ENVIRONMENTAL RESEARCH
LETTERS

ACCEPTED MANUSCRIPT • OPEN ACCESS

Global migration is driven by the complex interplay between environmental and social factors

To cite this article before publication: Venla Niva *et al* 2021 *Environ. Res. Lett.* in press https://doi.org/10.1088/1748-9326/ac2e86

Global migration is driven by the complex interplay between environmental and social factors

Niva, Venla [1]; Kallio, Marko [1,2]; Muttarak, Raya [3]; Taka, Maija [1]; Varis, Olli [1]; Kummu, Matti [1]

[1] Water and Environmental Engineering Research Group, School of Engineering, P.O. Box 15200, FI-00076, Aalto University, Finland

[2] Geoinformatics Research Group, School of Engineering, P.O. Box 11000, FI-00076, Aalto University, Finland

[3] Population and Just Societies Program, International Institute for Applied Systems Analysis (IIASA), Austria

Key words: human migration, adaptive capacity, drivers of migration, environmental stress, integrated approach, random forest analysis, grid cell

Abstract

Migration manifests an important response and adaptation measure to changes in the environment and socioeconomic conditions. In a time when environmental stressors and risks are unprecedentedly increasing, understanding the interplay between the underlying factors driving migration is of high importance. While the relationships between environmental and socioeconomic drivers have been identified conceptually, the comprehensive global-scale spatial quantification of their interactions is in its infancy. Here, we performed a geospatial analysis of gridded global net migration from 1990 to 2000 using a novel machine learning approach which analyzes the interplay between a set of societal and environmental factors simultaneously at the place of origins (areas of net-negative migration) and destinations (areas of net-positive migration). We diagnosed the importance of eight environmental and societal factors in explaining migration for each country, globally. Nearly half of global in- and out-migration took place in the areas characterized by low adaptive capacity and high environmental stress. Regardless of the income level, income was the key factor in explaining net-migration in half of the countries. Slow-onset environmental factors, drought and water risk, were found to be the dominant environmental variables globally. Our study highlights that factors representing human capacity need to be incorporated into the quantitative diagnosis of environmental migration more rigorously.

1 Introduction

Recent events such as migrant caravans from Central America to the United States in 2019, the Venezuelan migrant and refugee crisis in 2019-20 and the 2015 crisis of large refugee flows from the Middle East and North Africa to Europe have been frequently linked with preceding severe drought episodes in the country of origin (Chemnick 2019, Gustin and Henninger 2019, Markham 2019, Podesta 2019). Indeed, a stereotypical view that environmental change would induce mass-migration fluxes towards the "Global North" has been repeated in both research and policy-making for decades (Boas *et al* 2019). The empirical evidence supporting such claims however is inconsistent (Abel *et al* 2019, Selby *et al* 2017). Accordingly, investigating the fundamental, manifold role of environmental stress as a trigger and driver of migration has substantially gained both scholarly and public attention. Not only do various environmental factors influence migration in different directions and magnitudes (see e.g. Gray and Mueller 2012, Cattaneo and Peri 2016, Kubik and Maurel 2016), other societal factors and their interactions also play an important role. The understanding of human migration therefore needs to account for complex interactions between different drivers of migration at the micro, meso and macro levels (Boas *et al* 2019, Abel *et al* 2019, Borderon *et al* 2019, Hoffmann *et al* 2021).

A traditional gravity-based 'push-pull' model has often been used to identify the macro-level factors underlying migration decisions by analyzing spatial disparities between the place of origin (as pushing factors) and destination (presumably more attractive conditions, i.e. pulling factors) (de Haas 2011, Lee 1966). Despite their conceptual clarity, the push-pull model has been criticized for its simple assumption on the linear relationship between environmental change and migration dynamics (Jónsson 2010). The literature is dominated by the assumptions that environmental changes are the primary pushing factors that linearly lead to migration whereas in reality individuals and households employ diverse responses to environmental shocks based on their social, economic, demographic and political capital (Nelson *et al* 2007). Environmental stress thus may influence migration through affecting other migration drivers such as through exacerbating conflict, reducing agricultural production and income change (Abel *et al* 2019, Beine and Parsons 2015). On the other hand, migration is a costly process and people with little social and economic resources generally have lower capacity to move, thus the majority of migration is internal or between low- and middle-income countries (Hoffmann *et al* 2020). This non-linear pattern follows the prediction of the migration hump theory which holds that migration has an inverted U-shaped relationship with socioeconomic development (Martin and Taylor 1996). International migration hence is low in low income and the least developed countries because their populations cannot afford to emigrate.

Establishing the relationship between environmental change and migration response requires a comprehensive account of all other factors and contextual effects which could determine the migration-environment association (Borderon *et al* 2019). One commonly used approach for coupling the societal and environmental dimensions in studying migration on a conceptual level is introduced by Black et al. (2011b) and the Foresight report on Migration and Global Environmental Change (2011). Their approach depicts migration through a relationship between dimensions of human capacity and vulnerability to environmental change (Figure 1) and thus combines objective circumstances with subjective perceptions influencing migration. In addition to addressing vulnerability to environmental change, their widely used conceptual framework incorporates a diversity of psychosocial and socioeconomic factors (e.g. education, income, individual's intentions and cultural identity) that influence people's mobility-decisions and capacities to move. Failing to account for socioeconomic drivers and their interplay with other factors in influencing migration can provide a biased estimate of the role of environmental change and stressors.

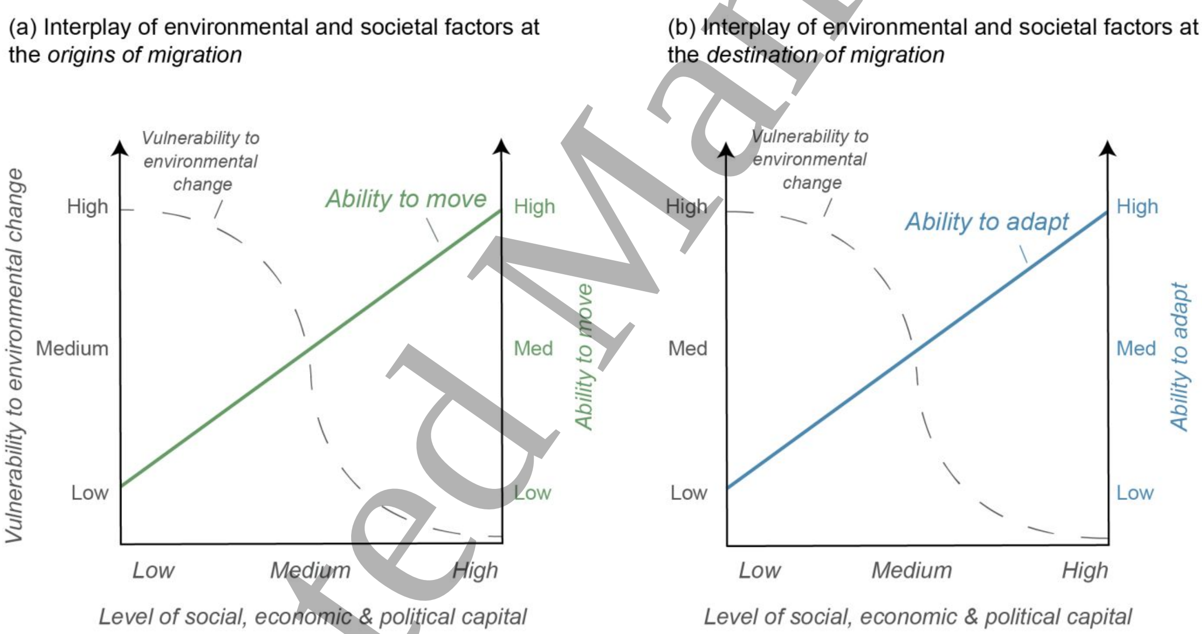

Figure 1. A typology characterizing (a) the interplay between adaptive capacity and vulnerability to environmental change underlying ability to migrate from the area of origin, and (b) to adapt to the destination. The dashed line illustrates how the vulnerability to environmental change depends on the level of social, economic and political capital; i.e. when the capacity is high, the vulnerability to environmental change and thus the ability or desire to move/adapt are low. Adapted from Foresight report on Migration and Global Environmental Change (2011) and Black et al (2011b).

There are, however, only few studies that provide *quantitative* global assessments of the interplay between societal and environmental factors underlying human migration. Marotzke *et al* (2020) and

Lilleør and Van den Broeck (2011) explored the poverty-climate-migration nexus in a laboratory setting considering only economic factors in less developed countries. De Sherbinin *et al* (2012) and Neumann *et al* (2015) studied global spatial patterns of environmentally induced migration but excluded socio-economic drivers from their analysis. Studies which include both environmental change and socioeconomic factors are mainly regional ones (see e.g. Wiederkehr *et al* (2018) on Sub-Saharan Africa and Kluger *et al* (2020) for Peru). Furthermore, studies on environmentally induced migration typically focus on the place of origin and their characteristics while much less attention is paid to conditions in the destination areas (Ayeb-Karlsson *et al* 2020, Findlay 2011), despite the fact that societal and environmental factors also reflect the ability of the destination area to absorb (or attract) migrants (Niva *et al* 2019). For policy planning, it is highly relevant to identify where environmentally induced migrants may move to, as well as to understand the characteristics of both the origins and destinations in order to assess migrants' vulnerability at both ends of migration. Moreover, quantitative global assessments of migration can be directly incorporated into other modelling frameworks such as the Integrated Assessment Models (IAMs) which are designed to describe key interactions between physical and social systems. Changes in drivers of migration would influence migration patterns and consequently population size, income distribution and emissions (Liang *et al* 2020, Benveniste *et al* 2021). The quantitative assessment of environmental and socioeconomic drivers of global migration thus can substantially improves our understanding of future socioeconomic development which can have considerable implications on the global climate system.

