IST 707 Applied Machine Learning

Spring 2023 Final Project

ESG Implementation for Flooding Issues and Consequential Policy Decision-Making

Chien-Yu Lin, Hendi Kushta, Jiachen Li and Somia Abdelrahman

1. Introduction

Flooding is among the leading climatic threats to people's livelihoods, affecting development prospects worldwide – and floods can also reverse years of progress in poverty reduction and development. While the threat is already substantial, climate change and rapid urbanization in flood zones are likely to further drive up flood risks. The latest Intergovernmental Panel on Climate Change report affirms the urgency of addressing the intensifying impacts of climate change and ensuring the adaptation and resilience of the most vulnerable (Rentschler et al. 2022).

In October 2020, the World Bank presented a working paper that offered insight into global flood risk exposure and its intersection with poverty. Published in *Nature Communications*, the report estimates that 1.81 billion people face significant flood risk worldwide. It also estimates that 170 million extremely poor people are facing flood risk and its devastating long-term consequences. Together, these findings provide alarming insights into the scale of people's exposure and their vulnerabilities to flooding hazards (Rentschler et al. 2022).

Although flooding risk is still exacerbated on the Earth with parts of regions starting critical actions, we cannot impede the speed of urbanization, population growth, and migration. Considering the equilibrium between the global environment and governance decision-making, individual government agencies and representatives start to modify rules and policies in urban zoning for mitigating the exponential increment of climatic issues, especially flooding.

Collaborating with environmental scientists, arborists and landscape architects, there are agreements approving the positive impacts of green spaces. According to Dr. Kim in South Korea, analysis results indicate that having a higher proportion of green space inside a city could have a reduced effect on flooding risks in Ulsan. 1 km^2 of increase in green space could reduce financial insurance payment for flooding by $44,099 (Kim 2021). Pointed out by a partnership institute, Panorama, the soil and roots of trees improve groundwater infiltration, whilst their branches and leaves intercept rainfall and evapotranspirate water back into the atmosphere. This green solution attenuates the intensity of pluvial flooding by slowing and storing water during intense rain events (Koehorst 2020). Furthermore, the United States Environmental Protection Agency has worked to reduce runoff and improve water quality by implementing stormwater management with green spaces named Green Infrastructure (EPA 2012).

Objective

Such a complex subject, mixed with environmental and social-economics issues, requires the support of the theory of Environmental, Social and Governance (ESG). According to Chartered Financial Analyst (CPA) Institute, investors are increasingly applying these non-financial factors as part of their analysis process to identify material risks and growth opportunities. Companies are increasingly making disclosures in their annual report or in standalone sustainability reports. Numerous institutions, such as the Sustainability Accounting Standards Board (SASB), the Global Reporting Initiative (GRI), and the Task Force on Climate-related Financial Disclosures (TCFD) are working to form standards and define materiality to facilitate the incorporation of these factors into the investment process (CFA Institute 2023).

In order to apply the theory, appropriate data analysis and prediction will help decision makers to interdisciplinary collaborate for regional and/or global climatic issues. In this study project, we aim to reference the theory of ESG to explore relevant factors to the flooding issues associated with consequential policy decision-making for future urbanization development. Focused points are listed below to be studied with the methodologies of applied machine learning.

- **Environmental Aspect** – Classification and Detection of Urban Green Spaces Factors and Management
- **Governance Aspect** – Detection of Green Spaces in an Urban Region for Future Urban Revitalization

Data Description

In this study, we apply two datasets for each aspect of the data mining process. Both datasets are collected by the Green Expo Park Center located in Zhengzhou City, Henan Province, China. The study site of collected datasets is in the middle and lower reaches of the Yellow River basin, about 15 kilometers away from the south bank of the river, and the soil quality is sandy loam.

For the Environmental Aspect, LiDAR point cloud datasets are provided by the Green Expo Park Center. Stored point cloud data were subsequently transformed through a hierarchical process from spatial data combination, denoising and normalization, classification and pixilation to mathematical equations referenced from the physical-based Urban Forest Effects-Hydrology (UFORE-Hydro) model and data manipulation as csv. files (Lin et al. 2022).

