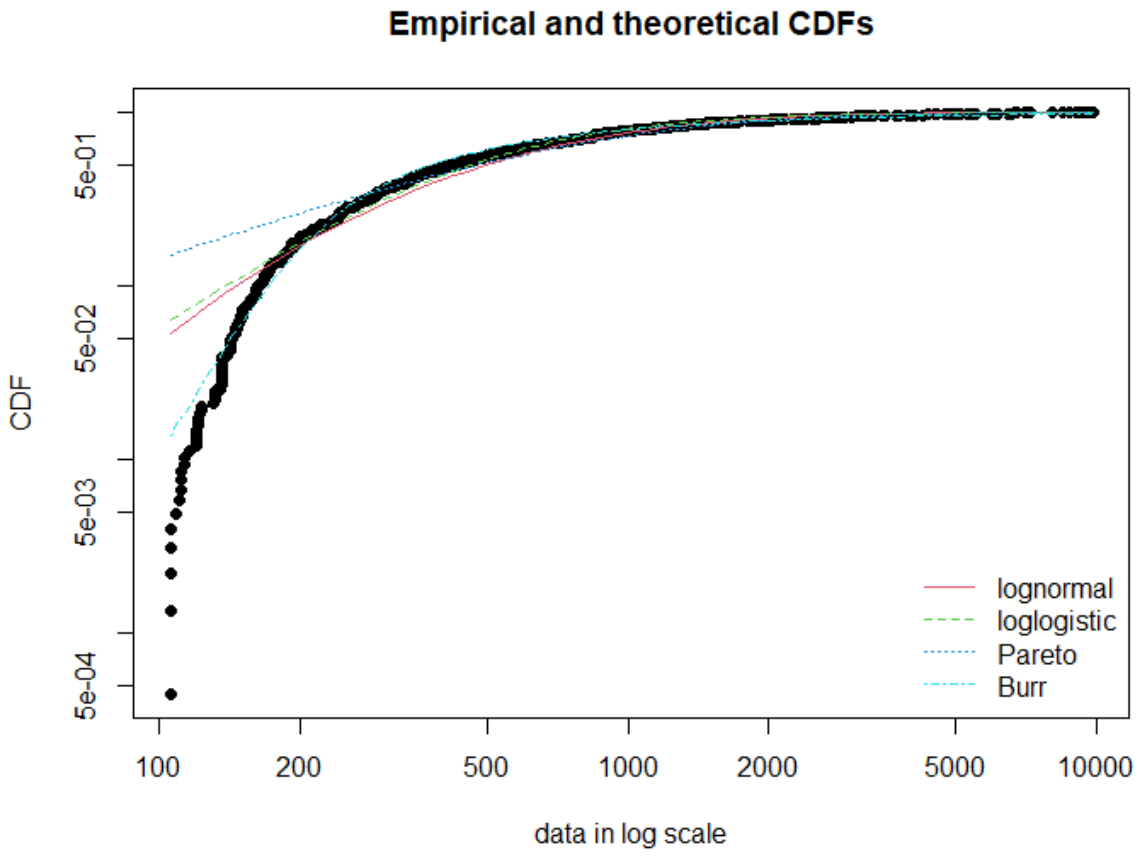


Figure 53. Severity Distribution Fit2 - Athens

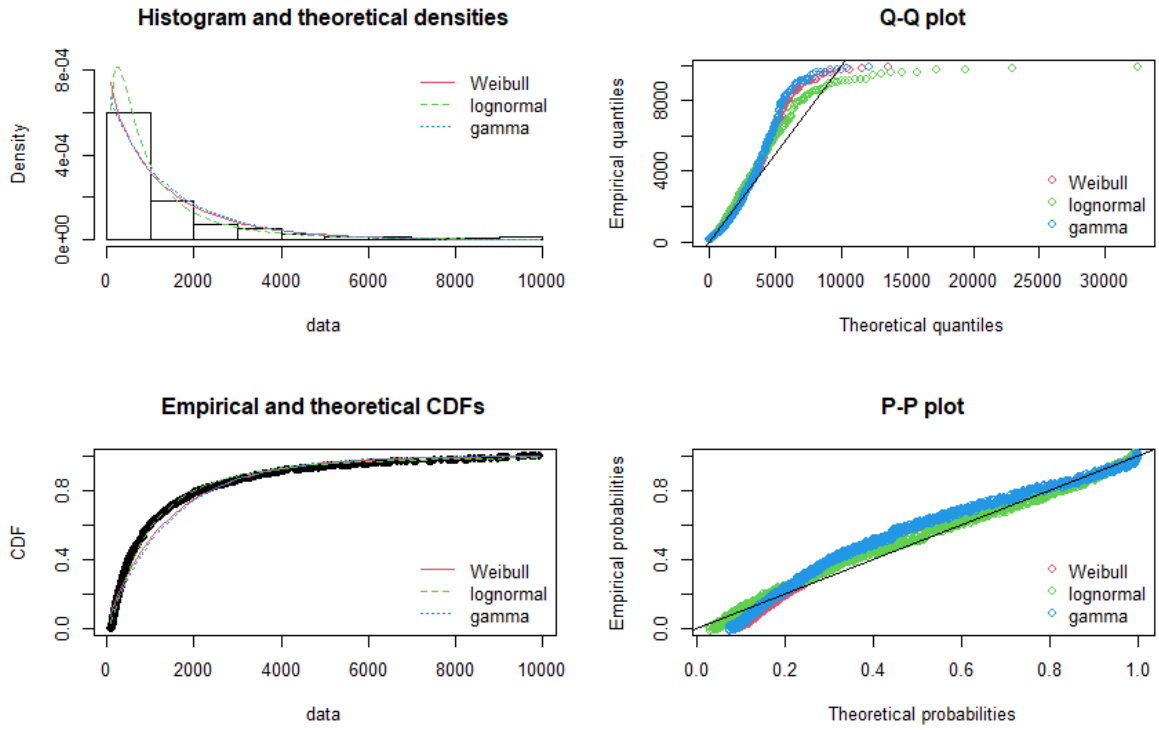


Figure 54. Severity Distribution Fit - Other Regions

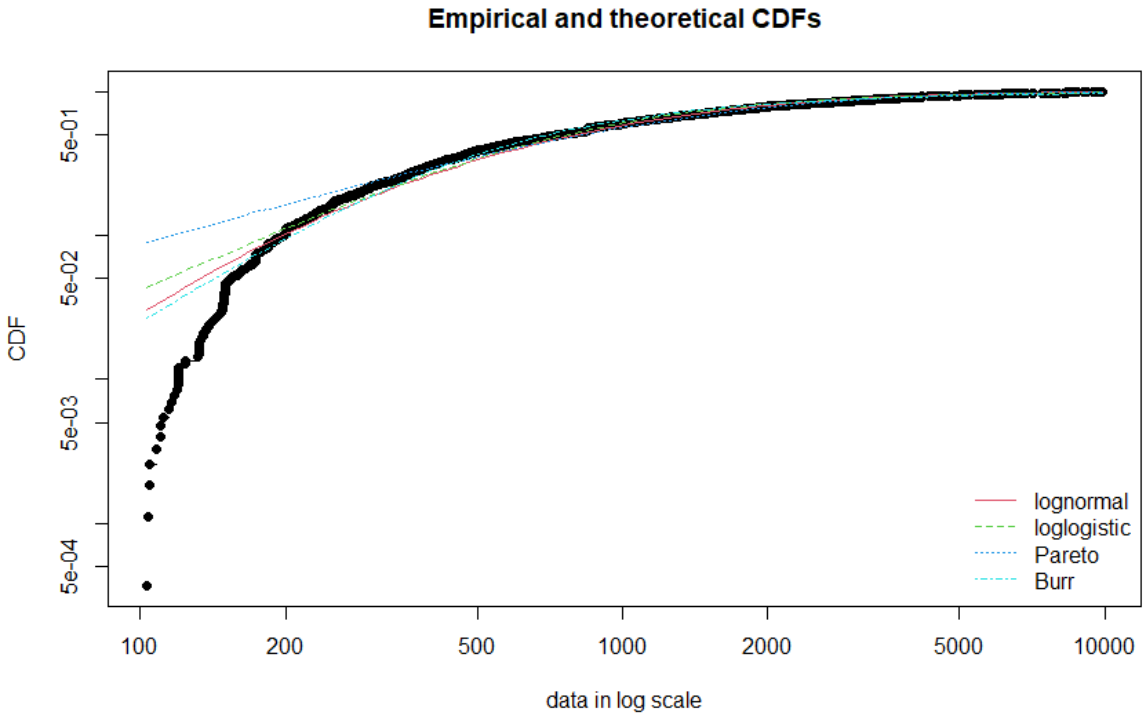


Figure 55. Severity Distribution Fit2 - Other Regions

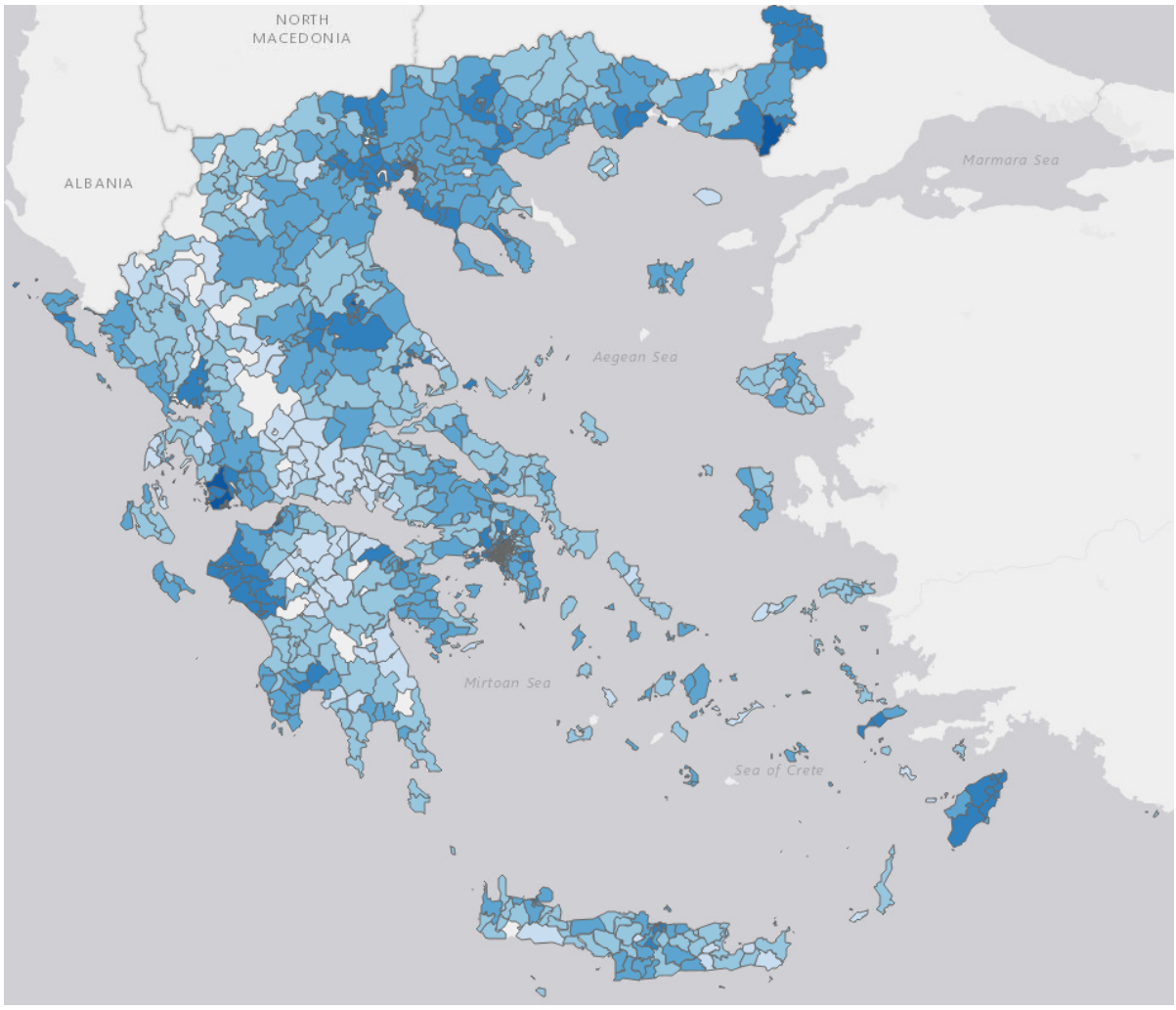


Figure 56. Flood Vulnerability Map of Greece

In the above map we can see the flood vulnerability spatial distribution in Greece according to meteorological input. We will also examine the contribution of this input to the corresponding weather related risk distribution.

Western and Northern areas of Greece, Attica and Central Macedonia seem to be more vulnerable to severe flood events.

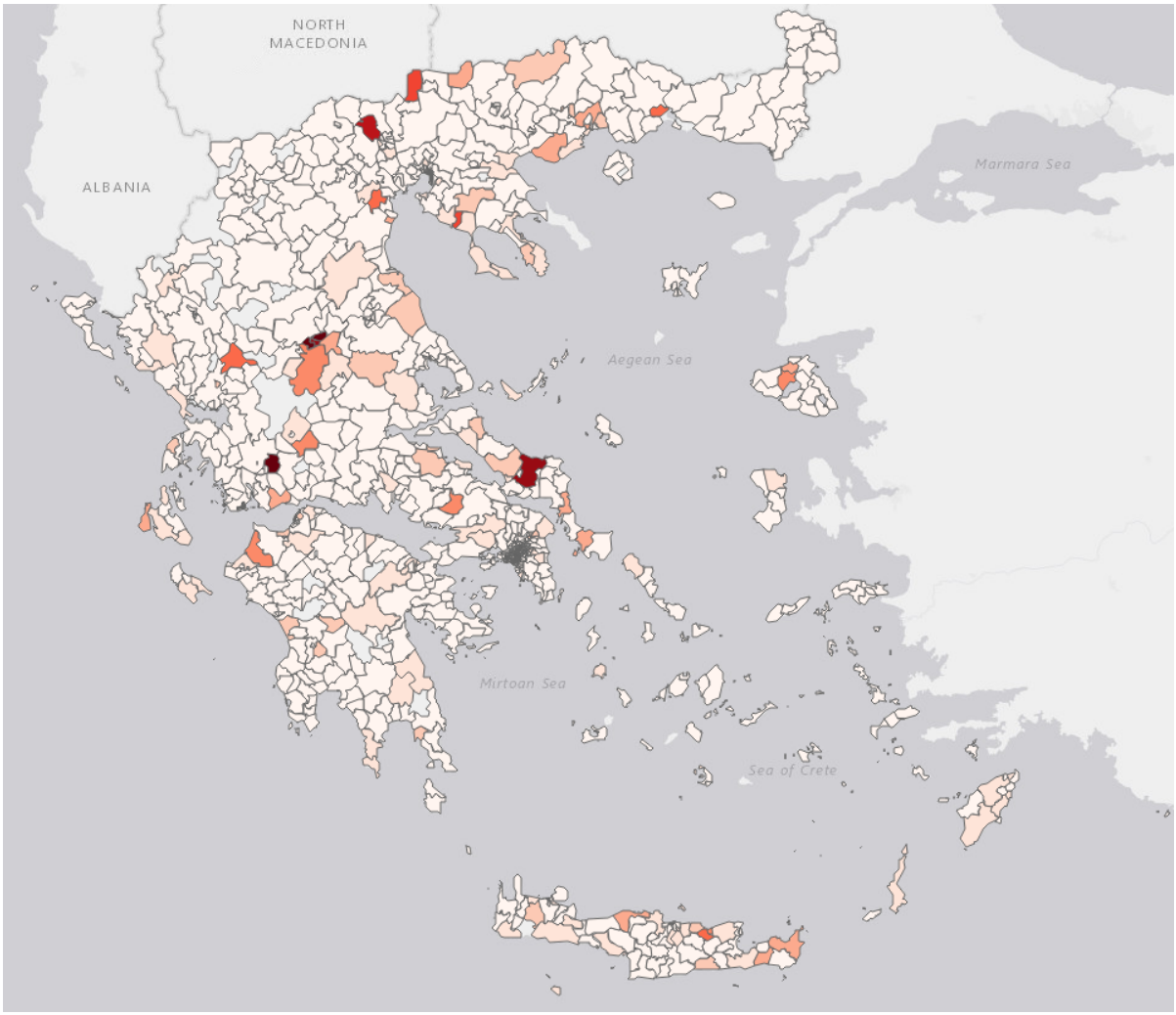


Figure 57. Frequency Map of weather related claims

In the above map we can observe the frequency of the insurance claims which result from the relevant database. We can trace specific regions with darker color which are mainly affected by meteorological events. This result is actually a function between the distribution of the exposure and the number of claims occurred per region.

Specific regions are highlighted, for example, North-West Attica, North-West Peloponnese, the Northern area of Crete and other regions in Central and Northern

Greece.

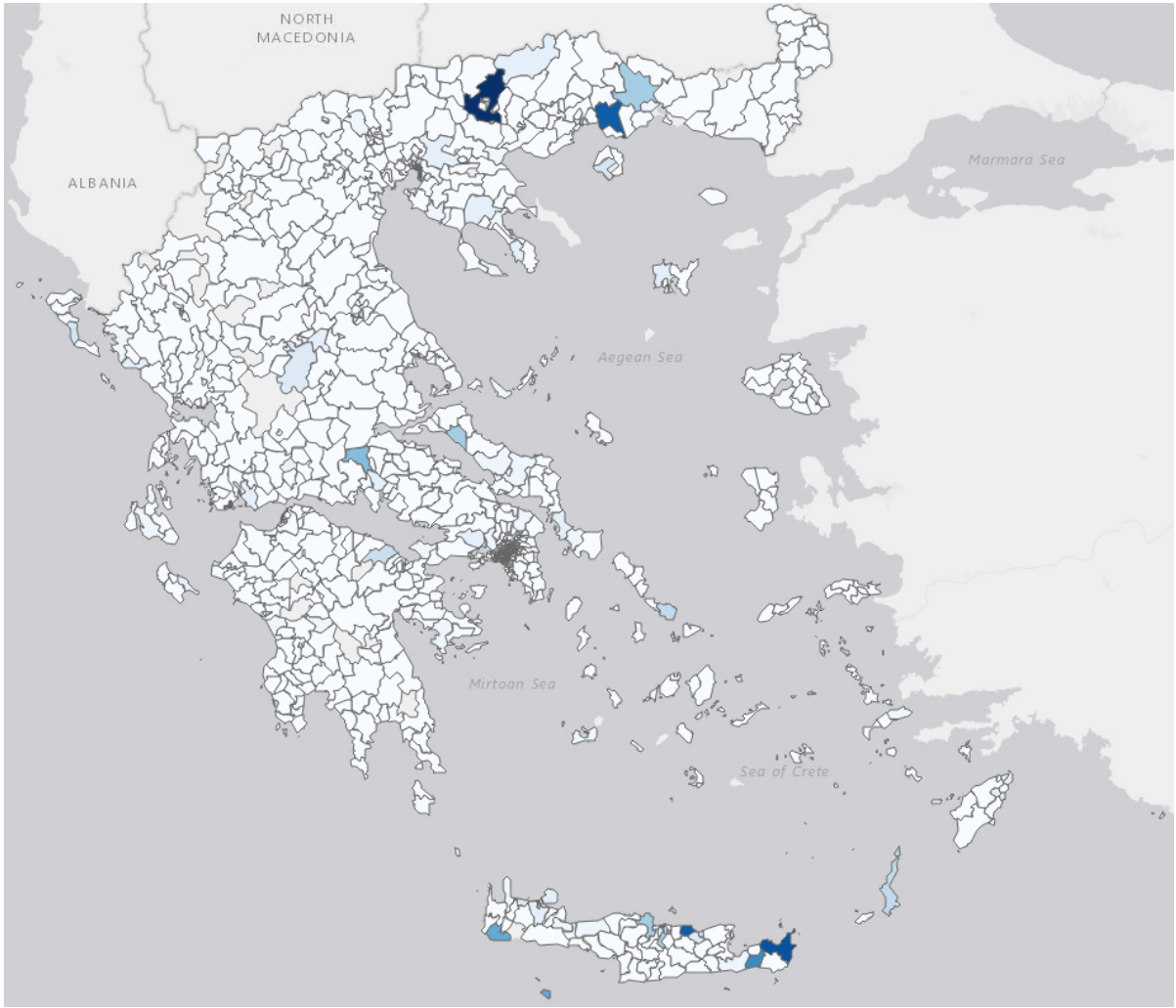


Figure 58. Severity Map of weather related claims

In terms of severity, Crete and regions of Central and North Greece seem to be mainly affected. Severity corresponds to the average claim cost of the claims reported at the level of the postal code analysis.

The specific severity view is also connected with the years of observation and the regions which were affected by severe weather related events in this specific time

period. Moreover, severity is always relative to portfolio characteristics, such as the capital insured etc.

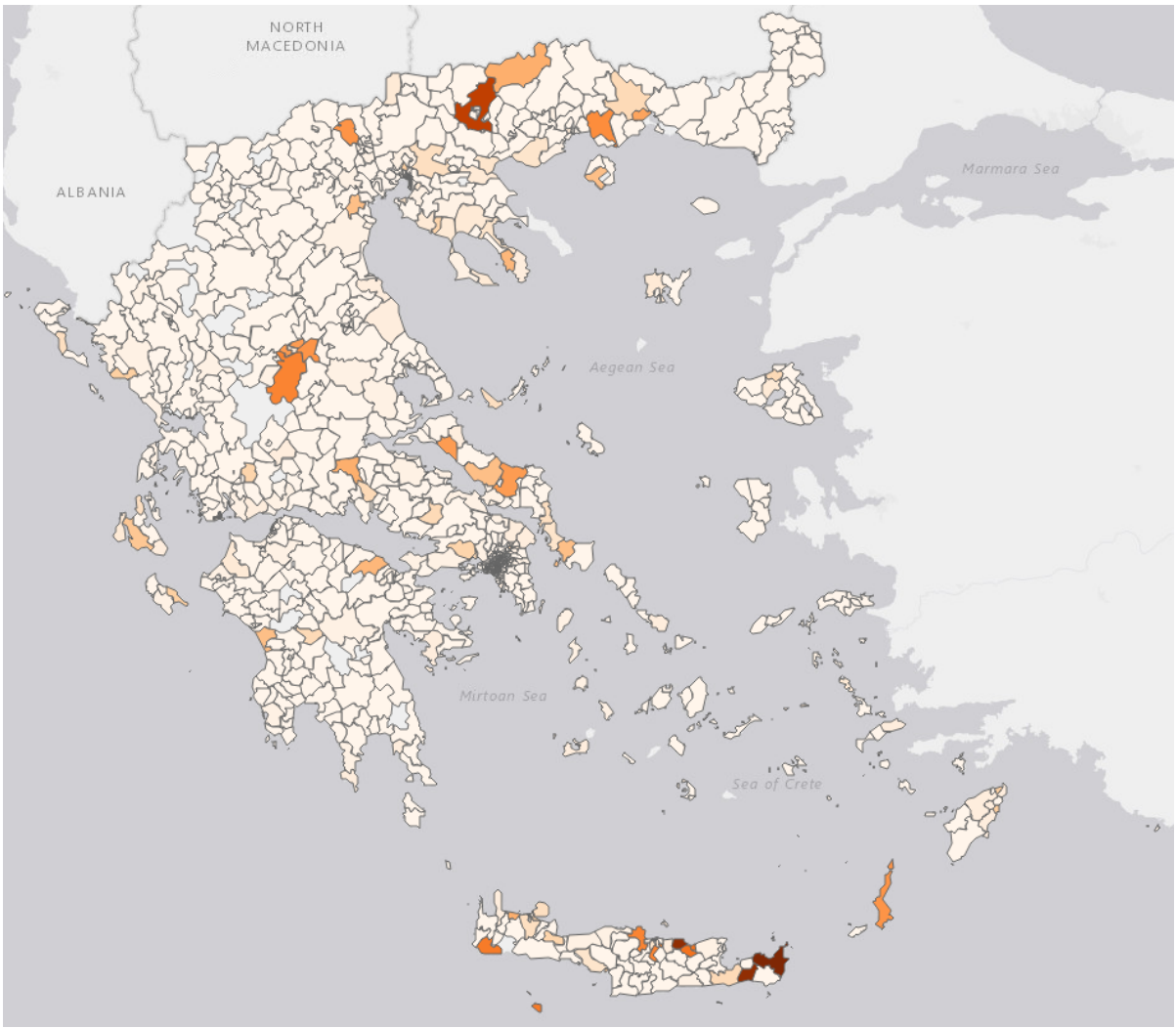


Figure 59. Risk Premium Map of weather related claims

Combining the frequency and the severity view we obtain the risk premium estimation which is actually the premium which reflects adequately the risk and is equal to:

$$Risk\ Premium = Frequency * Severity$$

Climate Related Risks Classification

As we have seen in the aforementioned literature, the use of "ensemble learning" methods using "decision trees" as "base learners" for studying climate-related claims in non-life insurance is a popular option, see for instance [64],[38], and [32]. However, the focus has been predominantly on "random forest" methodology and other powerful ensemble learning approaches such as boosting were not explored. In this case, we adopt a simplified approach and suggest a well known decision tree based ensemble method called gradient boosting to classify the number of claims caused by storms in Greece, where the data has been taken from insurance companies in Greece, and finally use a machine learning algorithm to help categorise the number of claims which have been occurred by a "storm event" into 3 buckets: "no claims", "1 claim", "2 or more claims". In particular, the "gradient boosting" process we use in this part is described in this section involves the following steps. Firstly, we fit a "shallow decision tree" on to the data. By "shallow" we mean a "decision tree" with very few splits on the predictors with highest predictive power. These splits are determined by a heuristic approach, as finding the optimal set of splits is a computationally difficult problem [27]. These trees are so called *weak learners* and give just a lower prediction error rate than the one that would result from random guessing. This has to do with the main idea behind ensemble modelling, see for example [4], which is to build a number of weak models and then combine them into a stronger one in order to reduce bias, and increase model accuracy. After we have fit a "shallow decision tree" to the data, we then fit the next "decision tree" to the residuals of the first one. The residuals are the differences between the prediction and the actual class labels, and thus the second decision tree addresses the shortcomings of the first one. Then, we add this second tree into the algorithm. We continue the process and fit a third tree to the residuals of the second tree and then we proceed by adding the third tree into the algorithm and so on. The process stops when adding more decision trees does not result in a more "powerful" model. [19]

[19]The model can easily overfit the training data and thus the stopping criterion is decided on a validation dataset that is not used for training, called "early stopping".