We address these gaps by providing a global quantitative assessment of 1) the interplay of environmental-societal characteristics in both sending (negative net-migration) and receiving (positive net-migration) areas globally, and 2) the importance of different environmental and socio-economic indicators underlying net-negative and net-positive migration by utilizing a machine learning method (random forests). This paper thus contributes to the current migration research by studying both out- and in-migration locations simultaneously by utilizing spatially explicit global data sets covering a range of relevant environmental, socio-economic and demographic indicators (see Table 1) as well as gridded net-migration data (de Sherbinin *et al* 2012). Furthermore, the use of random forests to quantitatively define the nexus between environmental change, socioeconomic factors and migration on a global scale is novel in the field. The number of international and internal migrants is constantly growing with rapidly changing environment around the globe (Xu *et al* 2020). It is thus of prime scholarly and policy importance to understand the characteristics and interplay of both environmental and societal factors behind human migration.

2 Materials and Methods

All analyses were conducted globally on 5 arc-minute resolution grid cell level (Figure 2, Table 1). For the random forest analysis, individual models for net-negative and net-positive were created for 178 countries in total, i.e., each model is based on the grid cells of the country in question (n varies from 1, in very small countries such as Vatican City or Gibraltar, to 3435160 cells in Russia, global median 4447 cells). Models were used to study the importance of each variable in explaining net-positive and net-negative migration, i.e. which variable had the highest explanatory power on the response variable. Feature importance distributions of each variable are illustrated for 12 groups based on the United Nations (UN) geoscheme (Statistics Division of the United Nations Secretariat 2021). Country classification is presented in Supplementary materials (Table S2).

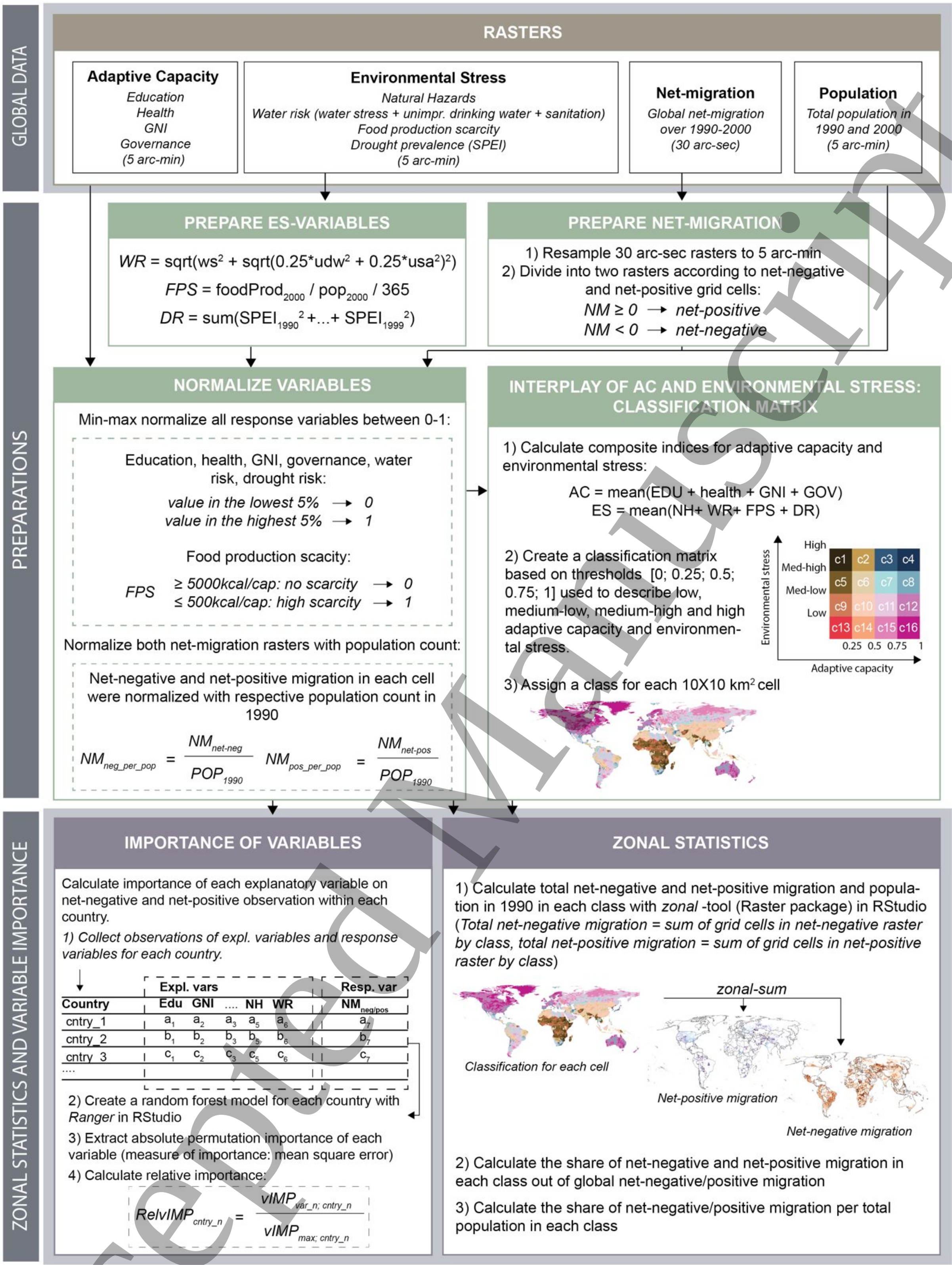

Figure 2. Overview of the workflow illustrating the general structure of the analysis. For more detailed information regarding the data and workflow, see Section 2, Table 1 and the Supplements. AC = adaptive capacity; GNI = Gross National Income; EDU = education; GOV = governance; ES = environmental stress; NH = natural hazards; WR = water risk (ws = water stress, udw =

unimproved drinking water, usa = unimproved sanitation); FPS = food production scarcity; DR = drought risk; NM = net migration; POP = population, RelvIMP = Relative variable importance, vIMP = variable importance.

Table 1. Description of data and their sources. See more detail explanation in Table S1.

Data	**Abbreviation**	**Description (year, resolution)**	**Source**
		Adaptive capacity	
Income level	GNI	1990-2000; 5 arcmin res. Values downscaled from sub-national to grid level (see SI for more details). Gross National Income.	Based on Smits and Permanyer (2019), missing values interpolated and extrapolated using method from Kummu et al (2018). Downscaled to grid level based on night lights and agricultural land use, using linear multiple regression model.
Education	EDU	1990-2000; 5 arcmin res. Gridded subnational data. Combined Mean years of schooling and Expected years of schooling.	Based on Smits and Permanyer (2019), missing values interpolated and extrapolated using method from Kummu et al (2018).
Health	Health	1990-2000; 5 arcmin res. Gridded subnational data. Measured as life expectancy at birth.	Based on Smits and Permanyer (2019), missing values interpolated and extrapolated using method from Kummu et al (2018).
Governance	GOV	1990-2000; National data resampled to 5 arcmin res. World Governance Index for government effectiveness.	Varis *et al* (2019a) adapted from WGI (2018)
		Environmental stress	
Natural hazards	NH	1990-2000; 2.5 arc-min gridded data resampled to 5 arc-min res. Multiple hazard index.	Varis *et al* (2019a) adapted from Dilley et al (2005)
Drought risk	DR	1990-2000; 1° gridded data resampled to 5 arc-min res. Measured as Standardized Precipitation-Evapotranspiration Index, SPEI.	Calculated from (Vicente-Serrano *et al* 2010)
Food production scarcity	FPS	2000; 5arcmin res. Food production per capita per day (kcal/capita/day).	Annual food production data (kcal) from (Mueller *et al* 2012) and population data from Klein Goldewijk *et al* (2010).
Water risk	WR	Compiled from ws and udw & usa (see below). WR was calculated so that it combines quantitative risk (water stress) and qualitative risk (drinking water and sanitation coverage) as follows: 1) two components of qualitative risk are first combined by taking a root of their summed squares, 2) qualitative risk is combined to quantitative risk by taking a root of their summed squares.	
Baseline water stress	ws	1990-2000; 5 arc-min res. Gridded Hydrological sub-basin (HydroBASINS 6) data. Use to availability ratio	(Hofste *et al* 2019)

		reported as risk levels between 1-5 (1: low, 5: extremely high)	
Unimproved/No Drinking water & Unimproved/No Sanitation	udw, usa	2015; 5 arc-min res. Gridded Hydrological sub-basin (HydroBASINS 6) data. Level of drinking water & sanitation coverage reported as risk levels between 1-5 (1: low, 5: extremely high).	(Hofste *et al* 2019)
		Population	
Net-migration	NM	1990-2000; 30 arc-sec gridded data aggregated to 5 arc-min.	de Sherbinin *et al* (2015)
Population	POP	1990, 2000; Gridded population count with 5 arc-min res.	Klein Goldewijk *et al* (2010), HYDE 3.1

2.1 Indicators of environmental stress and societal factors

Our indicator approach for analyzing the interplay of environmental and societal characteristics behind human migration has been extended from Varis et al. (2019b) who studied the resilience of human-natural systems through considering both adaptive capacity and environmental vulnerability. This approach allows a geospatial analysis of *environmental stress* factors in parallel with factors indicating societal *adaptive capacity* to cope with environmental and other stress factors. For the purposes of this study, some of the indicators were modified. We defined four societal factors: *governance effectiveness*, *level of income*, *health* and *education* as components of adaptive capacity, of which the last three are also the components of the Human Development Index used as a composite index in Varis et al. (2019b). Income was downscaled to grid level based on night lights and agricultural land use, using linear multiple regression model from Kummu et al (2018).

