# Climate change and its impact on landslides in High Mountain Asia region of Nepal and India

## 1.Introduction

The High Mountain Asia region is particularly susceptible to landslides due to its steep terrain, closeness to glaciers and the intense monsoon season (Jessica Merzdorf, 2022). In 2020 NASA's Goddard Space Flight Centre and National Oceanic and Atmospheric Administration published a paper suggesting that as the effects of climate change increase this will lead to an increase in landslides in High Mountain Asia (Kirschbaum, Kapnick, Stanley and Pascale, 2020). In this study, alternative methods will be employed to determine if this hypothesis can be corroborated using a reduced sample of landslide data from the Cooperative Open Online Landslide Repository (COOLR) covering an area including Jammu and Kashmir in India and finishing at the border of the Sikkim province with Bhutan. Spatial-temporal scan statistics permutation modelling (STPSS) will be used to detect cluster locations and estimate their statistical significance (Tonini and Cama, 2019). Autoregressive integrated moving average modelling (ARIMA) will be used to describe autocorrelations in the existing landslide data to forecast the future landslides. In addition, Random Forest analysis will be employed to construct a series of independent decision trees which when averaged will give us a prediction of the causes of landslides.

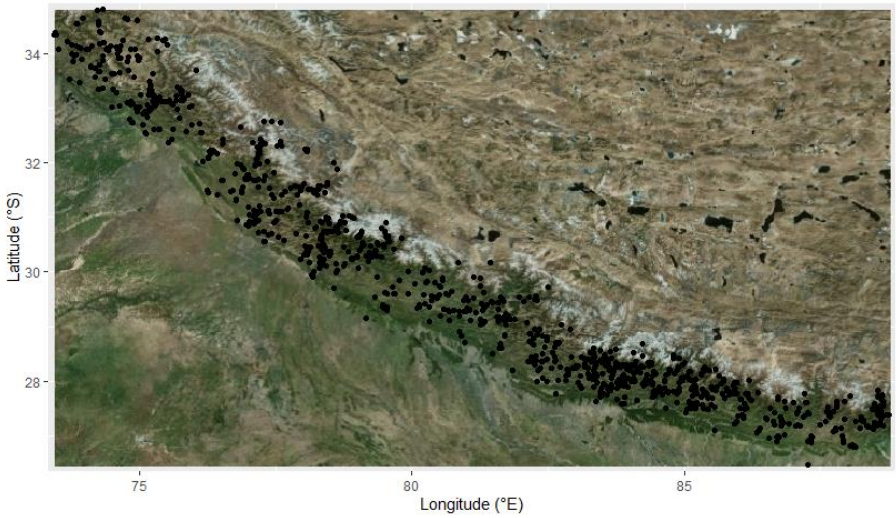



*Figure 1.1 Map of study area with landslide points plotted*

## 2.Exploratory Spatio-temporal data analysis

### 2.1 Point Pattern Analysis

#### 2.1.1 Visual Analysis

When plotted on a map the landslides appeared to show clustering. The most visually significant appears at the longitude of 75 and a latitude of 33(NW Cluster) and at a longitude of 84 and a latitude of 28(SE Cluster).

#### 2.1.2 Tests for Complete Spatial Randomness (CSR)

To understand if the landslides are spatially random, they were tested against complete spatial randomness. Performing the chi squared test based on the landslide's points with a 10 by 10 grid gave a p value of 2.2e-16, which suggests the data does not follow the characteristics of a homogeneous Poisson process. In addition, the Spatial and Density Kolmogorov-Smirnov tests, which compare the expected (CSR) to the observed dataset, both have significant p-values suggesting the data was not drawn from a distribution with CSR.

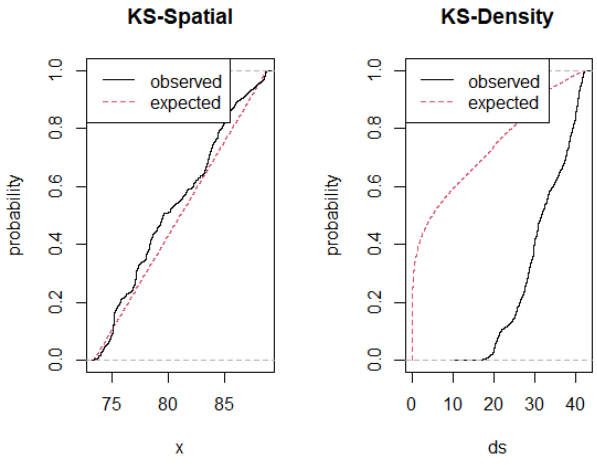



*Figure 2.1.2.1 a Kolmogorov Smirnov test on the x coordinates of the data. B. Kolmogorov Smirnov test on the density of the data.*

