Forecasting Forced Displacement Flows Using Machine Learning with Text Data

BSE Working Paper | 1573 April 2026

Laura Mayoral, Hannes Mueller, Christopher Rauh, RamónTalvi Robledo, Ben Seimon

bse.eu/research

Forecasting Forced Displacement Flows Using Machine Learning with Text Data

Laura Mayoral* Hannes Mueller† Christopher Rauh‡
Ramón Talvi Robledo§ Ben Seimon¶

April 14, 2026

Abstract

Forced displacement is an important policy challenge, yet forecasting is hindered by sparse, annually observed flow data and reporting delays. This article proposes a forecasting method for country outflows and dyadic flows tailored to this sparse data setting. We combine slow-moving structural predictors with high-frequency text-based signals, compress high-dimensional news into low-dimensional topic representations via Latent Dirichlet Allocation to mitigate overfitting, and estimate a stacked ensemble of gradient-boosted trees that captures non-linear origin–destination interactions while making optimal use of the available data. We further apply conformal prediction to construct statistically valid prediction intervals for bilateral flows. Analyzing the text component yields that destination-specific search intensity of migration terms is a central predictor of subsequent dyadic displacement flows.

Keywords: Forced displacement, Early warning, Forecasting, Google Trends, Dyadic, Machine learning, Conformal prediction

JEL Codes: P16, C53, D72

*Institut d'Anàlisi Econòmica (IAE-CSIC), Barcelona School of Economics, CEPR. mayoralaura@gmail.com

†Institut d'Anàlisi Econòmica (IAE-CSIC), Barcelona School of Economics, CEPR. h.mueller.uni@gmail.com

‡Institut d'Anàlisi Econòmica (IAE-CSIC), Barcelona School of Economics, University of Cambridge, CEPR. cr542@cam.ac.uk

§Fundació d'Economia Analítica. ramon.talvi@bse.eu

¶Fundació d'Economia Analítica. benjamin.seimon@bse.eu

‖We gratefully acknowledge support from the German Federal Foreign Office. Rauh acknowledges financial support from AEI/MICINN (ATR2023-144291). Mayoral, Mueller and Rauh acknowledge financial support from the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2024-001476-S), funded by MCIN/AEI/10.13039/501100011033. Mueller and Rauh acknowledge support from the European Research Council project ERC-AdG 101055176 (ANTICIPATE).

1 Introduction

Forced displacement is at its highest level on record, with over 117 million people forcibly displaced worldwide as of mid-2025 according to the United Nations Refugee Agency (UNHCR). Whether driven by conflict, persecution, or socio-economic turmoil, the scale of these displacement events strains the capacity of governments and humanitarian organizations, while their complex dynamics remain poorly understood. Many humanitarian organizations face critical funding and resource allocation decisions that require global forecasting models capable of frequent updates. Recent forecasting work has increasingly targeted Europe-bound irregular migration, where operational indicators are available at high frequency, enabling short-horizon prediction and evaluation. Yet the only dataset with global origin–destination coverage—UNHCR's annual flows panel—is released with a six-month lag and provides just one observation per country-pair per year. This impedes data-driven early humanitarian action and limits ethically responsible academic research.

To address this challenge, we develop a global Early Warning System (EWS) that forecasts forced displacement flows using machine learning and natural language processing based on the UNHCR flow data. The system produces two complementary one-year-ahead forecasts from information available at the time of prediction. First, it delivers risk forecasts for the origin of displacement. For each country, we estimate the probability that displacement outflows in year $t+1$ exceed a crisis threshold, providing a signal for strategic planning and anticipatory action like negotiating humanitarian access to the crisis country. Second, we develop dyadic flow forecasts. For each origin–destination pair, we predict the magnitude of flows in year $t+1$ and quantify forecast uncertainty via prediction intervals, yielding corridor-level projections of where displaced people are likely to go. A key distinction from existing early warning systems is that we forecast flows rather than refugee stocks. Stocks describe where displaced populations currently reside, but flows are the forward-looking object that captures new arrivals and therefore most directly informs pre-positioning and capacity decisions.

Our approach confronts a fundamental tension between the annual resolution of the target data and the need for monthly forecast updates. We resolve this through four innovations. First, we combine slow-moving structural predictors with high-frequency text-based signals that can be updated on a monthly basis. Second, we compress high-dimensional text data into low-dimensional topic representations using Latent Dirichlet Allocation, mitigating overfitting in a small-sample setting. Third, we employ a stacked ensemble of gradient-boosted trees that learns complex origin–destination interactions while accommodating heterogeneous data availability across predictors. Fourth, we apply conformal prediction to generate statistically valid prediction intervals for bilateral flows, allowing decision-makers to assess forecast uncertainty.

We introduce two complementary models, both designed for a one-year forecasting horizon. The *outflows* model is a binary classifier that predicts whether a country's displacement outflows will exceed 500 displaced per million inhabitants—a threshold calibrated to

capture well-documented historical crises. This model produces onset probabilities that support prioritizing attention toward countries at risk. The *dyadic* model is a regression that predicts bilateral flow magnitudes for each origin–destination pair, providing granular forecasts of displacement patterns across all country corridors.

Developing a model that adapts quickly to monthly information at the dyad level poses technical challenges. The common covariates used in traditional gravity models, such as distance and GDP, are either time-invariant, slow-moving, released with long delays or do not offer global coverage. To tackle this problem we rely on a large set of predictors, including historical displacement data, conflict metrics, and, most importantly, high-frequency text-based predictors. We integrate two sources: a corpus of over six million news articles dating back to 1989, and Google Trends search data spanning 80 migration-related terms (e.g., "refugee", "visa") across 107 languages. In the Google Trends data, we weight search terms by the share of the population speaking each language and adjust for internet penetration, ensuring that the signal is appropriately attenuated in low-connectivity settings. We also query destination-country names directly (e.g., searches for "Germany" originating in Syria), capturing revealed interest in specific destinations. These text-based features provide sensitivity to emerging crises that may leave little trace in conventional conflict or economic indicators.

Our work builds on several strands of prior research. A first strand has focused on constructing global displacement datasets. The UNHCR's annual flows panel covers 209 territorial units since 1962 (UNHCR, 2024), while the World Bank–UNHCR Joint Data Center has compiled microdata for 53 low- and middle-income countries (Masaki and Madson, 2023). Existing early warning systems, such as the Danish Refugee Council's Foresight (DRC, 2023) and UNHCR's Nowcasting model (UNHCR, 2023), forecast refugee stocks in selected countries. From a forecasting perspective, a gap remains for globally comprehensive dyadic flow forecasts with an explicit real-time evaluation design and uncertainty quantification.

