Scalable coastal inundation mapping using machine learning

Ophélie Meuriot
IBM Research Europe
ophelie.meuriot@ibm.com

Anne Jones
IBM Research Europe
anne.jones@ibm.com

Abstract

Coastal flooding is a significant climate hazard with impacts across economic sectors and society. This study provides a proof of concept for data-driven models for coastal flood inundation at the country scale, incorporating storm dynamics and geospatial characteristics to improve upon simpler geomorphological models. The best fit machine learning model scores an AUC of 0.92 in predicting flooded locations. For a case study storm event in December 2013 we find that all models over-predict flood extents, but that the machine learning model extents were closest to those observed.

1 Introduction

Climate change-driven global sea-level rise increases the risk of coastal flooding through higher return frequencies of coastal water levels during storms. Although the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report found limited evidence for an increase in storminess due to climate change, measurable regional rises in mean sea level have already led to increases in the frequency of inundation (IPCC 2014). The most recent climate change risk assessment for the UK projects winter weather will be dominated by more mobile cyclonic systems which cause coastal flooding, exacerbated by 5-10cm higher rises in sea levels than previously projected (Betts, Pearson 2021). Coastal flooding is a substantial climate change concern for both societal and economic impacts: a significant fraction of global infrastructure (power plants, ports, housing, road and rail) is located at the coast, and, a combination of sea-level rise, changes in storm surge intensity and land subsidence combined with economic growth pose immense threats to coastal and low-lying communities (Oddo et al. 2020). There is consequently demand for better quantification of coastal flood risk for current and future climates across multiple industries and the public sector.

Coastal flood risk quantification poses numerous challenges, including substantial modelling uncertainties and the computational intensity of physical simulation models for storm surge and inundation, which limit both the resolution and extent over which they can be applied. In this paper we will address the latter component, and consider the suitability of computationally "lightweight" approaches to derive inundation from water levels and geospatial features for coastal flood risk mapping at the country scale, a particular requirement for industry sectors such as transport, energy and utilities who operate across large geographic domains.

The increase in data availability in the field of environmental sciences has led to the emergence of data-driven models to represent complex physical processes. Studies have recently shown the use of machine learning (ML) techniques to evaluate the flood susceptibility of an area. Studies so far have mainly focused on fluvial and pluvial flooding rather than coastal flooding and have been limited to smaller regions (Tehrany et al. 2014; Chen et al. 2020; Avand et al. 2021).

The aim of this paper is to derive a machine learning model for coastal flood mapping using historical flood records in England. The model will be tested on the December 2013 Storm Xavier event and

Tackling Climate Change with Machine Learning workshop at NeurIPS 2021

compared to a static inundation model to understand the relative strengths and weaknesses of each approach. Our work, which is ongoing, will contribute to improved coastal flood risk quantification and climate change resilience across multiple impacted sectors.

2 Data and methods

The aim of this study is to use a data-driven approach to estimate extents of coastal flooding. The methodology is summarised in Figure 1.

