Journal of Computational Social Science (2021) 4:77–100
https://doi.org/10.1007/s42001-020-00066-9

RESEARCH ARTICLE

Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh

Kelsea B. Best[1] · Jonathan M. Gilligan[1] · Hiba Baroud[2] · Amanda R. Carrico[3] · Katharine M. Donato[4] · Brooke A. Ackerly[5] · Bishawjit Mallick[6]

Received: 17 July 2019 / Accepted: 26 February 2020 / Published online: 11 March 2020
© Springer Nature Singapore Pte Ltd. 2020

Abstract

The decision to migrate is complex and is often influenced by a combination of economic, social, political, and environmental pressures. Household survey instruments can capture detailed information about migration histories and their contexts, but it can be challenging to identify important predictors from large numbers of covariates with standard statistical methods, such as regression analyses. Machine learning techniques are well suited to pattern identification and can identify important covariates from large datasets. We report on the application of machine learning approaches to two large surveys collected from a total of more than 2800 households in southwestern Bangladesh. We applied random forest classification and regression models to identify significant covariates with the greatest predictive power for household migration decisions. The results show that random forest models are able to identify nuances in predictors of different types of migration and migration in different communities. Random forests also outperform logistic regression and support vector machines in predicting migration in all cases analyzed. Therefore, random forest models and other machine learning methods can be useful for improving the predictive accuracy of migration models and identifying patterns in complex social datasets. Future work should continue to explore the potential of machine learning techniques applied to questions of environmental migration.

Keywords Random forest · Machine learning · Migration · Climate change · Bangladesh

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42001-020-00066-9) contains supplementary material, which is available to authorized users.

✉ Kelsea B. Best
kelsea.b.best@vanderbilt.edu

Extended author information available on the last page of the article

Introduction

Climate change poses a wide variety of threats to human health and well-being [1]. This is especially true in low-lying coastal communities, where climate change is likely to affect a variety of natural phenomena including storms, sea-level rise, coastal inundation, erosion, and precipitation [2]. In addition, climatic change and other environmental stressors will combine to affect livelihood opportunities in vulnerable coastal areas [3].

Migration as a possible response to climate change and other environmental stresses has received a lot of attention in both the scholarly literature and the press. Discussions of climate-induced refugees have traditionally been framed around a looming crisis of "climate refugees" [4]. However, this narrative has been challenged as oversimplifying migration. Recent work has shown that, although climate change and environmental pressure can affect migration, those impacts can have complex and nonlinear interactions with other factors, so that some environmental stresses can actually reduce migration [5–8]. Migration is a complex, multi-causal phenomenon that is impacted by both "push" factors such as political instability, lack of economic opportunity, and lack of natural resources in the location of origin, as well as "pull" factors related to the destination location including availability of employment, resources, and social capital. Intervening factors such as transportation networks, social ties, and cultural norms can further complicate the decision to migrate [8, 9].

Past work has studied specific environmental drivers of migration such as sea level rise [10], impacts on agriculture [11], extreme weather events [12], and temperature increases [13]. Within this work, certain authors focus on the dynamics of temporary migration, while others question the causes of permanent migration specifically [14, 15]. An additional challenge in studying environmental migration is that the findings vary significantly by location [16]. Specific research has focused on climate variability in South America [17], drought in Ethiopia [18], land use in Ecuador [19], heat stress in Pakistan [20], soil quality in Kenya [21], tsunamis in Sumatra [22], weather anomalies across Africa [23], to name a few. Individual perceptions, preferences, and demographics may shape environmentally induced migration. For example, some research has focused on the impacts of individual risk perception on migration [24, 25], while other research suggests that preference for a certain type of climate is a more significant driver [26]. Demographic factors, such as gender [27, 28], legal status [29], and household wealth [30], also influence migration.

The complexity of human migration poses a challenge for researchers [31]. How to best model human migration to account for this complexity, as well as how to obtain appropriate and accurate data to test these models, remains an open and contested question [32]. Current work uses a wide range of methods and models, including strictly conceptual models [10, 33], logistic regression [25], multivariate regression [34], and a few agent-based models [35–39]. Additionally, some researchers choose to control for demographic variables, while others do not [40].

The complexity of the migration process often makes it difficult or impossible to isolate one or a few dominant driving factors of migration while controlling for all other variables. Researchers who study migration often use expert judgement or theory to select which variables to focus on. This approach can provide insights into how certain drivers may impact migration decisions, but it presents some limitations. First, there is not always a clear theoretical basis for selecting a few variables out of dozens or even hundreds collected by a survey. Second, there is a risk that by focusing only on migration dynamics predicted by current theory, researchers may miss novel dynamics not accounted for in current theory.

This paper aims to address this gap through a data-driven approach that applies random forest models to two large household surveys of communities in Bangladesh to identify which variables have the greatest statistical importance for predicting migration. Random forest models and other statistical learning methods have several advantages over more traditional regression analysis when analyzing large data sets with many covariates and no clear theoretical model of the processes being studied. Random forests in particular often display high predictive accuracy, ability to determine variable importance, and the ability to model complex and nonlinear interactions among variables [41]. Random forest models have been shown to perform well in environmental and ecological contexts [41, 42]. To our knowledge, this is a novel application of random forest algorithms to the topic of environmental migration, and to social survey datasets in general. Using this approach, this work aims to demonstrate the usefulness of random forest algorithms for identifying salient variables in complex social datasets for the study of migration and developing more powerful predictive models.

Study area

The study area for this research is located primarily in the southwest of rural Bangladesh. Bangladesh is located on the low-lying deltaic floodplain of the Ganges–Brahmaputra–Jamuna Delta, which includes the Ganges, Brahmaputra, Jamuna, Padma, and Meghna Rivers [43]. Bangladesh is commonly considered one of the most vulnerable countries to climate change in the world [44, 45]. As in other delta regions, future climate change is expected to create additional stress and uncertainty in Bangladeshi communities through its interactions with natural hazards such as cyclones, flooding, waterlogging, salinity encroachment, and land erosion, as well as with natural resources, such as accreting land and freshwater supplies [3, 46–52]. Environmental conditions in Bangladesh pose severe challenges to rural communities, where approximately two-thirds of workers (representing nearly half of all workers in Bangladesh) depend on agriculture as their primary source of livelihood [53].

Migration has long been a way of life in Bangladesh, where it serves to diversify livelihood activities and to adapt to environmental and economic stress [8, 24, 54, 55]. Rural to urban migration is the most prevalent form of migration in Bangladesh [56], especially temporary migration to adapt to seasonal poverty [57]. Environmentally induced migration has also been widely studied in Bangladesh [5, 6, 29, 56, 58–61]. Much of this research focuses on extreme weather events, such as cyclones and floods [6, 62]. Other research considers slower environmental change such as salinity encroachment, temperature change, and precipitation [5, 10, 63]. Recently, Chen and Mueller found that salinity encroachment could be a powerful driver of migration in Bangladesh due to impacts on agriculture [59]. While much of the existing body of work in migration has focused on individual factors, this paper is taking a holistic look at how different factors contribute to migration.

