FH Salzburg

Data and Satellite Image Based Methods for Flood Detection and Prediction

Maximilian Erlmoser, Florian Lechner, Heiko Lehrer, Christoph Lenzbauer, Christoph Renzl & Natasha Troth

25th June 2021

Contents

1 Challenge Definition

Countless phenomena such as floods, fires, and algae blooms routinely impact ecosystems, economies, and human safety. Our challenge is to use satellite data to create a machine learning model that detects floods and build a web interface that not only displays the detected floods, but also layers it alongside ancillary data to help researchers and decision-makers better understand its impacts and scope.

2 What is our goal?

- Detect floods using weather data and satellite imagery data
- Create a machine learning model (e.g. decision trees or deep data) that reads in raw data and analyses it and returns if there are floods detected in certain areas
- Help detect floods soon enough to provide in time rescue missions
- Provide a detailed overview of detected floods around the world
- Create the following prototypes:
  - Shows areas on a map where floods are detected
  - Comparing real time satellite images for flood detection and overview
  - Mobile app for flood notifications

3 Flood Prediction/Detection Model

According to The European Space Agency[1], floods are the world's most expensive natural disaster type. As an example, floods in Europe in 2002 cost dozens of lives and billions of Euros. They are capable of destroying housing, agriculture and communications. Furthermore, floods and the damage caused by them are to increase in the future. The World Health Organization[2] further explains that floods are the most frequent type of natural disaster. They can occur when water submerges land that is usually dry, when there is heavy rainfall, rapid snowmelt or a storm surge (tropical cyclone or tsunami). The three common types of floods are:

- Flash floods (rapid and excessive rainfall that raises water levels quickly)
- River floods (due to consistent rain of snow melt)
- Coastal floods (caused by storm surges from cyclones and tsunamis)

In the past 10 years, 80-90% of documented disasters that were a result of natural hazards were due to floods, droughts, tropical cyclones, heat waves and severe storms. Due to climate change, floods expected to become more frequent and more intense.

Between 1998-2017 more than 2 billion people worldwide were affected by floods. People most at risk are people who live in floodplains or non-resistant buildings, or have a lack warning systems and awareness of flooding hazard. The risk of drowning in a flood is higher in low- and middle-income countries in flood prone areas. This is due to the fact that ability to warn, evacuate or protect the inhabitants from floods is weak or only just developing.

For these reasons, we have designed machine learning models capable of automatically detecting and predicting floods. These predictions can be used to notify alert the population all over the world to upcoming potential dangers due to floods. These models can be custom made for each use case, or a more general, but less effective model could be produced. Our ideal dataset would consist of any data of forces that lead to floods. For this project we have chosen the following data attributes:


```
{
  "Location": "Salzburg",
  "Date": "13.05.2021",
  "Amount of Rainfall (mm)": 50,
  "Temperature of Rainfall ( C )": 5,
  "Storm": true,
  "Type of Storm": "Thunder",
  "Duration of Storm (min)": 65,
  "Intensity of Storm": 0.6,
  "Electromagnetic Radiation (kHz)": 10,
  "Soil Texture": "Ts2",
  "Ground Moisture": 0.3,
  "Amount of Vegetation [m ]": 50,
  "Type of Vegetation": "pine",
  "Terrain Steepness": 0.6,
  "Temperature of Ground ( C )": 13,
  "Air Temperature ( C )": 15,
```


[1] http://www.esa.int/Applications/Observing_the_Earth/Securing_Our_Environment/Flood_monitoring

[2] https://www.who.int/health-topics/floods#tab=tab_1, https://www.who.int/news-room/fact-sheets/detail/drowning

```
  "Flood": true
}
```

These attributes were selected, since they are common metrics that can be utilised to predict floods (according to The National Severe Storms Laboratory[3] and USGS[4]). This is merely a subset of recorded metrics that can be used to detect floods. It can be presumed, that the higher the variety of metrics recorded, the more accurate a prediction might be.

For our machine learning algorithm to work, the data for the above-mentioned metrics would need to have been recorded over a period of time, long enough to be able to predict repeating patterns. Such a dataset should be able to be conceived, since weather stations have recorded data for the past decades. Such records are for example provided through NOAA[5].
In order for supervised machine learning to be implemented, it is necessary to have the target variable flood, which in this case could be a boolean, describing weather there was a flood or not. This target variable could for example also be a float, which could indicate the degree or severity of flooding. Further target variables could also be implemented, depending on the necessity for their prediction, on the use case, or on the availability of such data. Examples for other target variables include number of people or houses affected, area affected, cost in damage, etc.

