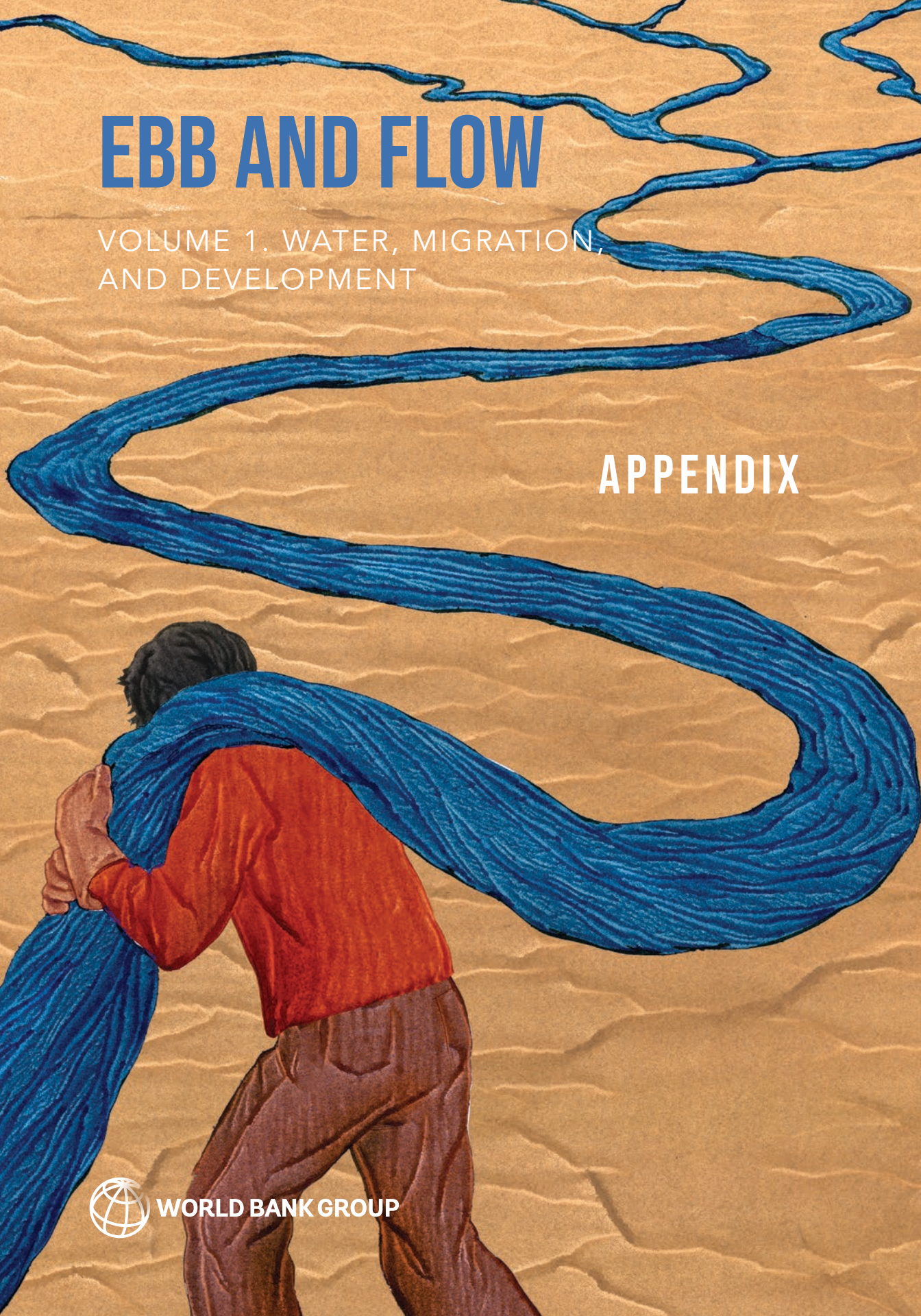
EBB AND FLOW

VOLUME 1. WATER, MIGRATION, AND DEVELOPMENT

APPENDIX



WORLD BANK GROUP

EBB AND FLOW

VOLUME 1. WATER, MIGRATION, AND DEVELOPMENT

APPENDIX

WORLD BANK GROUP

1818 H Street NW, Washington, DC 20433
Telephone: 202-473-1000; Internet: www.worldbank.org

Cover image: Alex Nabaum, via theispot.com.

CONTENTS

FIGURES

MAPS

TABLES

TECHNICAL APPENDIX OF RESULTS

This section describes the main empirical and identification techniques used in the report. Following the description, regression results that form the basis for those described in the report are displayed. This section is meant only to facilitate easy access to the main results and will not be printed in the final documents.

CHAPTER ONE

TRANSITIONS AND TRANSFORMATIONS

LEARNING ABOUT WATER'S ROLE IN GLOBAL MIGRATION FROM HALF A BILLION INDIVIDUAL RECORDS

MIGRATION DATA AND SOCIO-DEMOGRAPHIC CHARACTERISTICS

Individual data on migration and socio demographic characteristics were obtained from harmonised census microdata samples from the Integrated Public Use Microdata Series International (IPUMSI) database (Minnesota Population Center, 2015). IPUMSI provides the world's largest archive of publicly available census samples, with variables harmonised across countries and over time to facilitate comparative research. Samples in IPUMSI are typically close to 10 percent of the entire census; see Table A1.1 in the appendix for the sampling fractions of the IPUMSI samples used in our study.

Over 280 census samples in IPUMSI contained variables related to migration. We selected only the samples with variables associated with responses to questions on the individual's previous region of residence. The identification of migrants, given the availability of a response to a question

in individuals' past region of residence was derived in either one of the two following approaches. In the first approach we used an IPUMSI variable on an individual's region of residence at a fixed period of time prior to the census, such as one or five years ago. If the region was different from the region of the respondent's household (provided in another IPUMSI variable on the region of respondent at the time of the census) then we coded the respondent as a migrant. If a variable on a fixed period of time prior to the census was unavailable, we used a combination of three variables; the first variable contains the respondent's previous region of residence, the second variable indicates if the previous residence of the respondent was in the same or different region to the current location and the third variable gives the number of years the respondent has resided in their current location. If the respondent had changed their location within a year, and the previous residence was different to the current location (based on the second and third variables), then we coded the respondent as a migrant. We used the first variable on the previous region of residence to identify the climatic conditions for the migrant origin region, as discussed later. For the IPUMSI census samples where we derived migrants using the latter approach are indicated using the 1* in Table A1.1, to indicate an implied period of migration within a year from the respective census.

Under both approaches to deriving migrants we are identifying migration based on transitions rather than the movements. In other words, we derived migrant events from a comparison of the regions of residence at the start and end of a given interval for each individual. We do not capture multiple changes of residence that might have occurred within the period or region, and so, for example we would miss a migrant who moved to a different region and then returned to their original region within the migration interval. At a population level, the difference between the number of migration events and migration transitions will be small when the interval length is short, for example one-year. The discrepancy increases in a non-linear fashion as the interval length increases; see for example Rees (1977) or Courgeau (1973) for further details on the differences between migration transitions and movements.

MEASUREMENT OF MIGRATION

We were able to create a binary migrant variable for individuals in 189 IPUMSI census samples listed in Table A1.1. Although there were more IPUMSI census samples with migrant variables, an additional variable on the origin region, that we used for identifying climate conditions of migrants in our analysis were not available in other samples. In total our selection of IPUMSI census samples cover 441.6 million individual records, from which 18.7 million were coded as migrants. Summary details on the number of records and migrants by country and census year can be found in Table A1.1.

Direct cross-national comparisons between numbers of migrants, migration rates or the probability of migration are not possible due to the different census sample sizes and the different set of regional geography in each country, where countries are subdivided into different numbers of regions and regions themselves can vary greatly in both geographic and demographic scales. Consequently, the number of migrants is likely to be relatively higher in countries with more regions and large population in comparison to similar country with fewer regions and a smaller population (Bell, Charles-Edwards, Kupiszewska, et al. 2015; Bell, Charles-Edwards, Ueffing, et al. 2015) – a feature related to the Modifiable Areal Unit Problem (MAUP). In all countries we used a harmonised first level administrative regional geography provided by IPUMSI.

MEASUREMENT OF SOCIO-DEMOGRAPHIC CHARACTERISTICS

In addition to deriving migrant status for all individual in the census samples we also used five harmonized IPUMSI variables on individuals available in all 189 census samples; age, sex, education, marital status and the number of persons living in their household. These are the variables relevant to the migration decision identified in previous literature. Age and the number of persons living in the household (household size) are continuous variables. Sex is a binary variable classified into male and female. The harmonised education variable contains four categories: less than primary, primary completed, secondary completed and university completed. The harmonised marital status variable also contains four categories: single/never married, married/in union, separated/divorced/spouse absent and widowed.

ENVIRONMENTAL DATA

CLIMATE VARIABLES

Climatic shocks are measured as rainfall deviations from the long-term panel mean of rainfall or temperature for a given country. The deviations are measured as standardised z-scores which allow us to measure climatic differences in terms of the historical range of climatic variability across heterogeneous areas. Z-scores are considered to be better predictors of migration outcomes compared to raw climate data (Gray and Wise 2016; Mueller, Sheriff, et al. 2020). We calculate subnational-level annual rainfall z-scores for each census-year. Positive z-scores indicate wetter than normal conditions while negative values indicate drier than normal conditions.

Rainfall data are obtained from the gridded monthly time series data from Matsuura and Willmott (2018). The time-series data for rainfall consist of monthly average precipitation calculated on high-resolution (0.5 x 0.5 degree) grids. The precipitation grids are used to calculate rainfall z-scores

for each administrative unit matching the place of origin indicated by the previous place of residence of the migrants.

Depending on the migration interval defined in each census (place of residence 1, 5 or 10 years preceding the census), corresponding climatic conditions were calculated based on the length of the migration period and an additional two years before, i.e., 3, 7 and 12 years prior to the census respectively. For instance, for the Iraq Population and Housing Census 1997, the participants were asked about their place of residence 10 years ago. Accordingly, the rainfall z-scores for the entire population (migrants and non-migrants) are calculated based on average rainfall of the period 1984-1996 and the long run average based on the complete historical series.

METHODS

We work with 442 million individual records from 1,329 regions across 189 censuses covering 64 countries. This exceptionally large volume of data makes it inappropriate to estimate migration drivers using classical regression methods. Instead, we make use of random forest models on a country level to identify populations that were particularly sensitive to rainfall shocks in their migration behaviour. This non-parametric technique allows us to identify patterns in the data without imposing any statistical assumptions and being sensitive to misuse of significance testing like in conventional multiple regression models (Attewell et al. 2015).

RANDOM FOREST ANALYSIS

A random forest model is performed for each of the 64 countries in our dataset. Random forest is a robust machine learning algorithm that can be used for a variety of analyses including regression and classification. The technique is suitable for our context with a large dataset where we aim to identify important covariates underlying migration decisions. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, which each produce their own predictions (Hastie 2009).

The individual decision trees apply a systematic truncation of a data sample with regard to the distribution of an outcome variable (target feature). The target feature in our case is internal migration, which distinguishes each individual observation between migrants (labelled 1) and non-migrants (labelled 0). With regard to this target feature, the initial sample of each country's population has a certain distribution, e.g., 10 percent of a country's population migrated and 90 did not migrate. For this initial distribution, a measure of data purity, in our

case the Gini impurity, is calculated. The Gini impurity is a measure of how often a randomly chosen element from the sample would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the sample. This measure serves as a benchmark for the subsequent steps.

In each decision tree, our features, personal characteristics and regional climatic shocks, are considered for the truncation of the initial sample. The data purity of the two resulting subsets (nodes) is compared. The external feature that yields the highest level of data purity in the remaining subsets is chosen as a splitting criterion. After this split, the initial procedure is repeated for each of the resulting subsamples, testing all features: a tree-like, cascading structure emerges after repeating the procedure several times.

More formally, at each node τ within the decision tree the optimal split is sought using the Gini impurity G(τ). If we have C possible outcomes of our target feature, in our case 0 (for no migration) and 1 (for migration), and $p(i)$ is the probability of choosing a datapoint with class i, then the Gini impurity is calculated as:

$$G(\tau) = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Formally, we can measure the change in our Gini impurity measure $\Delta G(\tau)$ after a truncation by one of our input features.

