A data-mining approach to investigate *El Niño* damage in Peru

Fabio Brill[1,2,*], Silvia Passuni-Pineda[3], Bruno Espichán-Cuya[3], Heidi Kreibich[1]

[1]Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Section Hydrology
[2]Institute for Environmental Sciences and Geography, University of Potsdam
[3]INDECI/CEPIG Instituto Nacional de Defenca Civil, Lima

What?

Statistical investigation of nationwide damage survey by Peruvian authorities after the *El Niño* 2017, using explanatory features derived from topography, remote-sensing, and open data.

Why?

Neither damage models, nor statistical investigations with real observational data exist for such compound events. We aim to gain knowledge about damage processes during *El Niño* events, which is necessary to develop damage models and risk assessment approaches.

How?

1. Unsupervised clustering: grouping data into regions of different dominant processes
2. Supervised classification: learning patterns of ordinal damage grades
3. Model inspection: importance rankings and partial dependence plots reveal drivers of damage

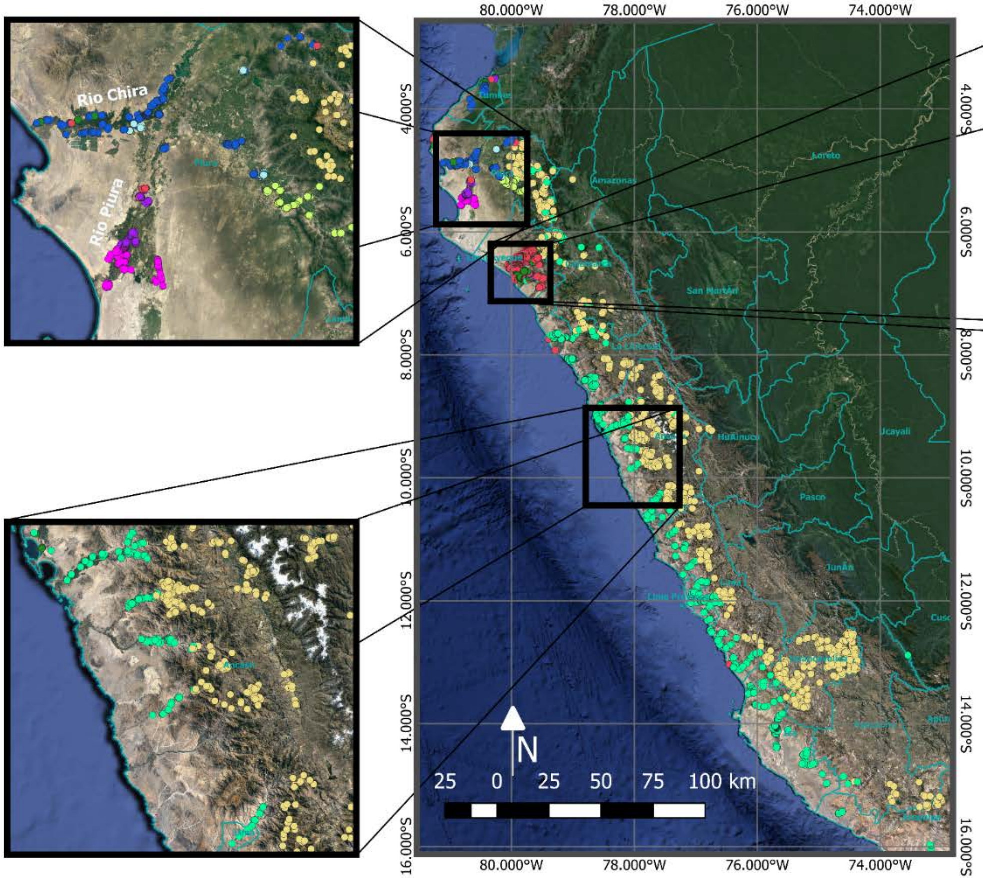
Raw Data

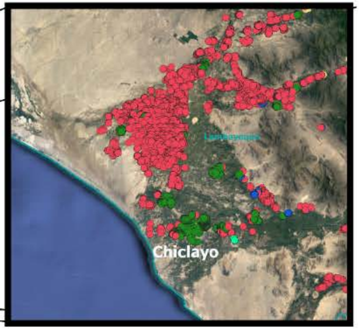
Damage: 119,675 buildings in 4 ordinal damage classes (D1-D4) from a field survey by COFOPRI