Data

Data for this work come from two distinct social surveys collected from households in southwestern Bangladesh. The first survey, Survey 1, comes from Integrated Social Environmental Engineering Bangladesh (ISEE-B), a multi-disciplinary collaborative project to study community resilience to environmental change in coastal Bangladesh [47]. The data were collected in household interviews in 26 communities in the southwest region of Bangladesh from March through April 2014. The 26 communities are a sample from a set of 75 communities identified based on the properties of their aquifer which gave them inadequate access to fresh safe water [47]. The 26 study communities were identified for variability on three dimensions: nongovernment organization partner, geographic dispersion, and ground water quality. Each community was a neighborhood of approximately 100 households, generally sharing a common water source. After doing a geolocated photo census of each community, 20–50% of the households were randomly selected for study. In total, 1204 heads of household were interviewed about their household's demographics, sources of livelihood, sources of water, environmental stressors, and other factors. Additional questions measured their individual risk perception, sense of social cohesion, and political trust. The original dataset consists of 1204 observations and 1456 variables.

The second survey of households, Survey 2, was also collected in the southwest region of Bangladesh by the Bangladesh Environment and Migration Survey (BEMS) (Online Resources 1). This survey contains migration, employment, and livelihood histories on more than 3000 individuals affiliated with 1695 households. The data represent 1695 randomly sampled households in nine sites in Bangladesh, which were surveyed in 2014. The survey specifically asks for histories of migration within Bangladesh, to India, and to any other country [29]. A full description of Survey 2 design and implementation can be found in previously published work [29, 60]. Here, we focus only on each household's reported migrations internal to Bangladesh. The original dataset consists of 1695 observations of 1997 distinct variables.

Because of their distinct purposes, Survey 1 and Survey 2 ask different questions and include data from different communities, so they present two unique opportunities to identify salient variables and test the performance of machine learning methods in discerning and predicting migration. Figure 1 shows the geographic locations of households surveyed in Survey 1 and Survey 2.

The structure of Survey 1 is such that the response variables are Boolean variables indicating the respondent's answer to yes or no questions about migration: "Have

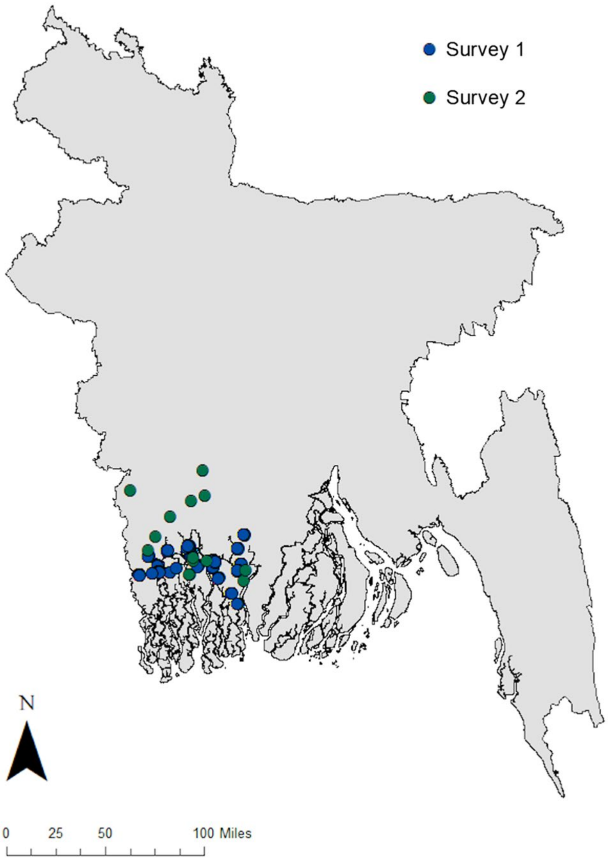

Fig. 1 Map of Bangladesh with locations of households surveyed by Survey 1 and Survey

you ever moved your household temporarily to another place within this village because of an environmental event?"; "Has anyone in your household ever moved for education?"; "Has anyone in your household ever moved for health care?"; "Has anyone in your household ever moved for commerce/ trading?"; and "Has anyone in your household ever moved to visit relatives?" These questions were used to assess migration for environmental reasons, for education, for health, for trade, and to visit relatives, respectively. Thus, Survey 1 also allows us to assess random forests' ability to compare the salient variables associated with migration for different reasons.

Survey 2 asks respondents to recall the total number of migrations that any member of the household has made, without attributing underlying motivation. This provides the total number of migration trips per household, normalized by total person-years. Person-years were calculated for each member of the household, beginning at age 11, which is the age that many Bangladeshis begin migrating for livelihood opportunities, until 2014 when the survey was collected [29]. Our analysis of Survey 2 considers the response variable to be annual probability of making a migration, which is represented as a continuous variable at the household level, and identifies salient variables that predict this probability.

Methods

Model selection

The first step of our analysis was to compare different approaches to analyze the survey data. We compared random forest models, multiple logistic regression, and support vector machines (SVMs) with a radial basis kernel function. Random forest modeling, which is a tree-based method, is described in more detail in the next section. Multiple logistic regression is a generalized linear model that fits coefficients to predictors to fit the logit transformation of the probability of the event of interest, which is then converted to a dichotomous prediction of the outcome variable [64]. SVMs are another class of supervised machine learning methods used for classification and regression. SVMs use hyperplanes in a high-dimensional feature space to optimally divide the data into different classes based on the response variable [65].

All three models were fit to each of the five motivations of migration in Survey 1: environmental, education, health, trade, and to visit relatives, for a total of 15 models. Each model was trained on a random sample comprising 80% of the data set, and tested on the remaining 20% to assess predictive accuracy. For random forests and SVMs, relevant model parameters were tuned by minimizing out-of-sample error. Table 1 shows the prediction error for each model on the test data in percent error.

Imputing missing data in Survey 1

Before further analyzing Survey 1, data related to the household respondent were selected from the household roster, and summary variables related to household size, household education, and livelihood were developed. We also eliminated questions that only applied to part of the sample, keeping only variables that were relevant to the full data set. The remaining variables were then screened manually, and variables that were likely not missing at random, or for which there were known problems during data collection, were dropped.

Table 1 Test data prediction error (percent error) for models fit to each type of migration in Survey 1

	Environmental	Education	Health	Trade	Visit relatives
Null	40.5	15.6	38.5	19.9	44.4
Logistic regression	47.1	44.9	44.1	43.4	42.6
SVM	36.0	16.2	36.0	19.9	41.2
Random forest	35.5	14.7	33.1	19.9	33.8

The resulting subset of data consisted of 1184 observations of 730 variables. Within this subset of the original survey, approximately 1.5% of data across all variables and rows were missing. Even after dropping columns that were not relevant to all households from the subset of Survey 1, restricting the analysis to complete cases would have needlessly lost information in the partial cases. We imputed missing variables in partial cases using multiple imputation, which allowed us to assess the stochastic uncertainty associated with the imputation process [66, 67].

Before imputing, the data were filtered to consider only variables with less than 12% missingness, which was a threshold that maintained 711 of the 730 variables. Imputations of missing data were then conducted using the *mice* (Multivariate Imputations by Chained Equations) package in R [67]. To accommodate both categorical and continuous data, a random forest imputation method was used to impute missing data 10 times. This resulted in 10 unique, complete datasets to be used in analysis.

Survey 2 did not have significant missing data, and therefore imputations were not necessary before assessing variable importance.

Random forest models for variable importance

Random forest models are an ensemble method of decision trees. They work by fitting many decision trees with random subsets of the data and then averaging across them for a final prediction. Thus, random forest models can achieve high predictive accuracy while avoiding overfitting the data. As previously mentioned, one strength of random forest models, especially over other "black box" statistical models, is their ability to assess variable importance and account for complex, nonlinear interactions between variables. They are also able to take inputs of categorical, factored, or continuous data without requiring dummy variables or scaled data. This makes them especially appealing tools for analyzing large social surveys and studying complex challenges such as migration.