In a random forest model, randomly selected decision trees are combined (Hastie 2009). Each node in the decision tree works on a random subset of features to calculate the output. The random forest then combines the output of individual decision trees to generate the final output. In this ensemble of decision trees, an exhaustive search over all features θ available at the node the maximal $\Delta G(\tau)$ is determined. The decrease in Gini impurity resulting from this optimal split $\Delta G_{\theta}(\tau, T)$ is recorded and accumulated for all nodes τ in all trees T in the forest, individually for all variables θ:

$$MDG\ (\theta) = \sum_{T} \sum_{\theta} G_{\theta}(\tau, T)$$

This quantity, the Mean Decrease Gini (MDG), calculates by how much the performance (Gini) of the random forest model would deteriorate, on average, if a respective feature were to be excluded from the model (Hastie 2009). A high MDG indicates that the respective characteristic is important when truncating the sample with regard to the target feature, similar to a change in R^2 in a regression model.

Our random forest model is fed with a ten percent sample of each country's population and the accuracy of the random forest with 1,000

tree combinations is evaluated on a 25 percent test set, a typical split for cross validation. With MDGs for all features – individual and climate characteristics – at hand, the relative MDG (RMDG) can be calculated for each feature θ.

$$RMDG_{\theta,\eta} = MDG_{\theta}/MDG_{\eta} * 100$$

This value of the RMDG assesses the importance of a given feature θ, relative to the importance of another reference feature. For our purpose, the MDG of the climatic shock characteristic is compared with the MDG of the education variable, as education is known to be a relevant factor for migration behaviour (Ginsburg et al. 2016). A RMDG of more than 100 indicates that a given climatic shock is more relevant for explaining migration patterns than educational attainment. As a selection threshold, only cases with a climatic shock MDG of higher than 50, e.g., climatic shocks being at least half as important as education, are marked as sensitive populations and considered in the following regional pairwise comparison.

WHY MACHINE LEARNING

No doubt multiple regression analysis is popular in policy-related research. Regression models have the advantage of quantifying relationships between variables of interest in a custom and interpretable way, e.g., beta coefficients that allow for making statements about the magnitude (size of the coefficient) and relevance (significance level) of the relationship. However, a very large sample size impedes an application of regression models because the expressiveness of significance levels diminishes for samples larger than 10.000 observations (Lin et al. 2013). Classification techniques in machine learning techniques, such as random forests, on the other hand, do not rely on variances and are applicable also on larger samples. Among machine learning classification techniques, random forests come with a data purity metric i.e. the mean decrease in Gini that allows us to compare the influence of different features for the classification, R squares in a multivariate regression analysis.

RESULTS

INDIVIDUAL AND ENVIRONMENTAL DRIVERS OF MIGRATION

Each individual in the dataset is classified into two categories: migrants and non-migrants; our outcome of interest. Random forest classification models

are applied to identify important covariates with the greatest predictive power for explaining individual migration decisions.

In addition to migration, the dataset contains the individual characteristics i.e. gender, age, education and marital status, and household characteristics like household size. Furthermore, with the information on the region of origin of migrants, we are able to identify the environmental conditions i.e. occurrence of a climatic shock in each region and census year in our dataset.

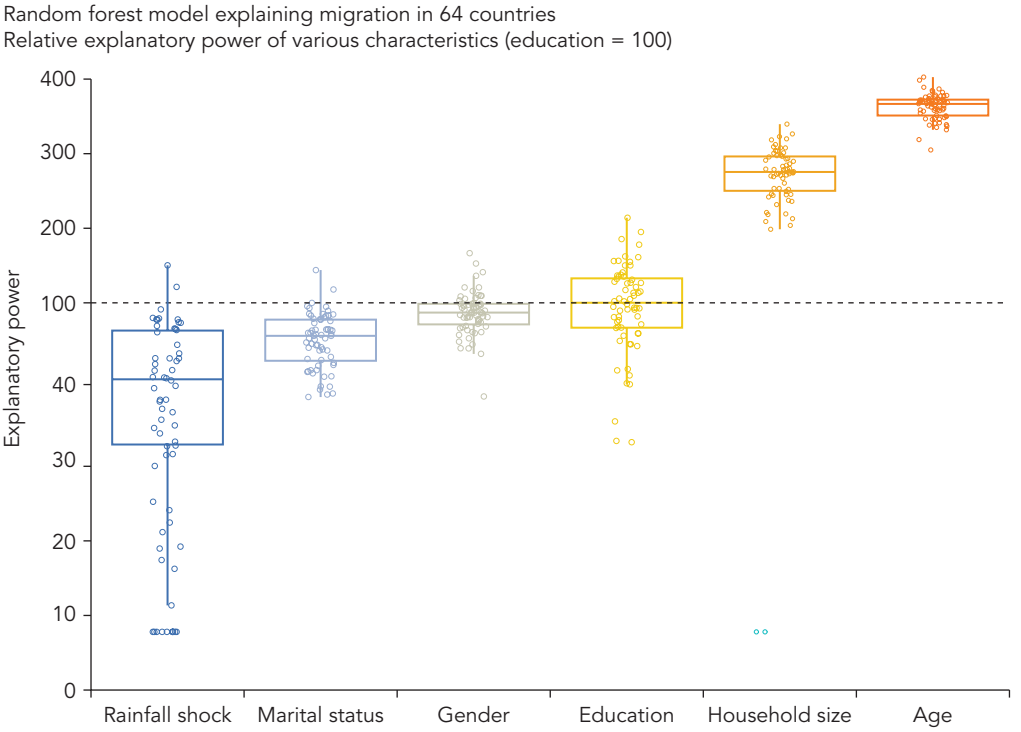
As part of our first analysis round, we apply a random forest model that allows us to compare the importance of each of the individual characteristics and exposure to climatic shocks for migration for 68 countries in our dataset.

Figure A1.1 summarises the results of our random forest model for all six characteristics (five individual features and a climatic shock occurrence) across 64 countries. Each dot shows how important a respective characteristic is for explaining migration in a given country. As we know that education is a relevant determinant of migration decisions (Dustmann and Glitz 2011), the 'explanatory power' (RMDG) of each characteristic in a given country is shown relative to the importance that education has, on average, across all countries (100 = dotted line).

We see a clear difference in explanatory power across characteristics. While age and household size, on average, are more important than education, other characteristics like sex or marital status are, on average, as relevant as education. The climatic shock analysed in this model, i.e., the occurrence of a negative precipitation shock, is slightly less important than education (51% of education's importance), on average. However, the widespread distribution of the climate characteristic indicates that negative precipitation shocks are as important as sex, marital status or education, in some countries.

For each country, we are able to compare the importance of each of the six features, i.e., five individual characteristics and one contextual variable i.e. rainfall shock, when explaining individual migration. A ranking of the importance of each feature shows that age and household size usually have the highest MDG, followed by education, gender, and marital status. The rainfall shock feature is usually ranked as the least relevant of all features. This is of little surprise as the five individual characteristics are known to be relevant drivers of migration. However, in contrast to individual characteristics, climate shocks vary strongly in their importance in explaining migration. For a set of countries shown in Figure A1.1, precipitation shocks even have a higher feature importance than some individual characteristics. For some of these countries, mostly less economically affluent countries, the rainfall shock was even ranked as the fourth or third most important explanatory factor of all features.

FIGURE A1.1: The Importance of Various Characteristics in Explaining Migration



DISCUSSION

Applying machine-learning techniques to analyse drivers of migration, this is the first study to comprehensively and consistently investigate the role of climatic shocks on internal migration for a large number of countries (64 countries). As highlighted in the conceptual framework explaining environmental drivers of migration from Black et al. (2011), migration is determined by the interactions among individual, meso and macro-level factors. Using individual educational attainment as a reference point to compare the importance of each factor underlying migration, we find that age and household size are key determinants of migration as well-documented in the literature (Borderon et al. 2019; Tsegai 2007). In addition, we find that relative to education and other individual characteristics including gender and marital status, climatic shocks also play an important role in influencing migration decisions.

Unlike traditional regression analysis which is typically driven by theoretical concepts to select which variables to focus on, in random forest

models drivers of migration are identified through a data driven approach (Best et al. 2020). Furthermore, being non-parametric models, random forest algorithms allow us to identify salient variables in large-scale and complex datasets without imposing any assumptions on the data distribution. Our finding that climatic shocks are as important as some individual characteristics in determining internal migration provides a solid evidence on the importance of environmental variables in influencing migration.

Future extensions of this study could explore alternative machine learning methods, such as support vector machines that are well suited for a truncation of high-dimensional feature space, and combine them with inferential statistics, like regression analysis. Again, machine learning isolates relevant population samples on a national, regional, and individual level, which could then be analysed by the means of regression analysis. This could help further reduce the limitations that increasingly popular big social datasets impose on established multivariate regression analysis.

The major limitations for this study are related to migration data:

First, relying on the migration-related question in the population censuses which asks the individuals about their previous place of residence, it is not possible to distinguish between short- and long-term migration and temporary and permanent migration. Previous studies have shown that different types of migration are adopted to cope with temperature and rainfall anomalies with temporary migration being used as an adaptive response for climate-vulnerable households (Bohra-Mishra et al. 2014; Joarder and Miller 2013; Mueller, Sheriff, et al. 2020; N. E. Williams and Gray 2020). Climate-related migration could, therefore, be underestimated in our case because we only capture migration that occurred within at least a one year interval.

Second, the size of administrative boundaries varies considerably across countries and time. Movements in a country with a higher number of subnational regions are more likely to be classified as migration because it is naturally easier to cross a regional boundary. Although this affects our cross-national comparison of climate-related migration, the number of administrative regions is distributed randomly across income groups. Therefore, the bias in our comparison of climate-related migration across countries by their GDP per capita is limited.

Third, our study is limited to 64 countries in the sample where countries in Asia (9) and Europe (6) are underrepresented. The findings, therefore, cannot be generalised as representing the climate-related migration patterns for the whole world.

Despite the limitations, this is the first large-scale individual-level study that comprehensively and systematically assess the role of climatic factors on internal migration. The application of machine learning techniques allows us to identify important factors driving migration without imposing any theoretical nor statistical assumptions.

REFERENCES

Bell, M., Charles-Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J., & Zhu, Y. (2015). Internal Migration Data Around the World: Assessing Contemporary Practice. *Population, Space and Place*, *21*(1), 1–17. https://doi.org/10.1002/psp.1848

Bell, M., Charles-Edwards, E., Ueffing, P., Stillwell, J., Kupiszewski, M., & Kupiszewska, D. (2015). Internal Migration and Development: Comparing Migration Intensities Around the World. *Population and Development Review*, *41*(1), 33–58. https://doi.org/10.1111/j.1728-4457.2015.00025.x

Best, K. B., Gilligan, J. M., Baroud, H., Carrico, A. R., Donato, K. M., Ackerly, B. A., & Mallick, B. (2020). Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. Journal of Computational Social Science. https://doi.org/10.1007/s42001-020-00066-9

Bohra-Mishra, P., Oppenheimer, M., & Hsiang, S. M. (2014). Nonlinear permanent migration response to climatic variations but minimal response to disasters. Proceedings of the National Academy of Sciences, 111(27), 9780–9785. https://doi.org/10.1073/pnas.1317166111

Courgeau, D. (1973). Migrants et migrations. *Population*, *28*(1), 95–129. https://doi.org/10.2307/1530972

Gray, C., & Wise, E. (2016). Country-specific effects of climate variability on human migration. *Climatic Change*, *135*(3–4), 1–14. https://doi.org/10.1007/s10584-015-1592-y

Hastie, T., Tibshirani, R., & Friedman, J.(2009). Random forests. In *The elements of statistical learning* (pp. 587-604). Springer, New York, NY. https://doi.org/10.1007%2F978-0-387-84858-7_15

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Joarder, M. A. M., & Miller, P. W. (2013). Factors affecting whether environmental migration is temporary or permanent: Evidence from Bangladesh. Global Environmental Change, 23(6), 1511–1524. https://doi.org/10.1016/j.gloenvcha.2013.07.026

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906-917. https://doi.org/10.1287/isre.2013.0480

Lutz, W. (2014a). A Population Policy Rationale for the Twenty-First Century. *Population and Development Review*, *40*(3), 527–544. https://doi.org/10.1111/j.1728-4457.2014.00696.x

Matsuura, K., & Willmott, C. J. (2018). Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900–2017), http://climate.geog.udel.edu/~climate/html_pages/download.html

Minnesota Population Center. (2015). *Integrated Public Use Microdata Series, International: Version 6.4 [Machine-readable database]*.