A second strand applies predictive methods to displacement. Approaches range from structural models (Martineau, 2010) to agent-based simulations (Suleimenova et al., 2017, 2021) and gravity-based frameworks that model flows as proportional to origin mass, destination attractiveness, and inverse distance (e.g. Welch and Raftery, 2022; Moraga and López Molina, 2024). Gravity models achieve strong performance for labor migration but typically fail to capture the abrupt regime shifts characteristic of forced displacement. We use gravity-style predictors to produce a benchmark model, but our forecasting objective emphasizes rare escalations and regime shifts, motivating a nonlinear ensemble model using text features.

A third strand exploits high-frequency data sources. Böhme et al. (2020) pioneered the use of Google Trends for migration forecasting, with subsequent work confirming its value for EU-bound asylum flows (Carammia et al., 2022; Boss et al., 2023). News data have also proven useful: Carammia et al. (2022) used GDELT event categories to identify push factors, and Mueller and Rauh (2018) introduced the large news corpus we employ

here. However, this prior work has relied on Eurostat's monthly asylum-application series, which offers fine temporal resolution but covers only European destinations. We extend this approach to UNHCR flow data, sacrificing temporal resolution for global coverage, and introduce methods to extract predictive signal from annual data despite its coarse granularity. Methodologically, we show how high-frequency leading indicators can improve forecasts of temporally aggregated annual outcomes when the target is sparse and delayed.

We evaluate both models in a pseudo-out-of-sample, rolling-origin forecasting design over 2019-2024. For each evaluation year t, we train the models using only data available up to t and generate predictions for year t+1, repeating this procedure sequentially across the test period. For the outflows classifier, we report performance both on all crisis *onsets*, defined as threshold crossings preceded by at least one year below threshold, and on *hard onsets*, defined as crossings after three consecutive years below threshold. Arguably, hard onsets are the most policy-relevant events, because they occur after sustained low displacement and therefore provide minimal recent signal for learning and preparedness (Mueller and Rauh, 2022). For the dyadic flow model, we benchmark against a strong persistence baseline that forecasts future flows using the most recently observed annual flow. Because bilateral flows are highly persistent, this baseline is difficult to outperform; improvements therefore indicate the model's ability to anticipate changes, including escalations and de-escalations, beyond what is implied by inertia alone.

The outflows classifier achieves a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.84 for onset detection, which implies that in 84% of randomly chosen pairs of onset and non-onset observations, the model correctly ranks the onset higher. This is a substantial improvement over random guessing (0.5) and indicates strong discriminative performance. For harder onset cases, the precision–recall AUC of 0.18 compared to 0.15 for a baseline represents a 20% relative improvement, but also underscores that performance remains low in absolute terms—consistent with the rarity and unpredictability of these events. In terms of magnitude prediction, the dyadic model reduces mean absolute error by nearly 50%, meaning that typical prediction errors are roughly halved relative to the naive benchmark. The largest gains in mid-range flow bins suggest that the model is particularly effective in predicting common, moderate changes, while extreme values remain more difficult.

Point estimates capture only part of the information provided by the model. This is particularly evident in the coverage results: while conformal prediction intervals achieve reliable coverage for stable flows, coverage deteriorates for large escalations, indicating that the model struggles to fully capture tail risks. As a result, point forecasts alone can give an incomplete picture of predictive performance, especially in more volatile regimes. At the same time, the behavior of the intervals provides additional insight. Interval width increases systematically prior to escalations, suggesting that the model detects rising uncertainty even when it cannot accurately predict the magnitude of the change. This implies that uncertainty measures are informative in their own right, helping to identify periods of heightened instability beyond what is captured by point predictions.

When we decompose the predictive gains from Google Trends within the dyadic model, we find that destination-side search intensity is one of the most informative corridor-level signals. This pattern is consistent with the migration-networks literature, which emphasizes that established networks lower information and mobility frictions and thereby shape destination choice (Munshi, 2003; McKenzie and Rapoport, 2010). It is also aligned with recent work showing that Google Search data can be used to infer the origins of migrants observed at destination (Ponticelli et al., 2024). In forecasting terms, incorporating these high-frequency signals improves our ability to allocate predicted outflows across destinations beyond what is achievable with slow-moving corridor attributes such as destination fixed effects or lagged stock shares.

To summarize, we make four contributions. First, we shift the forecasting target from refugee stocks to displacement flows, which are more directly tied to the onset and escalation of crises and therefore well suited for anticipatory action. Second, we develop a global forecasting system with full origin–destination coverage that supports monthly updates despite relying on annually observed flow outcomes. Third, we show that high-frequency text-based predictors—derived from a large news corpus and from Google Trends—provide incremental out-of-sample forecasting gains, with particularly pronounced improvements for hard-to-predict onset events. We operationalize these sources through a post-processing and alignment pipeline tailored to real-time forecasting. Fourth, we complement point forecasts with distribution-free uncertainty quantification via conformal prediction, and we release an open-source Python package (*LPCI*) to facilitate the construction of prediction intervals in panel data forecasting settings.[1]

2 Data and methodology

We predict outflows and bilateral flows using an ensemble of boosted-tree models combined through a greedy stacking procedure. A key challenge in this setting is that the effective training sample is small and different predictors are available over different periods, leading to incomplete data overlap. In practice, this means we train several base learners, each specialized in different sets of predictors: for example, a model that uses conflict fatalities ('fatalities model') or a model that uses news and Google Trends features ('text model'). The stacking step then builds a weighted average of these base learners, so that learners are weighted according to their contribution to predictive performance.