Original variables are factors of UFORE-Hydro model. In order to transfer the dataset applicable to machine learning, we request on-site experts' and the project holder's advice to remove irrelevant variables such as coefficients and representative constants and maintain influential variables. The refined dataframe was generated with variables including rainfall hours, the amount of rainfall, leaf area index (LAI), canopies' storage capacity, canopies' penetration capability, canopies' evaporation capacity, overloaded storage water to the next stage, water reaches to the ground, soil infiltration capacity and surface runoff. The last variable, surface runoff

is the key predictor to infer the potential of flooding. The Environmental dataframe is then set as training data and test data both containing 10 variables with 54,360 observations for a proposed study of ordinal and numeric attributes classification for surface runoff indicating the spatial flooding risk.

For the Governance Aspect, we are supported by the high-resolution image collected by the Green Expo Park Center through an unmanned aerial vehicle (UAV). Samples are randomly selected based on on-site experts' and the project holder's advice. Each sample is a 30 by 30 pixels image which is categorized into five classes, bldg (building and structure), park (any green spaces), pav (paving like plazas, sidewalks, etc.), water (lakes, ponds and etc.) and trans (transportation spaces like roads, streets, etc.). The Governance dataframe is then generated by extracting pixel values from image samples. The dataframe is set as training data and test data both containing 500 samples in 5 classes which each observation obtains 2,700 pixel values as features. The proposed study is for numeric attributes classification of green space detection.

2. Data Preprocessing and Observation

Since we have two datasets for each ESG aspect, we elaborate on details in separate sections in this project's final report.

Environmental Aspect

For Environmental Aspect datasets, we first plot histograms to observe environmental factors. We then manipulate data preprocessing for association rules application to further implicate significant variables before machine learning model training. We then process a separate data preprocessing for machine learning model training. Each data preprocessing is described in the following sub-sections.

- *Data Observation*

  By plotting histograms and scatter plots graphics, we first observe the relationship between environmental factors. According to the Figure 1, we can understand that areas with high density of tree canopies have the potential for lower flooding issues. However, we can not deduce the advantage of tree canopies yet since the number of areas of denser tree canopies areas are 1,992 less than bare ground areas in terms of Table 1. Therefore, we further apply scatter plots to understand the water storage capacities of trees.

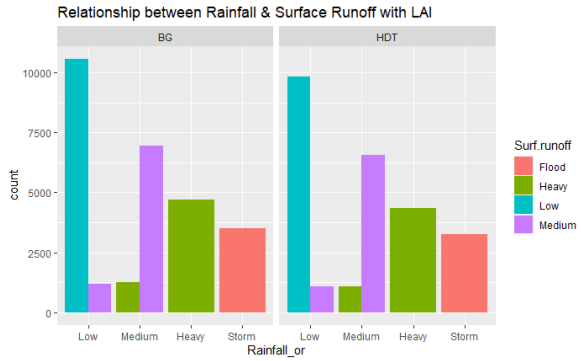

Figure 1 Relationship between Rainfall & Surface Runoff with LAI

LAI	BG (Bare Ground)	HDT (High-Density Trees)
Quantity	28,176	26,184

Table 1: Quantity of BR (Bare Ground) and HDT (High-Density Trees)

Based on Figure 2, we can further understand that the water storage capacity of trees at high-density canopies areas is more stable than bare ground areas. We also verified that the soil infiltration rate only varies during low surface runoff. This indicates that the soil has a limitation to hold the water. The site needs other natural infrastructures such as tree canopies to store water for reducing the potential of flooding.

