Modern Data Analytics

[G0Z39a]

Coordinator: Prof. Jan De Spiegeleer

WATER SECURITY

May 2021

Authors:

Alonso, Pedro Leite (r0773505)
Manikantan, Venkatesh Viswanathan (r0825919)
Hashim , Hani Mustafa (r0827025)
Tseng, Mitchikou Pearl (r0821342)
Valerio, Thea (r0821343)
Vandermeersch, Lili (r0691855)

Introduction

Water crises are believed to be the most impactful global risks within the next decade according to the World Economic Forum. In fact, US Intelligence Community Assessment of Global Water Security considers that water issues will contribute to social disruptions that can result in state failure when combined with poverty, social tensions, environmental degradation, ineffectual leadership, and weak political institutions.

Considering water sources and risks vary by region, it is therefore crucial to take into account country-specific conditions and profiles in finding ways to manage and protect water resources. After reviewing several data sources, the following were obtained and collated based on data quality and extensiveness: (1) global population and GDP scenarios from SSP to account for the socio-economic aspect of the analysis, (2) a comprehensive global hydrological model describing water flows and storage (and consequently water resources on land areas) from WaterGAP, (3) global land surface variables from NASA, (4) water poverty index (WPI) in 2002 which assesses water stress and scarcity, and (5) city-specific water security risks and actions in 2020 from CDP. Figure 1 illustrates the research pipeline from data retrieval to processing, and finally to analysis. Further data and processing details are described in the Appendix.

In this analysis, we use unsupervised and supervised learning methods to: 1) explore the data to group countries that find themselves under similar water forces, 2) use past data to predict a water stress score for countries in the future.

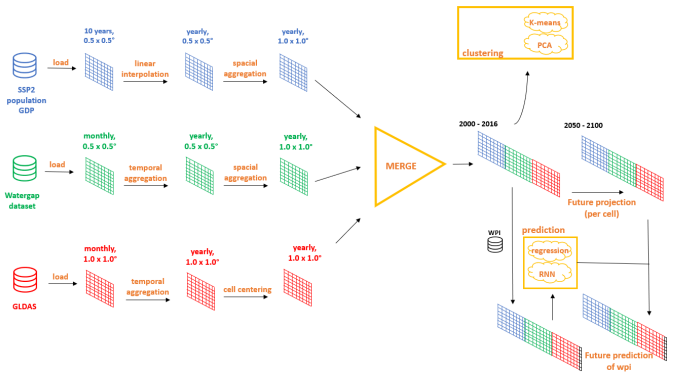


Figure 1. Data Pipeline.

1: Exploring Global Water Characteristics using K-means Clustering

K-means clustering was performed on the entire dataset of all the three models, and the result showcases a pattern where the clusters are able to identify areas of desert, huge river basins such as the Amazon and the Yellow River, followed by capturing the extreme weather condition in Greenland. This is further illustrated in the images provided in the Appendix where K-means was performed with individual countries and a pattern emerges which closely resembles the historic geographical/weather patterns, especially when clustering the U.S.A. and China. As the clusters closely resemble patterns of water availability in a real world scenario, it further gives gravity to the data collected to analyze the water situation around the planet.

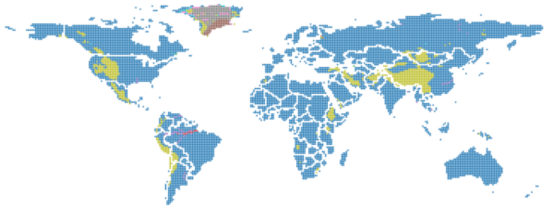


Figure 2: K-means (10-C) with SSP2,GLDAS,WaterGAP Data (years 2000-2016)