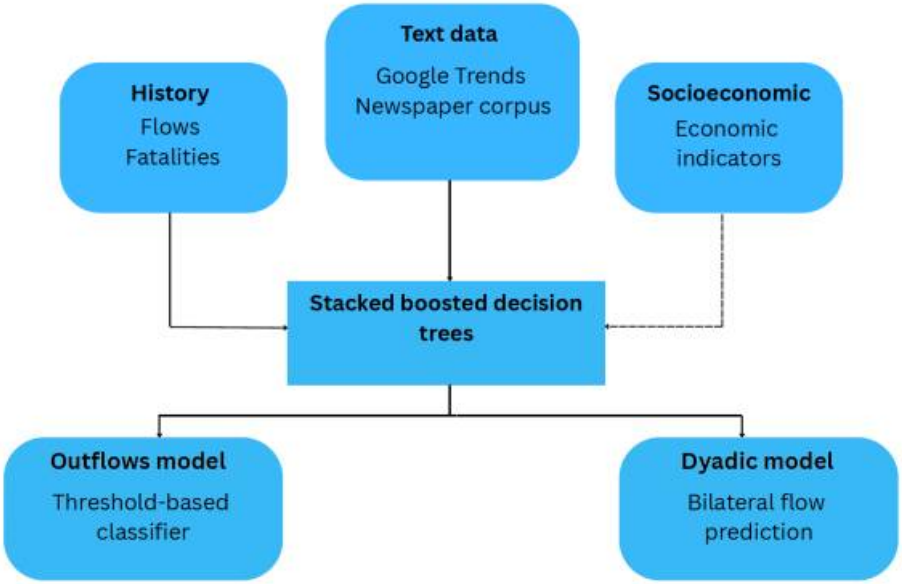
A key advantage of using a base-learners stacking approach lies in its flexibility to leverage heterogeneous sources of information across different training timelines. Because each base learner draws from a distinct set of features and has different data availability (e.g. the baseline model begins in 1989, whereas Google Trends-based text models only start in 2004, the first year Google Trends data is available), the stacking method enables us to capture complementary signals from model variants and the integration of all available information without discarding early data. Moreover, as the greedy stacking approach

[1]The package is distributed under the MIT license at `https://github.com/EconAIorg/LPCI`.

assigns stacking weights by iteratively improving the best-performing learner, it combines these complementary signals while filtering out redundant base learners.

As illustrated in Figure 1, the system processes multiple feature streams to produce two outputs: (i) an outflows classification model that detects large-scale displacement, and (ii) a dyadic regression model that predicts the magnitude and direction of bilateral displacement flows. We first explain the data sources and the steps we take to construct and harmonize these inputs, as well as the prediction methodology.

Figure 1: Schematic illustration of prediction pipeline



2.1 Dataset construction

In what follows we describe how the target variables and predictors are sourced and aggregated.

Refugee flows

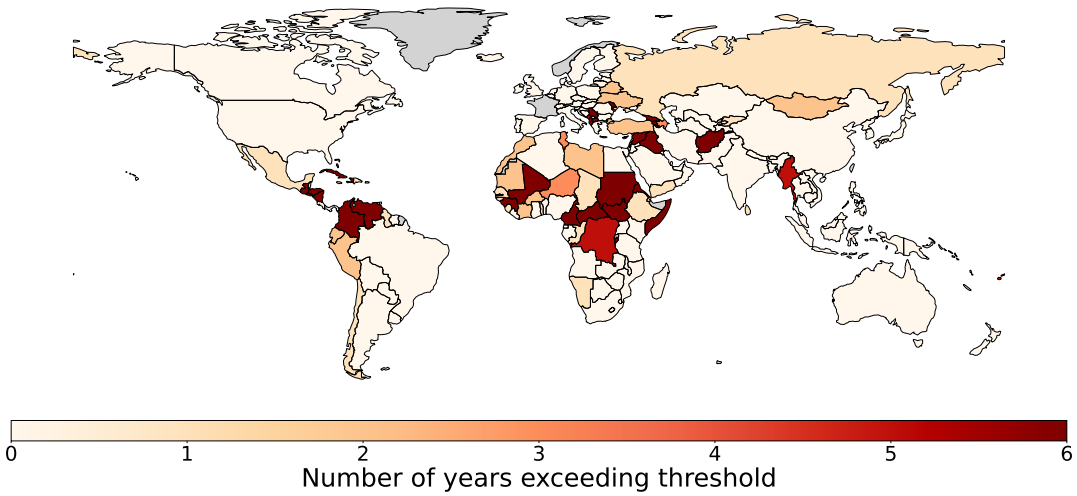
Our outcome variable is derived from the Office of the United Nations High Commissioner for Refugees (UNHCR) flow dataset (UNHCR, 2024), a panel dataset that reports bilateral forced displacement flows for over 209 territorial units spanning the years 1962 to 2024.[2] We define forced displacement as the sum of flows across all four categories reported by UNHCR: refugees, new asylum-seeker applications, people in refugee-like situations, and others in need of international protection. The dataset is updated twice a year, aligned with UNHCR's mid-year and end-year official statistics releases. Because the database is published with a built-in time lag, the figures for a given calendar year are released only

[2]These 209 units include the 193 UN official member states, UN observer states, non-sovereign dependencies, among other statistical reporting units.

halfway through the next one: for example, the full 2024 flow data was not available until June 2025. Our results rely on the December 2024 release, which includes observations up to the first semester of 2024 and aligns with the Google Trends data extraction period.[3] For pseudo–out-of-sample evaluation, however, we additionally use the second-semester flows of 2024 that were released later in June 2025. These observations are used solely for evaluation purposes and are not available to the model when producing the operational out-of-sample forecasts. When generating predictions, we ensure that forecasts rely strictly on information that would have been available at the time of prediction.

For the outflows model, we binarize the outcome by applying a threshold of 500 outflows per 1 million inhabitants. This classification threshold is chosen to capture well-documented instances of past crises effectively.[4] Figure 2 shows how many years between 1989 and 2024 a country has experienced forced displacement outflows in excess of this threshold. These crises are generally concentrated in Africa, the Middle East and northern South America. For the dyadic model, the target is a continuous variable that quantifies the bilateral count of cross-border displacement flows between country pairs.

Figure 2: Outflows classification threshold world map



Notes: Plot shows how many years between 1989 and 2023 a country has experienced outflows in excess of 500 people per 1mn inhabitants. Light gray shading indicates no data is available.

Based on this data we construct our targets: onset and hard onset. An onset is the change from below threshold to above threshold. A hard onset occurs when this happens after three consecutive years below threshold. Descriptive statistics of the flows data since 1989 for the two proposed models are presented in Table 1. Panel A shows that the outflows model faces substantial class imbalance: onsets account for only 3.2% of all observations, of which 59% correspond to hard onsets. It is important to keep this in mind:

[3]The EWS –if operational– can be easily updated monthly at a low cost with latest data.

[4]While a per-capita cutoff could in principle miss large absolute flows in very populous countries, we find no severe cases among the most populous, except Nigeria (population around 200 million), supporting a per-capita approach that favors cross-country comparability.

targeting policy at onsets or even hard onsets comes with a very low baseline likelihood and therefore with the danger of wasting resources on false positives. This makes preparedness and anticipatory action extremely costly in practice. The descriptive statistics for the dyadic model (Panel B) highlight that a large majority of dyadic observations have zero flows (95%) and, analogous to the outflows model, escalations are infrequent (3.9%). For the non-zero flows the mean is 1,092 but the standard deviation is very large with 17,312.

