Socioeconomic Impacts of the Kerala Flooding in 2018: A High Spatial Resolution and Multi-time Step Analysis

Hikari Murayama[a,b,1,2], **Wenqi Luo**[c,1], **and Yin Qiu**[c,1]

[a]Energy and Resource Group, University of California - Berkeley, Berkeley, CA 9720; [b]Global Policy Lab, Goldman School of Public Policy, University of California - Berkeley, Berkeley, CA 9720; [c]School of Information, University of California - Berkeley, Berkeley, CA 9720

This manuscript was compiled on May 3, 2021

Kerala, India experienced a massive flooding event that displaced millions and killed nearly 500 people in 2018. Although work has been conducted to assess socioeconomic impacts of flood events, research is needed to assess these factors at a spatially resolute scale and at multiple time steps. We investigated flooding impacts through the lens of multiple time steps (pre-event, post-event, and present day). To increase future usability of this workflow, all data are free and publicly available. We used an open-source CNN model with MAXAR and Google Static Imagery and observed that 1.3% of buildings had minor damage, 0.07% had major damage, and 2.5% were destroyed. A large portion of the destruction is in Alappuzha city. Conversely, we observed only 0.33% of buildings with minor, 0.01% with major, and 0.23% with complete recovery. Regressing the log of a weighted damage score against Demographic Health Survey (DHS) wealth index while holding Lasso selected climate variables (evaporation from open water, evaporation from vegetation, elevation) constant showed significant correlation (slope=-1.39, p=0.084) with wealth. Regressing the log of damage ratio with the log of recovery ratio yields a coefficient of 0.48 (p=0.017). Damage and degree of flood inundation by cluster share a weak relationship. In short, we observed a disproportionate immediate impact on poorer populations and minimal recovery within a two year span. Further work is required in conjunction with stakeholder engagement to further improve the model and expand the study region.

Building Damage | Natural Disaster | Convolutional Neural Network | High Resolution Imagery

Kerala, a densely populated Indian state located on the Southwestern coast, is highly vulnerable to natural disasters which are exacerbated by climate change. The most impending, common danger in the area is from floods that also cause landslides. Torrential rainfall from June to August 2018 caused massive flooding triggering hundreds of landslides and affected the majority of the villages in the state. Although the local government initially estimated damages to be between 3 and 4 billion USD, these estimates are believed to be lower than in actuality (European Commission's Directorate-General for European Civil Protection and Humanitarian Aid Operations, 2018 (1)). It seems especially crucial to quantify impacts on different demographics within a population since there is a bidirectional relationship between poverty and climate hazard (i.e. poorer households are more likely to be located in flood prone areas, living in areas of high flood risk may exacerbate poverty) (Kawasaki et al., 2020 (2)). Having a streamlined, quickly executable methodology to quantifying impacts to direct humanitarian efforts would serve impacted communities in both the short and long term (Oddo et al., 2018 (3)).

There are many studies that have tried to assess how this devastating event transpired and the magnitude of its impacts using modeling techniques. One group used a 4km resolution Weather Research and Forecasting in conjunction with hydrological models to assess the root causes. Although they found that flooding impacts were less than what could have occurred prior to the industrial period, they concluded that continued changing climate patterns would increase flood risk in the future. By modeling outcomes for both pre-industrial, a control, and using RCP 8.5 climates they found that the monsoon pressure systems would increase by 36%, six major reservoirs would need 34% more capacity to handle rainfall, and the southern boundary of flooding would be extended (Hunt & Menon, 2020) (4). This study was able to pinpoint the major factors that contribute to flooding in Kerala, but did not extend beyond climate variables to assess socioeconomic impact. Furthermore, there is a large amount of variance in results from flood damage models as they are sensitive to exposure (i.e. asset values) and vulnerability. There is a great need for expert knowledge to calibrate these models and aggregate land use data and adjust asset values according to the local economic landscape (Jongman et al., 2012 (5)). There are multiple research groups that have attempted to assess the impacts and damages caused by the event capitalizing on the plethora of satellite remote sensing data that is available. Vishnu et al. (2019) (6) used Sentinel-1A and Sentinel-2A radar data to calculate a Modified Normalized

Significance Statement

Flood impact studies can be conducted in granular detail without proprietary data, expensive computational models, or extensive fieldwork. We address this gap by using an open source Convolutional Neural Network (CNN) model and high-resolution satellite imagery from MAXAR and Google to conduct a pixel-by-pixel classification of damage severity. We focused on the 2018 flood in Kerala, India and assessed these impacts on multiple time steps (pre-event, post-event, and present day). Our model showed that the extent of damage far surpasses the amount of recovery thus far. Moreover, the flood damage disproportionately impacts poorer populations.

HM conducted project scoping, MAXAR image downloading and processing, climate data downloading and processing, and flood inundation comparison. WL downloaded and processed Google Static Imagery and modified/ran CNN models. YQ downloaded DHS data, took part in WL's analysis, and conducted regression analysis.

The authors declare no conflict of interest.

[1]HM, WL, YQ contributed equally to this work

[2]To whom correspondence should be addressed. E-mail: hikari_murayamaberkeley.edu

Difference Water Index (MNDWI) prior to and during the event. They estimated a 90% increase in water cover. Lal et al. (2020) (7) also used Sentinel-1A synthetic aperture radar data to assess flooding extent in conjunction with Landsat 7 and 8 data. They observed over 1100 km^2 in flooding, with the three most impacted districts being Alappuzha, Thrissur, and Kottayam. Settlements were the second most impacted land use type. Although these two papers are sufficient in assessing flooding impact, they are too spatial irresolute (10m) to assess specific flooding building damages and quantify socioeconomic impact. Parallel methodologies have been deployed in flood events outside of Kerala. One group also used Sentinel-2 Multispectral Instrument (MSI) to also use MNDWI in addition to conducting an independent component analysis (ICA) to delineate flooding extent areas (Li et al., 2018) (8).