For each survey and each model, the data were split into a training set, which consisted of a random subset representing 80% of the data and a holdout set, comprising the remaining 20%. The *randomForest* package in R was used to fit random forest models to the training data [68]. For Survey 1, 10 random forest classification models were fitted (one for each imputed dataset) for each of the five types of migration (environmental, education, health, trade, and to visit relatives), using a binary outcome variable indicating whether or not a respondent migrated. Each model used the same subset of the data as the training and holdout sets. For Survey 2, 10 random forest regression models, this time each using a different random subset of the data as the training and holdout sets, were fitted to the continuous outcome variable of total internal migration trips per household normalized by person-years. For each model, the parameter for the number of variables to be randomly sampled at each split was tuned by minimizing the out-of-sample error.

For each of the five types of migration in Survey 1, variable importance was ranked by averaging across the 10 imputed datasets. Variable importance for the model of Survey 2 was also ranked and averaged across the 10 complete models.

Tuned and fitted models were then validated using the testing data, which consisted of the remaining 20% of the data that were not used for training.

For Survey 1, variable importance is given by mean decrease in Gini index, which measures a specific variables ability to correctly classify an outcome. This use of a Gini index is completely different from the common use of the index for assessing economic inequality. For classification random forest models, the Gini index is a weighted measure of how much a specific variable contributes to decreasing the variance in the outcome, and, therefore, measures how much of the final model's predictive power can be attributed to that variable. The Gini index is not comparable from one analysis to another, but is useful as a comparative measure between variables within a given analysis, and allows us to rank variables by their contribution to the overall model [69]. For regression random forest models, which were used to analyze Survey 2, importance is calculated using node impurity, which also represents the contribution of a specific variable to decreasing the variance in the outcome [69].

Results

Variable importance

Figure 2 shows the 15 most important variables for random forest models predicting environmental migration, migration for education, migration for health, migration for trade, and migration to visit relatives from Survey 1. Fifteen variables are displayed because it was consistently found that below this cutoff there was very little difference in variable importance. However, full results from this analysis provide a ranked list of the variable importance of every survey variable.

In these figures, variable importance decreases from top to bottom. Colors in the figures are used to show similarities and differences across the five types of migration studied and to highlight the uniqueness of the variable. Colors represent occurrence: the number of times a variable occurs in the top 15 most important variables for all types of migration: An occurrence of one (red) means that a variable was important only for that type of migration. An occurrence of five (blue) means that the variable was important for models of all five types of migration.

In addition, Table 2 shows salient variables from all model results grouped into higher level categories of variables related to migration ("Migration"), livelihood and wealth ("Livelihood"), community-level variables ("Community"), infrastructural support ("Infrastructural Support"), level of trust in others including community and government ("Trust"), personal- and household-level demographics ("Personal"), and perceptions of locus of control ("Control"). These categories are useful to begin to identify differences in salience across the different types of migration. Table 2 lists the English translations of the actual survey questions corresponding to the variables in Fig. 2, together with the high-level categories and the models in which the variable appears.

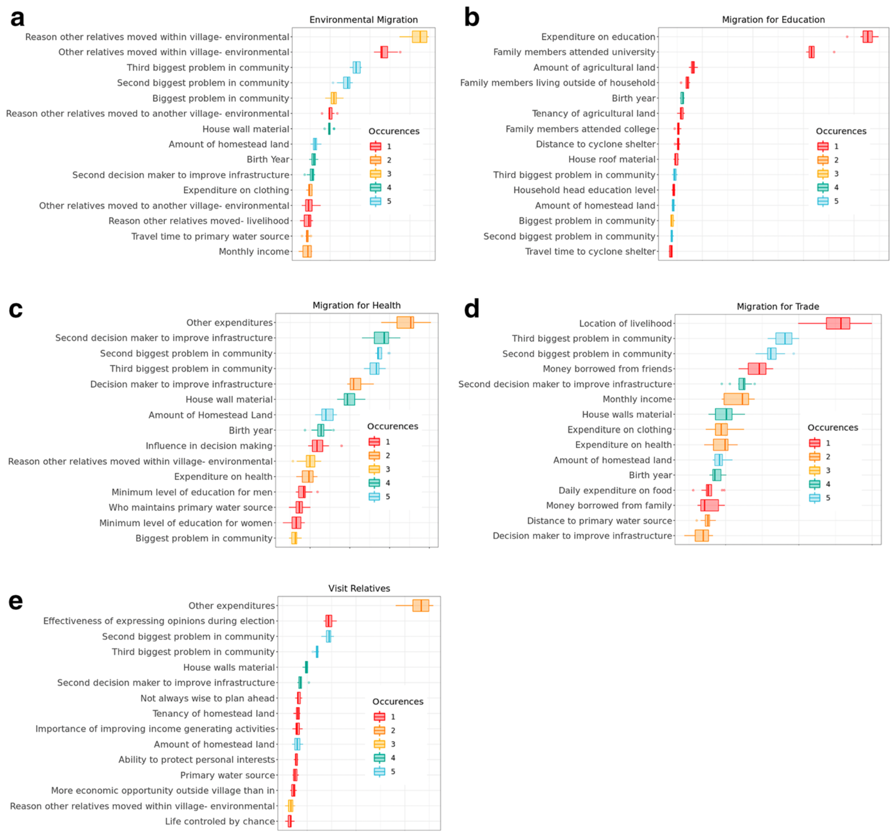

Fig. 2 Top 15 variables of importance identified by random forest models from top to bottom by mean decrease in Gini Index for environmental migration (**a**), migration for education (**b**), migration for health (**c**), migration for trade (**d**), and migration to visit relatives (**e**). Colors represent how many times a specific variable was in the top 15 most important variables for another model

Figure 3 shows the 15 most important variables in the random forest model of Survey 2. Table 3 presents English translation of the survey questions corresponding to the variables listed in Fig. 3.

Predictive accuracy

For Survey 1, predictive accuracy was assessed by error rates when the models were applied to the holdout test data. The error rate represents the fraction of total predictions that the random forest predicted incorrectly and a lower error rate indicates better model performance. Figure 4 shows the error rates for each of the five types of migration assessed in Survey 1. The model for predicting migration for education performs the best, with a mean error rate of 14.8%; while, the model for predicting environmental migration performs the most poor with a mean error rate of 33.8%.

Table 2 Variables of importance identified by random forest models of Survey 1

Category	Survey question	Variable name	Models
Migration	Have any of your other relatives not living with you now ever moved their whole household temporarily to another place within this village because of an environmental event?	Other relatives moved within village—environmental	Environmental
Migration	Thinking of the event that caused your family to move as you have just said, what was it? (1988 flood, Bhola cyclone, Sidr cyclone, Aila cyclone, other)	Reason other relatives moved within village—environmental	Environmental, health, visit relatives
Migration	Have any of your other relatives not living with you now ever moved their whole household temporarily to another village near here because of an environmental event?	Other relatives moved to another village—environmental	Environmental
Migration	Thinking of the event that caused your family to move as you have just said, what was it? (1988 flood, Bhola cyclone, Sidr cyclone, Aila cyclone, other)	Reason other relatives moved to another village—environmental	Environmental
Migration	Have any of your other relatives not living with you now ever moved their whole household permanently to another place because they could not make a livelihood here?	Reason other relatives moved—livelihood	Environmental
Community	As a community member here, what problems do you face regularly? In other words, what are the top 3 problems of this community for you?	Biggest problem in community	Environmental
Community	As a community member here, what problems do you face regularly? In other words, what are the top 3 problems of this community for you?	Second biggest problem in community	All
Community	As a community member here, what problems do you face regularly? In other words, what are the top 3 problems of this community for you?	Third biggest problem in community	All