Table 1: Descriptive statistics for target variable

Panel A: Outflows	Shares			
N	0s	1s	Onsets	Hard onsets
7,920	0.82	0.18	0.032	0.019
Panel B: Dyadic	Mean [SD]	Shares		
N	when positive	0s	Escalations	De-escalations
1,816,290	1,092 [17,312]	0.95	0.039	0.013

Notes: Onsets are country–years in which outflows per million inhabitants exceed 500 after remaining below that threshold in the previous year; hard onsets require remaining below it for the previous three years. Escalations (de-escalations) denote year-to-year movements up (down) between the bins 0; 1–100; 101–1,000; 1,001–10,000; and 10,001+.

Google Trends

The Google Trends Index (GTI) provides the relative frequency of search terms on a scale from 0 to 100, where 100 represents the peak popularity of the term in a given country for a specific timeframe. In other words, the GTI presents normalized search interest rather than absolute counts. The data start in 2004 and is updated in near real-time. To collect the GTI across countries and search terms, we rely on a paid subscription service from SerpApi (SerpApi, 2024). This provides a simple integration to obtain the GTI for search terms across languages, countries and terms. In total, we collect monthly GTI for 80 migration-related keywords–which include single and multi-word terms–, and 192 destination-country queries, spanning 107 languages and covering 198 countries.

Yearly aggregation. Prior to topic modeling, we map the monthly GTI into annual features by averaging either the most recent three months (outflows classifier) or six months (dyadic regressor) of data. For example, an outflows forecast updated in June 2024 uses the mean of the Google Trends indices from April to June 2024. This rolling window aggregation retains high-frequency signal while producing country-year inputs for the LDA.[5]

Latent Dirichlet Allocation (LDA). For the outflows model, we focus on capturing the search behavior within each origin country. We select a subset of 80 key terms iden-

[5]Boss et al. (2023) demonstrate that the explanatory power of Google Trends declines for leads beyond six months.

tified by Boss et al. (2023) by hand. Common examples include terms such as "refugee", "asylum" and "visa". The keywords are chosen based on their relevance to forced displacement and migration-related inquiries. The complete list of terms for our outflows model are presented in Appendix A.

To account for the linguistic diversity across different countries, we weight each search term by the proportion of the population that speaks the corresponding language in a given country. For example, in a country where both French and Arabic are widely spoken, we query the French and Arabic translation of a given keyword, and weight them based on the population that speaks each language according to the Central Intelligence Agency (2024). We consider 107 languages for each of our searches.

The weighted Google Trends data are summarized using a Latent Dirichlet Allocation (LDA) topic model. We treat Google search term counts as a document–term matrix where country/months are the "documents". Using the LDA we estimate 10 topics, thereby reducing the dimensionality of the data. Since all terms are migration-related, the resulting topics reflect sub-themes related to migration. By imposing only weak priors on the LDA, we obtain relatively distinct topics with some inevitable overlapping. Table 2 reports the main keywords for each topic for the December 2024 update. After adjusting for internet penetration and computing rolling, discounted token stocks to generate a smoother time series, the topic distributions or shares are incorporated as features in our models.

Table 2: Summary of LDA Google Trends topics for December 2024

Topic	**Theme**	**Top words**
1	Asylum eligibility	nationality, eligible, asylum
2	Precarious legal status	undocumented, persecute, immigrate, require documents, verification
3	Mobility barriers	screening, sponsor, discrimination, restriction, militia
4	Punitive measures	resettle, evacuee, sanctions, detain
5	Administrative gatekeeping	checkpoints, consulate, foreigner
6	Travel regulation procedures	camps, custom, passports, immigrate
7	Eligibility checks	waiver, verification, advisor, permits, eligibility
8	Family reunification	spouse, emigrate, immigrate, arrival
9	Irregular mobility	traffic, permits, flee
10	Bureaucratic mobility barriers	restrict, customs, unauthorized, advisor, seeker, visa free

Notes: "Top words" lists the most probable words for each topic such that their cumulative probability exceeds 0.2, subject to a minimum of 3 and a maximum of 6 words per topic if criteria are not met.

For the dyadic approach, we model both origin-country push and destination-country pull factors. On the push side, we include origin LDA topic distributions (as in the outflows model) summarizing migration intentions at the origin. On the pull side, we use (i) destination LDA topic distributions constructed from migration-related queries capturing the destination's migration information environment and (ii) raw GTI measures that track

how often destination names are queried from a given origin (e.g., "United States" searched in Mexico). For each origin, we systematically query all possible destination countries in the dataset (e.g., "Germany", "France"), thereby effectively characterizing cross-country search behavior.

LDA post-processing I: Internet penetration. The predictive value of Google Trends searches—and of the topics derived from them—depends critically on a country's level of internet penetration. In countries with low internet penetration, the model should place less weight on Google Trends topics when generating forecasts, and more weight when internet use is widespread. The underlying rationale is that in low-penetration settings, the observed topic distribution is measured with noise and should therefore converge toward a uniform prior.

Let $\text{TS}_{c,t}^n$ be the raw Google Trends derived topic share for topic n and country c at year t. Let $IPR_{c,t}$ be the internet penetration rate for a given country at a given time period. Hence, the adjusted Google Trends topic shares $x_{c,t}^n$ is given by:

$$x_{c,t}^n = \frac{C + TS_{c,t}^n \cdot IPR_{c,t}}{C \cdot N + IPR_{c,t}},$$

where C is a scaling constant in the range $(0, 1)$ that controls the strength of pull toward the uniform prior and N is the total number of topics. Note that when internet IPR is zero, the adjusted topic share converge to the prior $1/N$, whereas when its close to 1 the topic shares more closely reflect the observed Google Trends topic distribution.

LDA post-processing II: Temporal smoothing. In order to smooth fluctuations in search-query volumes over time, we convert flows into exponentially weighted stocks. For each country c and year t, let $w_{c,t}$ denote the total number of tokens in year t (i.e., the sum of GTI-scaled query tokens for that country–year). The token stock up to year τ is

$$W_{c,\tau} \;=\; \sum_{t=1}^{\tau} \delta^{\,\tau - t}\, w_{c,t}, \qquad \delta \in (0,1),$$

which differs from the flow $w_{c,t}$ by accumulating (and discounting) all past years up to τ.

Similarly, let $x_{c,t}^n$ be the share of topic n in year t for country c. The topic stock for topic n at year τ is the token-weighted, exponentially discounted average:

$$X_{c,\tau}^n \;=\; \frac{\sum_{t=1}^{\tau} \delta^{\,\tau - t}\, w_{c,t}\, x_{c,t}^n}{W_{c,\tau}}.$$

In essence, the weighting scheme assigns greater influence to the most recent periods and to years with higher search activity.

