Università
Ca'Foscari
Venezia
Dipartimento di Scienze Ambientali,
Informatica e Statistica

Master's Degree Programme in
Environmental Sciences
D.M. 270/2004

Final Thesis

Assessing risks at the land-sea interface: the case study of the Veneto coastal area

Supervisor
Ch. Prof. Andrea Critto

Co-tutor
Dr. Silvia Torresan

Graduand
Anna Pasquali
Matriculation Number 888225

Academic Year
2021-2022

TABLE OF CONTENTS

List of figures

List of tables

Glossary

ARPAV	*Agenzia Regionale Prevenzione Ambiente Veneto*
BN	*Bayesian Network*
CDD	*Consecutive dry days*
CMEMS	*Copernicus Marine Environment Monitoring Service*
DPGR	*Decreti Del Presidente della Giunta Regionale*
DRR	*Disaster Risk Reduction*
DSS	*Decision Support System*
DT	*Decision Tree*
EDA	*Exploratory Data Analysis*
EEA	*European Environment Agency*
EPA	*U.S. Environmental Protection Agency*
ESV	*Eastward seawater velocity*
ESWD	*European Severe Weather Database*
HuxWF	*Number of days in a year with mean daily Humidex value equal or higher than 35 °C for almost 3 consecutive days*
HWN	*Number of heatwaves in a year*
HWTXdx	*Heatwave temperature*
IPCC	*Intergovernmental Panel On Climate Change*
ISTAT	*Istituto Nazionale Di Statistica (Italian National Institute Of Statistics)*
ML	*Machine Learning*
MSSH	*Maximum sea surface height*
MWAH	*Maximum significant wave height*
MWIH	*Maximum significant wind wave height*
NOAA	*National Oceanic and Atmospheric Administration*
NSV	*Northward seawater velocity*
QGIS	*Quantum Geographic Information System*
RF	*Random Forest*
RH	*Relative humidity*
RX-1day	*Maximum cumulative precipitation in 1 day*
RX-5day	*Maximum cumulative precipitation in 5 consecutive days*
SSH	*Sea surface height*
TR	*Number of tropical nights*
TX90p	*Number of hot days*

UNEP	*United Nations Environmental Program*
UNISDR	*United Nations Office for Disaster Risk Reduction*
WAD	*Wave direction from*
WAH	*Significant wave height*
WAP	*Sea surface wave mean period*
WID	*Wind wave direction from*
WIH	*Significant wind wave height*
WIP	*Sea surface wind wave mean period*

SUMMARY

Extreme weather events are causing severe threats all over the world, posing increasing environmental and socio-economic risks, which are amplified by the effect of climate change. Coastal areas are particularly vulnerable to extreme marine and weather events (e.g., storm surges, extreme rainfall) given the high exposure of population, settlements, and economic activities at the land-sea interface.

Therefore, understanding the main risk factors of these extreme events is necessary to implement suitable disaster risk management measures, which could guide coastal authorities and policy-makers in improving the resilience of coastal communities to natural hazards and climate change.

Nevertheless, identifying the triggering factors of such risks has always been challenging, since the complex dynamics driving the coastal systems. In this regard, in order to unveil relations between hazards and their cascading effects, in recent years, Machine Learning (ML) algorithms have gained popularity due to their ability to extract information from a huge quantity of data, by overcoming the limits of traditional physical-mathematical models. However, the outcomes of these advanced methods, to be reliable, must be corroborated through traditional statistical analysis and scientific reasoning.

Based on these needs, this Thesis is aimed at investigating the factors that play a key role in the occurrence of damages (e.g., damages to people, buildings and infrastructures, agriculture, tertiary sector) generated by extreme weather events in the coastal municipalities of the Veneto region, focusing on the 2009-2019 timeframe. Accordingly, the aim was achieved by reviewing the scientific literature concerning the state-of-the-art ML algorithms implemented for assessing risks and impacts caused by natural hazards in coastal areas, as well as by applying traditional and ML-driven techniques of data science to find relations between the analyzed factors and the damage occurrences.

In particular, the scientometric and the systematic review revealed ML algorithms based on decision trees (e.g., Bayesian Network and Random Forests) as the main implemented models, given their high predictive ability and easy interpretation. Moreover, the majority of these models adopted, as input variables, indicators related to sea surface level, wave regime and precipitation.

Building upon the findings of the review, a comprehensive data analysis process was applied to the dataset made available from the AdriaClim project, in order to explore trends and relations between the collected atmospheric, oceanographic and territorial indicators with extreme weather-driven damages, both at the regional and local scale of the Veneto coastal area. Two data analysis techniques were used for accomplishing the study: a Random Forest (RF) algorithm for selecting the most important features related to the damage occurrence, and traditional Exploratory Data Analysis (EDA) both for an initial pre-analysis of the dataset and to evaluate the results of the RF.

The pre-analysis of the dataset was performed to identify the criticalities and main characteristics of the data, allowing a better design of the RF algorithm. Specifically, the obtained information regarding the presence

of similar patterns between the damage trends and those of some hazard indicators, and the presence of significant differences in the hazard and territorial indicators at the local scale, served to test several combinations of input variables for implementing the RF.

Consequently, after having balanced the initial dataset, due to the high discrepancy between the number of observations with and without damages, the set-up RF model was run gaining a F1 score of 95% and identifying sea surface height, precipitation, temperature, and wave characteristics as the most relevant features. Then, a further examination of these variables at the regional scale, through EDA techniques, permitted to assess their effective relevance when damages occurred, confirming the reliability of the RF. Additionally, for some of the selected features (e.g., mean and maximum sea surface height, significant wave height), it was possible to find threshold values associated with the damage occurrence and their relative annual and seasonal variations, information that could be helpful for the application of early warning systems. However, the same analyses executed at the municipal scale revealed different local characteristics for some hazard indicators recorded in presence of damage.

Overall, the developed methodology has pointed out some interesting relationships between the triggering factors and the damages occurred in the case study area within the 2009-2019 timeframe. These findings can pave the way for guiding decision-makers and local stakeholders in the development of suitable disaster risk reduction and climate adaptation measures, aimed at increasing the resilience of coastal communities to extreme weather events.

Finally, although the research encountered some limits due to the type and the resolution of the data, especially concerning the damage data and the exposure and vulnerability indicators, the results and the criticalities evidenced by this study could be useful for the implementation of advanced ML algorithms (e.g., Graph Neural Networks, Artificial Neural Networks) intended to predict damage occurrences in coastal areas.

OBJECTIVES AND MOTIVATIONS

In the last decade, extreme weather events have occurred with an increased frequency and intensity, worldwide and particularly in coastal areas (Seneviratne et al., 2012; EEA, 2022), presenting significant challenges to understanding, evaluating, and predicting the environmental risk (Zhou et al., 2022).

Moreover, the regions at the land-sea interface have recorded higher costs for damages and losses caused by such extreme events than inland zones (Li et al., 2022; EEA, 2022b), costs that, due to climate change, are expected to increase in the upcoming years (Coronese et al., 2019; Roudier et al., 2016). The reasons behind these severe consequences are determined by the greater number of natural hazards and socio-economic assets in coastal areas. In particular, these zones are affected by both atmospheric and marine hazards, which amplify the magnitude of the effects when combined. Additionally, they are inherently more vulnerable and exposed to risks (Nicholls et all., 2007) due to the several anthropogenic pressures, including population growth, tourism and numerous buildings and infrastructures.

Hence, coastal communities are increasingly requiring mitigation and adaptation plans to improve their resilience against the growing number of natural disasters.

In order to implement suitable disaster risk management, guided by the principles of the Sendai Framework for Disaster Risk Reduction 2015-2030 (UNISDR, 2015), the identification of the triggering factors of the damage risks is of paramount importance. However, the comprehension of such risks has always been demanding because of the multiple complex and non-linear interactions driving the coastal systems.

Building on these needs and for overcoming the limits of traditional physical-mathematical models, in recent years, Machine Learning (ML) algorithms have gained popularity in several natural hazard-related issues (Arinta & Andi, 2019; Wendler-Bosco & Nicholson, 2022), including extreme weather events (Qi & Majda, 2020). Specifically, ML models are powerful tools that can extract information from the input dataset by identifying structures, patterns, and relationships among variables. Moreover, even when working with a huge quantity of data and variables (Kuhn & Johnson, 2013), they are able to determine the most relevant factors driving the risk (Genuer et al., 2010).

On the other hand, these algorithms have been described as “black boxes”, capable of providing excellent predictions, but whose outcomes should be accurately evaluated, considering the physical and environmental aspects (Jones & Linder, 2015). Consequently, the combination of traditional data analysis techniques (i.e., descriptive statistics and Exploratory Data Analysis - EDA) with ML methods has been suggested for providing a comprehensive understanding of the investigated phenomena (Hafen & Critchlow, 2013). This is particularly important in the context of extreme weather events occurring at the land-sea interface, since they are determined by complex dynamics, exhibiting peculiar characteristics in every manifestation.

Based on this knowledge background, the main objectives of this Thesis are:

i) reviewing the scientific literature concerning the state of the art of ML algorithms implemented for assessing natural hazard risks in coastal areas;

ii) applying a series of data science techniques, which combine traditional statistics with ML methods, to identify the most influential factors in damage occurrences caused by extreme weather events, for the coastal municipalities of the Veneto region within the 2009-2019 timeframe.

These main objectives are reached by subdividing the research in more detailed operative tasks, which constitute the theoretical, methodological, and operative assets for this Thesis. In particular:

- Scientometric and systematic literature review regarding the peer-reviewed publications, of the last twenty years, dealing with the application of ML methods to assess coastal risks caused by natural hazards;
- Description of the case study area in terms of geomorphological, territorial, and climatological characteristics, as well as collection of atmospherical, oceanographical, territorial and damage data for the case study area within the 2009-2019 timeframe;
- Development of a conceptual scheme for guiding the prediction of damages, given a set of indicators, through the application of ML methods;
- Design a methodological process of data analysis, which combines traditional EDA techniques with a Random Forest (RF) algorithm, to detect trends, most relevant features and relations between boundary conditions and damage occurrences, both at the regional and local scale;
- Discussion of the results, to provide the environmental understanding of the factors that have driven the occurred damages, by highlighting the strengths and limitations of the study.

The Thesis was developed in the frame of the Interreg IT-HR AdriaClim project (www.italy-croatia.eu/adriaclim), in collaboration with the Foundation *Centro Euro-Mediterraneo sui Cambiamenti Climatici* (CMCC, www.cmcc.it). The project aims to support, in the cooperation area, the development of science-based regional and local climate change adaptation plans based on up-to-date meteorological and oceanographical information, derived from advanced observing and modelling systems for the Adriatic Sea.

THESIS STRUCTURE

This Thesis is structured in two main sections: **Section A** provides a picture of the theoretical background at the base of Machine Learning (ML) methods for assessing risks and impacts caused by natural hazards in coastal areas; **Section B** describes the process and the techniques of data analysis applied for evaluating the factors that contributed to the manifestation of damages, during extreme weather events, in the case study area of the Veneto coastal municipalities.

Specifically:

Section A, following a brief introduction of the main concepts and terminologies in the ML field, performs a scientometric and systematic review of the scientific literature in relation to the state of the art of ML applications for assessing risks in coastal environments caused by natural hazards, such as extreme weather and climate change.

Section B, on the other hand, is organized following subsequent phases. In particular, the description of the investigated case study area (i.e., the Veneto coastal municipalities) outlines the geomorphological, territorial, and climatological characteristics, providing information regarding the data collection. The terminology adopted in the field of risk assessment is then introduced, operationalizing the conceptual scheme designed to identify the main relationships between hazard, exposure and vulnerability factors contributing to determining risk. The scheme is then used as a starting point to investigate the role of each risk factor in the damage manifestation, by implementing a data analysis methodology combining traditional EDA techniques with a RF model. Accordingly, the main results of these applications are discussed in order to detect relations between hazard and territorial indicators with damage occurrences, both at the regional and local scale.

Conclusions aim at providing a comprehensive summary of the results obtained from the data analysis process, applied to identify the main factors which influenced the manifestation of extreme weather-driven damages in the investigated area, highlighting criticalities and limitations of the research as well as future improvements.

SECTION A: Review of Machine Learning algorithms to assess risks caused by natural hazards in coastal areas

1. State of art of Machine Learning applications to assess impacts caused by natural hazards at the land-sea interface

Natural hazards have always been a source of risk for human communities. However, in the last decades, climate change has intensified the occurrence of these phenomena both in frequency as well as in magnitude (López et al., 2015), and future projections reveal a worsening of the current conditions (Roudier et al., 2016). Coastal areas are even more affected by such disruptive events due to the strong interplay between atmospherical and marine hazards. Moreover, these regions have an inherent high exposure and vulnerability to risk, determined by the elevated concentration of natural and socio-economical assets (e.g., population, infrastructures, economical activities). Therefore, to support policymakers and government authorities in identifying suitable management strategies to cope with the increasing manifestation of damaging events, a deep understanding of the factors contributing to coastal risks is required.

In recent years, ML methods have gained popularity to predict short-term risks caused by natural hazards. They have been exploited for their ability to extract information from the data by identifying structures, patterns, and relationships among variables (Wendler-Bosco & Nicholson, 2022), and to overcome the issue of working with a huge quantity of data (Kuhn & Johnson, 2013), by selecting the most important predictors and discharging the not relevant ones (Genuer et al., 2010).

Additionally, new developments are implementing ML methods to evaluate the best mitigation and adaptation strategies toward natural extreme events (Milojevic-Dupont & Creutzig, 2021; Biesbroek et al., 2020; Huntingford et al., 2019), serving as important tools for decision-makers. Nevertheless, these kinds of studies, especially if combined with long-term climate change scenarios, are still very few (Zennaro et al., 2021).

In the frame of this thesis, in order to acquire an overall understanding of the state of the art of ML methods applied in coastal environments to assess the risks of natural hazards, a detailed review of the existing publications has been conducted, through scientometric and systematic analysis, by evaluating the evolution and the limitations of this topic. Accordingly, the following sections, after a brief description of the main terminology adopted in the ML field (*Section 1.1*), describe the methods applied for carrying out the scientometric and systematic review (*Section 1.2*) and present the relative results and the main findings (*Section 1.3*).

1.1. Machine Learning: definition and characteristics

Machine learning (ML), defined by Arthur Samuel in 1959 as the "*field of study that gives computers the ability to learn without being explicitly programmed",* represents a branch of Artificial Intelligence (AI) that implements algorithms capable of learning their parameters from data by finding statistically significant patterns among them (Awad, 2015).

There are three main categories of machine learning algorithms: supervised learning algorithms, unsupervised learning algorithms, and reinforcement learning algorithms (Heidenreich, 2018).

Supervised learning algorithms are algorithms that try to detect the relations between input variables (often referred as "*features*") and output variables (often referred as "*labels*") using labeled examples (often referred as "*samples*" or "data points"), i.e. examples for which both the input and the output variables are known. In supervised learning algorithms the optimum parameters of the model are selected by finding the values that minimize the loss function (i.e. the distance between known and predicted labels) on the train set. Supervised learning comprehends two distinct types of methods namely regression and classification. In regression algorithms the output variables are continuous; examples of regression problems can be the forecasting of weather parameters, the estimate of natural hazards' impacts, and the prediction of future carbon emissions trends (Kumar, 2022). On the other hand, classification methods estimate the class to which the input value belongs to; examples of classifiers can be models that predict the expected flooded areas as a consequence of the sea-level rise (Park & Lee, 2020), the change of land cover due to earthquakes and tornados (Volke & Abarca-Del-Rio, 2020) and many others. Most supervised learning algorithms, like Support Vector Machines, Neural Networks, Decision Trees, Random Forests, and Bayesian Networks can be used for both regression and classification problems.

Unsupervised learning includes all those algorithms which draw inferences and find patterns from input data without the knowledge of labeled outcomes. The main unsupervised learning algorithms are *clustering methods,* which aim to detect the presence of clusters among the data.

In **Reinforcement Learning** the model trains itself continually, through trial and error processes, to acquire from the past the best possible knowledge to make accurate decisions; the main method falling in this ML category is the Markov Decision Process.

The main important feature of ML is the capacity to estimate a previously unknown relationship between input and output data that can be used to estimate the output of new input data. In order to assess the quality of the estimated relationship, the input data are divided into three sets: a **training set**, used by the algorithm to estimate the parameters of the model, a **validation set** used to compare different algorithms and to tune the hyperparameters, and finally, a **test set**, once the hyperparameters have been tuned, to verify if the machine has well learned the patterns or relationships among the data (algorithm performance). Finally, in the ML's terminology, there are two important concepts: *bias* and *variance.* Bias represents the inability of a ML method to capture the true relationship between the variables, while variance indicates the

change in the model's results if the model itself is trained with different portions of the training dataset (Gutta, 2020). If a model, trained with one dataset, has a low bias but high variance, it is said to be *overfitted*; a model is said to be *underfitted* if it cannot capture the relations among the variables (high bias). The best model is the one having low bias and low variance.

1.2. Review methods

1.2.1. Data collection

The bibliometric research of peer-review literature published between 2001-2021, related to the state of the art of machine learning methods for assessing the impacts of natural hazards in coastal areas, was performed by consulting as a source of information the open-free Scopus database (Elsevier; http://www.scopus.com). Scopus is among the largest curated abstract and citation databases, with high precision and recall (Baas et al., 2020). The wider range of scientific publication coverage (Darko et al., 2019) allows the selection of targeted publications through the formulation of a query in which the research keywords are specified. In addition, the bibliographic data can be exported in the R environment for the scientometric analysis. The query formulated in the frame of this review, schematized in Table 1 and reported extensively in ANNEX I, was structured in four blocks expressing the main concepts behind this research, which were related to the methodology, the study area, the presence of scenarios, and the typology of hazards/risks. In each block, the keywords were identified to capture the broad spectrum of publications in the context of ML methods applied in coastal environments to assess natural hazard impacts. Specifically, the four blocks were linked with the boolean operator AND, whereas the keywords internally to each block were linked with the boolean operator OR, in this way, publications containing at least one keyword for each block were retrieved.

The application of the query to "title, abstract and keywords" of all the papers present in the Scopus database (search date: 5th March 2022) selected 651 publications. Then, the records of the retrieved publications were exported in a BibTeX file (by keeping all the information regarding 'citation', 'bibliography', 'abstract & keywords', 'funding details', and 'other') to be uploaded into the bibliometrix R Package (R version 3.6.1, Bibliometrix package version 3.2.1; Aria & Cuccurullo, 2017) for a more in-depth investigation of the results.

Table 1: Scheme of the formulated query to retrieve the publications related to the research topic

Type of method	Type of study area	Scenarios	Type of hazard/assessment endpoints
("ML" OR "machine learning") OR ("deep learning") OR ("AI" OR "artificial intelligence") OR ("decision tree" OR "DT") OR ("random forest" OR "RF") OR ("Bayesian network" OR "BN")	"coast*" OR "marine*" OR "sea"	"climate change" OR "scenario*"	"erosion" OR ("water quality" OR "turbidity" OR "eutrophication") OR "storm surge" OR ("slr" OR "sea level*") OR "extreme event*" OR "pluvial flood" OR "flood*" OR "inundation" OR "drought" OR "heat wave*" OR ("risk*" OR "vulnerability" OR "exposure")

1.2.2. Scientometric analysis

Scientometrics was defined by David J. Hess in 1997 as a "quantitative study of science, communication in science and science policy", aiming to analyze the bibliographic records of a research topic to provide an overall picture of the current knowledge, the relative evolution and gaps (Chen, 2017; Darko et al., 2019), through a quantitative-based analysis, in order to be less influenced by the results' interpretation given by the researchers (De-Toledo et al., 2022).

The scientometric analysis of the investigated topic was performed through the open-source bibliometrix R-package, developed in 2017 by Massimo Aria and Corrado Cuccurullo for quantitative research and science mapping of literature review, created out of the need to have an effective tool for summarizing the information of the increasing academic publications. To favor the use of bibliometrix, biblioshiny (a R Shiny app) was designed to provide an easy interactive web interface.

Bibliometrix enables the implementation of numerous high-quality routines which can be gathered in three groups: i) *data collection*, to import bibliographic databases such as Scopus in R environment; ii) *data analysis*, comprehending both descriptive analysis and science mapping, this latter one related to networks of bibliographic coupling such as co-citation, collaboration, and co-occurrence analyses; iii) *data visualization*, which allows a better understanding of the given information such as the identification of the principal themes of the topic (Aria & Cuccurullo, 2017). The use of bibliometrix tool, in this review study, has revealed the main trends and evolutions of publications dealing with ML for assessing natural hazards' impacts in coastal areas, providing a detailed overview of the most important themes and dynamics of the field (results reported in *Section 1.3.1*).

1.2.3. Systematic review

A systematic review is a secondary research methodology, which aims to synthesize and evaluate the best available scientific findings of a specific field of research, by clearly answering the question of the study (Cajal et al., 2020). In the frame of this research, the systematic review was conducted to summarize and facilitate the understanding of the publications, related to the investigated topic, retrieved from the Scopus database by applying the keywords query specified in *Section 1.2.1*.

To provide a complete and transparent reporting of the systematic review, the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) statement was adopted (Moher et al., 2009), whose guidelines were devised to facilitate the selections of the relevant papers during the review process. The steps followed for obtaining the papers reported in *Section 1.3.2* were:

i. collection of bibliometric records through Scopus database as described in *Section 1.2.1*;
ii. screening of papers by reading the title and abstract to eliminate the publications not pertinent to the topic of study (e.g., papers dealing with maritime traffic or off-shore incidents, studies not applying ML algorithms);

iii. screening of the papers obtained from the previous point, by reading the methodological section, to remove publications not compliant with the eligibility criteria of the review (e.g., studies forecasting only the evolution of the hazard sources like sea-level rise);

iv. reading of the full text of the papers to keep only the most appropriate and relevant publications for the scope of this review, hereafter addressed as "key papers";

v. classification and discussion of the key papers through a set of devised comparison criteria aiming at simplifying the understanding of the publications, by defining the type and the scope of the applied ML algorithm, the spatial scale of the analysis, the type of evaluated risks (e.g., coastal inundation, coastal erosion), the adopted variables of hazard, exposure and vulnerability, the specific receptors under investigation and, finally, the implementation or not of climate-change, socio-economic and management scenarios.

1.3. Results of the review

1.3.1. Results of the scientometric analysis

The selection of publications (specified in *Section 1.2.1*), related to ML applications for assessing risks caused by natural hazards in coastal environments, identified 651 papers for the 2001-2021 timeframe. These articles were analyzed through the Biblioshiny app (Aria & Cuccurullo, 2017), which allows both to draw some descriptive information related to the sources, the authors, and the type of documents of the selected publications, as well as, to analyze the science mapping in terms of conceptual (relations between words or concepts), intellectual (relations between different nodes to understand the evolution of a topic), and social (relations between authors, institutions and countries) structures characterizing the review topic (Forliano et al., 2021). Specifically, in the following paragraphs, the most comprehensive bibliographic metrics are reported, which are: i) the annual scientific production; ii) the most relevant disciplines; iii) the most productive countries; iv) the collaboration networks for different time slices (i.e., 2001–2006, 2006–2011; 2011–2016; 2016–2021); v) the analysis of the most frequent keywords; and vi) the keyword co-occurrence networks for different time slices (i.e., 2001–2006, 2006–2011; 2011–2016; 2016–2021).

Annual scientific production

The annual scientific production analysis reported in Figure 1 provides a comprehensive understanding of the number of papers, annually published, related to the application of ML methods for assessing natural hazards' impacts in coastal areas. By considering all the set of publications under investigation, for the 2001-2021 timeframe, over the years, there has been a continuously increasing trend (annual growth rate: 21.92%) with the exceptions of some years (i.e., 2008, 2014, and 2019) where the publications slightly decreased from the antecedent year. However, Figure 1 clearly shows how the production started to considerably increase only from 2011 (18 publications), revealing an exponential trend from 2016 (37 publications; 2016-2021

annual growth: 33.69%). In particular, publications of 2019 (53) nearly tripled in 2021 (158). The rising number of publications over the years reflects the increased interest of the scientific community in exploiting ML methods to investigate natural hazards' effects in coastal areas. The interest was also due to some limitations of physical-mathematical models to provide an overall comprehension of the system in a cost-effective way, as well as, from the urgency to formulate mitigation and adaptation solutions against the magnification of climate change-related phenomena.

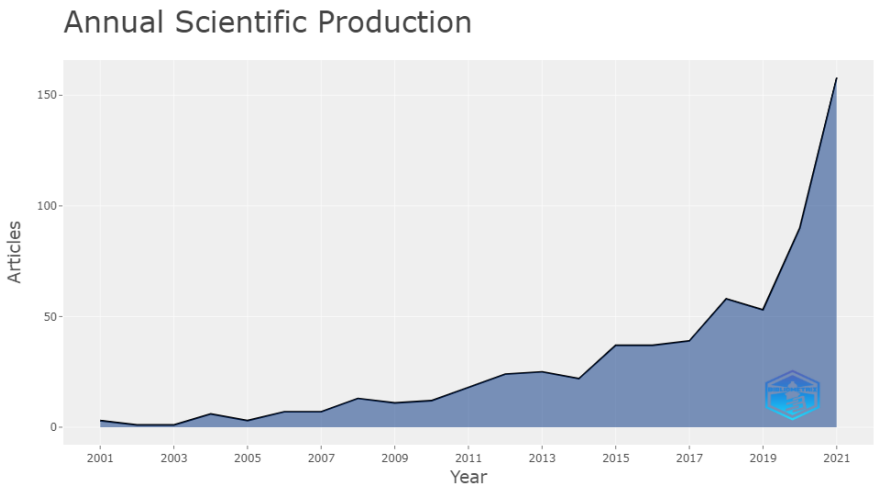


Figure 1: Annual scientific production of the publications dealing with the application of ML methods for assessing natural hazard risks in coastal environments within the 2001-2021 timeframe

Most relevant disciplines

The topic of ML methods applied for assessing natural hazard consequences in coastal areas is very broad and involves a variety of publications focused on specific themes, embracing different disciplines often interrelated with each other. In relation to the 651 publications retrieved from the Scopus database, Figure 2 reports the top ten subjects that have mainly contributed to incrementing the knowledge of the topic. Specifically, Environmental Sciences is the most productive discipline (279 papers), which was expected given its strong relation to the studied topic. It is followed by Earth and Planetary Sciences (212), Engineering (149), and Agricultural and Biological Sciences (143). Detached from the five most productive subjects, there are

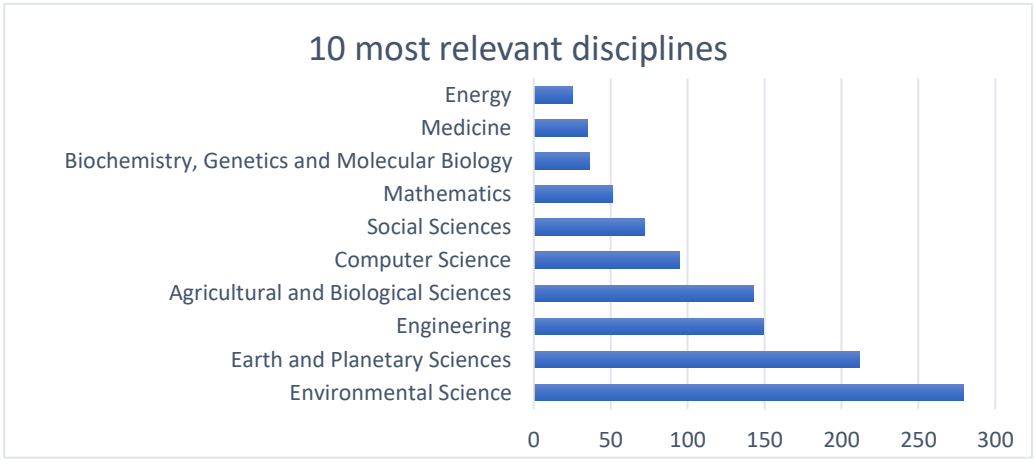


Figure 2: Barchart of the ten most relevant disciplines in the research topic

Computer Science (95), Social Sciences (72), Mathematics (51), Biochemistry, Genetics and Molecular Biology (36), Medicine (35), and Energy (25).

Most productive countries

The analysis of the most productive countries indicates the status of the research topic in different areas of the world. The map in Figure 3 visually represents the countries that have contributed more to the knowledge of the investigated topic; in the map, the shade of blue of a country is proportional to its number of publications (i.e., the darker the blue intensity is, the more productive a country is). In Bibliometrix R package, the publication frequency of the different countries is obtained by summing, for each country, the number of affiliated authors.

In the 2001-2021 timeframe, 77 countries and 589 different institutions contributed to publishing the 651 papers of the analyzed dataset.

As it can be seen from Figure 3, the most productive countries have ample coastal areas in their territory, as expected since the specific topic was filtered for the zones at the land-sea interface. Nevertheless, these countries have different issues related to natural hazards and therefore, with a further investigation, it can be found that specific themes have been developed by the different countries. In particular, the United States is the most productive country with a number of publications (440) which is more than double of that of Cina (185), which is in second place. Specifically, the USA is strongly impacted by extreme events (e.g., tornados, hurricanes) which are intensifying because of climate change and which are putting at stake multiple assets (e.g., infrastructure, houses) and activities (e.g., agriculture, transportation), by generating severe socio-economical damages (Collins et al., 2022). China's main focuses are related to the issue of flooding as a consequence of sea-level rise, a condition that threatens the majority of the population, concentrated in low-lying coastal areas (Yang et al., 2019). In the third place, in terms of productivity, there is Australia (146 papers), where the main theme is related to corals' vulnerability and their tendency to bleach because of the increasing ocean temperature and acidification. The following positions are covered mostly by European countries, including UK (136), Italy (102), Canada (93), Germany (86), Finland (63), Spain (62), and France (61). The presence of many European Mediterranean countries in the first positions (i.e., Italy, Spain, and France) could be explained by an increasing interest in the research topic due to a significant intensification of extreme events in the Mediterranean area. Finally, an important role is covered by north-latitude countries (e.g., Canada, Finland but also Netherland and Norway) where, generally, the management of fishery and aquaculture industries is the objective of the study, due to the risk posed by the modification of the fish stocks in terms of abundance and distribution.

Country	Frequency
Usa	440
China	185
Australia	146
Uk	136
Italy	102
Canada	93
Germany	86
Finland	63
Spain	62
France	61
Netherlands	45
Norway	41
Brazil	37
India	31
Belgium	25
Portugal	24
Sweden	24
Greece	21
New Zealand	21
Japan	20

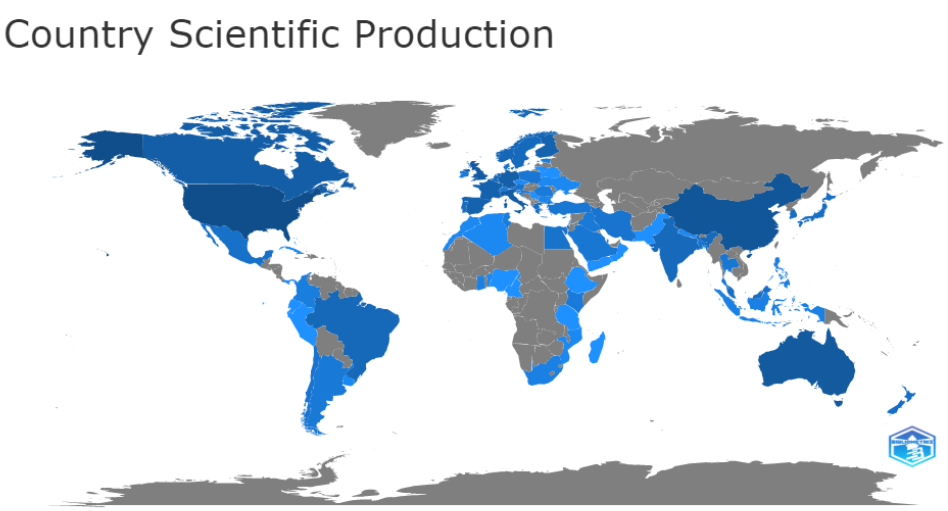


Figure 3: Country scientific production (2001-2021)

Collaboration networks for different time slices (i.e., 2001–2006, 2006–2011; 2011–2016; 2016–2021)

The analysis of the countries' collaboration networks is obtained by aggregating, for each country, the number of authors who are affiliated, for the same paper, with at least one other co-author from a different country. In the maps of Figure 4, the blue color intensity of a country is proportional to the number of its international collaborations (i.e., the darker is the blue, the more collaborations a country performs), whereas the width of the edge, linking two countries, is proportional to the number of papers published in collaboration between the authors of those countries. For the entire analyzed timeframe (2001-2021) USA results to be the most collaborative country, especially with the UK (18 frequency), Canada (17), Australia (15), and Cina (14). The UK is in second place in terms of collaboration, mainly with Cina, Germany, France, and Australia. It is interesting to see how the collaborations between countries have changed through time by dividing the analyzed timeframe into four time slices (i.e., 2001–2006, 2006–2011; 2011–2016; 2016–2021). In the first period (2001–2006; Figure 4a) collaborations were only six, with a maximum frequency of 2 (between USA-Australia and USA-New Zealand), involving just Anglo-Saxon countries except for Italy and Germany. However, in the following years, consistently with the growth of the scientific publications in the field of ML for assessing natural hazards in coastal areas, there has been an increment of the countries involved in collaborations as well as in the publications' frequency, especially with an increase of the Asian countries (precisely, Cina and India) if confronted with the first investigated years (i.e. 2001-2011).

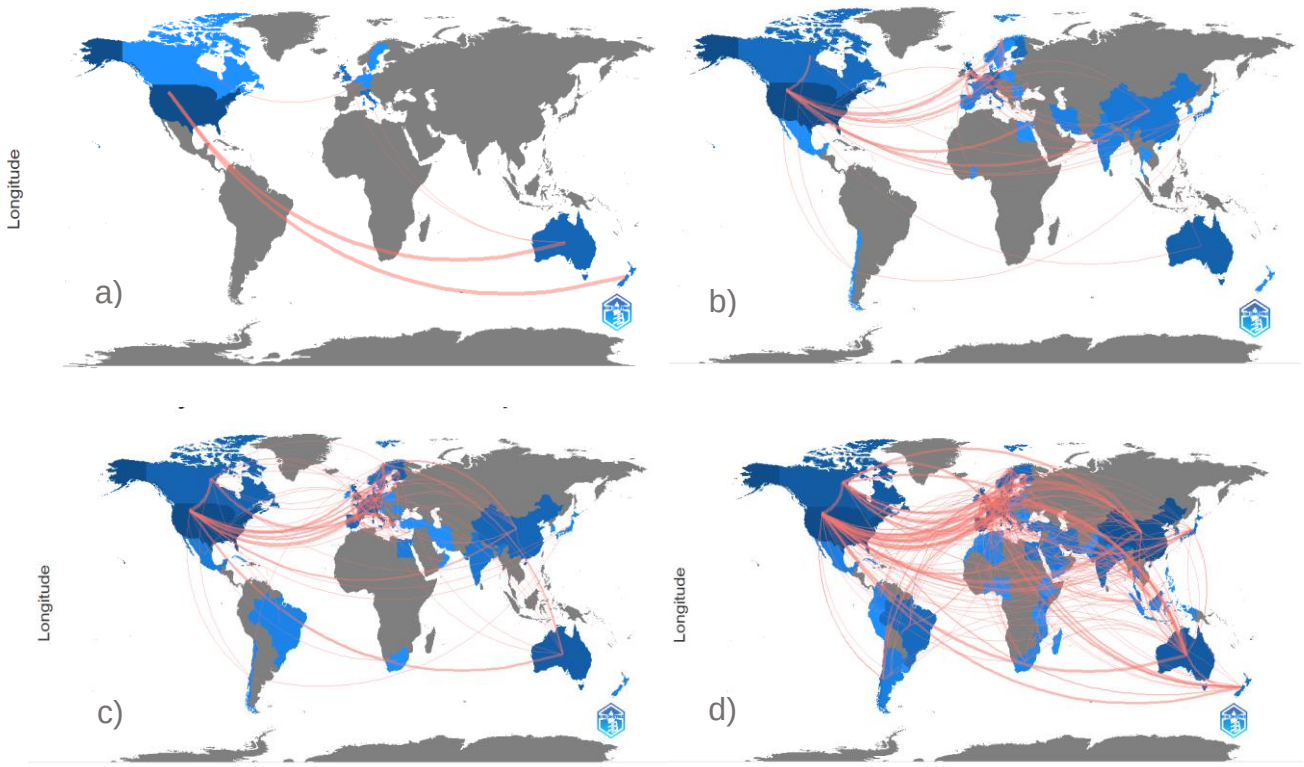


Figure 4: Collaboration maps under four time slices: a) 2001-2006; b) 2006-2011; c) 2011-2016; d) 2016-2021

Analysis of the most frequent keywords

The most relevant keywords, associated with the topic of ML methods for evaluating natural hazard impacts in coastal areas, are here examined. The unit of analysis is the author's keywords, which identifies the three to five keywords selected by the author to summarize its study. The keywords analysis has the potential to reveal the evolution and the trend of the themes associated with the main topic, both for the present and the past (Pesta et al., 2018). Concerning the selected 651, the most frequent author's keywords are reported in a word treeMap (Figure 5).

'Climate change' and 'Machine learning' are the most recurrent keywords, appearing respectively 16% and 9% in the selected publications, this result is also derived from the choice of using these words for the search query applied in the Scopus database (*Section 1.2.1*). However, they are key concepts of the topic under investigation so, their higher frequency reflects a correct selection of the 651 publications. In the following positions, two main ML algorithms are found namely 'bayesian network' (6% frequent) and 'random forest' (3% frequent). Their frequencies increase if synonym words are considered, such as "belief bayesian network" and "classification". These methodologies are representative of the rising widespread of these algorithms for assessing natural hazard impacts in order to find relations among the input variables. In fact, the identification of such relations is not always straightforward and this complexity has hindered the results of traditional physical-mathematical models, requiring alternatives such as ML algorithms. Then, among the most frequent keywords, also the word 'risk' and its declinations (e.g., 'risk-assessment', 'risk analysis') recurs very often and that is indicative of the wide implementation of ML methods for assessing a variety of climate

change-related risks. Concerning the hazards or their triggering sources, 'sea-level rise' is the most frequent (3%), since it is posing under threat the majority of the coastal areas in the world, both directly (i.e., expected sea-level rise due to ice melting and seawater expansion), or indirectly (i.e., sea-level rise associated to storm surges events). Other frequent words hazard-related are 'water quality' and 'eutrophication' since the increasing water temperature, combined with anthropic pressures, is putting at risk the health of the marine environment.

Other frequent keywords are 'remote sensing' (14%) and 'gis' (2%), and that is indicative of the diffusing implementation of satellite data for retrieving information, often in a more effective manner in terms of cost, time, and spatial coverage than traditional methods (e.g., in-situ measures). Finally, in the most 50 frequent keywords, words like 'decision support systems', 'adaptation', and 'resilience' occur, which are revelatory of the several studies, among the pool of 651 papers, dealing with the implementation of ML algorithms for detecting management strategies to adopt for increasing the resilience of coastal areas, or for formulating effective adaptation and mitigation plans to reduce the consequences of natural hazards' effects. Given these results, it must be specified how they could be partly biased since for 118 of the 651 papers Scopus did not report the author's keywords, and so the analysis performed through Bliblioshiny considered only 533 publications.

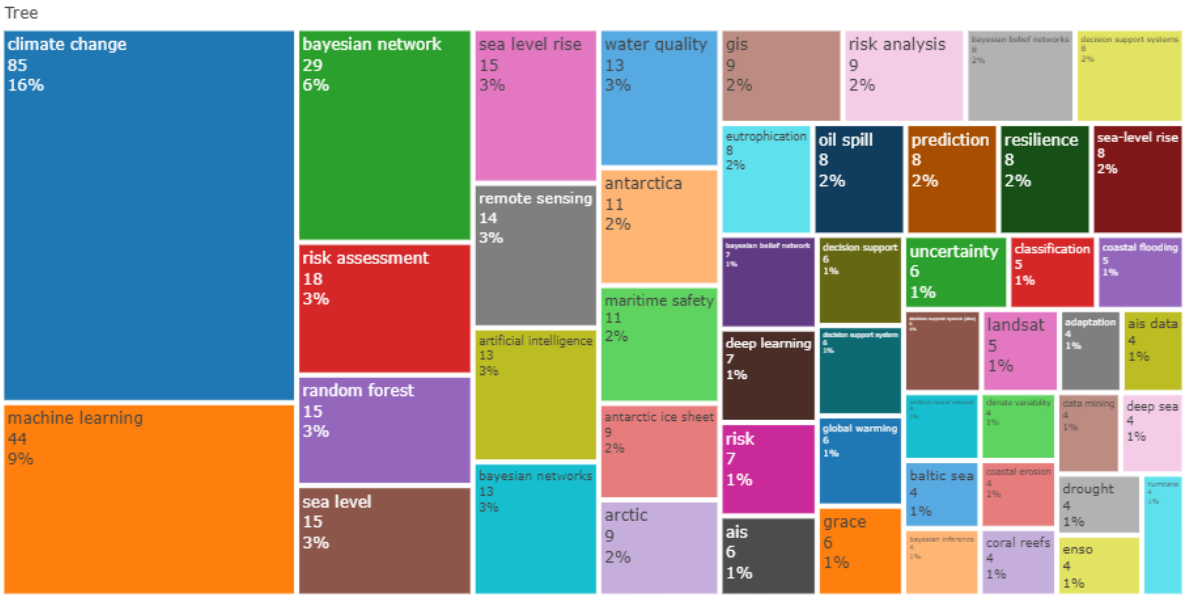


Figure 5: Keywords TreeMap

Keyword co-occurrence networks for different time slices (i.e., 2001–2006, 2006–2011; 2011–2016; 2016–2021)

The analysis of the keywords co-occurrence Network (KCN) is carried out in order to map the knowledge structure of the studied topic (Esfahani et al., 2019), and to inquire how different themes are linked together. Therefore, this survey provides additional information relative to the keywords TreeMap (Figure 5), that is the connection between the different keywords. A KCN is the graphical representation of a co-occurrence

matrix, that visually conveys the frequency with which two keywords appear together in different publications (in the frame of this review, having set the parameter 'minimum edge parameter' equal to 1, only co-occurrent keywords present in at least two publications were selected). In the KCN, each keyword represents a node of the network, whereas the edge (or link) connecting two keywords vehicles the co-occurrence between that pair of keywords (Radhakrishnan et al., 2017). There are some graphical characteristics that guide the interpretation of the KCN:

i) the biggerer is a node, the more the keyword co-occurs with other words;

ii) the distance between two nodes is inversely proportional to the number of times those keywords appear together in different publications (i.e., the more two keywords are closed, the more frequently they co-occurred);

iii) in the network, different colors represent different clusters, which means different themes comprehended in the same research topic;

iv) the centrality of a term/cluster in the network indicates the ability of that term/cluster to influence other clusters in the network and its interdisciplinary nature, conversely, the peripherical position of a term/cluster symbolizes the scarce influence on the network, or the development of that theme separately from the rest of the network.