Other methods have been developed to quantify flood building damage using spatially resolute data in conjunction with computationally intensive methods. One group presents a Convolutional Neural Network (CNN) based building damage detection network (BDD-Net) that can help conduct pixel level classification to recognize destroyed buildings during natural disasters (Shao et al., 2020 (9)). Specifically, the proposed general model was trained with a large amount of pre- and post-disaster satellite images as input for five main disaster types, including floods, across the globe. Compared with traditional methods, this new model is advantageous in several aspects: (1) it's capable of detecting different damage levels (instead of only the most severe damages); (2) it achieves higher prediction results (F1 scores) than using only post-disaster images because it can better extract building specific features and boundaries. For disaster type flooding, the model achieves a F1 score of 82.9.

Another study focused on washed-away building detection and suggested another CNN-based model to predict building damage during a tsunami in Japan (Fujita et al., 2017 (10)). Image pre-processing was relatively more complex than dealing with satellite images. Pictures were adjusted in size according to the actual building size and cropped to enclose the target building in the center of the picture. Three patch sizes were prepared: fixed-scale, size-adaptive, and resized. However in this study, using pre- and post-event image pairs didn't show any absolute advantage over only using post-event images. Both methods score about 95% in accuracy.

In addition to housing and road damages, numerous studies have also investigated floods' impact on socio-economic activities. Narayanan and Thakur (2020) (11) examined how Kerala flood impacted household finances, specifically in household budget and balance sheet items. Relative to the households in the bordering states, income for households in Kerala decreased by 16 percent (cumulative) between June and August 2018 but quickly rebounded. Household expenditure on the other hand decreased by 7 percent (cumulative) during the same period but the decline persisted for a longer period. In another study done by Sam et al. (2015) (12), researchers used the sociodemographic characteristics and Socio-economic Vulnerability Index (SeVI) to evaluate Indian rural households' vulnerability to flood hazards. Sociodemographic characteristics include literacy rate, dependency ratio and SeVI measures include items like "percent of household reporting damage to property/house due to flood in prior six years", "average durable asset diversification index", "crop/property loss". It's worth noting that both non-climate factors and flood incidence are contributive to household vulnerability to flood.

In this study we used an open-source CNN model developed by Eugene Khvedchenya, available at Github Repository, on high resolution satellite imagery (Google Static Maps, MAXAR Open Data) (13) to assess the degree of building damage in flooding-exposed areas (pre-event vs. immediate post-event), and then innovatively applied the same model to assess the recovery progress in housing patterns (immediate post-event vs. present day) (Figure 1). We then compared these results to a flood inundation map created in another study. To assess how flood damage differs by demographic attributes, we leveraged DHS survey data to conduct a regression analysis on the cluster level. By using open source data and models, we aim to demonstrate and leverage how impact assessments can be conducted without field work, and enable immediate response for humanitarian aid with little monetary investment.

Results and Discussion

Summary of damages and rebuilding. Throughout the entire region, out of 127,642,557 detected building pixels, we observed 96.08% of buildings had no damage, 1.34% had minor damage, 0.07% had major damage, and 2.5% were destroyed. On the other hand, out of 85,111,389 detected building pixels in the recovery phase, we saw 0.33% had minor recovery, 0.01% had major recovery, and 0.23% were rebuilt or new. When comparing the two, it is obvious that the amount of recovery is lacking compared to the level of damage that occurred during the flood (Figure 2). When assessing these changes by cluster, we created Figure 3 to show average damage and recovery scores, with the size of each point depicting the number of total damaged or recovered pixels. Clusters 170572, 170258 and 170284 are centered around the city of Alappuzha, the most damaged region. It is problematic that there were no observed recovery (no level 2, 3, 4) in 170572 and 170284, and only 191 pixels of minor recovery in 170258. It is counterintuitive since usually the most damaged areas should have had the most recovery (Figure 4). Cluster 170418 exhibited the most recovery (Figure 5). Clusters 170103, 170406 and 170158 are scattered in the top northwest and the bottom southeast, where the damages were relatively small so the recovery changes were very little to none. Additional work should be conducted to assess the possibility of misclassification by our model, or if the problem lies in the less granular resolution reduction of Google Static Imagery.

Correlation between pre-flooding wealth scores and building damages. The results of Regression 1, as shown in Table 1 and Figure 6 , indicated a negative correlation between wealth score and the log of damage ratio. The coefficient of wealth score is -1.39 and the p-value is 0.084. Holding the three climate variables constant, we are 90% confident that 1 unit increase in pre-flooding average wealth score of the cluster is associated with 1.4% reduction in the damage ratio. Since the DHS wealth score takes into account multiple aspects of demographic information of the household, including the quality of housing, we believe it is a good proxy for household clusters and has been used in numerous other studies. Thus we are able to conclude that wealthier populations in the Alappuzha district suffered less than the poorer counterparts