- **D1**: Non-structural damage, e.g. dented doors, broken windows, sanitation etc.
- **D2**: Moderate structural damage which is repairable; building is still habitable
- **D3**: Heavy structural damage which is repairabe; building is temporarily uninhabitable
- **D4**: Irreparable damage or collapse

Features:

- **Rainfall:** Tropical Rainfall Measurement Mission
- **Topography:** MERIT DEM
- **Water:** Global Surface Water, OpenStreetMap Waterways
- **Soil & Vegetation**: SoilGrids, TanDEM Forest/Non-Forest, Sentinel-2 spectral ratios
- **Urbanity**: Global Urban Footprint, WorldPop, OpenStreetMap Roads

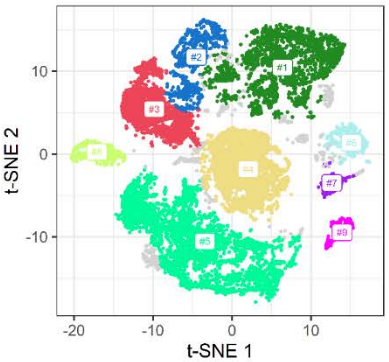




Method

t-SNE + OPTICS

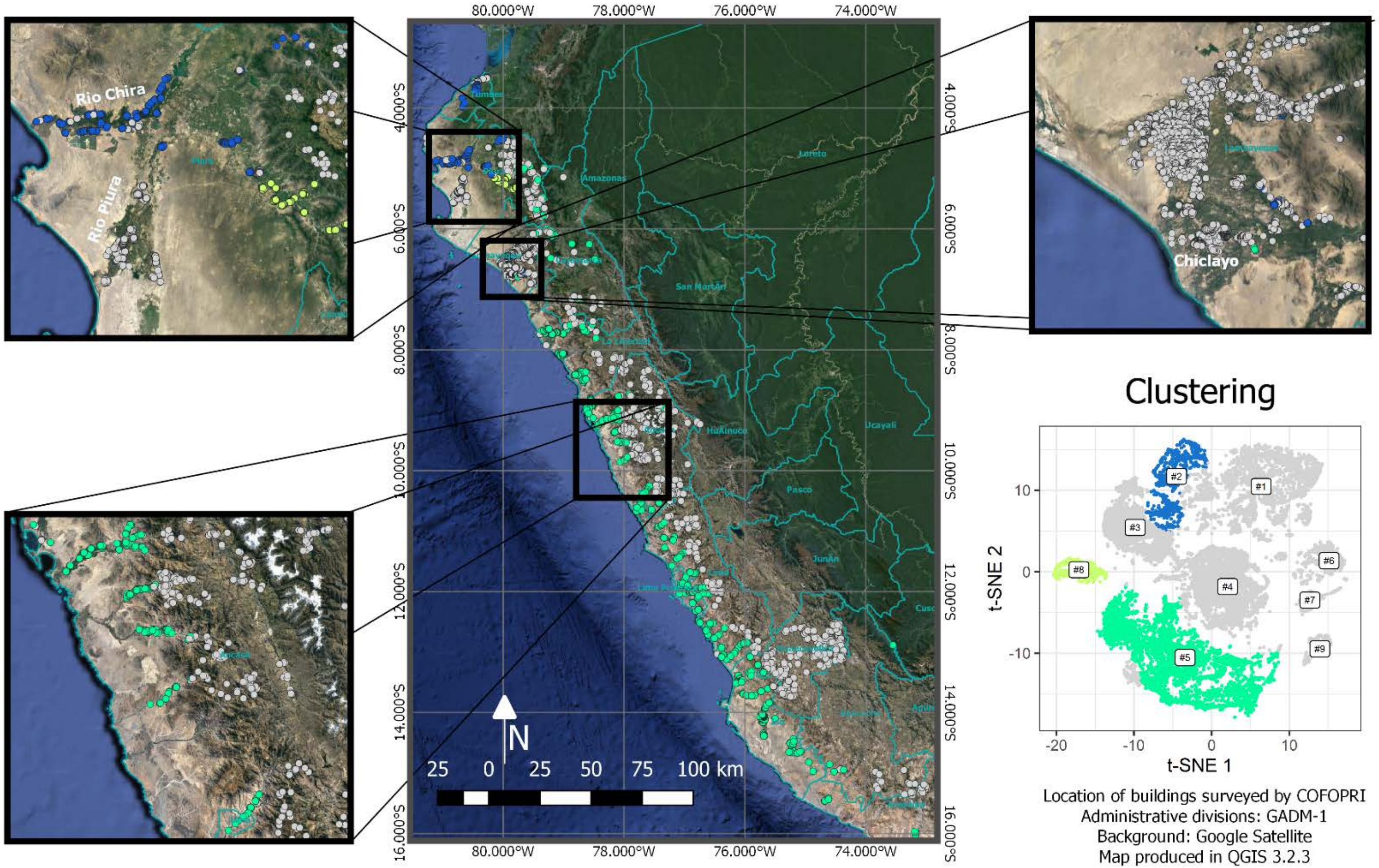
Clustering



Location of buildings surveyed by COFOPRI
Administrative divisions: GADM-1
Background: Google Satellite
Map produced in QGIS 3.2.3

Labels

#1 Urban
#2 Rivers/North
#3 Rural/Lambayeque
#4 Mountains
#5 Canyons
#6 Urban/Piura
#7 RioPiura/upper
#8 Max.Rainfall
#9 RioPiura/lower



