Modeling Climate-Induced Refugee Migration: An Explainable Machine Learning Approach

Haodong Qi*[a], Alina Sîrbu[b], Rahman Momeni[c], Enes Hisam[c], Carlos Arcila-Calderón[d], Tuba Bircan[e], and Stefano Iacus[f]

[a]Malmö University, Sweden
[b]Department of Computer Science, University of Pisa, Italy
[c]GMV, UK
[d]University of Salamanca, Spain
[e]Vrije Universiteit Brussels, Belgium
[f]Harvard University, USA

Introduction

While claims that climate change induces refugee flows continue to dominate headlines and surface in public discourse, the scientific community has yet to establish robust evidence supporting these assertions [1, 2]. The existing literature generally suggests that migration responses to climate and environmental changes are complex and heterogeneous. They can vary depending on climatic conditions considered, data and methodologies employed, and geographical areas covered [3]. They may also be influenced by people's ability and resources to migrate [4], as well as by policies that either facilitate or impede migration as a form of adaptation [5]. To foster a more informed public discourse, there is a need for more holistic methodological approaches that can better account for diverse migration behavior. Here, we propose a novel machine learning approach designed to model climate-induced migration as a complex yet explainable system.

Our approach builds upon the dynamic elastic net (DynENet) algorithm recently developed for forecasting asylum-related migration [6]. However, two new features are introduced into this algorithm. First, we introduce a novel metric, the *penalized deviance ratio* (or PDR), for tuning the hyper-parameter that determines which predictor should be included or excluded in DynENet regression. This is critical for preventing predictive models from over-fitting, as it balances the trade-off between prediction accuracy and model complexity. Second, we enriched the original DynENet model with multi-dimensional and high resolution climate indicators derived from Earth Observation (EO) data. Unlike previous models using country-level

*Address: Nordenskiöldsgatan 1, 205 06 Malmö, Sweden. Email: haodong.qi@mau.se

climate indicators to explain/predict cross-border mobility [7, 3], we develop more granular indicators at sub-national level. This higher spatial resolution allows us to quest not only whether changing climatic conditions in a source country induce migration, but also where these drivers are likely to emerge (i.e., predicting potential sub-national hotspots of climate-induced migration).

To ease the computational burden of processing high spatial and temporal resolution EO data, we demonstrate the use of our machine learning approach with a case study rather than a global analysis. Here, we focus on Somalia, a country with an estimated 72% of the population living below the international poverty line (2.5 USD per day) [8]. More pressingly, extreme weather events have been increasingly frequent, further damaging already fragile agriculture-based livelihood systems, and causing widespread food insecurity and internal displacement.

The predictors in our DynENet model are extracted from two sources of data. First, we process EO data to derive climate indicators at the district level including: i) Standardized Precipitation and Evapotranspiration Index (SPEI); and ii) Soil Moisture Index (SMI). Second, we derive a large set of district-level economic and socio-political indicators from the Global Database of Events, Language, and Tone (GDELT) – the largest, most comprehensive, and highest resolution open database of human society (https://www.gdeltproject.org). Specifically, we group different types of events curated in GDELT into five broad categories (social, economic, political, governance, conflict). The grouping is based on the Conflict and Mediation Event Observations (CAMEO) codebook [9]. These indicators allow our model to disentangle how worsening climate conditions may operate in tandem with economic, social, and political factors in shaping the migration patterns of Somali people.

The outcome of interest in our DynENet model is the intensity of Somali nationals to seek asylum in different European Union (EU) member states, measured by the asylum-seeking rate (ASR). To compute this measure, we normalize EUROSTAT's monthly asylum applications by the total population size in Somalia. The primary reason for using asylum applications, as opposed to migration statistics from other international organizations, is the higher time frequency (monthly). This is needed for DynENet to be operable, as the algorithm only exploits temporal variation which requires sufficiently long time-series data (more details in the Methods section). Secondarily, the focus on refugee migration may hold a myriad of implications for EU's asylum policy; since the unprecedented wave of refugee inflows in 2015, there has been an intensifying debate concerning whether changing climatic conditions have contributed to, and will amplify, asylum-related migration [7].