## 2.1.3 Spatial Dependence tests

In addition, the output of Ripley's K function presented values that lie above the line of a Poisson distribution, suggesting clustering in the dataset. The G and F function which measure distance of each point to neighbour and distance from point to empty space, also suggested the landslides are not spatially dependent.

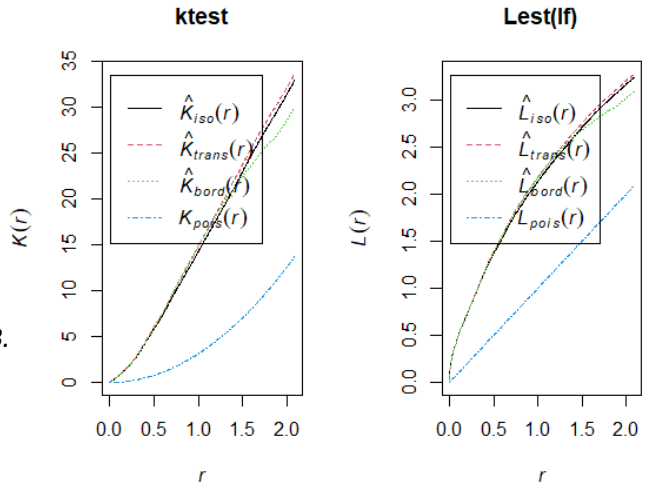



*Figure 2.1.3.1 A. K function graph comparing the distance between the observed landslides to a Poisson distribution. B. Plotting the K function on a straight line.*

## 2.1.4 Kernel Density Estimation

To measure autocorrelation in the landslide locations the *spatstat* package allowed density plotting and confirmed the visual analysis (figure 1.1). The SE Cluster showed a higher number of points over a wide area, whereas the NW Cluster presented more dense clustering. In addition, there was significantly less clustering of landslides in the region between 80 and 84 longitudes.

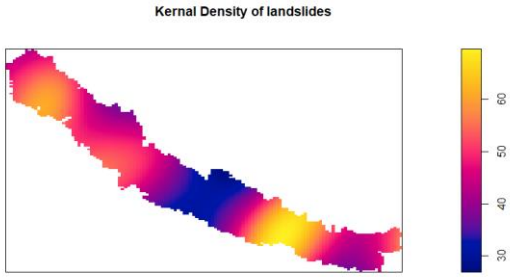



*Figure 2.1.4.1 Kernel density of landslide data*

The landslide data also includes the cause of the landslide; this was combined with the density plots to show clustering for each cause of landslide. From this, whilst monsoons are concentrated in the SE, it appears that landslides caused by other types of rainfall events are more distributed across the whole area. Landslides from all other causes are very localised.

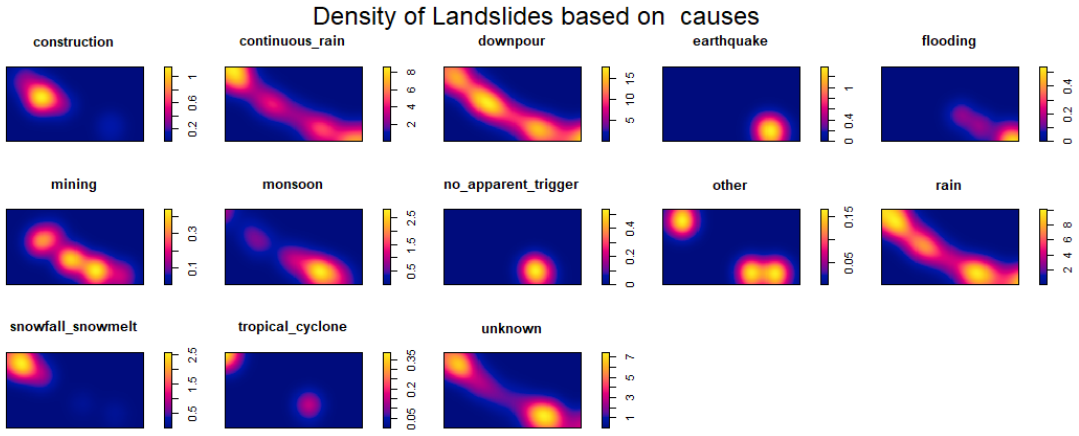



*Figure 2.1.4.2 Kernel Density plots for each cause of landslide with yellow corresponding to high density and purple low density*

## 2.2 Temporal Analysis

The original dataset contained the exact date of the landslide but for ease of analysis of temporal characteristics, the landslides were aggregated into the month and year of occurrence. The years before 2007 and after 2018 were removed due to data not being complete for those years.