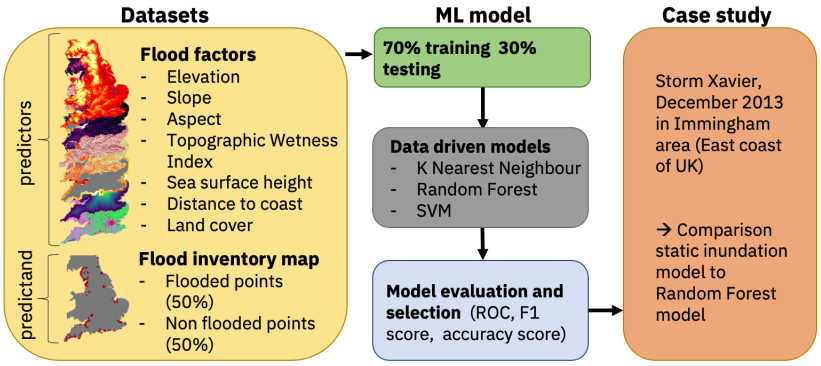

Figure 1: Methodology

2.1 Datasets

The variables used in this study to predict coastal flooding include: elevation, slope, aspect, topographic wetness index (TWI), land cover, sea surface height (SSH) and distance to the coast. The elevation, slope, aspect, TWI and land cover variables were selected based on a literature review of flood susceptibility (Avand et al. 2021; Park, Lee 2020). Past studies mainly focus on fluvial and pluvial flooding and include variables such as distance to river and rainfall (Avand et al. 2021). As the focus is here set on coastal flooding the choice of using distance to coast and a SSH is most suited. A historical record of flooding (Environment Agency 2018) was used to generate flooded and non flooded points (5099 points each) as labelling data for the ML model. A description of the variables and preprocessing steps are included in Appendix A.

2.2 Methodology

The flooded and non-flooded points were combined into a dataset sorted by date with a 1 assigned to flooded points and 0 to non flooded points. For each point (day, longitude, latitude) the corresponding elevation, slope, aspect, TWI, nearest SSH, distance to coast and land cover were assigned. The land cover was encoded using One-Hot encoding.The dataset was split temporally into training and testing with the first 70% of the dataset used for training and the last 30% of the dataset used for testing. The reason for performing a time-based splitting was to ensure the flooded points corresponding to a single event are not separated. Standard scaling was applied to the non categorical flood factors (elevation, slope, aspect, TWI, distance to coast and SSH). The ML models were trained using the flood factors as predictors (elevation, slope, aspect, TWI, land cover, distance to coast and SSH) and the binary 1 / 0 (flooded / non-flooded) as predictand.

Three ML models were assessed: the Support Vector Machine (SVM), k nearest neighbour (kNN) and random forest (RF) classification models. The Scikit-learn library was used to train and test the models. The models were tuned using the following hyperparameters: 5, 10, 15 and 20 neighbours for the kNN model and maximum depths of 5, 10, 15 and 20 for the RF model with 100 trees. Two kernel functions were implemented for the SVM model: the linear and radial basis function. After initial performance assessment on the training dataset, the ML models selected for further evaluation were the kNN model with 5 neighbours, RF model with 100 trees and a maximum depth equal to 15 and the SVM model with radial basis function. The best performing model was then selected by using metrics such as the AUC, the accuracy score and the F1 score on the testing set.

2.3 Case study event

The results of the best performing ML model were compared to a static inundation model during the Storm Xavier flooding event. On the 5th of December 2013, Storm Xavier caused severe flooding along the East Coast of England. The area specifically chosen for this case study is the Immingham area. The sea surface height dataset used for the case study is the UKCP18 Short Event Case Studies of Historical and Future Sea Surface Elevation around the UK which covers the event (3rd to 10th of December 2013) (Met Office Hadley Centre 2018b).

The static inundation model, also known as the "bathtub" model is a simple geomorphological model widely used in coastal inundation studies. The method used here is adapted from Vousdoukas et al. (2016). A flood water surface is interpolated using the sea surface height points. Locations where the elevation is lower than the flood water surface and which are connected to the sea are marked as flooded.

3 Results

3.1 Model evaluation

Here we present the results of the kNN, SVM and RF ML models. The ROC curves are shown in Figure 2. The highest AUC was obtained for the RF model with a value of 0.92 followed by 0.89 for the SVM model and 0.83 for the kNN model. The AUC results are in agreement with the F1 and accuracy scores with the RF model presenting the highest accuracy and F1 scores. The accuracy of the RF model is 0.84 compared to 0.8 and 0.79 for the SVM and kNN models respectively. The RF model has a F1 score of 0.74 compared to 0.71 and 0.67 for the SVM and kNN models respectively.