Table 2 (continued)

Category	Survey question	Variable name	Models
Community	Imagine that the village receives funds to invest in improving infrastructure in the village. A decision needs to be made about how the funds should be spent. Who would play the biggest role in resolving the dispute?	First decision maker to improve infrastructure	Health, trade
Community	Who would play the second biggest role?	Second decision maker to improve infrastructure	Environmental, health, trade, visit relatives
Community	What is the minimum level of education that a man can have in your village?	Minimum level of education for men	Health
Community	What is the minimum level of education that a woman can have in your village?	Minimum level of education for women	Health
Community	Who maintains the water source now?	Who maintains primary water source	Health
Community	What do you think should be implemented/improved to help you addressing your future needs related to disaster? Income generating activities	Importance of improving income generating activities	Visit relatives
Community	There is more economic opportunity outside my village than in it	More economic opportunity outside community than in	Visit relatives
Control	When decisions are made on issues that affect all villagers, do you feel that you are influential in determining the outcome?	Influence in decision-making	Health
Control	If you had concerns about how things were going in your village, tell me whether or not you think these things would help—Express opinions in elections	Effectiveness of expressing opinions during elections	Visit relatives
Control	It's not always wise for me to plan too far ahead because many things turn out to be a matter of good or bad fortune	Not always wise to plan ahead	Visit relatives

Table 2 (continued)

Category	Survey question	Variable name	Models
Control	I am usually able to protect my personal interests (I can usually look after what is important to me)	Ability to protect personal interests	Visit relatives
Control	To a great extent my life is controlled by accidental/chance happenings	Life controlled by chance	Visit relatives
Livelihood	Residential house wall construction material	House wall material	Environmental, health, trade, visit relatives
Livelihood	Residential house roof construction material	House roof material	Education
Livelihood	Land (homestead) in decimal	Amount of homestead land	All
Livelihood	Tenancy (homestead)	Tenancy of homestead land	Visit relatives
Livelihood	Expenses per year, clothing	Expenditure on clothing	Trade
Livelihood	Expenses per year, health	Expenditure on health	Health, Trade
Livelihood	Please let me know how much you expend daily for giving food to your family members?	Expenditure on food	Trade
Livelihood	Expenses per year, education	Expenditure on education	Education
Livelihood	Expenses per year, other	Other expenses	Health, visit relatives
Livelihood	How much did you make in taka per month?	Monthly income	Environmental, trade
Livelihood	Land (agriculture) in decimal	Amount of agricultural land	Education
Livelihood	Tenancy (agricultural)	Tenancy of agricultural land	Education
Livelihood	Source of Income, location	Location of livelihood	Trade
Personal	Year of birth	Birth year	Environmental, education, health, trade
Personal	Head of household level of education	Household head education level	Education
Personal	Level of education, university	Family members attended university	Education
Personal	Level of education, college	Family members attended college	Education
Personal	Lives in the household—no	Family members living outside of household	Education
Personal	Travel time to source (in minutes)	Distance to primary water source	Environmental, trade

Table 2 (continued)

Category	Survey question	Variable name	Models
Personal	What are the drinking water sources here?	Primary water source	Trade
Infrastructural Support	Number of minutes it takes to go to cyclone shelter on foot in day light?	Travel time to cyclone shelter	Education
Infrastructural Support	How far is your home from a cyclone shelter? (in km)	Distance to cyclone shelter	Education
Trust	If you suddenly needed a small amount of money, enough to pay for expenses for your household for one week, is there at least one person from the following groups that you could turn to who would be willing to provide this money?—Friends/ neighbors	Money borrowed from friends	Trade
Trust	If you suddenly needed a small amount of money, enough to pay for expenses for your household for one week, is there at least one person from the following groups that you could turn to who would be willing to provide this money?—Members of my extended family/relatives	Money borrowed from family	Trade

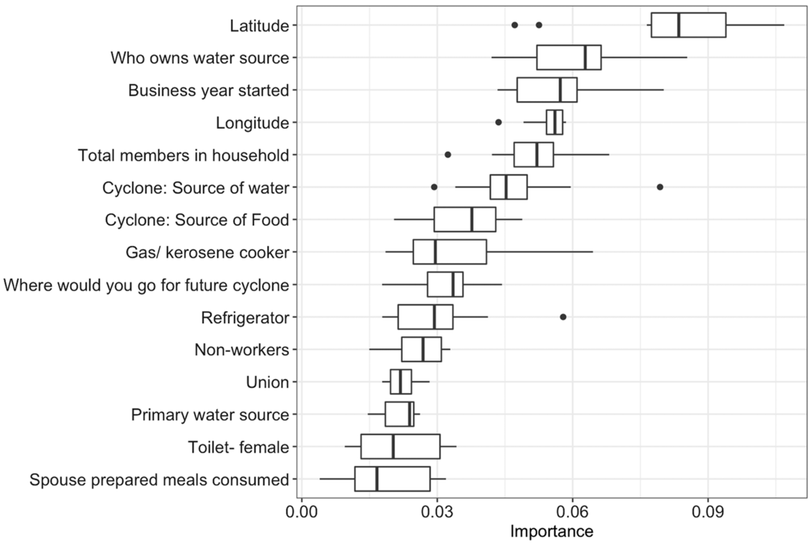

Fig. 3 Top 15 variables of importance in Survey 2 identified by random forest models of total household migrations normalized by person-years

Table 3 Variables of importance identified by random forest model of migration in Survey 2

Variable name	Survey question
Latitude	Household latitude
Water Sources: Who owns?	Who owns the primary water source?
Business: year started	What is the year that your business was started?
Longitude	Household longitude
Household: total number of members	How many household members are living in the home?
Cyclone: source of water	What was your principle source of water during the last cyclone?
Cyclone: Source of food	What was your principle source of food during the last cyclone?
Gas/kerosene cooker	Do you own a gas or kerosene cooker?
Where would you go for future cyclone	Where would you go if there was a future cyclone?
Refrigerator	Do you own a refrigerator?
Non-workers	What is the total number of non-workers in the household?
Union	Household union (indication of community)
Primary water source	What is the household's primary water source?
Toilet—female	What kind of toilet facility do female household members use?
Spouse-prepared meals consumed	Has household consumed prepared meals? If yes, who? Spouse

For Survey 2, error was assessed by root-mean-square error (RMSE) between the model predicted number of household migrations and the actual household migrations in the holdout data. To assess RMSE, the predicted number of migrations was

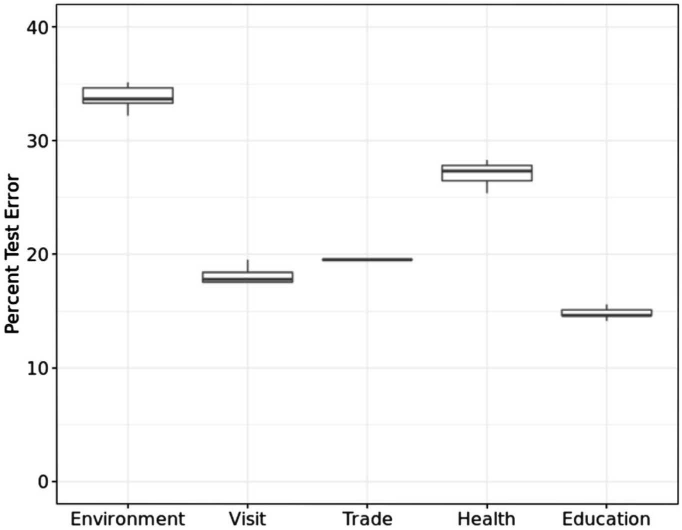

Fig. 4 Percent test errors for each random forest model of migration assessed in Survey 1. Test errors are calculated based on predictions of test data from models fitted with training data. The figure shows that the model of migration for education has the lowest test error, while the model of environmental migration has the highest percent test error. These differences represent that random forests' predictive abilities vary based on outcome variables and underlying patterns in data

rounded up to the nearest integer if the prediction was greater than one trip, and rounded down to zero if the prediction was less than one trip. This was to account for the fact that all of the holdout data migrations are reported as integer values. The mean RMSE of models was 2.22 with a standard deviation of 0.05. This is compared to an average mean of actual holdout set household migrations of approximately 3 migrations with a standard deviation of 3. For comparison, a naïve model that assumes that each household simply has the mean number of migrations has a mean RMSE of 2.27.

