Causal Analysis of climate indicators on extreme flooding with a case study on the Pakistan Floods of 2022

Tazbia Fatima, Columbia University, New York

December 15, 2022

1 Abstract

One third of Pakistan is currently flooding today, taking at least 1100 lives and displacing 33 million people. This comes after the country experienced a heatwave in May and June 2022 subsequently causing melting of the mountain ice caps which coupled with the heavy monsoon rains have caused these floods. This indicates a crippling global climate crisis. Multiple countries observed concurrent heatwaves this summer. In this study, we apply a four-step causal analysis model to find a pattern of "climate indicators" that can help predict floods. We analyzed data from 2000 to 2022 to capture both the 2010 and 2022 floods. We also studied two provinces of Pakistan separately as their terrain and hydrological systems are distinct. We found that snowmelt had an 81.4% positive causal effect on surface runoff (floods) in the northern province of Khyber Pakhtunkhwa. We also found that the moisture heavy winds coming from the monsoon depressions in Bay of Bengal have a positive causal effect on Precipitation in the Southern Balochistan and Sindh province. In 2022, these monsoon depressions became more frequent due to the global La Nina effect. (1)

2 Introduction

The year 2022 saw twenty-nine billion dollar worth of damage caused by climate related extreme events. There were a total of 14 severe weather events (damages done by thunderstorms, hail, and/or tornadoes), six floods, five droughts, three tropical cyclones, and one European windstorm this year just until October 2022. With December 2022 is expected to the third simultaneous La Nina winter in a row (2), the impacts of climate change are being strongly felt. For Pakistan in South Asia, 2022 was the year they experienced their second extreme flood. The first was in the year 2022. The Pakistan Floods of 2022 submerged around 10-12% of the country. Sindh, Balochistan and Khyber Pakhtunkhwa are the three regions that suffered the most damage in terms of human lives as well as area. In the Sindh and Balochistan region 1135 people were killed and 8422 injured. While in the Khyber Pakhtunkhwa region 309 people died and 600,000 others were displaced by floods. According to Climate Risk Country Profile by the World Bank, with the ongoing climate crisis, Pakistan will likely have an increased number of extreme river and coastal floods in the next few decades. Another report by the World Weather Attribution suggests that Pakistan's geographical location and landscape might be a factor as to why this particular country is facing extreme flooding.

Our motivation is to apply machine learning to this geo-climatic space to study if and which ground-level climate indicators that are heavily measured and monitored can be used as predictors or "symptoms" of extreme events. So, that governments can be better informed and prepared towards such extreme events.

3 Related Work

The Pakistan 2022 and 2010 floods have been studied intensively. Multiple scientists have collaborated to produce climate change attribution report. Pandey, 2021(5) used a classification and regression (CART) to model flood susceptibility. While Khoirunnisa, 2021 (6) used a GANN to model flood susceptibility. Khan, 2022 (7)reported a good relationship between geophysical (presence of crevasses and subsurface cavities) and meteorological (variations in temperature and precipitation) indicators as a cause of Glacial Lake Outburst Floods in Pakistan. Inspired from this work, our study aimed to narrow down these indicators that precede extreme floods and cause a change in the extent of these floods we use Causal Inference as a method to signify if a change in variable X causes a change in variable Y as a means of predicting that if X has changed by x% then Y will change by y%. For example, if snowmelt increases by 10% then by how much can we see a change in the surface run-off ? While a prediction or forecasting model can take the data and give us whats the next Y i.e. Y_1, we chose to use Causal Inference to be more specific and aid decision making.

4 Methodology

4.1 Study area

While both regions were at the forefront of the floods fury, the regions are quite different both geographically and geo-climatically.