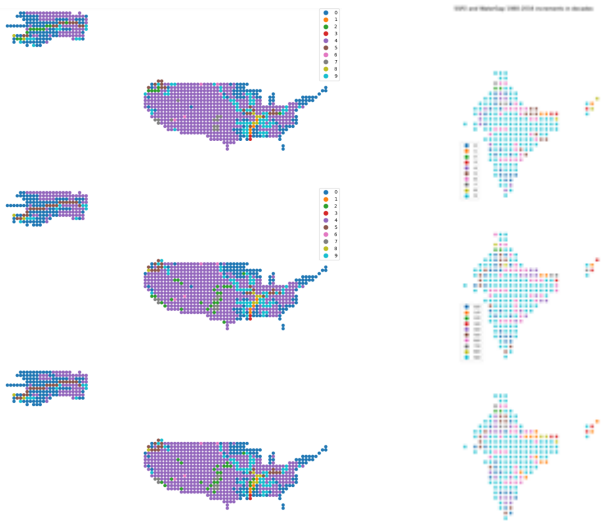


Figure 3: K-means(10-C) USA & India (SSP2 and WaterGap) each image showcasing a decade from 1980-1990,1990-2000 and 2000-2016 in that order from top to bottom.

Figure 2 showcases clusters formed using the data of only WaterGap and SSP. The GLDAS (NASA) data has not been included as the inclusion of the weather data overpowers the WaterGap data as it causes the clusters to lose the ability to identify water bodies when different climatic conditions are included. Figure 2 includes years from 1980-2016, and it can be observed in the case of the U.S.A. that the k-means clusters are able to identify the large water body such as the Mississippi river which splits vertically across the country and also the lake Michigan in the northern part of the country.

By splitting the data into three decades we were able to identify a new cluster forming in the middle of the country. As each decade passes, we can notice that the green patch is growing through this time period of 1990 to 2016.

Furthermore, the green patch showcases areas where the demand of water has significantly dropped and the total

availability of water has increased, as there are no major rivers going through the part of the map, the groundwater levels were checked and there was a noticeable amount of increase in its levels when comparing the decade 1990-2000 and 1980-1990 and this similar trend is carried forward when looking into the next years from 2000-2016 where the groundwater levels have increased along with the total available water in the area.

To contrast a water rich country like the U.S.A, we chose to do a similar clustering of a country which is poorer in terms of water resources like India to further understand if clustering can let us identify parts of the map which would be a cause of concern in terms of water security. When we have a higher number of clusters for India the clustering method is able to identify river bodies but this gives us less information in regards to areas which have problems.When clustering with lower numbers, we are able to easily identify the regions which have degraded which is showcased in the 3 image in the Figure 4, as the brown color growth which represents the high demand in water and a low supply from groundwater sources are clustered around namely two major cities of India i.e. Delhi NCR and Chennai. These cities face this challenge because of rapid increase in migration from people from the towns and villages to the cities and deforestation and pollution of the water bodies near these two cities resulting in them having one of the lowest sources of total water in India. The new information revealed is that Delhi NCR is not known to be in as poor condition as Chennai in the media but this clustering reveals that the Indian government should put more efforts in securing the water supply for the capital as it's also not close to any major water body.

Figure 4: India k-means clustering (3) with weather data

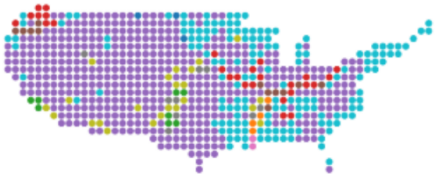


Figure 5: K-means(10-C) USA with projected data 2040-2050

Our 2050 forecast indicates that there could be some stress in the Messissippe river valley as there seems to be breakage of the link which is not in the real data.

2: Predicting WPI

The water poverty index is a composite score that captures water scarcity per country. We attempt to use the per-cell data we collected in order to predict WPI. For this we use Lasso against the variables of the last year in the data (around 80 columns). Each country is intersected against the earth cell lattice, and the corresponding cells are used as input. Using Lasso with cross-validation yields an adjusted R^2 of 0.44 for WPI variation per cell separately. In addition, only 24 of the 80 yearly variables were considered significant.

Similarly, we train an LSTM network on one year's variables again with the goal of predicting WPI. LSTM allows us to account for the variable number of cells that each country represents. It's a simple network of 3 layers of 5 nodes each with a final fully connected node. The network is able to predict WPI with a validation MSE of 3.4.