For environmental stress, we selected four variables representing diversity of environmental risks and stressors: *drought* and *water risk* were considered to be proxies for slow onset environmental change while *natural hazards* represent a more sudden change or shift in the environment. *Food production scarcity* was selected as a proxy of local food insecurity (see complete list of all indicator sources and their measurement in Table 1 and Table S1 in the Supplements). Spatial distributions of the indicators used are illustrated in Figures S2, S3 and S4.

Temporal average over 1990-2000 was used for all indicators which are available for the whole time period (except for food production which was measured in 2000 and drinking water and sanitation coverage measured in 2015 due to data availability). Drought risk (DR) was composed from the Standardized Precipitation-Evapotranspiration Index (SPEI) (Vicente-Serrano *et al* 2010) by computing a cumulative sum of negative index values (drier years than average) over the study period. Water risk (WR) was calculated based on quantitative risk factor, baseline water stress, and qualitative risk factor, the level of improved sanitation and drinking water, from Aqueduct Water Risk data (Hofste *et al* 2019). Food production scarcity (FPS) is the ratio between crop production and population (kcal per capita per day) and scaled between 0-1 based on kcal per capita level (FPS $\leq$ 500 kcal: high scarcity = 1; FPS $\geq$ 5000 kcal: no scarcity = 0). Finally, all indicators (except for FPS) were scaled between 0-1 with min-max normalization where the smallest and highest 5% were assigned values 0 and 1, respectively. Societal and environmental factors were then combined into two composite indices of adaptive capacity (AC) and environmental stress (ES), as the mean over their four components. The data were tested for cross-correlations: variables within AC index had strong correlation, while correlation between adaptive capacity and environmental stress variables was weak (see Figure S1).

2.2 Net-migration and population data

In the acquired dataset, decadal net-migration was defined as NM = total population change – (births – deaths), in each grid cell (de Sherbinin *et al* 2015). Net-negative migration illustrates areas with more emigrants than immigrants, and net-positive migration areas with more immigrants than emigrants over the time period. The NM data were aggregated from 30 arc-sec to 5 arc-min resolution to match other datasets, which were not available at higher resolution. Furthermore, de Sherbinin et al. (2015) data were not modelled with the 30 arc-sec resolution original input data. It is thus justified to aggregate the data to 5 arc-minute resolution without losing much information (see Figure S4 for the coefficient of variation in the aggregated data). The data were aggregated by summing over a 10x10 window by using the *aggregate*-tool in Raster-package in R (Hijmans 2019). For random forest analysis, the net-migration data were then normalized with the respective population count in the initial timestep (1990) in each grid cell in order to address the effect of population to net-migration count. Here it is important to note that net-migration accounts for all types of mobility and does not distinguish between voluntary and forced migration, for instance.

2.3 Interplay and importance of environmental and societal factors

We extend the conceptual typology introduced in Figure 1 to a quantitative tool by using the composite indicators of adaptive capacity (AC) and environmental stress (ES) (Varis et al (2019b); see above) to describe the relationship of environmental and societal factors driving migration (Figure 1). Accordingly, we created a four-by-four classification matrix representing the interplay at net-negative and net-positive migration locations (Figure 3) with four thresholds for low, medium-low, medium-high and high AC and ES as per the following breaks [0, 0.25, 0.5, 0.75, 1]. This framework was employed to both origins (net-negative migration) and destinations (net-positive migration) in order to define the interplay between AC and ES as the underlying conditions of migration at both ends. The matrix was used to calculate the sum of net-negative and net-positive migration in each class (e.g. total net-negative migration in class 1 would be the sum over-all net-negative grid cells within that class). Then the share of each class was calculated as the ratio to the total (global) net-negative/positive migration (sum of all net-negative/positive grid cells globally). Calculations were done by using the *zonal*-tool in the Raster package in R (Hijmans 2019).

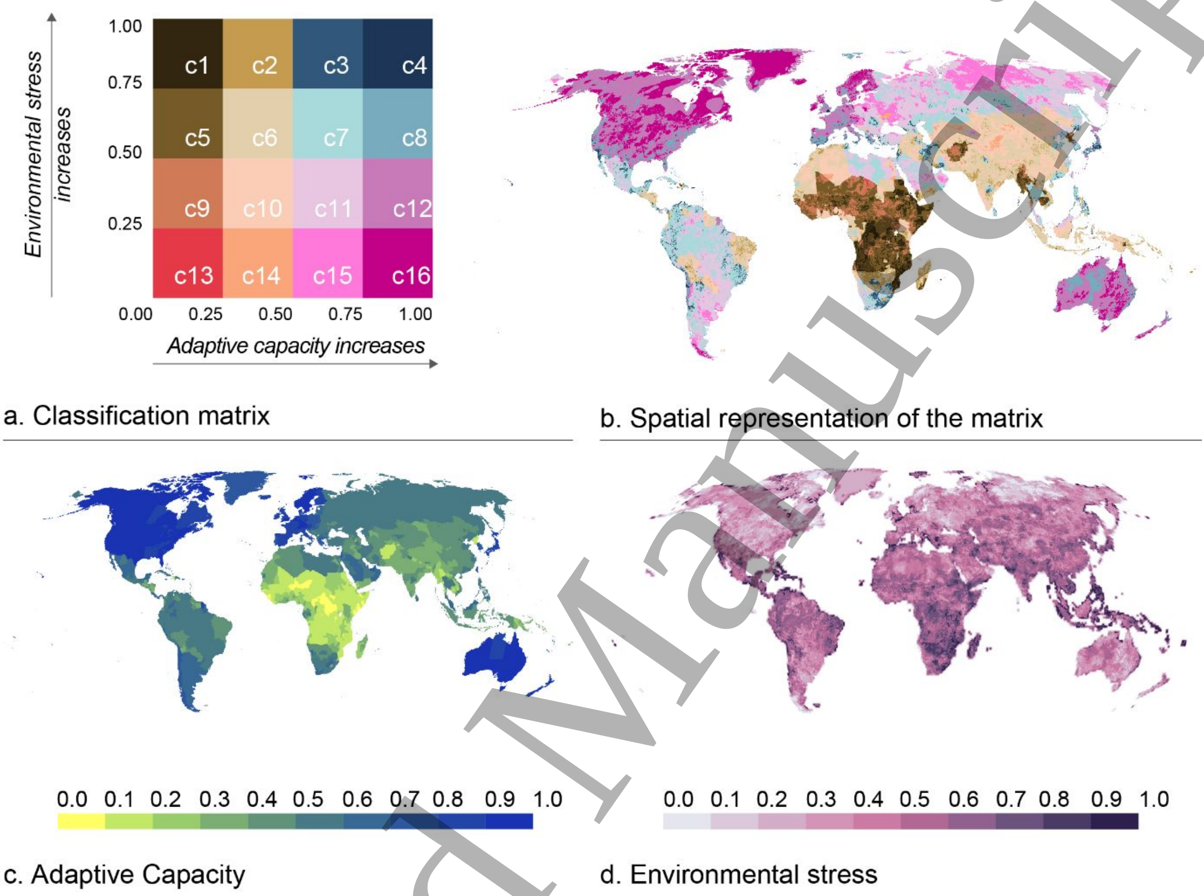

Figure 3. (a)., (b).: Classification matrix and its spatial representation. Thresholds in the matrix are defined with four thresholds for low, medium-low, medium-high and high adaptive capacity (AC) and environmental stress (ES) as per the following breaks [0, 0.25, 0.5, 0.75, 1]. The classes are named c1-c16. (c)., (d).: Geographic distribution of the composed AC and ES indices.

Random forest regression was utilized to quantitate the independent importance of each variable (Table 1) in explaining both net-negative and net-positive migration. Random forest regression is a machine learning algorithm that uses an ensemble of multiple bootstrap sample predictions (decision trees) to produce a consensus regression fit (Breiman 2001). This technique is suitable for identifying and ranking endogenous explanatory factors underlying migration decisions (Schutte *et al* 2021). It is also applicable to data with collinear explanatory variables and unique probability distributions as the method randomly splits or bags the data into multiple samples (and out-of-bag samples, i.e. the data left out of each sample) each containing only a subset of variables, i.e. potentially correlated variables are not represented in all decision trees (Cutler *et al* 2007). The importance of each variable describes the increase in prediction error (MSE from the out-of-bag sample) when the values of that variable are randomly permuted. High importance denotes high explanatory power in that specific model while negative importance indicates that the variable weakens the model's prediction power. Ultimately, relative feature importance (RI) is used to illustrate and rank how well a given feature predicts migration *in relation to the best feature with RI = 1*.

Country-specific regression models were created for relative net-negative (per population; 178 countries) and net-positive migration (per population; 178 countries) observations (response variables) and respective individual variables of adaptive capacity and environmental stress (explanatory variables) with the *Ranger* -package in R (Wright and Ziegler 2017). Regression was conducted for each country individually, as it represents a highly relevant scale for policy making. Grid cell values for both response and explanatory variables within each country were extracted and then used as individual observations for each model.