Mueller, V., Sheriff, G., Dou, X., & Gray, C. (2020). Temporary migration and climate variation in eastern Africa. *World Development*, *126*, 104704. https://doi.org/10.1016/j.worlddev.2019.104704

Rees, P. (1977). The measurement of migration, from census data and other sources. *Environment and Planning A*, *9*(3), 247–272. https://doi.org/10.1068/a090247

Williams, N. E., & Gray, C. (2020). Spatial and Temporal Dimensions of Weather Shocks and Migration in Nepal. *Population and environment*, *41*(3), 286–305. https://doi.org/10.1007/s11111-019-00334-5

TABLE A1.1: IPUMSI Census Samples

Census	Sample Size	Sampling Fraction	Regions	Migration Interval	Migrants Derived	ISO Code
Argentina 1970	466892	2	24	5	36240	ARG
Argentina 1980	2667714	10	24	5	159315	ARG
Argentina 2001	3626103	10	24	5	116180	ARG
Armenia 2001	326560	10	11	1*	2846	ARM
Armenia 2011	301831	10	11	1*	3464	ARM
Benin 1979	331049	10	12	1*	12797	BEN
Benin 1992	498419	10	12	1*	14747	BEN
Benin 2002	685467	10	12	1*	23630	BEN
Benin 2013	1009693	10	12	1*	28271	BEN
Belarus 1999	990706	10	6	1*	9196	BLR
Bolivia 1976	461699	10	10	5	21034	BOL
Bolivia 1992	642368	10	10	5	33005	BOL
Bolivia 2001	827692	10	10	5	46141	BOL
Brazil 1991	17045712	10	25	5	555209	BRA
Brazil 2000	20274412	10	25	5	609460	BRA
Brazil 2010	20635472	10	25	5	476703	BRA
Botswana 1981	97238	10	21	1	8815	BWA
Botswana 1991	132623	10	21	1	15923	BWA
Botswana 2001	168676	10	21	1	22724	BWA
Botswana 2011	201752	10	21	1	19580	BWA
Canada 1991	809654	3	10	1	9712	CAN
Canada 2001	801055	2.7	10	1	7743	CAN
Chile 1982	1133062	10	44	5	93498	CHL
Chile 1992	1335055	10	44	5	121945	CHL
Chile 2002	1513914	10	44	5	141448	CHL
China 1990	11835947	1	29	5	117122	CHN
China 2000	11804344	1	29	5	319091	CHN
Cameroon 1976	736514	10	7	1*	23168	CMR
Cameroon 1987	897211	10	7	1*	35580	CMR
Cameroon 2005	1772359	10	7	5	108935	CMR
Colombia 1964	349652	2	22	1*	14431	COL
Colombia 1973	1988831	10	22	1*	79254	COL
Costa Rica 1963	82345	6	7	1*	3760	CRI
Costa Rica 1973	186762	10	7	5	14244	CRI

Costa Rica 1984	241220	10	7	5	17412	CRI
Costa Rica 2000	381500	10	7	5	29118	CRI
Costa Rica 2011	430082	10	7	5	21107	CRI
Cuba 2002	1118767	10	29	1*	9017	CUB
Dominican Rep 1981	475829	8.5	26	5	29163	DOM
Dominican Rep 2010	943784	10	26	5	52314	DOM
Ecuador 1962	136443	3	14	1*	3082	ECU
Ecuador 1974	648678	10	14	1*	31618	ECU
Ecuador 1990	966234	10	14	5	47556	ECU
Ecuador 2001	1213725	10	14	5	52027	ECU
Ecuador 2010	1448233	10	14	5	70779	ECU
Egypt 1996	5902243	10	25	1*	21041	EGY
Egypt 2006	7282434	10	25	1*	26801	EGY
Spain 1981	2084221	5	19	10	115240	ESP
Spain 1991	1931458	5	19	1	82283	ESP
Spain 2001	2039274	5	19	10	114411	ESP
Spain 2011	4107465	10	19	1	141973	ESP
Fiji 1976	57214	10	4	5	4743	FJI
Fiji 1986	72158	10	4	5	5720	FJI
Fiji 1996	77382	10	4	5	6371	FJI
Fiji 2007	84323	10	4	5	8175	FJI
United Kingdom 1991	541894	1	12	1	11063	GBR
United Kingdom 2001	1843525	3	12	1	47480	GBR
Ghana 2000	1894133	10	10	5	56759	GHA
Guinea 1996	729071	10	52	1*	73261	GIN
Greece 1971	845483	10	54	5	68386	GRC
Greece 1981	923108	10	54	5	86145	GRC
Greece 1991	951875	10	54	1	26710	GRC
Greece 2001	1028884	10	54	1	35845	GRC
Greece 2011	1056607	10	54	1	19357	GRC
Guatemala 1964	210079	5	23	1*	6355	GTM
Guatemala 1973	289446	5.5	23	5	13315	GTM
Guatemala 1981	302106	5	23	5	11950	GTM
Guatemala 1994	833137	10	23	5	14992	GTM
Guatemala 2002	1121946	10	23	5	29031	GTM

Honduras 1974	278348	10	18	5	18079	HND
Honduras 1988	423971	10	18	5	18158	HND
Honduras 2001	608620	10	18	5	23585	HND
Haiti 1971	434869	10	4	1*	6693	HTI
Haiti 1982	128770	2.5	4	5	4389	HTI
Haiti 2003	838045	10	4	5	9438	HTI
Indonesia 1971	634642	0.54	27	1*	13351	IDN
Indonesia 1976	281170	0.22	27	5	5514	IDN
Indonesia 1980	7234577	5	27	5	173977	IDN
Indonesia 1985	605858	0.37	27	5	11937	IDN
Indonesia 1990	912544	0.51	27	5	27238	IDN
Indonesia 1995	718837	0.43	27	5	17489	IDN
Indonesia 2000	20112539	10	27	5	510538	IDN
Indonesia 2005	1090892	0.51	27	5	19286	IDN
Indonesia 2010	23603049	10	27	5	501808	IDN
Ireland 1981	344291	10	8	1	8271	IRL
Ireland 1986	355020	10	8	1	6343	IRL
Ireland 1991	353149	10	8	1	8652	IRL
Ireland 1996	365323	10	8	1	51987	IRL
Ireland 2002	410688	10	8	1	61677	IRL
Ireland 2006	440314	10	8	1	70644	IRL
Ireland 2011	474535	10	8	1	66725	IRL
Iraq 1997	1944278	10	18	10	1936272	IRQ
Israel 1983	403474	10	6	5	53032	ISR
Jamaica 1982	223668	10	14	1*	6413	JAM
Jamaica 1991	232625	10	14	1*	1476	JAM
Jamaica 2001	205179	10	14	1*	7437	JAM
Kenya 1979	1033769	6.7	8	1	37709	KEN
Kenya 1989	1074098	5	8	1	40008	KEN
Kenya 1999	1407547	5	8	1	62769	KEN
Kenya 2009	3841935	10	8	1	85157	KEN
Kyrgyz Republic 1999	476886	10	8	1*	14031	KGZ
Cambodia 1998	1141254	10	22	1*	1973	KHM
Cambodia 2004	102558	0.8	22	1*	1517	KHM
Cambodia 2008	1340121	10	22	1*	32087	KHM
Cambodia 2013	134964	0.9	22	1*	1881	KHM
Laos 2005	560480	10	18	10	165429	LAO

Morocco 2004	1482720	5	16	5	58795	MAR
Mexico 1960	502800	1.5	32	1*	7067	MEX
Mexico 1970	483405	1	32	1*	2802	MEX
Mexico 1990	8118242	10	32	5	358418	MEX
Mexico 1995	332061	0.4	32	5	14786	MEX
Mexico 2000	10099182	10.6	32	5	383343	MEX
Mexico 2005	10284550	10	32	5	339701	MEX
Mexico 2010	11938402	10	32	5	392571	MEX
Mexico 2015	11344365	9.5	32	5	298546	MEX
Mali 1998	991330	10	8	1*	9247	MLI
Mali 2009	1451856	10	8	1*	28410	MLI
Mongolia 2000	243725	10	21	5	19227	MNG
Mozambique 1997	1551517	10	11	1	29873	MOZ
Mozambique 2007	2047048	10	11	1	28368	MOZ
Malawi 2008	1341977	10	26	1*	82306	MWI
Malaysia 1991	347892	2	13	5	26890	MYS
Malaysia 2000	435300	2	13	5	21503	MYS
Nicaragua 1971	189469	10	13	5	23946	NIC
Nicaragua 1995	435728	10	13	5	59477	NIC
Nicaragua 2005	515485	10	13	5	66878	NIC
Nepal 2001	2067609	9	14	5	28746	NPL
Nepal 2011	3238842	12	14	5	75814	NPL
Panama 1960	53553	5	7	1*	1016	PAN
Panama 1980	195577	10	7	1*	4078	PAN
Peru 2007	2745895	10	25	5	154049	PER
Philippines 1990	6013913	10	76	1*	95429	PHL
Philippines 2000	7417810	10	76	5	266145	PHL
Philippines 2010	9411256	10	76	5	189869	PHL
Papua New Guinea 1980	296165	10	20	5	15161	PNG
Papua New Guinea 1990	359720	10	20	1	11918	PNG
Poland 2002	3824056	10	16	1	14148	POL
Portugal 1981	492289	5	20	1	42704	PRT
Portugal 1991	491755	5	20	1	39638	PRT
Portugal 2001	517026	5	20	1	43182	PRT
Portugal 2011	528870	5	20	1	164564	PRT
Paraguay 1972	233669	10	13	5	21341	PRY

Paraguay 1982	301582	10	13	5	35084	PRY
Paraguay 1992	415401	10	13	5	37312	PRY
Paraguay 2002	516083	10	13	5	37631	PRY
Romania 1977	1937021	10	39	1*	19238	ROU
Romania 1992	2238578	10	39	1*	10512	ROU
Romania 2002	2137967	10	39	1*	17960	ROU
Sudan 2008	5066530	16.6	15	1	68488	SDN
Senegal 1988	700199	10	11	5	29637	SEN
Senegal 2002	994562	10	11	5	39452	SEN
El Salvador 1992	510760	10	14	1*	10847	SLV
El Salvador 2007	574364	10	14	1*	6300	SLV
South Sudan 2008	542765	7	10	1	16739	SSD
Slovenia 2002	179632	10	12	1*	1944	SVN
Togo 2010	584859	10	3	1*	14261	TGO
Thailand 1970	772169	2	68	1*	28957	THA
Thailand 1980	388141	1	68	1*	5571	THA
Thailand 1990	485100	1	68	1*	9974	THA
Thailand 2000	604519	1	68	1*	9464	THA
Trinidad and Tobago 1990	113104	10	4	1	40498	TTO
Trinidad and Tobago 2000	111833	10	4	1	2908	TTO
Trinidad and Tobago 2011	116917	8.8	4	10	2154	TTO
Tanzania 1988	2310424	10	23	10	930169	TZA
Tanzania 2002	3732735	10	23	1	108109	TZA
Tanzania 2012	4498022	10	23	1	130410	TZA
Uruguay 1975	279994	10	19	5	18325	URY
Uruguay 1985	295915	10	19	5	23658	URY
Uruguay 1996	315920	10	19	5	23285	URY
Uruguay 2006	256866	8.4	19	5	2129	URY
Uruguay 2011	328425	10	19	5	14888	URY
United States 1970	2029666	1	51	5	198148	USA
United States 1980	11343120	5	51	5	609722	USA
United States 1990	12501046	5	51	5	1271083	USA
United States 2000	14081466	5	51	5	1422087	USA
United States 2005	2878380	1	51	1	76087	USA
United States 2010	3061692	1	51	1	74003	USA
United States 2015	3147005	1	51	1	88985	USA

Venezuela 1981	1441266	10	22	1*	53048	VEN
Vietnam 1989	2626985	5	38	5	126894	VNM
Vietnam 1999	2368167	3	38	5	56469	VNM
Vietnam 2009	14177590	15	38	5	345730	VNM
South Africa 2001	3725655	10	4	5	93301	ZAF
South Africa 2007	1047657	2	4	1*	8694	ZAF
South Africa 2011	4418594	8.6	4	1*	72660	ZAF
Zambia 1990	787461	10	8	1	19380	ZMB
Zambia 2000	996117	10	8	1	32076	ZMB
Zambia 2010	1321973	10	8	1	21554	ZMB
Zimbabwe 2012	654688	5	10	10	283158	ZWE

CHAPTER TWO

STAY OR GO?

DATA AND EMPIRICAL STRATEGY

As Box 2.1 describes, the net migration measure comes from the Global Estimated Net Migration Grids By Decade, v1 (1970-2000) (de Sherbinin et al., 2015). It provides estimates of net migration (in-migration minus out-migration) per one-kilometer grid cell for three decades, 1970s, 1980s and 1990s. The unit of the net migration measure in the original dataset is the net change of the number of people due to migration per square km. To arrive at net migration estimates, the compilation begins with a gridded distribution of population in the year 2000 based on census data drawn from the Global Rural–Urban Mapping Project, Version 1. The History Database of the Global Environment, Version 3.1, is then used to calculate population totals in 1970, 1980, and 1990 to arrive at population growth estimates for previous decades. Mortality rates specific to each nation, ethnicity group, and decade are applied to each grid cell to estimate decennial births and deaths. Finally, a measure for net population growth, defined as births minus deaths plus net migration, is used to estimate the net migration for each grid cell. We collapse the highly disaggregated observations to the 0.5 x 0.5 degree resolution. The original observations are aggregated by taking means. As a result, the unit of our net migration measure after aggregation is the number of people (due to migration) per square kilometer in a 0.5 degree gridcell. de Sherbinin et al. (2015) note that there could be measurement errors at a very local-level. Following Peri and Sasahara (2020), we aggregate the observations to a coarser resolution to mitigate these concerns.

