CUMMULATIVE DISSERTATION

Applications of Machine Learning and Open Geospatial Data in Flood Risk Modelling

by
Fabio Alexander BRILL

A thesis in fulfillment of the requirements for the degree 'Doctor rerum naturalium'

in

Natural Hazards Research

submitted to the Faculty of Science at the University of Potsdam

prepared at the Section Hydrology of the German Research Centre for Geosciences (GFZ)

Submitted: January 3, 2022
Defended: June 17, 2022

Supervisors & Reviewers:
Prof. Dr. Bruno MERZ
PD Dr. Heidi KREIBICH

External Reviewer:
Dr. Guy J.-P. SCHUMANN

Examination Board:
Prof. Dr. Axel BRONSTERT
Prof. Dr. Oliver KORUP
PD Dr. Heidi KREIBICH
Prof. Dr. Bruno MERZ
Dr. Guy J.-P. SCHUMANN
Prof. Dr. Annegret THIEKEN

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-55594
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-555943

Declaration of Originality

I, Fabio Brill, hereby declare that, to the best of my knowledge, this work does not bear resemblance to any other work in whole or in part and has been completed by myself. I did not use any other sources and means than specified. Furthermore, this work has not been previously submitted to any university. All sources have been referred to and this work gives adequate credit to others for their work. I, in no way, claim to have created this information myself.

Date, Signature: _____

Acknowledgements

I wish to thank all those people who have supported me, directly or indirectly, over the years during which this thesis was developed. First and foremost, my deepest gratitude goes to Heidi Kreibich for her supervision. We had an uncountable number of discussions on all sorts of topics, from the conceptual to the detailed, and I really learned to appreciate her forward-looking attitude. Her clear focus on the target has shaped many elements of this work. I am particularly greatful to Bruno Merz, who integrated me in the flood risk group at GFZ and provided me with many valuable insights about the subject matter as well as about the broader world of science. His honest judgement has guided my way of thinking. Special thanks to Guy Schumann, who agreed to invest his time and expertise as external reviewer for this thesis.

I am greatful to the RIESGOS project team and the international partners for all the exciting scientific exchange and for the experience of seeing the risk management in Peru, Chile, and Ecuador first-hand. In that context, I want to highlight Massimiliano Pittore, who I perceive as an inspiring character, and Camilo Gomez, who spent so much effort supporting me to establish local contacts and to better understand the region. My sincere gratitude goes to Bruno Espichán Cuya and Silvia Passuni Pineda of INDECI, who have been very nice and helpful during my trips to Peru and also in all the communication afterwards. Further, I would like to thank Stefan Jäger, André Assmann and Kerstin Büche of geomer for the professional discussions as well as for the good times abroad. Many thanks also to Stefan Schlaffer and Sandro Martinis for the professional collaboration.

As the beginning of a long project may feel like a journey to the unknown, I am particularly indepted to those who have helped me in the early stage of my doctoral studies, when the path was not yet clear. This includes Nivedita Sairam, Tobias Sieg, Viktor Rözer, Patric Kellermann, Max Steinhausen, Lars Gerlitz, Benjamin Kayatz, Kai Schröter and Stefan Lüdtke, who taught me so much about software development. I also want to express my gratitude to Mostafa Farrag, Dung Nguyen, Lukas Schoppa, Gustavo Speckhann, and to all the other colleagues at GFZ for the good atmosphere and the coffee breaks with strict cake rules. There are too many people to list all, but I want to thank the entire NatRiskChange group for expanding my view to the broader field of natural hazards research via conferences, retreats, etc.

Last but not least, I wish to thank my family, as well as my friends and my girlfriend Nina, who have supported me on a personal level through all these years. Thank you so much!

"To see what is general in what is particular,
and what is permanent in what is transitory,
is the aim of scientific thought."

Alfred North Whitehead

Zusammenfassung

Der technologische Fortschritt erlaubt es, zunehmend komplexe Vorhersagemodelle auf Basis immer größerer Datensätze zu produzieren. Für das Risikomanagement von Naturgefahren sind eine Vielzahl von Modellen als Entscheidungsgrundlage notwendig, z.B. in der Auswertung von Beobachtungsdaten, für die Vorhersage von Gefahrenszenarien, oder zur statistischen Abschätzung der zu erwartenden Schäden. Es stellt sich also die Frage, inwiefern moderne Modellierungsansätze wie das maschinelle Lernen oder Data-Mining in diesem Themenbereich sinnvoll eingesetzt werden können. Zusätzlich ist im Hinblick auf die Datenverfügbarkeit und -zugänglichkeit ein Trend zur Öffnung (open data) zu beobachten. Thema dieser Arbeit ist daher, die Möglichkeiten und Grenzen des maschinellen Lernens und frei verfügbarer Geodaten auf dem Gebiet der Hochwasserrisikomodellierung im weiteren Sinne zu untersuchen. Da dieses übergeordnete Thema sehr breit ist, werden einzelne relevante Aspekte herausgearbeitet und detailliert betrachtet.

Eine prominente Datenquelle im Bereich Hochwasser ist die satellitenbasierte Kartierung von Überflutungsflächen, die z.B. über den Copernicus Service der Europäischen Union frei zur Verfügung gestellt werden. Große Hoffnungen werden in der wissenschaftlichen Literatur in diese Produkte gesetzt, sowohl für die akute Unterstützung der Einsatzkräfte im Katastrophenfall, als auch in der Modellierung mittels hydrodynamischer Modelle oder zur Schadensabschätzung. Daher wurde ein Fokus in dieser Arbeit auf die Untersuchung dieser Flutmasken gelegt. Aus der Beobachtung, dass die Qualität dieser Produkte in bewaldeten und urbanen Gebieten unzureichend ist, wurde ein Verfahren zur nachträglichen Verbesserung mittels maschinellem Lernen entwickelt. Das Verfahren basiert auf einem Klassifikationsalgorithmus der nur Trainingsdaten von einer vorherzusagenden Klasse benötigt, im konkreten Fall also Daten von Überflutungsflächen, nicht jedoch von der negativen Klasse (trockene Gebiete). Die Anwendung für Hurricane Harvey in Houston zeigt großes Potenzial der Methode, abhängig von der Qualität der ursprünglichen Flutmaske.

Anschließend wird anhand einer prozessbasierten Modellkette untersucht, welchen Einfluss implementierte physikalische Prozessdetails auf das vorhergesagte statistische Risiko haben. Es wird anschaulich gezeigt, was eine Risikostudie basierend auf etablierten Modellen leisten kann. Solche Modellketten sind allerdings bereits für Flusshochwasser sehr komplex, und für zusammengesetzte oder kaskadierende Ereignisse mit Starkregen, Sturzfluten, und weiteren Prozessen, kaum vorhanden. Im vierten Kapitel dieser Arbeit wird daher getestet, ob maschinelles Lernen auf Basis von vollständigen Schadensdaten einen direkteren Weg zur Schadensmodellierung ermöglicht, der die explizite Konzeption einer solchen Modellkette umgeht. Dazu wird ein staatlich erhobener Datensatz der geschädigten Gebäude während des schweren *El Niño* Ereignisses 2017 in Peru verwendet. In diesem Kontext werden auch die Möglichkeiten des Data-Mining zur Extraktion von Prozessverständnis ausgelotet. Es kann gezeigt werden, dass diverse frei verfügbare Geodaten nützliche Informationen für

die Gefahren- und Schadensmodellierung von komplexen Flutereignissen liefern, z.B. satellitenbasierte Regenmessungen, topographische und hydrographische Information, kartierte Siedlungsflächen, sowie Indikatoren aus Spektraldaten. Zudem zeigen sich Erkenntnisse zu den Schädigungsprozessen, die im Wesentlichen mit den vorherigen Erwartungen in Einklang stehen. Die maximale Regenintensität wirkt beispielsweise in Städten und steilen Schluchten stärker schädigend, während die Niederschlagssumme in tiefliegenden Flussgebieten und bewaldeten Regionen als aussagekräftiger befunden wurde. Ländliche Gebiete in Peru weisen in der präsentierten Studie eine höhere Vulnerabilität als die Stadtgebiete auf. Jedoch werden auch die grundsätzlichen Grenzen der Methodik und die Abhängigkeit von spezifischen Datensätzen and Algorithmen offenkundig.

In der übergreifenden Diskussion werden schließlich die verschiedenen Methoden – prozessbasierte Modellierung, prädiktives maschinelles Lernen, und Data-Mining – mit Blick auf die Gesamtfragestellungen evaluiert. Im Bereich der Gefahrenbeobachtung scheint eine Fokussierung auf neue Algorithmen sinnvoll. Im Bereich der Gefahrenmodellierung, insbesondere für Flusshochwasser, wird eher die Verbesserung von physikalischen Modellen, oder die Integration von prozessbasierten und statistischen Verfahren angeraten. In der Schadensmodellierung fehlen nach wie vor die großen repräsentativen Datensätze, die für eine breite Anwendung von maschinellem Lernen Voraussetzung ist. Daher ist die Verbesserung der Datengrundlage im Bereich der Schäden derzeit als wichtiger einzustufen als die Auswahl der Algorithmen.

Summary

Technological progress allows for producing ever more complex predictive models on the basis of increasingly big datasets. For risk management of natural hazards, a multitude of models is needed as basis for decision-making, e.g. in the evaluation of observational data, for the prediction of hazard scenarios, or for statistical estimates of expected damage. The question arises, how modern modelling approaches like machine learning or data-mining can be meaningfully deployed in this thematic field. In addition, with respect to data availability and accessibility, the trend is towards open data. Topic of this thesis is therefore to investigate the possibilities and limitations of machine learning and open geospatial data in the field of flood risk modelling in the broad sense. As this overarching topic is broad in scope, individual relevant aspects are identified and inspected in detail.

A prominent data source in the flood context is satellite-based mapping of inundated areas, for example made openly available by the Copernicus service of the European Union. Great expectations are directed towards these products in scientific literature, both for acute support of relief forces during emergency response action, and for modelling via hydrodynamic models or for damage estimation. Therefore, a focus of this work was set on evaluating these flood masks. From the observation that the quality of these products is insufficient in forested and built-up areas, a procedure for subsequent improvement via machine learning was developed. This procedure is based on a classification algorithm that only requires training data from a particular class to be predicted, in this specific case data of flooded areas, but not of the negative class (dry areas). The application for hurricane Harvey in Houston shows the high potential of this method, which depends on the quality of the initial flood mask.

Next, it is investigated how much the predicted statistical risk from a process-based model chain is dependent on implemented physical process details. Thereby it is demonstrated what a risk study based on established models can deliver. Even for fluvial flooding, such model chains are already quite complex, though, and are hardly available for compound or cascading events comprising torrential rainfall, flash floods, and other processes. In the fourth chapter of this thesis it is therefore tested whether machine learning based on comprehensive damage data can offer a more direct path towards damage modelling, that avoids explicit conception of such a model chain. For that purpose, a state-collected dataset of damaged buildings from the severe *El Niño* event 2017 in Peru is used. In this context, the possibilities of data-mining for extracting process knowledge are explored as well. It can be shown that various openly available geodata sources contain useful information for flood hazard and damage modelling for complex events, e.g. satellite-based rainfall measurements, topographic and hydrographic information, mapped settlement areas, as well as indicators from spectral data. Further, insights on damaging processes are discovered, which mainly are in line with prior expectations. The maximum intensity of rainfall, for example, acts stronger in cities and steep canyons, while the sum of rain was found more informative

in low-lying river catchments and forested areas. Rural areas of Peru exhibited higher vulnerability in the presented study compared to urban areas. However, the general limitations of the methods and the dependence on specific datasets and algorithms also become obvious.

In the overarching discussion, the different methods – process-based modelling, predictive machine learning, and data-mining – are evaluated with respect to the overall research questions. In the case of hazard observation it seems that a focus on novel algorithms makes sense for future research. In the subtopic of hazard modelling, especially for river floods, the improvement of physical models and the integration of process-based and statistical procedures is suggested. For damage modelling the large and representative datasets necessary for the broad application of machine learning are still lacking. Therefore, the improvement of the data basis in the field of damage is currently regarded as more important than the selection of algorithms.

Contents

List of Figures

List of Tables

Abbreviations

AEP	Aggregated Exceedance Probability
AOI	Area of Interest
ATKIS	Amtliches Topographisch-Kartographisches Informationssystem (Dataset)
AUC	Area Under (the) Curve
BfG	Bundesanstalt für Gewässerkunde (Agency)
BKG	Bundesamt für Kartographie und Geodäsie (Agency)
BMBF	Bundesministerium für Bildung und Forschung (Agency)
BSI	Bare Soil Index
BSVM	Biased Support Vector Machine (Algorithm)
CNN	Convolutional Neural Network (Algorithm)
COFOPRI	Organismo de Formalización de la Propiedad Informal (Agency)
CORINE	Coordinated Information on the Environment (Dataset)
CPU	Central Processing Unit
CRU	Climate Research Unit (Agency)
CUDA	Compute Unified Device Architecture
DEM	Digital Elevation Model
DFG	Deutsche Forschungsgemeinschaft (Agency)
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
DWD	Deutscher Wetterdienst (Agency)
EAD	Expected Annual Damage
EB	Error Bias
EMSR	Emergency Management Service Rapid Mapping (Copernicus)
EUR	EURO (Currency)
FLEMO	Flood Loss Estimation Model (Software)
FN	False Negative
FNF	Forest/Non-Forest map of the TanDEM Project (Dataset)
FP	False Positive
G20	Group of Twenty (Intergovernmental Forum)
GDAL	Geospatial Data Abstraction Library (Software)
GDDR5	Graphics Double Data Rate 5
GEE	Google Earth Engine (Software)
GIS	Geo Information System
GPCC	Global Precipitation Climatology Centre (Agency)
GPM	Global Precipitation Monitoring (Dataset)
GPU	Graphical Processor Unit
GRASS	Geographic Resources Analysis Support System (Software)
GSW	Globas Surface Water (Dataset)
GUF	Global Urban Footprint (Dataset)
HAND	Heigth Above (the) Nearest Drainage
HCFCD	Harris County Flood Control District (Agency)
HEC-GeoRAS	Hydrologic Engineering Center's Geographical River Analysis System
HWM	High Water Mark

INDECI	Instituto Nacional de Defensa Civil (Agency)
LinSVM	Linear Support Vector Machine (Algorithm)
MAE	Mean Absoluve Error
MBE	Mean Bias Error
MERIT	Multi-Error-Removed Improved-Terrain (Dataset)
MGB	Modelo de Grandes Bacias (Software)
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology (Dataset)
NB	Naïve Bayes (Algorithm)
NDVI	Normalized Difference Vegetation Index
NED	National Elevation Data (Dataset)
NOAA	National Oceanic and Atmospheric Administration (Agency)
NSE	Nash-Sutcliffe Efficiency
NUTS	Nomenclature of Territorial Units for Statistics
NWS	National Weather Service (Agency)
OCC	One-Class Classification
OEP	Occurrence Exceedance Probability
OGC	Open Geospatial Consortium
OGD	Open Geospatial Data
OLR	Ordinal Logistic Regression (Algorithm)
OPTICS	Ordering Points to Identify the Clustering Structure (Algorithm)
OSM	OpenStreetMap
PDP	Partial Dependence Plot
PN	Positive (and) Negative
PU	Positive (and) Unlabeled
RAM	Random-Access Memory
RBF	Radial Basis Function
RF	Random Forest (Algorithm)
RFM	Regional Flood Model (Software)
RIM	Regional Inundation Model (Software)
RMSE	Root Mean Square Error
ROC	Receiver-Operator Characteristic
RWG	Regional Weather Generator (Software)
SAR	Synthetic Aperture Radar
SCS	Soil Conservation Service (Agency)
SDG	Sustainable Development Goals
SPI	Stream Power Index
SVM	Support Vector Machine (Algorithm)
SWE	Shallow Water Equations
SWIM	Soil and Water Integrated Model (Software)
TN	True Negative
TP	True Positive
TPI	Topographic Position Index
TRMM	Tropical Rainfall Measurement Mission (Dataset)
t-SNE	t-Distributed Stochastic Neighbor Embedding (Algorithm)
TVAR	Tail Value at Risk
TWI	Topographic Wetness Index
TX	Texas (USA)
USGS	United States Geological Survey (Agency)
VAR	Value at Risk
WPS	Web Processing Service

Chapter 1

Motivation & Objectives

1.1 The scope of flood risk modelling

Taming nature and using its resources for a safe and prosperous life has been the pursuit of mankind over the course of centuries. Natural hazards counteract these ambitions by threatening life and property. Risk management attempts to minimize their consequences. Providing actionable insights that allow for informed decision- making is the task of risk modelling.

Floods are the hazard affecting the most people globally and responsible for the gross of economic damage (Visser et al., 2014). While fatalities from floods have declined over the last decades (Merz et al., 2021; Luo et al., 2015), recent events made clear that still today hundreds of people can die of flooding, even in the most economically developed countries. Examples include the two major catastrophes that are investigated as case studies in this thesis: the coastal *El Niño* 2017 in Peru (Son et al., 2019), and hurricane Harvey in the USA the same year (HCFCD, 2018), as well as the 2021 flood event in Germany (Schäfer et al., 2021).

Risk is a quantity of potential negative consequences with a probability assigned - a definition that has its roots in the work of the early Chicago school economist Frank Knight, who separated calculable risk from uncertainty (Knight, 1921; Meyer et al., 2008). In the natural hazard context, the conceptual risk equation is typically denoted as a product of hazard intensity, exposure and vulnerability (Kron, 2005; Lavell et al., 2012), and all three components may change over time.

The globally observed increase in damage from all natural hazards over the last decades (CRED and Guha-Sapir, 2020; MunichRe, 2019) has been primarily attributed to changes in exposure, i.e. population growth, economic development and construction in hazardous areas (Merz et al., 2021; Bouwer, 2011; Visser et al., 2014). Hydro-meteorological hazards, such as floods, are additionally intensified by climate change, as a warmer atmosphere generally speaking can store and release more water and energy (Fischer and Knutti, 2016; Myhre et al., 2018; Zeder and Fischer, 2020). Therefore more damaging pluvial events are expected (Kaspersen et al., 2017; Bronstert et al., 2020). Although more precipitation alone does not necessarily translate to more river floods (Sharma et al., 2018), there are clear regional trends (Blöschl et al., 2019; Hirabayashi et al., 2021; Tabari, 2020). Also a shift in seasonality may play a role (Brunner et al., 2020; Rottler et al., 2021), as well as variations in the large-scale circulation patterns such as the Asian Monsoon or the *El Niño* Southern Oscillation (Loo et al., 2015; Cai et al., 2015). Whether a climate signal is already visible in observed flood damage statistics has been debated (Bouwer, 2011; Nicholls, 2011; Visser et al., 2014), but a recent study suggests that climate-induced trends are detectable in damage time series when disaggregating

regions according to hydrological criteria rather than by socio-economic similarity (Sauer et al., 2021). The Paris agreement of 2015 announced the political will to limit global warming to "well below 2°C above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5°C" (https://unfccc.int/sites/default/files/english_paris_agreement.pdf, accessed 06.09.2021). However, with the 1.5°C target almost out of reach (Jewell and Cherp, 2019; Warszawski et al., 2021), adaptation becomes more and more relevant. Through the lens of the risk equation, climate change mitigation (i.e. reduction of climate forcings like greenhouse gasses) translates to reducing the hazard component. Climate adaptation primarily attempts to adjust the other two components (Lavell et al., 2012), namely exposure and vulnerability, i.e. the amount of loss or damage suffered from the same hazard intensity. Vulnerability is expected to decrease globally, although with regional differences (Jongman et al., 2015; Formetta and Feyen, 2019; Sauer et al., 2021). Further reduction in exposure and vulnerability is envisaged to be achieved according to the concept of "building back better" (Clinton, 2006; Fernandez and Ahmed, 2019), risk-informed spatial planning, as well as improvements in the early warning and risk communication, which is expected to enhance private precaution (Sairam et al., 2019; Kreibich et al., 2021) – and as a last resort: risk transfer by insurance solutions. Of course some adaptation measures may also alter the local hazard characteristics, for example blue-green infrastructure (Hamel and Tan, 2021), sponge cities (Zevenbergen et al., 2018) and obviously dike construction.

The Sendai framework outlines current priorities for natural hazard risk reduction in the context of the United Nations Sustainable Development Goals (SDG) (https://www.preventionweb.net/files/43291_sendaiframeworkfordrren.pdf, accessed 06.09.2021). Natural hazards often act as setback for progress on the SDGs, either directly by increased mortality, spread of diseases, and destruction of relevant infrastructure, or indirectly. Even "mere" economic loss may lead to social, cultural, health and environmental damage, and on the personal level also cause psychological issues (e.g. Bubeck et al., 2017; Laudan et al., 2020). Natural hazard risk is therefore closely related to questions of humanitarian aid and displacement (Kakinuma et al., 2020), compensations (Calliari et al., 2020), social injustice (Penning-Rowsell et al., 2016), and systemic criticality (Kruse et al., 2021). Scientists and project initiatives are explicitly addressed in the Sendai final document to develop new models to assess hazard and risk and help strengthening local technical capacity. The use of geospatial information technology is encouraged, for example to share maps with potentially affected population and stakeholders. Improvement in risk modelling is thus identified as an important building block of risk management and adaptation.

Mathematical models are needed for various tasks along the risk management cycle, i.e. during the phases of monitoring, emergency response, recovery and reconstruction, risk analysis, and risk reduction (e.g. Kreibich et al., 2014; Spekkers et al., 2017). Arguably, all models developed and applied in this context may therefore be considered part of the broader field of flood risk modelling. In the narrow sense, the term risk modelling is often used to refer specifically to the prediction of negative consequences (e.g. number of affected people, damage to buildings or goods, business interruption time) in relation to the event return period (Merz et al., 2010). Especially in the insurance sector and related fields, risk is strictly defined as the expected value of monetary loss (or insurance claims) per time span of interest, typically a year (Olsen et al., 2015; Merz et al., 2010). Risk models in the latter sense are not only used for adjusting insurance premiums, but also for conducting cost-benefit analyses of adaptation measures (Hall et al., 2012) or to judge the regional distribution of governmental investments in risk reduction efforts (e.g. Sairam et al., 2021).

1.2 State of the art

Hazard assessment is the first step of any risk assessment. In the case of real flood events, hazard information can be taken from observations such as satellite or aerial footage, or even field work, potentially recreated in a hydrodynamic model (e.g. Jiménez-Jiménez et al., 2020; Wing et al., 2019). For statistical risk simulations, a catalogue of realistic scenarios needs to be generated, which can be achieved by coupling a precipitation time series, e.g. from a stochastic weather generator, to a hydrological and subsequent hydrodynamic model (e.g. Falter et al., 2015; Trigg et al., 2016). Although such model chains are by now even used on the global scale (Ward et al., 2013), the implemented process description details are limited in tradeoff for computational time demand, and the results consequently associated with large uncertainties (Trigg et al., 2016; Vorogushyn et al., 2018). The effect of model complexity on the resulting risk estimates is rarely quantified (Apel et al., 2008). Another form of hazard assessment can be found in literature under the term susceptibility mapping (Chapi et al., 2017; Tehrany et al., 2018; Yousefi et al., 2020; Madhuri et al., 2021; Islam et al., 2021), which consists of predicting the potential spatial extent of flooding, often in categorical classes of probability, by a statistical model trained on observed flood locations and geospatial information layers such as topography, hydrography, lithology and vegetation.

When the consequence of interest is building damage, exposure data is traditionally taken either from land use maps, or in small scale applications from individually mapped building types (de Moel et al., 2015). For damage estimation, several schools of thought are generally encountered: On the one hand the theoretical approach of assigning loss functions to types of buildings, either by experts in the field (e.g. Penning-Rowsell and Chatterton, 1977) or from laboratory measurements and analytical models of structural integrity of buildings under force of water (Mazzorana et al., 2014; Jalayer et al., 2016), which could also be termed the construction engineer approach. On the other hand are the empiricists, collecting damage data from real hazard events, and applying regression models (e.g. Thieken et al., 2008; Kreibich et al., 2010; Ettinger et al., 2016). Most established damage models are simple in design (Gerl et al., 2016) and reliability has been questioned in literature (Bubeck et al., 2011; Molinari et al., 2020), but probabilistic methods are now emerging to at least communicate the associated uncertainty (Schröter et al., 2014; Rözer et al., 2019; Schoppa et al., 2020). A completely different approach is indicator-based risk mapping in a geo information system (GIS) (Greiving et al., 2006; Fedeski and Gwilliam, 2007; Meyer et al., 2008; Müller et al., 2011; Aksha et al., 2020). While the cited approaches have the advantage of being able to account for multiple dimensions of risk, in some cases even multiple hazard types, they suffer from a severe drawback: these studies rely on expert-assigned weights in calculation (e.g. Delphi method, analytical hierarchy process).

Among all flood types, there has arguably been a lot of progress on the subtopic of fluvial floods, as they occur in central Europe, northern America, and eastern Asia – places where most internationally visible research on flood damage functions originates from (Gerl et al., 2016). These floods, although potentially disastrous, are characterized by relatively slow-rising water levels over a large catchment, where driving factors are either long-lasting precipitation over the same area, repeated cyclone tracks e.g. due to atmospheric blocking, intense snow melt or ice-jam events (e.g. Merz et al., 2021). Such events can be observed as they materialize, and consequently forecasted with at least several hours up to a few days lead time, allowing for emergency actions and evacuation orders to be issued (Kreibich et al., 2017). The potentially affected areas for a worst-case scenario can be coarsely estimated from a digital elevation model and/or

hydrodynamic modelling (Magnini et al., 2021; Mudashiru et al., 2021), therefore it is possible to inform inhabitants of risk-prone areas to create awareness and hopefully also preparedness (Osberghaus and Hinrichs, 2020). More difficult for modelling and management are the topics of urban pluvial flooding, flash flooding, and compound events. In the first case, extreme local convective precipitation – which is difficult to predict by large-scale models – leads to surface runoff and ponding water either directly on sealed ground or due to infiltration excess and overflowing sewers (Rözer et al., 2021). The second type of flood is similar in the making, but additionally requires water to be channeled in steep terrain, where a normally quiet stream rapidly turns into a raging current, picking up large objects such as boulders or cars, and wreaking havoc within a small town or suburb (e.g. Bronstert et al., 2018; Thouret et al., 2014). Compound events consist of multiple event types that occur in a cascade or simultaneously (Zscheischler et al., 2020). They often come as a surprise and therefore require utmost attention of risk managers (Merz et al., 2015). A merged risk assessment for fluvial and pluvial or other compound event types is still an open challenge (Couasnon et al., 2020; Thieken et al., 2021; Tilloy et al., 2019). It has been speculated that the sheer complexity of the task of multi-risk modelling has been a major obstacle preventing scientists and practitioners from developing models for such scenarios (Kappes et al., 2012). A project initiative that did attempt to develop such a model chain is the RIESGOS project (Pittore et al., 2020), (`www.riesgos.de/`), in which the author of this dissertation was included. The multi-risk model chain developed during RIESGOS is designed in a decentralized manner, so that experts of each discipline can develop and maintain their own model component. Instead of creating risk curves from a catalogue of events with explicit return periods, which are anyway difficult to estimate for rare multi-hazards, the project followed a scenario approach. The detailed scenarios are suited for exploration during exercises with potentially affected stakeholders or risk managers, and can serve as basis for communication and identification of critical elements, as shown by (Gomez-Zapata et al., 2021). However, such compound event scenarios remain hypothetical and associated with large uncertainties.

1.3 The era of open data and machine learning

We are currently witnessing rapid advances in the availability, accessibility and standardization of data and research software. First of all, vast streams of data about our planet are collected every day by sensors both on the ground and in space (Hart and Martinez, 2006; Laur and Liebig, 2010; Li et al., 2015). With recent initiatives such as the Copernicus program of the European Union (Aschbacher and Milagro-Pérez, 2012), that provides data free-of-charge, and many new commercial providers (Pabian, 2015), satellite imagery both in the optical and microwave spectrum is becoming available on a daily basis. Derived products from this kind of data include, among others, digital elevation models, land cover maps, soil moisture content, biomass estimates, flood masks and post-disaster damage assessments (Rizzoli et al., 2017; Wagner et al., 2013; Bégué et al., 2020; Taubenböck et al., 2011). Further, there are efforts to compile and interpolate existing measurements such as soil profiles (Hengl et al., 2017), and crowd-sourced mapping initiatives like OpenStreetMap (OpenStreetMap contributors, 2017).

The open data movement, visible at least since the 2007 gathering of tech visionaries at Sebastopol (OpenDataWorkingGroup, 2007), calls for free access to governmental and research data to support transparency and collaboration (e.g. Geiger and Lucke, 2012; Schrock and Shaffer, 2017). Geo- and environmental sciences have been at the core of the open data initiative, due to the trans-border nature of the field

that does not stick to political boundaries (e.g. Yu and Robinson, 2012; Turner et al., 2015). In the context of global change studies it is self-explanatory that data collected from different locations, and consequently by different institutions, needs to be merged to obtain a complete picture. A call for open data with standardized format is a logical consequence. Also, governmental data is often useful in the environmental science context (Roche et al., 2020), and more specifically in the flood risk modelling domain, for example cadastral information or geospatial maps of protected habitats, polder areas or dike lines. Since 2016, the G20 leaders explicitly support the clearly outlined FAIR principles: Findability, Accessibility, Interoperability, and Reuse (https://www.go-fair.org/fair-principles/, accessed 03.01.2022) (https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967, accessed 03.01.2022).

Simultaneously, the open access movement demands free-of-charge scientific publications (https://www.budapestopenaccessinitiative.org/read, accessed 30.09.2021). Specific journals are emerging to foster the sharing and citing of research data (e.g. Data, Scientific Data, Earth System Science Data). Hydrologically relevant data include an inventory of dams in Germany (Speckhann et al., 2021), a database on damaging floods (Paprotny et al., 2018), or the geospatial version of the emergency database GDIS (Rosvold and Buhaug, 2021). Also, mandatory public access to programming code developed in tax-funded projects has already been requested (https://publiccode.eu/, accessed 03.01.2022). Scientific code development for reproducible analyses is gradually professionalized by drawing upon established workflows from the tech industry, such as continuous integration on a version control software (Perez-Riverol et al., 2016).

As of today, the open geospatial consortium (OGC) issues standards and protocols for a variety of geodata formats, such as GeoTIFF (http://www.opengis.net/doc/IS/GeoTIFF/1.1, accessed 03.01.2022), and for software components like web processing services (WPS). Such conventions ease the integration of several components to a more complex modular processing architecture. An example with practical use in the natural hazard risk modelling domain is the previously mentioned RIESGOS model chain. Particularly for remote sensing imagery, a recent trend for data providers is to deliver Analysis Ready Data (Dwyer et al., 2018). This means that common preprocessing steps, some of which are computationally intensive or difficult to implement, are conducted by the data provider. The users can then skip these steps and directly start with data analysis and interpretation. Spatiotemporal data can further be organized in data cubes (Giuliani et al., 2019), such as https://www.opendatacube.org, accessed 03.01.2022). This concept is also emerging beyond remote sensing imagery, for example the Earth System Data Lab (Mahecha et al., 2020) used to study multivariate dynamics of the earth system. The most famous example up to now is Google Earth Engine (Gorelick et al., 2017), which comes with direct access to high performance computing power on Google servers.

Extracting information from these vast new data sources requires adequate analytical methods. Such have been developed by the statistical and computer science communities under the term machine learning (ML). A key element of these methods is that they discover patterns automatically and do not require manual inspection of endless variable combinations (e.g. Murphy, 2012). Algorithms can be grouped into supervised, unsupervised, and semi-supervised methods (e.g. Murphy, 2012; Géron, 2017). Supervised methods require labeled training data. Unsupervised methods are mainly clustering or dimensionality reduction, that operate without any labels. Semi-supervised refers to methods that use partially labeled examples, and a large pool of unlabeled data. Many of these methods are not entirely novel, e.g. the roots of Artificial

Neural Networks date back to the 1950s (Rosenblatt, 1958) and non-linear Support Vector Machines were invented in the 1990s (Cortes and Vapnik, 1995).

As Leo Breiman, inventor of the famous Random Forest algorithm (Breiman, 2001a), pointed out, there are two different cultures of statistical analysis: the more traditional school of statisticians who, given a classification or regression problem, aim at modelling the "data generating process"; and the computer science inspired algorithmic modellers, who try to solve the original problem directly with a focus on predictive accuracy in trade for model interpretability (Breiman, 2001b). This is more than a simple difference in methods, it is indeed a different philosophy. From the practical point of view, a traditional statistical model requires the user to manually define an equation including the interaction effects to be considered, and to make assumptions about residual distributions. This is also true for Bayesian inference models (e.g. Kruschke, 2015). The workflow typically includes hypothesis testing with accepting or rejecting certain elements. Algorithms used in ML rather rely on internal regularization procedures to handle variable selection, multicollinearity, non-linearity, and interactions in the explanatory features that may prove disturbing in a traditional regression (e.g. Murphy, 2012; Dormann et al., 2012). Thereby, ML algorithms free the user of dealing with these issues in detail, allowing for faster workflows when data is big in terms of sample size or number of variables. However, this is only true as long as predictive accuracy is indeed the only measure of interest. The exploratory counterpart of machine learning is termed data-mining (e.g. Géron, 2017). Some algorithms can be used for both tasks, therefore the separation of terms is not always clear. Besides using unsupervised methods to find structure in unknown data, data-mining can also mean to inspect the structure of predictive models, in search for process understanding (Friedman, 2001; Strobl et al., 2008; Debeer and Strobl, 2020; Zhao and Hastie, 2019).

1.4 Data science in the field of flood risk modelling

Also the geoscience and natural hazard risk domains are being transformed by these methods, however the field does not yet fully embrace the new possibilities (Blair et al., 2019; Reichstein et al., 2019; Wagenaar et al., 2020). Particularly in predictive flood impact modelling, Wagenaar et al. (2020) identify a "potential for significant change" by ML, which they claim to be largely untapped due to economically constrained data collection and a conservative mentality. Also Mack (2017) stretches the point that the skillset required to implement these methods, which usually come in a programming environment first, and the theoretical complexity of the algorithms, are obstacles for people trained as geoscientists or similar. In the broader field of flood risk modelling, the subdomain most advanced with respect to adopting ML is probably satellite image classification, e.g. for rapid flood mapping (Li, Martinis and Wieland, 2019; Bonafilia et al., 2020) or for exposure analysis (Geiß et al., 2017; Wieland and Pittore, 2016). In this context it is straightforward to apply these methods, as large amounts of data are collected in a systematic way by high-precision measurement instruments, and the target label is relatively clear.

In the prediction of flood hazard, process-based physical models are still state-of-the-art for a good reason (Wagenaar et al., 2020). Such type of models make sense when the process in question is well understood, which is the case for fluid dynamics (Stokes, 1880; Brutsaert, 2014). Here the questions are rather how much physical complexity is necessary (Neal et al., 2011), and how remotely sensed flood extents can be used in an assimilation procedure (Dasgupta et al., 2021). The picture is less clear for hydrological

models that make assumptions about the subsurface, for which comprehensive data is typically not available, and for which long short-term memory (LSTM) networks were proposed as an alternative (Clark et al., 2017; Kratzert et al., 2018). Also the field of susceptibility mapping uses geospatial data and machine learning to predict flood hazard, with some authors explicitly stating that hydrological models might be too simple for complex catchment processes (Chapi et al., 2017; Tehrany et al., 2018).

Most obstacles are encountered in the subdomain of damage modelling. Damage models based on ML algorithms have been designed in recent years, but usually trained on a very limited empirical dataset from local flood events (Merz et al., 2013; Laudan et al., 2017; Wagenaar et al., 2017). In addition to the low sample size, the data can contain further biases depending on the data collection method (Flanagin and Metzger, 2008; Stern et al., 2014). Even the influencing variables of damage, besides water depth, are not entirely clear (e.g. Kreibich et al., 2009; Cerri et al., 2021), although some data-mining studies have been conducted (Schröter et al., 2014; Sieg et al., 2017; Rözer et al., 2019).

1.5 Objectives and structure

Several topics have been identified in the introduction that require investigation: The potential of open geospatial data (OGD), and in particular remote sensing products, in the flood risk context is to be tested. With regard to methods, ML workflows in comparison to process-based models for prediction and process understanding seem worth exploring. In terms of application cases, risk estimation methods for compound events are urgently needed. In consideration of these topics, the following specific research questions to be answered in this thesis are defined:

1. Which open geospatial datasets contain useful information for flood hazard and risk modelling?
2. What is the quality of current satellite-based flood masks and how can they be improved?
3. How much are risk estimates by state-of-the-art process-based flood model chains influenced by process description details such as hydrodynamic interactions?
4. Can ML and OGD be used to construct damage models for compound events that are difficult to describe by explicit process-based model chains?
5. To what extent can process understanding be derived from either of these methods?

All in all, this thesis makes a novel contribution to the advancement of the field of flood risk modelling in the broad sense, by sketching new ways of how ML and OGD can be utilized. The identified scientific challenge is not only which algorithm to apply, but rather to develop workflows that enable the meaningful application of algorithms for specific unsolved problems. Further, the thesis presents improvements in the individual components of hazard observation, hazard modelling, and damage modelling. The three following chapters each present an aspect of flood risk modelling at its current limit and push the respective boundaries. The research presented has a methodological axis and a thematic axis (Fig. 1.1). Methods range from process-based to predictive machine learning and finally data-mining for process understanding. Thematically, the thesis evolves from hazard observations, over single hazard comprehensive risk modelling, to compound event damage modelling.

Specifically, Chapter 2 assesses the limitations of currently available satellite-based flood masks and demonstrates how these products can be turned into labeled training areas for subsequent modelling, with the aim of estimating the true flood extent in vegetated and urban areas. Such an improvement is needed for assimilation in hydrodynamic models and for rapid impact assessments. The study also shows that merging the domains of remote sensing, where traditionally satellite sensor data is analyzed, with indicators used in susceptibility mapping, can be fruitful.

In Chapter 3 a state-of-the-art process-based model chain for fluvial flood risk assessment is presented, to investigate the effect of process description details, namely hydrodynamic interactions, on economic risk estimates. Methodically, the study contrasts the ML approaches and demonstrates how a risk estimation could look like when processes are more or less known. However, it also becomes clear that currently implemented process description is not sufficient.

Chapter 4 then moves forward to the domain of damage modelling for a compound *El Niño* flood event, for which methods as in Chapter 3 do not yet exist, using similar methods as in Chapter 2. Classification algorithms are used here to directly model the resulting total structural building damage. By this approach, it is possible to leapfrog over the detailed hazard assessment, and directly derive empirical spatial risk estimates for compound events. The study is based on a comprehensive empirical survey conducted by state authorities, and therefore avoids sample bias. The potential and limitations of fully empirical methods for mapping risk and deriving process understanding are investigated.

The connections between Chapters 2 and 4 are on the methodical level. In both studies, machine learning methods are used in a detailed manner, multiple algorithms are compared and geomorphic indices and other typical and novel indicators from the geoscience domain are used as predictors. In both contexts, the usage of such indicators is not entirely uncommon, but the way of application is novel: In Chapter 2, the geo-features are used in a second step, i.e. not in combination with the satellite sensor data but afterwards. In Chapter 4, the features are used to directly predict damage probability, rather than hazard. Thereby they act as complex intensity metrics. Chapters 2 and 3 connect on the level of data assimilation and model calibration, as well as on the question of how much physical complexity is necessary to describe inundation. Chapters 3 and 4 are further connected via the thematic question of process-based model chains versus ML approaches for risk modelling. Finally, the individual findings are summarized and discussed in the light of the stated research questions in Chapter 5, ending with suggestions for further research and the overall conclusion.

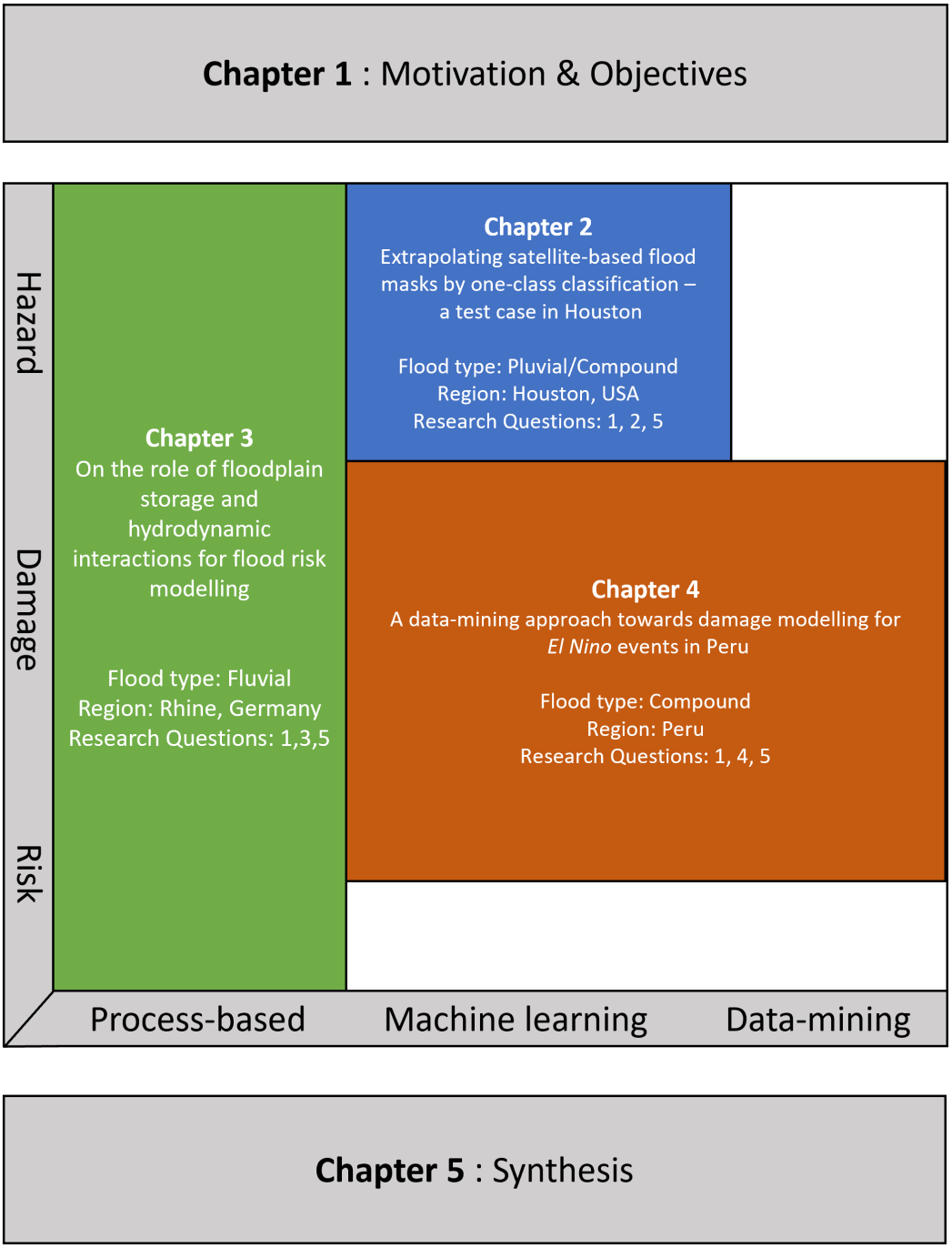

FIGURE 1.1: Structure of this dissertation.

1.6 Author contributions

Chapters 2 and 4 are printed here in the version as published. Chapter 3 has been accepted for publication, but not undergone official proofreading and typesetting at the journal, and is printed here in the latest version at the time this thesis was finalized. The explicit contibutions by the author of this thesis (F.B.) are given below:

Chapter 2: **Brill, F.**, Schlaffer, S., Martinis, S., Schröter, K., Kreibich, H. (2021). Extrapolating Satellite-Based Flood Masks by One-Class Classification—A Test Case in Houston. *Remote Sensing* **13**:2042. https://doi.org/10.3390/rs13112042

Conceptualization: **F.B.**, H.K., K.S.; methodology: **F.B.**, S.S.; formal analysis, investigation, validation, visualization and writing—original draft preparation: **F.B.**; supervision and writing—review and editing: S.S., S.M., K.S., H.K.; funding acquisition: H.K., S.M.; All authors have read and agreed to the published version of the manuscript.

Chapter 3: Farrag, M., **Brill, F.**, Nguyen, D. V., Sairam, N., Schröter, K., Kreibich, H., Merz, B., de Bruijn, K. M., Vorogushyn, S. (2021). On the role of floodplain storage and hydrodynamic interactions for flood risk estimation. *Hydrological Sciences Journal*, in press (accepted).

S.V., K.d.B., and B.M. conceptualised the study. M.F. developed the 1D hydrodynamic code and calibrated both RIM1.0 and RIM2.0 models, carried out 1D-2D model simulations, run the SWIM model, and, together with S.V., analyzed the results and drafted the initial manuscript. D.V.N. run the weather generator, calibrated the SWIM model, and contributed to the analysis of the results. N.S. estimated exposure values, and **F.B.** carried out damage and risk assessment supervised by K.S. and H.K.; All authors contributed to the final manuscript.

Chapter 4: **Brill, F.**, Passuni Pineda, S., Espichán Cuya, B., Kreibich, H. (2020). A data-mining approach towards damage modelling for *El Niño* events in Peru. *Geomatics, Natural Hazards and Risk* **11:1** https://doi.org/10.1080/19475705.2020.1818636

F.B and H.K. have designed the study; **F.B.** has performed the data processing, visualization, and writing of the first draft, supervised by H.K.; S.P.P. and B.E.C. have contributed to the interpretation with local expertise; All authors have made significant contributions to the writing of the final manuscript.

Besides the content of this dissertation, the author has contributed to the following additional publications:

Sairam, N., **Brill, F.**, Sieg, T., Farrag, M., Kellermann, P., Nguyen, V. D., Lüdtke, S., Merz, B., Schröter, K., Vorogushyn, S., Kreibich, H. (2021). Process-Based Flood Risk Assessment for Germany. *Earth's Future* **9**:10. https://doi.org/10.1029/2021EF002259

Gomez-Zapata, J.C., Parrado, C., Frimberger, T., Barragán-Ochoa, F., **Brill, F.**, Büche, K., Krautblatter, M., Langbein, M., Pittore, M., Rosero-Velásquez, H., Schoepfer, E., Spahn, H., Zapata-Tapia, C. (2021). Community Perception and Communication of Volcanic Risk from the Cotopaxi Volcano in Latacunga, Ecuador. *Sustainability* **13**:1714. https://doi.org/10.3390/su13041714

Sairam, N., Schröter, K., Carisi, F., Wagenaar, D., Domeneghetti, A., Molinari, D., **Brill, F.**, Priest, S., Viavattene, C., Merz, B., Kreibich, H. (2020). Bayesian Data-Driven approach enhances synthetic flood loss models. *Environmental Modelling & Software* **132**:104798. https://doi.org/10.1016/j.envsoft.2020.104798.

Chapter 2

Extrapolating satellite-based flood masks by one-class classification — A test case in Houston

2.1 Abstract

Flood masks are among the most common remote sensing products, used for rapid crisis information and as input for hydraulic and impact models. Despite the high relevance of such products, vegetated and urban areas are still unreliably mapped, sometimes even excluded from the analysis. The information content of synthetic aperture radar (SAR) images is limited in these areas due to the side-looking imaging geometry of radar sensors and complex interactions of the microwave signal with trees and urban structures. Classification from SAR data can only be optimized to reduce false positives, but cannot avoid false negatives in areas that are essentially unobservable to the sensor, e.g. due to radar shadows, layover, speckle and other effects. We therefore propose to treat satellite-based flood masks as intermediate products with true positives, and unlabeled cells instead of negatives. This corresponds to the input of a positive-unlabeled (PU) learning one-class classifier (OCC). Assuming that flood extent is at least partially explainable by topography, we present a novel procedure to estimate the true extent of the flood, given the initial mask, by using the satellite-based products as input to a PU OCC algorithm learned on topographic features. Additional rainfall data and distance to buildings had only minor effect on the models in our experiments. All three of the tested initial flood masks were considerably improved by the presented procedure, with obtainable increases in the overall κ score ranging from 0.2 for a high quality initial mask to 0.7 in the best case for a standard emergency response product. An assessment of κ for vegetated and urban areas separately shows that the performance in urban areas is still better when learning from a high quality initial mask.

2.2 Introduction

Satellite-based flood mapping is a central topic in applied remote sensing, due to the high relevance of accurate event maps in all phases of the disaster risk management cycle. Besides the use during emergency response, the observed flood extent is often necessary for post-event analysis, including modelling studies. An emerging field is also the assimilation of flood extents in near-real-time into hydrodynamic models (Hostache et al., 2018). The term flood mask refers to a binary geospatial data layer of flood water extent, where the permanent water bodies are excluded. Most products are currently

based on SAR sensors, which can operate day and night, independent of cloud cover. As the temporal coverage and free-of-charge availability of satellite imagery steadily increases, flood masks of varying quality and file format are produced, e.g. by the Copernicus Emergency Management Service of the European Commission (EMSR). However, there are still obvious limitations to these currently available products, which hamper their usage: (1) Urban flooding is usually underdetected, because built-up areas are difficult to observe from space, due to the occurrence of radar shadows, layover effects, and speckle (e.g. Mason et al., 2010; Pulvirenti et al., 2016). This is for example problematic for damage estimations, which are strongly influenced by the number of exposed buildings within the flood mask (Sieg et al., 2019). (2) Flooding below the vegetation canopy, although theoretically detectable on longer wavelength sensors depending on the density of the canopy (Henderson and Lewis, 2008; Plank et al., 2017), is typically omitted as well, even along river courses, which obscures the true land-water boundary. Algorithms for deriving the water depth from a mapped extent are available, but hinge on the precision of that land-water boundary (Zwenzner and Voigt, 2008; Cian et al., 2018; Cohen et al., 2017; Matgen et al., 2016) as well as on the quality of the elevation data (Schumann et al., 2008). Inundation depth is often required in applications, for example flood damage models usually rely on it as main explanatory feature (exceptions being crop damage models, which may use duration and timing). Therefore, a step towards more reliable flood extent is also a step towards the applicability in hydrodynamic and flood damage models. (3) The often undescribed uncertainty of satellite-based flood masks leads to further problems in applications, e.g. when assessing the performance of a hydraulic model (Stephens et al., 2012). Although some scientific studies provide uncertainty estimates (e.g. Giustarini et al., 2015; Martinis and Rieke, 2015), this is not yet operational standard, e.g. for the EMSR products.

A staggering amount of different methods has already been explored for water delineation from SAR images. Examples include automatic grey level thresholding (Martinis et al., 2009), active contour models (Horritt et al., 2001), fuzzy scoring (Pulvirenti et al., 2011; Twele et al., 2016), time series analysis (Schlaffer et al., 2015), Bayesian networks (Li, Martinis, Wieland, Schlaffer and Natsuaki, 2019) and, recently, convolutional neural networks (Li, Martinis and Wieland, 2019; Bonafilia et al., 2020). Nevertheless, the information content of single-date, single-polarization SAR amplitude data is limited in vegetated and especially in urban environments, which are the most interesting areas with respect to impact estimations. Acknowledging these limitations, the remote sensing community moves towards integration of additional information layers, such as interferometric coherence (Chini et al., 2019; Pulvirenti et al., 2016; Li, Martinis, Wieland, Schlaffer and Natsuaki, 2019), optical data (Wieland and Martinis, 2019), terrain elevation (Mason et al., 2010, 2012; Huang et al., 2017) and even social media content (Scotti et al., 2020). The cited approaches incorporating topographic information use this mainly to exclude false positives. Most notably for this study, a typical postprocessing step is to overlay the classified flood extent with so-called exclusion layers, to reduce false positives from material that exhibits low backscatter, like dry sand (Martinis, 2018). With this in mind, we conclude that flood masks from SAR data can be optimized to reduce false positives, by sophisticated classification methods and exclusion layers, but cannot avoid false negatives in areas which are unobservable for the sensor, e.g. due to the abovementioned effects.

The hydrological and geomorphological communities have developed advanced GIS approaches to delineate flood-prone areas without having to resort to numerical hydraulic models (Samela et al., 2017; Chapi et al., 2017; Tehrany et al., 2018), also with a focus on urban areas (Kelleher and McPhillips, 2019; Mukherjee and Singh, 2019).

While numerical models still have many advantages, and benefit from the increase of computation power, they require bathymetry and discharge or water level as boundary condition, which is not always available. Examples of indicators that have successfully been used in the context of flood susceptibility mapping, e.g. in the cited studies above, are the Height Above Nearest Drainage (HAND) index (Rennó et al., 2008) and the Topographic Wetness Indicator (TWI) (Quinn et al., 1991). In the following, we investigate whether these and other geomorphological features, precipitation, and distance to buildings, are suitable to identify flooded areas, which are not detected on remote sensing products. The approach consists of using a satellite-based flood mask as training area for a machine learning algorithm. The basic research question is: "if this is the satellite-based flood mask, where then should we expect water in reality?".