Each individual model is weighted according to some criterion and the ensemble is a ”stage-wise additive” model (i.e. adding ”decision trees” sequentially of all individual trees.

The structure of the analysis includes some relevant machine learning techniques, it is presented the novel data-set and the machine learning method we use to solve our classification problem with the relevant reasults.

Machine Learning Preliminaries

Interest and use of machine learning techniques are increasing each day - whether in our daily lives for example via smart technology, or in business, for example with a greater focus from insurers, and they look to incorporate such techniques to reduce internal costs and improve consumer experiences. Before we continue with our analysis, in this section, we will introduce some machine learning concepts that will hopefully be useful for the reader to understand the approaches used by us. The explanations given here are not meant to be complete nor exhaustive, but more to give the reader a better overview of the methods used and hence how and why they fit into this study.

”Supervised” and ”Unsupervised” Learning

“Supervised” and “unsupervised” learning are 2 fundamental approaches which we can take in machine learning. “Supervised learning” builds models where the data-set is labelled, and has input data and output variables. Hence, “supervised learning” can be used to calculate outcomes e.g. for future predictions of values based on historic data. “Unsupervised learning” builds models where the data-set outputs are not labelled, and hence can be used to find “hidden” patterns and relationships within the data e.g. for categorisation problems. Generally, “supervised learning” is computationally simpler, produces a more accurate output but requires external supervision to “train” (i.e. construct) the model when compared to “unsupervised learning”. However, it also require more human intervention and as such may potentially cost more to implement.

[20]

Algorithms and Data-sets

A typical process within machine learning is to use data to develop a model which hopefully mimics the outcome required, within a certain level of accuracy/materiality threshold. The original data-set itself is split into a "train" data-set and "test" data-set, where the "train" detest is the larger proportion of the data. The "train" data-set is used to develop an initial model, where the model is based on an underlying individual algorithm/set of algorithms, usually via an iterative process until a certain level of accuracy is reached. Once the model is "trained", we use the "test" data-set to check the accuracy of the model, and ensure that the model has not been "over-trained" and biased to the input data originally used. There is a question of which algorithm to use: we may choose a range of algorithms which we initially feel are appropriate for the problem we are analysing, and compare the accuracy of the outputs of each algorithm before decision which to choose.

Ensemble Learning

Above we mentioned the use of algorithms within machine learning. Ensemble learning is a collection of models which tend to produce a more accurate output i.e. "the whole is greater than the sum of the parts". For more details of merging predictions in such a manner please see for example [4], [10], [45] and [63]. The building blocks within this framework are called "base learners". Ensemble learning takes a "blend" or "average" of these "base learners" to hopefully remove any inherent bias produced by an individual "base learner" and to produce an overall better output. One popular base learner is a decision tree and it is discussed further below. For more background on "ensemble

[20] For more on general background to both approaches, please see for example www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

learning", including some background to mathematical detail, please see [14].

Decision Trees

A "decision tree" is a "supervised learning" approach where the output is decided based on a series of decisions at each "node" - similar to a branches on a tree where the decisions can branch out. For more background, include graphical explanations of "decision trees" please see [8]. We present a mathematical formulation of a classification decision tree. Suppose that we have a collection of n observations, each of which consists of a response variable and k predictors. For the i^{th} observation, we denote it with the notation $(y_i, x_{i1}, x_{i2}, ..., x_{ip})$. To ease notation, we will also denote the p-dimensional vector of predictor $(x_{i1}, ..., x_{ip})$ by $\mathbf{x}_i$. For the purpose of this paper, we assume that y_i is categorical and can take up to K distinct values, which we shall denote by $\{0, 1, \cdots, K-1\}$. Any of the predictors can be either continuous and numeric, or categorical. For ease of exposition, let us assume that they are all continuous and generalise to the categorical case later.

The goal of decision tree is to split the feature space into m non-overlapping regions, denoted $R_1, ...R_m$, such that all observations within the same region will have the same prediction. However - given the enormous number of ways one can partition the feature space, one needs to restrict the shape that these regions can take. One simple way to do that is to restrict the regions R_i to be be a *cylinder set*, i.e. sets of the form:

$$\{\mathbf{x} \in \mathbb{R} | L_1 \leqslant x_1 \leqslant U_1, L_2 \leqslant x_2 \leqslant U_2, ..., L_p \leqslant x_p \leqslant U_p\}, \tag{9}$$

where all L_i, U_i belongs to $[-\infty, \infty]$. We also allow any of the inequalities to be strict or non-strict. As it turns out, even under this restriction, to find in some sense an optimal collection of cylinder sets would present an impossible computational challenge. Therefore, we will need to restrict also the procedures by which these cylinder sets might

arise.

We will describe the procedure here in a somewhat informal way. We start with the entire feature space, i.e, the set of all possible values for the predictors, which in this case is simply $\mathbb{R}^p$. We will now find an index-threshold pair (l, t), where $t \in \{1, ..., p\}$ and $t \in \mathbb{R}$ and split $\mathbb{R}^p$ into two sets:

$$R_1 \triangleq \{x \in \mathbb{R}^p | x_l \leqslant t_l\}, \quad R_2 \{\triangleq \mathbb{R}^p - R_1\}. \tag{10}$$

All observations in R_1 (respectively, R_2) should have the same prediction. In the case where the response is categorical, we take the the estimate to be the most commonly occurring value. Formally, we define:

$$\hat{\mu}_1 = \text{Mode}\,\{y_i | \mathbf{x}_i \in R_i\} \tag{11}$$

$\hat{\mu}_2$ is defined similarly.

In order to optimise the choice of (l, t), we need some notion of optimality, a mechanism that punishes 'bad' estimates. Intuitively, a leaf node is good when it correctly classifies a large porportion of the responses, i.e. they coincide with the *modal class*. We say that this region is *pure* and seek to minimise the *impurity* of this region. Given a region R and an estimate $\hat{\mu}$ for the modal class, by an *impurity* function we mean a mapping $Q(R, \hat{\mu})$ that, informally speaking, is increasing with the *impurity* of this region. We shall minimise the following weighted sum (we abuse notations where and write , for example, $|R_1|$ for the number of observations with predictors in R_1):

$$L(l, t_l, R) \triangleq \frac{|R_1|}{|R|} Q(R_1, \hat{\mu}_1) + \frac{|R_2|}{|R|} Q(R_2, \hat{\mu}_1) \tag{12}$$

Thus, the optimal (l^*, t^*) is defined by:

$$(l^*, t^*) \triangleq argmin_{(l,t)} L(t, l_t)$$

There are a number of candidates loss functions that can play this role. Given a region R and a class k, we define:

$$\hat{p}(R, k) \triangleq \frac{1}{|R|} \sum_{y \in R} \mathbb{I}_{\{y=k\}}, \tag{13}$$

the proportion of class k observations in class m. One natural and easily interpretable candidate is the misclassification error rate.

$$E(R, k) = 1 - \max_k \hat{p}(R, k). \tag{14}$$

Two other options are are the Gini Index and the Cross Entropy (denoted G and D below, respectively). It can be shown that these functions will tend to zero as $\hat{p}(R, k)$ is either really close to 0 or 1, implying a high level of class purity.

$$G(R, k) = \sum_{k=1}^{K} \hat{p}(R, k)(1 - \hat{p}(R, k)) \tag{15}$$

$$D(R, k) = -\sum_{k=1}^{K} \hat{p}(R, k) \log(\hat{p}(R, k)) \tag{16}$$

Let us now turn our attention to the leaf node R_1. At this stage, we will go ahead and split each of the leaf nodes into further regions. For instance, at R_1, we will search for another pair (h, t_h) that splits the R_1 into two regions:

$$R_{1,1} \triangleq \{x \in R_1 | x_h \leqslant t_h\}, \quad R_{1,2}\{\triangleq \mathbb{R}_1 - R_{1,1}\}. \tag{17}$$

With $\hat{\mu}_{1,1}$ and $\hat{\mu}_{2,1}$ defined as above, we split R_1 by minimising $L(l, t, R_1)$. After this step, R_1 will no longer be the leaf node. Rather $R_{1,1}$ and $R_{1,2}$ will become the leaf node. R_2 will be split similarly. At each iteration, we can split a leaf node into two leaf nodes,

and each of them into two further leaf nodes, hence the name 'recursive binary splitting'.

At some point, we will want to terminate this splitting process. Clearly, it is impossible to split a region that has only one observation. However, such a small region would imply a very finely divided and overfitted tree, and we want to stop the process some time before that. Thus, we implement a *terminating rule* that inhibits a leaf node from splitting. Two simple terminating rules are: 1. *minimal leaf size*, which disallows splitting once a leaf node falls below a certain size and 2. *maxmium tree depth*, which terminates the algorithm once the tree has reached a certain number of layeres.

Now, let us assume that some of the predictors are categorical (Generalisation To Categorical Predictors). For the sake of concreteness, suppose that x_1 is categorical with L different levels. The feature space is now no longer $\mathbb{R}^p$. Let us denote it by Θ. Then given any region R, if we attempt to split it by x_1, we will search for a pair of index-subset (l, T), where $l \in \{1, .., p\}$ and T is a subset of $\{1, ..., L\}$ and divide R into two sub-regions:

$$R_1 \triangleq \{\mathbf{x} \in \Theta | x_1 \in T\}, \quad R_2 \triangleq R - R_1.$$

The rest of the algorithm proceeds exactly as before.

Boosting

The general idea behind "boosting" is that the "base learners" within an "ensemble" are developed sequentially in a way that each of them focuses on the data which produced worse predictions to now make more accurate predictions, such that the accuracy of the overall model increases. Hence, this should reduce the bias and variance of the outputs. The "base learners" are normally "decision trees" which are constructed through an iterative process namely "recursive partitioning", see [5], which splits the data into an initial partition, and then splits this partition into smaller groups with the process continuing up until a stopping criterion is reached. There are many variations of boosting

and gradient boosting, see [21] and [20] across a plethora of applications. The underlying philosophy is that we can minimise the overall prediction error by combining the best possible next base learner to the previous ones, see for example [19].

The additive mechanism of boosting implies that it is suited for prediction of continuous values. Thus, for classification trees, we must find a connection between the categorical outcome, and the continuous values returned by the additive emsemble method. The technique here is borrowed from GLM: instead of predicting the categorical outcome, we will predict the probabilities of the outcomes. Moreover, at each step, the deviance residual is utilised.

To simplify the exposition, let us consider the case where the response can take only two values: 0 and 1. In this case, the response can be interpreted as coming from a Bernoulli distribution, whose parameters depending on the values of the predictors (or rather pedantically, the region in the feature space it belongs to). We digress briefly to recall some facts about the logistic regression model, the classic binary outcome GLM. In this model, the connection between the Bernoulli success probability and the predictors is established canonically by considering the log-odd, i.e.:

$$\log\left(\frac{\pi}{1-\pi}\right) = \eta,$$

where η is the linear predictor. Moreover, π can be retrieved by inverting the relationship above to $\pi = \sigma(\eta)$, where $\sigma : \eta \to \frac{e^{\eta}}{e^{\eta}+1}$ is the *sigmoid function*.

Firstly, we consider a tree stump where all observations are predicted the same log-odds. That is:

$$\hat{\pi}_{\text{ini}} \triangleq \frac{\sum\limits_{i} y_i}{n}, \quad F_0(x_i) = \log\left(\frac{\hat{\pi}_{\text{ini}}}{1-\hat{\pi}_{\text{ini}}}\right).$$

Before the incrementing step, we need to borrow another concept from GLM. For a logistic regression mode, given a prediction $\hat{\pi}_i$ for each observation, the total deviance

residuals is given by:

$$D(y_i, \hat{\pi}_i) = \left\{ y_i \log(\hat{\pi}_{i)} + (1 - y_i) \log(1 - \hat{\pi}_{i)}) \right\}.$$

At each step, $F_0(x_i)$ is incremented in such a way that minimises the deviance. Mathematically, for $m = 1, \cdots, M$, we fit a tree h_m such that:

$$h_m = \text{argmin}_h \sum_{i=1}^{n} D\left(y_{i,}, \sigma\left[F_{m-1}(x_i) + h_m(x_i)\right]\right),$$

and define $F_m = F_{m-1} + h_m$. The probability of y_i belonging to class 1 is $\sigma(F_M(x_i))$.

Multiclass Classification Case.

Let us now turn into the case where the response can take an arbitrary number of values. For concreteness, suppose there are K possible levels, denoted $\{0, 1, ..., K-1\}$. We observe that, in the binary classification case, the log-odds can be interpreted as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\right).$$

We generalise it by considering the $K-1$ different log odds:

$$\log\left(\frac{\mathbb{P}(Y=j)}{\mathbb{P}(Y=0)}\right), \quad j = 1, \cdots, K-1.$$

The broad idea for this case is to treat these $K-1$ log-odds independently and build $K-1$ corresponding boosted ensemble, before combining them. Since the general framework has been described in details in the binary case, we take the liberty to describe this case rather briefly.

First, define for $j = 1, \cdots, K-1$

$$\hat{\pi}_{\text{ini}}^{(j)} = \frac{\sum_i \mathbb{I}_{\{y_i=j\}}}{n}, \quad F_0^{(j)}(x_i) = \log\left(\frac{\pi_{\text{ini}}^{(j)}}{1 - \sum_{l=1}^{K-1} \pi_{\text{ini}}^{(l)}}\right)$$

For $m = 1, \cdots, M$, define

$$h_m^{(j)} = \text{argmin}_h \sum_{i=1}^{n} D\left(y_i, \sigma\left[F_{m-1}^{(j)}(x_i) + h(x_i)\right]\right)$$

$$F_m^{(j)}(x_i) = F_{m-1}^{(j)}(x_i) + h_m^{(j)}(x_i)$$

At the end, we obtain the following system of $K-1$ equations:

$$\begin{aligned} P(Y_i = 1|x_i) &= P(Y_i = 0|x_i)e^{F_M^{(1)}(x_i)} \\ &\vdots \\ P(Y_i = K-1|x_i) &= P(Y_i = 0|x_i)e^{F_M^{(K-1)}(x_i)}, \end{aligned}$$

which can be simplified thanks to the relation $\sum_{k=0}^{K-1} P(Y_i = k|x_{i)} = 1$. This reduces to:

$$\begin{aligned} P(Y_i = 0|x_i) &= \frac{1}{1 + \sum_{k=1}^{K-1} e^{F^{(k)}(x_i)}} \\ P(Y_i = 0|x_j) &= \frac{e^{F^{(j)}(x_i)}}{1 + \sum_{k=1}^{K-1} e^{F^{(k)}(x_i)}}, \quad j = 1, \cdots, K-1. \end{aligned}$$

One form of boosting is Gradient Boosting. The ensemble here is made up of "base learners" where each "base learner" is a "decision tree". The "gradient boosting" element of this process refers to the underlying algorithm (a "gradient descent" algorithm) which

is used reduce the loss with each incremental additional to the overall "ensemble" model. It is a common choice for classification problems to combine "decision trees" with "gradient boosting" because of the scalability of the approach and optimal results it yields on tabular data, and currently, this seems popular in industry to use "gradient boosting". For more details on "gradient boosting", and detail of the open-source software used, including more mathematical background, please see for example here [31].

Model

Starting from a Greek P&C insurance data-set, described in details in the next section, $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we use gradient boosting to find an approximation, $\hat{F}(\mathbf{x})$, of a function $F^*(\mathbf{x})$, which is the mapping from instances $\mathbf{x}$ to their output values y, via minimizing the expected value of a given loss function, $L(y, F(\mathbf{x}))$. In our case, we solve a multi-class classification problem with the three classes presented in Figure **??**. The loss function is categorical cross–entropy defined as

$$\sum_{i=1}^{N} -y_i \log \tilde{y}_i$$

where $\tilde{y}_i = \Pr[F(x_i) == y_i]$ is the probability that example x_i belongs to class y_i. The probabilities are computed as described in the previous section. Note that the cross–entropy function is convex and the gradient with respect to each tree can be efficiently computed.

By using gradient boosting, we develop an additive approximation of $F^*(\mathbf{x})$ as a weighted sum of functions,

$$F_k(\mathbf{x}) = F_{k-1}(\mathbf{x}) + \rho_k g_k(\mathbf{x}) \tag{18}$$

where ρ_k reflects the weight of the k_{th} function, $g_k(\mathbf{x})$. It is worth noting that these functions are base learners within a model ensemble and these base learners, in our case, are decision trees, see [5]. The construction of this additive approximation of

$F*(\mathbf{x})$ happens iteratively. In particular, firstly, we obtain a constant approximation α of $F^*(\mathbf{x})$ as

$$F_0(\mathbf{x}) = \arg\min_{\alpha} \sum_{i=1}^{N} L(y_i, \alpha) \tag{19}$$

and then the succeeding models need to minimize

$$(\rho_k, g_k(\mathbf{x})) = \arg\min_{\rho,g} \sum_{i=1}^{N} L(y_i, F_{k-1}(\mathbf{x}_i) + \rho g(\mathbf{x}_i)). \tag{20}$$

The value α is constant in the sense that it is independent of the input features x_i. In fact, the optimal constant α optimizes the loss function for the given label distribution. In the case of categorical cross-entropy, α is the proportional distribution of the labels y_i, i.e., $\alpha_c = \sum_{i=1}^{N}(y_i == c)/N$ for each class c.

Nevertheless, the aforementioned optimization problem is not solved directly. In particular, we train each model g_k on a new data-set $D' = \{\mathbf{x}_i, r_{ki}\}_{i=1}^{N}$, where the residuals, r_{ki}, are computed by

$$r_{ki} = \left[\frac{\partial L(y_i, F(\mathbf{x}))}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{k-1}(\mathbf{x})} \tag{21}$$

and then the value of ρ_k is subsequently calculated using a line search optimization strategy.

Of course, as with other machine learning algorithms the prediction performance of gradient boosting ensemble is assessed using a separate test set with unseen data points or using cross validation techniques. An important advantage of gradient boosting compared to other methods such as logistic classification or a neural network model is that we don't need careful feature scaling which is appealing given our data-set. A

scheme of the gradient boosting method for our classification problem is presented in Figure 2. The gradient boosting modelling results for the Greek P&C data-set follow in the next section.

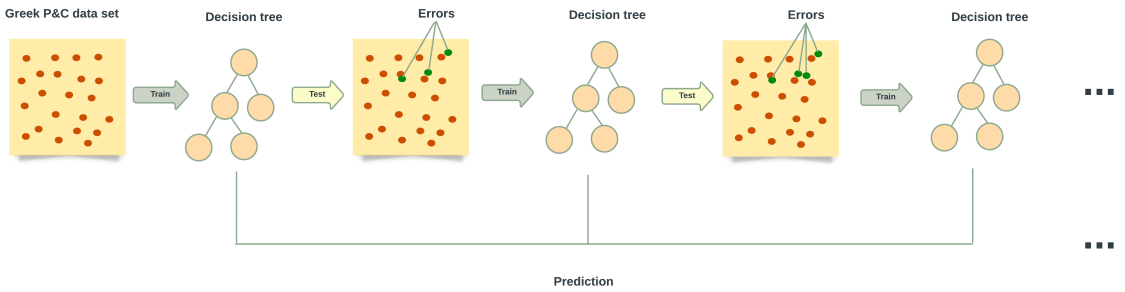


Figure 60. Basic diagram to illustrate the process of "gradient boosting" process.

Data-set Details and Manipulation

We have taken annual data from a local insurer in Greece, from different insurer branch locations across Greece. The data consists of over 60 "columns" of data. There was no material manipulation of the data before training - this is an advantage of using "decision trees" and "gradient boosting". There are around 30 columns of data used for classification. Each branch is described by a number of features including meteorological conditions, the number of policies for different types, and in particular the number of significant events with wind or rain during the year. Also, we have the number of claims as a feature. Again, these features are for each branch, aggregated over a full year. There are no timestamps in the data.

Specifically, the data-set represents the filed insurance claims as a result of weather anomalies in different geographic areas in Greece, represented by their postcodes, between the years 2012 and 2019 for a specific local insurer. There are 5216 unique postcodes, and in each postcode one or more insurer branches can be present. In total, there are six branch categories that handle different accidents such as properties, cargo, motors, etc. In total, there are 36939 examples. Each such example represents the number of

claims in a given branch with existing insurance policies, at a given postcode, for a given year. For example, in 2017 at postcode 41335 there are two insurance claims for damaged properties, one for a motor vehicle and no filed claims for yachts or cargo. The figure below represents the number of claims from such the data-set.

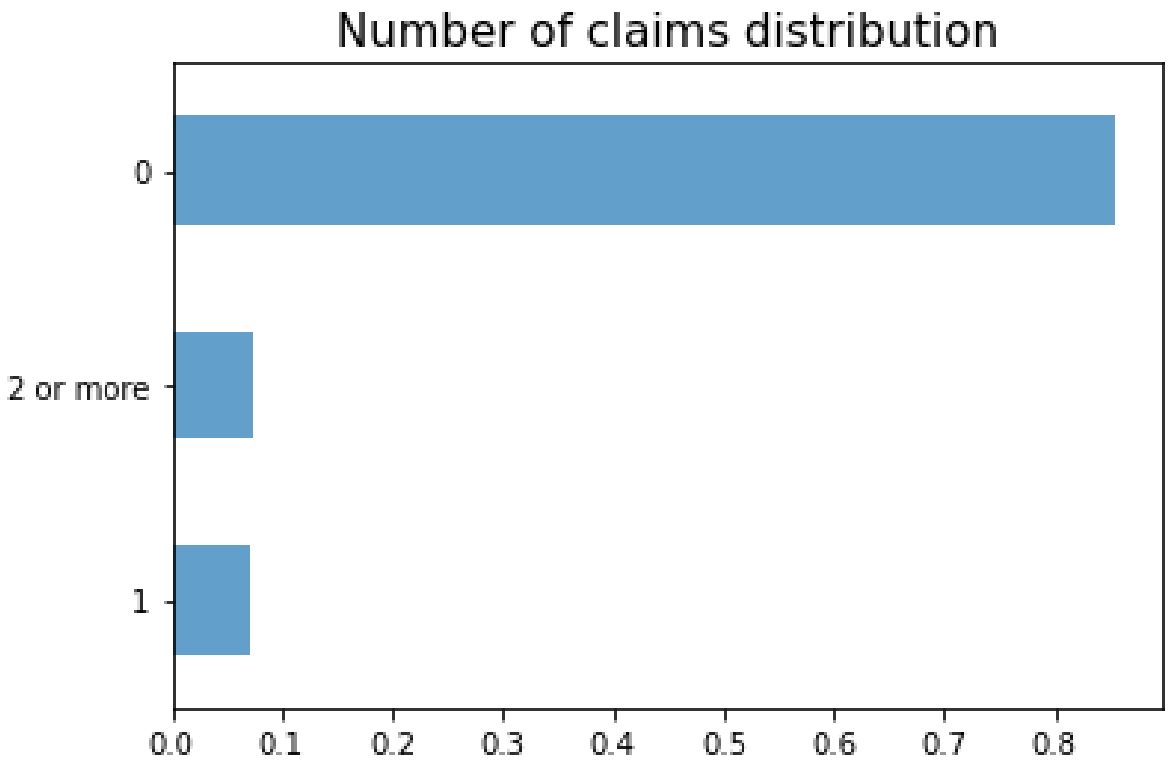


Figure 61. Number of Claims Distribution.

As it can be seen, the data-set is highly imbalanced, with 85.5% of the values being 0 (no claims), 7% with a single claim and 7.5% with two or more claims, see Figure **??**. In total, the data has 88 such features. (Note that many features are one-hot encoded categorical features. For example, the CORINE land cover classification introduces 44 unique features.) The postcodes are described by a number of features that can be divided into several categories:

- Geographic features. These features describe the postcode area and contain information such as mean, minimum and maximum altitude, the length of coast, surface roughness and slope, etc. Also, the CORINE land cover codes of the area, see [16].