Rainfall variability is measured in terms of local deviations from the long-run mean using weather data from Matsuura and Willmott (2018). This gridded dataset contains monthly observations of precipitation and average temperature at the 0.5 degree gridcell level. We transform this data into average monthly temperature, and total precipitation (mm), per year, for each gridcell. To define rainfall shocks, we calculate the long run mean and standard deviation of annual precipitation for each gridcell. We then define a positive (negative) shock in a given year/gridcell if annual precipitation in that year/gridcell is at least 1 standard deviation higher (lower) than the long run mean for that gridcell.

In the analysis, the rainfall shock variables measure the *number* of years in the decade for which rainfall was at least 1 standard deviation above or below the mean. The analysis focuses on the impact of repeated water shocks that occur over a decade on out-migration rates since a single rainfall episode may be misleading, as the decision to migrate often entails high cost and longer-term considerations.

The main empirical strategy is predicated on the fact that rainfall shocks are exogenous and consequently unpredictable with respect to net migration rates. The model tests how net migration rates differ in response to cumulative number of rainfall shocks for each decade in the dataset from 1970 to 2000. To do so, the following equation is estimated:

$$Y_{it} = \alpha_1 + \alpha_2 Prec10_{it}^{-} + \alpha_3 Prec10_{it}^{+} + + X_{it}'\lambda + f_c(t) + \theta_t + \gamma_i + \varepsilon_{it}$$

where *i* indicates a grid cell, *t* indicates years, and *c* indicates countries.

Following Peri and Sasahara (2019), the outcome variable Y_{it} is the net migration rate from year *t-10* to year *t* measured as:

$$\frac{NetMig_{it}}{Pop_{it} - NetMig_{it}}$$

where $NetMig_{it}$ denotes net migration in gridcell i for the period between year *t-10 and* year *t*. Dividing by the initial population provides a net migration rate: the percentage change in population due to mobility. $Prec10_{it}^{-}(Prec10_{it}^{+})$ is the number of negative (positive) rainfall shocks greater than 1 standard deviation within the last 10 years, $f_c(t)$ are country-specific time trends, θ_t are year fixed effects, and γ_i are gridcell fixed effects. X_{it} is a vector of control variables which includes an interaction between initial gridcell population and a linear time trend. This allows for differential trends by initial population. Also included is a control for mean annual temperature (°C) within the last 10 years. Although we do not focus on the impact of temperature, we do control for it, in order to obtain unbiased estimates of the effects of changes in precipitation. α_2 and α_3 are our coefficients of interest and measure how repeated negative or positive rainfall shocks, respectively, can impact the percentage change

in population in a gridcell. By using this set up, we are measuring how cumulative rainfall shocks, presumed to be exogenous and unpredictable, lead to a change in net migration rates that is over and above the change caused by trending growth patterns, and time invariant factors of the gridcell itself (which will thus also account for changes in policy at the regional and national levels).

Weighting the grid-level results by grid population would allow more weight on shocks in heavily populated areas. To eliminate excessive emphasis on large urban areas, the results estimate unweighted regressions, allowing each grid to contribute equally to the analysis. Standard errors are clustered at the level of a grid-cell and at the level of admin level 1 in robustness checks to account for spatial and serial correlation.

To understand the extent to which the relation between rainfall shocks and net migration rates is modulated by agricultural dependence of locations, we identify high-cropland gridcells. Gridcells whose share of cropland is greater than the 95th percentile of the distribution among all grid cells in each country are identified as high-cropland gridcells. This term is interacted with rainfall shocks to investigate the differential effect of rainfall shocks on net migration.

Similarly, to understand the extent to which the relation between rainfall shocks and net migration rates is modulated by baseline income, countries are classified into four groups based on the baseline GDP per capita in 1980: those in [0*th*,25*th*), [25*th*,50*th*), [50*th*,75*th*), and [75*th*,100*th*) percentile bins. The regression then introduces additional interaction terms between rainfall shocks and baseline income levels to study the differential effect of rainfall shocks by income. When presenting the results, the first two bins are combined to overcome sample size issues.

In order to assess the differential impacts of migration in places with low and high access to irrigation and forests, additional data sources are used. The analysis uses data on share of irrigated cropland at the start of the period over which migration occurs (Siebert et al. 2015), and forest cover data from the European Space Agency. These are used to construct baseline time-invariant shares of irrigated cropland and shares of forest cover for the sample of grid cells. High irrigation access and high forest access indicator variables are constructed to represent grid cells where the baseline share of irrigated cropland or share of forested area is above the global median. The analysis follows a similar methodology and adds an additional interaction term between rainfall shocks and irrigation access or rainfall shocks and forest access. The coefficients on the interactions show the extent to which irrigation and forest access attenuate the migration response. Furthermore, the analysis also examines the differential buffering impact of irrigation access in arid regions engaged in water-intensive cropping practices. The analysis combines climatic zone data along with data on the geographical distribution of agricultural crops from Monfreda et al. (2008).

RESULTS

The estimates underlying Figure 2.2a in the report is displayed in column (3) of table A2.1. The estimates in Column(1) represent the extent to which dry rainfall shocks have driven migration across the global sample.

Table A2.1: Impact of rainfall shocks on net migration rates, by Agricultural Dependence

	Dependent variable= Net-Migration rates= In-migration- Out-migration			
	1	2	3	4
	S.Es clustered at Grid	S.Es clustered at Admin 1	S.Es clustered at Grid	S.Es clustered at Admin 1
#1SD NegShock t-10	-0.3426***	-0.3426***	-0.3081***	-0.3081**
	(0.044)	(0.124)	(0.046)	(0.122)
#1SD PosShock t-10	-0.0789*	-0.0789	-0.0780*	-0.0780
	(0.044)	(0.090)	(0.045)	(0.090)
#NegShock X High-Cropland			-0.7247***	-0.7247**
			(0.180)	(0.335)
#PosShock X High-Cropland			0.1003	0.1003
			(0.184)	(0.280)
Observations	94275	94275	94114	94114
Adj. Rsq	0.707	0.707	0.708	0.708
RMSE	13.310	13.310	13.262	13.262
Rainfall Shock effects (linear combination of coefficients)				
NegShock for High-Cropland			-1.0328***	-1.0328***
			(0.175)	(0.362)
PosShock for High-Cropland			0.0224	0.0224
			(0.180)	(0.283)
grid FEs	y	y	y	y
Year FEs	y	y	y	y
CountryTrend	y	y	y	y
PopTrend	y	y	y	y
Temperature control	Y	Y	Y	Y

Note: Dependent variable is the decadal net migration rate over 3 decades from 1970 to 2000. Standard errors reported in parentheses are robust and clustered at the gridcell level (columns 1 and 3) or at the province-level (columns 2 and 4). Number of Positive (Negative) shocks in 10 years is a time-varying count variable indicating the number years, of the prior ten years, for which annual rainfall in the gridcell was at least 1 standard deviation higher (lower) than the long run mean of the gridcell. Statistical significance is given by *p<0.10 ** p<0.05 *** p <0.01

Since net-in-migration rates are defined as in-migration minus out-migration, a negative coefficient implies an increase in out-migration. For the figures in the report, these were changed to net-out-migrations rates to ease interpretation.

The estimates underlying Figure 2.2b in the report is displayed in table A2.2. The estimates in Column(1) represent the extent to which dry and wet rainfall shocks have driven migration across the global sample for each income-bin. The estimates in Column(2) represent the extent to which dry and wet rainfall shocks have driven migration across the rural sample for

Table A2.2: Impact of rainfall shocks on net migration rates, by baseline income and agricultural dependence

	(1)	(2)
	Full-sample	**High-Cropland sample**
#1SD NegShock t-10		
(0-50th] percentile	-0.1009	-0.3093
	(0.077)	(0.251)
(50-75th] percentile	-0.3428***	-0.8119*
	(0.084)	(0.419)
(75-100th] percentile	-0.6178***	-1.1605***
	(0.063)	(0.279)
#1SD PosShock t-10		
(0-50th] percentile	-0.3862***	-0.6430*
	(0.082)	(0.337)
(50-75th] percentile	-0.0811	-0.5222
	(0.087)	(0.384)
(75-100th] percentile	0.0930*	0.2249
	(0.054)	(0.230)
Observations	94275	5282
grid FEs	y	y
Year FEs	y	y
CountryTrend	y	y
PopTrend	y	y
Adj. Rsq	0.707	0.748
RMSE	13.306	12.707

Note: Dependent variable is the decadal net migration rate over 3 decades from 1970 to 2000. Standard errors reported in parentheses are robust and clustered at the gridcell level. Number of Positive (Negative) shocks in 10 years is a time-varying count variable indicating the number years, of the prior ten years, for which annual rainfall in the gridcell was at least 1 standard deviation higher (lower) than the long run mean of the gridcell. Gridcells whose share of cropland is greater than the 95th percentile of the distribution among all grid cells in each country are identified as high-cropland gridcells. Statistical significance is given by *p<0.10 ** p<0.05 *** p <0.01

each income-bin. Since net-in-migration rates are defined as in-migration minus out-migration, a negative coefficient implies an increase in out-migration.

The estimates underlying Figure 2.3 in the report are displayed in table A2.3 in Columns (1) and (2). The estimates in Column (1) represent the extent to which the impact of dry and wet rainfall shocks are modulated by high irrigation access. The estimates in Column (2) represent the extent

Table A2.3: Joint impact of rainfall shocks, forest access and irrigation on net migration rates

	Dependent variable= Net-Migration rates= In-migration- Out-migration				
	1	2	3	4	5
#1SD NegShock t-10	-0.4026***	-0.3575***	-0.4007***	-0.5545***	-0.6467***
	(0.087)	(0.071)	(0.078)	(0.096)	(0.104)
#1SD PosShock t-10	-0.2080**	-0.1525**	-0.2083**	-0.2715**	-0.3580***
	(0.090)	(0.077)	(0.081)	(0.109)	(0.115)
#NegShock X High-irrigation	0.3011***			0.3531***	0.4015***
	(0.111)			(0.109)	(0.108)
#NegShock X High-forest		0.1837*		0.2542**	
		(0.109)		(0.107)	
#NegShock X Share forest			0.3559*		0.5184***
			(0.190)		(0.189)
#PosShock X High-irrigation	0.2033*			0.2268*	0.2658**
	(0.114)			(0.117)	(0.118)
#PosShock X High-forest		0.0644		0.1152	
		(0.120)		(0.123)	
#PosShock X Share forest			0.2311		0.3355*
			(0.183)		(0.190)
Observations	61475	61475	61475	61475	61475
grid FEs	y	y	y	y	y
Year FEs	y	y	y	y	y
CountryTrend	y	y	y	y	y
PopTrend	y	y	y	y	y
Adj. Rsq	0.636	0.636	0.636	0.636	0.637
RMSE	14.752	14.753	14.752	14.751	14.750

Note: Dependent variable is the decadal net migration rate over 3 decades from 1970 to 2000. Standard errors reported in parentheses are robust and clustered at the gridcell level. Number of Positive (Negative) shocks in 10 years is a time-varying count variable indicating the number years, of the prior ten years, for which annual rainfall in the gridcell was at least 1 standard deviation higher (lower) than the long run mean of the gridcell. High irrigation access and high forest access refers to those grid cells where the baseline share of irrigated cropland or share of forested area is above the global median. Statistical significance is given by *p<0.10 ** p<0.05 *** p <0.01

Table A2.4: Unit cost of irrigation expansion by region (2005 USD per hectare)

EAP	9,236
SAR	3,812
FSU (Former Soviet Union)/ECA	9,237
SSA	16,226
MEN	9,748
LAC	5,512

Source: Inocencio et al. (2007)

to which the impact of dry and wet rainfall shocks are modulated by high forest access. Column (3) also shows the extent to which the impact of dry and wet rainfall shocks are modulated by forest access but uses a continuous share in place of the indicator variable.