4.1.1 Geographically

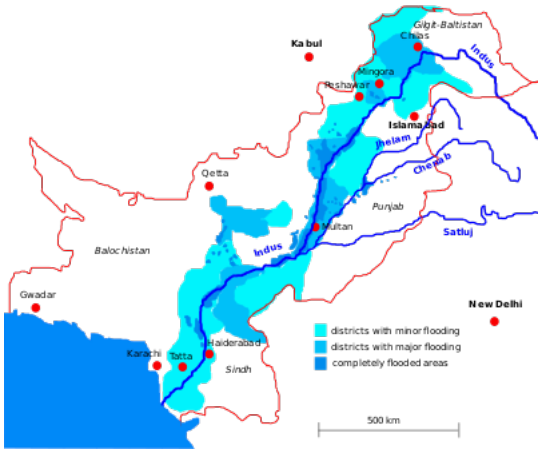
- Southern Balochistan and Sindh Region:

  This region forms the South and Mid-Western region of Pakistan. Balochistan is an plateau of rough terrain divided into basins by ranges. It is bounded by the Arabian Sea on the south and the climate is on the arid side. (8) Similarly, Sindh is also a dry region. The Indus River runs 1,000 miles vertically through this region and is the main source of water. (8)

- Khyber Pakhtunkhwa Region: This comprises of the great mountain ranges of the Karakorams and the Hindu Kush and forms the northern province of Pakistan. Over 17,870 square kilometers (6,900 square miles) of the range is covered by permanent glaciers. The steepness of the Karakorams causes some of these glaciers to advance with astonishing speed, adding to the glacial lake's volume due to snowmelt.

4.1.2 Geoclimatically

- Southern Balochistan and Sindh Region Rainfall in the Indus basin is however extremely variable from year to year, due to the strong correlation with the ENSO cycle. In the 2010 floods of Pakistan, the mid-latitudinal jet stream which flows westward from the Bay of Bengal to the Arabian Sea and then north towards Pakistan. The La Nina of 2022, exacerbated these winds due to the increased number of monsoon depressions in the Bay of Bengal caused by increase in Sea Surface Temperatures. (SSTs), according to World Weather Attribution Report. Since the year 2000, this region experienced fifty Glacial Lake Outburst Floods, with the latest one taking place on May 5th, 2022.



In 2022, both regions experienced a heatwave in the months of April and May followed by heavy precipitation which preceded the extreme floods. The two southern provinces each experienced their wettest August ever recorded, receiving 7 and 8 times their usual monthly totals. Furthermore, the Indus River turned into a 62 mile wide lake in the Sindh Area. While a combination of Glacial Lake Outburst Floods and intense rainfall caused landslides and flash floods in the Khyber Pakhtunkhwa region. World Weather Attribution Report

Since, the geographical reasons and the hydrological systems that added to the impact of the precipitation are different for both, we chose to separately study the climate indicators of both the regions. For example, the snowmelt equivalent is useless when determining causality to surface runoff in the Southern Balochistan and Sindh region as there are no snowcaps there. Similarly, bare-soil evaporation value is not the most relevant in our study with respect to the Khyber Pakhtunkhwa region.

4.2 Data

4.2.1 Collection and Processing

As our study aims to check if regular climate indicators like temperature or humidity can be used as predictors or "symptoms" of an extreme event. We chose to use surface-level data. Wind components, soil variables like soil layer temperature, wetness, evaporation from baresoil, etc and snow variables like snowfall, snow depth water equivalent, snowmelt, etc were sourced from the ERA5-Land dataset. This dataset is available on Copernicus Open Access Hub The data is monthly averaged for each coordinate in the region. The data is standardized using a MinMaxScaler from sklearn. Categorical columns were created for each selected feature variable by mean.

For the Khyber Pakhtunkhwa region, in addition to the snow variables, data on Glacial Lake Outburst Floods (GLOFs) and their occurrence date was collected from global glacial lake outburst flood database project by Institute of Environmental Science and Geography, University of Potsdam, Potsdam-Golm, Germany.

Date of Glacial Lake Outburst Flood for the region was extracted and a GLOF occur binary variable was created next to each date in the ERA-5 dataset where 1 corresponds to GLOF occurred in that month and 0 means no GLOF.

4.2.2 Temporal and Spatial range of the Data

Temporal: From the Exploratory Data Analysis we concluded that the 2000-2022 time period is good to use for regression and causal inference because it captures the decade before the first extreme floods of Pakistan in 2010 and also captures the "normal" values that lead up to the 2022 extreme Floods. This is done with the positive assumption that the regional climate variables of

Pakistan encapsulate the global climate patterns like the increase in global temperatures and the La Nina effect.

Spatial: Since the ERA5-Land data is provided for each coordinate i.e. latitude-longitude set. We decided to narrow down our study area according to specific latitude and longitude range.