We then build a per-variable per-cell model that attempts to forecast the variable in this region. We use a simple linear regression for this purpose. Even though more sophisticated models can be used for this purpose, the linear regression model captures the general trend of the variables well.

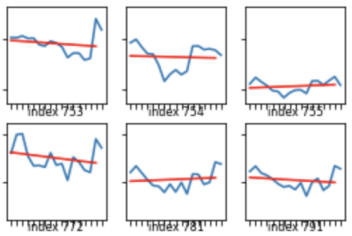


Figure 6: 6 adjacent 1-degree cells with different forecasts

Based on these models, we forecast the variables in 2050 and we predict WPI using both the lasso and LSTM models on these new variables for each country. The Lasso model predicts that Denmark, Canada, Finland, Russian Federation, Kazakhstan, Sweden, Belarus, Poland, Ireland, and Norway would be doing well. Meanwhile Equatorial Guinea, Guyana, Eritrea, Yemen, Colombia, Malaysia, Sierra Leone, Guatemala, Guinea, Oman would have a low WPI. This roughly maps to northern countries and equatorial regions respectively.The LSTM model similarly predicts northern countries for countries that have a high WPI, and African and equatorial countries as ones with lower water scores.

3: Grouping variables using PCA

The projected future data for the year 2050 was aggregated by country by averaging the values of the country's geographical cells. Further, the data was standardized to avoid scaling issues. A PCA analysis was

conducted for dimensionality reduction and to identify clusters.

The first 10 principal components accounted for 81.6% of the total variation in the data. The elbow method, used to select the number of principal components, did not work well as the scree plot showed a smooth curve with an ambiguous k.

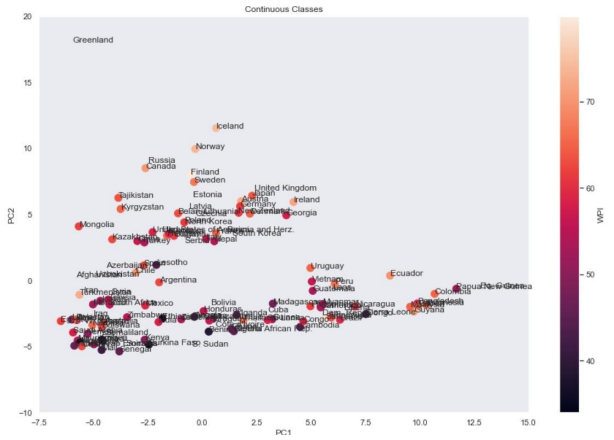


Figure 7: 2-dimensional biplot based on 2002 data

Countries have been plotted according to their first and second component values. Countries in the top half of the plot on the PC2 axis represent the well-off countries that score high on WPI. The far right has grouped countries with tropical climates. The bottom left represent the countries that are under high water stress in Africa and the Middle East.

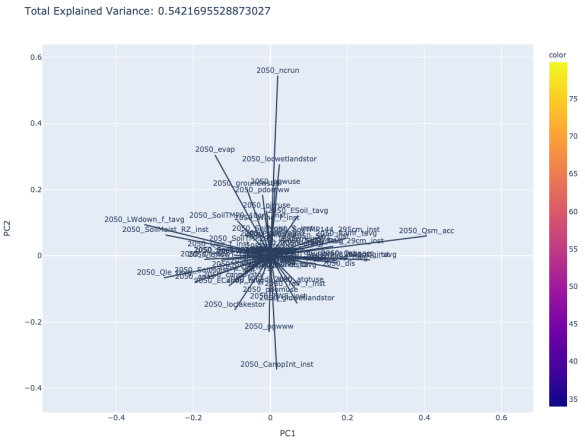


Figure 8: Vector representation of 2-dimensions

PC1 seems to be closely related to water evaporation as diffuse groundwater recharge, monthly precipitation, latent heat net flux and evapotranspiration had the highest absolute correlation in this component. It also appears to be directly proportional to WPI. PC2 captures net cell runoff and potential water consumption from WorldGAP.