The analysis of the KCN for different time slices allows to understand how the themes, in the domain of ML methods for assessing natural hazard impacts in coastal areas, have evolved through time. In the first timeslice (2001-2006; ANNEX II; Figure A), no interesting information appears: there is only one cluster dominated by the keyword 'microbial transport', relative to the risks of contamination from maritime transport. In the second time slice (2006-2011; ANNEX II; Figure B) 'climate change' keyword is introduced in the network, with the highest number of co-occurrences with other keywords. However, the climate change cluster, mainly associated with 'gis' and 'DDS', is isolated, without edges with the other two most relevant clusters, one related to 'DDS' in coastal environments especially for 'water quality', and the other related to 'bayesian network' for 'oil spill' detection. During the 2011-2016 timeframe (ANNEX II; Figure C) the number of thematic clusters increases (11) likewise the edges between different keywords. Nevertheless, the two most important clusters which are 'climate change' and 'risk assessment' are still separated one from the other. Specifically, 'climate change' cluster has intra-relations involving mainly 'sea-level rise' and 'invasive species' (threat emerging from the increasing temperature) and shows some level of co-occurrence with the cluster related to 'water quality' and 'algal bloom' (which is, again, associated to the increasing temperature). The second most important cluster sees the 'bayesian network' implemented for 'risk assessment' (like in the second time slice 2006-2011). The third most important group, separated from the others, is related to maritime hazards due to ship traffic, with a strong intra-cluster co-occurrence between 'hazardous chemical', 'chemical spill', and 'bayesian network'. The network changes completely its appearance for the fourth and last time slice (2016-2021; ANNEX II; Figure D), as a consequence of the abrupt

development of the topic and the relative number of scientific publications (Figure 1) in these last years. In this time slice, 'machine learning' enters the network and together with 'climate change' is the most frequent co-occurrent keyword. 'Machine learning' and 'climate change' lead their respective clusters, whose centrality and proximity show how the two themes have become extremely interdisciplinary and interrelated among them. In particular, in the 'climate change' cluster, along with words indicating the common two effects of it (i.e., 'sea level' and 'global warming'), 'resilience' and 'adaptation' appear, which are generally the objectives of the investigated publications. Moreover, in this same cluster, two of the most threatened environments/habitats are found namely 'coral reefs' (jeopardized by the increasing temperatures and ocean acidification) and 'Antarctica' (where the erosion risk of the shoreline, left free from ice, is increasing). Concerning the 'machine learning' cluster, together with 'bayesian networks', another ML algorithm gains traction namely 'random forest'. The contemporary presence of 'remote sensing' and 'sentinel 2' indicates the rising use of satellite data in the ML field, often applied to monitor 'eutrophication' which is posing several risks in coastal areas. Finally, these two main clusters are in relation to the one concerning the theme of 'risk assessment', revealing a general tendency to implement ML algorithms for evaluating the risks posed by climate change. The considerations made for the last time slice are similar to the ones that can be made for the KCN applied to the entire timeframe (2001-2021) (Figure 6), this fact highlights how the knowledge of this review topic has been developed only in these recent years.

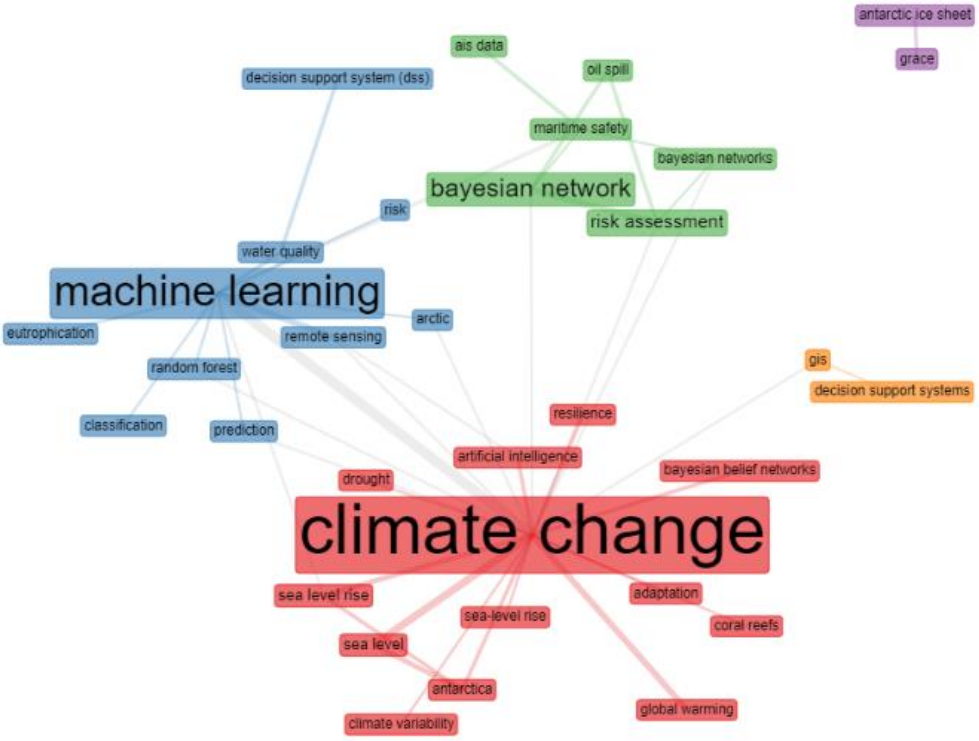


Figure 6: Keywords co-occurrence network for the 2001-2021 timeframe

1.3.2. Results of the systematic review

In this section, the results of the systematic review are discussed in detail. In particular, the application of the PRISMA statement (Moher et al., 2009) on the 651 publications retrieved from the Scopus database (*Section 1.2.1*), through a series of skimming passages (specified in *Section 1.2.3*), has brought to the final selection of 17 publications. These selected publications are specified as "key papers" of the research topic concerning the application of ML methods for assessing risks of natural hazards in coastal environments.

Table 2 summarizes the main characteristics of the 17 key papers in relation to the comparison criteria adopted for guiding the selection, which permitted to clarify i) the scale of analysis of the study and the ii) case study area; iii) the applied ML algorithm and iv) relative aim; v) the type of natural hazard and vi) relative hazard variables; vii) the variables of exposure and vulnerability; viii) the receptors; ix) the type of data; and finally, x) the application of climate change scenarios or xi) management and socio-economic scenarios.

Table 2: Key papers selected from the systematic analysis

Reference	Title	Scale of analysis	Location	Type of ML-method	Model aim	Hazard type	Hazard variables	Exposure/vulnerability variables	Receptors	Type of data (measured, satellite, modeled)	Climate change scenario	Management scenario / Socio-economic scenario
Jäger et al. (2018)	A BAYESIAN NETWORK APPROACH FOR COASTAL RISK ANALYSIS AND DECISION MAKING	Local	North Norfolk (United Kingdom)	Bayesian Network	i) To estimate, for any storm scenario, the percentage of affected receptors, through the prediction of the hazard's impact on the receptors and the relative damages; and ii) to evaluate several DRR measures	Flooding caused by storm surge	Maximum water level; Maximum wave height	Residential damages; Commercial damages; Risk to life; Saltmarsh damages	Residential properties; Commercial properties; People; Saltmarsh	Modeled (2D TELEMAC and SWAN models); socio-economic data;	RCP8.5 (for 2060)	DRR measures: 1) Construction of an extended flood wall; 2) Increasing the height of the flood wall in combination with a movable barrier; 3) Placement of a series of display boards to sensibilize the population regarding the storm surge risks
Plomaritis et al. (2018)	USE OF BAYESIAN NETWORK FOR COASTAL HAZARDS, IMPACT AND DISASTER RISK REDUCTION AT A COASTAL BARRIER (RIO FORMOSA, PORTUGAL)	Local	Faro Becah (Ria Formosa, Portugal)	Bayesian Network	i) To predict the impacts caused by erosion and overwash hazards on infrastructures and houses; and ii) to evaluate several DRR measures	Overwash and erosion due to storm surge	Tide above the mean sea level; Maximum significant wave height; Wave period	Erosion state of houses; Overwash state of houses	Houses and Infrastructures	Modeled (Xbeach)		DRR measures: 1) Beach replenishment; 2) House removal; 3) Improvement of communication channels with residents; 4) Combination of different measures
Sanuy & Jiménez (2021)	PROBABILISTIC CHARACTERIZATION OF COASTAL STORM-INDUCED RISKS USING BAYESIAN NETWORK	Local	Tordera delta (Spain)	Bayesian Network	To assess storm-induced risks (due to erosion and inundation) at a local scale, considering also the location of the receptors	Inundation and erosion due to storm surge	Significant wave height; Wave period; Wave direction; Water level; Event duration	Distance to inner beach limit; Area of the receptor's location	~ 4000 receptors	Historical time series data; modeled (Xbeach)	Scenarios of shoreline retreat due to background erosion (evaluation of the inundation and erosion impacts within +5, +10 and +20 years from the baseline scenario)	
Park & Lee (2020)	PREDICTION OF COASTAL FLOODING RISK UNDER CLIMATE CHANGE IMPACTS IN SOUTH KOREA USING MACHINE LEARNING ALGORITHMS	National	South Korea	K-nearest neighbor; Random forest; Support vector machine	To analyze the future risk of coastal flooding (2030, 2050, and 2080) in order to support decision-making for ICZM	Flooding due to sea-level rise	Mean tide; Daily maximum rainfall	Slope; Urban area; Grassland	Coastal area	Measured; modeled (for future)	RCP 4.5/8.5 (from the 2030s to the 2080s)	

Reference	Title	Scale of analysis	Location	Type of ML-method	Model aim	Hazard type	Hazard variables	Exposure/vulnerability variables	Receptors	Type of data (measured, satellite, modeled)	Climate change scenario	Management scenario / Socio-economic scenario
Maina et al. (2021)	IDENTIFYING GLOBAL AND LOCAL DRIVERS OF CHANGE IN MANGROVE COVER AND THE IMPLICATIONS FOR MANAGEMENT	National	Western Indian Ocean	Random Forest	To investigate the impacts of environmental and human drivers on changes in mangroves, by considering also future climate scenarios	Climate change	CDD (Consecutive dry days); Tx90p (Percentage of days when warm days >90th percentile); Sea level anomaly; Tide; Coastal erosion	Nearshore coastal typology; Land development index; Access to market (human pressure)	Mangroves (proxy for the mangroves' status: NDVI and VCI)	Satellite; modeled	RCP8.5 (2050/2060)	
Jakariya et al. (2020)	ASSESSING CLIMATE-INDUCED AGRICULTURAL VULNERABLE COASTAL COMMUNITIES OF BANGLADESH USING MACHINE LEARNING TECHNIQUES	Local	Bangladesh	Linear regression; Bayesian ridge regression; Regression random forest; Regression XGB algorithm; Extremely randomized tree regression	Identification of significant factors which influence the crop yield vulnerability	Climate change and extreme weather conditions	Humidity; Temperature; Rainfall	Crop disease; Soil quality; Water availability; Crop loss; Availability of rain; Adaptive capacity	Crop yield	Stakeholder questionnaires; historical data		
Zahura et al. (2020)	TRAINING MACHINE LEARNING SURROGATE MODELS FROM A HIGH-FIDELITY PHYSICS-BASED MODEL: APPLICATION FOR REAL-TIME STREET-SCALE FLOOD PREDICTION IN AN URBAN COASTAL COMMUNITY	Local	Norfolk, Virginia (USA)	Random Forest	To replace physics-based models with ML methods for a real-time flood prediction, at a street scale, of the surface water depth on the road	Urban flood	Total, maximum, and cumulative rainfall; Tide level	Elevation; Topographic wetness index; Depth to water index	Roadway	Measured; modeled; crowdsourced data		

Reference	Title	Scale of analysis	Location	Type of ML-method	Model aim	Hazard type	Hazard variables	Exposure/vulnerability variables	Receptors	Type of data (measured, satellite, modeled)	Climate change scenario	Management scenario / Socio-economic scenario
Cai et al. (2018)	MODELING THE DYNAMICS OF COMMUNITY RESILIENCE TO COASTAL HAZARDS USING A BAYESIAN NETWORK	Local	Lower Mississippi river basin (USA)	Bayesian Network (optimized with a Genetic Algorithm and trained with the Expectation-Maximization EM learning algorithm)	To study the interdependencies of 10 resilience variables on the global resilience community (addressed in terms of population change)	Extreme events (e.g., hurricanes); anthropic hazards (e.g., urbanization)	Hazard threat; % flood zone area	Socioeconomic and demographic variables (employment rate, % owner-occupied house units, population density, % housing units built before 1970, % female householder); Damage per capita; Distance to the coastline; % agricultural land	Community resilience (addressed in terms of '% population change')	Measured; data form the literature		
Rohmer et al. (2021)	UNRAVELLING THE IMPORTANCE OF UNCERTAINTIES IN GLOBAL-SCALE COASTAL FLOOD RISK ASSESSMENTS UNDER SEA LEVEL RISE	Global		Random Forest	To evaluate the influence of different variables' uncertainties regarding flood hazard in the determination of two risk metrics namely *expected annual damage* (EAD) and *adaptation costs* (AC) for a coastal dyke	Coastal flooding due to sea-level rise			Coastal dyke	Modeled (from DIVA -Dynamic Interactive Vulnerability Assessment model)	RCP2.6, 4.5, 8.5; Magnitude of the regional sea-level rise; r-largest annual value (rGEV); Subsidence in delta region (SUBS)	Socio-economic development (SSP): 1-5; Global population distribution (POP); Assets-to-GDP ratio (A:GDPr); Logistic depth-damages curves (DF)
Praharaj et al. (2021)	ESTIMATING IMPACTS OF RECURRING FLOODING ON ROADWAY NETWORKS: A NORFOLK, VIRGINIA CASE STUDY	Local	Norfolk, Virginia (USA)	Linear regression classification; Regression trees; Random forest	To determine the impacts of low-intensity recurring flooding on the transportation sector (addressed in terms of "traffic volume")	Pluvial flooding	Hydrological data (Rainfall, flood incidents, tidal level)	Roadway data (n° of lanes, speed limits, per lane capacity); Traffic flow data (Trip counts, speed, time of day, type of day)	Roadway	Crowdsourced data; measured;		
Taramelli et al. (2020)	ASSESSING PO RIVER DELTAIC VULNERABILITY USING EARTH OBSERVATION AND A BAYESIAN BELIEF NETWORK MODEL	Local	Po Delta area (Italy)	Belief Bayesian Network	To investigate the vulnerability of the Po Delta coastal area (in terms of ecological, morphological and social factors) in relation to the future sea-level rise	Sea level rise	Vertical velocity; Wave height; Wave frequency; Wave regime; Sea-level rise	Vulnerability variables (Protected natural area, resilience index); Pathway variables (dune, coverage factor, protective distance, geomorphology, global surface water dynamic, surface water occurrence change intensity, water transition, elevation)	Coastal zone under the RICE area (extended 1000 m inland)	Satellite (Copernicus data)	SLR scenarios (IPCC 2014): up to 2100	

Reference	Title	Scale of analysis	Location	Type of ML-method	Model aim	Hazard type	Hazard variables	Exposure/vulnerability variables	Receptors	Type of data (measured, satellite, modeled)	Climate change scenario	Management scenario / Socio-economic scenario
Ferreira et al. (2019)	EFFECTIVENESS ASSESSMENT OF RISK REDUCTION MEASURES AT COASTAL AREAS USING A DECISION SUPPORT SYSTEM: FINDINGS FROM EMMA STORM	Local	Faro Becah (Ria Formosa, Portugal)	Bayesian Network	i) To determine the potential impacts of storms, in coastal areas, in terms of overwash and erosion hazards on houses and infrastructures; and ii) to assess the performance of DRR measures (evaluated through the effectiveness index)	Overwash and erosion due to storm surges	Wave height; Peak period; Water level	Distance of the house/infrastructure from the coast (to determine damage or potential damage); Morphology (to consider nourishment DDR)	Houses; Infrastructures	Modeled (2D TELEMAC and SWAN models); measured (related to Emma storm)		DRR measures: 1) Beach nourishment including the construction of a circa 45 m wide berm; 2) Removal of the houses placed at the ocean side of Faro Beach; 3) Beach nourishment (1) + house removal (2)
Tolo et al. (2017)	RISK ASSESSMENT OF SPENT NUCLEAR FUEL FACILITIES CONSIDERING CLIMATE CHANGE	Local	Nuclear Power Station of Sizewell B, (East Anglia, United Kingdom)	Bayesian Network	To evaluate the exposure risk of a spent nuclear fuel stored in a facility to flood hazards. Specifically, the BN aims to model the interaction between extreme weather conditions and the technological installation, as well as the propagation of failures within the system itself by considering also possible human error	Flooding due to extreme events and future sea-level rise	Natural variables (Extreme precipitation, sea water level, sea wave period, sea wave height); Failures directly triggered by natural events (drainage system, flooding surrounding, outfall, wave overtopping)	On-site substation, external power grid, emergency hydrant system, emergency power supply, offsite, reservoirs, closure, planned outrage, unplanned outrage, onsite AC, cooling system, spent fuel exposure, emergency supplies, delay in reaction, human error	A spent nuclear fuel stored in a facility	Modeled	SRES A1B (medium-emission scenario) for 2015,2055,2099	Scenarios were evaluated for 4 types of the Power Plant subsystems (On-site Flooding, Cooling System, Spent Fuel Exposure, Flooding in Surroundings). Specifically for the Spent fuel exposure risk, other forced scenarios have been evaluated (Cooling system failed, station flooded, failure drainage system, surroundings flooded, human error, human error and failure drainage system) forced with different what-if scenarios
Tolo et al. (2015)	ENHANCED BAYESIAN NETWORK APPROACH TO SEA WAVE OVERTOPPING HAZARD QUANTIFICATION	Local	Liverpool Bay (Irish Sea)	Enhanced Bayesian Networks (i.e., Bayesian Networks enhanced with System Reliability methods)	To evaluate the expected risk of wave overtopping on a seawall structure, due to future sea-level rise	Wave overtopping hazard (due to SLR)	Wind wave peak period; Swell peak period; Significant wind wave height; Significant swell height; Surge level; Tide level; Sea level rise; Still Water Level	Characteristics of the slope of the seawall (sea wall inclination, slope roughness, crest level, mean permissible discharge); Incident significant height; Incident peak period	Hypothetical seawall	Historical time-series data; modeled (for the future)	B1 (low emission), A1B (medium emission) and A1f1 (high emission) for every decade between 2020 and 2100	

Reference	Title	Scale of analysis	Location	Type of ML-method	Model aim	Hazard type	Hazard variables	Exposure/vulnerability variables	Receptors	Type of data (measured, satellite, modeled)	Climate change scenario	Management scenario / Socio-economic scenario
Xie et al. (2017)	EVACUATION ZONE MODELING UNDER CLIMATE CHANGE: A DATA-DRIVEN METHOD	Local	Manhattan, (NYC, USA)	Random Forest; Classification Tree	To confront different ML methods (e.i., RF and DT) in finding relationships between grid cell-specific features related to evacuation (i.e. geographical, evacuation mobility and demo-economic features) and current hurricane risk zone categories; the most performant ML method is then used to predict future hurricane evacuation zones under sea level rise scenarios	Flooding due to hurricanes	Average elevation above sea level	Evacuation mobility (distance to the nearest evacuation center, distance to the nearest subway station, distance to the nearest bus stop, distance to the nearest expressway); Demo-economic features (total population, population below the poverty level, population with disability); Geographical characteristics (DEM, distance to the coast)	Transportation system resilience (= ability of the transportation systems to maintain a certain level of service under hurricane evacuation scenarios)	Measured; modeled	RCP 4.5 and RCP 8.5 for 2050/2090	Change of demographical-economic features for 2050 /2090 (growth rates of different age groups are considered constant for the analyzed future timeframe)
Bolle et al. (2018)	AN IMPACT-ORIENTED EARLY WARNING AND BAYESIAN-BASED DECISION SUPPORT SYSTEM FOR FLOOD RISKS IN ZEEBRUGGE HARBOUR	Local	Zeebrugge harbour (Bruges, Belgium)	Bayesian Network	i) To assess the damage to infrastructure due to flooding; and ii) to evaluate several DRR measures	Overwash due to storm surge	Peak water level; Significant wave height	Location of: Roads, Houses, Containers, Roll-on/off areas, Buildings, Gas areas, Bluk, Railways	Roads; Furniture; Industries; Cars; Houses; Containers	Modeled		DRR measures: 1) Master Plan for Coastal Safety; 2) Mobile flood-barriers; 3) Moving assets out of risk
Moftakhari et al. (2017)	TRANSLATING UNCERTAIN SEA LEVEL PROJECTIONS INTO INFRASTRUCTURE IMPACTS USING A BAYESIAN FRAMEWORK	Local	Two sites: Orange County and Marin County (California)	Bayesian Model Averaging	The model aims to combine surge predictions (NTRs), from 8 climate models, with tidal predictions and sea-level rise projections to statistically characterize the expected lengths of roads exposed to coastal flooding	Flooding due to sea-level rise and storm surge	Hourly water level		Coastal roads	Historical time series; modeled (storm surge modeled from 8 climate scenarios)	50th and the 95th percentile of the projected mean sea level under the climate scenario RCP4.5 and RCP8.5; prediction for the near future (1998–2063) and mid-future (2018–2083)	

Given the complexity of coastal systems, ML methods have started to be implemented for capturing the relations between the natural and anthropic pressures, aiming to understand the consequences of such interactions and to support decision-makers in developing suitable adaptation and mitigation plans.

Specifically, 4 of the 17 selected papers (Bolle et al., 2018; Ferreira et al., 2019; Jäger et al., 2018; Plomaritis et al., 2018), all developed under the frame of RISC-KIT (Resilience-Increasing Strategies for Coasts toolKIT) project, aimed to estimate damages due to overwash or/and erosion, originated by storm surge events. The general idea behind this group of papers was the use of a Bayesian Network (BN) as a surrogate of more complex physical-mathematical models (which require an elevated computational cost), for translating marine offshore hazards into damages at the coastal receptors, by following the source-pathway-receptor (SPR) concept. In this context, the BN served as a decision support system (DSS), since different disaster risk reduction (DRR) measures were tested in the network to evaluate how they affected the hazard impacts on the receptors. Specifically, the general scheme of the BN proposed in these studies comprehended 5 categories of variables which were 'boundary conditions', 'receptor type', 'hazard', 'impacts' and 'DDR measures'.

Jäger et al. (2018) and Bolle et al. (2018) specified the extreme event boundary conditions in terms of 'peak water level' and 'maximum significant wave height', which were calculated for different storm scenarios through a model train that exploited 2D TELEMAC and SWAN models. Jäger et al. (2018), by studying the surge risks at the North Norfolk coast, considered four receptors (e.i., residential properties, commercial properties, people, and saltmarshes), which were affected by different hazards (e.g., flood depth, wave height) and which experienced diverse consequences in terms of damages and risks. Moreover, three DRR measures were tested, which comprised both physical constructions (e.g., a flood wall, a mobile dam) as well as information campaigns to inform the population about the dangers of storm-surge intensification in the area, hoping to increase the adoption of property-level protections. Bolle et al. (2018) implemented a similar scheme for the Zeebrugge harbour (Belgium), where the receptors were buildings and infrastructures (e.g.,

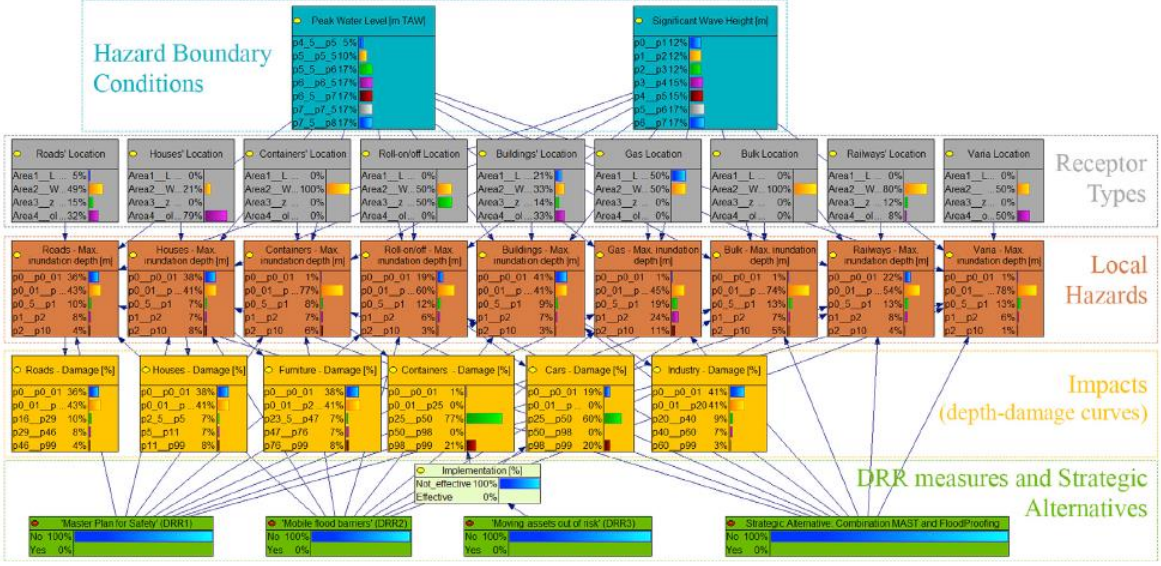


Figure 7: BN developed by Bolle et al. (2018) for the case study of Zeebrugge harbor (Belgium)

railways, roads, gas stations, houses, buildings, containers; Figure 7), whose relative nodes in the BN were divided into four bins, representing the four different districts of the port. In this study, all the receptors were affected only by the 'maximum inundation depth' hazard, and the damage was calculated with a damage-depth curve.

The main structure of the BN proposed in the previous two studies was adopted also by Plomaritis et al. (2018) however, in this case, the BN was trained with 124 synthetic data obtained from the morphodynamic Xbeach model, simulating not only the overwash but also the erosion hazard at Faro beach (Ria Formosa, Portugal). The impacts of the hazards on the receptors (i.e., houses and infrastructures) were evaluated under different DDRs measures, showing how, for the considered storm surge events, beach nourishment was the most efficient DDR measure, followed by house removal. Nevertheless, in case of a larger wave period, overwash to houses could have been efficiently reduced only with a combination of the two measures. In the frame of Plomaritis et al. (2018) research, Ferreira et al. (2019), for the same case study area, trained the BN with data similar to those of Emma storm, which severely hit the Portuguese coast in February-March of 2018. The BN performed well in the estimation of damages if confronted with in-field observations, confirming the erosion hazard to be the most dangerous under storm conditions. In addition, for some of the DDR measures implemented by Plomaritis et al. (2018), Ferreira et al. (2019) calculated the correspondent effectiveness index. In relation to Emma Strom, the index was almost 100% for overwash when the combination of beach nourishment and house removal was applied, whereas for the single erosion hazard the beach nourishment alone had an effectiveness of 54-100%.

Sanuy & Jiménez (2021) further perfectioned the studies presented so far, by training the BN with 179 real storms data, consisting of hourly evolution of wave parameters, decreasing the uncertainty associated to the use of synthetic events. That allowed the construction of a fully probabilistic BN for characterizing the risk of erosion and inundation at the Tordera delta (Spain). For each storm simulation, the resulting hazard map was transformed into a risk value at each one of the almost 4000 receptors. The novelty of the study resigned also in splitting the BN into two parts, one part considering the variability of the forcing conditions (solving the source–consequence relationships) while the other the spatial distribution of the receptors (by linking the receptors' impacts to the receptor's locations). Moreover, the study revealed how under future morphological scenarios of the coast, affected by decadal-scale background erosion, an intensification and extension of both erosion and inundation are expected (inundation risk increased from 2-6% to 8-13%, erosion risk increased from 1-3% to 3-7%).

Among the 17 papers, there was a group that dealt mainly with flooding risk, caused by rapid events like extreme weather events and storm surges, or due to the increasing sea-level rise. Park & Lee (2020) developed a risk probability map for evaluating flooding in the coastal areas of South Korea. Different classification ML methods were confronted, namely k-nearest neighbor (kNN), random forest (RF), and support vector machine (SVM) with radial basis function (RBF), to predict the presence or the absence of

flooding, taking as input the variables of tide, rainfall, elevation, slope, urban area, and grassland. In relation to observed values, kNN performed better (gaining the highest value of ROC-receiver operating characteristic) while rainfall and tides resulted being the most influential predictors. These findings were used to forecast future flooding risk under the RCP 4.5 and RCP 8.5 scenarios, which revealed an increased flooding risk over time, especially on the southern coast. The author suggested that the inclusion of the shoreline changes, as an input variable, could increase the reliability of the forecast.

Then, several publications evaluated flood risks related to specific receptors (Moftakhari et al., 2017; Praharaj et al., 2021; Zahura et al., 2020; Xie et al., 2017; Tolo et al., 2017; Tolo et al., 2015), which were mainly related to the transportation sector. This is probably due to the importance of this sector in driving national and local economies, however, being extremely vulnerable to natural disasters and requiring elevated costs to be repaired, studying the causes of damages is fundamental to prevent them.

In particular, Moftakhari et al. (2017) adopted a Bayesian Model Averaging to weight an ensemble of eight climate models for storm prediction which, combined with predicted astronomical tides and mean sea level rise projections, allowed to forecast the roadway flooding under RCP 4.5 and RCP 8.5 for near-future (1998-2063) and mid-future (2018-2018) at two Californian sites: Orange County and Marin County. The length of road subjected to flooding, for each analyzed scenario, was estimated by considering the amount of time that the total water level TWL (sum of astronomical tide height, mean sea level, and surge component) exceeded the mean sea water level MSWL. The results revealed that if no adaptation measures were put in place, the risks intensified over time.

Both Praharaj et al. (2021) and Zahura et al. (2020) implemented a RF to forecast, spatially and temporally, real-time recurring flooding (addressed as "nuisance flooding"), caused by rainfall and tides, for the Norfolk town (Virginia, USA), with empirical data. Precisely, Praharaj et al. (2021) used agency-provided data for evaluating the impact, on the traffic volume, of daily road floodings on a city scale, while, for a more localized evaluation of the same impact, real-time data from a crowdsourced database were used. For both the analyses, the general structure of the model (Figure 8) was to use input variables related to roadway characteristics (number of lanes, speed limit, and the lane capacity), flow traffic data (trip counts, speed, time of day and type of day) and hydrological data (rainfall, tide level, flood incident data) to estimate the effects on traffic volume. In

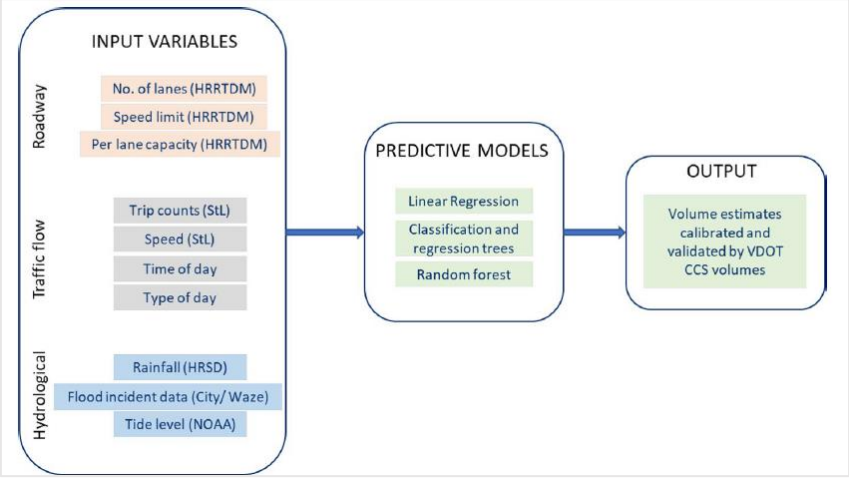


Figure 8: Traffic volume prediction process developed by the study of Praharaj et al. (2021)

predicting the traffic volume, RF performed better among the other tested algorithm (i.e., linear regression and decision tree), having the lowest value of RMSE (Root-mean-square deviation) and NRMSE (normalized RMSE). Moreover, the comparison of two RFs, one considering the hydrological data and one not considering them, revealed how the first had a better predictive capacity, emphasizing the importance of the hydrological data in the prediction. The results showed how, during flood events, freeways saw a decreased traffic, whereas principal arterials manifested an increased traffic volume.

Zahura et al. (2020) tested the efficiency of RF to serve as a surrogate of the TUFLOW model, to reduce the computational cost for forecasting real-time flood prediction at street level. 20 storm events were used for the RF, trained to find relations between environmental (i.e., tides, rainfall) and topographic features with hourly water depth simulated by the TUFLOW model. Specifically, two RF were trained, one considering only the data of the road segments most prone to flood risk, the other one considering all the roads of Norflok city, showing a better prediction for the first case.

Indicators regarding the transportation sector can be integrated to predict the future evacuation zones in Manhattan city under different climate change scenarios (Xie et al., 2017). Specifically, the study of Xie et al., (2017) firstly confronted RF and DT (decision tree) to predict the evacuation zone categories due to hurricane risk over the baseline scenario, by considering geographic features (distance to the coast, average elevation above sea level), evacuation mobility (distance to the nearest evacuation center, distance to the nearest subway station, distance to the nearest bus stop, distance to the nearest expressway) and demo-economic features (total population, population below the poverty level, and population with disability). Then, the RF, which performed better in terms of accuracy and Kappa statistics, was used to predict how the evacuation zone categories will change under future sea-level rise scenarios for RCP 8.5 and RCP4.5 in the 2050s and 2090s. The results displayed a widespread increase of the zones at the higher risk. Therefore, this kind of application can help to better design evacuation planning, as well as to improve the resilience of the transportation system (i.e., maintain the service despite the intensification of hazards).

In the retrieved pool of papers, other types of receptors, affected by flooding, were investigated. Tolo et al. (2015) proposed a framework for predicting the level of overwash (Figure 9), due to sea-level rise, on seawall defenses, by adopting an Enhanced Bayesian Networks (EBNs) methodology, which integrated a Bayesian Networks (BNs) with a Structural Reliability method.

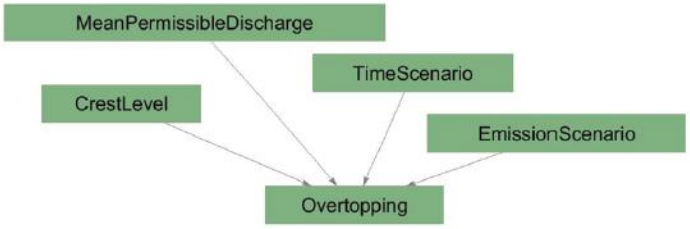


Figure 9: Overview of the reduced BN proposed by Tolo et al. (2015) for assessing the overwash hazard over a hypothetical seawall

The same author Tolo et al. (2017) elaborated a BN model to assess the risk exposure of a spent nuclear fuel, stored in a facility pond subject to flood hazard, an example of Natech risk (natural technological disaster), whose interactions are still scarcely studied by the scientific community. The model, aiming to evaluate the

overall risk failure of the nuclear power station of Sizewell B in East Anglia (United Kingdom), was composed of 37 nodes, divided into three interacting modules: the natural-technological module, to model the effects of natural events on the nuclear facility and its surroundings; the human operators module, to analyze the human error; and the technological interface module to consider the effects that natural events could have on the cooling system or on the emergency one.

The assessment of future coastal flooding is fundamental given the high number of assets present in coastal regions, but there is the need to understand also the uncertainty of these predictions, in order to support efficient adaptation plans and avoid maladaptations. Rohmer et al. (2021) investigated the uncertainty of two risk metrics, namely expected annual damage (EAD) and adaptation cost (AC) of a coastal dyke, in relation to flooding risk. In the research, the uncertainty of these two metrics, for the years 2020-2100, was obtained through different combinations of the variables adopted in the DIVA model, which in the study was replaced by a regression RF, to decrease the computational cost. Then a variance-based global sensitivity analysis was performed to determine the variables that mostly contributed to the uncertainty of the two metrics, exhibiting how, for the long-term scenario (>2050), the main uncertainties of the two metrics were due to RCP-SSP uncertainty.

To adopt suitable management strategies, for increasing the resilience of a community, it is necessary, not only to identify the uncertainties of a prediction, but also to understand the influence that different stressors have on targeted receptors. Maina et al. (2021) evaluated the exposure of mangroves, which provide multiple natural and social-economical ecosystem services, but which are extremely endangered to different human and natural drivers in the Western Indian Ocean region. The author addressed the status of mangroves in terms of two indexes, namely NDVI (Normalized Difference Vegetation Index) and VCI (Vegetation Condition Index), and for both of them, through a RF, he calculated their relations with 11 human and natural variables, from satellite and spatial databases. Future scenarios analysis revealed how, under the RCP 8.5, by 2050, mangroves resulted to be extremely compromised, especially because of sea-level rise and drought.

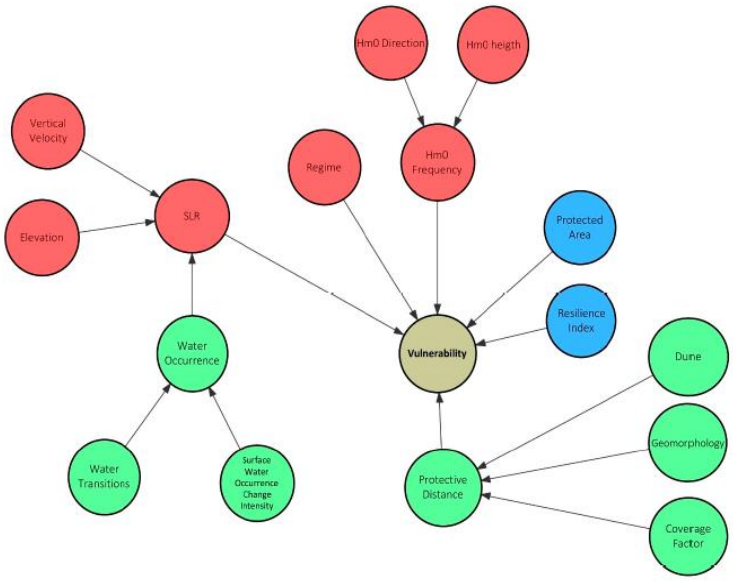


Figure 10: BN developed by Taramelli et al. (2020) for the case study of the Po River Delta

The studies concerning the exposure and the adaptation capacity of a system are fundamental for investigating the relative vulnerability. Taramelli et al. (2020) evaluated the vulnerability of the Po River Delta under different sea-level

rise (SLR) scenarios, by considering the interaction of natural and human variables. He adopted a Belief Bayesian Network where the nodes were divided into four categories (Figure 10): sources (driver factors), pathway (land cover factors), receptors (land use factors which are resilience index and protected natural areas), and consequences (vulnerability), by integrating the results in a GIS environment. The results indicated that under the worst SLR scenario 35% of the study area was in a vulnerable status, where the spatial heterogeneities of the case study were mainly related to the land use and land cover variables. A sensitivity analysis demonstrated how the wave regime was the variable mostly influencing the vulnerability. Jakariya et al. (2020) assessed the vulnerability of three coastal districts of Bangladesh in relation to the agricultural practice, which is the principal source of revenue in the area. In particular, to find the variables (retrieved through interviews with the locals) that mainly contributed to the crop yield vulnerability, referenced in terms of Vulnerability Livelihood Index for agriculture, different ML models were tested. The Bayesian ridge regression algorithm performed better (highest R^2) and was used for the development of an app in which the farmers could log-in and choose, among the 9 most important selected vulnerability factors, those requiring immediate intervention from the government.

All these studies were important to acquire information aiming at fostering the resilience of coastal communities. Cai et al. (2018) explored the interactions between the variables related to community resilience, to determine the overall disaster resilience against natural hazards. The study focused on the Lower Mississippi river basin, an area subjected both to intense land-use change, as well as, severely impacted by hurricanes. A BN, optimized with a genetic algorithm and trained with the expectation-maximization EM learning algorithm, considered different categories of variables related to the hazard threat, the regional socio-economy, the environmental characteristics, and disaster damage, to understand their effect on the overall community resilience, specified in terms of population change. Finally, through the Junction tree algorithm, a type of belief updating, it was investigated how the key variables influenced independently the population change, identifying how, for example, distance to coastline or % of employment rate were inversely proportional to the decrease of population change, whereas increasing hazard threat or the per capita damages had the opposite effect.

To summarize, the 17 selected key papers showed high heterogeneity in the investigated themes, whose main characteristics are statistically represented in Figure 11. The systematic review revealed that, up to now, the core area of the research devoted to the application of ML methods for assessing natural hazards in coastal communities is focused on estimating flooding risks. Important efforts are made to understand how natural and human variables interact together, to evaluate, in the end, the risks on different receptors. In this regard, several studies praised ML methods for overcoming some of the limits of traditional models in finding relations among the aforementioned variables, since they do not require the knowledge of exact equations linking the variables themselves.