Newspaper text

To enhance our models with contextual information, we leverage a large news-text corpus of over 6 million articles since 1989. This corpus is also processed using a LDA topic model, generating 15 distinct topics that capture the thematic structure of the text. Table 3 presents the topic themes and associated keywords for the December 2024 update.[6] Instead of using the topic proportions directly from the LDA model, we compute a stock topic share, i.e. a cumulative sum with temporal decay, in a procedure analogous to the one discussed above.[7] The objective is twofold: (1) to let months with higher news volume weigh more heavily in the yearly stock-topic share, and (2) to smooth out short-term fluctuations while preserving the long-term signal contained in the text data.

Table 3: Summary of LDA news article topics for December 2024

Topic	Theme	Top words
1	Sports	win, game, play
2	Social welfare	health, food, education, school
3	Urban violence	militant, injure, town
4	Democracy & elections	opposition, parliament, vote
5	Military conflict	missile, air, drone
6	Energy infrastructure	energy, oil, production
7	International alliances	ally, decade, push, turn, organization
8	Judiciary proceedings	arrest, court, charge
9	Middle-east Gaza crisis	gaza, islamic, hamas
10	Communism	communist, communist party, broadcaster
11	Economics	economy, market, financial
12	Humanitarian aid	summary, urge, humanitarian
13	Civilian life	life, family, outside, history, school
14	Bilateral diplomatic relations	strengthen, tie, bilateral
15	Political commentary	possible, fact, territory, western, position

Notes: "*Top words*" lists the most probable words for each topic such that their cumulative probability exceeds 0.02, subject to a minimum of 3 and a maximum of 5 words per topic if criteria are not met.

2.2 Prediction method

Recall that we train and evaluate two forecasting models: (i) an outflows binary classifier that predicts whether next-year forced-displacement outflows from each origin exceed a threshold, and (ii) a dyadic regressor that predicts bilateral flow magnitudes for each origin–destination pair. Each model is implemented as an ensemble of base learners trained

[6] Refer to Appendix Figure B1 for a concrete example showing the November 2024 topic shares for Syria and their most-common associated words (note that topic themes may vary slightly between different update periods).

[7] Here we follow Mueller et al. (2023). In the Google Trends approach, we aggregate monthly data to yearly before processing through the LDA, whereas for the newspaper topics we perform an ex-post aggregation when computing stocks.

on complementary feature blocks (historical displacement dynamics, conflict and socioeconomic covariates, and text-derived indicators). We obtain a final ensemble forecast by stacking base-learner predictions, where the combination weights are learned via greedy stacking to optimize calibration-period performance.

To support model development and evaluation, the full dataset is partitioned into three distinct time blocks: a training set (1989/2004–2011, depending on feature availability), a calibration set (2012–2018), and a test set (2019–2024). Given this partition, the pipeline separates model development into three stages—an offline experiment stage, an offline forecasting stage, and an online forecasting stage—as outlined in Algorithm 1. In the *offline experiment stage* (training + calibration), base learners are first trained on the initial training window. Hyperparameters are then selected using time-respecting cross-validation on the calibration window and fixed thereafter. Calibration predictions are used to estimate stacking weights, which remain fixed in subsequent stages; for the dyadic model, an interval-calibration model is also fitted on the calibration residuals. In the *offline forecasting stage*, we perform a pseudo out-of-sample evaluation using an expanding window: for each forecast origin year T, base learners are refitted on all data available up to T with the fixed hyperparameters, one-year-ahead predictions for T+1 are generated, and forecasts are combined using the fixed stacking weights. Finally, in the *online forecasting stage*, operational predictions are produced by refitting the base learners on all information available up to the latest period and generating the corresponding one-step-ahead forecast.

We generate each base learner's predictions using CatBoost (Prokhorenkova et al., 2018), an algorithm which builds an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones by minimizing a loss function through gradient descent. It uses ordered boosting to reduce overfitting and handles categorical variables natively via a technique called "ordered target statistics," which prevents target leakage. CatBoost is well-suited for predicting refugee flows due to its ability to model complex, non-linear relationships between numerous predictors, including economic, political, and conflict-related variables. Moreover, its robustness to overfitting and strong performance on imbalanced data allow it to capture rare but critical displacement events, improving predictive accuracy.

Algorithm 1 Expanding-window forecasting design

Inputs: Full data sample $D = \{d_t\}_{t=1989}^{T^*=2024}$; base learners $\{F_k\}_{k=1}^{K}$

Stage 1: Offline experiment stage (training + calibration)

Fit each base learner F_k on training data $\{d_t\}_{t\leq 2011}$.

Tune hyperparameters θ_k for each F_k via time-respecting CV on calibration data $\{d_t\}_{2012\leq t\leq 2018}$; fix θ_k.

Generate calibration predictions $\{\hat{y}_t^{(k)}\}_{2012\leq t\leq 2018}$ using $F_k(\theta_k)$

Estimate stacking weights $W = (w_1, \ldots, w_K)$ by greedy stacking on the calibration period; fix W.

(Dyadic model only) Fit interval-calibration model on calibration residuals & save object.

Stage 2: Offline forecasting stage (pseudo out-of-sample evaluation)

Note: evaluated years are $T = t + 1$: hence, with $t \in [2018, 2023]$ we evaluate forecasts for $T \in [2019, 2024]$

for $t = 2018$ to 2023 **do**

  $D_{\text{train}} \leftarrow \{d_i\}_{i\leq t}$.

  Refit each base learner $F_k(\theta_k)$ on D_{train}

  Obtain base forecasts $\hat{y}_{t+1}^{(k)}$ for each $k = 1, \ldots, K$.

  Combine forecasts using fixed stacking weights:

$$\hat{y}_{t+1} \leftarrow \textstyle\sum_{k=1}^{K} w_k \, \hat{y}_{t+1}^{(k)}.$$

  Append $\hat{y}_{t+1}$ to $\hat{Y}^{\text{eval}}$.

end for

Stage 3: Online forecasting stage (operational out-of-sample prediction)

Note: The latest forecasting period in this study is $T^* =$ December 2024, corresponding to the $T^* + 1 = 2025$ forecast

Use all information available up to T^* to form the training set:

$D_{\text{train}} \leftarrow \{d_t\}_{t\leq T^*}$

Refit each base learner $F_k(\theta_k)$ on D_{train}

Obtain base forecasts $\hat{y}_{T^\star+1}^{(k)}$ for each $k = 1, \ldots, K$.