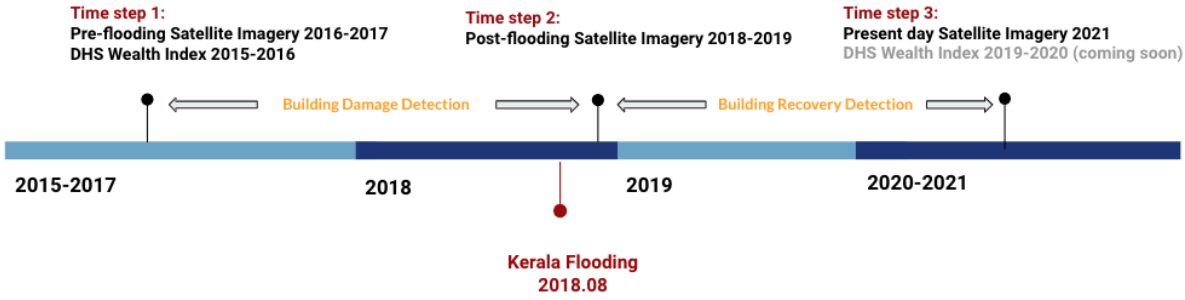

Fig. 1. Timeline of study period defining the three time steps relative to the 2018 flooding event

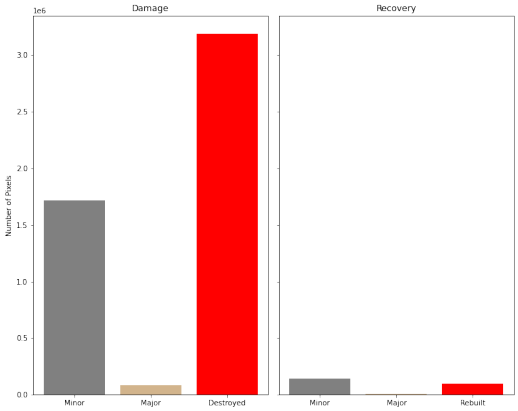

Fig. 2. Bar chart comparing number of pixels and their damage level in the pre/post comparison (left) and the number of pixels and their recovery level in the post/current day comparison (right)

in terms of housing damages in the Kerala flooding.

The coefficients of the climate control variables are also statistically significant. We find that the lower the elevation of the cluster, the more damage it will have. This agrees with our intuition that lower elevations are more prone to flooding. We also find that the more evaporation from open water surfaces (except oceans) the less damage the cluster will have, while the more evaporation from vegetation transpiration the more damage the clusters will have.

Focusing on water surface evaporation, it is important to note that the drivers of the rate are water temperature, air temperature, air humidity, and wind speed. It is difficult to parse out the direct relationship with building damage without conducting in-depth field work in the region comparing various water bodies in the region over time. Furthermore, we can think about this in the context of evaporation from soil as well. One study observed that flooding rapidly increases evaporation rates and only returns after a soil salt crust is formed after a dry period (Li & Shi, 2019) (14). This correlation may indicate that the monsoon season had already sufficiently saturated the ground soil and thus enabling more water to flood the area with subsequent rainfall. Therefore, the households living surrounded by more inland open water surfaces (e.g. lakes, rivers) are associated with less building damages, potentially because, first, the lakes and rivers were able to alleviate the flooding impacts in some way, and second, our detection model was less capable of detecting damages when closer to water surfaces.

Table 1. Regression 1: Correlation between Damage and Wealth

Independent Variables	Coefficient	Std. Err.	P-value
Constant	-1.465*	0.769	0.057
Wealth score	-1.393*	0.807	0.084
Evaporation from open water	-2744.593**	1147.957	0.017
Evaporation from vegetation	36970***	11800	0.002
Elevation	-0.074**	0.033	0.025
Dependent Variable	Log (Damage Ratio)		
R-squared	0.342		
No.observations	35		

*, **, *** indicates significance at 90%, 95% and 99% level

For considering vegetation transpiration, if there is an increase in transpiration rate, there may be less water runoff into streams and rivers, which could lead to more damage from floods. However, this is not the relationship we see here– we attribute this to the 3 month period we have aggregated under with excessive moisture may be complicating this relationship. To put it more plainly, for the households living in more vegetated areas (e.g. forest, farm lands), they could be more likely to live in less urbanized areas. Therefore, our model hypothesizes that the more vegetation found in the clusters the more damages they may have encountered during the flooding.

Correlation between building damages and building recoveries. The results of Regression 2, as shown in Table 2 and Figure 7, indicated a positive correlation between the log of damage ratio and the log of recovery ratio. The coefficient of damage ratio is 0.48 and the p-value is 0.017. We can be 95% confident that a 1% increase damage ratio is associated with 0.48% increase in recovery ratio. This is a non-trivial and exciting observation. In the relative terms, those more damaged areas have witnessed more rebuilding and recoveries, including some improvement to the houses and reconstructing new houses. Over the two to three years after the flooding, it seems resources have been targeted to the clusters that are in greater need for post-disaster recovery. However, we will need additional research to understand how the post-flooding rebuilding needs in different areas have been met in the absolute terms. Future work could expand on this analysis to be on the pixel after model calibration for the recovery phase and with sufficient storage to save all model outputs as images.