Model Parameter Estimates

As our models are flow-specific (i.e., each model represents a flow from Somalia to a EU country), the resulting parameter estimates for each predictor can vary across EU destinations. Moreover, as our predictors are constructed at the district level in Somalia, the estimate for a given factor (say soil moisture) can vary across these districts. To facilitate the interpretation of our estimates below, we put forward the

definitions of different types of parameters:

- Type I: a destination-district-specific parameter corresponds to how migration intensity to a specific destination would respond to a change in a given district-level predictor (e.g., how much ASR from Somalia to Germany would increase/decrease as a result of a change in soil moisture in the district of Eyl);
- Type II: a district-specific parameter represents how migration intensity to the EU would respond, on average, to a change in a given district-level predictor (e.g., how much ASR from Somalia to the EU would increase/decrease as a result of a change in soil moisture in the district of Eyl);
- Type III: a destination-specific parameter refers to how migration intensity to a specific EU destination would respond, on average, to changes in a predictor across all Somali districts (e.g., how much ASR from Somalia to Germany would increase/decrease as a result of changing soil moisture across all Somali districts).

Unlike standard least square regression, DynENet does not provide the statistical significance of parameter estimates. Instead, through the least absolute shrinkage and selection operator (LASSO), it retains the regressors that are important in predicting the outcome variable and eliminates (zeros out) the ones that are unimportant. Hence, we interpret those retained predictors as statistically important factors driving forced migration.

District-Specific Parameters

By aggregating parameters at sub-national level (i.e., converting parameters from Type-I to Type-II), we can explore how migration responses may differ across localities within a country of origin. Such differential responses can help predict potential hotspots of climate-induced refugee migration within an origin country. Figure 1 depicts the average elasticity of the Somalia–EU migration intensity with respect to different district-level stressors; red/green colors indicate that migration increases/decreases with worsening climatic, economic and/or socio-political conditions, i.e., positive/negative migration responses to adversities. White (or light yellow) indicates that the parameters are zeroed out by DynENet (or close to zero), i.e., inelastic responses.

A key pattern in Figure 1 is that soil moisture index (SMI), SPEI, and conflict have notably more non-zero district-level estimates (see N.Dist), compared to the remaining predictors. This difference implies that climate factors coupled with conflict situations are more quantitatively important in explaining refugee flows from Somalia to the EU. However, as most estimates of these three variables are close to zero, the averaged coefficients (Avg.Coef) are small. As a result, the qualitative importance of climatic conditions and conflict is low, particularly compared to the economic and governance variables.

Another important note in Figure 1 is the pervasive heterogeneity in terms of how the Somalia-EU refugee migration may respond to the district-level climate factors. Specifically, the directions of the SMI and SPEI elasticity estimates differ