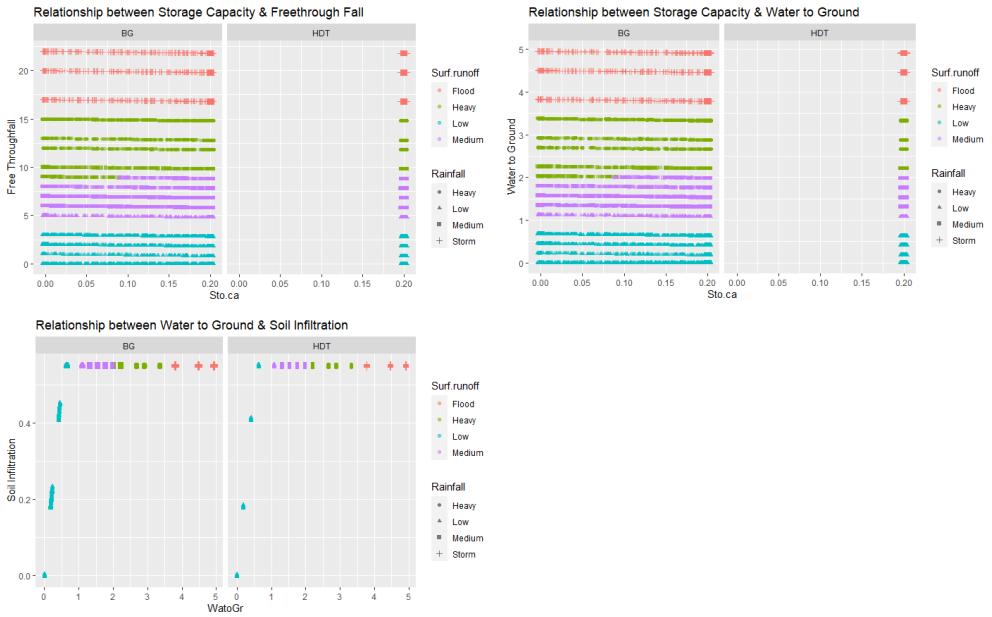

Figure 2 Top Left: Relationship between Storage Capacity & Freethrough Fall; Top Right: Relationship between Storage Capacity & Water to Ground; Bottom Left: Relationship between Water to Ground & Soil Infiltration

- *Association Rules Data Preprocessing*

  Association rules require categorical variables. Therefore, we request suggestions from experts and the project holder to further categorize numeric variables into critical levels. After applying the suggestions from the profession, we apply to mutate() to revise numeric variables to be ordinal variables for association rules mining.

- *Machine Modeling Data Preprocessing*

  We first remove Hour variable since time is not regarded as an environmental factor in this case. We further remove Surf.runoff (surface runoff and flooding potential) because it is regarded as the target label for machine learning detection which we will apply in modeling training. In order to make the dataset more applicable for machine learning, we then transform all variables to be numeric. For ordinal variables, Rainfall (the amount of rainfall) and LAI (leaf area index), we transform them to be factors and assign levels as numbers for them. As all variables except for Sur.runoff become numeric, we use scale() to generalize all numeric variables to be in the same scale in case of overfitting. The same steps are also applied to the test dataset. For the machine learning model, we will apply Naïve Bayes, SVM and Random Forest for data mining and model evaluation for finely approaching metrics. Different from SVM and Random Forest, we further reaffirm a dataset without the target labels, Surf.runoff, and also generate a vector that is the target labels, Surf.runoff for Naïve Bayes model training. Lastly, we separate the training dataset into 70% sub-training data and 30% sub-test data to follow common statistic model evaluation.

Governance Aspect

In order to compare green space detection, we propose to apply machine learning models and deep learning models based on accuracy and precision because of an expectation of precise image classification. For machine learning, we apply the data frame while we use images for deep learning. Both methods were processed with scaling to keep pixel values from 0 to 1. For machine learning the dataset was divided into 50% training dataset and 50% for testing while for deep learning it was divided into 50% for training(80%) and validation (20%) and the other 50% for testing. We then build parameters tuning elaborated in the next section.

3. Data Mining, Model Training and Evaluation

Following the section on Data Preprocessing and Observation, we apply the same format to elaborate on Data Mining, Model Training and Evaluation for the Environmental aspect and the Governance aspect.

Environmental Aspect

For Environmental Aspect datasets, we apply our knowledge based on the data observation and explore further relationships between factors by association rules mining. We then apply machine learning to detect the test dataset which doesn't contain labeled Surf.runoff. We propose to compare types of machine learning models and discuss their performance for better detection of flooding potential based on environmental factors.

- *Association Rules*

  We did a naïve test of association rules with support of 0.001 and confidence of 0.8 to demonstrate the overall factors' relationships. We mostly received trivial rules such as low water storage capacity at the bare ground area and the storm rainfall event causing flooding. Therefore, we propose to further focus on Surf.runoff in the following association rules mining. We further set the right-hand side as flooding potential to explore the possible situation. Except for rainfall storms, we observe other critical variables impacting flooding issues. Based on Table 2 to Table 4, we set the association rule mining with the confidence of 0.8 and support of 0.001, 0.01 and 0.1 to separately analyze the relationship. We can verify that the high penetration capability (Free Through Fall) of trees in certain areas will cause flooding regardless of the LAI factor.