This question can be expressed as a supervised learning task, in which the labels necessary for training are taken from the initial mask. Supervised models are able to learn complex relationships from the explanatory features by optimizing an objective function that penalizes misclassification of the provided labels in the training samples. However, correct labels are required for training. Regular binary classifiers require positive and negative (PN) examples. We argue in this paper that positive and unlabeled (PU), rather than PN, is the appropriate description of state-of-the-art satellite-based flood masks, as long as the limitations of these masks are not clearly communicated, e.g. in a validity layer. This leads us to formulate our research question as a one-class classification (OCC) problem. OCC algorithms require only one class to be labeld, termed the positive (P) class. They may use either only P or PU training data, thereby avoiding to wrongly treat unknown labels as true negatives. Such methods are commonly used in habitat modelling (Phillips et al., 2006) as well as for specific remote sensing questions like mapping raised bogs (Mack et al., 2016), invasive tree species (Piiroinen et al., 2018), Bark Beetle infestation (Ortiz et al., 2013) or damaged maize fields (Jozani et al., 2020). Mack & Waske (Mack and Waske, 2016) investigated the discriminative power of the well-known PU algorithms MaxEnt (Phillips et al., 2006) and Biased Support Vector Machine (BSVM, (Liu et al., 2003)) in comparison to a P classifier and a PN benchmark model for a variety of classification tasks. PU learning is generally considered more promising than P learning, especially when classes are not perfectly separable, because PU algorithms may learn about the overall distributional characteristics. When using an OCC on satellite-based flood masks, there is no need for a validity layer, as long as false positives have been minimized during the creation of the flood mask (depending on the algorithm, some violation of this assumption is acceptable).

The aim of this study is to improve satellite-based flood masks by reducing false negatives in areas where the satellite sensor has low sensitivity, such as vegetated and urban areas. Our investigation requires a flood event covered by multiple satellite-based flood masks of different quality, relatively high resolution topography, gridded rainfall measurements, and mapped building footprints. Additionally, we use high-quality flood extent maps ("ground truth") for testing the performance of the proposed approach. We chose the well-documented event of 2017 hurricane Harvey in Houston, TX, as test case. We present a novel methodology for extrapolation by OCC, and test it with three different initial satellite-based masks on different spatial scales. The paper is organized as follows: Section 2.3 gives a description of the flood event and used datasets, followed by details on the algorithms, performance metrics, and experimental setup. In Section 2.4, the skill of the BSVM and MaxEnt models is compared, and the effect of a region-growing postprocessing is quantified. Example maps of spatial predictions are shown for selected models. The results are then discussed in a broader context in Section 2.5.

2.3 Materials & Methods

2.3.1 Study area and datasets

Hurricane Harvey ranks among the costliest disasters that affected the United States during the last decades (NHC, 2018), with Houston in particular suffering severe damage in the final days of August 2017. Although considered primarily a pluvial flood event, with implications for modelling (Wing et al., 2019), the vast spatial extent and long duration of the rainfall also caused all major river basins to overflow. According to the Harris County Flood Control District (HCFCD), 70,370 out of 154,170 flooded homes were located beyond the official 500-year flood hazard zone (Lindner and Fitzgerald, 2018). Water levels in the San Jacinto River exceeded all historical records, with estimated return periods above 500 years in many places. In the western part of Houston, two large-scale flood control structures, the Barker reservoir and the Addicks reservoir (Figure 2.1) were forced to open their release gates on August 28, but the water level within continued to rise until August 30 to the point of local overtopping, despite the open gates (HCFCD, 2018). The combined outflow of both reservoirs led to a massive flooding of the Buffalo Bayou. It is reported that about 14,000 homes were even located within the reservoirs themselves.

Flood masks for training and validation

The following products were used in our study as initial flood masks for training the OCC models: The EMSR released a mapping of areas inundated by Hurricane Harvey on 31st of August 2017 (EMSR_229), based on Cosmo-SkyMed data. This is a typical standard product, designed for rapid response. The EMSR_229 mask covers the entire urban area of Houston and surroundings. Li et al. (Li, Martinis, Wieland, Schlaffer and Natsuaki, 2019) further classified parts of a Sentinel-1 scene from August 30th, including interferometric coherence with previous scenes, by a Bayesian Network fusion technique (DLR_BN). Li et al. (Li, Martinis and Wieland, 2019) also processed TerraSAR-X images by a convolutional neural network (DLR_CNN), with the flooded scene dating to September 1st. The latter is only available for a rather small region within the city, along the Buffalo Bayou. Both DLR_BN and DLR_CNN can be regarded as "high quality" masks, with reported κ coefficients of 0.68 in both cases from comparison to a labeled aerial image. However, we observed some flaws in this labeling when comparing it to the raw aerial image.

Validation in our study is based on two independent products: Firstly, we downloaded the original 50 cm resolution aerial image acquired by the National Oceanic and Atmospheric Administration (NOAA) on 08/30/2017 (`https://storms.ngs.noaa.gov/storms/harvey/index.html#9/29.8430/-95.0729`) and manually labeled all flooded areas on the image (NOAA_labeled) in three categories: open flood water, flooded vegetation, and flooded urban area. The land cover classes allow for calculating the model skill in a stratified manner, providing numbers for vegetated and urban areas separately. The guiding principle for assigning these land cover classes was to consider what is visible from the point of view of a satellite. Small patches of open water within built-up environment were still labeled "urban", as the SAR signal in these locations would most likely be influenced by the surrounding buildings. The main channel of the Buffalo Bayou was labeled "vegetation", as there are mainly tree canopies visible from a satellite's perspective. Great care was taken to only include buildings in this reference map where it was obvious, e.g. from the color of the swimming pools, that at least

the ground floor of the building got affected – otherwise we only delineated the visible water on the roads. Permanent lakes within the urban areas were intentionally not mapped, only the flood waters surrounding the regular lake extents. While some residual ambiguity remained between the assigned land cover classes, especially between open water and flooded vegetation inside the large reservoirs, we are confident that this manually labeled image is a very precise reference for the situation on 08/30/2017. This reference map is publicly available as online supplement to this publication. Secondly, we obtained a mapping by the United States Geological Survey (USGS) for the San Jacinto River (USGS_SJ). The USGS has released flood extents for major river catchments (Watson et al., 2018), based on interpolated field measurements of high water marks (HWM), which have been used by the company Fathom (Wing et al., 2019) as "ground-truth" for validating their hydraulic model simulation of the event. Watson et al. (Watson et al., 2018) acknowledge that some uncertainties remain in areas where the coverage of the HWM is sparse and that the mapped boundary was manually extended to anthropogenic structures such as roads or bridges. We overlayed all masks with OpenStreetMap (OSM) waterways polygons (OpenStreetMap contributors, 2017), which include categorized water bodies in high spatial detail, and removed all of these areas from the masks, thereby equally converting all masks to flood masks.

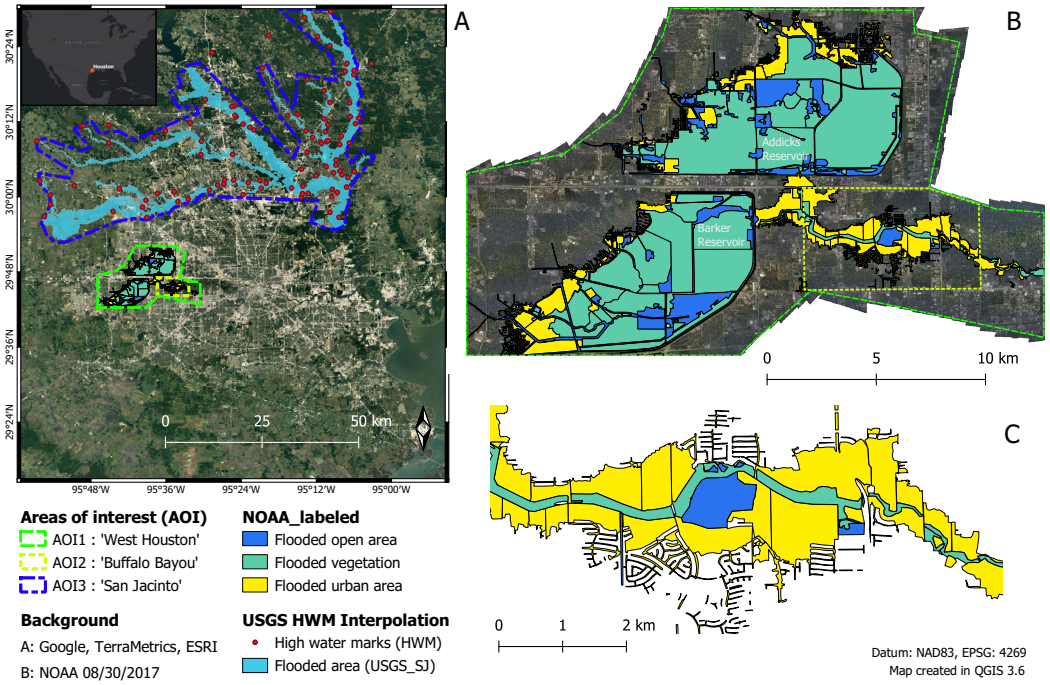

FIGURE 2.1: Overview map including the three AOIs, the USGS interpolation of high water marks (HWM), and the manually labeled aerial image from August 30, 2017. (**A**) Full extent. (**B**) West Houston extent. The large vegetated areas are the Addicks reservoir (north) and Barker reservoir (south). (**C**) Buffalo Bayou extent. Note the detailed mapping of streets. EMSR_229 covers the entire area depicted on subplot A, while DLR_BN is available for AOI1 and DLR_CNN for AOI2.

Explanatory features

An overview of used datasets and features is given in Table 2.1 and Table 2.2. A digital elevation model (DEM) called the National Elevation Dataset (NED) is available from the USGS, based on the best available data source per area (Arundel et al., 2015). We used the 1/3 arc seconds (~10 m) version. From the DEM, different features have been derived: slope, curvature, topographic wetness index (TWI) and topographic position index (TPI). The TPI is a geomorphological measure derived by focal window operations, which in machine learning terminology can be considered a manual convolution on the DEM. TPI has a clear physical meaning, as it indicates local hills and depressions. Combining TPI on multiple scales allows for identifying more complex landscape morphologies (Reu et al., 2013). We used the implementation in the R library spatialEco (Evans, 2021) and computed TPI on the scale of 11, 51, and 101 cells, which corresponds to about 50, 250, and 500 m in all directions. The OSM water layer (`http://hydro.iis.u-tokyo.ac.jp/~yamadai/OSM_water/`) distinguishes 5 types of water bodies in this area, namely 'Ocean', 'Large Lake & River', 'Major River', 'Small Stream' and 'Canal'. We discarded the ocean and merged 'Small Stream' and 'Canal' as these labels appeared to have been used interchangeably from visual inspection in the Houston area. This left us with three different stream layers, for which we computed the HAND and Euclidean distance separately (by GRASS r.watershed and GDAL Proximity). OSM buildings had very limited coverage in Houston at the time of this study, therefore we used Microsoft USBuildingFootprints (`https://github.com/microsoft/USBuildingFootprints`). The Euclidean distance was computed on rasterized shapes, which corresponds to the distance to the closest building cell. Gridded rainfall data was downloaded from the US National Weather Service (NWS) website (`https://water.weather.gov/precip/download.php`). We used the sum of August 26-30, where most of the rainfall occurred in Houston. The accumulated rainfall was computed via the GRASS GIS tool r.accumulate, with the rainfall sum as input. Features were separated into 3 groups for our experiments. Most features were derived from the DEM and/or stream location data, and therefore called 'Topo' features. These were always used. The 'Rain' and 'Buildings' features were added separately to test the effect of the additional data. To keep it simple during processing, we resampled all datasets to the resolution of the DEM, so that all layers could be converted to a raster stack.

TABLE 2.1: Floodmasks for training and validation

Floodmask	Data source	Date of image	Resolution	Usage
EMSR_229	Cosmo-SkyMed	31 August 2017	30 m	Training
DLR_BN	Sentinel-1	30 August 2017	15 m	Training
DLR_CNN	TerraSAR-X	01 September 2017	40 m (32 x 1.25)	Training
NOAA_labeled	Aerial image	30 August 2017	0.5 m	Validation
USGS_SJ	HWM	Maximum extent	3 m	Validation

TABLE 2.2: Datasets and features

Feature	Data source	Category
HAND_large_lake_river	NED + OSM	Topo
HAND_major_river	NED + OSM	Topo
HAND_small_stream_canal	NED + OSM	Topo
Dist_large_lake_river	OSM	Topo
Dist_major_river	OSM	Topo
Dist_small_stream_canal	OSM	Topo
Slope	NED	Topo
Curvature	NED	Topo
TWI	NED	Topo
TPI 11x11	NED	Topo
TPI 51x51	NED	Topo
TPI 101x101	NED	Topo
Rainfall_sum	NWS	Rain
Rainfall_acc	NWS + NED	Rain
Dist_to_buildings	Microsoft USBuildingFootprints	Buildings

2.3.2 Algorithms and performance metrics

OCC algorithms

Two commonly used PU learning algorithms are tested in this study for the purpose of extrapolating satellite-based flood masks from the abovementioned features. BSVM (Liu et al., 2003) is a discriminative algorithm, originally developed for text classification. It was found superior to previous multi-step OCC procedures, and also to other P and PU learners, for classification of remote sensing images (Mack and Waske, 2016). Essentially, it is a support vector machine with radial basis function (RBF) kernel and unequal misclassification penalty terms in the cost function. By assigning higher penalty to misclassified positive samples, the unlabeled samples are considered "negotiable" during training. The biased cost function is given as (Eq. 2.1)

$$\begin{aligned} &\text{Minimize } \frac{1}{2}w^T w + C_+ \sum_{i=1}^{k-1} \xi_i + C_- \sum_{i=k}^{n} \xi_i \\ &\text{Subject to } y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, ..., n \\ &\xi_i \geq 0, i = 1, 2, ..., n \end{aligned} \tag{2.1}$$

where C_+ and C_- are the cost of misclassification for positive and unlabeled samples, respectively. C_+ is in practice parametrized by $C_{Multiplier}$ times C_-. w is the weight vector, x is the feature vector, and y is the corresponding label. ξ is the slack variable used to evaluate potential hyperplanes for k-1 positive and k-n unlabeled samples. Superscript T denotes the inner product.

The so constructed hyperplane has by definition a value of 0, which can be regarded as the 'default' threshold ($\theta_{Default}$) for binary classification of BSVM. The continuous output of BSVM gives the distance to this hyperplane, where higher values indicate

samples associated with the positive class (i.e. flood in our case), and lower values indicate samples associated with the negative class. However, a threshold for binary classification can be set by the user at any value, and it is sometimes recommended in literature not to rely on the default in application, (e.g. Mack and Waske, 2016). For the PN benchmark models, we used a regular (unbiased) support vector machine (SVM), in which the misclassification costs for both classes are equal.

MaxEnt (Phillips et al., 2006) is a generative algorithm with solid roots in information theory and probabilistic reasoning (Jaynes, 1957). The implementation in a stand-alone software, which is now open source (Phillips et al., 2017), is commonly used in ecological modelling as well as for mapping rare land cover classes. The developers phrase the objective of the maximum entropy principle as estimating a distribution that agrees with everything that is known, and at the same time avoiding any assumptions about what is unknown. More specifically, the procedure searches for a Gibbs distribution, under the constraints that the expectation of every feature corresponds roughly to the empirical feature mean, while pertaining a shape as close to the prior distribution as possible. MaxEnt internally computes variance features, product features, threshold features, and hinge features. This allows the algorithm to learn complex responses and interactions, but requires regularization to avoid overfitting. The optimal value of the regularization parameter β is accordingly determined over a grid search. Note that the original formulation by (Phillips et al., 2006) is in geographic space, and in that space the prior distribution is a uniform distribution, i.e. all locations are a-priori equally likely to contain the positive class. More in line with machine learning literature is the formulation in feature space, where the prior is the marginal feature distribution, and MaxEnt estimates the distribution of the positive class by minimizing the relative entropy (Kullback-Leibler divergence) between the positive and marginal distributions under the constraints imposed by the feature means (Elith et al., 2010). The formulation is unconditional, so that only positive and unlabeled data is required. In other words, MaxEnt models the ratio of presence to background, which results in a relative probability. The cost function (Eq. 2.2), in the notation of the authors, can be shown to be the negative log-likelihood with an L1 penalty term.

$$\begin{aligned} &\text{Minimize } \tilde{\pi}[-ln(q_\lambda)] + \sum_j \beta_j |\lambda_j| \\ &\text{Subject to } |\hat{\pi}[f_j] - \tilde{\pi}[f_j]| \leq \beta_j \end{aligned} \tag{2.2}$$

where $\tilde{\pi}$ is the prior distribution, $\hat{\pi}$ the resulting MaxEnt distribution, q_λ the Gibbs distribution, square brackets [] denote the expectation, ln the natural logarithm, β is the cost parameter, and λ the weights, over j features f.

The result of MaxEnt is a relative occurrence rate, sometimes termed “suitability”, which can be obtained in different transformed (monotonically related) output formats. Similar to (Ortiz et al., 2013), we use a value 0.5 on the so-called logistic output format as "default" threshold for MaxEnt, because this is the default value for the internal parameter used to create the logistic output (Elith et al., 2010) – despite strong arguments in literature stating that this output format should not be carelessly treated as absolute probability of presence (Guillera-Arroita et al., 2014; Merow et al., 2013). This theoretical issue is not of interest to us here, since we do not apply any probabilistic interpretation. For further mathematical details, the interested reader is referred to the

abovementioned original literature. In this study, we relied on the R library oneClass (https://github.com/benmack/oneClass), which contains a BSVM implementation, as well as an R wrapper of MaxEnt that calls the Java source file. Both implementations internally scale the data.

Post processing by region growing

To restrict the predicted flood extent to those areas that have a spatial connection to the initial extent, we applied the ConnectedThreshold method from the Python module SimpleITK (Lowekamp et al., 2013). The procedure starts at given seed points and checks whether neighboring raster cell values fall within or outside a user-defined range. As seed, the original SAR-derived flood extents were used. If a cell is discarded, its neighbors are not considered and the propagation in that direction stops. When providing a binary raster (dry or flooded, denoted as 0 or 1) and setting the user-defined threshold to 1, then the result is simply a cut-back binary raster, on which all flood cells unconnected to the initial flood extent are reset to non-flooded.

Performance metrics

Two different types of metrics are needed for this study: training metrics based on PU data to select the best model during the parameter grid search, and validation metrics based on the PN reference to evaluate the final extrapolations. With only positive and unlabeled data, the quantities that can reliably be estimated are the True Positives (TP, prediction and observation are positive), the False Negatives (FN, prediction misses positive observation), and the model's probability of positive predictions among all predictions. From these quantities, various metrics have been proposed in literature (see e.g. Lee and Liu, 2003; Mack et al., 2014). However, most of these metrics are depending on the binarization threshold. For threshold-independent evaluation of binary classifiers, it is common to compute the area under the curve (AUC) of the receiver-operator characteristic (ROC) (Hanley and McNeil, 1982; Fawcett, 2006). The AUC indicates how well the algorithm ranks the instances. For PU data, the best obtainable AUC value is theoretically lower than 1, as some unlabeled samples should get ranked among the positive class, but Phillips et al. (Phillips et al., 2006) have claimed that the difference in AUC_{PU} is still a valid measure to compare the discriminative power of multiple models. In line with Phillips et al., we argue that AUC_{PU} is a consistent metric for model selection, as it has the same meaning for any algorithm (BSVM has a different default threshold than MaxEnt), and is adequate for any purpose. The user can later decide to put more emphasis on sensitivity or specificity during threshold selection, depending on the intended application of the model. We verified that AUC_{PU} indeed correlates with AUC_{PN}, which denotes the same metric based on PN reference data (Figure 2.2). While even high PU performance is no guarantee for high PN performance, and the very best model on test set might not be the rank #1 on training set, AUC_{PU} generally selects good models, which makes it a reasonable choice in the absence of PN test data. This behavior has been previously reported by (Mack et al., 2014), who suggest a manual inspection of several candidate models. However, as we present a method rather than a specific classification, manually inspecting several candidate models for each experimental setup was deemed unfeasible and too subjective for a methodological study. It is worth to note that we have conducted similar checks with other PU metrics in the early stage of this study, but only present AUC_{PU} here, due to the abovementioned consistency of this metric.

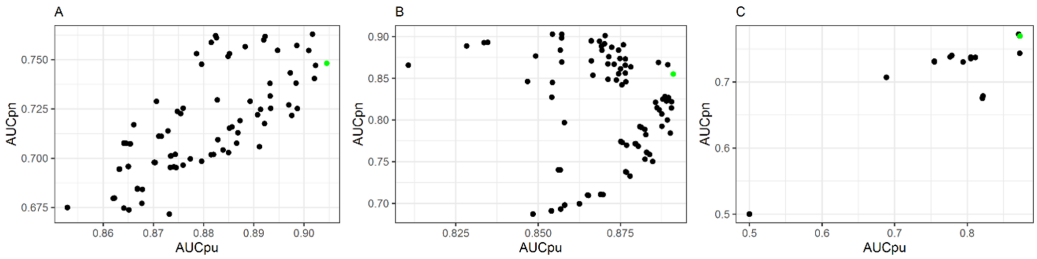

FIGURE 2.2: PU vs PN performance of all candidate models during the grid search for selected setups. Each point represents a model trained on the same data, but with different parameters. AUC_{PU} is given as the mean of a 5-fold cross validation, AUC_{PN} is a single score computed on an independent test set of the reference data on the corresponding AOI. **A** shows a BSVM trained on EMSR_229 and using USGS_SJ as test set. **B** shows a BSVM trained on DLR_BN and using NOAA_labeled as test set. **C** shows MaxEnt models trained on EMSR_229 and using USGS_SJ as test set. The green dot signalizes the selected model by the criterion of maximum AUC_{PU}, which has been the basis of model selection for this study.

Validation metrics for PN data are more standard. We measure the commonly used κ score by Cohen (Cohen, 1960) as well as the sensitivity (true positive rate, Eq. 2.3), specificity (true negative rate, Eq. 2.4), and error bias (EB, Eq. 2.5). To evaluate the initial masks, we further provide the percentages of detected open water, flooded vegetation, and flooded urban areas. The PN performance is given for the entire images, i.e. all pixels, stratified by the manually assigned land cover class.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{2.3}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{2.4}$$

$$EB = \frac{FP}{FN} \tag{2.5}$$

2.3.3 Experimental setup

The presented extrapolation procedure by OCC, as visualized in Figure 2.3, works in four steps plus validation: (1) feature engineering, by which we mean the derivation of explanatory variables (e.g. topographic indicators) from the raw data (e.g. DEM). (2) Training data sampling, (3) model learning and (4) prediction. It requires a stack of features in raster format, and an initial satellite-based flood mask. The learning step includes a parameter grid search with cross validation, where AUC_{PU} is used as metric for model selection. After a first coarse grid search, the fine tuning in each model run was restricted to the following parameter grid: BSVM: σ = {0.1, 0.5, 1, 2}, C_{-} = {0.1, 1, 5, 10, 25, 50, 250}, $C_{Multiplier}$ = {2,4,6,8}. SVM: σ = {0.1, 0.5, 1, 2, 5}, C = {0.1, 1, 5, 10, 25, 50, 250, 1000}. MaxEnt: fc = {D, LQ, LQP, H}, β = {0.001, 0.01, 0.1, 1, 10, 50, 100, 500}. The selected model is then re-trained with the full training data and applied to the entire feature stack. This results in a single raster with continuous values, which represent the raw output of the algorithms (i.e. distance to the hyperplane for BSVM, and relative probability for MaxEnt) for each raster cell. To obtain a binary prediction (flooded or not), a threshold has to be applied to this continuous prediction. Subsequent region-growing removes areas without connection to the initial mask, which makes the result appear like an inter-/extrapolation. The binary predictions, raw and postprocessed, are then validated

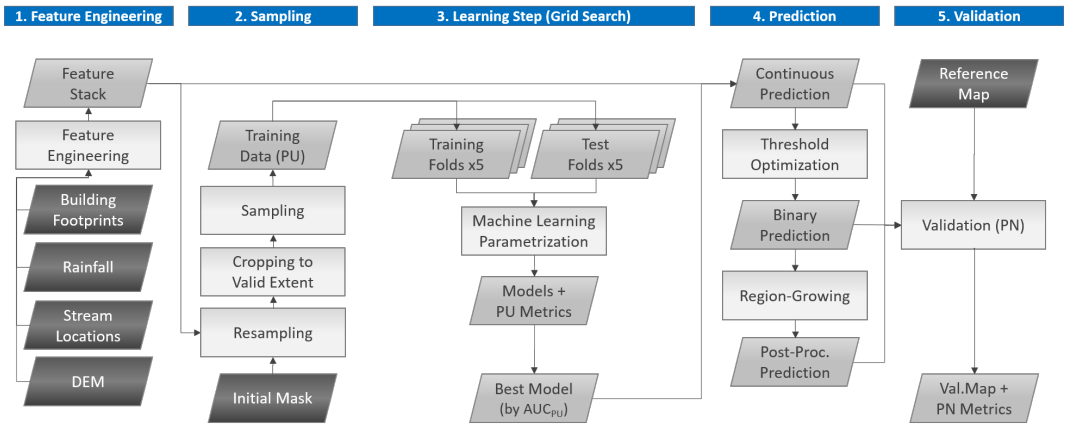

FIGURE 2.3: Flowchart of the presented procedure

by comparison to the independent reference maps NOAA_labeled and USGS_SJ. The difference between the binary predictions and corresponding binary reference results in a validation map with the 4 classes TP, FP, TN, and FN.

Samples for training the models were drawn from the valid extent of the respective initial mask, i.e. AOI1 for DLR_BN and NOAA_labeled, AOI2 for DLR_CNN, and AOI3 for the USGS_SJ benchmark. In the case of EMSR_229, which has by far the largest extent, it was tested how the sampling area during training affects the skill. Eventually we used the entire area covered by the feature stack as training area ("Full Extent") for the presented results.

OCC methods have been applied to problems with very few positive training samples, because these occur rarely or are expensive to obtain. In our case, obtaining positive samples does not constitute a problem since we can potentially use the entire flood extent as training area. The number of unlabeled samples should be high enough so that the feature space during training is representative for the feature space in the application case, i.e. more is better, limited only by concerns about computation time (Mack et al., 2014). For each PU classification problem, we randomly sampled (without replacement) 2000 positive and 8000 unlabeled pixels. The PN benchmark models were trained with 5000 positive and negative samples each. Further, we tested two sampling modes, named "regular" and "urban". In regular mode, samples were drawn entirely random. In urban mode, samples were drawn in equal parts from a distance up to 20 m, 100 m and above 100 m distance to buildings. The idea behind this urban sampling was to provide the algorithms with more of those samples which we consider to be difficult and of primary interest. DLR_CNN, like the manually labeled reference, contains distinct labels for flooded open water and flooded urban areas, so in that case for the urban mode we instead only used the urban class.

Models were further trained on 4 different feature subsets as denoted in Table 2.2, guided by the question of potential application. Both algorithms use regularization, so theoretically there is no need for manual feature selection. However, models including rainfall or distance to buildings require that additional data to be available, and might potentially learn different types of patterns. Therefore we investigated these choices separately. The subsets are: only topographic data and distance to streams (“Topo”), the aforementioned plus rainfall data (“Topo+Rain”), topographic data plus distance to buildings (“Topo+Buildings”) and all data combined (“All”).

For the sake of providing consistent numbers, two thresholds were considered for all models: the default ($\theta_{Default}$), i.e. 0 for BSVM and 0.5 for MaxEnt, which is learned from the PU training data, and the optimal threshold (θ_{Opt}) at maximum κ, which requires PN reference data. In practical application, the user would most likely inspect the continuous prediction of the best models (selected by AUC_{PU}), before deciding on the threshold. However, as we present a novel procedure here, we cannot inspect all models in detail, and want to provide the maximum obtainable skill.

2.4 Results

2.4.1 Skill of the initial masks

To evaluate whether the proposed procedure is able to improve the initial masks, we first quantified the quality of the original products by the same measures as used for the models and using the same reference data (Table 2.3). EMSR_229, despite detecting essentially no flooded vegetation or urban areas at all, still obtains a tolerable accuracy score, due to its outstanding specificity (0.999), i.e. no false positives. The higher overdetection in the San Jacinto area might also hint at errors in the USGS_SJ reference. Also DLR_BN and DLR_CNN exhibit 0.99 and 0.98 specificity, respectively, while detecting just 20-40% of the flooded vegetation and urban areas. This clearly underlines our hypothesis, that these products should be regarded as positive and unlabeled. EB consequently ranges between 0.001 for EMSR_229 to 0.13 for DLR_CNN, indicating underdetection. Note that DLR_BN only achieves an overall κ score of 0.34 (0.51 in urban areas) on our manually labeled reference, as opposed to 0.68 on the inconsistently labeled reference used in the original study by (Li, Martinis, Wieland, Schlaffer and Natsuaki, 2019). It is still a high quality product, judged by the specific skill on urban areas.

TABLE 2.3: Skill of the initial masks on the AOIs used in this study. Reference data for AOI1 and AOI2 is the manually labeled aerial image NOAA_labeled, reference for AOI3 is the USGS HWM interpolation USGS_SJ. The metrics EB, Sens., Spec., Acc., and κ are calculated over all landcover classes, while $\kappa_{veg.}$ and κ_{urban} were derived using only the flooded vegetation and flooded urban areas, respectively.

Product - AOI	%open	%veg.	%urban	EB	Sens.	Spec.	Acc.	κ	$\kappa_{veg.}$	κ_{urban}
EMSR_229 - 1/West Houston	32.06	1.16	0.43	0.001	0.06	0.999	0.63	0.07	0.01	0.01
EMSR_229 - 2/Buffalo Bayou	0	1.16	0	-	0	0	0	0	0	0
EMSR_229 - 3/San Jacinto	-	-	-	0.01	0.05	0.99	0.76	0.06	-	-
DLR_BN - 1/West Houston	69.01	19.60	41.36	0.03	0.32	0.99	0.73	0.34	0.24	0.51
DLR_BN - 2/Buffalo Bayou	3.53	6.93	23.27	0.04	0.21	0.99	0.82	0.28	0.06	0.31
DLR_CNN - 2/Buffalo Bayou	63.77	46.84	42.41	0.13	0.44	0.98	0.86	0.51	0.27	0.50

2.4.2 Skill of the extrapolation models

A full list of model setups and the threshold-independent ranking performance AUC_{PN}, as well as the training performance AUC_{PU}, can be found in the Appendix Table 2.4. The setup of our experiments (feature selection and sampling mode) apparently had only minor impact on the results. The only remarkable finding in this context is that the spatial transfer application of DLR_CNN models to the entire AOI1 gave much better results with distance to buildings included. The best EMSR_229 models are those trained on all features, and the urban sampling mode did slightly improve these models on the urban AOI2 (Buffalo Bayou) — however, the same cannot be stated for the other initial flood masks. The effect of feature selection on the benchmark models was also negligible.

We interpret this as indication that the most important features are already included in the "Topo" selection. In the following, we therefore analyze the models from different setups together, as we consider them to rather show random variation than meaningful differences. This adds a rough estimation of variance to our results and helps to visualize the effect of algorithm selection, threshold selection and postprocessing more clearly.

The κ score on validation data over all land cover classes (Figure 2.4) shows that all initial flood masks can be considerably improved by the presented approach, with differences to the best models ranging from about 0.2 (DLR_CNN) to 0.6 (EMSR_229 on AOI1). Learned models are clearly performing best in their respective area of training: the West Houston AOI for DLR_BN, and the Buffalo Bayou for DLR_CNN. In San Jacinto, the best models are those learned from EMSR_229, which is the only initial mask that is defined in all three AOIs. The skill obtained when extrapolating from the EMSR product is mediocre on the Buffalo Bayou, where no flood was initially detected, better in the San Jacinto basin, and surprisingly high in West Houston. Predictions of the other models in San Jacinto, and also the application on the entire West Houston AOI for models learned from DLR_CNN, are spatial transfer. It is unsurprising that performance is lower in these cases, and not aim of the paper to improve this spatial transfer performance. The overall skill of the best extrapolation from the EMSR_229 mask on AOI2 is similar to the original DLR_BN product, and on AOI1 even competitive with the models learned from DLR_BN and DLR_CNN — however, the improvements on AOI1 stem primarily from correct detection of flooded vegetation, while the specific skill on urban flooding is still relatively low. This can potentially be explained by the fact that AOI1 is dominated by forest, while AOI2 is almost exclusively urban area, therefore the models are optimized on different conditions. It is encouraging to see that all models learned from DLR_CNN further improve this high quality initial flood mask in urban areas. Differences in κ between the best PU and PN models account to 0.15 on AOI1, 0.16 on AOI2 and 0.38 on AOI3.

On first glance, both algorithms perform similarly well, with MaxEnt often showing larger variance, meaning it appears to be more sensitive towards setup than BSVM. One notable difference is the skill on urban areas: MaxEnt models learned from DLR_BN perform worse on urban areas than the initial mask. All MaxEnt models on AOI2 perform worse than their BSVM counterparts. At the same time, performance of MaxEnt models for flooded vegetation on AOI1 is higher. Both algorithms were trained with identical data, therefore the differences have to result from the model structure. It is reasonable to assume that topography in vegetated areas behaves differently than in urban areas. The training scores (Table 2.4) show that BSVM in general fits closer to the training data. The initial flood masks DLR_BN and DLR_CNN already cover significant areas of urban flooding, so the close fit could be one reason for the good performance on urban areas in these cases. However, the case of EMSR_229 is less clear.

A remarkable difference was observed in robustness of the optimal classification threshold (Figure 2.5). The optimal threshold value for BSVM varies considerably in our experiments. This behavior may be a drawback for use cases without reference data, and for integration into automatic processing chains. MaxEnt is slightly less affected by this problem. Keep in mind, though, that the continuous output of both algorithms has a different meaning and scale (unbounded distance to the hyperplane for BSVM, and probability between 0 and 1 for MaxEnt). The average loss of skill for the PU models is below 0.1, but in individual cases considerably higher. The suitability of the default threshold may dependent on the representativeness of the training samples: For the reference models, training and application data were drawn from the same underlying

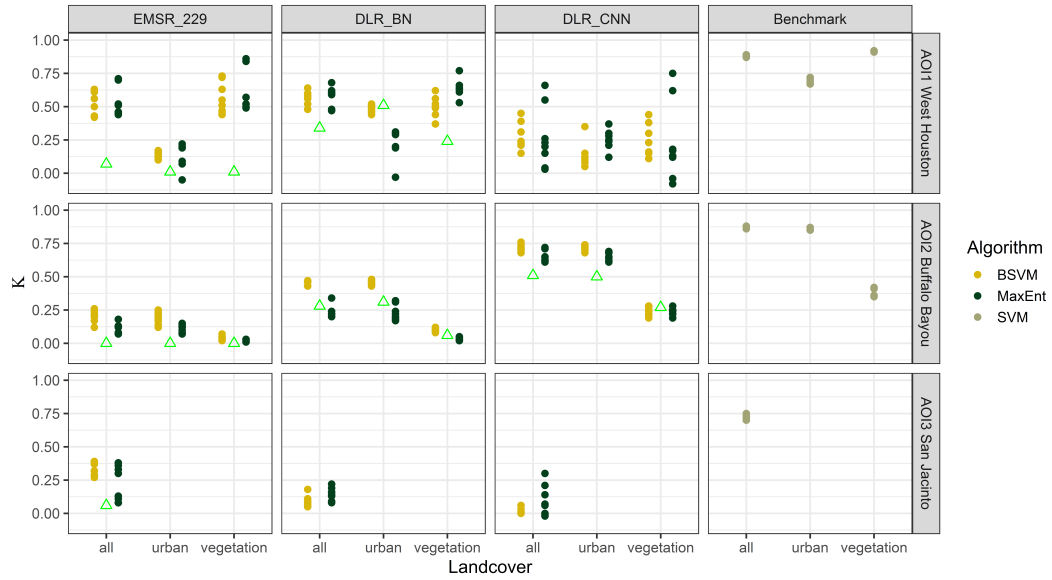

FIGURE 2.4: κ score on validation data at θ_{Opt} without postprocessing. The green triangle denotes the skill of the original product (initial mask) if the product exists on that AOI. Each point represents a model with different setup. BSVM and MaxEnt have been trained with identical data.

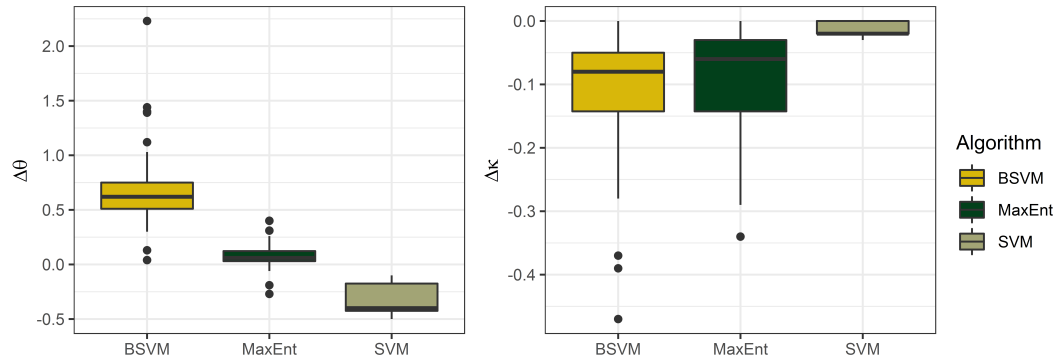

FIGURE 2.5: Difference of default and optimal threshold. $\Delta\theta$ denotes the difference in the threshold value and $\Delta\kappa$ the respective difference in skill.

distribution, and in that case $\theta_{Default}$ and θ_{Opt} are closer, with the skill being almost identical ($\Delta\kappa$ below 0.025).

Classification on pixel level may lead to noisy results and in some cases detect possible flood in areas that were not affected by the event in question. Postprocessing, as expected, increased the specificity in tradeoff for sensitivity, but overall κ was raised as well (Figure 2.6). Beyond the intended effect, we also observed significantly reduced noise from the initial mask, because random errors are unlikely to occur in the same spot twice (meaning the satellite image classification and the classification from topography as presented in this paper), so that these areas are removed. Specificity of the best EMSR_229-derived extrapolations is again close to 1 after the postprocessing, meaning that the derived flood extent is reliable. Obviously, the region-growing, which checks for connectivity with the initial flood extent, only makes sense for those areas where the initial mask is defined, not for spatial transfer (DLR_BN and DLR_CNN to San Jacinto, DLR_CNN to aerial).

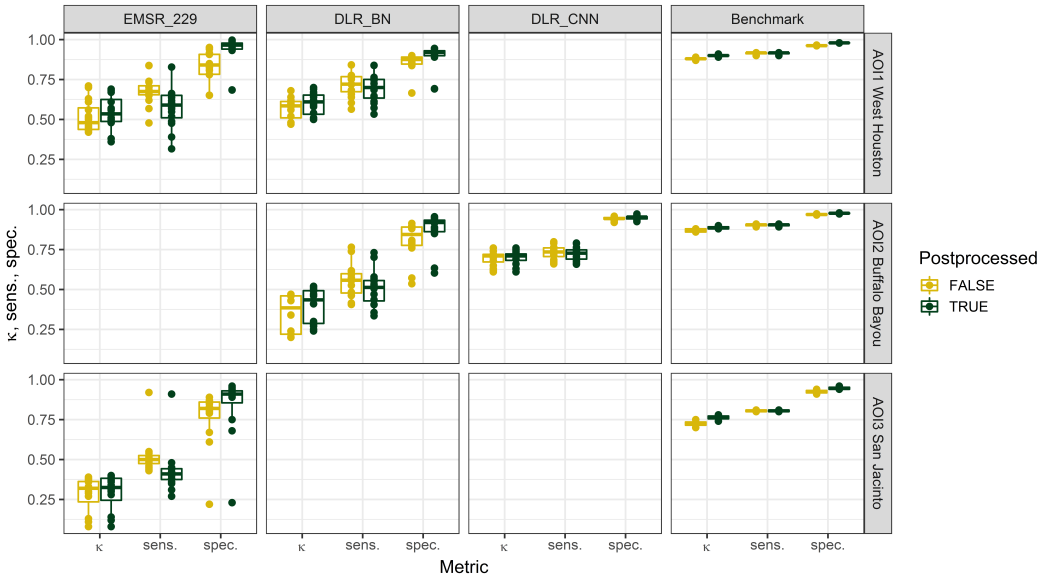

FIGURE 2.6: Overall effect of postprocessing on κ, sensitivity and specificity at θ_{Opt}. The range of the boxplots includes both BSVM and MaxEnt models to visualize the general trend. Empty boxes indicate that postprocessing is not possible because the initial mask does not exist on that extent. Note that EMSR_229 is theoretically defined on AOI2, but there was no flood detected in that area, therefore the region-growing would remove all predictions there.

2.4.3 Spatial comparison of predicted flood extents

The large-scale comparison (Figure 2.7) visualizes the general behavior of an OCC model learned from the EMSR initial flood mask. The initial mask used for training (green) is primarily located outside the test areas. Some disagreement between the training mask and the validation mask is visible, especially in the west. The overestimation (yellow area) is explainable given the training data, which are learned as true extent. Note that the NOAA_labeled reference has been created by us, and we are accordingly confident about the quality, while the USGS_SJ mapping on the other hand is based on interpolated high water marks, and could contain errors which we cannot further evaluate. Note also that the underestimation visible on the map (red) stems to large parts from the postprocessing, which removes predicted flood without spatial connection to the initial mask. This is especially obvious for the channel of the Buffalo Bayou, which is completely missing on the postprocessed version. The continuous prediction outside the validation areas shows that the model has indeed learned quite smooth and understandable patterns along the rivers. Also it is obvious that the models correctly learned to exclude the permanent river channels.

An example model trained on DLR_BN (Figure 2.8) exhibits large fractions of correctly identified flooded vegetation and coarse coverage of the affected urban areas. Visually disturbing is the buffer around the channels that has been classified as non-flood, which is also seen on the initial mask. This is probably an artifact from training on flood masks instead of water masks. The area around these streams is covered by dense forest. Note that the previously shown EMSR_229 model performed better along these channels, presumably because it could learn the relationships of flooding along other streams, which are less obstructed by vegetation. The DLR_BN model performs much better on urban areas, though. There is underestimation visible along the Buffalo Bayou settlements, yet the affected urban areas in the north and south-west are captured quite well. These areas are colored mainly in yellow (overdetection) because the model did

not restrict the predictions to the streets, which are visible as fine blue patterns, but the affected area seems reasonable. The overestimation along the western channel is not removed during postprocessing due to spatial connection with unluckily distributed noise on the initial mask. Even with the highest quality initial mask, DLR_CNN (Figure 2.9), water in the streets remains mostly undetected. Still, the extrapolation outside the training area, visible in dark colors, appears smooth and connected. Noise from the initial mask has been entirely eliminated. The land-water boundary appears quite sharp.

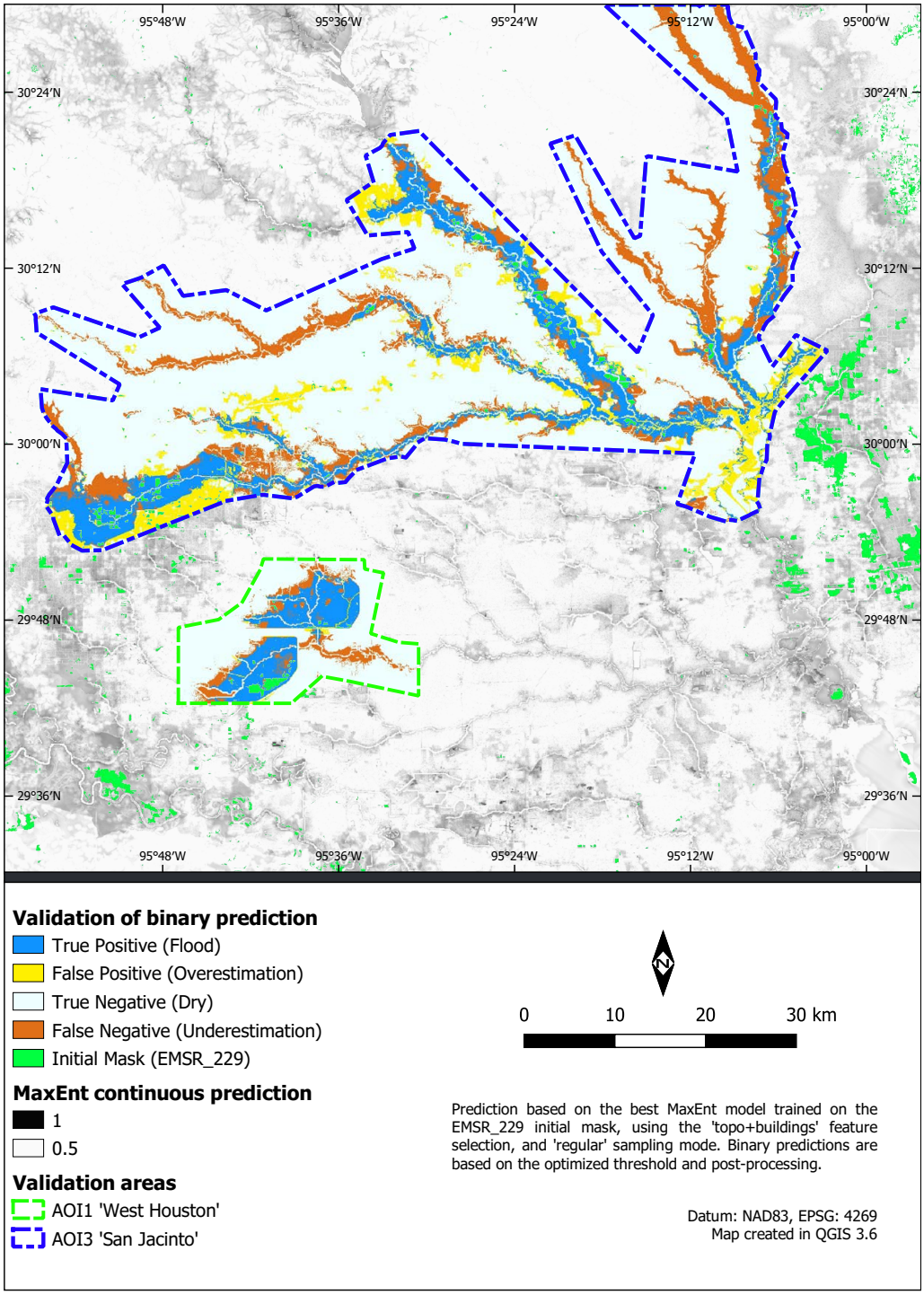

FIGURE 2.7: Example of the spatial prediction of a MaxEnt model learned from EMSR_229

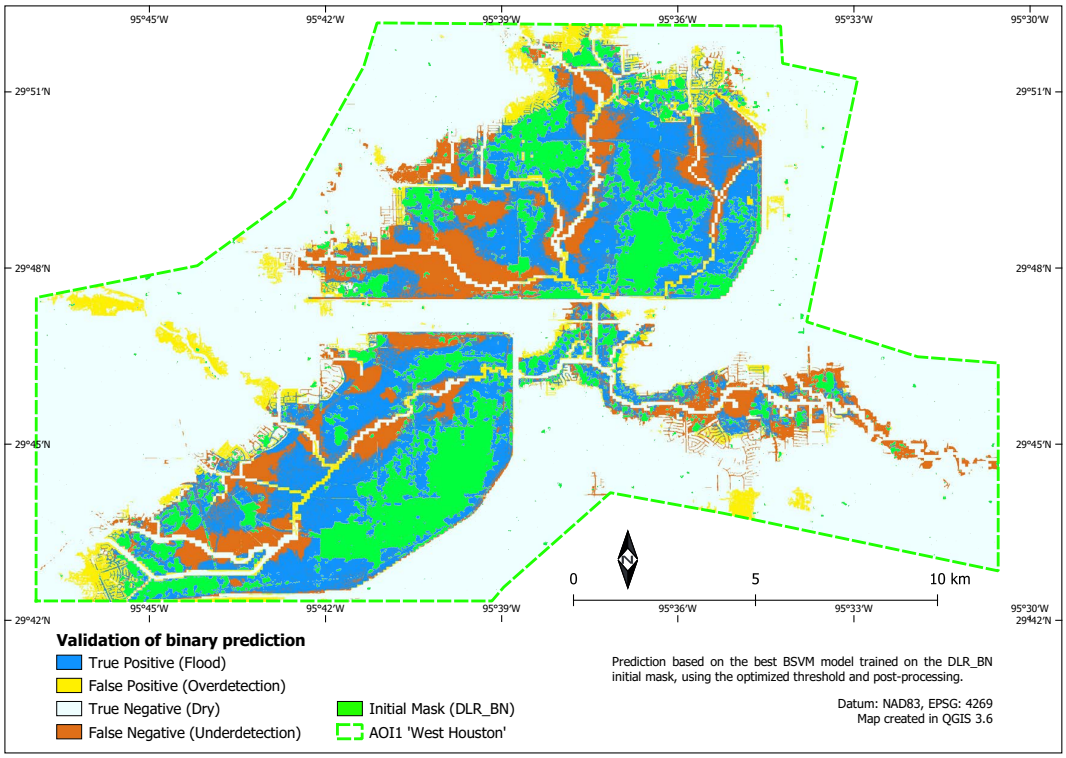

FIGURE 2.8: Example of the spatial prediction of a BSVM model learned from DLR_BN.

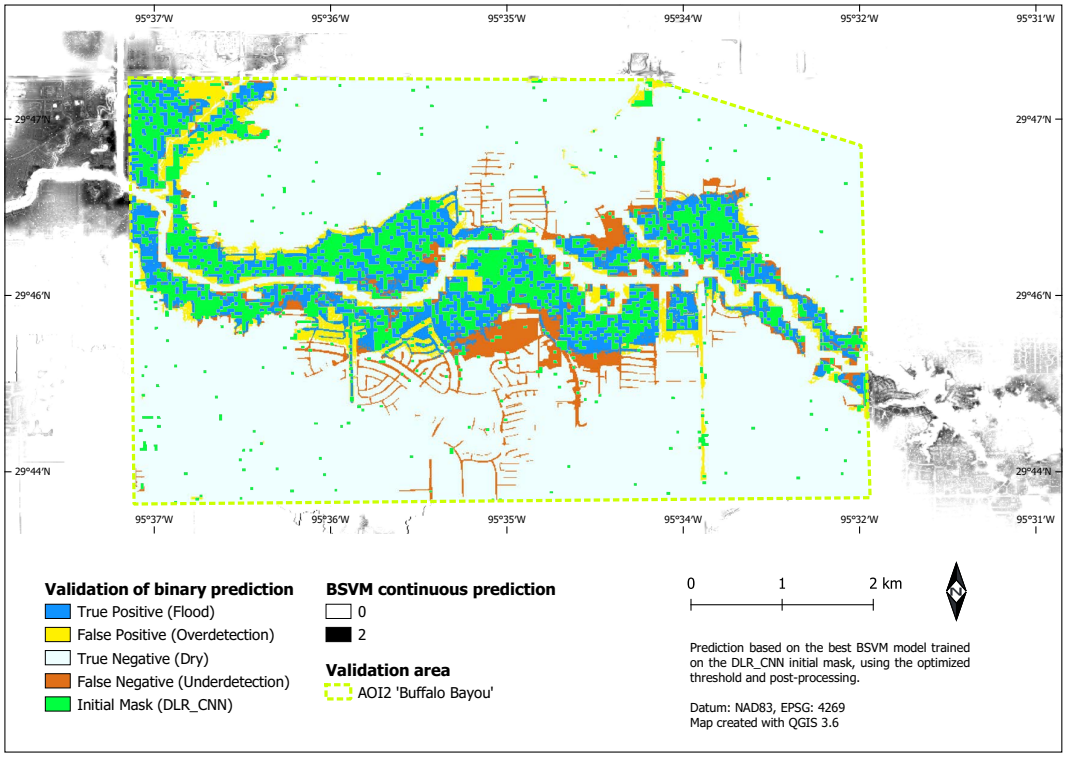

FIGURE 2.9: Example of the spatial prediction of a BSVM model learned from DLR_CNN. Water in the streets is not detected, although some fine patters are visible in the continuous prediction.

2.5 Discussion

2.5.1 Aim and overall success

The assessed satellite-based flood masks exhibit very low (EMSR_229) to moderate (DLR_BN, DLR_CNN) detection skill in vegetated and urban areas. This is to be expected, due to the various effects which constrain the information content of SAR images in these cases, i.e. volume scattering, layover, oblique viewing geometries, and others. The specificity of all these products is very high, though, meaning that those areas, which are identified as flooded, indeed represent true flood. We therefore propose to treat such products as PU data. Our study demonstrates how these satellite-based flood masks can then be improved in vegetated and urban areas by an OCC procedure. A critical point for such studies is the reliability of the reference data. We present a performance evaluation on a precisely labeled aerial image, which is of higher quality than what is frequently used in other studies (e.g. Giustarini et al., 2013; Matgen et al., 2011). For the larger scale, we use the extent in the San Jacinto river as published by the USGS, which is based on interpolated HWM and has been used as reference by (Wing et al., 2019).

According to the performance metrics, all initial flood masks can be considerably improved by the presented procedure. For the EMSR product, κ over all classes rose from 0.06 to 0.76 in the best case with postprocessing, and from 0.00 to 0.25 in urban areas. The high quality initial masks, DLR_BN and DLR_CNN, have also been successfully enhanced up to about 0.2 points. Although the raw classification may at first lead to some overestimation far off the initial mask, the postprocessing improved the specificity as well as the visually perceived quality of the results, by suppressing uncorrelated errors of the initial SAR classification and our classification from topographic data. In a pluvial event, the formation of disconnected puddles is possible. The region-growing may delete such correctly predicted puddles from the classification. However, if the initial satellite mask contains a single pixel of that puddle, the area is kept. Whether or not to apply this postprocessing is therefore also a question of the quality of the initial mask. Overestimation after the postprocessing occurs mainly in places where the reference mask disagrees with the input mask, meaning either false positives in the input or false negatives in the reference data. For USGS_SJ, some uncertainty is to be expected. For NOAA_labeled, minor differences might be induced by the different acquisition times of the aerial and satellite images. The models were explicitly trained on flood masks. For many applications it might be more suitable to generate water masks, which include the permanent water. This should also solve the visible underdetection along the streams in the DLR_BN models. We refer to extrapolation as growing areas of flood detected on the initial datasets. Spatial transfer (e.g. DLR_BN to USGS_SJ) did not work well. A local approach is necessary, because event characteristics differ spatially. Although some extrapolation outside the extent of the initial mask is possible, the predictions far off the original extent, and especially in different river basins, are therefore deemed unreliable (note the difference here is between undetected flood on the original extent of the satellite image, and areas outside the satellite image). Whether the area in between two or more satellite images could be modelled by this approach has not been investigated, but could be an interesting question to try in future studies.