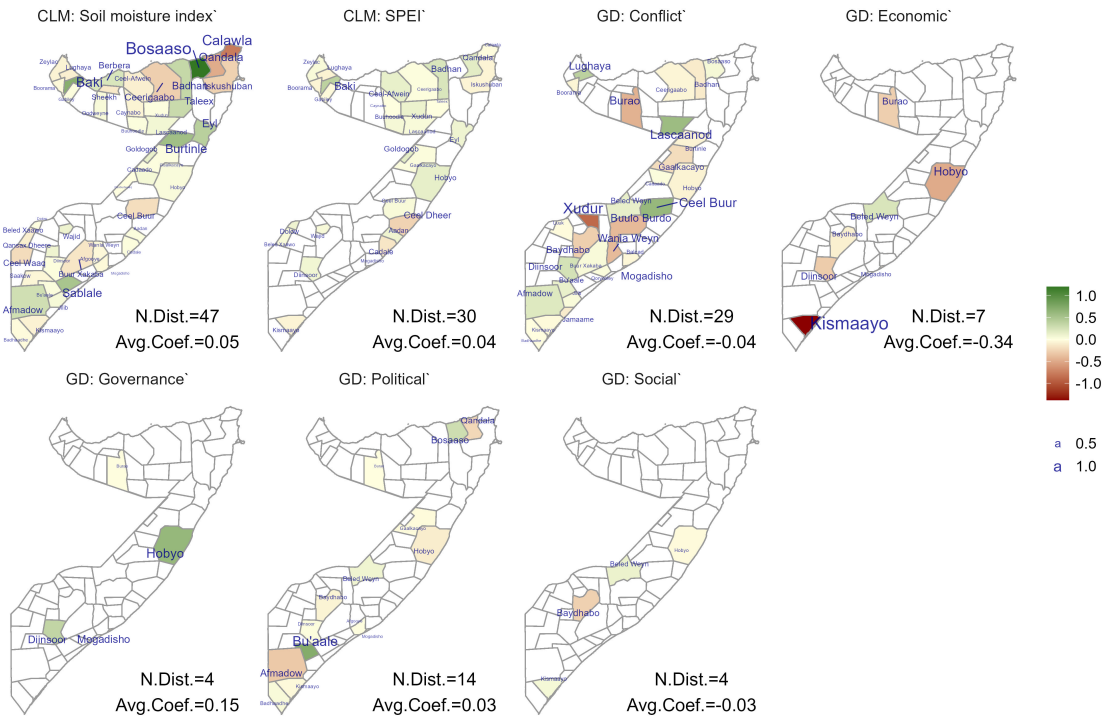

Figure 1: Average Parameter Estimates by Somali Districts

substantially across districts. These varying estimates indicate that SMI and SPEI may act as both push and trapping factors, when they drop (i.e., land becomes drier) in some districts, people might be forced to migrate, while in other places, people might be increasingly constrained to move. It could also be true that the livelihood impact of climate conditions might differ depending on locations. In some areas, drier land (particularly in the aftermaths of excessive rainfall or flood) may improve agriculture production, hence retain people in place. In other districts, however, it may indicate severe drought which is detrimental for agriculture, and therefore push people away.

Given the pervasive heterogeneity of elasticity, it is evident that the forces driving Somalis to seek asylum in the EU are not evenly distributed, rather they tend to be concentrated in a small number of districts. If we use -0.5 as a threshold elasticity to define strong migration response to adversities, merely three districts standout: Kismaayo, Xudur, and Calawla. As a commercial capital, Kismaayo's economic conditions played an important role in inducing refugee migration to the EU; a 10% decline in the economy of Kismaayo is associated with a 13.8% increase in the Somalia-EU ASR. The conflict situations in Xudur tend to exert strong impact on asylum-related flows to the EU; when the severity of conflict intensifies by 10%, ASR would increase by 8.9%. Finally, the Somalia-EU refugee flows tend to be strongly associated with changing soil conditions in the North-East part of Somalia; for a 10% drop in SMI in Calawla, the ASR to EU tends to increase by 7.8%.

The results presented above hold two important implications. First, there is only