- Meteorological features. Information about the weather conditions such as wind and rain intensity, flood vulnerability index, etc.

- Insurance features. The number and different kinds of insurance policies.

LGBM Model Results

We use the Python framework LightGBM [31] , which is free and open source. It was developed in part by Microsoft. LightGBM is a *gradient boosting framework that uses tree based learning algorithms,* according to their main Github [21]. For more commentary on LightGBM see for example [31]. LightGBM contains highly efficient implementations of various gradient boosting algorithms for classification and classification tasks.

We divide the data into three subsets using the train, validation, test paradigm as follows:

- Train data. Examples between the years 2012 and 2016, consisting of 20143 examples. The data is used to learn a sequence of decision trees.

- Validation data from 2017 and 2018, in total 11150 examples. This data-set is used to validate the performance of the model and avoid over-fitting.

- Test data from 2019, in total 5,646 examples. The data is used to report the performance of the model on unseen data and simulate the behaviour of the model on new data.

[21] Github page url: https://github.com/microsoft/LightGBM

The model itself outputs predicted probabilities allocating to one of the three categories ("0 claims", "1 claim", "2+ claims". An example output could be [0.2, 0.7, 0.1] - where the model is inferring this the second category in our list "1 claim".)

We train the model using early stopping on the validation data-set for regularization. More precisely, if the loss function (in this case a log loss function, called "categorical cross entropy" - see below) has not improved on the validation data-set for 50 consecutive iterations, i.e., 50 new decision trees, we stop training. The plot depicting the evolution of training and validation error is shown in the figure below. As usual, the loss improves for both data-sets for some time and then the model starts to over-fit the training data.

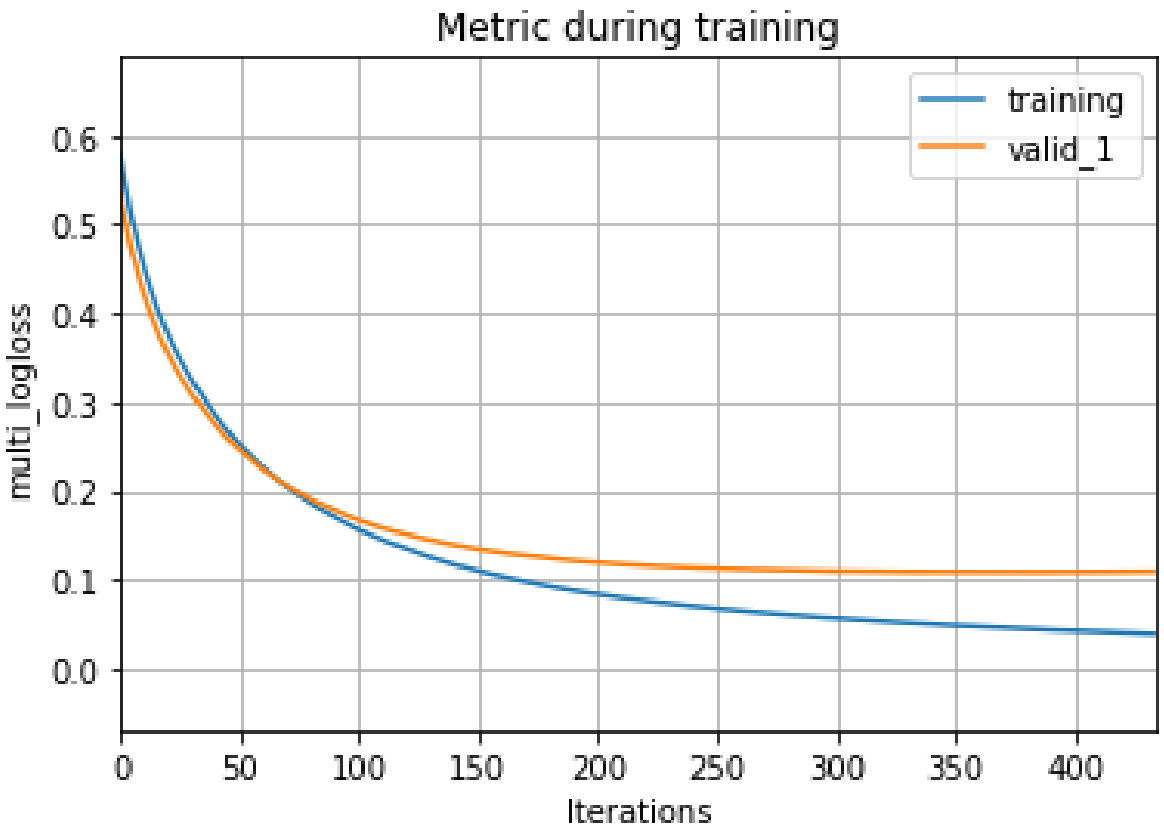


Figure 62. Metric during training.

The overall accuracy of the model is 0.983. The accuracy only for class 1 and 2, i.e., one claim vs. 2 or more claims, is 0.893. In Figure below we show the ROC curves and the corresponding AUC scores for the three classes. We see that the model is highly accurate.

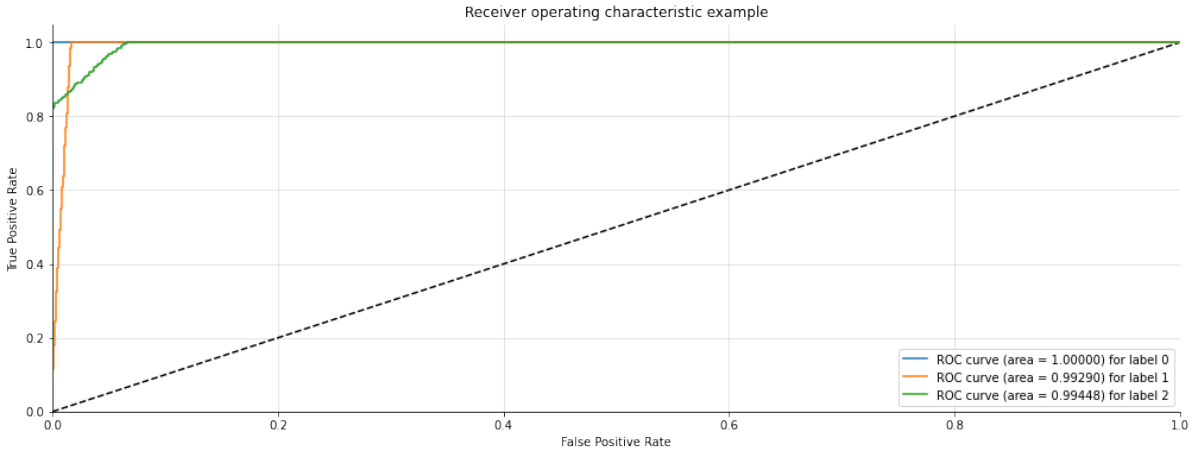


Figure 63. The AUC plots for the three classes.

In Figure below we plot the most important features. As expected, the meteorological conditions have the highest impact on the number of insurance claims, followed by some policy features, and the type of area. Wind vulnerability is very significant as high wind gusts are usually combined with severe weather outbreaks like thunderstorms which cause serious insurance losses.

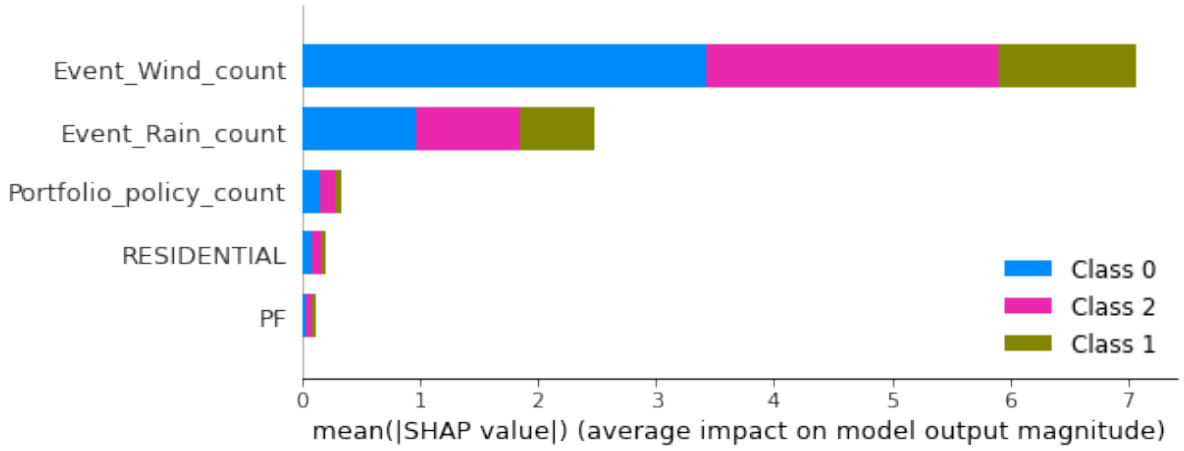


Figure 64. The most important features and their contribution to the classifying each example to a given class.

Insurers may categorise the impact of climate-related claims due to a single issue: a ”storm event”. Trying to understand the influences of each underlying climate-related peril in part can be challenging, due to complexity of the model and claims accuracy, the amount of data it would require to better understand such underlying influences and the cost involved from a business perspective to produce such a model. If we could, however, produce a realistic approach for insurers to categorise the underlying perils and the influence of each peril, this could lead to better pricing and reserving models for an insurer, as well as more appropriate policy premiums being paid by an insured. Given the complexities and potential volumes of data involved, we discussed an example of how machine learning techniques could be used towards this direction. Given the complexity of this problem, as a first step, we analysed insurer data, using “supervised learning” techniques to categorise if a ”storm event” would result in a claim, and if so, how many claims. We used a machine learning algorithm called “gradient boosting” to categorise the claim as ”0 claims”, ”1 claim” and ”2+ claims” based on local insurer data from Greece.

Weather Related Risk - Frequency and Severity Models

Attritional Claims Modeling

Probability of Event Occurrence Estimation

In order to make an estimation about the related capitals to the flood related risk we group portfolio information per underwriting year, business branch and postal code.

Firstly our aim is to estimate the expected annual loss in terms of attritional claims (apart from large). We first model the probability of event occurrence using the below GLM[22] model of the logistic regression.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

For our prediction we use X_1 as the flood vulnerability index and X_2 the policy counts. So the above equation transforms to

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Using the "gamlss" package in R we have the following estimators:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2043  -0.4478  -0.4301  -0.4181   2.2274

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
```

[22] Generalized Linear Model

```
(Intercept)            -2.217e+00  5.548e-02 -39.954   <2e-16 ***
Vln_Ind                -3.593e-02  1.454e-02  -2.472   0.0134 *
Portfolio_policy_count  2.940e-03  5.575e-05  52.744   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25624  on 31214  degrees of freedom
Residual deviance: 21195  on 31212  degrees of freedom
AIC: 21201

Number of Fisher Scoring iterations: 5
```

Both independent variables, vulnerability index and Portfolio policy count are statistically significant (p-value<0.05). In order to model the probability of occurrence we use data for the years 2012-2019 as “training input” for the model.

Using the ROC[23] curves we see that AUC[24] equals to 0.89 which implies a good predictive performance of the model above.

[23] Receiver Operating Characteristic
[24] Area Under Curve

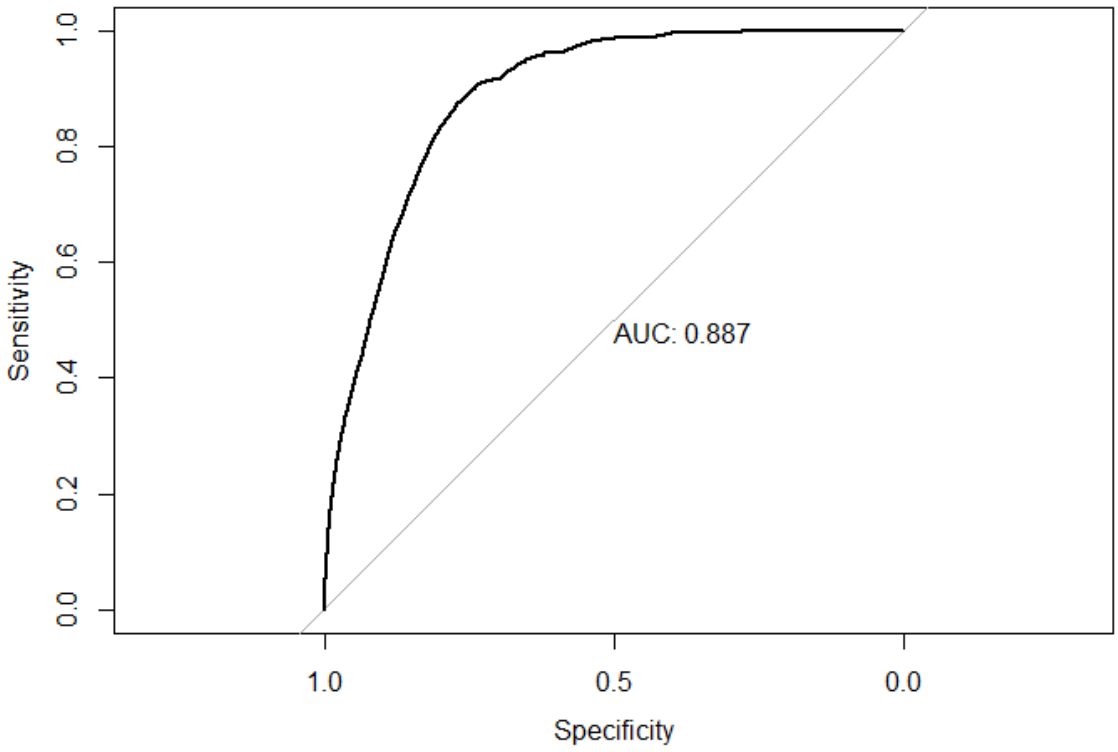


Figure 65. ROC-AUC plot for the above logistic regression model.

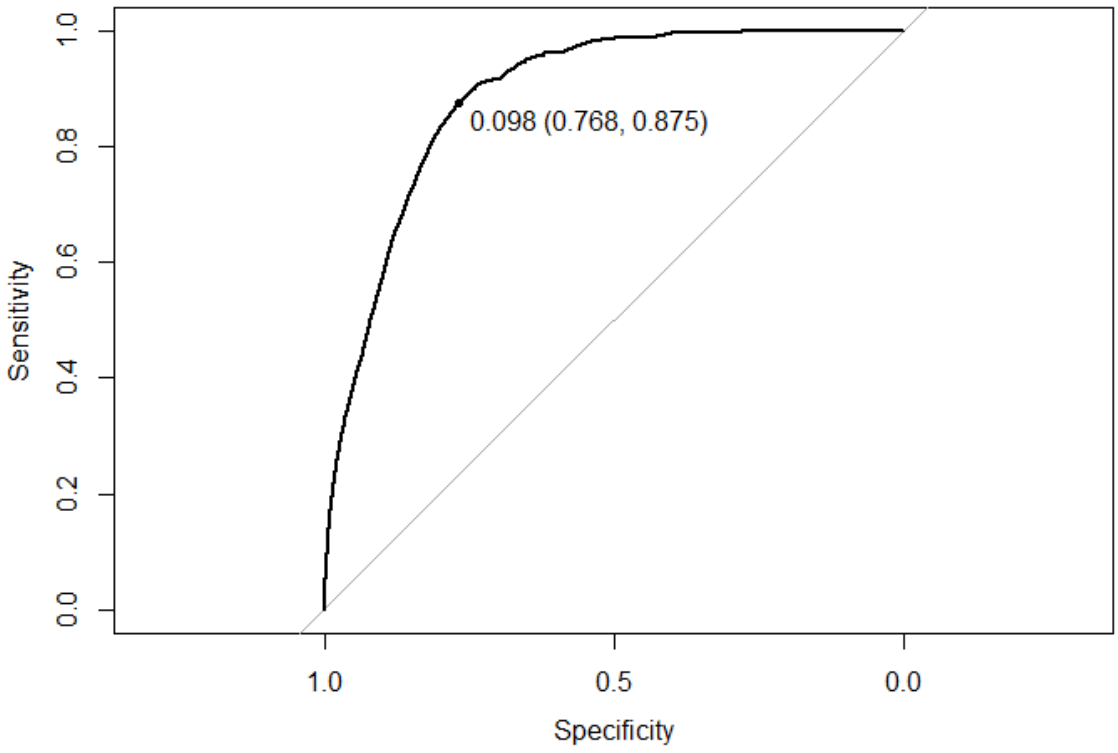


Figure 66. ROC-Threshold selection for the above logistic regression model.

Taking into account the information above and business insights we conclude on a final probability of occurrence above which we expect at least one claim. The final threshold selection is an assumption which determines the "prudency" margin of the final estimation and back-testing should always be made.

Number of Claims Estimation

Secondly we model the number of claims occurred at the same level of aggregation using the Poisson distribution. For our prediction we use again as independent variables the flood vulnerability index and the policy counts. Using the R "gamlss" package we have the following results:

```
glm(formula = Claims_count ~ Portfolio_policy_count + Vln_Ind,
    family = poisson(link = "log"), data = Data_GLM_FRselection)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.843  -1.209  -0.877  -0.096  33.035

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.192e+00  2.856e-02   41.74   <2e-16 ***
Portfolio_policy_count  3.955e-04  6.319e-06   62.59   <2e-16 ***
Vln_Ind                -9.277e-02  7.637e-03  -12.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22329  on 4490  degrees of freedom
```

```
Residual deviance: 19540  on 4488  degrees of freedom
AIC: 31065

Number of Fisher Scoring iterations: 6
```

Severity Estimation

Furthermore, we model the average cost of claims[25] occurred at the same level of aggregation using the Gamma distribution. For our prediction we use as independent variables the flood vulnerability index and the policy premiums. Using the R "gamlss" package we have the following results:

```
glm(formula = Avg_Cost ~ Total_Premium + Vln_Ind, family = Gamma,
    data = Data_GLM_SEVselection)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.78563  -0.43452  -0.24248   0.05657   2.52079

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.707e-04  2.370e-05  15.638  < 2e-16 ***
Total_Premium 1.188e-09  1.403e-10   8.462  < 2e-16 ***
Vln_Ind       2.116e-05  6.812e-06   3.106  0.00193 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.4275545)
```

[25] Average Claims Cost is adjusted for inflation

```
    Null deviance: 502.64  on 1660  degrees of freedom
Residual deviance: 453.89  on 1658  degrees of freedom
AIC: 27472