Method

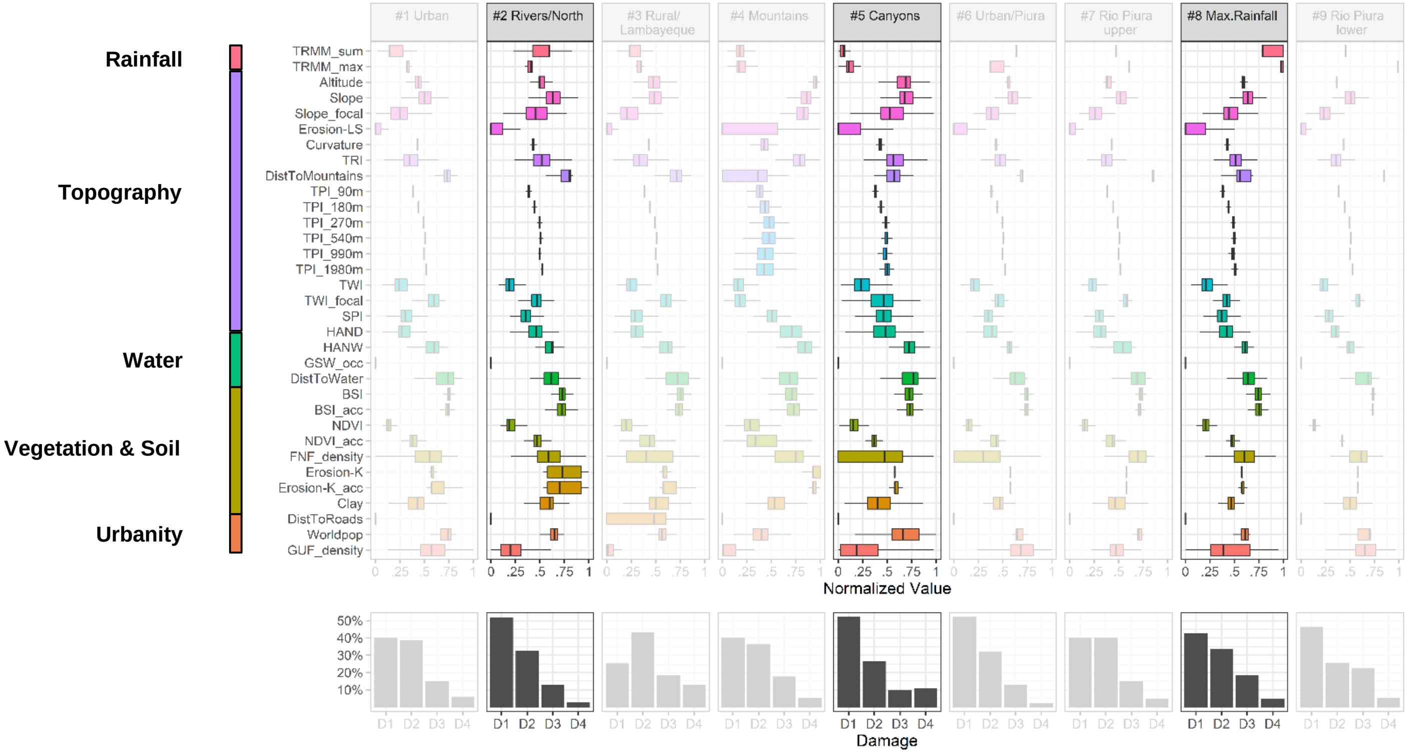
t-SNE + OPTICS

Labels

#1 Urban
#2 Rivers / North
#3 Rural/Lambayeque
#4 Mountains
#5 Canyons
#6 Urban/Piura
#7 RioPiura/upper
#8 Max.Rainfall
#9 RioPiura/lower

Feature distributions and damage frequency per cluster

Engineered features based on



More D4 cases in the canyons than in the area of maximum rainfall