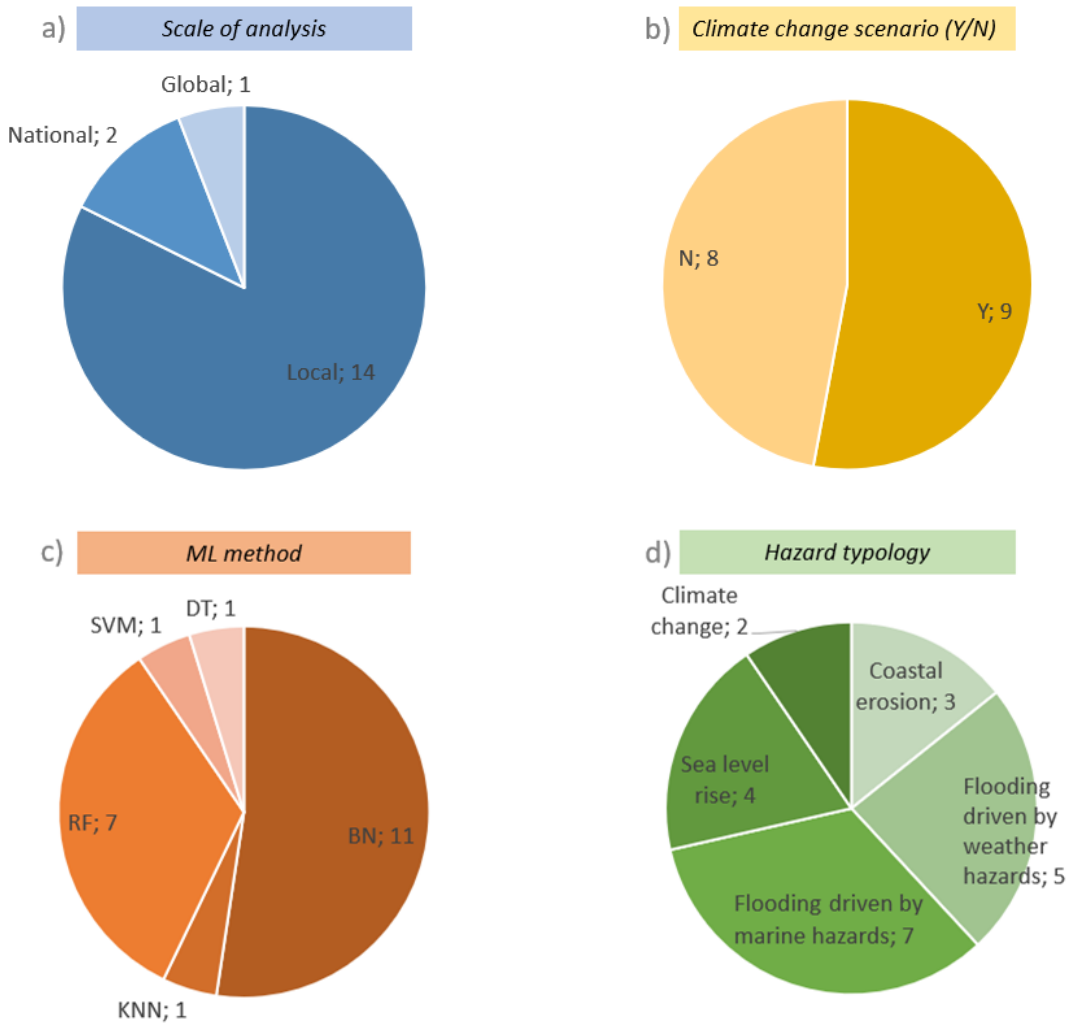


Figure 11: Main statistics of the 17 key papers: a) scale of analysis; b) consideration of climate change scenarios; c) type of ML method; d) hazard typology

That represents a step forward in understanding the dynamics of coastal areas, which are extremely complex socio-ecological systems.

Accordingly, most of the selected key papers studied these kinds of relations, mainly investigated through the implementation of two ML algorithms: Bayesian Networks (present in 11 studies) and Random Forest (present in 7 studies), both applied for their ability to process different data types (e.g., continuous, ordinal, categorical, Boolean) retrieved from various sources (e.g., in-situ observations, models, satellites).

Moreover, these algorithms were often used as surrogates to replace complex physical-mathematical models (which are highly demanding in terms of computational cost), proving, in the end, how the combination of conventional models and ML methods can efficiently give accurate results, which is a pivotal aspect in risk prediction.

Half of the selected studies evaluated the coastal risks under future climate change scenarios (Figure 11b). Specifically, it has been demonstrated that, if no mitigation and adaptation measures will be implemented in

the near future, the risk will increase due to the expected intensification of natural hazards. On the other hand, the testing of a variety of DDR measures against current and future climate risk has revealed how some management strategies can effectively reduce the exposure and vulnerability of the receptors.

Finally, the investigated studies outlined their prevalent development at the local level (14 studies), rather than at higher scales (i.e., national, global). This aspect is due to the fact that hazards' effects on the land part of coastal areas depend on geomorphological characteristics of the territory, as well as on the type of exposed assets, which vary greatly from one place to another. Therefore, a local scale analysis is necessary for designing effective adaptation plans.

SECTION B: Data analysis process to assess the factors influencing the damage occurrences in the Veneto coastal municipalities

2. Characterization of the case study area

2.1. Interreg IT-HR AdriaClim Project

AdriaClim (*Climate change information, monitoring, and management tools for adaptation strategies in Adriatic coastal areas*)[1] is a European research project started in 2020 and funded by the Italy-Croatia Interreg Cooperation Programme 2014-2020 under the EU Strategy for the Adriatic Ionian Region (EUSAIR).

AdriaClim is devoted to supporting the development of science-based climate change adaptation plans, for increasing climate resilience in the cooperation area, by turning potential threats into economic opportunities. This aim will be reached by designing mitigation strategies based on high resolution, more accurate, and reliable climate information for the coastal and marine areas, with particular attention to the economic sectors and the ecosystem services of the Adriatic region.

In brief, the focuses of AdriaClim project are:

- The building of an up-to-date harmonized base knowledge of meteorological and oceanographical information, acquired through newly implemented and more accurate observing and modeling systems for the Adriatic Sea;
- The elaboration of future climate scenarios and methodologies to assess climate change-related impacts, vulnerabilities, and risks. This step allows the development of maps and indexes for the nine pilot case studies in relation to the blue economy (aquaculture, tourism), the marine ecosystems services and Marine Protected Areas (MPA), the coastal towns, and the ports;
- The design of adaptation plans at different spatial scales (i.e., local and regional) to support coastal authorities and stakeholders in the implementation of suitable management strategies against the climate change threat.

Furthermore, the project promotes the cooperation between regional actors, by contributing to increasing both the commitment in planning adaptation strategies and the need to improve climate policies.

This thesis was carried out within the AdriaClim project by implementing a data analysis process for assessing the factors which influenced the damage occurrences, caused by extreme weather events, in the AdriaClim pilot case of the coastal municipalities of the Veneto region (Italy). The study has to be considered as a preliminary analysis to support the design of ML-methods capable of predicting damages, which was developed in collaboration with the Euro-Mediterranean Centre on Climate Change (CMCC)[2].

[1] https://www.italy-croatia.eu/web/adriaclim
[2] https://www.cmcc.it/it

2.2. Case study area

Veneto region is located in the northeastern part of Italy with a surface area of 18378 km^2 and a total perimeter of 1104 km. From the northeast direction, anti-clockwise, Veneto is bordered by Friuli-Venezia Giulia, Austria, Trentino-Alto Adige, Lombardia, Emilia Romagna, and finally by the Adriatic Sea in the south-east direction. Specifically, the 169 km of the Veneto coastline (Ruol et al., 2016) overlook the North Adriatic Sea sub-basin, which is the northernmost region of the Mediterranean Sea (Cushman-Rosin, 2001).

The Veneto coastal area, case study of this thesis, belongs to the provinces of Venice and Rovigo, comprehending eleven municipalities which are respectively seven for the province of Venice (municipality of San Michele al Tagliamento, Caorle, Eraclea, Jesolo, Cavallino-Treporti, Venice, and Chioggia), and four for the province of Rovigo (municipalities of Rosolina, Porto Viro, Porto Tolle and Ariano nel Polesine) (Figure 12). Comprehensively, the coastal municipalities cover a surface of 1573,94 km^2 with a population that counts 412735 inhabitants (ISTAT, 2021).

Figure 12: Case study area: the coastal municipalities of the Veneto Region

Climatologically, according to the Köeppen classification, the littoral of the Veneto region belongs to the sub-continental temperate zone (Barbi et al., 2013), showing the typical mesoclimate of the plain, characterized by a certain degree of continentality with moderately rigid winters and warm summers. This feature derives

from its status of transitional region between the continental Central Europe climate and the Mediterranean one (Lionello et al., 2012). In the coastal areas, the annual mean temperatures are around 14°C, higher than the average 13°C of the internal zones (Barbi et al., 2013). Precipitations are quite homogeneously distributed throughout the year, with an average annual value of rainfall between 800 and 1000 mm, and winters that are slightly dry in comparison to the other seasons. Anyhow, the coastal area, compared to the internal one, is characterized by fewer rain days, lower rainfall accumulations, yet more days with heavy precipitation (Barbi et al., 2012).

From a geomorphological point of view, the low-lying coast, fragmented by the presence of seven river mouths (from north to south: Tagliamento, Livenza, Piave, Sile, Brenta, Adige, Po), presents gentle-slope and sandy beaches resulting from alluvial plain coasts that evolved during the Holocene in lagoons (i.e., Caorle lagoon, Venice lagoon, lagoons of the Po River Delta), barrier beaches, deltas, and spits. Moreover, due to the combined effect of available grain size (fine sand) and the onshore wind regime, coastal dunes are favorable to be formed (Bezzi et al., 2018).

Additionally, by considering the morphological characteristics, the sedimentary shore can be subdivided into a northern, central, and southern trait. The northern trait is delimited northward by the Tagliamento river and southward by the Lido inlet; it comprises straight littoral coasts, where the longshore transport has a south-westerly direction, increasing progressively from ca. 38,700 m^3/y to 99,100 m^3/y (Ruol et al., 2018). The central trait is associated with the Venice littorals (Lido and Pellestrina) with sandy barriers and barrier islands; it is a convergent site with quite null net longshore sediment transport (Bezzi et al., 2018). The southern trait comprehends the Po Delta system, the largest wetland area of Italy, covering an area of 610 km^2 and 60 km of coast, which extends from the Porto Caleri inlet to the mouth of the River Po di Goro (Ruol et al., 2018), comprising several river outlets and salt marshes (Torresan et al., 2008; Regione del Veneto, 2012).

The natural evolution of the coast has been modified since Roman times to allow human settlement. However, starting from 1950, the coastal area has been subjected to abrupt urbanization and anthropic pressure, a condition that, over time, has brought to a considerable change in land use. In fact, the coastal build-up area showed a progressive increase in the years 1990-2000 and 2006 with a buffer area from the coastline which incremented from 1 km to 10 km. From 1988 to 2012, 11 km of the coast were altered to make space for industrial and touristic buildings. Of the 169 km of the Veneto coastline, the 36% (61 km) has been irreversibly transformed for urban and infrastructural uses, specifically, 4 km are occupied by infrastructure works, 24 km encompass urban landscapes of high density, and 33 km urban landscapes of medium density. The remaining 109 km can be considered still “preserved”, in detail, 49 km are devoted to agriculture and 60 km have conserved their natural status, mainly because they cover lagoonal and fluvial estuary areas, which are difficult to urbanize (Legambiente, 2012).

The cementification of the coast, together with an intensive water withdrawal from the rivers for agricultural and industrial purposes, the practice of gravel and sand excavation along the riverbeds (i.e., Piave and Po River; Ferretti et al., 2003), the presence of several dams which intercept the riverine solid load, and inappropriate dune protections (Bezzi et al., 2018), have decreased the sedimentary budget for beach and dune accretion. As consequence, coastal erosion is posing a serious issue along the Veneto shoreline. The erosion phenomenon was denounced already in 1970 by a monitoring study of *Studio della Commissione De Marchi* (*Rapporto spiagge 2021,* 2021) where 20 km of the littoral (15% of the total beaches) were reported to be under erosion threat. However, in recent times, there has been a period of recovery: if between 1960-1994 the coastal surface of Veneto was for 17.9 km^2 in retreatment and for 6.6 km^2 in accretion, between 1994-2012 the situation changed, with a coastal area in retreatment for 1.9 km^2 and in accretion for 2.9 km^2. This improvement was possible thanks to the beach nourishment interventions accomplished between 1997-2011, which brought 7.3 Mmc of sand in the area (MATTM, 2017). Specifically, beach nourishment of the Veneto coast, between 2003-2015, constituted the 25% (4.8 Mmc of sand) of the national interventions. To protect coasts from erosion, Veneto region invested 60 million euros between 2014-2018 and 25 million euros in 2019 for the implementation of structural works, preferring, over time, soft defenses such as foredune restoration (e.g. ReDune Project) or beach nourishment (Bezzi et al., 2018) instead of coastal hard defenses (e.g., revetments, seawalls, groins), solutions adopted especially in the 1960s-1970s. Regardless of these management improvements, according to the *National Guidelines on coastal erosion* (MATTM, 2018), in the period 2007-2012, 52 km (37% of the total) of the 169 km of the regional coast, were under erosion process, with an estimated annual loss of sandy shores of 870.000 m^2, a value that increases under storm surge conditions.

The coastal erosion phenomenon, already accelerated by human pressures, is aggravated by the sea level rise, which has always been a problem for the area (e.g., Venice lagoon saw an increase of 30 cm in the water level in just 129 years). In fact, observations related to the sea-level rise variation of the Northern Adriatic Sea present higher relative sea-level rates (i.e., from 1.2 mm/year in Trieste to 2.5 mm/year in Venice) compared to the average rates of the other regions of the Mediterranean Sea (ranging from 1.1 to 1.3 mm/year)(Gallina et al., 2019). In particular, in this zone, the relative sea-level rise is due to the combined effects of natural eustacy and subsidence (Camuffo, 2021). Precisely, *eustacy*, which is a natural phenomenon where the sea level increases because of the change in the ocean water volume (due to ice melting, thermal expansion of the water, or change in the ocean floor consequent to tectonic activity), has been recorded in the Adriatic Sea since 1890 and, in the last century, it accounted for 10 cm to the sea level increment. *Subsidence* is related to the downward vertical movement of the bottom level, which along the Veneto coast has both natural and human causes (Cavalieri, 2021). Particularly, the natural or geological subsidence is linked to the type of substrate characterizing the area, mainly made of alluvial and soft soils (e.g., sand, gravel, and silt), which through time get compacted and originate the sinking of the surface (between 0.7 and 0.9

mm per year). On the contrary, the human-driven subsidence is related to the exploitation of artesian aquifers for agricultural and industrial water supply (Brambati et al., 2003), and to the extraction of gas-bearing water that started in 1930 with the development of the Marghera industrial site (Madricardo et al., 2019; Gatto & Carbognin, 1981) and continued from 1938 to 1961 in the Po River Delta area (Fabris, 2019). Considering the subsidence rate affecting the Veneto coastal municipalities, for the period 2009-2020, the values varied greatly along the coast. However, except for Venice, which had 19.3% of the municipal area affected by a subsidence rate higher than 2 mm/y, all the others registered values above the 30%, with Cavallino-Treporti having 91% of the area subjected to downward subsidence, followed by Eraclea and Jesolo (with values around 60%). The most endangered area is that of the Po Delta system where subsidence has reached peaks of 3-5 mm/year (Ruol et al., 2018), for these reasons, polders, artificially enclosed by embankments, have been built.

Coming to the natural and socio-economic resources of the Veneto coastal areas, they present an invaluable capital. Specifically, these zones comprehend several natural protected areas, regional parks and reserves (e.g., Bocche di Po, Valle Averto, Delta Po regional park, Bosco Nordio), and areas included in the European ecological network Natura 2000 specified as Sites of Community Importance (SCI) and Special Protection Areas (SPA), which are are respectively 14 and 9 (Regione del Veneto, 2012; Ruol et al., 2016).
In relation to the socio-economic capital, the main activities are related to maritime traffic (Venice harbor is one of the first in Italy for the amount of trade and passenger traffic), fisheries, aquaculture, agriculture, industrial activities (Porto Marghera zone), offshore activities and tourism (Torresan et al., 2012). The study of Modica et al. (2017) identifies that, in the coastal municipalities, the primary sector (i.e., agriculture, forestry, and fishing) covers an average share of 7.42% of the total regional employment, against the 0.59% of the non-coastal areas. The tertiary sector, comprehending mainly tourism, is very important: it is particularly associated to the city of Venice, which before the COVID-19 pandemic counted alone more than 25 million visitors per year (Madricardo et al., 2019), as well as to beach destinations, which are chosen for the high water quality (Rizzi et al., 2016). According to the estimates of Unionmare Veneto, the recreational-touristic sector of the coastal areas generates 20 billion euros every year.

From this overall picture, the Veneto coastal area, for its characteristics, can be recognized as an extreme-city-territory (Aerts et al., 2018), subjected to multiple natural and anthropic pressures. However, this complex system is now threatened by several natural hazards which are intensified by climate change.
Precisely, in relation to climate change in Veneto, from 1993 to 2020, the average temperature rose by +0.55°C per decade, a higher value compared to the global trend, with summer and autumn seasons recording the highest increment of +0.7 °C. Rising temperature determines, on one hand, the widespread of intense rainfalls with strong wind gusts, flooding, and storm surge, on the other, the magnification of heatwaves which create health risks for the population and droughty conditions. The growing number of

tropical nights (days with a minimum temperature higher than 20°C) is of +5 days for decade, and days with rainfall higher than 20 mm have increased by 10% for decade (Regione del Veneto, 2021).

Up to now, the most evident impacts of climate change are related to the frequent and intense manifestations of extreme weather, which have disruptive consequences in terms of damages and losses (Figure 13; Figure 14; Figure 15).

Figure 13: Damages to the beaches of Bibione caused by the storm surge event on the 1st of November 2021. Source: Il Gazzettino (www.ilgazzettino.it)

Figure 14: Damages to the sandy shores of Cortellazzo beach (Jesolo), caused by the storm surge event on the 5th and 6th of December 2020. Source: Il Gazzettino (www.ilgazzettino.it)

Figure 15: Coastal flooding of the Bibione beaches after the storm event on the 13th of November 2016. Source: Il Gazzettino (www.ilgazzettino.it)

In addition, these extreme events aggravate the coastal inundation phenomenon, which is one of the principal natural hazards affecting the case study area. In fact, Veneto's coastal shoreline, being part of the Northern Adriatic Sea, has always been one of the Mediterranean areas more vulnerable to inundation due to the presence of: large river mouths; frequent storm surges, driven by meteorological forcings (i.e., pressure gradients and wind velocities) (Mel et al., 2014); relatively large tides compared to the rest of the Mediterranean (average of 1 m of tidal range), due to the semi-enclosed nature of the North Adriatic Sea; and presence of seiches (Gallina et al., 2019). Among the aforementioned factors, which trigger coastal flooding, storm surge is the main one, showing, in the Western North Adriatic, larger heights than other Mediterranean subbasins (Međugorac et al., 2018), with values higher than 1 m during exceptional cases. Generally, the meteorological conditions that cause storm surge flooding, which is intensifying in recent years, are associated with low atmospheric pressure, combined with a strong southeast Sirocco wind (Cushman-Rosin, 2001). Moreover, these conditions can be worsened when combined with low pressure on the upper Adriatic there is a contemporary high-pressure center on the lower Adriatic, and when, simultaneously with the Scirocco wind, there is a strong blowing north-east Bora wind: this configuration creates a convergence of wind-induced marine currents towards the Western Adriatic coast (Di Nunno et al., 2021). These occurrences of extreme storm surge levels, along the North Adriatic, are expected to increase in the upcoming years, with events having a return period of 1000 years that could have a surge level higher than 3.5 m (Rizzi et al., 2017).

Despite rapid extreme events are the class of natural phenomena which is posing higher threats to the region, another relevant natural hazard is the sea level rise, which, although “naturally” determined by the territorial characteristics (e.g., subsidence), is magnified by climate change, putting the entire area at risk of disappearing (Cavalieri, 2021). According to the projections of the 2018 ENEA report, related to the future inundation of the Mediterranean area (ENEA, 2018), the coast of the Pianura Padana-Veneta (comprising Friuli-Venezia Giulia, Veneto, Emilia-Romagna) in 2100 will have 5451 km^2 affected by permanent inundation risk, and for what concerns Venice harbor, the increase of sea-level rise in 2100 is expected to reach +1,064 m and +2,064 m under storm surges conditions.

Lastly, in recent years, a rising problem is that of saltwater intrusion, related to the Adriatic saltwater reaching the mainland because of the drop of the river water level (e.g. Po River). The issue is aggravated due to the intensification of the before-mentioned extreme events like droughts and downpours, which combined with sea-level rise and land subsidence generate a severe condition that is putting at risk several agricultural and industrial activities, creating serious environmental and socio-economic impacts (Da Lio et al., 2015).

2.3. Data collection for the case study area

In order to implement a ML method capable of predicting the occurrence of damages and selecting the most important factors contributing to the damage itself, a series of heterogeneous input variables, related to hazard, exposure, and vulnerability indicators are needed.

In the frame of the Interreg IT-HR AdriaClim project, the collected data can be gathered into four classes of indicators namely, atmospheric and oceanographic indicators (hazard indicators), territorial indicators, and lastly, damage indicators. Specifically, the metadata of the collected indicators are summarized in Table 3, by reporting information concerning the: i) indicator class; ii) indicator's macro-category; iii) data source; iv) data spatial domain and v) spatial resolution; vi) temporal resolution; vii) data timeframe; and viii) data format.

Table 3: Summary of the metadata of the collected indicators

Class of indicator	Macro-category indicator	Data source	Spatial domain	Spatial resolution	Temporal resolution	Data timeframe	Data Format
Atmospheric indicators	• Temperature • Precipitation • Humidity • Solar radiation • Wind	ARPAV	Veneto region	20 stations	Hourly, daily, monthly, yearly	2000-2019	Csv, NetCDF
Oceanographic indicators	Sea surface	CMEMS	Mediterranean area	0.0625*0.0625	Hourly	2000-2019	NetCDF
	Wave regime	CMEMS	Mediterranean area	0.042*0.042	Hourly	2000-2019	NetCDF
Territorial indicators	Land use	Carta Copertura Suolo Veneto	Veneto region		Triennial	2009,2012, 2015,2018	Shp
	River discharge	ARPAV – ARPAE	Veneto region		Hourly	2000-2019	Xlsx
	Permeability	Carta permeabilità dei suoli ARPAV	Veneto region			2016	Shp
	Geomorphology (Soil type)	Carta suoli ARPAV	Veneto region			2019	Shp
	Topographic	DEM INGV Tinitaly	Veneto region	10 m		2007	Tiff
	• Shoreline length • Coastal dunes	Geodatabase gestionale delle coste venete (GCV - Progetto coste)	Veneto coastal area			2013	Shp
	Subsidence	Geodatabase gestionale delle coste venete (GCV - Progetto coste)	Veneto coastal area	50 m		2002-2010	Tiff

Damage occurrence	Presence/absence of damages	DPGR (Decreto del Presidente della Giunta Regionale) documents	Veneto coastal zone	Municipality	Daily	2009-2021	Xlsx

Atmospheric indicators were provided by ARPAV (Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto)[3] and were obtained from 9 meteorological stations placed in the 11 municipalities under investigation, which are located in Bibione (S. Michele al Tagliamento), Lugugnana (Portogruaro, Caorle), Eraclea, Cavallino-Treporti, Favaro Veneto (Venice), Venice – Istituto Cavanis, Sant'Anna (Chioggia), Po di Tramontana (Rosolina), and Padron (Porto Tolle). These stations collect data of precipitation, temperature, humidity, wind, and solar radiation on an hourly or daily basis. Three municipalities of the investigated case study (i.e., Jesolo, Porto Viro, and Ariano nel Polesine) do not have meteorological stations. Therefore, the nearest neighbors' rule was applied to infer their atmospheric conditions. Moreover, Venice counted two meteorological stations but just the data from Istituto Cavani were kept since the dataset was more complete.

The oceanographic indicators were downloaded from the CMEMS (Copernicus Marine Environment Monitoring Service) database[4], the marine division of the Copernicus Programme of the European Union. CMEMS provides free data of physical, chemical, and biological oceanic variables derived from satellites, in situ measurements, or models for the global ocean and the European marine waters. In particular, the dataset taken into consideration for this study is based on parameters related to the sea surface levels and state (wave regime).

Territorial indicators provide a picture of the geo-morphological and anthropic characteristics of the territory under investigation. They can be considered as elements of exposure and vulnerability as well as triggering factors depending on the context. To retrieve these variables different sources and georeferenced maps were consulted such as the *Carta Copertura Suolo Veneto*[5] for the land use cover, ARPAV and ARPAE[6] (Agenzia Regionale per la Prevenzione e Protezione Ambientale Emilia-Romagna) databases for the river discharge, ARPAV for the soil permeability and soil category chart, the *National Institute of Geophysics and Volcanology* (INSV) for the topographical indicators[7], the *Geodatabase gestionale delle coste venete*[8] for the information related to shoreline length, dune extensions, and subsidence.

[3] https://www.arpa.veneto.it/previsioni/it/html/index.php
[4] https://marine.copernicus.eu/it
[5] https://idt2.regione.veneto.it/
[6] https://www.arpae.it/it
[7] http://tinitaly.pi.ingv.it/
[8] http://sistemavenezia.regione.veneto.it/sites/default/files/documents/08_Shape/RelazioneGCV-rev-ott2015_0.pdf

The damages indicators were obtained from the Veneto Region historical database by retrieving the DPGR ("*Decreto del Presidente della Giunta Regionale"*) documents[9], drawn up after the activation of the "*stato di crisi*" at the regional level, which is claimed subsequently to the manifestation of severe weather events on the territory. These documents reported qualitatively the occurrence of some damages' typology such as damages related to flooding in cities, damage to agriculture/fisheries, population problems (e.g., deaths, displacements), damages to beaches (e.g., shoreline erosion, debris accumulation), damages to structures/infrastructures and economic activities (e.g., tertiary sector problems). Since the DPGR documents did not always report the municipalities where a specific type of damage happened, to all the municipalities involved in an extreme event on a certain date, all the indicated types of damage, for that day, were attributed. In order to gain more information on the occurrence of extreme events and the relative damages, two free databases were consulted: the Italian MeteoNetwork database[10] and the European Severe Weather Database (ESWD)[11]. The MeteoNetwork database was examined to corroborate the DPGR reportings on the presence of extreme weather events that occurred on the Veneto territory, while the ESWD database was consulted to obtain more detailed information regarding not only the extreme events but also, when possible, the damage manifestations.

[9] https://www.regione.veneto.it/web/protezione-civile/archivio-emergenze-anno-2021
[10] https://www.meteonetwork.it/tt/stormreport/
[11] https://eswd.eu/cgi-bin/eswd.cgi

3. ML-based methodology for assessing damages caused by extreme events in the case study area: development of a ML-driven coastal risk conceptual scheme

Natural hazards, such as extreme climate and weather events, have always trigged human communities. However, in the last centuries, due to the development of human society and economy, the number of assets that could be adversely impacted has exponentially increased (e.g., population, cities, infrastructures). Additionally, in the last decades, climate change has intensified the occurrence of these phenomena both in frequency as well as in magnitude (EEA, 2022), and projections are worsening the scenarios for the upcoming years. As consequence, the combination of natural and anthropogenic pressures is posing severe risks to communities all around the world, as documented by the study of Coronese et al. (2019), which revealed an increment in extreme economic damages due to natural disasters, whose trend is consistent with the climate change signal.

Accordingly, to mitigate the effects of such events and allow the communities to enhance their resilience against future hazards intensification, suitable adaptation plans must be adopted to prevent or reduce risks. In order to determine the best measures to apply, a risk assessment must be performed, which is the process of anticipating probable damages before they happen (Lee et al., 2020), or in other words, the study of the causes of possible hazards and probable undesirable events, and the potential damage (or consequences) that they may produce (Barandiaran et al., 2018).

Therefore, the aim of a risk assessment analysis, in the field of natural hazards, is to identify the disaster risk, which is defined as the potential loss of life, injury, or destroyed or damaged assets that could occur to a system, society or a community in a specific period of time (UNDRR, 2022). Moreover, the comprehension of the disaster risk is reported as the phase with the highest priority in the Sendai Framework for Disaster Risk Reduction 2015-2030 (UNISDR, 2015). In fact, only if the causes of the risks are known, a disaster risk management (DRM) can be developed.

Specifically, disaster risk is a function of three elements which are:

- **Hazard**. *A phenomenon that can have negative social and/or economic consequences or cause environmental damage.*
- **Exposure**. *The spatial and temporal coexistence of people or assets (both physical and environmental) and natural hazards, with the potential to suffer damage.*
- **Vulnerability**. *The characteristics and circumstances of a community, system, or asset, which make them susceptible to the harmful effects of hazards.*

In the context of risk assessment, coastal areas are a particular case of study. In fact, according to the last IPCC report (IPCC, 2022), coastal cities and settlements are in a more precarious situation if confronted with inland zones, since the presence of a higher population (in 2020, almost 11% of the global population resided in Low Elevation Coastal Zone), economic activities, and infrastructures.

Furthermore, in addition to global climatic hazards, the regions at the land-sea interface are affected also by ocean-driven hazards (e.g., shoreline erosion, harmful algal blooms, severe storms and storm surges, flooding, tsunamis, and sea-level rise; NOAA, 2022) which, by interacting with the before-mentioned socio-economic assets, make coastal communities extremely exposed to multiple risks, especially to those associated with extreme weather events (Li et al., 2022).

Nowadays, in relation to the complex interactions between elements of hazard, exposure, and vulnerability characterizing coastal areas, an important factor is determined by the responses, in terms of adaptation and mitigation strategies, adopted by the coastal communities, as emphasized by the framework proposed by Simpson et al. (2021) and modified for coastal environments (Figure 16). Indeed, the human response factor is fundamental to determining the level of risk (Calliari et al., 2019), for example, maladaptations or unsuitable coastal management can compromise, even more, an already fragile condition (Marone et al., 2017).

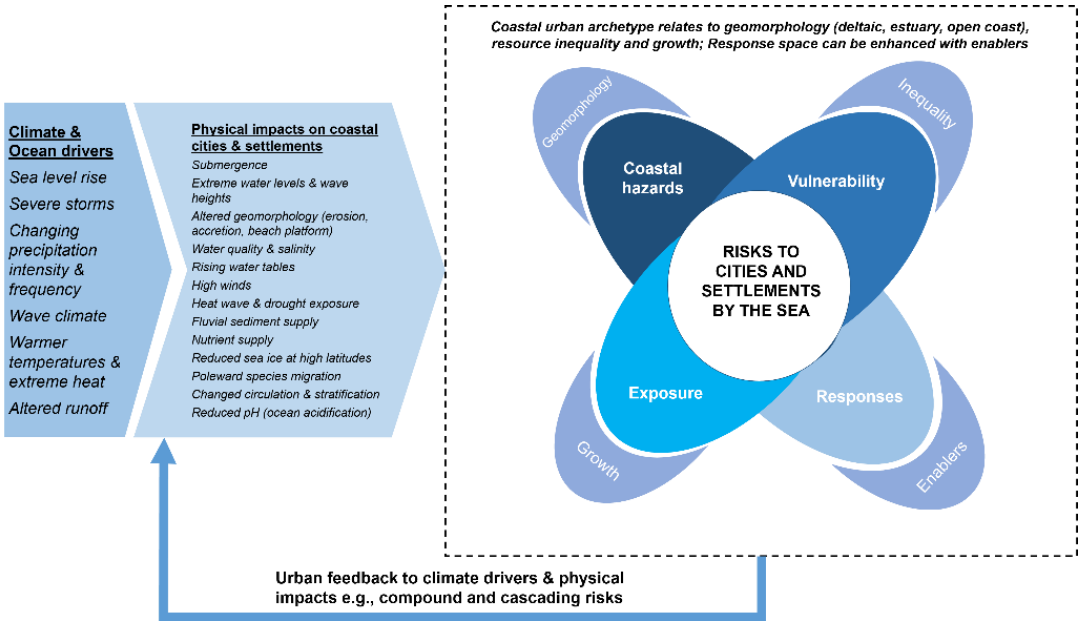


Figure 16: Coastal risk framework (IPCC, 2022)

In the frame of this thesis, a coastal risk conceptual scheme was designed to support the implementation of a classification ML algorithm capable of predicting the occurrence of damages, caused by extreme weather events in the Veneto coastal area, which, being part of the Mediterranean region, is particularly exposed to these phenomena (Pereira et al., 2021). Specifically, given a set of input variables recorded on a certain date, the ML algorithm aims to predict the presence or the absence of damage.

In the envisioned conceptual scheme (Figure 17), all the different categories of input variables are referred as “triggering factors” (since a classification ML does not distinguish the distinct categories of indicators i.e. hazard, exposure, and vulnerability categories), whereas the algorithm’s output (presence/absence of damage) is mentioned as “response factor”.

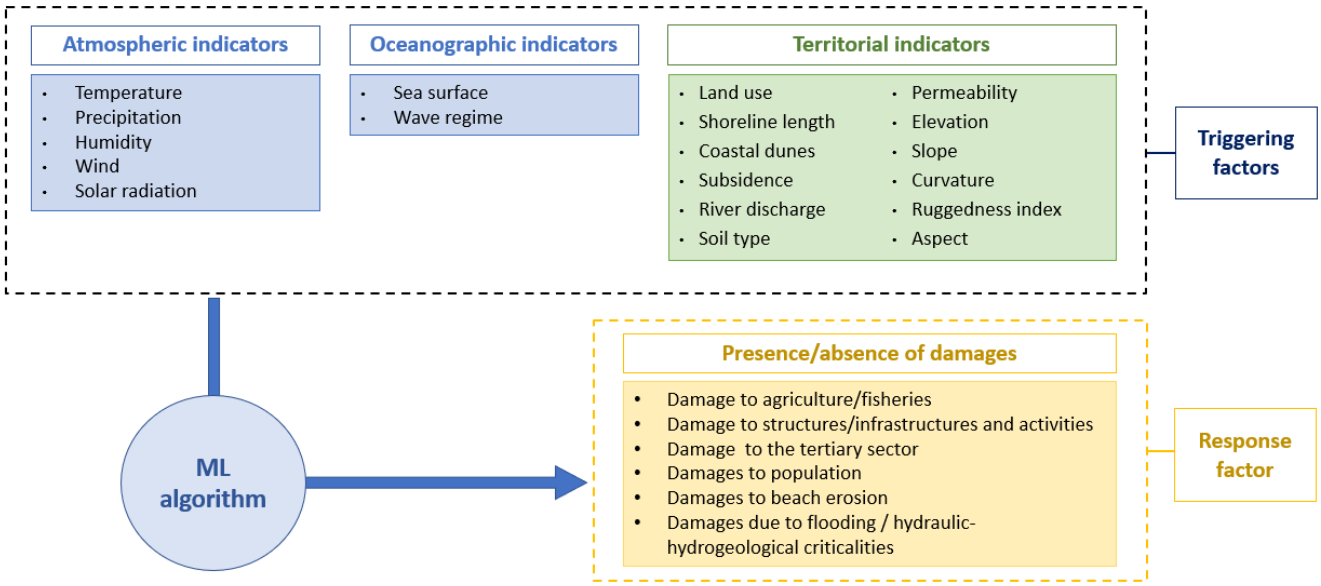


Figure 17: ML-driven coastal risk conceptual scheme

The triggering factors were selected according to the main indicators adopted in the literature for the prediction and evaluation of coastal damages and risks related to natural events (specified in Table 2), as well as to available data (specified in *Section 2.3*). To simplify the comprehension of the motivation that brought to the selection of the input variables, the triggering factors can be divided into three classes: atmospheric indicators, oceanographic indicators (which pose additional risks in coastal areas), and territorial indicators. It is important to remember that all these indicators strongly interact together and so, often, it is not a single variable that determines the damage occurrence but a combination of factors and their relations (Rutgersson et al., 2022; Wazneh et al., 2020).

In the following paragraphs, the choice of the indicators is motivated for each class of the triggering factors.

Atmospheric indicators. The selected 21 atmospheric indicators are related to:

Temperature. Temperature is one of the main variables used to assess extreme weather events since its increase can deeply change the hydrological cycle, affecting the atmospheric water vapor content and consequently the precipitation intensity (Aleshina et al., 2021). Higher temperature means more energy in the Earth's system, which in turn can increase the evaporation and therefore the formation of clouds (UNEP, 2022). Traditional temperature indicators such as the value of the mean, maximum (used to assess heat stress), and minimum temperature (used to assess cold stress) can indicate annual and seasonal trends. On the other hand, extreme temperature indicators can provide additional understanding of the pattern of extreme events, specifically in relation to drought and heatwaves (a phenomenon where high temperatures occur for several days), whose consequences can provoke several dangers to civil society (e.g. human health), economy (e.g., agriculture, electrical and technology's sector) and environment (Crespi et al., 2020). In the frame of this thesis, the chosen temperature indicators are:

- **Mean daily temperature** [°C]: average daily temperature;

- **Minimum daily temperature** [°C]: minimum temperature in a day;
- **Maximum daily temperature** [°C]: maximum temperature in a day;
- **Number of tropical nights** (TR): number of days in a year with a temperature of 20 °C or higher;
- **Number of hot days** (TX90p): monthly number of days with maximum daily temperature higher than the 90th percentile of the maximum temperatures, based on a mobile window of 5 days in the reference period 1991-2020;
- **Number of heat waves** (HWN): number of days in a year in which for at least 3 consecutive days the 90th percentile of the maximum temperatures is overcome, based on a mobile window of 31 days in the reference period 1991-2020;
- **Heatwave temperature** (HWTXdx) [°C]: maximum value between the averages of the maximum temperatures of each heatwave event.

Precipitation. Together with temperature, precipitation is a key variable to determine weather and climate regime, and therefore possible extreme conditions. Nevertheless, sometimes, records of daily mean precipitation do not allow to find changes in extreme precipitation conditions, which could be due to two opposite phenomena: heavy precipitation (defined as the maximum annual 5-day consecutive precipitation; EEA, 2021) or drought (a period of abnormally dry weather, long enough to cause a serious hydrological imbalance). Especially these latter mentioned extreme conditions are those that generally lead to severe damages and losses (EEA, 2022). Therefore, the adoption of these indicators is even more necessary in relation to climate change, since extreme precipitations are expected to increase over Europe up to 5% by 2050, although the annual precipitation is supposed to decrease (Pereira et al., 2021). The precipitation indicators selected for this study are:

- **Daily precipitation** [mm]: total daily amount (sum) of hourly precipitation;
- **Maximum precipitation** [mm]: maximum daily amount of hourly precipitation;
- **RX1day** [mm]: monthly maximum cumulative precipitation in 1 day;
- **RX5day** [mm]: monthly maximum cumulative precipitation in 5 consecutive days;
- **R95pDAY**: monthly number of days with cumulative daily rainfall exceeding the 95th percentile of the distribution of cumulative rainy days (precipitation ≥ 1 mm) in the reference period 1991-2020;
- **CDD**: maximum number of consecutive dry days (precipitation < 1 mm) in a year.

Humidity. Humidity is intrinsically related to temperature and pressure and affects precipitations. In fact, higher temperatures allow to increase the moisture content in the atmosphere, and therefore the intensity of precipitations. Specifically, the adopted humidity indicators are:

- **Humidex** [°C]: mean monthly humidex index;
- **Maximum humidity** [%]: maximum daily relative humidity;
- **Minimum humidity** [%]: minimum daily relative humidity;

- **HuxWF**: number of days in a year with mean daily Humidex value equal or higher than 35 °C for almost 3 consecutive days.

Wind. In addition to temperature and precipitation, winds are identified as key indicators when extreme weather events are investigated. In fact, extreme wind speeds can cause severe impacts to infrastructures and activities. Specifically, the study of wind conditions is particularly important in coastal environments, as wind intensity and direction deeply affect wave regimes, which can have cascading effects on coastal erosion and flooding (Seneviratne, 2012). In relation to global warming, higher heat content in the atmosphere and warmer ocean surface allow the formation of highly energetic storms with consequently high wind intensity. The wind indicators used in this study are:

- **Daily average wind velocity** [m/s]: mean wind speed at 10 m;
- **Daily maximum wind velocity** [m/s]: maximum wind speed at 10 m (wind flurry);
- **Daily mean wind direction** [°]: mean wind direction at 10 m**.**

Solar radiation. Solar radiation is the driver of the cyclic component of the variations of the terrestrial atmosphere's thermodynamic state (Battinelli, 1997). Therefore, it has a strong impact on the average temperature and consequently on the weather/climatic machine. To consider this factor, in this thesis, the adopted indicator is:

- **Solar radiation** [W/m^2]: daily solar global radiation.

Oceanographic indicators. The selected 12 oceanographic indicators are related to:

Sea-surface indicators. In coastal environments, the change in the sea surface level and in the marine currents regime, particularly in relation to storm surge events (generated by the drop in atmospheric pressure and strong winds), can cause disastrous effects on coastal communities, principally along the shoreline. If future scenarios are considered, higher mean sea levels, consequent to climate change, will modify permanently the wave height in the surf zones, a condition that will be aggravated in presence of extreme events (Seneviratne, 2012). In the frame of this thesis the selected sea-surface related indicators are:

- **Sea surface height** (SSH) [m]: the height of the sea surface above a reference ellipsoid;
- **Maximum sea surface height** (MSSH) [m]: maximum registered SSH value;
- **Eastward seawater velocity** (ESV) [m/s]: eastward component of the seawater velocity current, detailed in earth coordinates relative to North True;
- **Northward seawater velocity** (NSV) [m/s]: northward component of the seawater velocity current, detailed in earth coordinates relative to North True.

Wave. Wave parameters are generally used in several studies aiming at assessing the damages caused by extreme hazards in coastal environments (Table 2). Extreme waves can threaten the safety of coastal inhabitants and those involved in maritime activities (Seneviratne, 2012). In addition, waves are decisive in

shaping littoral areas, since the energy dissipation of the breaking wave on the coastline can severely contribute to the erosion process (Seneviratne, 2012). The selected variables related to wave regime are:

- **Significant wave height** (WAH) [m]: average height of the highest one-third of all waves measured;
- **Maximum significant wave height** (MWAH) [m]: maximum registered value of WAH;
- **Significant wind wave height** (WIH) [m]: height of waves that are the direct result of the local wind;
- **Max significant wind wave height** (MWIH) [m]: maximum registered value of WIH;
- **Sea surface wave mean period** (WAP) [s]: time interval between two consecutive wave crests to reach a fixed point;
- **Sea surface wind wave mean period** (WIP) [s]: time interval between two consecutive wind wave crests to reach a fixed point;
- **Wind wave direction from** (WID) [degree]: wind wave direction expressed as North (0°) and East (90°) component in earth coordinates relative to the Magnetic North;
- **Wave direction from** (WAD) [degree]: wave direction expressed as North (0°) and East (90°) component in earth coordinates relative to the Magnetic North.

Territorial indicators.

The territorial indicators play an important role in the generation of damages caused by extreme weather, since they can act as pathways or exposure and vulnerability elements, increasing or reducing the risk according to the circumstances and the hazard type. A variety of territorial indicators are used for evaluating risks. For example, flooding risk, which is one of the main hazards threatening coastal areas, is assessed through different indicators: for instance, Ha & Kang (2022) used permeability and river discharge, Collins et al. (2022) adopted also curvature and elevation, Park & Lee (2020) considered land use and urban area indicators. In relation to the land use indicator, some scientific studies have demonstrated how the decreased vegetation cover would increase the flood risk (Apollonio et al., 2016; Bae & Chang, 2019; UNEP, 2022). In detail, for this study, 18 territorial indicators have been chosen:

- **River discharge** [m^3/h]: volumetric flow rate of water that is transported through a given cross-sectional area;
- **Soil type** [% of municipal area]: in the frame of this thesis only two soil types are retained, that count as individual indicators: i) CL1 (soils on dune ridges and lagoon islands, formed by sands, from highly to extremely calcareous); ii) CL2 (soils on reclaimed lagoon areas, artificially drained, formed by highly to extremely calcareous silts). Specifically, for each municipality, the two soil classes are expressed in % of covered municipal area;
- **Permeability** [% of municipal area]: identifies the surface capacity to absorb or reject water, such capacity can be classified in 5 categories: very low (0,036-0,36 mm/h), low (0,36-3,6 mm/h), medium (3,6-36 mm/h), high (36-360 mm/h), very high (360+ mm/h). In the frame of this thesis, the

permeability indicator is expressed in terms of % of covered muncipal area with a permeability lower than 3,6 mm/h;

- **Elevation** [m]: height value (z-axis) at each cell, extracted from the DEM;
- **Slope** [°]: represents the steepness of a terrain;
- **Curvature** [°]: represents the distortion of the slope surface;
- **Ruggedness index** [m]: expresses the amount of elevation difference between adjacent cells of a DEM;
- **Aspect** [°]: is the direction of the maximum slope of a surface; the values of aspect range from 0° to 360°, degrees identified with the North direction;
- **Land use** [% of municipal area]: % of municipal area covered by 5 land use categories namely natural, beach, internal water, agricultural and fisheries area, and anthropic classes. Each land use category counts as an individual indicator;
- **Shoreline length** [m]: length of the municipal shoreline;
- **Coastal dunes** [m]: length of the municipal shoreline covered by dunes;
- **Subsidence** [m]: vertical land movement of the Earth surface.