Correlation between pre-flooding wealth scores and building recoveries. The results of Regression 3, as shown in Table 3, indicated a slightly negative correlation between the pre-

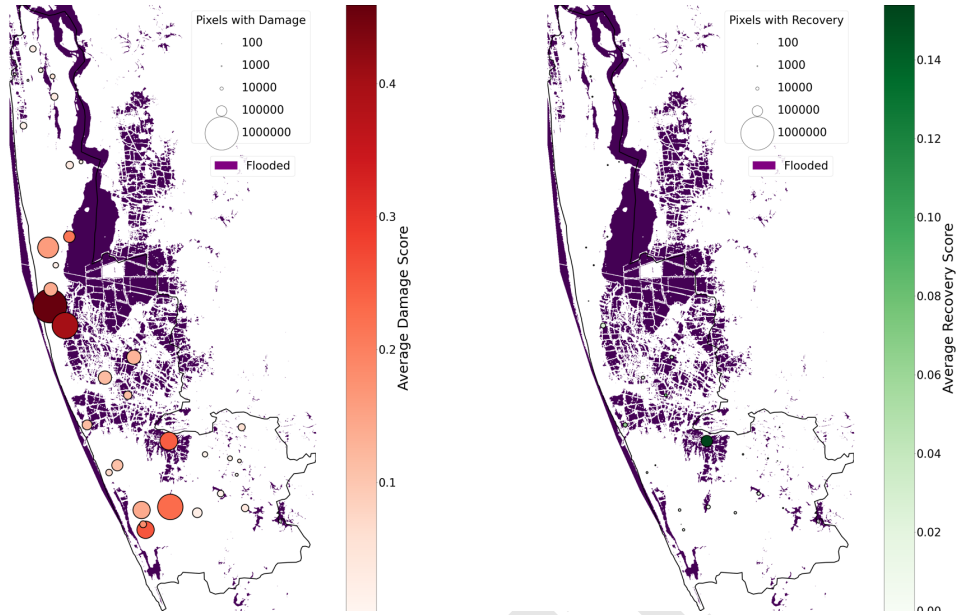

Fig. 3. (Left) Map of Alappuzha colored by average damage score and sized by number of damaged pixels. (Right) Map of Alappuzha colored by average recovery score and sized by number of recovered pixels.

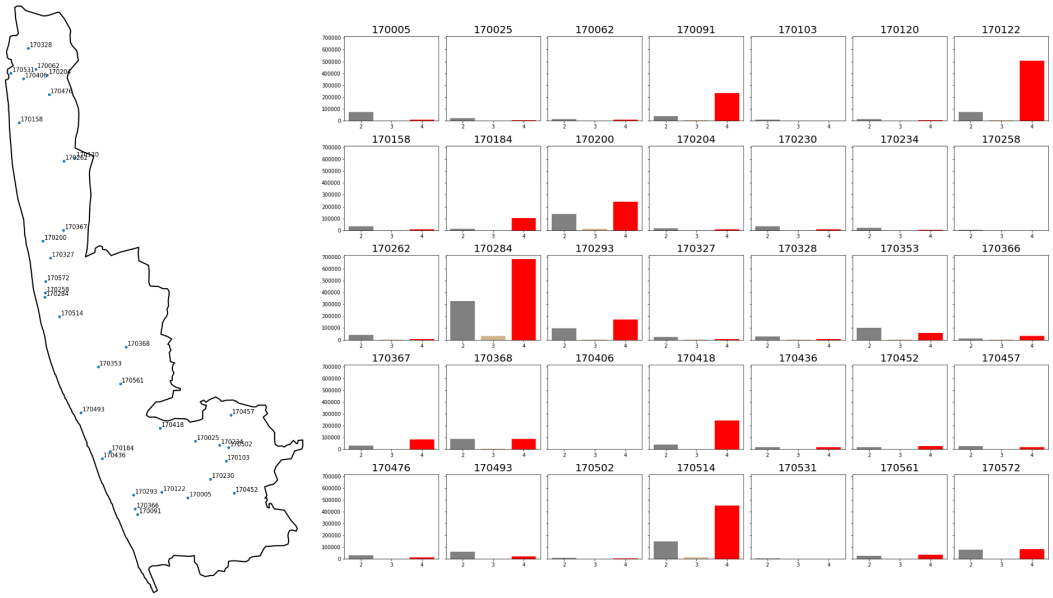

Fig. 4. Damage level by cluster with 2 indicating minor damage, 3 indicating major damage, and 4 indicating destruction

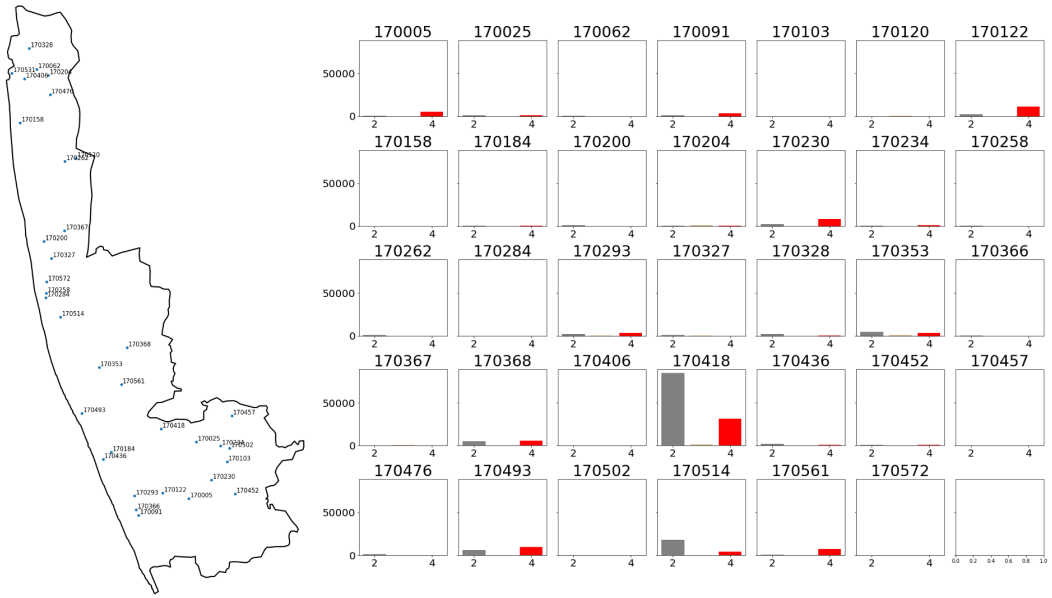

Fig. 5. Recovery level by cluster with 2 indicating minor rebuilding, 3 indicating major rebuilding, and 4 indicating newly rebuilt

Fig. 6. Regression plot between pre-flooding wealth score and log(damage ratio).