The estimates in Column (4) and (5) in Table A2.3 represent a regression model that includes interactions of rainfall shocks with both irrigation and forests together. The estimates in Column (5) are combined with summary measures of forest loss and irrigation costs to conduct back-of-the-envelope calculations to measure the extent of irrigation that would be needed to compensate for the drought-buffering effects of lost forest. To quantify the cost of irrigation expansion, data from Inocencio et al. (2007) as seen in table A2.4 are used. These are the most widely used estimates on the costs of irrigation expansion at the global level and provide unit costs of irrigation expansion per hectare by world region (for recent applications, see Rozenberg and Fay 2019; Rosegrant et al. 2017).

REFERENCES

de Sherbinin, Alex, M. Levy, S. Adamo, K. MacManus, G. Yetman, V. Mara, L. Razafindrazay, B. Goodrich, T. Srebotnjak, C. Aichele, and L. Pistolesi. 2015. *Global Estimated Net Migration Grids by Decade: 1970–2000*. Palisades, NY, United States of America: NASA Socioeconomic Data and Applications Center (SEDAC).

Inocencio, A., M. Kikuchi, M. Tonosaki, A. Maruyama, D. Merrey, H. Sally, and I. de Jong. 2007. *Cost of Performance of Irrigation Projects: A Comparison of Sub-Saharan Africa and Other Developing Regions*. Colombo, Sri Lanka: International Water Management Institute.

Matsuura, K., & Willmott, C. J. (2018). Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900–2017), http://climate.geog.udel.edu/~climate/html_pages/download.html

Monfreda, Chad, Navin Ramankutty, and Jonathan A. Foley. 2008. "Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000." *Global biogeochemical cycles* 22(1) doi:10.1029/2007GB002947

Peri, Giovanni, and Akira Sasahara. 2019. *The Impact of Global Warming on Rural-Urban Migrations: Evidence from Global Big Data*. Cambridge, MA, United States of America: National Bureau of Economic Research.

Rosegrant, M. W., T. B. Sulser, D. Mason-D'Croz, N. Cenacchi, A. Nin-Pratt, S. Dunston, T. Zhu, C. Ringler, K. D. Wiebe, S. Robinson, D. Willenbockel, H. Xie, H. Y. Kwon, T. Johnson, T. S. Thomas, F. Wimmer, R. Schaldach, G. C. Nelson, and B. Willaarts. 2017. *Quantitative Foresight Modeling to Inform the CGIAR Research Portfolio*. Washington, DC: International Food Policy Research Institute.

Rozenberg, J., and M. Fay, eds. 2019. *Beyond the Gap: How Countries Can Afford the Infrastructure They Need While Protecting the Planet*. Washington, DC: World Bank.

Siebert, S., M. Kummu, M. Porkka, P. Döll, N. Ramankutty, and B. R. Scanlon. 2015. "A Global Data Set of the Extent of Irrigated Land from 1900 to 2005." *Hydrology and Earth System Sciences* 19 (3): 1521–45.

CHAPTER THREE

WATER, MIGRATION, AND HUMAN CAPITAL SPILLOVERS

DATA

The data analyzed in Box 1 of chapter 3 in the main report utilizes data on internal migrants from 403 subnational regions from across 21 developing countries assembled by Gennaioli et al. (2014), who assemble data on the economic and demographic characteristics of native and migrant populations originating from each of these regions as documented in the most recent census available prior to 2010. Since the census data covers individuals that are within the country, this data covers internal migrants only. Importantly, these data contain separately estimates of the education levels of migrants and natives from these regions. Migrants are found to have higher human capital than natives for their place of origin, and on average have 1 year of additional schooling. The data collected by authors also contain regional data on GDP per capita, population and urbanization, which serve as important control variables in the analysis. The list of countries is given in Table A3.1.

This data, which are mainly available at the first subnational administrative region (usually province or state), are matched to subnational admin region boundary files available from the University of Minnesota's IPUMS project, and then matched to climate data from Center for Climatic Research of the

Table A3.1: Sample of Regions in Cross Sectional Analysis

Country	Regions	Year of Census Data
Chile	10	2002
Colombia	23	2005
Ecuador	20	2001
El Salvador	7	2007
Indonesia	24	2010
Kenya	5	1999
Kyrgyz Republic	6	1999
Malaysia	11	2000
Mexico	32	2010
Mongolia	20	2000
Nepal	5	2001
Panama	8	2000
Peru	23	2007
Philippines	4	1990
Romania	38	2002
South Africa	4	2007
Tanzania	16	2002
Thailand	55	2000
Turkey	55	2000
Uruguay	17	1996
Venezuela	20	1990

University of Delaware (Matsuura and Willmott, 2018). The climate data is available in a gridded format, with data on annual average temperature and total rainfall available for each 50 x 50 km gridcell between 1915 and 2017. For each province, the climate variables are averaged, first by taking the mean across all grid-cells that have their center contained within the province for each year, and taking the mean of these province-level averages over the entire annual period.

This yields a measure of the average annual rainfall and temperature experienced in each of these provinces. Additionally, the presence of primate cities is inferred by overlaying the province level boundaries over a spatial database of global urban footprint boundaries from Khan et al (2019) and identifying the regions which contain the top 3 cities within a country. The geographic coverage of the sample, with administrative boundaries of the regions used, are depicted on the map in Map A3.1.

For the second set of results for Mexico, Brazil and Indonesia, data on internal urban migrants in each of these countries is extracted from census data available from the University of Minnesota's IPUMS project. In particular, the analysis is restricted to the subsample of male individuals aged between 15 and 65 in each census year and surveyed in urban areas. Each of these countries has data from three recent censuses available,

corresponding to the decades of 1990, 2000 and 2010. These countries are chosen in particular because their censuses ask the respondents to disclose their place of residence 5 years prior to the census, thus allowing for the identification of their migration status and place of origin. The availability of data from multiple census rounds and migrant's place of origin allows the analysis to exploit panel data techniques. For each of these countries, individuals are defined as being "high-skilled" if they have completed middle school, or in other words have completed the 8th grade. Map A3.2 shows the administrative boundaries of the subnational regions from Mexico, Brazil and Indonesia for which census data were used.

This census data for the 3 countries is again augmented with climate data from the University of Delaware (Matsuura and Willmott, 2018). To exploit the time dimension from the multiple census rounds, the province-level climate data is not averaged over time. For each province, the rainfall for each year are first averaged spatially, by taking the mean across all grid-cells that have their center contained within the province. Then for each year, rainfall shocks are defined based on deviations from the long term mean rainfall. Specifically, a negative (positive) rainfall shock is defined as occurring in a given province in a certain year if rainfall in that province is 1 standard deviation lower (higher) than mean rainfall, where mean and standard-deviation are calculated as based on annual rainfall between 1915 and 2017 for all provinces in the country. For each urban migrant observed in a census, the number of positive and negative rainfall shocks occurring in

MAP A3.1: Regions used in the cross-sectional analysis

Note: The map shows 403 subnational regions from across 21 developing countries for which demographic and economic characteristics were analyzed in conjunction with climate data. Socioeconomic data were obtained from Gennaioli et al. (2014) and augmented with climate data from Matsuura and Willmott (2018).

MAP A3.2: Regions used in the cross-sectional analysis

The map shows the subnational regions of Mexico, Brazil and Indonesia. Data on internal urban migrants from these regions observed in the censuses of 1990, 2000 and 2010 was combined with climate data from Matsuura and Willmott (2018) to examine how the education levels of migrants moving in response to rainfall shocks differed from other migrants. It is found that migrants moving in response to frequent dry-shocks are less likely to have attended high-school.

their place of origin in the 5 years prior to the earliest possible date of arrival in the host region.

METHODOLOGY & RESULTS

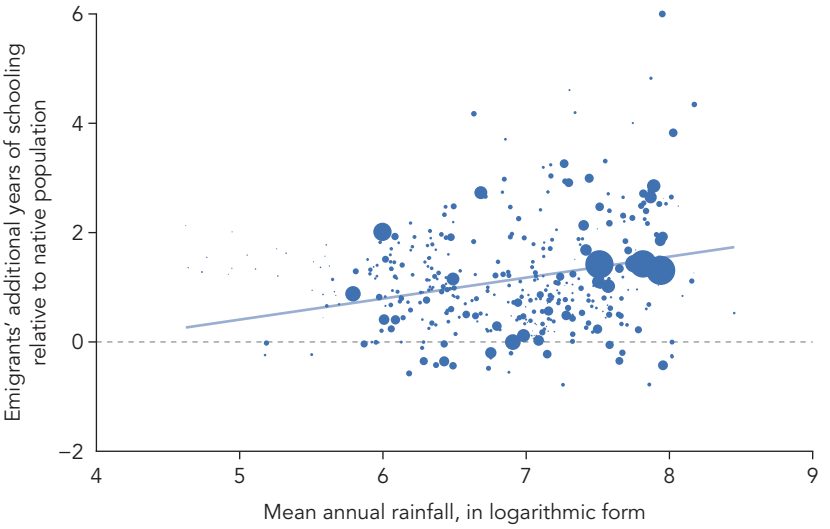
The first result builds on the cross-sectional relationship between average rainfall conditions in a region and the skill levels of migrants that originate from region. As shown in Figure A3.2, a strong cross-sectional correlation is observed between the average rainfall - emigrants from regions with higher rainfall have higher education levels compared to natives from that region.

To control empirical test this relationship, and account for possible confounding vraiables, the relationship is tested using the below regression specification:

$$EmigrantsSchooling_{ij} = \alpha + \beta \ln(MeanRainf\,all_{ij}) + \lambda\, NativesSchooling_{ij} + \gamma \boldsymbol{X}_{ij} + \delta_j + \varepsilon_{ij}$$

For each region I in country *j*, the dependent variable is the years of schooling completed by the average emigrants originating from that region. The main explanatory variable of interest is the natural logarithm of average annual rainfall in that region. The regression specification needs to control for the education levels of natives in the regions, since the overall human capital investment in a region is directly related to the education of migrants leaving that region and can drive the result. Additionally, the specification also controls for country fixed effects δ_j and a set of regional characteristics contained in the control vecto in X_{ij}, which include regional measures of urbanization, GDP, temperature and population. The regression is estimated using regional

FIGURE A3.2: Rainfall and Migrants' Education



Note: Figure A3.1 shows that (internal) migrants originating from regions with higher average rainfall levels tend to have higher years of schooling compared to the natives of their place of origin. Each bubble represents a subnational region, with the size of the bubble proportional to the population of that region, and only within-country migrants are considered in this analysis.

domestic population weights to make it representative for the average individual in the underlying data and robust standard errors are estimated to account for heteroskedasticity in the data.

The results are presented below in table A3.2. The first column presents the relationship between rainfall endowment and the education of emigrants leaving a region is robust to a regression controlling for natives education, and the second column additionally includes country fixed effects. A positive and statistically significant relationship is found to exist between rainfall and emigrants education levels. The next columns sequentially add controls for urbanization, logarithm GDP, temperature and population. The coefficient remains robust to these additional controls. A coefficient of 0.3 means that a doubling of rainfall (which equals the interquartile range of rainfall in this sample) is associated with 0.3 additional years of schooling for migrants.

Additionally, the data allows for the exploration of local characteristics that may mediate this relationship are explored through heterogeneities. To this end an important regional characteristics that can interact with the climate to affect the skills of workers migrating from a region are the income level of a country. In low income settings, workers may face a binding cost constraint (or a lack of urban opportunities due to low structural transformation) that breaks down the relationship between climate and education. Similarly, the presence of large urban centers such as primate cities within a region would reduce incentives for educated workers to move out of a region, for instance if they simply choose to move to the city within their home region.

Table A3.2: Global Cross-Section

	Emigrant Skill and Rainfall					
Dependent Variable	**Average Years of Education of Emigrants**					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
ln(Mean Annual Rainfall 1900-1990)	0.347***	0.293*	0.375**	0.375**	0.293*	0.337**
	(0.119)	(0.164)	(0.170)	(0.174)	(0.172)	(0.162)
Years of Education of Natives	0.854***	0.847***	0.746***	0.720***	0.776***	0.782***
	(0.0383)	(0.0608)	(0.0821)	(0.125)	(0.138)	(0.123)
Urbanization Rate in 1990			0.714***	0.686***	0.632***	0.590***
			(0.252)	(0.219)	(0.221)	(0.213)
ln(GDP per capita)				0.108	0.0585	0.129
				(0.232)	(0.244)	(0.226)
Mean Annual Temperature (1900-1990)					0.0470***	0.0379**
					(0.0168)	(0.0192)
ln(Population)						-0.241***
						(0.0814)
Country FE	No	Yes	Yes	Yes	Yes	Yes
Observations	403	403	403	403	403	403
R-Squared	0.711	0.816	0.822	0.823	0.841	0.841
Adjusted R-Square	0.71	0.805	0.811	0.811	0.83	0.83

Note: Observation corresponds to subnational regions from across 21 countries. Estimates are from regressions weighted using domestic population of the region, with robust standard errors are presented in parenthesis.