As it will be described in *Chapter 4*, this thesis has to be considered as a preliminary step for the evaluation of the factors which mainly contributed to the damages occurred within the 2009-2019 timeframe in the case study area. Accordingly, more accurate ML algorithms can be designed starting from the identification of the most relevant factors. In fact, the presence of negligible variables can decrease the predictive capacity of the model and, at the same time, anomalies in the dataset can bias the predictions.

4. Data analysis methodology to evaluate the factors influencing damages caused by extreme events

The objective of this chapter is to describe the data analysis methodology implemented to find, and analyze, the factors that mostly contributed to the damage occurrences in the case study area. In order to do so, traditional statistical methods of exploratory data analysis (EDA) and a classification Random Forest (RF) were applied both at the regional and municipal scale, for evaluating the presence of local differences. Accordingly, the following sections report the methodology regarding: the pre-processing of the data, to homogenize in space and time the input variables (*Section 4.1*); the methods of descriptive statistics and EDA tools adopted to identify trends and associations among the variables (*Section 4.2*); the application of the RF model, which aimed to select the most important features causing damages (*Section 4.3*); and the investigation of the selected features, by exploring their differences in presence and absence of damage (*Section 4.4*).

4.1. Data pre-processing

Data pre-processing is a phase of the data preparation and follows the data collection (presented in *Section 2.3*). It is a fundamental aspect of data analysis since the raw dataset is prepared and elaborated in order to extract the desired information. Precisely, two main concepts of data pre-processing were applied: data transformation and data cleansing. Data transformation consists in manipulating the dataset through strategies concerning the homogenization of the data format as well as the creation of new variables (e.g., data aggregation, attribute/feature constructions, normalization) (Mushtaq, 2019), instead data cleansing identifies inaccurate, incomplete or incorrect parts of the data which are then modified, replaced or deleted (Rahman, 2019). The manipulation of the collected data was particularly necessary to obtain variables with the same spatial and temporal resolution. Moreover, additional variables were created to gain more understanding in predicting the damage occurrences.

In the frame of this thesis, the spatial resolution was set at the municipal scale, whereas the temporal resolution was daily for the period 2009-2019.

The final dataset was in a tabulated format and it comprehended four different classes of indicators (i.e., atmospheric, oceanographic, territorial, and damages) requiring a distinct pre-processing which was performed in both QGIS and Python environments (i.e., *pandas* library).

Concerning the atmospheric dataset, a portion of it required to be homogenized, as variables had hourly (i.e., precipitation, temperature, relative humidity), monthly (e.g., RX-1day, RX-5day) or yearly resolution (e.g., heat waves and dry days indexes), for the entire Veneto region, and not just for the coastal municipalities.

To resolve the heterogeneous spatio-temporal resolution, the input variables were converted into averaged values, however, taking into account also daily maximum and minimum values. On the other hand, in order to overcome the spatial issue, the tabular dataset was transformed into a vector in the QGIS environment for framing the dataset by taking into account the administrative boundaries of the Veneto municipalities.

For the oceanographic variables, each of them was computed by calculating the average daily value for each of the investigated municipalities.

The main pre-processing methodologies of the territorial data regarded the creation of new variables and the aggregation of others. Specifically, from the digital elevation model (DEM) layer, through the *SaGa-terrain analysis-Morphometry tool,* variables such as slope, plan curvature, aspect, and ruggedness Index (RI) were generated. For variables consisting of several classes (e.g., soil type, permeability, land use), the area of each class was summed over the respective municipal boundary.

As described in *Section 2.3* the damages data were provided by the DPGR ("*Decreto del Presidente della Giunta di Regione*") documents where six damage categories were reported, namely, hydrological damages, damages to agriculture and fisheries, damages to beaches, damages to infrastructures and activities, damages to the tertiary sector, and damages to population. However, since this information was not so detailed for the analysis carried out in this thesis, a unique damage variable was created called "any damage" indicator, to which was attributed value 1 if at least one type of damage occurred on a certain day in a considered municipality, otherwise 0.

After having consulted the national MeteoNetwork and European Storm Weather Database (ESWD) for validating the DPGR documents, it was noticed that sometimes extreme events were recorded some days before the damages reported in the regional papers, hence new variables were created. Precisely, from the dataset, the information of 12 main hazard variables related to 1, 2, and 3 days before the event were retrieved. These main hazard variables were: solar radiation, wind direction, maximum wind velocity, mean wind velocity, R1-day, precipitation sum, precipitation max, maximum temperature, sea surface height (SSH), maximum sea surface height (MSSH), maximum significant wind wave height (MWIH), and maximum significant wave height (MWAH). Regardless of the fairness of the DPGR documents, these additional variables could increase the ability to predict damages since in coastal areas floods and damages are the direct effects of meteorological forcings that produce, for a period of time spanning from hours to days, anomalous values of metrics such as sea level and waves (Lionello et al., 2012).

In this study, the data cleansing phase was mainly related to the treatment of missing values, which were replaced with the mean value of the associated variable for the considered timeframe. Extreme values of the dataset were not deleted, since, in dealing with damages caused by extreme events, these values could be correct and their elimination could bais the entire analysis.

4.2. Explorative data analysis of the dataset

In this section, the main techniques of data science, followed for analyzing the dataset, are described. In particular, after having homogenized in space and time the initial dataset, during the pre-processing phase (*Section 4.1*), descriptive statistics and exploratory data analysis (EDA) techniques were applied. EDA is a term coined by John W. Tukey for describing the act of looking at data to see what it seems to say (Morgenthaler, 2009) through visualization and manipulation of the initial data. In addition, implementing EDA techniques in ML-driven studies provides multiple information: it helps to understand the fairness of the input data, to have a better comprehension of the observed results, and to detect anomalies (Hafen & Critchlow, 2013).

In the frame of this thesis, analyses of the dataset were executed both at the regional *(Section 4.2.1)* and municipal scale *(Section 4.2.2).*

4.2.1. Regional-scale analysis

Initially, the historical yearly time series of the extreme weather-related damages, occurred in the Veneto coastal municipalities in the 2009-2019 timeframe, was analyzed. Since the dataset contained daily data for each of the investigated municipalities, to perform a regional assessment, the data had to be "normalized". Therefore, for each day, it was detected if at least one damage had occurred in one of the 11 municipalities, if so, the variable "any damage" (see *Section 4.1*) was modified and attributed with value 1, if not with value 0. Then, through the *groupby* function of the Python's *pandas* library, the number of damages was summed according to the year.

Separately, the yearly mean of the variables of the initial dataset was calculated. This allowed to perform a correlation matrix between the yearly mean values of the oceanographic and atmospheric indicators and the yearly damage occurrence. The correlation matrix, which provides an indication of the association between the variables (Senthilnathan, 2019), was based on the Pearson's coefficient (ρ) which measures, precisely, the strength of the linear relationship between two variables X and Y.

The formula is:

$$\rho_{(x,y)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sigma_{(x)}\sigma_{(y)}}$$

where *cov(x,y)* indicates the covariance between the two variables and *σ* the standard deviation.

For each indicator's macro-category (see Table 3), the variable having the highest correlation index with the yearly damage variable was plotted together with this latter one, to visually inspect similar patterns (e.g., similar peaks behavior). The same operation was repeated by keeping, from the original dataset, only the dates reporting the presence of damage, and from that, the yearly mean value of the hazard variables was calculated.

Finally, the presence of similar seasonal and monthly patterns, between the damages and the hazard indicators, was assessed by performing the same kind of analysis executed for the years.

4.2.2. Municipal-scale analysis

To understand if, at the local scale, the values of the hazard indicators and the damages recordings showed significant differences in respect to those found at the regional scale, a proper investigation was carried out by confronting the different municipalities. Since the number of dates in which damages occurred was very small, and it became even smaller if disaggregated for the municipalities, the analysis conducted at the local scale considered values averaged for the entire 2009-2019 timeframe. In particular, for each municipality, the mean values of the hazard indicators and the number of damages were examined to detect local variations. The number of occurred damages was obtained by summing, for each municipality, the variable "any damage" (see *Section 4.1*).

In addition, for the variables that resulted highly correlated with the yearly number of damages at the regional scale (*Section 4.2.1*), an ANOVA test was performed to evaluate significant statistical differences among the municipalities, which could have been differently impacted by the same atmospheric and oceanographic variables.

With a municipal-scale analysis, it was possible to investigate if territorial indicators were correlated with the damage occurrences; that was visually assessed through a scatterplot in which the number of damages and the territorial variables, of each municipality, were compared.

A focused analysis was specifically applied for the land use indicator with its relative categories (i.e., anthropic, agricultural, natural, internal waters, and beach coverage), whose values were provided on a triannual base. The aim was to detect if the change of a land-use category, over the years, could have influenced the damage generation. In fact, some studies demonstrated how land cover affects the occurrence of damages induced by natural hazards like flooding (Apollonio et al., 2016; Bae & Chang, 2019) or storms (Frazier et al., 2019); additionally, other studies discovered how land use can play an important role in the mitigation of extreme events themselves (Findell et al., 2017). To inspect this possible association (i.e., between the number of damages and land use evolution), for each municipality, the number of damages was summed on a triannual basis to meet the temporal resolution of the land use category, and finally, the relations among the variables were visualized through a scatterplot.

If significant differences both related to hazard as well as to territorial indicators, among the municipalities, were not found, that could support the reliability to execute a single RF for the regional scale, without performing a specific RF for the 11 municipalities (*Section 4.3*). Nevertheless, the obtained results must be carefully read, since the recordings of the damages at the municipal scale could be not so accurate (see *Section 2.3*).

4.3. Random Forest for feature selection

4.3.1. Introduction to Random Forest

To evaluate the variables that mostly influenced the damage occurrence, a classification random forest (RF) or random decision forest (RDF) was applied. A RF is a machine learning algorithm used both for regression and classification devised by Leo Breiman in 2001, which as output provides not only the prediction of an event but also the related most important features.

It is an ensemble method (multi-classifier) obtained by creating multiple decision trees (the single decision tree is the single classifier) during the training phase; the final result of a RF is determined by the results of the individual decision trees.

A decision tree (DT) is a ML algorithm aiming at predicting the values of a variable by giving in input a set of features; the scheme of a DT is reported in Figure 18.

In a DT there are three main node types:

- **Root node**. The node at the top of the tree;
- **Internal Nodes** (or simply "nodes"). They have arrows pointing to them and arrows pointing away from them. According to certain rules (e.g. Gini impurity), at the internal nodes, data are split;
- **Leaf Nodes** (or just "leaves"). They have arrows pointing to them but not arrows pointing away from them; they represent the terminal part of the DT.

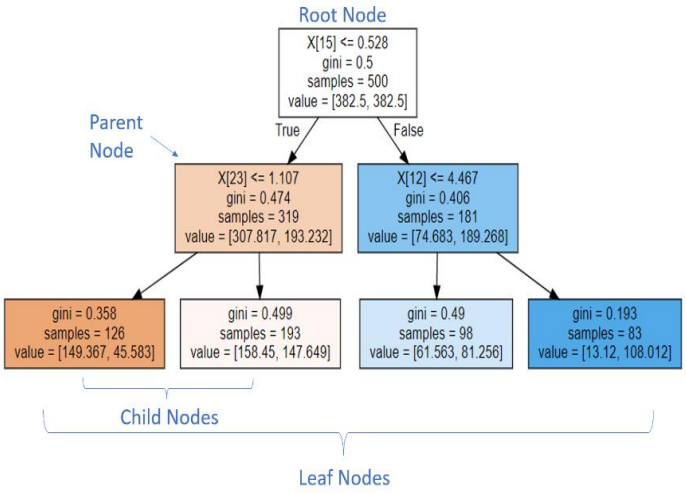


Figure 18: Scheme of a Decision Tree

Root node and internal nodes are associated with the input variables, leaf nodes with the output variables.

To construct a decision tree, the training samples are recursively split, by using the input variables. The splitting is performed in order to create groups/sets (or sub-populations) that are the most possible homogeneous internally and the most possible heterogeneous between each other. In other words, the objective of the tree is to find the values of the variable for which the best split is obtained. To split the population into heterogeneous groups, different methods can be used; in this study, the Gini impurity was chosen.

DTs have many pros such as being scale-invariant, robust to irrelevant features, and easily interpretable also by non-experts. However, they have a big con which is the tendency to overfitting (Jackson, 1988), so they perform very well for the training set but they are not good at generalizing the results, therefore they have high variance (terminology specified in *Section 1.1*).

To overcome the overfitting problem, a possible solution is the adoption of a RF algorithm. The combination of several decision trees is effective when the single classifiers (the DTs) are independent among them; to obtain independent decision trees two ideas are applied:

- Bootstrapped samples. Each DT is trained with a *bootstrapped dataset* that has the same size of the original one (e.i. every sample has the same number of variables of the initial dataset), by randomly selecting samples from the original dataset. The same sample can be sampled more than one time.
- Random feature subset (or feature sampling). For each DT only a *random subset of variables (features)* of the training bootstrapped dataset is selected. To decide which variable, from the random subset of variables, goes into each node, the algorithm selects the one for which the best splitting is obtained. There are different strategies aimed to decide the number of features used in each decision tree; a common choice is the square root of the total number of features.

This entire procedure (the creation of a new bootstrapped dataset and the building of a tree by taking into account only a subset of variables) must be repeated several times in order to have different independent trees; the ensemble of them constitutes the RF.

For each new input data each DF is evaluated: the RF's result is the one occurring more times among the DTs. In other words, the prediction of the category (class) associated with a certain leaf node, is the *mode* of all the categories falling in that leaf considering all the DTs constituting the RF (Hastie, 2009). This procedure, the bootstrapping of the data, and the aggregation of the results obtained from the individual DTs, is called *bagging* (Boostrap AGGregaTING).

RF for feature selection

Determining which predictors should be included in a model is becoming one of the most pivotal questions, as data are becoming increasingly high-dimensional (Kuhn & Johnson, 2013).

In many empirical analyses, a crucial problem is the presence of a set of variables not significatively contributing to explaining the analyzed phenomenon; that creates a random noise which prevents recognizing the main effects and the relevant predictors (Genuer et al., 2010). For the analysis performed in this thesis, a RF was adopted to evaluate the most important features in determining the damage occurrence since RF is considered a quite successful method for high-dimensional datasets and/or highly correlated input features (Chen et al., 2020; Zhou et al., 2022). Features selection with RF is considered an *embedded method*, which exploits the positive characteristics of the traditional feature selection methods namely *filtering* (methods that evaluate the relevance of the predictors outside of the predictive models, and subsequently model only the predictors that pass some criterion) and *wrapping* (in which multiple models are assessed by using procedures that add and/or remove predictors) methods (Kuhn & Johnson, 2013).

To determine the most important features, for each DT constituting the RF, the algorithm calculates how much a feature decreases the impurity of the leaf; then, the value of impurity decrease is averaged over all

the DTs to obtain the importance of that feature in the RF (Dubey, 2018). The more the presence of a feature reduces the impurity, the more important is the feature in the estimate of the output.

Among other feature selection methodologies, RF was chosen because is less affected by multicollinearity since, for each DT, a random subset of the initial features is selected (Raj, 2019). Anyways, in dealing with lots of variables the probability of having correlated variables in the *random features subset* increases, and that could partially obscure correlated variables in the outcome of the features' importance. To avoid the problem of having correlated features in the dataset, several studies (Strobl et al., 2007; Strobl et al., 2008; Parr, 2018) suggest using methods based on permutation importance (which exploit the concept of out-of-bag-samples and out-of-bag-errors) instead of on impurity indexes; however, in the frame of this thesis, the traditional method (impurity-based feature selection) was applied.

4.3.2. Data preparation and RF set-up

The most important task during a ML project is the correct formulation of the problem to solve. In order to estimate the importance of the features contributing to the damage occurrence, the question that we wanted to ask was: "Given these set of features, for day and municipality, which is the probability to have a damage?". To have a proper response, the dataset had to be balanced with a similar number of dates in which damages occurred and in which they did not (or the algorithm itself can take care of the skewness of the data by applying different weights to each sample). If the dataset is highly imbalanced (e.g., a class of the response variable is more represented than others) the results could be misleading because the more represented class would obscure the other response's variable classes, hindering the predictive ability of the model to find relations (Javaheri et al., 2013). Since, for this analysis, the problem was present (i.e., higher number of dates with no damages compared to the number of dates with damages), the dataset obtained after the pre-processing phase was manipulated again, specifically for the RF application.

Then, as described in *Section 1.1,* to run any type of supervised ML algorithm, the initial dataset must be split into three sets, called *training set*, *validation set* and *testing set.* The validation phase, to tune the hyperparameters, was conducted in the frame of the AdriaClim project by ML experts and was out of the scope of this analysis.

Therefore, after having randomized the order of the observations in the dataset, the dataset itself was balanced for having the same number of dates with and without damage, and finally split into training and test set, having respectively 75% and 25% of the dataset data. Since the splitting of the balanced dataset was executed through the function *train_test_split* of the *sklearn* package (implemented in Python), which every time splits randomly the train and test dataset, to guarantee the replication of the results, a fixed 'random_state' parameter (seed) was chosen.

Once the dataset is balanced and split into train and test sets, the RF can be run.

Initially, for each municipality, a RF was run to detect local differences in the prediction of damage occurrences. However, fearing the issue related to having an exiguous dataset if divided for the individual

municipality, which is a typical problem in ML (Zahura et al., 2020), a RF was run also for the whole Veneto coastal area, by combining all the municipalities' data. This last option permitted to take into account variables related to territorial characteristics, which were constant for each municipality within the 2009-2019 timeframe. The RF model was obtained from the *RandomForestClassifier* function of the *sklearn.ensemble* package.

Then, to correctly perform a RF, the hyperparameters have to be set up. As described before, the choice of the hyperparameters' values was based on pre-analyses made by ML experts within the AdriaClim project, by using a validation set.

The hyperparameters modified from default values were: 'n_estimators' (representing the number of trees), 'max_depth' (representing the maximum depth of the tree), 'min_samples_split' (minimum number of samples required to split an *internal node*), and 'class_weight' (weight associated to the classes of the response variable; in this case, since the dataset was previously balanced, the same weight was attributed to the response variable's classes namely damage occurrence and damage absence). For the RF run on a municipal scale the following hyperparameters were set: *n°estimators* = 150, *max depth* = 5, *min samples split* = 2; instead, for the RF run on a regional scale the hyperparameters were: *n°estimators* = 200, *max depth* = 8 and *min samples split* =2.

Once the random forest is run, the feature selection comes as an inherited result.

Finally, four RFs for the whole region, with different input variables, were compared through the F1 score (see next *Section 4.3.3*) and the best one was kept for the next analyses.

4.3.3. Evaluation of the Random Forest performance

To assess the performance of the model, therefore to summarize how a ML method (e.g., RF) performs on the testing data, a confusion matrix is calculated, and from the confusion matrix different metrics can be extracted. The advantage of the metrics over the confusion matrix is that each metric provides a scalar value for each configuration of parameters, hence allowing to determine if a configuration is better than another. For this reason, it is of paramount importance that the chosen metric represents the result that the user would like to obtain.

A confusion matrix is a square matrix where the dimension depends on the number of variables that the algorithm has to predict. The *correct predictions* are contained in the cells of the matrix diagonal whereas the *not correct predictions* are found in the cells not belonging to the matrix diagonal. The results of the confusion matrix for binary classification fall into four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). If the model aims to classify the presence or the absence of a certain event, for example, the presence or absence of damages, and the positive case is related to the presence of the damages, TP are the n° of damages that the model correctly predicts to happen, FP are the n° of damages that the model incorrectly predicts to happen, TN are the n° of no damages that the model correctly predicts not to happen and FN are the n° of no damages that the model incorrectly predicts not to happen.

From the confusion matrix several important metrics can be calculated (Kanstrén, 2020):

- **Sensitivity** or **Recall** = TP/(TP + FN). It expresses the ratio between the TP and the total actual positive cases;
- **Specificity** = TN/(TN + FP). It expresses the ratio between the TN and the total actual negative cases;
- **Accuracy** = (TN + TP)/(TN + TP + FP + FN). It expresses the ratio between the correct predictions and the total amount of predictions;
- **Precision** = TP/(TP + FP). It expresses the ratio between the TP and the total predicted positive cases;
- **F1 score** = 2/(1/Recall + 1/Precision).

All these metrics return a value between 0 (highly erroneous prediction) and 1 (perfect prediction). In this study, F1 score was selected to test the predictive ability of the RF, since it is considered the most comprehensive metric (Frasca, 2018).

The confusion matrix and the relative metrics were computed by importing *sklearn.metrics* package in Python environment.

4.4. Analysis of main indicators influencing damage occurrence

4.4.1. Regional-scale analysis

ML algorithms are powerful tools to investigate complex data but should not be considered a substitute for good research design and scientific reasoning (Jones & Linder, 2015). They are perceived as "black boxes", which can provide excellent predictions but whose outcomes should be accurately examined. Therefore, to understand if the results obtained from the RF feature selection were meaningful, and consequently, to provide a physical interpretation in the light of the observations, the indicators associated with high feature importance values in damage prediction were evaluated with traditional statistics and visual tools.

For this final analysis, all the records of the original dataset were kept. The initial objective was related not only to validating the RF's results but also to evaluating the differences in the variables' value during normal and extreme weather conditions, since the RF told just the relative importance of a variable but not how it changed (e.g., increase or decrease) during the tested events.

For this aim, for each selected variable, two probability density functions, for the two different situations (i.e., damage and not damage), were compared. The probability density function (PDF) is a non-negative function of a continuous random variable, that, when integrated across an interval, it gives the probability that the random variable takes a value that lies in that interval (Ibe, 2014). For the analysis conducted in this research, the PDF was automatically derived and visualized through the kernel density estimation (KDE) of the *seaborn* library in Python. The comparison of the two curves allowed to detect if there were differences in the two distributions (one related to the variable's values in damage presence and the other in damage absence), the associated variability and if, on average, the value of the variables increased or decreased during events with damage. If no differences were observed, probably that variable was not so important in

predicting damages and so not strongly associated with them. However, it must be remembered that the most important features, selected for predicting damages, were obtained by training the RF with a small fraction of the original number of observations, to overcome the problem of the unbalanced dataset (*Section 4.3.2*). Therefore, the random training samples could have been not representative of the entire statistical population, biasing the RF's predictions and consequently the retrieved selected features.

Since the ordinary KDE is good to describe the mode of a probability distribution but not very precise in modeling the distribution tails (Matsueda & Nakazawa, 2015), contemporary to the PDF, the 1-D distribution of the data, through a rug plot, was visualized. This simple graph is very powerful since it easily allowed to spot if, when there were damages, the variables' values were higher or lower than a certain threshold and if, by overpassing that threshold, no damage events were still present or not. In fact, if by exceeding a certain threshold only damages were observed, it could be said that the variable, directly or indirectly, was strongly associated with the damage occurrence. Nowadays, the identification of operational thresholds, defined as "levels of weather conditions at which a facility or piece of infrastructure experiences disruption, damage, or other impacts" (Asariotis et al., 2020), is fundamental to design suitable mitigation strategies against weather or climatic hazards (UNCTAD, 2017), and to operationalize early-warning systems (Papagiannaki et al., 2022; Young et al., 2021).

Then, to assess if the values of the most important variables, for predicting damages, changed over time, a specific investigation was done by executing the previous analysis on a yearly scale. In particular, since extreme weather events remarkedly increased in the last ten years, both in frequency as well as in intensity, globally and specifically in Italy (Osservatorio Nazionale clima e città, 2021), the analysis wanted to evaluate if the phenomenon was noticeable also in the case study area for the period 2009-2019.

Finally, it was explored how, over the seasons, the distribution of the main hazard variables changed during damage events. Since most of the atmospheric and oceanographic variables were seasonally driven, the difference in their respective values, during damage and no damage events, could give insights into the possible role played by the variable in determining the damage in a certain season, and how that role could mutate by changing season.

4.4.2. Municipal-scale analysis

The atmospheric and oceanographic variables which resulted to change significantly during damage events at the regional scale (*Section 4.4.1*) were investigated at the local scale.

The aim was to understand if the distinct municipalities, during damage events, were affected differently by the same hazard indicators and if, during these events, the variables showed the same municipal pattern present in no damage conditions.

Initially, for the 11 municipalities, the mean value of the selected variables, during damage and no damage events, was computed and confronted. Then, in parallel with the analyses conducted at the regional scale, the investigation was performed also on a seasonal basis to evaluate the seasonal differences on a local scale. To conclude, it must be remembered that the analysis of the damages conducted at the local scale could be biased by the presence of an insufficient number of data, as the damages recorded in each municipality were very few. Therefore, the results could be not indicative of municipal differences, and more detailed examinations should be made.

5. RESULTS: Data analysis of the indicators influencing damage occurrences in the coastal area of Veneto region, Italy

In this section, the main findings of the data analysis process, applied for determining the most important variables related to extreme weather-driven damages, are reported. After a brief description of the outcomes related to the raw dataset pre-processing (*Section 5.1*), the principal results derived from the pre-survey of the indicators through EDA (*Section 5.2*), the classification RF for feature selection (*Section 5.3*), and the final analysis which combines the previous two techniques (*Section 5.4*) are discussed both at the regional and municipal scale.

5.1. Data pre-processing

Following the pre-processing methodology reported in *Section 4.1*, the final homogenized dataset consisted of 44187 observations (samples) and 92 input variables (features). The observations were recorded for each of the 11 investigated Veneto coastal municipalities, for each day of the 2009-2019 timeframe.

The 92 variables comprised all the indicators described in *Chapter 3*, which were subdivided into 21 atmospheric variables, 12 oceanographic variables, 18 territorial variables, 1 damage variable, and 36[12] additional variables reporting, for a specific daily observation, the values of the most important oceanographic and atmospheric indicators registered 1, 2 and 3 days before the observation itself, as discussed *in Section 4.1*. Finally, four other variables accounted for the municipality index, the date, the season, and the month of the observation. The name of the 11 municipalities and the related identification index, adopted in the frame of this thesis, is clarified in Table 4, whereas the months associated with the season identification index in Table 5.

Table 4: Municipality name and relative identification index

Municipality name	Municipality index
San Michele al Tagliamento	0
Caorle	1
Eraclea	2
Jesolo	3
Cavallino-Treporti	4
Venezia	5
Chioggia	6
Rosolina	7
Porto Viro	8
Porto Tolle	9
Ariano nel Polesine	10

Table 5: Months and relative season index

Months	Season index
January/February/March	Season 1 (Winter)
April/May/June	Season 2 (Spring)
July/August/September	Season 3 (Summer)
October/November/December	Season 4 (Autumn)

The dataset presented 34437 missing values, which were replaced with the average value of the associated variable by considering the entire timeframe.

[12] The list of the atmospheric, oceanographic and territorial variables is reported in *Chapter 3*. The information related to the damage indicator and to the variables recorded in the days preceding the observations is in *Section 4.1*.

5.2. Explorative data analysis of the dataset

5.2.1. Regional-scale analysis

Annual analysis

The information related to the days which registered the occurrence of at least one damage in the Veneto coastal municipalities, within the 2009-2019 timeframe, is summarized in Figure 19. Specifically, among the 4015 days of the dataset, only 95 of them saw the manifestation of at least one damage.

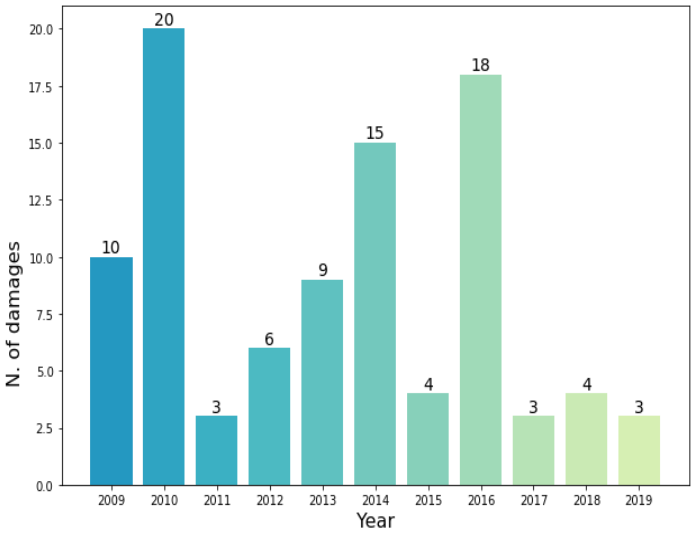


Figure 19: Yearly distribution of the occurred damages in the coastal area of Veneto region within the 2009-2019 timeframe

2010 recorded the highest number of damages (20); in fact, in that year the Veneto region was severely hitten by a series of extreme meteorological events, among which the dreadful flooding episode occurring between the 31st of October and 2nd of November, known as "*Alluvione dei Santi*" (Regione del Veneto, 2011). 2016 was in the second position for the number of damage occurrences (18), followed by 2014 (15) while 2011, 2017, and 2019 accounted for the smallest number (3).

To gain more understanding of possible relations existing between the damage occurrences and the atmospheric and oceanographic indicators, a correlation matrix (reported in ANNEX III) was calculated by confronting the mean yearly values of the hazard indicators with the number of yearly damages (procedure reported in *Section 4.2.1*). For each macro-category of the hazard indicators defined in Table 3, the variables mostly correlated with the yearly damage occurrences are reported in Table 6.

Table 6: Variables highly correlated with the number of yearly damages within the 2009-2019 timeframe

Indicator macro-category	Variable having the highest absolute value of correlation with the yearly damages variable	Correlation index value
Temperature	Maximum temperature	-0.62
Precipitation	Daily precipitation	0.676
Humidity	Minimum humidity	0.579
Wind	Maximum velocity	0.48
Solar radiation	Solar radiation	-0.701
Sea surface	MSSH (maximum sea surface height)	0.449
Wave regime	WAP (sea surface wave mean period)	0.609

To visually inspect the correlation between the mean yearly values of the described variables and the yearly damages, the annual trends are reported in Figure 20. For most of the cases, the positive and negative peaks

of the damages and the ones of the examined variables showed a significant similar or opposite pattern, except for 2018 and 2019. That could signify a scarce association between the variables and the damages in the last two years of the dataset. Anyways, the yearly association between the selected variables and the damage occurrences is quite undeniable, meaning that the annual fluctuation of the hazard variables could play a role in creating the conditions for damage manifestation.

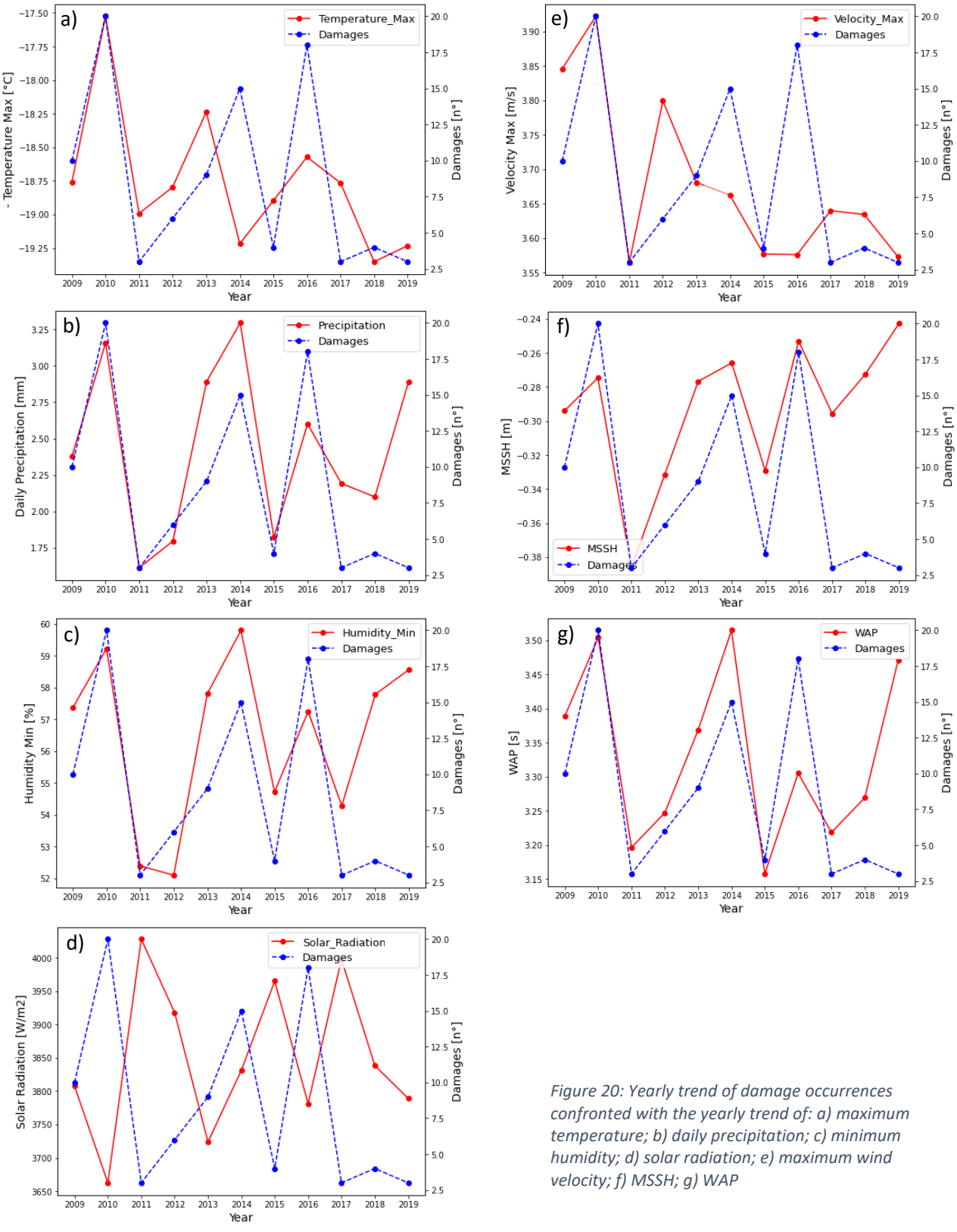


Figure 20: Yearly trend of damage occurrences confronted with the yearly trend of: a) maximum temperature; b) daily precipitation; c) minimum humidity; d) solar radiation; e) maximum wind velocity; f) MSSH; g) WAP

The same type of analysis was performed by keeping only the dates in which damages occurred, and from that, the yearly mean of the hazard variables was calculated. However, in this case, the same graphs plotted in Figure 20 did not show any kind of similar pattern between the peaks of the variables and those of the number of yearly damages. The controversial result could be explained by the fact that there could have been a delay in the damage reporting (i.e., the damage was not recorded on the same date of its happening) or simply, the values of the hazard variable during days in which damage happened were not yearly correlated with the number of damage occurrences.

By analyzing the entire correlation matrix interesting findings were discovered. Neglecting the expected correlation between the same type of indicator class (i.e., oceanographical and atmospheric) such as, for instance, the high correlation between temperature and humidity, humidity and precipitation, considerable correlation values were found between other types of variables. For example, mean (SSH) and sea surface height (MSSH) were highly related to the indicators of precipitation (0.591 - 0.8), humidity (0.84 - 0.93), and solar radiation ((- 0.715) - (-0.743)), in turn, these latter variables were highly correlated with wind wave characteristics, especially with the mean wave period (WAP), e.g. WAP and precipitation indicators registered a correlation of 0.74 - 0.92. These values revealed how atmospheric and oceanographic variables strongly interacted with each other. Therefore, also in the context of damages caused by extreme weather events in coastal areas, it must be remembered that often it is the combination of different variables which allows explaining a certain occurrence rather than a variable alone (Wazneh et al., 2020).

A secondary correlation analysis was then executed to test the information of the main oceanographic and atmospheric variables recorded 1, 2, and 3 days before the observation's date (described in *Section 4.1*). On average, the variables registered 2 days before the event's date were the ones having the highest correlation with the damages. This information was used for running the RF (*Section 5.3*).

Seasonal and monthly analysis

The same damages analysis executed for detecting the yearly trend was proposed for seasons (Figure 21) and months (Figure 22). It emerged how the spring and the summer seasons had the highest number of damages, reaching their peaks respectively in the months of June and September. These seasons showed also the highest values of minimum, mean and maximum temperature (respectively 17.58 °C, 22.71 °C, and 28.10 °C), solar radiation (5589 W/m^2), and extreme precipitation (maximum precipitation: 1.34 mm; RX-1day: 29.52 mm); oppositely, in these seasons, the lowest values of minimum humidity and of those associated to the main oceanographic variables (e.g., SSH, MSSH, WAH, WID, WIP) were reached.

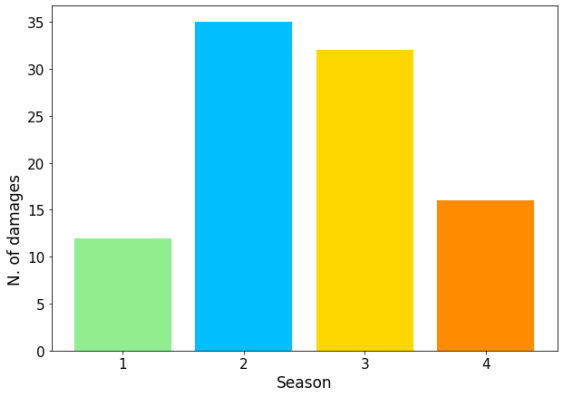


Figure 21: Seasonal distribution of damage occurrences within the 2009-2019 timeframe

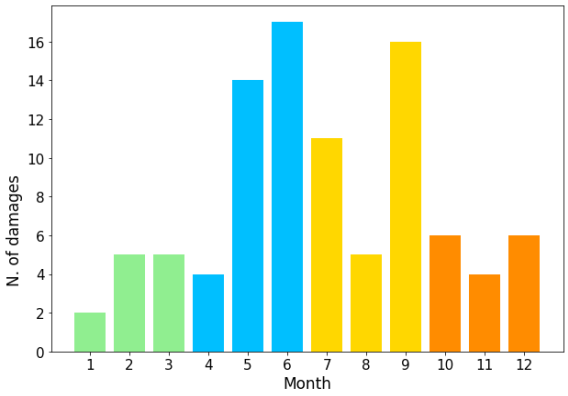


Figure 22: Monthly distribution of damage occurrences within the 2009-2019 timeframe

While some of the hazard variables (e.g., temperature, precipitation) had mean seasonal values that exhibited a similar pattern to that of the seasonal damages, for the monthly analysis just the precipitation indicator presented these characteristics. The main interesting observations between the seasonal and monthly damage trends with the hazard indicators are reported in ANNEX IV.

Nevertheless, an in-depth analysis based on the seasons and months of every individual year of the 2009-2019 timeframe, regarding both the damage occurrences (distribution reported in ANNEX V) and the hazard indicators, has brought no evidence of a clear seasonal/monthly trend (i.e., the mean values of the variables had not a recognizable seasonal/monthly pattern in the different years). Therefore, the role played by the seasons and months in the damage occurrences is difficult to assess.

5.2.2. Municipal-scale analysis

For every investigated municipality Figure 23 shows the sum of days in the 2009-2019 timeframe that presented at least one damage. It can be seen how, except for municipality 5 (Venice), the number of damages decreases by going from north to south, shifting from 55 to 28. That could be due to a variety of reasons related to different hazards intensity and frequency among the municipalities, different territorial features as well as a diverse municipal area, or different local characteristics of the assets' exposure and vulnerability (this latter information was not available in an exhaustive way for the analysis conducted in the frame of this thesis).

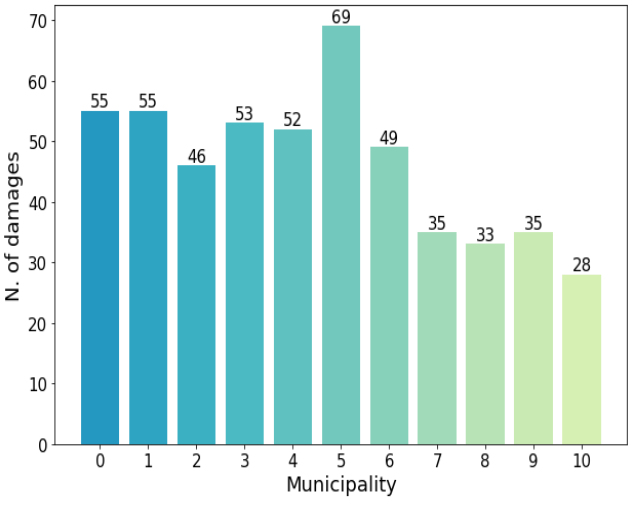


Figure 23: Damages occurred in the 11 investigated municipalities within the 2009-2019 timeframe

Concerning the averaged values of the hazard indicators for the 2009-2019 timeframe, computed for each municipality, it was found that the atmospheric indicators related to solar radiation, temperature, and

humidity were quite similar. In particular, solar radiation had values around 3613-4025 W/m^2, maximum temperature around 18.4-19.1 °C, mean temperature around 13.3-14.5 °C, and maximum humidity of 95.3-98.3% RH. The only slight difference concerned municipality 5 where the minimum temperature (8.1 °C) and the minimum humidity (53.2% RH) showed the lowest values compared with the other municipalities, whose values were instead quite homogeneous.

CCD (cumulative dry days) and HuxWF (number of days in a year with Humidex value higher than 35 °C for three consecutive days) variables displayed a similar pattern, with municipalities 4 and 5 having the highest values, followed by the southern municipalities and then by the northern ones. In detail, CDD reached a value of 43 days for municipalities 4 and 5, around 19-24 days for municipalities 6-10, and 11-26 days for municipalities 0-3; while HuxWF counted 12 and 15 days for municipality 4 and 5, 10 days for municipalities 6-10 and about 5-9 days for municipalities 0-3. All this means that the central municipalities suffered more heatwave and droughty conditions, whereas the northern ones were less affected.

All the precipitation indexes evidenced decreased values going from north to south, for example, the average daily precipitation passed from the 2.99 mm of municipality 0 (San Michele al Tagliamento) to the 1.91 mm of municipality 10 (Ariano nel Polesine); the northern municipalities resulted wetter because they were influenced by the Pre-alpine barriers (Barbi et al., 2012).