Combine forecasts using fixed stacking weights:

$$\hat{y}_{T^\star+1}^{\text{oos}} \leftarrow \textstyle\sum_{k=1}^{K} w_k \, \hat{y}_{T^\star+1}^{(k)}.$$

return $\left\{\hat{Y}^{\text{eval}},\ \hat{y}_{T^\star+1}^{\text{oos}}\right\}$

Note: T^* = December 2024 is the latest forecasting period used to construct predictors (aligned with latest Google Trends extraction period). At that moment of prediction, UNHCR flow outcomes for 2024 are only available for the first semester. Nonetheless, given in June 2025 UNHCR released flows for the full 2024 year, when evaluating model performance we extend evaluation years to include 2024, exploiting information that only becomes available after December 2024, but that is not included as input in the online forecasting stage when generating the operational out-of-sample predictions.

Outflows model

We construct four base learners, each incorporating different sources of information and with different starting training years: i) Baseline model: includes only past outflows as features, ii) Fatalities model: includes lagged outflows and fatalities, integrating historical conflict data, iii) Text model A: includes outflows, fatalities, and Google Trends topic shares, and iv) Text model B: includes outflows, fatalities, and newspaper topic shares. Features for outflows base learners are detailed in Panel A of Table 4. Finally, each model produces a predicted probability or risk, that is, the likelihood of exceeding the 500-per-million outflows threshold next year.

The predictions from multiple base learners are integrated through a greedy stacking algorithm (Kurz et al., 2020). This algorithm sequentially adds base learners to generate the final ensemble prediction by assigning a relative weight to each base learner based on their contribution to predictive performance on calibration set. The greedy stacking assigns weights of 1/3 each to the Fatalities model, Text model A, and Text model B. This greedy selection ensures the ensemble is never worse than the strongest base learner, and improves performance by aggregating non-overlapping signal across learners while suppressing those with redundant information.[8] Any improvements of the final *ensemble model* over the baseline and fatalities base learners, suggests leveraging text-enhanced base learners yields incremental predictive power beyond what is captured by outflows and fatalities trends alone.

Dyadic model

For predicting bilateral counts of displacement flows we train three base learners which are also combined using the greedy stacking algorithm: i) Baseline model: relies on past displacement flows and conflict-related fatalities, ii) Text model: adds to baseline migration-related search behavior from Google Trends and media coverage from the large news corpus, and iii) Socioeconomic model: include past outflows, fatalities, text features and additionally socioeconomic indicators such as GDP and V-Dem indices. The ensemble assigns weights of 1/6 to the Baseline model, 1/6 to the Text model and 2/3 to Socio-economic model.

A key aspect of this bilateral model is the differentiation between origin features, which capture conditions in the country of departure (e.g. prior displacement trends, conflict intensity, and search behavior); destination features, which reflect conditions in potential host countries (e.g. economic stability and historical asylum trends); and dyadic features, which encode relational characteristics between country pairs (e.g. google trends destination-country searches, geographic proximity, linguistic ties and historical bilateral migration links). This structured approach allows the model to effectively capture both

[8]The greedy stacking algorithm we employ optimises performance over all base learners with respect to the ROC-AUC for incidence. Whilst, this guarantee that the final ensemble will outperform all base learners with respect to this target/metric, it does not guarantee the same for onset or the F1 score for example.

Table 4: Predictors for outflows and dyadic base learners

Panel A: Outflows base learners features

Category	Feature	Description	Baseline	Fatalities	Text model A	Text model B
Historical flows	UNHCR outflows	past migration flows, rolling-mean windows	✓	✓	✓	✓
Conflict data	UCDP fatalities	conflict-related deaths, rolling-mean & persistence inds.		✓	✓	✓
Google Trends	GTI stock topics	weighted stock of google trends topic shares (10 topics)			✓	
News corpus	LDA stock topics	weighted stock of LDA-derived news topics (10 topics)				✓
Demographics	Population	total population per year	✓	✓	✓	✓

Panel B: Dyadic base learners features

Type	Feature	Description	Baseline	Text	Socio-economic
Origin					
	UNHCR outflows	past migration outflows, rolling-mean windows	✓	✓	✓
	UCDP fatalities	conflict-related deaths, rolling-mean & persistence inds.	✓		✓
	GTI stock topics	weighted stock of Google Trends topic shares (10 topics)		✓	✓
	LDA news topics	weighted stock of LDA-derived news topics (15 topics)		✓	✓
	Socio-economic ind.	GDP per capita, inflation, V-Dem			✓
	Population	total population of origin	✓	✓	✓
Destination					
	UNHCR inflows	past migration inflows, rolling-mean windows	✓	✓	✓
	UCDP fatalities	conflict-related deaths, rolling-mean & persistence inds.	✓		✓
	GTI stock topics	weighted stock of google trends topic shares (10 topics)		✓	✓
	LDA news topics	weighted stock of LDA-derived news topics (15 topics)		✓	✓
	Socio-economic ind.	GDP per capita, inflation, V-Dem			✓
	Population	total population of destination	✓	✓	✓
Dyadic					
	GTI raw	relative frequency of destination-country searches at origin		✓	✓
	CEPII distance	log-transformed geographic distance	✓	✓	✓
	Rolling flow shares	past dyadic migration flow shares	✓	✓	✓
	Rolling stock shares	past dyadic migration stock shares	✓	✓	✓

Notes: Panel A presents the features of the outflows models and Panel B of the dyadic models. Each column represents one base learner. Predictions of these models are weighted for the final prediction.

push-and-pull factors influencing forced displacement patterns. Features for dyadic base learners are detailed in Panel B of Table 4.

Prediction intervals. To convey the uncertainty around point estimates of our dyadic flow forecasts, we implement the Longitudinal Prediction Conformal Inference (LPCI) algorithm for panel data (Batra et al., 2023). This approach allows us to obtain statistically valid prediction intervals where data point exchangeability, which is a common assumption in conformal prediction framework, does not hold. As a part of this project we have open-sourced a general code base for implementing LPCI for panel data in regression settings.[9]

Prediction intervals through LPCI are obtained through a sequence of steps. Initially, we use our final ensemble model to generate predictions for calibration and test set ($\hat{y}_t(g)$). On the calibration window, we compute residuals and treat them as non-conformity scores.[10] We then fit a Quantile Regression Forest (QRF) (Meinshausen, 2006) that uses the last observed residuals as the target and lagged residuals, a group identifier (one-hot encoding), and control variables (representative features from original model) as predictors. Under a rolling forecast design, the QRF predicts conditional quantiles $(\tilde{Q}_{t,\beta}(g), \tilde{Q}_{t,1-\alpha+\beta}(g))$ of future residuals for each group/country g, and these quantiles define the bounds of the prediction interval for each test point:

$$C_t(g) = [\hat{y}_t(g) + \tilde{Q}_{t,\beta}(g),\ \hat{y}_t(g) + \tilde{Q}_{t,1-\alpha+\beta}(g)]. \tag{1}$$

Here, $1-\alpha$ denotes the nominal confidence level (95%), and $\beta \in [0, \alpha]$ determines how the total miscoverage α is distributed across the lower and upper tails. We select β on the calibration set by searching over candidate quantile pairs (β, $1-\alpha+\beta$) -each covering a central probability mass of $(1-\alpha)$-, and choosing the pair with minimal interval width. The main purpose is to obtain the narrowest possible prediction intervals that guarantee coverage.