For the Khyber Pakhtunkhwa region we focus on the north latitudes in the range 35.95 to 36.7 (86 miles distance)and the east longitudes 71to 74(4 miles distance) because this area captures of most of the glacial lakes where GLOFS occurred since 2000-2022.

For the Southern Balochistan and Sindh region there are five north latitudes that pass through the area starting from 25.55 North to 29.35 North. We chose to analyze the latitude 26.55East, because it captures the middle ground between the coastal 25.55 North and the inland 27.55 North. The north latitude of 29.35 showed anomalies during the Exploratory Data Analysis where each variable was plotted by latitude. The latitude 28.55 is just 69 miles south of the anomalous latitude. Hence rejected and 27.55 North is far from the coastal region and might not capture the effect of the variables there, hence that one was rejected too.

There are 7616 observations for the Khyber Pakhtunkhwa region while there are 16048 observations or samples for the Southern Balochistan and Sindh Region.

4.2.3 Selected Features for the two study areas

From the thirty-odd features, we chose to focus on the following common features(units):

Temperature (Kelvin) : Temperature of air at 2m above the surface of land, sea or in-land waters.

Dew point (Kelvin) : Temperature to which the air, at 2 meters above the surface of the Earth, would have to be cooled for saturation to occur.It is a measure of the humidity of the air.

Precipitation(meters) : Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface.

Surface Runoff (meters) : Equivalent to the amount of water that drains away over the surface after

- Specific to the Khyber Pakhtunkhwa region:

Snowmelt (meters of water equivalent) : The units given measure the depth the water would have if the snow melted and was spread evenly over the grid box.

Glacial Lake Outburst Flood occurrence : Boolean variable created next to each date in the ERA-5 dataset where 1 corresponds to truth value that GLOF occurred in that month and 0 means no GLOF occurred in that month.

- Specific to the Southern Balochistan and Sindh region:

Evaporation from bare soil (meters of water equivalent) : The amount of evaporation from bare soil at the top of the land surface. Is an indicator of the ground water table of the soil and is used with the assumption that it indicates the aridity of soil.

Wind-direction : This is a boolean variable created by checking for the condition if the u-component (Eastward component of the wind) is negative and the v-component (Northward component of the wind) is positive i.e. is the wind flowing from East to West and is the wind blowing from South to North. The unit of both the components if the horizontal speed measured in meter / second. This variable is used with the assumption that the direction of the wind is an indicator of the presence of the mid-latitudinal jet stream's effect.

4.3 Models

As we've established before, the motivation for this research was to study if and which ground-level climate indicators that are heavily measured and monitored can be used as predictors or "symptoms" of extreme floods. To do so, it is important to study the relationship between these indicators including the flood variables. In our study we use Causal Inference as a method to signify if a change in variable X causes a change in variable Y as a means of predicting that if X has changed by x% then Y will change by y%. For example, if snowmelt increases by 10% then by how much can we see a change in the surface run-off ? While a prediction or forecasting model can take the data and give us whats the next Y i.e. Y_1, we chose to use Causal Inference to be more specific and aid decision making.

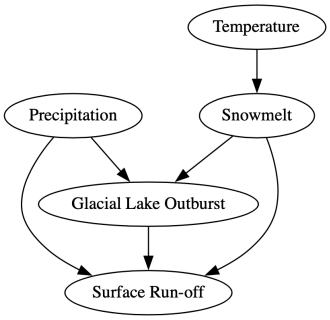
Our study does so in four steps:

- Creating a Directed Acyclic Graph (DAG) from the initial correlations between the variables found through Exploratory Data Analysis
- Performing Mixed Effect Multi-linear Regression with the DAG outcome as the 'dependent' variable and a random effects dummy variable
- Performing Granger Causality on a few of the edges in the DAG to see if Causal Discovery also supports our assumed causal model
- Performing Causal Inference using Microsoft DoWhy package to test our hypothesis of the edges and see if what is found in step 2 and 3 hold true and as a test of our assumptions in the DAG

4.3.1 Creating Direct Acyclic Graph

Runoff is a measure of the availability of water in the soil, and can, for example, be used as an indicator of drought or flood. So, in our model for both regions we chose Surface Runoff as the Outcome variable. Precipitation was not used as it is established that there was high precipitation that caused the floods. World Weather Attribution Report