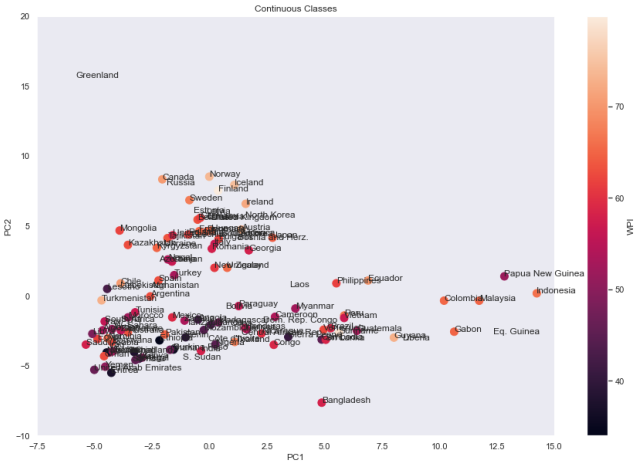


Figure 9: 2-dimensional biplot based on 2050 data

Using PCA with our 2050 forecast, we can see that Bangladesh stands out as a country that will experience an increasing amount of water stress.

Conclusions

Geological and water models capture complex planet-level variables that provide a useful snapshot of the world. Unsupervised clustering techniques on these variables show geographical patterns that match reality and provide insights to the real underlying geological and hydrological changes.

Using supervised learning methods with these variables as well as socioeconomic data (population, GDP) as predictors and carefully curated water scores as the response variable is a viable path for building prediction models.

While these variables are complex on a global scale, they show local trends that can potentially be modelled for a small region. Potential areas for improvement are using more sophisticated models, or using scenario based forecasting where we have 3 different scenarios that reflect a negative, neutral and positive projection.

Exploring the CDP city responses shows that very few cities (e.g Cape Town, Tokyo, Dubai São Paulo, Singapore and Los Angeles) are being conscientious in addressing identified water issues through mitigation plans. As a matter of fact, in most of the cities disclosing, water security and scarcity were not identified as risks. It could very well be that these cities are water secure but this may also indicate unawareness to potential risks and existing issues, which is most detrimental.

Only 4 out of the 10 identified “lowest” countries from the WPI predictions have few cities disclosing to CDP (i.e., Tanzania, India, Kenya, Morocco from LSTM and Colombia, Malaysia, Sierra Leone, and Guatemala from Lasso). For example, Dar es Salaam in Tanzania has already started implementing actions to address inadequate

or ageing water supply infrastructure and increased water stress while two cities from India have been implementing and monitoring different adaptation actions to address inadequate or aging water supply infrastructure, increased water stress, scarcity, and demand and pollution incidents. Kisumu, Kenya has been implementing watershed preservation, stormwater management and investment in existing water supply infrastructure. Lastly, Le Grand Casablanca in Morocco is now assessing the impact of investment in existing water supply infrastructure and implementing conservation awareness and education.

On the other hand, almost all LSTM 'winning' countries are disclosing to CDP with the exception of Belarus and Paraguay. The city of Gdynia in Poland has no identified water security risks. While, only 1 of 4 disclosing cities from Norway has identified the risk of inadequate or ageing water supply infrastructure for which investment in existing infrastructure is already being implemented; other three cities have answered "Not Applicable" on the question of any water security risk identified. For the Russian Federation, only the increased water stress is the identified risk for which water metering is done as an adaptive action. In Germany, 2 cities have acknowledged various water risks but mitigation plans are already being implemented. Lastly, the remaining LSTM 'winning' countries, Finland (6 cities), Brazil (61 cities), US (132 cities) and Indonesia (7 cities), have noted different water risks and actions. In general, these countries with the highest predicted WPIs have several cities reporting or have completed efforts to resolve inadequate or ageing water supply infrastructure. This is not surprising as these countries are considered rich and are therefore, more progressive and capable of implementing actions.