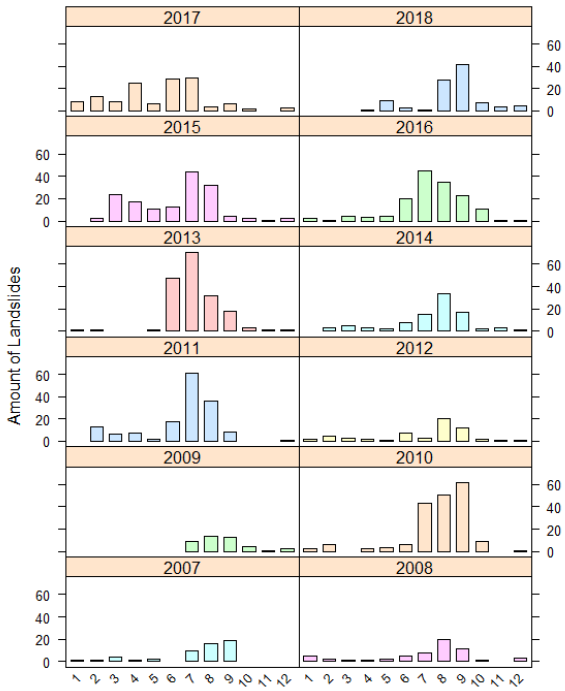



*Figure 2.2.1 Count of landslides for every moth of every year from 2007 to-2018*

Inspecting the landslide count through time showed a strong seasonality in the dataset with landslide count increasing to June – September then dropping off to the end of the year. This will need to be accounted for when modelling. In addition,2010 and 2013 have large landslides in their monsoon seasons compared to all the other years. In 2010 there was a peak in landslides, potentially due to the impact of El Nino on the monsoon season that year (NASA - NASA's TRMM Satellite Confirms 2010 Landslides, 2022). Furthermore in 2007 and 2008 very few landslides occurred outside of the monsoon window (June to September). However, 2015,2016 and 2018 show more landslides occurring prior to June.

Plotting the lagged variables of the landslide count gives a value of autocorrelation coefficient of 0.579 suggesting landslides in previous months are strongly correlated and suggest further modelling of the dataset can be achieved.

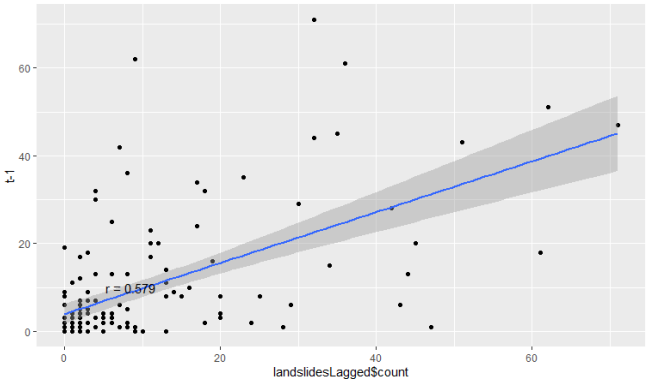



*Figure 2.2.2 Scatter plot with count of landslide per month on the x axis and the lagged landslide count on the y axis.*

## 3.Methodology and Results

### 3.1 STPSS Modelling: Understanding landslide cluster over space and time

STPSS Modelling is based on creating cylindrical moving windows which scan the study area across space and time counting the observed and expected occurrences (based on a Poisson distribution) of an event and compute a Likelihood ratio. The significance is then evaluated by Monte Carlo hypothesis testing and given a p-value. For this dataset it allowed observations in the clustering of landslides over space and time, to investigate if there has been an increase in clustering in a region or longer lasting clusters, which would suggest an overall increase in the activity of landslides in the past, a trend that could be expected to increase in the future.