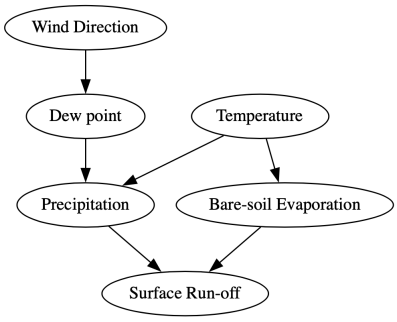
Khyber Pakhtunkhwa region



For this region, the treatment is "Glacial Lake Outburst" The confounders are "Precipitation" and "Snowmelt" which are also the 'common cause' of the outcome and the treatment. Since,

"Temperature" impacts "Snowmelt", in our model we define "Temperature" as the Causal Grandparent.

Southern Balochistan and Sindh region



For this region, the treatments are "Evaporation from Bare Soil" and "Precipitation" The Causal Grandparents are Dew Point which is in turn impacted by the Wind Direction. Here "Temperature" is the common cause of "Evaporation from Bare Soil" and "Precipitation"

4.3.2 Mixed Effect Multi linear regression

Linear Regression on a single dependent variable can be used as the preceding step to inferring causality. Linear Mixed Effects models are used for regression analyses involving dependent data when more than one observation is made for each data point. In our case we implemented the Mixed Effect Multi Linear Regression using the statsmodels.formula.api python packages's mixedlm model. This model gives the result in terms of correlation coefficient for each Interventions and the associated p-value. The expected outcome to reject the null hypothesis is that the p-value should be below 0.05 For both of out study areas: Our dependent variable is the outcome i.e. surface runoff, while every other node in the DAG is the intercept. To account for the random effects, we groupby "Longitude" to see if there is any significant change due to change in Longitude.

$$runoff - GLOFoccur + precipitation + snowmelt + temp$$

$$runoff - evapfrombaresoil + precipitation + temp + dewpoint + wind_direction$$

Here we run the model on both numerical version of the variables as well as categorical version of the variables. The formula for our model Mixed Effect Multi Linear Regression is:

$$y = b0 + b1X1 + b2X2 + b3X3 + e$$

The coefficient here b0, b1, b2, b3 can be considered as the estimate of causality but we can only infer correlation and cannot define causality without performing our Step 4 i.e. Causal Inference.

4.3.3 Granger Causality for Causal Discovery

Causal Discovery is a method to get the causal structure from the data between two time-series such as in our case. The result of a Granger Causality does not imply actual causal effect but implies the predictability of variable Y by variable X. It results in a p-value and a F-test statistic that can be used to accept or reject a null hypothesis. The following were our hypothesis for this step:

- The amount of snowmelt is not predictive of the future surface runoff
- The amount of snowmelt is not predictive of the occurrence of a Glacial Lake Outburst Flood.

- The amount of Bare-soil evaporation does not 'Granger-cause' or is predictive of Surface run-off
- The wind-direction does not 'Granger-cause' or is not predictive of Precipitation

4.3.4 Causal Inference - Causal graph

We used Microsofts DoWhy Library to perform estimation of causal effect or Causal Inference. We used the categorical versions of our features and used two methods: propensity score matching and linear regression with a "Backdoor" criterion. We then performed Refute Tests on the estimate using the same package and with three methods:

- Random Common Cause: If assumption is true, then no change in estimate
- Data Subset Refuter: If assumption is true, then no change in estimate
- Placebo Effect: If assumption is true, then no change in estimate

Our hypothesis during the Causal Inference were either accepted or rejected after the three refute tests on the causal effect estimate

5 Results

5.0.1 Khyber Pakhtunkhwa region

In this region, we found that Snowmelt which we had considered to be a confounder has a stronger causal effect as "treatment" on the Surface runoff. Even though Granger Causality rejected the predictability of surface runoff from snowmelt. Causal Inference using propensity score matching method defended the causal effect which is estimated to be higher than the estimate found from the linear regression thereby also defending the linear regression. We also see that GLOFs have a 1.8% positive causal effect on surface runoff.