Table A3.3 explores these heterogeneities. Columns 1 shows that the relationship between rainfall and migrants' education levels is driven by the middle-income country sample, by using an interaction term with an indicator variable to identify the countries that belong to the low-income group. Column 2 shows the interaction term for regions without primates, term with a dummy variable to identify regions containing the largest three cities within any given country. For both characteristics, the interaction term is negative, statistically significant, and large enough to reverse the positive sign on the coefficient on rainfall. Columns 3 and 4 show the coefficient as estimated for the sample after excluding these regions.

The preferred specification in column 4 uses regressions weighted by the size of the region's population, to account for the precision of estimation using regional aggregates and make them representative of the average migrant in our sample. Following suggestions by Solon et al. (2015), the results are also confirmed to be robust to the use of alternative weights. Table A3.4 shows the robustness of these results to alternative weighting

Table A3.3: Global Cross-Section

Emigrant Skill and Rainfall – Regional Heterogeneities				
Dependent Variable	**Average Years of Education of Emigrants All Regions**		**Excluding Primate Regions**	**Excluding Low Income**
	(1)	**(2)**	**(3)**	**(4)**
ln(Mean Annual Rainfall)	0.474*** (0.165)	0.408** (0.162)	0.429** (0.179)	0.429*** (0.164)
ln(Mean Annual Rainfall) x Primate Region	-0.619*** (0.172)			
ln(Mean Annual Rainfall) x Low Income		-1.229*** (0.423)		
Country FE	No	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	403	403	352	351
R-Squared	0.85	0.844	0.847	0.825
Adjusted R-Square	0.839	0.833	0.835	0.813

Each observation corresponds to one subnational regions from across 21 countries. Estimates are from regressions weighted using domestic population of the region, with robust standard errors are presented in parenthesis. All columns contain controls for urbanization rate, log of GDP, mean annual temperature and log of population.

approaches. Column 1 reproduces the estimates shown in column 4 of table A3.3, which uses domestic population as regression weights. Column 2 shows the coefficient remains positive and statistically significant when using the stock of emigrants to weight the regression, or no weighting at all.

In addition to the cross-sectional analysis, data from multiple census rounds from Mexico, Brazil and Indonesia were also employed to test through panel-data approach whether rainfall shocks in the region of origin result in less educated urban migrants arriving from that region. To do so, for each urban migrant observed in a given census the number of rainfall shocks that occurred in their home region during the five years prior to the earliest possible year of migration. Then the below regression is used to test whether more frequent rainfall shocks affect the likelihood of an urban migrant being high-school educated.

$$SkillDummy_{ijky} = \beta_1(NumberNegShocks_{jy}) + \beta_2(NumberPosShocks_{jk}) + \Gamma\ Age_{ijky} + \alpha_j \delta_{ky} + \varepsilon_{ijk}$$

The dependent variable here is a dummy variable equal to one if middle school (i.e. grade 8) was completed by individual *i*, from province of origin *j*, observed in destination *k*, in census year *y*. A set of age dummies is included in

Table A3.4: Global Cross-Section

Emigrant Skill and Rainfall – Alternative Weighting Approaches			
Dependent Variable	**Average Years of Education of Emigrants**		
	(1)	**(2)**	**(3)**
ln(Mean Annual Rainfall 1900-1990)	0.429*** (0.164)	0.347*** (0.110)	0.232* (0.121)
Weights	Domestic population	Stock of Emigrants	None
Country FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes
Observations	351	351	351
R-Squared	0.825	0.899	0.848
Adjusted R-Square	0.813	0.893	0.838

Each observation corresponds to a subnational regions from the 16 middle-income countries in the full sample. Estimates are from regressions weighted as indicated, with robust standard errors are presented in parenthesis.

the regression specification as well as origin fixed effects. The specification also controls for destination fixed effect, census-year fixed effects and the interaction of these two. This allows for a flexible specification of labor market trends across different regions of a country. Standard error are clustered at the origin-year level, as this is the unit of variation over which rainfall shocks are exploited.

The results are presented in Table A3.5 for the three countries. For each country, the first column shows estimates with controls for age, survey-year and origin fixed effects, and the next two columns then add on destination fixed effects and then the interaction of these with the survey-year dummies. A robust negative coefficient is consistently found to exist for negative rainfall shocks across all the countries, but not for positive shocks across all countries. This confirms that the relationship between rainfall and migrant's education earlier observed in the cross-section is robust to estimation using panel data, where rainfall is measured using frequent and large deviations from the average rainfall over time. The estimates suggest, each additional negative rainfall shock in the five years preceding migration is associated with a 1 to 2 percentage point reduction in the probability of an urban migrant being "high-skilled". Relative to typical migrants moving to urban areas in these countries, migrants escaping regions that receive negative rainfall shocks bring lower levels of human capital, as proxied by their education levels.

Table A3.5: Rainfall Shocks at Origin and Urban Migrant Skill

Dependent Variable	High-Skill Dummy (above median migrant's education level)								
	Brazil			Mexico			Indonesia		
	1	2	3	4	5	6	7	8	9
# of Neg Shocks at origin over last 5 years	-0.0186***	-0.0200***	-0.0164**	-0.0190***	-0.0197***	-0.0213***	-0.0112***	-0.0123***	-0.00852**
	-0.00632	-0.00595	-0.00684	-0.00685	-0.00727	-0.00744	-0.0029	-0.00328	-0.00394
# of Pos Shocks at origin over last 5 years	-0.00296	-0.00289	-0.00907	-0.000173	0.000671	-0.00714	-0.00471	-0.0132***	-0.0101*
	-0.0177	-0.0161	-0.0115	-0.00651	-0.00677	-0.00752	-0.00604	-0.00449	-0.00541
Observations	546676	546676	546676	262206	262206	262206	296683	296683	296683
Adjusted R-sq	0.118	0.126	0.128	0.084	0.095	0.099	0.141	0.16	0.162
Mean of Dependent variable	0.57	0.57	0.57	0.64	0.64	0.64	0.63	0.63	0.63
Survey Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Origin FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Destination FE	N	Y	Y	N	Y	Y	N	Y	Y
Destination X Year FE	N	N	Y	N	N	Y	N	N	Y

All regressions include Standard errors in parentheses clustered at the origin-year level. + p<0.15, * p<0.1, ** p<0.05, *** p<0.01

REFERENCES

Gennaioli, N., La Porta, R., De Silanes, F.L. and Shleifer, A., 2014. Growth in regions. *Journal of Economic growth*, *19* (3), pp.259-309.

Khan, A., Selod, H., & Blankespoor, B. (2019). *The Two Tails of Cities. A (More) Exhaustive Perspective on Urban Population Growth and City Spatial Expansion*. Technical Report.

Matsuura, K. and Willmott, C. J., 2018. Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900–2017), http://climate.geog.udel.edu/~climate/html_pages/download.html

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What are we weighting for?. *Journal of Human resources*, *50* (2), 301-316.

CHAPTER FOUR

THE COST OF DAY-ZERO EVENTS

EMPIRICAL STRATEGY

The main empirical strategy is predicated on the fact that rainfall, and therefore water supply shocks are exogenous and consequently unpredictable with respect to urban growth. Thus, the model tests how the growth of NTLs, a proxy for economic activity, changes in years following large water supply shocks. To do so, equation 1 is estimated:

$$\Delta\log(NTL_{it}) = \alpha_1 + \alpha_2 Water\ Shock_{it-1,t-3} + X_{it}'\lambda + f_c(t) + \theta_t + \gamma_i + \varepsilon_{it} \quad (4.1)$$

where i indicates urban areas, t indicates years, and c indicates countries. The outcome variable $\Delta\log(Y_{it})$ is the first difference of log mean luminosity (i.e. the growth rate) inside the urban area. $Water\ Shock_{it}$ is an indicator of a water supply shock for urban area i from years t-1 to t-3. X_{it} is a vector of control variables which include $\Delta\log(Population_{it})$ (i.e. the growth rate of population), measures of contemporaneous weather (precipitation and temperature) in the urban area itself, and a measure of temperature in the water supply regions. $f_c(t)$ are country-specific time trends, θ_t are year fixed effects, and γ_i are urban area fixed effects. By using

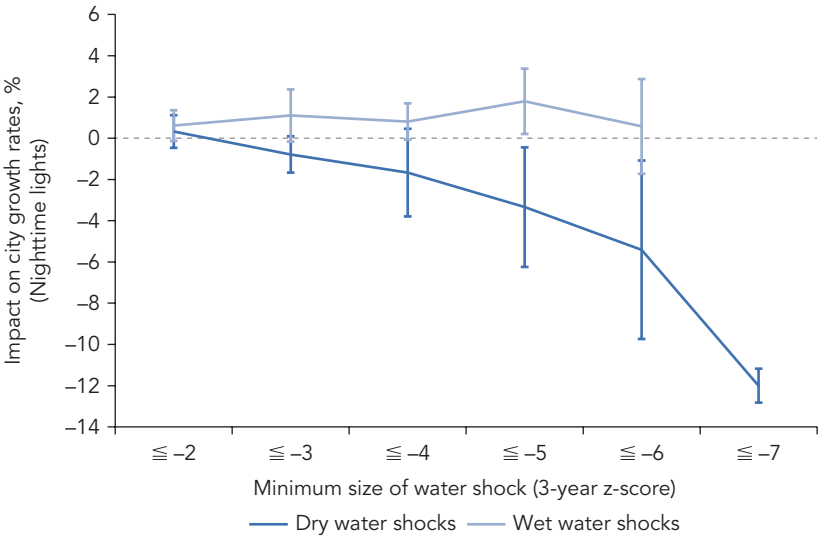
this set up, we are measuring how water supply shocks, presumed to be exogenous and unpredictable, lead to a change in economic growth that is over and above the change caused by trending growth patterns, time invariant factors of the city itself (which will thus also account for changes in policy at the regional and national levels), population growth. In addition, we control for contemporaneous weather changes in the city itself which is critical given the spatial autocorrelation in weather, and the fact that weather in the city has been shown to impact productivity (Sudarshan and Tewari 2014; Graff Zivin and Neidell 2014; Desbureaux and Rodella 2019).

To measure a water supply shock we use the 3-year cumulative z-score of precipitation in the region of the water supply locations, as supplied by TNC and McDonald (*2016*). The 3-year time period was chosen, as a drought that occurs over that time period would generally exceed the length of time that urban water supplies are designed to be resilient against. The benefits of using the cumulative z-score, rather than, for instance, a dummy indicator of drought over that time period, are three-fold. First, it allows an unusually dry year to be canceled out by an unusually wet year. This is important as water storage facilities can be replenished after droughts. Second, it allows for one to distinguish between relatively mild droughts and much deeper droughts to estimate heterogeneous impacts. Finally, it allows one to measure the intensity of the drought along two dimensions—in depth, as well as over time—as a very harsh but short drought can be just as likely to lead to a water supply shock as a relatively milder but longer drought.

RESULTS

The results for estimating equation 4.1 are shown in Figure A4.1 (identical to figure 4.1 in the main report) which plots the coefficients from 6 different regressions which are identical except for the cutoff for shock size in the water point area. As one moves from the left to right, the cutoff increases by 1 standard deviation. The results show that positive rainfall shocks in the water point areas have positive, but muted and statistically noisy impacts. The results for negative shocks, however, are much clearer, with an increasingly large impact as the shock size increases. For instance, when the 3 year cumulative shock is more than 3 standard deviations below the long run mean, luminosity growth falls by 0.8%. This increases to 1.7%, 3.3%, 5.4%, and 12.0% as the minimum shock size increases to 4, 5, 6, and 7 standard deviations, respectively.