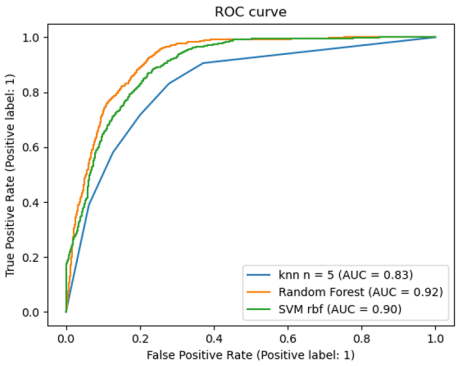

Figure 2: ROC curve

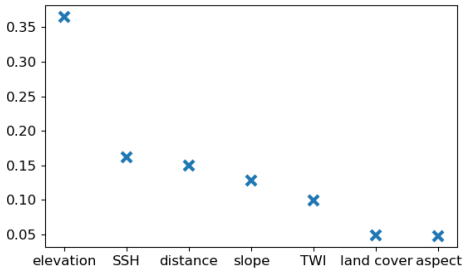

Figure 3: RF model feature importance

The model selected for the case study was the RF model as it performs the best according to the AUC, F1 and accuracy scores. The feature importance of the different flood factors are shown in Figure 3 and provide an indication of the relative influence of each factor. The feature with the highest importance is elevation (0.37), followed by SSH (0.16), distance to the coast (0.15), slope (0.13) and TWI (0.10). The features with the lowest importance are land cover (0.05) and aspect (0.05). The results are in agreement with Avand et al. (2021) where elevation presents the highest relative importance followed by distance to the river, slope, TWI and aspect.

3.2 Case study

The RF model and the static inundation model were run using the maximum SSH recorded during the Storm Xavier event. The results are shown in Figure 4 and compared to the recorded flood extents from the Environment Agency shown as hashed areas (Environment Agency 2018). The flood extent is shown in yellow for the RF model and in red for the static inundation model. Areas where both the flood extents from the RF model and static inundation models overlap are represented in orange.

Compared to the recorded flood extents from the Environment Agency, both the RF and static inundation models correctly mark the flooded extents as flooded. They however both over predict the

flood extents. The RF model prediction is closer to the Environment Agency recorded flood extents with a smaller flooded area compared to the static inundation model. The similar shapes in flood extents for both the static inundation model and RF model can be explained by the importance of the elevation in both models. The results suggest that the additional flood factors in the RF model improve the flood extent prediction.

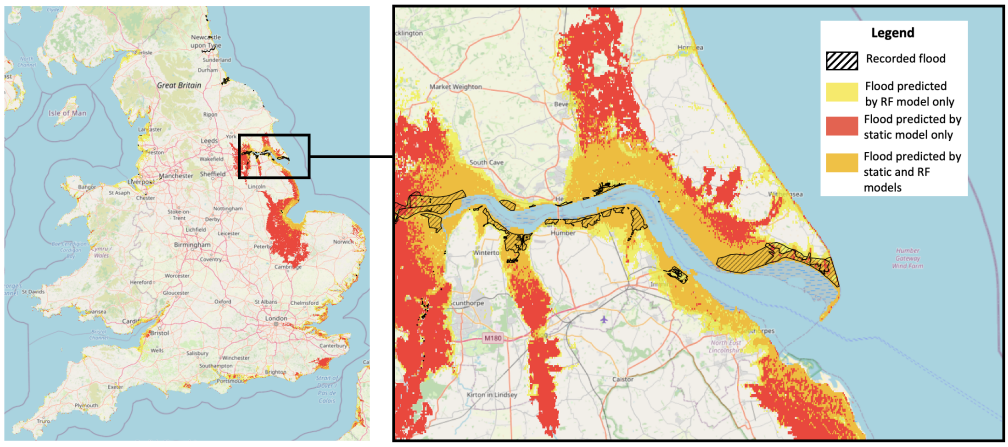

Figure 4: Flood extents from the static inundation model and the RF model compared to the Environment Agency Recorded Flood Outlines (Environment Agency 2018)

4 Implications for coastal flood mapping and future steps

Although the RF model over predicts the flood extent during the Storm Xavier event, this case study demonstrates the potential of data-driven modelling approaches for coastal flood mapping. The focus was here set on the Immingham area for which historical flood records were available, however the model can be applied to a country wide scale. The model was run using a 250m grid based on the DEM resolution. Higher resolution satellite or LIDAR-derived DEMs are available in the UK, and will be evaluated in future work.