3 Results

3.1 Interplay of AC and EV

Our analysis shows that in 1990-2000, the majority of net-negative and net-positive migration occurred in areas characterized by high environmental stress (ES). Globally, 58% of the total net-negative migration took places in areas with medium-low to medium-high adaptive capacity (AC) and ES. Further, 32% of global net-negative migration originated in just one class (c6), with medium-high to high ES but medium-low AC (Figure 4a) while neighboring class c7 (with higher AC) and c10 (with lower ES) together accounted for 27% of global net-negative migration.

Despite the majority of global net-negative migration being concentrated in intensively populated areas (35% of world's population lived in c6, c7 and c10 in 1990) migration-to-population ratio shows a slightly different pattern. For instance, the net-negative migration-to-population ratio (total net-negative migration per population per class) in the abovementioned c6 was very low, around 69 emigrants per 1000 inhabitants, compared to the highest net-negative ratio of 5860 emigrants per 1000 inhabitants in c13 with globally lowest ES and AC (Figure 4b); however, the populated areas in c13 represent a very small share of global land and population as they include only a handful of cells e.g. in rural Kenya and Afghanistan (see Figure 3).

The clusters accommodating the majority of global net-positive and net-negative migration were characterized by similar profiles (Figure 4). A total of 80% of global net-positive migration took place in five classes of which c7 alone accommodated 22% of global net-positive migration (Figure 4c). Yet, the median net-positive migration-to-population ratio across all observations in c7 was only 96 immigrants per 1000 inhabitants. The highest net-positive migration-to-population ratio was found in c3 with 147 immigrants per 1000 inhabitants (Figure 4d).

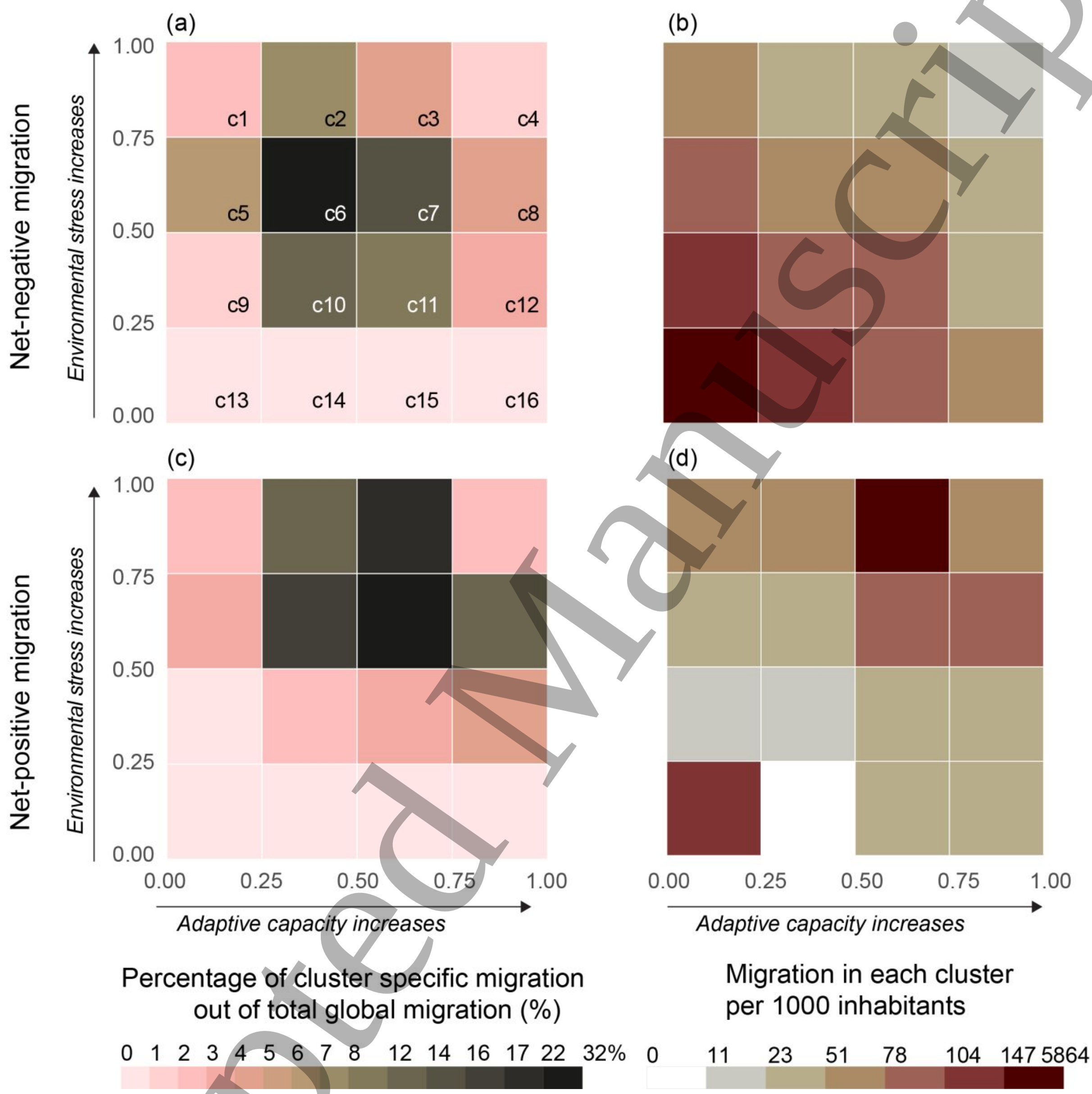

Figure 4. Heatmaps representing the share of each class in terms of corresponding variable. Share of (a) net-negative and (b) net-positive migration out of total global net-migration in each class in 1990-2000; (c) net-negative and (d) net-positive migration per 1000 inhabitants in each class in 1990-2000.

3.2 Relative importance of explanatory variables

The analysis of the variables' importance and explanatory power highlights the following three points. Firstly, Ethiopia, Georgia, Jordan, Bangladesh, Demographic Republic of Congo and Papua New Guinea stood out with the strongest explanatory power for net-negative migration (R^2= 0.63, 0.61, 0.58, 0.52, 0.51 and 0.5 respectively), compared to moderate global predictions (global median of R^2 = 0.17) (Figure 5). In terms of net-positive migration, explanatory power was moderately strong (R^2 > 0.50) in 10 countries (e.g. R^2 = 0.72 in Tanzania; 0.67 in Eritrea, 0.66 in Guyana, 0.58 in Mali), while global median remained very low (global median R^2 = 0.14). Noteworthy, the selected variables could not explain any of net-negative migration in 14% of all countries, or any of the net-positive migration 28% of the countries (R2 = 0). See Figure S5 for the overall out-of-bag prediction error for each model.

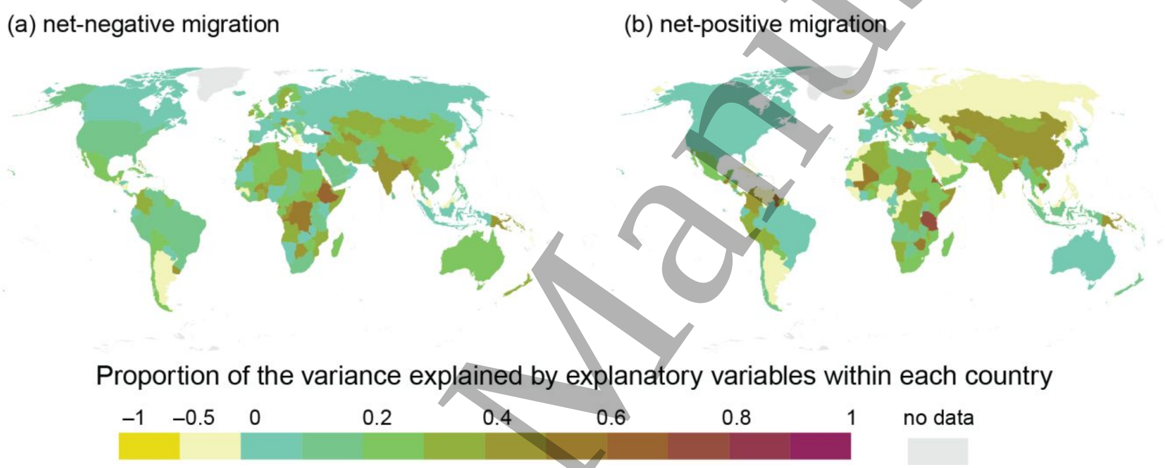

Figure 5. Proportion of the variance (R^2, from the out-of-bag sample) of a) net-negative and (b) net-positive migration, explained by all the studied explanatory variables together within each country.

Secondly, income level was the key determinant for both net-negative (**Error! Reference source not found.**a) and positive migration (**Error! Reference source not found.**a), illustrating a globally mutual feature importance even when other societal and environmental factors were included in the models. Given that the income data were downscaled with night-lights data, this also indicates a strong effect of urbanization. In other words, income was the best variable in describing the internal variation of both net-positive and net-negative migration across the low to high income gradients in around half of the countries (58 and 60% of the countries for net-positive and negative migration, respectively).

Notably, education and health were the second most important societal features, by ranking highest in 8% and 6% of the countries in terms of net-negative migration, respectively (Figure 6b-c). Importantly, the global median relative importance (RI) of education (global median RI = 0.41) and health (global median RI = 0.39) in explaining net-negative migration were a third of the most important factor income level (global median RI = 1.00), being higher than the global median RI of any of the environmental variables (Figure 8a, Figure S6). To mention a few, education was the most important feature in Kyrgyz Republic (absolute feature importance AFI = 736; R^2 = 0.33; mean square error MSE = 555), Syria (AFI= 290; R^2 = 0.19; MSE = 188) and Colombia (AFI= 283; R^2 = 0.32; MSE = 67) for net-negative migration (Figure S8).