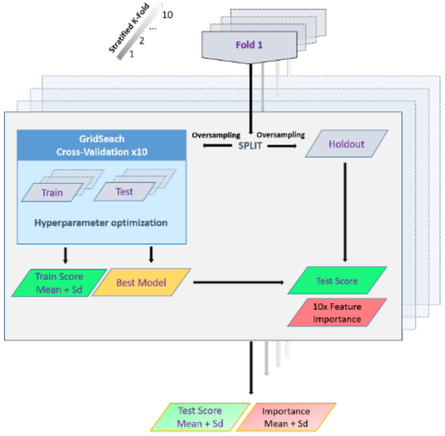
Classification

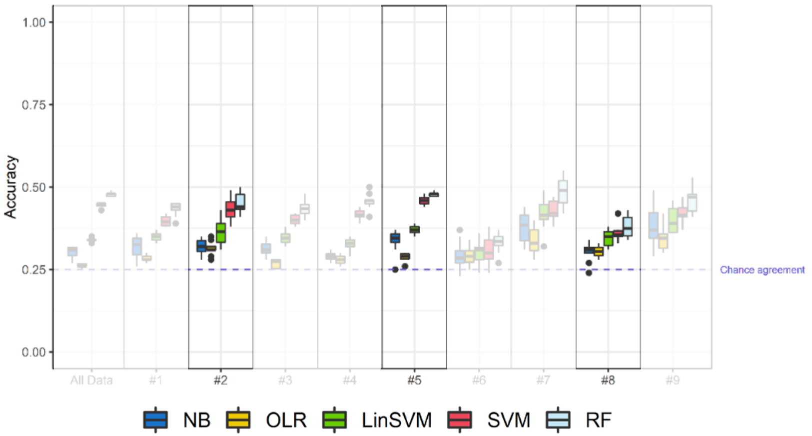
Sampling: nested cross-validation
Class balance: equal (oversampling)

Algorithms:

- Ordinal Logistic Regression (OLR)
- Naive Bayes (NB)
- Linear Support Vector Machine (LinSVM)
- Radial Support Vector Machine (SVM)
- Random Forest (RF)