By using already existing data, the model could be built in the near future. It might be advisable to implement new and more advanced measuring infrastructure, tailored for flood prediction. This could lead to more accurate detection and prediction results. However, it could take several years to gather enough data to make accurate prediction.

Datasets exist that have recorded flood events in the past. An example for a European Floods Database is provided by the European Environment Agency[6].
The flood data can be combined with the weather data leading up to the flood time, at the flood location (and any surrounding location that might influence it, for example when a river overflows upstream, it might also affect the area downstream). Thus, we have our feature variables and target variables. Using supervised machine learning algorithms, patterns in feature variables can be detected, which lead to similar target variables. The algorithm can recognise these and determine the target variable. This is how the flood detection and prediction will work.

In order to obtain the best possible prediction/detection of floods, it would be best practice to compare different models, created using different machine learning algorithms. The results could be measured and compared using quality metrics, for example such as RMSE (Root Mean Square Error) or Precision and Recall. For the purpose of this project, we will concentrate on using decision tree algorithms.

According to Han, Pei, and Kamber (2011)[7] decision trees are flowchart-like tree structures. Each branch has a node with a test (for example, is the value greater than 5). The result of the test decides if the algorithm continues on the left side of the tree or the right, which in turn

[3] https://www.nssl.noaa.gov/education/svrwx101/floods/forecasting/
[4] https://www.usgs.gov/faqs/how-are-floods-predicted?qt-news_science_products=3#qt-news_science_products
[5] https://www.ncdc.noaa.gov/cdo-web/datatools/records, https://www.ncdc.noaa.gov/cdo-web/
[6] https://www.eea.europa.eu/data-and-maps/data/external/european-floods-database
[7] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." The Morgan Kaufmann Series in Data Management Systems 5, no. 4 (2011): 83-124.

leads to more decisions until the leaf of the tree is reached. This is the target. Classification models predict categorical labels (class labels, e.g., true or false, strings). Regression models predict continuous values (numerical values). In our example, prediction whether a flood has occurred or not (true or false) would be classification. Predicting how severe a flood is (float number, e.g., 0.6) would require a regression model.

As explained in Larose and Larose (2015)[8], it is important to pre-process raw data to make it suitable for machine learning algorithms. Raw data often has missing values, is noisy (contains outliers) and can contain old or redundant data. To handle missing values, data can be replaced. This however can introduce bias, so it might be more advisable to remove incomplete data samples. Furthermore, our dataset has 16 different features, which means there are 16 dimensions. It is important to use as little data as possible, since high dimensional data can be sparse and more difficult to work with (curse of dimensionality). The features that are used should be relevant for the prediction. Feature selection can be used to prune out the irrelevant redundant features. Irrelevant features are features with no predictive value. Including them in the model does not improve the ability to predict the target. This means there is no or little correlation between the feature and the target variables. These features can instead reduce the accuracy of the model and should be removed. Redundant features are data attributes whose predictive value is already contained in another feature. These add no new predictive value to model and correlate strongly with other features. By removing irrelevant and redundant features, you should be left with less features, but the same predictive value.

To perform feature selection in our model, we would use the Random Forest algorithm to create a model, sort the features by their feature importances, and remove features with an importance under a given threshold. The Random Forest algorithm[9] creates multiple decision trees and returns the winning (classification) or average (regression) decision made by all the trees algorithm. It builds multiple decision trees on the given features.

Moreover, to remove redundant features, if there are two features with a high correlation (above a given threshold), one of these features (the one with a lower feature importance) will be removed. Dimensionality techniques such as Principal Component Analysis (PCA)[10] and T-SNE[11] (T-Distributed Stochastic Neighbor Embedding) can be used to further reduce the dimensionality (if necessary).

Good features in our dataset that we predict will most likely have high importance, are for example:

- Location: climates are different, some places have rivers/seas that can lead to flooding.
- Date: weather patterns are often similar at a certain time of year, in several years. E.g., there are often thunderstorms in the summer which can lead to dry grounds and sudden

[8]Larose, Daniel T, and Chantal D Larose. 2015. Data Mining and Predictive Analytics. 2nd Edition.Wiley Series on Methods and Applications in Data Mining. Hoboken, New Jersey: John Wiley Sons. ISBN: 9781118116197.