2.5.2 Features and algorithms

Our analysis was based on features that have commonly been suggested in literature for the purpose of flood susceptibility mapping (Samela et al., 2017; Chapi et al., 2017;

Tehrany et al., 2018; Jalayer et al., 2014; Kelleher and McPhillips, 2019), like HAND, TWI, distance to streams, and descriptors of the local topographic situation. The skill of the PN benchmark models suggests that these features are indeed useful, also in urban and vegetated areas, given a representative training set. In addition we tested whether rainfall data and distance to buildings help to improve the models. The rainfall sum for hurricane Harvey had very little spatial variance over the Houston area, therefore it is rather unsurprising that it does not lead to an improvement here. We hesitate to draw a general conclusion from this result, as the effect may be different for an event with more heterogeneous rainfall distribution. The results of our investigation did neither show clear improvements from using distance to buildings as a feature, nor from drawing more training samples from urban areas, when learning from the DLR_BN initial mask. However, the skill of the models learned from EMSR_229 did improve slightly, and the transfer skill of models learned from DLR_CNN to AOI1 did improve strongly, when including the distance to buildings. The sampling further seems to have at least a small effect on the skill on urban areas for the EMSR_229 models. A possible explanation is that DLR_BN already covers significant parts of urban and non-urban areas alike, while DLR_CNN covers primarily urban flood, and EMSR_229 almost exclusively non-urban flood. Therefore, the distance to buildings might be more useful in these models to describe how feature distributions of flooded areas differ closer to and further away from the city, respectively. Since the results do not indicate any negative effect of including the distance to buildings, we suggest to include it when available. To further improve the feature engineering, automating this step via deep learning might be an idea worth investigating in future studies. Especially local context features, as generated by a CNN, have been successful in improving various land cover classifications, including detection of water (Yu et al., 2017; Wu et al., 2019).

Both tested OCC algorithms, BSVM and MaxEnt, performed similarly well in the overall statistics. BSVM exhibits a closer fit to the training data, and is less affected by feature selection and sampling. Ng & Jordan (Ng and Jordan, 2002) state that discriminative algorithms often perform better than generative algorithms for complex classification problems. This might partially explain why BSVM in most cases performs better than MaxEnt in detecting urban flooding. However, explaining this finding remains speculative to a certain extent. The best models on AOI1 and AOI2 also came close to the PN benchmark, but there is still a significant margin which indicates potential for improvement. While we assume that our positive training labels are mostly correct, there will for sure be some violation of this assumption. BSVM can theoretically handle this problem to a certain extent, because outliers in the positive training samples will be classified as negative if they are so far in the "negative realm" that the biased penalty term is overruled. MaxEnt assumes positive samples to be clean from errors (Elith et al., 2010), so a preprocessing of the initial masks might be an option to consider. Instead of performing classification in one step, it is also possible to iteratively single out the reliable negatives (Mack and Waske, 2016). As the amount of available training samples in our task is relatively high, we did not implement such an iterative refinement, but rather relied of the effectiveness of data. The effect of training label distribution is debated and difficult to estimate without doing systematic tests for each dataset, as the naturally occurring class distribution – even in cases where there is such a distribution – is often not the most appropriate (Weiss and Provost, 2001). Besides this, also other PU algorithms are available in literature, (e.g. Li et al., 2011; Ruff et al., 2018). If a validity layer for the initial flood mask is available, an alternative approach would be to train any regular PN classifier on the valid areas. Hydrodynamic simulations are able to model flooding in vegetated and urban areas as well, e.g. Wing et al. (Wing et al., 2019) for the

event in question. A drawback of our presented approach in comparison to a physical model is that the machine learning models do not account for hydrodynamic effects, or in general a closed water balance (no more water predicted than available). However, we argue that a hydrodynamic simulation could make use of the improved flood masks from our approach via data assimilation.

2.5.3 Threshold selection

As this paper presents a novel approach, rather than a particular classification, we provide the threshold-independent score AUC_{PN}, further performance metrics at θ_{Opt}, and the loss $\Delta\kappa$ when resorting to $\theta_{Default}$. We are fully aware that optimal threshold selection in the absence of PN reference data is tricky. By which metric the user optimizes the threshold selection will depend on the application case, i.e. how much sensitivity or specificity is required. Maximum κ may not be the desired quantity. Mack et al. (Mack and Waske, 2016) further suggested a manual approach (i.e. not automated) to derive a maximum a-posteriori threshold from a Gaussian mixture model analysis of the posterior density of the continuous prediction. However, that procedure is based on the assumptions that the posterior can be described by a combination of Gaussians, and that the component with the highest mean value is equal to the positive class, while all other components belong to the negative class. Another assumption in their approach is that the classes do not overlap at a specified point used to estimate the prior probabilities. These assumptions are certainly violated for some of our models, and this approach is not feasible in the context of this paper, as we compare many models to get an idea of the upper bound of performance of our procedure. MaxEnt also provides a different form of output, called the cumulative format, which allows setting a threshold based on the accepted omission rate (Phillips and Dudík, 2008). Depending on the application, this may be a more desirable way of threshold selection. In cases where the training data is representative, the most straightforward approach is to use the default threshold, or to optimize a PU performance metric of choice on the training data. For the benchmark models, training and application data were drawn from the same underlying distribution, and in that case the skill at $\theta_{Default}$ and θ_{Opt} is almost identical. This proves that the procedure is in principle able to obtain very good results, given a representative training set. In the application using satellite-based flood masks, a bias in the feature distributions of the positive training samples is to be expected, as we know that the areas detected from satellite imagery are not entirely representative of the true flood extent. Elith et al. (Elith et al., 2010) claim that PU models are even stronger affected by sample bias than PN models, because sample bias affects both positive and negative records in the PN case, but only the positive samples in the PU case. In our case, this "sample bias" corresponds to the representativeness of the initial flood mask. This leads us to assume that including additional positive class examples from within the urban area could make the positive training data more representative, and thereby improve the performance of the PU models at $\theta_{Default}$. Such data could potentially be taken from sources such as social media content or street camera footage, which is only punctually available but provides data from within the city center.

2.6 Conclusion

We presented an extrapolation technique for satellite-based flood masks to unobservable areas, by using OCC algorithms. Especially vegetated and urban areas still pose a challenge to currently available remote sensing products, the latter of which are of major

importance for impact estimation. The quality of the initial EMSR_229 mask was found to be poor, detecting almost exclusively open water. Although it does exhibit very high specificity, a map with extreme specificity but very low sensitivity is trivial (only few easy-to-find spots detected) and of limited practical value. As long as the spatial validity of satellite-based flood masks is not clearly communicated, e.g. by a separate validity layer, we suggest treating them as positive and unlabeled in this context. OCC is then the adequate tool, avoiding to explicitly train unobservable areas as "non-flooded". Using supervised machine learning for extrapolation is straightforward once using an OCC, as the necessary positive labels for training are readily available from the initial mask. Our procedure allows for predicting a continuous score of how likely flood is to be expected per pixel, given the original mapping and the used features. A threshold can then be applied to derive a binary classification, and a subsequent region-growing raises the specificity of the extrapolation. From the user's perspective, the presented method is relatively simple to use, as the entire initial mask can be processed without the need to exclude any areas from sampling. The most important features can already be generated from a DEM and stream locations (which can also be derived from a DEM if necessary). Distance to building footprints and gridded rainfall data did not consistently improve the results, although positive effects were observed for some models.

We conclude that all three of the tested satellite-based products have been improved to a certain extent. The absolute quality of the extrapolation, as well as the suitability of the default threshold in application, hinges on the representativeness of the initial mask. The features used in this study are not sufficient for a full separation of flooded and dry locations, but a model trained on representative training data still achieves high performance (AUC_{PN} 0.91-0.98 in the benchmark case, 0.94 for the best PU model). The method in its current form may be useful for statistical applications on a scale where satellite imagery is utilized. It is not yet fit for analysis of individual streets, although the results with high quality input seem promising. Potential application of the presented method is not limited to masks from SAR data — it could also be used to fill holes from clouds in masks from optical data, or tested for social media derived extents. In particular, we see potential for future studies in the fusion of satellite-based flood masks with spottily mapped flood locations within a city center, e.g. by social media or street camera footage. Such a fused dataset is expected to provide more representative coverage in feature space, which should lead to a more reliable default threshold. The presented approach could be tested in this direction with the aim of deriving more reliable flood extents in vegetated and urban areas.

Data Availability: Most data used in this study is publicly available and has been cited in the manuscript. This includes the NED DEM (https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map, accessed on 21 July 2020), NWS rainfall data (https://water.weather.gov/precip/download.php, accessed on 25 August 2020), Microsoft USBuilding- Footprints (https://github.com/microsoft/USBuildingFootprints, accessed on 26 August 2020), OSM water (http://hydro.iis.u-tokyo.ac.jp/~yamadai/OSM_water/, accessed on 20 July 2020), the flood masks USGS_SJ (https://www.sciencebase.gov/catalog/item/5aa023ebe4b0b1c392e6881b, accessed on 11 September 2020) and EMSR_229 (https://emergency.copernicus.eu/mapping/ems-productcomponent/EMSR229_01HOUSTON_01DELINEATION_MONIT02/1, accessed on 11 September 2020), as well as the raw aerial image by NOAA (https://storms.ngs.noaa.gov/storms/harvey/index.html#8/29.822/-94.823, accessed on 3 December 2020). The manually labeled reference data is available as online supplement to this publication. The flood masks DLR_BN and DLR_CNN have kindly been made available to us by Yu Li.

Acknowledgments: The authors wish to thank Yu Li for providing the high quality flood masks, all abovementioned providers of publicly available datasets, as well as OliverWing of FATHOM and the company JBA for access to hydraulic simulations that were used for comparison in the early stage of this study.

Funding: The research presented in this paper was conducted in the framework of the project "Multirisk analysis and information system components for the Andes region (RIESGOS)" funded by the German Ministry of Education and Research (BMBF); (contract no: 03G0876B).

2.7 Appendix

TABLE 2.4: All model setups and threshold-independent ranking skill. Setup IDs 1-8 have been excluded for the plots in the main publication, to ensure the same number of points for all initial flood masks.

Setup ID	Algorithm	Flood Mask	Training Extent	Sampling	Features	AUC_{PU}	AUC_{PN}-AOI1	AUC_{PN}-AOI2	AUC_{PN}-AOI3
1	BSVM	EMSR_229	AOI1	regular	Topo	0.977	0.59	0.47	0.54
1	MaxEnt	EMSR_229	AOI1	regular	Topo	0.913	0.79	0.63	0.49
2	BSVM	EMSR_229	AOI1	regular	Topo+Rain	0.984	0.78	0.62	0.52
2	MaxEnt	EMSR_229	AOI1	regular	Topo+Rain	0.942	0.78	0.62	0.56
3	BSVM	EMSR_229	AOI1	regular	Topo+Buildings	0.981	0.75	0.54	0.58
3	MaxEnt	EMSR_229	AOI1	regular	Topo+Buildings	0.928	0.74	0.67	0.54
4	BSVM	EMSR_229	AOI1	regular	All	0.988	0.75	0.58	0.49
4	MaxEnt	EMSR_229	AOI1	regular	All	0.946	0.73	0.64	0.58
5	BSVM	EMSR_229	AOI3	regular	Topo	0.896	0.76	0.68	0.77
5	MaxEnt	EMSR_229	AOI3	regular	Topo	0.848	0.79	0.6	0.75
6	BSVM	EMSR_229	AOI3	regular	Topo+Rain	0.911	0.75	0.69	0.76
6	MaxEnt	EMSR_229	AOI3	regular	Topo+Rain	0.844	0.79	0.58	0.75
7	BSVM	EMSR_229	AOI3	regular	Topo+Buildings	0.917	0.86	0.73	0.77
7	MaxEnt	EMSR_229	AOI3	regular	Topo+Buildings	0.872	0.89	0.76	0.78
8	BSVM	EMSR_229	AOI3	regular	All	0.927	0.84	0.69	0.76
8	MaxEnt	EMSR_229	AOI3	regular	All	0.866	0.85	0.61	0.76
9	BSVM	EMSR_229	Full Extent	regular	Topo	0.861	0.76	0.65	0.7
9	MaxEnt	EMSR_229	Full Extent	regular	Topo	0.82	0.84	0.61	0.71
10	BSVM	EMSR_229	Full Extent	regular	Topo+Rain	0.881	0.82	0.59	0.75
10	MaxEnt	EMSR_229	Full Extent	regular	Topo+Rain	0.834	0.83	0.61	0.72
11	BSVM	EMSR_229	Full Extent	regular	Topo+Buildings	0.89	0.84	0.64	0.74
11	MaxEnt	EMSR_229	Full Extent	regular	Topo+Buildings	0.86	0.89	0.59	0.76
12	BSVM	EMSR_229	Full Extent	regular	All	0.905	0.84	0.62	0.76
12	MaxEnt	EMSR_229	Full Extent	regular	All	0.87	0.89	0.6	0.76
13	BSVM	EMSR_229	Full Extent	urban	Topo	0.85	0.77	0.67	0.7
13	MaxEnt	EMSR_229	Full Extent	urban	Topo	0.645	0.72	0.56	0.58
14	BSVM	EMSR_229	Full Extent	urban	Topo+Rain	0.88	0.8	0.68	0.72
14	MaxEnt	EMSR_229	Full Extent	urban	Topo+Rain	0.664	0.74	0.58	0.59
15	BSVM	EMSR_229	Full Extent	urban	Topo+Buildings	0.86	0.75	0.68	0.67
15	MaxEnt	EMSR_229	Full Extent	urban	Topo+Buildings	0.595	0.82	0.64	0.66
16	BSVM	EMSR_229	Full Extent	urban	All	0.889	0.78	0.71	0.72
16	MaxEnt	EMSR_229	Full Extent	urban	All	0.599	0.82	0.64	0.66
17	BSVM	DLR_BN	AOI1	regular	Topo	0.891	0.86	0.81	0.6
17	MaxEnt	DLR_BN	AOI1	regular	Topo	0.804	0.87	0.65	0.6
18	BSVM	DLR_BN	AOI1	regular	Topo+Rain	0.904	0.83	0.81	0.54
18	MaxEnt	DLR_BN	AOI1	regular	Topo+Rain	0.807	0.87	0.65	0.6
19	BSVM	DLR_BN	AOI1	regular	Topo+Buildings	0.905	0.82	0.82	0.59
19	MaxEnt	DLR_BN	AOI1	regular	Topo+Buildings	0.809	0.86	0.64	0.61
20	BSVM	DLR_BN	AOI1	regular	All	0.914	0.8	0.81	0.53
20	MaxEnt	DLR_BN	AOI1	regular	All	0.812	0.86	0.64	0.61
21	BSVM	DLR_BN	AOI1	urban	Topo	0.866	0.84	0.83	0.56
21	MaxEnt	DLR_BN	AOI1	urban	Topo	0.672	0.85	0.73	0.67
22	BSVM	DLR_BN	AOI1	urban	Topo+Rain	0.882	0.81	0.82	0.52
22	MaxEnt	DLR_BN	AOI1	urban	Topo+Rain	0.672	0.85	0.73	0.66
23	BSVM	DLR_BN	AOI1	urban	Topo+Buildings	0.886	0.77	0.83	0.54
23	MaxEnt	DLR_BN	AOI1	urban	Topo+Buildings	0.538	0.87	0.64	0.69
24	BSVM	DLR_BN	AOI1	urban	All	0.895	0.77	0.82	0.51
24	MaxEnt	DLR_BN	AOI1	urban	All	0.539	0.87	0.64	0.68
25	BSVM	DLR_CNN	AOI2	regular	Topo	0.902	0.63	0.9	0.49
25	MaxEnt	DLR_CNN	AOI2	regular	Topo	0.876	0.65	0.94	0.63
26	BSVM	DLR_CNN	AOI2	regular	Topo+Rain	0.904	0.62	0.91	0.47
26	MaxEnt	DLR_CNN	AOI2	regular	Topo+Rain	0.879	0.59	0.93	0.57
27	BSVM	DLR_CNN	AOI2	regular	Topo+Buildings	0.904	0.81	0.91	0.5
27	MaxEnt	DLR_CNN	AOI2	regular	Topo+Buildings	0.876	0.88	0.94	0.72
28	BSVM	DLR_CNN	AOI2	regular	All	0.906	0.62	0.94	0.55
28	MaxEnt	DLR_CNN	AOI2	regular	All	0.88	0.84	0.93	0.64
29	BSVM	DLR_CNN	AOI2	urban	Topo	0.918	0.69	0.91	0.5
29	MaxEnt	DLR_CNN	AOI2	urban	Topo	0.889	0.67	0.92	0.52
30	BSVM	DLR_CNN	AOI2	urban	Topo+Rain	0.923	0.58	0.91	0.53
30	MaxEnt	DLR_CNN	AOI2	urban	Topo+Rain	0.901	0.68	0.93	0.56
31	BSVM	DLR_CNN	AOI2	urban	Topo+Buildings	0.926	0.49	0.88	0.5
31	MaxEnt	DLR_CNN	AOI2	urban	Topo+Buildings	0.9	0.72	0.9	0.6
32	BSVM	DLR_CNN	AOI2	urban	All	0.931	0.58	0.91	0.44
32	MaxEnt	DLR_CNN	AOI2	urban	All	0.909	0.71	0.91	0.63
33	SVM	NOAA_labeled	AOI1	regular	Topo	0.966	0.97	0.92	0.63
34	SVM	NOAA_labeled	AOI1	regular	Topo+Rain	0.971	0.97	0.93	0.6
35	SVM	NOAA_labeled	AOI1	regular	Topo+Buildings	0.970	0.97	0.93	0.64
36	SVM	NOAA_labeled	AOI1	regular	All	0.973	0.98	0.94	0.62
37	SVM	NOAA_labeled	AOI2	regular	Topo	0.976	0.57	0.98	0.54
38	SVM	NOAA_labeled	AOI2	regular	Topo+Rain	0.980	0.57	0.98	0.49
39	SVM	NOAA_labeled	AOI2	regular	Topo+Buildings	0.978	0.59	0.98	0.47
40	SVM	NOAA_labeled	AOI2	regular	All	0.982	0.62	0.98	0.43
41	SVM	USGS_SJ	AOI3	regular	Topo	0.912	0.84	0.66	0.92
42	SVM	USGS_SJ	AOI3	regular	Topo+Rain	0.925	0.85	0.68	0.94
43	SVM	USGS_SJ	AOI3	regular	Topo+Buildings	0.923	0.86	0.67	0.93
44	SVM	USGS_SJ	AOI3	regular	All	0.934	0.85	0.7	0.94

Chapter 3

On the role of floodplain storage and hydrodynamic interactions for flood risk estimation

3.1 Abstract

Hydrodynamic interactions, i.e., the floodplain storage effects due to inundations upstream on flood wave propagation, inundation areas, and flood damage downstream, are important but often ignored in large-scale flood risk assessments. Though new methods considering these effects emerge, they are often limited to small or mesoscale. In this study, we investigate the role of hydrodynamic interactions and floodplain storage on flood hazard and risk in the German part of the Rhine basin. To do so, we compare a new continuous 1D routing scheme within a flood risk model chain to the piece-wise routing scheme, which largely neglects floodplain storage. The results show that floodplain storage is significant and lowers water levels, discharges, and reduces risks by over 50%. Therefore, for accurate risk assessments, a system approach must be adopted, and floodplain storage and hydrodynamic interactions must carefully be considered.

3.2 Introduction

Large-scale flood risk assessments are needed for national policy developments, disaster management planning, and the insurance industry (de Moel et al., 2015). However, studies presenting large-scale flood risk assessments are still rare and suffer from data and methodological limitations regarding the realistic representation of hydrodynamic interactions (Vorogushyn et al., 2018).

With hydrodynamic interactions, we understand changes in discharge, water levels, and, consequently, hazard and risk downstream the river network due to inundation and water storage in the upstream reaches. Hence, the term includes lateral interactions between the river channel and the floodplains and longitudinal interactions between upstream and downstream locations. Concrete examples are the discharge reducing effects of flooding due to the overflow of embankments or breaches and detention areas. Such interactions are complex and depend on several factors, including floodplain topography, the presence of dikes, and their failures. Floodplain topography determines the amount of stored water and thus flood wave attenuation. Dike failures result in flooding of protected floodplains which, in turn, affects flood hazards of areas downstream.

A common approach for large-scale flood mapping and risk assessment is based on the estimation of peak discharge values associated with certain return periods at gauge locations, which are subsequently used as boundary conditions for hydrodynamic simulations. The results of these local simulations are mosaicked into a large-scale picture ('mosaicked approach'). Some approaches used uniform return periods to estimate peak flows for entire countries (Bradbrook et al., 2005; Merz et al., 2008) or even continents (Alfieri et al., 2013). The assumption of a uniform return period, if used for risk assessments of larger areas, contravenes the basic spatial dependence structure of floods and leads to an overestimation of flood damage for high return periods (Nguyen et al., 2020; Metin et al., 2020; Vorogushyn et al., 2018). The expected annual damage is, however, not necessarily affected by the homogeneous return period assumption (Metin et al., 2020). Recently, several studies used multivariate extreme value models for estimating peak discharges (Keef et al., 2009; Lamb et al., 2010; Ward et al., 2013; Jongman et al., 2014; Wyncoll and Gouldby, 2015; Quinn et al., 2019; Winter et al., 2019). These approaches generate patterns of spatially dependent peak flows at multiple gauges with heterogeneous return periods.

To use peak flows (with homogenous or heterogeneous return periods) for unsteady hydrodynamic simulations, assumptions about the flow hydrographs are needed, e.g., typical hydrograph shapes scaled to the predefined peak magnitude. Hydrodynamic simulations are then carried out for individual reaches, i.e., piece-wise, whereas a new boundary condition is assigned at the next downstream gauge (e.g. Alfieri et al., 2016; Falter et al., 2016; Quinn et al., 2019). This approach is valid only for a single river reach. In a larger river network, it results in an inconsistent, not mass conservative sets of flood hydrographs, inundation areas and risk estimates (Curran et al., 2019). Reach-wise or piece-wise routing is still common for many large-scale risk assessment studies (Alfieri et al., 2016; Quinn et al., 2019).

Flood hazard and risk assessment approaches considering hydrodynamic interactions have emerged in recent years and deploy continuous hydrodynamic simulations along a single river reach or in a larger river network. These approaches account for floodplain storage effects caused by inundation considering or disregarding dike failures. Apel et al. (2004) and Vorogushyn et al. (2010, 2012) indicated the effect of hydrodynamic interactions on flood hazard and risk at the Lower Rhine and a small reach of the Elbe River, respectively, due to dike failures. Apel et al. (2009), for instance, showed that the design flood at the German-Dutch border is significantly reduced when dike breaches are stochastically considered, compared to the gauge-based extreme value statistics, which does not fully account for hydrodynamic interactions. van Mierlo et al. (2007) indicated the effect of hydrodynamic interactions for a dike ring in the Rhine-Meuse delta. de Bruijn et al. (2014) compared scenarios with and without hydrodynamic interactions for the entire Rhine-Meuse delta in terms of the number of simulated dike breaches and resulting estimates of flood fatalities. They showed that the estimated annual probability of life loss is more than doubled in the scenario without considering dike failures. Curran et al. (2019) continued their work and further improved the method by better representing floodplain flows. Recently, Ciullo et al. (2019) demonstrated how hydrodynamic interactions resulting from dike breaches in the IJssel River in the Netherlands affect the optimal dike height in a cost-benefit framework. They concluded that disregarding the dike failures mutual dependence results in suboptimal design height and increased overall costs. Dupuits et al. (2019) came to the same conclusion using a similar setting applied for a case study in a small region in the Netherlands. These approaches, however, have used one or very few locations to define the boundary conditions for hydrodynamic models, and do not consider the upstream part of the

catchment with several tributaries and rainfall distribution issues. Hydrodynamic interactions are expected to gain importance with increasing spatial scale and number of tributaries and in lowland rivers with large floodplain storage capacities (Vorogushyn et al., 2018). Hence, there is a need to consider hydrodynamic interactions for large-scale risk assessments.

Although the role of the floodplain in attenuating floods in large-scale basins is generally understood (de Paiva et al., 2013; Yamazaki et al., 2011, 2012) the implications for flood risk are not well investigated. Fleischmann et al. (2019) have investigated the role of the floodplains in large scale catchments using the MGB inline coupled hydrologic-hydrodynamic model. They concluded that the absence of right floodplain representation in the model would lead to an expressive discharge overestimation with higher peaks and faster recession limbs.A continuous simulation approach represents an alternative way to estimate flood hazard and risk for large-scale basins compared to the mosaicked approach. It is based on a model chain covering the whole flood hazard/risk process cascade from heterogeneous patterns of precipitations, runoff generation in the catchments, river discharge down to inundation and damage. This approach, termed 'Derived Flood Risk Analysis' (Falter et al., 2015), extends the 'Derived Flood Frequency Approach based on continuous simulation' (e.g. Blazkova and Beven, 2004), often used for flood design estimation. Hence, depending on the goal, the model chain can be truncated and focus on discharge or flood frequency estimation (Hegnauer et al., 2014; Haberlandt and Radtke, 2014), inundation area estimation (Grimaldi et al., 2013) or flood damage and risk assessment (Falter et al., 2015, 2016). The model chain is driven with observed or synthetic climate data, e.g., provided by a stochastic weather generator. The rainfall-runoff modelling provides flow hydrographs at specific locations, which are used as boundary conditions for hydrodynamic flood simulations. Contrary to the gauge-based multisite models, this approach delivers spatially consistent, mass-conservative, and time-continuous flow hydrographs at multiple locations (depending on the structure and resolution of the hydrological models). These continuous flow hydrographs are used as upstream and lateral boundary conditions for continuous hydrodynamic simulations considering hydrodynamic interactions. Continuously coupled rainfall-runoff and hydrodynamic simulations are state-of-the-art (e.g., (Biancamaria et al., 2009; de Paiva et al., 2013; Hegnauer et al., 2014), but they are rarely deployed so far for large-scale flood risk assessments due to their complexity and computational constraints. Winsemius et al. (2013) used a global hydrological model coupled to a kinematic wave routing scheme with simplified inundation approximation for past damage reanalysis. Particularly, two-dimensional hydrodynamic models are computationally demanding when it comes to the long-term simulation of multiple scenarios needed for risk assessments. To our knowledge, there exists no large-scale risk analysis based on a continuous hydrologic-hydrodynamic simulation driven by a long-term weather generator.

In the presented study, we analyse the effect of hydrodynamic interactions on flood hazard and risk assessment for the Rhine basin (185,260 km^2). In particular, we focus on the importance of these interactions with increasing basin scale by comparing (1) flood flows, (2) overtopping volumes, (3) inundation patterns and (4) damage to residential, commercial and agricultural sectors. The comparison is done between assessments with and without consideration of floodplain storage and interactions at different locations upstream and downstream of the basin by comparing two model chains: one based on a sub-basin wise river routing and a fully continuous hydrodynamic approach. Indeed, there are several studies focussing on modelling floodplain flow, but they do not quantify the effect on flood risk. The contribution of our study is to address how this

interaction result in redistribution not only the hazard part (similar to previous studies) but also the estimated damage and risk.

The paper is organised as follows. Section 2 provides a detailed description of the RFM model chain using a piece-wise and a newly implemented continuous routing. The model setup for the Rhine case study, input data for each component of the RFM model chain, and the design of the computational experiments are detailed in section 3, followed by a discussion of the results and conclusions. The cross-section data derivation, extensive model validation, including a comparison of simulated maximum annual discharge/water levels against observations, are presented in detail in the Supplement.

3.3 Methods

For the analysis of hydrodynamic interactions, we use the Regional Flood Model (RFM) (Falter et al., 2015, 2016), representing a model chain for time-continuous, spatially consistent simulation of flood processes from atmospheric input to flood damage and risk. It consists of several components, including the multisite, multivariate Regional Weather Generator (Hundecha et al., 2009; Nguyen et al., 2021), the SWIM rainfall-runoff model (Krysanova et al., 1998), a 1D river routing model based on diffusive wave approximation coupled to a 2D raster-based hinterland inundation model (Falter et al., 2015, 2016) and the FLEMO flood loss estimation model (Kreibich et al., 2010; Thieken et al., 2008) Figure 3.1

3.3.1 Regional Weather Generator (RWG)

The multisite, multivariate weather generator RWG (Regional Weather Generator) is a stochastic model which generates daily time series of precipitation at multiple locations. Based on the state of the generated precipitation (dry/wet), the RWG then generates non-precipitation variables such as temperature (minimum, average, and maximum), relative humidity, and solar radiation. RWG was introduced by Hundecha et al. (2009) and was recently comprehensively evaluated by Nguyen et al. (2021). RWG assumes that extreme precipitation events have different stochastic behaviour compared to the normal precipitation regime. Hence, it uses a mixed distribution, i.e., Gamma distribution for bulk precipitation and a more heavy-tailed Generalised Pareto distribution for extreme precipitation. The spatial and temporal dependence is represented by a first-order multivariate autoregressive model considering the spatial covariance structure. The distribution of the full range of precipitation (including zero precipitation) is formulated by combining the mixed distribution and the frequency of non-zero precipitation. For relative humidity and the temperature variables, the normal distribution is used. For solar radiation, the square-root transformed data is fitted to the normal distribution due to its relatively strong right skewness. RWG is parameterised on a monthly basis to account for seasonality.

3.3.2 Rainfall-runoff model (SWIM)

The hydrological model SWIM (Soil and Water Integrated Model) (Krysanova et al., 1998) is a conceptual semi-distributed rainfall-runoff model for mesoscale catchments. SWIM computes average daily runoff at sub-basin spatial discretisation. Each sub-basin is further subdivided into hydrological response units or hydrotopes, where soil type, land use, and average water table depth are assumed to be homogeneous. Runoff

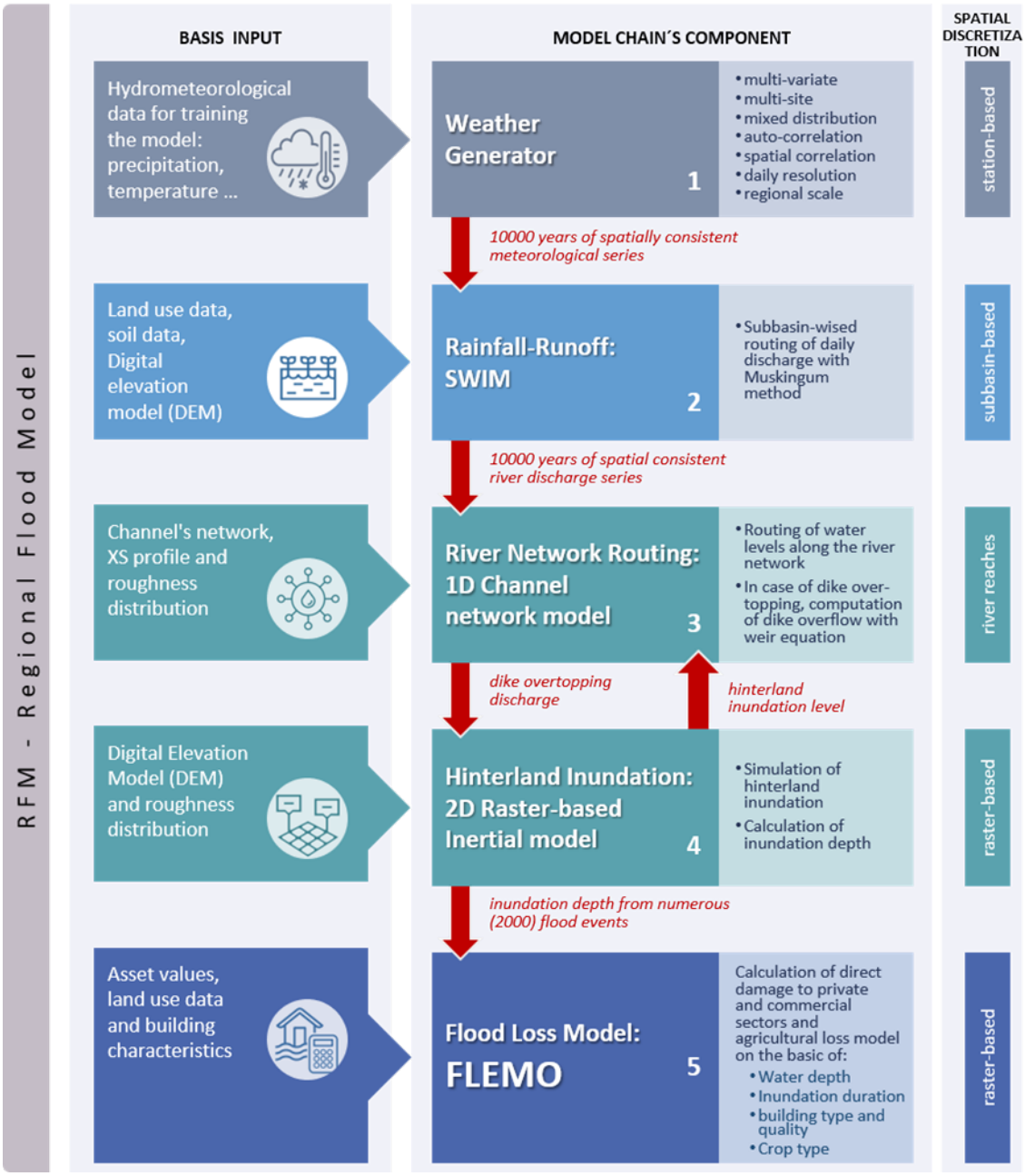

FIGURE 3.1: RFM modelling framework: components, input data requirements, and spatial discretisation.

is calculated for each hydrotope and then aggregated on a sub-basin scale. SWIM is driven by daily precipitation, maximum and minimum air temperature, relative humidity, and solar radiation. SWIM relies on the water balance equation considering snow, precipitation, evapotranspiration, percolation, surface and subsurface runoff, recharge, and capillary rise. The model applies the SCS curve number method for surface runoff volume estimation. For the routing between sub-basins, the Muskingum routing approach is used (Cunge, 1969).

3.3.3 Regional Inundation Model (RIM)

RIM model consists of a 1D hydrodynamic routing model coupled with a 2D hinterland inundation model. The 1D hydrodynamic model solves the diffusive wave approximation of the shallow water equations (SWEs) (Eq. 3.1 & 3.2), where the first two terms- (local and advective acceleration) are neglected in the momentum equation (Eq. 3.1).

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x}\left(\frac{Q^2}{A}\right) + gA\frac{\partial h}{\partial x} - gA(S_o - S_f) = 0 \tag{3.1}$$

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = 0 \tag{3.2}$$

where Q is the discharge ($\frac{m^3}{s}$), A is the cross-section area (m^2), g is the gravitational acceleration ($\frac{m}{s^2}$), S_o is the bed slope ($\frac{1}{m}$), S_f is the friction slope ($\frac{1}{m}$), x is the distance (m), and t is time (s).

An explicit finite difference scheme is used to solve the equations with adaptive time steps following the Courant-Friedrichs-Lewy criterion to secure numerical model stability. However, numerical instabilities may occur due to a sudden increase in the value of the wetted perimeter (P) with a minimal increase in water depth. Such instabilities may happen when the water level exceeds the bankfull area, and overbank flow occurs. Fread (1976) and Smith (1978) solved this problem by dividing the system into two separate conveying systems, i.e., a compound channel. We adopted this solution following Eq. 3.3:

$$\frac{\partial Q_{bf}}{\partial x} + \frac{\partial A_{bf}}{\partial t} + \frac{\partial Q_{Overbank}}{\partial x} + \frac{\partial A_{Overbank}}{\partial t} = 0 \tag{3.3}$$

where Q_bf is the discharge conveyed within the channel ($\frac{m^3}{s}$), Abf is the bankfull (m^2) area and Q_Overbank is the discharge ($\frac{m^3}{s}$) conveyed through Aoverbank the overbank area (m^2).

The deployed solution of the SWEs uses water depth as an upstream boundary condition. Therefore, the flow hydrograph resulting from summing up all the upstream flow (lateral & upstream river branch) is converted into the water depth using Manning's equation for the very first cross-section in each sub-basin. To establish a stage-discharge relationship, the bed slope and hence cross-section bankfull depth are calibrated.

The 1D model uses simplified cross-section data that consist of two components: overbank river geometry and bankfull area. The cross-sections represent the river geometry between left and right dike crests or elevated banks. The overbank geometry is represented by the trapezoidal shape described by six points. Model discretisation and derivation of cross-sections are detailed in the Supplement. In this study, we use two

implementations of the 1D hydrodynamic model to quantify the effect of hydrodynamic interactions. In the original model version (RIM1.0), a piece-wise routing of overbank flow was implemented (Falter et al., 2015, 2016) (Supplement Figure 3.12A). The bankfull depth is not explicitly represented in the cross-sections (Supplement Figure 3.13). Discharge corresponding to the bankfull depth is subtracted at the upstream node of each sub-basin, where runoff from the SWIM model is assigned as a boundary condition. Hence, floodplain storage effects due to dike overtopping and inundation, i.e., hydrodynamic interactions, are considered only within single sub-basins and are not propagated across the sub-basin boundaries.

Further, a continuous routing approach (RIM2.0) is implemented (Supplement Figure 3.12B), which is based on the full cross-section geometry, i.e. rectangular bankfull area and trapezoidal overbank geometry (Supplement Figure 3.13). The flow is continuously routed through the entire river network. In this way, storage effects due to dike overtopping and floodplain inundation on downstream flow hydrographs can be explicitly accounted for. RIM2.0 uses the same computational engine as RIM1.0 based on the explicit solution of the diffusive wave equation. Derivation of cross-sections for both RIM1.0 and RIM2.0 models is detailed in the Supplement.

The 2D hinterland inundation model is deployed when water levels in the river channel overtop dike crests. The outflow into the hinterland is calculated with the broad crested weir equation and provided as a boundary condition for the inundation model. The connection between 1D and 2D models is one-way, where no return flow from the 2D model to the 1D channel is currently accounted for. As soon as the water level in the river channel drops below the crest level, no further overtopping flow is simulated. The 2D hinterland inundation model solves Eq. 3.1 and Eq. 3.2 in two dimensions with a neglected advection acceleration term. The 2D model domain is discretised into a regular grid to compute fluxes q per unit width in both directions, x and y and update water depths h in each raster cell (i, j). The explicit solution for q proposed by Bates et al. (2010) (Eq. 3.4) reads:

$$q_{t+\Delta t} = \frac{q_t - g\, h_{flow}\, \Delta t\, S_f}{\left(1 + \frac{g\, h_{flow}\, \Delta t\, n^2\, |q_t|}{h_{flow}^{10/3}}\right)} \tag{3.4}$$

$$\frac{\Delta h^{i,j}}{\Delta t} = \frac{q_x^{i-1,j} - q_x^{i,j} + q_y^{i-1,j} - q_y^{i,j}}{\Delta x} \tag{3.5}$$

$$\Delta t_{max} = \alpha \frac{\Delta x}{\sqrt{g\, h_{flow}}} \tag{3.6}$$

where n is Manning's roughness ($\frac{s}{m^{1/3}}$), h_{flow} is the flow depth between cells (m), and Δt is the time step (s).

Then water depth is calculated based on the continuity equation (Eq. 3.5). An adaptive time step (Eq. 3.6) is used to improve the stability of the numerical scheme, where the value of parameter α ranges from 0.2 to 0.7, as suggested by Bates et al. (2010)). The 2D hydrodynamic code using an explicit finite-difference scheme is implemented in the CUDA Fortran environment to enable simulations on highly parallelised NIVIDIA Graphical Processor Units (GPUs) (Falter et al., 2016).

For each flood event, where dike overtopping and hinterland inundation occur, two grids are produced: maximum water depth and inundation duration. The extracted grids are used as input for the flood loss model. The season of the event is also recorded since it is required for the agricultural loss calculation. To limit the 2D computational time, water depths are set to zero ten days after the last overtopping day of a dike. It is expected that the maximum depths and extent relevant for damage estimation have been reached within this period.

3.3.4 Flood damage and risk estimation

Flood damage models

The loss estimation component of the model chain is based on the FLEMO models for private and commercial sectors and an agricultural loss model. These models are developed from empirical damage data of German river floods and validated in previous modelling studies (Klaus et al., 2016; Kreibich et al., 2010; Seifert et al., 2010; Kuhlmann, 2010; Thieken et al., 2008). For the private and commercial sector, damage functions to estimate relative losses to buildings and content are based on water depth, discretised into 6 levels (<0.21, 0.21–0.6, 0.61–1, 1.01–1.5, >1.5 m). The private sector model, FLEMOps, additionally distinguishes 3 building types (single-family, semi-detached, multi-family) and 2 building quality levels (low/medium quality, high quality) (Thieken et al., 2008). The commercial sector model FLEMOcs considers additionally 4 sub-sectors (producing industry, trade, corporate services, public & private services) and 3 company size classes according to the number of employees (1–10, 11–100, > 100) (Kreibich et al., 2010). For the agricultural sector, the relative damage functions are based on four inundation duration classes (1–3, 4–7, 8–11, >11 days), the month of the flood, and 7 crop types (canola, maise, potatoes, sugar beet, barley, rye, wheat (Kuhlmann, 2010; Klaus et al., 2016). The damage models use the inundation results produced by RIM to perform spatial intersection with the exposed assets and apply the FLEMO damage functions specific to each sector, implemented via table joins on a PostgreSQL database.

Exposure estimation

Residential building asset values for Germany are estimated according to the approach of (Kleist et al., 2006), who based their estimates on official statistical data, i.e. a total living area for three classes of residential buildings per district provided by the Federal Statistical Office of Germany and the standard construction costs per square meter gross floor space published by the German Federal Ministry of Transport, Building and Urban Development. The asset values are disaggregated based on the ATKIS land cover data, according to Wünsch et al. (2009). Based on Paprotny et al. (2020), residential content values are derived from the building values by dividing them by 5.09 (ratio between the household building and content values in Germany at the 2018 price level). Commercial building and content values are estimated according to Paprotny et al. (2020a) using Eurostat data (e.g., distribution of company sectors per NUTS3 region, gross value added for the NUTS3 region by economic activity, fixed assets by the economic activity) and disaggregation to ATKIS land cover. The agricultural exposure, i.e. revenues, are estimated by multiplying the yield with the sales price (Kuhlmann, 2010; Klaus et al., 2016). The revenue (in Euro per hectare) of a particular crop in a region, averaged over five years (to equalise strong annual fluctuations), is determined with the help of an agricultural statistics database. Regional differentiation considers 38 districts in Germany.

Risk metrics

The occurrence exceedance probability (OEP) is used as a risk metric. OEP is the probability that the maximum loss from a single event in a given year exceeds a certain amount. OEP is calculated by ranking the most severe loss event per simulated year and counting the number of years in the entire time series, in which the loss of a given event is exceeded. These damage-exceedance probabilities are plotted in a flood loss curve. The expected annual damage (EAD) equally distributes the risk over the time series and is computed as the area under the OEP curve (e.g. Merz et al., 2009). The value at risk (VAR) gives the expected damage for a specific exceedance probability, while the tail-value at risk (TVAR) is defined as the average expected damage above the threshold used for VAR. We provide these measures based on the 100-year event, which corresponds to a simulated 0.99 probability of non-exceedance in a given year.

3.4 Setup of the Regional Flood Model for the Rhine basin

3.4.1 Study area

The river Rhine is one of the largest rivers in Europe, with a total length of 1,233 km and a total catchment area of 185,260 km^2 (Figure 3.2). It originates in the Swiss Alps and flows through Germany, France, and the Netherlands. Most of Luxemburg and some parts of Austria and Belgium drain into the Rhine river. Major tributaries include Neckar, Main, Moselle, Lahn, Sieg, Lippe, and Ruhr. About 58 million inhabitants live in the river basin, with 10.5 million in flood-prone areas (ICPR, 2013). The basin topography ranges from the high alpine regions with elevations up to 2,500 $masl$ to lowland floodplains at the Lower Rhine. Major floods in the upstream parts of the basin are caused by snowmelt combined with rain-on-snow, whereas the Middle and Lower Rhine floods are dominated by long-lasting frontal rainfalls in winter and early spring. The average discharge at the gauge Lobith at the border between Germany and the Netherlands is 2,200 $\frac{m^3}{s}$, and the maximum observed discharge is 12,600 $\frac{m^3}{s}$ in 1926 (Pinter et al., 2006). In the past, flood management in the densely populated floodplains was focused on dike reinforcement with a design discharge corresponding to return periods between 1/200 to 1/500 in Germany and 1/1250 to 1/2000 in the Netherlands (te Linde et al., 2011; VenW, 2007). Consequently, floods may occur in the German parts of the basin, while the Dutch Rhine reaches will likely receive significantly reduced flood flows due to hydrodynamic interactions (Apel et al., 2009).

3.4.2 Weather generator setup

The regional weather generator (RWG) is set up for a large region embracing entire Germany and parts of the neighbouring countries and covering five major river basins: Ems, Weser, Upper Danube, Elbe, and the Rhine. The setup is based on daily climate observations for the period 1950-2003. The dataset contains six variables (precipitation, minimum, average, and maximum temperature, relative humidity, and solar radiation) at 528 locations, of which 465 are climate stations (Österle et al., 2016) and 63 are grid points for the French part of the Rhine basin from E-OBS gridded dataset (Haylock et al., 2008). In this study, we use the RWG to generate 1,000 years of synthetic time series of the six mentioned variables, which are then used to drive the hydrological model SWIM.

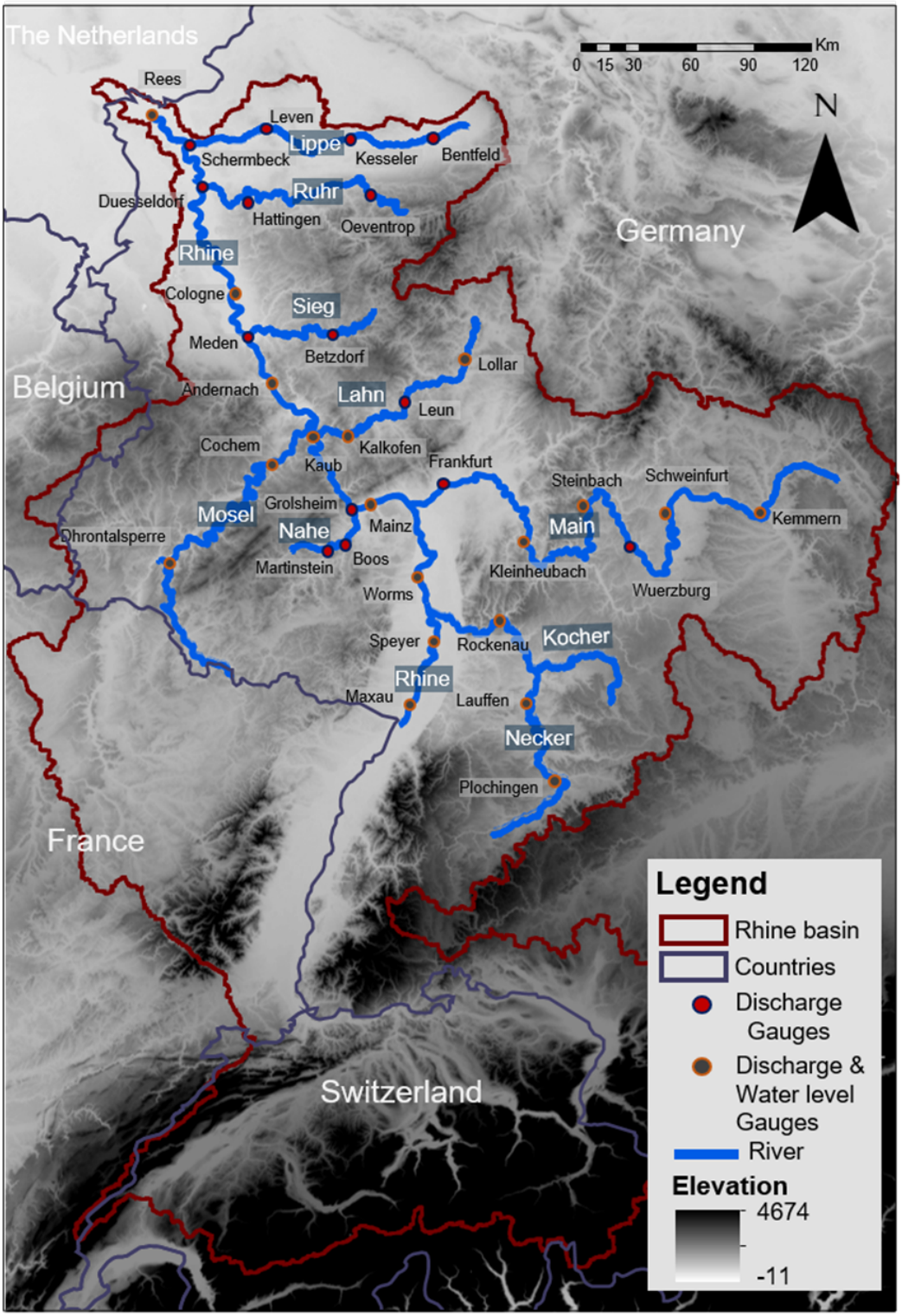

FIGURE 3.2: Rhine basin upstream of Rees with the river network used in the 1D RIM model component. Labelled gauges are used for calibration and validation of the 1D hydrodynamic model.

3.4.3 SWIM model setup and calibration

In this study, the Rhine catchment is divided into 936 sub-basins based on the digital elevation data provided by the Federal Agency for Cartography and Geodesy in Germany (BKG). Soil and land-use data are derived from the soil map for Germany (BÜK 1000 N2.3), obtained from Bundesanstalt für Geowissenschaften und Rohstoffe (BGR) and the European Soil Database map, obtained from the European Commission's Land Management and Natural Hazards unit and the CORINE (COoRdinated INformation on the Environment) land cover map. SWIM is driven by meteorological data, either observations or RWG-generated data. These sub-basins are grouped into eight sub-regions (Supplement Figure 3.11): Upper Rhine, Nahe, Neckar, Mosel, Lahn, Sieg, Lippe, and Main & Rhine. Nine parameters are calibrated for each sub-region automatically using the SCE-UA optimisation algorithm (Duan et al., 1994) (Supplement).

3.4.4 Setup and calibration of the RIM 1D model

RIM 1D (both RIM1.0 and RIM2.0) is set up for the river network represented in Figure 2. The network includes the main channel starting at the weir Iffezheim and major tributaries with a catchment area of at least 500 km^2. Overbank cross-section geometry is derived from the 10x10m DEM with a vertical accuracy of $\pm$0.5–2 m provided by the Federal Agency for Cartography and Geodesy in Germany (BKG). For the RIM2.0 model, the bankfull depth d_{BF} is initially estimated using the power-law relationships calibrated on more than 1200 surveyed cross-sections provided by the Federal Institute of Hydrology (BfG) for the Rhine, Mosel, and Necker rivers (Supplement). However, given the uncertainty related to these estimates and the uncertainties of the DEM, we calibrate the estimated bankfull depth in the RIM 2.0 model by varying d_{BF} in the range $\pm$0.5 m to maximise NSE_m (Eq. 3.7) for the simulated discharge and water level within the calibration period. Bankfull depth calibration was also previously applied by Neal et al. (2012) and Wood et al. (2016). Additionally, the slope used to compute the stage-discharge relation to convert the inflow for each sub-basin into the water level is adjusted in the calibration process to maintain the water balance. The model is manually calibrated based on the period 1996-2003, and the validation is carried out for the whole period between 01-01-1950 and 30-12-1995 . Several metrics are considered to assess the performance of the RIM model by comparing the simulated hydrograph by RIM and SWIM models with observed flows at 34 gauging stations and observed water levels at 19 gauges (Figure 3.2). The dike heights represented in the edge points of the cross-section are estimated initially from the 10x10m DEM but are likely to be underestimated due to smoothing effects. Dikes are designed for a specific return period flood (te Linde et al., 2011; VenW, 2007). Since no consistent large-scale information on the dike heights is presently available, a regionalised 200-year return period flow in combination with Manning's equation was used to estimate the elevation of the dikes along the Rhine main channel, while a return period of 100-year was used for other tributaries, according to te Linde et al. (2011). This step is carried out after the model is calibrated. The estimated dike elevation is compared with the extracted dike levels from the DEM. The higher of both values is finally taken as a dike crest elevation. The performance of the model in the calibration and validation period is evaluated based on the modified Nash-Sutcliffe efficiency NSE_m (Eq. 3.7), which emphasises peak flows.

$$NSE_m = 1 - \frac{\Sigma_t Q_{obs}(Q_{obs} - Q_{sim})^2}{\Sigma_t Q_{obs}(Q_{obs} - Q_{avg})^2} \tag{3.7}$$

where Q_{obs} is the observed discharge at time t, Q_{sim} is the simulated discharge, and Q_{avg} is the mean of observed discharges. Model performance regarding water levels is assessed by mean absolute error (MAE) and mean bias error (MBE), which provide values in meters and thus better guidance on model performance than NSE_m.