Performance:
Consistently above chance agreement
Non-linear models (SVM, RF) perform better
Limitation: resolution of data, no building attributes

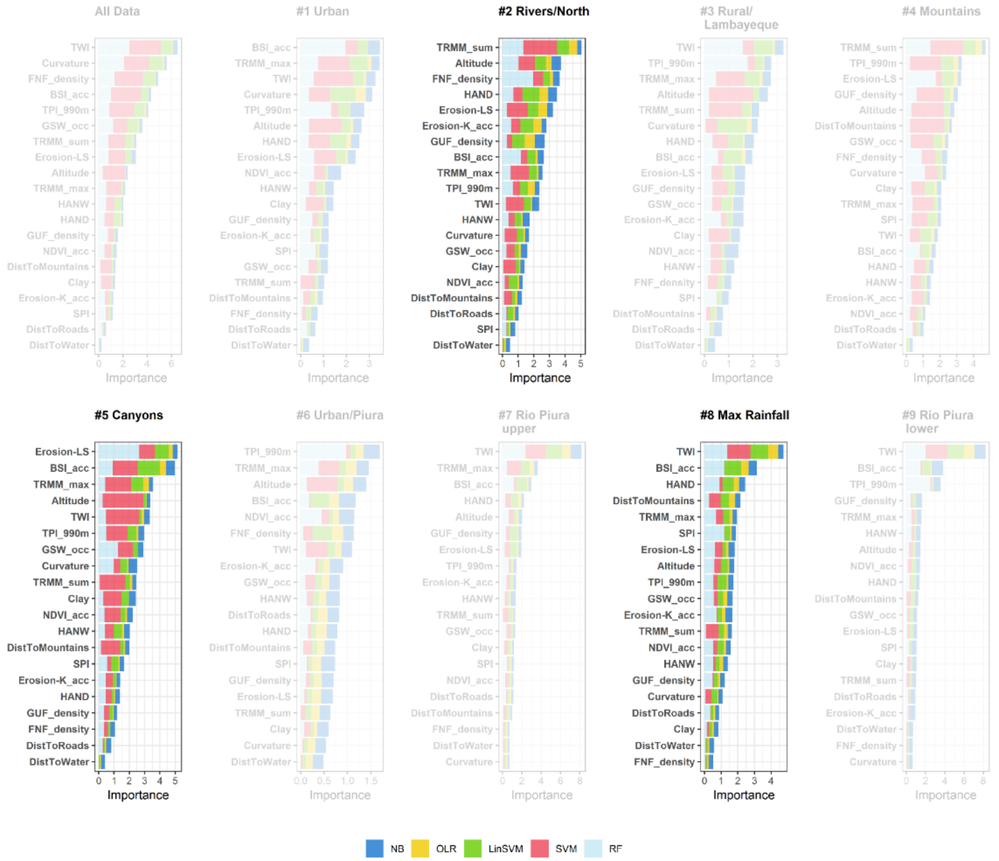




Feature Importance

- **TRMM_sum** & **TRMM_max** = Rainfall
  Consistently selected by all algorithms. Sum is more important for fluvial systems, maximum for canyons
- **TWI** = Topographic Wetness Index
  Most important for low elevation and high rainfall
- **HAND** = Heigth Above Nearest Drainage
- **Erosion-LS** = Slope length and steepnes factor
- **BSI_acc** = Bare Soil Index, weighted along the flow accumulation raster
- **FNF_density** = Forest cover within 1km²

Importance was computed as drop of model skill, when features are randomly permuted. This initial score was normalized for all algorithms and weighted by the model skill to create an aggreated ranking, while preserving the individual rankings in the visualization. Note that those feature which define a cluster have low variance within this same cluster, and will not be „important“, e.g. rainfall maximum is not dominating in cluster #8