Discussion

The analysis of variable importance from random forest models reveals similarities and differences between the variables associated with different types of migration in Survey 1: environmental migration, migration for education, migration for trade, migration for health, and migration to visit relatives. For all five of the models, possible proxies for wealth or socio-economic status such the amount of homestead land owned were among the most important variables that predict the migration outcome variable. This analysis only establishes associations among variables and cannot speak to causal connections, but we can compare the salient variables identified here with causal relationships discussed in other work. The material of the respondent's home was important in models of four of the five types of migration. Previous research has indicated that livelihood and economic opportunity can greatly motivate or limit movement [70, 71]. Perceived issues in the community were also important across all of the models ("Biggest problem in community", "Second biggest problem in community", and "Third biggest problem in community"), suggesting that perceptions and satisfaction with one's home are also importantly associated with migration, regardless of the dominant motivation. Birth year of the head of household was also important for many of the models, suggesting that age is likely an important personal factor associated with migration decisions.

Assessing the individual models more closely provides additional insights into the differences between the models of migration as a result of varying self-reported motivations. Figure 5 shows high-level differences between models based on the frequency of important variables categorized by theme. From this analysis, we see that the response variable for environmental migration is uniquely associated with knowing others who have also migrated for environmental reasons, as the first two most important variables reflect this. Past research has shown that the barrier to migrate can be significantly lowered by potential migrants having social connections with others who have migrated in the past [9, 72, 73]. It is also noteworthy that there were not any explicitly environmental variables amongst the most important variables for environmental migration. This reinforces the common understanding that even when environmental pressures impact migration, they are rarely the only driver [7].

In contrast to environmental migration, the model of migration for education is uniquely influenced by variables that relate to household education level, such as the annual household expenditures on education and the number of household members who attended college and university. In addition to these variables related

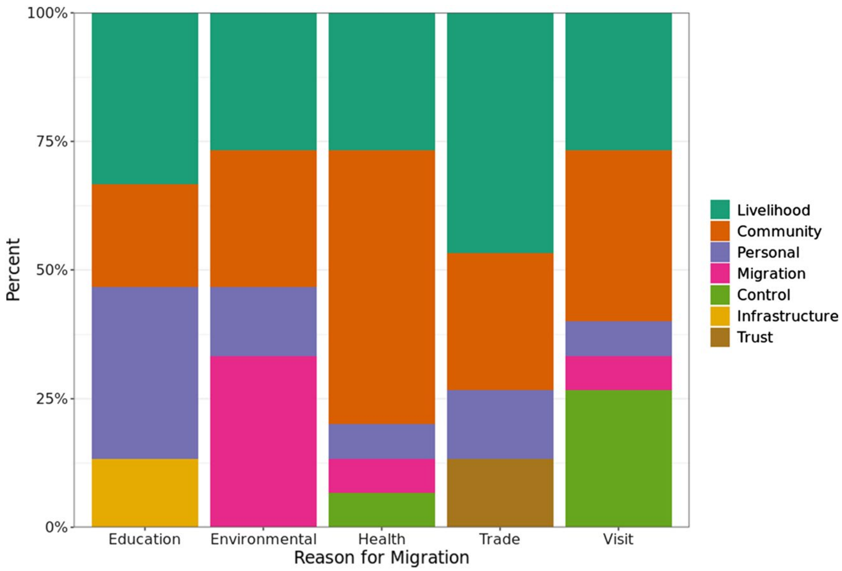

Fig. 5 Overview of categories of variables present in top 15 variables of importance for models of migration assessed with Survey 1. The legend shows the colors corresponding to the categories of variables

to education, the models have identified variables related to socioeconomic status and access to infrastructure, such as tenancy of agricultural land and distance to a cyclone shelter as important variables.

Other nuances in the important variables in the models of migration for health care, migration for trade, and migration to visit relatives further demonstrate the strength of random forest models in identifying nuances and complex relationships in data. For instance, migration for trade or commerce is strongly associated with the location of the respondent's primary source of livelihood. Migration for health is uniquely associated with factors that reflect community-level conditions, such as the minimum level of education that a community member would be able to obtain. Migration to visit relatives is uniquely associated with variables related to locus of control such as the respondent's faith in their ability to plan ahead and ability to express themselves during elections.

The test error rates for the random forest models for each type of migration may also be telling. Models of migration for education have the lowest error rate, suggesting that this type of migration is easier for a model to predict. Models of environmental migration, in contrast, have the highest error rate, followed by migration for health care. There seems to be a clear divide between models to predict migration for education, migration to visit relatives, and migration for trade, which performed relatively well; and models for environmental migration and migration for health care, which were considerably less accurate. One possible explanation is that environmental events or health challenges are responses to specific negative events, such as severe storms or illness, that cannot be predicted in advance. Education, visiting relatives, and economic opportunities, on the other hand, may be driven more by ongoing positive conditions at a destination that draw someone to move and that are more predictable and subject to advance planning [8, 36].

The analysis of Survey 2 also demonstrates the power of random forest models for providing insights into important variables associated with migration in Bangladesh, this time where the outcome variable is continuous rather than dichotomous. Here, the first and fourth most important variables are latitude and longitude, which are important even when controlling for the survey community. We speculate that the importance of latitude and longitude may reflect spatial variations in environmental and geographic conditions that affect migration. For example, soil salinity varies strongly on a north–south gradient, and salinity has been shown to be an important factor influencing migration in Bangladesh [59]. Longitude correlates both with proximity to the Indian border and with proximity to a number of major rivers whose orientations are largely north–south.

In addition to latitude and longitude, several variables that suggest socio-economic status are important in the model. The second and thirteenth most important variables reflect ownership and the type, respectively, of the household's primary water source. In addition, the year a business was started as well as owning a refrigerator or a gas cooker all correlate with household wealth. Toilet facilities used by female members of the household, and whether or not the spouse of the household head consumes prepared meals are also likely to reflect socioeconomic status as well as gender empowerment, education, household wealth, and household employment [74]. As with the results from Survey 1, this supports the common understanding that livelihood and economic opportunity are important to migration decisions [70, 71]. Results from Survey 2 also suggest that household composition is important to migration: the size of the household and the number of non-workers in the household are both important, even though the number of migrations is normalized by the number of person-years for the household.

Finally, several variables related to the most recent cyclone were important in this analysis, including sources of food and water during the latest cyclone, as well as where a household would go in the event of a future cyclone. Cyclones have been widely studied as a driver of migration in Bangladesh [62, 75, 76]. Mallick and Vogt found that that male household members migrated towards cities to access livelihood opportunities after the cessation of emergency aid after the 2009 cyclone Aila [62].