Number of Fisher Scoring iterations: 6
```

Given the above three models and characteristics of the portfolio for a year period per postal code (number of policies, total premiums) we can estimate the expected annual claims cost in terms of flood events. Using the above characteristics we can predict the claims cost as:

$$E_k(Attr\ Claims\ Cost) = Pr_k(occurence) * E_k(Claims\ Count) * E_k(Avg\ Cost) \quad (22)$$

for each "k" postal code. And the total claims cost as the:

$$Total\ Attr\ Cost = \sum_k E_k(Claims\ Cost) \quad (23)$$

Actually, the total cost calculated above is the expected claims cost which is usually experienced for every year for small and medium claims. The prediction result can be easily checked when compared with the actual figures realised within periods with similar portfolio characteristics.

Large Claims Modelling

Frequency of Large Claims Estimation

In order to estimate the risk it is also essential to run scenarios about the probable large losses that may occur which actually determine the final risk distribution. Large claims are modelled as a whole for all regions because of large claims count restrictions in order to achieve statistical significance and more robust models. The usual issue regarding the distribution fitting of climate related data is the short period of observations.

The actual large claims observed during a eight year period are described below:

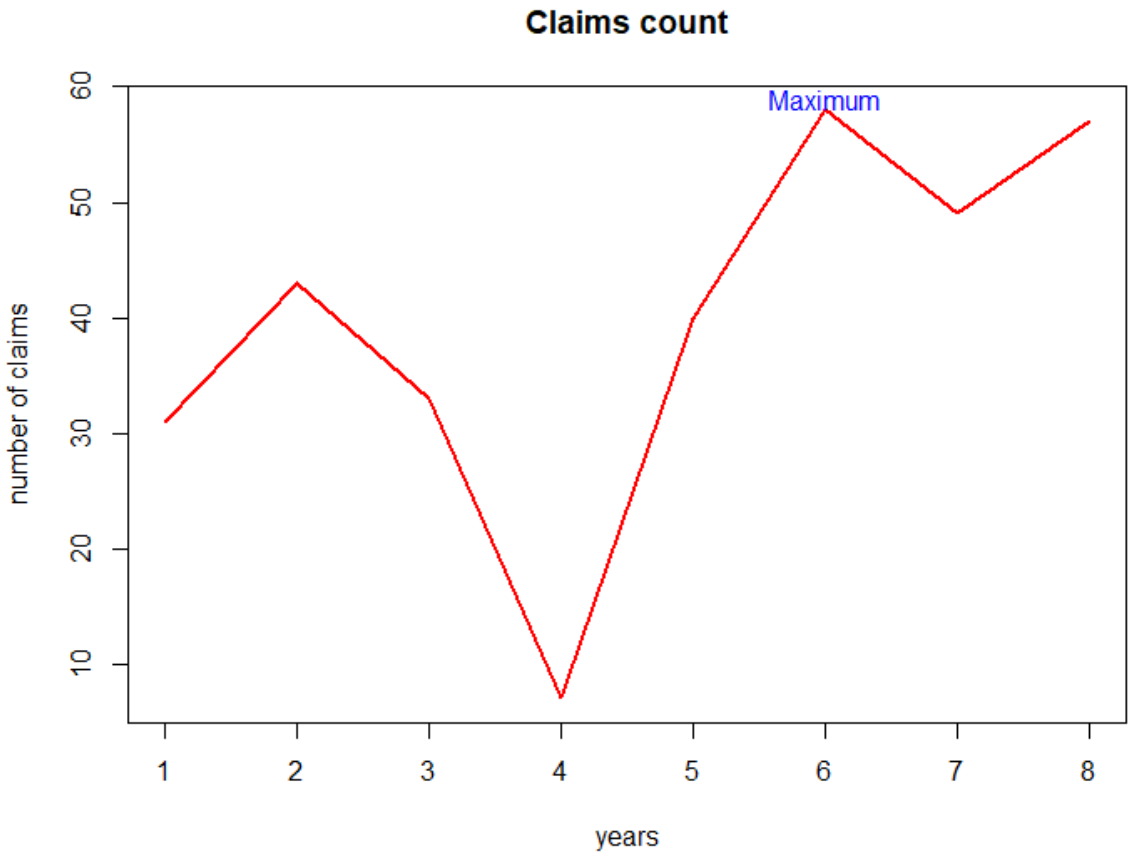


Figure 67. Flood Large Claims observed 2012-2019.

Using a "Cullen and Frey" Graph we conclude that the frequency distribution of claims converges to a negative binomial distribution.

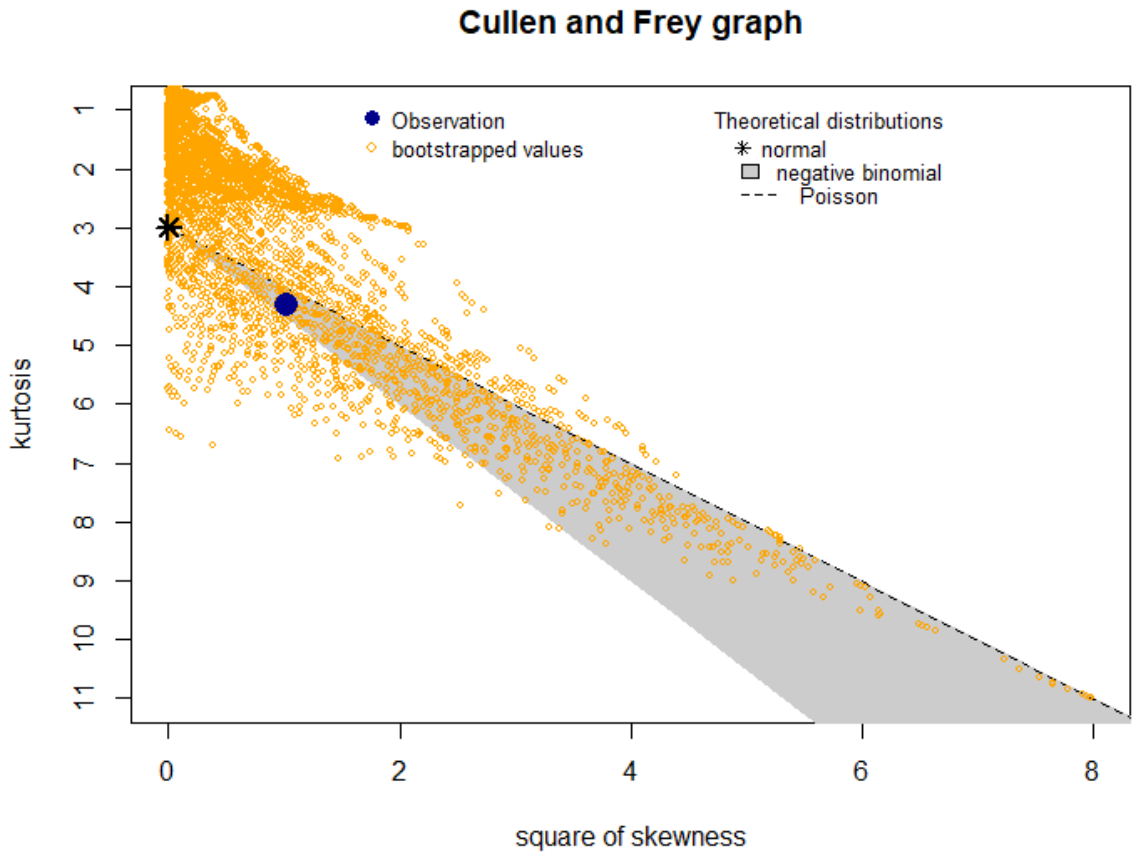


Figure 68. Cullen and Frey graph for the observed number of claims.

Using a Greek insurance market database regarding the claim counts triggered by "NAT-CAT" events and making an exposure adjustment regarding the specific insurance company, we can generate values for a 60 year period as depicted in the graph below:

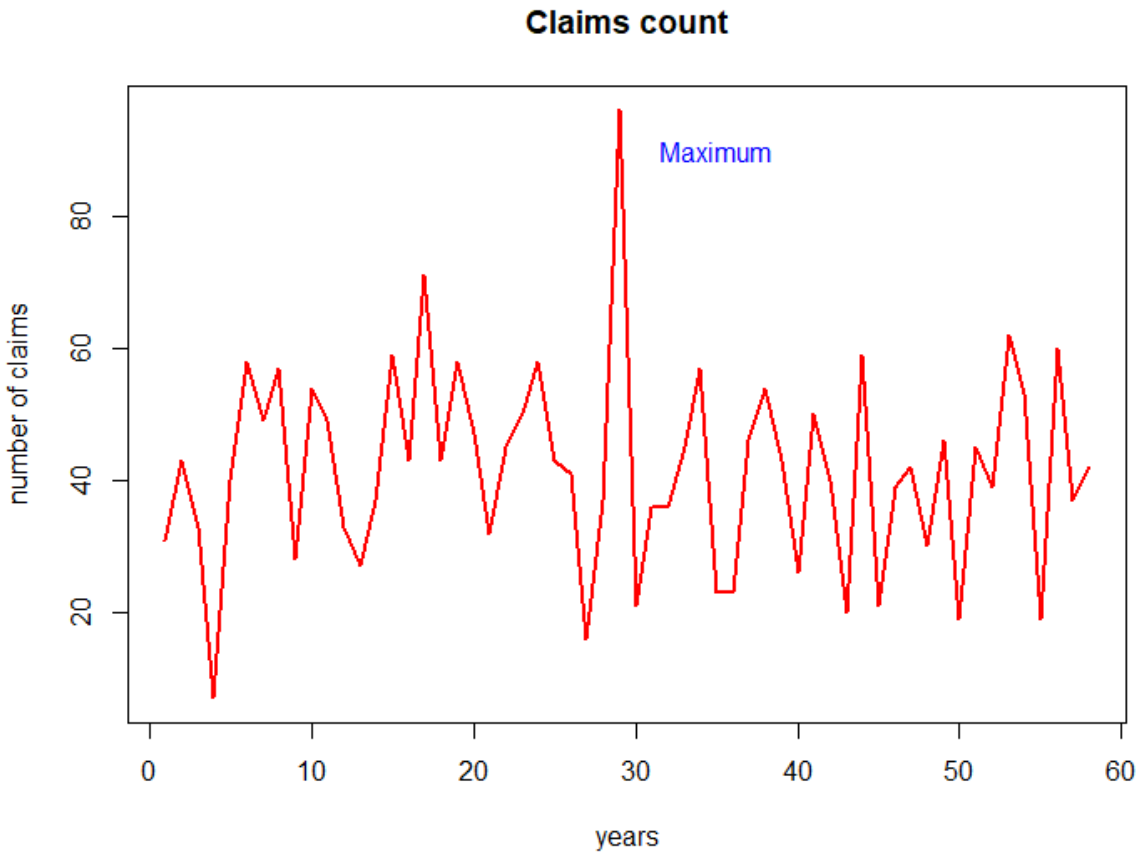


Figure 69. Simulated Large Claims Counts for a sixty year period.

The above time series is produced from a negative binomial distribution taking also into account the actual observations of a decade. The negative binomial distribution fits as follows:

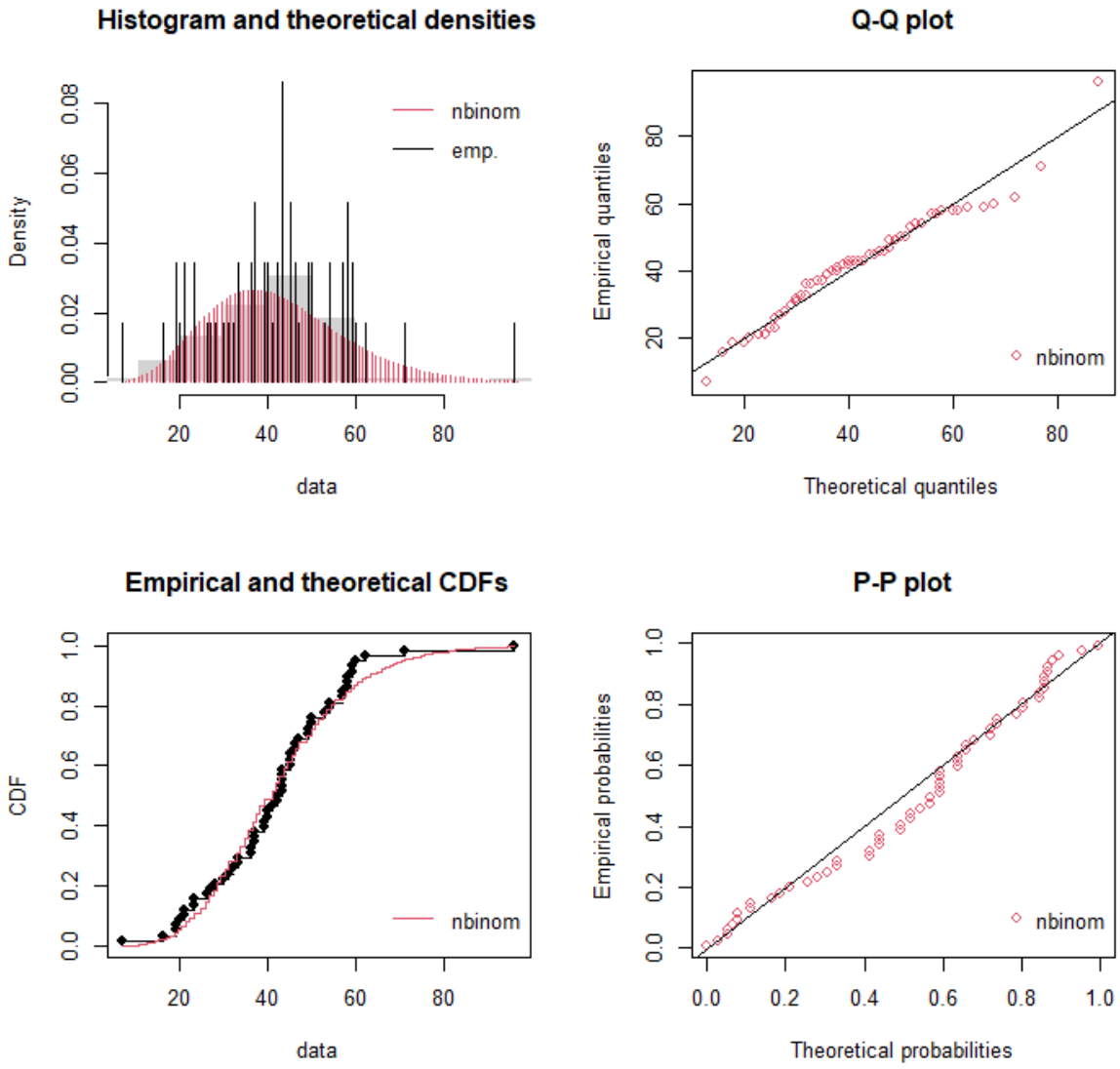


Figure 70. Negative Binomial Distribution Fit for Large Claims related to flood events.

```
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
     estimate Std. Error
size  8.28643   1.865567
mu   41.68482   2.081638
```

```
Loglikelihood:  -240.5728   AIC:  485.1456   BIC:  489.2665
Correlation matrix:
             size           mu
size 1.0000000000 0.0004358663
mu   0.0004358663 1.0000000000
```

Severity of Large Claims Estimation

The Pareto distribution fits flood severity data well:

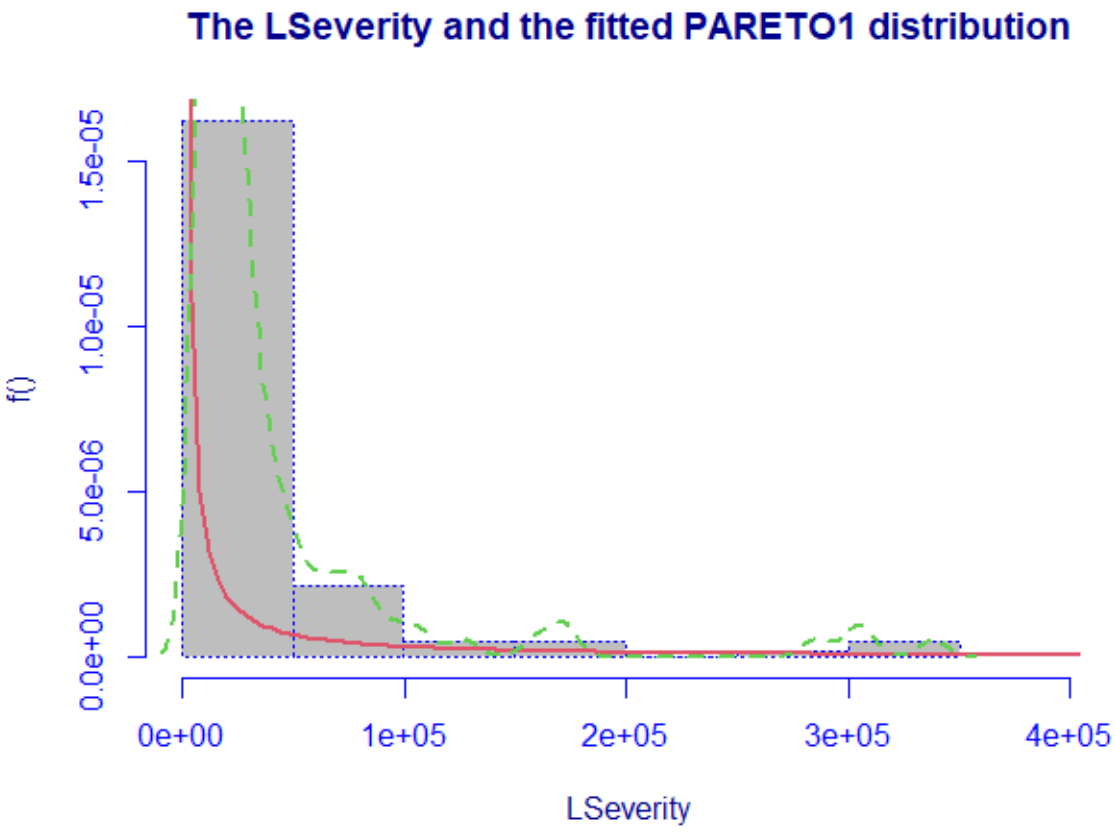


Figure 71. Histogram of Pareto Distribution Fit for Large Claims severity related to flood events.

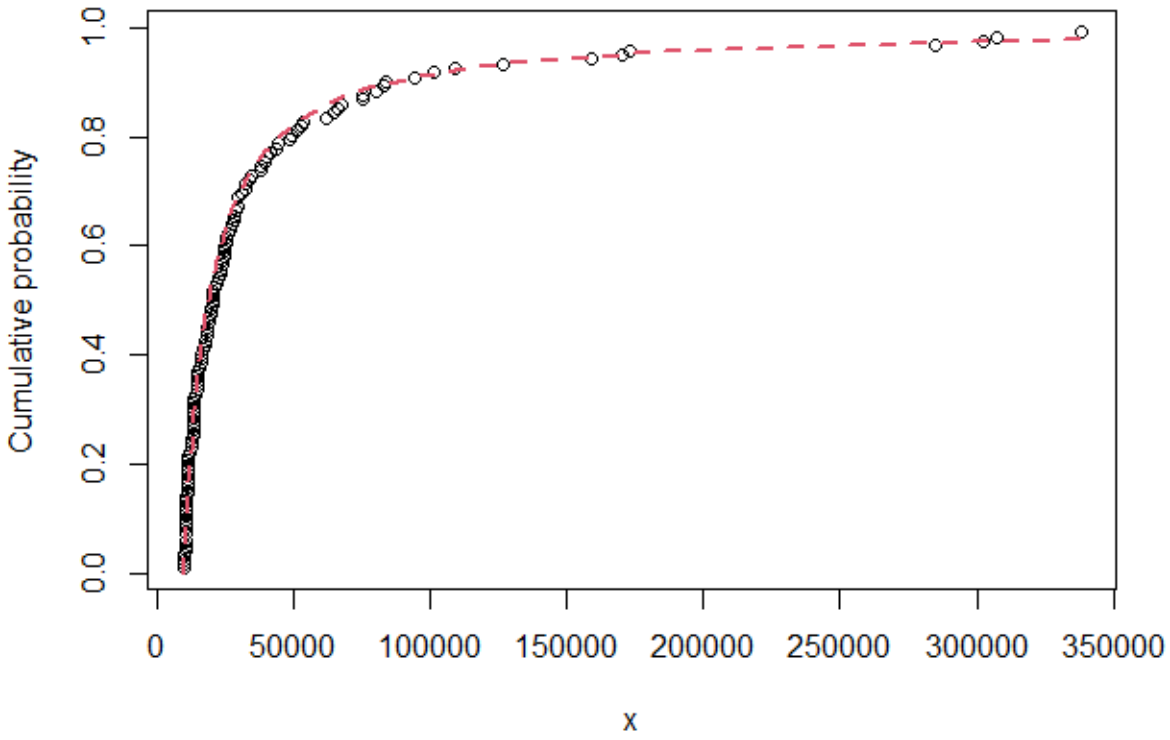


Figure 72. Cumulative Probability of Pareto Distribution Fit for Large Claims severity related to flood events.

```
**********************************************************************
Family:  c("PARETO1", "Pareto one Y>0")

Call:  gamlssML(formula = LSeverity, family = "PARETO1")