Partial dependence plots

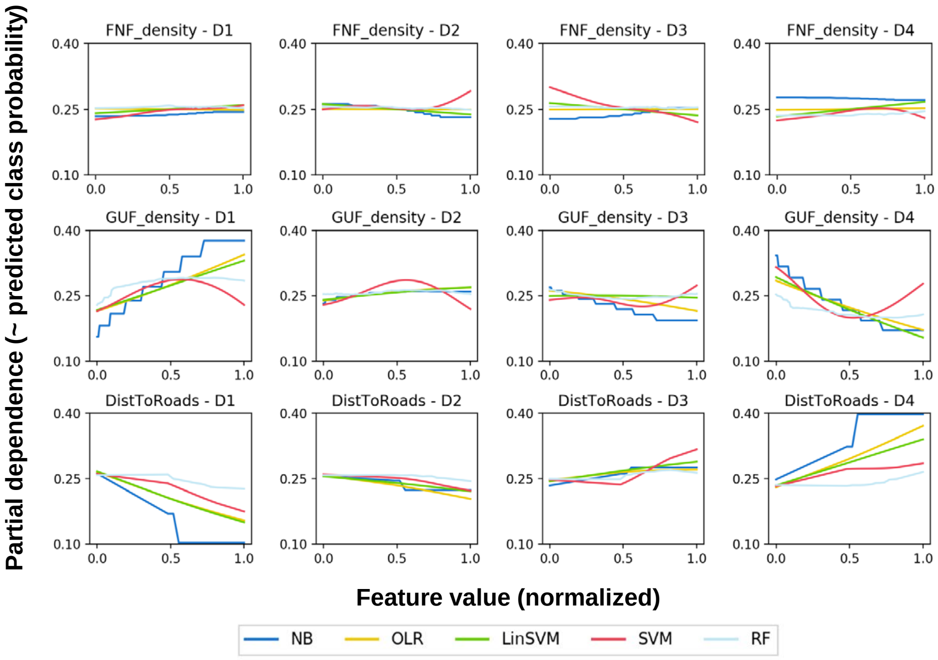
- Using **all data**, i.e. without clustering, to better understand the model behaviour

- **Forest cover**, as indicated by **FNF_density,** exhibits no meaningful net-effect in the PDP. This contradicts the importance ranking*, but is more in agreement with our expectations.

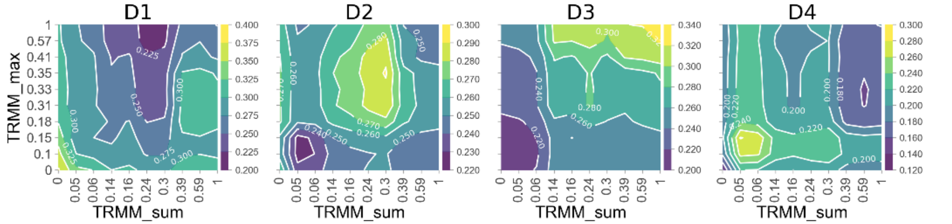
- **Urbanity**, as indicated by **GUF_density** and **DistToRoads**, was not among the top features, but has a strong effect on the predictions: the more urban, the lower(!) the damage of individual buildings.

- 2D interaction plots further show that **Rainfall** (**TRMM_sum & TRMM_max**) seems to cause damage D1-D3 in ascending order, but fails to explain D4 (collapsed).

*The importance ranking for „All Data" is dominated by complex algorithms, due to the low performance of linear algorithms, and therefore rather difficult to interpret.