Region: Khyber Pakhtunkhwa		**Methods**			
Outcome	**Treatment**	**Mixed Effect Linear Regression**	**Granger Causality**	**Causal Inference (Do Why)**	
		Estimate	**Accepted/Rejected**	**Estimate**	**Refuted/Defended**
Surface Runoff	Snowmelt	**0.799** (SRO increases by 79.9% when Snowmelt increases by 1 unit)	**Rejected** i.e., Causality or Predictivity rejected	**0.814** (SRO increases by 81.4% when Snowmelt increases by 1 unit)	**Defended**
Surface Runoff	Glacial Lake Outburst Flood Occurrence	**0.098** (SRO increases by 9.8% when GLOF Occurrence increases by 1 unit)	**Rejected** i.e., Causality or Predictivity rejected	**0.018** (SRO increases by 1.8% when GLOF Occurrence increases by 1 unit)	**Defended**
Glacial Lake Outburst Flood Occurrence	Snowmelt	**0.644** (GLOF Occurrence increases by 64.4% when Snowmelt increases by 1 unit)	Not performed but this gives an idea that Snowmelt is the stronger Confounder in the DAG – effecting both Surface Runoff and GLOF.		
Glacial Lake Outburst Flood Occurrence	Precipitation	**-0.017** (GLOF Occurrence decreases by 1.7% when Precipitation increases by 1 unit)	Not performed as mixedMLM estimate of -0.017 shows association but does not explain how GLOF can decrease with increase in Precipitation even though rainfall should add to any lake's volume. This does not follow the laws of ground truth physics.		

It was interesting to note that precipitation only had an 18.6% causal effect estimate from the mixed effect multi-linear regression. This value is less than that of Snowmelt(79.9%). Further Causal Inference on this association was not performed since in case of extreme events like the 2022 Floods, Precipitation and Surface Runoff are expected to have a cyclic correlation.

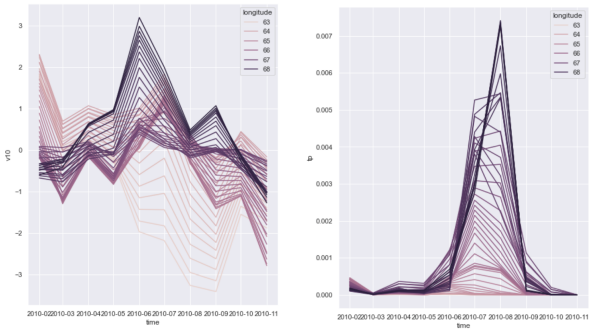
The results from this region show that due to the orography in this province, snowmelt is a greater cause of the floods experienced in this region. It indicates that snowmelt data collected from the time of the heatwave can be used to predict the occurrence of surface runoff(floods), so officials can be better prepared for damage control.

5.0.2 Southern Balochistan and Sindh region

Region: Southern Baluchistan and Sindh		**Methods**			
Outcome	**Treatment**	**Mixed Effect Linear Regression**	**Granger Causality**	**Causal Inference (Do Why)**	
		Estimate	**Accepted/Rejected**	**Estimate**	**Refuted/Defended**
Surface Runoff	Evaporation from Bare Soil	**- 0.100** (SRO decreases by 10. 0% when Evaporation f rom Bare Soil increases by 1 unit)	**Rejected** i.e., Causality or Predictivity rejected	**- 0.14** (SRO decreases by 14.0% when Evaporation from Bare Soil increases by 1 unit)	**Defended** (Defended the causal effect i.e., defended the causality)
Surface Runoff	Precipitation	**0.229** (SRO increases by 22.9% when Precipitation increases by 1 unit)	**Rejected** i.e., Causality or Predictivity rejected	**0.210** (SRO increases by 21.0% when Precipitation increases by 1 unit)	**Defended** (Defended the causal effect i.e., defended the causality)
Precipitation	Wind Direction (Indicative of the winds traveled from the East i.e., due to the monsoon depressions in Bay of Bengal)	**0.190** (Precipitation increases by 19.0% when Wind-direction (i.e., Heavy winds from Bay of Bengal reaching Pakistan) increases by 1 unit)	**Accepted!** (Granger Causality accepted that these Northward winds coming from the East that carry moisture **can be used to predict** Precipitation)	**0.237** (Precipitation increases by 23.7% when Wind-direction (i.e., Heavy winds from Bay of Bengal reaching Pakistan) increases by 1 unit)	**Defended** (Defended the causal effect i.e., defended the causality)

Here, we see that evaporation from bare soil has a causal effect on surface runoff(flooding) by only 10%. The main motivation to test for Evaporation from Bare Soil was to understand if an increase in the aridity of the region will cause an increase in the extent or power of the surface runoff(flooding). Our results do lean towards this indication and perhaps this variable can be used to not only predict the extent of surface runoff but also guide the official s in which area of their geography can be improved by human interventions and land use.