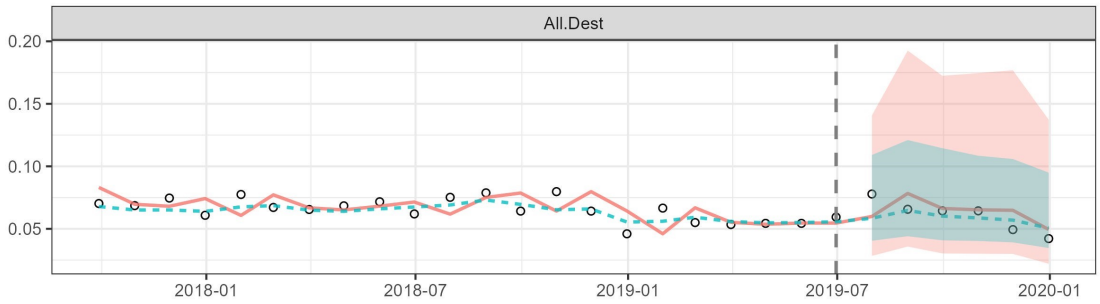

Figure 2: Predicted vs. Observed Asylum Seeking Rate

one district in Somalia that can be classified as a hotspot of climate-induced refugee migration, Calawla. However, the soil moisture elasticity is smaller in magnitude, compared to the conflict elasticity in Xudue and the economic elasticity in Kismaayo. Given these differences, we conjecture that, should the observed migration responses and the dire situations persist, the scale of climate-induced migration from Somalia to the EU is unlikely to be as profound as flows driven by conflict and economic stress. Furthermore, it is clear that in most Somali districts, worsening climatic, economic, and socio-political conditions do not or only mildly contributed to refugee migration. This implies that the majority of Somalis do not move to the EU or perhaps to other countries in responses to a variety of adversities, a phenomenon known as (resource-constrained) immobility [10, 11, 4].

Model Performances

To evaluate the predictive performance of our DynENet model, we compare its prediction errors with a benchmark model – first-order autoregressive or AR(1). Note, as our DynENet also contains an AR(1) component, the benchmark model here can be regarded as a restricted DynENet regression, i.e., imposing zeros on all coefficients in the DynENet model, except for the autoregressive component. Such a comparison essentially informs how well the climate, economic, and/or socio-political predictors retained by DynENet can explain and predict refugee migration from Somalia to EU member states.

Figure 2 illustrates how well models' predictions can resemble the actual trends of ASR from Somalia to the EU. The 95% confidence intervals are obtained through a bootstrapping procedure. DynENet models' forecasts are more stable and accurate (i.e., with lower variability and bias), compared to the benchmark model. The uncertainty associated with the forecast is also lower (due to smaller training errors). Given these results, we argue that our DynENet approach is more comprehensive, compared to time series extrapolation methods. It can provide nuanced insights into the climate-migration nexus, and, at the same time, accurately predict possible futures based on these insights.

Conclusions

In this article, we introduced a novel machine learning model which seeks to explain and predict climate-induced migration. To demonstrate its performance, we applied our model to a case study: asylum-related migration from Somalia to the EU. Leveraging satellite imagery data, we developed a large set of district-level climate indicators (soil moisture and precipitation-evaporation balance). With these indicators, we examined how various climate conditions drove Somali refugees to different EU member states, and more importantly, where these drivers emerged. We also tested our model's ability to forecast possible futures of refugee migration from Somalia to the EU.

A key finding from our analysis is that the factors that pushed Somalis to seek asylum in the EU are not evenly distributed, but concentrated in several districts. In particular, the flows are strongly linked to the level of soil moistrue in Calawla, the economic condition in Kismaayo, and to the conflict situation in Xudur. These results are non-trivial; while previous studies showed how migration may respond to environmental and/or socio-economic changes at the country-level [3], here we demonstrated that these responses can differ within a country of origin. Such differential responses are useful for detecting hotspots of climate-induced refugee migration. For example, by calibrating these differential responses, we can simulate the scale of refugee migration for each district in case of a climate shock, and map where the largest refugee flows might come from.

Moreover, the predictive performance of our DynENet model is satisfactory. Compared to an auto-regressive AR(1) model, DynENet exhibits better training and testing (forecasting) results for the vast majority of Somalia–EU refugee flows. Most importantly, when aggregating flow-specific predictions, our model can resemble the overall intensity of Somalis to seek asylum in the EU more closely, and provide more reliable assessment of the uncertainties associated with the forecasts. These results underscore the added value of our machine learning model, namely it can capture the complexity of climate-induced migration, but at the same time be explainable and predictive.

Methods

Data and Preprocessing

The case study of Somalia-EU refugee migration presented above relies on various sources of data. To measure the outcome variable, asylum seeking rate (ASR), we make use of the information on the number of first–time asylum applications lodged every month in different EU countries which is routinely compiled by EUROSTAT. These numbers are then normalized by the population size in Somalia obtained from the World Bank. ASR is defined as the number of first–time asylum seekers per 1000 people remained in Somalia. To align with the time window of our earth observation data, we use the asylum application data for the period January 2016 – December 2020. Moreover, some EU destinations are dropped from our analysis if the ASR time-series has too many missing values and/or insufficient variability (i.e.

no statistical information).