Regarding the wind velocity, both the mean and the maximum velocity values manifested the same trend over the municipalities, with municipalities 7 and 8 having the highest wind values (mean wind velocity respectively of 2.87 and 2.39 m/s and maximum wind velocity of 5.34 and 4.65 m/s). These data could be related to the different shoreline orientation of the municipalities 7 and 8, which could influence the wind intensity.

Most of the oceanographic variables, both related to the sea surface height as well as to the wave regime, exhibited a clear pattern by going from north to south of the investigated area, in fact, except for WAD (wave direction), eastward (ESV) and northward seawater velocity (NSV), all the other indicators increased in their values. Specifically, SSH went from -0.39 to -0.35 m, MSSH from -0.30 to -0.28 m, MWAH (maximum significant wave height) from 0.49 to 0.81 m, MWIH (maximum significant wind wave height) from 0.35 to 0.49 m, WIH from 0.13 to 0.20 m, WID (wind wave direction) from 136 to 154 degrees, WAP (wave period) from 3.01 to 3.56 s and WIP (wind wave period) from 1.35 to 1.70 s. Conversely, WAD decreased from municipality 0 to municipality 8 (respective values of 138 and 94 degrees, shifting towards a more northern direction of the wave regime) to increase again for municipalities 9 and 10. An atypical behaviour was manifested by the variables related to the velocity of ocean currents (ESV and NSV), which had a quite heterogeneous pattern among the municipalities. ESV reached the maximum values for municipality 6 (0.051 m/s) followed by municipality 7 (0.034 m/s) and 8 (0.034 m/s), NSV reached the maximum positive value (e.i. a northward direction of the sea current) for municipality 6 (0.034 m/s) and the highest negative value (e.i. a southward direction of the current) for municipality 10 (-0.07 m/s).

If the average trend of the analyzed variables is compared with the ones of the damages (Figure 23), visually, it seems that only precipitation indexes and those related to oceanographic variables could have had an association with the damages at the municipal scale, since the gradual increase or decrease.

The variables which resulted to be higher correlated with the damages at the regional scale were investigated for each individual municipality (Figure 24), to detect differences or similarities among the municipalities themselves.

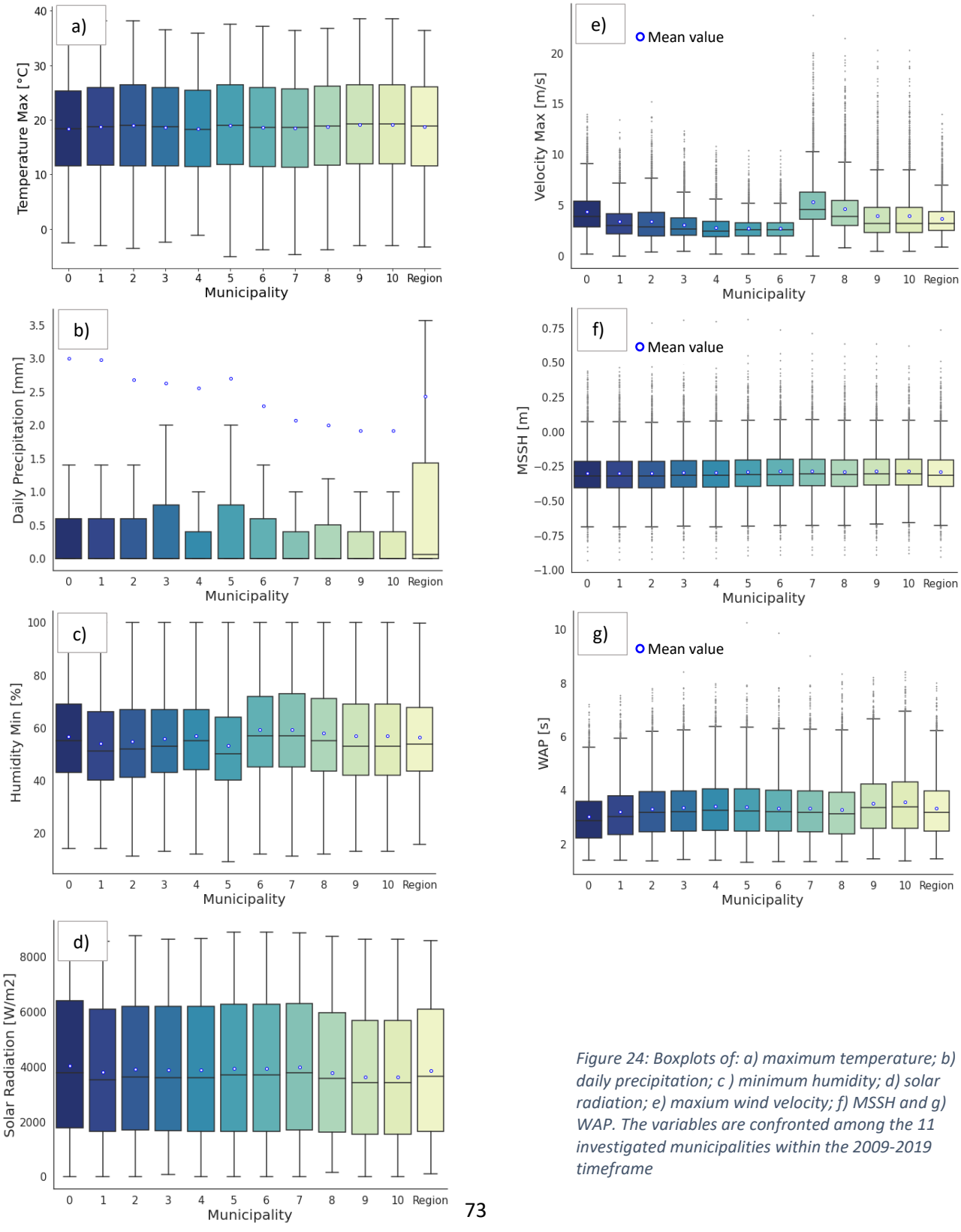


Figure 24: Boxplots of: a) maximum temperature; b) daily precipitation; c) minimum humidity; d) solar radiation; e) maxium wind velocity; f) MSSH and g) WAP. The variables are confronted among the 11 investigated municipalities within the 2009-2019 timeframe

Unfortunately, to investigate if the variables among the 11 municipalities were statistically different, the ANOVA test could not be performed since most of the variables did not respect the ANOVA assumptions, which are the normal distribution of the residual errors and the homoscedasticity requirement. The test was applied only to mean and maximum sea surface height (SSH and MSSH) that respected the assumptions, but the p-value was always lower than 0.01 so the variables for the different groups (municipalities) were statistically different. Even though from the boxplots (Figure 24) MSSH seemed quite similar among the municipalities, the information from the ANOVA test allows to carefully interpret the results and not make hasty conclusions. So in the end, the hazard variables among the municipalities could be statistically different and could have differently affected the damage occurrences in presence of extreme weather events.

Concerning the territorial indicators, the municipalities showed significant divergences, although the retrieved data were mean values, calculated over the entire municipal area, which as a matter of fact could have had heterogeneities inside its territories.

The municipalities' area had a value spanning from about 45 km^2 (municipality 4; Cavallino Treporti) to 416 km^2 (municipality 5; Venice).

The percentage of the municipal area with a permeability lower than 3.6 mm/h was highest in the municipalities 0, 1, 2, 5, 8, 9 with a value around 0.81%, lower for the others, and null for municipality 4, meaning that the soil in this latter municipality was pretty permeable.

The percentage of the municipal area with a soil type falling in the CL1 class (calcareous soil) reached the highest value for municipality 4 (0.64%) followed by municipality 11 (Ariano nel Polesine) (0.41%); the others had a value lower than 0.4%. In relation to the soil type belonging to the CL2 class (calcareous silty soil), municipalities 1, 2, and 9 registered the highest values (respectively 0.66%, 0.67%, and 0.56%) whereas it was practically absent for municipalities 4 and 5.

The percentage of the subsiding area with a subsidence rate higher than 2 mm/y was the lowest for municipality 5 (19.3%) and the highest for municipality 4 (91.9%).

Elevation had heterogenous values over the municipalities with the maximum positive value for municipality 1 (Caorle) (1.75 m) and the maximum negative value for the southern municipalities (the minimum elevation was recorded in municipality 9 with -1.19 m); these latter municipalities are part of the Po River Delta, which is located below the mean sea level.

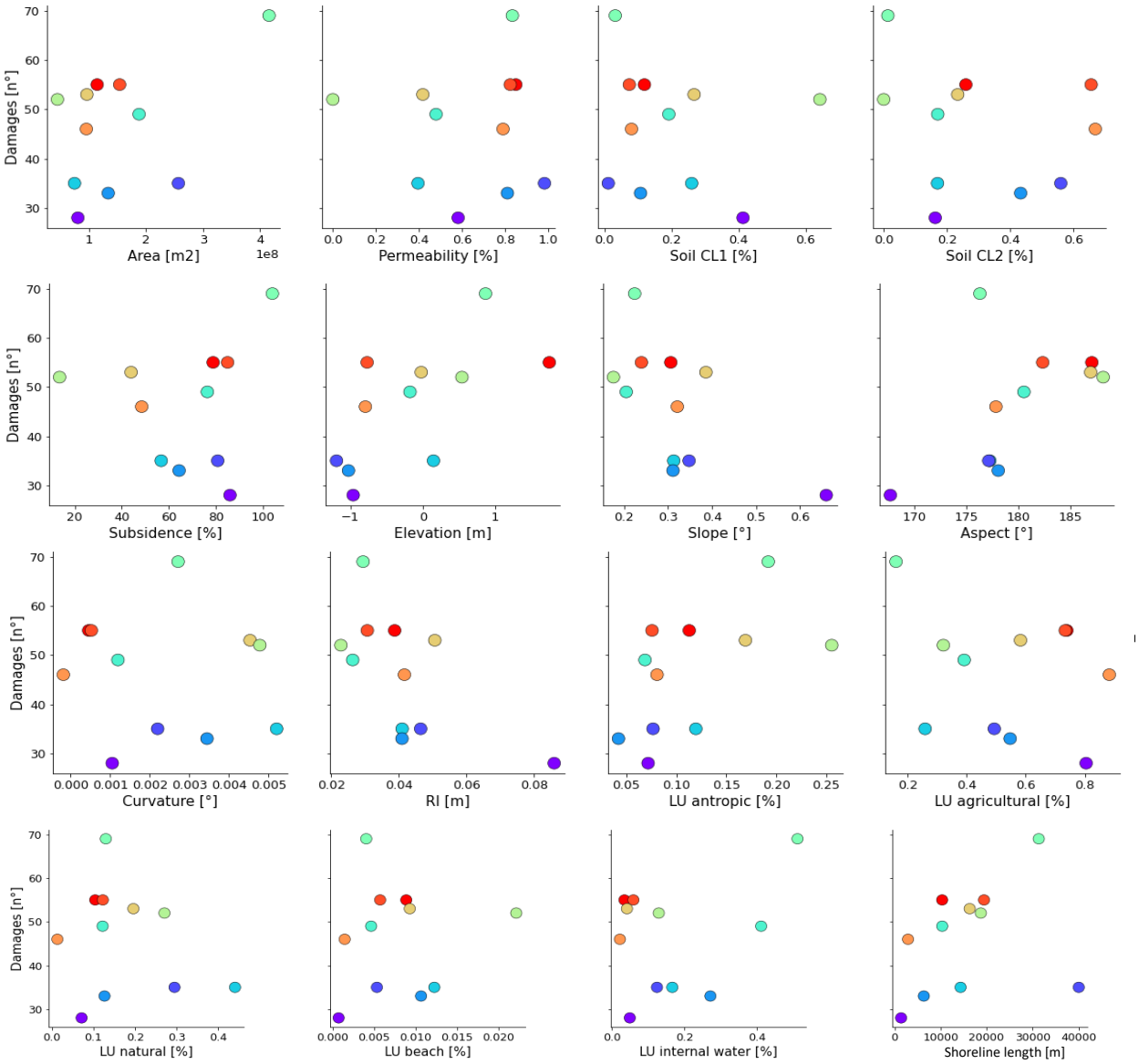
Slope and RI index seemed to increase by going from north to south of the study area with values respectively of 0.31° and 0.04 m for municipality 1 whereas of 0.66° and 0.09 m for municipality 10.

The aspect indicator decreased by moving southward (from 187° of municipality 0 to 168° of municipality 10), outlining the tendency of the slope to be exposed towards a more eastern direction.

The length of the shoreline and the coastal dunes coverage had a similar pattern over the municipalities, with the highest values for municipality 9 (respectively 40 km and 23 km), followed by municipality 5 and 1; the lowest values were reached by municipality 10 and 2.

In relation to the percentage of land use categories: the central municipalities had the highest values of anthropic coverage (26-19%), a pattern that was reversed for agricultural coverage (i.e., lower values for the central municipalities; 32-16%), municipality 7 had the highest % of natural cover (44%), municipality 4 the highest % of beach coverage (2%), and finally, municipality 5 and 6 had the highest % of internal water (51 and 41%).

A scatterplot between the damages that occurred in the 11 municipalities and the territorial indicators (Figure 25) did not reveal the existence of recognizable correlations. The only exceptions that seemed to be positively associated with the number of damages were the elevation and the anthropic land-use coverage indicator, this latter correlation is quite reasonable since more assets could be negatively impacted by extreme weather events. However, we must keep in mind that several factors can play a role simultaneously, and correlation does not mean causation.



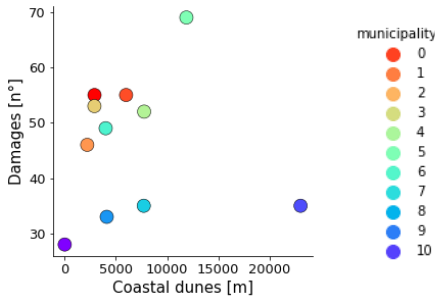


Figure 25: Scatterplot between the number of damages occured in the municipalities within the 2009-2019 timeframe and the relative territorial indicators

It must be highlighted how the relations between the territorial indicators and the number of occurred damages could be more informative if specific damage categories would have been considered (information not available for this thesis). For example, in the case of flooding, generally, the higher the elevation is, the lowest the number of damages is, oppositely to the slope index since, with a higher slope, the more rapid the water flow is, which consequently has higher energy (Collins et al., 2022; Ha & Kang, 2022).

Finally, given the availability of land use indicators on a triannual basis for the studied timeframe, it was explored if the change of the land-use coverage over the years, for the different municipalities, somehow, could have influenced the damage occurrences; the findings are reported in Figure 26. What can be deduced is that – except for some cases – the number of damages changed even though the land use categories for the 11 municipalities remained quite constant for the period 2009-2019.

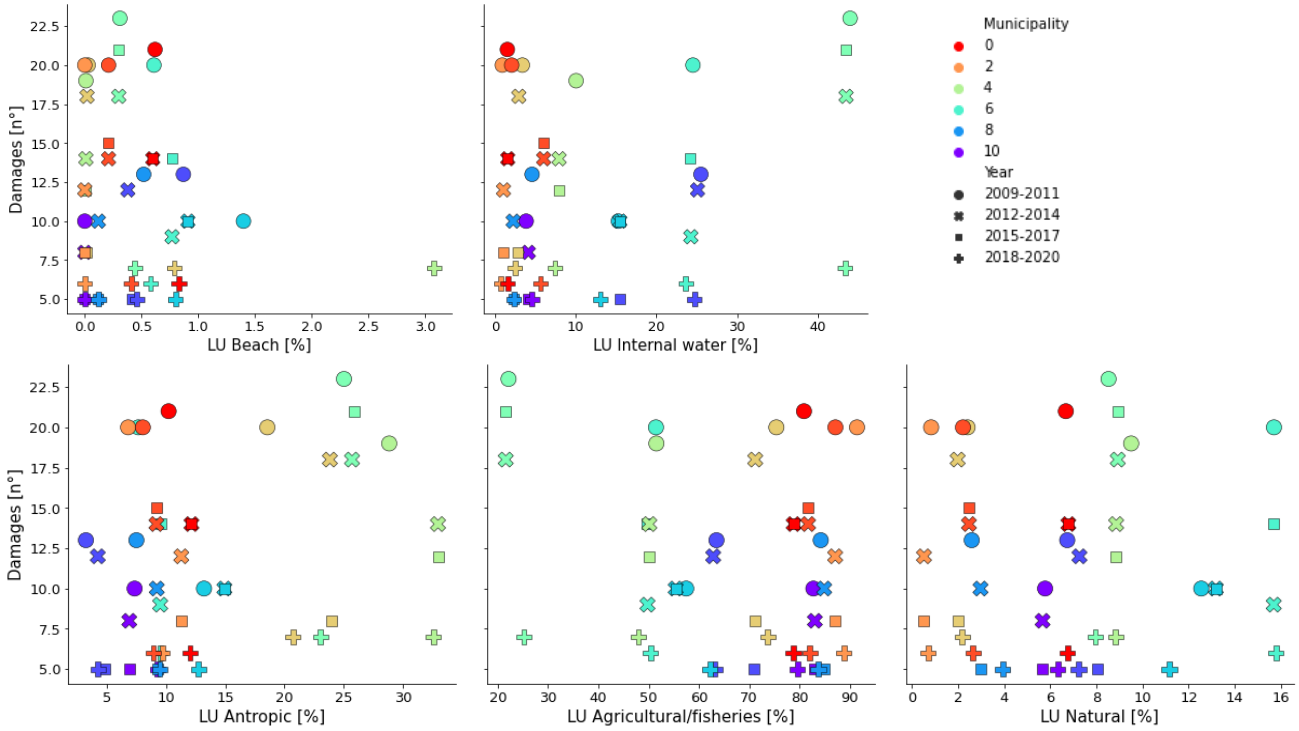


Figure 26: Damage occurrences and evolution of land use categories over the years 2009-2019 for the different investigated municipalities

5.3. Random Forest for feature selection

5.3.1. Balanced dataset and tests of different input combinations

Balanced dataset

As specified in *Section 4.3.2*, the composition of the dataset is a key factor for the outcome of the analysis. The initial dataset, obtained after the pre-processing phase (*Section 5.1*) consisted of 44187 observations, of which only 510 (1,2 %) recorded the presence of damage, whereas 43677 (98,8 %) did not detect any damage; in other words, the dataset was highly imbalanced (Figure 27a).

In order to attribute the same importance to damage and no damage data, the dataset was manipulated to have the same number of damages (510) and non-damages (510) recordings (Figure 27b).

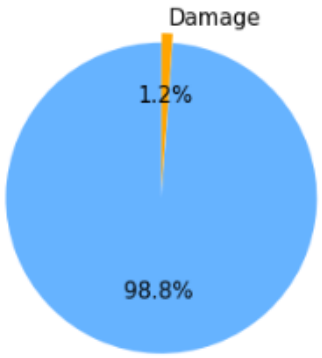


Figure 27a: Percentage of damage recordings before balancing the dataset

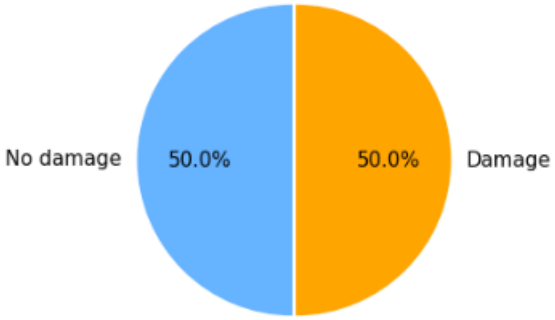


Figure 27b: Percentage of damage recordings after having balanced the dataset

Then, the balanced dataset, consisting of 1020 observations was split into training (75% of the observations) and test (25% of the observations) sets. The validation of the trained models was performed in the framework of the AdriaClim project and was beyond the scope of the thesis, hence in the following only the train and test datasets will be discussed. So, the final training set had 765 observations (with 384 recordings of no damages and 381 recordings of damages) while the test set had 255 observations (with respectively 126 no damages and 129 damages recordings).

Testing different combinations of input variables

Initially, an individual RF was run for each one of the municipalities, with 36 input features. The considered features were: the atmospheric (21) and oceanographic (12) variables registered in the date of the observation, two categorical features related to the season and the month of the observation, and the river discharge. The other territorial indicators were dismissed because they would have been meaningless for the RF since at the municipal scale they were constant for the entire analyzed timeframe. The results showed that, generally, the SSH and MSSH (mean and maximum sea surface height), as well as RX-1day and RX-5day (extreme precipitation indicators), were the most important features for all the municipalities, having a relative importance of 8-10%; however, some differences were observed. On average, extreme precipitation

variables resulted in the first positions for the northern municipalities whereas SSH and MSSH were always the first features for the southern municipalities. Municipality 1 (Caorle) had river discharge among the most important features; municipalities 3 (Jesolo) and 6 (Chioggia) had indicators of wind and daily precipitation among the first positions. The information of seasons and months had always relative importance lower than 1.5%.

Nevertheless, every time that the "municipal" RFs were run, the results and the order of the most important features changed a lot as well as the F1 score of each independent RF, which spanned between 0.73 to 0.99. These highly variable results were probably related to the scarce number of observations (samples) with which the "municipal" RFs were trained. If more observations, for each municipality, would have been available, running individual RFs could really give an understanding of how hazard indicators contribute to local damage occurrence since some of the variables presented heterogeneous values at the municipal level (*Section 5.2.2*).

Hence, due to these not reliable outcomes, it was decided to run a "regional RF", comprising all the observations of the balanced dataset, by testing different input variables' combinations.

In total, four initial RFs were tested, the first one had 54 input variables (21 atmospheric, 12 oceanographic, 18 territorial, and 3 other variables regarding season, month, and municipality); the second combination was equal to the first except that the territorial variables were not considered; the third combination was equal to the first with the addition of 9 variables related to atmospheric and oceanographic indicators (i.e., RX1-day, precipitation maximum, maximum temperature, wind direction, maximum wind speed, solar radiation, MWIH, MSSH, and SSH) registered 2 days before the observation and, finally, the last fourth combination was equal to the third with the exclusion of territorial indicators.

Evaluated on the test dataset, the first combination obtained a F1 score of 0.93, the second of 0.94, the third of 0.94, and finally the fourth of 0.95.

From the outcomes it seems that the presence of territorial indicators (first and third option) reduced the RF performance, in fact, they resulted always in the last positions of the features' relative importance, so their presence created just noise in the RF, decreasing the performance. However, among the territorial indicators, elevation and slope had the highest relative importance; the municipality feature was always at the end of the features ranking as well as indexes of TX90p, maximum humidity, CDD, and other wave indicators; the month feature was always more important than the season.

In general, in all the performed RFs, SSH and MSSH had the highest importance (usually higher than 8%), followed by parameters related to precipitation, than temperature; other oceanographical features emerging among the first 15 were MWAH and WAH (maximum and mean significant wave height). Since the fourth combination got the best F1 score, it was considered as the final result on which further analyses were based (*Section 5.4*). In the next section, in-depth observations related to this final result are discussed.

5.3.2. Validation of the Random Forest and feature selection

As previously mentioned, the fourth combination of variables implemented in the RF gained the best performance according to the F1 score metric with a respective value of 0.95, a recall of 0.99 and a precision of 0.91 (definitions provided in *Section 4.3.3*). The confusion matrix (Figure 28) showed how the model correctly predicted the damages' presence or absence 95% of the time, by correctly evaluating 123 damage occurrences (TP) and 119 no damage occurrences (TN). By looking at the mistaken predictions, it appears that the model had the tendency to overestimate the number of occurred damages (12 FP), while just one time it failed in not predicting correctly the presence of damage (1 FN).

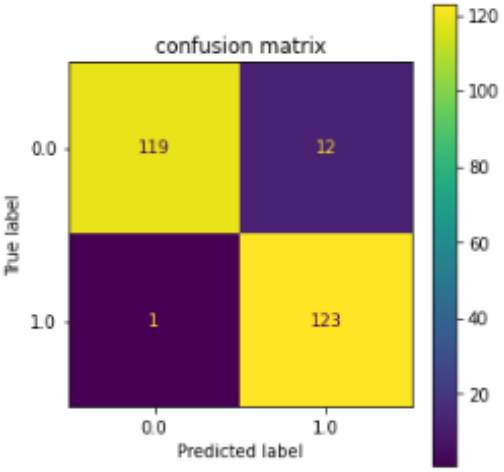


Figure 28: RF confusion matrix (0 = damage absence; 1= damage presence)

A possible cause of the tendency to overestimate the damage presence could be due to the fact that, since the RF was trained mainly with hazard data and not with data of exposure and vulnerability (which are very important when damages are assessed), the damage information could have been interpreted, by the RF, more as the presence or absence of an extreme event (in fact the recorded damages were generally caused by very extreme events). So, it could have happened that in some dates the hazard characteristics were those of an extreme event that however did not cause damage.

Figure 29 displays the relative importance of the input features of the RF, while Table 7 reports only the ones having relative importance higher than 2% (18 features) hereafter called "*selected features*". What is evident is the role of SSH and MSSH (mean and maximum significant sea surface height) in determining the damage prediction as they had an importance that was double if confronted with that of the features coming next. SSH and MSSH were followed by precipitation indicators, as resulted also by running the "municipal" RFs. Then, other important features were related to temperature and WAH/MWAH (mean and maximum significant wave height). Except for the features in the first positions, it was difficult to assess the role of the others since they had similar low importance. To this concern, two additional RFs were run to evaluate possible increments of the F1 score, by reducing the number of input variables. For the first tentative only the selected features were kept, for the second one the last 10 features were eliminated, but in both the cases the F1 score decreased. That means how all these features could play a role in determining the damage prediction, so also the detection of irrelevant features is not straightforward. It must be remembered that the results could be biased by the presence of correlated features, but a further investigation was out of the scope of this analysis.

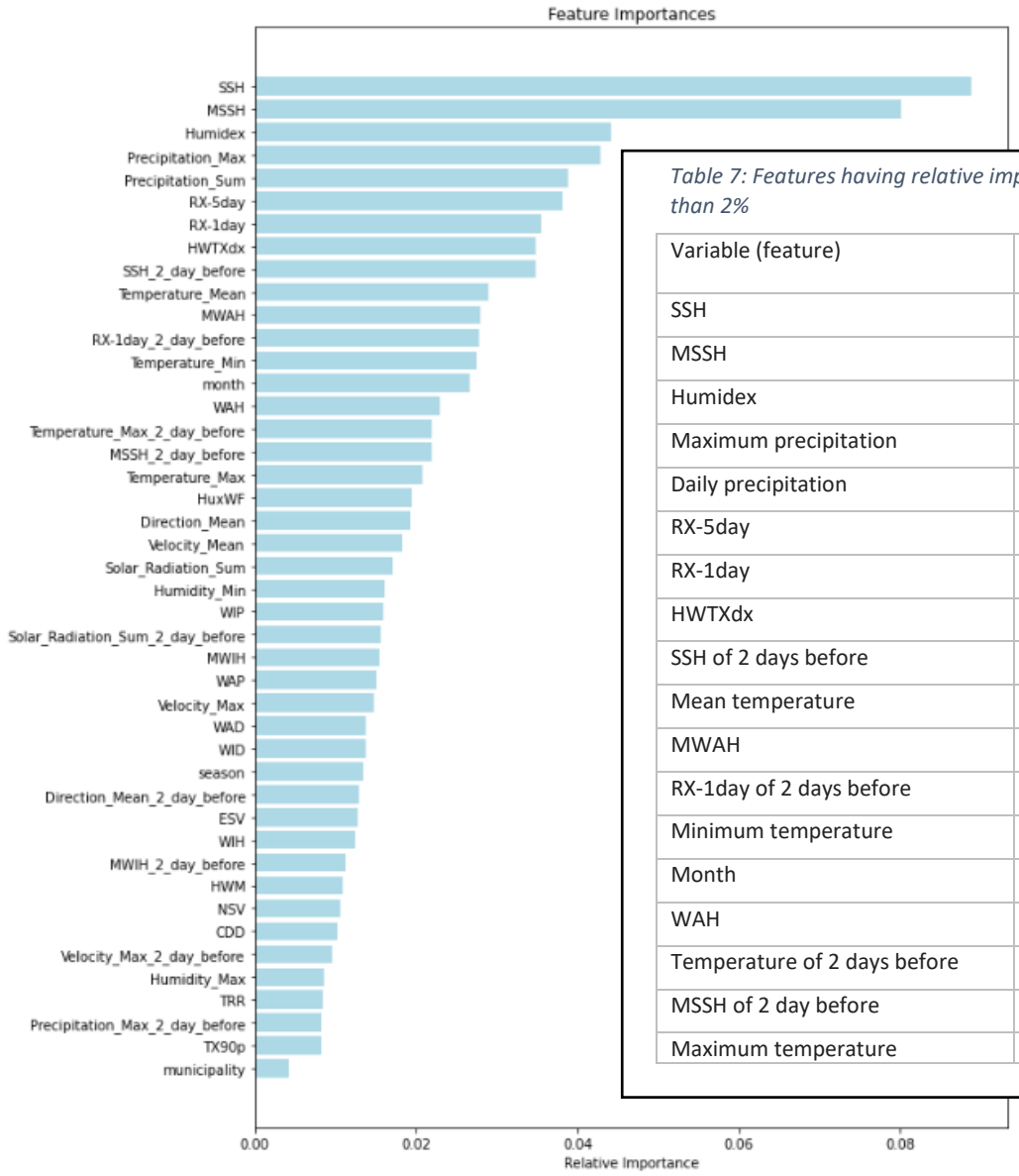


Table 7: Features having relative importance higher than 2%

Variable (feature)	Relative importance
SSH	0.089
MSSH	0.080
Humidex	0.044
Maximum precipitation	0.043
Daily precipitation	0.039
RX-5day	0.038
RX-1day	0.035
HWTXdx	0.035
SSH of 2 days before	0.035
Mean temperature	0.029
MWAH	0.028
RX-1day of 2 days before	0.028
Minimum temperature	0.027
Month	0.027
WAH	0.023
Temperature of 2 days before	0.022
MSSH of 2 day before	0.021
Maximum temperature	0.020

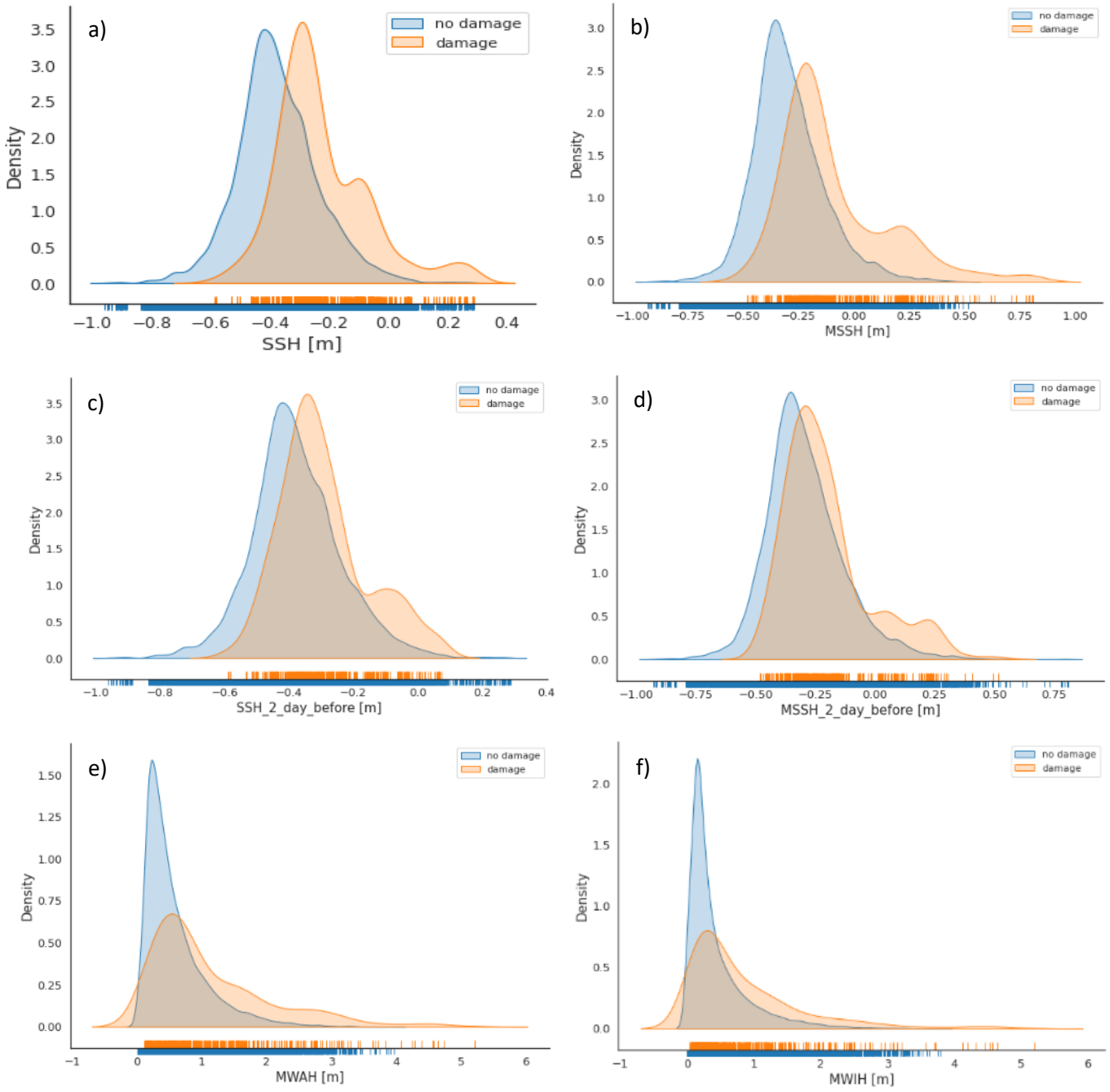
Figure 29: Relative importance of the input features form the RF feature selection

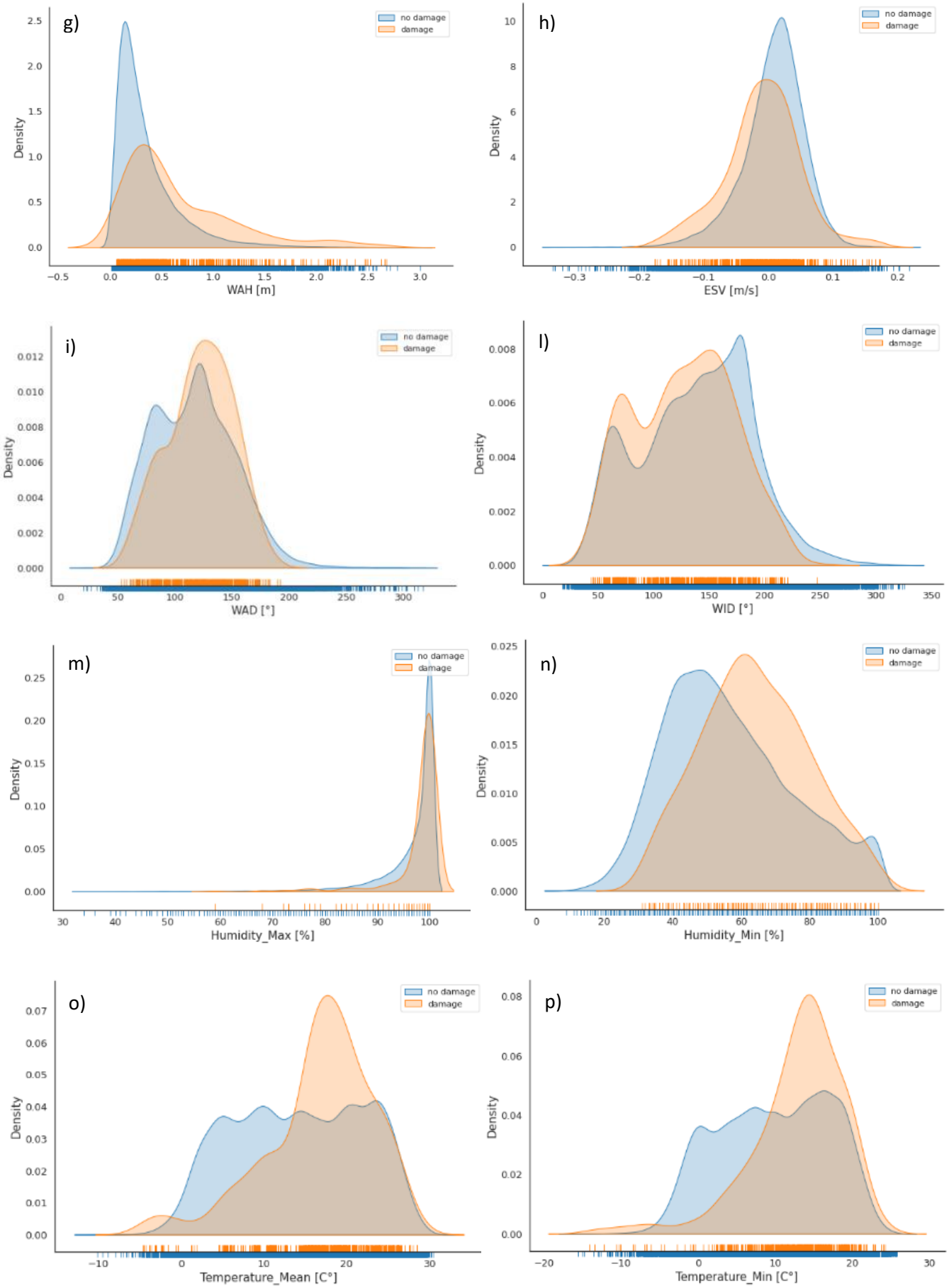
To summarize, the RF correctly classified the majority of the events. The feature selection associated with the RF presented more dubious outcomes, which were probably related to the elevated correlation between the variables. Anyhow, some variables, every time, resulted to have the highest relative importance, specifically SSH, MSSH, and precipitation indicators.

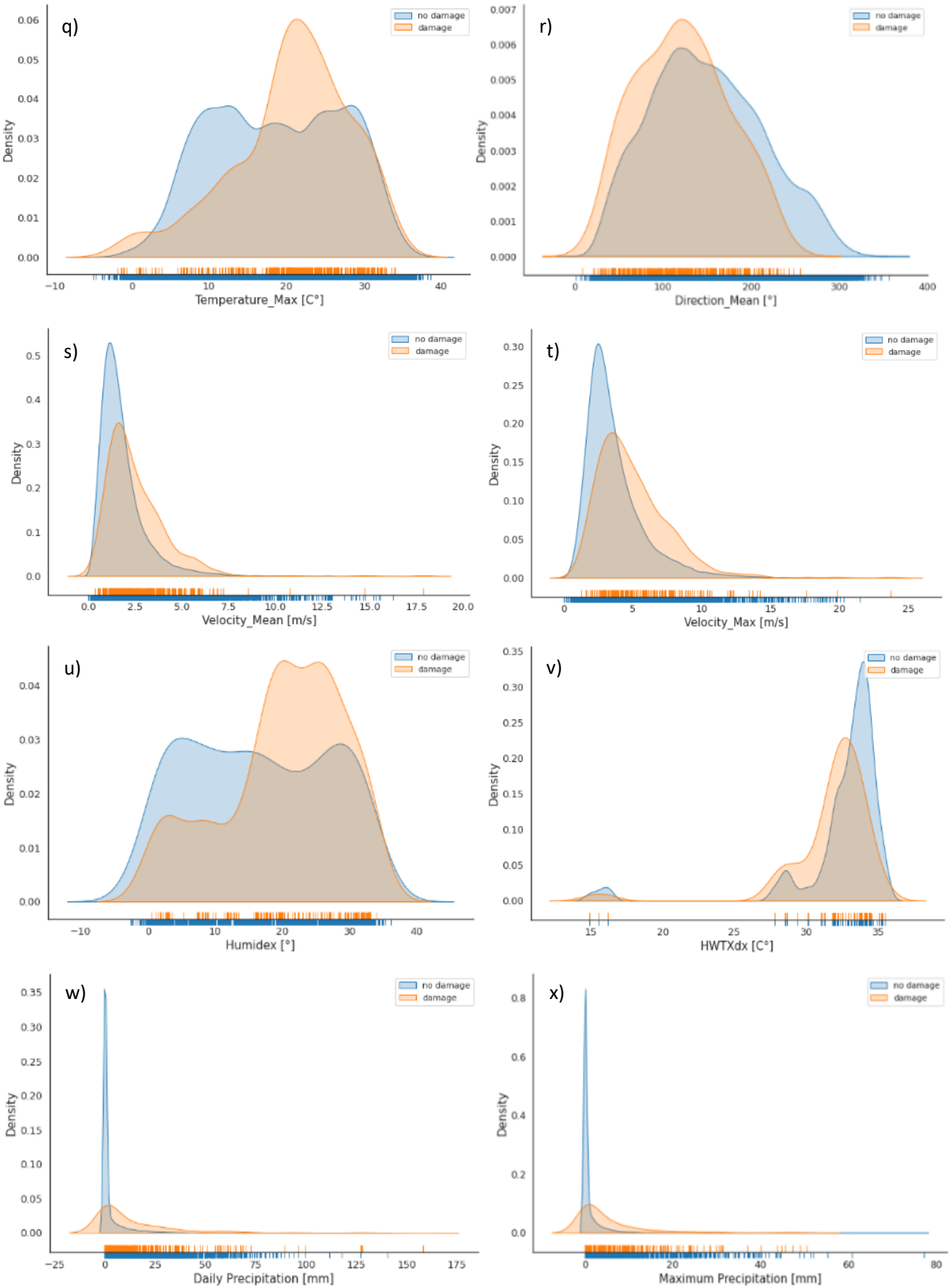
5.4. Analysis of the most influential variables associated with the damage occurrence

5.4.1. Regional-scale analysis

In this section, the behavior of the RF's selected variables (Table 7) is analyzed and the distribution curves of the indicators on two datasets, i.e. the set comprising the observations associated with no damages and the set comprising the observations associated with damages are compared. The most interesting comparisons are graphically reported in Figure 30.




g)
no damage
damage
Density
WAH [m]
h)
ESV [m/s]
i)
WAD [°]
l)
WID [°]
m)
Humidity_Max [%]
n)
Humidity_Min [%]
o)
Temperature_Mean [C°]
p)
Temperature_Min [C°]


q)
no damage
damage
Density
Temperature_Max [C°]
r)
Direction_Mean [°]
s)
Velocity_Mean [m/s]
t)
Velocity_Max [m/s]
u)
Humidex [°]
v)
HWTXdx [C°]
w)
Daily Precipitation [mm]
x)
Maximum Precipitation [mm]

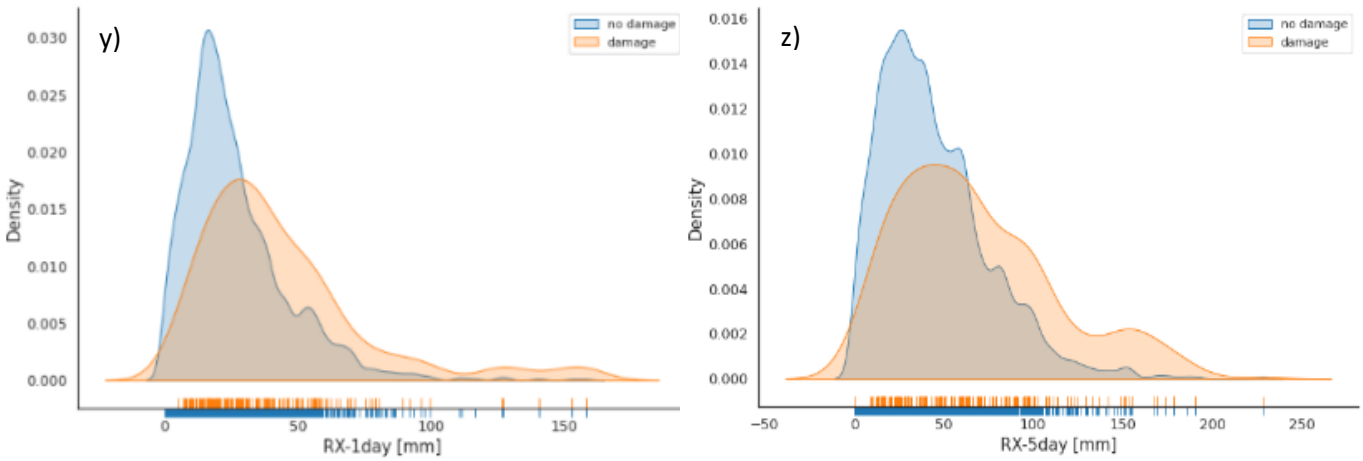


Figure 30: Probability density distribution, for observations with and without damages, of: a) SSH; b) MSSH; c) SSH of 2days before; d) MSSH of 2 days before; e) MWAH; f) MWIH; g) WAH; h) ESV; i) WAD; l) WID; m) maximum relative humidity; n) minimum relative humidy; o) mean temperature; p) minimum temperature; q) maximum temperature; r) wind direction; s) mean wind velocity; t) maximum wind velocity ; u) Humidex; v) HWTXdx; w) daily precipitation; x) maximum precipitation; y) RX-1day; z) RX-5day

The variables mean (SSH) and maximum sea surface height (MSSH) showed similar distributions for the two datasets (presence and absence of damage), but the distribution of the dataset associated to damage events presented a clear shift towards higher values (respectively Figure 30a and Figure 30b). This behavior is reasonable since during extreme weather events the meteorological conditions, especially related to the atmospheric circulations (Bergant et al., 2005), trigger the formation of storm surges, defined as "an abnormal rise of water generated by a storm, over and above the predicted astronomical tides" (NOAA, 2022b). This shift in the sea surface's values (although smaller) between the two datasets, was observed also comparing the distribution of the two indicators recorded 2 days before the event (Figure 30c for SSH and Figure 30d for MSSH); information like this could be useful for activating early warning systems and preparing possible action plans.