In terms of net-positive migration, health was the most important determinant after income, by ranking the highest in 8% of the countries, while education ranked the highest in only 4% of the countries (Figure 7b-c). Yet, the global median RI of education and health were around a third (RI = 0.34; 0.32, respectively) of income level (RI = 1.00) (Figure 8b, Figure S7). To mention a few, health was the best variable in Madagascar, (AFI = 67; R^2 = 0.28; MSE = 47), India (AFI = 16; R^2 = 0.39; MSE = 1.7) and Lao (AFI = 12; R^2 = 0.36; MSE = 4.9) for net-positive migration. Expectedly, governance ranked the lowest in explaining both net-negative and positive migration; data for governance were on a country level and thus do not explain well variation within a country. See Figure S9 for country specific results regarding absolute feature importance.

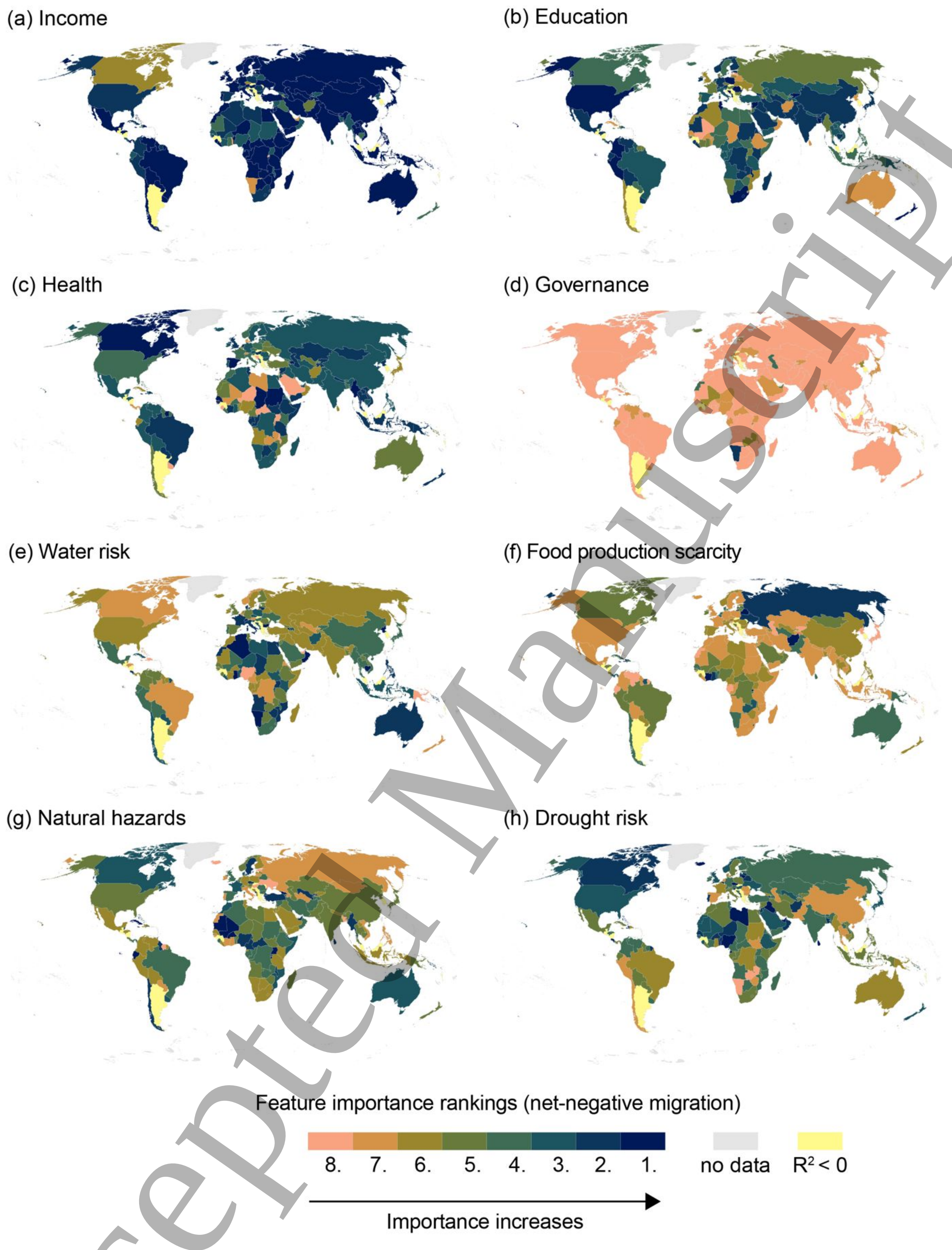

Figure 6. Feature importance ranking in each country (N=178). Importance of each feature on net-negative migration is ranked so that the most and least important variables in each country's model are assigned values 1 and 8, respectively. The higher the importance, the better the variable is in explaining net-negative migration in each country.

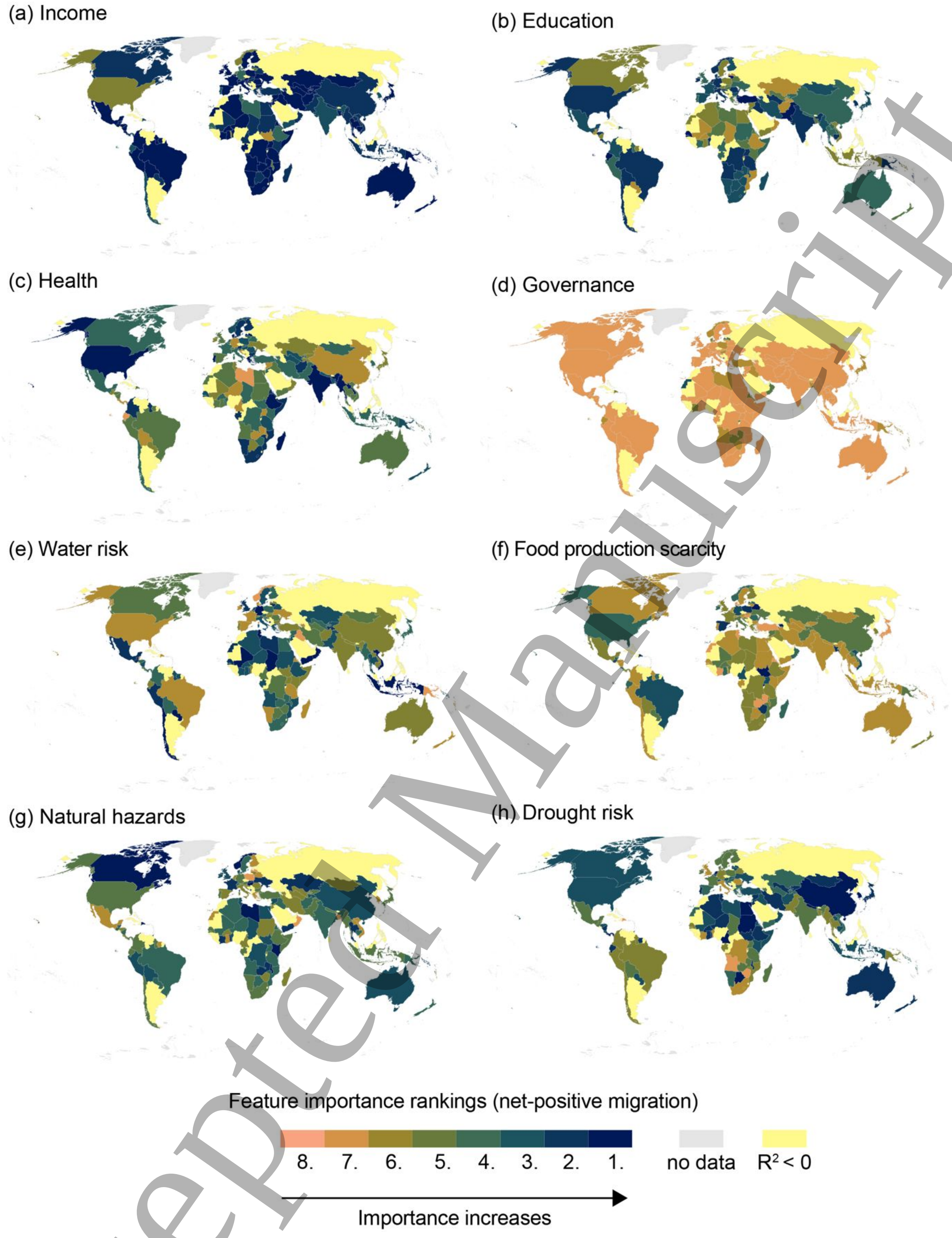

Figure 7. Feature importance ranking in each country (N=178). Importance of each feature on net-negative migration is ranked so that the most and least important variables in each country's model are assigned values 1 and 8, respectively. The higher the importance, the better the variable is in explaining net-negative migration in each country.

Thirdly, another collective feature is shown by slow-onset environmental stressors and natural hazards which were globally the dominant environmental variables in explaining net-negative and net-positive migration in almost all country groups (**Error! Reference source not found.**b; **Error! Reference source not found.**b: Figure 8). Drought risk and natural hazards ranked the highest in explaining net-negative migration in 7% of the countries each (Figure 6g-h). Drought risk was the best feature in Iraq (AFI = 6278, R^2 = 0.33, MSE = 3977) and Libya (AFI = 0.008, R^2 = 0.37, MSE = 0.01) while natural hazards ranked the highest in Georgia (AFI =248, R^2 = 0.61, MSE = 111) and Mali (AFI = 15, R^2 = 0.30, MSE = 11), to mention few (See Figure S8 for country specific results). Yet, the global median RI of drought risk and natural hazards were less than 30% (global median RI = 0.28; 0.21, respectively) of the most important variable income (RI = 1.0) (Figure 8a), indicating that their importance in relation to the most important variable was relatively low in the countries where the variables did not rank the highest (Figure 6g-h). The importance of water risk and food production was lower, by being the best variable in only 6% and 4% of the countries, respectively.