This work provides initial evidence that random forest models can be promising tools for predicting migration from a large collection of covariates. Assessed by the accuracy of out-of-sample predictions, random forests outperformed logistic regressions and SVM models for Survey 1. This is likely due to the fact that tree-based models allow us to identify nonlinear and multimodal (nonmonotonic) interactions between variables that we cannot effectively model with generalized linear models or SVMs. When assessing the predictive accuracy of the model for Survey 2, we see that the random forest models only slightly outperform a naïve model. This is in large part due to the structure of the outcome data in Survey 2, which is a continuously distributed variable that is strongly zero inflated (most households do not migrate at all) with a very long tail (a few households have very high levels of migration). This complex structure of the data that the model is trying to predict means that it is very difficult to model and even define an appropriate measure of predictive accuracy. For these complex, continuous data, results show that the skill of the random forest model is very limited. However, future work will continue to investigate more sophisticated ways of modeling continuous data from a zero-inflated distribution.

In general, we see significant potential in the random forest models for improving predictive accuracy, primarily from the results of Survey 1. Predictive ability is valuable for informing future climate policy and adaptation strategies that aim to address migration [9, 58, 77, 78]. Where prediction might be more important than understanding underlying drivers, especially when providing information to policymakers, random forest should be explored further as a possible tool. As modelers and researchers continue to work to improve their ability to predict migration, random forest models have much to offer for future analysis. However, modelers should continue to develop more sophisticated methods, as our results show that even the random forest model still has limited predictive accuracy, especially when predicting the number of migrations versus a dichotomous classification of a household's mobility. It is possible that the limited predictive power of our random forest models was partially due to the absence of potential push factors (such as cyclones or illnesses) in models of migration for environmental and health reasons. Explicitly including push factors in future analyses may improve model predictive accuracy. These push factors are not themselves very predictable: there is great stochastic uncertainty in the distribution of cyclone strikes. However, even when the push variables are not themselves predictable, predictive models that incorporate them via conditional probabilities (e.g., what is the probability of migration in the event that a cyclone strikes?) can be useful for disaster planning and "what-if" policy analyses.

This work demonstrates that random forest models can help researchers identify salient variables from large social surveys when studying migration. This is especially useful when dealing with large, complex datasets from social surveys, where it can be challenging to decide which variables are worthwhile for further investigation. For cases of categorical as well as continuous outcome variables, random forest models were able to identify a small number of important predictors of migration from sets of more than 700 independent variables. From these important predictor variables, we were able to provide insights into the underlying patterns in the datasets and, thus, identify nuances in the possible drivers of different kinds of migration in southwestern Bangladeshi communities.

Conclusion

This work is a novel application of machine learning techniques to social survey data, and our results highlight the usefulness of such methods to identify important variables in such large, complex datasets. By applying these methods to the study of environmental migration from household surveys, our results show that random forest models can be useful tools for researchers studying migration, especially environmental migration, where the theory is not clearly established or varies considerably by location.

One downside of the random forest models, however, is that although they can quantify variable importance, they do not provide simple explanations of the ways in which the individual predictors connect to the outcome variable. For example, we know from random forest outputs that latitude is important to predicting normalized total number of household internal migration trips, but there is not a simple relationship in which either higher or lower latitudes are associated with migration. Once important variables are identified, a combination of theory and traditional regression methods may be useful in identifying more directly how those variables relate to migration. Thus, random forest models are not the final answer to assessing or modeling environmentally induced migration, but they can serve a valuable role in exploratory data analysis by providing insights into datasets, informing hypotheses, or testing theories.

Future work should continue to develop modeling methods that are able to assess and explain the complex relationship between environmental, economic, and political factors and how they contribute to migration decisions. Predictive accuracy should remain a priority for future research, but data-driven methods, such as random forest modeling, should work together with theoretical analysis that can provide fundamental understanding of the socio-environmental systems [22, 79, 80].

Acknowledgements This work was supported by the National Science Foundation Coupled Human-Natural Systems Grant No. 1716909 and Office of Naval Research through Grant No. N00014-11-1-0683.

Data availability More information about Survey 1 and Survey 2 is available upon request from the corresponding author. The questionnaire for Survey 2 is included in full in the Electronic Supplementary Materials (Online Resource 1).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Intergovernmental Panel on Climate Change. (2018). *Global warming of 1.5 °C*. Retrieved from https://www.ipcc.ch/report/sr15/.
2. Nicholls, R. J., Wong, P. P., Burkett, V., Codignotto, J. O., Hay, J., McLean, R. F., … Woodroffe, C. D. (2007). *Coastal systems and low-lying areas* (pp. 315–356). Cambridge University Press. Retrieved from https://pubs.er.usgs.gov/publication/70204340.
3. Nicholls, R. J., Wong, P. P., Burkett, V., Woodroffe, C. D., & Hay, J. (2008). Climate change and coastal vulnerability assessment: Scenarios for integrated assessment. *Sustainability Science, 3*(1), 89–102. https://doi.org/10.1007/s11625-008-0050-4.
4. Myers, N. (2002). Environmental refugees: A growing phenomenon of the 21st century. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 357*(1420), 609–613. https://doi.org/10.1098/rstb.2001.0953.
5. Call, M. A., Gray, C., Yunus, M., & Emch, M. (2017). Disruption, not displacement: Environmental variability and temporary migration in Bangladesh. *Global Environmental Change, 46*, 157–165. https://doi.org/10.1016/j.gloenvcha.2017.08.008.
6. Gray, C. L., & Mueller, V. (2012). Natural disasters and population mobility in Bangladesh. *Proceedings of the National Academy of Sciences United States of America, 109*(16), 6000–6005. https://doi.org/10.1073/pnas.1115944109.
7. Obokata, R., Veronis, L., & McLeman, R. (2014). Empirical research on international environmental migration: A systematic review. *Population and Environment, 36*(1), 111–135. https://doi.org/10.1007/s11111-014-0210-7.
8. Amrith, S. S. (2013). *Crossing the Bay of Bengal: The furies of nature and the fortunes of migrants*. Cambridge, MA: Harvard University Press.
9. Black, R., Adger, W. N., Arnell, N. W., Dercon, S., Geddes, A., & Thomas, D. (2011). Migration and global environmental change. *Global Environmental Change, 21*, S1–S2. https://doi.org/10.1016/j.gloenvcha.2011.10.005.
10. Perch-Nielsen, S. L., Bättig, B. M., & Imboden, D. (2008). Exploring the link between climate change and migration. *Climatic Change, 91*(3–4), 375–393. https://doi.org/10.1007/s10584-008-9416-y.
11. Cai, R., Feng, S., Oppenheimer, M., & Pytlikova, M. (2016). Climate variability and international migration: The importance of the agricultural linkage. *Journal of Environmental Economics and Management, 79*, 135–151. https://doi.org/10.1016/j.jeem.2016.06.005.
12. Black, R., Arnell, N. W., Adger, W. N., Thomas, D., & Geddes, A. (2013). Migration, immobility and displacement outcomes following extreme events. *Environmental Science & Policy, 27*, S32–S43. https://doi.org/10.1016/j.envsci.2012.09.001.
13. Cattaneo, C., & Peri, G. (2016). The migration response to increasing temperatures. *Journal of Development Economics, 122*, 127–146. https://doi.org/10.1016/j.jdeveco.2016.05.004.
14. 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.
15. 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 of the United States of America, 111*(27), 9780–9785. https://doi.org/10.1073/pnas.1317166111.