The difference in the mean value of MSSH, during damage and no damage events, was about 0.2 m (-0.295 m during no damage events and -0.09 m in damage presence), whereas it was observed that damages occurred only when MSSH exceeded -0.48 m, and how values higher than 0.51 m were associated only with the presence of damage. A similar difference of 0.2 m between the mean values of the two datasets was recorded also for SSH (values of -0.38 m without damages and -0.23 m with damages) but in this case, if a lower threshold below which no damage occurred (-0.59 m) was detectable, the presence of a higher threshold above which only damages occurred was not present.

For the values of 2 days before the observation, both for MSSH and SSH, the difference in the mean values between damage and no damage events was about 0.1-0.15 m (MSSH recorded a mean value of -0.29 m during no damage and -0.20 m during damage whereas, for SSH, values were respectively of -0.38 m and -0.29 m).

The probability distributions of MWAH (maximum significant wave height) and WAH (significant wave height) presented two right-skewed curves for the two datasets, with the one related to damages having a higher

variability (respectively Figure 30e and Figure 30g). The median values of the two distributions differed of more than 0.3 m for MWAH (respectively 0.43 m in damage absence and 0.77 m damage presence) and about 0.2 m for WAH (respectively 0.25 m in damage absence and 0.46 m in damage presence). As for the MSSH, the MWAH indicator, after reaching the value of 3.97 m, was associated only with damage occurrence. These findings highlight how the "maximum" condition of the sea, rather than mean values, could provide some additional information regarding the occurrence of damages.

In relation to precipitation indicators (Figure 30, w-z), the distributions of the variables were right-skewed for both datasets although the dataset with damages presented higher variability. For extreme precipitation indicators, the median values differed of about 20 mm of rain (RX-1day's median values were respectively 22 mm in damage absence and 35 mm in damage presence; for RX-5day's respectively 39 mm and 61 mm), for daily precipitation, the median values varied of 3.3 mm (0.1 mm in damage absence and 3.4 mm in damage presence). However, a specific lower/upper threshold below/above which only damages or only no damages were recorded was not detected.

All indicators of temperature (mean, minimum, and maximum) showed similar behaviors: in absence of damage, the curve presented a plateau, while in damage presence the curve presented a left-skewed distribution (Figure 30, o-q). On average, when damages occurred, all mean values of temperature indicators were higher than 2 °C if compared with normal conditions (maximum temperature: 18.7 °C for no damage and 20.6 °C for damage events; minimum temperature: 9.7 °C for no damage and 12.9 °C for damage events; mean temperature: 14.6 °C for no damage and 16.8 °C for damage events). These results were probably related to the fact that damages occurred prevalently in spring and summer (Figure 21) when the temperature was higher in the case study area. Temperature itself contributes to creating the conditions of more severe extreme weather because higher temperature means higher energy in the atmosphere that can fuel the convective cells, by provoking intense precipitation or generating strong winds (Liu et al., 2019). The same behavior of the temperature indicators was found also for the humidex indicator (Figure 30u).

The indicators examined so far validated the results of the RF: they effectively presented significant changes in damage presence or absence, whether in their mean or extreme values. More subtle is the difference in the behavior of the dispersion curve for the heatwave temperature (HWTXdx) indicator (Figure 30v), which also scored high in the random forest algorithm.

In order to assess the overall RF reliability, also the variables with relative importance lower than 2% were investigated, to see if their distributions were effectively similar in damage presence and absence. The fact that the RF's F1 score decreased if these less important features were removed from the input variables, could signify their importance in damage prediction, even though to a smaller extent than the previous ones. Among this group of variables of lower importance, all wind indicators were present. The distribution of these indicators (e.i., mean and maximum wind velocity, wind direction), the mean, and the extreme values for both datasets looked similar (Figure 30, r-t). The only slight difference was found for the wind direction, for

which after 255° no damages were detected, revealing how damages were more associated with winds blowing from a north-eastern direction.

The investigation of humidity indicators showed how the maximum humidity brought no information in determining the presence or the absence of the damage (identical curves; Figure 30m), but that was not true for the minimum humidity (Figure 30n). In fact, damages occurred only with a minimum humidity value higher than 31% RH, and typically, for the damage dataset minimum humidity was higher by 10% RH than the values of the no damage dataset (mean value of 54% RH in damage absence and 63% RH in damage presence).

The oceanographic variables NSV (northward seawater velocity), WAP (sea surface wave mean period), WIH (significant wind wave height), and WIP (wind wave mean period) presented no different distribution for the two datasets, as the RF predicted. However, the wave direction indicators (Figure 30, i-l), in damage presence, had a distinct range of values: for WID (wind wave direction from) between 48° and 255°, and for WAD (wave direction from) lower than 190°. A slight difference between the two distributions was noted for maximum significant wind wave height MWIH (after 3.8 m only damages occurred) and for the eastward seawater velocity ESV (minimum, maximum, and median values were lower in damage presence, for example, the mean value in normal condition was 0.008 m/s whereas during extreme events was -0.010 m/s) reveling an oceanic current that shifted from an offshore direction in absence of damage to an onshore direction in presence of damage.

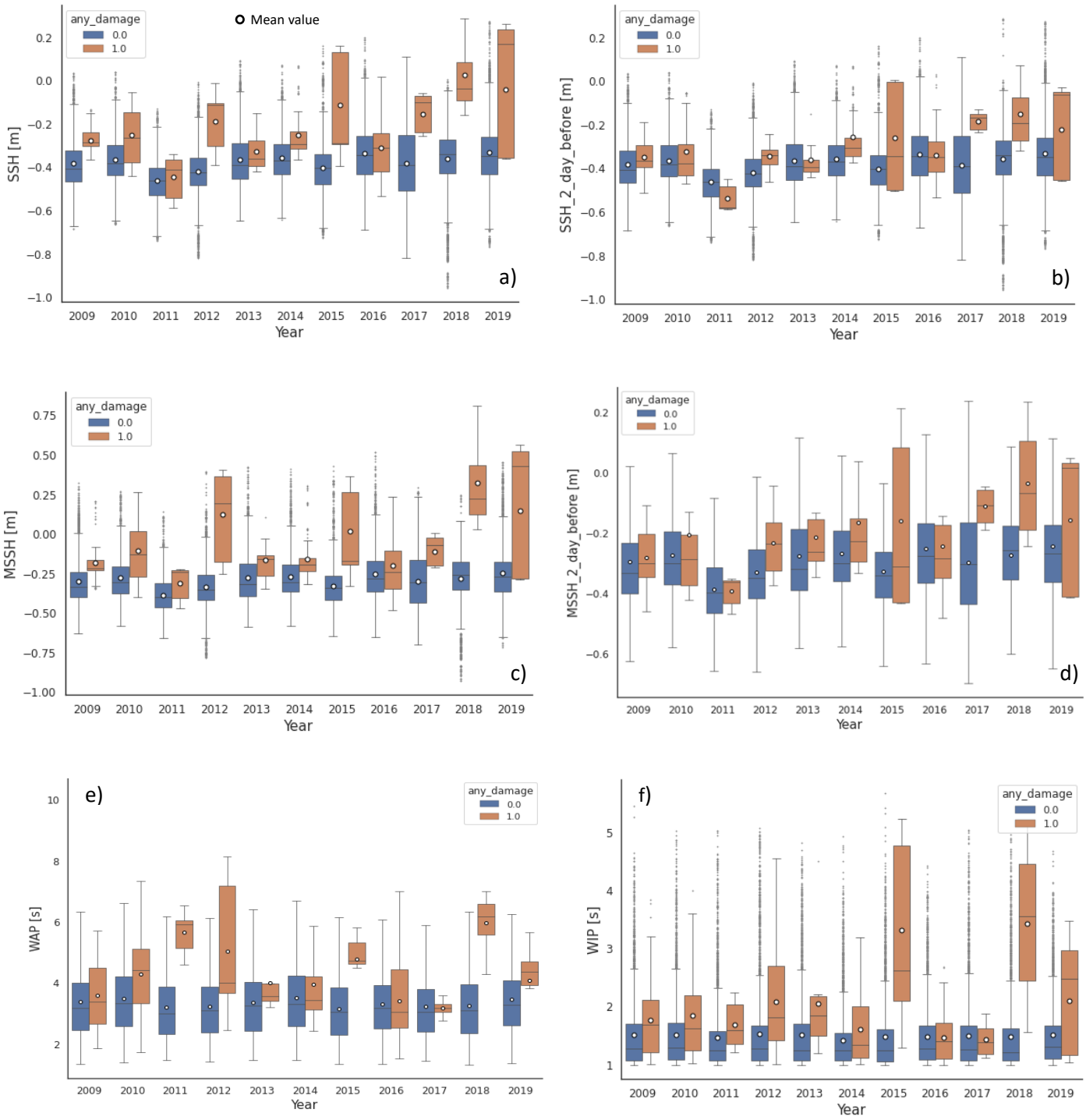
Summarizing, it can be said that the features selected as the most important by the RF were reliable since these variables displayed differences in the two evaluated conditions. However, it must be remembered how often it is the combination of multiple parameters that provokes damage rather than a single one. Hence it was explored how the interaction of two variables could add some more information in the determination of damage occurrence. Specifically, ANNEX VI reports scatterplots[13] that confront the main hazard variables together, by distinguishing observations with and without damages; what can be deduced is that, in some cases, already the interaction of two variables creates more identifiable clusters of damage presence and absence.

Yearly analysis

The investigation regarding the change of the values of the main variables in the years between 2009 and 2019, both in normal and damage-provoking extreme conditions, led to some general results. On average, the range of values of a variable in normal conditions (damage absence) did not register significant changes, with a variability quite constant over the years. On the contrary, for almost all the hazard variables, the values assumed by data belonging to the damage dataset varied remarkedly over the years, sometimes not showing

[13] *Note*. Since the initial dataset was highly imbalanced, this analysis was executed on the balanced dataset, obtained for the RF preparation (Section 5.3.1).

differences with the values assumed by data belonging to the no damage dataset, sometimes having a strong discrepancy. This evidence could suggest that, in certain years, those hazard variables could have played an important role in causing the damage (relatively to other years), or they were correlated to other parameters which strongly contributed to the damage occurrence. Specific differences found for the individual hazard variables are discussed, and the main ones are visually reported in Figure 31.



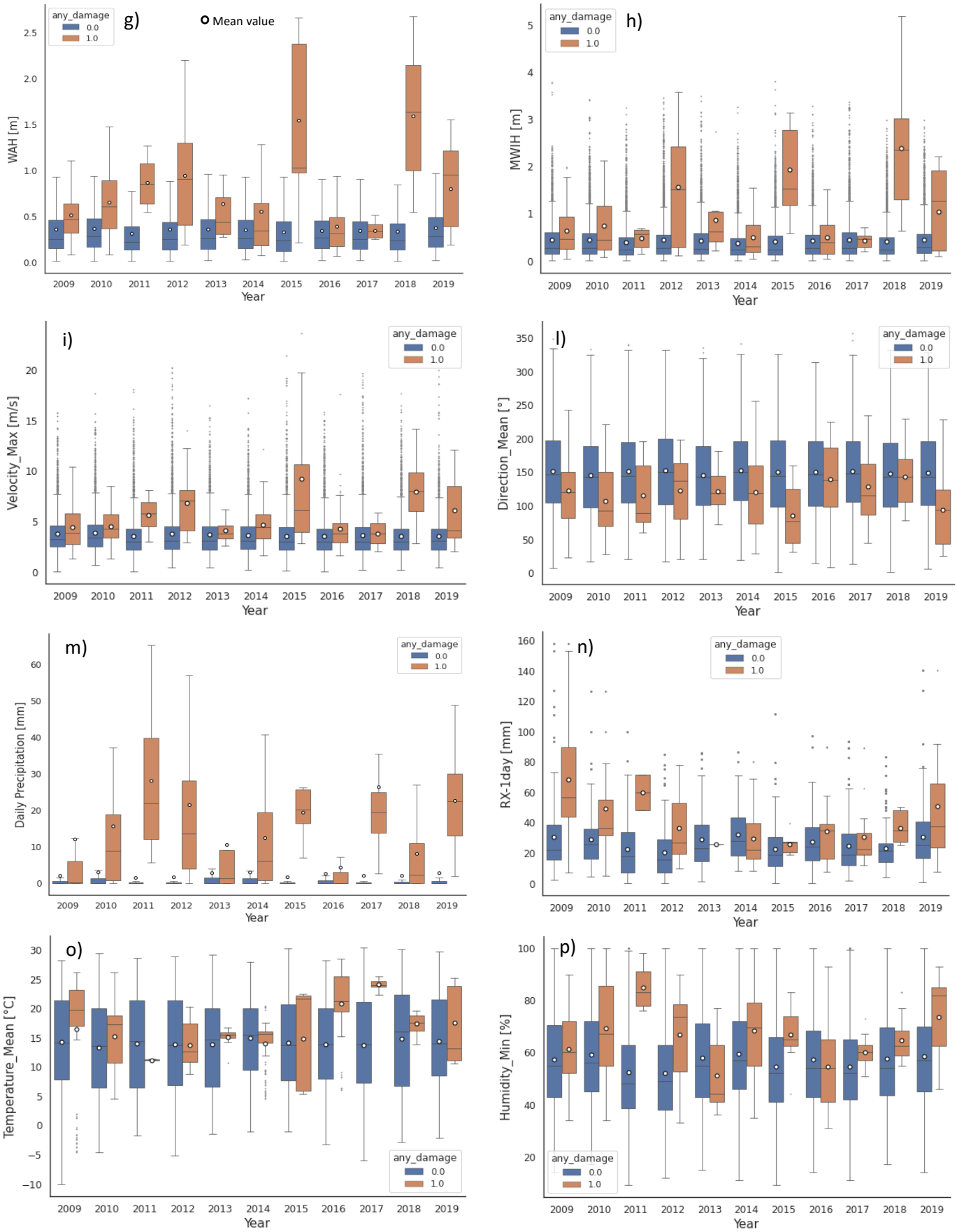


Figure 31: Boxplots, for yearly observations with and without damages, of: a) SSH; b) SSH of 2 days before; c) MSSH; d) MSSH of 2 days before; e) WAP; f) WIP; g) WAH; h) MWIH; i) maximum wind velocity; l) mean wind direction; m) daily precipitation; n) RX-1day; o) mean temperature; p) minimum humidity

Sea surface height parameters (SSH, MSSH, SSH of 2 days before the observation, MSSH of 2 days before the observation) in the no damage dataset showed an oscillatory pattern over the years which could be explained by the influence that inter-annual, decadal, and multidecadal climate fluctuations have on sea level, which can vary several tenths of mm/year (Meli et al., 2021).

The SSH mean values in no damage dataset ranged from -0.46 m to -0.33 m, and those of MSSH from -0.39 m to -0.24 m, assuming lower values compared to the same indicators evaluated on the damage dataset. Such difference is however negligible in some years, and more significant in others, like 2012, 2015, and 2017-2019 when very high values of SSH and MSSH were observed in damage presence, reaching their maximum mean value in 2018 (MSSH: 0.32 m; SSH: 0.26 m). In addition, it seemed that for the last three years of the dataset, the values of extreme sea surface increased (Figure 31a, 31c). All these same considerations were detected also for SSH and MSSH registered two days before the observation, only with a slightly reduced gap between values in damage presence and absence (Figure 31b, 31d).

Mean (WAH) and maximum significant wave height (MWAH), mean (WIH) and maximum significant wind wave height (MWIH), sea surface wave (WAP) and wind wave mean period (WIP), exhibited patterns similar to those of SSH/MSSH in the variation over the years (representative variables reported in Figure 31, e-h). That was somehow expected since oceanographic parameters strongly affect one another and, especially during extreme events, wind-waves modify the total water-level elevation (Pranavam et al., 2022).

Extreme precipitation indicators of maximum cumulative precipitation in 1 (RX-1day) and 5 days (RX-5days) displayed considerable differences between the dataset of no damage and that of damage particularly in the years 2009-2012 and 2018-2019, where higher values were observed in damage presence (Figure 31n). A similar pattern was discovered also for the maximum and daily precipitation (Figure 31m), although for these indicators the difference in the values of the two datasets was clearer for every analyzed year; in addition, a more oscillating behavior over the years was spotted for the values associated to damage.

Temperature indicators for the no damage dataset showed quite constant values over the years (mean values of mean temperature ranging from 13.3 °C (2010) to 14.9 °C (2014); maximum temperature ranging from 17.5 °C (2011) to 19.3 °C (2014) and minimum temperature ranging from 9.2 °C (2010) to 10.7 °C (2014)). Very different values between damage and no damage sets were found for 2009, 2016, and 2017 (Figure 31o). These latter two years registered the highest mean temperature values in presence of damage (for 2016: minimum temperature of 16.4°C, mean temperature of 25.4°C, and maximum temperature of 29.9°C; for 2017: minimum temperature of 18.4°C, mean temperature of 25.53°C and maximum temperature of 29.9°C). The higher mean values of temperature indicators during damage events found for 2016 and 2017 were probably related to the presence, in these two years, of a higher number of damages in the spring and summer seasons (see ANNEX V).

Wind velocity indicators (velocity mean and maximum) revealed a pretty constant range of values for the no damage dataset; for the damage dataset the wind velocity was generally higher, with maximum variations

recorded in 2015 followed by 2018-2019 (Figure 31i). Mean wind direction showed a prevalently north-eastern direction in damage presence (Figure 31l).

Minimum humidity presented higher values on the damage dataset with a quite oscillating pattern over the years (Figure 31p). Except for 2009, 2013, and 2016, the change in the values for the two datasets was significant, with the maximum mean difference reached in 2011 (52.85% RH in damage absence and 85% RH in damage presence).

As evidenced by the previous comprehensive assessment, also the yearly analysis of heatwaves temperatures (HWTXdx) reported no particular differences between the damage and no damage datasets.

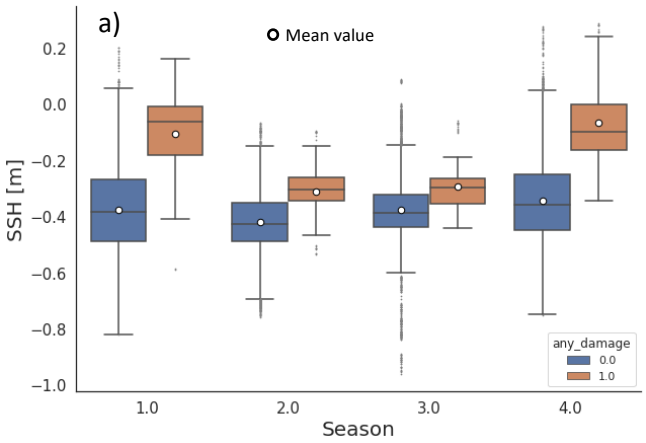
In general, it can be said that the yearly analysis presented results similar to those of the previous assessment (Figure 30), with the values of the main hazard variables on the damage dataset diverging from the ones on the damage dataset, and presenting higher variability. However, the annual analysis cast light on the fact that, for some of the years, variables did not change that much between the two datasets, a detail that could not emerge from the previous analysis. This information is important also for running the RF because, if for the construction of the training dataset, the random selection of the observations keeps only the observations falling in these years with "anomalies", the RF could misinterpret the role played by the variables in the prediction of damages.

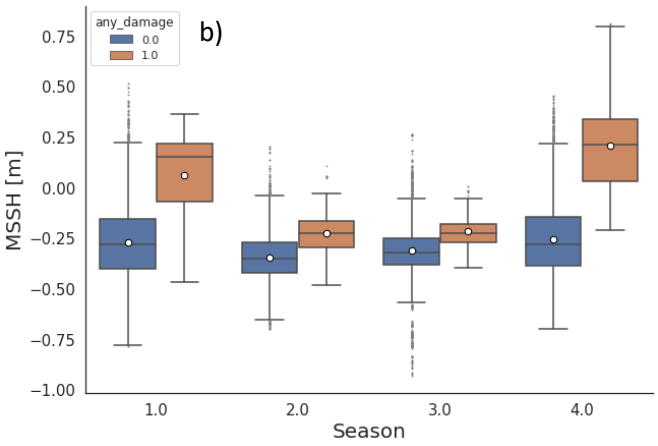
In damage presence, the annual variation in the value of the variables could be associated with the fact that, in different years, more damages happened in certain months/seasons rather than others (see ANNEX V), and the variable itself could have had a seasonal oscillation which could have influenced the results obtained in the yearly analysis. For example, SSH presented a higher discrepancy for the two datasets in the years 2012, 2015, and 2018, and it was found that in those years more damages occurred in the winter and autumn seasons, seasons in which the SSH values, for the damage dataset, were higher, as it will be reported in the next paragraph. Oppositely, the year 2016 registered more damages in the hotter seasons, seasons in which the SSH assumed similar values in both datasets.

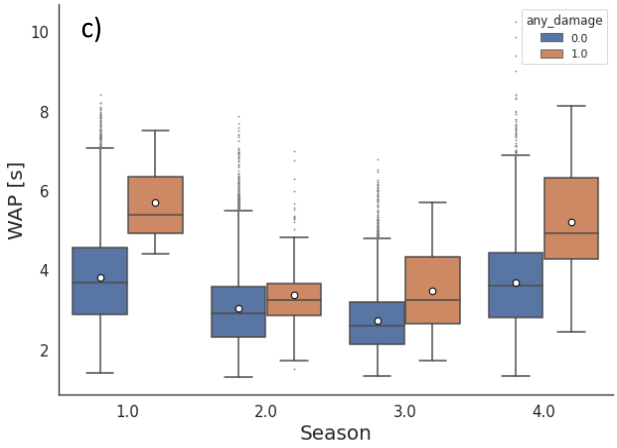
The yearly analysis had the secondary aim to detect a possible intensification, over the years, of the values of the hazard variables during extreme events causing damages, since these phenomena are increasing both in frequency as well as intensity. However, by confronting minimum, maximum, and mean values associated with damage presence, this trend was not observable for the considered dataset and area of study.

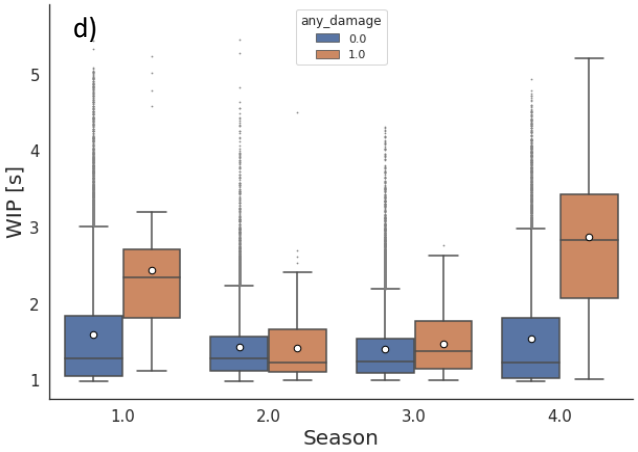
Seasonal analysis

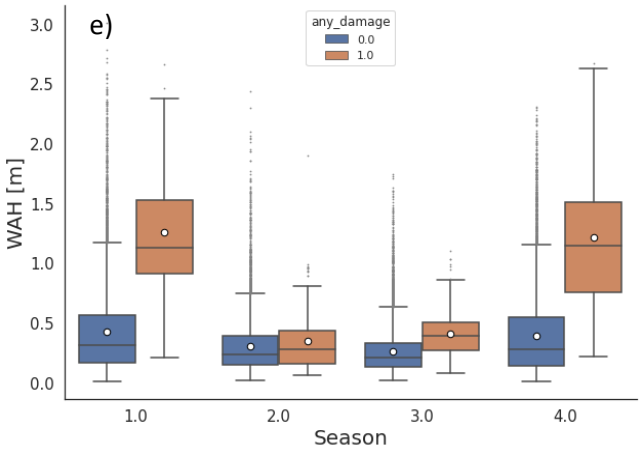
By comparing the seasonal values of the main hazard variables for the damage and no damage datasets, it was found that for the no damage dataset only temperature (higher in spring and summer) and minimum humidity (higher in winter and autumn) had a clear seasonal pattern. The main observations are reported in Figure 32.

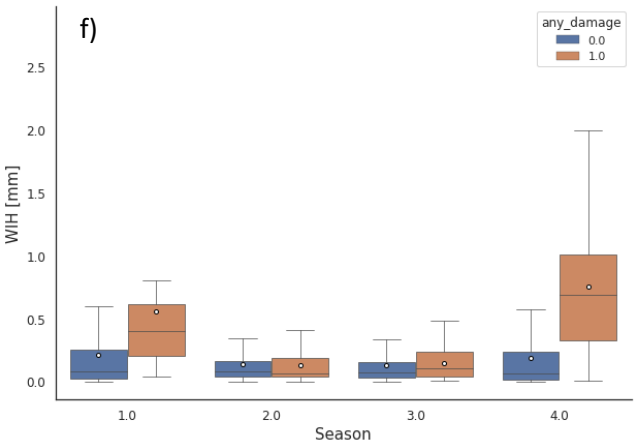

a)
Mean value
SSH [m]
Season
any_damage
0.0
1.0

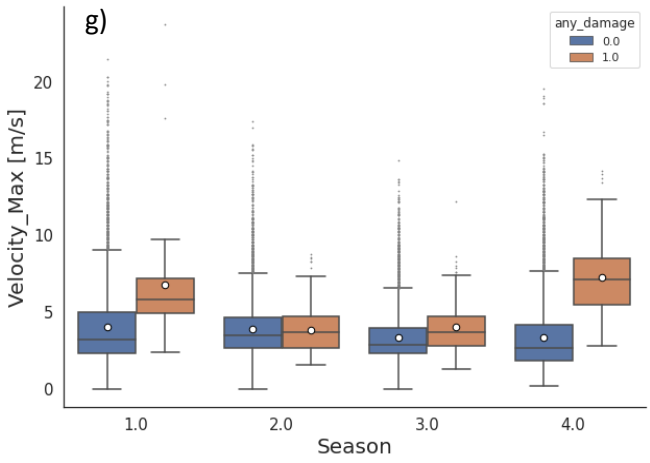

b)
MSSH [m]
Season
any_damage
0.0
1.0

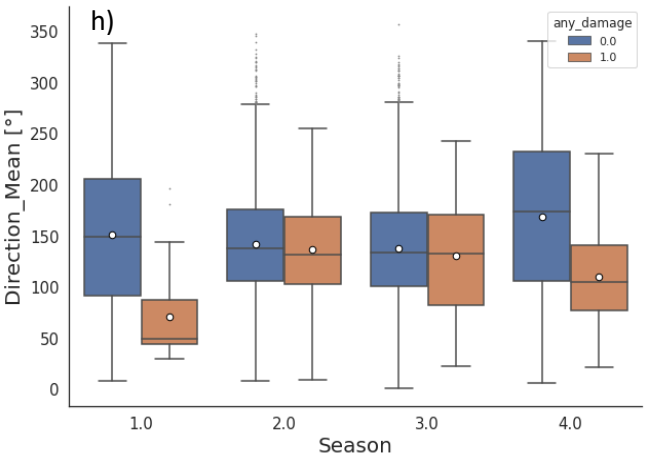

c)
WAP [s]
Season
any_damage
0.0
1.0


d)
WIP [s]
Season
any_damage
0.0
1.0


e)
WAH [m]
Season
any_damage
0.0
1.0


f)
WIH [mm]
Season
any_damage
0.0
1.0


g)
Velocity_Max [m/s]
Season
any_damage
0.0
1.0


h)
Direction_Mean [°]
Season
any_damage
0.0
1.0

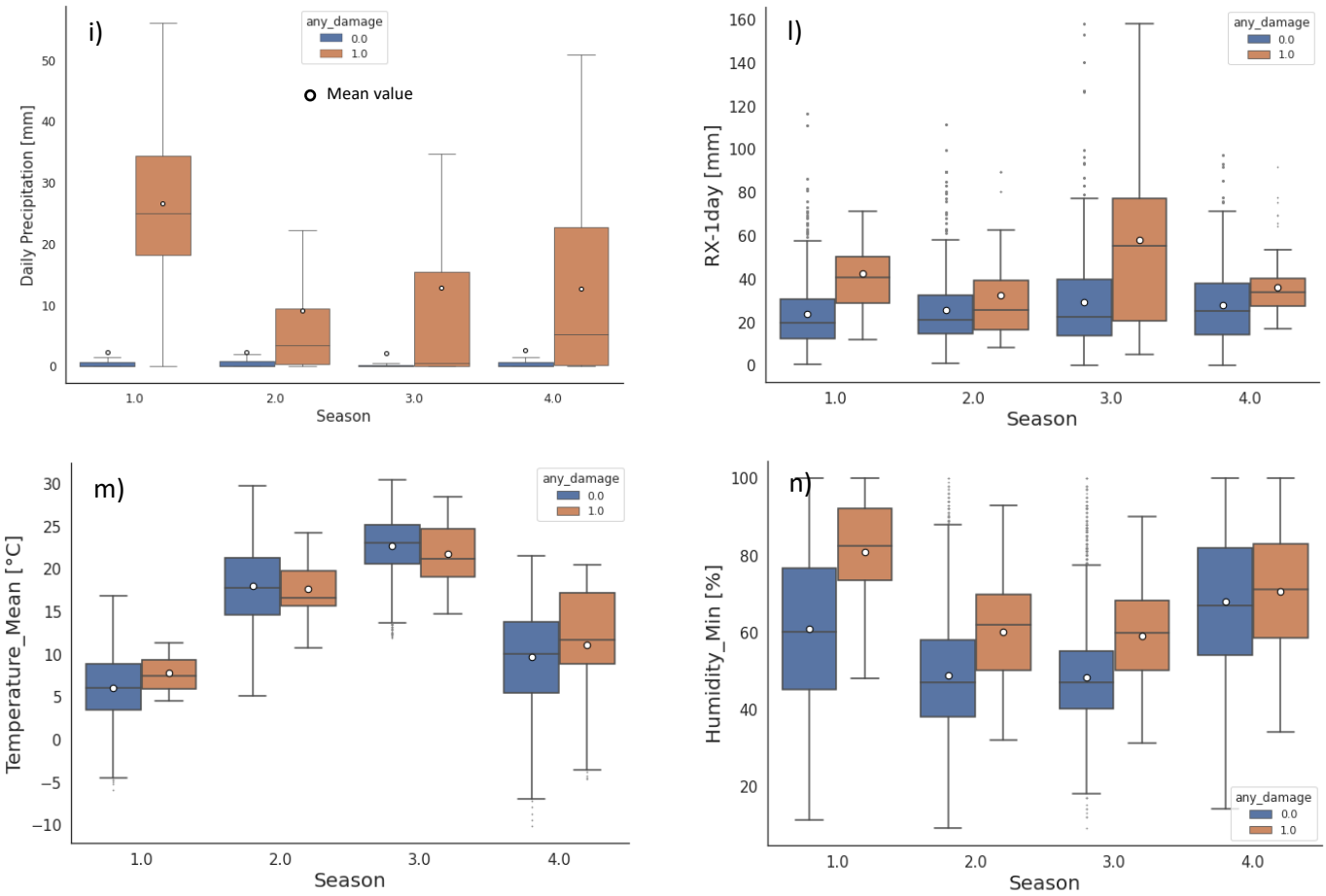


Figure 32: Boxplots, for seasonal observations with and without damages, of: a) SSH; b) MSSH; c) WAP; d) WIP; e) WAH; f) WIH; g) maximum wind velocity; h) mean wind direction; i) daily precipitation; l) RX-1day; m) mean temperature; n) minimum humidity

As the global (Figure 30) and yearly analyses (Figure 31) revealed, also the seasonal comparison of SSH and MSSH displayed higher values in the damage dataset (Figure 32, a-b). However, if for spring and summer these ranges of values did not substantially differ from the no damage dataset, for autumn and winter seasons there was a remarkable discrepancy, and a higher variability (mean winter values of SSH were respectively -0.38 m for the no damage dataset and -0.1 m for the damage dataset, while mean autumn values were -0.42 m and -0.06 m; mean winter values of MSSH were respectively -0.27 m for the no damage dataset and -0.06 m for the damage dataset, while mean autumn values respectively -0.25 m and 0.20 m; in spring and summer all these differences where lower than 0.07 m). These findings could highlight a different role played by the sea level in determining a possible damage occurrence, over the seasons. These seasonal patterns of the sea surface indicators during extreme weather were concordant with those described by Bergant et al. (2005).

The same seasonal pattern of SSH and MSSH was found not only for other oceanographic variables (i.e., WAH, WAP, WID, WIH, MWIH, and WIP) but also for precipitation indicators and wind speed. Wind velocity (both mean and maximum values) registered always higher values in the damage dataset but, like SSH and MSSH, with a significant difference from the no damage dataset especially in winter and autumn (e.g., mean values

of maximum wind velocity in winter were respectively 4.04 m/s for the no damage dataset and 6.80 m/s for the damage dataset, in spring respectively 3.89 m/s and 3.84 m/s, in summer 3.36 m/s and 4.01m/s, in autumn 3.37m/s and 7.27m/s). The higher values of oceanographic and wind parameters in autumn and winter during extreme conditions may be related, on those days, to the presence of the north-easterly Bora wind (Dorman et al., 2007).

Precipitation indicators, for each season, showed always higher mean values in the damage dataset than in the no damage one (Figure 32, i-l). In particular, the maximum difference in the mean values of precipitation was reached in summer by the RX-1day indicator (29 mm for the no damage dataset and 58 mm for the damage dataset) and in winter by the daily precipitation (2 mm for the no damage dataset and 27 mm for the damage dataset).

ESV manifested always lower values, for all the seasons, in the damage dataset than in the no damage one, meaning that the oceanic currents tended to move from an offshore to an onshore direction (in this latter case the values of ESV become negative).

Minimum humidity (Figure 32n) was always higher in the damage dataset than in the no damage one, with the maximum difference in the winter season, where the respective mean values were 80.7% RH and 60.8% RH. For the other seasons, mean values differed by less than 10% RH.

Values of temperature for the two datasets manifested no substantial variation for each individual season (Figure 32m). However, during winter and autumn, the mean temperatures were higher in the damage dataset than in the no damage one, and vice-versa for summer and spring (e.g., for mean temperature, the mean values in the no damage dataset and in the damage one were respectively: 6.1 °C and 7.8 °C in winter, 18.0 °C and 17.7 °C in spring, 22.7 °C and 21.8 °C in summer, 6.7 °C and 11.1 °C in autumn).

The seasonal analysis illustrated how the values of several hazard variables, during extreme events, changed considerably. Therefore, the importance of a variable in influencing the damage occurrence could change according to the season.

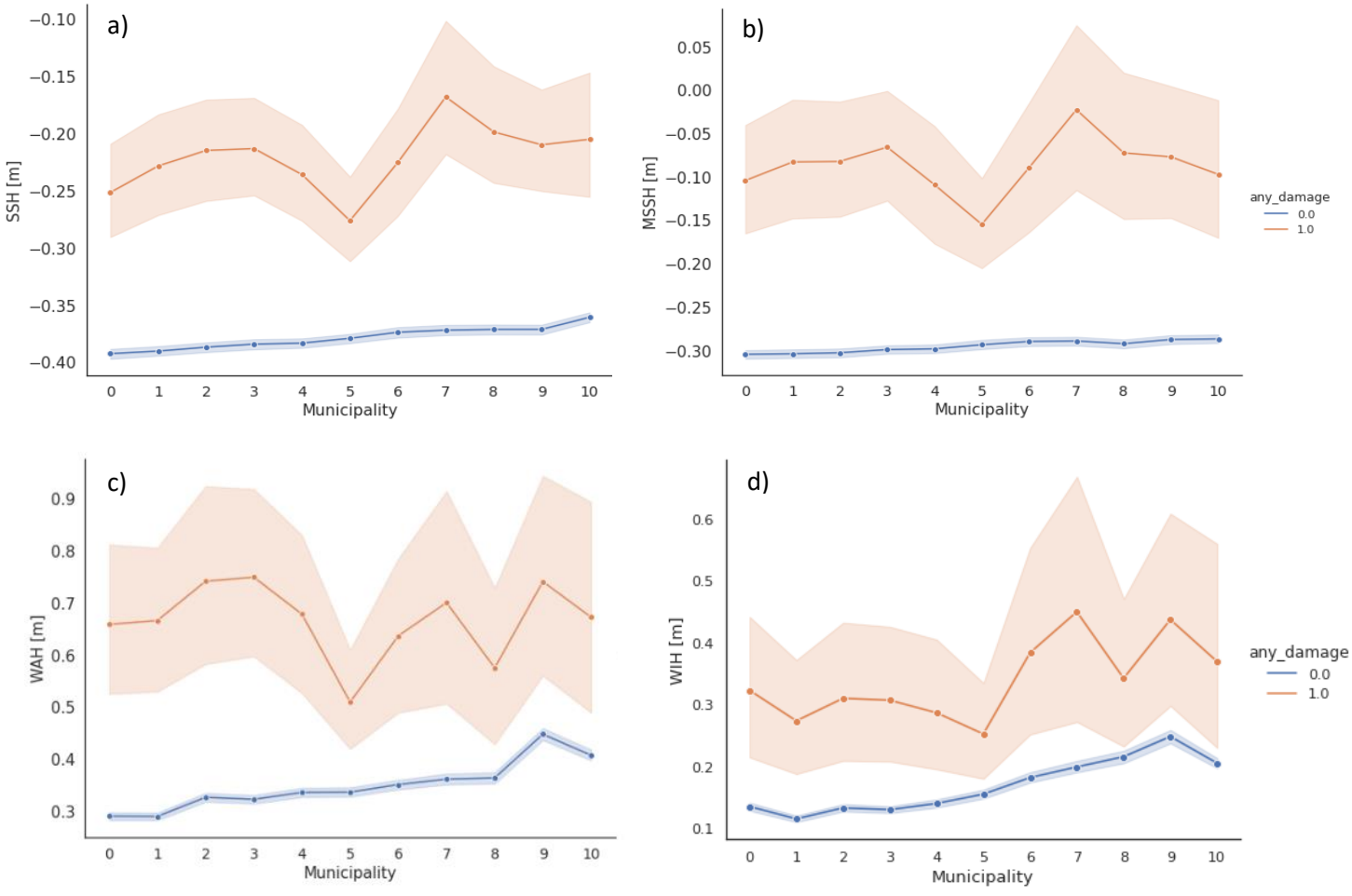
If all this information is analyzed in the light of future scenarios, having a narrative picture of what could happen becomes complex. On one hand, extreme precipitations are projected to increase over the all Mediterranean area (Zittis et al., 2021); however, for the north-western Adriatic coast the annual precipitations are predicted to decrease by 3% by the end of the century with a marked seasonality and an associated temperature increase of 3.2 °C (Lionello, 2012). On the other hand, it is expected a reduction in climatic extreme wind waves (Denamiel et al., 2020), but an increased storm surge risk under sea-level rise scenarios (Rizzi et al., 2017). Nevertheless, the future sea level of the north Adriatic sea is very uncertain (in the frame of RESPONSe project[14], recent models predicted a slight decrease in SSH until 2040 and a modestly increase afterward). Seasonal differences could be amplified in the next years, determining a variable risk

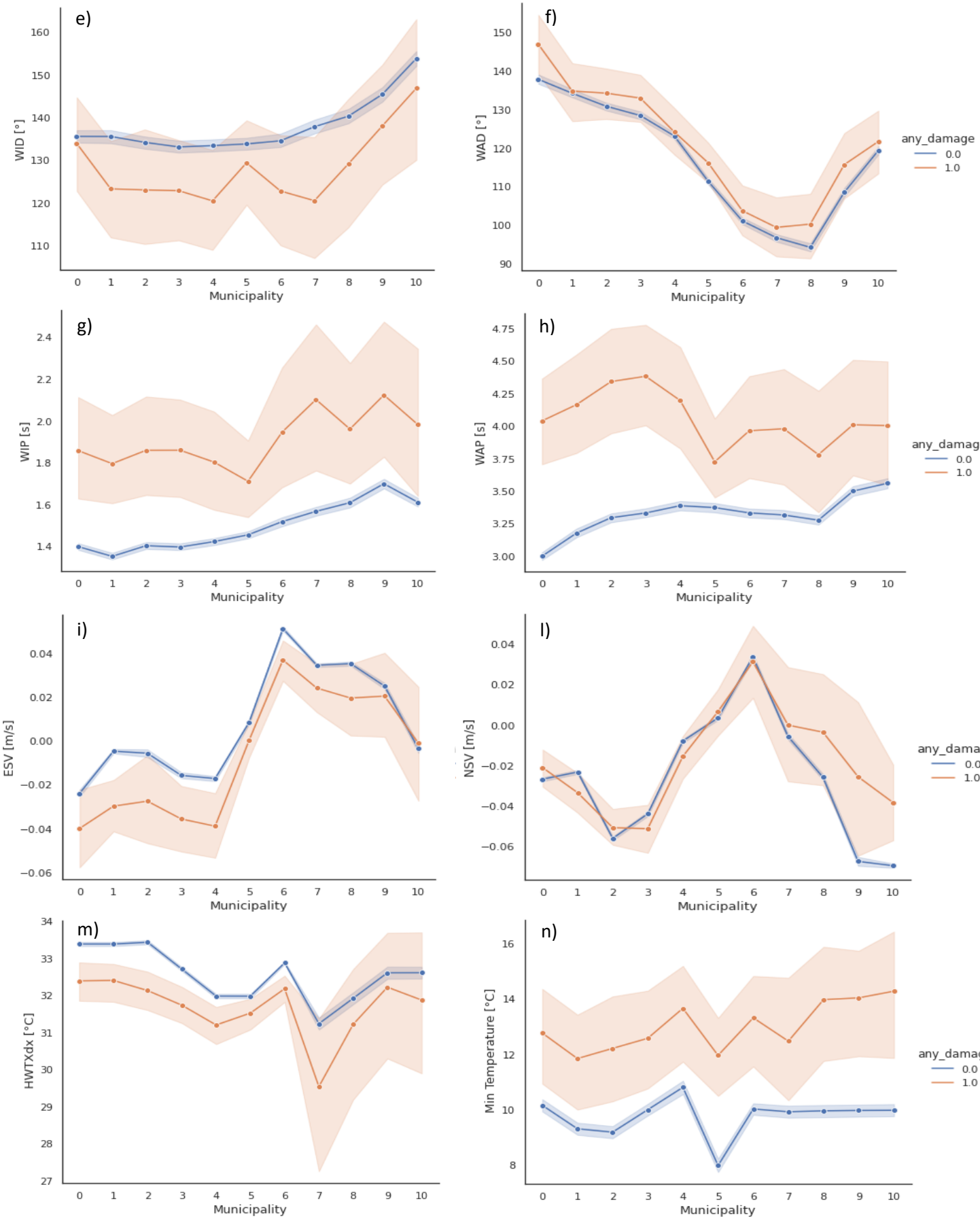
[14] https://www.italy-croatia.eu/documents/275198/2777230/RESPONSe_D321.pdf/6c5fed68-72eb-e726-b015-cde70ed0c613?t=1613552795746

associated with extreme waves (Torresan et al., 2019). Moreover, the generation of damages is related not only to physical hazards but also to exposure and vulnerability features, and the study of Furlan et al. (2021) demonstrated how, if all these parameters are taken into account, the north Adriatic Italian coast is extremely vulnerable to future inundation risk (the main risk of coastal areas), aggravated during extreme weather events.