Fitting method: "nlminb"
```

```
Coefficient(s):
         Estimate  Std. Error  t value   Pr(>|t|)
eta.mu -2.3167335   0.0909092  -25.484 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Degrees of Freedom for the fit: 1 Residual Deg. of Freedom   120
Global Deviance:     3257.13
            AIC:     3259.13
            SBC:     3261.92
```

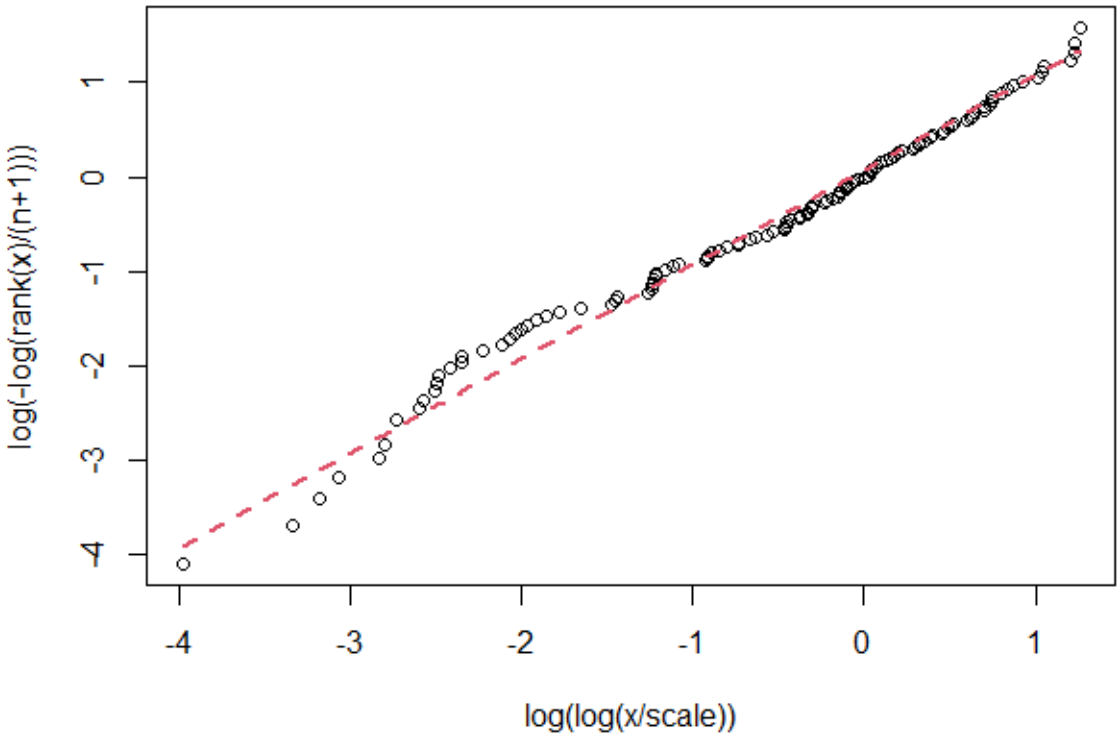


Figure 73. Pareto Distribution Fit for Large Claims severity related to flood events.

Risk Distribution and Capitals Calculation

The total cost distribution of the flood events will be:

$$Final\ Cost_i = Total\ Attr\ Cost_i + Total\ Large\ Cost_i \tag{24}$$

where i is the scenario of the simulations produced for attritional and large claims as follows:

Table 28: Risk Distribution - Simulations Table

Scenario	FR_ur	SEV_ur	Ult_Cost	LC1	LC2	LC3	LC4	LC5	...	LCn	LargeTotal	Final_Cost
1	2676	1803	4824828	25360	16527	9397	23620	98409	...	...	1124807	5949635
2	2603	1929	5021187	47341	59228	94751	17891	64587	...	...	659544	5680731
3	2565	1870	4796550	62547	9505	14028	29464	16929	...	15000	2449732	7246282
4	2576	1761	4536336	9817	12205	35497	9039	122668	...	...	2303975	6840311
5	2556	1889	4828284	9758	25024	50727	18288	18352	...	...	809645	5637929
6	2615	1944	5083560	35885	10452	16107	13588	13578	...	...	1685126	6768686
7	2629	2036	5352644	46198	17746	51716	36992	14130	...	...	2023867	7376511
8	2634	1939	5107326	14320	23700	126739	96339	42260	...	...	1760767	6868093
9	2595	1740	4515300	17497	33560	17595	34339	21431	...	...	1364455	5879755
...	...	...	...	...	...	...	...	...	...	...	...	...
99992	2602	1897	4935994	37006	17930	9896	23835	9494	...	...	1863970	6799964
99993	2514	1646	4138044	9780	12564	13812	15938	21127	...	...	1269876	5407920
99994	2545	1910	4860950	17346	15002	42662	10807	11719	...	...	5351494	10212444
99995	2614	1950	5097300	9664	10278	13482	58669	14470	...	...	826993	5924293
99996	2579	1776	4580304	23548	20696	10807	12495	12397	...	...	2417650	6997954
99997	2657	2028	5388396	34169	13211	401206	15754	18857	...	...	1372911	6761307
99998	2562	1738	4452756	15439	44613	19001211	15470	9639	...	...	20258774	24711530
99999	2582	1961	5063302	71906	9542	9434	13870	12655	...	...	2371739	7435041
100000	2574	1650	4247100	67313	225015	19840	13876	18596	...	...	3047261	7294361

Simulations with 100000 scenarios are produced and for each scenario an expected frequency for attritional claims (Fr ur), an average cost for attritional claims (SEV ur) and a number of Large Claims are generated. These are depicted separately, mainly for reinsurance reasons.

$$Final\ Cost_i = Frequency_i(Attr) * Severity_i(Attr) + \sum_{k=1}^{n} LC_k \tag{25}$$

Table 29: Risk Distribution - Percentiles

Quantiles	Claims Cost
0	4109252
0.05	5205671
0.1	5396650
0.15	5537848
0.2	5652501
0.25	5764613
0.3	5868221
0.35	5972332
0.4	6075012
0.45	6182608
0.5	6296814
0.55	6419843
0.6	6558555
0.65	6714471
0.7	6897483
0.75	7122886
0.8	7433280
0.85	7883690
0.9	8666834
0.95	10781697
0.995	45618570
1	82154000

A risk distribution with a mean of €7.83 mil. approximately has an 99.5% percentile value of €45.6 mil. as depicted in the table above. This means that the required capitals for the flood risk should be €37.78 mil. (99.5% percentile minus the distribution mean)

In order to understand better the risk spatial distribution, and assuming that large claims have a similar pattern with attritional claims, we could dis-aggregate capital requirements with the estimated risk premium driver.

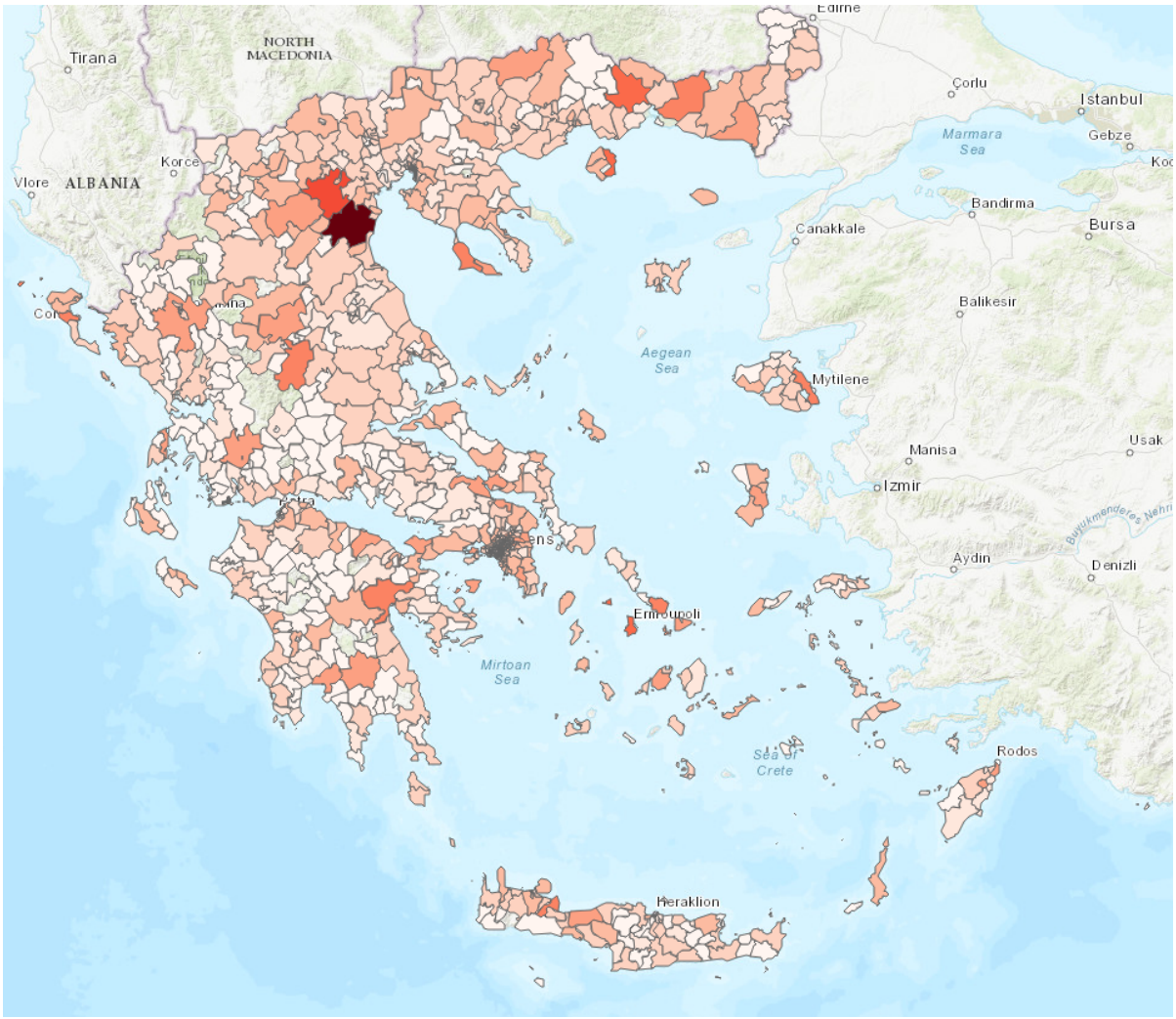


Figure 74. Insurance weather related risk spatial distribution in Greece.

Regions of Central-West Macedonia, Thrace, Crete and Western regions of Greece are highlighted in terms of insurance risk significance. The metropolitan Area of Attica is also considered a hot spot.

Portfolio characteristics, claims experience and realization and weather phenomena are the refined components which finally produce the above map.

Multivariate Regression Models on Weather Related Events

The links between climate change and insurance sector are well recognized in the literature both for the impact that the former has on the latter, see, for example, [47], [49], [53], [7], [52], [59], [46], [40], [58], [30], [23] and the growing importance of insurance sector's role in addressing the effects of climate change, see, for example, [55], [29], [6], [50]. With regards to relevant statistical research on the topic of climate change and general insurance, see, for instance [62]. We see that the focus lies on traditional insurance modelling tools used to assist decision-makers in estimating and managing catastrophe insurance risk in the past three decades adapted to assess the impact of these changes to the insurance landscape. However, to the best of our knowledge, there is no relevant literature on multivariate regression models for quantifying the impact hazards due to climate change to the number of claims in the general insurance sector.

In this chapter, a study is motivated by the stylized characteristics of a climate-related motor insurance data-set of a major insurance company operating in Greece, we introduce a number of alternative compound claim frequency models with the aim to jointly model the actual storm and claim arrival processes and understand how they are jointly impacted by various geo-spatial features. In what follows, we provide a detailed discussion of our contributions:

- Our main achievement is to unravel the joint distribution of the actual number of storms and the number of claims triggered by each storm by constructing multivariate regression models. The key challenge is that the observations are incomplete, so it is not appropriate to model the joint distribution directly.

- Also, we include geospatial covariates since we aim to understand their impacts on the storm and claim frequencies, which are important for both actuarial ratemaking and capital estimation methodology. Also, note that the covariate influence on the storm frequency has a different implication from the covariate influence on the number of claims per storm for an insurance company. For instance, consider two

geographical regions, one with a high storm frequency but a low claim frequency per storm, and vice versa for another area. The latter case may be more prone to climate change (i.e., storms become more frequent) because an additional storm will result in many claims.

- Furthermore, we reveal the characteristics and particularities of various exposures (i.e., different types of vehicles) and how these affect the reported claims frequency. Consequently, these explicitly help insurance company to determine specific features of a "NAT-CAT" reinsurance treaty, which is activated to cover an excess amount attributed to natural catastrophic events.

- Finally, we uncover the actual intrinsic dependence between the storm and claim frequencies. After accounting for the overestimated correlations between the empirical storm and claim frequencies, we discover a negative correlation between the actual storm frequency and per-storm claim frequency. This may be explained by the situation where drivers adapting to more frequent storms may change their driving behaviors due to more storms. For example, drivers may become more careful and drive less when they expect poorer weather conditions. As a result, climate change may not be that bad in triggering excessive insurance claims because of the diversifying effects on policyholders.

The remainder of this analysis proceeds as follows. In Section 7.1, we provide a detailed description of the motivating climate-related Greek insurance dataset. The proposed methodology and the alternative compound claim frequency models we derive for jointly modelling the actual storm and claim arrival processes are presented in Section 7.2. In Section 7.5, we present the analyses of the fitting results including estimated parameters associated with the standard errors and a goodness-of-fit comparison for the alternative model specifications. Finally, concluding remarks are given in Section 8.

Regression Models - Data-set description

In this section, we present a climate-related insurance data-set kindly provided by a Greek insurance company, which underwrites and insures a portfolio of policyholders with the natural phenomena coverage. The data-set records the number of climate events and storm-triggered motor claims filed by the policyholders for the years 2012 to 2019, grouped by the policy year and postcode (geographical location). There are 9,302 entries in total, and each observation contains the following grouped information:

In this section, we present a climate-related insurance data-set kindly provided by a Greek insurance company, which underwrites and insures a portfolio of policyholders with the natural phenomena coverage. The data-set records the number of climate events and storm-triggered motor claims filed by the policyholders for the years 2012 to 2019, grouped by the policy year and postcode (geographical location). There are 9,302 entries in total, and each observation contains the following grouped information:

1. **Storm and claim information**

   (a) The total number of storms (rain or wind events) that triggers at least one claim in a policy year within a particular postcode is denoted as M. The count concerns significant meteorological events regarding heavy rain, flood, and strong wind. An event is considered significant by meteorologists based on their expert judgments considering both qualitative and quantitative criteria. For example, total rainfall greater than 50 mm, maximum precipitation rate greater than 50mm/h, wind gust greater than 80 km/h, incidents reported to police or fire department, and/or expanded reported infrastructure problems.

   (b) The total number of claims attributed to a severe wind or rain storm in a policy year within a particular postcode is denoted as N. In the dataset, reported insurance claims per postcode are relevant to the corresponding magnitude of the exposure per area and its characteristics. For instance, we may have a significant number of severe meteorological events for a specific

postcode but minimal exposure (only a few policies), and consequently the climate events will bring no material impact.

The empirical mean and variance for M are 0.5396 and 1.1871 respectively, while these values are 1.4497 and 49.4345 respectively for N. Hence, the storm counts M and claim counts N are both overdispersed, but the overdispersion of N is much more severe. The largest M and N observed from the data are 14 and 182 respectively. We plot a matrix table showing the frequency of (M, N) in Table 30, where the vertical and horizontal axes refer to M and N respectively. We observe several relationships between M and N:

- Firstly, $M = 0$ implies $N = 0$, and vice versa. The storms that trigger no claims are not reported to the insurance company and hence they are not counted as M. On the other hand, N contains only the claims that are relevant to a significant meteorological event. In practice, it is difficult for an insurance company to obtain the full storm data, which includes the storms that do not trigger any claims. It is because insurance company is not a weather observatory expert that extensively collects climate data day-by-day. Instead, they are more concerned about underwriting the policies and keeping track on the claim history from the policyholders, which are more relevant to the insurance pricing perspective. Therefore, they usually have richer information on policyholder characteristics and claims history, while they are historically less incentivized to collect granular climate information for each specific region, which can be very costly given that the judgement on the significance of climate events involves extensive qualitative and quantitative assessments from the meteorologists (as mentioned above).
- Secondly, N is always greater than or equal to M. This is a direct consequence of the incomplete nature of the data-set, where every storm that is counted towards M must trigger at least one claim.
- Thirdly, the empirical distribution N is severely heavy-tailed. There are substantial number of observations where the number of claims N is greater

than 9.

- Fourthly, M and N are empirically highly positively correlated (Pearson correlation: 0.5628; Spearman's correlation: 0.9922; Kendall's tau: 0.9627). The high correlation reflects the triangular distributional structure of (M, N) (see Table 30), which is attributed to the truncated nature of data, where M only counts the storms that trigger claims N. Therefore, solely based on the empirical correlation, one cannot conclude that the storm frequency is intrinsically correlated to the claim frequency per storm.

	N=0	1	2	3	4	5	6	7	8	9	>9
M=0	6529	0	0	0	0	0	0	0	0	0	0
1	0	1191	194	73	37	24	10	6	9	8	26
2	0	0	361	129	42	25	23	14	7	6	55
3	0	0	0	102	51	34	21	9	10	9	48
4	0	0	0	0	18	18	13	16	8	10	31
5	0	0	0	0	0	5	10	6	2	6	41
6	0	0	0	0	0	0	1	5	2	2	23
7	0	0	0	0	0	0	0	0	0	0	15
8	0	0	0	0	0	0	0	0	0	0	7
9	0	0	0	0	0	0	0	0	0	1	4
>9	0	0	0	0	0	0	0	0	0	0	5

Table 30: Frequency table of (M, N) for the observations under the "Motors" branch.

2. **Geospatial features**

Several geospatial features, including *vulnerability*, *coast_ratio*, *log_alt_mean*, *residential*, *industrial*, *agricultural* and *wetlands*, may provide predictive power on the storm and claim frequencies. The summary statistics of these explanatory variables are provided in Table 31. Figure 75 presents how the covariates are distributed in the Greek climate insurance data-set. Note that *forest* is the reference level for the proportions of land uses.

(a) *Vulnerability* is an indicator of flood vulnerability due to rainfall and wind storm. Meteorological, geophysical and socio-demographic conditions are

taken into account by meteorologists and the insurance company, and each postcode is rated from 1.0 (least vulnerable) to 4.0 (most vulnerable) subject to the insurance company's discretion.

(b) *Coast_ratio* is the ratio of perimeter of a region that is adjacent to the coast. For example, if a postcode refers to an island, then the corresponding *coast_ratio* is 1. Larger *coast_ratio* implies that the region might be more exposed to the flood risk.

(c) *Log_alt_mean* is the log-transformed value of mean altitude of a region. Due to its definition, it tends to be inversely proportional to *coast_ratio*.

(d) *Residential* represents the proportion of land area within a postcode that are mainly occupied by dwellings and buildings used by administrative or public utilities, including residential buildings, individual houses, parking lots, transport networks and urban greenery. The distribution of this variable is heavily bi-polarized. 62.1% of the postcode regions have $residential < 0.05$, meaning that these regions are mainly rural or industrial areas that may be rather sparsely populated. On the other hand, 12.4% of the postcode regions have $residential > 0.95$, meaning that these regions are urban and densely populated.

(e) *Industrial* represents the proportion of land area within a postcode that are mostly occupied by buildings, other built-up structures and artificial surfaces, which include road and rail networks, sea/airports, and mineral extraction sites. This class is assigned for land units that are under industrial or commercial use or serve for public service facilities.

(f) *Agricultural* represents the proportion of land area within a postcode that is mainly used for agricultural purposes, which include rice fields, vineyards, olive groves, and pastures.

(g) *Wetlands* represents the proportion of land area within a postcode that is flooded either permanently or seasonally, which include inland marshes, peatbogs, and coastal lagoons.

Table 31: Summary statistics for the explanatory variables

Variable index	Variable name	Min	Median	Mean	Max
1	*vulnerability*	1.0	2.5	2.413	4.0
2	*coast_ratio*	0.000	0.000	0.1886	1.000
3	*log_alt_mean*	-0.091	5.298	5.173	7.259
4	*residential*	0.009	0.024	0.250	1.000
5	*industrial*	0.000	0.006	0.056	1.000
6	*agricultural*	0.000	0.261	0.300	0.978
-	*forest*	0.000	0.403	0.383	0.999
7	*wetlands*	0.000	0.001	0.012	0.450

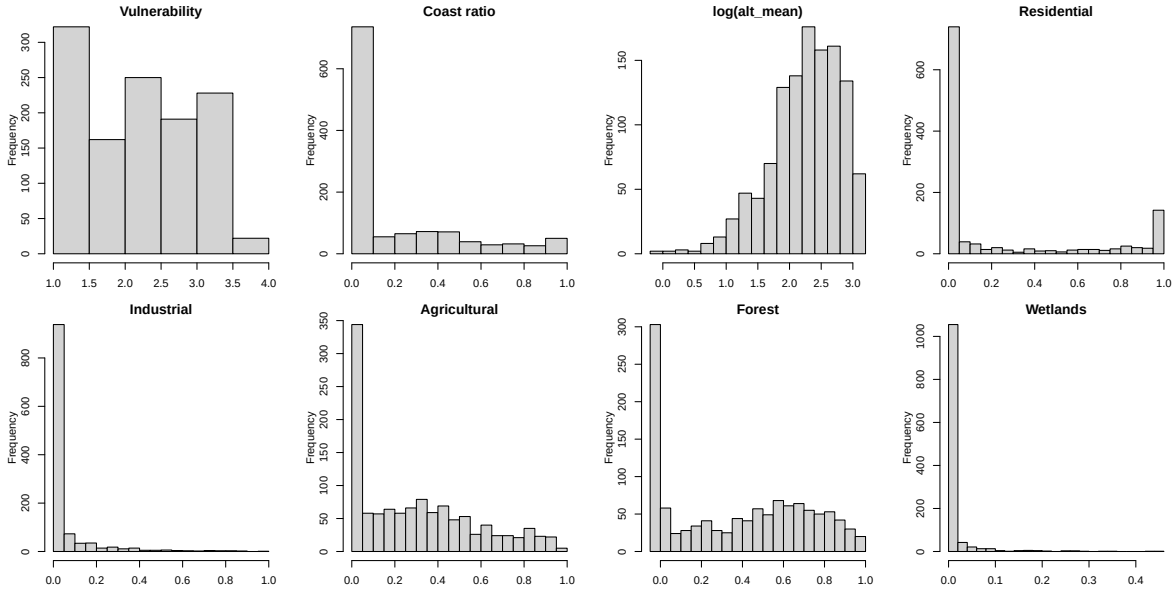


Figure 75. Distribution of observed postcodes per each explanatory variable.

3. **Exposure information** Since the data-set is grouped, there are multiple policyholders within a specific postcode in a given year. There are also three types of vehicles insured by the company: motorcycle, car and truck. Different types of vehicles may be impacted differently by a storm. For example, poor weather conditions may be more likely to trigger a motorcycle crash than an ordinary car crash. In the insurance perspective, we call an insured vehicle an "exposure unit", and in this

data-set there are three exposure types. We expect that the number of exposure units for each exposure type is directly influential to the claim frequency per storm, not the actual number of storms. Figure 76 shows the distribution of the number of each type of exposure units along with the covariates.

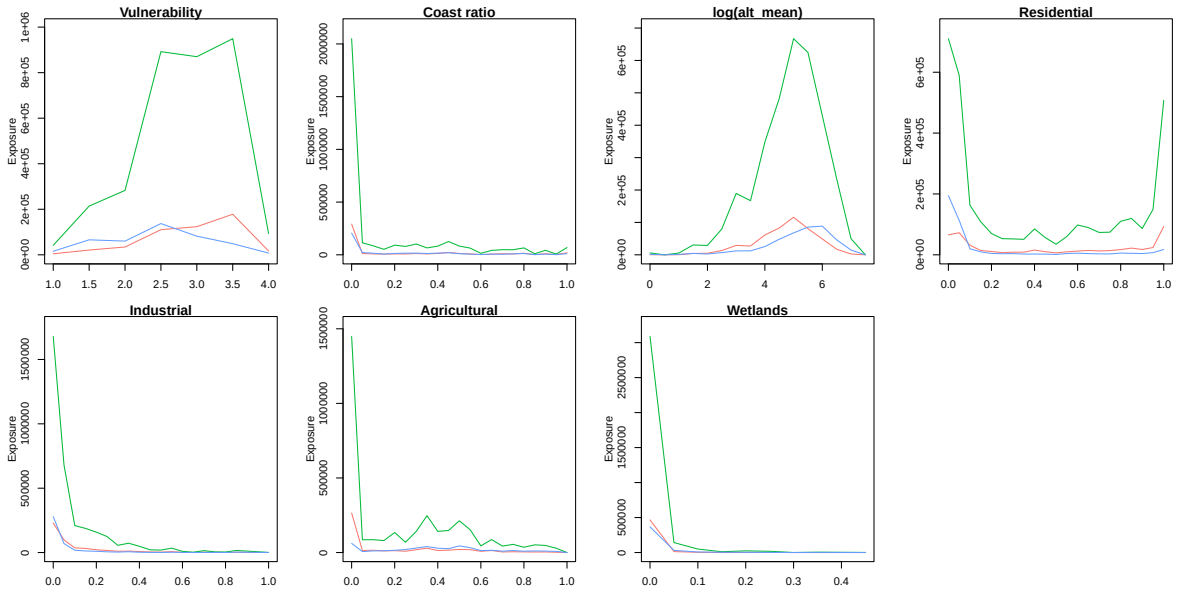


Figure 76. Distribution of exposure units per each explanatory variable (red: motorcycle, green: car, blue: truck)

Note that the following challenges need to be considered in a statistical modelling of the data-set. Here the observed M and N are incomplete, so it is inappropriate to directly model the distribution of M and consider its relationship to the distribution of N. More specifically, the observed storm frequency underestimate the true number of storms due to the reporting issue mentioned above. Besides, due to the grouped data structure, only the total number of claims are recorded for each policy year and postcode, not the number of claims per storm. Such truncation mechanism of the data-set leads to a disproportionately high correlation between the observed M and N. Therefore, it is problematic to perform regression analyses directly to observed M and N, which necessitates development a model that captures the intrinsic actual storm and claim arrival processes and adjusts for the biases resulted from data incompleteness. Indeed,

because of the overestimated correlation, direct regression analyses will likely show that every covariate has similar impacts to M and N, even if e.g. a variable has a positive effect to the number of storm but negative effect to the claim frequency triggered by a storm.

Methodology

As described in the previous section, we aim to model jointly the actual storm and claim arrival processes, and understand how they are jointly impacted by various geospatial features. Therefore, it is desirable to propose multivariate regression models on the actual storm and claim frequencies that can be inferenced by the incomplete data (M, N). Before that, it is essential to mathematically understand and formulate the truncation mechanism of the data-set as follows. Denote $\tilde{M}$ as the actual number of storms occurred within a particular geographical region and policy year. Also, let $\tilde{\boldsymbol{N}} := (\tilde{N}_1, \ldots, \tilde{N}_{\tilde{M}})$ be a vector of claim counts triggered by each storm. The mathematical relationships between the observed data (M, N) and the complete data $(\tilde{M}, \tilde{\boldsymbol{N}})$ are

$$M = \sum_{t=1}^{\tilde{M}} 1\{\tilde{N}_t > 0\}, \quad \text{and} \quad N = \sum_{t=1}^{\tilde{M}} \tilde{N}_t. \tag{26}$$

Also, each observation is equipped by K exposure units, representing the total number of policy counts in a postal region. Denote $\tilde{N}_{tk}$ $(k = 1, \ldots, K)$ as the number of claims triggered by storm t for exposure unit k. Therefore, $\tilde{N}_t$ is the aggregated annual number of claims in the postal region, and the relationship between $\tilde{N}_t$ and $\{\tilde{N}_{tk}\}_{k=1,\ldots,K}$ is

$$\tilde{N}_t = \sum_{k=1}^{K} \tilde{N}_{tk}. \tag{27}$$

As mentioned previously, there are multiple exposure types in the dataset, so $\tilde{N}_{tk}$ and $\tilde{N}_{tk'}$ $(k \neq k')$ may not be identically distributed if the k-th and k'-th exposure units belong to different exposure types because they may have heterogeneous risk

characteristics. For example, the claim rate of a motorcycle (one type of exposure) may be higher than that of an ordinary car (another type of exposure). In this section, we study several candidate models for the complete data $(\tilde{M}, \{\tilde{N}_t\}_{t=1,\ldots,\tilde{M}})$ such that the corresponding joint distribution of the incomplete data (M, N) is still analytically tractable.

Models

Poisson-Poisson model

We first consider a Poisson-Poisson (PP) model to jointly model the actual storm and claim frequencies. Under this model, $\tilde{M}$ and $\tilde{N}_{tk}$ are both assumed to follow Poisson distributions, denoted as $\tilde{M} \sim \text{Poisson}(\lambda_{(M)})$ and $\tilde{N}_{tk} \sim \text{Poisson}(\lambda_k)$, where $\lambda_{(M)}$ and λ_k are the mean parameters of the Poisson distribution. The subscript k in λ_k reflects the heterogeneity among various exposure types. For simplicity, we further assume that $(\tilde{M}, \{\tilde{N}_{tk}\}_{t=1,\ldots,\tilde{M},k=1,\ldots,K})$ are jointly independent, so that $\tilde{N}_t \sim \text{Poisson}(\sum_{k=1}^{K} \lambda_k)$. Suppose that there are only J types of exposures, and let K_j $(j = 1, \ldots, J)$ be the number of exposure units belonging to the j^{th} type. For example, in this dataset there are $J = 3$ types of vehicles, and hence K_1, K_2 and K_3 are respectively the number of motorcycles, cars and trucks. Denote $\boldsymbol{x} \in \mathbb{R}^P$ as the geospatial features (covariates). Then, we write $\lambda_{(M)} = \exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(M)}\}$, and $\lambda_k = \alpha_j \exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(N)}\}$ if the k^{th} exposure unit belongs to the j^{th} type of exposure, where $\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)} \in \mathbb{R}^P$ are the vectors of regression coefficients. We fix $\alpha_J = 1$ because of the identifiability issue. Therefore, we have

$$\tilde{M} \sim \text{Poisson}\Big(\lambda_{(M)} := \exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(M)}\}\Big); \quad \tilde{N}_t \sim \text{Poisson}\left(\lambda_{(N)} := \left[\sum_{j=1}^{J} \alpha_j K_j\right] \exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(N)}\}\right). \tag{28}$$

Because of the assumed independence between $\tilde{M}$ and $\tilde{\boldsymbol{N}}$, the complete data joint pmf of $(\tilde{M}, \tilde{\boldsymbol{N}})$ is given by

$$\tilde{f}_{\text{PP}}(\tilde{m}, \tilde{\boldsymbol{n}}; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) := P(\tilde{M} = \tilde{m}, \tilde{\boldsymbol{N}} = \tilde{\boldsymbol{n}}; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) = \frac{{\lambda_{(M)}}^{\tilde{m}} e^{-\lambda_{(M)}}}{\tilde{m}!} \prod_{t=1}^{\tilde{m}} \frac{{\lambda_{(N)}}^{\tilde{n}_t} e^{-\lambda_{(N)}}}{\tilde{n}_t!}, \tag{29}$$

where $\tilde{m}$ and $\tilde{\boldsymbol{n}} := (\tilde{n}_1, \ldots, \tilde{n}_{\tilde{m}})$ are the realizations of $\tilde{M}$ and $\tilde{\boldsymbol{N}}$ respectively, and $\boldsymbol{\Phi}_{\text{PP}} = (\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)}, \boldsymbol{\alpha})$ contains all model parameters with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$.

For the purpose of statistical inference, it is necessary to derive the joint distribution of the observed incomplete data (M, N) based on the PP model for the complete data $(\tilde{M}, \tilde{\boldsymbol{N}})$. We first write the observed data pmf as