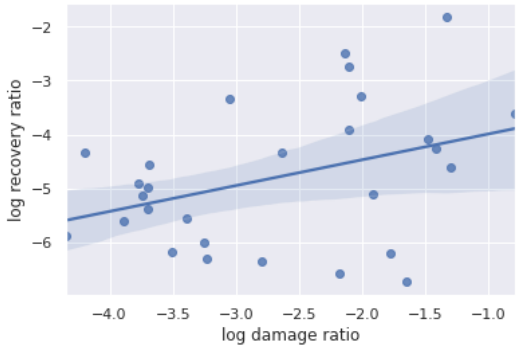

Fig. 7. Regression plot between log(recovery ratio) and log(damage ratio).

Table 2. Regression 2: Correlation between Recovery and Damage

Independent Variables	Coefficient	Std. Err.	P-value
Constant	-3.512***	0.691	0.000
Log (Damage Ratio)	0.480**	0.202	0.017
Dependent Variable	Log (Recovery Ratio)		
R-squared	0.141		
No.observations	28		

, * indicates significance at 90%, 95% and 99% level

Table 3. Regression 3: Correlation between Recovery and Wealth

Independent Variables	Coefficient	Std. Err.	P-value
Constant	0.020**	0.008	0.015
Wealth score	-0.001	0.036	0.984
Dependent Variable	Log (Recovery Ratio)		
R-squared	0		
No.observations	28		

, * indicates significance at 90%, 95% and 99% level

flooding wealth score and the log of recovery ratio. However, the magnitude of coefficient for wealth score is very small (0.0007) and the result is not statistically significant (p-value is 0.975). We were not able to draw any conclusions. We expect to finalize this regression analysis when the 2019-2020 DHS survey data becomes available.

Comparison with flood inundation map. We compared our damage results to the flood inundation map created by Tiwari et al. (2020) (15) and saw no significant correlation between the number of pixels by damage level with the mean flood

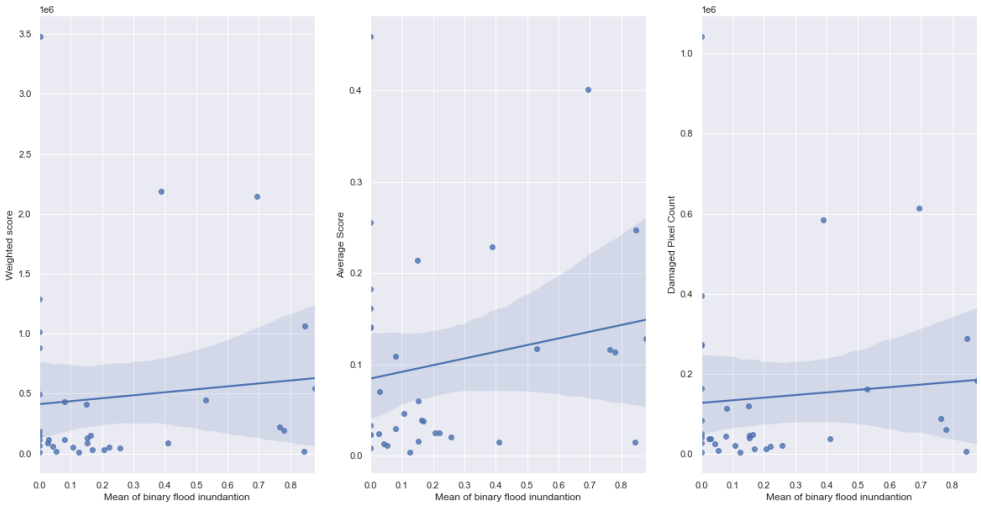

Fig. 8. Correlation between weighted damage score (left), average damage score (center), and damage pixel count (right) against the mean of the binary flood inundation indicator.

inundation (Figure 8). We also assessed this correlation by looking at the total weighted damage, mean damage score, and damaged pixel count by cluster. All three correlations exhibited a positive correlation, but their R^2 values were low (0.157, 0.288, and 0.163 respectively) showing a weak relationship.

Although there are no clusters that are square in these inundated areas, it is clear that the recovery that has occurred since the flooding in 2018 does not cover the extent of the damaged area (Figure 3). Furthermore, our results show that the flood inundation map may be accurate to highlight flooding extent, but may not be sufficient in highlighting areas that have building damage.

Limitations

Due to the limited availability of post-flooding MAXAR images, we were not able to examine the economic impact of the flood for the entire state of Kerala. Alappuzha district was selected for its relative abundance of both pre-/post-flooding images, and images that were predominantly clouds were also filtered out. Consequently, in our correlation analysis we only got a small sample size of 34 DHS clusters covered by these images. Moreover, both pre-/post-flooding MAXAR images were also taken over the course of a few years. This could be an especially critical issue for post-flooding images as building damage and inundation might be differently represented between 1 month and 5 months after the flood. In our analysis of building damage assessment, it's hard to control for such variances in the time that image were taken, which might contribute to the final damage level predictions. Other caveats in the current image-based analysis include the difference in the quality of images between MAXAR (on average 2000 pixels by 2000 pixels) and Google Static Images (400 pixels by 400 pixels). There are certainly rooms for additional image processing (i.e. cloud removal). In the present study, the limited RAM space only allowed the model outputs to be saved as data frames rather than images, which made it hard to validate prediction results at pixel level. Though the inundation map (Tiwari et al., 2020) (15) was used as a great reference to compare against the model's building damage predictions, they are in nature very hard to validate at cluster level. It would be helpful to use other proxies to further assess those results as prediction classes are pretty imbalanced, with the "major damage" level extremely rare to occur. Furthermore, DHS survey clusters are not geographically precise; we could solely rely on each cluster's geospatial centroid to aggregate image pairs by the cluster centroid they are closest to, assuming it's also the cluster they were located in, which however, needs further validation.