3.4.5 Setup of the RIM 2D model

The RIM 2D hydrodynamic model is set up on the resampled DEM with a resolution of 100x100 m for the areas behind dikes. Floodplains between the dikes or elevated banks are masked in the DEM and are not used for 2D computations. This significantly reduces the computational load. A uniform Manning's roughness value of 0.03 is assigned to each raster cell, as suggested by Falter et al. (2013) and is not calibrated. The major flood characteristics at this scale are found to be not very sensitive to the floodplain roughness but rather determined by the 1D-2D model interface, and dike heights, i.e. how much water enters the 2D model domain from the 1D model Bates et al. (2010). Furthermore, no inundation due to dike overtopping occurred in the Rhine basin over the past simulated period, which would allow model calibration.

3.4.6 Design of computational experiments

To investigate the effect of the hydrodynamic interactions, the RIM2.0 results based on continuous routing are compared to those of RIM1.0. Daily synthetic meteorological data of 1000 years from the regional weather generator are used to drive the SWIM model in order to generate a set of events exceeding the dike design level. To account for the fact that hydrodynamic models require calibration, which may compensate for the methodological difference between both models, the calibration of both models is done using the same procedure to make the comparison meaningful and to make sure that the differences between the results of both models are due to the new conceptual changes. To analyse the effect of hydrodynamic interactions, we compare simulated discharge along the river network. Comparison of overtopping volume resulting in inundation areas and damages elucidates the effect of interactions on hazard and risk. We trace the effects along the channel profile and investigate the role of spatial scale on the effect of hydrodynamic interactions.

3.5 Results and Discussion

3.5.1 Performance of SWIM, RIM1.0, and RIM2.0 flood routing

In this section, the performance of the hydrological model SWIM and both hydrodynamic models (RIM1.0 and RIM2.0) is presented and discussed. SWIM and RIM2.0 are evaluated in terms of the simulated discharge time series for the period 1950-2003, excluding the calibration period (1996-2003). Further, the simulation of water level hydrographs for RIM 2.0 is evaluated. RIM1.0 output cannot be directly compared to SWIM and RIM2.0 output since it routes only overbank flow (flow exceeding bankfull depth) and thus does not produce continuous hydrographs. For this reason,

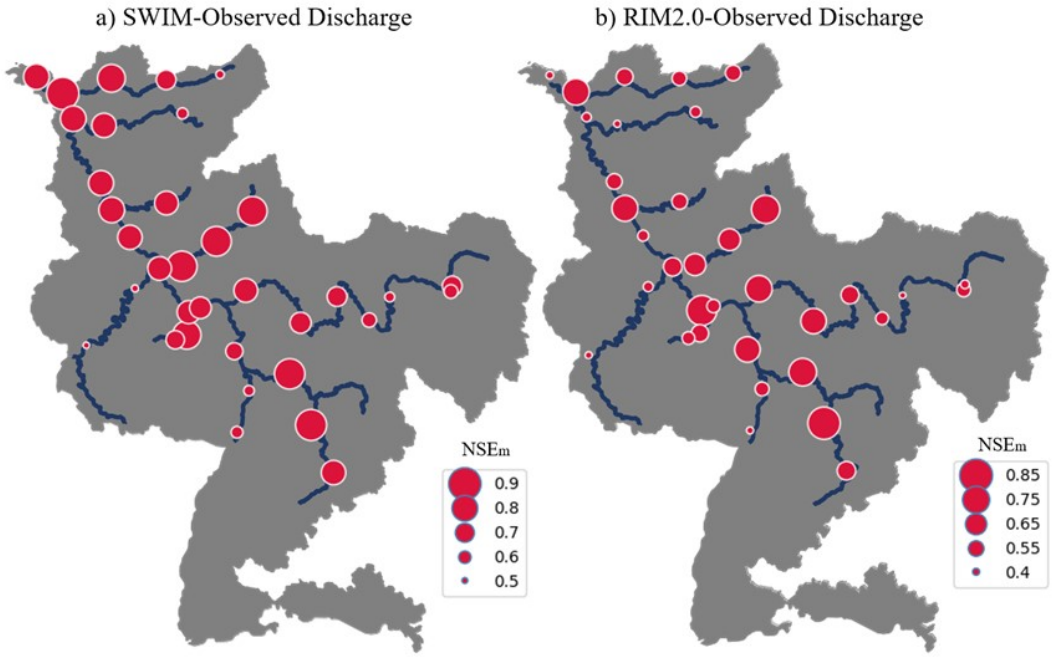

FIGURE 3.3: Performance of SWIM a) and RIM2.0 b) (NSE_m values) with respect to observed discharges.

the performance of RIM1.0 is checked visually at the gauge locations, and the ability to reproduce peak flows and water levels is assessed.

Figure 3.3a shows the performance of SWIM over the entire catchment in terms of the NSE_m values. SWIM performance is overall very good, with values ranging from about 0.9 at the gauge Schermbeck (Lower Rhine) down to 0.51 at the gauge Dhrontalsperre on the Mosel river. The latter is likely influenced by the reservoir operation. Further, the performance of RIM2.0 is compared to observed discharge (Figure 3.3b). Generally, RIM2.0 has a good performance all over the catchment compared to observed hydrographs with NSE_m values ranging from 0.39 to 0.84. However, the performance of RIM2.0 deteriorates compared to SWIM in the downstream parts. We attribute this to the representation of the bankfull depth in the lowland parts of the catchment. Furthermore, floodplains between dikes at the Lower Rhine become much wider compared to upstream parts and are less well represented by the 1D simplified cross-sections. Further improvements in the representation of the cross-section geometry of the downstream reaches are required in the future.

In the Mosel river (Figure 3.2), RIM2.0 performs comparably poor to SWIM with NSE_m around 0.5, as RIM2.0 is not intended to compensate for errors of preceding modules in the model chain. This is likely a consequence of substantial errors in the precipitation input in the French part of the basin, where coarse resolution gridded data had to be used. In the Ruhr tributary (Figure 3.2), the performance of SWIM is acceptable with NSE_m between 0.6 and 0.8, whereas RIM2.0 drops below 0.55, likely due to misestimation of river conveyance in this heavily trained reach.

The inspection of the hydrographs for selected flood events, including a major flood in 1993, reveals a very similar performance of RIM1.0 and RIM2.0 at gauges Steinbach, Lauffen, Leun and Chochem, while SWIM and RIM1.0 slightly better match the observed peak flows compared to RIM2.0 at Cologne and Duesseldorf (Supplement Figure 3.16). All three models show comparable performance in relation to annual

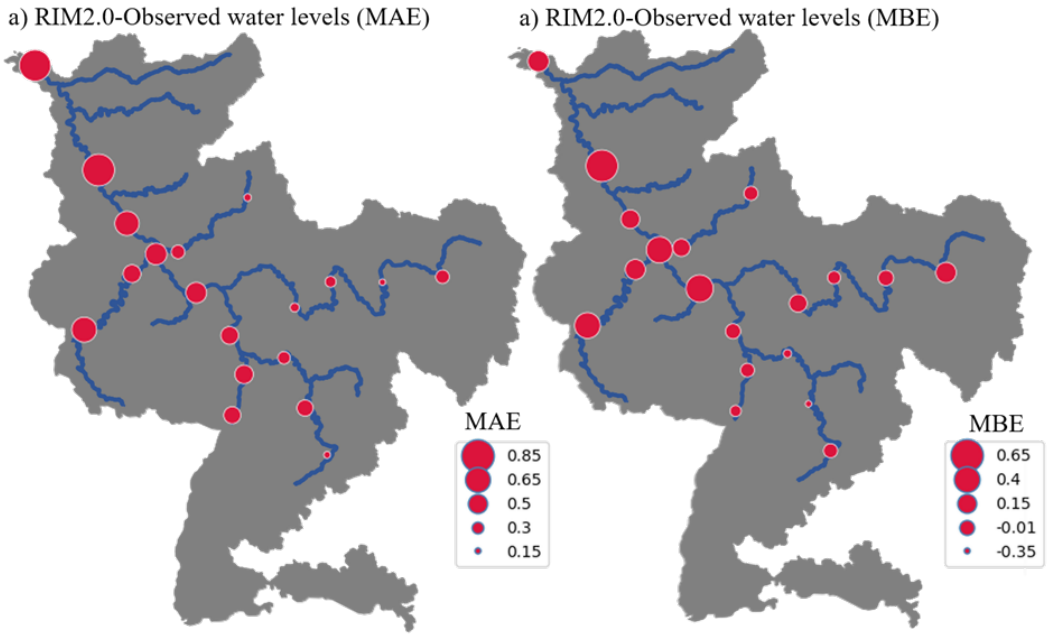

FIGURE 3.4: Performance of RIM2.0 against observed water level in the validation period. MAE and MBE are calculated as measured minus modelled values. Negative MBE indicates overestimation, and positive indicates underestimation.

maximum observed flows at most of the selected gauges (Supplement Figure 3.17). However, RIM2.0 overestimates some high flow peaks at gauge Lauffen in the Neckar basin and underestimates high flow peaks in the Lower Rhine. The latter is likely due to an underestimation of bankfull depth, which results in more water conveyed above bankfull depth and stronger attenuated peaks. For water level hydrographs, MAE ranges between 0.83 m at gauge Cologne to values below 0.2 m in the Main and Neckar tributaries (Figure 3.4). While the MBE ranges between 0.33 m (overestimation) at Lauffen to 0.63 m (underestimation) at Cologne.

Further, we evaluate the performance of RIM1.0 and RIM2.0 with regards to water level simulation at selected gauges in a period in December 1993 – March 1994, including a major winter flood (Supplement Figure 3.18), and compare the simulated water levels to observations for annual maximum events (Supplement Figure 3.19). RIM2.0 outperforms RIM1.0 in the upstream parts of the Rhine basin and the tributaries. Both models underestimate observed maximum water levels of high flow events at the Lower Rhine, whereas, at Cologne, RIM2.0 underestimation is more severe than in RIM1.0. A more detailed performance evaluation of all models is provided in the Supplement. Overall, the models perform reasonably well considering the scale of the basin, the limited information about the cross-section geometry and the degree of anthropogenic influences on river geometry and flow conveyance. The performance of large-scale hydrodynamic models for river networks with uncertain bathymetry reported in the literature is comparable to the one achieved in our study. For instance, for the Amazon River and its tributaries, (de Paiva et al., 2013) achieved comparable discharge simulation performance with a combined Muskingum-Cunge and a full Saint-Venant hydrodynamic model with NSE values mostly between 0.6 and 0.9 but also dropping to 0.2-0.6 for some reaches. Most of the reaches exhibited a bias in the simulated water depths between 3 and 15 m, which is by far larger than in our study being in the range of +0.35 to -0.65 m. Impact of hydrodynamic interactions on flood hazard.

In this section, we analyse the impact of hydrodynamic interactions, which result from inundation and storage effects in the river network, on flood hazard characteristics. First, we compare the overtopping volume over the dike crests and total inundation area in the Rhine basin and present the results for the first 100 years of the 1000-year simulation period for the sake of brevity (Figure 3.5). The tendency is similar for the remaining period. Both models show the same major events. Besides, RIM1.0 simulates a large number of small events with small inundation areas. On the contrary, RIM2.0 simulates overtopping and inundation for only 5 events in the selected 100-year period. Taking into account the protection level of dikes in the Rhine basin, the overtopping frequency of RIM2.0 appears much more realistic, suggesting that the representation of cross-section geometry, dike height, and water levels is closer to reality in RIM2.0. For the simulated major flood events, the overtopping flow and inundation areas from RIM1.0 are considerably larger than simulated by RIM2.0.

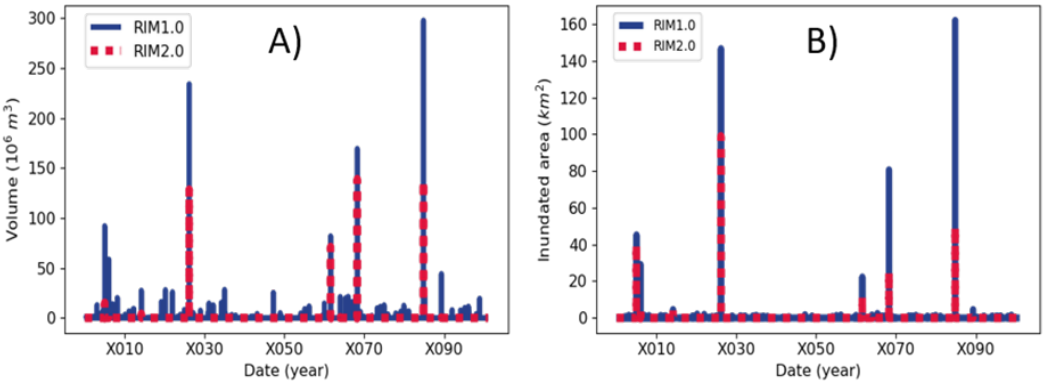

FIGURE 3.5: Overtopping volume A) and inundation area B) simulated by RIM1.0 and RIM2.0 for a 100-year synthetic simulation period. Years in synthetic simulations are denoted with 'X'.

Figure 3.6 compares inundation areas for the most severe event Nov-X084, i.e. the event with the highest overtopping volume and inundation area (Figure 3.5). The inundation areas tend to occur in the same river reaches, however, RIM1.0 inundations are larger compared to RIM2.0. We observe this behaviour for all major events in Figure 5. Not only inundation areas but also inundation depths are overestimated by RIM1.0, as shown in the histograms of inundation depth for the regions around Trier (Mosel), Leverkusen (Rhine), and Mainz/ Wiesbaden (Main). These results suggest that the hydrodynamic interactions, which are insufficiently represented in RIM1.0, considerably reduce water levels downstream. This is further illustrated by analysing the longitudinal profiles of maximum water level and overtopping volume for the river Moselle and the Rhine during the Nov X084 and April X068 events (Figure 3.7). The profiles start at the upstream node in the Moselle river and continue along the Middle and Lower Rhine. We select this course since inundation mainly occurs along these tributaries in these two events.

Figure 3.7 shows that the difference in maximum water levels between RIM1.0 and RIM2.0 fluctuates around zero at the upstream part of Mosel and is controlled by overtopping in both models. The maximum difference happens upstream of the confluence with the Rhine river at Trier (see Figure 3.2), which results in the highest overtopping (spatially during both events), and consequently highest difference in inundation depth and extent (Figure 3.6) between both models. The overtopping frequency and volume are noticeably higher in RIM1.0 at the downstream reaches compared to RIM2.0 for both events.

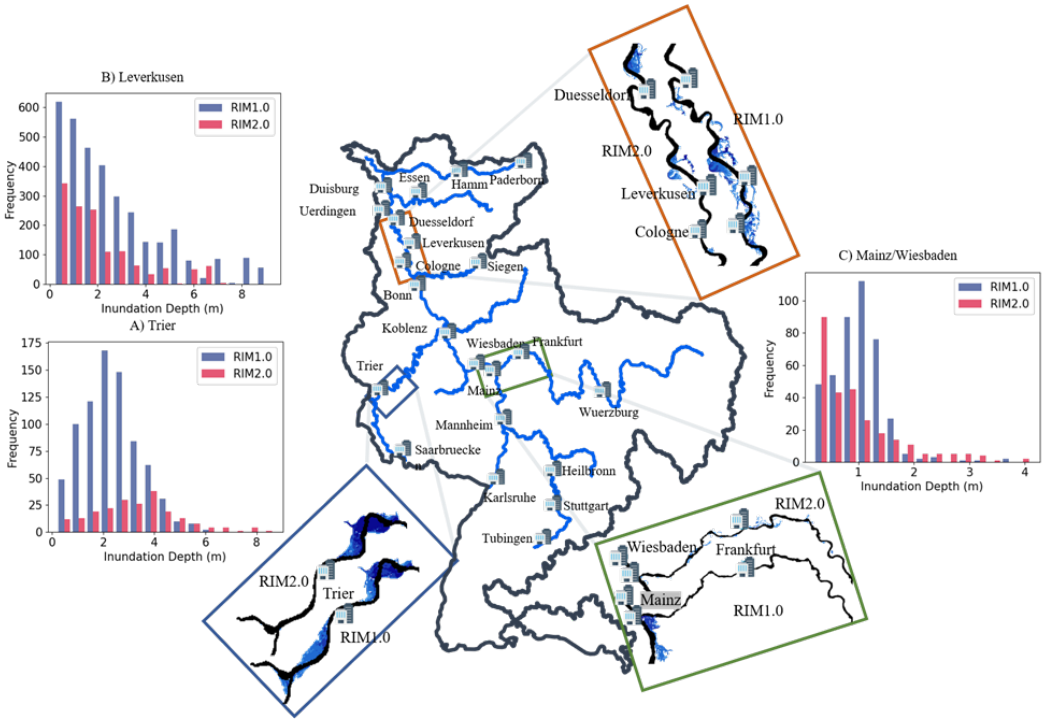

FIGURE 3.6: Comparison of the inundation extent and depth histogram from RIM1.0 and RIM2.0 during the strongest event (Nov-X084) in the first 100-year simulation period. Inundation depth for RIM1.0 and RIM2.0 at Trier a), Leverkusen b) and Mainz/Wiesbaden c) for the Nov-X084 event.

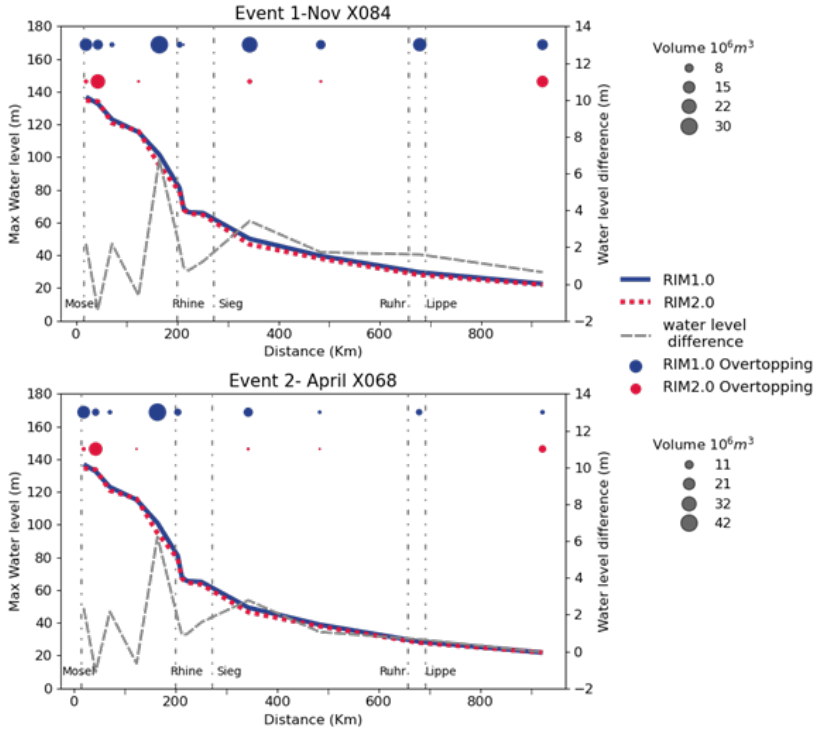

FIGURE 3.7: Longitudinal profile of the maximum water level (left y-axis), the difference in maximum water level between RIM1.0 and RIM2.0 (right y-axis), and the overtopping locations represented by points of different sizes corresponding to the overtopping volume in a million m^3. The profile corresponds to events Nov X084 and April X068 starting from the Moselle river till Rees at the outlet of the Rhine catchment. Confluences of major tributaries such as Saar, Rhine, Sieg, Ruhr and Lippe are indicated with vertical dashed lines.

The maximum water levels in RIM1.0 are much less sensitive to overtopping compared to RIM2.0. In consequence, RIM1.0 produces more overtopping volume along the entire reach. We explain this pattern by the hydrodynamic interactions. RIM1.0 considers hydrodynamic interactions only within individual sub-basins, i.e., dike overtopping and inundation cause discharge and water level reductions only within the sub-basin where they occur. This storage effect is not translated to the sub-basin downstream. The discharge boundary condition in the subsequent sub-basin is updated using SWIM output, which is not aware of the overtoppings upstream. Hence, the total floodwater volume in the river system is overestimated, causing stronger overtopping and larger inundation areas in RIM1.0. To some extent, this effect also occurs in the mosaicked flood mapping approach, in which upstream boundary conditions of piece-wise hydrodynamic models are updated from the local regionalised flood frequency curves (e.g. Quinn et al., 2019). There may be cases where the streamflow observations comprise events with upstream inundations and, thus, downstream flow reductions, which then may leave a trace in the flood frequency curve. However, given the length of measured discharge series and the rarity of dike overtopping and inundations along embanked rivers, the degree to which the effect of hydrodynamic interactions is contained in the results of the extreme value statistics is limited.

3.5.2 Impact of hydrodynamic interactions on damage and risk

In this section, we investigate how decisive hydrodynamic interactions are for flood damage and risk estimates. For this, we integrate the risk from all flood events simulated in the 1000-year period of synthetic simulations. The comparison of estimated economic risk from both models (Figure 3.8) indicates that the chain, including RIM2.0, generally simulates less damage. This corresponds to the overall smaller inundation areas and depths discussed in the previous section. We attribute this difference in particular to the effect of hydrodynamic interactions. We analyse the overall simulated damages and flood risk in the Rhine basin and discuss the results in view of previous studies.

Simulated absolute damage numbers for the residential, commercial, and agricultural sectors are in reasonable agreement with actual reported losses, though a direct comparison between single observed and simulated events is not expedient. For instance, VAR, i.e., the 100-year total loss amounts to EUR 1.94 billion in RIM2.0 compared to EUR 3.14 billion in RIM1.0, while the reported loss of the 1993 event (about 100-year flood at Cologne) in the German Rhine was DM 1.5 billion (Engel, 1997), which corresponds to roughly EUR 1.2 billion today. The estimated EAD for the two model variants (EUR 0.16 billion by RIM1.0 and EUR 0.08 billion by RIM2.0) is considerably lower than estimated by te Linde et al. (2011) (EUR 0.79 billion for the German Rhine). This difference likely originates from the differences in the approaches to compute flood hazard, asset values and different damage models. First, the approach of te Linde et al. (2011) is based on the homogeneous assumption of an “extreme” inundation scenario provided by ICPR (2001) without an associated return period. Probabilities of inundation and losses in that study were retrospectively assigned by using the nominal protection levels in the respective parts of the Rhine reach. Hence, EAD was estimated as summed probability-weighted losses corresponding to the return period of dike overtopping, i.e., dike design level. Damage estimations corresponding to a specific high return period under homogenous assumptions are found to be largely overestimated (Metin et al., 2020). Second, the exposure values used in the Damage Scanner model in te Linde et al. (2011) include infrastructure and other land-use types not considered in our approach. Damage Scanner is found to overestimate loss values by more than 3 times compared to

the FLEMO damage model (Jongman et al., 2012). Finally, damage calculated by te Linde et al. (2011) comprises a share of, on average, 5% indirect damage, which is not accounted for in our study.

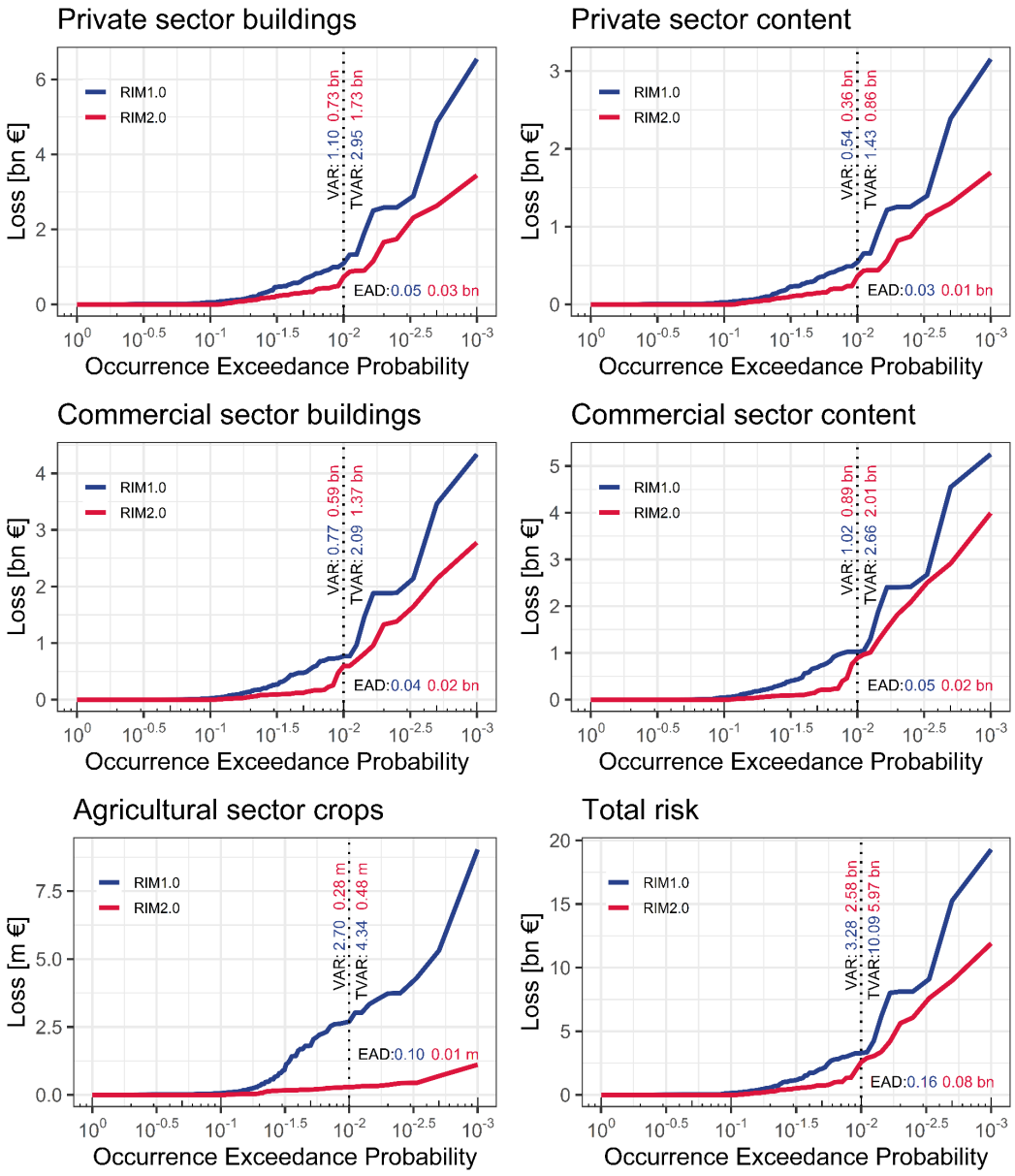

FIGURE 3.8: Risk per economic sector derived from both model chains, the VAR and TVAR and the EAD values are displayed for both models in blue (RIM1.0) and red (RIM2.0. The agriculture loss is given in million Euro in contrast to other sectors.

The private and commercial sectors contribute equally to the overall financial risk in our simulation. However, the loss in the agricultural sector differs largely between both models due to the difference in the inundation extent, it only makes a small fraction in the overall loss (Figure 3.8). This result is expected for the highly industrialised Rhine catchment, particularly the Lower Rhine. The difference in TVAR is larger than the difference in VAR, indicating that the models show less agreement in the upper tail. We find that events above the 100-year return period contribute significantly to the difference in risk curves between the RIM1.0 and RIM2.0 model chains. Different protection levels around 100 and 200-year return periods can be observed in the loss

curves. A steep growth of the curve is visible at these thresholds, depending on whether damages occur along the main course with the higher protection standard (~200 years) or in the tributaries, where the protection standard is likely lower (~100 years). OEP considers the most severe event per year. However, there is a possibility of having multiple flood events in one year, which can then be reflected in the risk curves by the Aggregated Exceedance Probability (AEP). As AEP and OEP are mostly identical (not shown), which means that both models, RIM1.0 and RIM2.0, seldom encounter two major events per year, only OEP values are presented.

To analyse the impact of hydrodynamic interactions on flood damage and risk, we subdivide the river network into different segments based on the location of major cities and confluences (Figure 3.9). The results of both RIM1.0 and RIM2.0 are aggregated and used to derive the risk curves for each segment. At the upstream segments 1, 7, and partially 10, the risk curves by RIM2.0 are higher than the RIM1.0 curves, which indicates higher losses produced by RIM2.0. Further downstream, the risk curves converge (segments 2 and 3). At the lower part of the Rhine (segments 4, 5 and 6), the modelled risk by RIM1.0 consistently exceeds the estimations by RIM2.0. In these segments with extended floodplains and high exposure, there is a strong potential for widespread inundation and high losses. In segment 9 i.e. Lippe tributary, the RIM2.0 calibration results with respect to discharge are poor (Figure 3.3), presumably due to anthropogenic effects on river geometry and discharge in this strongly affected reach. Moreover, water levels were not available for model calibration, which affects water level estimations. Hence, significant differences resulted between RIM1.0 and RIM2.0 in this reach that explain the resulting risk curves. The discharge contribution of this segment to the main Rhine channel is in the order of a few percent, so the effect on the risk downstream is expected to be small.

Figure 3.10 compares the spatial distribution of inundation areas and damages for two large events (Nov-X084 and April-X068, see Figure 3.5). The inundation areas and damages are aggregated for the sub-areas shown in Figure 3.10. To present the results on a proper scale, the hydrological sub-basins of SWIM are aggregated to larger areas. The total damage of the Nov-X084 event is 7.9 billion, estimated from RIM1.0 and 5.6 billion from RIM2.0, while the April-X068 event simulation results in 3.2 billion and 1.0 billion for RIM1.0 and RIM2.0, respectively. Locating these numbers on the risk curves would result in return periods of 200 and 166 years for the Nov-X084 event simulated by RIM1.0 and RIM2.0, respectively, and 100 and 66 years for the April-X068 event by RIM1.0 and RIM2.0, respectively. The shading of the sub-area indicates the absolute inundation area/damage. Nov-X084 event originates in the Neckar, Main, and Mosel tributaries, whereas April-X068 event strikes Main and Mosel (see Figure 3.2 for orientation), and the Lower Rhine is affected by both floods. For both events, inundation areas and damage simulated by RIM1.0 are often similar at upstream sub-areas but clearly dominate the Lower Rhine compared to the RIM2.0 results. The Lower Rhine is hardly affected by Event 2 in the RIM2.0 model. We attribute the difference in inundation/damage patterns between both models mainly to the difference in the routing scheme and, consequently, to the consideration of hydrodynamic interactions. However, underestimation of flow and water level in RIM2.0 compared to RIM1.0 due to differences in cross-section geometry and parameterization may result in less inundation in RIM2.0 and thus represents a limitation for the achieved results. The higher sensitivity of RIM2.0 simulations to dike overtopping due to the consideration of hydrodynamic interactions becomes pronounced in the Lower Rhine, where several overtopping occur during these events. Consequently, less inundation and damage occur during both events at the Lower Rhine, when hydrodynamic interactions are accounted for.

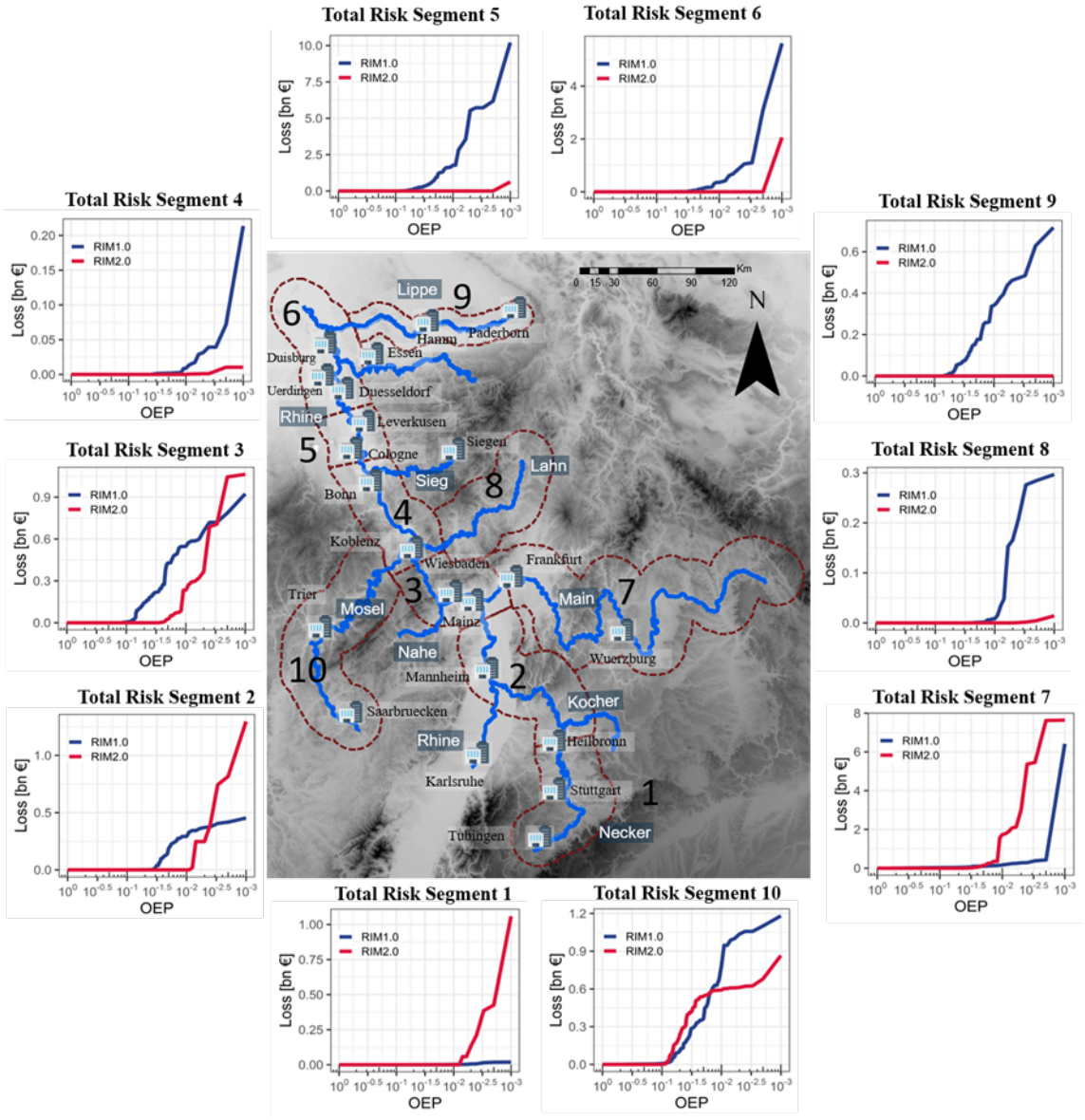

FIGURE 3.9: Risk curves derived from RIM1.0 and RIM2.0 for river basin segments at different locations across the main Rhine and the tributaries.

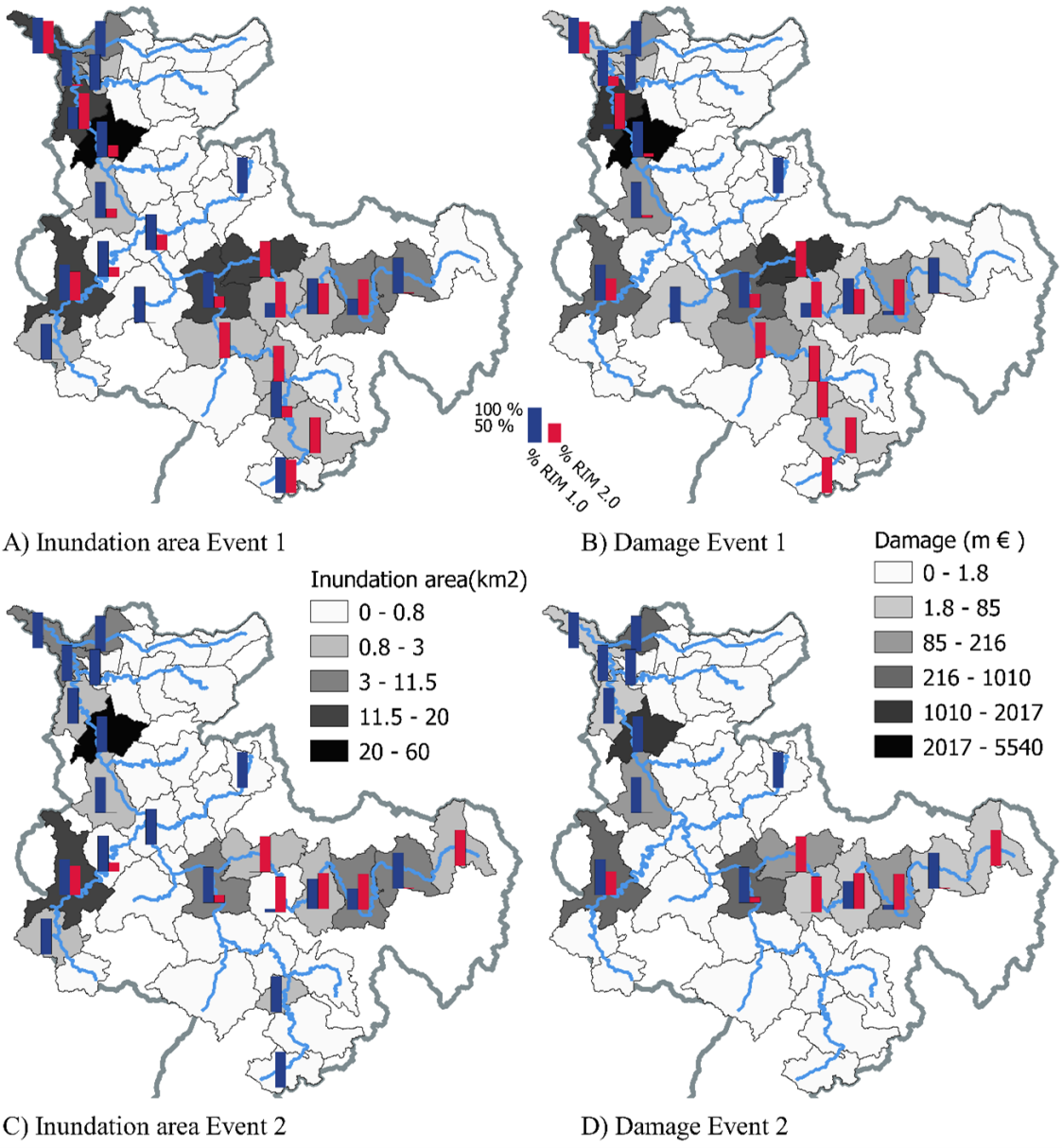

FIGURE 3.10: Comparison of inundation area and damage between RIM1.0 and RIM2.0 for two selected events in the synthetic series. The shaded colour in the background refers to the maximum inundation area of both models in the left panels and the maximum damage of both models in the right panels for the same sub-area. The bars indicate the share (%) of each model from the maximum value of the sub-area, i.e., one of the two bars refers to 100% in each sub-area.

3.5.3 Computational time performance

Besides considering hydrodynamic interactions and a more plausible representation of dike overtopping events leading to inundation, we achieve a considerable reduction of the computational time with RIM2.0. The model runtime for the 1000-year simulation run was nearly halved (from about 60 to 30 days on the GPU-cluster (NVIDIA Tesla K80 with 24 Gigabyte of GDDR5 RAM). The computational load in RIM2.0 caused by the 1D component increased nearly 20-fold compared to RIM1.0. 1D routing in RIM2.0 runs continuously for all flows, whereas RIM1.0 routes only flow exceeding the 2-year discharge. This happens every second year on average. However, due to the lack of hydrodynamic interactions in RIM1.0, the overall larger inundation computations result in a much longer 2D computational time (Figure 3.20).

3.6 Conclusions

In this work, we analyse the effects of hydrodynamic interactions on flood hazard and risk estimates. To this end, two versions of the 1D-2D coupled hydrodynamic diffusive wave models driven by a weather generator and a hydrological catchment model were compared– RIM2.0 considering hydrodynamic interactions, and RIM1.0 considering interactions only within individual sub-basins or river reaches, but not beyond sub-basin boundaries. The analysis of peak water levels, overtopping volumes, inundation areas and depths reveals that the piece-wise routing approach is largely insensitive to overtopping. Unconsidered hydrodynamic interactions and updated upstream boundary conditions at each sub-basin from the hydrological model largely contribute to this effect. The hydrological model is unaware of upstream inundations and provides hydrologically routed flow, including lateral input as if no upstream overtopping has occurred. In this way, the overall mass balance is violated in the piece-wise routing approach and leads to larger inundation areas, depths and damages. At some reaches this overestimation is however partly caused by differences in simulated discharges between two models resulting from calibration and uncertainties in cross-section geometry. We conclude that considerable risk overestimation can be expected when using mosaicking of piece-wise hydrodynamic simulations driven by hydrological models or extreme value statistics as an upstream boundary condition. This approach is still very common for fluvial flood risk assessments. Even if extreme value statistics accounts for multisite spatial dependence structure, hydrodynamic interactions are only accounted for along the reach considered in the hydrodynamic model. Our results for the Rhine basin suggest that interaction aware model produces smaller inundation areas and hence requires half of the computational time for the 2D component compared to the piece-wise approach. The overall risk estimates, i.e. the EAD, derived by the RIM2.0 model version is just about half the value provided by RIM1.0. This highlights the practical relevance of the continuous routing considering hydrodynamic interactions for large-scale flood risk analysis.

Acknowledgements: We would like to acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 676027. This work contributes to the DFG Research Unit FOR 2416 "Space-Time Dynamics of Extreme Floods (SPATE)".

3.7 Supplement

3.7.1 Calibration of the hydrological model SWIM

SWIM has a set of nine most sensitive parameters which need to be estimated. Two parameters are related to the channel routing, three parameters to snowfall and snowmelt processes, two parameters control the subsurface flow contribution to the streamflow, a correction parameter for potential evaporation, and a correction parameter for saturated conductivity. SWIM is calibrated in two phases. In the first phase, we calibrate and validate the model for all sub-regions except Main & Rhine independently (Figure 3.11). In the second phase, the Main & Rhine sub-region is calibrated and validated. It uses the input from other already calibrated sub-regions as tributary inflow. Observed daily streamflow at 79 selected stations is used to calibrate the model based on the modified Nash-Sutcliffe Efficiency NSE_m (Eq. 3.7). The period between 01-01-1981 and 31-12-1989 is used for the calibration, while the period between 01-01-1950 and 03-12-2003 is considered for validation (excluding the calibration period).

3.7.2 Description of the RIM1.0 and RIM2.0 hydrodynamic model discretization and cross-section geometry

Discretization and coupling to the hydrological model SWIM

The discretization of depends mainly on the river topography and SWIM modelsub-basins, Each cross-section in the 1D model is linked to a specific SWIM sub-basin, whereby several cross-sections can be linked to one sub-basin. There can also be sub-basins, which contain no cross-sections, i.e., no river channels are discretized within these sub-basins (Figure 3.12). In the latter case, the flow between sub-basins is routed by SWIM using the Muskingum routing method. For RIM1.0, at the upstream point of each sub-basin, the 2-year return perion discharge is first subtracted from the SWIM flow hydrographs (explicitly representing the bankfull depth), then used as a boundary conditions for the RIM 1D component. For RIM2.0 the flow hydrographs are used as upstream boundary conditions at the most upstream nodes of the discretized river network. Otherwise, the sub-basin discharge is distributed as lateral inflow at every time step among the cross-sections linked to a specific sub-basin.

River cross-section data

Overbank cross-section area

The 1D hydrodynamic model (both RIM1.0 and RIM2.0) requires river cross-section data to describe the river channel geometry. In our approach, the entire river channel and adjacent overbank areas between the dikes or elevated banks are characterized by the river cross-sections and represented in the 1D model domain. Surveyed cross-sections are often not available for large river networks. Furthermore, they usually do not cover the entire floodplain between dikes as represented in the RIM approach. Therefore, RIM relies on cross-sections derived from a digital elevation model. However, the bankfull depth, i.e. river bathymetry below the water level, is not represented in the DEM. The cross-sections for the RIM 1.0 are directly extracted from the DEM. For the RIM 2.0 model, we additionally estimate the bankfull depth using three options based on power-law functions that are detailed below.

First, the overbank component of the river cross-sections, including dike location and elevation along the river network, is derived from the DEM perpendicular to the flow

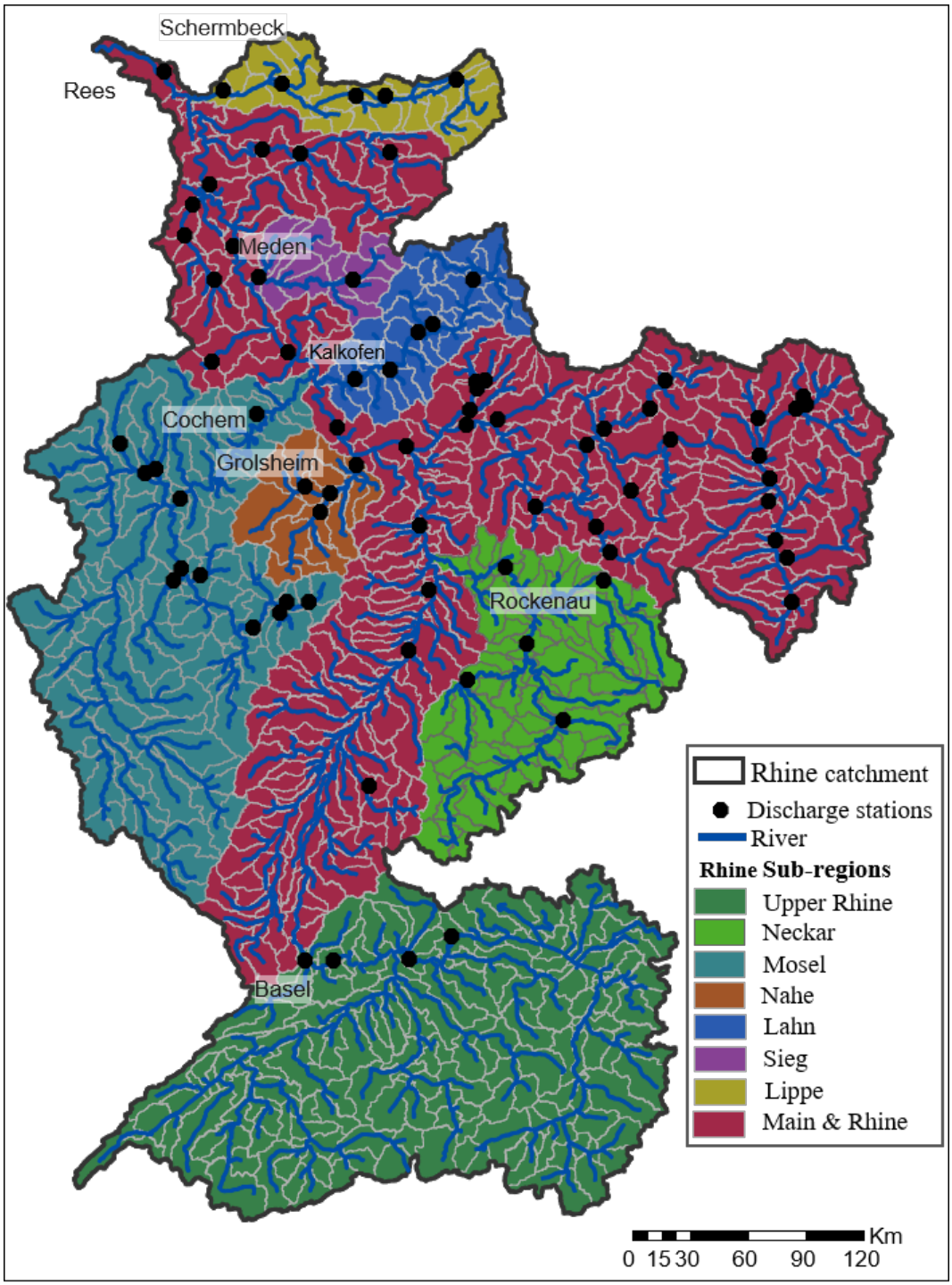

FIGURE 3.11: SWIM model setup with eight calibration sub-regions (each sharing the same parameter values). The river network shown is the result of the catchment delineation and is used in SWIM. It differs from the river network used in RIM (Figure 3.2).

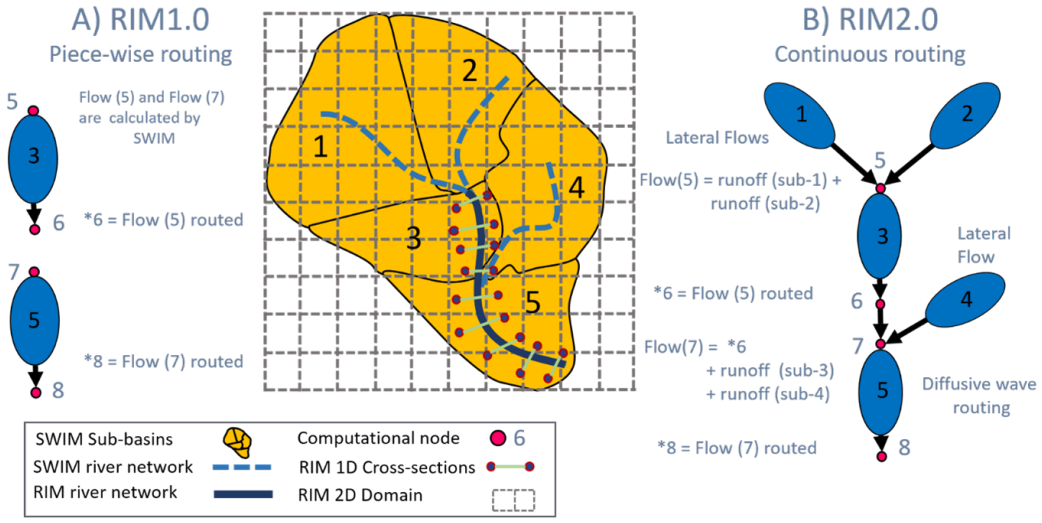

FIGURE 3.12: Difference between the piece-wise routing (A) and continuous routing (B) approaches considered in RIM1.0 and RIM2.0, respectively. Computational nodes are imaginary points at which discharge and water levels are computed. The calculation at points with the notation "Flow (#)" refers to the addition of the upstream flow, while calculation with the notation "*#" refers to routing using the diffusive wave approximation

direction, with the GIS integrated tool HEC-GeoRas. Additional information on dike locations and channel width can be taken from the digital basic landscape model (Base DLM). Subsequently, cross-sections were simplified into 6-point cross-sections by retaining the cross-sectional area (Figure 3.13) using an optimization procedure, i.e. a point is varied vertically on each floodplain side until the optimized vertical location of those points corresponds to the minimal difference between the resulting area and original cross-section area (Figure 3.13). Typically, low-resolution DEMs do not well resolve dike heights and tend to underestimate them. Therefore, dike height derived from the DEM is further calibrated based on a design flow return period. The highest of both estimates is adopted as a crest height in the cross-sections.

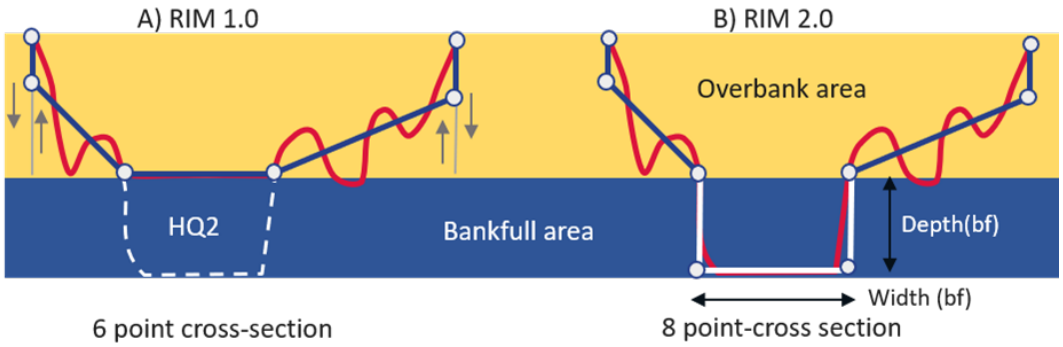

FIGURE 3.13: Difference between cross-section shapes considered in both model versions RIM1.0 (A) and RIM2.0 (B). The overbank area is represented by a 6-point cross-section derived from the 10x10 m DEM for both models. In RIM1.0, the bankfull area is represented by 2- year flow, while in RIM2.0, it is derived statistically from hydraulic geometry relationships

Cross-section bankfull depth in RIM2.0

The bankfull depth for all cross-sections in the RIM2.0 model is derived using hydraulic geometry relationships (Leopold and Maddock, 1953). Three different approaches to derive bankfull depth based on the power-law functions (Eq. 3.8, 3.10 and 3.11) can be used to estimate the bankfull depth at each derived cross-section from the DEM. The first one relates bankfull depth to the 2-year discharge.

$$d_{BF} = \alpha H Q_2^b \tag{3.8}$$

$$W_{BF} = c H Q_2^d \tag{3.9}$$

where W_{BF} is the bankfull width (m), d_{BF} the mean bankfull depth (m), and HQ_2 the 2-year discharge ($\frac{m^3}{s}$).