In terms of net-positive migration, water risk was the best variable in 9% of the countries, the global median RI being one third (RI = 0.3) of income (RI = 1.0). Notably, the global median relative importance of drought risk was higher, 37% of the best feature, indicating it had a moderate importance even when not ranking as the best feature (Figure 7e, Figure 8b). Natural hazards ranked highest in 8% of the countries, including Libya (AFI = 36, R^2 = 0.20, MSE = 53), Kenya (AFI = 1.2, R^2 = 0.22, MSE = 2.6) and Lesotho (AFI = 0.36, R^2 = 0.41, MSE = 0.28) but also Norway (AFI = 7.3, R^2 = 0.19, MSE = 5.3), where the conditions regarding the risk to natural hazards as well as adaptive capacity range from low to high (See Figure S9 for country specific results). Food production scarcity ranked highest in 5% of the countries, with the global median RI being 0.1.

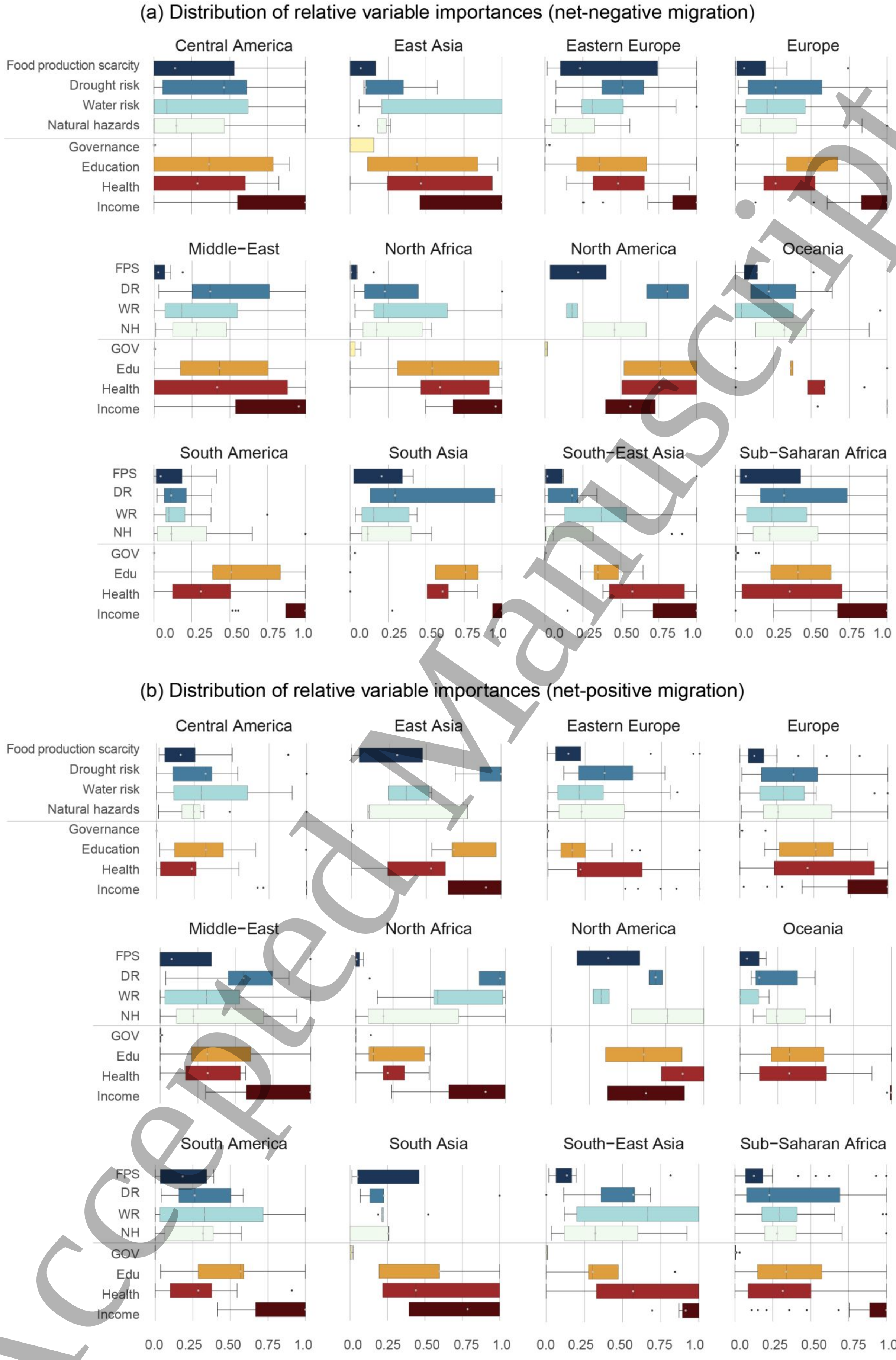
(a) Distribution of relative variable importances (net-negative migration)
Central America
East Asia
Eastern Europe
Europe
Food production scarcity
Drought risk
Water risk
Natural hazards
Governance
Education
Health
Income
Middle−East
North Africa
North America
Oceania
FPS
DR
WR
NH
GOV
Edu
Health
Income
South America
South Asia
South−East Asia
Sub−Saharan Africa
0.0 0.25 0.50 0.75 1.0
(b) Distribution of relative variable importances (net-positive migration)

Figure 8. Relative importance (0-1) of each examined variable by country groups for a. net-negative (178 models) and b. net-positive migration (178 models). Minimum and maximum of each variable are shown with whiskers, while the box represents first and third quartiles over median. Values above and below those, i.e. outliers are shown as points. Relative importance tells the importance of each feature in explaining migration in relation to the most important feature.

4 Discussion

4.1 Importance of societal factors on environmental migration

The majority of global migration in our study period occurred in areas with a risky combination of high environmental stress (ES) and low to medium adaptive capacity (AC). Income level was the key factor in explaining net-migration, interestingly across the global income groups from low to high. Slow-onset environmental variables, drought and water risk, had the highest importance amongst environmental stress for both net-positive and net-negative migration especially in dry regions like South and East-Asia and North-Africa. Here net-positive refers to situations where in-migration exceeds out-migration while net-negative refers to situations where out-migration exceeds in-migration. Our global synthesis with sixteen classes successfully illustrated the spatial heterogeneity of the different factors underlying migration and their interplay. While the global prediction power with the selected factors was moderate, we were able to identify geographical heterogeneities of migration patterns.

A clear majority of global net-negative migration originates from environmentally stressed and hazardous areas (in agreement with de Sherbinin *et al* 2012) with medium-low to medium-high environmental stress and medium level of adaptive capacity. This aligns the previous literature showing that environmental migration is more common among the middle-level income countries, not among the poorest nor the richest (Cattaneo and Peri 2016, Hoffmann *et al* 2020). Our results indicate that income level, followed by drought risk and education have a primary importance in explaining net-negative migration in areas with high environmental stress (Figure 6; Figure 8a). In fact, aligned with our finding, Neumann and Hermans (2017) observed economic and social aspects to be the predominant reasons for out-migration whereas environmental factors, such as droughts, were found to drive migration indirectly through "economic deterioration" in areas like the Sahel. Our results suggest that environmental pressures alone are unlikely to cause migration through simple linear linkages, despite the fact that the presence of environmental pressures in the sending areas of migration is evident (Black *et al* 2011b, 2011a, de Sherbinin *et al* 2012, Neumann *et al* 2015, Abel *et al* 2019). The role of the environment in driving migration should thus be investigated critically (Boas *et al* 2019, Murphy 2015, Betts and Pilath 2017), and socioeconomic variables should be factored in in the attempts to quantify environmental migration.

We found that the majority of global net-positive migration was characterized by high environmental stress and medium level of adaptive capacity (Figure 4c). This finding is in line with the empirical evidence that both voluntary and forced migration tend to occur between neighboring countries or within the same region (Abel *et al* 2019, Abel and Sander 2014). African migrants, for instance, predominantly move within Africa so the high environmental stress observed in the destinations may reflect the fact that most migration is short-distance. The characteristics of the destination areas, on the other hand, have received less attention in the environmental-migration nexus literature (Cattaneo and Peri 2016, Hoffmann *et al* 2020). A combination of high environmental stress and low-to-medium capacity potentially exposes migrants to a twofold risk at both origin and destination: firstly, they are also exposed to numerous social and ecological vulnerabilities in the destination (de Sherbinin *et al* 2012, Adri and Simon 2018), and secondly, such conditions might prevent people with low capabilities from moving to a more desired location or relocating back to their origin (Ayeb-Karlsson *et al* 2020). Environmental hazards combined with numerous inadequacies in terms of human development, economy and governance may trap in-coming migrants with increasing vulnerabilities (Ayeb-Karlsson *et al* 2020) and thus hamper the positive gains from migration.

Despite the fact that our global analysis does not distinguish between rural and urban areas in terms of origins and destinations of migration, our income data capture the importance of regional disparities in producing migration. These data were downscaled from sub-national income data to 5 arc-min (ca 10 km in the equator) resolution by using night lights and agricultural land use data and thus illustrate the difference in income levels between rural and urban areas within a country. Considering the importance of income in explaining both net-negative and net-positive migration, it is likely that it is the difference between income-levels of the origin and destination areas that explains migration instead of income itself. This finding aligns well with the classic gravity-model theories of migration (de Haas 2011, Lee 1966).