Conclusion

Through this study we have found that the 2018 Kerala flooding caused damage across Alappuzha, with the city of Alappuzha showing the most loss. We observed that wealthier populations suffered less building damage than their poorer counterparts. Our analysis also shows that resources have been targeted to rebuild areas that were more heavily damaged by the floods. However, the extent of recovery is far less than the extent of damage that was ravaged. This building damage analysis, when compared to a flood inundation map, showed that flooded areas may have more damage than non-flooded areas but showed no strong correlation. This may be an indication that flood inundation maps are insufficient in determining the extent of destruction.

Beyond this study, we would like to expand the use of the current model to study the entire state of Kerala. Since Alappuzha was heavily impacted by the flooding across the district, we hypothesize that expanding the study region may yield stronger results for correlating with the DHS wealth index. Moreover, the additional step to examine the possible shift in socio-economic makeup with the 2020 DHS data could be insightful to assess resettlement by demographic characteristics. Following the work of other research studies such as Jean et al.

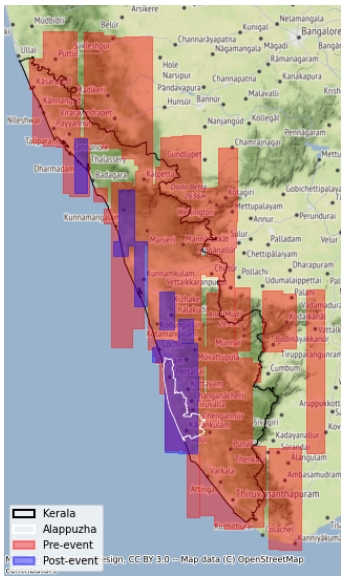

Fig. 9. Map of Kerala showing MAXAR pre-event image extent (red) and post-event image extent (blue)

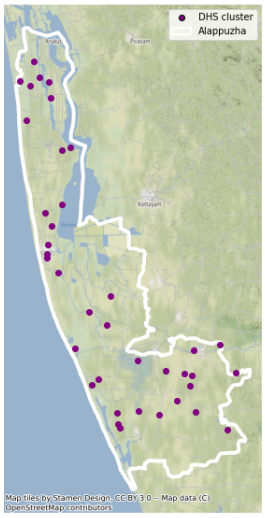

Fig. 10. Map of Alappuzha showing DHS cluster locations

(2016) (16) and Blumenstock et al. (17), it would be vital to adopt their methodologies to create spatially resolute wealth index maps to further improve this study.

Although there are limitations to this study, this methodology is novel in its application of a single model to assess damage and recovery. Furthermore, because this model is open source and the data we used are all free and publicly available, it is more easily accessible to stakeholders who may not have funds to apply proprietary methods. Future improvements to the model and in-depth discussions with policymakers and humanitarian aid organizations are vital to bring this methodology to the forefront.

Materials and Methods

Data Processing. MAXAR, a private space and technology company, provides free and publicly accessible high resolution imagery for "sudden and onset crises". For the 2018 Kerala flood pre-event imagery was taken between November 22, 2016 and July 12, 2018, while post-event imagery was taken between August 24, 2018 and February 3, 2019. There are 92 images (127 GB) and 15 images (15GB) available respectively, and these are non-uniformly sized and did not cover the entire study region 9. We cropped these GeoTiffs into one km^2 squares for easier processing. We also conducted additional compressions via GDAL to increase performance for later steps. Furthermore, we narrowed down our study region to areas that had post-event imagery since our study area cover was limited for this time frame. We also focused on the district of Alappuzha since it not only had almost complete cover of images for both time frames, but also since it was the most economically impacted district during the flooding event (European Commission's Directorate-General for European Civil Protection and Humanitarian Aid Operations, 2018) (1).

For post-flood recovery detection, we obtained present-day daytime satellite imagery of the entire Alappuzha district using Google Static Maps api. We used Kerala and Alappuzha shapefiles to determine boundaries of our downloaded area. The images were taken in 2021 and were downloaded at zoom level 16, with a pixel resolution of about 2.5m. The image size was set to 400 pixels by 400 pixels and each image covered one square kilometers, to match up with the sizes of MAXAR images. In total we got 1489 google static images that correspond to the regions covered in post-flooding MAXAR imagery. We used India Demographic and Health Survey (DHS) 2015-2016 as the primary source for household economic and demographic statistics at cluster level 10. In addition to household wealth index, the India DHS includes a variable on housing type counting the number of kachha (made from mud, thatch, or other low-quality materials), semi-pucca (use partly low-quality and partly high-quality materials), and pucca houses (made with high quality materials throughout, including the floor, roof, and exterior walls). Using a nearest neighbor algorithm, we assigned a cluster id and the corresponding cluster centroid to each image. 35 out of all 38 DHS clusters in Alappuzha were covered by all available pre-/post-flooding satellite images, and 34 out of 38 clusters were covered by present day/post-flooding satellite images for later analysis. The rest of the 4 clusters are located in the south-east of Alappuzha, where there are no MAXAR images available.