[9]Ho, Tin Kam. "Random decision forests." In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, pp. 278-282. IEEE, 1995

[10]Pearson, Karl. 1901. "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11): 559–572. doi:10.1080/14786440109462720. And Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. Baltimore: Warwick York.

[11]Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." Journal of Machine Learning research 9 (Nov): 2579–2605.

heavy rainfalls. Or in the spring there is often a lot of snow melting in some places which can lead to higher flowing rivers.

- Amount of rainfall: the more rain, the more likely there will be flooding.
- Storm: storms often bring sudden heavy rainfalls.
- Air Temperature: this can be important to distinguish between snow and rain.

Examples for features that would be bad in this dataset are:

- Ids: each row in the dataset may have an id. Each is unique and has no predictive value.
- Anything that has no connection to flooding and does not lead to it, e.g., number of sales in a fashion shop, etc.

The resulting features can then be used to create an XGBoost[12] model. The XGBoost machine learning algorithm uses decision trees and gradient boosting to return predictions and reduce errors.

The model should be built on training data and then tested on test data. A way to achieve this, is by utilising k-fold cross-validation. This is accomplished by splitting the data into k (often 5) parts. 4/5 of the data is training data, the rest is test data. The model is built by using the training data features and their target variables. It is then tested on the test data, by feeding in the features without the target variables. It must then predict these. The predicted target variables are then compared to the actual target variables. RMSE can be used to compare these and see how well the model has performed. This process is repeated k times, rotating the data until all of the data has been used as training and test data. The average RMSE from the k models can be computed and used to determine how accurate the model was.

The model should be optimised and improved (by changing features, using different machine learning algorithms) until a suitable RMSE is reached. It can be presumed, that flood detection will be quite predictable, as there are several known factors that result in it (e.g. heavy sudden rain, dry floors, melting snow, storms, etc.). A good RMSE value can therefore be expected.

Once the model is complete, weather stations or other measuring institutions can use it to predict floods, hopefully with good accuracy. They can measure their data, enter it into the system and receive a result of how likely a flood might be. With each passing day, new weather data is recorded and can be periodically added to the model to update it and make it more accurate.

It is important that the same features that are used to train the model, are then fed into it to create a prediction. If features are missing for the prediction, this will most likely lead to inaccurate or even false predictions. For this reason, it is important that the model for a specific location is built using the data available from that region. For example, it would make little sense to include snow fall in a prediction model for Hawaii. But it is a vital feature in somewhere like Austria, where snow can have a big impact on flooding.

[12]Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

The advantage of using a computed model, is that flood prediction/detection can be automated. Every time the features are measured and entered into the database, they could automatically be fed into our prediction model. If a value is predicted that indicates flooding will occur, it can alert the proper institutions immediately (authorities, weather stations...) so that they can immediately access the situation and take action. Satellite imagery can used to support this process (see below).

The downside to this model, is that only one set of measures can contribute to the prediction. If you want to predict a flood based on multiple days of weather, the average prediction/values would need to be used. A further disadvantage of this model is that a flood cannot be predicted very far in advance. Weather data is needed for the prediction, which cannot usually be acquired very far in advance. A weather forecast can also be rather inaccurate, the further away it is. A final downside to this model, is that it does not really provide much overview of the situation of a current flood. Our model should however be substantial in detecting floods in real-time.

In order to predict flooding further in advance, multi linear regression could for example be used to predict weather patterns, depending on previous years. It is however worth considering, that with global warming, weather patterns might be changing, which could make them hard to predict.

3.1 Clustering

Another approach to predict/detect flooding, is by using unsupervised machine learning. While supervised machine learning predicts given targets, unsupervised machine learning data has no targets. Clustering is a method of unsupervised machine learning used to group similar data together. This could be used to group together similar weather patterns and to determine which ones result in flooding.