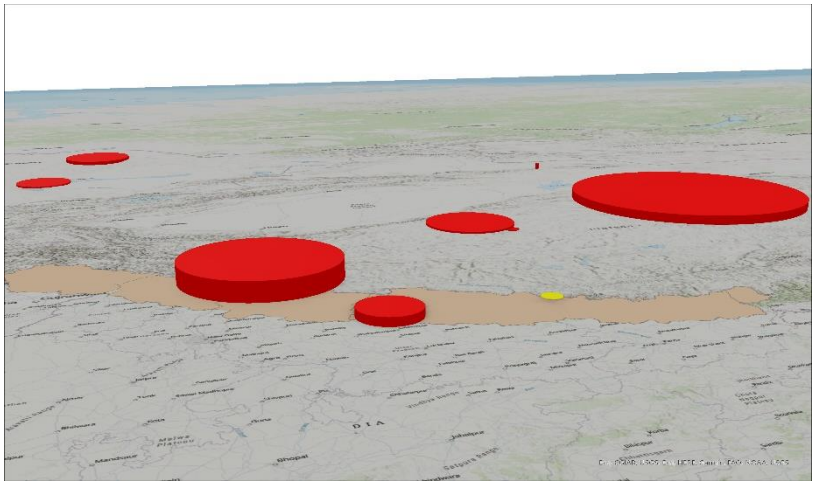
To run the SaTScan model requires 3 files: a case file containing the timing of each landslide, a coordinates file which contains the cartesian coordinates for each landslide and a session file which contains all the parameters to analyse the data.

These were created in R using the *rsatscan* package. This package enables files to be written in the correct format to be run in SaTScan. The case file contained the event date for each landslide and was formatted into year/month/day with a count added to signify each date a landslide occurred. The coordinates file contained the longitude and latitude columns which contained the location of each landslide in the WGS 84 Coordinate System.

The session file contained the start date, which was the date of the first landslide, and the end date was set as the date of the last landslide. Retrospective space-time analysis was chosen for the type of analysis and the probability model used was space-time permutation. High rates were chosen to focus on where clusters of landslides are occurring, and time aggregation was set to month to look at patterns at the monthly scale.

The clusters were presented in ARCGIS Pro where the output of the SaTScan was converted into buffers defined by the radius of the clusters and projected in 3D with the length of the cylinder defining the length in time.

Figure 3.1.1 SaTScan Model of Clusters in ArcGIS Pro. Cylinders in red have a p-value of less than 0.05.



The Longest Cluster of the dataset starts in July 2010 and continues till end of August 2013, corresponding with the 2010 El Nino event impacts (NASA - NASA's TRMM Satellite Confirms 2010 Landslides, 2012) and is focussed on the mouth of the Ganges. Since this event there have not been any similar long clusters. However there appears to be a trend of large radius significant clusters from 2013 to 2018 which only last for approximately a year but are not neighbouring each other, suggesting a more global factor affecting the landslide clustering from 2013 onwards.

## 3.2 ARIMA Modelling: Modelling Landslides using previous years

Arima modelling is a linear model composed of 3 statistical components: an autoregressive component which uses past values to forecast future values, an integrated component which allows incorporation of differencing due to seasonal or periodical time series, and a moving average component which uses past forecasting errors to forecast future values. For this dataset it allowed the analysis of the temporal pattern of the landslides in the past and to project this into the future to understand how effectively landslides can be predicted.

To achieve ARIMA modelling the landslide dataset had to be split into subsets containing a count of how many landslides occurred in each month. These were then merged to give a single data frame with a count which contained the number of landslides for each month from January 2007 to December 2018.