The surprise result is the confirmation of wind-direction which in our case was an indicator of the presence of the moisture heavy winds coming from the monsoon depressions in the Bay of Bengal (refer to the diagram). The causal effect of these winds on precipitation was accepted by all the three methods. Causal Inference by Do Why showed a higher causal effect of 23%. We saw the above relationships in the 2010 and 2022 dataset during our Exploratory Data Analysis and these were confirmed by the Causal Analysis.



On the left: Northward wind component in the Southern Balochistan Region in 2022. On the right: Precipitation in the Southern Balochistan Region in 2022. We can see that Northward wind preceded in time Precipitation

6 Discussion

- Overall the results of the four step causal analysis have been promising and results show which indicators can be used as "predictors" of extreme floods or at least which indicator's causal effects can be studied further.
- Many assumptions were made during the design of the model and localization of the study areas for this research. In the future, having a domain expert like a geo-climatic scientist who studies a specific region in question can be really helpful in doing away any doubts or uncertainties while making the assumptions.
- Surprisingly Mixed Effect Multilinear Regression performed quite well in estimating causal relationships which were confirmed by the implementation of the Do Why operator.
- We found that Causal Discovery did not yield much concrete results. Perhaps a white-box implementation of Causal Discovery using PC and Lingam models followed by implementation of sensitivity analysis for multicollinearity and SCM for Causal Inference can be a much more robust models.

7 Acknowledgements

I thank Professor Alp Kucukelbir and the course assistant Dafne Papaefthymiou for their expertise and feedback throughout all aspects of our study. I acknowledge that the author has no Conflict of Interest.

References

[1] "Salient features of monsoon 2022," *IMD.gov*, Oct 2022. [Online]. Available: https://internal.imd.gov.in/press_release/20221001_pr_1849.pdf

[2] "October 2022 la niña update: snack size," *Climate.gov*, Oct 2022. [Online]. Available: https://www.climate.gov/news-features/blogs/october-2022-la-ni~na-update-snack-size

[3] "Climate risk country profile: Pakistan (2021)," *World Bank*. [Online]. Available: https://climateknowledgeportal.worldbank.org/sites/default/files/2021-05/15078-WB_Pakistan%20Country%20Profile-WEB.pdf

[4] "Climate change likely increased extreme monsoon rainfall, flooding highly vulnerable communities in pakistan," *World Weather Attribu-*

tion, Sep 2022. [Online]. Available: https://www.worldweatherattribution.org/climate-change-likely-increased-extreme-monsoon-rainfall-flooding-highly-vulnerable-communities-in-pakistan/

[5] M. Pandey, A. Arora, A. Arabameri, R. Costache, N. Kumar, V. N. Mishra, H. Nguyen, J. Mishra, M. A. Siddiqui, Y. Ray *et al.*, "Flood susceptibility modeling in a subtropical humid low-relief alluvial plain environment: application of novel ensemble machine learning approach," *Frontiers in Earth Science*, p. 1091, 2021.

[6] N. Khoirunisa, C.-Y. Ku, and C.-Y. Liu, "A gis-based artificial neural network model for flood susceptibility assessment," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, p. 1072, 2021.

[7] S. Khan, U. B. Nisar, A. Hussain, N. Ahmad, and B. Siddique, "Hydrological investigation of subsurface glacial lake outburst floods at bindo gol valley under changing climate, pakistan," *Journal of Applied Geophysics*, vol. 201, p. 104633, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926985122001045

[8] Wikipedia contributors, "Balochistan, pakistan — Wikipedia, the free encyclopedia," 2022, [Online; accessed 16-December-2022]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Balochistan,_Pakistan&oldid=1126293852