In the coming decades, African countries, in particular, are expected to experience fast urbanization resulting from a combination of natural population growth and in-migration driven by the disparities between rural and urban areas (Farrell 2018, Awumbila 2017). Rapidly expanding urban areas with low capacity in terms of income level, governance and basic services, in particular, tend to generate informal settlements that often function as "waiting rooms" for in-coming migrants with low capabilities (Tacoli *et al* 2015, Andrews 2020, Niva *et al* 2019). Meanwhile, the population living under water stress is expected to grow by half up to double in the coming decades due to climate change (Munia *et al* 2020). In fact, there is already some evidence showing that some urban agglomerates are facing a dual-risk from both droughts and floods (Cai *et al* 2018). Notably, our results show drought and water risk had the highest or second highest importance in explaining net-positive migration in numerous areas with low-to-medium adaptive capacity and high environmental stress, reflecting the evidence from other studies as well as showing further research needs; Future studies should pay elevated attention to the conditions of where people move to (Ayeb-Karlsson *et al* 2020, Findlay 2011), especially in urban destination.

4.2 Limitations of this study

This work has analysis and data -related limitations commonly faced in global analyses. Firstly, the results are prone to uncertainty, because the migration data obtained from de Sherbinin *et al* (2015) themselves are a product of modelling: the original migration dataset contained a minor built-in error of around (-) 400,000 migrants, (ca. 0.1% of global net-migration). The same issue applies to the environmental data of which many are originally modelled (water stress, SPEI index and natural hazards), and may thus contain and result in inaccuracies especially in remote locations.

Secondly, while our global analysis was conducted at high resolution grid, it should be noted that the net-migration data used here represent the world in the past. Here, the dataset from de Sherbinin *et al* (2012) at 10 km spatial resolution were selected over a recent net-migration dataset by Alessandrini et al (2020). While Alessandrini et al (2020) data has a fine temporal resolution, they used only gridded national values on a coarse spatial resolution (25 km) instead of using downscaled sub-national values, as done in de Sherbinin (2015). Notably, despite we utilized the best available data for building our indicators, water stress and food production scarcity were comprised with data from varying years.

Thirdly, the explanatory variables could explain up to 60% of the variance in any of the models, and notably, income outperformed all other variables systematically across the globe. While this aligns with many studies highlighting the role of income as a primary driver of migration, the results may be biased. The data of income were downscaled to grid level by using a proxy for rural-urban division (see Supplement) thus potentially overriding other variables that were gridded from sub-national data. Moreover, some of the indicators used here (NH, WR, FPS) comprise of multiple indices and thus do not provide information on the importance of their individual components on migration.

It should also be noted that studying a complex phenomenon such as migration by using quantitative indices is prone to uncertainty as global indicators and the data cannot capture decision-making processes at an individual level, or in very small countries. Despite the population living in countries where the number of cells is 20 or less is only 0.1% of the global population, it can be presumed that the data do not fully capture migration dynamics in micro-states, such as Liechtenstein or Andorra. Moreover, it should be noted that our data only illustrate net-migration and thus do not separate voluntary from forced migration. While it is not entirely possible to make a clear-cut distinction between forced and voluntary migration since in fact migration decisions do have a certain degree of volition (Erdal and Oeppen 2018), different types of migrants are protected by different bodies of international law as well as non-legally binding best practices and principles (Martin 2017). Therefore, in practice, migration policy and regulations need to distinguish between types of migration which unfortunately is not possible in the net-migration data used here.

Nevertheless, our analysis does tap into various indicators such as governance, education and health that have previously been identified as being fundamental in reducing vulnerability and enhancing adaptive capacity (Andrijevic *et al* 2020, Lutz *et al* 2014). The novel machine learning approach which helps identify the importance of each variable in explaining migration thus allows for pinpointing which societal factor is highly relevant and can be used as an empirical ground in policy making processes. Furthermore, our analysis provides useful insights on the relationship between the used variables as well as variation of relative feature importance in terms of migration globally, by country groups, and by similarity classes. That the variables featured very different level of explanation power between neighboring countries indicates that selecting variables for future studies is sensitive to location.

4.3 Ways forward

Our results and limitations partly reflect the availability, accuracy and development needs of migration and socioeconomic indicator data. Demand for high-resolution spatiotemporal data on detailed subnational net-migration is urgent. To our knowledge there are altogether two gridded datasets of global net-migration of which both compromise with either temporal or spatial scale and the scale of input data (national vs. sub-national) (see section 4.2). This significantly hinders the production of accurate and comparable spatiotemporal estimates of migration. For instance, the simplistic narratives of mass-migration fluxes and portraying migration as a security hazard has been repeated in both research and policy-making for decades (Boas *et al* 2019), but data for investigating these recent developments lag behind.

Noteworthy, identifying local characteristics underlying migration is equally difficult. Globally comparable fine-scale socio-economic data are scarce and typically sub-national scale data require downscaling if a more refined scale is desired. For instance, education, governance and health were outperformed by downscaled and spatially more detailed income data income in explaining net-negative and net-positive migration. We thus call for high-resolution spatiotemporal data for producing consistent and up-to-date predictions of human migration and its conditions globally.

5 Conclusions

We provided a global assessment of the interplay of environmental and societal characteristics underlying migration in sending (negative net-migration) and receiving (positive net-migration) areas by creating a novel classification-matrix. Furthermore, we assessed the importance of eight environmental and socio-economic indicators on net-negative and net-positive migration at national scale using a machine learning method. Our findings extend the current knowledge on three fronts:

- Within the study period 1990-2000, the majority of global net-negative and net-positive migration was concentrated in areas with rather similar profiles; a combination of both low-to-medium adaptive human capacity and medium-to-high environmental stress, and low migration-to-population ratio.
- Income outperformed all other variables in circa half of both sending and receiving areas. Education and health were also significant local factors in explaining migration, especially net-negative, with global median importance being around 40% of the most important factor, income. Drought and water risk had the highest importance among environmental variables, globally.
- The combination of the novel matrix approach, an ensemble of national-level models, and machine computational methods allowed us to identify new global patterns on both net-positive and net-negative migration, thus significantly improving the knowledge on important drivers of in- and out-migration.

Finally, we highlight the urgency for adapting integrative approaches in the quantitative analysis of environment-migration nexus more rigorously. A phenomenon that is ultimately based on individual and human decision-making simply cannot and should not be studied without the inclusion of societal dimension: human capacity and agency. In order to study the complex causalities between migration and its underlying conditions further in both research and policy-making, it is of urgent importance to produce detailed and timely spatiotemporal data regarding migration and its drivers. In the time when environmental vulnerabilities are on the surge, it is indeed fundamental to understand how human populations respond and adapt to them.

Acknowledgements

This study was funded by Maa- ja vesitekniikan tuki r.y., Aalto University (Base Funds and the Doctoral Programme of the School of Engineering), Academy of Finland funded project WATVUL (grant no. 317320), European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819202) and IIASA including the National Member Organizations that support the institute.

References

Abel G J, Brottrager M, Crespo Cuaresma J and Muttarak R 2019 Climate, conflict and forced migration *Global Environmental Change* **54** 239–49

Abel G J and Sander N 2014 Quantifying Global International Migration Flows *Science* **343** 1520–2

Adri N and Simon D 2018 A tale of two groups: focusing on the differential vulnerability of "climate-induced" and "non-climate-induced" migrants in Dhaka City *Climate and Development* **10** 321–36

Alessandrini A, Ghio D and Migali S 2020 *Estimating net migration at high spatial resolution, EUR 30261 EN* (Luxembourg: European Union) Online: doi:10.2760/383386

Andrews C J 2020 Toward a research agenda on climate-related migration *Journal of Industrial Ecology* **24** 331–41

Andrijevic M, Crespo Cuaresma J, Muttarak R and Schleussner C-F 2020 Governance in socioeconomic pathways and its role for future adaptive capacity *Nat Sustain* **3** 35–41

Awumbila M 2017 Drivers of Migration and Urbanization in Africa: Key Trends and Issues United Nations Expert Group Meeting on Sustainable Cities, Human Mobility and International Migration (New York: United Nations: Department of Economic and Social Affairs, Population Division)

Ayeb-Karlsson S, Kniveton D and Cannon T 2020 Trapped in the prison of the mind: Notions of climate-induced (im)mobility decision-making and wellbeing from an urban informal settlement in Bangladesh *Palgrave Communications* **6** 1–15

Beine M and Parsons C 2015 Climatic Factors as Determinants of International Migration *Scand. J. of Economics* **117** 723–67

Benveniste H, Cuaresma J C, Gidden M and Muttarak R 2021 Tracing international migration in projections of income and inequality across the Shared Socioeconomic Pathways *Climatic Change* **166** 39

Betts A and Pilath A 2017 The politics of causal claims: the case of environmental migration *J Int Relat Dev* **20** 782–804

Black R, Adger W N, Arnell N W, Dercon S, Geddes A and Thomas D 2011a The effect of environmental change on human migration *Global Environmental Change* **21, Supplement 1** S3–11

Black R, Bennett S R G, Thomas S M and Beddington J R 2011b Migration as adaptation *Nature* **478** 447–9