$$f_{\text{PP}}(m, n; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) := P(M = m, N = n; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) = P(M = m; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) \times P(N = n | M = m; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}). \tag{30}$$

Note that M marginally follows Poisson distribution with parameter $\lambda^*_{(M)} = \lambda_{(M)} \times P(\tilde{N}_m > 0)$ due to the thinning property. Also, given $M = m$, N follows the same distribution as a sum of m independent zero-truncated Poisson random variables. Applying these observed properties, Equation (30) can be rewritten as

$$f_{\text{PP}}(m, n; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) = \begin{cases} \frac{1}{m!}\left[\lambda_{(M)}(1 - e^{-\lambda_{(N)}})\right]^m e^{-\lambda_{(M)}(1-e^{-\lambda_{(N)}})} g_{\text{ZTP}}(n; m), & n \geqslant m;\ m, n \in \mathbb{N}^+, \\ e^{-\lambda_{(M)}(1-e^{-\lambda_{(N)}})}, & m = n = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{31}$$

where $g_{\text{ZTP}}(n; m)$ is the pmf of a zero-truncated Poisson sum, given by

$$g_{\text{ZTP}}(n; m) := g_{\text{ZTP}}(n; m, \boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}) = \frac{\lambda^n_{(N)}}{n!(e^{\lambda_{(N)}} - 1)^m} \sum_{t=0}^{m} (-1)^t (m - t)^n \binom{m}{t} \tag{32}$$

for $n = m, m + 1, \ldots$. Equation (32) can be proved by induction, see Section 3 of [54] for details.

To better understand the distributional behavior of the observed data (M, N) under the PP model, we study the moment properties of (M, N). The mean $E[\cdot]$, variance

$Var[\cdot]$ and dispersion ratio $\mathcal{D}[\cdot] := Var[\cdot]/E[\cdot]$ of M and N are respectively given by:

$$E[M|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = Var[M|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = \lambda_{(M)}(1 - e^{-\lambda_{(N)}}), \quad \mathcal{D}[M|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = 1, \tag{33}$$

$$E[N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = \lambda_{(M)}\lambda_{(N)}, \quad Var[N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = (1+\lambda_{(N)})\lambda_{(M)}\lambda_{(N)}, \quad \mathcal{D}[N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = (1+\lambda_{(N)}). \tag{34}$$

Also, the covariance and correlation between M and N are:

$$Cov[M, N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = \lambda_{(M)}\lambda_{(N)}, \quad Corr[M, N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}] = \sqrt{\frac{\lambda_{(N)}}{(1 + \lambda_{(N)})(1 - e^{-\lambda_{(N)}})}}. \tag{35}$$

The moment properties above are directly derived using the double expectation and conditional covariance formulas, so the details are omitted. There are two interesting observations from the moment properties. Firstly, the dispersion ratio of M is equal to one, showing that the PP model is unable to cater for the overdispersion of the observed number of storms implied by the data-set described in Section 7.1. On the other hand, since N follows a compound distribution, the dispersion of N is greater than one, allowing for some flexibilities to deal with the overdispersion of the empirical N. Secondly, even if the actual storm and claim arrival processes $\tilde{M}$ and $\tilde{\boldsymbol{N}}$ are assumed to be independent, Equation (35) shows that the observed (M, N) are positively correlated, aligning with the characteristics of the empirical data. The correlation function depends only on $\lambda_{(N)}$. Plotting $Corr[M, N|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{PP}}]$ against $\lambda_{(N)}$, we find a very strong correlation between M and N regardless of $\lambda_{(N)}$. Precisely, the minimum correlation is 0.8776 when $\lambda_{(N)} = 1.7933$, which unfortunately overstates the empirical pearson correlation of 0.5628 from the real dataset. The analysis of moments reveals several shortcomings of the PP model, motivating us to develop more versatile models in the subsequent subsections.

Negative binomial-Poisson model

Since the PP model fails to capture the overdispersion of the number of storms M, we hereby incorporate a negative binomial distribution on $\tilde{M}$ to improve the model flexibility. The corresponding negative binomial-Poisson (NP) model is the same as the PP model with the following extensions. It is assumed that $\tilde{M}|W \sim \text{Poisson}(\lambda_{(M)}W)$ and $W \sim \text{Gamma}(1/\phi, \phi)$, where the Gamma density is given by

$$f_W(w;\phi) = \frac{1}{\Gamma(1/\phi)\phi^{1/\phi}} w^{\phi-1} e^{-w/\phi}. \tag{36}$$

Note that the mean of W is 1 and ϕ measures the dispersion of the W. Then, $\tilde{M}$ is unconditionally a negative binomial distributed random variable with the following pmf:

$$\tilde{f}_{\text{NP}}(\tilde{m};\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}) = \binom{\tilde{m}+1/\phi-1}{\tilde{m}} \left(\frac{\lambda_{(M)}\phi}{1+\lambda_{(M)}\phi}\right)^{\tilde{m}} \left(\frac{1}{1+\lambda_{(M)}\phi}\right)^{1/\phi}, \tag{37}$$

where $\boldsymbol{\Phi}_{\text{NP}} = (\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)}, \boldsymbol{\alpha}, \phi)$ contains all parameters of the NP regression model. Given that $\tilde{M}$ and $\tilde{\boldsymbol{N}}$ are independent, the complete data joint pmf of $(\tilde{M}, \tilde{\boldsymbol{N}})$ can be written as

$$\tilde{f}_{\text{NP}}(\tilde{m},\tilde{\boldsymbol{n}};\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}) = \binom{\tilde{m}+1/\phi-1}{\tilde{m}} \left(\frac{\lambda_{(M)}\phi}{1+\lambda_{(M)}\phi}\right)^{\tilde{m}} \left(\frac{1}{1+\lambda_{(M)}\phi}\right)^{1/\phi} \prod_{t=1}^{\tilde{m}} \frac{{\lambda_{(N)}}^{\tilde{n}_t} e^{-\lambda_{(N)}}}{\tilde{n}_t!}. \tag{38}$$

Applying the same arguments as Equation (30), the observed data joint pmf is given by

$$f_{\text{NP}}(m,n;\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}) = \begin{cases} \binom{m+1/\phi-1}{m} \left(\frac{\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi}{1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi}\right)^{m} \left(\frac{1}{1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi}\right)^{1/\phi} g_{\text{ZTP}}(n;m), & n \geqslant m;\ m,n \in \mathbb{N}^+, \\ \left(\frac{1}{1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi}\right)^{1/\phi}, & m = n = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{39}$$

where $g_{\text{ZTP}}(n;m)$ is given by Equation (32). Various moments of M and N are given by

$$E[M|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=\lambda_{(M)}(1-e^{-\lambda_{(N)}}),\quad Var[M|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=\lambda_{(M)}(1-e^{-\lambda_{(N)}})(1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi), \tag{40}$$

$$E[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=\lambda_{(M)}\lambda_{(N)},\quad Var[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=(1+\lambda_{(N)}+\lambda_{(M)}\lambda_{(N)}\phi)\lambda_{(M)}\lambda_{(N)}, \tag{41}$$

$$\mathcal{D}[M|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi,\quad \mathcal{D}[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=1+\lambda_{(N)}+\lambda_{(M)}\lambda_{(N)}\phi, \tag{42}$$

and the correlation between M and N can be expressed as

$$Corr[M,N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NP}}]=\sqrt{\frac{\lambda_{(N)}(1+\lambda_{(M)}(1-e^{-\lambda_{(N)}})\phi)}{(1+\lambda_{(N)}+\lambda_{(M)}\lambda_{(N)}\phi)(1-e^{-\lambda_{(N)}})}}\geqslant Corr[M,N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{PP}}], \tag{43}$$

where $Corr[M,N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{PP}}]$ is the correlation in Equation (35) under the PP model. The proof of the inequality in Equation (43) is leveraged to Section **??** of the supplementary materials. From the moment properties above, we observe that the dispersion ratio of M is greater than one and is governed by ϕ. This mitigates the inflexibility of the PP model in capturing the overdispersed observed storm count in the dataset. Also, the dispersion of N under the NP model (Equation (42)) is greater than that under the PP model (Equation (34)), making the NP model more suitable in modelling the empirical observed claim frequencies that are excessively overdispersed. However, as compared to the PP model, the NP model does not reduce the correlation between M and N. Given that the PP model already overestimates the empirical pearson correlation, the NP model does not address the problem of overstated correlation.

Negative binomial - Negative binomial model

Motivated by the excessive overdispersion of the claim frequencies N observed from the dataset, we extend the NP model by allowing for a further increased dispersion ratio for N. Building upon the NP model, the so-called negative binomial-negative binomial (NN) model assumes that $\tilde{N}_t|V\sim\text{Poisson}(\lambda_{(N)}V)$ and $V\sim\text{Gamma}(1/\psi,\psi)$, where ψ is

the dispersion parameter. The complete data pmf is given by

$$\tilde{f}_{\text{NN}}(\tilde{m}, \tilde{\boldsymbol{n}}; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}) = \binom{\tilde{m} + 1/\phi - 1}{\tilde{m}} \left(\frac{\lambda_{(M)}\phi}{1+\lambda_{(M)}\phi}\right)^{\tilde{m}} \left(\frac{1}{1+\lambda_{(M)}\phi}\right)^{1/\phi} \times \prod_{t=1}^{\tilde{m}} \binom{\tilde{n}_t + 1/\psi - 1}{\tilde{n}_t} \left(\frac{\lambda_{(N)}\psi}{1+\lambda_{(N)}\psi}\right)^{\tilde{n}_t} \left(\frac{1}{1+\lambda_{(N)}\psi}\right)^{1/\psi}, \quad (44)$$

where the parameter set is $\boldsymbol{\Phi}_{\text{NN}} = (\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)}, \boldsymbol{\alpha}, \phi, \psi)$, and the observed data pmf is given by

$$f_{\text{NN}}(m, n; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}) = \begin{cases} h_{\text{NB}}(m; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}) g_{\text{ZTN}}(n; m, \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}), & n \geqslant m;\ m, n \in \mathbb{N}^+, \\ h_{\text{NB}}(0; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}), & m = n = 0, \\ 0, & \text{otherwise}, \end{cases} \quad (45)$$

where $h_{\text{NB}}(m; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}})$ is a negative binomial pmf, given by

$$h_{\text{NB}}(m; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}) = \binom{m + 1/\phi - 1}{m} \left(\frac{\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})\phi}{1+\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})\phi}\right)^{m} \left(\frac{1}{1+\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/} \quad (46)$$

for $m \in \mathbb{N}$, and $g_{\text{ZTN}}(n; m, \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}})$ is the pmf of a zero-truncated negative binomial sum, given by

$$g_{\text{ZTN}}(n; m, \boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}) = \left(\frac{(1+\lambda_{(N)}\psi)^{-1/\psi}}{1-(1+\lambda_{(N)}\psi)^{-1/\psi}}\right)^{m} \left(\frac{\lambda_{(N)}\psi}{1+\lambda_{(N)}\psi}\right)^{n} \sum_{t=1}^{m} (-1)^{m-t} \binom{m}{t} \binom{n + t/\psi - 1}{n}. \quad (47)$$

The derivation of Equation (47) is leveraged to Section **??** of the supplementary materials. Several moments of (M, N) are presented as follows:

$$E[M|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}] = \lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi}), \quad (48)$$

$$Var[M|\boldsymbol{x}, \boldsymbol{\Phi}_{\text{NN}}] = \lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})(1+\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})\phi), \quad (49)$$

$$E[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NN}}]=\lambda_{(M)}\lambda_{(N)},\quad Var[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NN}}]=(1+\lambda_{(N)}(1+\lambda_{(M)}\phi+\psi))\lambda_{(M)}\lambda_{(N)}, \tag{50}$$

$$\mathcal{D}[M|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NN}}]=1+\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})\phi,\quad \mathcal{D}[N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NN}}]=1+\lambda_{(N)}(1+\lambda_{(M)}\phi+\psi), \tag{51}$$

$$Corr[M,N|\boldsymbol{x},\boldsymbol{\Phi}_{\text{NN}}]=\sqrt{\frac{\lambda_{(N)}(1+\lambda_{(M)}(1-(1+\lambda_{(N)}\psi)^{-1/\psi})\phi)}{(1+\lambda_{(N)}(1+\lambda_{(M)}\phi+\psi))(1-(1+\lambda_{(N)}\psi)^{-1/\psi})}}. \tag{52}$$

From Equation (51), we observe that the excessive dispersion of N can be modelled by increasing ψ. Also, the correlation between M and N in Equation (52) may reduce by increasing ψ, and hence the NN model provides a greater flexibility in capturing a broader range of dependence between M and N.

Finite mixture models

The aforementioned models assume independence between the actual number of storms $\tilde{M}$ and the claim counts $\tilde{\boldsymbol{N}}$, enabling us to analytically derive the joint distributions of the observed data (M,N) and ensuring that the statistical inference on these models is feasible (see the next subsection for details). However, this assumption obstructs us to uncover the dependence structure between $\tilde{M}$ and $\tilde{\boldsymbol{N}}$, which is one of the main research goals highlighted in Section 7.1.

Dependence of random variables is often modelled through copulas [18, 36, 28]. However, because of the incompleteness of the observed data (M,N), it is difficult to write the joint distribution of (M,N) as a closed-form expression if $\tilde{M}$ and $\tilde{\boldsymbol{N}}$ are linked by a copula. Moreover, applications of the copulas to discrete variables inherit several issues, including a lack of model flexibility, an increased computational burden and a loss of model identifiability (see, e.g., [24]).

To this end, we propose a finite mixture approach to flexibly model the joint distribution of $(\tilde{M},\tilde{\boldsymbol{N}})$. Under this approach, we assume that there are B latent subgroups, representing B distinguishable environmental statuses (or risk profiles) governing the

occurrence of storms and claims. Each observation (the aggregated storms and claims grouped by policy year and postcode) belongs to one of the B subgroups. Define $\boldsymbol{Z} := (Z_1, \ldots, Z_B)$ as a vector for subgroup classification, where an observation belongs to the b-th subgroup if $Z_b = 1$, and $Z_b = 0$ otherwise. We assume that $\boldsymbol{Z}$ follows a categorical distribution with $P(Z_b = 1) = \pi_b(\boldsymbol{x}; \boldsymbol{\beta}_{(\pi)})$ for $b = 1, \ldots, B$, such that $\sum_{b=1}^{B} \pi_b(\boldsymbol{x}; \boldsymbol{\beta}_{(\pi)}) = 1$. Here, $\boldsymbol{\beta}_{(\pi)} = (\boldsymbol{\beta}_{(\pi)1}, \ldots, \boldsymbol{\beta}_{(\pi)B})$ represents B sets of regression parameters for the mixture probabilities $\{\pi_b(\boldsymbol{x}; \boldsymbol{\beta}_{(\pi)})\}_{b=1,\ldots,B}$. We assume a logit-linear link for the mixture probability, given by

$$\pi_b(\boldsymbol{x}; \boldsymbol{\beta}_{(\pi)}) = \frac{\exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(\pi)b}\}}{\sum_{b'=1}^{B} \exp\{\boldsymbol{x}^T \boldsymbol{\beta}_{(\pi)b'}\}}, \tag{53}$$

where we set $\boldsymbol{\beta}_{(\pi)B} = \mathbf{0}$ because of the identifiability issue. The choice of the logit-linear link is motivated by the theoretical results of [22] that the resulting finite mixture model is able to accurately approximate a broad range of regression links between $\boldsymbol{x}$ and $(\tilde{M}, \tilde{\boldsymbol{N}})$, including non-linear or even interactive covariates effects. Conditioned on the latent subgroup $b \in \{1, \ldots, B\}$, we assume that $(\tilde{M}, \tilde{\boldsymbol{N}})$ follows PP, NP or NN models introduced in Sections 7.3.1 to 7.3.3. The complete data pmf is

$$\tilde{f}_{\text{mix},k}(\tilde{m}, \tilde{\boldsymbol{n}}; \boldsymbol{x}, \boldsymbol{\Phi}_{\text{mix},k}) = \sum_{b=1}^{B} \pi_b(\boldsymbol{x}; \boldsymbol{\beta}_{(\pi)}) \times \tilde{f}_k(\tilde{m}, \tilde{\boldsymbol{n}}; \boldsymbol{x}, \boldsymbol{\Phi}_{k,b}), \tag{54}$$

where $k \in \{\text{PP}, \text{NP}, \text{NN}\}$. For example, if $k = \text{PP}$, then $\tilde{f}_k(\cdot)$ corresponds to the PP model in Equation (29), and $\boldsymbol{\Phi}_b := (\boldsymbol{\beta}_{(M)b}, \boldsymbol{\beta}_{(N)b}, \boldsymbol{\alpha})$ are denoted as the PP model parameters corresponding the b^{th} mixture component. Also, $\boldsymbol{\Phi}_{\text{mix},k} = (\boldsymbol{\beta}_{(\pi)}, \boldsymbol{\Phi}_{k,1}, \ldots, \boldsymbol{\Phi}_{k,B})$ contains all mixture model parameters. We assume a universal exposure coefficients $\boldsymbol{\alpha}$ across all B mixture components. For $k = \text{NP}$ or $k = \text{NN}$, we further fix the dispersion parameters (ϕ and/ or ψ) uniform across mixture components. With this regards, the observed data

pmf becomes

$$f_{\text{mix},k}(m,n;\boldsymbol{x},\boldsymbol{\Phi}_{\text{mix},k}) = \sum_{b=1}^{B} P(Z_b=1;\boldsymbol{x},\boldsymbol{\Phi}_{k,b}) \times P(M=m,N=n|Z_b=1,\boldsymbol{x},\boldsymbol{\Phi}_{k,b})$$
$$= \sum_{b=1}^{B} \pi_b(\boldsymbol{x};\boldsymbol{\beta}_{(\pi)}) \times f_k(m,n;\boldsymbol{x},\boldsymbol{\Phi}_{k,b}), \tag{55}$$

which is still a finite mixture pmf with analytically tractable component functions. While $\tilde{M}$ and $\tilde{\boldsymbol{N}}$ are independent conditional on the latent status $\boldsymbol{Z}$, they are unconditionally dependent. In particular, Remark 2.5 of [3] shows that a finite mixture of multivariate (independent) Poisson distributions covers a wide range of dependence structures, and the Pearson's correlation coefficient of multivariate Poisson mixtures can span from -1 to 1. This justifies the flexibility of the proposed model (Equation (53)) in terms of dependence modelling. Also, equipped by multiple distributional components, finite mixture models are more appropriate in fitting complicated distributional characteristics, such as heavy-tailedness and multimodality.

Inference

The parameters of the models introduced above are estimated using a maximum likelihood estimation approach. Suppose that we have I independent observations in the dataset, where each observation corresponds to a specific policy year and postcode. We denote the observed data as $\mathcal{D} := \{(m_i, n_i, \boldsymbol{x}_i)\}_{i=1,\dots,I}$. Here, i is the observation index, while m_i, n_i and $\boldsymbol{x}_i$ represent (the realizations of) the observed number of storms, observed aggregated claim count and geospatial features of observation i respectively. The observed data log-likelihood function is

$$\mathcal{L}(\boldsymbol{\Phi}_k;\mathcal{D}) = \sum_{i=1}^{I} \log f_k(m_i,n_i;\boldsymbol{x}_i,\boldsymbol{\Phi}_k), \tag{56}$$

where $k \in \{\text{PP}, \text{NP}, \text{NN}, \text{mixPP}, \text{mixNP}, \text{mixNN}\}$ represents one of the six model classes chosen from Section 7.3. Thanks to the closed-form expression of $f_k(m_i,n_i;\boldsymbol{x}_i,\boldsymbol{\Phi}_k)$ regardless of the model k selected, the log-likelihood function can be directly maximized with respect to $\boldsymbol{\Phi}_k$ to obtain the estimated model parameters.

Analysis of fitting results

Distribution		Mixture	Regression on		Goodness-of-fit					
M	N	components	$\lambda_{(M)}$, $\lambda_{(N)}$	π_b	log-likelihood	AIC	BIC	χ^2 for (M,N)	χ^2 for M	χ^2 for N
$\mathcal{P}$	$\mathcal{P}$	1			-21999.63	44007.25	44035.80	536447.01	39393.01	342.18
$\mathcal{NB}$	$\mathcal{P}$	1			-21584.37	43178.75	43214.44	954.50	37.59	137.99
$\mathcal{NB}$	$\mathcal{NB}$	1			-13348.19	26708.38	26751.21	1462.58	48.47	131.37
$\mathcal{P}$	$\mathcal{P}$	2			-14563.89	29141.79	29191.75	37442.47	2503.44	142.86
$\mathcal{NB}$	$\mathcal{P}$	2			-14306.73	28629.45	28686.56	302.86	24.40	22.29
$\mathcal{NB}$	$\mathcal{NB}$	2			-12097.37	24212.75	24276.99	104.65	10.82	15.28
$\mathcal{P}$	$\mathcal{P}$	4			-12622.89	25271.79	25364.58	1953.35	175.79	61.25
$\mathcal{NB}$	$\mathcal{P}$	4			-12534.52	25097.04	25196.97	125.81	16.75	20.47
$\mathcal{NB}$	$\mathcal{NB}$	4			-11920.65	23871.30	23978.37	108.27	5.87	20.98
$\mathcal{P}$	$\mathcal{P}$	1	✓		-21352.23	42740.45	42868.93	106839.62	7035.66	275.74
$\mathcal{NB}$	$\mathcal{P}$	1	✓		-21004.53	42047.05	42182.67	989.88	23.02	154.52
$\mathcal{P}$	$\mathcal{P}$	2	✓		-13996.88	28063.77	28313.60	3103.25	583.70	92.01
$\mathcal{NB}$	$\mathcal{P}$	2	✓		-13796.07	27664.15	27921.11	213.09	10.71	24.87
$\mathcal{P}$	$\mathcal{P}$	4	✓		-12189.25	24516.50	25009.02	319.59	50.37	26.32
$\mathcal{NB}$	$\mathcal{P}$	4	✓		-12129.53	24399.05	24898.71	92.82	7.85	13.10
$\mathcal{P}$	$\mathcal{P}$	2		✓	-14542.43	29112.86	29212.79	37059.83	2318.91	139.37
$\mathcal{NB}$	$\mathcal{P}$	2		✓	-14289.98	28609.96	28717.03	308.42	25.38	21.08
$\mathcal{P}$	$\mathcal{P}$	4		✓	-12539.70	25147.39	25390.08	1622.85	120.22	47.23
$\mathcal{NB}$	$\mathcal{P}$	4		✓	-12487.60	25045.19	25295.02	143.62	14.03	21.01
$\mathcal{P}$	$\mathcal{P}$	2	✓	✓	-13965.10	28014.20	28314.00	3086.30	597.78	83.13
$\mathcal{NB}$	$\mathcal{P}$	2	✓	✓	-13779.19	27644.38	27951.31	229.70	10.80	26.85
$\mathcal{P}$	$\mathcal{P}$	3	✓	✓	-12525.03	25182.06	25653.16	1089.45	193.39	41.09
$\mathcal{NB}$	$\mathcal{P}$	3	✓	✓	-12422.87	24979.75	25457.99	123.96	8.83	10.92
$\mathcal{P}$	$\mathcal{P}$	4	✓	✓	-12080.94	24341.87	24984.29	349.41	62.80	19.20
$\mathcal{NB}$	$\mathcal{P}$	4	✓	✓	-12049.17	24280.34	24929.90	91.02	8.70	12.55
$\mathcal{P}$	$\mathcal{P}$	5	✓	✓	-11955.44	24138.88	24952.61	118.90	18.24	20.60
$\mathcal{NB}$	$\mathcal{P}$	5	✓	✓	-11937.84	24105.67	24926.54	79.19	6.32	15.45
$\mathcal{P}$	$\mathcal{P}$	6	✓	✓	-11882.09	24040.18	25025.23	97.31	17.68	15.76
$\mathcal{NB}$	$\mathcal{P}$	6	✓	✓	-11872.89	24023.77	25015.95	71.46	7.31	13.26

Table 32: Summary of goodness-of-fit for various model specifications

The models described in the previous sections are fitted to the dataset and their goodness-of-fits are showcased in Table 32, which include the log-likelihood, the AIC/BIC, and the χ^2 test statistics to measure the deviations of the fitted distributions from the empirical distribution. We classify the models in terms of used covariates as follows:

- Without covariates: All the regression coefficients, $\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)}$, and $\boldsymbol{\beta}_{(\pi)}$, are set to 0 except for the intercepts.

- With covariates: In the case of single-component compound frequency models ($B = 1$), $\boldsymbol{\beta}_{(\pi)}$ is automatically a zero-vector. In the case of mixture compound frequency models ($B > 1$), we may have three different specifications:

  - Heterogeneous π_b (logit-reduced mixture-of-experts models): $\boldsymbol{\beta}_{(M)}$ and $\boldsymbol{\beta}_{(N)}$ are fixed as 0 except for the intercepts.
  - Heterogeneous $\lambda_{(M)}$ and $\lambda_{(N)}$ (finite mixture of regression models): $\boldsymbol{\beta}_{(\pi)}$ is fixed as 0 except for the intercept.
  - Heterogeneous $\lambda_{(M)}$, $\lambda_{(N)}$, and π_b (mixture-of-experts models): No restriction is given on $\boldsymbol{\beta}_{(M)}, \boldsymbol{\beta}_{(N)}$, and $\boldsymbol{\beta}_{(\pi)}$.

Based on the goodness-of-fits in Table 32, it is observed that Poisson-Poisson model fits the data rather poorly especially the covariates are not used, as reflected by low log-likelihood. The extremely high χ^2 statistics reflect that the Poisson-Poisson model severely under-extrapolates the tails of the empirical distribution, especially for M. However, such tendency is alleviated when the available covariates are used and finite mixture of Poisson-Poisson distributions are considered. NB-NB model significantly improves upon the Poisson-Poisson model in terms of fitting performance (log-likelihood, χ^2 statistics, etc.). However, the fitted parameters fall on the boundary of the parameter space, especially when the covariates are not used. The estimated $\lambda_{(M)}$ and ψ are extremely large (371.53 and 3001.89, respectively), while the fitted $\lambda_{(N)}$ is very small (0.00693). This makes the NB-NB model hardly interpretable even though the fitting performance is superior, e.g., it does not make sense that 371 storms occur during a year while a claim occurs per $144 \simeq 1/0.00693$ storms on average. In that regard, NB-NB model was not considered afterwards with mixture components and/or covariates to avoid possible overfitting issues. Lastly, NB-Poisson model improves the fit significantly in terms of both log-likelihood and χ^2 statistics as long as mixture components and/or covariates are considered. Indeed, mixture models significantly improve the fitting performances in all cases and the NB-Poisson mixture model with 5-components and covariates is the most reliable model.