Based on Rawlins (1995), flows corresponding to 1-2.5 years are representative to bankfull discharge for stable natural channels; therefore, the two year return period discharge (HQ_2) is calculated for all locations with observed discharge by fitting the Gumbel distribution to the annual maximum discharge series at the observational gauges. Since the value of HQ_2 is required at all cross-section locations along the river, not only at gauges, we fit the power-law curve $HQ_2 = \alpha A^\beta$, where A is the upstream drainage area (m^2), to the data via regression and used this curve to estimate HQ_2 at all cross-sections. To estimate the target variable d_{BF}, coefficients a and b need to be approximated. Using Manning's equation, we obtain the bankfull depth (d_{BF}). The bankfull width W_{BF} and slope S_o are estimated from the DEM. Based on the estimated d_{BF} with the very wide cross-sections, the a and b parameters in Eq. 3.8 are obtained by the regression analysis. Using Eq. 3.8, the bankfull depths can directly be derived.

Furthermore, the relationship between channel width and 2-year discharge (Eq. 3.9) was derived through regression analysis to get the parameters c and d; this relationship is not used to calculate the bankfull width as it is extracted from the DEM for all cross-sections, therefore the role of the relationship in Eq. 3.9 is to obtain the parameters c and d and use them in Eq. 3.10, and 3.4 By multiplying Eq. 3.8 and Eq. 3.9, another relationship based on two predictors WBF and HQ_2 can be obtained (Eq. 3.10). Then the third option to derive d_{BF} is to substitute HQ_2 from Eq. 3.8 into Eq. 3.9, resulting in Eq. 3.11

$$d_{BF} = \frac{\alpha c}{W_{BF}} H Q_2^{b+d} \tag{3.10}$$

$$d_{BF} = \alpha \left(\frac{W_{BF}}{c} \right)^{\frac{b}{d}} \tag{3.11}$$

Derived bankfull depths using three different approaches are assessed against the bankfull depths of the surveyed cross-sections at about 1100 locations at the Rhine, Mosel, and Neckar using mean absolute error (MAE) and the mean bias error (MBE). Different Manning's roughness values between 0.025 and 0.05 are considered to estimate the value of d_{BF}, but the differences are found insignificant. Ultimately, the uniform value of n = 0.03 is used across the river network. Considering different options to derive bankfull depth, Eq. 3.12 is found to deliver the best fit with the overall MAE of 1.01 m and MBE of 0.12 m (Figure 3.14). It used measured bankfull width from the DEM as a single predictor. The error in the calculated bankfull depth in the Rhine main channel

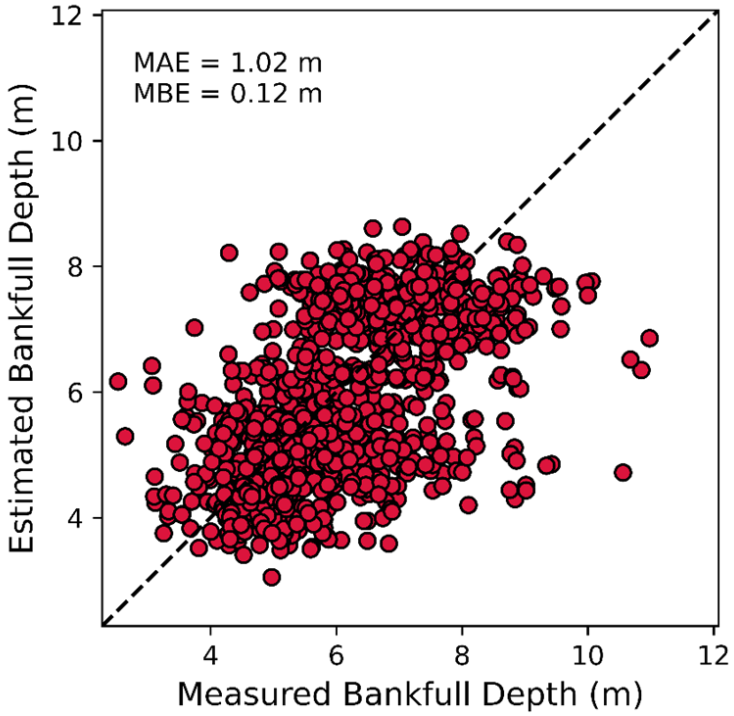

FIGURE 3.14: Comparison of estimated and measured bankfull depth at the locations of the surveyed cross-sections, MAE, and MBE is calculated as measures value-calculated value.

(MAE of 1.02 and MBE of -0.37) differs slightly from the Mosel (MAE of 1.03 and MBE of 0.46) and Neckar (MAE of 0.99 and MBE of 0.63) tributaries.

$$d_{BF} = 0.6354 \left(\frac{W_{BF}}{0.7093} \right)^{0.3961} \tag{3.12}$$

Figure 3.14 shows considerable variability in the estimation of bankfull depth used in RIM2.0 for single cross-sections compared to observations. This highlights the need to calibrate the model by adapting the bankfull depth to match the water level and discharge hydrographs. In order to analyze the effect the cross-section estimation method has on the calculation of bankfull capacity and thus on flow and water level, we compare the bankfull capacity of the cross-sections in RIM1.0 and RIM2.0 at the locations of the gauging stations with the estimates of the 2-year flood based on the gauge records (Figure 3.15). The bankfull discharge (cut-off value) in RIM1.0 is estimated for each sub-basin based on the simulated flows with the SWIM hydrological model. For the purpose of this comparison, the bankfull discharge of RIM2.0 cross-sections is estimated with Manning's equation. The basin average bankfull discharge is then compared to the respective gauge-based estimations. Figure 3.15 shows the difference between the observations and both model estimations. At most locations, the bankfull discharge in the RIM2.0 model (red) marginally differs from the observed 2-year discharge, which indicates a sound estimation of the bankfull-depth and conveyance. Only in the upstream parts of the Neckar and the Nahe tributary, the bankfull conveyance is considerably overestimated. The RIM1.0 tends to slightly overestimate the observed 2-year discharge in the downstream part of the Rhine and in the main tributary.

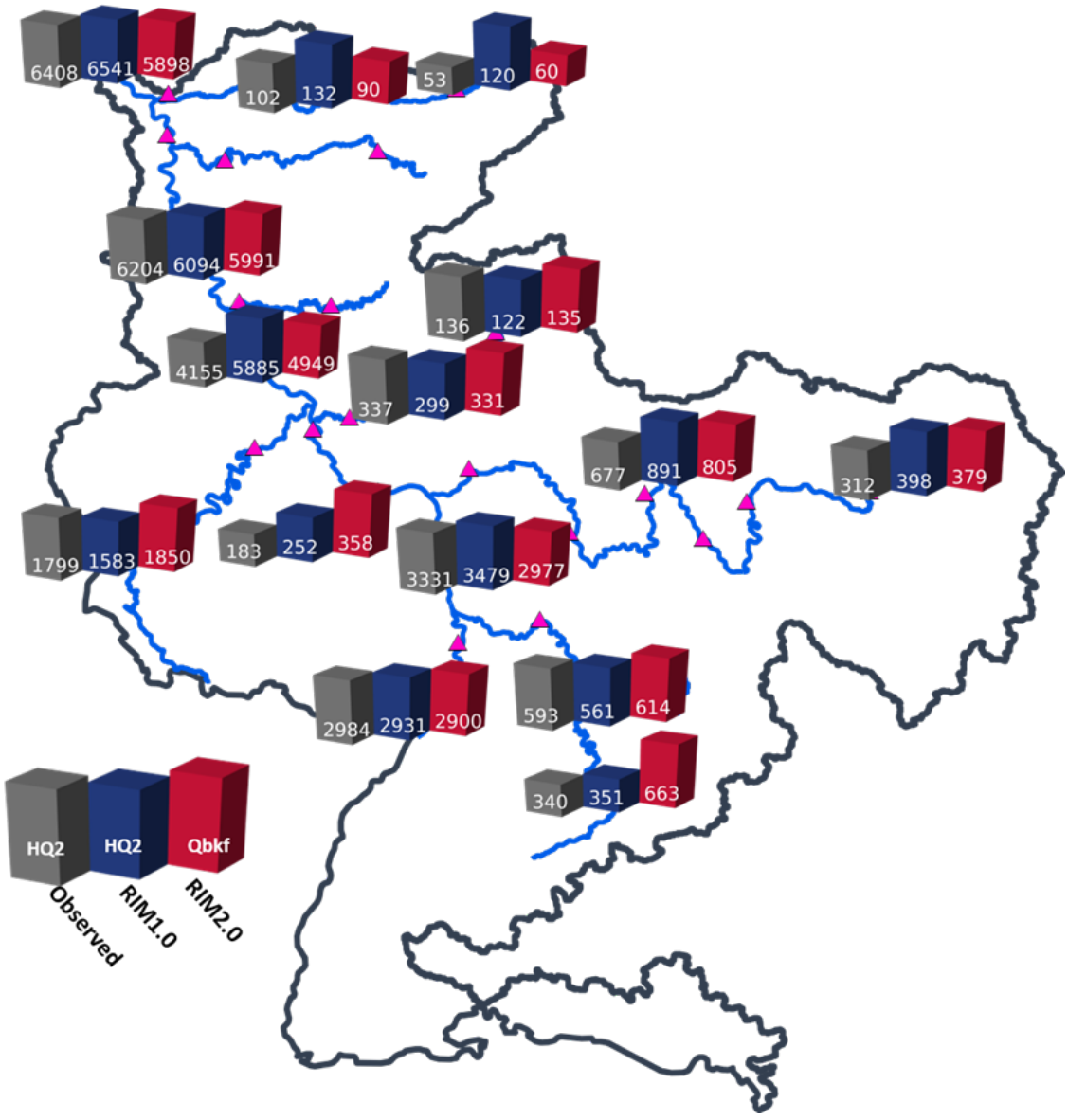

FIGURE 3.15: Comparison between the 2-year return period discharge derived from gauge observations (grey column), cut-off value (HQ_2) in RIM1.0 (blue column), and estimated bankfull cross-section capacity in RIM2.0 (red column) in $\frac{m^3}{s}$. The bar height is related to the maximum value at each location. Bars at one location cannot be compared with those at other locations.

3.7.3 Performance of SWIM, RIM1.0, and RIM2.0 flood routing

Performance of discharge simulation

Figure 3.16 presents a comparison of a few selected flood events exceeding 2-year bankfull discharge in the observation period at 6 gauges (for the locations of gauges see Figure 3.2) in the Rhine basin, including the comparison with RIM1.0, where the cut-off value is represented by the dotted horizontal blue line. These peaks belong to a major flood event which hit the Rhine catchment in December 1993 - January 1994. These gauges are selected to cover the main channel and major tributaries, including potential risk hot spots at the Lower Rhine (Cologne, Duesseldorf). The inspection of the hydrographs in Figure 3.16 reveals a very close performance of both model versions at the Steinbach, Lauffen, and Leun and Cochem gauges, while the performance at Cologne and Duesseldorf slightly differs, where SWIM and RIM1.0 better match the observed peak flows.

The comparison of the modelled and observed hydrographs reveals that the performance of the RIM models with respect to observations is affected by the input from SWIM. For example, at gauge Leun, SWIM underestimates the peak flows, and this error inherently appears in the RIM results. In general, the comparison reveals that RIM2.0 mostly captures the peak magnitudes, but shows a slight shift in phase of the hydrographs at nearly all gauges. The peaks come earlier and appear to be less attenuated than in other models, and this explains the poorer performance compared to SWIM routing. The performance of RIM2.0 compared to observations tends to be worse compared to the performance of the SWIM routing. This might be due to the fact that the calibration of the cross-section bankfull depth changes the bed slope between cross-sections. This controls the water levels and, consequently, discharge calculations. Hence, a certain trade-off is expected between the performance with respect to the water levels and discharge hydrographs. In other words, the performance with respect to discharge can deteriorate compared to SWIM in order to improve the simulation of water levels.

Figure 3.17 presents a comparison between the maximum annual discharges in the validation period modelled by RIM1.0, RIM2.0, and SWIM vs observed peak flow of the corresponding flood events. It can be noticed that RIM1.0 does not simulate flows below the HQ_2 threshold. Therefore, the number of simulated peaks by the RIM1.0 is roughly half of those delivered by the other two models. The points for all models are scattered along the 45-degree line at most of the selected gauges, which indicates an overall similar model performance with regards to peak flows. At gauge Steinbach, RIM1.0, RIM2.0, and SWIM perform very similar but slightly overestimate the peaks. At Lauffen, RIM2.0 overestimates a few high flow peaks. At gauges Duesseldorf and Cologne, RIM2.0 slightly underestimates the peaks above 8000 $\frac{m^3}{s}$, whereas RIM1.0 and SWIM show very good performance. As discussed above, the representation of the cross-sectional geometry in the lowland river reaches might be inadequate and requires further improvement in the future.

Evaluation of RIM1.0 and RIM2.0 with regards to water levels To benchmark the performance of RIM1.0 and RIM2.0, we analyze the water level simulations at 6 selected gauges with available water level observations for the major flood event in 1993 analogously to the analysis of discharge hydrographs (Figure 3.18). RIM1.0 and RIM2.0 are mostly consistent in the simulation of water level dynamics and match the peak levels similarly well (or poor). As for discharge, RIM2.0 water level hydrographs are slightly shifted in time compared to RIM 1.0 and observed hydrographs, though peak levels are met mostly well. The constant water level in RIM 1.0 (horizontal line) is

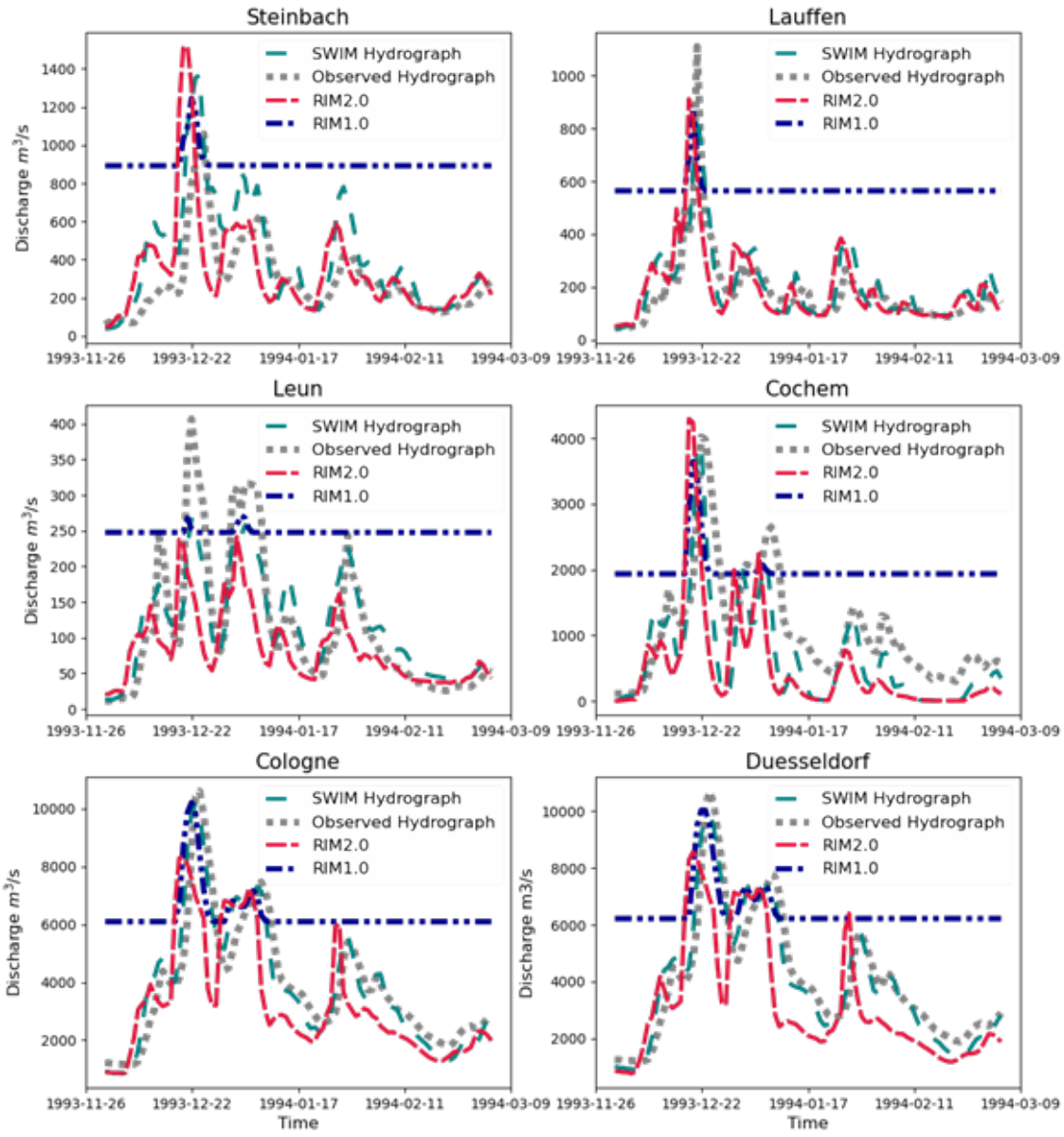

FIGURE 3.16: Comparison of discharge hydrographs between RIM1.0, RIM2.0, SWIM, and observations at selected gauges. The selected dates show the highest peak of the time series in the observational period. The horizontal blue line represents the cut off value (HQ_2) in RIM1.0 and is not simulated.

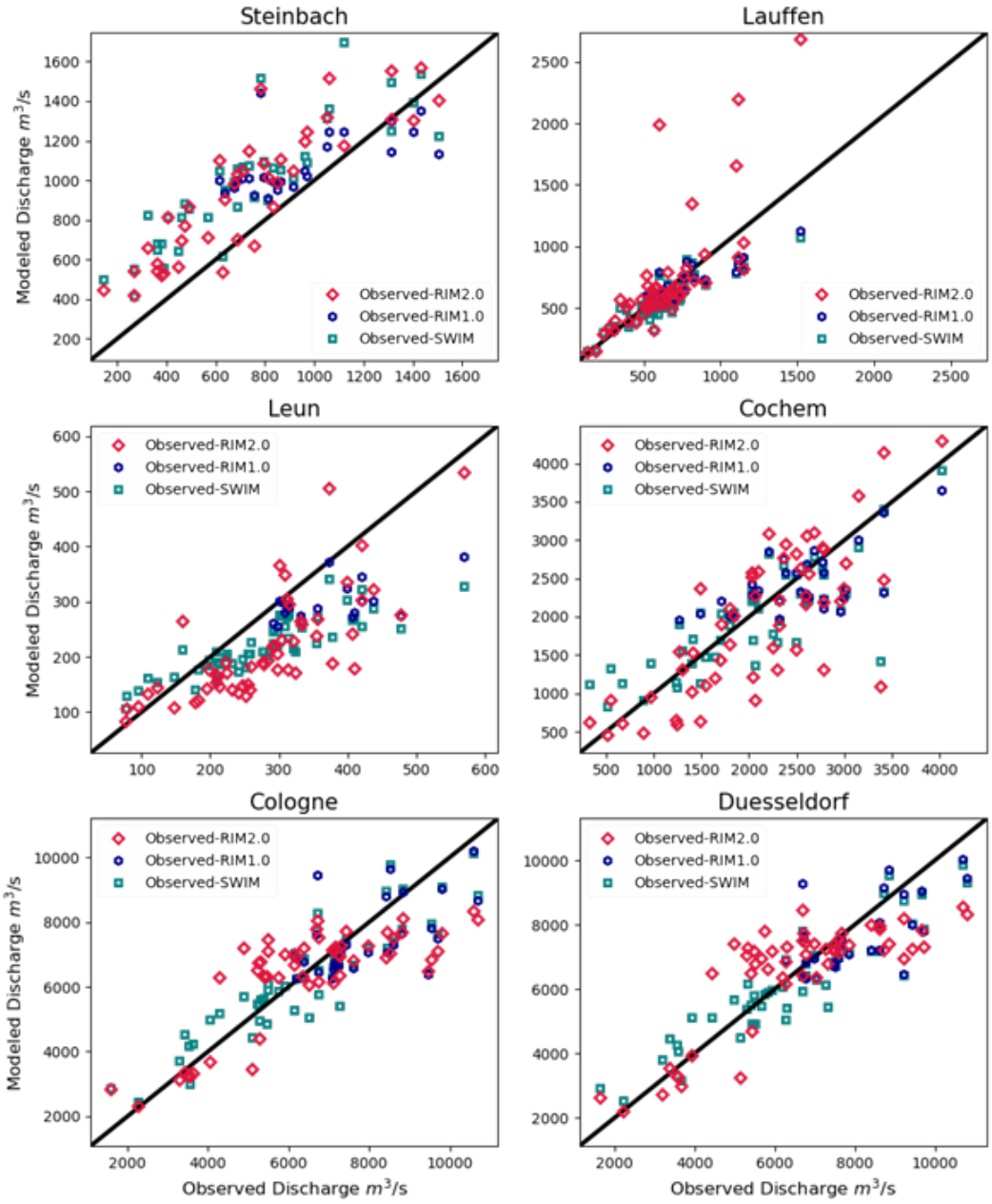

FIGURE 3.17: Comparison of the maximum annual modelled and observed discharge peaks in the historical period from 1950 till 2003.

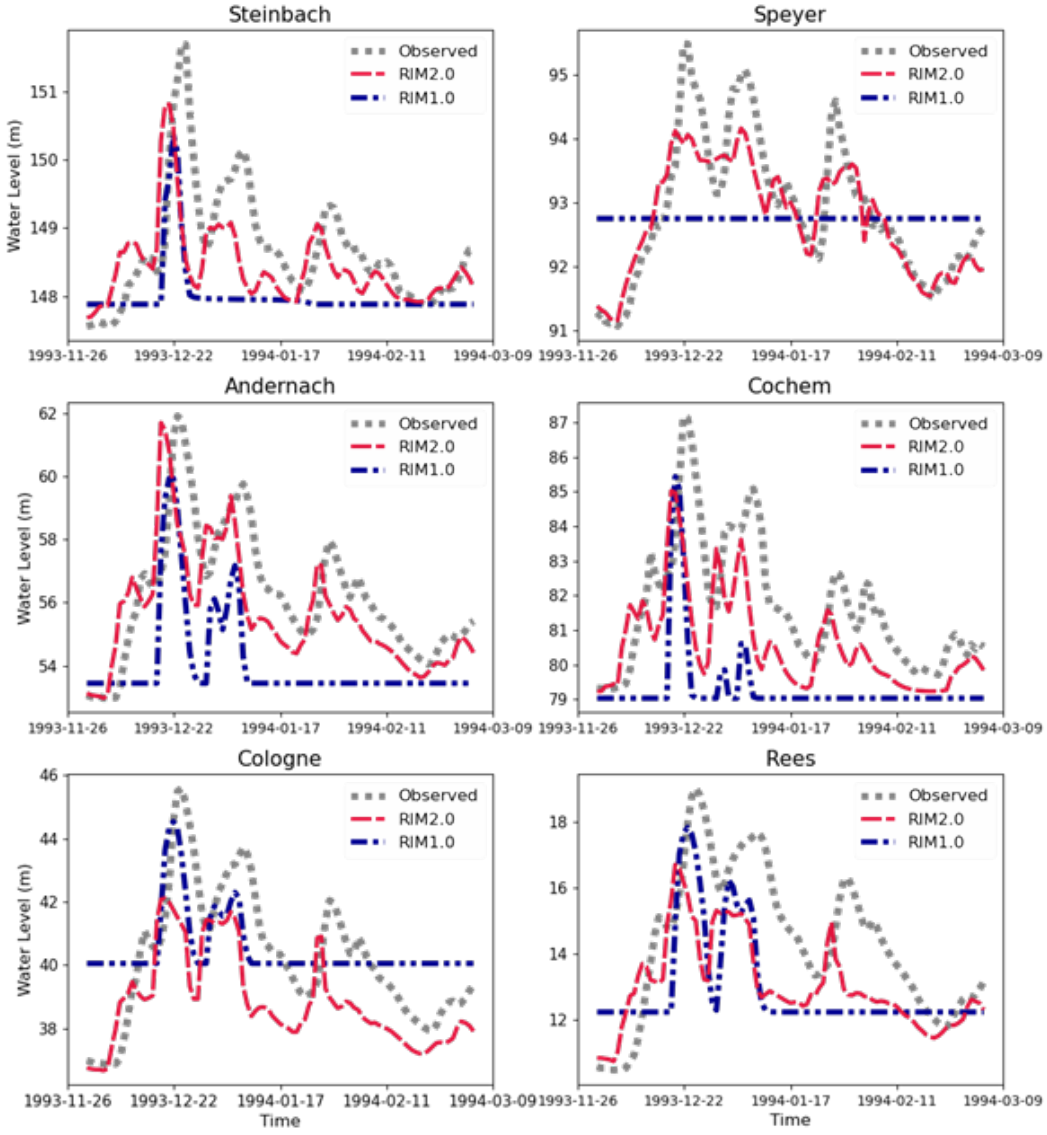

FIGURE 3.18: Comparison of simulated water levels between RIM1.0 and RIM2.0 for selected historical flood events in December 1993 – January 1994 at 6 selected gauges.

the water level corresponding to the cut off value (HQ_2). At the gauge station Speyer, water levels simulated by RIM2.0 show that both high and low flows are well presented in addition to the timing of the flood wave, with 0.48 m MAE and -0.05 m MBE (overestimation). In contrast, RIM1.0 misses the event, which indicates a problem in RIM1.0 for estimation bankfull capacity. At the gauge Andernach, both models predict a slightly earlier peak compared to the observed data, but RIM2.0 matches the observed peak water levels much better than RIM1.0 At gages Cologne and Rees, RIM2.0 simulates slightly smaller water levels than RIM1.0 for some peaks, whereas other peaks are well attained by RIM2.0 and are entirely missed by the RIM1.0 model. At the gauge Cochem, RIM2.0 underestimates the peaks with MBE of 0.19 and MAE of 0.46. The degradation of the water level performance at Cochem indicates a problem in the bankfull depth estimation at the cross-sections in this reach and should be revised in future. The gauge Steinbach in the middle of the river Main shows that RIM1.0 underestimates smaller events, while RIM2.0 manages to capture the observed peaks and delivers good performance with MBE of -0.08 m and MAE of 0.28 m.

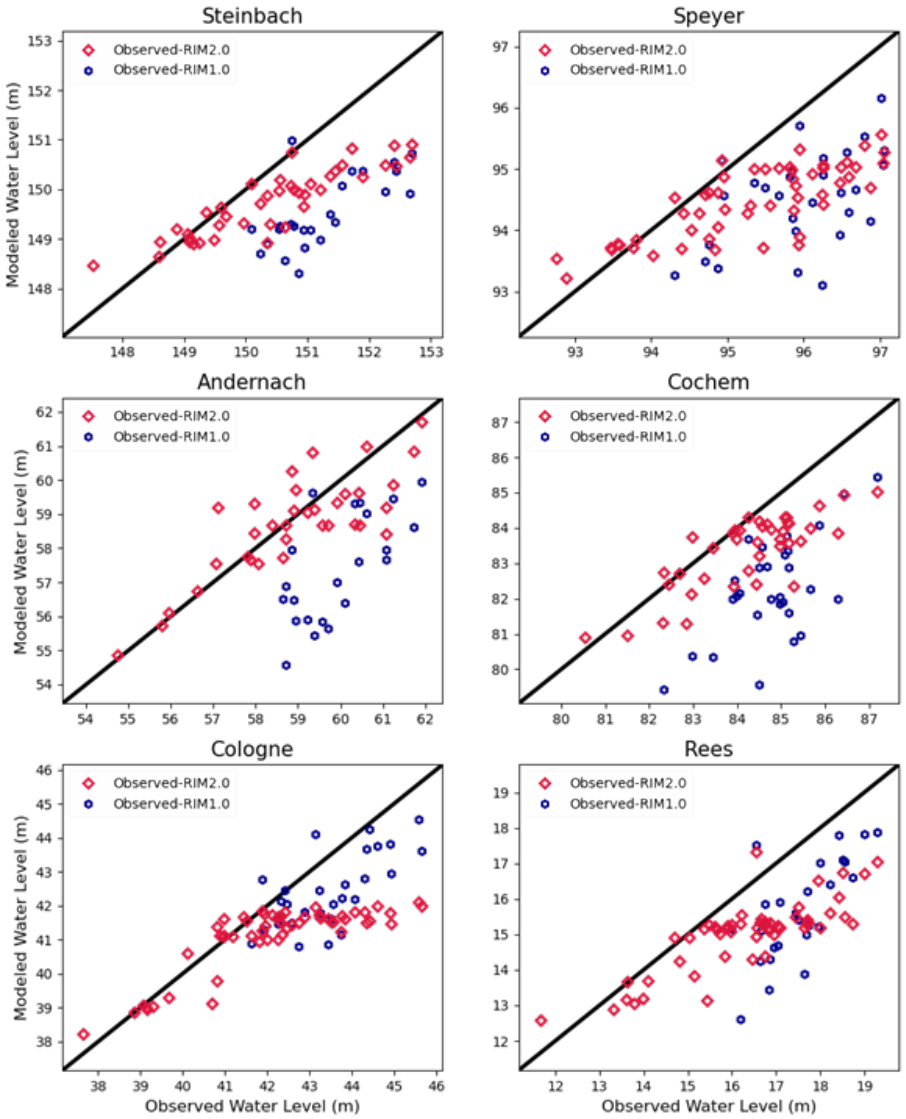

FIGURE 3.19: Comparison of the maximum annual modelled and observed highest water level in the historical period from 1950 till 2003

Figure 3.19 presents a comparison between the water levels modelled by RIM1.0 and RIM2.0 for flood events corresponding to events with the maximum annual observed water level peaks. Similar to the simulated discharge by RIM1.0 and RIM2.0, the simulated water levels by both models have an overall similar model performance but are underestimated at most stations, as shown in Figure 3.19. RIM2.0 seems to have problems replicating the peak water levels at Cologne. Nonetheless, the performance of both models is very similar at Duesseldorf (not shown). This highlights the effect of the cross-sections on water levels estimates and stresses the importance of the robust estimation of the river profile geometry.

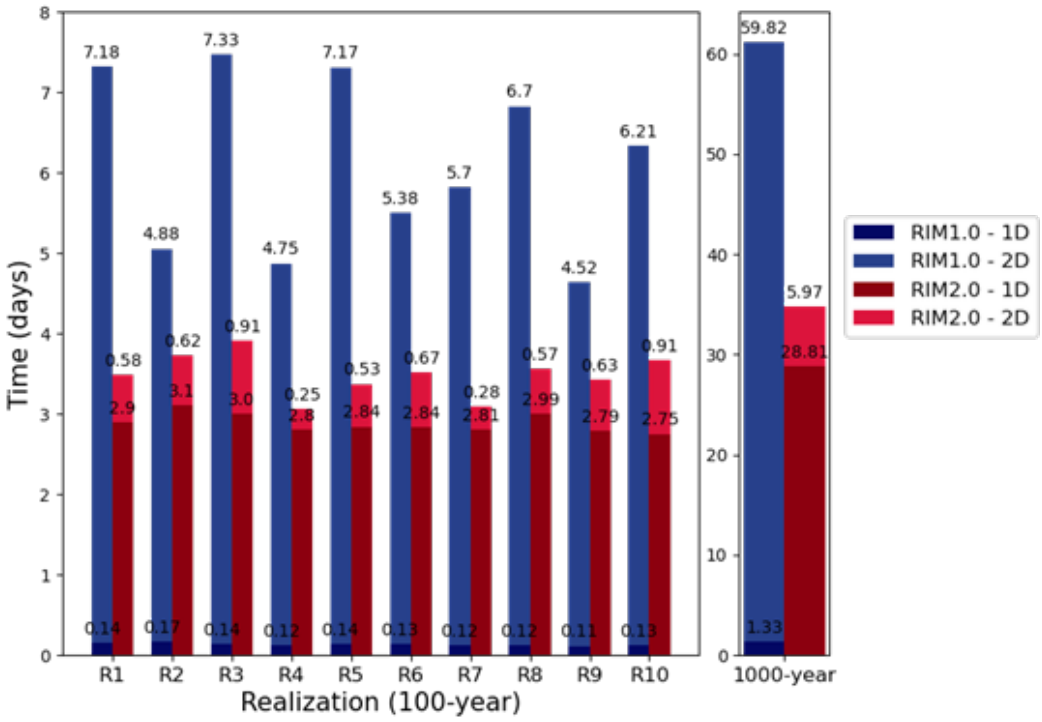

FIGURE 3.20: Computational time required by RIM1.0 and RIM2.0 for simulating 1000 years. The 1000-years period is divided into 10 realisations, each modelled using one NVIDIA Tesla K80 card.

Chapter 4

A data-mining approach towards damage modelling for *El Niño* events in Peru

4.1 Abstract

Compound natural hazards like *El Niño* events cause high damage to society, which to manage requires reliable risk assessments. Damage modelling is a prerequisite for quantitative risk estimations, yet many procedures still rely on expert knowledge, and empirical studies investigating damage from compound natural hazards hardly exist. A nationwide building survey in Peru after the *El Niño* event 2017 – which caused intense rainfall, ponding water, flash floods and landslides – enables us to apply data-mining methods for statistical groundwork, using explanatory features generated from remote sensing products and open data. We separate regions of different dominant characteristics through unsupervised clustering, and investigate feature importance rankings for classifying damage via supervised machine learning. Besides the expected effect of precipitation, the classification algorithms select the topographic wetness index as most important feature, especially in low elevation areas. The slope length and steepness factor ranks high for mountains and canyons. Partial dependence plots further hint at amplified vulnerability in rural areas. An example of an empirical damage probability map, developed with a random forest model, is provided to demonstrate the technical feasibility.

4.2 Introduction

El Niño events are compound events, comprising torrential rainfall, fluvial and flash floods, mudslides, and other processes difficult to disentangle. Relevant for society is the total damage resulting from consecutive or cascading hazards during several months. Modelling studies and measurements still do not provide a clear image about the effect of climate change on the *El Niño* Southern Oscillation, but the general expectation is that more extreme states will occur more frequently (Cai et al., 2015, 2018; Freund et al., 2019). For this reason, a stronger shift from the "emergency response"-mindset towards integrated disaster risk management would be advantageous for Latin America. This shift requires risk analyses including damage models for cost-benefit analysis, and for the planning of needed relief aid. However, to the best of our knowledge, damage models for such compound events are still non-existent, and even statistical analyses of damage processes are scarce (compare Gerl et al., 2016).

Methodologies in damage modelling range from very coarse national average curves with focus on economic damage (Jongman et al., 2012) to detailed local engineering approaches with focus on structural failure of buildings (Risi et al., 2013; Jalayer et al., 2016). Engineers have defined structural damage grades of buildings for various kinds of hazards, e.g., floods (Ester et al., 1996), tsunamis (Suppasri et al., 2012), or debris (Jenkins et al., 2015). These damage grades usually imply certain types of observable characteristics, like cracks in the wall or dented windows. Previous attempts to predict structural damage from empirical surveys had limited explanatory power due to the low number of samples used, especially for buildings with high damage grades (Laudan et al., 2017). For Peru, the most comprehensive published investigations were made about flash floods in Arequipa (Thouret et al., 2013, 2014; Ettinger et al., 2016). Although the cited studies in Arequipa included environmental covariates like distance to the channel and soil impermeability, most other studies typically use a single intensity metric, like water depth for floods, which is not straight-forward for compound events like *El Niño* situations.

Topographic indicators, on the other hand, have been successfully used to predict flood and landslide hazard, (e.g. Cao et al., 2016; Chapi et al., 2017; Pour and Hashim, 2017; Tehrany et al., 2018). For example the Topographic Wetness Index (TWI) and Stream Power Index (SPI) are measures of potential water accumulation and erosive power of a stream (Moore et al., 1991; Quinn et al., 1991). Curvature is in the focus of recently published geomorphological studies trying to explain the occurrence of surface water – potentially due to correlations of surface morphology to underlying bedrock stress (Clair et al., 2015; Moon et al., 2017; Prancevic and Kirchner, 2019). The slope length and steepness factor (Erosion-LS) is measure for erosion potential, which is commonly used in the Universal Soil Loss Equation and has been identified as the most important parameter in respect to quantifying erosion at least in Europe (Panagos et al., 2015). For a multi-hazard risk context, (Aksha et al., 2020) summarized recent literature and concluded that a geospatial approach can be more realistic than a chain of dynamic models, which still hinge on availability of data and model scenarios. While academic literature is divided in specific domains for individual hazards, the multi-hazard aspect is of practical relevance to spatial planners, who are responsible for an area in its entirety (Greiving et al., 2006). However, (Aksha et al., 2020) and similar studies use expert-assigned weights for risk calculations (via an analytic hierarchy process). We argue that this procedure should eventually be replaced by a sound empirical basis. Therefore we directly train models on observed damage, which could be called a geoscientific take on damage modelling.

Data-driven approaches, as opposed to analytical, are recently on the rise due to increased availability of computational resources and suitable datasets. Especially satellite remote sensing and voluntary geographic information now enable the application of such methods in formerly data-scarce regions, which are often of interest for environmental studies (e.g., remote wilderness) as well as for risk studies (e.g., high vulnerability of economically poor regions) (Malgwi et al., 2019). Well-known examples of novel data sources include the Copernicus mission (Aschbacher and Milagro-Pérez, 2012) and OpenStreetMap (OSM) project (OpenStreetMap contributors, 2017). Beyond these, there are other more specific global datasets: For example, GlobalSurfaceWater (GSW) is an earth observation product based on the entire Landsat archive, which identifies the maximum extent, seasonality, and recurrence of water bodies since 1984 (Pekel et al., 2016). Globally consistent gridded soil information is provided by the SoilGrids database (Hengl et al., 2017). A novel by-product of the TanDEM mission makes a global forest/non-forest (FNF) layer available (Martone et al., 2018).

WorldPop (Tatem, 2017) is another gridded statistical product, providing information on population density, adjusted to official country census data. The Global Urban Footprint (GUF) (Esch et al., 2012) is a binary layer of detected settlements from satellite imagery, distributed by the German Aerospace Centre (DLR). The Tropical Rainfall Measurement Mission (TRMM) and the follow-on Global Precipitation Monitoring (GPM) provide a satellite-based global precipitation product with up to 3h temporal resolution. (Erazo et al., 2018) compared different rainfall datasets for Ecuador and the Pacific Slope, and concluded that TRMM represented *El Niño* events better than the products from the Climate Research Unit (CRU), Global Precipitation Climatology Centre (GPCC), and European Centre for Medium-Range Weather Forecasts (ERA-Interim).

Drawing upon a large-scale field survey by Peruvian authorities in the aftermath of the coastal *El Niño* 2017 we aim to identify drivers of damage based on open geodata and machine learning (ML) methods. As opposed to traditional statistics, which typically suffer from strong assumptions and require manual parametrization, ML algorithms can deal with complex non-linear and interacting relationships. In the field of damage modelling, ML has been applied to individual hazards on event-level (e.g. Terti et al., 2017; Alipour et al., 2020) and on object-level (e.g. Rözer et al., 2019; Sieg et al., 2017; Wagenaar et al., 2017; Nafari et al., 2016). Most of the cited authors used tree-based algorithms in supervised mode. Although these methods are often called black-boxes, there is currently a trend towards more interpretable ML, even to the point of making causal interpretations (Zhao and Hastie, 2019). The main objective of this study is to test the potential and limitations of such a data-mining approach, mainly based on remote-sensing products, to characterise damage processes during *El Niño* events. A fitted damage model with sufficient skill can be used to generate empirical damage probability maps. This paper is structured in the style of a ML project: data preparation and feature engineering, feature selection, unsupervised clustering, supervised classification, model inspection via feature importance ranking and partial dependence plots. An example of a classified damage probability map is presented to prove the technical feasibility.

4.2.1 Study area and event 2017

Peru is located in the central and western part of South America and has an estimated area of 1,285,215.60 km2. This is distributed in three macro geographical areas known as the coast, mountains and jungle. Although the territory of the coast represents only approx. 10.6% of the national territory, more than 58% of the country's population are concentrated there (INEI, 2018). Being a tropical country, the presence of the Andes Mountains significantly influences the climatology and morphology of the territory. Due to this, the coast of Peru is a mainly arid area, with little precipitation and vegetation, whose geography is abruptly changing towards the Peruvian highlands that reach altitudes of more than 6700 m.a.s.l. East of the Andes is the Peruvian jungle, a territory mainly flat where the Amazon river originates. Between 2003 and 2018, the National Institute of Civil Defence (INDECI) has registered around 75,000 emergencies, with more than 53,000 occurring due to natural hazards and 52% of these were classified as mass movements, floods and other dangers triggered by heavy rains (INDECI, 2019). According to the susceptibility map of the Peruvian territory to mass movements of the Geological, Mining and Metallurgical Institute of Peru (INGEMMET), surfaces like hillsides with slopes greater than 35°, fault zones, rock masses with moderate to intensely weathered erosion and unconsolidated surface deposits are especially susceptible for mass movement (Villacorta et al., 2012). On average, over the last 15 years there have been 4,800 emergencies per year, excluding extraordinary years. However, in

2017, the year of *El Niño* phenomenon, the amount of emergencies increased by about 57%. The Multisectoral Commission in Charge of the National Study of the *El Niño* Phenomenon in Peru (ENFEN) has classified the phenomenon that occurred in 2017 as the third most intense in the last hundred years. However, its manifestation was abrupt and surprising, since the monthly anomaly in December 2016 was estimated by the Oceanic Niño Index of -0.6, indicated for this area the presence of a weak magnitude *La Niña* phenomenon. (ENFEN, 2017; NOAA, 2020). The event was compared to the coastal *El Niño* of 1925 (Son et al., 2019). The impact statistics of the period of the rainy season 2016–2017 in which the *El Niño* phenomenon occurred, shows that the impact on the population and public infrastructure was severe, registering a total of 283,137 inhabitants who lost their homes and/or their livelihoods, due to which they were considered as victims. 65,942 homes collapsed or were uninhabitable due to severe damage (a map is provided in the result section) and around 51,854 hectares of crop were lost. There was a significant impact on public infrastructure, registering 456 educational institutions and 71 collapsed or uninhabitable health centres. From the 25 regions of the country, nine were significantly affected. Severe damage occurred mainly in the coastal regions Tumbes, Piura, Lambayeque, La Libertad, Ancash, Lima, Ica, and Arequipa but also in the Cajamarca region located in the country's highlands. The regions with the largest number of affected population were Piura (98,894 people) and La Libertad (78,978 people) (INDECI, 2019).

4.3 Data and methods

4.3.1 Data sources and feature engineering

In the framework of the actions ordered by the Peruvian State in the presence of the *El Niño* phenomenon 2017, the Organisation for the Formalization of Informal Property of Peru (COFOPRI) received the assignment of carrying out "... the gathering of information about the affected, collapsed and uninhabitable homes and the cadastre of damages of the areas declared in emergency, and identifying the owners and/or occupants of the homes and other properties in said areas" (Emergency Decree No. 004-2017, (Government of Peru, 2017). In this framework, trained personal by COFOPRI collected information about the affected homes nationwide. While other entities also recorded damage, this data of COFOPRI is the only one available with precise geolocation, which is necessary for this study. In addition to data on the property conditions and general characteristics of the dwelling, the physical damage to the houses was identified, establishing four ordinal classes: mildly affected, i.e., non-structural damage to doors, windows, or sanitation (D1); moderately affected, i.e., structural damage which is repairable, and the building is still habitable (D2); uninhabitable, i.e., structural damage which is repairable, but currently presents a threat to the inhabitants (D3); and collapsed, i.e., irreparable damage (D4). The available raw data includes only coordinates and damage according to these categories. The database was analysed, ruling out unaffected buildings and inconsistent data, resulting in a sample of 119,675 buildings that have been used in the study. At the closing date of this document, i.e., in April 2020 the damage data was available on the COFOPRI website (Geollaqta) for download in csv format (comma delimited) which can be easily transformed into vector format (shapefile or geojson) using any geographic information system. To frame the problem as an ML task we used feature engineering, i.e., we derived predictor variables (features, Table 4.1) from remote sensing products and open data, and assigned these features in numerical vector format to the target classes (in our case the observed damage). Candidate features were taken from literature, as cited in the introduction, and

theoretical considerations. All topographic features in this study are based on the MERIT elevation model (Yamazaki et al., 2017). A binary layer of steep terrain was computed by applying Otsu's threshold (Otsu, 1979) to a smoothened slope raster. The algorithm automatically maximizes the between-class variance from a bimodal histogram. We used a focal mean window of 5 × 5 cells, so that small patches of flat regions within the mountains mainly fall into the steep category as well. This layer was then used to derive the Euclidean distance to mountains (i.e., steep terrain).

GSW extent and OSM waterways were merged to obtain a product of potential surface water occurrence. We observed that OSM provides surprisingly comprehensive information about river courses and even underground channels. While GSW is especially helpful for large inaccessible areas, OSM appears more accurate for urban areas, which cannot be properly observed by Landsat. This merged layer was used as input for GRASS r.stream.distance to derive height above and distance to water, as alternative to the Height Above Nearest Drainage (HAND) raster provided by (Yamazaki et al., 2019). Since SoilGrids is a statistical product, the variation in depth appears quite smooth and for our purpose not relevant. We therefore decided to use the grain size fractions from the top layer for calculating the erodibility (Erosion-K). Pre-event mosaics of the Bare Soil Index (BSI) and Normalized Difference Vegetation Index (NDVI) from Seninel-2 data have been processed in Google Earth Engine (Gorelick et al., 2017). Both indicators are commonly used spectral ratios, yielding high values for bare soil or vegetated regions, respectively. Rather than using the values at the location of the buildings, we suppose that such indices gain more meaning in a natural hazard context when weighted along the flow accumulation raster (i.e., using the features as weight in GRASS r.accumulate, and dividing the weighted accumulation by the unweighted accumulation). These engineered features, denoted as Erosion-K_acc, BSI_acc, and NDVI_acc are expected to provide information on erosion processes and sediment load. FNF and GUF are binary remote sensing products, mapping forest and urban areas, respectively. These binary layers naturally contain some misclassifications. We used a focal mean within 11×11 cells (roughly $1\,km^2$) as proxy for forest density and urbanity, i.e., a cell has the value 1 when all cells within this area are classified as forest/urban. While WorldPop is a product on population density, the GUF_density can be interpreted as urbanity: high values identify urban centres, medium values suburbs or small towns, and low values rural areas. Relatively high values of population per pixel are also possible in rural areas (i.e., compact satellite towns). All features were prepared in the form of stacked rasters, on the resolution of the DEM ($90\,m$). Resampling, where necessary, was done by bilinear interpolation. Only WorldPop, which represents counts, was resampled manually by disaggregation and re-aggregation to preserve the correct total amount of people, and co-registered by nearest neighbour. In cases where rasters contained missing values (e.g., clouded pixels in BSI), holes were interpolated by GDAL.fillnodata – this should not have any effect on the results presented here, except for the classified map, since the statistical analysis in this paper is based on extracted points at the location of buildings, not on the entire rasters. Extraction of points was done using the velox library in R. Values were then rescaled to the range of 0-1 and very steep distributions were log-transformed for better performance of the distance-based methods. Google Earth Engine was used to access, process, and download Sentinel-2 spectral ratios, as well as TRMM precipitation data. Demanding computations were performed on a Windows Server with 72 CPUs and 256 GB RAM. Data analysis and visualization was carried out on a regular Lenovo ThinkPad laptop.

TABLE 4.1: Candidate features.

Feature	Explanation	Data Source (Resolution)	Reference	Software
Altitude	Absolute elevation above sea level	MERIT DEM (90 m)	Yamazaki et al. (2017)	–
Slope	Derivative of elevation in steepest direction			R spatialEco
Slope_focal	Focal mean of Slope			
Curvature	Second derivative of elevation. Total curvature contains planform and profile directions		Zevenbergen and Thorne (1987)	R spatialEco
TRI	Terrain Ruggedness Index. Information on heterogeneity of elevation		Riley et al. (1999)	R spatialEco
TPI x6	Topographic Position Index. Calculated as elevation difference between a cell and the average value in a defined neighbourhood. Single-scale TPI identifies local hills or depressions. Multi-scale TPI can be used to classify complex landscapes. We used 6 scales: 90, 180, 270, 540, 990, 1980 [m]		Reu et al. (2013) Weiss (2001)	R spatialEco
SPI	Stream Power Index. Represents the force of a flow, with high values in steep areas		Moore et al. (1991)	GRASS r.watershed
TWI	Topographic Wetness Index. Represents potential accumulation of water, high values in flat areas		Quinn et al. (1991)	GRASS r.watershed
TWI_focal	Focal mean of TWI			R raster
Erosion-LS	Slope length and steepness (LS factor) of the Universal Soil Loss Equation			GRASS r.watershed
DistToMountains	Distance to steep terrain. Computed as Euclidean distance to a binary layer from Otsu's threshold on Slope_focal		Otsu (1979)	Python scikit-image, GDAL proximity
HAND	Height Above Nearest Drainage. Normalization of the terrain, where streams are set to 0 and height difference is computed along the flow direction	MERIT Hydro (90 m)	Yamazaki et al. (2019)	–
HANW	Height Above Nearest Water. Same as HAND for different stream raster from GSW and OSM	GSW + OSM (30 m, vector)	Rennó et al. (2008)	GRASS r.stream.distance
DistToWater	Distance to water, along the flow direction			
GSW_occ	Global Surface Water occurrence	Global Surface Water (GSW) (30 m)	Pekel et al. (2016)	–
BSI	Bare Soil Index. Spectral ratio which identifies bare soil	Sentinel-2 Multi-Spectral Instrument (10 m)		Google Earth Engine
BSI_acc	BSI weighted along flow accumulation			GRASS r.accumulate
NDVI	Normalized Difference Vegetation Index. Spectral ratio which identifies vegetation			
NDVI_acc	NDVI weighted along flow accumulation			
Clay	Percent clay in topsoil	SoilGrids (250 m)		–
Erosion-K	Erodibility (K factor) of the Universal Soil Loss Equation			GRASS r.uslek
Erosion-K_acc	Erosion-K weighted along flow accumulation			GRASS r.accumulate
FNF_density	Forest density. Maximum when all FNF cells within an 11x11 window are classified as forest	TanDEM Forest / Non-Forest (FNF) (50 m)	Martone et al. (2018)	R raster
GUF_density	Urbanity. Maximum when all GUF cells within an 11x11 window are classified as urban	Global Urban Footprint (GUF) (84 m)	Esch et al. (2012)	R raster
Worldpop	Population density in people per pixel	WorldPop (100 m)	Tatem (2017)	R raster, GDAL warp
DistToRoads	Euclidean distance to roads	OpenStreetMap (OSM) (vector)	OpenStreetMap contributors (2017)	GDAL proximity
TRMM_max	Rainfall 3-hour maximum	Tropical Rainfall Measurement Mission (TRMM) (30 km)		Google Earth Engine
TRMM_sum	Rainfall sum from January to April 2017			

Σ 33

Feature selection

For the purpose of clustering, it is often recommended to restrict the number of features to the necessary minimum, since clustering algorithms tend to be sensitive to duplicated information content, noise, and shape of the feature value distributions (Milligan, 1996). We chose the rainfall, information about the streams from HAND, potential accumulation from TWI, soil from Erosion-K, vegetation from NDVI, and settlements from GUF_density. TPI was included on small, medium, and large scale, together with altitude and slope, for delineating geomorphological regions as originally presented by (Weiss, 2001). Collinear features are not necessarily a problem for statistical learning algorithms, if the aim is simply to produce the most accurate predictive model. However, if we aim at estimating feature importance these interactions may lead to wrong conclusions, since the information content of two collinear features is still available to the algorithm even when one is removed – therefore judging both as not important. For inspection we plotted a dendrogram of spearman rank correlation (Figure 4.1) between the features, which shows how closely the features are hierarchically related to each other. We selected one feature per group at threshold 1 – with two exceptions: NDVI was discarded for consistency with BSI and Erosion-K, where only the weighted versions were used. Rainfall sum and maximum were both kept, despite some correlation, since we consider this to be different information and wanted to investigate the potential of TRMM for this purpose. The selection resulted in 20 features used for classification.

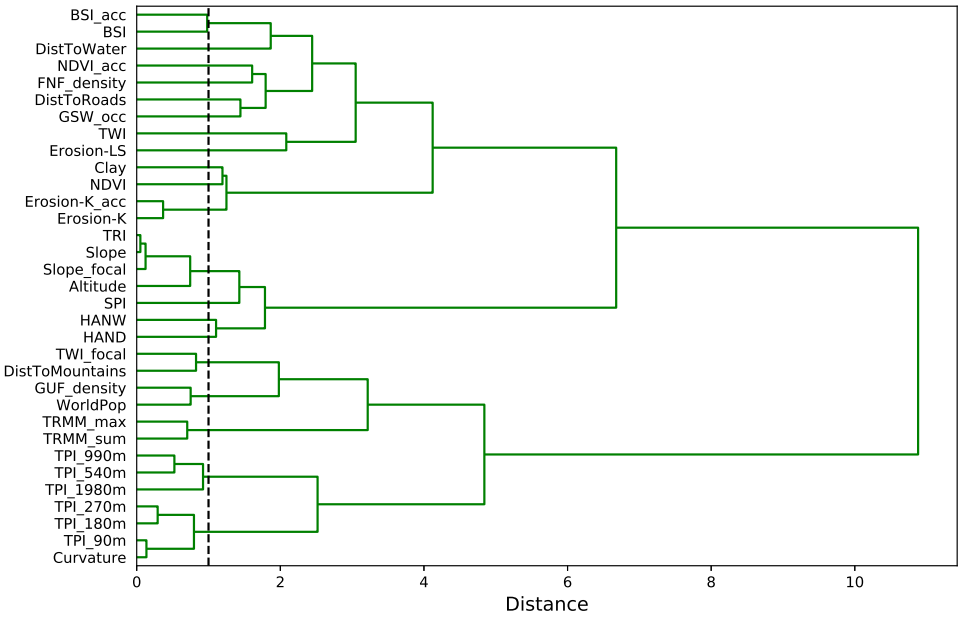

FIGURE 4.1: Hierarchical clustering on Spearman rank correlation of the candidate features. The dashed line marks the threshold used for defining groups.