To conduct our regression analysis and normalize against climate and environment related factors, we downloaded three raster datasets from Google Earth Engine. Each dataset was averaged over a three month period, to cover the monsoon period prior to the flooding event, and clipped using the Kerala district shapefile downloaded from datameet.org/maps. The first dataset is the ERA5-Land hourly ECMWF climate reanalysis which has a 0.1 degrees that contains information such as Dewpoint, surface temperature, soil temperature, heat flux, albedo, evaporation, runoff, precipitation, and vegetation. CHIRPS Pentad: Climate Hazards Group InfraRed Precipitation with Station Data provided precipitation data at a spatial resolution of 0.05 degrees. WWF HydroSHEDS Hydrologically

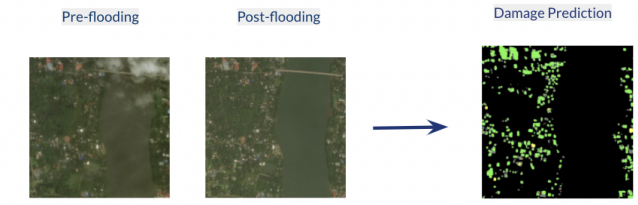

Fig. 11. Damage detection: Pre-flooding and Post-flooding image pair example.

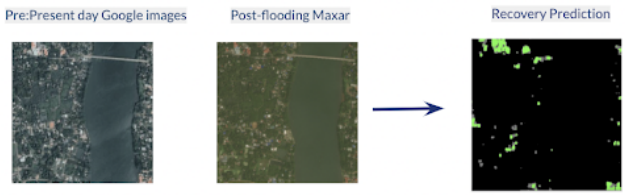

Fig. 12. Recovery detection: Post-flooding and Present-day image pair example.

Conditioned DEM provided digital elevation at at a resolution of 0.00083 degrees. A full list of these variables can be found in Table S1. After downloading, these data were averaged within a radius of 0.07°, 0.03°, and 0.006°respectively to take into account 1 to 4 neighbors but capitalize on the spatial resolution.

Damage Model. In order to quantify building damage and recovery from the flood, we adapted and applied a pretrained CNN-based neural network model from a winner solution of xView2 building damage assessment challenge (Github Repository). The original model was an ensemble of thirteen semantic segmentation submodels, and was pre-trained on xBD dataset, "one of the largest and highest-quality public dataset" of satellite imagery containing 850,736 buildings before and after 6 types of natural disasters including flooding (Gupta et al., 2019) (13). The training dataset also comes with human annotation of building location and building damage scores, which are the two main outputs of model prediction. Due to the long runtime of the original computationally expensive ensemble model, we selected one high performing submodel with a F1 score of 0.75 and the postprocessing mode of "floodfill" for our building damage and recovery detection tasks.

For building damage detection, the model took 1522 chucked pairs of pre-/post-flooding MAXAR image pairs, covering 35 out of all 38 DHS clusters. The model returned pixel-level prediction labels of building localization and damage levels (on a scale from 1 through 4) of previously located buildings. Figure 11 illustrates the process. The pixel-level predictions of each pre-/post-flooding image pair were then aggregated at DHS clusters based on the images' associated geo-coordinates.

Recovery Model. Similarly for building recovery detection, the model was applied to take present-day google static images (as the new normal days references) and post-flooding MAXAR image pairs to again predict building location and damage. Figure 12 illustrates the process. The predicted building damage here, however, was used as a proxy of building recovery (rebuilding) from the post-flood phase. In this second time step prediction, a greater predicted damage suggested a greater rebuilding of demolished/destroyed houses from post-flood damages. No predicted damage instead suggested no difference between post-flooding and now, and thus no recovery. Both time steps were compared against a flood inundation map, which was using Sentinel-1 Synthetic Aperture RADAR (SAR) data and an Otsu algorithm for August 21, 2018 with an accuracy of 94.3% by Tiwari et al. (2020) (15). We conducted a simple regression between the amount of damage and flood inundation.

Regression Preparation. With the output of the building damage detection model and the pre-flooding wealth score, we prepared the data to run a regression model to understand the correlation between wealth and damage levels. The first step was to aggregate the predicted damage levels (on a scale from 1 through 4) by the cluster id assigned to each image pair. The second step was to compute the weighted damage ratio by this formula to represent the damage level of each cluster:

$$weighted\,damage\,ratio = \frac{(destroyed \times 4 + major\,damage \times 3 + minor\,damage \times 2)}{no\,damage}$$

Since the number of buildings and characteristics of landscape for each cluster are very different, it is not sensible to compare the absolute number of pixels of damage areas across different clusters. We created a ratio between damage and no damage only among the area of located buildings so that this ratio can be more comparable across clusters. The weights added in the formula can capture information about the severity of the damages. The weighted damage ratio was then used as the dependent variable in the first regression analysis.

Similarly, with the output of the building recovery detection model, we prepared the data to run a second regression model to understand the correlation between damage levels and recovery levels in Alappuzha district. The first step was to aggregate the predicted recovery levels (on a scale from 1 through 4) by the DHS cluster id assigned to each image pair. The second step was to compute the weighted recovery ratio by this formula to represent the recovery level of each cluster:

$$weighted\,recovery\,ratio = \frac{(new\ \times 4 + major\,modification \times 3 + minor\,modification \times 2)}{no\,rebuilding}$$

Similar to our methodology to quantify damages as explained above , we created a ratio between recovery and no recovery only among the area of located buildings so that this ratio can be more comparable across clusters. The weights added in the formula can capture information about the degree of the rebuilding or modification. We call this change as recovery in general. The weighted recovery ratio was then used as the dependent variable in the second regression analysis.