To measure climatic conditions, we developed various indicators at the district-level in Somalia. The standardized precipitation and evaporation index (SPEI) is a normalized indicator for the intensity of extreme climate conditions [12]. A value of +1/-1 indicates a wet/dry condition that is one standard deviation away from the normal condition. The index is considered to be more comprehensive than a single measure of temperature, drought, or rainfall, as it captures the overall balance between precipitation and the sum of evaporation and transpiration. The soil moisture index (SMI) are derived from earth observation data. This data is acquired from the European Space Agency Climate Change Initiative Soil Moisture Climate Copernicus.

To measure conflict situations, as well as economic and socio-political conditions, we make use of the Global Database of Events, Language, and Tone (GDELT). GDELT curates event documents from broadcast, print, and web news in nearly every corner of every country and at every second of every day. These events are grouped into 316 event categories based on the CAMEO codebook [9]. Following [6], we further aggregate these categories into five macro-categories: political events (GD:Political), social unrest (GD: Social), conflicts (GD: Conflict), economic events (GD: Economic), governance-related events (GD: Governance). The CAMEO codebook offers several mechanisms for assessing the "importance" or immediate-term "impact" of an event. Here, we use the average "tone" of all documents related to a given event, which ranges from -100 (extremely negative impact) to +100 (extremely positive impact), with zero being neutral.

The outcome variable and predictors are preprocessed as follows. Monthly ASR is transformed by taking the natural logarithm. All predictor variables are aggregated to monthly frequencies to be aligned with the outcome.

Lead-Lag Analysis and Pre-selection of Predictors

Before training the DynENet model, we conducted a lead-lag analysis to i) select the predictors that have significant correlations with the outcome; and ii) find the optimal lag length for each selected predictor, an approach inspired by [6]. Currently, the lead-lag analysis can only be conducted through the `yuima` R package [13]. In our analysis, we use `yuima` to decide which predictors to be mapped into the DynENet model and to identify the optimal lag of each predictor.

Empirical Migration Model

We specify our migraiton model as,

$$Y_{i,t} = b_{0,i} + \sum_{k=1} b_{k,i} X_{k,i,t-\widehat{\theta_{k,i}}} + \sum_{j=1} c_{j,i} Y_{j,t-\widehat{\theta_{j,i}}} + d_i Y_{i,t-1} + \epsilon_{i,t},\ i \neq j \qquad (1)$$

where, i and j are indices for origin-destination dyad flows, t is a time index, $Y_{i,t}$ is the outcome variable, $X_{k,i,t-\widehat{\theta_{k,i}}}$ is k^{th} predictor at $\widehat{\theta_{k,i}}$ lag, $Y_{j,t-\widehat{\theta_{j,i}}}$ is the flow to

j^{th} destination at $\widehat{\theta_{j,i}}$ lag, $Y_{i,t-1}$ is an auto-regressive term, and $\epsilon_{i,t}$ is an error term assumed to be normally distributed with zero mean and a constant variance.

Eq.(1) essentially entails multiple time-series models stratified by origin-destination flows, and hence is also known as the Flow-Specific Temporal Gravity (FTG) model, a new class of migration model that seeks to better explain and predict temporal patterns of migration flows [14]. A key feature of FTG is that all parameters are no longer fixed, rather they vary across flows. Such a parameterization attempts to isolate the spatial correlation between the predictors and the migration outcome. Hence, the parameter estimates are only identified by exploiting the temporal variations in the data.

Dynamic Elastic Net Algorithm

The Dynamic Elastic Net (DynENet) is a relatively new type of regularisation method [6]. It is similar to the classic Elastic Net (ENet) with the objective to find an optimal model specification (i.e., a set of predictors and their weights) that can best predict the outcome variable. However, a key difference is that the DynENet is trained on a rolling fold or time window, rather than on the entire time-series data.