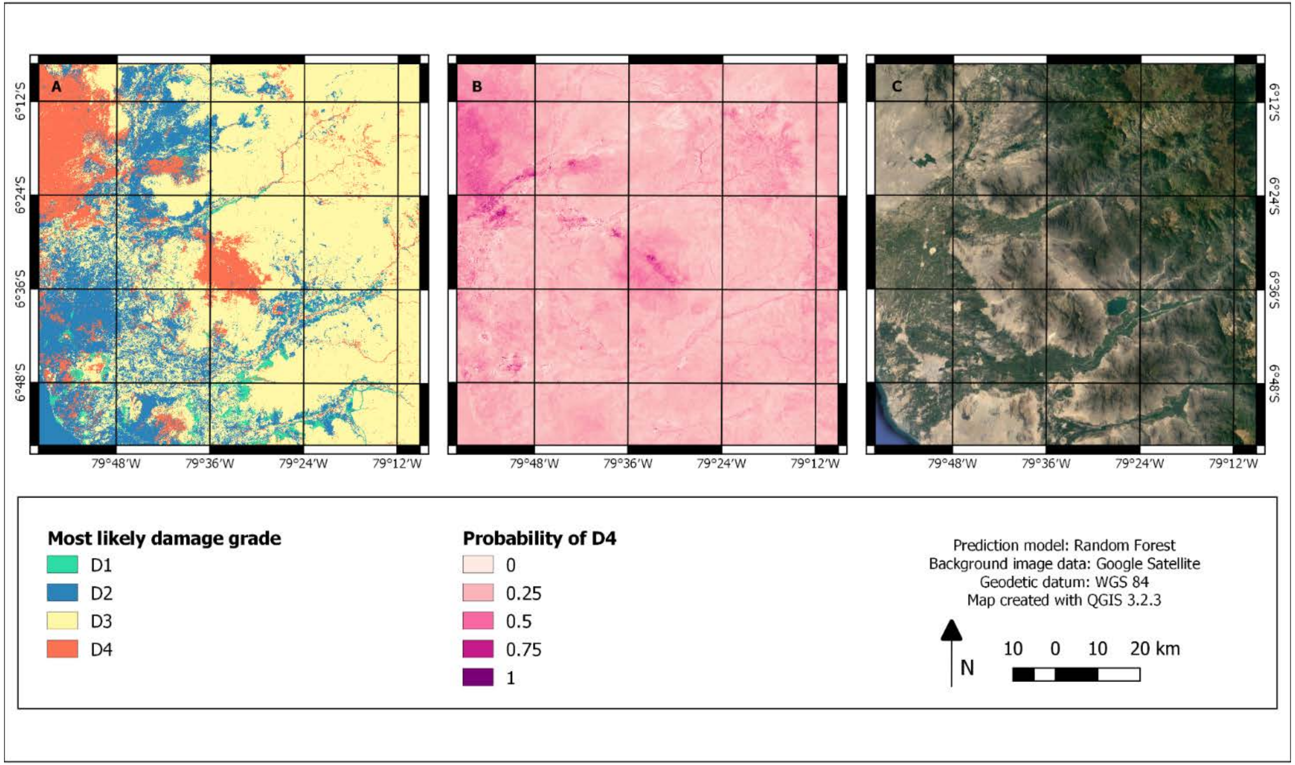


Only RF



Damage probability map (example based on RandomForest)

- Entirely data-driven
- Visualizes model behaviour: in this case, channels and desert areas were learned to be dangerous, while urban areas seem rather safe in case of *El Niño*
- Potential application of a damage model, e.g. to help identify critical areas for spatial planning. Could be intersected with exposure.
- Limitation: this example is event-specific for 2017, due to the used rainfall data