4.3.2 Clustering methods

We visualized the data via t-distributed stochastic neighborhood embedding (t-SNE, van der Maaten and Hinton, 2008), and eventually applied density-based clustering OPTICS (Ankerst et al., 1999) on the t-SNE results. This yielded almost perfect separation of the visible shapes and understandable meaning when visualized on a geographical

map (Figure 4.3). Contrary to some theoretical caveats, it has been observed that this procedure is able to handle highly complex real world datasets like genome expression profiles (Macosko et al., 2015) or hand-written digits (MNIST, Lecun et al., 1998), on which standard algorithms like k-means fail. This seems to be the case here as well (Appendix).

4.3.3 Classification methods

The fraction of variability in observed damage explained by our datasets was estimated via machine learning algorithms of varying complexity, implemented in Python using scikit-learn (Pedregosa et al., 2011). To account for the structural differences of algorithms, a comparison of different types of models is presented (linear generative and discriminative, non-linear tree-based and kernel-based). Ordinal logistic regression (OLR) is explicitly designed for handling ordinal data. Therefore, this approach would be the most appropriate solution from a theoretical point of view. We used the all-threshold version of the MORD package (Pedregosa et al., 2017). Naïve Bayes (NB) is the generative counterpart of multinomial regression (Ng and Jordan, 2002), which has been shown to perform surprisingly well in classification despite its oversimplified assumptions (Ester et al., 1996; Hand and Yu, 2001). Since our data is far from normally distributed, and scikit-learn does not support other distributions, we applied equal frequency discretization and used multinomial NB. Support Vector Machine (SVM) (Vapnik, 2000) is a frequently used discriminative algorithm in statistical learning. It selects actual samples from the data as support vectors of a maximum margin separation plane. For non-linear classification, SVM relies on a kernel function to increase the dimensionality of the input data to potentially infinite degrees, where classes ideally become linearly separable. When mapped back into the original feature dimension, this hyperplane appears non-linear. Parametrization affects the kernel function as well as margin criteria (support vectors are those which generate the maximum margin plane, but some violation may be allowed by the user to avoid overfitting). We have used a radial basis function (based on LIBSVM) and a linear kernel with l1 penalty (LinSVM) (based on LIBLINEAR) (Chang and Lin, 2011; Abdiansah and Wardoyo, 2015). To provide probability estimates for these models, we applied Platt scaling, which is essentially a logistic regression on the distance to the SVM hyperplane (Platt, 1999). Random Forest (RF) (Breiman, 2001a) can be described as ensemble of decision trees, constructed via bagging, bootstrapping, and random feature selection. The method is very flexible, computationally efficient, and not too sensitive on hyperparameters. RF is less prone to overfit than individual decision trees (the generalization error converges), but in practical application the training data is often not sufficiently representative, and regularisation is most effectively implemented via restricting the maximum depth of the trees. Probability estimates of RF are generated from the class distributions in the terminal nodes. We also scaled the RF probability estimates by the same procedure as for SVM and LinSVM.

Performance metrics

Since the test scores were computed on equally distributed classes, only two metrics are shown: total accuracy is the most common classification metric, simply counting the percentage of correctly classified labels (Eq. 4.1). Random guessing, or always predicting the same out of 4 classes, would result in a score of 0.25, which is marked in the plot as chance agreement. Values above this threshold indicate a true skill of a model (comparable to the kappa statistic). Log loss (Eq. 4.2), sometimes called cross-entropy, is a

metric for probabilistic classification, which penalizes "confident wrong" decisions. It is also the internal cost function of multinomial logistic regression. Random guessing will simply result in a score of log(K), and lower log loss values indicate better probability estimates.

$$Total Accuracy = \frac{Correct\ labels}{Number\ of\ samples} \tag{4.1}$$

$$-logPr(Y|P) = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} y_{i,k}\, log(p_{i,k}) \tag{4.2}$$

where Y is a matrix of true labels in 1-of-K encoding with elements $y_{i,k}$ and P the associated matrix of estimated probabilities with elements $p_{i,k}$

Model inspection

Feature importance metrics are used to create a ranking of features which to consider for a certain classification task. We computed the permutation importance, defined as variation of the skill of a fitted model in response to permutations of individual features, as implemented in scikit-learn. Permutation importance has been shown to be theoretically more reliable than measures based on impurity decrease (Strobl et al., 2008) and already been applied for RF models in the field of flood damage data (Sieg et al., 2017; Rözer et al., 2019). Since the approach is model-agnostic, it can be used to estimate importance even for algorithms which do not intrinsically provide such a measure (e.g., non-linear SVM). Also, the importance measure for various algorithms should be comparable when consistently computed by the same approach. Similar to (Rözer et al., 2019) we further normalized the mean scores per algorithm, and weighted them by the classification accuracy, to merge the rankings into a single plot while retaining the information from the individual algorithms. Partial dependence plots (PDP) (Friedman, 2001) were used to visualize the shape of the marginal effect of a feature on the prediction of the trained models. PDP average the predictions for a set of samples (the training set) when only the specific feature of interest is altered across its entire value range, while all other features of these samples stay fix. Finally, an RF model was used to develop a spatial damage probability map, based on the feature raster stack.

Sampling

Training and performance evaluation was conducted via a nested cross-validation (Figure 4.2). In this scheme, the inner loop optimizes the hyperparameters based on 10-fold train/test-splits, while the outer loop provides independent holdout sets. As the number of buildings with high damage is naturally lower than the number of buildings with low damage, we applied oversampling to balance the training set. In statistical learning, this class imbalance would otherwise cause a problem: when an algorithm is trained to optimize total accuracy, it often learns to trivially predict the most frequent class all of the time (Krawczyk, 2016). Our aim, though, was not to obtain the highest predictive accuracy, but to derive insight on the question what separates these classes. The oversampling was implemented after the outer train/test-split, to avoid replicated samples falling into both folds to bias the performance score on test set. The best estimator of every inner cross-validation loop was used to compute the importance score. Also the permutation process was repeated 10 times for the same fitted model, resulting in 100 importance scores per model per feature.

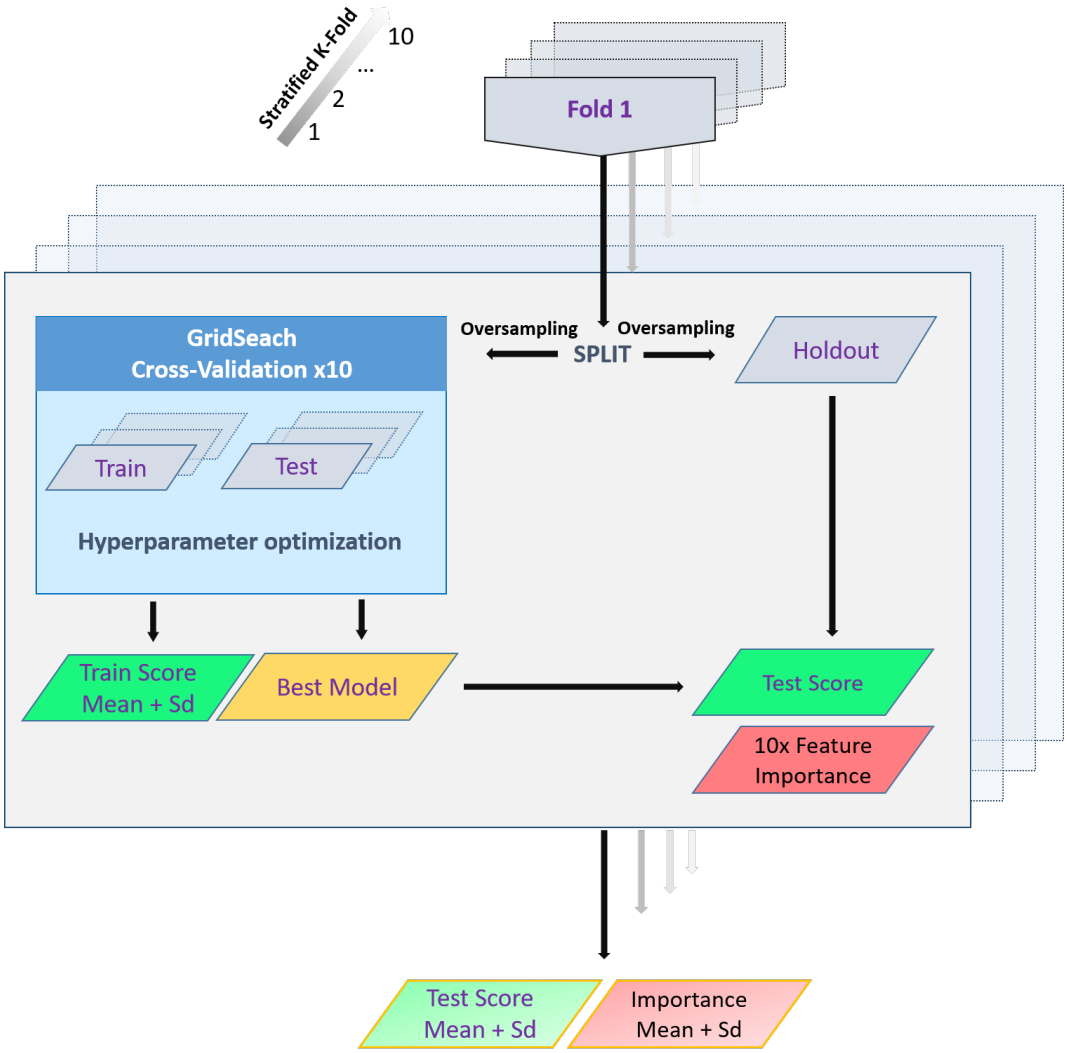

FIGURE 4.2: Sampling workflow (nested cross-validation).

4.4 Results and discussion

4.4.1 Clustering results

The damaged buildings were grouped into 9 clusters, which exhibit a regionally consistent distribution on the map (Figure 4.3) and can be interpreted in a meaningful way. Boxplots of feature value distributions per cluster (Figure 4.4) help to understand the environmental characteristics, while the corresponding relative frequency of damage grades indicates the severity of impact. The largest clusters are the high mountain ranges #4 and steep canyons #5, clearly distinguished from the rest by altitude, slope, and multi-scale TPI. While #4 is characterized by highest absolute altitude, #5 exhibits the highest range of altitude, as well as of HAND and SPI. It is noteworthy that the mountains in Peru also have more forest cover than the coastal areas, which is reflected in FNF_density and NDVI. Vegetation is scarce in the arid coastal areas, where the soil has poor humidity. In contrast, the mountain forests fully occupy the middle part of the western flank of the Andes mountains in Peru and inter-Andean valleys, at a soil with moderate moisture deficit.

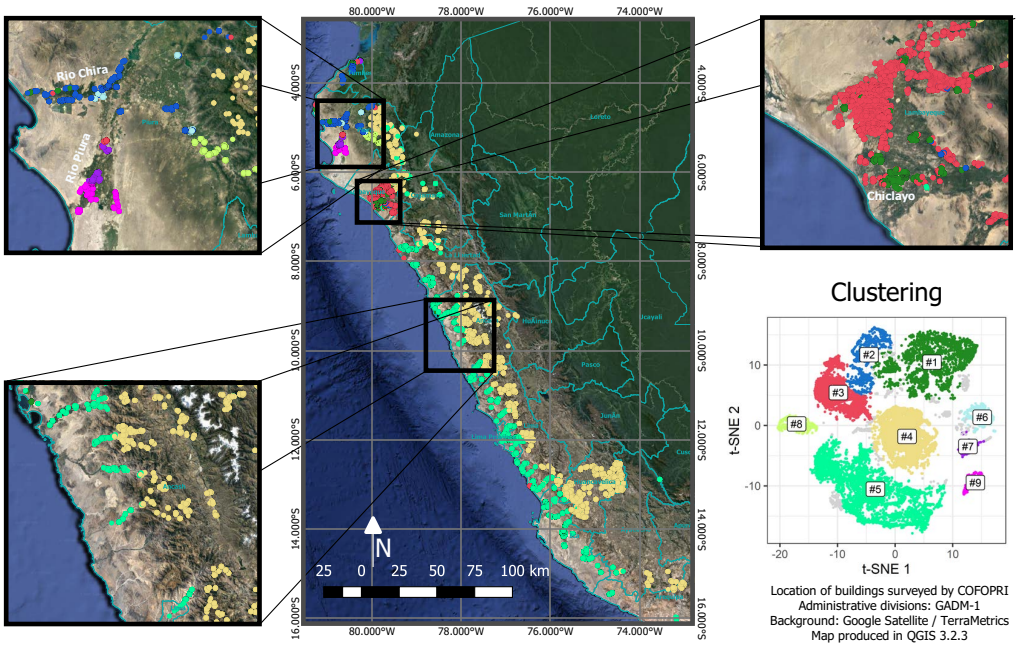

FIGURE 4.3: Clusters on a geographic map and in t-SNE space. Points in grey colour have been excluded by the clustering algorithm as "noise", and not been used in the further analysis.

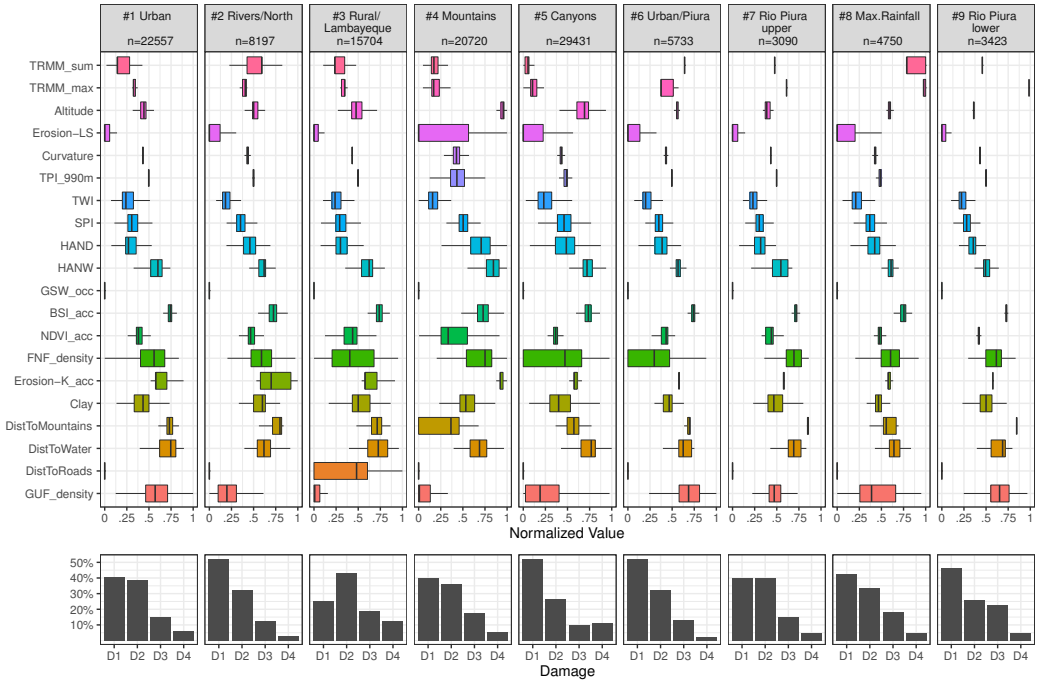

FIGURE 4.4: Distribution of feature values and relative frequency of damage grades per cluster: mildly affected (D1), moderately affected (D2), uninhabitable (D3) and collapsed (D4). Number of buildings per cluster denoted as n.

Rural Lambayeque #3 and the otherwise similar urban areas #1 are separated by DistToRoads and GUF_density. Cluster #2 contains buildings along the northern rivers, including Rio Chira (lowest DistToWater). Piura, the region where the event hit the hardest, is further subdivided by the algorithm into 4 small but dense clusters, based

mainly on variations in TRMM rainfall measurements: #6 consists of Piura's cities, which were affected by significantly more rain than the cities in #1. Clusters #7 and #9 appear spatially connected in the Rio Piura basin, with #9 showing higher maximum rainfall. Finally, #8 is identified as the zone where highest 3-hour maximum and highest sum of rainfall coincided. Already from comparing the boxplots and histograms the impression arises that measured intensity of rainfall does not directly relate to damage. Clusters #8 and #9 do not stick out with respect to damage grade D4 (collapsed), but rather with D3 (inhabitable). This is most obvious when comparing the very similar clusters #7 and #9. A large D4 fraction is found in #5 with low rainfall, although the fraction of D1 is also large in this cluster. It is reasonable to assume that rainfall, locally and upstream, triggers flash floods or landslides causing heavy destruction within narrow channels. The urban cluster #1 exhibits less relative damage compared to the rural areas in #3 at similar rainfall. Interesting is cluster #3, which does not only have the highest D4 fraction of all clusters; it is also the only region where more D2 cases were observed than D1 cases. We assume that the area was partly hit by fluvial floods in addition to long-lasting inundation from ponding water. This would fit to the relatively high values of TWI and low values of HAND. We might further speculate that building quality and shielding effects of the urban environment lead to less damage in cities in comparison to rural areas. This observation is in accordance with Ettinger et al. (2016) who state that dense urban areas were less damaged. From a hydraulic point of view, urban morphology is expected to affect the flow velocity and consequently the hydraulic pressure on a building, as for example sketched in Maiwald and Schwarz (2015).

4.4.2 Classification results

Various machine learning algorithms were trained to predict the observed damage, using all data at once, as well as data from each cluster separately. The measured classification accuracy was moderate for all models, but consistently better than chance agreement (Figure 4.5). The skill of the linear models is very low for the entire dataset, but apparently increases for smaller clusters – potentially, linear relationships are present in smaller subsets of the data. The highest scores are obtained for the clusters #5, #7 and #9, but RF and SVM show the most stable performance when using all data at once. The highest total accuracy of all models was 55% correct labels with 4 classes in cluster #7; RF trained on all data yielded a very stable mean of 48%. To put this into perspective: the spatial resolution of our features is relatively coarse, and empirical studies on damage data typically report rather low scores, even when intensity metrics and building details are included in the analysis. Ettinger et al. (2016) claim to predict 67% correct labels, also with 4 classes, in a detailed local flash flood study in Arequipa using high resolution data, Jakob et al. (2011) report about 50% for debris with 4 classes. In Germany, Laudan et al. (2017) produced an imbalanced model, which failed for high damage grades. Compare also the regression approaches of Merz et al. (2013) for percental damage to residential buildings with an RMSE around 0.1, Sieg et al. (2017) for damage to commercial buildings and goods with RMSE around 0.25.

The log-loss scores imply quite unreliable probability estimates of the classifiers. NB is known to be weak in this respect due to the independence assumption. OLR should provide more reasonable estimates, and we also observed this tendency, but it suffered from its very low overall skill in this task. Interestingly, SVM achieved the lowest (i.e., best) log-loss values, despite not even being a probabilistic method – in fact, the probability is only generated via additional logistic regression on the distance to the hyperplane (Platt, 1999). It is worth to mention that RF without calibration (not shown)

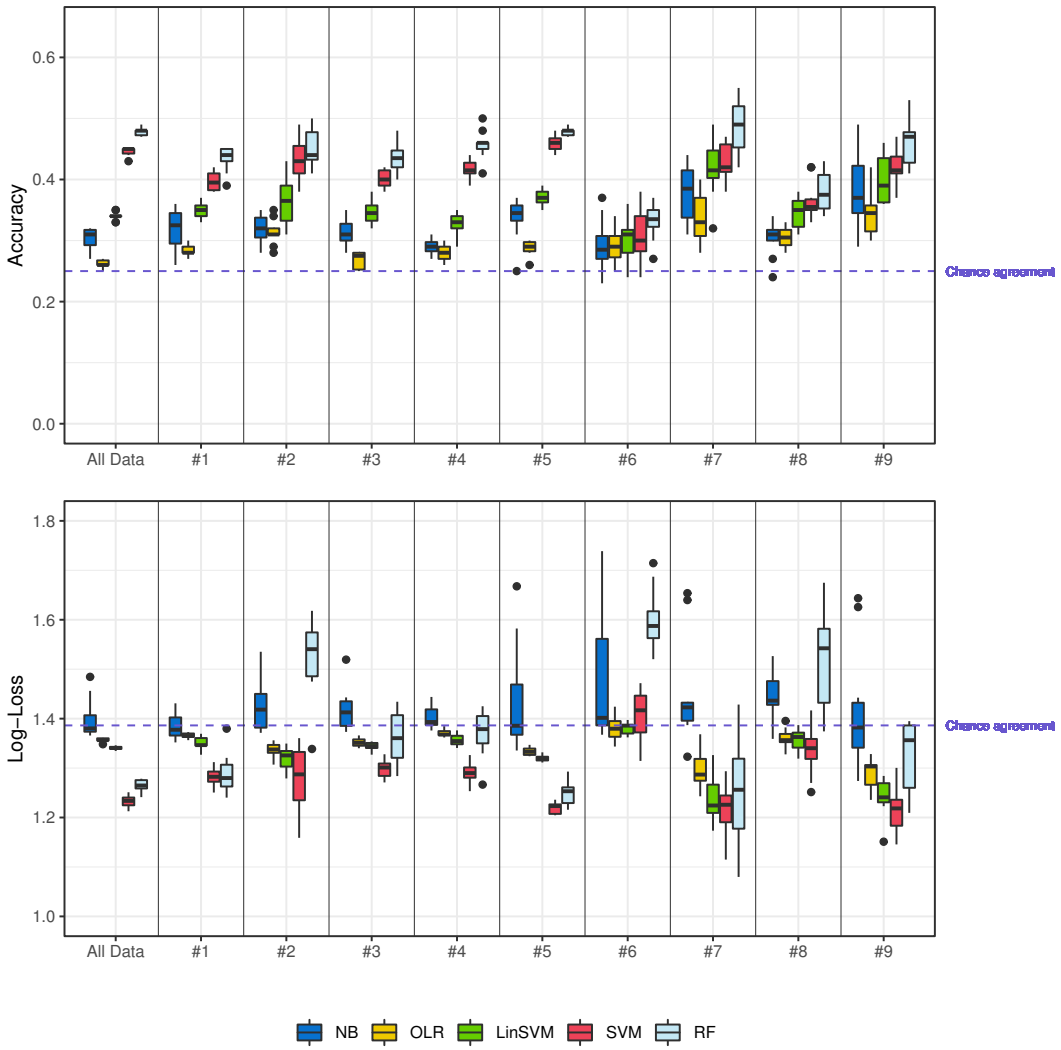

FIGURE 4.5: Skill of different algorithms trained on selected features for all data and per cluster. For accuracy, higher values better, for log-loss, lower values are better. Dashed blue lines indicate chance agreement.

performed an order of magnitude worse, with log-loss scores up to 8 in cluster #6. RF may sometimes assign confident but wrong labels on unseen data, since tree-based models are by design unable to extrapolate. Although the scores get more stable as more training samples are used, it seems recommendable to also calibrate the RF output when probability estimates are desired (as suggested by Dankowski and Ziegler (2016) among others).

Feature importance

As already observed by Rözer et al. (2019) among others, the individual algorithms rarely agree on the ranking of features (Figure 4.6), except in the case of a single very important predictor – it is therefore first of all interesting to look at the consensus: TWI is very clearly selected as the most important feature for all data, clusters #7, #8, #9, and – despite less contribution from SVM – also #3. These happen to be mainly areas of low topography and intense rainfall. The precipitation data is in most cases assigned medium to high importance by all algorithms equally, with TRMM_sum ranked first in clusters #2 and #4, and TRMM_max in the top three for #1, #3, #5, #6, and #7. In the

canyons, #5, all algorithms agree on the importance of Erosion-LS and BSI_acc, and in the high mountains, #4, GUF_density receives similar contributions. Other high-ranked features are mainly selected by a single algorithm: BSI_acc in #1, FNF_density in #2, and TPI_990m in #1, #3, #6, and #9 were all chosen by RF. SVM assigned relatively high values to altitude in #1, #3, #5, and #6, while LinSVM pushed curvature up in #1 and #3. For all data (i.e., without clustering), the ranking is dominated by the complex algorithms, due to the low skill of NB and OLR, and explaining this ranking is not straight-forward. In fact, it seems that RF, despite being the strongest classifier in terms of accuracy, shows the most unexpected behaviour in the importance ranking: even the very clear importance of TWI in cluster #9 would turn to a mere second place after TPI_990m, if relying solely on RF. Still, it is possible that RF detects real interaction effects with TPI_990m which the other algorithms fail to discover. Already in the correlation dendrogram (Figure 4.1) the linkage between curvature and small-scale TPI is strong. Besides their intended use, curvature and BSI might carry textural and spectral information, respectively, which an algorithm of sufficient complexity might exploit to learn patterns e.g., of landforms, urban structure, or other characteristics.

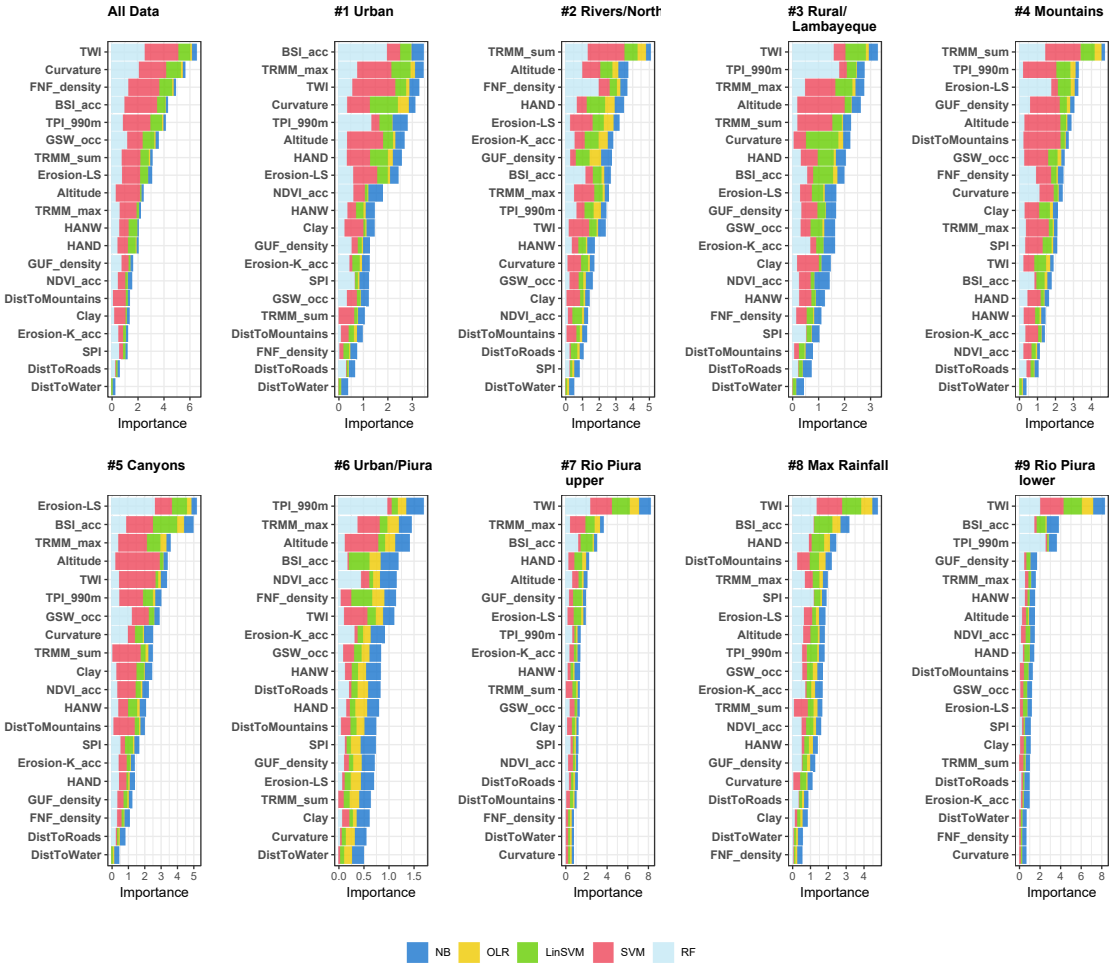

FIGURE 4.6: Importance ranking by 5 different algorithms. Shown is the mean value from 100 repetitions, rescaled to the range [0, 10] per cluster per algorithm, and weighted by the classification accuracy.

Partial dependence plots

PDP are presented to further investigate the model behavior when training on the entire dataset (Figure 4.7). The stepwise behavior of NB reflects the discretization levels. Since the discretization followed an equal frequency strategy, these steps give a visual impression of the distribution of the data. In some cases, like Curvature, the vast majority of samples fall in a narrow value range, and the behavior of the models outside that range is rather not credible.

A strong effect is found in the PDP for GUF_density and DistToRoads, which are apparently not among the top features in the importance ranking. Higher urbanity indeed leads to less prediction of high damage (D4) and more prediction of low damage (D1). TWI and HAND exhibit the expected trend (higher TWI and lower HAND lead to more predicted damage). A very remarkable finding, best seen for RF and SVM, is that increasing intensity of rainfall, TRMM_sum and TRMM_max, leads to damage grades D1-D3 in the expected order, but D4 (collapse) is caused by different factors. The rainfall maximum seems to be more damaging than the sum. The interpretation in the case of Curvature and FNF_density is that there is no net effect of these features (i.e., we cannot state that higher Curvature leads to higher or lower damage). While this finding seems to contradict the importance ranking, it is more in agreement with our expectations. Both results together, the importance ranking and the PDP, imply that a feature may be selected as important by a learning algorithm, taking into account various interaction effects, without containing any clear message for process understanding. Since the PDP only displays the marginal effect, it is possible, especially for complex algorithms like RF, that individual effects average out. The plot in this case would look flat, despite high variability for individual samples. Effects visible in the PDP are causal for the models, not necessarily causal in nature. PDP can potentially be computed for every feature of every model for every cluster and every target class. In principle, PDP can be computed for the interaction of multiple features. Unfortunately, visualization beyond 2 features, or the comparison of different algorithms, is challenging. An example of a 2D plot, based on RF, is provided (Figure 4.8) to underline a central thematic finding of this study: The intensity of rainfall, sum and maximum, leads to damage grades D1-D3 in the expected order, but D4 (collapse) is caused by different factors. Although there is a slight increase of probability for D4 in the extreme cases, maximum seems to be more damaging than sum. Damage grade 4 has higher probability of occurrence in areas of high SPI and BSI_acc/low HAND and low GUF_density.

From the abovementioned observations, we draw the following thematic conclusions: TWI is the most useful of the tested topographic features, especially in case of pluvial floods in low elevation terrain. Rainfall from satellite measurements, despite the coarse spatial resolution, can be useful for damage modelling. The sum of rainfall seems more important for vegetated mountains and low riverbeds, while the maximum seems more important in steep canyons and urban areas. Erosion-LS and BSI are good indicators in steep terrain or desert areas. This is in line with our expectation.

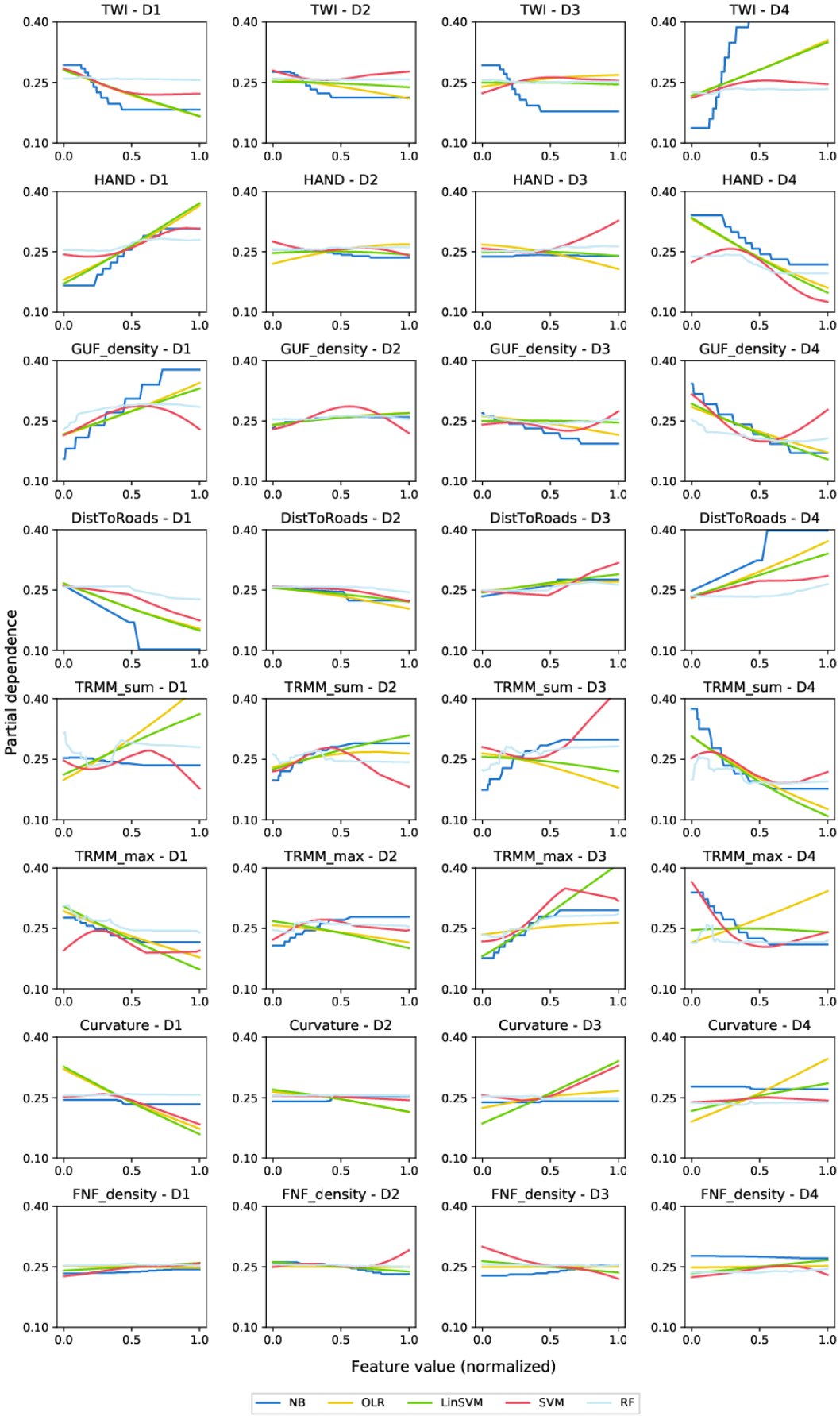

FIGURE 4.7: PDP for using all data. Y-axis is the predicted probability of the class.

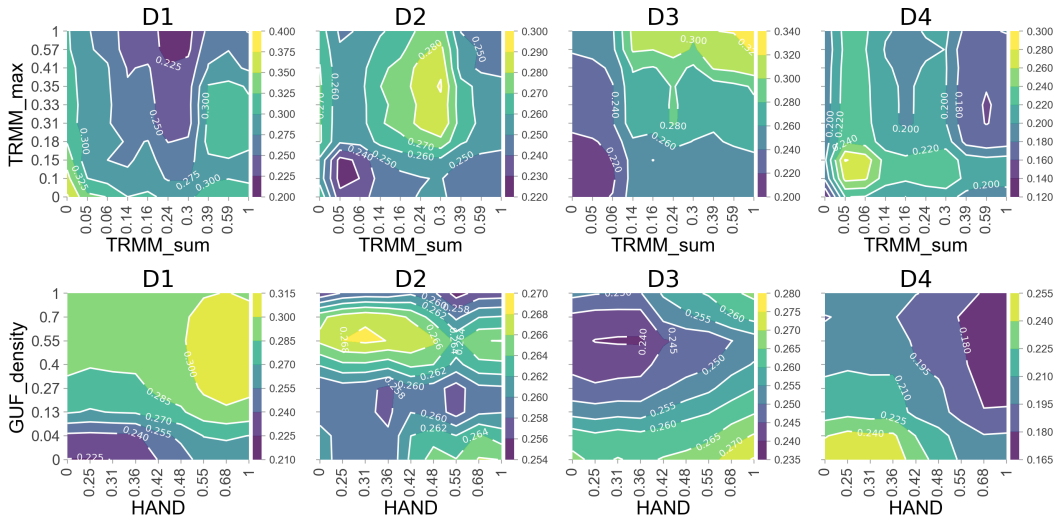

FIGURE 4.8: PDP interaction of 2 features, generated using PDPbox.

Damage probability map

Figure 4.9 presents an example of a classified damage probability map, based on RF, to demonstrate the technical feasibility. Figure 4.9A shows the most likely damage grade, while in B the probability of D4 is plotted, and C is a natural colours image (Google Satellite) for comparison. RF was chosen due to its classification accuracy, as well as computational efficiency. Such a spatial prediction can also be useful to visualize what the ML model has learned, making the decisions more transparent, although detailed evaluation for every model is beyond the scope of this study. In the presented example, rivers and desert areas were learned to be dangerous in case of *El Niño*, while urban areas seem relatively safe places. A map produced by a damage model with sufficient skill could help to identify critical areas for risk planning.

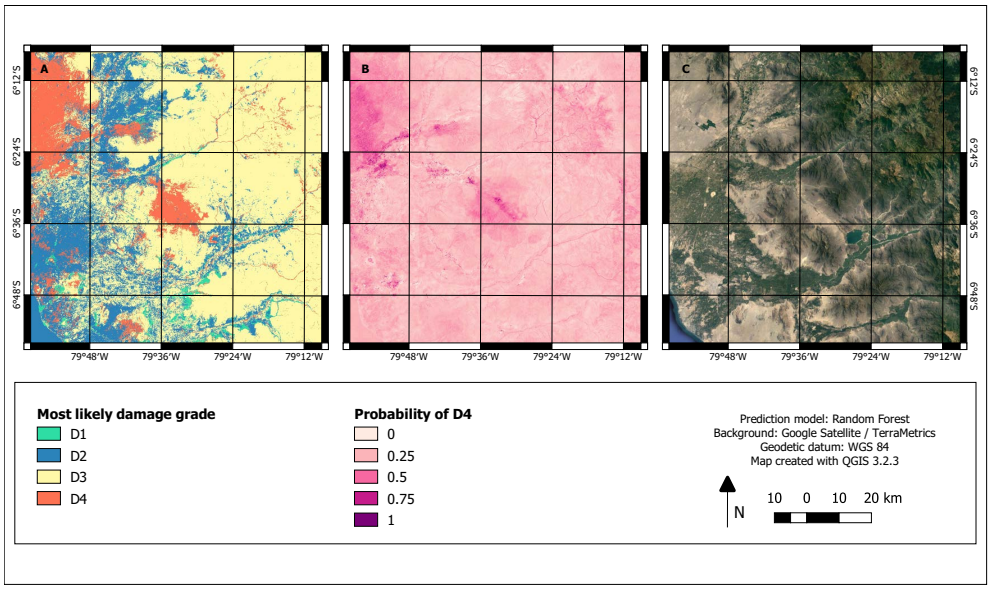

FIGURE 4.9: Example of a damage probability map, produced by RF on 20 features.

4.5 Conclusion

This study has demonstrated how a research question from the natural hazard risk domain can be formulated as data-mining problem, and the potential of ML methods in combination with open geodata, to identify drivers of observed building damage from a compound *El Niño* event. We found t-SNE a very valuable technique to visualize the high dimensional dataset, and clustering via OPTICS to yield a good separation of regions with different dominant processes. Plotting the clusters spatially on a map, and comparing the distribution of feature values to the relative frequencies of damage grades, already helped to discover some key findings. Classification accuracy of the non-linear models (RF, SVM) was within the expected range for an object-specific damage model, and given the resolution of the data. Performance of the linear classification algorithms (OLR, NB, LinSVM) was just slightly above chance agreement when using all data at once, and a little higher in smaller clusters. Apparently, linear relationships might be found when subdividing the dataset, but complex models are necessary to capture the entire event. SVM with radial kernel is computationally very expensive, especially in the optimization via grid search, therefore RF seems the algorithm of choice for such a task: it shows consistently good accuracy, and scales well with high numbers of observations due to efficient parallelization. Since damage processes at the object level are still difficult to predict, there is currently a trend towards the development of probabilistic damage models, which capture the stochastic nature of the observations. Probability estimates in this study were rather unsatisfactory, although SVM with Platt scaling was found superior to RF in this respect. We clearly recommend to use probability calibration for RF as well. Gaussian Process would be a classifier rooted in solid probability theory, but it turned out as computationally unfeasible for the amount of data at hand (complexity $O(n^3)$). Besides the selection of algorithms, the preprocessing of data, and especially our manual feature engineering, might be a limiting factor for the model skill. If the aim is a predictive model with maximum accuracy, a possible solution would be to try deep learning via a convolutional neural network, which performs feature engineering automatically. In that case, however, process understanding will be more difficult to extract.

Previous studies presenting feature importance rankings often relied on a single algorithm. By weighting the normalized scores for different algorithms, and providing the individual contributions, a more facetted picture appears. Computing feature importance scores per cluster yielded more understandable results than the ranking for the entire dataset. It is necessary to keep in mind that those variables which define the cluster have lower variance and will accordingly not be as useful in separating samples within this same cluster anymore. In areas of strong precipitation we therefore do not find the precipitation as most important, but rather local topography – those features that make a difference in case of strong precipitation. However, some results are difficult to understand, e.g., the high ranking of curvature for the entire dataset. We would like to point out that studies presenting such feature importance rankings still have to be taken with care, and strongly suggest to provide partial dependence plots as complementary information about the direction of the influence. A potential explanation for the sometimes contradicting results is the moderate model skill. It is possible that models with better generalization performance would also yield more stable feature importance rankings. A classified map may further visualize which patterns a model has learned, although this study only presented a single example to prove the technical feasibility. A clear limitation is the resolution of the elevation model. Finer local segmentation, and higher classification accuracy, might be possible by the

same methodology, when using high resolution topography.

Thematically, we could show that topography, settlement indicators, and rainfall partially explain the observed damage. The datasets used allow for separating regions with different dominant damage processes, like mountain regions, canyons, floodplains, urban areas, and regions affected by exceptionally strong rainfall, via unsupervised techniques. These clusters exhibit different histograms of damage grades – most vulnerable appear to be the canyons and rural areas, while urban centres were less damaged at similar rainfall. TWI is clearly selected as the most important topographic feature, especially in regions of strong rainfall, which should be considered for hazard maps. Slope length and steepness (Erosion-LS) is identified as main driver of damage within canyons and lower mountain ranges. TRMM_sum and TRMM_max are often selected despite the coarse spatial resolution. We are therefore optimistic that satellite-based rainfall data can be valuable for damage modelling. We could further show that the models learn from observational data to predict damage grades D1-D3 as expected with increasing intensity of rainfall, but D4 (collapse) is driven by different factors like the location close to a stream in a rural environment (low HAND and GUF_density). We have sketched the concept of empirical damage probability maps. However, the presented map is meant as prove of the technical feasibility, and by no means a result to use in practical risk management, which to create would be a study on its own. So far, we have shown what is possible when using openly available data. We are optimistic that a similar methodology with higher resolution data, and including characteristics of the affected buildings, could obtain better performance in terms of accuracy and probability estimation, and consequently help to improve risk assessments.

Acknowledgements: The authors wish to thank El Organismo de Formalización de la Propiedad Informal (COFOPRI) for providing the damage data of the *El Niño* event in Peru in 2017. The damage data is available via the COFOPRI website (Geollaqta). All other data sets used for the study are openly available and cited in the text.

Funding: The research presented in this paper was conducted in the framework of the project "Multi-risk analysis and information system components for the Andes region (RIESGOS)" funded by the German Ministry of Education and Research (BMBF); (contract no: 03G0876B).

4.6 Appendix

Although clustering is by definition an unsupervised procedure, many methods require the user to specify the expected number of clusters. This is usually not possible when real data is analysed. Global criteria for estimating the optimum number of clusters are arbitrary and often contradicting (the R package NbClust includes 30 metrics). Arguably, for geographical data, the most suitable validation is on a spatial map. Depending on the size of the data, it may be unfeasible to manually inspect many different clustering approaches in detail. In that case, an embedding technique like t-SNE is helpful.

At first, we tested simple k-means and hierarchical clustering, applied in the original data space, with various numbers of clusters. Results were difficult to interpret on a geographical map. We then visualized the data via t-SNE, iteratively tuning the parameters to test the stability of the resulting shapes. The labels from k-means were in moderate agreement with the t-SNE shapes for about 6 clusters, but apparently contained some confusion, and additional clusters were visible which could not be separated by any tested approach (Figure 4.10). From 8 clusters upwards, k-means started to split the central cluster in t-SNE space, which corresponds to the high mountain ranges in geographical space. Therefore a density-based clustering (OPTICS) was applied to label the t-SNE shapes. OPTICS also requires the user to set a threshold, but provides a visualization to help making this decision. Two thresholds appeared meaningful, resulting in 7 and 9 clusters, respectively. The slightly lower threshold produced a large cluster composed of urban areas, rural Lambayeque, and the northern rivers, i.e. on a geographical map the decision was obvious (Figure 4.3).

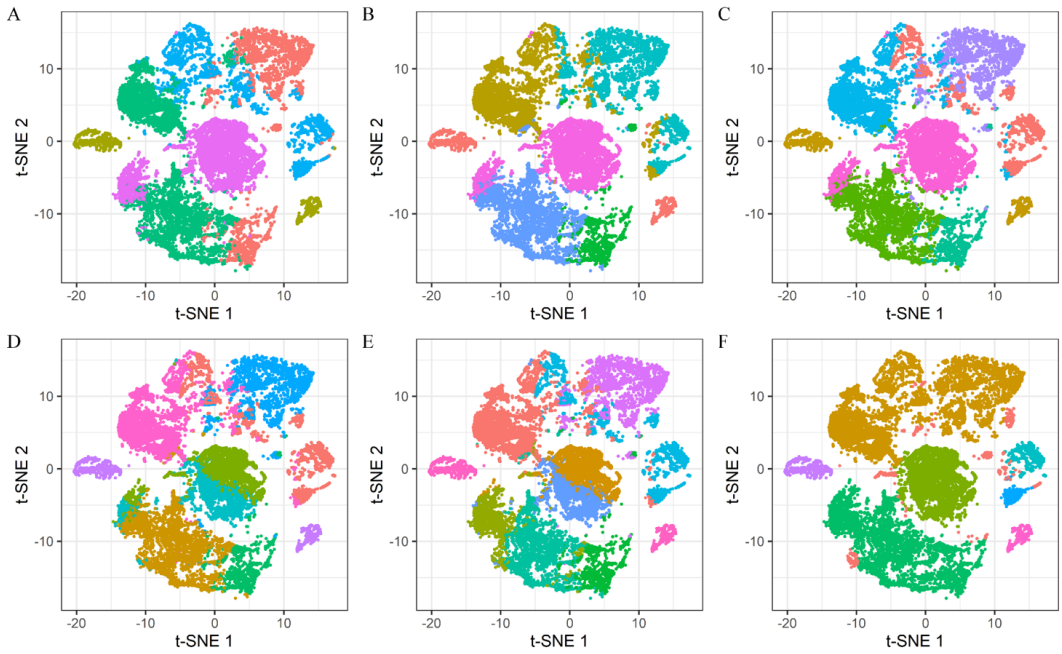

FIGURE 4.10: Example of clustering attempts using k-means in the original data space for 5-9 clusters (A-E), and OPTICS in t-SNE space with a threshold resulting in 7 clusters (F). Clusters are represented by colors, no labels were assigned.

Chapter 5

Synthesis

5.1 Answering the research questions

5.1.1 Q1: Which open geospatial datasets contain useful information for flood hazard and risk modelling?

All studies presented in the previous three Chapters required a digital elevation model. Although not novel per se, DEMs are increasingly available free of charge in high spatial resolution (e.g. 10 m and below in the USA, even 0.5 m in the Netherlands https://www.ahn.nl/, accessed 03.01.2022). Refined datasets with error correction have improved the quality of older products, in the case of MERIT even with explicit focus on hydrological applications (Yamazaki et al., 2019). The DEM created by the TanDEM mission is also openly available in a 90 m version (German Aerospace Center (DLR), 2018), while the high resolution version is distributed commercially. The German DGM10 that was used in Chapter 3 is only free of charge for state agencies according to §4 V GeoBund (https://sg.geodatenzentrum.de/web_public/gdz/lizenz/Vertrag_GeoBund.pdf, in German, accessed 18.10.2021), and fee-based for third parties. Higher resolution DEMs from LiDAR sources have been made available as open data by individual federal states, but not for entire Germany yet. Hydrodynamic models, such as used in Chapter 3, depend heavily on a DEM. Explanatory features derived from a DEM have also been found most relevant for the ML-based hazard and damage models presented in Chapters 2 and 4.

Obviously useful in the flood risk context is information on hydrography. If necessary, such information can be derived from a DEM, as done for the SWIM hydrological model in Chapter 3. However, hydrography is also openly available from the OSM water layer. The contained data is crowd-sourced, but from visual inspection against background images the product appeared to be of good quality in Houston and large parts of Peru. Mapping of water bodies and even underground channels seems to be more comprehensive than the mapped streets and buildings in OSM for these regions. Another tested product is GSW, derived from almost the entire Landsat archive. As noted in Chapter 4, a combination of OSM water and GSW could overcome the limitations of both products, as OSM is more comprehensive in urban areas, which are difficult to observe from space, and GSW more reliable in the remote wilderness that is difficult to access by human cartographers. Results of Chapters 2 and 4 hint at the importance of HAND and distance to channels for flood hazard and damage modelling.

Spatially explicit precipitation measurements can typically be obtained from national institutions such as the NWS in the USA, SENAMHI in Peru, or the DWD in Germany. However, not all countries have the same density and reliability of measurement

stations. The aggregated feature importance rankings in Chapter 4 show that satellite-derived rainfall measurements were useful for explaining patterns of damage during an *El Niño* event in Peru. Such products are available quasi-globally in high temporal resolution (currently 30 *min* for GPM), while spatial resolution might be improved by sophisticated downscaling techniques (Sharifi et al., 2019).

Land cover information in a static format can be taken from products such as CORINE in Europe (Varga et al., 2021), or from coarse global datasets that are still rather inconsistent (Hua et al., 2018; Chirachawala et al., 2020). Near real time information on vegetation cover, soil or sealed surfaces can be derived from spectral ratios such as NDVI, BSI, and others, as conducted in Chapter 4 to estimate pre-event conditions. Less dynamical but still difficult to observe is the subsurface. Soils can be very heterogeneous on the small scale and only extensive fieldwork can clarify the true conditions. SoilGrids attempts to compile, harmonize, and interpolate data from fieldwork around the globe – however, the information is associated with large uncertainties, especially in areas with low density of soil profiles (Poggio et al., 2021). The BSI was selected as useful predictive feature by some damage models in Chapter 4. Other soil information was never among the most important predictors, though, and not even considered in Chapter 2 as the study area was primarily urban.

Mapping of exposed objects has reached the building level even on national scale (OSM in Europe, Microsoft USBuildingFootprints in the USA). Aggregated information is at least globally consistent (GUF, WorldPop). Further methodologies exist for extracting detailed vulnerability-related information, for example on construction material and building height, for large areas (Pelizari et al., 2021; Geiß et al., 2019; Paprotny et al., 2020).

The standardized access via data cubes such as GEE is promising, but still today many sources of information need to be downloaded individually and merged. Besides the datasets used in this thesis, additional promising sources include social media and street view camera footage. A very prominent data source in the flood risk context are flood extent maps of real events, which are for example distributed rapidly and free-of-charge by the Copernicus EMSR service. Mapped inundation extent can be useful in emergency response, when available during or directly after a flood event (Chung et al., 2015; Oddo and Bolten, 2019). In principle, it should also be possible to generate rapid impact assessments by applying damage models on this satellite-derived hazard information, and to use the flood masks as input for calibrating hydrodynamic models (Taubenböck et al., 2011; Dasgupta et al., 2021).

5.1.2 Q2: What is the quality of current satellite-based flood masks and how can they be improved?

The presented evaluation of different flood masks for the hurricane Harvey event in Houston, 2017, clearly shows that currently available products have major problems in detecting flood in vegetated and urban areas. The Copernicus EMSR product detected only 1% of flooded vegetation and even less flooded urban areas. High quality products correctly detected about 20% to 47% of flooded vegetation and up to 42% of flooded urban areas. The primary reasons for this low skill are specific characteristics of the sensors (typically SAR) and viewing geometry, i.e. issues that can hardly be solved during the creation of flood masks from the sensor data. This finding has implications for subsequent use of these products in hydrodynamic or damage models. Reliable

derivation of water depth is accordingly difficult, as algorithms require a precise land-water boundary (Zwenzner and Voigt, 2008; Cohen et al., 2017). In the early stage of this research project, the limitations of these flood masks in the case of the *El Niño* event investigated in Chapter 4 became obvious. Therefore it was not possible to include such products in an empirical study of flood damage.