Regression Analysis.

Regression 1: wealth score vs damage ratio. The first regression model is to test our hypothesis that the building damage due to the Kerala flooding is different for populations of different pre-flooding wealth levels. The independent variable of interest is the DHS wealth scores by clusters, and the dependent variable

is the damage ratio by clusters computed from our building damage detection predictions.

To get a more accurate coefficient of the wealth score, we decided to add control variables for climate, housing types, and geographic information (longitude, latitude, elevation) to the regression. We first used a MinMaxScaler to rescale all the 75 candidate independent variables to the range from 0 to 1 Then, we selected 15 variables from 75 variables using Lasso linear regression, and used the variance inflation factor function to further identify and address the multilinearity of these 15 variables. We ended up selecting three controls (evaporation from open water surfaces excluding oceans mean, evaporation from vegetation transpiration hourly mean, and elevation) to add to the right side of the equation together with the wealth score as independent variables. Since the distribution of damage ratio of clusters is right skewed, we applied log transformation to the damage ratio to make it more normally distributed before feeding into the regression. Given the wealth scores has a mean very close to its median for Alappuzha, we did not apply log transformation. We then ran an OLS linear regression.

Regression 2: damage ratio vs. recovery ratio. The second regression model served the purpose to find out the correlation between damages due to flooding and recovery. The independent variable of interest is damage ratio by clusters and the dependent variable is recovery ratio by clusters. Since both ratios are right skewed, we applied log transformation to both variables before running the model. We also excluded 6 clusters with no rebuilding or extremely low rebuilding from the dataset in order to measure a more accurate slope between damage and recovery. We then used an OLS linear regression to find out the correlation between damage and recovery.

Regression 3: wealth score vs. recovery ratio. The third regression model related the DHS reported wealth index with recovery. Since the more recent DHS survey data for 2019-2020 is not yet available, this served as an initial test to determine how the wealth of a neighborhood correlates with building repair. Assuming the wealth distribution as not changed substantially between 2015-2016 and 2019-2020, we temporarily used the 2015-2016 data. We applied a log transformation for the recovery ratio and then ran an OLS linear regression between wealth scores and log(recovery ratio) by clusters. This will be repeated with the 2019-2020 data once it is released.

ACKNOWLEDGMENTS. Modified data from the MAXAR Open Data Program via https://www.maxar.com/open-data. Climate data from ERA5-Land hourly ECMWF climate reanalysis, CHIRPS Pentad: Climate Hazards Group InfraRed Precipitation with Station Data, WWF HydroSHEDS Hydrologically Conditioned DEM were acquired through Google Earth Engine. Kerala district shapefiles were obtained from datameet.org/maps. DHS survey data is available via https://dhsprogram.com. We thank Eugene Khvedchenya for allowing us to use his model. Most importantly, we thank Joshua Blumenstock and Suraj Nair for support throughout this process.

References

1. ECDG for European Civil Protection, HA Operations, G of India, UD Programme, W Bank, Kerala post disaster needs assessment floods and landslides - august 2018, Technical report (2018).
2. A Kawasaki, G Kawamura, WW Zin, A local level relationship between floods and poverty: A case in myanmar. *International Journal of Disaster Risk Reduction* **42**, 101348 (2020).
3. PC Oddo, A Ahamed, JD Bolten, Socioeconomic impact evaluation for near real-time flood detection in the lower mekong river basin. *Hydrology* **5**, 23 (2018).
4. KM Hunt, A Menon, The 2018 kerala floods: a climate change perspective. *Climate Dynamics* **54**, 2433–2446 (2020).
5. B Jongman, et al., Comparative flood damage model assessment: towards a european approach. *Natural Hazards and Earth System Sciences* **12**, 3733–3752 (2012).
6. C Vishnu, et al., Satellite-based assessment of the august 2018 flood in parts of kerala, india. *Geomatics, Natural Hazards and Risk* (2019).
7. P Lal, et al., Evaluating the 2018 extreme flood hazard events in kerala, india. *Remote Sensing Letters* **11**, 436–445 (2020).
8. J Li, X Yang, C Maffei, S Tooth, G Yao, Applying independent component analysis on sentinel-2 imagery to characterize geomorphological responses to an extreme flood event near the non-vegetated río colorado terminus, salar de uyuni, bolivia. *Remote Sensing* **10**, 725 (2018).
9. J Shao, et al., Bdd-net: A general protocol for mapping buildings damaged by a wide range of disasters based on satellite imagery. *Remote Sensing* **12**, 1670 (2020).
10. A Fujita, et al., Damage detection from aerial images via convolutional neural networks in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*. (IEEE), pp. 5–8 (2017).
11. A Narayanan, GM Thakur, Natural calamities and household finance: Evidence from kerala floods. (2019).
12. AS Sam, R Kumar, H Kächele, K Müller, Vulnerabilities to flood hazards among rural households in india. *Natural hazards* **88**, 1133–1153 (2017).
13. R Gupta, et al., xbd: A dataset for assessing building damage from satellite imagery (2019).
14. X Li, F Shi, The effect of flooding on evaporation and the groundwater table for a salt-crusted soil. **11**, 1003 (2019).
15. V Tiwari, et al., Flood inundation mapping- kerala 2018; harnessing the power of sar, automatic threshold detection method and google earth engine. *Plos One* **15** (2020).
16. N Jean, et al., Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
17. J Blumenstock, G Cadamuro, R On, Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).