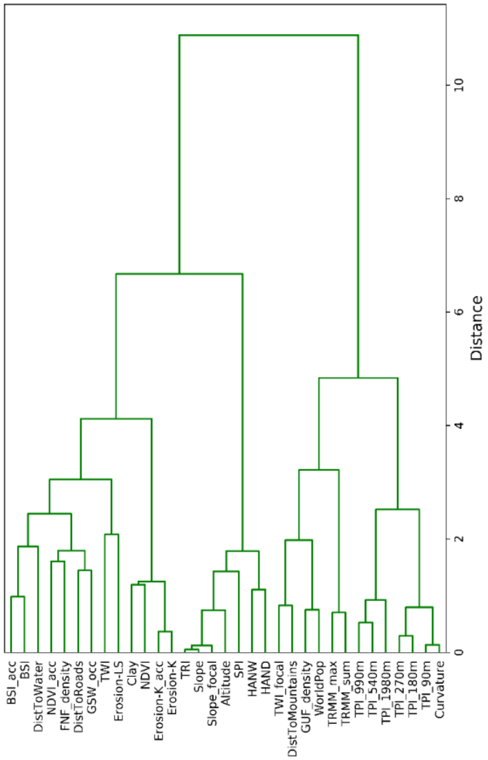
Thank you

Appendix

Candidate features, data, and software

Feature	Explanation	Data Source	Software
Altitude	Absolute elevation above sea level	MERIT DEM	–
Slope	Derivative of elevation in steepest direction	MERIT DEM	R spatialEco
Slope_focal	Focal mean of Slope	MERIT DEM	R raster
Curvature	Second derivative of elevation. Total curvature contains planform and profile directions	MERIT DEM	R spatialEco
TRI	Terrain Ruggedness Index. Information on heterogeneity of elevation	MERIT DEM	R spatialEco
TPI x6	Topographic Position Index. Calculated as elevation difference between a cell and the average value in a defined neighbourhood. Single-scale TPI identifies local hills or depressions. Multi-scale TPI can be used to classify complex landscapes. We used 6 scales: 90, 180, 270, 540, 990, 1980 [m]	MERIT DEM	R spatialEco
SPI	Stream Power Index. Represents the force of a flow, with high values in steep areas	MERIT DEM	GRASS r.watershed
TWI	Topographic Wetness Index. Represents potential accumulation of water, high values in flat areas	MERIT DEM	GRASS r.watershed
TWI_focal	Focal mean of TWI	MERIT DEM	R raster
Erosion-LS	Slope length and steepness (LS factor) of the Universal Soil Loss Equation	MERIT DEM	GRASS r.watershed
DistToMountains	Distance to steep terrain. Computed as Euclidean distance to a binary layer from Otsu's threshold on Slope_focal	MERIT DEM	Python scikit-image GDAL proximity
HAND	Height Above Nearest Drainage. Normalization of the terrain, where streams are set to 0 and height difference is computed along the flow direction	MERIT Hydro	–
HANW	Height Above Nearest Water. Same as HAND for different stream raster from GSW and OSM	Globasl Surface Water (GSW) + Open Street Map (OSM)	GRASS r.stream.distance
DistToWater	Distance to water, along the flow direction	Globasl Surface Water (GSW) + Open Street Map (OSM)	GRASS r.stream.distance
GSW_occ	Global Surface Water occurrence	Global Surface Water (GSW)	–
BSI	Bare Soil Index. Spectral ratio which identifies bare soil	Sentinel-2	Google Earth Engine
BSI_acc	BSI weighted along flow accumulation	–	GRASS r.accumulate
NDVI	Normalized Difference Vegetation Index. Spectral ratio which identifies vegetation	Sentinel-2	Google Earth Engine
NDVI_acc	NDVI weighted along flow accumulation	–	GRASS r.accumulate
Clay	Percent clay in topsoil	SoilGrids	–
Erosion-K	Erodibility (K factor) of the Universal Soil Loss Equation	SoilGrids	GRASS r.uslek
Erosion-K_acc	Erosion-K weighted along flow accumulation	–	GRASS r.accumulate
FNF_density	Forest density. Maximum when all FNF cells within an 11x11 window are classified as forest	TanDEM Forest / Non-Forest (FNF)	R raster
GUF_density	Urbanity. Maximum when all GUF cells within an 11x11 window are classified as urban	Global Urban Footprint (GUF)	R raster
Worldpop	Population density in people per pixel	WorldPop	R raster, GDAL warp
DistToRoads	Euclidean distance to roads	OpenStreetMap (OSM)	GDAL proximity
TRMM_max	Rainfall 3-hour maximum	Tropical Rainfall Measurement Mission (TRMM), Product TRMM/3B42	Google Earth Engine
TRMM_sum	Rainfall sum from January to April 2017	Tropical Rainfall Measurement Mission (TRMM), Product TRMM/3B42	Google Earth Engine

Total: 33



Hierarchical clustering on Spearman correlation