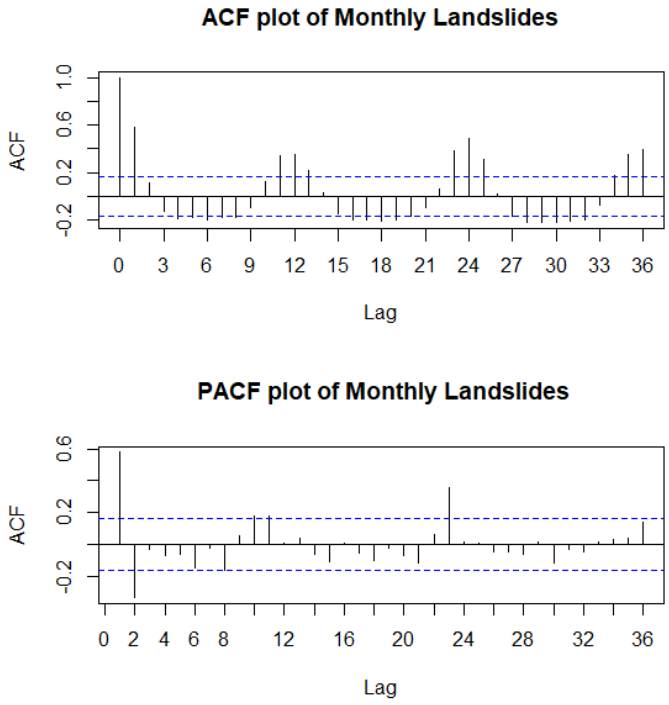
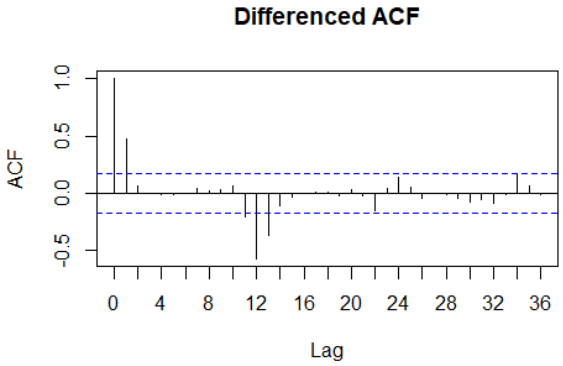



*Figure 3.2.1 A. plot of autocorrelation of the landslides plot to 36 lags B. plot of partial autocorrelation of the landslide count to 35 lags*

### 3.2.1 Identifying order of differencing

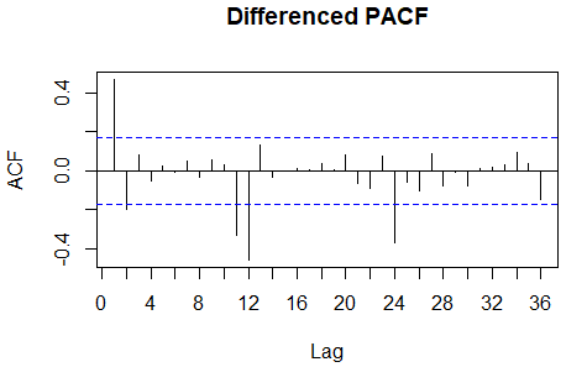
Applying differencing was the first step introduced to make the dataset more stationary. Observing the ACF presented a strong seasonal pattern in the autocorrelation, potentially due to changes in rainfall throughout the year. In Figure 3.2.1.1 applying a seasonal differencing of 12 seemed to remove most of the seasonal components.




### 3.2.2 Finding the Moving average terms

*Figure 3.2.2.1 Differenced ACF plot*

The observed differenced ACF showed significant autocorrelation at lags 1,11,12,13. This suggested an ARIMA model of (0,0,3) (0,1,1)12 model due to 3 nonseasonal MA terms and 1 seasonal MA term with a single difference at lag 12.



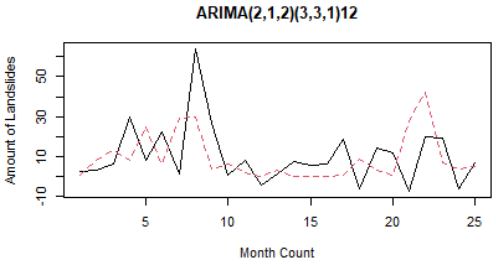

### 3.2.3 Find the AR

*Figure 3.2.3.1 Differenced PACF plot*

To determine the autoregressive terms the differenced PACF for the Landslides was observed showing significant lags at 1,2,11,12 and 24. This suggested an ARIMA model of (2,0,3) (2,1,1)12 due to 2 nonseasonal PACF values and 2 seasonal PACF values.

### 3.2.4 Fitting the ARIMA Model

To understand the significance of the observed ARIMA model it was compared to an Automated Arima model (2,0,0) (0,1,0) and a model that minimised the AIC (2,1,2) (3,3,1)12. Each ARIMA model was trained on the landslides counts from 2007 to 2016 and then validated using the final 2 years to test its predictive ability.