lhs		rhs	support	confidence	lift
2 { Rainfall=Storm, Sto.nx.stg=Low}	=>	{ Surf.runoff=Flood}	0.001	1	8
4 { Rainfall=Storm, Sto.ca=Low}	=>	{ Surf.runoff=Flood}	0.004	1	8
5 { Rainfall=Storm, Sto.ca=Medium}	=>	{ Surf.runoff=Flood}	0.008	1	8
6 { Rainfall=Storm, Free.thrfall=High}	=>	{ Surf.runoff=Flood}	0.125	1	8

Table 2: Association critical rule mining result for Surf.runoff at support set as 0.001

lhs		rhs	support	confidence	lift
2 { Rainfall=Storm, Free.thrfall=High}	=>	{ Surf.runoff=Flood}	0.125	1	8
6 { Rainfall=Storm, Soil.inf=High}	=>	{ Surf.runoff=Flood}	0.125	1	8
7 { Rainfall=Storm, Sto.ca=High}	=>	{ Surf.runoff=Flood}	0.113	1	8
8 { Rainfall=Storm, Sto.nx.stg=High}	=>	{ Surf.runoff=Flood}	0.124	1	8
10 { Rainfall=Storm, LAI=HDT, Free.thrfall=High}	=>	{ Surf.runoff=Flood}	0.060	1	8

Table 3: Association critical rule mining result for Surf.runoff at support set as 0.01

lhs		rhs	support	confidence	lift
4 { Rainfall=Storm, Sto.ca=High}	=>	{ Surf.runoff=Flood}	0.125	1	8
5 { Rainfall=Storm, Sto.nx.stg=High}	=>	{ Surf.runoff=Flood}	0.113	1	8
7 { Rainfall=Storm, Free.thrfall=High, WatoGr=High}	=>	{ Surf.runoff=Flood}	0.124	1	8
8 { Rainfall=Storm, Sto.nx.stg=High}	=>	{ Surf.runoff=Flood}	0.125	1	8
9 { Rainfall=Storm, Free.thrfall=High, Soil.inf=High}	=>	{ Surf.runoff=Flood}	0.125	1	8
10 { Rainfall=Storm, Sto.ca=High, Free.thrfall=High}	=>	{ Surf.runoff=Flood}	0.113	1	8

Table 4: Association critical rule mining result for Surf.runoff at support set as 0.1

- *Machine Learning Evaluation*

  We trained models based on specific parameters tunning as Table 5 shows below. There are three types of machine learning models, Naïve Bayes, SVM and Random Forest trained for options to predict unclassified areas for flooding risks. Among the options, we have the Random Forest model presenting the best quality according to Table 6.

Parameters / Model	Fold Number	Repeats Number	tuneGrid	Best Option
Naïve Bayes	**5**	**10**	**usekernel (TURE, FALSE); laplace = c(0,1); adjust = c(0,1,2)**	**usekernel = TRUE; laplace = 0; adjust = 1**
SVM	**5**	**10**	**C = seq(0.1, 2, length = 20)**	**C = 0.1**
Random Forest	**5**	**10**	**.mtry = (2:9)**	**mtry = 2**

Table 5: Parameters Tunning for Naïve Bayes, SVM and Random Forest

Recall / Model	Flood	Heavy	Low	Medium
Naïve Bayes	1.00	0.99	1.00	1.00
NB Accuracy	99.77% with Sub-Test Data (99.81% Pure Model Evaluation)			
SVM	1.00	1.00	1.00	0.86
SVM Accuracy	95.96% with Sub-Test Data (96.02% Pure Model Evaluation)			
Random Forest	1.00	1.00	1.00	1.00
RF Accuracy	100.00% with Sub-Test Data (100.00% Pure Model Evaluation)			

Table 6: Model Evaluation and Comparison between Naïve Bayes, SVM and Random Forest

Governance Aspect

For Governance Aspect datasets, we applied the two types of data, labeled pixel values and images, for machine learning model training and deep learning model training. We propose to compare types of machine learning models and discuss their performance for better detection of green spaces based on pixel values. We also plan to discuss possible improvements to the model architecture. For machine learning models, we applied label pixel values and trained models based on parameters tunning as Table 7. For deep learning models, we use images for further exploration. Among the options, we have the Refined CNN model presenting the best quality according to Table 8 and Figure 3.