It is important to note however that the few disclosing cities do not represent the conditions of the entire country. Further, the mitigation plans and present actions, although laudable, may not guarantee complete resolution later on. In any case, recognizing and disclosing water risks together with proactive planning and committed actions are important in ensuring water security, today and in the future.

APPENDIX I. DATA SOURCES

Shared Socioeconomic Pathways (SSP)

The SSP database was developed with the aim of documenting and exploring different scenarios of how the world might change over the years. There are five pathways under this framework, each designed to account for various uncertainties in the future. The data used in this research is SSP2, described as the Middle of the Road with the society facing medium challenges to mitigation and adaptation. Particular variables used in the SSP2 database are the global population (in millions) and GDP (in PPP, Billion US$2005/yr) given from 1980 to 2100 by 10 years in 0.5 × 0.5 degree grids by country.

Since the data is provided in 10-year increments, the values were linearly imputed to obtain yearly data and transformed to a spatial resolution of 1 x 1 degree grids.

Water – Global Assessment and Prognosis (WaterGAP)

WaterGAP is a global hydrological model that quantifies human use of groundwater and surface water as well as water flows and water storage and thus water resources on all land areas of the Earth (Müller Schmied, 2021). Outputs of the WaterGAP 2.2d model which includes 40 water-related variables provided either monthly or yearly from 1901 to 2016 in 0.5 × 0.5 degree grids by country were included in the analysis.

To achieve consistent yearly reporting, variables provided monthly were aggregated to yearly data by taking the average values. The data was also transformed to a spatial resolution of 1 x 1 degree grids.

NASA Global Land Data Assimilation System (GLDAS)

The goal of the Global Land Data Assimilation System (GLDAS) is to ingest satellite- and ground-based observational data products, using advanced land surface modeling and data assimilation techniques, in order to generate optimal fields of land surface states and fluxes (Rodell et al., 2004a). The GLDAS Catchment Land Surface Model v2.1, which was used in this study, includes 38 monthly global land surface variables from 2000 to 2021 in 1 x 1 degree grids. The monthly data was averaged to obtain a value representative for a year.

CDP

CDP Cities disclosure cycle 2020 was used to explore the water security risk drivers together with the actions each responding city is taking to reduce the risks. Specifically, responses from Section 14.2a of the questionnaire provides the information on water security risk drivers while Section 14.3 provides the data on the adaptation action and status of action taken per risk identified. Only the disclosing cities from 75 countries were included in the scatter geomaps visualization.

WPI

The Water Poverty Index (2002) is a water management tool created to assess water stress and scarcity in 141 countries around the world. The indicator links household welfare with water availability and is calculated using 5 components: resources, access, capacity, use and environment.

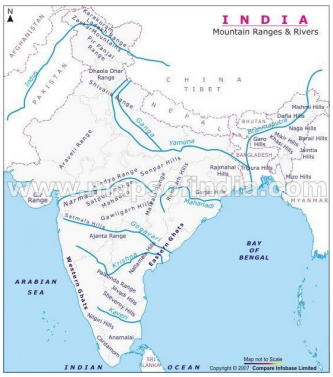
Integration of Data Sources

The different data sources were merged together to arrive at a single data frame spanning from 1980 to 2016, with the GLDAS data reporting missing from 1980 to 1999. The intersection of all gridded cells resulted in 10,880 rows. Subsequent data analyses were based on this merged data frame.

APPENDIX II. IMAGES

A. US Major River Image

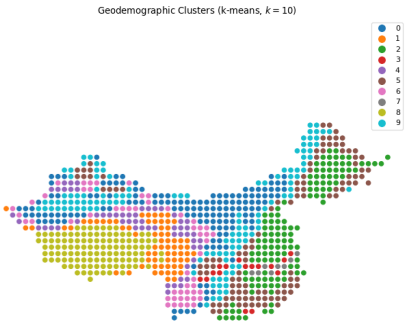
B. Indian River



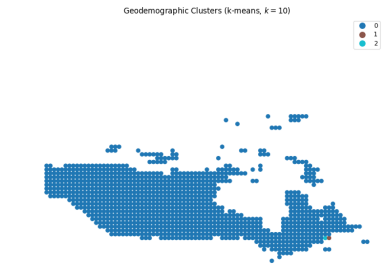
C. Further Clustering, with NASA Data

i. USA k-means clustering with weather data

ii. China k-means clustering with weather data



iii. Canada k-means clustering with weather data



REFERENCES

Center for Global Environmental Research, Tsukuba International Office, Global Carbon Project. Population. http://db.cger.nies.go.jp/dataset/gcp/population-and-gdp/pop_ssp2.csv