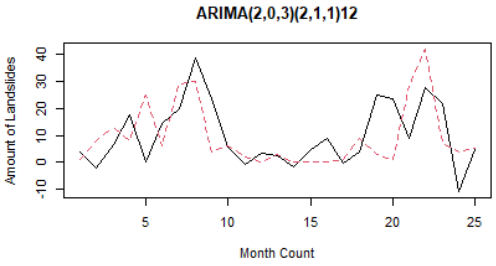




*Figure 3.2.4.1 Plots of ARIMA models the red dashed line is the observed landslide count and the solid black line is the prediction by the ARIMA model*

Visual inspection of the auto Arima model appeared to be less significant in predicting the landslides of 2017 and 2018. Comparing our AIC model against our observed model they both correctly predicted the peak in the summer of 2017 and 2018 however with different degrees of magnitude. Using evaluating statistics the AIC model appeared to show the lowest AIC suggesting the model fits better. However, the NRMSE and P-value from the Ljung-Box Piece are higher for the observed ARIMA modelled which suggested the model performance is better and that the residuals are more accounted for.

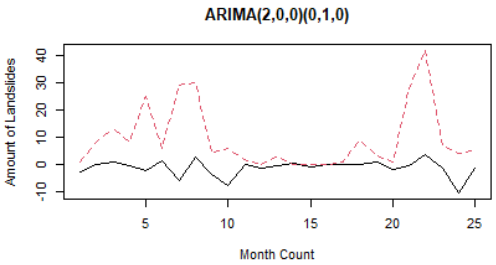



*Figure 3.2.4.2 Table of tests to evaluate performance and fit of Arima models*

| | AIC Minimised ARIMA | Observed ARIMA | Auto Arima |
|---|---|---|---|
| AIC | 960.7832 | 991.4442 | 1231.7147 |
| NRMSE | 0.899 | 0.635 | 1.192 |
| Box Pierce P-value | 0.95 | 0.958 | 2.22e-16 |

3.3 Random Forest: Can we understand the factors that cause landslides?

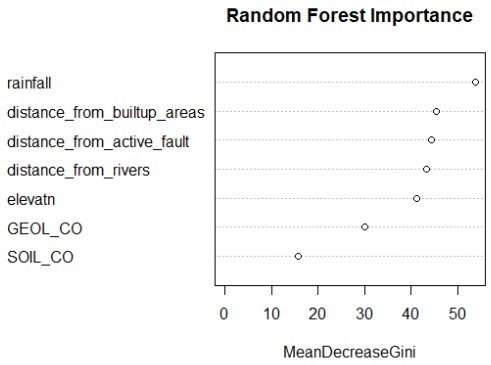
Random Forest is a nonlinear ensemble model which creates a series of decision trees independent from each other. This is achieved due to sampling with replacement and each split in the trees is made with random chosen features. These trees are then averaged to obtain the result. For the landslide data a Random Forest was used to classify the causes of landslides based on geological, geomorphic, human and climate factors. In addition, it allowed the plotting of an importance of each variable to observe which are the most significant in classifying landslides.

The original landslide dataset contained little information about conditions when landslides occurred. Therefore, for our model geological, geomorphic, climate and human factors such as geology, distance from closest river, rainfall that occurred in the month of the landslide aggregated from development regions and the distance from built up areas. The study area for the Random Forest area was reduced to just Nepal as more data was available for this area.

The data was trained on 80% of the samples and validated on the remaining 20%. The number of trees was set to 500 with each branching randomly selecting 3 features. These parameters were chosen as they gave us the lowest validation error. The result gives an error value of 25%. In addition, the F1 scores suggested many of the causes were not present in our validation such as mining and construction which caused a 100% error in prediction.

*Figure 3.3.1 Ranked Variable importance in classifying the landslide cause by the decrease in mean Gini*

Inspecting our variable importance, rainfall appeared to be the most important factor for classifying the different landslide causes. Distance from the built-up areas also seems an important variable. However, this data is questionable since the COOLR reporting is a combination between citizen reporting, which is more likely in built up areas, and satellite observation. Another interesting feature was that Geological and Geomorphic factors have less of an impact on the classification of landslides, suggesting preparatory factors such as climate and human factors have a stronger influence over causing landslides.