Given Eq.(1), the objective function of DynENet for each flow i can be expressed as,

$$\min_{\beta_i} \left\{ \frac{1}{T} \sum_{t=1}^{T} L(Y_{i,t}, X_{i,t}\beta_i) + \frac{\lambda_i}{2} \left[(1-\alpha_i)\beta_i^2 + 2\alpha|\beta_i| \right] \right\} \tag{2}$$

where, $X_{i,t}\beta_i = b_{0,i} + \sum_{k=1} b_{k,i} X_{k,i,t-\widehat{\theta_{k,i}}} + \sum_{j=1} c_{j,i} Y_{j,t-\widehat{\theta_{j,i}}} + d_i Y_{i,t-1}$. T is the length of time-series data, $L(.)$ is a loss function, λ_i determines the magnitude of penalty on β_i, and α_i is a mixing factor determining the fraction of penalty applied to β_i^2 and to $|\beta_i|$, respectively.

Eq.(2) combines two types of penalized regression: Ridge and LASSO (Least Absolute Shrinkage and Selection Operator). For $\alpha_i = 0$, Eq.(2) is a Ridge regression which will shrink the coefficients through the penalty factor $\frac{\lambda_i}{2}\beta_i^2$. For $\alpha_i = 1$, Eq.(2) becomes a LASSO regression which will zero out the coefficients through the penalty factor $\lambda_i|\beta_i|$. When $\alpha_i = 0.5$, the model becomes the DynENet with half Ridge and half LASSO regression. This mix is considered a good compromise in terms of prediction and interpretation [6].

Hyper-parameter Tuning

Having set $\alpha_i = 0.5$, DynENet has one hyper-parameter to be tuned, λ_i. Unlike in classic Elastic Net, λ_i in DynENet is adaptive, as it is tuned based on a rolling fold (time window) cross-validation. Specifically, within each fold, we estimate 100 Elastic Net regressions with different values of λ_i. Typically, the best-tuned λ_i is chosen from the regression which produces the smallest prediction errors in the validation set (i.e., the last six months of each training fold). However, one of the pitfalls of this conventional approach is that smaller prediction errors might be driven by an increased model complexity, e.g., DynENet may favor a smaller λ_i, and

thus zero out fewer predictors, to improve model's training performance. The risk of this approach is that the model might overfit the training data with small bias but yield high variability in out-of-sample predictions. This so-called variance-bias tradeoff is undesirable, as it may lead to poor extrapolation (or generalization) into the future. To avoid overfitting, we introduce a new metric to evaluate the model's training performance: the penalized deviance ratio (PDR). For each flow i in each sub-period w, PDR is computed as,

$$PDR_{i,w} = \left[1 - \frac{\sum_{t_w=1}^{T_w}\left(Y_{i,t_w} - \widehat{Y_{i,t_w}}\right)^2}{\sum_{t_w=1}^{T_w}\left(Y_{i,t_w} - \frac{\sum_{t_w=1}^{T_w} Y_{i,t_w}}{T_w}\right)^2}\right] \times \frac{T_w - 1 - k_{i,w}}{T_w - 1} \tag{3}$$

where, T_w is the length of the training data within each fold, $k_{i,w}$ is the number of selected features for each flow in each fold.

In essence, $PDR_{i,w}$ measures the fraction of null deviance in Y_{i,t_w} (i.e., the sum of squared deviations from the unconditional mean of Y_{i,t_w}) explained by the model after adjusting for the number of selected regressors $k_{i,w}$. It is important to note that $PDR_{i,w}$ is a decreasing function of $k_{i,w}$, hence the deviance ratio is penalized when model complexity increases. The optimal $\lambda_{i,w}$ is chosen from the tuning regression that produces the highest value of $PDR_{i,w}$.

Forecasting Climate-Induced Refugee Migration

For each refugee flow i between Somalia and different EU destinations, the optimal λ_i values tuned through rolling fold cross-validation are used for the final training of our forecasting model. The predictors retained by the final DynENet model are matched with those in the testing data. These matched predictors, together with their estimated weights, are then mapped to the forecasting function to predict the ASR during the last six months of our data (July–December 2020).