16. Gray, C., & Wise, E. (2016). Country-specific effects of climate variability on human migration. *Climatic Change, 135*(3–4), 555–568. https://doi.org/10.1007/s10584-015-1592-y.
17. Thiede, B., Gray, C., & Mueller, V. (2016). Climate variability and inter-provincial migration in South America, 1970–2011. *Global Environmental Change, 41*, 228–240. https://doi.org/10.1016/j.gloenvcha.2016.10.005.
18. Gray, C., & Mueller, V. (2012). Drought and population mobility in rural Ethiopia. *World Development, 40*(1), 134–145. https://doi.org/10.1016/j.worlddev.2011.05.023.
19. Gray, C. L., & Bilsborrow, R. E. (2014). Consequences of out-migration for land use in rural Ecuador. *Land Use Policy, 36*, 182–191. https://doi.org/10.1016/j.landusepol.2013.07.006.
20. Mueller, V., Gray, C., & Kosec, K. (2014). Heat stress increases long-term human migration in rural Pakistan. *Nature Climate Change, 4*(3), 182–185. https://doi.org/10.1038/nclimate2103.
21. Gray, C. L. (2011). Soil quality and human migration in Kenya and Uganda. *Global Environmental Change, 21*(2), 421–430. https://doi.org/10.1016/j.gloenvcha.2011.02.004.
22. Gray, C., Frankenberg, E., Gillespie, T., Sumantri, C., & Thomas, D. (2014). Studying displacement after a disaster using large scale survey methods: Sumatra after the 2004 tsunami. *Annals of the Association of American Geographers. Association of American Geographers, 104*(3), 594–612. https://doi.org/10.1080/00045608.2014.892351.
23. Marchiori, L., Maystadt, J.-F., & Schumacher, I. (2012). The impact of weather anomalies on migration in sub-Saharan Africa. *Journal of Environmental Economics and Management, 63*(3), 355–374. https://doi.org/10.1016/j.jeem.2012.02.001.
24. Alam, G. M. M., Alam, K., & Mushtaq, S. (2017). Climate change perceptions and local adaptation strategies of hazard-prone rural households in Bangladesh. *Climate Risk Management, 17*, 52–63. https://doi.org/10.1016/j.crm.2017.06.006.
25. Koubi, V., Spilker, G., Schaffer, L., & Bernauer, T. (2016). Environmental stressors and migration: Evidence from Vietnam. *World Development, 79*, 197–210. https://doi.org/10.1016/j.worlddev.2015.11.016.
26. Svart, L. M. (1976). Environmental preference migration: A review. *Geographical Review, 66*(3), 314–330. https://doi.org/10.2307/213888.
27. Farris, S. R. (2010). Interregional migration: The challenge for gender and development. *Development, 53*(1), 98–104. https://doi.org/10.1057/dev.2009.89.
28. Gioli, G., Khan, T., Bisht, S., & Scheffran, J. (2014). Migration as an adaptation strategy and its gendered implications: a case study from the Upper Indus Basin. *Mountain Research and Development, 34*(3), 255–265. https://doi.org/10.1659/MRD-JOURNAL-D-13-00089.1.
29. Donato, K. M., Massey, D. S., Donato, K. M., Carrico, A. R., Sisk, B., & Piya, B. (2016). Different but the same: How legal status affects international migration from Bangladesh. *The Annals of the American Academy of Political and Social Science, 666*(1), 203–218. https://doi.org/10.1177/0002716216650843.
30. Dustmann, C., & Okatenko, A. (2014). Out-migration, wealth constraints, and the quality of local amenities. *Journal of Development Economics, 110*, 52–63. https://doi.org/10.1016/j.jdeveco.2014.05.008.
31. McLeman, R. (2013). Developments in modelling of climate change-related migration. *Climatic Change, 117*(3), 599–611. https://doi.org/10.1007/s10584-012-0578-2.
32. Neumann, K., & Hilderink, H. (2015). Opportunities and challenges for investigating the environment-migration nexus. *Human Ecology, 43*(2), 309–322. https://doi.org/10.1007/s10745-015-9733-5.
33. Renaud, F. G., Dun, O., Warner, K., & Bogardi, J. (2011). A decision framework for environmentally induced migration: Framework for environmentally induced migration. *International Migration, 49*, e5–e29. https://doi.org/10.1111/j.1468-2435.2010.00678.x.
34. Hino, M., Field, C. B., & Mach, K. J. (2017). Managed retreat as a response to natural hazard risk. *Nature Climate Change, 7*(5), 364–370. https://doi.org/10.1038/nclimate3252.
35. Cai, R., & Oppenheimer, M. (2013). An Agent-based model of climate-induced agricultural labor migration. In *2013 Annual meeting, August* (pp. 4–6).
36. Hassani-Mahmooei, B., & Parris, B. W. (2012). Climate change and internal migration patterns in Bangladesh: An agent-based model. *Environment and Development Economics, 17*(06), 763–780. https://doi.org/10.1017/S1355770X12000290.
37. Kniveton, D., Smith, C., & Wood, S. (2011). Agent-based model simulations of future changes in migration flows for Burkina Faso. *Global Environmental Change, 21*, S34–S40. https://doi.org/10.1016/j.gloenvcha.2011.09.006.