Parameters / Model	Fold Number	Repeats Number	tuneGrid	Best Option
Naïve Bayes	**5**	**10**	**usekernel (TURE, FALSE); laplace = c(0,1); adjust = c(0,1,2)**	**usekernel = TRUE; laplace = 0; adjust = 1**
SVM	**5**	**10**	**C = seq(0.1, 2, length = 20)**	**C = 0.1**
Random Forest	**5**	**10**	**.mtry = (2:9)**	**mtry = 8**

Table 7: Parameters Tunning for Naïve Bayes, SVM and Random Forest

Model	**Accuracy**
Naïve Bayes	80%
SVM	59.78%
Random Forest	86.16%
Standard CNN	80%
Refined CNN	100%

Table 8: Model Evaluation and Comparison between Naïve Bayes, SVM, Random Forest Standard CNN and Refined CNN

The graph below shows that after running 600 epochs, both training and validation accuracy reached 100% for the refined CNN model.

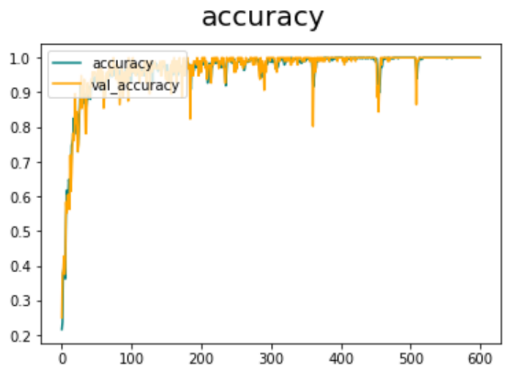

Figure 3 Refined CNN Accuracy Plot

4. Result

Following the previous sections, we also apply the same format to elaborate on the result and the performance for the Environmental aspect and the Governance aspect.

Environmental Aspect

Model performance and detection of environmental factors are demonstrated in Table 9. We can understand that all models' performances are of high quality based on metrics of accuracy and recall. We also apply recall to avoid the low possibility of flooding and any risk to people's livelihoods. Among types of machine learning models, we select Random Forest as the best detection due to its 100% accuracy. Additionally, the detection of flooding potential is in high accuracy in terms of given environmental factors. In the test datasets, there are 6,795 cases of flooding issues assumed to be at high risk during the day of the annual leaf-on period.

Model \ Surf.runoff	Flood	Heavy	Low	Medium
Naïve Bayes	**6,795**	**11,325**	**20,385**	**15,855**
Acc. 99.77%	**Rec. 1.00**	**Rec. 0.99**	**Rec. 1.00**	**Rec. 1.00**
SVM	**6,795**	**11,371**	**20,385**	**15,809**
Acc. 94.33%	**Rec. 1.00**	**Rec. 1.00**	**Rec. 1.00**	**Rec. 0.86**
Random Forest	**6,795**	**11,434**	**22,564**	**13,567**
Acc. 100%	**Rec. 1.00**	**Rec. 1.00**	**Rec. 1.00**	**Rec. 1.00**

Table 9: Machine Learning Models Performance for the Environmental Aspect

Governance Aspect

After running the best options models from the previous section on the testing dataset as demonstrated on Table 10 and Figure 4, it was found that Refined CNN has the best detection due to its highest 80% accuracy and higher precise detection to park and water bodies, 98% and 100% respectively.

Model	**Accuracy on testing data**	**Precision (Park class)**
Refined CNN	80%	98%
Random Forest	69.6%	92%
Naïve Bayes	61.2%	88%
SVM	57.4%	89%

Table 10: Models performance on the testing dataset for the Government Aspect

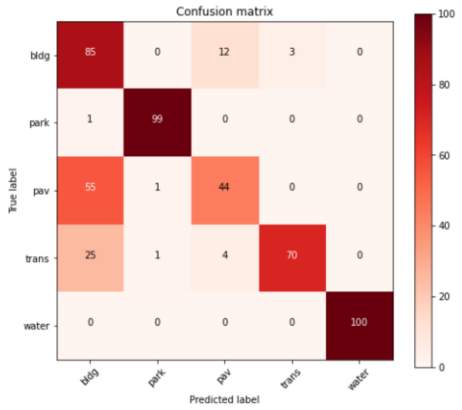

Figure 4 Refined CNN Confusion Matrix

5. Discussion

Based on the result of the model's evaluation and performance, we select Random Forest machine learning models for the Environmental aspect and the refined CNN model for the Governance aspect.

For the Environmental aspect, all models are of high quality. The Random Forest model performs with higher accuracy in all aspects. A possible assumption can be the consideration of

both independence and generalization from the major vote function which seems to be more comprehensive.

For the Governance aspect, we focus on the discussion about the deep learning models. We purify the architecture of layers to reduce the possibility of overfitting. The performance was significantly raised to be even higher than all machine learning models. Hence, the time consumption was comparatively efficient.