5.4.2. Municipal-scale analysis

The initial analysis conducted at the local scale aimed to determine the differences in the mean values of the hazard variables for the damage and the no damage sets over the 11 municipalities; the main results are graphically reported in Figure 33, where the mean values are represented with their associated 95% confidence interval. Generally, for the damage set, the municipalities recorded more homogeneous values for the investigated variables. On the contrary, the mean values of the main hazard indicators in the damage set, not only changed significantly over the municipalities but had also a higher variability. In some cases, at the local scale, the differences in mean values between municipalities, already present in the no damage set, were amplified in the damage set (e.g. for wind speed).




e)
WID [°]
Municipality
f)
WAD [°]
Municipality
any_damage
0.0
1.0
g)
WIP [s]
Municipality
h)
WAP [s]
Municipality
i)
ESV [m/s]
Municipality
l)
NSV [m/s]
Municipality
m)
HWTXdx [°C]
Municipality
n)
Min Temperature [°C]
Municipality

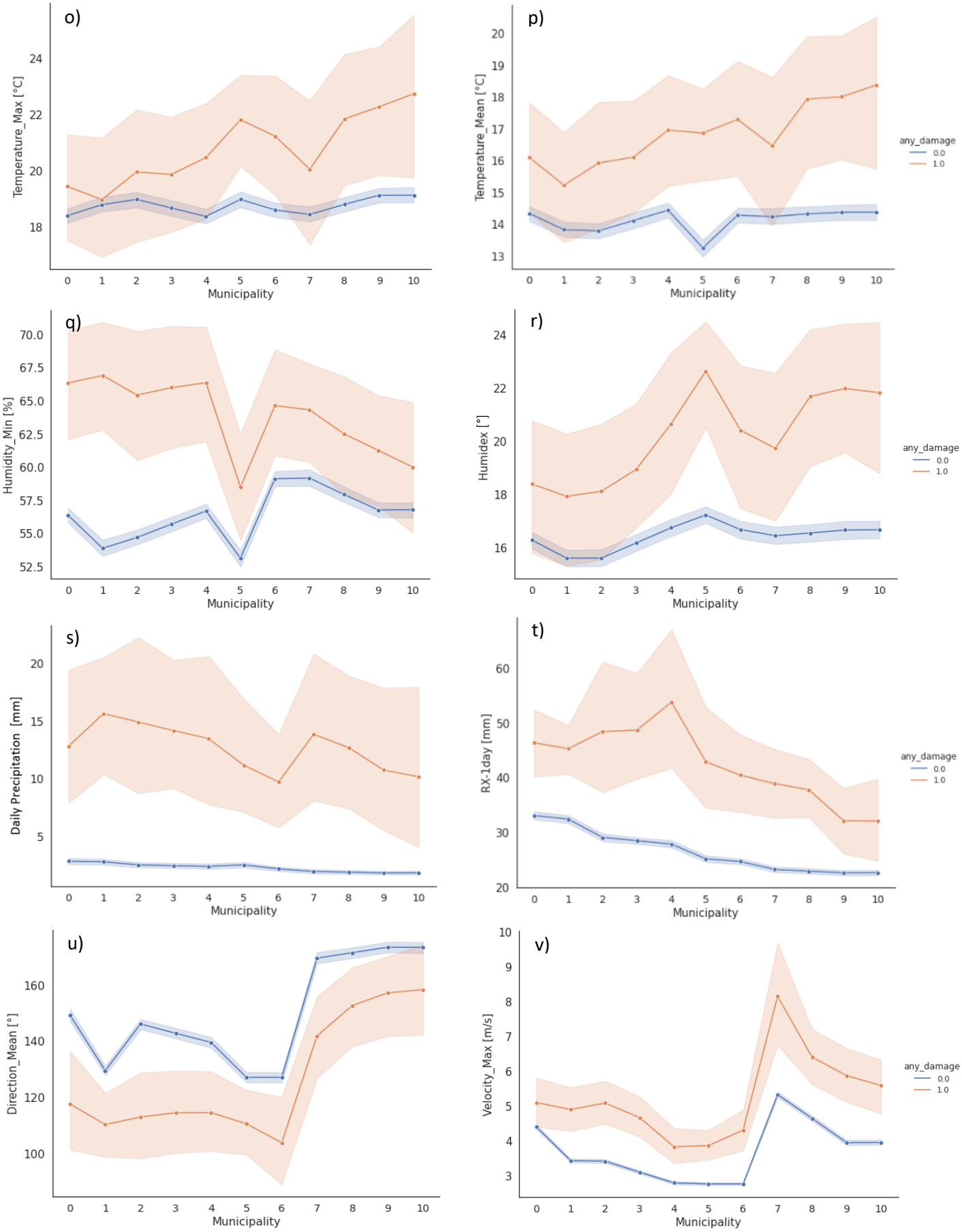


Figure 33: Mean municipal values of observations with and without damages for the variables: a) SSH; b) MSSH; c) WAH; d) WIH; e) WID; f) WAD; g) WIP; h) WAP; i) ESV; l) NSV; m) HWTXdx; n) minimum temperature; o) maximum temperature; p) mean temperature; q) minimum humidity; r) Humidex; s) daily precipitation; t) RX-1day; u) mean wind direction; v) maximum wind velocity

As already reported in *Section 5.2.2*, in the no damage set the sea level height did not show important differences over the municipalities, with values varying respectively between -0.39 and -0.35 m for SSH and between -0.32 and -0.29 m for MSSH, by increasing homogeneously from north to south of the case study area. In the damage dataset, the values changed heterogeneously over the municipalities: SSH between -0.28 and -0.16 m and MSSH between -0.15 and 0.23 m (Figure 33, a-b). The lowest values were reached by municipality 5 and the highest by municipality 7, meaning that, for some municipalities, already when the sea surface parameters recorded a slight deviation from the values associated with no damage, the conditions could lead to damage.

With minimal differences, the same patterns of SSH/MSSH for the two datasets (damage and no damage) were found also for the sea surface wave (WAP) and wind wave mean period (WIP), and for the significant wave and wind wave height (main variables reported in Figure 33, c-d and g-h). Specifically, maximum mean values of significant wave (MWAH) and wind wave height (MWIH), during damage events, were reached by municipality 9 with respective values of 1.31 m and 1.48 m. The indicators regarding the direction of waves (WAD) and wind waves (WID) did not show remarkable differences between the two datasets, although their mean values varied over the municipalities (Figure 33, e-f).

The extreme precipitation indicator RX-1day (Figure 33t) exhibited higher values associated with damage occurrence. Except for municipality 4, the same precipitation pattern decreasing from north to south of the Veneto coastal area was found for both datasets, with values ranging from 33 mm to 22 mm for the no damage dataset and from 46 to 32 mm for the damage dataset. The daily precipitation indicator (Figure 33s) presented a similar behavior for the no damage dataset, while in the damage dataset the mean values decreased up to municipality 6 and raised again at municipality 7. In all cases, the variations of the mean values were minimal for the no damage dataset (ranging from 1.84 mm (municipality 10) to 2.8 mm (municipality 0)) and significant for the damage dataset (ranging from 9.72 mm (municipality 6) to 15.65 mm (municipality 7)).

Temperature indicators were on average pretty constant over the municipalities for the no damage dataset, while more variable for the damage dataset (Figure 33, n-p) with mean values which seemed to increase by going southward (e.g. for mean temperature: in the no damage dataset values varied between 14.43°C and 13.2°C, for the damage dataset between 15.22°C and 18.37°C).

The mean and the maximum wind velocity in the no damage dataset had quite constant values for municipalities 1 to 6, with an increase from municipality 7. The same “municipal” pattern of the wind velocity in the no damage dataset was found also for the damage dataset, although all the values were higher, with municipality 7 reaching the highest mean maximum velocity (e.g., municipality 7 reached a mean value of maximum velocity of 5.32 m/s for the no damage dataset and 8.15 m/s for the damage dataset; Figure 33v). Analogous considerations regarding the similar municipal pattern between the dataset with and without

damages can be drawn for the wind direction, which however showed a prevalently northern-eastern direction for the damage dataset (Figure 33u).

Minimum humidity values increased from north to south of the case study area for the no damage dataset, while in the damage dataset this trend was reversed (Figure 33q). Possibly, humidity behavior in the damage dataset was associated with the decreasing precipitation by going southward of the investigated area.

In general, the characteristics of the indicators found at the regional scale were observed also at the local scale (e.g., precipitation, temperature, and sea surface height indicators had higher values in the dataset with damages). However, the local scale allowed to notice differences in the mean values of the hazard variables over the municipalities, outlining that especially wind velocity and direction varied remarkably. These differences in the hazard indicators at the local scale, amplified during damage conditions, could represent an indication of how the 11 investigated municipalities were differently affected by the same hazard indicator, which could play a different role in the generation of the damage. Therefore, for this kind of study, a local assessment of the factors associated with the damage occurrences should be preferred to assessments with a higher scale of analysis.

Finally, to investigate if the seasonal analysis executed at the regional scale, by confronting the damage and no damage dataset, exhibited the same characteristics at the municipal scale, for the four main indicators associated with extreme weather events causing damages, namely MSSH, RX-1day, mean temperature and maximum wind velocity, this further analysis was performed and respectively represented in Figure 34, 35, 36, 37.

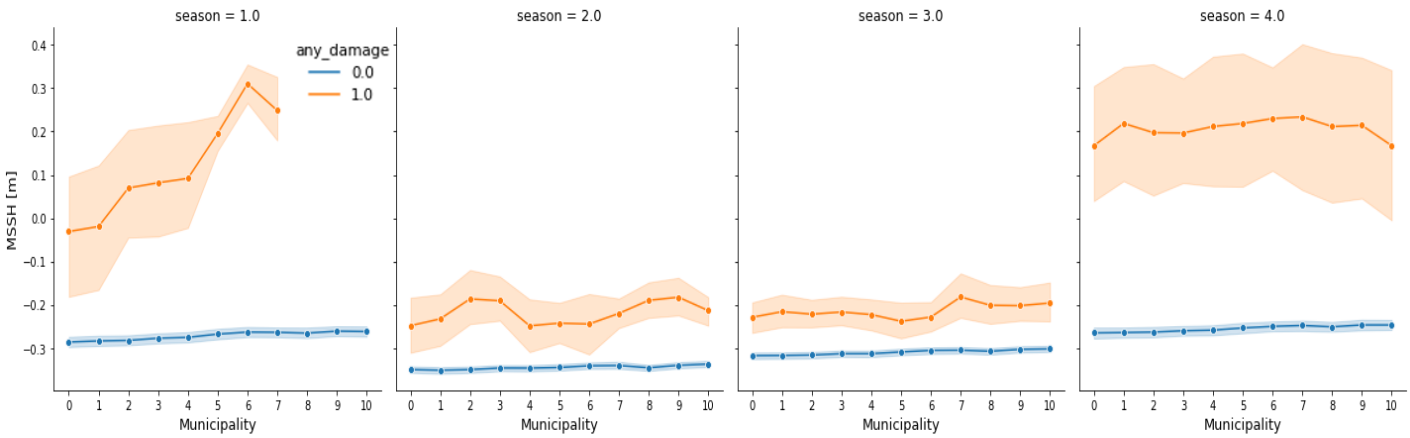


Figure 34: Seasonal analysis of the mean values of MSSH for the 11 investigated municipalities

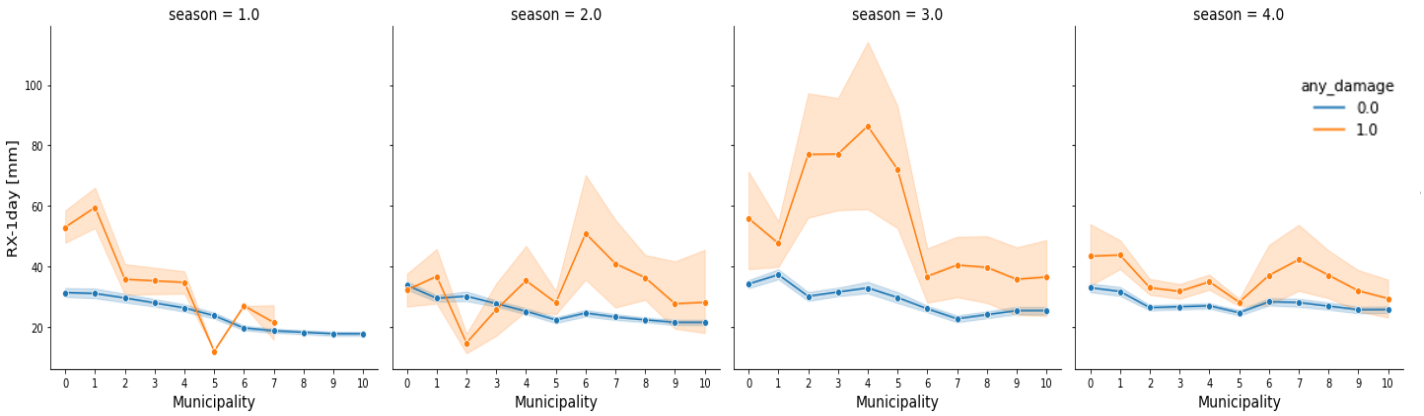


Figure 35: Seasonal analysis of the mean values of RX-1day for the 11 investigated municipalities

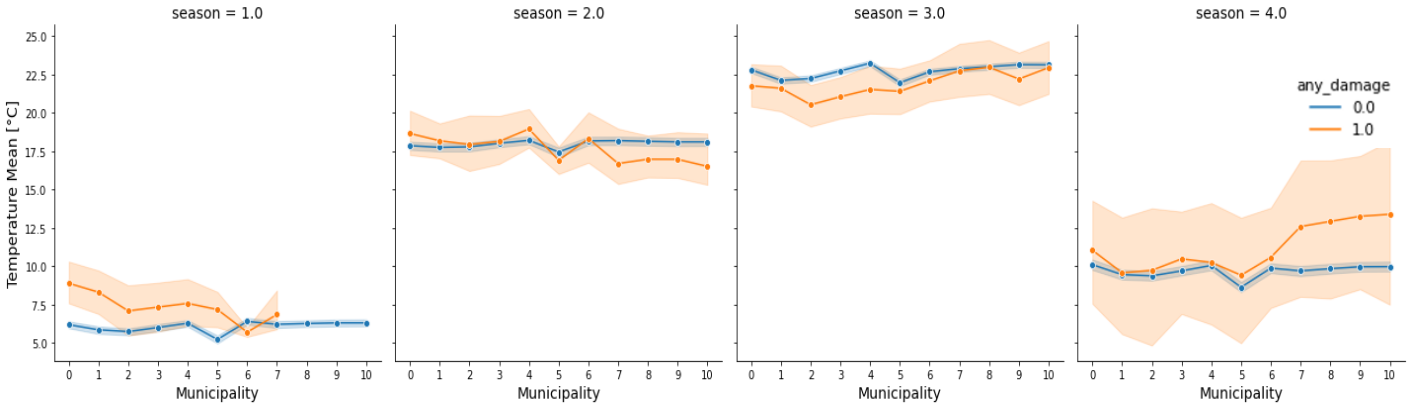


Figure 36: Seasonal analysis of the mean values of mean temperature for the 11 investigated municipalities

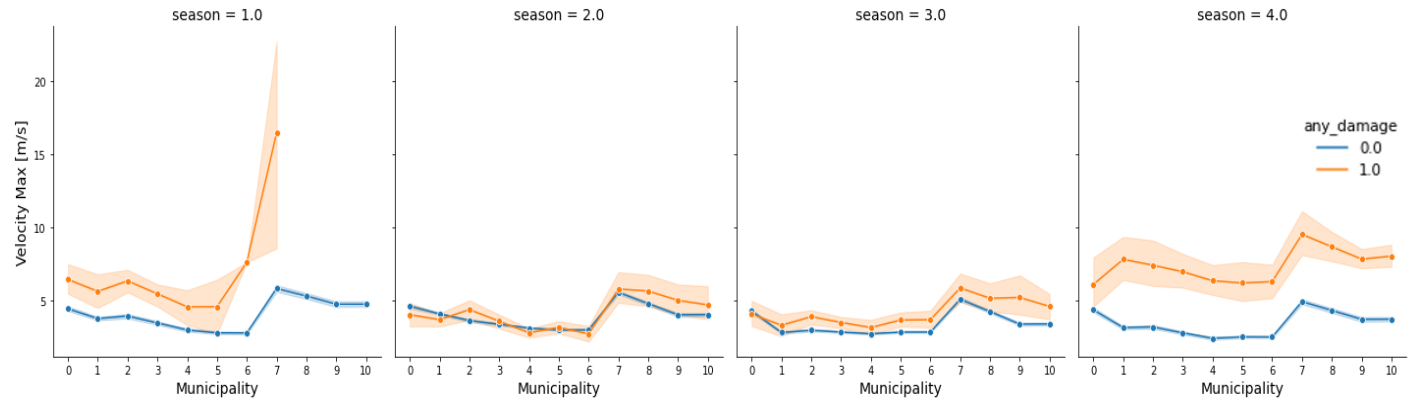


Figure 37: Seasonal analysis of the mean values of maximum wind velocity for the 11 investigated municipalities

What is evident is that, for all the considered indicators, the seasonal analysis executed with mean municipal values had the same characteristics observed at the regional scale. For example, MSSH and maximum wind velocity presented higher differences between the damage and no damage dataset in the winter and autumn seasons, while extreme precipitation (RX-1day) and temperature in the spring and summer seasons.

Even though this seasonal analysis revealed some heterogeneities among the variables at the municipal scale, they were not so relevant. The only slight difference in the seasonal damage dataset, at the local scale, was related to the winter season as clearly visible for MSSH (Figure 34). A detailed investigation of the seasonal distribution of the damages occurred in the 11 municipalities (reported in ANNEX VII), in fact, exhibited that, for all the seasons, the municipalities recorded a similar number of days that presented damage.

In the end, it can be said that, albeit the hazard variables seemed to present not negligible variabilities over the municipalities, these same variabilities were not determined by a different seasonal influence.

CONCLUSIONS

This Thesis was aimed at understanding the triggering factors of extreme weather-driven damages, that occurred in the coastal municipalities of the Veneto region within the 2009-2019 timeframe. Specifically, the study served as a preliminary analysis of the historical dataset for designing Machine Learning (ML) algorithms capable of predicting damages.

The Thesis' objective was achieved by reviewing the scientific literature investigating the state of the art of ML models in the research topic, and by applying multiple data science techniques combining traditional statistical methods (i.e., Exploratory Data Analysis) with ML algorithms (i.e., Random Forest model) in order to find trends and relations between the analyzed variables and the damage occurrence.

In particular, both the scientometric and systematic review helped to examine the improvements of ML algorithms to assess coastal risks due to natural hazards, as well as their limitations. The systematic review led to selecting a quite diversified group of studies, although mainly focused on quantifying the risk of coastal inundation originated by extreme storm surges or sea-level rise. The main applied ML models resulted to be decision trees (i.e., Random Forest and Bayesian Network), which allowed to obtain high predictive accuracy of assessment endpoints and to combine various types of data (i.e., data of different sources and spatio-temporal resolution). Besides, most of these models adopted oceanographic (e.g., sea surface height, wave regime) and atmospheric variables (e.g., precipitation) for assessing and predicting coastal risks.

Based on the knowledge acquired from the literature review, and given the availability of data, a set of atmospheric, oceanographic, and territorial indicators was explored to determine the relations of these indicators with the extreme weather-driven damages occurred in the investigated case study, both at the regional and municipal scale.

The preliminary analysis of the dataset revealed a heterogeneous distribution of the damages, both on an annual and seasonal basis (e.g., damages occurred mainly in spring and summer seasons), following the patterns of the main hazard indicators (i.e., temperature, precipitation, sea surface level, wave regime) as evaluated through a correlation analysis.

The local assessment at the municipal scale outlined a decreased number of damages when going from the northern area of the Veneto region southwards, tracking the trend of atmospheric indicators, such as precipitation.

Building on these results, the designed RF algorithm, aimed at predicting the occurrence of damages in order to unveil the predominant driving features, gained a F1-score of 95% and identified the variables of sea surface height, precipitation, temperature, and significant wave height as the most important factors for the damage manifestation. A further investigation of these features, intended to ascertain their effective relevance when damages occurred, confirmed the reliability of the RF model. Then, for some of the selected features (e.g., mean and maximum sea surface height, significant wave height), it was possible to detect threshold values associated with the damage occurrence, information that could support the coastal

authorities in the activation of early-warning systems.

In addition, the regional-scale analysis of these variables showed how, during damage events, their values changed considerably over the years and the seasons. Anyway, these same analyses performed at the local scale presented different dynamics between damages and hazard variables among the investigated municipalities.

Overall, the study has pointed out some interesting relations between the triggering factors and damage occurrences in the coastal area of the Veneto region. Specifically, by combining traditional statistics and advanced ML algorithms, the proposed data analysis methodology permitted to have a better comprehension of the historical dataset and the associated criticalities.

The followed analysis has produced an effective tool that can help the identification of the most influencing factors in damage occurrence, which, in turn, can serve to guide policy-makers as well as civil protection in adopting suitable management strategies (e.g., Disaster Risk Reduction measures) to cope with extreme weather events.

However, the uncertainty related to the damage data represented a significant limitation, particularly at the municipal scale. Moreover, the scarce information regarding the damages, provided in terms of absence or presence over an entire municipality for a certain date, hindered a thorough understanding of the phenomenon. In fact, having geo-referenced damage data would have allowed a more precise interpretation of the relationships between the damages and the territorial indicators, which for this study had to be averaged over the entire municipal area, reducing the information on exposure and vulnerability. Similarly, having detailed information concerning the type of occurred damage (e.g., damage due to flooding or beach erosion) would have been useful to better evaluate the different influences of hazard indicators on the various types of damage.

Finally, the results and criticalities evidenced by the performed analysis can be used to develop more accurate ML-models such as Artificial Neural Networks and Graph Neural Networks to predict damages in coastal areas. Additionally, these kinds of models have the potential to improve the comprehension of the relations between the triggering factors and damage occurrences, guaranteeing a better estimation of the most important features, which is a pivotal starting point for making reliable predictions under future climate change scenarios.

Bibliography

Aerts, J. C. J. H., Barnard, P. L., Botzen, W., Grifman, P., Hart, J. F., De Moel, H., Mann, A. N., de Ruig, L. T., & Sadrpour, N. (2018). Pathways to resilience: Adapting to sea level rise in Los Angeles. *Annals of the New York Academy of Sciences*, *1427*(1), 1–90. https://doi.org/10.1111/nyas.13917

Aleshina, M. A., Semenov, V. A., & Chernokulsky, A. V. (2021). A link between surface air temperature and extreme precipitation over Russia from station and reanalysis data. *Environmental Research Letters*, *16*(10). https://doi.org/10.1088/1748-9326/ac1cba

Apollonio, C., Balacco, G., Novelli, A., Tarantino, E., & Piccinni, A. F. (2016). Land use change impact on flooding areas: The case study of Cervaro Basin (Italy). *Sustainability (Switzerland)*, *8*(10). https://doi.org/10.3390/su8100996

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Arinta, R. R., & Andi, E. W. R. (2019). Natural disaster application on big data and machine learning: A review. *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019*, *6*, 249–254. https://doi.org/10.1109/ICITISEE48480.2019.9003984

Asariotis, R., Kruckova, L., & Naray, V. M. (2020). *Climate Change Impacts and Adaptation for Coastal Transport Infrastructure: A Compilation of Policies and Practices TRANSPORT AND TRADE FACILITATION Series No 12* (Issue 12).

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Bae, S., & Chang, H. (2019). Urbanization and floods in the Seoul Metropolitan area of South Korea: What old maps tell us. *International Journal of Disaster Risk Reduction*, *37*(March). https://doi.org/10.1016/j.ijdrr.2019.101186

Barandiaran, M., Esquivel, M., Lacambra, S., Suarez, G., & Zuloaga, D. (2018). Executive Summary of the Disaster and Climate Change Risk Assessment Methodology for IDB Projects. A technical reference document for IDB project teams. *Technical Note No. IDB-TN-01583*, *December*. https://publications.iadb.org/publications/english/document/Executive_Summary_of_the_Disaster_and_Climate_Risk_Assessment_Methodology_for_IDB_Projects_A_Technical_Reference_for_IDB_Project_Teams.pdf

Barbi, A., Monai, M., Racca, R., & Rossa, A. M. (2012). Recurring features of extreme autumnall rainfall events on the veneto coastal area. *Natural Hazards and Earth System Science*, *12*(8), 2463–2477. https://doi.org/10.5194/nhess-12-2463-2012

Barbi, A., Mariani, P. L., & Cola, G. (2013). Inquadramento climatico del Veneto. *ARPAV*, *1*.

Battinelli, P. (1997). *The Relation between Extreme Weather Events and the Solar Activity*. Joint European and National Astronomical Meeting.

Bergant, K., Sušnik, M., Strojan, I., & Shaw, A. G. P. (2005). Sea level variability at Adriatic coast and its relationship to atmospheric forcing. *Annales Geophysicae*, *23*(6), 1997–2010. https://doi.org/10.5194/angeo-23-1997-2005

Bezzi, A., Pillon, S., Martinucci, D., & Fontolan, G. (2018). Inventory and conservation assessment for the management of coastal dunes, Veneto coasts, Italy. *Journal of Coastal Conservation*, *22*(3), 503–518. https://doi.org/10.1007/s11852-017-0580-y

Biesbroek, R., Badloe, S., & Athanasiadis, I. N. (2020). Machine learning for research on climate change adaptation policy integration: an exploratory UK case study. *Regional Environmental Change*, *20*(3). https://doi.org/10.1007/s10113-020-01677-8

Bolle, A., das Neves, L., Smets, S., Mollaert, J., & Buitrago, S. (2018). An impact-oriented Early Warning and Bayesian-based Decision Support System for flood risks in Zeebrugge harbour. *Coastal Engineering*, *134*(January 2017), 191–202. https://doi.org/10.1016/j.coastaleng.2017.10.006

Brambati, A., Carbognin, L., Quaia, T., Teatini, P., & Tosi, L. (2003). The Lagoon of Venice: Geological setting, evolution and land subsidence. *Episodes*, *26*(3), 264–265. https://doi.org/10.18814/epiiugs/2003/v26i3/020

Cai, H., Lam, N. S. N., Zou, L., & Qiang, Y. (2018). Modeling the Dynamics of Community Resilience to Coastal Hazards Using a Bayesian Network. *Annals of the American Association of Geographers*, *108*(5), 1260–1279. https://doi.org/10.1080/24694452.2017.1421896

Cajal, B., Jiménez, R., Gervilla, E., & Montaño, J. J. (2020). *Clínica y Salud*. *31*, 77–83.

Calliari, E., Surminski, S., & Mysiak, J. (2019). *The Politics of (and Behind) the UNFCCC's Loss and Damage Mechanism*. https://doi.org/10.1007/978-3-319-72026-5_6

Camuffo, D. (2021). Four centuries of documentary sources concerning the sea level rise in Venice. *Climatic Change*, *167*(3–4), 1–16. https://doi.org/10.1007/s10584-021-03196-9

Cavalieri, C. (2021). Extreme-city-territories. Coastal geographies in the Veneto region. *Journal of Urbanism*, *14*(2), 185–203. https://doi.org/10.1080/17549175.2020.1801490

Chen, C. (2017). Science Mapping: A Systematic Review of the Literature. *Journal of Data and Information Science*, *2*(2), 1–40. https://doi.org/10.1515/jdis-2017-0006

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Collins, E. L., Sanchez, G. M., Terando, A., Stillwell, C. C., Mitasova, H., Sebastian, A., & Meentemeyer, R. K. (2022). Predicting flood damage probability across the conterminous United States. *Environmental Research Letters*, *17*(3). https://doi.org/10.1088/1748-9326/ac4f0f

Coronese, M., Lamperti, F., Keller, K., Chiaromonte, F., & Roventini, A. (2019). Evidence for sharp increase in the economic damages of extreme natural disasters. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21450–21455. https://doi.org/10.1073/pnas.1907826116

Crespi, A., Terzi, S., Cocuccioni, S., Zebisch, M., Julie, B., & Füssel, H.-M. (2020). *Climate-related hazard indices for Europe*. 64. https://doi.org/10.25424/cmcc/climate

Cushman-Rosin, B. (2001). *Physical Oceanography of the Adriatic Sea. Past, Present and Future*. Springer. https://doi.org/10.1007/978-94-015-9819-4 The

Da Lio, C., Carol, E., Kruse, E., Teatini, P., & Tosi, L. (2015). Saltwater contamination in the managed low-lying farmland of the Venice coast, Italy: An assessment of vulnerability. *Science of the Total Environment*, *533*, 356–369. https://doi.org/10.1016/j.scitotenv.2015.07.013

Darko, A., Chan, A. P. C., Huo, X., & Owusu-Manu, D. G. (2019). A scientometric analysis and visualization of global green building research. *Building and Environment*, *149*(December 2018), 501–511. https://doi.org/10.1016/j.buildenv.2018.12.059

Denamiel, C., Pranić, P., Quentin, F., Mihanović, H., & Vilibić, I. (2020). Pseudo-global warming projections of extreme wave storms in complex coastal regions: the case of the Adriatic Sea. *Climate Dynamics*, *55*(9–10), 2483–2509. https://doi.org/10.1007/s00382-020-05397-x

Di Nunno, F., Granata, F., Gargano, R., & de Marinis, G. (2021). Forecasting of extreme storm tide events using narx neural network-based models. *Atmosphere*, *12*(4). https://doi.org/10.3390/atmos12040512

Dorman, C. E., Carniel, S., Cavaleri, L., Sclavo, M., Chiggiato, J., Doyle, J., Haack, T., Pullen, J., Grbec, B., Vilibić, I., Janeković, I., Lee, C., Malačič, V., Orlić, M., Paschini, E., Russo, A., & Signell, R. P. (2007). *February 2003 marine atmospheric conditions and the bora over the northern Adriatic*. Journal of Geophysical Research: Oceans. https://doi.org/10.1029/2005JC003134

Dubey, A. (2018). *Feature Selection Using Random forest*. https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

EEA. (2021). *Heavy precipitation in Europe*. https://www.eea.europa.eu/data-and-maps/indicators/precipitation-extremes-in-europe-3

EEA. (2022). *Economic losses from climate-related extremes in Europe*. https://www.eea.europa.eu/ims/economic-losses-from-climate-related

ENEA. (2018). *Innalzamento del Mar Mediterraneo in Italia. Aree costiere e porti a rischio inondazione al 2100 Il*.

Esfahani, H. J., Tavasoli, K., & Jabbarzadeh, A. (2019). Big data and social media: A scientometrics analysis. *International Journal of Data and Network Science*, *3*(3), 145–164. https://doi.org/10.5267/j.ijdns.2019.2.007

Fabris, M. (2019). Coastline evolution of the Po River Delta (Italy) by archival multi-temporal digital photogrammetry. *Geomatics, Natural Hazards and Risk*, *10*(1), 1007–1027. https://doi.org/10.1080/19475705.2018.1561528

Ferreira, Ó., Plomaritis, T. A., & Costas, S. (2019). Effectiveness assessment of risk reduction measures at coastal areas using a decision support system: Findings from Emma storm. *Science of the Total Environment*, *657*, 124–135. https://doi.org/10.1016/j.scitotenv.2018.11.478

Ferretti, O., Delbono, I., Furia, S., & Barsanti, M. (2003). *Elementi di Gestione Costiera Parte Seconda: Erosione Costiera Lo stato dei litorali italiani*. *October 2014*.

Findell, K. L., Berg, A., Gentine, P., Krasting, J. P., Lintner, B. R., Malyshev, S., Santanello, J. A., & Shevliakova, E. (2017). The impact of anthropogenic land use and land cover change on regional climate extremes. *Nature Communications*, *8*(1), 1–9. https://doi.org/10.1038/s41467-017-01038-w

Forliano, C., De Bernardi, P., & Yahiaoui, D. (2021). Entrepreneurial universities: A bibliometric analysis within the business and management domains. *Technological Forecasting and Social Change*, *165*(January), 120522. https://doi.org/10.1016/j.techfore.2020.120522

Frasca, M. (2018). *Data Mining and Machine Learning Lab . Lezione 5 Master in Data Science for Economics , Business and Finance 2018*.

Frazier, A. E., Hemingway, B. L., & Brasher, J. P. (2019). Land surface heterogeneity and tornado occurrence: an analysis of Tornado Alley and Dixie Alley. *Geomatics, Natural Hazards and Risk*, *10*(1), 1475–1492. https://doi.org/10.1080/19475705.2019.1583292

Furlan, E., Pozza, P. D., Michetti, M., Torresan, S., Critto, A., & Marcomini, A. (2021). Development of a Multi-Dimensional Coastal Vulnerability Index: Assessing vulnerability to inundation scenarios in the Italian coast. *Science of the Total Environment*, *772*, 144650. https://doi.org/10.1016/j.scitotenv.2020.144650

Gallina, V., Torresan, S., Zabeo, A., Rizzi, J., Carniel, S., Sclavo, M., Pizzol, L., Marcomini, A., & Critto, A. (2019). Assessment of climate change impacts in the North Adriatic coastal area. Part II: Consequences for coastal erosion impacts at the regional scale. *Water (Switzerland)*, *11*(6), 1–20.

https://doi.org/10.3390/w11061300

Gatto, P., & Carbognin, L. (1981). The lagoon of venice: Natural environmental trend and man-induced modification. *Hydrological Sciences Bulletin*, *26*(4), 379–391. https://doi.org/10.1080/02626668109490902

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Glavovic et al. (2022). Cross-Chapter Paper 2: Cities and Settlements by the Sea. In B. R. (eds.). [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem (Ed.), *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Gutta, S. (2020). *Bias and Variance in simple terms!* https://medium.com/analytics-vidhya/difference-between-bias-and-variance-in-machine-learning-fec71880c757

Ha, J., & Kang, J. E. (2022). Assessment of flood-risk areas using random forest techniques: Busan Metropolitan City. *Natural Hazards*, *111*(3), 2407–2429. https://doi.org/10.1007/s11069-021-05142-5

Hafen, R., & Critchlow, T. (2013). EDA and ML-A perfect pair for large-scale data analysis. *Proceedings - IEEE 27th International Parallel and Distributed Processing Symposium Workshops and PhD Forum, IPDPSW 2013*, 1894–1898. https://doi.org/10.1109/IPDPSW.2013.118

Hastie, T. (2009). Random Forest. In *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7_15

Heidenreich, H. (2018). *What are the types of machine learning?* https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

IPCC. (2022). *Climate Change 2022 - Impacts, Adaptation and Vulnerability - Summary for Policymakers.* Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. In Press.

ISTAT. (2021). *Popolazione residente al 1° Gennaio 2021*.

Jackson, A. H. (1988). Machine learning. In *Expert Systems* (Vol. 5, Issue 2). https://doi.org/10.1111/j.1468-0394.1988.tb00341.x

Jäger, W. S., Christie, E. K., Hanea, A. M., den Heijer, C., & Spencer, T. (2018). A Bayesian network approach for coastal risk analysis and decision making. *Coastal Engineering*, *134*(January), 48–61. https://doi.org/10.1016/j.coastaleng.2017.05.004

Jakariya, M., Alam, M. S., Rahman, M. A., Ahmed, S., Elahi, M. M. L., Khan, A. M. S., Saad, S., Tamim, H. M., Ishtiak, T., Sayem, S. M., Ali, M. S., & Akter, D. (2020). Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. *Science of the Total Environment*, *742*, 140255. https://doi.org/10.1016/j.scitotenv.2020.140255

Javaheri, S. H., Sepehri, M. M., & Teimourpour, B. (2013). Response Modeling in Direct Marketing. A Data Mining-Based Approach for Target Selection. *Data Mining Applications with R*, 153–180. https://doi.org/10.1016/B978-0-12-411511-8.00006-2

Jones, Z., & Linder, F. (2015). Exploratory Data Analysis using Random Forests. *73rd Annual MPSA Conference*, 1–31.

Kanstrén, T. (2020). *A Look at Precision, Recall, and F1-Score*. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, A. (2022). *Machine Learning Use Cases for Climate Change*. https://vitalflux.com/machine-learning-use-cases-climate-change/

Lee, S. hyeok, Kang, J. E., Park, C. S., Yoon, D. K., & Yoon, S. (2020). Multi-risk assessment of heat waves under intensifying climate change using Bayesian Networks. *International Journal of Disaster Risk Reduction*, *50*(June), 101704. https://doi.org/10.1016/j.ijdrr.2020.101704

Legambiente. (2012). *Il consumo delle aree costiere italiane. La COSTA VENETA, da Bibione a Porto Tolle: l'aggressione del cemento e i cambiamenti del paesaggio*.

Li, C., Huang, M., Tian, J., & Isabella Anak Gisen, J. (2022). Research and application of the mathematical model for extreme weather event in coastal urban areas. *Physics and Chemistry of the Earth*, *125*(September 2021), 103075. https://doi.org/10.1016/j.pce.2021.103075

Lionello, P. (2012). The climate of the Venetian and North Adriatic region: Variability, trends and future change. *Physics and Chemistry of the Earth*, *40–41*(October 2008), 1–8. https://doi.org/10.1016/j.pce.2012.02.002

Lionello, P., Cavaleri, L., Nissen, K. M., Pino, C., Raicich, F., & Ulbrich, U. (2012). Severe marine storms in the Northern Adriatic: Characteristics and trends. *Physics and Chemistry of the Earth*, *40–41*, 93–105. https://doi.org/10.1016/j.pce.2010.10.002

Lionello, P., Galati, M. B., & Elvini, E. (2012). Extreme storm surge and wind wave climate scenario simulations at the Venetian littoral. *Physics and Chemistry of the Earth*, *40–41*, 86–92. https://doi.org/10.1016/j.pce.2010.04.001

Liu, M., Vecchi, G. A., Smith, J. A., & Knutson, T. R. (2019). Causes of large projected increases in hurricane precipitation rates with global warming. *Npj Climate and Atmospheric Science*, *2*(1), 1–5. https://doi.org/10.1038/s41612-019-0095-3

López, R. E., Thomas, V., & Troncoso, P. (2015). *Climate Change and Natural Disasters*.

Madricardo, F., Foglini, F., Campiani, E., Grande, V., Catenacci, E., Petrizzo, A., Kruss, A., Toso, C., & Trincardi, F. (2019). Assessing the human footprint on the sea-floor of coastal systems: the case of the Venice Lagoon, Italy. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-43027-7

Maina, J. M., Bosire, J. O., Kairo, J. G., Bandeira, S. O., Mangora, M. M., Macamo, C., Ralison, H., & Majambo, G. (2021). Identifying global and local drivers of change in mangrove cover and the implications for management. *Global Ecology and Biogeography*, *30*(10), 2057–2069. https://doi.org/10.1111/geb.13368

Awad, M. (2015). *Efficient Learning Mashines. Theories, Concepts, and Applications for Engineers and System Designers* (Vol. 59). Apress Berkeley, CA. https://doi.org/https://doi.org/10.1007/978-1-4302-5990-9

Marone, E., Camargo, R. De, & Castro, J. S. (2017). Coastal Hazards , Risks , and Marine Extreme Events : What Are We Doing About It ? *Oxford Handbooks Online*, *October*, 1–19. https://doi.org/10.1093/oxfordhb/9780190699420.013.34

Matsueda, M., & Nakazawa, T. (2015). Early warning products for severe weather events derived from

operational medium-range ensemble forecasts. *Meteorological Applications*, *22*(2), 213–222. https://doi.org/10.1002/met.1444

MATTM-Regioni. (2018). *Linee Guida per la Difesa della Costa dai fenomeni di Erosione e dagli effetti dei Cambiamenti climatici. Versione 2018*. 305.