Boas I, Farbotko C, Adams H, Sterly H, Bush S, van der Geest K, Wiegel H, Ashraf H, Baldwin A, Bettini G, Blondin S, de Bruijn M, Durand-Delacre D, Fröhlich C, Gioli G, Guaita L, Hut E, Jarawura F X, Lamers M, Lietaer S, Nash S L, Piguet E, Rothe D, Sakdapolrak P, Smith L, Tripathy Furlong B, Turhan E, Warner J, Zickgraf C, Black R and Hulme M 2019 Climate migration myths *Nature Climate Change* **9** 901–3

Borderon M, Sakdapolrak P, Muttarak R, Kebede E, Pagogna R and Sporer E 2019 Migration influenced by environmental change in Africa: A systematic review of empirical evidence *Demographic Research (Special Collection)* **11** 491–544

Breiman L 2001 Random Forests *Machine Learning* **45** 5–32

Cai J, Kummu M, Niva V, Guillaume J H A and Varis O 2018 Exposure and resilience of China's cities to floods and droughts: a double-edged sword *International Journal of Water Resources Development* **34** 547–65

Cattaneo C and Peri G 2016 The migration response to increasing temperatures *Journal of Development Economics* **122** 127–46

Chemnick J 2019 Where Climate Change Fits into Venezuela's Ongoing Crisis *Scientific American* Online: https://www.scientificamerican.com/article/where-climate-change-fits-into-venezuela-rsquo-s-ongoing-crisis/

Cutler D R, Edwards T C, Beard K H, Cutler A, Hess K T, Gibson J and Lawler J J 2007 Random Forests for Classification in Ecology *Ecology* **88** 2783–92

Dilley M, Chen R S, Deichmann U, Lerner-Lam A, Arnold M, Agwe J, Buys P, Kjekstad O, Lyon B and Yetman G 2005 *Natural disaster hotspots: A global risk analysis* (Washington, DC: The World Bank)

Erdal M B and Oeppen C 2018 Forced to leave? The discursive and analytical significance of describing migration as forced and voluntary *Journal of Ethnic and Migration Studies* **44** 981–98

Farrell K 2018 An Inquiry into the Nature and Causes of Nigeria's Rapid Urban Transition *Urban Forum* **29** 277–98

Findlay A M 2011 Migrant destinations in an era of environmental change *Global Environmental Change* **21** S50–8

Foresight 2011 *Migration and Global Environmental Change: Final Project Report* (Government Office for Science) Online: https://www.gov.uk/government/collections/migration-and-global-environmental-change

Gray C and Mueller V 2012 Natural disasters and population mobility in Bangladesh *PNAS* **109** 6000–5

Gustin G and Henninger M 2019 Central America's choice: Pray for rain or migrate *NBC News* Online: https://www.nbcnews.com/news/latino/central-america-drying-farmers-face-choice-pray-rain-or-leave-n1027346

de Haas H 2011 *The determinants of international migration: Conceptualizing policy, origin and destination effects* (Oxford: International Migration Institute, IMI Working Paper)

Hijmans R J 2019 *raster: Geographic Data Analysis and Modeling* Online: https://CRAN.R-project.org/package=raster

Hoffmann R, Dimitrova A, Muttarak R, Crespo Cuaresma J and Peisker J 2020 A meta-analysis of country-level studies on environmental change and migration *Nature Climate Change* 1–9

Hoffmann R, Sedova B and Vinke K 2021 Improving the Evidence Base: A Methodological Review of the Quantitative Climate Migration Literature *Global Environmental Change*

Hofste R, Kuzma S, Walker S and Sutanudjaja E H 2019 Aqueduct 3.0: Updated Decision- Relevant Global Water Risk Indicators

Jónsson G 2010 The environmental factor in migration dynamics: A review of African case studies *IMI Working Paper Series* **21** Online: https://www.imi.ox.ac.uk/publications/wp-21-10

Klein Goldewijk K, Beusen A and Janssen P 2010 Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1 *The Holocene* **20** 565–73

Kluger L C, Alff H, Alfaro-Córdova E and Alfaro-Shigueto J 2020 On the move: The role of mobility and migration as a coping strategy for resource users after abrupt environmental disturbance – the empirical example of the Coastal El Niño 2017 *Global Environmental Change* **63** 102095

613 Kubik Z and Maurel M 2016 Weather Shocks, Agricultural Production and Migration: Evidence from Tanzania *The Journal of Development Studies* **52** 665–80

615 Kummu M, Taka M and Guillaume J H A 2018 Gridded global datasets for Gross Domestic Product and Human Development Index over 1990–2015 *Scientific Data* **5** 180004

617 Lee E S 1966 A Theory of Migration *Demography* **3** 47–57

618 Liang S, Yang X, Qi J, Wang Y, Xie W, Muttarak R and Guan D 2020 CO2 Emissions Embodied in International Migration from 1995 to 2015 *Environ. Sci. Technol.* **54** 12530–8

620 Lilleør H B and Van den Broeck K 2011 Economic drivers of migration and climate change in LDCs *Global Environmental Change* **21** S70–81

622 Lutz W, Muttarak R and Striessnig E 2014 Universal education is key to enhanced climate adaptation *Science* **346** 1061–2

624 Markham L 2019 How climate change is pushing Central American migrants to the US *The Guardian* Online: https://www.theguardian.com/commentisfree/2019/apr/06/us-mexico-immigration-climate-change-migration

627 Marotzke J, Semmann D and Milinski M 2020 The economic interaction between climate change mitigation, climate migration and poverty *Nature Climate Change* **10** 518–25

629 Martin P L and Taylor J E 1996 The anatomy of a migration hump *Development strategy, employment and migration : insights from models*

631 Martin S F 2017 Forced Migration and Refugee Policy *Demography of Refugee and Forced Migration* **13** 271–303

633 Mueller N D, Gerber J S, Johnston M, Ray D K, Ramankutty N and Foley J A 2012 Closing yield gaps through nutrient and water management *Nature* **490** 254–7

635 Munia H A, Guillaume J H A, Wada Y, Veldkamp T, Virkki V and Kummu M 2020 Future Transboundary Water Stress and Its Drivers Under Climate Change: A Global Study *Earth's Future* **8** e2019EF001321

638 Murphy D W A 2015 Theorizing climate change, (im)mobility and socio-ecological systems resilience in low-elevation coastal zones *Climate and Development* **7** 380–97

Nelson D R, Adger W N and Brown K 2007 Adaptation to Environmental Change: Contributions of a Resilience Framework *Annual Review of Environment and Resources* **32** 395–419

Neumann K and Hermans F 2017 What Drives Human Migration in Sahelian Countries? A Meta-analysis *Population, Space and Place* **23** e1962

Neumann K, Sietz D, Hilderink H, Janssen P, Kok M and van Dijk H 2015 Environmental drivers of human migration in drylands – A spatial picture *Applied Geography* **56** 116–26

Niva V, Taka M and Varis O 2019 Rural-Urban Migration and the Growth of Informal Settlements: A Socio-Ecological System Conceptualization with Insights Through a "Water Lens" *Sustainability* **11** 3487

Podesta J 2019 *The climate crisis, migration and refugees* (Washington, DC.: The Brookings Institution) Online: https://www.brookings.edu/research/the-climate-crisis-migration-and-refugees/#footnote-19

Schutte S, Vestby J, Carling J and Buhaug H 2021 Climatic conditions are weak predictors of asylum migration *Nature Communications* **12** 2067

Selby J, Dahi O S, Fröhlich C and Hulme M 2017 Climate change and the Syrian civil war revisited *Political Geography* **60** 232–44

de Sherbinin A, Levy M, Adamo S, MacManus K, Yetman G, Mara V, Razafindrazay L, Goodrich B, Srebotnjak T, Aichele C and Pistolesi L 2015 Global Estimated Net Migration Grids by Decade: 1970-2000 Online: https://doi.org/10.7927/H4319SVC

de Sherbinin A, Levy M, Adamo S, MacManus K, Yetman G, Mara V, Razafindrazay L, Goodrich B, Srebotnjak T, Aichele C and Pistolesi L 2012 Migration and risk: net migration in marginal ecosystems and hazardous areas *Environ. Res. Lett.* **7** 045602

Smits J and Permanyer I 2019 The Subnational Human Development Database *Scientific Data* **6** 190038

Statistics Division of the United Nations Secretariat 2021 Standard country or area codes for statistical use (M49) - Geographic Regions. Online: https://unstats.un.org/unsd/methodology/m49/

Tacoli C, Mcgranahan G and Satterthwaite D 2015 Urbanization, rural–urban migration and urban poverty *IIED Working Paper*

669 Varis O, Taka M and Kummu M 2019a Data from: The planet's stressed river basins: too much pressure or too little adaptive capacity? 495366684 bytes Online: http://datadryad.org/stash/dataset/doi:10.5061/dryad.h2v2398

672 Varis O, Taka M and Kummu M 2019b The Planet's Stressed River Basins: Too Much Pressure or Too Little Adaptive Capacity? *Earth's Future* **7** 1118–35

674 Vicente-Serrano S M, Beguería S and López-Moreno J I 2010 A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index *Journal of Climate* **23** 1696–718

677 WGI 2018 *The worldwide governance indicators* (Washington, DC) Online: http://info.worldbank.org/governance/wgi/#home

679 Wiederkehr C, Beckmann M and Hermans K 2018 Environmental change, adaptation strategies and the relevance of migration in Sub-Saharan drylands *Environ. Res. Lett.* **13** 113003

681 Wright M N and Ziegler A 2017 ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **77** 1–17

683 Xu C, Kohler T A, Lenton T M, Svenning J-C and Scheffer M 2020 Future of the human climate niche *Proc Natl Acad Sci USA* **117** 11350–5