Specificity of all tested products was found to be high, though, meaning that detected flood corresponds to true flood. Therefore it makes sense to treat satellite-based flood masks as PU data, as long as no spatial validity layer is provided. In Chapter 2 it was further demonstrated how these products can then be improved by a secondary modelling step, using OCC methods to re-learn the flood extent from topographic and hydrographic features. Including additional features from rainfall measurements and distance to buildings led to mixed results and cannot be recommended in general, although this finding might need to be re-evaluated in different case studies. It is imaginable that especially topographic features have more explanatory power for detecting flooded vegetation than flooded urban areas, which might be influenced by additional factors such as sealed areas or a sewer system. In principle, the developed procedure can use any kind of explanatory features, so there is potential for future studies.

According to the performance metrics, the so-obtained refinement of the flood masks can be very successful. However, in a practical application the user has to select a model in the absence of reference data, and define a classification threshold to the continuous output of the algorithms. The main issue with these extrapolations is the "sample bias" from unrepresentative coverage of the initial masks. This could potentially be solved by including additional positive class samples (flood locations) from within the city, e.g. taken from street camera footage or social media. A promising way forward would be data fusion of satellite-derived initial masks and spottily mapped urban flood extent by street cameras or social media. For the derivation of water depth, methods are needed that allow a certain tolerance of error in the land-water boundary, or at least quantify the uncertainty. The author of this thesis has also developed an (unpublished) procedure to derive a water surface elevation based on kriging. In principle, other intensity metrics than water depth can be derived from satellite imagery, for example flood duration and affected area in the vicinity. Also the detected flood can be used as "stream raster", and a height above it can be computed, similar to the HAND index. Future studies could explore the improvement in assimilation performance and damage estimation by the extrapolated flood masks. However, users of the procedure should keep in mind that the ML-based extrapolation models do not necessarily follow the laws of physics. For example they do not enforce a closed water balance (mass conservation).

5.1.3 Q3: How much are risk estimates by state-of-the-art process-based flood model chains influenced by process description details such as hydrodynamic interactions?

Physical processes governing fluvial floods are more or less known in theory (e.g. Brutsaert, 2014), and process-based models are consequently well-established in fluvial hazard modelling. The more detailed the representation of processes is implemented, the higher the computational demand gets, though. While scientific literature contains plenty of studies using coupled hydrological-hydrodynamic models (e.g. de Paiva et al., 2013; Hoch et al., 2017; Curran et al., 2019), the effect of hydrodynamic process description on damage is usually not investigated. The modelled expected annual damage (EAD) in Chapter 3 was reduced by about 50% when considering hydrodynamic

interactions for the entire German Rhine basin. This aspect is therefore not negligible for models on that spatial scale. Also the spatial occurrence of predicted flood damage differs significantly when using a continuous routing approach.

Since a major catastrophic flood event has occurred in Germany, including parts of the Rhine basin, during the time as this thesis was written, it is appropriate to also discuss the published results in the light of this recent event. Total compensations issued by the federal government were set at a stunning EUR 30 billion (http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl121s4147.pdf, accessed 18.10.2021). According to the simulations presented in Chapter 3, numbers of such magnitude should only be expected for return periods > 1000 years in a major catchment such as the Rhine. However, the damage did in fact occur along much smaller tributaries (sub-basins), not along the main course of the Rhine itself. Estimated local return periods of rainfall were mainly in the range of 100 years (in some areas up to 1000 years), and there are even historical records of a similar magnitude flooding of the Ahr river in 1804 (Schäfer et al., 2021). Discrepancies to the presented model results potentially arise from the following issues: (1) the stream network considered in the model chain of Chapter 3 is not including "minor" streams such as the Ahr, and the even smaller ones, many of which are ungauged. (2) Flash flood processes with sediment transport and failure of protective structures are not implemented in the hydrodynamic module. (3) The damage models do not represent the impact of boulders or other transported objects, and completely exclude the infrastructure sector. Destroyed bridges, roads, drainage networks, power grids, etc., are simply not accounted for. (4) A stochastic weather generator only produces realistic precipitation patters when based on truly representative data. If not enough historical data is available, or if climate change alters future precipitation, the weather generator needs to be adjusted as well.

It is therefore evident that complexity of process description needs to be increased to capture dynamics such as the abovementioned. Whether ML-based modelling of the total damage could make sense for these flash flood events is probably worth further investigation, but hinges on the availability of comprehensive data. Data collection of damage from the event 2021 in Germany, in combination with observations of flood extent, velocity and sediment transport, would provide a good opportunity to improve the understanding of processes and accuracy of current models for such situations.

5.1.4 Q4: Can ML and OGD be used to construct damage models for compound events that are difficult to describe by explicit process-based model chains?

Model chains are already quite complex for a single hazard like fluvial floods. Similar tools for compound or cascading events are still rare, especially when including multi-hazard damage (Marzocchi et al., 2012; Kappes et al., 2012). A step in that direction is the RIESGOS demonstrator of decentralized WPS models (Pittore et al., 2020). However, the complexity and associated uncertainty of these chains is very high and difficult to capture. How to generate realistic scenarios with return periods remains largely unclear.

Imagine how such a model chain would have to be designed to address the *El Niño* event from Chapter 4: The first element would have to be an *El Niño* rainfall generator or historical rainfall measurements. Secondly, a landslide model would have to assess whether the rainfall alone might trigger mass movements, and estimate the damage thereof. A pluvial damage model would be required to capture the direct damage

induced by rainfall on buildings, or by ponding water in the cities. Then, regular hydrological routing and hydrodynamic modelling would be needed, but these models would have to include sediment transport, which is a relevant driver of damage in parts of Peru (Thouret et al., 2014). The mixed processes of flash floods, slow river floods and long-lasting inundation from ponding water might each require an individual damage model component. In the case of multiple hazards hitting the same place, the damage functions would theoretically need to be adjusted based on previous damage (Gomez-Zapata et al., 2020). It is very unlikely that sufficient empirical evidence for such state-dependent vulnerability functions will ever be available, as the data collection would have to take place in between the different hazard occurrences, i.e. during a time when people are struggling with the impact of the first hazard (not necessarily expecting a second hit).

Therefore the more direct approach developed in Chapter 4 seems worth exploring. The explicit description of the individual processes is substituted by a statistical description of total damage probability from a representative sample (full enumeration). It is worth to note here again that the confidence placed in this specific analysis is backed by the fact that the survey of damaged buildings was conducted by state authorities with the aim to identify every damaged residential building in the country for issuing financial compensations. We can therefore expect this dataset to be more or less unaffected by sampling bias.

In accordance with literature on susceptibility mapping (e.g. Cao et al., 2016; Chapi et al., 2017; Pour and Hashim, 2017; Tehrany et al., 2018), Chapter 2 has demonstrated that inundation from a hurricane event can to some extent be modelled from topographic and hydrographic features. In Chapter 4 it could be shown that ML models for *El Niño* damage can obtain a skill similar to what has been reported in literature for single hazard damage (Ettinger et al., 2016; Jakob et al., 2011; Laudan et al., 2017). The best performance was observed for the algorithms RF and SVM. Linear models performed poor on the entire dataset. Skill improved when subdividing the data into clusters of similar dominant damage processes, which was possible by unsupervised techniques. The resulting ML models can in principle be used to derive forecasts based on different scenarios of rainfall (sum and maximum) or different land-use (urbanity, forest density, bare soil index, etc.), however the validation would be virtually impossible without data from further events. This hints at the main drawback of ML models: they require data, which is often not available.

There is still a long way to go to bring such ML-based risk analyses on the level of the model chain used in Chapter 3. More promising is the concept of an event-based empirical damage probability map, to be used for risk communication and potentially spatial planning. The idea is similar to a susceptibility for damaging events, as recently proposed for landslides (Steger et al., 2021). In fact, there are reasons to prefer such an approach for catastrophic compound events: Financial risk modelling definitely has its eligibility for large-scale planning, governmental spending prioritization and regulatory requirements on backup reserves (e.g. Sairam et al., 2021). A measure like EAD, especially on highly aggregated scales, fails to capture the drastic consequences of compound hazards like *El Niño* floods, though. EAD can be the same for a hazard regime of frequently occurring low impact events as for rare high impact events (Merz et al., 2009). Also, the most vulnerable are the poorest with the lowest quality housing – which accordingly will contribute only little monetary value in the financial risk calculations. Engineering-type models, or statistical models predicting structural building damage, may therefore be more adequate in these cases. The population at

risk needs to be provided with information that allows for sustainable risk-mitigating spatial planning and emergency preparation. Besides hazards zoning, the potential number of uninhabitable residential houses is an important piece of information, as it allows the local authorities to plan with sufficient amounts of emergency shelters, storage for food, clean drinking water, and other relevant supplies. The methodology developed in Chapter 4 might be further developed in that direction. Also, information on critical damage processes and dangerous areas can potentially be derived from model inspection. Such information could be of value for creating more empirically-grounded risk maps, as opposed to relying solely on expert opinion.

5.1.5 Q5: To what extent can process understanding be derived from either of these methods?

Physical models with clear process behaviour are termed white-boxes (e.g. Ghonchepour et al., 2021). Hydrodynamic modelling by solving the diffusive wave equation may be regarded as such a white-box approach. The entire RFM chain is rather a grey-box, though, as empirical elements are included and calibration to data may introduce unexpected effects. Nevertheless, model inspection in Chapter 3 clearly shows how flood hazard and risk are shifted upstream (reduced downstream) when using a continuous routing scheme. Plotting the risk curves spatially allows for understanding the model behaviour. If the behaviour is accurate, such visualizations can be used to learn about the effect of different measures on flood risk. Self-evidently, process-based models are only as good as their internal process description. The general knowledge on the physics of fluid dynamics has largely been derived from laboratory experiments and analytical mathematics (Johnson, 2016).

ML models learn patterns and interaction effects from the data. Model inspection tools, such as feature importance rankings and partial dependence plots, can then be used to uncover the structure of the fitted models. However, the cost function used during training and the user-defined performance metric for model selection matter. ML models adjust their internal structure to optimize a given metric! This can consequently lead to very different results during model inspection. In addition, it has been hypothesized that the problem of equifinality, known from hydrological modelling, may similarly affect ML models (Schmidt et al., 2020). Whether process understanding can be derived from them is therefore up to debate. In the ML attitude, a model has the primary purpose of making accurate predictions. It is not of interest whether that model discovered any laws of nature or whether it fulfills certain mathematical properties. This also implies that some research questions are simply not of interest to ML advocates. In the context of flood damage models: as long as the model predicts the correct damage grade, it is not of interest whether the cause was water depth, flow velocity, or impact of boulders – and whether the model got the right answer for the wrong reason (Kirchner, 2006).

Model inspection is valuable when choosing a single model (or a few particular models) to be used in an operational procedure. In that case it is of interest to know how exactly the model functions, and if for example it contains a bias from a feature that it is not supposed to use. Infamous examples of such effects are observed in natural language processing models that learn racial or gender biases when trained on unfiltered texts from internet forums (Garrido-Muñoz et al., 2021). Similarly, in the risk modelling domain we might want to avoid model biases towards neglecting the most vulnerable population due to low building values. Such behavior would be uncovered during model inspection, and could subsequently be corrected.

The question remains whether it is possible, by doing data-mining "the right way", to discover process understanding "in general", according to the aim of science formulated by Alfred North Whitehead? Zhao and Hastie (2019) sketch a way of obtaining causal inference information from ML models using PDP, however they also mention that this requires domain knowledge in addition, to constrain the potential effects e.g. via a structural equation model. The problems of empirical research are even more basic, though: When attempting to quantify the drivers of damage or hazard occurrence from empirical data, we can mainly learn something about that particular dataset. However, the dataset itself may not contain the answer, as Steger et al. (2021) phrased it (quoting Tukey, 1986). The dataset reflects the occurrence of processes in a certain place and time, with a specific sampling bias. If empirically deriving process understanding is the goal, then comprehensive or truly representative data is necessary.

For the study presented in Chapter 2, the method t-SNE was originally used to detect structure in the training data, and individual OCC models were inspected by PDP. These plots did not make it into the final publication, for the following reason: The models were trained for maximum predictive accuracy, disregarding collinearity of features, potential autocorrelation, and representativeness of the training data. Model inspection in this case only provides insight in the structure of specific models, which depends heavily on the coverage of the initial flood mask and characteristics of the features and processes in that region, as well as on the characteristics of the OCC algorithms themselves. It does not offer general process understanding, neither about pluvial floods in general, nor about satellite-based flood masks in general, and not even about hurricane Harvey in general. Chapter 2 is therefore conceptually a pure ML study.

Chapter 4 gave a detailed analysis of 5 different algorithms trained in a nested cross validation on comprehensive damage data. Class distributions were balanced in training to focus the ML models on separating classes rather than on maximum predictive accuracy. Data was subdivided into thematic clusters for more detailed inspection, which greatly enhanced interpretability. The feature importance was computed for all models consistently by feature permutation, using only features of low collinearity, and with 100 repetitions per algorithm. Results were visually merged in stacked barplots to retain as much information as possible, and thereby avoid over-interpreting feature importance rankings from single algorithms. PDPs were used in addition to uncover the direction of the marginal feature effects. When making this effort, it is at least possible to gain insight about a particular event. Most of the thematic results obtained from this data-mining study were in line with prior expectations. Rainfall maximum was found more relevant in urban areas and canyons, rainfall sum along low elevation rivers and in vegetated mountain ranges. The learned patterns suggest that collapse of buildings rather occurred under relatively low rainfall in canyons and areas with low HAND. An interesting finding was that vulnerability appears higher in rural areas compared to urban centres. More detailed processes might be discovered when using higher resolution data, which was not openly available for the region. It is also worth to note that the complex models probably detected many more interactions than have been visualized. We are simply not able to intuitively make sense of multidimensional arrays. New tools for visualization are needed to make the full information content of such models accessible. The machine has already learned a lot, the additional challenge in data-mining is human learning.

5.2 Outlook

5.2.1 Further development of the presented research

First of all, the individual components developed in this thesis can be improved. Satellite-based flood masks should be accompanied by spatial uncertainty or validity layers, indicating where the sensor did not obtain a useful signal. The extrapolation procedure by OCC offers at least two additional opportunities: it can be used for merging satellite observations with data from social media or street cameras, and it might be tested whether the area between several satellite images can be interpolated. Derivation of water depth from the modelled flood extent is the next logical step. Procedures exist, but they need to be adjusted to handle the uncertainty associated with the land-water boundary. Probabilistic methods might be useful in this respect. This information could then be used for rapid damage estimation.

Hydrodynamic models could make use of improved satellite-based flood masks for calibration as well. Besides this, they need to incorporate additional damage-relevant processes such as dike failures and sediment transport. Damage models consequently need to be adjusted as well to capture these processes. Substituting open channel hydrodynamic models by ML does not seem logical, as the physical equations are known. This might be different for precipitation, subsurface hydrology, and damage models. The merging of both worlds by physically-constrained ML is worth investigating.

A direct shortcut from rainfall and environmental descriptors to damage via ML can work if comprehensive data is available. The clear advantage is that the complexity of model design is outsourced to a learning algorithm. Data of a single event is not sufficient to evaluate model predictions, though. The concept of damage probability maps might be further studied in regions where data from multiple events is available. The next attempts in that direction should definitely include characteristics of the damaged buildings, such as construction material or number of storeys. For the 2017 *El Niño* event in Peru, this kind of data was not made publicly available, although it has been recorded. There is also potential for improvement of models by using even more complex algorithms. Particularly the spatial context should be considered in geospatial modelling, therefore the use of Convolutional Neural Networks (CNNs) is encouraged. Last but not least, the target label in damage modelling is ambiguous: damage to buildings needs to be quantified during data collection either in categories of structural damage, or in terms of value loss. The first may be up to subjective judgement, the latter requires an estimation of pre- and post-disaster building value, or a comprehensive list of repair costs. It is clear that both options are fuzzy. Future research should consider methods that deal with this ambiguity in training.

5.2.2 Data is key

Obviously, but noteworthy, data sciences approaches first and foremost require data. The fields that have really been revolutionized by ML are those fields where truly comprehensive data is available, for example computer vision or natural language processing. Due to open data from satellite missions, and other large-scale geospatial mapping projects, we are approaching this point for environmental monitoring, including hazard observations. However, the field of damage modelling is still far from that stage. Systematic data collection, compilation, harmonization, and potentially data augmentation is needed to advance the development of generalized models. Open

data is necessary to break the silos of research and establish internationally accepted standard references. Surveys such as conducted in Peru provide the opportunity to apply strong learning algorithms for modelling and process understanding, but such surveys are rare. Data collection on the required scale is clearly beyond the capacity of individual scientists. Conceivable options are either that surveys are conducted by state authorities or insurance companies in labor-intensive fieldwork, during disaster response and reconstruction, or that affected people report the status of their buildings and potentially other categories of impact in an online database of standardized format, similar to the OSM project. The response rate of such a voluntary mapping project would probably depend on the trust in the issuing organization. Therefore a carefully designed information campaign might be needed in addition.

5.3 Conclusion

The overall aim of this thesis was to investigate the potential and limitations of open geospatial data and machine learning (ML) methods for improvement of flood hazard observation, damage modelling, process understanding and for substituting or complementing process-based risk model chains, particularly in the case of compound events. Plenty of hazard-related geospatial information is openly available, including digital elevation models (DEM), hydrography, land cover, precipitation, roads, building footprints, and flood extents mapped by satellites. The latter are unfortunately of limited quality in vegetated and urban areas, as shown for the case of hurricane Harvey in Houston. A procedure for significantly improving these products by a secondary ML step has been developed in this thesis. The procedure treats the data as positive and unlabeled (PU) and uses a one class classifier (OCC) to re-learn the flood extent from different features. Most important was topographic and hydrographic information, but the workflow is independent of the specific data and could in the future be used for merging satellite data with other sources of flood mapping. The opportunities for improving hazard observations by ML are bright, as operational data collection is already quite advanced. The focus here should be on designing creative workflows to merge different data sources, and on testing cutting-edge algorithms.

In principle, there is no need to substitute process models by ML if the physical equations governing the respective process are known. Such is theoretically the case for hydrodynamics. Also process-based flood risk model chains still need to be improved to include more damage-relevant processes, though. A continuous routing approach and consideration of hydrodynamic interactions along the entire river system have led to a reduction of modelled expected annual damage (EAD) by about 50%, indicating that such process description details cannot be neglected. Still, there are more aspects that need to be included to reproduce real events, such as dike failure and sediment transport. Accordingly, process-based model chains can get very complex and associated with high uncertainty, which poses the question whether such model chains could turn out to be unfeasible for compound or cascading events.

Impacts like building damage might therefore be modelled more directly by ML algorithms that learn complex relationships from empirical data, if the necessary amount of training data for that purpose is available. Compound damage modelling for an *El Niño* flood event in Peru has been tested by 5 different supervised ML algorithms trained on open data and a nationwide comprehensive damage survey. Performance was in the range that has been reported in literature for single hazard damage models. Damage data could be separated in thematic clusters of different dominant processes

by unsupervised methods. The ML-based damage estimation is still far from the level of insurance-type risk estimates obtained from a process-based model chain, but the method could be used for risk communication and potentially for spatial planning. Derivation of process understanding from these models was attempted by rigorous model inspection. Key thematic results include the finding that building collapse was not associated to intensity of rainfall, but rather low height above the nearest drainage (HAND) or location in a canyon were informative. Also vulnerability appeared higher in rural areas as opposed to urban centers. The topographic wetness index (TWI) was selected as the overall most informative predictive feature. Such insights may appear rather coarse, which reflects the spatial resolution of the open topographic data in Peru, and challenges in visualizing the full quantitative model results. Studies aiming at process understanding require high standards in data acquisition. Simple importance rankings by individual algorithms trained on arbitrary data are not to be trusted in a generalized interpretation. Comprehensive damage data such as used in this thesis is quite exceptional. Scientists and practitioners alike have to deal with what is available – still, in the domain of damage modelling it appears recommendable to direct more resources towards the collection, compilation and harmonization of data. In the long run, risk assessments from hazard to damage estimation should stand on a robust empirical basis, so that risk management decisions can be taken with confidence.

Bibliography

Abdiansah, A. and Wardoyo, R. (2015). Time complexity analysis of support vector machines (SVM) in LibSVM, *International Journal of Computer Applications* **128**(3): 28–34. https://doi.org/10.5120/ijca2015906480

Aksha, S. K., Resler, L. M., Juran, L. and Carstensen, L. W. (2020). A geospatial analysis of multi-hazard risk in dharan, nepal, *Geomatics, Natural Hazards and Risk* **11**(1): 88–111. https://doi.org/10.1080/19475705.2019.1710580

Alfieri, L., Feyen, L., Salamon, P., Thielen, J., Bianchi, A., Dottori, F. and Burek, P. (2016). Modelling the socio-economic impact of river floods in europe, *Natural Hazards and Earth System Sciences* **16**(6): 1401–1411. https://doi.org/10.5194/nhess-16-1401-2016

Alfieri, L., Salamon, P., Bianchi, A., Neal, J., Bates, P. and Feyen, L. (2013). Advances in pan-european flood hazard mapping, *Hydrological Processes* **28**(13): 4067–4077. https://doi.org/10.1002/hyp.9947

Alipour, A., Ahmadalipour, A., Abbaszadeh, P. and Moradkhani, H. (2020). Leveraging machine learning for predicting flash flood damage in the southeast US, *Environmental Research Letters* **15**(2): 024011. https://doi.org/10.1088/1748-9326/ab6edd

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J. (1999). OPTICS, *Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD '99*, ACM Press, p. 49–60. https://doi.org/10.1145/304182.304187

Apel, H., Aronica, G. T., Kreibich, H. and Thieken, A. H. (2008). Flood risk analyses—how detailed do we need to be?, *Natural Hazards* **49**(1): 79–98. https://doi.org/10.1007/s11069-008-9277-8

Apel, H., Merz, B. and Thieken, A. (2009). Influence of dike breaches on flood frequency estimation, *Computers & Geosciences* **35**(5): 907–923. https://doi.org/10.1016/j.cageo.2007.11.003

Apel, H., Thieken, A. H., Merz, B. and Blöschl, G. (2004). Flood risk assessment and associated uncertainty, *Natural Hazards and Earth System Sciences* **4**(2): 295–308. https://doi.org/10.5194/nhess-4-295-2004

Arundel, S., Phillips, L., Lowe, A., Bobinmyer, J., Mantey, K., Dunn, C., Constance, E. and Usery, E. (2015). Preparing the national map for the 3d elevation program – products, process and research, *Cartography and Geographic Information Science* **42**(sup1): 40–53. https://doi.org/10.1080/15230406.2015.1057229

Aschbacher, J. and Milagro-Pérez, M. P. (2012). The european earth monitoring (GMES) programme: Status and perspectives, *Remote Sensing of Environment* **120**: 3–8. https://doi.org/10.1016/j.rse.2011.08.028

Bates, P. D., Horritt, M. S. and Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling, *Journal of Hydrology* **387**(1-2): 33–45. https://doi.org/10.1016/j.jhydrol.2010.03.027

Bégué, A., Leroux, L., Soumaré, M., Faure, J.-F., Diouf, A. A., Augusseau, X., Touré, L. and Tonneau, J.-P. (2020). Remote sensing products and services in support of agricultural public policies in africa: Overview and challenges, *Frontiers in Sustainable Food Systems* **4**. https://doi.org/10.3389/fsufs.2020.00058

Biancamaria, S., Bates, P. D., Boone, A. and Mognard, N. M. (2009). Large-scale coupled hydrologic and hydraulic modelling of the ob river in siberia, *Journal of Hydrology* **379**(1-2): 136–150. https://doi.org/10.1016/j.jhydrol.2009.09.054

Blair, G. S., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S. and Young, P. J. (2019). Data science of the natural environment: A research roadmap, *Frontiers in Environmental Science* **7**. https://doi.org/10.3389/fenvs.2019.00121

Blazkova, S. and Beven, K. (2004). Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the czech republic, *Journal of Hydrology* **292**(1-4): 153–172. https://doi.org/10.1016/j.jhydrol.2003.12.025

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K. and Živković, N. (2019). Changing climate both increases and decreases european river floods, *Nature* **573**(7772): 108–111. https://doi.org/10.1038/s41586-019-1495-6

Bonafilia, D., Tellman, B., Anderson, T. and Issenberg, E. (2020). Sen1floods11: a georeferenced dataset to train and test deep learning flood algorithms for sentinel-1, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, pp. 835–845. https://doi.org/10.1109/cvprw50498.2020.00113

Bouwer, L. M. (2011). Have disaster losses increased due to anthropogenic climate change?, *Bulletin of the American Meteorological Society* **92**(1): 39–46. https://doi.org/10.1175/2010bams3092.1

Bradbrook, K., Waller, S. and Morris, D. (2005). National floodplain mapping: Datasets and methods – 160,000 km in 12 months, *Natural Hazards* **36**(1-2): 103–123. https://doi.org/10.1007/s11069-004-4544-9

Breiman, L. (2001a). Random forests, *Machine Learning* **45**(1): 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical Science* **16**(3). https://doi.org/10.1214/ss/1009213726

Bronstert, A., Agarwal, A., Boessenkool, B., Crisologo, I., Fischer, M., Heistermann, M., Köhn-Reich, L., López-Tarazón, J. A., Moran, T., Ozturk, U., Reinhardt-Imjela, C. and Wendi, D. (2018). Forensic hydro-meteorological analysis of an extreme flash flood:

The 2016-05-29 event in braunsbach, SW germany, *Science of The Total Environment* **630**: 977–991. https://doi.org/10.1016/j.scitotenv.2018.02.241

Bronstert, A., Crisologo, I., Heistermann, M., Ozturk, U., Vogel, K. and Wendi, D. (2020). Flash-floods: More often, more severe, more damaging? an analysis of hydro-geo-environmental conditions and anthropogenic impacts, *Climate Change Management*, Springer International Publishing, pp. 225–244. https://doi.org/10.1007/978-3-030-37425-9_12

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W. and Clark, M. P. (2020). Future streamflow regime changes in the united states: assessment using functional classification, *Hydrology and Earth System Sciences* **24**(8): 3951–3966. https://doi.org/10.5194/hess-24-3951-2020

Brutsaert, W. (2014). *Hydrology*, Cambridge University Press, University Printing House, Cambridge CB2 8BS, United Kingdom. URL: https://www.ebook.de/de/product/3270950/wilfried_brutsaert_hydrology.html

Bubeck, P., de Moel, H., Bouwer, L. M. and Aerts, J. C. J. H. (2011). How reliable are projections of future flood damage?, *Natural Hazards and Earth System Sciences* **11**(12): 3293–3306. https://doi.org/10.5194/nhess-11-3293-2011

Bubeck, P., Otto, A. and Weichselgartner, J. (2017). Societal impacts of flood hazards, *Oxford Research Encyclopedia of Natural Hazard Science*, Oxford University Press. https://doi.org/10.1093/acrefore/9780199389407.013.281

Cai, W., Santoso, A., Wang, G., Yeh, S.-W., An, S.-I., Cobb, K. M., Collins, M., Guilyardi, E., Jin, F.-F., Kug, J.-S., Lengaigne, M., McPhaden, M. J., Takahashi, K., Timmermann, A., Vecchi, G., Watanabe, M. and Wu, L. (2015). ENSO and greenhouse warming, *Nature Climate Change* **5**(9): 849–859. https://doi.org/10.1038/nclimate2743

Cai, W., Wang, G., Dewitte, B., Wu, L., Santoso, A., Takahashi, K., Yang, Y., Carréric, A. and McPhaden, M. J. (2018). Increased variability of eastern pacific el niño under greenhouse warming, *Nature* **564**(7735): 201–206. https://doi.org/10.1038/s41586-018-0776-9

Calliari, E., Serdeczny, O. and Vanhala, L. (2020). Making sense of the politics in the climate change loss & damage debate, *Global Environmental Change* **64**: 102133. https://doi.org/10.1016/j.gloenvcha.2020.102133

Cao, C., Xu, P., Wang, Y., Chen, J., Zheng, L. and Niu, C. (2016). Flash flood hazard susceptibility mapping using frequency ratio and statistical index methods in coalmine subsidence areas, *Sustainability* **8**(9): 948. https://doi.org/10.3390/su8090948

Cerri, M., Steinhausen, M., Kreibich, H. and Schröter, K. (2021). Are OpenStreetMap building data useful for flood vulnerability modelling?, *Natural Hazards and Earth System Sciences* **21**(2): 643–662. https://doi.org/10.5194/nhess-21-643-2021

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**: 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., Bui, D. T., Pham, B. T. and Khosravi, K. (2017). A novel hybrid artificial intelligence approach for flood susceptibility assessment, *Environmental Modelling & Software* **95**: 229–245. https://doi.org/10.1016/j.envsoft.2017.06.012

Chini, M., Pelich, R., Pulvirenti, L., Pierdicca, N., Hostache, R. and Matgen, P. (2019). Sentinel-1 InSAR coherence to detect floodwater in urban areas: Houston and hurricane harvey as a test case, *Remote Sensing* **11**(2): 107. https://doi.org/10.3390/rs11020107

Chirachawala, C., Shrestha, S., Babel, M. S., Virdis, S. G. and Wichakul, S. (2020). Evaluation of global land use/land cover products for hydrologic simulation in the upper yom river basin, thailand, *Science of the Total Environment* **708**: 135148. https://doi.org/10.1016/j.scitotenv.2019.135148

Chung, H.-W., Liu, C.-C., Cheng, I.-F., Lee, Y.-R. and Shieh, M.-C. (2015). Rapid response to a typhoon-induced flood with an SAR-derived map of inundated areas: Case study and validation, *Remote Sensing* **7**(9): 11954–11973. https://doi.org/10.3390/rs70911954

Cian, F., Marconcini, M., Ceccato, P. and Giupponi, C. (2018). Flood depth estimation by means of high-resolution SAR images and lidar data, *Natural Hazards and Earth System Sciences* **18**(11): 3063–3084. https://doi.org/10.5194/nhess-18-3063-2018

Ciullo, A., de Bruijn, K. M., Kwakkel, J. H. and Klijn, F. (2019). Accounting for the uncertain effects of hydraulic interactions in optimising embankments heights: Proof of principle for the IJssel river, *Journal of Flood Risk Management* **12**(S2). https://doi.org/10.1111/jfr3.12532

Clair, J. S., Moon, S., Holbrook, W. S., Perron, J. T., Riebe, C. S., Martel, S. J., Carr, B., Harman, C., Singha, K. and d. Richter, D. (2015). Geophysical imaging reveals topographic stress control of bedrock weathering, *Science* **350**(6260): 534–538. https://doi.org/10.1126/science.aab2210

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W. and Peters-Lidard, C. D. (2017). The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, *Hydrology and Earth System Sciences* **21**(7): 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Clinton, W. (2006). Lessons learned from tsunami recovery: key propositions for building back better, *Technical report*, Office of the UN Secretary-General's Special Envoy for Tsunami Recovery, New York, USA. URL: https://reliefweb.int/report/thailand/lessons-learned-tsunami-recovery-key-propositions-building-back-better

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1): 37–46. https://doi.org/10.1177/001316446002000104

Cohen, S., Brakenridge, G. R., Kettner, A., Bates, B., Nelson, J., McDonald, R., Huang, Y.-F., Munasinghe, D. and Zhang, J. (2017). Estimating floodwater depths from flood inundation maps and topography, *JAWRA Journal of the American Water Resources Association* **54**(4): 847–858. https://doi.org/10.1111/1752-1688.12609

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3): 273–297. https://doi.org/10.1007/bf00994018

Couasnon, A., Eilander, D., Muis, S., Veldkamp, T. I. E., Haigh, I. D., Wahl, T., Winsemius, H. C. and Ward, P. J. (2020). Measuring compound flood potential from river discharge and storm surge extremes at the global scale, *Natural Hazards and Earth System Sciences* **20**(2): 489–504. https://doi.org/10.5194/nhess-20-489-2020

CRED and Guha-Sapir, D. (2020). Em-dat: The emergency events database - université catholique de louvain (ucl), online. URL: www.emdat.be

Cunge, J. A. (1969). On the subject of a flood propagation computation method (musklngum method), *Journal of Hydraulic Research* **7**(2): 205–230. https://doi.org/10.1080/00221686909500264

Curran, A., de Bruijn, K., Klerk, W. and Kok, M. (2019). Large scale flood hazard analysis by including defence failures on the dutch river system, *Water* **11**(8): 1732. https://doi.org/10.3390/w11081732

Dankowski, T. and Ziegler, A. (2016). Calibrating random forests for probability estimation, *Statistics in Medicine* **35**(22): 3949–3960. https://doi.org/10.1002/sim.6959

Dasgupta, A., Hostache, R., Ramsankaran, R., Schumann, G. J.-P., Grimaldi, S., Pauwels, V. R. N. and Walker, J. P. (2021). On the impacts of observation location, timing, and frequency on flood extent assimilation performance, *Water Resources Research* **57**(2). https://doi.org/10.1029/2020wr028238

de Bruijn, K. M., Diermanse, F. L. M. and Beckers, J. V. L. (2014). An advanced method for flood risk analysis in river deltas, applied to societal flood fatality risk in the netherlands, *Natural Hazards and Earth System Sciences* **14**(10): 2767–2781. https://doi.org/10.5194/nhess-14-2767-2014

de Moel, H., Jongman, B., Kreibich, H., Merz, B., Penning-Rowsell, E. and Ward, P. J. (2015). Flood risk assessments at different spatial scales, *Mitigation and Adaptation Strategies for Global Change* **20**(6): 865–890. https://doi.org/10.1007/s11027-015-9654-z

de Paiva, R. C. D., Buarque, D. C., Collischonn, W., Bonnet, M.-P., Frappart, F., Calmant, S. and Mendes, C. A. B. (2013). Large-scale hydrologic and hydrodynamic modeling of the amazon river basin, *Water Resources Research* **49**(3): 1226–1243. https://doi.org/10.1002/wrcr.20067

Debeer, D. and Strobl, C. (2020). Conditional permutation importance revisited, *BMC Bioinformatics* **21**(1). https://doi.org/10.1186/s12859-020-03622-2

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D. and Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography* **36**(1): 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Duan, Q., Sorooshian, S. and Gupta, V. K. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology* **158**(3-4): 265–284. https://doi.org/10.1016/0022-1694(94)90057-4

Dupuits, E., Klerk, W., Schweckendiek, T. and de Bruijn, K. (2019). Impact of including interdependencies between multiple riverine flood defences on the economically optimal flood safety levels, *Reliability Engineering & System Safety* **191**: 106475. https://doi.org/10.1016/j.ress.2019.04.028

Dwyer, J., Roy, D., Sauer, B., Jenkerson, C., Zhang, H. and Lymburner, L. (2018). Analysis ready data: Enabling analysis of the landsat archive, *Remote Sensing* **10(9)**: 1363. https://doi.org/10.20944/preprints201808.0029.v1

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. and Yates, C. J. (2010). A statistical explanation of MaxEnt for ecologists, *Diversity and Distributions* **17**(1): 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

ENFEN (2017). Extraordinary technical report n 001-2017/enfen, *Technical report*, Comisión Multisectorial Encargada del Estudio nacional del Fenómeno El Niño. URL: www.imarpe.pe/imarpe/archivos/informes/imarpe_inftco_informe_tecnico_extraordinario_001_2017.pdf

Engel, H. (1997). The flood events of 1993/1994 and 1995 in the rhine river basin, *IAHS-AISH publication* pp. 21–32.

Erazo, B., Bourrel, L., Frappart, F., Chimborazo, O., Labat, D., Dominguez-Granda, L., Matamoros, D. and Mejia, R. (2018). Validation of satellite estimates (tropical rainfall measuring mission, TRMM) for rainfall variability over the pacific slope and coast of ecuador, *Water* **10**(2): 213. https://doi.org/10.3390/w10020213

Esch, T., Taubenböck, H., Roth, A., Heldens, W., Felbier, A., Thiel, M., Schmidt, M., Müller, A. and Dech, S. (2012). TanDEM-x mission—new perspectives for the inventory and monitoring of global settlement patterns, *Journal of Applied Remote Sensing* **6**(1): 061702–1. https://doi.org/10.1117/1.jrs.6.061702

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., *Kdd*, Vol. 96, pp. 226–231.

Ettinger, S., Mounaud, L., Magill, C., Yao-Lafourcade, A.-F., Thouret, J.-C., Manville, V., Negulescu, C., Zuccaro, G., Gregorio, D. D., Nardone, S., Uchuchoque, J. A. L., Arguedas, A., Macedo, L. and Llerena, N. M. (2016). Building vulnerability to hydro-geomorphic hazards: Estimating damage probability from qualitative vulnerability assessment using logistic regression, *Journal of Hydrology* **541**: 563–581. https://doi.org/10.1016/j.jhydrol.2015.04.017

Evans, J. S. (2021). *spatialEco*. R package version 1.3-6. URL: https://github.com/jeffreyevans/spatialEco

Falter, D., Nguyen, D. V., Vorogushyn, S., Schröter, K., Hundecha, Y., Kreibich, H., Apel, H., Theisselmann, F. and Merz, B. (2016). Continuous, large-scale simulation model for flood risk assessments: proof-of-concept, *Journal of Flood Risk Management* **9**(1): 3–21. https://doi.org/10.1111/jfr3.12105

Falter, D., Schröter, K., Dung, N. V., Vorogushyn, S., Kreibich, H., Hundecha, Y., Apel, H. and Merz, B. (2015). Spatially coherent flood risk assessment based on long-term continuous simulation with a coupled model chain, *Journal of Hydrology* **524**: 182–193. https://doi.org/10.1016/j.jhydrol.2015.02.021

Falter, D., Vorogushyn, S., Lhomme, J., Apel, H., Gouldby, B. and Merz, B. (2013). Hydraulic model evaluation for large-scale flood risk assessments, *Hydrological Processes* **27**(9): 1331–1340. https://doi.org/10.1002/hyp.9553

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters* **27**(8): 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fedeski, M. and Gwilliam, J. (2007). Urban sustainability in the presence of flood and geological hazards: The development of a GIS-based vulnerability and risk assessment methodology, *Landscape and Urban Planning* **83**(1): 50–61. https://doi.org/10.1016/j.landurbplan.2007.05.012

Fernandez, G. and Ahmed, I. (2019). "build back better" approach to disaster recovery: Research trends since 2006, *Progress in Disaster Science* **1**: 100003. https://doi.org/10.1016/j.pdisas.2019.100003

Fischer, E. M. and Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models, *Nature Climate Change* **6**(11): 986–991. https://doi.org/10.1038/nclimate3110

Flanagin, A. J. and Metzger, M. J. (2008). The credibility of volunteered geographic information, *GeoJournal* **72**(3-4): 137–148. https://doi.org/10.1007/s10708-008-9188-y

Fleischmann, A., Collischonn, W., Paiva, R. and Tucci, C. E. (2019). Modeling the role of reservoirs versus floodplains on large-scale river hydrodynamics, *Natural Hazards* **99**(2): 1075–1104. https://doi.org/10.1007/s11069-019-03797-9

Formetta, G. and Feyen, L. (2019). Empirical evidence of declining global vulnerability to climate-related hazards, *Global Environmental Change* **57**: 101920. https://doi.org/10.1016/j.gloenvcha.2019.05.004

Fread, D. L. (1976). Theoretical development of an implicit dynamic routing model, *Hydrologic Research Laboratory, Office of Hydrology, U.S. Department of Commerce, NOAA, NWS, Silver Spring, MD., presented at Dynamic Routing Seminar, Lower Mississippi River Forecast Center, Slidell, LA., 13-17 Dec 1976* .

Freund, M. B., Henley, B. J., Karoly, D. J., McGregor, H. V., Abram, N. J. and Dommenget, D. (2019). Higher frequency of central pacific el niño events in recent decades relative to past centuries, *Nature Geoscience* **12**(6): 450–455. https://doi.org/10.1038/s41561-019-0353-3

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine., *The Annals of Statistics* **29**(5): 1189–1232. https://doi.org/10.1214/aos/1013203451

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F. and Ureña-López, L. A. (2021). A survey on bias in deep NLP, *Appl. Sci.* **11**(7): 3184. https://doi.org/10.3390/app11073184

Geiger, C. P. and Lucke, J. V. (2012). Open government and (linked) (open) (government) (data), *JeDEM - eJournal of eDemocracy and Open Government* **4**(2): 265–278. https://doi.org/10.29379/jedem.v4i2.143

Geiß, C., Leichtle, T., Wurm, M., Pelizari, P. A., Standfus, I., Zhu, X. X., So, E., Siedentop, S., Esch, T. and Taubenbock, H. (2019). Large-area characterization of urban morphology—mapping of built-up height and density using TanDEM-x and sentinel-2 data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**(8): 2912–2927. https://doi.org/10.1109/jstars.2019.2917755

Geiß, C., Thoma, M., Pittore, M., Wieland, M., Dech, S. W. and Taubenbock, H. (2017). Multitask active learning for characterization of built environments with multisensor earth observation data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **10**(12): 5583–5597. https://doi.org/10.1109/jstars.2017.2748339

Gerl, T., Kreibich, H., Franco, G., Marechal, D. and Schröter, K. (2016). A review of flood loss models as basis for harmonization and benchmarking, *PLOS ONE* **11**(7): e0159791. https://doi.org/10.1371/journal.pone.0159791

German Aerospace Center (DLR) (2018). Tandem-x - digital elevation model (dem) - global, 90m, https://doi.org/10.15489/ju28hc7pui09.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Sebastopol, CA.

Ghonchepour, D., Sadoddin, A., Bahremand, A., Croke, B., Jakeman, A. and Salmanmahiny, A. (2021). A methodological framework for the hydrological model selection process in water resource management projects, *Natural Resource Modeling* **34**(3). https://doi.org/10.1111/nrm.12326

Giuliani, G., Camara, G., Killough, B. and Minchin, S. (2019). Earth observation open science: Enhancing reproducible science using data cubes, *Data* **4**(4): 147. https://doi.org/10.3390/data4040147

Giustarini, L., Hostache, R., Matgen, P., Schumann, G. J.-P., Bates, P. D. and Mason, D. C. (2013). A change detection approach to flood mapping in urban areas using TerraSAR-x, *IEEE Transactions on Geoscience and Remote Sensing* **51**(4): 2417–2430. https://doi.org/10.1109/tgrs.2012.2210901

Giustarini, L., Vernieuwe, H., Verwaeren, J., Chini, M., Hostache, R., Matgen, P., Verhoest, N. and Baets, B. D. (2015). Accounting for image uncertainty in SAR-based flood mapping, *International Journal of Applied Earth Observation and Geoinformation* **34**: 70–77. https://doi.org/10.1016/j.jag.2014.06.017

Gomez-Zapata, J. C., Parrado, C., Frimberger, T., Barragán-Ochoa, F., Brill, F., Büche, K., Krautblatter, M., Langbein, M., Pittore, M., Rosero-Velásquez, H., Schoepfer, E., Spahn, H. and Zapata-Tapia, C. (2021). Community perception and communication of volcanic risk from the cotopaxi volcano in latacunga, ecuador, *Sustainability* **13**(4): 1714. https://doi.org/10.3390/su13041714

Gomez-Zapata, J. C., Pittore, M., Brinckmann, N. and Shinde, S. (2020). Dynamic physical vulnerability: a multi-risk scenario approach from building- single- hazard fragility- models, *EGU2020-18379* . https://doi.org/10.5194/egusphere-egu2020-18379

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone, *Remote Sensing of Environment* **202**: 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Government of Peru (2017). Autorizan transferencia de partidas en el presupuesto del sector público para el año fiscal 2017 a favor del pliego organismo de formalización de la propiedad informal (in spanish), Online. URL: http://www3.vivienda.gob.pe/popup/Documentos_pdf/DS-287-2017-EF.pdf

Greiving, S., Fleischhauer, M. and Lückenkötter, J. (2006). A methodology for an integrated risk assessment of spatially relevant hazards, *Journal of Environmental Planning and Management* **49**(1): 1–19. https://doi.org/10.1080/09640560500372800

Grimaldi, S., Petroselli, A., Arcangeletti, E. and Nardi, F. (2013). Flood mapping in ungauged basins using fully continuous hydrologic–hydraulic modeling, *Journal of Hydrology* **487**: 39–47. https://doi.org/10.1016/j.jhydrol.2013.02.023

Guillera-Arroita, G., Lahoz-Monfort, J. J. and Elith, J. (2014). Maxent is not a presence-absence method: a comment on thibaud et al., *Methods in Ecology and Evolution* **5**(11): 1192–1197. https://doi.org/10.1111/2041-210x.12252

Haberlandt, U. and Radtke, I. (2014). Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows, *Hydrology and Earth System Sciences* **18**(1): 353–365. https://doi.org/10.5194/hess-18-353-2014

Hall, J. W., Brown, S., Nicholls, R. J., Pidgeon, N. F. and Watson, R. T. (2012). Proportionate adaption, *Nature Climate Change* **2**: 833–834.

Hamel, P. and Tan, L. (2021). Blue–green infrastructure for flood and water quality management in southeast asia: Evidence and knowledge gaps, *Environmental Management* . https://doi.org/10.1007/s00267-021-01467-w

Hand, D. J. and Yu, K. (2001). Idiot's bayes? not so stupid after all?, *International Statistical Review* **69**(3): 385–398. https://doi.org/10.1111/j.1751-5823.2001.tb00465.x

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve., *Radiology* **143**(1): 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hart, J. K. and Martinez, K. (2006). Environmental sensor networks: A revolution in the earth system science?, *Earth-Science Reviews* **78**(3-4): 177–191. https://doi.org/10.1016/j.earscirev.2006.05.001

Haylock, M. R., Hofstra, N., Tank, A. M. G. K., Klok, E. J., Jones, P. D. and New, M. (2008). A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, *Journal of Geophysical Research* **113**(D20). https://doi.org/10.1029/2008jd010201

HCFCD (2018). Hurricane harvey: Impact and response in harris county, *Technical report*, Harris County Flood Control District. URL: https://www.hcfcd.org/Portals/62/Harvey/harvey-impact-and-response-book-final-re.pdf

Hegnauer, M., Beersma, J., van den Boogaard, H., Buishand, T. and R., P. (2014). Generator of rainfall and discharge extremes (grade) for the rhine and meuse basins. final report of grade 2.0, *Technical report*, Deltares, Delft. URL: https://cdn.knmi.nl/system/data_center_publications/files/000/069/731/original/1209424004zws0018rgenerator_of_rainfall_and_discharge_extremes_grade_for_the_rhine_and_meuse_basins_definitief.pdf?1495621963

Henderson, F. M. and Lewis, A. J. (2008). Radar detection of wetland ecosystems: a review, *International Journal of Remote Sensing* **29**(20): 5809–5835. https://doi.org/10.1080/01431160801958405

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S. and Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning, *PLOS ONE* **12**(2): e0169748. https://doi.org/10.1371/journal.pone.0169748

Hirabayashi, Y., Alifu, H., Yamazaki, D., Imada, Y., Shiogama, H. and Kimura, Y. (2021). Anthropogenic climate change has changed frequency of past flood during 2010-2013, *Progress in Earth and Planetary Science* **8**(1). https://doi.org/10.1186/s40645-021-00431-w

Hoch, J. M., Neal, J. C., Baart, F., van Beek, R., Winsemius, H. C., Bates, P. D. and Bierkens, M. F. P. (2017). GLOFRIM v1.0 – a globally applicable computational framework

for integrated hydrological–hydrodynamic modelling, *Geoscientific Model Development* **10**(10): 3913–3929. https://doi.org/10.5194/gmd-10-3913-2017

Horritt, M. S., Mason, D. C. and Luckman, A. J. (2001). Flood boundary delineation from synthetic aperture radar imagery using a statistical active contour model, *International Journal of Remote Sensing* **22**(13): 2489–2507. https://doi.org/10.1080/01431160116902

Hostache, R., Chini, M., Giustarini, L., Neal, J., Kavetski, D., Wood, M., Corato, G., Pelich, R.-M. and Matgen, P. (2018). Near-real-time assimilation of SAR-derived flood maps for improving flood forecasts, *Water Resources Research* **54**(8): 5516–5535. https://doi.org/10.1029/2017wr022205

Hua, T., Zhao, W., Liu, Y., Wang, S. and Yang, S. (2018). Spatial consistency assessments for global land-cover datasets: A comparison among GLC2000, CCI LC, MCD12, GLOBCOVER and GLCNMO, *Remote Sensing* **10**(11): 1846. https://doi.org/10.3390/rs10111846

Huang, C., Nguyen, B. D., Zhang, S., Cao, S. and Wagner, W. (2017). A comparison of terrain indices toward their ability in assisting surface water mapping from sentinel-1 data, *ISPRS International Journal of Geo-Information* **6**(5): 140. https://doi.org/10.3390/ijgi6050140

Hundecha, Y., Pahlow, M. and Schumann, A. (2009). Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes, *Water Resources Research* **45**(12). https://doi.org/10.1029/2008wr007453

ICPR (2001). Atlas 2001. atlas of flood danger and potential damage due to extreme floods of the rhine, International Commission on the Protection of the Rhine, Koblenz, Germany. http://www.iksr.org/.

ICPR (2013). The rhine and its catchment: An overview, *Technical report*, International Commission for the Protection of the Rhine. URL: https://www.iksr.org/fileadmin/user_upload/DKDM/Dokumente/Broschueren/EN/bro_En_2013_The_Rhine_and_its_catchment.pdf

INDECI (2019). Statistical compendium of emergencies 2019, *Technical report*, Instituto Nacional de Defensa Civil, Lima, Peru.

INEI (2018). Definitive results of the census 2017. URL: www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1544

Islam, A. R. M. T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K. U., Pham, Q. B., Kuriqi, A. and Linh, N. T. T. (2021). Flood susceptibility modelling using advanced ensemble machine learning models, *Geoscience Frontiers* **12**(3): 101075. https://doi.org/10.1016/j.gsf.2020.09.006

Jakob, M., Stein, D. and Ulmi, M. (2011). Vulnerability of buildings to debris flow impact, *Natural Hazards* **60**(2): 241–261. https://doi.org/10.1007/s11069-011-0007-2

Jalayer, F., Carozza, S., Risi, R. D., Manfredi, G. and Mbuya, E. (2016). Performance-based flood safety-checking for non-engineered masonry structures, *Engineering Structures* **106**: 109–123. https://doi.org/10.1016/j.engstruct.2015.10.007

Jalayer, F., Risi, R. D., Paola, F. D., Giugni, M., Manfredi, G., Gasparini, P., Topa, M. E., Yonas, N., Yeshitela, K., Nebebe, A., Cavan, G., Lindley, S., Printz, A. and Renner, F.

(2014). Probabilistic GIS-based method for delineation of urban flooding risk hotspots, *Natural Hazards* . https://doi.org/10.1007/s11069-014-1119-2

Jaynes, E. T. (1957). Information theory and statistical mechanics, *Physical Review* **106**(4): 620–630. https://doi.org/10.1103/physrev.106.620

Jenkins, S. F., Phillips, J. C., Price, R., Feloy, K., Baxter, P. J., Hadmoko, D. S. and de Bélizal, E. (2015). Developing building-damage scales for lahars: application to merapi volcano, indonesia, *Bulletin of Volcanology* **77**(9). https://doi.org/10.1007/s00445-015-0961-8

Jewell, J. and Cherp, A. (2019). On the political feasibility of climate change mitigation pathways: Is it too late to keep warming below 1.5°c?, *WIREs Climate Change* **11**(1). https://doi.org/10.1002/wcc.621

Jiménez-Jiménez, S. I., Ojeda-Bustamante, W., Ontiveros-Capurata, R. E. and de Jesús Marcial-Pablo, M. (2020). Rapid urban flood damage assessment using high resolution remote sensing data and an object-based approach, *Geomatics, Natural Hazards and Risk* **11**(1): 906–927. https://doi.org/10.1080/19475705.2020.1760360

Johnson, R. W. (2016). *Handbook of Fluid Dynamics*, CRC Press. URL: https://doi.org/10.1201/b19031

Jongman, B., Hochrainer-Stigler, S., Feyen, L., Aerts, J. C. J. H., Mechler, R., Botzen, W. J. W., Bouwer, L. M., Pflug, G., Rojas, R. and Ward, P. J. (2014). Increasing stress on disaster-risk finance due to large floods, *Nature Climate Change* **4**(4): 264–268. https://doi.org/10.1038/nclimate2124

Jongman, B., Kreibich, H., Apel, H., Barredo, J. I., Bates, P. D., Feyen, L., Gericke, A., Neal, J., Aerts, J. C. J. H. and Ward, P. J. (2012). Comparative flood damage model assessment: towards a european approach, *Natural Hazards and Earth System Sciences* **12**(12): 3733–3752. https://doi.org/10.5194/nhess-12-3733-2012

Jongman, B., Winsemius, H. C., Aerts, J. C. J. H., de Perez, E. C., van Aalst, M. K., Kron, W. and Ward, P. J. (2015). Declining vulnerability to river floods and the global benefits of adaptation, *Proceedings of the National Academy of Sciences* **112**(18): E2271–E2280. https://doi.org/10.1073/pnas.1414439112

Jozani, H. J., Thiel, M., Abdel-Rahman, E. M., Richard, K., Landmann, T., Subramanian, S. and Hahn, M. (2020). Investigation of maize lethal necrosis (MLN) severity and cropping systems mapping in agro-ecological maize systems in bomet, kenya utilizing RapidEye and landsat-8 imagery, *Geology, Ecology, and Landscapes* pp. 1–16. https://doi.org/10.1080/24749508.2020.1761195

Kakinuma, K., Puma, M. J., Hirabayashi, Y., Tanoue, M., Baptista, E. A. and Kanae, S. (2020). Flood-induced population displacements in the world, *Environmental Research Letters* **15**(12): 124029. https://doi.org/10.1088/1748-9326/abc586

Kappes, M. S., Keiler, M., von Elverfeldt, K. and Glade, T. (2012). Challenges of analyzing multi-hazard risk: a review, *Nat Hazards* **64**(2): 1925–1958. https://doi.org/10.1007/s11069-012-0294-2

Kaspersen, P. S., Ravn, N. H., Arnbjerg-Nielsen, K., Madsen, H. and Drews, M. (2017). Comparison of the impacts of urban development and climate change on exposing european cities to pluvial flooding, *Hydrology and Earth System Sciences* **21**(8): 4131–4147. https://doi.org/10.5194/hess-21-4131-2017

Keef, C., Svensson, C. and Tawn, J. A. (2009). Spatial dependence in extreme river flows and precipitation for great britain, *Journal of Hydrology* **378**(3-4): 240–252. https://doi.org/10.1016/j.jhydrol.2009.09.026

Kelleher, C. and McPhillips, L. (2019). Exploring the application of topographic indices in urban areas as indicators of pluvial flooding locations, *Hydrological Processes* **34**(3): 780–794. https://doi.org/10.1002/hyp.13628

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resources Research* **42**(3). https://doi.org/10.1029/2005wr004362

Klaus, S., Kreibich, H., Merz, B., Kuhlmann, B. and Schröter, K. (2016). Large-scale, seasonal flood risk analysis for agricultural crops in germany, *Environmental Earth Sciences* **75**(18). https://doi.org/10.1007/s12665-016-6096-1

Kleist, L., Thieken, A. H., Köhler, P., Müller, M., Seifert, I., Borst, D. and Werner, U. (2006). Estimation of the regional stock of residential buildings as a basis for a comparative risk assessment in germany, *Natural Hazards and Earth System Sciences* **6**(4): 541–552. https://doi.org/10.5194/nhess-6-541-2006

Knight, F. H. (1921). *Risk, Uncertainty and Profit*, University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship. URL: https://ssrn.com/abstract=1496192

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M. (2018). Rainfall-runoff modelling using long-short-term-memory (LSTM) networks, *Hydrology and Earth System Sciences Discussions* . https://doi.org/10.5194/hess-2018-247

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence* **5**(4): 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kreibich, H., den Bergh, J. V., Bouwer, L., Bubeck, P., Ciavola, P., Green, C., Hallegatte, S., Logar, I., Meyer, V., Schwarze, R. and Thieken, A. (2014). Costing natural hazards, *Nature Climate Change* **4**: 303–306.