The model was here trained and tested using Environment Agency Recorded Flood Outline dataset. The dataset presents the advantage of covering a large time period (1946 to 2020) and over 1300 coastal and tidal floods have been recorded. However not all flood events during that period have been recorded and the extents may not be precise. Future steps in the model development will involve using satellite derived flood extents such as Sen1Floods11 (Bonafilia et al. (2020)) or validation of existing flood extents using satellite data.

The model was developed using six static predictors (elevation, distance to coast, slope, TWI, land cover and aspect) and one dynamic predictor (sea surface height). The aim was to study coastal flooding independently from fluvial and pluvial flooding. However, storm surge events are often coincident with high precipitation resulting in rivers overflowing. While modelling such processed with physics-based models is incredibly complex, The data-driven approach taken here can easily be adapted to study compound flooding and increase prediction accuracy.

5 Conclusion

ML models (kNN, SVM and RF) were trained and tested to detect coastal flooding using historical recorded flood outlines between 1970 and 2006 in England. Seven predictors were used including elevation, distance, SSH, TWI, aspect, slope and land cover. The RF model performed best (AUC of 0.92). The predictions of the RF model and a static inundation model were compared for the Storm Xavier event in December 2013 against recorded flood extents. Although both models over predicted the flood extents, the RF flood extents were closer to the observed flood extent. This study provides a proof of concept for data-driven models for scalable coastal flood mapping and future works will aim to improve and expand the capabilities of the current method.

Acknowledgments

This work was supported by the Hartree National Centre for Digital Innovation, a collaboration between STFC and IBM.

References

Avand Mohammadtaghi, Moradi Hamidreza, Lasboyee Mehdi Ramazanzadeh. Using machine learning models, remote sensing, and GIS to investigate the effects of changing climates and land uses on flood probability // Journal of Hydrology. apr 2021. 595. 125663.

Betts Haward A.B. R.A., Pearson K.V.(eds.). The Third UK Climate Change Risk Assessment Technical Report. 2021.

Bonafilia Derrick, Tellman Beth, Anderson Tyler, Issenberg Erica. Sen1Floods11: A Georeferenced Dataset to Train and Test Deep Learning Flood Algorithms for Sentinel-1. 2020. 210–211.

Chen Wei, Li Yang, Xue Weifeng, Shahabi Himan, Li Shaojun, Hong Haoyuan, Wang Xiaojing, Bian Huiyuan, Zhang Shuai, Pradhan Biswajeet, Ahmad Baharin Bin. Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods // Science of The Total Environment. jan 2020. 701. 134979.

Environment Agency . Recorded Flood Outlines // https://data.gov.uk/dataset/16e32c53-35a6-4d54-a111-ca09031eaaaf/recorded-flood-outlines. 2018.

IPCC . Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC, 2014.

Met Office Hadley Centre . UKCP18 Historical Simulations of Gridded Sea Surface Elevation around the UK from 1970-2006. // Centre for Environmental Data Analysis https://catalogue.ceda.ac.uk/uuid/610a8d5b32fc4c51b8a2ed64de95ed73. 2018a.

Met Office Hadley Centre . UKCP18 UKCP18 Short Event Case Studies of Historical and Future Sea Surface Elevation around the UK // Centre for Environmental Data Analysis https://catalogue.ceda.ac.uk/uuid/58c393f773504caaad48cdb6310e17b2. 2018b.