FIGURE A4.1: Impact of Water Supply Shocks on City Growth Rates



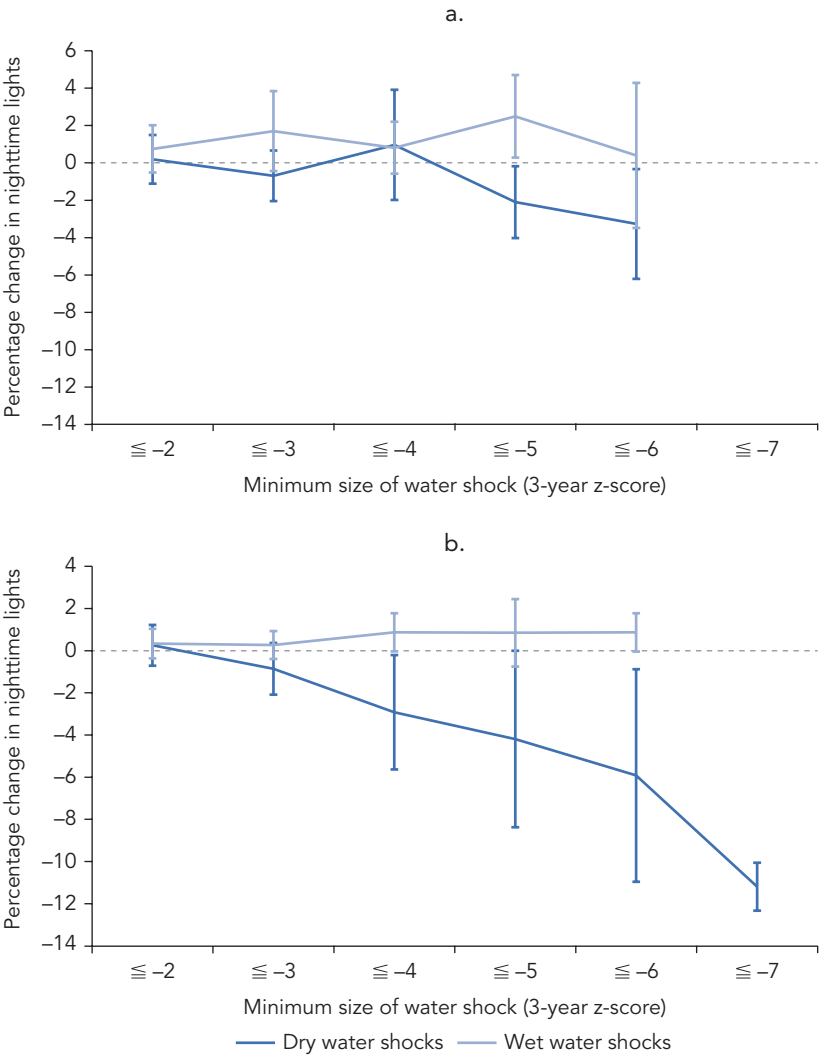
Source: World Bank figure based on analysis using weather data from Matsuura and Willmott (2018); Nighttime Lights Time Series Version 4, from NOAA National Centers for Environmental Information, Earth Observation Group; and data on urban water sources from The Nature Conservancy and McDonald (2016)
Note: Figure shows point estimates with 95% confidence intervals from regressing equation 4.1 six different times, each time increasing the minimum threshold for a water shock, as one moves from left to right.

Next, identical regressions to those run in Figure A4.1 are tested, but with a split sample. Figure A4.2 shows the sample split by average rainfall in the city, with the panel A including cities below the median rainfall level for the sample, and the panel B including cities above the median rainfall level.

Next, the procedure is repeated but the sample is split based on the population size of the city. Figure A4.3 shows the sample split by average population in the city, with the panel A including cities below the median population level for the sample, and the panel B including cities above the median population level.

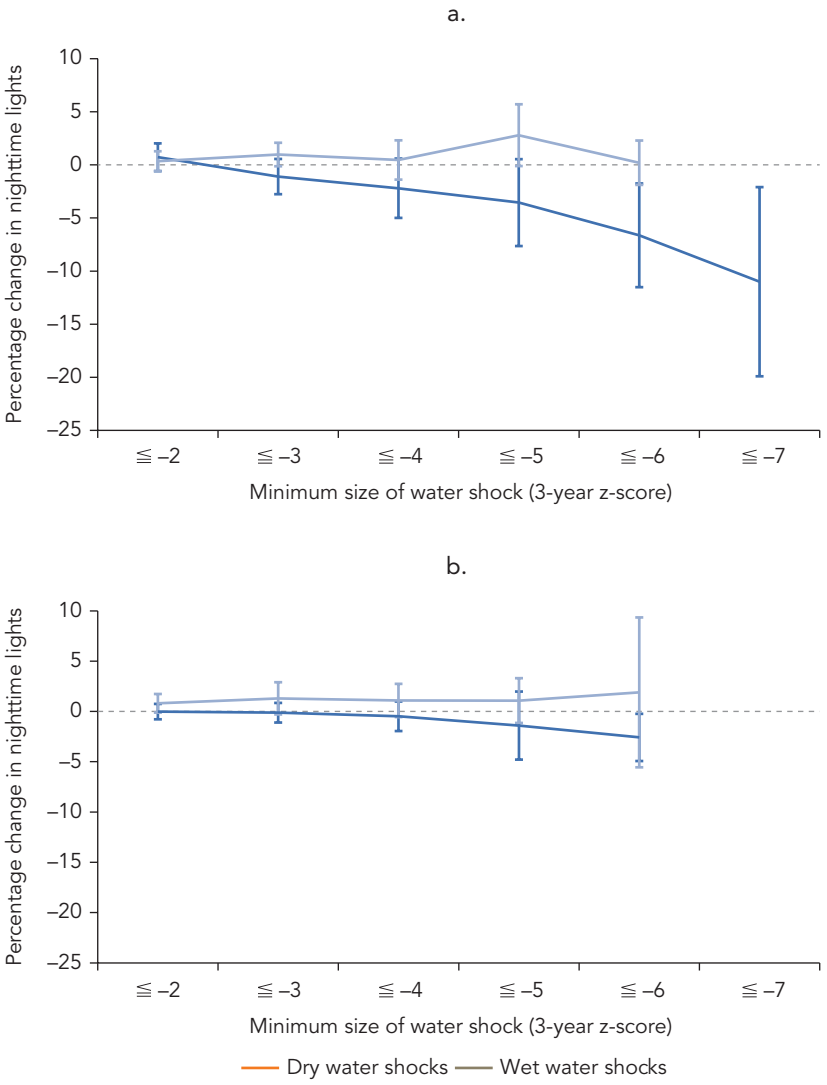
Finally, to test if the methodology employed here is misattributing the estimated impacts to water supply shocks, a placebo test is conducted. In this test, the sample is restricted to cities which do not have surface water supply points, but instead rely on desalination, groundwater, or other forms of water supply. Results from this test are in figure A4.4.

FIGURE A4.2: Impact of Water Supply Shocks on Urban Luminosity Growth Rate, by Climate



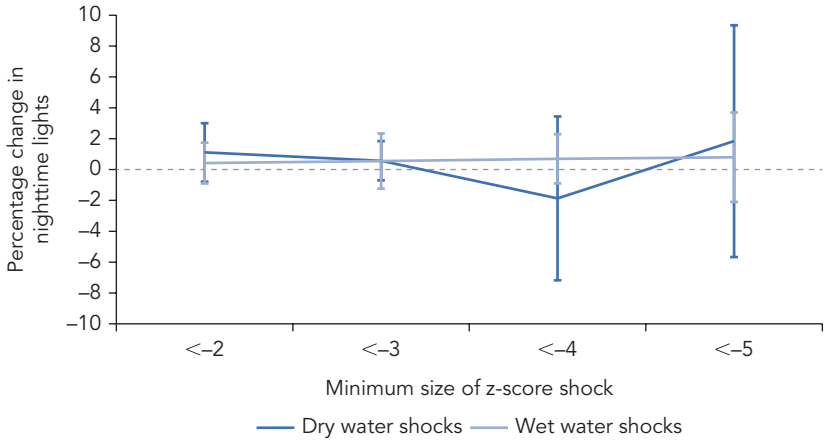
Source: World Bank figure based on analysis using weather data from Matsuura and Willmott (2018); Nighttime Lights Time Series Version 4, from NOAA National Centers for Environmental Information, Earth Observation Group; and data on urban water sources from The Nature Conservancy and McDonald (2016)
Note: Panel A show results for cities in the bottom 50th percentile of long-run mean rainfall, and panel B shows results for cities in the top 50th percentile of long-run mean rainfall.

FIGURE A4.3: Impact of Water Supply Shocks on Urban Luminosity Growth Rate, by City Population Size



Source: World Bank figure based on analysis using weather data from Matsuura and Willmott (2018); Nighttime Lights Time Series Version 4, from NOAA National Centers for Environmental Information, Earth Observation Group; and data on urban water sources from The Nature Conservancy and McDonald (2016)
Note: Panel A show results for cities in the bottom 50th percentile in terms of population size, and panel B shows results for cities in the top 50th percentile in terms of population size.

FIGURE A4.4: Impact of Weather at Nonsurface Urban Water Points on Urban Luminosity Growth Rate, Placebo Test



Source: World Bank figure based on analysis using weather data from Matsuura and Willmott (2018); Nighttime Lights Time Series Version 4, from NOAA National Centers for Environmental Information, Earth Observation Group; and data on urban water sources from The Nature Conservancy and McDonald (2016).
Note: Figure shows results of estimating the impact of weather at urban water supply points for cities that do not have surface water-based water supply points.

REFERENCES

Desbureaux, S., & Rodella, A. S. (2017). Shocks in the Cities: The Economic Impact of Water Shocks in Latin American Metropolitan Areas. Technical Background Paper of *Uncharted Water: The New Economics of Water Scarcity and Variability*

Graff Zivin, J., & Neidell, M. (2014). Temperature and the allocation of time: Implications for climate change. *Journal of Labor Economics*, *32*(1), 1–26.

Matsuura, K., & Willmott, C. J. (2018). Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900–2017), http://climate.geog.udel.edu/~climate/html_pages/download.html

Sudarshan, A., & Tewari, M. (2014). *The economic impacts of temperature on industrial productivity: Evidence from indian manufacturing* (No. 278). Working paper.

The Nature Conservancy and Robert McDonald. 2016. "City Water Map (version 2.2). KNB Data Repository. doi:10.5063/F1J67DWR." Accessed through Resource Watch, (date). www.resourcewatch.org

CHAPTER FIVE

GOING WITH THE FLOW

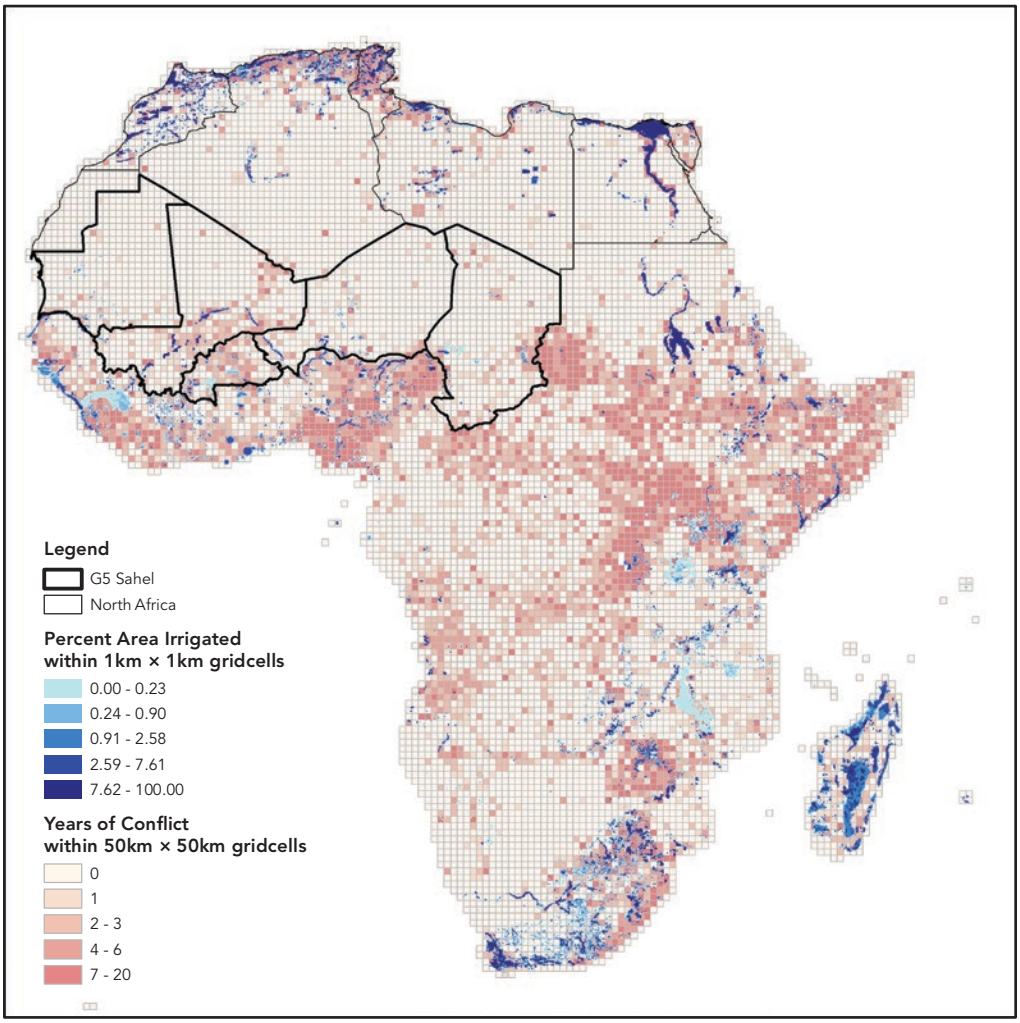
DATA

This section presents data utilized in the empirical examination of conflict breakout in the G5 Sahel countries and countries in Northern Africa, before and after the Arab Spring (circa 2011), building on the approach taken in by Khan and Rodella (forthcoming). Data on conflict events occurring between 1997 and 2011 is compiled from the Armed Conflict Location and Event Dataset (ACLED), which records a wide range of conflict events such as protests, battles and rebel activities derived from war zone media reports, humanitarian agencies and research publications. As is standard in the conflict literature, grid-cells (i) are coded as a dummy equal to 1 if a grid-cell experienced any conflict event in a given year (t).