Kreibich, H., Hudson, P. and Merz, B. (2021). Knowing what to do substantially improves the effectiveness of flood early warning, *Bulletin of the American Meteorological Society* **102**(7): E1450–E1463. https://doi.org/10.1175/bams-d-20-0262.1

Kreibich, H., Müller, M., Schröter, K. and Thieken, A. H. (2017). New insights into flood warning reception and emergency response by affected parties, *Natural Hazards and Earth System Sciences* **17**(12): 2075–2092. https://doi.org/10.5194/nhess-17-2075-2017

Kreibich, H., Piroth, K., Seifert, I., Maiwald, H., Kunert, U., Schwarz, J., Merz, B. and Thieken, A. H. (2009). Is flow velocity a significant parameter in flood damage modelling?, *Natural Hazards and Earth System Science* **9**(5): 1679–1692. https://doi.org/10.5194/nhess-9-1679-2009

Kreibich, H., Seifert, I., Merz, B. and Thieken, A. H. (2010). Development of FLEMOcs – a new model for the estimation of flood losses in the commercial sector, *Hydrological Sciences Journal* **55**(8): 1302–1314. https://doi.org/10.1080/02626667.2010.529815

Kron, W. (2005). Flood risk = hazard • values • vulnerability, *Water International* **30**(1): 58–68. https://doi.org/10.1080/02508060508691837

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis*, Elsevier LTD, Oxford. URL: https://www.ebook.de/de/product/22836901/john_k_kruschke_doing_bayesian_data_analysis.html

Kruse, P. M., Schmitt, H. C. and Greiving, S. (2021). Systemic criticality—a new assessment concept improving the evidence basis for CI protection, *Climatic Change* **165**(1-2). https://doi.org/10.1007/s10584-021-03019-x

Krysanova, V., Müller-Wohlfeil, D.-I. and Becker, A. (1998). Development and test of a spatially distributed hydrological/water quality model for mesoscale watersheds, *Ecological Modelling* **106**(2-3): 261–289. https://doi.org/10.1016/s0304-3800(97)00204-4

Kuhlmann, B. (2010). Schäden in der landwirtschaft, *in* B. M. A. H. Thieken, I. Seifert (ed.), *Hochwasserschäden – Erfassung, Abschätzung und Vermeidung (in German)*, oekom, Munich.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., Dunning, P. and Batstone, C. (2010). A new method to assess the risk of local and widespread flooding on rivers and coasts, *Journal of Flood Risk Management* **3**(4): 323–336. https://doi.org/10.1111/j.1753-318x.2010.01081.x

Laudan, J., Rözer, V., Sieg, T., Vogel, K. and Thieken, A. H. (2017). Damage assessment in braunsbach 2016: data collection and analysis for an improved understanding of damaging processes during flash floods, *Natural Hazards and Earth System Sciences* **17**(12): 2163–2179. https://doi.org/10.5194/nhess-17-2163-2017

Laudan, J., Zöller, G. and Thieken, A. H. (2020). Flash floods versus river floods – a comparison of psychological impacts and implications for precautionary behaviour, *Natural Hazards and Earth System Sciences* **20**(4): 999–1023. https://doi.org/10.5194/nhess-20-999-2020

Laur, H. and Liebig, V. (2010). Earth observation satellite missions and data access, *Handbook of Geomathematics*, Springer Berlin Heidelberg, pp. 69–92. https://doi.org/10.1007/978-3-642-01546-5_3

Lavell, A., Oppenheimer, M., Diop, C., Hess, J., Lempert, R., Li, J., Muir-Wood, R., Myeong, S., Moser, S., Takeuchi, K., Cardona, O.-D., Hallegatte, S., Lemos, M., Little, C., Lotsch, A. and Weber, E. (2012). Climate change: New dimensions in disaster risk, exposure, vulnerability, and resilience, *in* C. B. Field, V. Barros, T. F. Stocker and Q. Dahe (eds), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, pp. 25–64. https://doi.org/10.1017/cbo9781139177245.004

Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, pp. 2278–2324.

Lee, W. S. and Liu, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC*, Vol. 20, pp. 448–455. URL: https://www.aaai.org/Papers/ICML/2003/ICML03-060.pdf

Leopold, L. B. and Maddock, T. (1953). The hydraulic geometry of stream channels and some physiographic implications, *Technical report*, US Geological Survey. URL: https://www.doi.org/10.3133/pp252

Li, W., Guo, Q. and Elkan, C. (2011). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data, *IEEE Transactions on Geoscience and Remote Sensing* **49**(2): 717–725. https://doi.org/10.1109/tgrs.2010.2058578

Li, Y., Martinis, S. and Wieland, M. (2019). Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-x intensity and interferometric coherence, *ISPRS Journal of Photogrammetry and Remote Sensing* **152**: 178–191. https://doi.org/10.1016/j.isprsjprs.2019.04.014

Li, Y., Martinis, S., Wieland, M., Schlaffer, S. and Natsuaki, R. (2019). Urban flood mapping using SAR intensity and interferometric coherence via bayesian network fusion, *Remote Sensing* **11**(19): 2231. https://doi.org/10.3390/rs11192231

Li, Z., Yang, C., Jin, B., Yu, M., Liu, K., Sun, M. and Zhan, M. (2015). Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework, *PLOS ONE* **10**(3): e0116781. https://doi.org/10.1371/journal.pone.0116781

Lindner, J. and Fitzgerald, S. (2018). Hurricane harvey - storm and flood information, *Technical report*, Harris County Flood Control District (HCFCD). URL: https://www.hcfcd.org/Portals/62/Harvey/immediate-flood-report-final-hurricane-harvey-2017.pdf

Liu, B., Dai, Y., Li, X., Lee, W. and Yu, P. (2003). Building text classifiers using positive and unlabeled examples, *Third IEEE International Conference on Data Mining*, IEEE Comput. Soc. https://doi.org/10.1109/icdm.2003.1250918

Loo, Y. Y., Billa, L. and Singh, A. (2015). Effect of climate change on seasonal monsoon in asia and its impact on the variability of monsoon rainfall in southeast asia, *Geoscience Frontiers* **6**(6): 817–823. https://doi.org/10.1016/j.gsf.2014.02.009

Lowekamp, B. C., Chen, D. T., Ibáñez, L. and Blezek, D. (2013). The design of SimpleITK, *Frontiers in Neuroinformatics* **7**. https://doi.org/10.3389/fninf.2013.00045

Luo, P., He, B., Takara, K., Xiong, Y. E., Nover, D., Duan, W. and Fukushi, K. (2015). Historical assessment of chinese and japanese flood management policies and implications for managing future floods, *Environmental Science & Policy* **48**: 265–277. https://doi.org/10.1016/j.envsci.2014.12.015

Mack, B. M. (2017). *Applied One-Class Classification of Remote Sensing Data*, PhD thesis, Freie Universität Berlin. URL: https://dx.doi.org/10.17169/REFUBIUM-10885

Mack, B., Roscher, R., Stenzel, S., Feilhauer, H., Schmidtlein, S. and Waske, B. (2016). Mapping raised bogs with an iterative one-class classification approach, *ISPRS Journal of Photogrammetry and Remote Sensing* **120**: 53–64. https://doi.org/10.1016/j.isprsjprs.2016.07.008

Mack, B., Roscher, R. and Waske, B. (2014). Can i trust my one-class classification?, *Remote Sensing* **6**(9): 8779–8802. https://doi.org/10.3390/rs6098779

Mack, B. and Waske, B. (2016). In-depth comparisons of MaxEnt, biased SVM and one-class SVM for one-class classification of remote sensing data, *Remote Sensing Letters* **8**(3): 290–299. https://doi.org/10.1080/2150704x.2016.1265689

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. and McCarroll, S. A. (2015). Highly parallel genome-wide

expression profiling of individual cells using nanoliter droplets, *Cell* **161**(5): 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Madhuri, R., Sistla, S. and Raju, K. S. (2021). Application of machine learning algorithms for flood susceptibility assessment and risk management, *Journal of Water and Climate Change* . https://doi.org/10.2166/wcc.2021.051

Magnini, A., Lombardi, M., Persiano, S., Tirri, A., Conti, F. L. and Castellarin, A. (2021). Machine-learning blends of geomorphic descriptors: value and limitations for flood hazard assessment across large floodplains, *Natural Hazards and Earth System Sciences* . https://doi.org/10.5194/nhess-2021-253

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P. and Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics, *Earth System Dynamics* **11**(1): 201–234. https://doi.org/10.5194/esd-11-201-2020

Maiwald, H. and Schwarz, J. (2015). Damage and loss prognosis tools correlating flood action and building's resistance-type parameters, *International Journal of Safety and Security Engineering* **5**(3): 222–250. https://doi.org/10.2495/safe-v5-n3-222-250

Malgwi, M. B., Fuchs, S. and Keiler, M. (2019). An adaptive regional vulnerability assessment model: Review and concepts for data-scarce regions, *Natural Hazards and Earth System Sciences Discussions* . https://doi.org/10.5194/nhess-2019-366

Martinis, S. (2018). A sentinel-1 times series-based exclusion layer for improved flood mapping in arid areas, *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE. https://doi.org/10.1109/igarss.2018.8518604

Martinis, S. and Rieke, C. (2015). Backscatter analysis using multi-temporal and multi-frequency SAR data in the context of flood mapping at river saale, germany, *Remote Sensing* **7**(6): 7732–7752. https://doi.org/10.3390/rs70607732

Martinis, S., Twele, A. and Voigt, S. (2009). Towards operational near real-time flood detection using a split-based automatic thresholding procedure on high resolution TerraSAR-x data, *Natural Hazards and Earth System Sciences* **9**(2): 303–314. https://doi.org/10.5194/nhess-9-303-2009

Martone, M., Rizzoli, P., Wecklich, C., González, C., Bueso-Bello, J.-L., Valdo, P., Schulze, D., Zink, M., Krieger, G. and Moreira, A. (2018). The global forest/non-forest map from TanDEM-x interferometric SAR data, *Remote Sensing of Environment* **205**: 352–373. https://doi.org/10.1016/j.rse.2017.12.002

Marzocchi, W., Garcia-Aristizabal, A., Gasparini, P., Mastellone, M. L. and Ruocco, A. D. (2012). Basic principles of multi-risk assessment: a case study in italy, *Natural Hazards* **62**(2): 551–573. https://doi.org/10.1007/s11069-012-0092-x

Mason, D., Schumann, G.-P., Neal, J., Garcia-Pintado, J. and Bates, P. (2012). Automatic near real-time selection of flood water levels from high resolution synthetic aperture radar images for assimilation into hydraulic models: A case study, *Remote Sensing of Environment* **124**: 705–716. https://doi.org/10.1016/j.rse.2012.06.017

Mason, D., Speck, R., Devereux, B., Schumann, G.-P., Neal, J. and Bates, P. (2010). Flood detection in urban areas using TerraSAR-x, *IEEE Transactions on Geoscience and Remote Sensing* **48**(2): 882–894. https://doi.org/10.1109/tgrs.2009.2029236

Matgen, P., Giustarini, L., Chini, M., Hostache, R., Wood, M. and Schlaffer, S. (2016). Creating a water depth map from SAR flood extent and topography data, *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE. https://doi.org/10.1109/igarss.2016.7730991

Matgen, P., Hostache, R., Schumann, G., Pfister, L., Hoffmann, L. and Savenije, H. (2011). Towards an automated SAR-based flood monitoring system: Lessons learned from two case studies, *Physics and Chemistry of the Earth, Parts A/B/C* **36**(7-8): 241–252. https://doi.org/10.1016/j.pce.2010.12.009

Mazzorana, B., Simoni, S., Scherer, C., Gems, B., Fuchs, S. and Keiler, M. (2014). A physical approach on flood risk vulnerability of buildings, *Hydrology and Earth System Sciences* **18**(9): 3817–3836. https://doi.org/10.5194/hess-18-3817-2014

Merow, C., Smith, M. J. and Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter, *Ecography* **36**(10): 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J. C. J. H., Bates, P., Bertola, M., Kemter, M., Kreibich, H., Lall, U. and Macdonald, E. (2021). Causes, impacts and patterns of disastrous river floods, *Nature Reviews Earth & Environment* **2**(9): 592–609. https://doi.org/10.1038/s43017-021-00195-3

Merz, B., Elmer, F. and Thieken, A. H. (2009). Significance of 'high probability/low damage' versus 'low probability/high damage' flood events, *Natural Hazards and Earth System Sciences* **9**(3): 1033–1046. https://doi.org/10.5194/nhess-9-1033-2009

Merz, B., Kreibich, H. and Lall, U. (2013). Multi-variate flood damage assessment: a tree-based data-mining approach, *Natural Hazards and Earth System Science* **13**(1): 53–64. https://doi.org/10.5194/nhess-13-53-2013

Merz, B., Kreibich, H., Schwarze, R. and Thieken, A. (2010). Review article "assessment of economic flood damage", *Natural Hazards and Earth System Science* **10**(8): 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010

Merz, B., Vorogushyn, S., Lall, U., Viglione, A. and Blöschl, G. (2015). Charting unknown waters-on the role of surprise in flood risk assessment and management, *Water Resources Research* **51**(8): 6399–6416. https://doi.org/10.1002/2015wr017464

Merz, R., Blöschl, G. and Humer, G. (2008). National flood discharge mapping in austria, *Natural Hazards* **46**(1): 53–72. https://doi.org/10.1007/s11069-007-9181-7

Metin, A. D., Nguyen, D. V., Schröter, K., Vorogushyn, S., Guse, B., Kreibich, H. and Merz, B. (2020). The role of spatial dependence for large-scale flood risk estimation, *Natural Hazards and Earth System Sciences* **20**(4): 967–979. https://doi.org/10.5194/nhess-20-967-2020

Meyer, V., Scheuer, S. and Haase, D. (2008). A multicriteria approach for flood risk mapping exemplified at the mulde river, germany, *Natural Hazards* **48**(1): 17–39. https://doi.org/10.1007/s11069-008-9244-4

Milligan, G. W. (1996). Clustering Validation: Results and Implications for Applied Analyses, *Clustering and Classification*, World Scientific, pp. 341–375. https://doi.org/10.1142/9789812832153_0010

Müller, A., Reiter, J. and Weiland, U. (2011). Assessment of urban vulnerability towards floods using an indicator-based approach – a case study for santiago de chile, *Natural Hazards and Earth System Sciences* **11**(8): 2107–2123. https://doi.org/10.5194/nhess-11-2107-2011

Molinari, D., Scorzini, A. R., Arrighi, C., Carisi, F., Castelli, F., Domeneghetti, A., Gallazzi, A., Galliani, M., Grelot, F., Kellermann, P., Kreibich, H., Mohor, G. S., Mosimann, M., Natho, S., Richert, C., Schroeter, K., Thieken, A. H., Zischg, A. P. and Ballio, F. (2020). Are flood damage models converging to "reality"? lessons learnt from a blind test, *Natural Hazards and Earth System Sciences* **20**(11): 2997–3017. https://doi.org/10.5194/nhess-20-2997-2020

Moon, S., Perron, J. T., Martel, S. J., Holbrook, W. S. and Clair, J. S. (2017). A model of three-dimensional topographic stresses with implications for bedrock fractures, surface processes, and landscape evolution, *Journal of Geophysical Research: Earth Surface* **122**(4): 823–846. https://doi.org/10.1002/2016jf004155

Moore, I. D., Grayson, R. B. and Ladson, A. R. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications, *Hydrological Processes* **5**(1): 3–30. https://doi.org/10.1002/hyp.3360050103

Mudashiru, R. B., Sabtu, N., Abustan, I. and Balogun, W. (2021). Flood hazard mapping methods: A review, *Journal of Hydrology* **603**: 126846. https://doi.org/10.1016/j.jhydrol.2021.126846

Mukherjee, F. and Singh, D. (2019). Detecting flood prone areas in harris county: a GIS based analysis, *GeoJournal* **85**(3): 647–663. https://doi.org/10.1007/s10708-019-09984-2

MunichRe (2019). Natcatservice database, Munich Reinsurance Company, Geo Risks Research, Munich. URL: http://www.munichre.com/natcatservice

Murphy, K. (2012). *Machine learning : a probabilistic perspective*, MIT Press, Cambridge, Mass.

Myhre, G., Samset, B. H., Hodnebrog, Ø., Andrews, T., Boucher, O., Faluvegi, G., Fläschner, D., Forster, P. M., Kasoar, M., Kharin, V., Kirkevåg, A., Lamarque, J.-F., Olivié, D., Richardson, T. B., Shawki, D., Shindell, D., Shine, K. P., Stjern, C. W., Takemura, T. and Voulgarakis, A. (2018). Sensible heat has significantly affected the global hydrological cycle over the historical period, *Nature Communications* **9**(1). https://doi.org/10.1038/s41467-018-04307-4

Nafari, R. H., Ngo, T. and Mendis, P. (2016). An assessment of the effectiveness of tree-based models for multi-variate flood damage assessment in australia, *Water* **8**(7): 282. https://doi.org/10.3390/w8070282

Neal, J., Schumann, G. and Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas, *Water Resources Research* **48**(11). https://doi.org/10.1029/2012wr012514

Neal, J., Villanueva, I., Wright, N., Willis, T., Fewtrell, T. and Bates, P. (2011). How much physical complexity is needed to model flood inundation?, *Hydrological Processes* **26**(15): 2264–2282. https://doi.org/10.1002/hyp.8339

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, *in* T. G. Dietterich, S. Becker and Z. Ghahramani (eds), *Advances in Neural Information Processing Systems 14*, MIT Press,

pp. 841–848. URL: http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf

Nguyen, V. D., Merz, B., Hundecha, Y., Haberlandt, U. and Vorogushyn, S. (2021). Comprehensive evaluation of an improved large-scale multi-site weather generator for germany, *International Journal of Climatology* **41**(10): 4933–4956. https://doi.org/10.1002/joc.7107

Nguyen, V. D., Metin, A. D., Alfieri, L., Vorogushyn, S. and Merz, B. (2020). Biases in national and continental flood risk assessments by ignoring spatial dependence, *Scientific Reports* **10**(1). https://doi.org/10.1038/s41598-020-76523-2

NHC (2018). Costliest u.s. tropical cyclones tables updated, *Technical report*, National Hurricane Center. URL: https://www.nhc.noaa.gov/news/UpdatedCostliest.pdf

Nicholls, N. (2011). Comments on "have disaster losses increased due to anthropogenic climate change?", *Bulletin of the American Meteorological Society* **92**(6): 791–791. https://doi.org/10.1175/2011bams3167.1

NOAA (2020). Calculation of oceanic niño index (oni) between the years 1950 to 2020, *Technical report*, National Oceanic And Atmospheric Administration of the United States of America. URL: cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v4.shtml

Oddo, P. C. and Bolten, J. D. (2019). The value of near real-time earth observations for improved flood disaster response, *Frontiers in Environmental Science* **7**. https://doi.org/10.3389/fenvs.2019.00127

Olsen, A., Zhou, Q., Linde, J. and Arnbjerg-Nielsen, K. (2015). Comparing methods of calculating expected annual damage in urban pluvial flood risk assessments, *Water* **7**(12): 255–270. https://doi.org/10.3390/w7010255

OpenDataWorkingGroup (2007). Open government data principles, *Retrieved August 20, 2021*, Sebastopol. URL: https://opengovdata.org/

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org , https://www.openstreetmap.org.

Ortiz, S., Breidenbach, J. and Kändler, G. (2013). Early detection of bark beetle green attack using TerraSAR-x and RapidEye data, *Remote Sensing* **5**(4): 1912–1931. https://doi.org/10.3390/rs5041912

Osberghaus, D. and Hinrichs, H. (2020). The effectiveness of a large-scale flood risk awareness campaign: Evidence from two panel data sets, *Risk Analysis* **41**(6): 944–957. https://doi.org/10.1111/risa.13601

Österle, H., Gerstengarbe, F. W. and Werner, P. C. (2016). Ein meteorologischer datensatz für deutschland, 1951–2003 (in german), *in* F. W. V. H. S. K. M. V. B. H. P. Gräfe (ed.), *Die Elbe im Globalen Wandel*, SchweizerbartScience Publishers, Stuttgart, Germany:, pp. 81–84.

Otsu, N. (1979). A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1): 62–66. https://doi.org/10.1109/tsmc.1979.4310076

Pabian, F. (2015). Commercial satellite imagery as an evolving open-source verification technology: Emerging trends and their impact for nuclear nonproliferation analysis, *Technical report*, Publications Office of the European Union, Luxembourg (Luxembourg). URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC97258

Panagos, P., Borrelli, P. and Meusburger, K. (2015). A new european slope length and steepness factor (LS-factor) for modeling soil erosion by water, *Geosciences* **5**(2): 117–126. https://doi.org/10.3390/geosciences5020117

Paprotny, D., Kreibich, H., Morales-Nápoles, O., Castellarin, A., Carisi, F. and Schröter, K. (2020a). Exposure and vulnerability estimation for modelling flood losses to commercial assets in Europe, *Science of The Total Environment* **737**: 140011. https://doi.org/10.1016/j.scitotenv.2020.140011

Paprotny, D., Kreibich, H., Morales-Nápoles, O., Terefenko, P. and Schröter, K. (2020). Estimating exposure of residential assets to natural hazards in Europe using open data, *Natural Hazards and Earth System Sciences* **20**(1): 323–343. https://doi.org/10.5194/nhess-20-323-2020

Paprotny, D., Morales-Nápoles, O. and Jonkman, S. N. (2018). HANZE: a pan-european database of exposure to natural hazards and damaging historical floods since 1870, *Earth System Science Data* **10**(1): 565–581. https://doi.org/10.5194/essd-10-565-2018

Pedregosa, F., Bach, F. and Gramfort, A. (2017). On the consistency of ordinal regression methods, *Journal of Machine Learning Research* **18**(55): 1–35.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**: 2825–2830.

Pekel, J.-F., Cottam, A., Gorelick, N. and Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes, *Nature* **540**(7633): 418–422. https://doi.org/10.1038/nature20584

Pelizari, P. A., Geiß, C., Aguirre, P., María, H. S., Peña, Y. M. and Taubenböck, H. (2021). Automated building characterization for seismic risk assessment using street-level imagery and deep learning, *ISPRS Journal of Photogrammetry and Remote Sensing* **180**: 370–386. https://doi.org/10.1016/j.isprsjprs.2021.07.004

Penning-Rowsell, E. C. and Chatterton, J. B. (1977). *The Benefits of Flood Alleviation: A Manual of Assessment Techniques*, Gower Technical Press.

Penning-Rowsell, E. C., Priest, S. J. and King, D. (2016). Flood risk management and 'fairness': aspirations and reality, *E3S Web Conf.* **7**: 24001. https://doi.org/10.1051/e3sconf/20160724001

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., da Veiga Leprevost, F., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K. and Vizcaíno, J. A. (2016). Ten simple rules for taking advantage of git and GitHub, *PLoS Computational Biology* **12**(7): e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E. and Blair, M. E. (2017). Opening the black box: an open-source release of maxent, *Ecography* **40**(7): 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions, *Ecological Modelling* **190**(3-4): 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J. and Dudík, M. (2008). Modeling of species distributions with maxent: new extensions and a comprehensive evaluation, *Ecography* **31**(2): 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Piiroinen, R., Fassnacht, F. E., Heiskanen, J., Maeda, E., Mack, B. and Pellikka, P. (2018). Invasive tree species detection in the eastern arc mountains biodiversity hotspot using one class classification, *Remote Sensing of Environment* **218**: 119–131. https://doi.org/10.1016/j.rse.2018.09.018

Pinter, N., van der Ploeg, R. R., Schweigert, P. and Hoefer, G. (2006). Flood magnification on the river rhine, *Hydrological Processes* **20**(1): 147–164. https://doi.org/10.1002/hyp.5908

Pittore, M., Zapata, J. C. G., Brinckmann, N., Weatherill, G., Babeyko, A., Harig, S., Mahdavi, A., Proß, B., Velasquez, H. F. R., Straub, D., Krautblatter, M., Frimberger, T., Langbein, M., Geiß, C. and Schoepfer, E. (2020). Towards an integrated framework for distributed, modular multi-risk scenario assessment, *EGU2020-19097* . https://doi.org/10.5194/egusphere-egu2020-19097

Plank, S., Jüssi, M., Martinis, S. and Twele, A. (2017). Mapping of flooded vegetation by means of polarimetric sentinel-1 and ALOS-2/PALSAR-2 imagery, *International Journal of Remote Sensing* **38**(13): 3831–3850. https://doi.org/10.1080/01431161.2017.1306143

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *ADVANCES IN LARGE MARGIN CLASSIFIERS*, MIT Press, pp. 61–74.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E. and Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, *SOIL* **7**(1): 217–240. https://doi.org/10.5194/soil-7-217-2021

Pour, A. B. and Hashim, M. (2017). Application of landsat-8 and ALOS-2 data for structural and landslide hazard mapping in kelantan, malaysia, *Natural Hazards and Earth System Sciences* **17**(7): 1285–1303. https://doi.org/10.5194/nhess-17-1285-2017

Prancevic, J. P. and Kirchner, J. W. (2019). Topographic controls on the extension and retraction of flowing streams, *Geophysical Research Letters* **46**(4): 2084–2092. https://doi.org/10.1029/2018gl081799

Pulvirenti, L., Chini, M., Pierdicca, N. and Boni, G. (2016). Use of SAR data for detecting floodwater in urban and agricultural areas: The role of the interferometric coherence, *IEEE Transactions on Geoscience and Remote Sensing* **54**(3): 1532–1544. https://doi.org/10.1109/tgrs.2015.2482001

Pulvirenti, L., Pierdicca, N., Chini, M. and Guerriero, L. (2011). An algorithm for operational flood mapping from synthetic aperture radar (SAR) data using fuzzy logic,

Natural Hazards and Earth System Sciences **11**(2): 529–540. https://doi.org/10.5194/nhess-11-529-2011

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J. and Heffernan, J. (2019). The spatial dependence of flood hazard and risk in the united states, *Water Resources Research* **55**(3): 1890–1911. https://doi.org/10.1029/2018wr024205

Quinn, P., Beven, K., Chevallier, P. and Planchon, O. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models, *Hydrological Processes* **5**(1): 59–79. https://doi.org/10.1002/hyp.3360050106

Rawlins, C. L. (1995). A view of the river by luna b. leopold, *Western American Literature* **29**(4): 386–386. https://doi.org/10.1353/wal.1995.0126

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat (2019). Deep learning and process understanding for data-driven earth system science, *Nature* **566**(7743): 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J. and Waterloo, M. J. (2008). HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in amazonia, *Remote Sensing of Environment* **112**(9): 3469–3481. https://doi.org/10.1016/j.rse.2008.03.018

Reu, J. D., Bourgeois, J., Bats, M., Zwertvaegher, A., Gelorini, V., Smedt, P. D., Chu, W., Antrop, M., Maeyer, P. D., Finke, P., Meirvenne, M. V., Verniers, J. and Crombé, P. (2013). Application of the topographic position index to heterogeneous landscapes, *Geomorphology* **186**: 39–49. https://doi.org/10.1016/j.geomorph.2012.12.015

Riley, S. J., DeGloria, S. D. and Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity, *Intermountain Journal of Sciences* **5**(1-4): 23–27.

Risi, R. D., Jalayer, F., Paola, F. D., Iervolino, I., Giugni, M., Topa, M. E., Mbuya, E., Kyessi, A., Manfredi, G. and Gasparini, P. (2013). Flood risk assessment for informal settlements, *Natural Hazards* **69**(1): 1003–1032. https://doi.org/10.1007/s11069-013-0749-0

Rizzoli, P., Martone, M., Gonzalez, C., Wecklich, C., Tridon, D. B., Bräutigam, B., Bachmann, M., Schulze, D., Fritz, T., Huber, M., Wessel, B., Krieger, G., Zink, M. and Moreira, A. (2017). Generation and performance assessment of the global TanDEM-x digital elevation model, *ISPRS Journal of Photogrammetry and Remote Sensing* **132**: 119–139. https://doi.org/10.1016/j.isprsjprs.2017.08.008

Roche, D. G., Granados, M., Austin, C. C., Wilson, S., Mitchell, G. M., Smith, P. A., Cooke, S. J. and Bennett, J. R. (2020). Open government data and environmental science: a federal canadian perspective, *FACETS* **5**(1): 942–962. https://doi.org/10.1139/facets-2020-0008

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain., *Psychological Review* **65**(6): 386–408. https://doi.org/10.1037/h0042519

Rosvold, E. L. and Buhaug, H. (2021). GDIS, a global dataset of geocoded disaster locations, *Scientific Data* **8**(1). https://doi.org/10.1038/s41597-021-00846-6

Rottler, E., Bronstert, A., Bürger, G. and Rakovec, O. (2021). Projected changes in rhine river flood seasonality under global warming, *Hydrol. Earth Syst. Sci.* **25**(5): 2353–2371. https://doi.org/10.5194/hess-25-2353-2021

Rözer, V., Kreibich, H., Schröter, K., Müller, M., Sairam, N., Doss-Gollin, J., Lall, U. and Merz, B. (2019). Probabilistic models significantly reduce uncertainty in hurricane harvey pluvial flood loss estimates, *Earth's Future* **7**(4): 384–394. https://doi.org/10.1029/2018ef001074

Rözer, V., Peche, A., Berkhahn, S., Feng, Y., Fuchs, L., Graf, T., Haberlandt, U., Kreibich, H., Sämann, R., Sester, M., Shehu, B., Wahl, J. and Neuweiler, I. (2021). Impact-based forecasting for pluvial floods, *Earth's Future* **9**(2). https://doi.org/10.1029/2020ef001851

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E. and Kloft, M. (2018). Deep one-class classification, *in* J. Dy and A. Krause (eds), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 4393–4402. URL: http://proceedings.mlr.press/v80/ruff18a.html

Sairam, N., Brill, F., Sieg, T., Farrag, M., Kellermann, P., Nguyen, V. D., Lüdtke, S., Merz, B., Schröter, K., Vorogushyn, S. and Kreibich, H. (2021). Process-based flood risk assessment for germany, *Earth's Future* **9**(10). https://doi.org/10.1029/2021ef002259

Sairam, N., Schröter, K., Lüdtke, S., Merz, B. and Kreibich, H. (2019). Quantifying flood vulnerability reduction via private precaution, *Earth's Future* **7**(3): 235–249. https://doi.org/10.1029/2018ef000994

Samela, C., Troy, T. J. and Manfreda, S. (2017). Geomorphic classifiers for flood-prone areas delineation for data-scarce environments, *Advances in Water Resources* **102**: 13–28. https://doi.org/10.1016/j.advwatres.2017.01.007

Sauer, I. J., Reese, R., Otto, C., Geiger, T., Willner, S. N., Guillod, B. P., Bresch, D. N. and Frieler, K. (2021). Climate signals in river flood damages emerge under sound regional disaggregation, *Nature Communications* **12**(1). https://doi.org/10.1038/s41467-021-22153-9

Schäfer, A., Mühr, B., Daniell, J., Ehret, U., Ehmele, F., Küpfer, K., Brand, J., Wisotzky, C., Skapski, J., Rentz, L., Mohr, S. and Kunz, M. (2021). Hochwasser mitteleuropa, juli 2021 (deutschland) : 21. juli 2021 – bericht nr. 1 „nordrhein-westfalen & rheinland-pfalz", *Technical report*, Center for Disaster Management and Risk Reduction Technology (CEDIM). URL: https://www.doi.org/10.5445/IR/1000135730

Schlaffer, S., Matgen, P., Hollaus, M. and Wagner, W. (2015). Flood detection from multi-temporal SAR data using harmonic analysis and change detection, *International Journal of Applied Earth Observation and Geoinformation* **38**: 15–24. https://doi.org/10.1016/j.jag.2014.12.001

Schmidt, L., Heße, F., Attinger, S. and Kumar, R. (2020). Challenges in applying machine learning models for hydrological inference: A case study for flooding events across germany, *Water Resources Research* **56**(5). https://doi.org/10.1029/2019wr025924

Schoppa, L., Sieg, T., Vogel, K., Zöller, G. and Kreibich, H. (2020). Probabilistic flood loss models for companies, *Water Resources Research* **56**(9). https://doi.org/10.1029/2020wr027649

Schrock, A. and Shaffer, G. (2017). Data ideologies of an interested public: A study of grassroots open government data intermediaries, *Big Data & Society* **4**(1): 205395171769075. https://doi.org/10.1177/2053951717690750

Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F. and Merz, B. (2014). How useful are complex flood damage models?, *Water Resources Research* **50**(4): 3378–3395. https://doi.org/10.1002/2013wr014396

Schumann, G., Pappenberger, F. and Matgen, P. (2008). Estimating uncertainty associated with water stages from a single SAR image, *Advances in Water Resources* **31**(8): 1038–1047. https://doi.org/10.1016/j.advwatres.2008.04.008

Scotti, V., Giannini, M. and Cioffi, F. (2020). Enhanced flood mapping using synthetic aperture radar (SAR) images, hydraulic modelling, and social media: A case study of hurricane harvey (houston, TX), *Journal of Flood Risk Management* . https://doi.org/10.1111/jfr3.12647

Seifert, I., Thieken, A. H., Merz, M., Borst, D. and Werner, U. (2010). Estimation of industrial and commercial asset values for hazard risk assessment, *Natural Hazards* **52**(2): 453–479. https://doi.org/10.1007/s11069-009-9389-9

Sharifi, E., Saghafian, B. and Steinacker, R. (2019). Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques, *Journal of Geophysical Research: Atmospheres* **124**(2): 789–805. https://doi.org/10.1029/2018jd028795

Sharma, A., Wasko, C. and Lettenmaier, D. P. (2018). If precipitation extremes are increasing, why aren't floods?, *Water Resources Research* **54**(11): 8545–8551. https://doi.org/10.1029/2018wr023749

Sieg, T., Vogel, K., Merz, B. and Kreibich, H. (2017). Tree-based flood damage modeling of companies: Damage processes and model performance, *Water Resources Research* **53**(7): 6050–6068. https://doi.org/10.1002/2017wr020784

Sieg, T., Vogel, K., Merz, B. and Kreibich, H. (2019). Seamless estimation of hydrometeorological risk across spatial scales, *Earth's Future* . https://doi.org/10.1029/2018ef001122

Smith, R. H. (1978). *Development of a dynamic flood routing model for small meandering rivers*, PhD thesis, University of Missouri–Rolla. URL: 172.https://scholarsmine.mst.edu/doctoral_dissertations/172

Son, R., Wang, S.-Y. S., Tseng, W.-L., Schuler, C. W. B., Becker, E. and Yoon, J.-H. (2019). Climate diagnostics of the extreme floods in peru during early 2017, *Climate Dynamics* **54**(1-2): 935–945. https://doi.org/10.1007/s00382-019-05038-y

Speckhann, G. A., Kreibich, H. and Merz, B. (2021). Inventory of dams in germany, *Earth System Science Data* **13**(2): 731–740. https://doi.org/10.5194/essd-13-731-2021

Spekkers, M., Rözer, V., Thieken, A., ten Veldhuis, M.-C. and Kreibich, H. (2017). A comparative survey of the impacts of extreme rainfall in two international case studies, *Natural Hazards and Earth System Sciences* **17**(8): 1337–1355. https://doi.org/10.5194/nhess-17-1337-2017

Steger, S., Mair, V., Kofler, C., Pittore, M., Zebisch, M. and Schneiderbauer, S. (2021). Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling – benefits of exploring landslide data collection effects, *Science of The Total Environment* **776**: 145935. https://doi.org/10.1016/j.scitotenv.2021.145935

Stephens, E., Bates, P., Freer, J. and Mason, D. (2012). The impact of uncertainty in satellite data on the assessment of flood inundation models, *Journal of Hydrology* **414-415**: 162–173. https://doi.org/10.1016/j.jhydrol.2011.10.040

Stern, M. J., Bilgen, I. and Dillman, D. A. (2014). The state of survey methodology, *Field Methods* **26**(3): 284–301. https://doi.org/10.1177/1525822x13519561

Stokes, G. G. (1880). On the theories of the internal friction of fluids in motion, and of the equilibrium and motion of elastic solids, *Mathematical and Physical Papers vol.1*, Cambridge University Press, pp. 75–129. https://doi.org/10.1017/cbo9780511702242.005

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

Suppasri, A., Mas, E., Charvet, I., Gunasekera, R., Imai, K., Fukutani, Y., Abe, Y. and Imamura, F. (2012). Building damage characteristics based on surveyed data and fragility curves of the 2011 great east japan tsunami, *Natural Hazards* **66**(2): 319–341. https://doi.org/10.1007/s11069-012-0487-8

Tabari, H. (2020). Climate change impact on flood and extreme precipitation increases with water availability, *Scientific Reports* **10**(1). https://doi.org/10.1038/s41598-020-70816-2

Tatem, A. J. (2017). WorldPop, open data for spatial demography, *Scientific Data* **4**(1). https://doi.org/10.1038/sdata.2017.4

Taubenböck, H., Wurm, M., Netzband, M., Zwenzner, H., Roth, A., Rahman, A. and Dech, S. (2011). Flood risks in urbanized areas – multi-sensoral approaches using remotely sensed data for risk assessment, *Natural Hazards and Earth System Science* **11**(2): 431–444. https://doi.org/10.5194/nhess-11-431-2011

te Linde, A. H., Bubeck, P., Dekkers, J. E. C., de Moel, H. and Aerts, J. C. J. H. (2011). Future flood risk estimates along the river rhine, *Natural Hazards and Earth System Sciences* **11**(2): 459–473. https://doi.org/10.5194/nhess-11-459-2011

Tehrany, M. S., Kumar, L., Jebur, M. N. and Shabani, F. (2018). Evaluating the application of the statistical index method in flood susceptibility mapping and its comparison with frequency ratio and logistic regression methods, *Geomatics, Natural Hazards and Risk* **10**(1): 79–101. https://doi.org/10.1080/19475705.2018.1506509

Terti, G., Ruin, I., Gourley, J. J., Kirstetter, P., Flamig, Z., Blanchet, J., Arthur, A. and Anquetin, S. (2017). Toward probabilistic prediction of flash flood human impacts, *Risk Analysis* **39**(1): 140–161. https://doi.org/10.1111/risa.12921

Thieken, A. H., Mohor, G. S., Kreibich, H. and Müller, M. (2021). Compound flood events: different pathways–different impacts–different coping options?, *Natural Hazards and Earth System Sciences Discussions* . https://doi.org/10.5194/nhess-2021-27

Thieken, A. H., Olschewski, A., Kreibich, H., Kobsch, S. and Merz, B. (2008). Development and evaluation of FLEMOps – a new flood loss estimation MOdel for the private sector, *Flood Recovery, Innovation and Response I*, WIT Press. https://doi.org/10.2495/friar080301

Thouret, J.-C., Enjolras, G., Martelli, K., Santoni, O., Luque, J. A., Nagata, M., Arguedas, A. and Macedo, L. (2013). Combining criteria for delineating lahar- and flash-flood-prone hazard and risk zones for the city of arequipa, peru, *Natural Hazards and Earth System Sciences* **13**(2): 339–360. https://doi.org/10.5194/nhess-13-339-2013

Thouret, J.-C., Ettinger, S., Guitton, M., Santoni, O., Magill, C., Martelli, K., Zuccaro, G., Revilla, V., Charca, J. A. and Arguedas, A. (2014). Assessing physical vulnerability in large cities exposed to flash floods and debris flows: the case of arequipa (peru), *Natural Hazards* **73**(3): 1771–1815. https://doi.org/10.1007/s11069-014-1172-x

Tilloy, A., Malamud, B. D., Winter, H. and Joly-Laugel, A. (2019). A review of quantification methodologies for multi-hazard interrelationships, *Earth-Science Reviews* **196**: 102881. https://doi.org/10.1016/j.earscirev.2019.102881

Trigg, M. A., Birch, C. E., Neal, J. C., Bates, P. D., Smith, A., Sampson, C. C., Yamazaki, D., Hirabayashi, Y., Pappenberger, F., Dutra, E., Ward, P. J., Winsemius, H. C., Salamon, P., Dottori, F., Rudari, R., Kappes, M. S., Simpson, A. L., Hadzilacos, G. and Fewtrell, T. J. (2016). The credibility challenge for global fluvial flood risk analysis, *Environmental Research Letters* **11**(9): 094014. https://doi.org/10.1088/1748-9326/11/9/094014

Tukey, J. W. (1986). Sunset salvo, *The American Statistician* **40**(1): 72. https://doi.org/10.2307/2683137

Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J., Herold, M., Koh, L., Leimgruber, P., Taubenboeck, H., Wegmann, M., Wikelski, M. and Woodcock, C. (2015). Free and open-access satellite data are key to biodiversity conservation, *Biological Conservation* **182**: 173–176. https://doi.org/10.1016/j.biocon.2014.11.048

Twele, A., Cao, W., Plank, S. and Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain, *International Journal of Remote Sensing* **37**(13): 2990–3004. https://doi.org/10.1080/01431161.2016.1192304

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE, *Journal of Machine Learning Research* **9**: 2579–2605.

van Mierlo, M., Vrouwenvelder, A., Calle, E., Vrijling, J., Jonkman, S., de Bruijn, K. and Weerts, A. (2007). Assessment of flood risk accounting for river system behaviour, *International Journal of River Basin Management* **5**(2): 93–104. https://doi.org/10.1080/15715124.2007.9635309

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*, Springer New York.

Varga, O. G., Kovács, Z., Bekő, L., Burai, P., Szabó, Z. C., Holb, I., Ninsawat, S. and Szabó, S. (2021). Validation of visually interpreted corine land cover classes with spectral values of satellite images and machine learning, *Remote Sensing* **13**(5): 857. https://doi.org/10.3390/rs13050857

VenW (2007). Hydraulische randvoorwaarden primaire waterkeringen voor de derde toetsronde 2006-2011 (hr 2006) (in dutch), *Technical report*, Ministerie van Verkeer en Waterstaat, The Hague. URL: https://puc.overheid.nl/rijkswaterstaat/doc/PUC_128421_31/

Villacorta, S., Fidel, L. and Zavala, B. (2012). Mapa de susceptibilidad por movimientos en masa del perú, *Revista de la Asociacion Geologica Argentina* **69**: 393–399.

Visser, H., Petersen, A. C. and Ligtvoet, W. (2014). On the relation between weather-related disaster impacts, vulnerability and climate change, *Climatic Change* **125**(3-4): 461–477. https://doi.org/10.1007/s10584-014-1179-z

Vorogushyn, S., Bates, P. D., de Bruijn, K., Castellarin, A., Kreibich, H., Priest, S., Schröter, K., Bagli, S., Blöschl, G., Domeneghetti, A., Gouldby, B., Klijn, F., Lammersen, R., Neal, J. C., Ridder, N., Terink, W., Viavattene, C., Viglione, A., Zanardo, S. and Merz, B. (2018). Evolutionary leap in large-scale flood risk assessment needed, *WIREs Water* **5**(2). https://doi.org/10.1002/wat2.1266

Vorogushyn, S., Lindenschmidt, K.-E., Kreibich, H., Apel, H. and Merz, B. (2012). Analysis of a detention basin impact on dike failure probabilities and flood risk for a channel-dike-floodplain system along the river elbe, germany, *Journal of Hydrology* **436-437**: 120–131. https://doi.org/10.1016/j.jhydrol.2012.03.006

Vorogushyn, S., Merz, B., Lindenschmidt, K.-E. and Apel, H. (2010). A new methodology for flood hazard assessment considering dike breaches, *Water Resources Research* **46**(8). https://doi.org/10.1029/2009wr008475

Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., Sarica, G. M., Ruangpan, L., Molinario, G. and Lallemant, D. (2020). Invited perspectives: How machine learning will change flood risk and impact assessment, *Natural Hazards and Earth System Sciences* **20**(4): 1149–1161. https://doi.org/10.5194/nhess-20-1149-2020

Wagenaar, D., de Jong, J. and Bouwer, L. M. (2017). Multi-variable flood damage modelling with limited data using supervised learning approaches, *Natural Hazards and Earth System Sciences* **17**(9): 1683–1696. https://doi.org/10.5194/nhess-17-1683-2017

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., de Rosnay, P., Jann, A., Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang, Y., Blöschl, G., Eitzinger, J. and Steinnocher, K. (2013). The ASCAT soil moisture product: A review of its specifications, validation results, and emerging applications, *Meteorologische Zeitschrift* **22**(1): 5–33. https://doi.org/10.1127/0941-2948/2013/0399

Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., van Beek, R., Bierkens, M. F. P., Ligtvoet, W. and Winsemius, H. C. (2013). Assessing flood risk at the global scale: model setup, results, and sensitivity, *Environmental Research Letters* **8**(4): 044019. https://doi.org/10.1088/1748-9326/8/4/044019

Warszawski, L., Kriegler, E., Lenton, T. M., Gaffney, O., Jacob, D., Klingenfeld, D., Koide, R., Costa, M. M., Messner, D., Nakicenovic, N., Schellnhuber, H. J., Schlosser, P., Takeuchi, K., Leeuw, S. V. D., Whiteman, G. and Rockström, J. (2021). All options, not silver bullets, needed to limit global warming to 1.5 °c: a scenario appraisal, *Environmental Research Letters* **16**(6): 064037. https://doi.org/10.1088/1748-9326/abfeec

Watson, K. M., Harwell, G. R., Wallace, D. S., Welborn, T. L., Stengel, V. G. and McDowell, J. S. (2018). Characterization of peak streamflows and flood inundation of selected areas in southeastern texas and southwestern louisiana from the august and september 2017 flood resulting from hurricane harvey, *Technical report*, US Geological Survey. URL: https://www.doi.org/10.3133/sir20185070

Weiss, A. (2001). Topographic position and landforms analysis, Poster Presentation, ESRI Users Conference, San Diego, CA. http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf.

Weiss, G. M. and Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study, *Technical report*, Rutgers University. URL: https://www.doi.org/10.7282/T3-VPFW-SF95

Wieland, M. and Martinis, S. (2019). A modular processing chain for automated flood monitoring from multi-spectral satellite data, *Remote Sensing* **11**(19): 2330. https://doi.org/10.3390/rs11192330

Wieland, M. and Pittore, M. (2016). Large-area settlement pattern recognition from landsat-8 data, *ISPRS Journal of Photogrammetry and Remote Sensing* **119**: 294–308. https://doi.org/10.1016/j.isprsjprs.2016.06.010

Wing, O. E., Sampson, C. C., Bates, P. D., Quinn, N., Smith, A. M. and Neal, J. C. (2019). A flood inundation forecast of hurricane harvey using a continental-scale 2d hydrodynamic model, *Journal of Hydrology X* **4**: 100039. https://doi.org/10.1016/j.hydroa.2019.100039

Winsemius, H. C., Beek, L. P. H. V., Jongman, B., Ward, P. J. and Bouwman, A. (2013). A framework for global river flood risk assessments, *Hydrology and Earth System Sciences* **17**(5): 1871–1892. https://doi.org/10.5194/hess-17-1871-2013

Winter, B., Schneeberger, K., Dung, N., Huttenlau, M., Achleitner, S., Stötter, J., Merz, B. and Vorogushyn, S. (2019). A continuous modelling approach for design flood estimation on sub-daily time scale, *Hydrological Sciences Journal* **64**(5): 539–554. https://doi.org/10.1080/02626667.2019.1593419

Wünsch, A., Herrmann, U., Kreibich, H. and Thieken, A. H. (2009). The role of disaggregation of asset values in flood loss estimation: A comparison of different modeling approaches at the mulde river, germany, *Environmental Management* **44**(3): 524–541. https://doi.org/10.1007/s00267-009-9335-3

Wood, M., Hostache, R., Neal, J., Wagener, T., Giustarini, L., Chini, M., Corato, G., Matgen, P. and Bates, P. (2016). Calibration of channel depth and friction parameters in the LISFLOOD-FP hydraulic model using medium-resolution SAR data and identifiability techniques, *Hydrology and Earth System Sciences* **20**(12): 4983–4997. https://doi.org/10.5194/hess-20-4983-2016

Wu, G., Guo, Y., Song, X., Guo, Z., Zhang, H., Shi, X., Shibasaki, R. and Shao, X. (2019). A stacked fully convolutional networks with feature alignment framework for multi-label land-cover segmentation, *Remote Sensing* **11**(9): 1051. https://doi.org/10.3390/rs11091051

Wyncoll, D. and Gouldby, B. (2015). Integrating a multivariate extreme value method within a system flood risk analysis model, *Journal of Flood Risk Management* **8**(2): 145–160. https://doi.org/10.1111/jfr3.12069

Yamazaki, D., Baugh, C. A., Bates, P. D., Kanae, S., Alsdorf, D. E. and Oki, T. (2012). Adjustment of a spaceborne DEM for use in floodplain hydrodynamic modeling, *Journal of Hydrology* **436-437**: 81–91. https://doi.org/10.1016/j.jhydrol.2012.02.045

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. and Pavelsky, T. (2019). MERIT hydro: A high-resolution global hydrography map based on latest topography

datasets, *Water Resources Research* **55**(6): 5053–5073. https://doi.org/10.1029/2019wr024873

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S. and Bates, P. D. (2017). A high-accuracy map of global terrain elevations, *Geophysical Research Letters* **44**(11): 5844–5853. https://doi.org/10.1002/2017gl072874

Yamazaki, D., Kanae, S., Kim, H. and Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model, *Water Resources Research* **47**(4). https://doi.org/10.1029/2010wr009726

Yousefi, S., Pourghasemi, H. R., Emami, S. N., Pouyan, S., Eskandari, S. and Tiefenbacher, J. P. (2020). A machine learning framework for multi-hazards modeling and mapping in a mountainous area, *Scientific Reports* **10**(1). https://doi.org/10.1038/s41598-020-69233-2

Yu, H. and Robinson, D. G. (2012). The new ambiguity of 'open government', *UCLA Law Review Discourse* **59**: 178–208. https://doi.org/10.2139/ssrn.2012489

Yu, L., Wang, Z., Tian, S., Ye, F., Ding, J. and Kong, J. (2017). Convolutional neural networks for water body extraction from landsat imagery, *International Journal of Computational Intelligence and Applications* **16**(01): 1750001. https://doi.org/10.1142/s1469026817500018

Zeder, J. and Fischer, E. M. (2020). Observed extreme precipitation trends and scaling in central europe, *Weather and Climate Extremes* **29**: 100266. https://doi.org/10.1016/j.wace.2020.100266

Zevenbergen, C., Fu, D. and Pathirana, A. (2018). Transitioning to sponge cities: Challenges and opportunities to address urban water problems in china, *Water* **10**(9): 1230. https://doi.org/10.3390/w10091230

Zevenbergen, L. W. and Thorne, C. R. (1987). Quantitative analysis of land surface topography, *Earth Surface Processes and Landforms* **12**(1): 47–56. https://doi.org/10.1002/esp.3290120107

Zhao, Q. and Hastie, T. (2019). Causal interpretations of black-box models, *Journal of Business & Economic Statistics* **39**(1): 272–281. https://doi.org/10.1080/07350015.2019.1624293

Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., Maraun, D., Ramos, A. M., Ridder, N. N., Thiery, W. and Vignotto, E. (2020). A typology of compound weather and climate events, *Nat Rev Earth Environ* **1**(7): 333–347. https://doi.org/10.1038/s43017-020-0060-z

Zwenzner, H. and Voigt, S. (2008). Improved estimation of flood parameters by combining space based SAR data with very high resolution digital elevation data, *Hydrology and Earth System Sciences Discussions* **5**(5): 2951–2973. https://doi.org/10.5194/hessd-5-2951-2008