38. Silveira, J. J., Espindola, A. L., & Penna, T. J. P. (2006). An agent-based model to rural-urban migration analysis. *Physica A: Statistical Mechanics and its Applications, 364*, 445–456. https://doi.org/10.1016/j.physa.2005.08.055.
39. Smith, C. D. (2014). Modelling migration futures: Development and testing of the Rainfalls Agent-Based Migration Model—Tanzania. *Climate and Development, 6*(1), 77–91. https://doi.org/10.1080/17565529.2013.872593.
40. Fussell, E., Hunter, L. M., & Gray, C. L. (2014). Measuring the environmental dimensions of human migration: The demographer's toolkit. *Global Environmental Change, 28*, 182–191. https://doi.org/10.1016/j.gloenvcha.2014.07.001.
41. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1.
42. Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1.
43. Passalacqua, P., Lanzoni, S., Paola, C., & Rinaldo, A. (2013). Geomorphic signatures of deltaic processes and vegetation: The Ganges-Brahmaputra-Jamuna case study. *Journal of Geophysical Research: Earth Surface, 118*(3), 1838–1849. https://doi.org/10.1002/jgrf.20128.
44. Walsham, M. (2010). *Assessing the Evidence: Environment*. International Organization for Migration: Climate Change and Migration in Bangladesh.
45. Black, R., Kniveton, D., Skeldon, R., Coppard, D., Murata, A., & Schmidt-Verkerk, K. (2008). *Demographics and climate change: Future trends and their policy implications for migration. Development Research Centre on Migration, Globalisation and Poverty*. Brighton: University of Sussex.
46. Nicholls, R. J., Wong, P. P., Burkett, V., Codignotto, J., Hay, J., McLean, R., … Arblaster, J. (2007). *Coastal systems and low-lying areas*. Cambridge : Cambridge University Press.
47. Ackerly, B. A., Anam, M. M., & Gilligan, J. (2015). Environment, political economies and livelihood change. In B. Mallick & B. Etzold (Eds.), *Environment, migration and adaptation: Evidence and politics of climate change in Bangladesh*. Dhaka, Bangladesh: AH Development Publishing House (AHDPH). Retrieved from https://eprints.qut.edu.au/84192/.
48. Auerbach, L. W., Goodbred Jr, S. L., Mondal, D. R., Wilson, C. A., Ahmed, K. R., Roy, K., … Ackerly, B. A. (2015). Flood risk of natural and embanked landscapes on the Ganges–Brahmaputra tidal delta plain. *Nature Climate Change*, *5*(2), 153–157. https://doi.org/10.1038/nclimate2472.
49. Benneyworth, L., Gilligan, J., Ayers, J. C., Goodbred, S., George, G., Carrico, A., … Piya, B. (2016). Drinking water insecurity: water quality and access in coastal south-western Bangladesh. *International Journal of Environmental Health Research*, *26*(5–6), 508–524. https://doi.org/10.1080/09603123.2016.1194383.
50. Brammer, H. (2014). Bangladesh's dynamic coastal regions and sea-level rise. *Climate Risk Management, 1*, 51–62. https://doi.org/10.1016/j.crm.2013.10.001.
51. Tessler, Z. D., Vorosmarty, C. J., Grossberg, M., Gladkova, I., Aizenman, H., Syvitski, J. P. M., et al. (2015). Profiling risk and sustainability in coastal deltas of the world. *Science, 349*(6248), 638–643. https://doi.org/10.1126/science.aab3574.
52. Xu, J., Grumbine, R. E., Shrestha, A., Eriksson, M., Yang, X., Wang, Y., et al. (2009). The melting himalayas: Cascading effects of climate change on water, biodiversity, and livelihoods. *Conservation Biology, 23*(3), 520–530. https://doi.org/10.1111/j.1523-1739.2009.01237.x.
53. World Bank. (2016). Bangladesh: Growing the economy through advances in agriculture. *World Bank*. Text/HTML. Retrieved from https://projects-beta.worldbank.org/en/results/2016/10/07/bangladesh-growing-economy-through-advances-in-agriculture
54. Black, R., Natali, C., & Skinner, J. (2005). *Migration and inequality*. Washington, DC: World Bank.
55. Martin, M., Billah, M., Siddiqui, T., Abrar, C., Black, R., & Kniveton, D. (2014). Climate-related migration in rural Bangladesh: A behavioural model. *Population and Environment, 36*(1), 85–110. https://doi.org/10.1007/s11111-014-0207-2.
56. Afsar, R. (2003). Internal migration and the development nexus: The case of Bangladesh. In *Regional Conference on Migration, Development and Pro-Poor Policy Choices in Asia* (pp. 22–24).
57. Khandker, S. R. (2012). Seasonality of income and poverty in Bangladesh. *Journal of Development Economics, 97*(2), 244–256. https://doi.org/10.1016/j.jdeveco.2011.05.001.
58. Ahsan, R., Karuppannan, S., & Kellett, J. (2011). Climate migration and urban planning system: A study of Bangladesh. *Environmental Justice, 4*(3), 163–170. https://doi.org/10.1089/env.2011.0005.
59. Chen, J., & Mueller, V. (2018). Coastal climate change, soil salinity and human migration in Bangladesh. *Nature Climate Change*. https://doi.org/10.1038/s41558-018-0313-8.

60. Carrico, A. R., & Donato, K. (2019). Extreme weather and migration: Evidence from Bangladesh. *Population and Environment*. https://doi.org/10.1007/s11111-019-00322-9.
61. Bernzen, A., Jenkins, J. C., & Braun, B. (2019). Climate change-induced migration in coastal Bangladesh? A critical assessment of migration drivers in rural households under economic and environmental stress. *Geosciences, 9*(1), 51. https://doi.org/10.3390/geosciences9010051.
62. Mallick, B., & Vogt, J. (2014). Population displacement after cyclone and its consequences: Empirical evidence from coastal Bangladesh. *Natural Hazards, 73*(2), 191–212. https://doi.org/10.1007/s11069-013-0803-y.
63. Rakib, M. A., Sasaki, J., Matsuda, H., & Fukunaga, M. (2019). Severe salinity contamination in drinking water and associated human health hazards increase migration risk in the southwestern coastal part of Bangladesh. *Journal of Environmental Management, 240*, 238–248. https://doi.org/10.1016/j.jenvman.2019.03.101.
64. Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. New York: Wiley.
65. Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293–300. https://doi.org/10.1023/A:1018628609742.
66. Liu, Y., & De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International journal of statistics in medical research, 4*(3), 287–295. https://doi.org/10.6000/1929-6029.2015.04.03.7.
67. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(1), 1–67. https://doi.org/10.18637/jss.v045.i03.
68. Cutler, F. original by L. B. and A., & Wiener, R. port by A. L. and M. (2018). *randomForest: Breiman and Cutler's random forests for classification and regression*. Retrieved from https://CRAN.R-project.org/package=randomForest.
69. James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: with applications in R*. New York: Springer.
70. Adger, W. N., Arnell, N. W., Black, R., Dercon, S., Geddes, A., & Thomas, D. S. G. (2015). Focus on environmental risks and migration: Causes and consequences. *Environmental Research Letters, 10*(6), 060201. https://doi.org/10.1088/1748-9326/10/6/060201.
71. Bennett, G., Thomas, S. M., & Beddington, J. R. (2011). Migration as adaptation. *Nature, 478*, 447–449.
72. Haug, S. (2008). Migration networks and migration decision-making. *Journal of Ethnic and Migration Studies, 34*(4), 585–605. https://doi.org/10.1080/13691830801961605.
73. Hunter, L. M., Murray, S., & Riosmena, F. (2013). Rainfall patterns and U.S. migration from rural Mexico. *The International Migration Review, 47*(4), 874–909. https://doi.org/10.1111/imre.12051.
74. Sraboni, E., Malapit, H. J., Quisumbing, A. R., & Ahmed, A. U. (2014). Women's empowerment in agriculture: What role for food security in Bangladesh? *World Development, 61*, 11–52. https://doi.org/10.1016/j.worlddev.2014.03.025.
75. Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, Md., Wetter, E., Iqbal, A., … Bengtsson, L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, *38*, 1–7. https://doi.org/10.1016/j.gloenvcha.2016.02.002.
76. Islam, M. R., & Hasan, M. (2016). Climate-induced human displacement: a case study of Cyclone Aila in the south-west coastal region of Bangladesh. *Natural Hazards, 81*(2), 1051–1071. https://doi.org/10.1007/s11069-015-2119-6.
77. Stern, N. (2006). The price of change. *IAEA Bulletin, 48*(2), 25.
78. Biermann, F., & Boas, I. (2010). Preparing for a warmer world: Towards a global governance system to protect climate refugees. *Global Environmental Politics, 10*, 60–88. https://doi.org/10.1162/glep.2010.10.1.60.
79. Hugo, G. (2011). Future demographic change and its interactions with migration and climate change. *Global Environmental Change, 21*, S21–S33. https://doi.org/10.1016/j.gloenvcha.2011.09.008.
80. Shayegh, S. (2017). Outward migration may alter population dynamics and income inequality. *Nature Climate Change, 7*(11), 828–832. https://doi.org/10.1038/nclimate3420.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Kelsea B. Best[1] · **Jonathan M. Gilligan**[1] · **Hiba Baroud**[2] · **Amanda R. Carrico**[3] · **Katharine M. Donato**[4] · **Brooke A. Ackerly**[5] · **Bishawjit Mallick**[6]

[1] Department of Earth and Environmental Sciences, Vanderbilt University, Nashville, TN, USA

[2] Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

[3] Environmental Studies Program, University of Colorado at Boulder, Boulder, CO, USA

[4] Institute for the Study of International Migration, Georgetown University, Washington, DC, USA

[5] Department of Political Science, Vanderbilt University, Nashville, TN, USA

[6] Environmental Sciences, Technische Universität Dresden, Dresden, Germany