3 Results

In this section, we present the *pseudo-out-of-sample* forecasting performance of our models over the sample $T \in \{2019, ...2024\}$. Our results show that we can effectively predict onsets, that text feature improve hard-onset detection significantly and that text features do this by "suppressing" false alarms.

Our evaluation method aims to simulate the intended use case. In the discussion of the results, we report the forecasted year $T = t + 1$ rather than the information year t used to generate the prediction. For example, forecasts produced using information available

[9] Available as a Python package named *LPCI*, distributed under the MIT license at `https://github.com/EconAIorg/LPCI`.

[10] In the conformal prediction framework, the non-conformity score of how unusual or unexpected a prediction is according to the previous examples in the data.

in 2022 to predict outcomes in 2023 are reported as 2023.[11]

For the outflows model, forecast performance is systematically assessed against the most difficult cases to predict: onsets and hard onsets. Recall that we define an onset as a threshold crossing in which the last observed outflows remain below the threshold, while a hard onset is a threshold crossing preceded by three consecutive years of outflows below the threshold. Within the 2019–2024 test window, the evaluation sample comprises 56 onsets of which 42 are hard onsets.[12]

Our dyadic model is also evaluated over the 2019–2024 test period now against a naive benchmark, which predicts last observed flows $(t-1)$ for next year $(t+1)$. Even though the naive benchmark fails to capture (de)escalations by design, it constitutes a hard-to-beat benchmark given bilateral flows have inertia and hence tend to exhibit minimal year-to-year variation.

3.1 Outflows model

The receiver operating characteristic (ROC) curve traces the true-positive rate (TPR or recall) against the false-positive rate (FPR) as the decision threshold moves from fully conservative (threshold = 1, labeling every case negative) to fully permissive (threshold = 0, labeling every case positive). The area under this curve (ROC-AUC) then condenses the model's overall ability to distinguish positive instances (those exceeding the 500-per-million cutoff) from negative ones (those falling below it) into a single number.

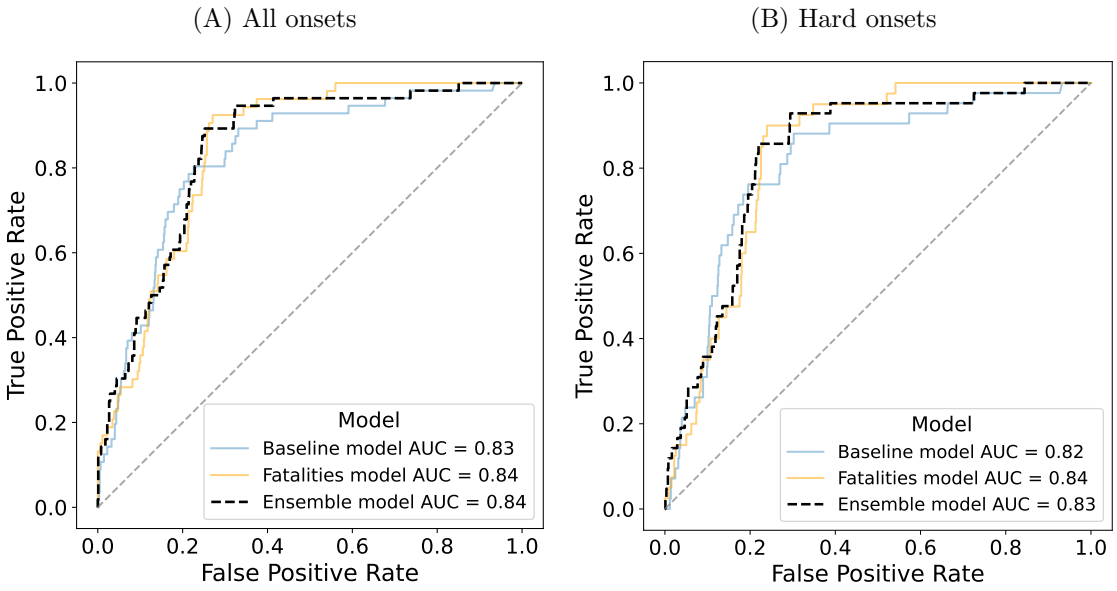
Figure 3 shows the ROC-AUC for onset and hard-onset events for the baseline (blue), fatalities (orange), and final ensemble (dashed black) models. All models exhibit robust discriminatory power in ranking crisis events above non-crisis ones, achieving ROC-AUC scores above 0.8 for both standard onsets (Panel 3A) and the more stringent hard-onset subset (Panel 3B). Augmenting the baseline with fatalities data lifts ROC-AUC by one percentage point in the standard-onset (rising from 0.83 to 0.84) and two percentage points in the hard-onset cases (from 0.82 to 0.84). The ensemble model, which leverages text-based base learners, attains very similar ROC-AUC values as the fatalities-enhanced classifier. Nonetheless, a closer look at the high-threshold region of the ROC curves in Panel 3 suggests that the ensemble attains slightly higher TPR at comparable low FPR values, indicating improved recall without additional false alarms—a nuance that the aggregate ROC–AUC metric obscures.

For this reason we complement the ROC curves with precision-recall curves. By plotting precision against recall for different decision thresholds, we put the focus on the problem of false positives that most policy makers face. Catching even a few onsets will trigger many false alarms. The area under that precision–recall curve (PR-AUC) collapses this

[11] The last year with complete flow data in our sample is 2024. Full 2024 flows are used for evaluation, while operational forecasts based on the December 2024 release observe only first-semester flows. See Section 2.1 for details on the UNHCR flow-dataset release schedule.

[12] Both 2022 and 2023 were turbulent years where the hard onsets significantly increased from below 3% in previous test years to above 8%.