Figure A5.2 plots the spatial distribution of the frequency of battle events in Africa at the 50km x 50 km resolution, which is the resolution of the main analysis in section 5.2. Additionally, to measure the share of each grid-cell that is irrigated data at the start of the analysis period gridded, gridded data on the average area equipped for irrigation between 1990 and 2005 is taken from the FAO. The construction of this data set is described in Siebert et al (2015). Figure A2 also plots this data at the 1km x 1km resolution at which the raw data is available. To conduct the analysis, this irrigation data is aggregated to the 50km x 50km by taking the average of percentage area irrigated in all smaller cells that fall within a larger grid-cell.

In addition to these main data, the analysis detailed in section 5.2 also employs various other time-invariant grid-cell level characteristics to control for possible confounds of climate and irrigation availability that may also be correlated with conflict. These include: the share of area with cropland and pastureland taken from Ramankutty et al (2008); population density in the year 2000 derived from the Gridded Population of the World dataset

MAP A5.1: Conflict trends in Irrigated and Non irrigated regions in Africa



Source: Figure A5.2 shows Irrigated Area from FAO (pre-2005), and conflict events data from Armed Conflict & Event Data Project (ACLED), https://www.acleddata.com, for the whole continent of Africa.
Note: The five G5 Sahel countries are Burkina Faso, Chad, Mali, Mauritania, and Niger. North Africa refers to Algeria, Djibouti, Arab Republic of Egypt, Libya, Morocco, and Tunisia.

(GPW v4); the presence of wetlands from the Global Lakes and Wetlands Database (GWLD v3) from WWF; and the presence of ethnic homelands of ethnicities whose historic economic activity is identified as being agriculture (see (Michalopoulos and Papaioannou, 2013) for data description).

METHODOLOGY AND RESULTS

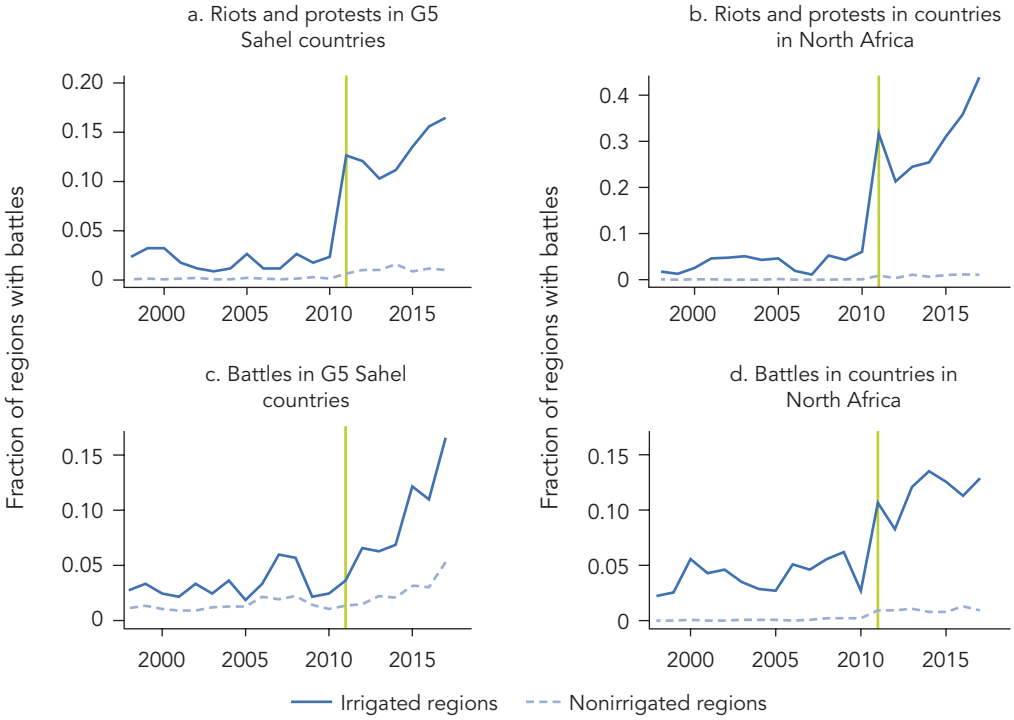
A difference-in-difference approach is taken to compare the trends in conflict occurrence between irrigated and unirrigated areas. The approach used here builds on evidence compiled by Khan and Rodella (forthcoming), who conduct a deeper analysis comparing the countries of the G5 Sahel and the rest of West Africa. Grid cells are classified as irrigated if they contain any irrigated land, and the trends in conflict occurrence in the irrigated and non-irrigated grid cells is presented in Figure 5.2 in the main report, also presented in Figure A5.1 below. The differential trends suggest that when compared to non-irrigated regions irrigated areas in the G5 Sahel and MENA regions experienced a much higher likelihood of conflict events following the exogenous increase in fragility in the aftermath of the Arab Spring. To focus the analysis on fragility arising from violent political clashes over control of irrigated territory, the rest of this section focuses on a narrower definition of conflict which captures only the occurrence of "battles" in ACLED, defined as "a violent interaction between two political organized armed groups at a particular time and location", with a grid-cell i being coded as a dummy equal to one if a battle occurred in year t.

The presence of irrigation, however, may be correlated with various characteristics such as the presence of crops, local climate conditions, population density, which may be driving the observed differential trends in Figure A5.2. To control for these, the relationship between conflict and the presence of irrigation during the post-Arab Spring period of heightened fragility is next examined in a difference-in-difference framework through an empirical specification that that controls for covariates of irrigation that might be influencing the observed differential trends. Specifically, the following regression is estimated:

$$Conflict_{ict} = \beta \ln(irrigarea_i + 1) \times Post_t + \gamma_t \boldsymbol{X_i} + \omega \boldsymbol{W_{it}} + \delta_{ct} + \alpha_i + \in_{ict}$$

where the outcome of interest $conflict_{ict}$ is an indicator variable that equals to 1 if a battle event occurs in grid-cell i in country c in a given year t. The variable $irrigarea_i$ is a time-invariant measure of the area in each grid-cell i that was irrigated in the initial period, while $Post_t$ takes is a dummy variable that takes on a value of 1 if t refers to the years after 2011.[1] The coefficient β captures the differential trend in conflict between cells with high-irrigation and low-irrigation grid-cells in the post-Arab spring period. Additionally, $\boldsymbol{X_i}$ is a set of time-invariant cell-specific characteristics that may be correlated with the presence of irrigation, and includes log of population at the start of

FIGURE A5.2: Conflict trend in Irrigated and Non irrigated regions, before and after the Arab Spring



Source: World Bank calculations, based on data on irrigated areas from Siebert et al. (2015), and conflict data from Armed Conflict & Event Data Project (ACLED), https://www.acleddata.com. Calculations are based on the approach from Khan and Rodella (forthcoming).
Note: Figure A5.1 shows the share of grid cells in North Africa (MENA) and G5-Sahel countries that experienced different types of conflict events, separated by the presence of irrigation. This figure is similar to Figure 5.2 in the main report. The five G5 Sahel countries are Burkina Faso, Chad, Mali, Mauritania, and Niger. North Africa refers to Algeria, Djibouti, Arab Republic of Egypt, Libya, Morocco, and Tunisia.

the period as well as dummies to indicate the presence of wetlands, croplands, pasture lands, rivers and the homelands of ethnic groups that historically relied on agricultural livelihood strategies. γ_t allows for the effects of these characteristics to vary in flexible form over the period of the analysis. W_{it} is a vector of time-varying controls for rainfall and temperature to control for differences in climate over time and across grid-cells. Lastly δ_{ct} is a set of country-year fixed effects, and α_i a set of grid-cell fixed effects that account for average differences in other unobserved characteristics.

The results are presented in Table A5.1 separately for the G5 Sahel countries, MENA countries, and the Rest of Sub-Saharan Africa. For the G5 Sahel and MENA countries, a statistically significant and large positive increase in the likelihood of conflict breaking out in grid-cells where more irrigated land is observed and unchanged by multiple controls. Additional checks of the data also confirm the existence of parallel trends in the period up until 2011. In the post-Arab spring period, there is an increase in the

TABLE A5.1: Conflict and Rainfall Shocks

Dependent Variable:	Dummy variable equal to one if any battle event occurred in a year.								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	G5 Sahel			Rest of SSA			North Africa (MENA)		
Ln(Share of Cell Irrigated) X Post	0.0582***	0.0510***	0.0481***	0.00483	-0.00513	-0.00391	0.0504***	0.0292***	0.0293***
	(0.0108)	(0.0112)	(0.0111)	(0.00322)	(0.00344)	(0.00352)	(0.00651)	(0.00683)	(0.00722)
Observations	34160	34160	34160	127800	127580	127580	40560	40540	40540
R-sq	0.226	0.233	0.233	0.369	0.372	0.372	0.389	0.403	0.403
Country-Year FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Cell FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls X Time		Y	Y		Y	Y		Y	Y
Climate Controls			Y			Y			Y
No Of Clusters/Cells	1708	1708	1708	6379	6379	6379	2027	2027	2027
Mean	0.018	0.018	0.018	0.064	0.064	0.064	0.0234	0.0234	0.0234

Standard errors in parentheses clustered at the gridcell level. * p<0.1, ** p<0.05, *** p<0.01"
Note: The five G5 Sahel countries are Burkina Faso, Chad, Mali, Mauritania, and Niger. North Africa refers to Algeria, Djibouti, Arab Republic of Egypt, Libya, Morocco, and Tunisia. Time-invariant cell-specific characteristics included in controls are log of population at the start of the period, as well as dummies to indicate the presence of wetlands, croplands, pasture lands, rivers and the homelands of ethnic groups that historically relied on agricultural livelihood strategies.

likelihood of conflict where higher irrigation is observed specifically in the G5 Sahel and MENA countries, but the not in the rest of Western Africa. This provides further support to the hypothesis that highly-irrigated land in the G5 Sahel, relative to non-irrigated land in the same region, is more prone to conflict during periods of heightened fragility.

NOTE

1. The use of $ln(irrigarea_i+1)$ instead of $ln(irrigarea_i)$ is to ensure that gridcells with zero irrigation are not dropped from the analysis. Alternate specifications such as $ln(irrigarea_i+0.1)$, $ln(irrigarea_i+0.01)$, the use of inverse-hyperbolic transformations, or the inclusion of a simple dummy to indicate the presence of any irrigation also provide similar results.

REFERENCES

Khan, A., and A.-S. Rodella. Forthcoming. "A Hard Rain's a-Gonna Fall: New Insights on Water Security and Fragility in the Sahel". Washington, DC: World Bank.

Michalopoulos, S. and Papaioannou, E., 2013. Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, *81*(1), pp.113-152.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., & Scanlon, B. R. (2015). A global data set of the extent of irrigated land from 1900 to 2005. Hydrology and Earth System Sciences, 19(3), 1521-1545.

ECO-AUDIT

Environmental Benefits Statement

The World Bank Group is committed to reducing its environmental footprint. In support of this commitment, we leverage electronic publishing options and print-on-demand technology, which is located in regional hubs worldwide. Together, these initiatives enable print runs to be lowered and shipping distances decreased, resulting in reduced paper consumption, chemical use, greenhouse gas emissions, and waste.

We follow the recommended standards for paper use set by the Green Press Initiative. The majority of our books are printed on Forest Stewardship Council (FSC)–certified paper, with nearly all containing 50–100 percent recycled content. The recycled fiber in our book paper is either unbleached or bleached using totally chlorine-free (TCF), processed chlorine–free (PCF), or enhanced elemental chlorine–free (EECF) processes.

More information about the Bank's environmental philosophy can be found at http://www.worldbank.org/corporateresponsibility.

GWSP
GLOBAL WATER
SECURITY & SANITATION
PARTNERSHIP

WORLD BANK GROUP

SKU W21032