Our model training is precisely applicable to Random Forest for detecting flooding issues with associated environmental factors. The refined CNN model can be applied to detecting green spaces based on image samples and their pixel values. However, we may diversify classes or locations of image samples to increase accuracy. Nevertheless, the capability of distinguishing green spaces from other classes is acceptable in this project study.

6. Conclusion

By data mining and the exploration of machine learning and deep learning model training and evaluation, we can further understand the environmental factors of flooding issues for designing and planning green infrastructure. We can also detect green spaces from other classes to help the agency focus on areas of green infrastructure management. In addition, we reveal possible reasons for inapplicable models. We aim to further refine the data sources and fit the model with a more suitable architecture.

For the ESG Environmental aspect, the detection of flooding risk can help environmental experts and decision-makers with strategies and rules of green infrastructure management such as tree planting, species selection, and seasonal maintenance. On the other hand, for the ESG Governance aspect, we believe the selected model detection can distinguish green spaces for the focus on green infrastructure management in urban areas at flooding zones.

As with any model training, there is likely to be some uncertainty in the results. We cannot claim that these outcomes are perfect predictions of what could happen in a current or future situation. However, given our confidence in both the mathematical models as well as data mining and machine learning model detection, we feel that this research reduces the amount of uncertainty compared to pure simulations that focus solely on water flow, LAI, and the site topography without linking collected data. Our next steps will aim to improve several areas of the research related to accuracy.

Reference

CFA Institute. (2023). *ESG investing and analysis*. CFA Institute Research & Analysis. from https://www.cfainstitute.org/en/research/esg-investing#:~:text=ESG%20stands%20for%20Environmental%2C%20Social,material%20risks%20and%20growth%20opportunities

Cukier, K., & Mayer-Schönberger, V. (2014). The rise of Big Data: How it's changing the way we think about the world. *The Best Writing on Mathematics 2014*, 20–32. https://doi.org/10.1515/9781400865307-003

EPA. (2012). *2012 Green Infrastructure Technical Assistance Program*. EPA. from https://www.epa.gov/guidance

Harari, Y. N. (2016, August 26). *Yuval Noah Harari on Big Data, Google and the end of free will*. Financial Times. https://www.ft.com/content/50bb4830-6a4c-11e6-ae5b-a7cc5dd5a28c

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet Classification. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.123

Huang, G., Liu, Z., Maaten, L. van der, & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks . *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. https://doi.org/10.1109/csci46756.2018.00084

Kim, H. Y. (2021). Analyzing green space as a flooding mitigation – storm chaba case in South Korea. *Geomatics, Natural Hazards and Risk*, *12*(1), 1181–1194. https://doi.org/10.1080/19475705.2021.1920478

Koehorst, M. (2020, November 5). *The effect of green spaces and urban trees on reducing flood risk*. PANORAMA. from https://panorama.solutions/en/solution/effect-green-spaces-and-urban-trees-reducing-flood-risk#:~:text=The%20soil%20and%20roots%20of,water%20during%20intense%20rain%20events

Lin, C.-Y., Ackerman, A., Johnston, D., Tian, G., & Liu, Y. (2022, June 12). LiDAR Operation and Digital Modeling Visualization to Communicate Stormwater Management at Green Spaces in Developing Regions. *Workshop on Visualization in Environmental Sciences (EnvirVis)*. https://doi.org/10.2312/envirvis.20221056

Mazzoleni, M., Dottori, F., Cloke, H. L., & Di Baldassarre, G. (2022). Deciphering human influence on annual maximum flood extent at the global level. *Communications Earth & Environment*, *3*(1). https://doi.org/10.1038/s43247-022-00598-0

PricewaterhouseCoopers. (2023). *Turn ESG theory into action From meaningful change to measurable value*. Environmental, Social and Governance (ESG). from https://www.pwc.com/gx/en/issues/esg.html

Rentschler, J., Salhab, M., & Jafino, B. A. (2022, June 28). *Flood risk already affects 1.81 billion people. climate change and unplanned urbanization could worsen exposure*. World Bank Blogs. from https://blogs.worldbank.org/climatechange/flood-risk-already-affects-181-billion-people-climate-change-and-unplanned

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Neural Networks for Large-scale Image Recognition. *2015 International Conference on Learning Representations*. https://doi.org/10.1109/slt.2016.7846307