Center for Global Environmental Research, Tsukuba International Office, Global Carbon Project. GDP. http://db.cger.nies.go.jp/dataset/gcp/population-and-gdp/gdp_ssp2.csv

Climate Disclosure Panel. *2020 - Full Cities Dataset: CDP Open Data Portal*. 2020 - Full Cities Dataset | CDP Open Data Portal. https://data.cdp.net/Governance/2020-Full-Cities-Dataset/eja6-zden.

Explainer: How 'Shared Socioeconomic Pathways' explore future climate change. Carbon Brief. (2019, February 21). https://www.carbonbrief.org/explainer-how-shared-socioeconomic-pathways-explore-future-climate-change.

Global dataset of gridded population and GDP scenarios: GCP Tsukuba International Office. Global dataset of gridded population and GDP scenarios | GCP Tsukuba International Office. (n.d.). http://www.cger.nies.go.jp/gcp/population-and-gdp.html.

Intergovernmental Hydrological Programme. (2017, May 30). *Water Poverty Index (2002)*. http://ihp-wins.unesco.org/layers/geonode_ihp_data:geonode:wpif.

Li, B., H. Beaudoing, and M. Rodell, NASA/GSFC/HSL (2020), GLDAS Catchment Land Surface Model L4 monthly 1.0 x 1.0 degree V2.1, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 30-May-2021, 10.5067/FOUXNLXFAZNY

Müller Schmied, Hannes; Cáceres, Denise; Eisner, Stephanie; Flörke, Martina; Herbert, Claudia; Niemann, Christoph; Peiris, Thedini Asali; Popat, Eklavyya; Portmann, Felix Theodor; Reinecke, Robert; Shadkam, Somayeh; Trautmann, Tim; Döll, Petra (2020): The global water resources and use model WaterGAP v2.2d - Standard model output. PANGAEA. https://doi.org/10.1594/PANGAEA.918447.

Müller Schmied, Hannes; Cáceres, Denise; Eisner, Stephanie; Flörke, Martina; Herbert, Claudia; Niemann, Christoph; Peiris, Thedini Asali; Popat, Eklavyya; Portmann, Felix Theodor; Reinecke, Robert; Schumacher, Maike; Shadkam, Somayeh; Telteu, Camelia-Eliza; Trautmann, Tim; Döll, Petra (2021): The global water resources and use model WaterGAP v2.2d: Model description and evaluation. Geoscientific Model Development, 14(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021.

Murakami, D. and Yamagata, Y. (2016) Estimation of gridded population and GDP scenarios with spatially explicit statistical downscaling, ArXiv, 1610.09041. https://arxiv.org/abs/1610.09041.

NASA GESDISC Data Archive. Land Data Assimilation System (LDAS). https://hydro1.gesdisc.eosdis.nasa.gov/data/GLDAS/GLDAS_CLSM10_M.2.1/.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., van Ruijven, B. J., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., & Solecki, W. (2017). The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century. Global Environmental Change, 42, 169–180. https://doi.org/10.1016/j.gloenvcha.2015.01.004

Rodell, M., P.R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J.K. Entin, J.P. Walker, D. Lohmann, and D. Toll, 2004: The Global Land Data Assimilation System, Bull. Amer. Meteor. Soc., 85, 381-394. doi:10.1175/BAMS-85-3-381

SSP Database. (n.d.). https://secure.iiasa.ac.at/web-apps/ene/SspDb/dsd?Action=htmlpage&page=about.

United Nations. (n.d.). Goal 6 | Department of Economic and Social Affairs. United Nations. https://sdgs.un.org/goals/goal6.