Mohamedou Cheikh, Korhonen Lauri, Eerikaïnen Kalle, Tokola Timo. Using LiDAR-modified topographic wetness index, terrain attributes with leaf area index to improve a single-tree growth model in south-eastern Finland // Forestry. jul 2019. 92, 3. 253–263.

Oddo Perry C., Lee Ben S., Garner Gregory G., Srikrishnan Vivek, Reed Patrick M., Forest Chris E., Keller Klaus. Deep Uncertainties in Sea-Level Rise and Storm Surge Projections: Implications for Coastal Flood Risk Management // Risk Analysis. 2020. 40, 1. 153–168.

Park Sang-Jin, Lee Dong-Kun. Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms // Environmental Research Letters. aug 2020. 15, 9. 094052.

Tehrany Mahyat Shafapour, Pradhan Biswajeet, Jebur Mustafa Neamah. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS // Journal of Hydrology. may 2014. 512. 332–343.

Vousdoukas Michalis I., Voukouvalas Evangelos, Mentaschi Lorenzo, Dottori Francesco, Giardino Alessio, Bouziotas Dimitrios, Bianchi Alessandra, Salamon Peter, Feyen Luc. Developments in large-scale coastal flood hazard mapping // Natural Hazards and Earth System Sciences. 2016. 16, 8. 1841–1853.

A Appendix

Table 1: Dataset description and preprocessing steps

Dataset	Description
Elevation	The Global Multi-resolution Terrain Data Digital Elevation Model (GMTED DEM) is used for elevation and has a resolution of 250m. The dataset is obtained through IBM Physical Analytics Integrated Data Repository and Services (PAIRS). Low elevation coastal areas are most prone to coastal flooding.
Slope and aspect	The slope affects the velocity of flow with steeper areas increasing the velocity. The aspect indicates which direction the slopes face. The slope and aspect are derived from the DEM using WhiteboxTools with the same grid resolution as the GMTED DEM.
Topographic Wetness Index (TWI)	The topographic wetness is defined as $ln(a/tan(b))$ where a is the local upslope area and b is the slope angle . It provides an estimation of where water will accumulate and is commonly used as an indication of soil moisture (Mohamedou et al. (2019)). The TWI is calculated using WhiteboxTools with the same grid resolution as the GMTED DEM.
Land cover	The copernicus land cover raster is used here and is obtained through IBM Physical Analytics Integrated Data Repository and Services (PAIRS). The 25m resolution land cover raster is aligned to the 250m GMTED DEM grid using QGIS tools. There are 23 different land cover categories including, multiple vegetation types, open sea, permanent water bodies, wetland and urban areas.
Distance to coast	The coastline is extracted from the GMTED DEM as the limit between ocean values (indicated as 0) and land. A raster is created with the minimum distance to the coastline calculated for each grid point.
Sea surface height (SSH)	The sea surface height is obtained from the United Kingdom Climate Projections 2018 (UKCP18) historical simulations of gridded sea surface elevation using the Met Office HadGEM2-ES (Met Office Hadley Centre 2018a). The dataset provides hourly values along the British coast from 1970 to 2006. The daily maximum value is calculated and saved for each coastal point.
Flooded points	The flooded points are obtained from the Recorded Flood Outlines provided by the Environment Agency (Environment Agency 2018). The dataset includes flood records from 1946 to 2020. The floods are classified as coastal, tidal and fluvial and are provided as polygons. The coastal and tidal polygons from the time period between 1970 and 2006 are filtered to match the time range of the sea surface height series. The polygons are converted to points using QGIS with points spaced by a minimum distance corresponding to the GMTED DEM resolution. A total of 5099 points from historical flood events are generated.
Non flooded points	5099 non flooded points are generated to match the number of flooded points. The non flooded points are selected randomly and assigned a unique location (longitude, latitude) and time (day between 1970 and 2006) combination. A maximum distance to the coast is set to ensure the selected non flooded points are not further away than the furthest flooded point.