3.2 Using Satellite Imagery

The problem however with only using data from weather stations, is that you can only detect and predict floods where the weather station is located. Not every town, square km of river or sea has a weather station. Furthermore, these measurements do not take into account how well certain parts of a river/sea have been protected against floods (e.g. made river deeper, defences against floods). According to The European Space Agency[13], satellite images can be utilised to detect where a river is most likely to burst its banks. Furthermore, it can be highly detailed digital elevation models of areas at risk. These models can be used for computerised flood simulations. Moreover, satellite images are used to provide an overview of the extent of a current flood. They can also show the development of the floods over time. Thus, further regions in danger can be detected before they are hit. The images can thus be used to create a rapid and authoritative damage assessment estimate and calculate the damage cost.

Satellite imagery can also be used to predict flooding. Deep learning models, such as convolutional neural networks can be used to analyse images and detect similar patterns in the pixels. These images are being utilised more and more to access damages, show the overview and extent of flooding. Examples for such implementations are outlined in Application of Satellite Data for

[13] `http://www.esa.int/Applications/Observing_the_Earth/Securing_Our_Environment/Flood_monitoring`

Flood Monitoring[14] and in The Use of Satellite Imagery in Crisis Management after Flooding[15].

The European Space Agency[16] outlined the usage of the Sentinel-1 satellite in Balkan floods in 2014. Through its radar data, it was able to map floods in Bosnia and Herzegovina. The radar has the ability to 'see' through clouds, rain and in darkness. It is therefore useful in monitoring floods. Images were compared before and after the flood to determine the extent of the flood and assess property and environmental damage. Through this imagery, supervisors of the Copernicus Emergency Management Service (EMS) were able to discover an area of flooding that had not yet been detected.

Sentinel-1 further assisted in the recent flooding in Australia in April 2021[17]. Along with other missions (RADARSAT-2, TerraSAR-X and COSMO-SkyMed), Sentinel-1 mapped the flooded areas and therefore helped with relief efforts. Since Austrailia was unable to deploy aircraft for aerial surveys, they relied on RADAR systems, like those used by Sentinel-1, to see the floods.

Satellite data monitors floods by comparing before and after images of a specific area. An example of such comparisons is provided here[18]. This comparison shows the flooding in eastern Australia caused by the Tropical Cyclone Debbie on 1st April 2017. The second images shows that the water had receded by the 11th April 2017. The water flooded roads and homes. These satellite images were taken by the Copernicus Sentinel-2 satellite mission.

These images are used to monitor a flood in real time. Perhaps these images could assist in predicting floods before they happen. The Copernicus Sentinels[19] provide data needed to predict floods, such as radar imaging of all global landmasses, and oceans. They also provide ocean forecasting systems, environmental monitoring and climate monitoring. This data, combined with weather data, could be used as in the section above, to predict patterns that lead to floods.

Our machine learning model would work in a similar way described by Anolytics[20]. We would use deep learning, namely convolutional neural networks, to scan satellite data to detect buildings, houses, rivers, road, etc. It could compare images over time and detect changes such as rising water levels and water spreading. It could also detect receding water that might indicate tsunamis. Furthermore, by being combined with annotated satellite objects (through annotation services) to detects which areas are buildings, roads, trees, water etc.. Through such data, it can be automatically detected roughly how many people are in danger, where people are in danger, which inhabited areas are in danger and what damage is to be expected. Authorities can then begin with preventative actions such as evacuation in a timely matter. It could also aid in monitoring a current flood. Moreover, satellite images of areas that can directly or indirectly lead to floods, or that have led to floods in the future, should also be monitored. For example,

[14] https://www.researchgate.net/profile/Heike-Bach/publication/267988097_Application_of_satellite_data_for_flood_monitoring/links/54f6ebd80cf27d8ed7202226/Application-of-satellite-data-for-flood-monitoring.pdf

[15] https://medium.com/sentinel-hub/the-use-of-satellite-imagery-in-crisis-management-after-flooding-382be517224f

[16] http://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Sentinel-1_aids_Balkan_flood_relief

[17] https://sentinel.esa.int/web/sentinel/-/copernicus-sentinel-1-facilitates-australia-s-flood-extent-delineation/1.1

[18] http://www.esa.int/ESA_Multimedia/Videos/2017/04/2017_Queensland_floods

[19] https://sentinel.esa.int/web/sentinel/missions/

[20] https://medium.com/nerd-for-tech/what-are-the-ai-applications-for-satellite-imagery-in-machine-learning-3495b2abe5

in Austria when the snow melts on the mountains, this can lead to higher rivers. If the snow on the mountains is monitored, combined with the weather forecast, this could be used to detect recurring patterns that have led to floods in the past. Another example might be, if a river is high upstream, it could impact an area further downstream with lower riverbanks. The model can be constantly checking and comparing images automatically, using convolutional neural networks. When flood indications appear, these can be compared with the weather predictions and other measurements from the decision tree model. If the calculated probability of a flood is over a specific threshold, the authorities should then automatically be alerted. The images can be provided via a web interface (see prototypes section 4) along with the following ancillary data:

- Automatically scan number buildings, houses, etc. to determine how many people will be in affected
- Where will/could be affected
- Which areas will be affected next
- Supply information of that regions flood defence mechanisms
- Weather forecast for the following days
- Also monitor areas that can lead to floods (e.g. monitor snow melting, upstream)

This model could once again be used in each region instead of being monitored by one general place. People who live in a certain region will know more about it and know what kind of relief actions should be taken when.

Links to imagery data are provided in the "Further Sources" paragraph.

3.3 Social Media and Mobile Notification App

In addition to using satellite images to aid relief in floods, social media could also be used. In today's modern world, people tend to share videos, photos and information online in real-time when disasters occur. While this data is less reliable than satellite images and might contain fake statements, a lot of the data posted could prove to be useful. It might aid in showing almost live footage of the flooding situation and also alert the authorities to the places that need the most help. Another point to consider, is that it could alert authorities of floods that satellites may not have picked up.

It might worth considering creating a flood notification app where people can upload information on current floods in a specific schema. This way the data would be more organised, uniform and less chaotic. Thus, making it easier to sift through quickly in an emergency. People might also be discouraged to upload fake reports on an official app (maybe lead by the government). Once a certain number of reports are received from a certain area in a certain time window, the authorities will automatically be alerted, as these are most likely potential dangerous floods. This app could also be useful to quickly notify people in a risk area and to provide them with correct, trustworthy and custom information on how they should proceed.

3.4 Summary

This paper describes how a machine learning model can be built to not only detect but also predict floods. Depending on which data is available, multiple models can be built for each use case. The model also depends on the location that is being monitored. For example, a model built for monitoring floods in the Austrian Alps would need to consider factors such as different seasons, temperature, and snow fall. A weather station in Australia would not need to consider the features describing snow fall. A weather station in more tropical climates would need to take monsoon seasons into account. Therefore, while it would be possible to build a general model for flood prediction, it would not be very accurate. Best practice would be to build an individual model for each location in the world. This however would require a lot of resources and does not seem to be a very efficient solution. Perhaps a type of middle ground can be found by producing a general model for overall prediction and then some smaller models customised to specific locations. Certain areas with similar weather patterns can share models (for example in Austria, Salzburg, Tirol and Vorarlberg have similar environmental conditions and therefore similar weather patterns). There will however always be variables that are different in each location that can influence the prediction. For example, flooding depends on how well the rivers have been constructed/improved to handle flooding.

Detecting floods using imagery can provide further precision and more detail on the flooding and its effects. While the data measured from weather stations can tell you if flooding is likely, real-time satellite imagery can give you a precise overview of the current situation. This way you can see which areas are impacted, which areas are in immediate danger and need to be evacuated and which areas are in most need of emergency services.

4 Prototypes

- **Comparing real time satellite images for flood detection overview:** `https://www.figma.com/proto/xXCMzRhXGFf88YZJJ7CD5K/Satellite-Images?node-id=1%3A108&scaling=contain&page-id=0%3A1`
- **Show areas on a map where floods are detected:** `https://www.figma.com/proto/YzRK9AvQbXxOeLGsRCfIC1/MONKE?node-id=1%3A3&scaling=contain&page-id=0%3A1`
- **Mobile app for flood notifications:** `https://www.figma.com/proto/E8h2KzmAtvqXElwrLbhv0t/Notification-app?scaling=scale-down&page-id=0%3A1&node-id=1%3A3`

5 Further Sources

- `https://gpm.nasa.gov/data`
- `https://earthdata.nasa.gov/earth-observation-data/near-real-time/hazards-and-disasters/floods`
- `https://floodmap.modaps.eosdis.nasa.gov/`
- `https://earthdata.nasa.gov/earth-observation-data/near-real-time/mcdwd-nrt`
- `https://earth.esa.int/eogateway/search?text=using+satellite+data+to+predict+floods+and+droughts&category=Data`