MATTM. (2017). *L'Erosione Costiera in Italia - Le Variazioni Della Linea Di Costa Dal 1960 Al 2012*. https://www.minambiente.it/sites/default/files/archivio/biblioteca/monografia_variazioni_linea_costa_mar17.pdf

Međugorac, I., Orlić, M., Janeković, I., Pasarić, Z., & Pasarić, M. (2018). Adriatic storm surges and related cross-basin sea-level slope. *Journal of Marine Systems*, *181*(February), 79–90. https://doi.org/10.1016/j.jmarsys.2018.02.005

Mel, R., Viero, D. Pietro, Carniello, L., Defina, A., & D'Alpaos, L. (2014). Simplified methods for real-time prediction of storm surge uncertainty: The city of Venice case study. *Advances in Water Resources*, *71*, 177–185. https://doi.org/10.1016/j.advwatres.2014.06.014

Meli, M., Olivieri, M., & Romagnoli, C. (2021). Sea-level change along the emilia-romagna coast from tide gauge and satellite altimetry. *Remote Sensing*, *13*(1), 1–26. https://doi.org/10.3390/rs13010097

Milojevic-Dupont, N., & Creutzig, F. (2021). Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society*, *64*(June 2020), 102526. https://doi.org/10.1016/j.scs.2020.102526

Modica, M., Zoboli, R., & Pagliarino, E. (2017). Mapping the environmental pressure due to economic factors. The case of Italian coastal municipalities. *Argomenti*, *0*(8), 73–105. http://ojs.uniurb.it/index.php/argomenti/article/view/999

Moftakhari, H., AghaKouchak, A., Sanders, B. F., Matthew, R. A., & Mazdiyasni, O. (2017). Translating Uncertain Sea Level Projections Into Infrastructure Impacts Using a Bayesian Framework. *Geophysical Research Letters*, *44*(23), 11,914-11,921. https://doi.org/10.1002/2017GL076116

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 33–44. https://doi.org/10.1002/wics.2

Mushtaq, S. (2019). *Data preprocessing in detail*.

Nicholls, R.J., P.P. Wong, V.R. Burkett, J.O. Codignotto, J.E. Hay, R.F. McLean, S. R. and C. D. W. (2007). *Coastal systems and low-lying areas. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution ofWorking Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*.

NOAA. (2022a). *Coastal Hazard:Preparing for the Threats that Face our Coastal Communities*. https://oceanservice.noaa.gov/hazards/natural-hazards/

NOAA. (2022b). *What is a strom surge?*

Ibe, O. (2014). Fundamentals of Applied Probability and Random Processes (Second Edition). In *Fundamentals of Applied Probability and Random Processes (Second Edition)* (pp. 57–79). Academic Press. https://doi.org/10.1016/B978-0-12-800852-2.00002-X

Osservatorio Nazionale clima e città. (2021). *Il clima è già cambiato, Le città e le reti di fronte alla sfida dell'adattamento climatico*. Legambiente.

Papagiannaki, K., Kotroni, V., Lagouvardos, K., Bezes, A., Vafeiadis, V., Messini, I., Kroustallis, E., & 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), 994. https://doi.org/10.3390/w14060994

Park, S. J., & Lee, D. K. (2020). Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. *Environmental Research Letters*, *15*(9), 94052. https://doi.org/10.1088/1748-9326/aba5b3

Pawluk De-Toledo, K., O'Hern, S., & Koppel, S. (2022). Travel behaviour change research: A scientometric review and content analysis. *Travel Behaviour and Society*, *28*(March), 141–154. https://doi.org/10.1016/j.tbs.2022.03.004

Pereira, S. C., Carvalho, D., & Rocha, A. (2021). Temperature and precipitation extremes over the iberian peninsula under climate change scenarios: A review. *Climate*, *9*(9). https://doi.org/10.3390/cli9090139

Pesta, B., Fuerst, J., & Kirkegaard, E. O. W. (2018). Bibliometric keyword analysis across seventeen years (2000–2016) of intelligence articles. *Journal of Intelligence*, *6*(4), 1–12. https://doi.org/10.3390/jintelligence6040046

Plomaritis, T. A., Costas, S., & Ferreira, Ó. (2018). Use of a Bayesian Network for coastal hazards, impact and disaster risk reduction assessment at a coastal barrier (Ria Formosa, Portugal). *Coastal Engineering*, *134*(September), 134–147. https://doi.org/10.1016/j.coastaleng.2017.07.003

Praharaj, S., Chen, T. D., Zahura, F. T., Behl, M., & Goodall, J. L. (2021). Estimating impacts of recurring flooding on roadway networks: a Norfolk, Virginia case study. *Natural Hazards*, *107*(3), 2363–2387. https://doi.org/10.1007/s11069-020-04427-5

Pranavam, U., Pillai, A., Pinardi, N., Federico, I., Causio, S., Trotta, F., Unguendoli, S., & Valentini, A. (2022). *Wind-Wave Characteristics and extremes along the Emilia-Romagna coast*. *April*, 1–26.

Qi, D., & Majda, A. J. (2020). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(1), 52–59. https://doi.org/10.1073/pnas.1917285117

Radhakrishnan, S., Erbis, S., Isaacs, J. A., & Kamarthi, S. (2017). Correction: Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature (PLoS ONE (2017) 12:3 (e0172778) DOI: 10.1371/journal.pone.0172778). *PLoS ONE*, *12*(9), 1–16. https://doi.org/10.1371/journal.pone.0185771

Rahman, A. (2019). *What is Data Cleaning? How to Process Data for Analytics and Machine Learning Modeling?* https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45

Raj, S. (2019). *Effects of Multi-collinearity in Logistic Regression, SVM, Random Forest(RF)*. https://medium.com/@raj5287/effects-of-multi-collinearity-in-logistic-regression-svm-rf-af6766d91f1b

Regione del Veneto. (2011). 31 ottobre - 2 novembre 2010: l'alluvione dei Santi. In *Rapporto Statistico 2011* (pp. 410–425).

Regione del Veneto. (2012). *Analysis of ICZM practice in Italy: Veneto*.

Regione del Veneto. (2021). *Statistiche: numeri e grafici per capire il Veneto*. https://statistica.regione.veneto.it/Pubblicazioni/StatisticheFlash/statistiche_flash_ottobre_2021.pdf

Rizzi, J., Torresan, S., Critto, A., Zabeo, A., Brigolin, D., Carniel, S., Pastres, R., & Marcomini, A. (2016). Climate change impacts on marine water quality: The case study of the Northern Adriatic sea. *Marine Pollution Bulletin*, *102*(2), 271–282. https://doi.org/10.1016/j.marpolbul.2015.06.037

Rizzi, J., Torresan, S., Zabeo, A., Critto, A., Tosoni, A., Tomasin, A., & Marcomini, A. (2017). Assessing storm surge risk under future sea-level rise scenarios: a case study in the North Adriatic coast. *Journal of Coastal Conservation*, *21*(4), 453–471. https://doi.org/10.1007/s11852-017-0517-5

Rohmer, J., Lincke, D., Hinkel, J., Le Cozannet, G., Lambert, E., & Vafeidis, A. T. (2021). Unravelling the importance of uncertainties in global-scale coastal flood risk assessments under sea level rise. *Water (Switzerland)*, *13*(6), 1–18. https://doi.org/10.3390/w13060774

Roudier, P., Andersson, J. C. M., Donnelly, C., Feyen, L., & Greuell, W. (2016). *Projections of future floods and hydrological droughts in Europe under a + 2 ° C global warming*. 341–355. https://doi.org/10.1007/s10584-015-1570-4

Ruol, P., Martinelli, L., Favaretto, C., Pinato, T., Galiazzo, F., Patti, S., Anti, U., Piazza, R., Simonin, P., & Selvi, G. (2016). *Gestione integrata della Zona costiera studio e monitoraggio per la definizione degli interventi di difesa dei litorali dall'erosione nella Regione Veneto-linee guida*.

Ruol, P., Martinelli, L., & Favaretto, C. (2018). Vulnerability analysis of the venetian littoral and adopted mitigation strategy. *Water (Switzerland)*, *10*(8). https://doi.org/10.3390/w10080984

Rutgersson, A., Kjellström, E., Haapala, J., Stendel, M., Danilovich, I., Drews, M., Jylhä, K., Kujala, P., Larsén, X. G., Halsnæs, K., Lehtonen, I., Luomaranta, A., Nilsson, E., Olsson, T., Särkkä, J., Tuomi, L., & Wasmund, N. (2022). Natural hazards and extreme events in the Baltic Sea region. *Earth System Dynamics*, *13*(1), 251–301. https://doi.org/10.5194/esd-13-251-2022

Sanuy, M., & Jiménez, J. A. (2021). Probabilistic characterisation of coastal storm-induced risks using Bayesian networks. *Natural Hazards and Earth System Sciences*, *21*(1), 219–238. https://doi.org/10.5194/nhess-21-219-2021

Seneviratne, S. I. (2012). Changes in Climate Extremes and their Impacts on the Natural Physical Environment. In S. K. A. [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, and P. M. M. (eds.)]. A. S. R. of W. G. I. and I. of the I. P. on C. M. Tignor, & C. (IPCC). (Eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change.* (pp. 109-230.). Cambridge University Press, Cambridge, UK, and New York, NY, USA.

Senthilnathan, S. (2019). Usefulness of Correlation Analysis. *SSRN Electronic Journal*, *July*. https://doi.org/10.2139/ssrn.3416918

Simpson, N. P., Mach, K. J., Constable, A., Hess, J., Hogarth, R., Howden, M., Lawrence, J., Lempert, R. J., Muccione, V., Mackey, B., New, M. G., O'Neill, B., Otto, F., Pörtner, H. O., Reisinger, A., Roberts, D., Schmidt, D. N., Seneviratne, S., Strongin, S., … Trisos, C. H. (2021). A framework for complex climate change risk assessment. *One Earth*, *4*(4), 489–501. https://doi.org/10.1016/j.oneear.2021.03.005

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*. https://doi.org/10.1186/1471-2105-8-25

Taramelli, A., Valentini, E., Righini, M., Filipponi, F., Geraldini, S., & Xuan, A. N. (2020). Assessing po river deltaic vulnerability using earth observation and a bayesian belief network model. *Water (Switzerland)*, *12*(10). https://doi.org/10.3390/w12102830

Tolo, S., Patelli, E., & Beer, M. (2015). Enhanced Bayesian Network approach to sea wave overtopping hazard quantification. *Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*, *September*, 1983–1990. https://doi.org/10.1201/b19094-258

Tolo, S., Patelli, E., & Beer, M. (2017). Risk Assessment of Spent Nuclear Fuel Facilities Considering Climate Change. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *3*(2). https://doi.org/10.1061/ajrua6.0000874

Torresan, S., Critto, A., Rizzi, J., & Marcomini, A. (2012). Assessment of coastal vulnerability to climate change hazards at the regional scale: The case study of the North Adriatic Sea. *Natural Hazards and Earth System Science*, *12*(7), 2347–2368. https://doi.org/10.5194/nhess-12-2347-2012

Torresan, S., Critto, A., Dalla Valle, M., Harvey, N., & Marcomini, A. (2008). Assessing coastal vulnerability to climate change: Comparing segmentation at global and regional scales. *Sustainability Science*, *3*(1), 45–65. https://doi.org/10.1007/s11625-008-0045-1

Torresan, S., Gallina, V., Gualdi, S., Bellafiore, D., Umgiesser, G., Carniel, S., Sclavo, M., Benetazzo, A., Giubilato, E., & Critto, A. (2019). Assessment of Climate Change Impacts in the North Adriatic Coastal Area. Part I: A Multi-Model Chain for the Definition of Climate Change Hazard Scenarios. *Water*, *11*(1157). https://doi.org/doi:10.3390/w11061157

UNCTAD. (2017). *Climate change impacts on coastal transport infrastructure in the Caribbean: enhancing the adaptive capacity of Small Island Developing States (SIDS), Climate Risk and Vulnerability Assessment Framework for Caribbean Coastal Transport Infrastructure*. *December*, 119.

UNDRR. (2022). *Disaster risk: terminology.* https://www.undrr.org/terminology/disaster-risk#:~:text=The potential loss of life,%2C exposure%2C vulnerability and capacity.

UNEP. (2022). *How climate change is making record-breaking floods the new normal*. https://www.unep.org/news-and-stories/story/how-climate-change-making-record-breaking-floods-new-normal

UNISDR. (2015). *Sendai Framework for Disaster Risk Reduction 2015-2030.*

Volke, M., & Abarca-Del-Rio, R. (2020). Comparison of machine learning classification algorithms for land cover change in a coastal area affected by the 2010 Earthquake and Tsunami in Chile. *Natural Hazards and Earth System Sciences Discussions*, *2010*(March), 1–14.

Wazneh, H., Arain, M. A., Coulibaly, P., & Gachon, P. (2020). Evaluating the Dependence between Temperature and Precipitation to Better Estimate the Risks of Concurrent Extreme Weather Events. *Advances in Meteorology*, *2020*. https://doi.org/10.1155/2020/8763631

Wendler-Bosco, V., & Nicholson, C. (2022). Modeling the economic impact of incoming tropical cyclones using machine learning. In *Natural Hazards* (Vol. 110, Issue 1). Springer Netherlands. https://doi.org/10.1007/s11069-021-04955-8

Xie, K., Ozbay, K., Zhu, Y., & Yang, H. (2017). Evacuation Zone Modeling under Climate Change: A Data-Driven Method. *Journal of Infrastructure Systems*, *23*(4), 04017013. https://doi.org/10.1061/(asce)is.1943-555x.0000369

Yang, X., Yao, C., Chen, Q., Ye, T., & Jin, C. (2019). Improved estimates of population exposure in low-elevation coastal zones of China. *International Journal of Environmental Research and Public Health*, *16*(20). https://doi.org/10.3390/ijerph16204012

Young, A., Bhattacharya, B., & Zevenbergen, C. (2021). A rainfall threshold-based approach to early warnings in urban data-scarce regions: A case study of pluvial flooding in Alexandria, Egypt. *Journal of Flood Risk Management*, *14*(2). https://doi.org/10.1111/jfr3.12702

Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., & Behl, M. (2020). Training Machine Learning Surrogate Models From a High-Fidelity Physics-Based Model: Application for Real-Time Street-Scale Flood Prediction in an Urban Coastal Community. *Water Resources Research*, *56*(10). https://doi.org/10.1029/2019WR027038

Zanchini, E. (2021). *Rapporto spiagge 2021: La situazione ed i cambiamentiin corso nelle aree costiere italiane*. https://www.legambiente.it/wp-content/uploads/2021/07/Rapporto-Spiagge-2021.pdf

Zennaro, F., Furlan, E., Simeoni, C., Torresan, S., Aslan, S., Critto, A., & Marcomini, A. (2021). Exploring machine learning potential for climate change risk assessment. *Earth-Science Reviews*, *220*(June), 103752. https://doi.org/10.1016/j.earscirev.2021.103752

Zhou, H., Ren, H., Royer, P., Hou, H., & Yu, X. Y. (2022). Big Data Analytics for Long-Term Meteorological Observations at Hanford Site. *Atmosphere*, *13*(1). https://doi.org/10.3390/atmos13010136

Zittis, G., Bruggeman, A., & Lelieveld, J. (2021). Revisiting future extreme precipitation trends in the Mediterranean. *Weather and Climate Extremes*, *34*(July), 100380. https://doi.org/10.1016/j.wace.2021.100380

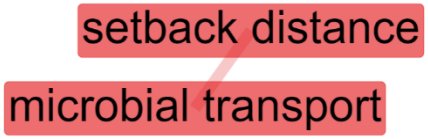
ANNEX I: Formulated query for selecting the publications related to the performed literature review

The query implemented in the Scopus database, for selecting the publications dealing with the application of ML methods to assess risks due to natural hazards in coastal environments, was the following:

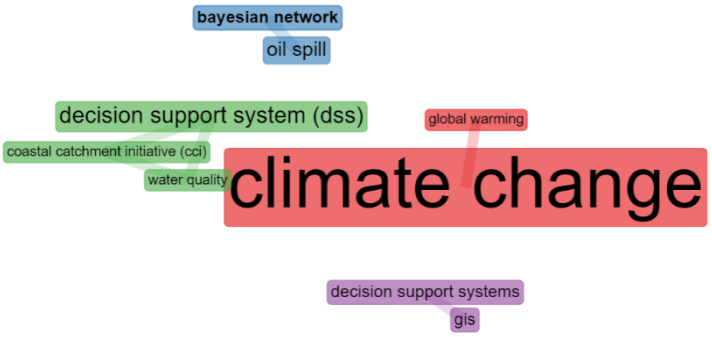
(((("ml" OR "machine learning") OR ("deep learning") OR ("ai" OR "artificial intelligence") OR ("decision tree" OR "DT") OR ("random forest" OR "RF") OR ("Bayesian network" OR "BN")) AND ("coast" OR "marine*" OR "sea") AND ("climate change" OR "scenario*") AND (("erosion") OR ("water quality" OR "turbidity" OR "eutrophication") OR ("storm surge") OR ("slr" OR "sea level*") OR ("extreme event*") OR ("pluvial flood") OR ("flood*") OR ("inundation") OR ("drought") OR ("heat wave*") OR ("risk*" OR "vulnerability" OR "exposure")))*

ANNEX II: Keywords Co-occurrence network graphs under four time slices A) 2001-2006, B) 2006-2011, C) 2011-2016, D) 2016-2021

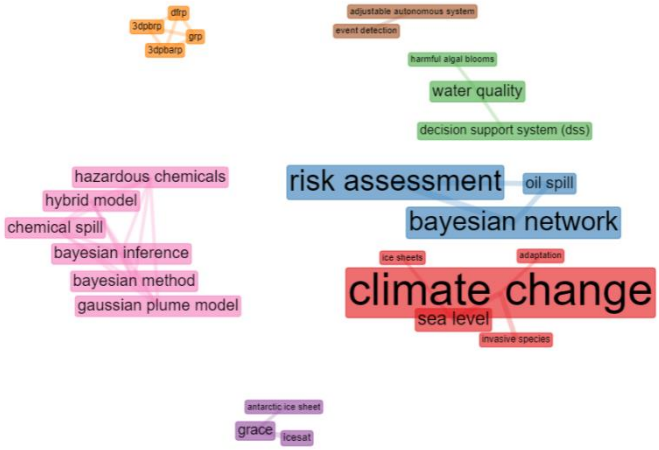
A) 2001-2006:



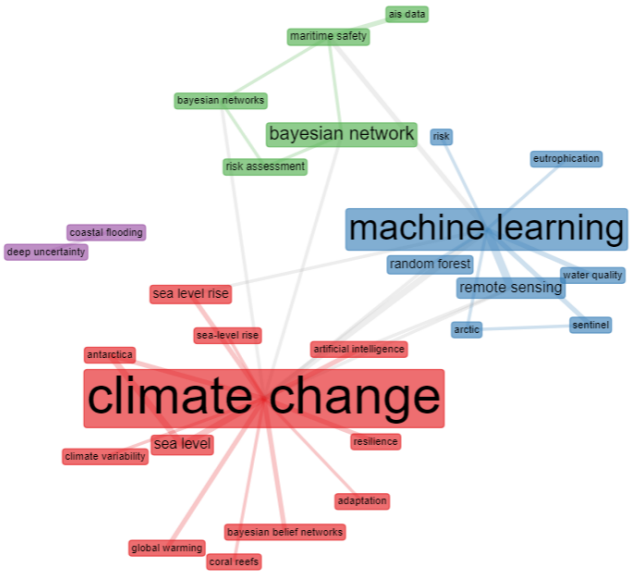
B) 2006-2011:



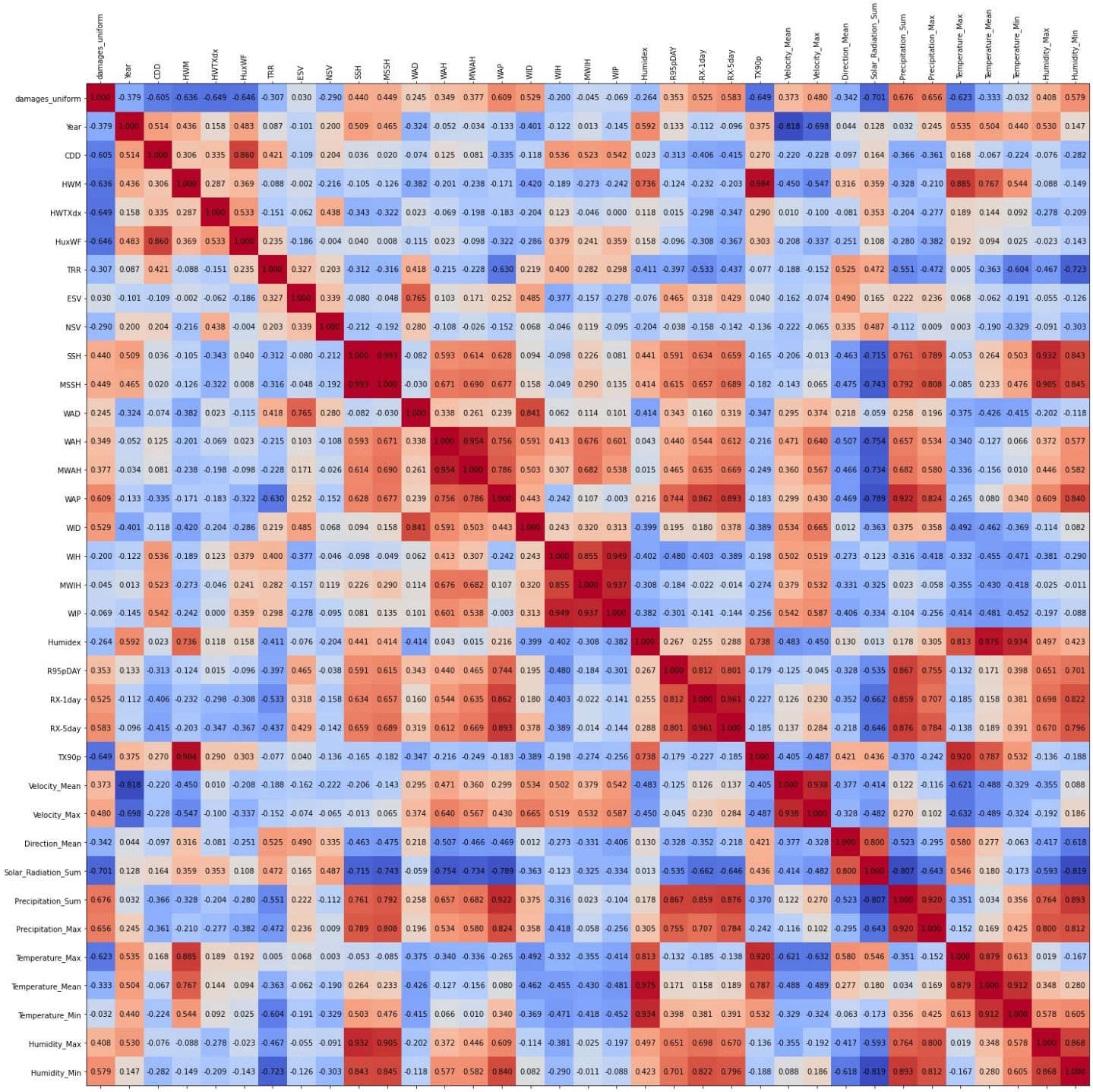
C) 2011-2016:



D) 2016-2021:



ANNEX III: Correlation matrix between the yearly number of damages and the yearly mean values of the main hazard variables



ANNEX IV: Seasonal and monthly trends of the variables (mean values) showing similar patterns to the seasonal and monthly trends of the damage occurrences

Seasonal trend

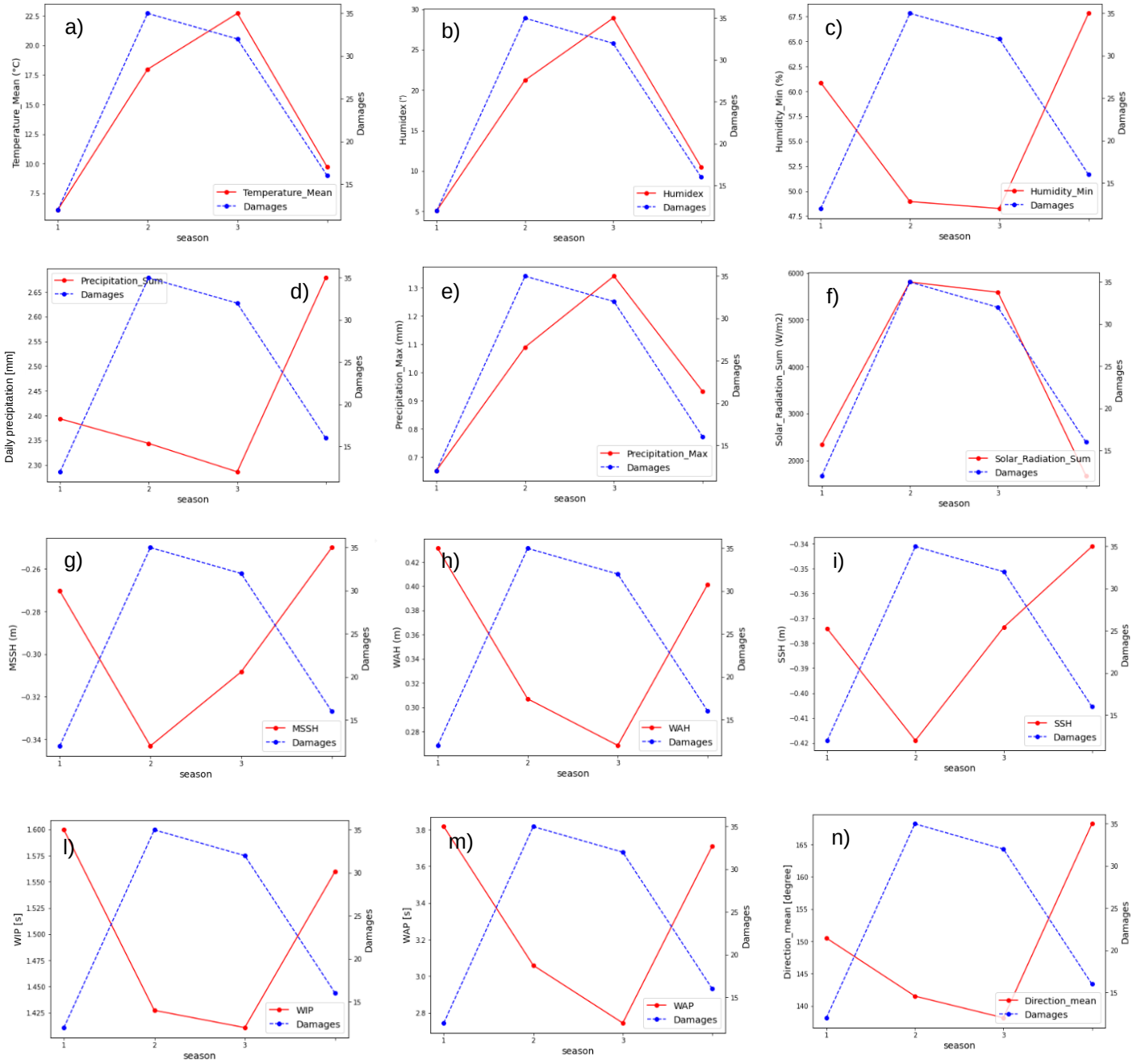


Figure III.1: Seasonal number of damages confronted with the mean seasonal values of the variables: a) mean temperature; b) Humidex; c) minimum humidity; d) daily precipitation; e) maximum precipitation; f) solar radiation; g) MSSH; h) WAH; i) SSH; l) WIP m) WAP; n) wind mean direction

Monthly trend

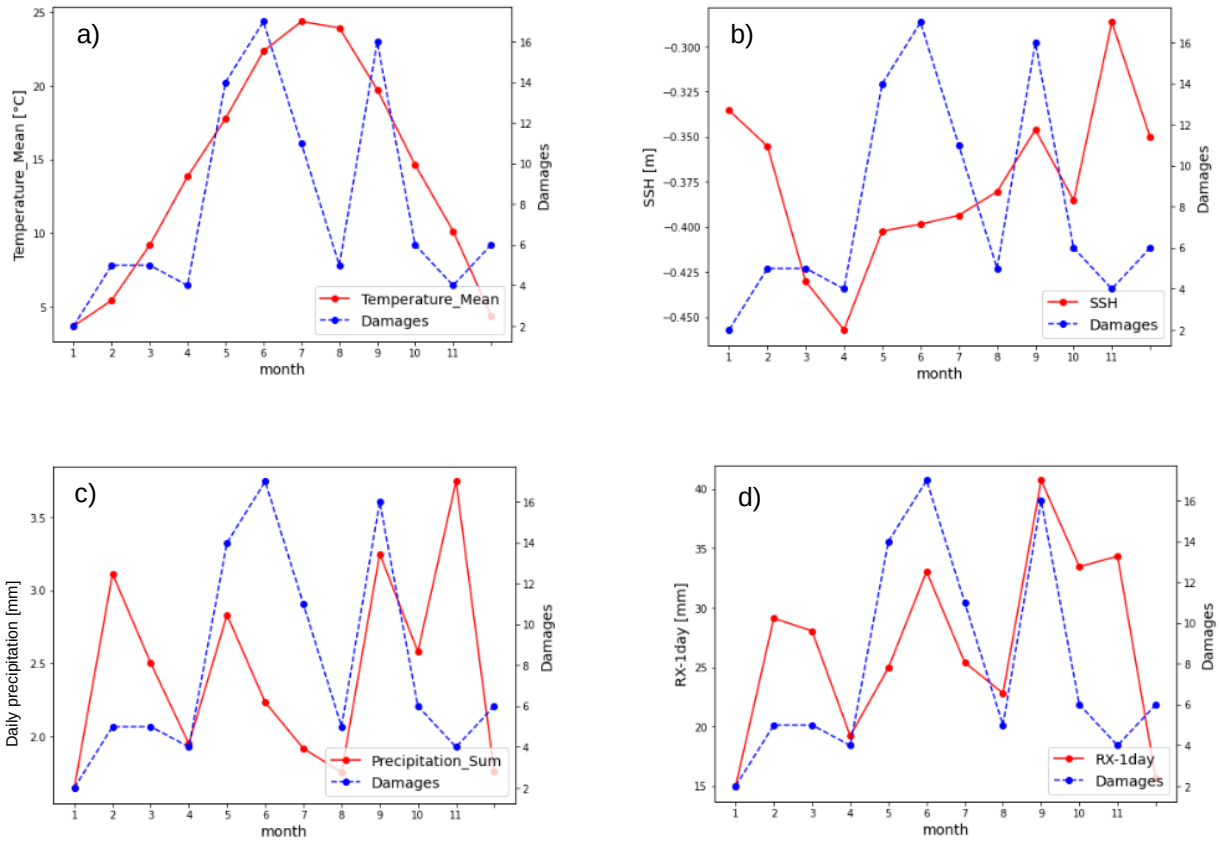
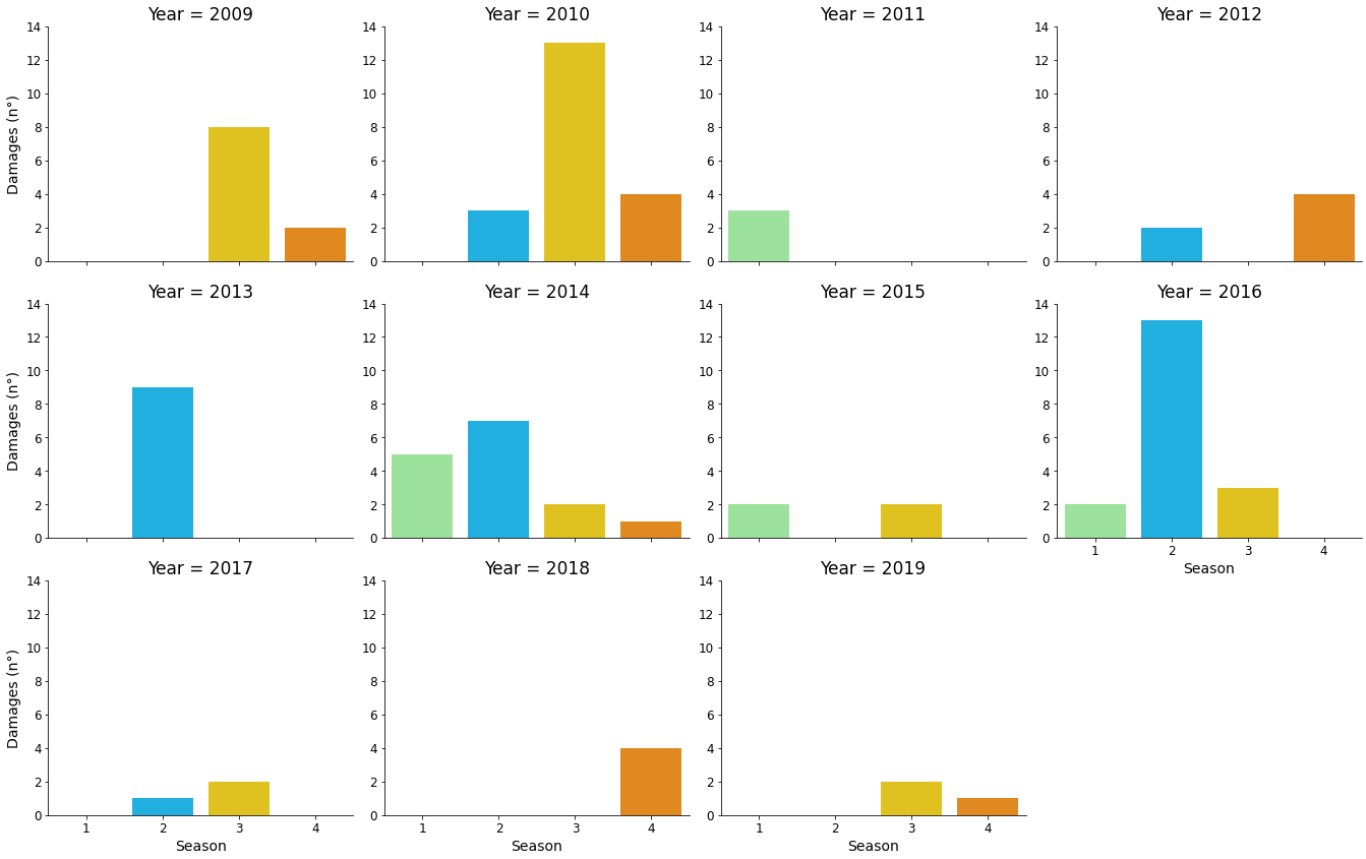


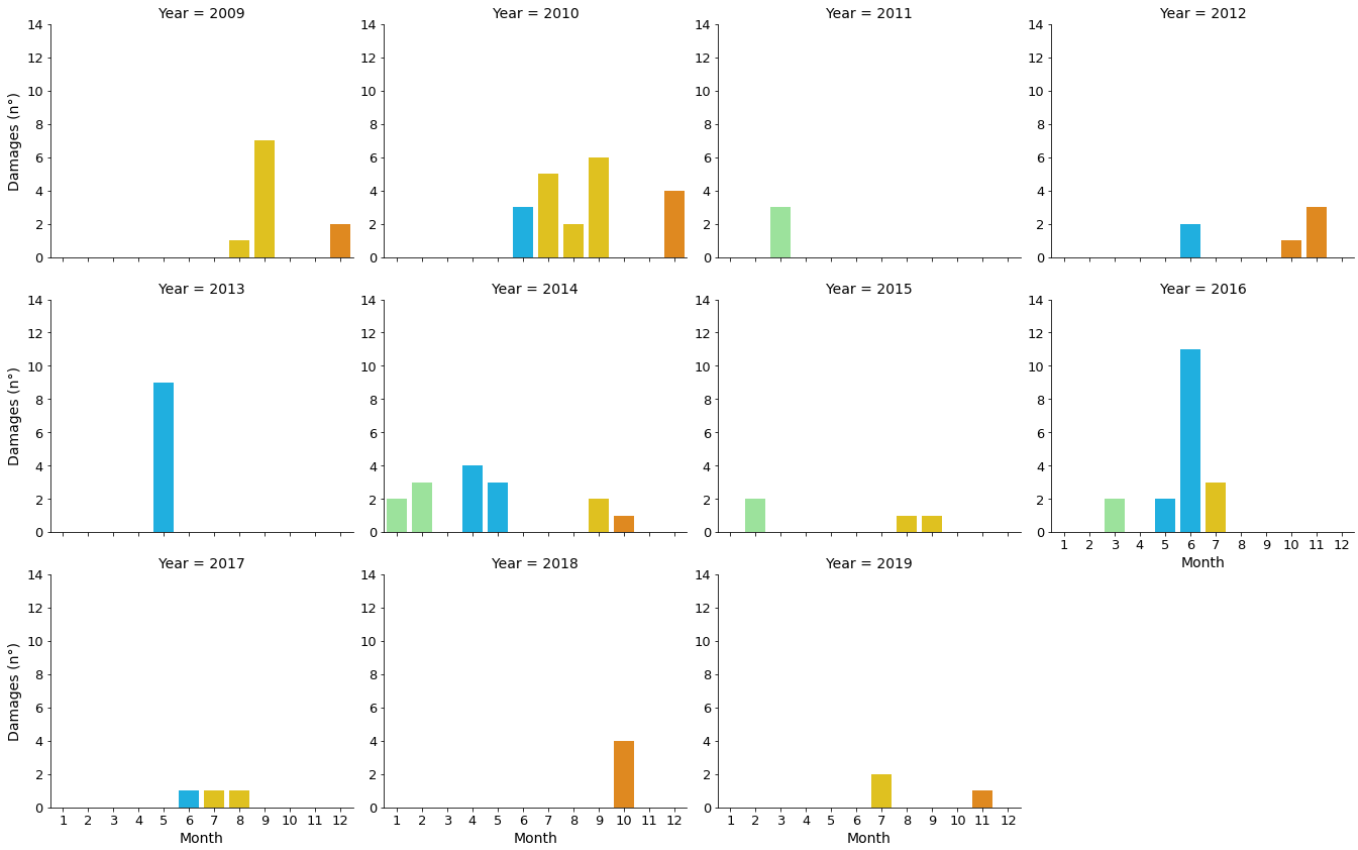
Figure III.2: Monthly number of damages confronted with the mean monthly values of the variables: a) mean temperature; b) SSH; c) daily precipitation; d) RX-1day

ANNEX V: Seasonal and monthly distribution of the damages in the years 2009-2019

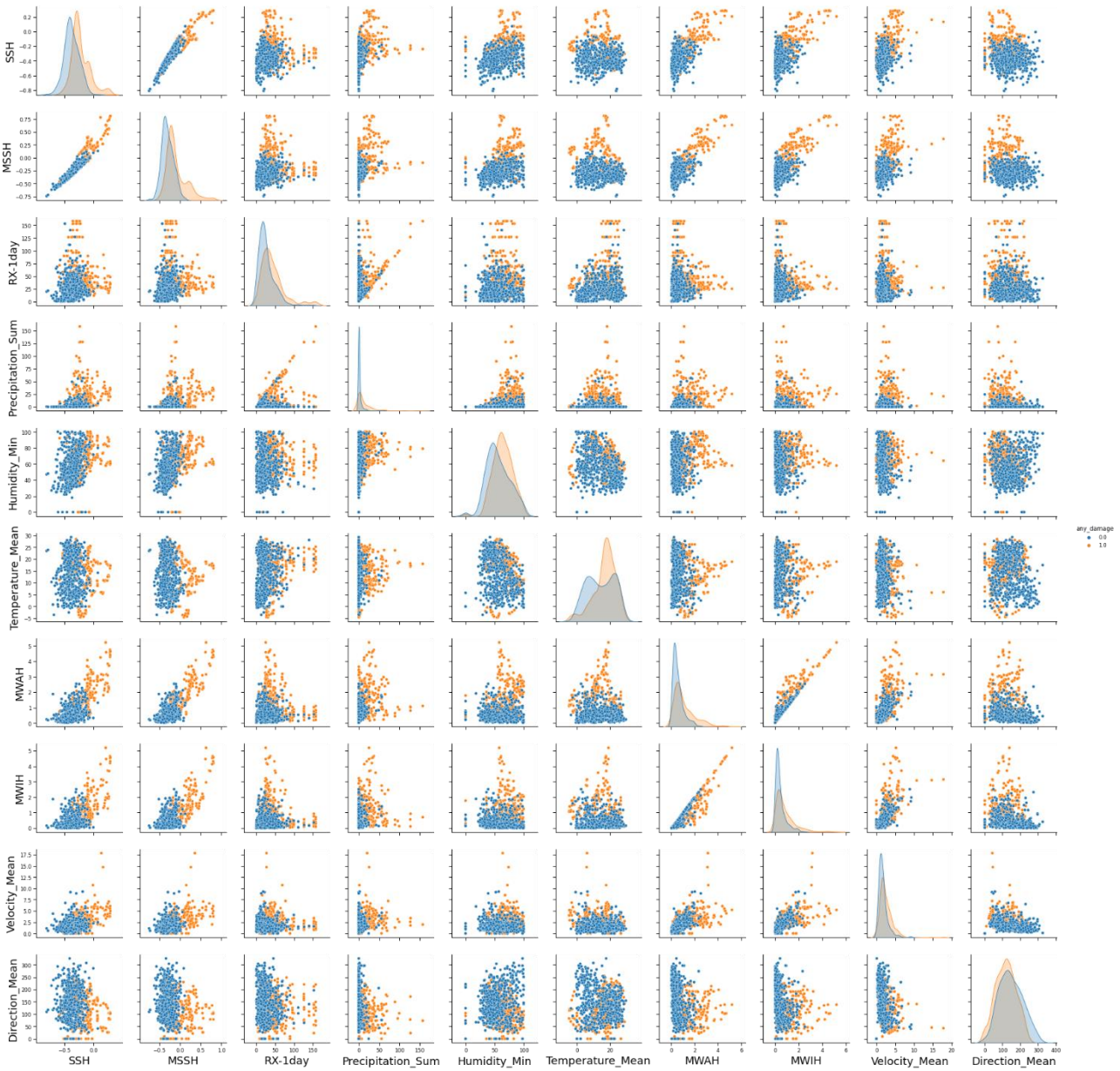
Seasonal analysis: seasonal distribution of the damages for the years within the 2009-2019 timeframe



Monthly analysis: monthly distribution of the damages for the years within the 2009-2019 timeframe



ANNEX VI: Scatterplots between the main hazard variables in damage presence and absence[15]



[15] Observations retrieved from the balanced dataset prepared from the Random Forest implementation (see *Section 5.3.1*)

ANNEX VII: Seasonal distribution of the damages in the 11 investigated municipalities