The forecasting performance is evaluated by comparing the prediction errors (Root Mean Squared Errors or RMSE) of the final DynENet model with those of a Benchmark model (first-order auto-regressive or AR(1) model). Uncertainties in model forecasts are assessed at the 95% confidence level. The prediction intervals are obtained through a bootstrapping procedure. Specifically, for each flow and for S forecasting steps, we draw S residuals from the training set. These residuals are then added to the forecasted ASR (on log scale). We repeat this procedure 1000 times and obtain a sample of possible futures, and the 2.5% and 97.5% of the sampled values constitute the 95% prediction interval. These bootstrapped intervals are then converted from natural logarithm to their original scale. The results are shown in Figure 2.

Acknowledgements

The article has benefited from valuable comments of Prof. Mathias Czaika and Dr. Başak Yavçan. Financial support by European Union's Horizon 2020 Programme under grant agreement 870661, 871042, and 101004535, and the Swedish Research Council Vetenskapsrådet (grant agreement 2022-06012-3) are gratefully noted.

References

[1] I. Boas, C. Farbotko, H. Adams, H. Sterly, S. Bush, K. Van der Geest, H. Wiegel, H. Ashraf, A. Baldwin, G. Bettini, *et al.*, "Climate migration myths," *Nature Climate Change*, vol. 9, no. 12, pp. 901–903, 2019.

[2] H. Wiegel, I. Boas, and J. Warner, "A mobilities perspective on migration in the context of environmental change," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 10, no. 6, p. e610, 2019.

[3] R. Hoffmann, A. Dimitrova, R. Muttarak, J. Crespo Cuaresma, and J. Peisker, "A meta-analysis of country-level studies on environmental change and migration," *Nature Climate Change*, vol. 10, no. 10, pp. 904–912, 2020.

[4] H. Benveniste, M. Oppenheimer, and M. Fleurbaey, "Climate change increases resource-constrained international immobility," *Nature Climate Change*, vol. 12, no. 7, pp. 634–641, 2022.

[5] R. Black, S. R. Bennett, S. M. Thomas, and J. R. Beddington, "Migration as adaptation," *Nature*, vol. 478, no. 7370, pp. 447–449, 2011.

[6] M. Carammia, S. M. Iacus, and T. Wilkin, "Forecasting asylum-related migration flows with machine learning and data at scale," *Scientific Reports*, vol. 12, no. 1, pp. 1–16, 2022.

[7] A. Missirian and W. Schlenker, "Asylum applications respond to temperature fluctuations," *Science*, vol. 358, no. 6370, pp. 1610–1614, 2017.

[8] World Bank, "Sub-saharan africa-macro poverty outlook: Country-by-country analysis and projections for the developing world, april 2022," 2022.

[9] P. Schrodt, "Conflict and mediation event observations event and actor codebook v. 1.1 b3," 2012.

[10] H. De Haas, "A theory of migration: the aspirations-capabilities framework," *Comparative Migration Studies*, vol. 9, no. 1, pp. 1–35, 2021.

[11] J. Carling and K. Schewel, "Revisiting aspiration and ability in international migration," *Journal of Ethnic and Migration Studies*, vol. 44, no. 6, pp. 945–963, 2018.

[12] J. H. Stagge, L. M. Tallaksen, C. Y. Xu, and H. A. Van Lanen, "Standardized precipitation-evapotranspiration index (spei): Sensitivity to potential evapotranspiration model and parameters," in *Hydrology in a changing world*, vol. 363, pp. 367–373, 2014.

[13] S. M. Iacus and N. Yoshida, "Simulation and inference for stochastic processes with yuima," *A comprehensive R framework for SDEs and other stochastic processes. Use R*, 2018.

[14] H. Qi and T. Bircan, "Modelling and predicting forced migration," *Plos One*, vol. 18, no. 4, p. e0284416, 2023a.