## 4.Discussion and Conclusion

This report has examined 3 different models to understand how climate change is affecting landslides in High Mountains Asia. Applying STPSS has enabled us to identify time frames with greater landslide activity. The longest cluster can be clearly linked to the impacts of El Nino in the region suggesting that worsening El Nino effects could correlate to increasing amounts of landslides. Limitations of this STSS analysis are that the geography is not incorporated and cylinders which are spreading outside of our observation area are suggesting clustering in areas of no observation.

ARIMA modelling has indicated the difficulty of predicting landslides in the future due to their high variance in time but that they correlate with a strong seasonality to the monsoon seasons. However, due to ARIMA's complexity compared to other soothing models, it makes it more difficult to explain trends as easily.

Random forest modelling has given an insight into the factors that affect landslide causes. However, the error margin is still very high and the accuracy so low that it cannot give any significant projections on how landslides might be affected by increasing climate change in the future.

In conclusion, STSS analysis using SaTScan is the strongest model in helping understand climate change's effect on landslides due to it providing us with actionable data that is easily understandable. For future work an increasing window of time is needed to reliably make predictions as major climate events such as El Nino only occur every 3-7 years with our landslide data only containing 1 such event. Furthermore, for a more reliable random forest more data is need for each cause and the attributes should be more spatially and temporally accurate to achieve more reliable results.

## References for Report

*Jessica Merzdorf, N., 2022. Climate Change Could Trigger More Landslides in High Mountain Asia – Climate Change: Vital Signs of the Planet. [online] Climate Change: Vital Signs of the Planet. Available at: <https://climate.nasa.gov/news/2951/climate-change-could-trigger-more-landslides-in-high-mountain-asia/> [Accessed 10 May 2022].*

*Kirschbaum, D., Kapnick, S., Stanley, T. and Pascale, S., 2020. Changes in Extreme Precipitation and Landslides Over High Mountain Asia. Geophysical Research Letters, 47(4).*

*Tonini, M. and Cama, M., 2019. Spatio-temporal pattern distribution of landslides causing damage in Switzerland. Landslides, 16(11), pp.2103-2113.*

*Climate Change: Vital Signs of the Planet. 2012. NASA's TRMM satellite confirms 2010 landslides – Climate Change: Vital Signs of the Planet. [online] Available at: <https://climate.nasa.gov/news/818/nasas-trmm-satellite-confirms-2010-landslides/> [Accessed 10 May 2022].*

## References for data

**Built up areas, Rivers and Nepal shapefiles data source:**

*2022. The Humanitarian Data Exchange. [online] Available at: <https://data.humdata.org/> [Accessed 10 May 2022].*

**Map Background for figure 1.1:**

*Bing Maps. 2022. Bing Maps. [online] Available at: <https://www.bing.com/maps> [Accessed 10 May 2022].*

**Rainfall data source:**

*Climateknowledgeportal.worldbank.org. 2022. World Bank Climate Change Knowledge Portal. [online] Available at: <https://climateknowledgeportal.worldbank.org/download-data> [Accessed 10 May 2022].*

**Landslide data sources:**

*Juang, C., Stanley, T. and Kirschbaum, D., 2019. Using citizen science to expand the global map of landslides: Introducing the Cooperative Open Online Landslide Repository (COOLR). PLOS ONE, 14(7), p.e0218657.*

*Kirschbaum, D., Adler, R., Hong, Y., Hill, S. and Lerner-Lam, A., 2009. A global landslide catalog for hazard applications: method, results, and limitations. Natural Hazards, 52(3), pp.561-575.*

*Kirschbaum, D., Stanley, T. and Zhou, Y., 2015. Spatial and temporal analysis of a global landslide catalog. Geomorphology, 249, pp.4-15.*

**Development Regions of Nepal, Geoloy and Soil of Nepal data source:**

*Rds.icimod.org. 2022. ICIMOD | RDS. [online] Available at: <https://rds.icimod.org/> [Accessed 10 May 2022].*

**Elevation data obtained from:**

*Registry of Open Data on AWS. 2022. Terrain Tiles. [online] Available at: <https://registry.opendata.aws/terrain-tiles.> [Accessed 10 May 2022].*

**Shapefile for Indian state boundaries obtained from:**

*Projects.datameet.org. 2022. States - Community Created Maps of India. [online] Available at: <http://projects.datameet.org/maps/states/> [Accessed 10 May 2022].*

**Active Faults data source:**

*Styron, R. and Pagani, M., 2020. The GEM Global Active Faults Database. Earthquake Spectra, 36(1_suppl), pp.160-180.*