That being said, we consider Poisson-Poisson (hereafter 5PP) and negative binomial-Poisson (hereafter 5NP) models with 5-components and the available covariates as our proposed models to fit given dataset. The fitted parameters of the chosen model specifications are presented in Table 33. Note that 5PP model is a special case of 5NP model so that the resulting parameter of 5PP model was used as an initial value of 5NP model, which explains similarity of the resulting coefficients between the two models. From Table 33, one can observe that among the three vehicle types (motorcycles, cars and trucks), trucks are most prone to a storm event, followed by motorcycles. Under the 5PP and 5NP models, whenever a storm occurs, the average numbers of claims filed by a motorcycle and a car are respectively around 47% and 26% as the number of claims filed by a truck.

Overall, there are multiple sets of regression parameters for the mixture probabilities and the mixture components, making it hard to interpret $\boldsymbol{\beta}_{(M:\cdot)}$, $\boldsymbol{\beta}_{(N:\cdot)}$ and $\boldsymbol{\beta}_{(\pi:\cdot)}$ directly. For example, in case of $\boldsymbol{\beta}_{(M:\cdot)}$ and $\boldsymbol{\beta}_{(N:\cdot)}$, there is no estimated regression coefficient that show the same sign along with the components, which makes understanding the impacts of available covariates difficult to $\tilde{M}$ and $\tilde{N}$ with the selected models. Therefore, it is instructive to construct certain forms of partial dependence plots (discussed later in this section) to visualize the individual impacts of each geospatial variable to the storm and claim frequencies.

	5PP					5NP				
	$b=1$	$b=2$	$b=3$	$b=4$	$b=5$	$b=1$	$b=2$	$b=3$	$b=4$	$b=5$
$\beta_{(M:Intercept)}$	-2.3310	-1.8640	1.1002	-4.6915	3.0756	-2.3323	-1.9123	0.8849	-4.5081	3.2509
	(1.4864)	(1.2207)	(0.7841)	(1.4652)	(1.4742)	(1.5774)	(1.3247)	(0.9261)	(1.3967)	(1.4299)
$\beta_{(M:vulnerability)}$	-0.1367	0.4019	0.2064	0.3667	0.0493	-0.1253	0.4220	0.2002	0.3553	-0.0198
	(0.2792)	(0.2262)	(0.1329)	(0.2042)	(0.3098)	(0.3081)	(0.2497)	(0.1520)	(0.1942)	(0.3181)
$\beta_{(M:coast_ratio)}$	1.2260	-0.7875	0.7265	0.7482	-0.1186	1.2002	-0.7630	0.6870	0.6886	-0.1528
	(0.5561)	(0.3977)	(0.2670)	(0.3541)	(0.5118)	(0.6009)	(0.4340)	(0.3065)	(0.3570)	(0.5179)
$\beta_{(M:log_alt_mean)}$	0.3885	0.3197	-0.0822	0.7924	0.0897	0.3734	0.3086	-0.0597	0.7553	0.1115
	(0.1641)	(0.1394)	(0.0961)	(0.1815)	(0.1100)	(0.1727)	(0.1514)	(0.1141)	(0.1716)	(0.1085)
$\beta_{(M:residential)}$	2.0381	-0.1489	0.1471	0.9518	-2.0432	2.0074	-0.1666	0.1741	0.8640	-2.0381
	(0.6571)	(0.4128)	(0.3292)	(0.3713)	(1.4069)	(0.7042)	(0.4658)	(0.3660)	(0.3760)	(1.3025)
$\beta_{(M:industrial)}$	1.4436	0.1786	-0.6290	2.4741	-2.0795	1.4500	0.1467	-0.4668	2.5433	-2.0920
	(0.8363)	(0.6371)	(0.4335)	(0.6002)	(1.3883)	(0.8753)	(0.7134)	(0.4982)	(0.6669)	(1.3014)
$\beta_{(M:agricultural)}$	0.8297	-0.1734	-0.3300	-1.2487	-3.1960	0.8002	-0.2256	-0.4640	-1.0739	-3.2413
	(0.6953)	(0.4271)	(0.4775)	(0.4611)	(1.4243)	(0.7373)	(0.4735)	(0.5100)	(0.4736)	(1.3246)
$\beta_{(M:wetlands)}$	1.0518	-0.3386	-8.0167	2.2268	-1.9245	1.0816	-0.2340	-7.9331	2.3298	-1.8873
	(1.5870)	(0.9694)	(3.5818)	(1.0672)	(11.6725)	(1.6015)	(1.1200)	(4.6650)	(1.1120)	(11.2087)
ϕ			-					0.1450 (0.0297)		
$\beta_{(N:Intercept)}$	-1.0039	-4.2227	-7.1314	-4.9872	-8.5758	-0.9865	-4.2091	-7.1554	-4.9793	-8.4462
	(0.3029)	(0.4115)	(0.5248)	(1.5918)	(1.6464)	(0.3032)	(0.4108)	(0.5485)	(1.7274)	(1.5641)
$\beta_{(N:vulnerability)}$	-0.3609	-0.4416	-0.2894	0.2171	-0.1460	-0.3609	-0.4416	-0.2894	0.2171	-0.1460
	(0.0604)	(0.0663)	(0.0706)	(0.2057)	(0.3441)	(0.0603)	(0.0670)	(0.0712)	(0.2247)	(0.3504)
$\beta_{(N:coast_ratio)}$	1.0763	0.2014	1.4691	-0.7058	0.0307	1.0763	0.2014	1.4691	-0.7058	0.0307
	(0.1136)	(0.1517)	(0.1804)	(0.4202)	(0.5980)	(0.1127)	(0.1576)	(0.1920)	(0.4301)	(0.5865)
$\beta_{(N:log_alt_mean)}$	-0.0926	0.2912	0.5237	-0.2783	-0.2414	-0.0926	0.2912	0.5237	-0.2783	-0.2414
	(0.0382)	(0.0561)	(0.0708)	(0.1883)	(0.1189)	(0.0383)	(0.0562)	(0.0740)	(0.2009)	(0.1154)
$\beta_{(N:residential)}$	0.2762	0.4508	0.2432	-0.3611	2.5021	0.2762	0.4508	0.2432	-0.3611	2.5021
	(0.1512)	(0.1367)	(0.1888)	(0.4207)	(1.4972)	(0.1505)	(0.1382)	(0.1924)	(0.4272)	(1.3673)
$\beta_{(N:industrial)}$	-0.8558	2.4727	1.8691	0.6074	3.3378	-0.8558	2.4727	1.8691	0.6074	3.3378
	(0.2076)	(0.1604)	(0.2938)	(0.5691)	(1.5306)	(0.2071)	(0.1616)	(0.3060)	(0.5740)	(1.4127)
$\beta_{(N:agricultural)}$	-0.4810	-0.9183	-1.5388	-0.5798	3.3704	-0.4810	-0.9183	-1.5388	-0.5798	3.3704
	(0.1555)	(0.1553)	(0.2464)	(0.5006)	(1.5973)	(0.1550)	(0.1555)	(0.2501)	(0.4981)	(1.4498)
$\beta_{(N:wetlands)}$	-8.6028	1.0765	-0.8119	1.7023	-3.0704	-8.6028	1.0765	-0.8119	1.7023	-3.0704
	(0.5367)	(0.3590)	(3.5260)	(1.0364)	(13.4240)	(0.5672)	(0.3608)	(3.9992)	(1.0254)	(12.8805)
α_1^*			0.4740 (0.1107)					0.4672 (0.1099)		
α_2^*			0.2658 (0.0323)					0.2584 (0.0310)		
$\beta_{(\pi:Intercept)}$	-0.8088	0.1071	3.2288	4.3943	-	-0.7586	0.1748	3.2530	4.2305	-
	(1.8549)	(1.6965)	(2.1161)	(2.5849)	-	(1.9117)	(1.7680)	(2.2759)	(2.7450)	-
$\beta_{(\pi:vulnerability)}$	0.0208	-0.0734	-1.1743	-1.4369	-	-0.0137	-0.0874	-1.1402	-1.4341	-
	(0.3850)	(0.3439)	(0.3874)	(0.4401)	-	(0.4137)	(0.3694)	(0.4273)	(0.4753)	-
$\beta_{(\pi:coast_ratio)}$	-1.2184	0.5685	-1.6711	0.4392	-	-1.1551	0.4222	-1.4914	0.5268	-
	(0.7529)	(0.6648)	(0.8094)	(0.7617)	-	(0.7956)	(0.7104)	(0.8857)	(0.8178)	-
$\beta_{(\pi:log_alt_mean)}$	-0.1262	-0.2737	-0.1903	0.0026	-	-0.0932	-0.2518	-0.1658	0.0533	-
	(0.2062)	(0.1969)	(0.2521)	(0.3373)	-	(0.2082)	(0.2027)	(0.2682)	(0.3471)	-
$\beta_{(\pi:residential)}$	-2.3105	-0.9263	-0.3349	-0.5398	-	-2.3548	-0.9963	-0.5638	-0.7098	-
	(0.9629)	(0.7939)	(0.8617)	(0.8912)	-	(0.9894)	(0.8348)	(0.9038)	(0.9384)	-
$\beta_{(\pi:industrial)}$	-0.9028	-4.3433	-0.5436	-4.0708	-	-0.9253	-4.3773	-0.6784	-4.0652	-
	(1.0978)	(1.2579)	(0.9825)	(1.4204)	-	(1.1132)	(1.2794)	(1.0222)	(1.5153)	-
$\beta_{(\pi:agricultural)}$	-1.9539	-0.4964	-1.1327	-1.0644	-	-1.9272	-0.4117	-1.0930	-1.1221	-
	(0.9537)	(0.8125)	(1.0986)	(0.8885)	-	(0.9774)	(0.8462)	(1.1261)	(0.9281)	-
$\beta_{(\pi:wetlands)}$	11.6636	6.7391	9.2854	3.6652	-	11.6978	6.7832	9.1640	3.7509	-
	(4.7365)	(3.4752)	(4.0600)	(3.6196)	-	(5.1693)	(3.6321)	(3.9485)	(3.6785)	-

Table 33: Estimated parameters (and standard errors) for the selected models.

As mentioned in the interpretation of Table 33, there are many regression coefficients in the selected models, 5PP and 5NP so that it might be hard to find intuitive interpretation of the impacts of covariates to the actual number of storms ($\tilde{M}$), the number of claims per each storm ($\tilde{N}$) per each exposure unit, and total number of storm-triggered claims (N) per each exposure unit. In that regard, we produce partial dependence plots, which visualize the marginal impact of each covariate to the response variables. By plotting the following functions against x, the partial dependence plots for the p-th covariate ($p = 1, \ldots, P$) $\mathbf{x}_{\cdot p}$ are obtained:

$$\hat{\mathbb{E}}[\tilde{M}|\mathbf{x}_{\cdot p} = x] = \frac{1}{I}\sum_{i=1}^{I}\sum_{b=1}^{B}\pi_b\left(\boldsymbol{x}_i^{(p)};\hat{\boldsymbol{\beta}}_{(\pi)}\right)\exp\left\{\boldsymbol{x}_i^{(p)T}\hat{\boldsymbol{\beta}}_{(M)b}\right\}, \tag{57}$$

$$\hat{\mathbb{E}}[\tilde{N}|\mathbf{x}_{\cdot p} = x] = \frac{1}{I}\sum_{i=1}^{I}\sum_{b=1}^{B}\pi_b\left(\boldsymbol{x}_i^{(p)};\hat{\boldsymbol{\beta}}_{(\pi)}\right)\exp\left\{\boldsymbol{x}_i^{(p)T}\hat{\boldsymbol{\beta}}_{(N)b}\right\}, \tag{58}$$

$$\hat{\mathbb{E}}[N|\mathbf{x}_{\cdot p} = x] = \frac{1}{I}\sum_{i=1}^{I}\sum_{b=1}^{B}\pi_b\left(\boldsymbol{x}_i^{(p)};\hat{\boldsymbol{\beta}}_{(\pi)}\right)\exp\left\{\boldsymbol{x}_i^{(p)T}\left(\hat{\boldsymbol{\beta}}_{(M)b}+\hat{\boldsymbol{\beta}}_{(N)b}\right)\right\}, \tag{59}$$

where $\hat{\boldsymbol{\beta}}_{(\pi)}$, $\hat{\boldsymbol{\beta}}_{(M)b}$ and $\hat{\boldsymbol{\beta}}_{(N)b}$ are the estimated parameters, and the adjusted covariates $\boldsymbol{x}_i^{(p)}$ are identical to the observed covariates $\boldsymbol{x}_i$ except that the p-th element of $\boldsymbol{x}_i^{(p)}$ is replaced by x. Figure 77 displays partial dependence plots of 5PP and 5NP models along with the available covariates, which provides the following observations:

- While the linear predictors are one of the main building blocks of the selected models, use of the mixture models allows enough flexibility so that marginal impacts of each covariates to $\tilde{M}$, $\tilde{N}$, and N might be neither linear nor monotonic, as shown in the partial dependence plots for *vulnerability*, *residential*, and *wetlands*.

- Some covariates affect $\tilde{M}$ and $\tilde{N}$ in the same direction. For example, both $\tilde{M}$ and $\tilde{N}$ are increasing (decreasing) functions of *industrial* (*agricultural*), which implies

higher proportion of industrial (agricultural) land use leads to more (less) storms and more (less) motor claims per each storm at the same time. On the other hand, most of the other covariates affect $\tilde{M}$ and $\tilde{N}$ in an opposite direction. In the case of *coast_ratio* and *log_alt_mean*, their positive impacts to $\tilde{M}$ outweigh their negative impacts to $\tilde{N}$ so that the total claim count due to the storms (i.e., N) is expected to increase as these covariates increase.

- Marginal impacts of *residential* to the response variables are quite interesting. Here, $\tilde{M}$ and $\tilde{N}$ are decreasing functions of *residential* as this variable falls between 0 to 0.4, while N is an increasing function of *residential*. While it might not be intuitive, such a situation may happen if there exists one mixture cluster associated with frequent storms but few claims per storm and another cluster with few storms but many claims per storm. This motivates us to further investigate a potential negative dependence between $\tilde{M}$ and $\tilde{N}$. Quantitative explanations on this situation are demonstrated in Section **??** of the supplementary material with a hypothetical numerical example.

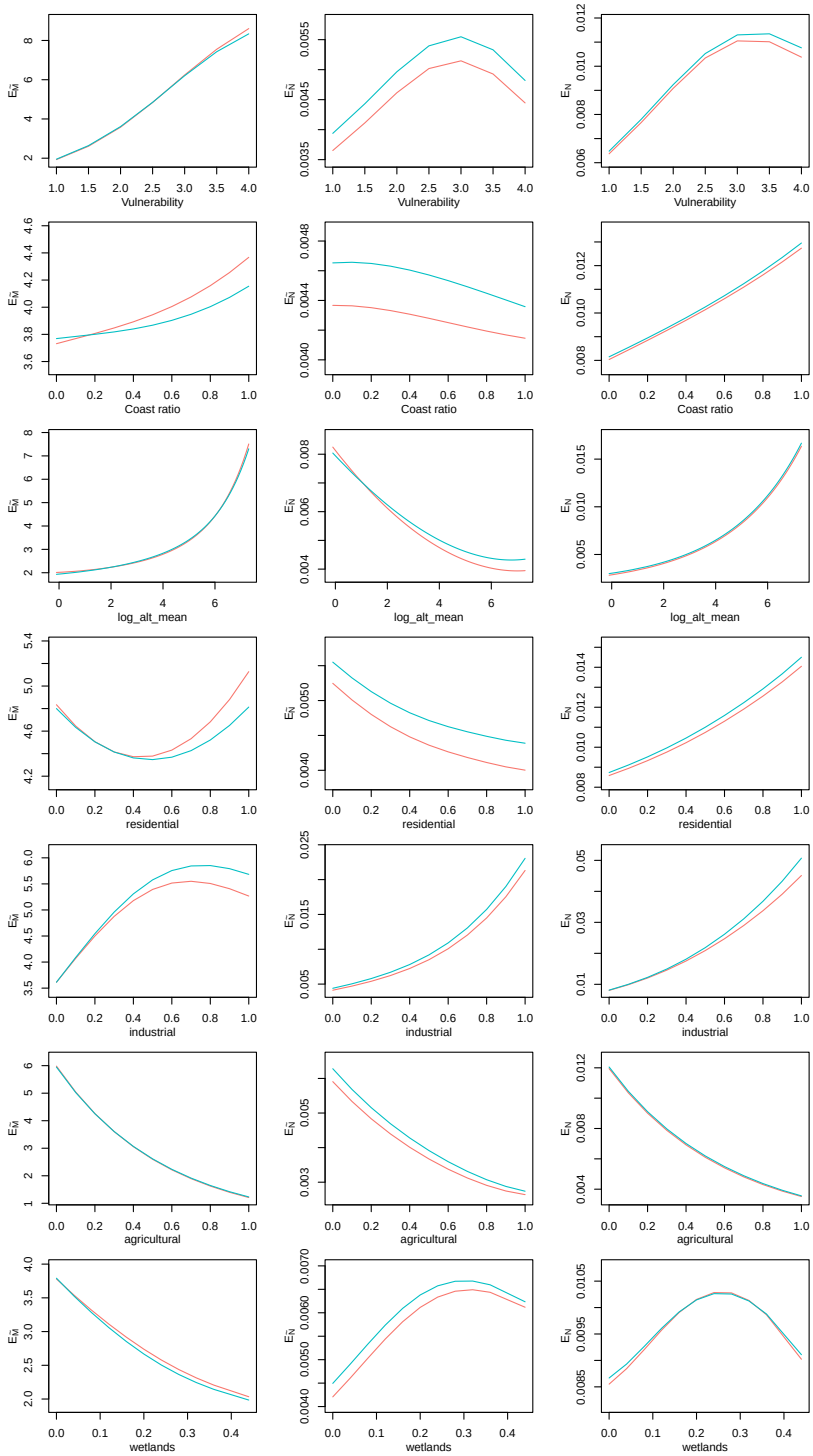


Figure 77. Partial dependence plots for the selected models (Green: 5PP, Red: 5NP). Left panel: $\hat{\mathbb{E}}[\tilde{M}|\mathbf{x}_{.p} = x]$; Middle panel: $\hat{\mathbb{E}}[\tilde{N}|\mathbf{x}_{.p} = x]$; Right panel: $\hat{\mathbb{E}}[N|\mathbf{x}_{.p} = x]$.

Tables 34 and 35 present the estimated frequency tables of (M, N) generated by the selected models, 5PP and 5NP. The table can also assess the goodness-of-fits of the selected models compared to the empirical data. While the parameter estimates from the 5PP and 5NP models are not dramatically different, one can observe that the estimated frequencies of (M, N)s under the 5NP model are much closer to the empirical frequencies in Table 30.

	N=0	1	2	3	4	5	6	7	8	9	>9
M=0	6481.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	1295.66	192.89	54.89	24.35	13.99	9.37	6.89	5.36	4.33	38.73
2	0.00	0.00	334.92	124.23	53.84	28.44	17.27	11.61	8.42	6.46	60.37
3	0.00	0.00	0.00	87.65	55.47	33.01	20.79	13.86	9.74	7.20	57.90
4	0.00	0.00	0.00	0.00	22.34	20.84	16.06	11.91	8.76	6.51	45.35
5	0.00	0.00	0.00	0.00	0.00	5.60	7.06	6.73	5.79	4.71	31.32
6	0.00	0.00	0.00	0.00	0.00	0.00	1.41	2.24	2.53	2.47	19.35
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.68	0.88	10.57
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.21	5.13
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	2.30
>9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.75

Table 34: Fitted frequency table of (M, N) under 5PP model.

	N=0	1	2	3	4	5	6	7	8	9	>9
M=0	6531.6	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.0	1250.78	192.8	57.15	26.01	15.12	10.18	7.50	5.85	4.72	41.67
2	0.0	0.00	327.6	119.06	51.79	27.67	16.92	11.39	8.25	6.32	58.08
3	0.0	0.00	0.0	91.69	53.72	30.88	19.37	12.95	9.12	6.74	53.16
4	0.0	0.00	0.0	0.00	26.11	21.31	15.28	11.07	8.12	6.05	41.75
5	0.0	0.00	0.0	0.00	0.00	7.54	7.92	6.81	5.62	4.53	30.08
6	0.0	0.00	0.0	0.00	0.00	0.00	2.23	2.85	2.85	2.63	20.15
7	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.68	1.02	1.15	12.43
8	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.22	0.37	7.08
9	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.07	3.81
>9	0.0	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.18

Table 35: Fitted frequency table of (M, N) under 5NP model.

To evaluate the performance of the proposed models in terms of dependence modelling, Table 36 presents three dependence measures, Pearson correlation, Spearman's

rho and Kendall's tau, between the number of storms reported M and the total number of climate-induced claims N under several fitted models. The figures are calculated by bootstrapping a very large number of covariates from the empirical dataset, simulating M and N based on the bootstrapped covariate sample and the fitted model, and computing the correlation measures from the simulated M and N. The bottom row of the table shows the correlation measures of the empirical data for comparison purpose. We find that both 5PP and 5NP models underestimate the Pearson correlation from the empirical data (0.3803 or 0.3834 from the fitted models versus 0.5628 from the data). However, such discrepancies are not unexpected because Pearson correlation is often distorted by outliers, which are prevalent in heavy-tailed empirical observed frequencies. On the other hand, Spearman's rho and Kendall's tau are more robust metrics because these rank-based measures are less sensitive to outliers. We observe that the Spearman's rho and Kendall's tau for the 5PP and 5NP fitted models are very close to those from the empirical data-set. As a benchmark comparison, the Spearman's rho and Kendall's tau for the PP and NP fitted models do not match well with the empirical values. Therefore, finite mixture models are apparently more preferred as they offer enhanced versatility to accurately resemble the dependence structure of the data.

Model	Pearson correlation	Spearman's rho	Kendall's tau
5PP	0.3803	0.9927	0.9648
5NP	0.3834	0.9927	0.9642
PP	0.7893	0.9835	0.9378
NP	0.8124	0.9896	0.9509
Empirical	0.5628	0.9922	0.9627

Table 36: Several correlation measures between the observed storm and claim frequencies (M, N) under various fitted models and the empirical data.

To unravel the intrinsic dependence structure between the unobserved actual number of storms occurred $\tilde{M}$ and the claim frequency attributed to a storm $\tilde{N}_t$, Table 37 showcases the three dependence measures between $\tilde{M}$ and $\tilde{N}_t$ under various fitted models. The correlations are evaluated in two ways: "with covariates' influence" and "without covariates' influence". The correlations with covariates' influence are computed

by bootstrapping and simulating many pairs of $(\tilde{M}, \tilde{N}_t)$ from the fitted model, similar to how the figures in Table 36 are produced. However, the correlations computed using this method absorb the impacts of covariates, e.g., the variable *agricultural* has negative impacts to both $\tilde{M}$ and $\tilde{N}_t$, contributing positively to the correlation. To understand the distributional dependence between $\tilde{M}$ and $\tilde{N}_t$ removing the effect of regressions, we also evaluate the marginal distribution functions of $\tilde{M}$ and $\tilde{N}_t$ using the simulated data and the fitted models and compute the correlations of the marginal distributions, resulting in the dependence measures "without covariates' influence". It is obvious that both PP and NP models produce zero correlations because of the assumed independence between $\tilde{M}$ and $\tilde{N}_t$ under these benchmark models. On the other hand, the fitted 5PP and 5NP mixture models reveal a negative relationship between $\tilde{M}$ and $\tilde{N}_t$ under all three dependence measures and two evaluation methods. Hence, there is an overall diversification benefit of fewer per-storm claims for a climate change triggering more storms. The diversification benefit cannot be fully explained by the geospatial variables.

Evaluation method	Model	Pearson correlation	Spearman's rho	Kendall's tau
With covariates' influence	5PP	-0.0621	-0.1133	-0.0948
	5NP	-0.0568	-0.1068	-0.0894
	PP/ NP	0.0000	0.0000	0.0000
Without covariates' influence	5PP	-0.1272	-0.1271	-0.0850
	5NP	-0.1230	-0.1231	-0.0823
	PP/ NP	0.0000	0.0000	0.0000

Table 37: Several correlation measures between the true storm frequency and per-storm claim count $(\tilde{M}, \tilde{N}_t)$ under various fitted models and evaluation methods.

Concluding Remarks

In the above research process there has been an analytical risk assessment regarding weather related phenomena and their impact on Greek insurance companies. Severe weather events were investigated and an analysis was made of the consequences on insurance companies' portfolios. Exposure characteristics and particularities on business practises were identified and pricing implications and difficulties were mentioned in context with significant possible underwriting options.

Moreover, the significance of the meteorological input was highlighted and its contribution to the prediction of claims occurrence of a severe weather related event with high financial impact on the insurance sector was estimated. Meteorological indices in combination with demographic and insurance data input were combined in order to reveal specific factors which may trigger a significant insurance loss.

Distributions regarding the frequency and the severity of insurance claims were examined regarding their "goodness of fit" characteristics, and best options were pointed out. Using machine learning techniques, a methodological approach regarding the proximate cause of meteorological risk was undertaken and a measurement model was deployed regarding the significance and contribution of specific meteorological parameters.

Furthermore, as Greece is not included in the Solvency II "Standard Formula" calculation for weather related perils, a Solvency Capital Requirement approach was examined in order to estimate and measure flood risk, identify regions with high vulnerability indication and highlight spatial geographical patterns for more effective underwriting and pricing processes. Throughout the whole process, weaknesses and restrictions were mentioned in order to be included as a guideline in sensitivity scenarios and back-testing techniques in order to result in a clear risk identification range.

Several compound frequency models were employed to jointly fit the storm and claim frequencies observed by a Greek insurance company. The proposed models reveal the intrinsic number of storms and number of claims attributed to each storm implied

by the Greek climate insurance data-set. After fitting the proposed models to the data set there were several findings. Firstly, compared to cars, motorcycles and trucks are more vulnerable and at a greater risk so they trigger more excess claims (per unit exposure) when a storm occurs. Further, as the selected models consist of multiple mixture components, marginal impacts of each covariates to the actual numbers of storms and number of claims per storm could be flexibly described via partial dependence plots, which are not necessarily either linear or monotonic. Finally, it was found that the actual number of storms is negatively correlated to the number of claims per storm, which is not obvious from the incomplete observations of M and N.

We also acknowledge that our research work can be extended in various directions in the future, which includes but is not limited to frequency-severity modelling of storm-triggered claims that explores how the storm frequencies influence not only the claim frequencies but also the claim amounts, analysis of the storm-triggered insurance claims from multiple lines of businesses, and/or spatial-temporal modelling of the data to understand how the storm arrival process from one geospatial region influences adjacent regions.

Finally, a potentially interesting line of further research would be to impact of weather related hazards due to climate change on different types of claims from the same and/or different types of policies based on multivariate claim frequency and severity models and copula based models.

Annex I

Main Distributions for Frequency Modeling

Negative Binomial Distribution

The negative binomial distribution is a discrete probability distribution that models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of successes (denoted r) occurs. An alternative approach is to model the number of total trials (instead of the number of failures). In fact, for a specified (non-random) number of successes (r), the number of failures (n − r) are random because the total trials (n) are random. The Pascal distribution (after Blaise Pascal) and Polya distribution (for George Pólya) are special cases of the negative binomial distribution. A convention among engineers, climatologists, and others is to use "negative binomial" or "Pascal" for the case of an integer-valued stopping-time parameter (r) and use "Polya" for the real-valued case. For occurrences of associated discrete events, like tornado outbreaks, the Polya distributions can be used to give more accurate models than the Poisson distribution by allowing the mean and variance to be different, unlike the Poisson. The negative binomial distribution has a variance μ/p, with the distribution becoming identical to Poisson in the limit $p \to 1$ for a given mean (i.e. when the failures are increasingly rare). This can make the distribution a useful overdispersed alternative to the Poisson distribution, for example for a robust modification of Poisson regression. In epidemiology, it has been used to model disease transmission for infectious diseases where the likely number of onward infections may vary considerably from individual to individual and from setting to setting.More generally, it may be appropriate where events have positively correlated occurrences causing a larger variance than if the occurrences were independent, due to a positive covariance term.

The probability mass function of the negative binomial distribution is

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{k} (1 - p)^k p^r \tag{60}$$

where r is the number of successes, k is the number of failures, and p is the probability of success on each trial.

The binomial coefficient is equal to:

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{(r-1)!\,(k)!} = \frac{(k+r-1)(k+r-2)\cdots(r)}{k!}. \tag{61}$$

The cumulative distribution function of the binomial distribution is:

$$F(k; r, p) = F_{\text{binomial}}(k; n = k + r, 1 - p). \tag{62}$$

Poisson Distribution

A discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$, if it has a probability mass function given by

$$f(k; \lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{63}$$

with $\lambda = \mathrm{E}(X) = \mathrm{Var}(X)$.

Main Distributions for Claims Severity Modeling

Gamma Distribution

Gamma distribution is a two-parameter family of continuous probability distributions. The exponential distribution, Erlang distribution, and chi-squared distribution are special cases of the gamma distribution. There are two equivalent parameterizations in common use:

- With a shape parameter k and a scale parameter θ.
- With a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = 1/\theta$, called a rate parameter.

A random variable X that is gamma-distributed with shape α and rate β is denoted $X \sim \Gamma(\alpha, \beta) \equiv \text{Gamma}(\alpha, \beta)$

The corresponding probability density function in the shape-rate parameterization is:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^{\alpha}}{\Gamma(\alpha)} \quad \text{for } x > 0 \quad \alpha, \beta > 0, \tag{64}$$

where $\Gamma(\alpha)$ is the gamma function. For all positive integers, $\Gamma(\alpha) = (\alpha - 1)!$.

The cumulative distribution function is the regularized gamma function:

$$F(x; \alpha, \beta) = \int_0^x f(u; \alpha, \beta)\, du = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}, \tag{65}$$

where $\gamma(\alpha, \beta x)$ is the lower incomplete gamma function.

If α is a positive integer (i.e., the distribution is an Erlang distribution), the cumulative distribution function has the following series expansion:

$$F(x;\alpha,\beta) = 1 - \sum_{i=0}^{\alpha-1} \frac{(\beta x)^i}{i!} e^{-\beta x} = e^{-\beta x} \sum_{i=\alpha}^{\infty} \frac{(\beta x)^i}{i!}.$$

A random variable X that is gamma-distributed with shape k and scale θ is denoted by $X \sim \Gamma(k,\theta) \equiv \text{Gamma}(k,\theta)$

The probability density function using the shape-scale parametrization is:

$$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k,\theta > 0. \tag{66}$$

The cumulative distribution function is the regularized gamma function:

$$F(x;k,\theta) = \int_0^x f(u;k,\theta)\,du = \frac{\gamma\left(k,\frac{x}{\theta}\right)}{\Gamma(k)}, \tag{67}$$

where $\gamma\left(k,\frac{x}{\theta}\right)$ is the lower incomplete gamma function.

It can also be expressed as follows, if k is a positive integer (i.e., the distribution is an Erlang distribution)

$$F(x;k,\theta) = 1 - \sum_{i=0}^{k-1} \frac{1}{i!}\left(\frac{x}{\theta}\right)^i e^{-x/\theta} = e^{-x/\theta} \sum_{i=k}^{\infty} \frac{1}{i!}\left(\frac{x}{\theta}\right)^i. \tag{68}$$

The mean of gamma distribution is given by the product of its shape and scale parameters:

$$\mu = k\theta = \alpha/\beta$$

The variance is:

$$\sigma^2 = k\theta^2 = \alpha/\beta^2$$

The square root of the inverse shape parameter gives the coefficient of variation:

$$\sigma/\mu = k^{-0.5} = 1/\sqrt{\alpha}$$

Pareto Distribution

If X is a random variable with a Pareto (Type I) distribution, then the probability that X>x, is given by:

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha} & x \geqslant x_{\mathrm{m}}, \\ 1 & x < x_{\mathrm{m}}, \end{cases} \tag{69}$$

where X_m is the minimum possible value of X, and α is a positive parameter. The Pareto Type I distribution is characterized by a scale parameter X_m and a shape parameter α, which is known as the tail index or Pareto index.

The cumulative distribution function of a Pareto random variable with parameters α and X_m is

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_\mathrm{m}}{x}\right)^\alpha & x \geqslant x_\mathrm{m}, \\ 0 & x < x_\mathrm{m}. \end{cases} \tag{70}$$

The probability density function is:

$$f_X(x) = \begin{cases} \frac{\alpha x_\mathrm{m}^\alpha}{x^{\alpha+1}} & x \geqslant x_\mathrm{m}, \\ 0 & x < x_\mathrm{m}. \end{cases} \tag{71}$$

The expected value of a random variable following a Pareto distribution is

$$\mathrm{E}(X) = \begin{cases} \infty & \alpha \leqslant 1, \\ \frac{\alpha x_\mathrm{m}}{\alpha - 1} & \alpha > 1. \end{cases}$$

The variance of a random variable following a Pareto distribution is

$$\mathrm{Var}(X) = \begin{cases} \infty & \alpha \in (1, 2], \\ \left(\frac{x_\mathrm{m}}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2} & \alpha > 2. \end{cases}$$

(If α 1, the variance does not exist.)

Weibull Distribution

The “Weibull” distribution is a continuous probability distribution. It models a broad range of random variables, largely in the nature of a time to failure or time between events.

The standard parameterization of the probability density function of a Weibull random variable is

$$f(x;\lambda,k)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}, & x\geqslant 0,\\ 0, & x<0,\end{cases} \tag{72}$$

where k >0 is the shape parameter and $\quad>0$ is the scale parameter of the distribution.

$$F(x;\lambda,k)=\begin{cases}1-e^{-(x/\lambda)^k}, & x\geqslant 0,\\ 0, & x<0.\end{cases} \tag{73}$$

The Mean is $\lambda\,\Gamma(1+1/k)$ and the Median $\lambda(\ln 2)^{1/k}$

$$\mathrm{E}(X)=\lambda\Gamma\left(1+\frac{1}{k}\right)$$

and

$$\operatorname{var}(X)=\lambda^2\left[\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]. \tag{74}$$

The skewness is given by:

$$\gamma_1=\frac{2\Gamma_1^3-3\Gamma_1\Gamma_2+\Gamma_3}{[\Gamma_2-\Gamma_1^2]^{3/2}} \tag{75}$$

where

$\Gamma_i=\Gamma(1+i/k)$, which may also be written as

$$\gamma_1 = \frac{\Gamma\left(1+\frac{3}{k}\right)\lambda^3 - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

where the mean is denoted by μ and the standard deviation is denoted by σ.

The excess kurtosis is given by:

$$\gamma_2 = \frac{-6\Gamma_1^4 + 12\Gamma_1^2\Gamma_2 - 3\Gamma_2^2 - 4\Gamma_1\Gamma_3 + \Gamma_4}{[\Gamma_2 - \Gamma_1^2]^2} \tag{76}$$

where

$$\Gamma_i = \Gamma(1 + i/k)$$

The kurtosis excess may also be written as:

$$\gamma_2 = \frac{\lambda^4\Gamma(1+\frac{4}{k}) - 4\gamma_1\sigma^3\mu - 6\mu^2\sigma^2 - \mu^4}{\sigma^4} - 3.$$

The Moment generating function is:

$$\mathrm{E}\left[e^{tX}\right] = \int_0^\infty e^{tx}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}\,dx.$$

Lognormal Distribution

Log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed.

A positive random variable X is log-normally distributed,

$X \sim \text{Lognormal}(\mu_x, \sigma_x^2)$, if the natural logarithm of X is normally distributed with mean μ and variance σ^2:

$\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$

Let Φ and φ be respectively the cumulative probability distribution function and the probability density function of the N(0,1) distribution, then we have that:

$$\begin{aligned} f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\Pr(X \leqslant x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(\ln X \leqslant \ln x) = \frac{\mathrm{d}}{\mathrm{d}x}\Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\ &= \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\ln x - \mu}{\sigma}\right) = \varphi\left(\frac{\ln x - \mu}{\sigma}\right)\frac{1}{\sigma x} \\ &= \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \end{aligned} \tag{77}$$

The cumulative distribution function is

$$F_X(x) = \Phi\left(\frac{(\ln x) - \mu}{\sigma}\right) \tag{78}$$

where Φ is the cumulative distribution function of the standard normal distribution.

This may also be expressed as below:

$$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right] = \frac{1}{2}\operatorname{erfc}\left(-\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)$$

Burr Distribution

The Burr (Type XII) distribution has probability density function:

$$\begin{aligned} f(x;c,k) &= ck\frac{x^{c-1}}{(1+x^c)^{k+1}} \\ f(x;c,k,\lambda) &= \frac{ck}{\lambda}\left(\frac{x}{\lambda}\right)^{c-1}\left[1+\left(\frac{x}{\lambda}\right)^c\right]^{-k-1} \end{aligned} \tag{79}$$

and cumulative distribution function:

$$F(x;c,k) = 1-(1+x^c)^{-k} \tag{80}$$

and

$$F(x;c,k,\lambda) = 1-\left[1+\left(\frac{x}{\lambda}\right)^c\right]^{-k}$$

Annex II

Hill's Tail Index Estimator

Let $(X_t, t \geqslant 1)$ be a sequence of independent and identically distributed random variables with distribution function F $\in D(H(\xi))$, the maximum domain of attraction of the generalized extreme value distribution H, where $\xi \in \mathbb{R}$. The sample path is $X_t : 1 \leqslant t \leqslant n$ where n is the sample size. If $\{k(n)\}$ is an intermediate order sequence, i.e. $k(n) \in \{1, \ldots, n-1\}$, $k(n) \to \infty$ and $k(n)/n \to 0$, then the Hill tail-index estimator is:

$$\xi^{\text{Hill}}_{(k(n),n)} = \left(\frac{1}{k(n)} \sum_{i=n-k(n)+1}^{n} \ln(X_{(i,n)}) - \ln(X_{(n-k(n)+1,n)}) \right)^{-1}, \tag{81}$$

where $X_{(i,n)}$ is the i-th order statistic of $X_1, \ldots, X_n$. This estimator converges in probability to ξ, and is asymptotically normal provided $k(n) \to \infty$ is restricted based on a higher order regular variation property. Consistency and asymptotic normality extend to a large class of dependent and heterogeneous sequences,irrespective of whether X_t is observed, or a computed residual or filtered data from a large class of models and estimators, including mis-specified models and models with errors that are dependent. Hill's tail-index estimators commonly make use of logarithm of the order statistics. [26]

In case of heavy tailed data Hill's method can be used in order to have an indication about a threshold in order to separate the large and the "attritional" claims. Using the projected ultimate claims amount an optimal threshold is determined using the above approach based on the Generalised Pareto Distribution (GPD).

Akaike Information Criterion - AIC

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other

models. Thus, AIC provides a means for model selection.

If k is the number of estimated parameters in the model and $\hat{L}$ the maximized value of the likelihood function then the AIC value of the model is:

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{82}$$

Baysian Information Criterion - BIC or SBC

Bayesian information criterion (BIC) or Schwarz information criterion (SBC) is a criterion for model selection among a finite set of models. Models with lower BIC are generally preferred. When fitting models, it is possible to increase the maximum likelihood by adding parameters, but doing so may result in over-fitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC for sample sizes greater than 7.

The BIC is formally defined as:

$$\text{BIC} = k\ln(n) - 2\ln(\widehat{L}). \tag{83}$$

where $\hat{L}$ the maximized value of the likelihood function of the model, n the sample size, k the number of parameters estimated by the model.

Kolmogorov–Smirnov Test

The Kolmogorov–Smirnov test (K–S test or KS test) is a non-parametric test of the equality of continuous or discontinuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

The empirical distribution function Fn for n independent and identically distributed (i.i.d.) ordered observations Xi is defined as:

$$F_n(x) = \frac{\text{number of (elements in the sample} \leqslant x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i),$$

where

$$1_{(-\infty,x]}(X_i)$$

is the indicator function, equal to 1 if $X_i \leqslant x$ and equal to 0 otherwise.

The Kolmogorov–Smirnov statistic for a given cumulative distribution function F(x) is

$$D_n = \sup_x |F_n(x) - F(x)| \tag{84}$$

Cramér–von Mises Criterion

The "Cramér–von Mises" criterion is a criterion used for judging the goodness of fit of a cumulative distribution function F^* compared to a given empirical distribution function F_n, or for comparing two empirical distributions. It is also used as a part of other algorithms, such as minimum distance estimation. It is defined as:

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 \, \mathrm{d}F^*(x)$$

Let $x_1, x_2, \ldots, x_n$ be the observed values, in increasing order. Then the statistic is:

$$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2 . \tag{85}$$

Anderson–Darling Test

If the hypothesized distribution is F, and empirical (sample) cumulative distribution function is F_n, then the quadratic EDF statistics measure the distance between F and F_n by

$$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \, w(x) \, dF(x) \tag{86}$$

where n is the number of elements in the sample, and w(x) is a weighting function.

The Anderson–Darling test assesses whether a sample comes from a specified distribution. It makes use of the fact that, when given a hypothesized underlying distribution and assuming the data does arise from this distribution, the cumulative distribution function (CDF) of the data can be assumed to follow a uniform distribution. The data can be then tested for uniformity with a distance test (Shapiro 1980). The formula for the test statistic A to assess if data $\{Y_1 < \cdots < Y_n\}$ (note that the data must be put in order) comes from a CDF F is:

$$A^2 = -n - S \tag{87}$$

where

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln(F(Y_i)) + \ln\left(1 - F(Y_{n+1-i})\right) \right].$$

References

[1] Agency, E. E. (2017). *Climate change adaptation and disaster risk reduction in Europe: enhancing coherence of the knowledge base, policies and practices*. Office for Official Publ. of the Europ. Communities.

[2] Ahmed, O. and Serra, T. (2015). Economic analysis of the introduction of agricultural revenue insurance contracts in spain using statistical copulas. *Agricultural Economics*, 46(1):69–79.

[3] Badescu, A., Lan, G., Lin, X. S., and Tang, D. (2015). Modeling correlated frequencies with application in operational risk management. *Journal of Operational Risk*, 10(1):1–43.

[4] Breiman, L. (1996). Stacked regressions. *Machine learning*, 24(1):49–64.

[5] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

[6] Broberg, M. (2020). Parametric loss and damage insurance schemes as a means to enhance climate change resilience in developing countries. *Climate Policy*, 20(6):693–703.

[7] Caselli, G. and Figueira, C. (2020). The impact of climate risks on the insurance and banking industries. In *Sustainability and Financial Risks*, pages 31–62. Springer.

[8] Charbuty, B. and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28.

[9] Choudhury, A., Jones, J., Okine, A., and Choudhury, R. (2016). Drought-triggered index insurance using cluster analysis of rainfall affected by climate change. *Journal of Insurance Issues*, pages 169–186.

[10] Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

[11] Daron, J. and Stainforth, D. (2014). Assessing pricing assumptions for weather index insurance in a changing climate. *Climate Risk Management*, 1:76–91.

[12] Diakakis, M., Andreadakis, E., Nikolopoulos, E., Spyrou, N., Gogou, M., Deligiannakis, G., Katsetsiadou, N., Antoniadis, Z., Melaki, M., Georgakopoulos, A., et al. (2019). An integrated approach of ground and aerial observations in flash flood disaster investigations. the case of the 2017 mandra flash flood in greece. *International Journal of Disaster Risk Reduction*, 33:290–309.

[13] Diakakis, M., Nikolopoulos, E., Mavroulis, S., Vassilakis, E., and Korakaki, E. (2017). Observational evidence on the effects of mega-fires on the frequency of hydrogeomorphic hazards. the case of the peloponnese fires of 2007 in greece. *Science of the total environment*, 592:262–276.

[14] Dietterich, T. G. et al. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125.

[European Environment Agency] European Environment Agency. Economic losses and fatalities from weather- and climate-related events in Europe. Available at https://www.eea.europa.eu/publications/economic-losses-and-fatalities-from (2022/02/03).

[16] European Environment Agency (2022). What is CORINE land cover? [Online]. Available from: https://www.eea.europa.eu/help/faq/what-is-corine-land-cover.

[17] Feyen, L., Ciscar Martinez, J. C., Gosling, S., Ibarreta Ruiz, D., Soria Ramirez, A., Dosio, A., Naumann, G., Russo, S., Formetta, G., Forzieri, G., et al. (2020). Climate change impacts and adaptation in europe. jrc peseta iv final report. Technical report, Joint Research Centre (Seville site).

[18] Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

[19] Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

[20] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[21] Friedman, J. H., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York, NY, USA.

[22] Fung, T. C., Badescu, A. L., and Lin, X. S. (2019). A class of mixture of experts models for general insurance: Theoretical developments. *Insurance: Mathematics and Economics*, 89:111–127.

[23] Furukawa, K., Ichiue, H., Shiraki, N., et al. (2020). How does climate change interact with the financial system? a survey. Technical report, Bank of Japan.

[24] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

[Hellenic Association of Insurance Companies] Hellenic Association of Insurance Companies. Property insurances 2020 year survey. Available at http://www1.eaee.gr/sites/default/files/review_property_2020.pdf.

[26] Hill, J. B. (2010). On tail index estimation for dependent, heterogeneous data. *Econometric Theory*, 26(5):1398–1436.

[27] Hyafil, L. and Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Inf. Process. Lett.*, 5(1):15–17.

[28] Jeong, H. and Dey, D. (2020). Application of a vine copula for multi-line insurance reserving. *Risks*, 8(4):111.

[29] Jóhannsdóttir, L., Wallace, J., and Jones, A. (2012). The primary insurance industry's role in managing climate change risks and opportunities. In *Managing climate change business risks and consequences: Leadership for global sustainability*, pages 51–79. Springer.

[30] Kahn, M. E. (2016). The climate change adaptation literature. *Review of Environmental Economics and Policy*, 10(1):166–178.

[31] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

[32] Knighton, J., Buchanan, B., Guzman, C., Elliott, R., White, E., and Rahm, B. (2020). Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity. *Journal of Environmental Management*, 272:111051.

[33] Lagouvardos, K., Karagiannidis, A., Dafis, S., Kalimeris, A., and Kotroni, V. (2022). Ianos—a hurricane in the mediterranean. *Bulletin of the American Meteorological Society*, 103(6):E1621–E1636.

[34] Lagouvardos, K., Kotroni, V., Giannaros, T. M., and Dafis, S. (2019). Meteorological conditions conducive to the rapid spread of the deadly wildfire in eastern attica, greece. *Bulletin of the American Meteorological Society*, 100(11):2137–2145.

[35] Leblois, A. and Quirion, P. (2013). Agricultural insurances based on meteorological indices: realizations, methods and research challenges. *Meteorological Applications*, 20(1):1–9.

[36] Lee, G. Y. and Shi, P. (2019). A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics*, 87:115–129.

[37] Lyubchich, V., Newlands, N. K., Ghahari, A., Mahdi, T., and Gel, Y. R. (2019). Insurance risk assessment in the face of climate change: Integrating data science and statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4):e1462.

[38] Mobley, W., Sebastian, A., Blessing, R., Highfield, W. E., Stearns, L., and Brody, S. D. (2021). Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast texas. *Natural Hazards and Earth System Sciences*, 21(2):807–822.

[39] Odening, M. and Shen, Z. (2014). Challenges of insuring weather risk in agriculture. *Agricultural Finance Review*, 74(2):188–199.

[40] Olarewaju, O. O. and Nomlala, B. C. (2021). Climate change and insurance business in developing economies. In *Handbook of Research on Climate Change and the Sustainable Financial Sector*, pages 190–201. IGI Global.

[41] Papagiannaki, K., Kotroni, V., Lagouvardos, K., Bezes, A., Vafeiadis, V., Messini, I., Kroustallis, E., and Totos, I. (2022). Identification of rainfall thresholds likely to trigger flood damages across a mediterranean region, based on insurance data and rainfall observations. *Water*, 14(6).

[42] Papagiannaki, K., Lagouvardos, K., and Kotroni, V. (2013). A database of high-impact weather events in greece: a descriptive impact analysis for the period 2001–2011. *Natural Hazards and Earth System Sciences*, 13(3):727–736.

[43] Papagiannaki, K., Lagouvardos, K., Kotroni, V., and Bezes, A. (2015). Flash flood occurrence and relation to the rainfall hazard in a highly urbanized area. *Natural Hazards and Earth System Science*, 15(8):1859–1871.

[44] Papavasileiou, G., Kotroni, V., Lagouvardos, K., and Giannaros, T. M. (2022). Observational and numerical study of a giant hailstorm in attica, greece, on 4 october 2019. *Atmospheric Research*, 278:106341.

[45] Perrone, M. P. (1993). *Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization*. PhD thesis, Physics Department, Brown University, Providence, RI.

[46] Phelan, L., Holley, C., Shearing, C., and du Toit, L. (2020). Insurance and climate change. In *Routledge International Handbook of Green Criminology*, pages 449–462. Routledge.

[47] Prudential Regulation Authority (2015). The impact of climate change on the UK insurance sector. *A Climate Change Adaptation Report*.

[48] Rädler, A. T., Groenemeijer, P. H., Faust, E., Sausen, R., and Púčik, T. (2019). Frequency of severe thunderstorms across europe expected to increase in the 21st century due to rising instability. *npj Climate and Atmospheric Science*, 2(1):30.

[49] Ranger, N. and Surminski, S. (2013). A preliminary assessment of the impact of climate change on non-life insurance demand in the BRICS economies. *International Journal of Disaster Risk Reduction*, 3:14–30.

[50] Robinson, P. J. and Botzen, W. (2022). Setting descriptive norm nudges to promote demand for insurance against increasing climate change risk. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 47(1):27–49.

[51] Rohrbeck, C., Eastoe, E. F., Frigessi, A., and Tawn, J. A. (2018). Extreme value modelling of water-related insurance claims. *The Annals of Applied Statistics*, 12(1):246–282.

[52] Savitz, R. and Dan Gavriletea, M. (2019). Climate change and insurance. *Transformations in Business & Economics*, 18(1).

[53] Soliman, M., Lyubchich, V., Gel, Y. R., Naser, D., and Esterby, S. (2015). Evaluating the impact of climate change on dynamics of house insurance claims. In *Machine learning and data mining approaches to climate science*, pages 175–183. Springer.

[54] Springael, L. and Van Nieuwenhuyse, I. (2006). On the sum of independent zero-truncated Poisson random variables. *Research paper 2006-011, June*, pages 1–15.

[55] Surminski, S., Bouwer, L. M., and Linnerooth-Bayer, J. (2016). How insurance can support climate resilience. *Nature Climate Change*, 6(4):333–334.

[Swiss Re Group] Swiss Re Group. Global insured catastrophe losses rise to usd 112 billion in 2021, the fourth highest on record, swiss re institute estimates. Available at https://www.swissre.com/media/press-release/nr-20211214-sigma-full-year-2021-preliminary-natcat-loss-estimates.html (2021/12/14).

[57] Tack, J. B. and Ubilava, D. (2015). Climate and agricultural risk: measuring the effect of enso on us crop insurance. *Agricultural Economics*, 46(2):245–257.

[58] Tesselaar, M., Botzen, W. W., and Aerts, J. C. (2020). Impacts of climate change and remote natural catastrophes on eu flood insurance markets: An analysis of soft and hard reinsurance markets for flood coverage. *Atmosphere*, 11(2):146.

[59] Thistlethwaite, J. and Wood, M. O. (2018). Insurance and climate change risk management: rescaling to look beyond the horizon. *British Journal of Management*, 29(2):279–298.

[60] Tolika, K., Maheras, P., and Tegoulias, I. (2009). Extreme temperatures in greece during 2007: Could this be a "return to the future"? *Geophysical Research Letters*, 36(10).

[61] Tsakiris, G. (2014). Flood risk assessment: concepts, modelling, applications. *Natural Hazards and Earth System Sciences*, 14(5):1361–1369.

[62] Walker, G. R., Mason, M. S., Crompton, R. P., and Musulin, R. T. (2016). Application of insurance modelling tools to climate change adaptation decision-making relating to the built environment. *Structure and Infrastructure Engineering*, 12(4):450–462.

[63] Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

[64] Yang, Q., Shen, X., Yang, F., Anagnostou, E. N., He, K., Mo, C., Seyyedi, H., Kettner, A. J., and Zhang, Q. (2022). Predicting flood property insurance claims over conus, fusing big earth observation data. *Bulletin of the American Meteorological Society*, 103(3):E791–E809.