Figure 3: Pseudo out-of-sample ROC-AUC across models



Notes: The sample in Panel A is restricted to all observations that are not above the 500-per-million threshold at the time the forecast is made and in Panel B that have not experienced an episode above the threshold in previous three-year period.

entire trade-off into a single number providing an overall metric of model performance.[13]

Figure 4 represents precision-recall curves across test years, with Panel A focusing on all onsets and Panel B on the hard-onset subset. For all onset events, the baseline model attains a PR-AUC of approximately 0.23, which increases to 0.31 when incorporating fatality derived features (fatalities model). The ensemble precision-recall curve largely overlaps with the fatalities model across most recall levels, though between recall values of 0.2 and 0.3 the ensemble achieves a higher precision.

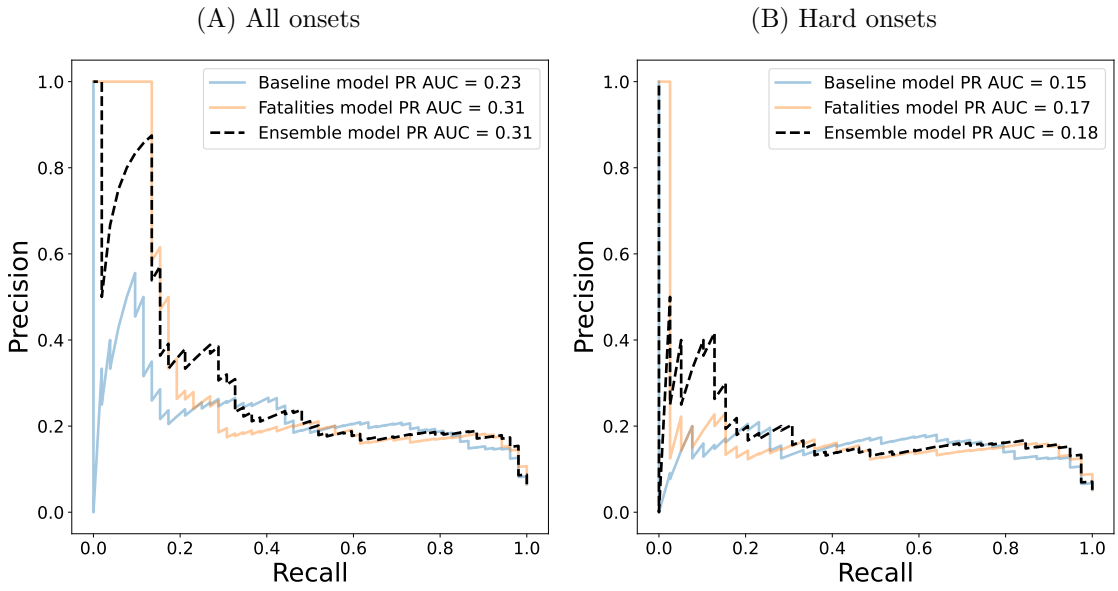
The advantage of the text-enhanced model becomes more evident on the hardest-to-predict events shown in Panel 4B. The ensemble's PR-AUC reaches 0.18, outperforming both the baseline (0.15) and fatalities model (0.17). At 10% recall, the ensemble achieves a precision of around 0.40, compared to roughly 0.20 for the baseline model and fatalities model. This pattern shows that textual features deliver the greatest boost precisely where precision matters most: when users focus on a small set of highest-risk countries, the ensemble model cuts false alarms by a factor of two.[14]

To illustrate this added value of the ensemble we compare its SHAP contributions across feature types to those of the fatalities-only model. SHAP values quantify a feature's contribution to a prediction—that is, how much the feature pushes the prediction above

[13]Figure D1A in Appendix Figure D1B shows that the onset ROC-AUC is relatively stable across all the years analyzed. The onset PR-AUC exhibits larger differences ranging from below 0.2 in 2019 to above 0.6 in 2021.

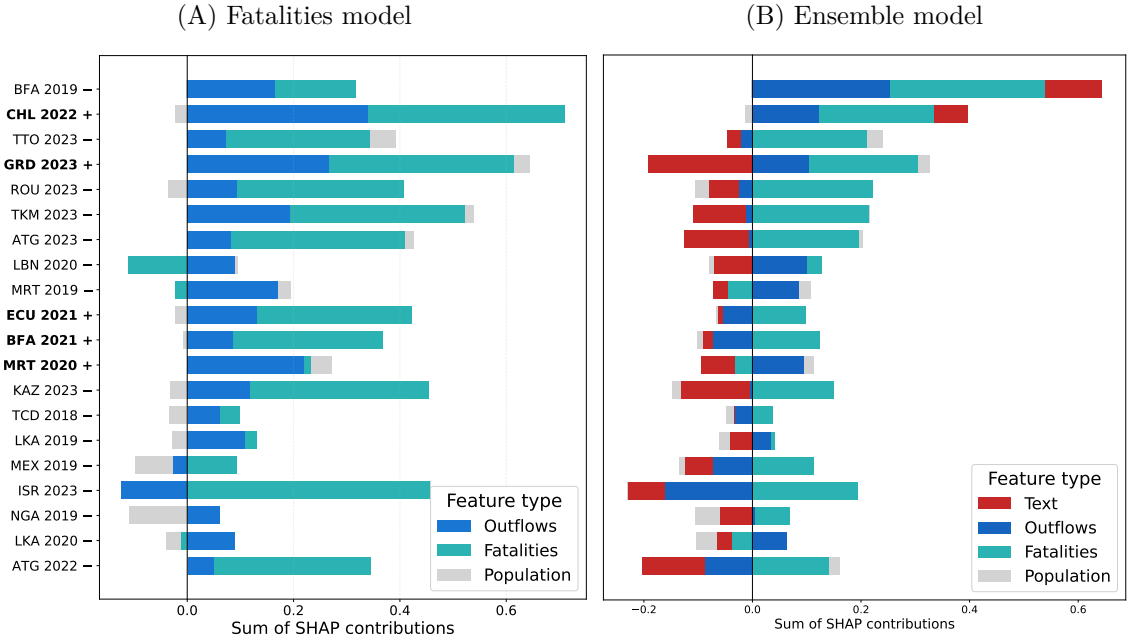
[14]See Appendix Figure D2 for the calibration plot for the final ensemble predictions.

Figure 4: Pseudo out-of-sample precision-recall AUC across models



Notes: The sample in Panel A is restricted to all observations that are currently not above the 500-per-million threshold and in Panel B that have not experienced an episode above the threshold in previous three-year period.

Figure 5: SHAP contributions by feature type for hard-to-predict countries (test set)



Notes: sample includes country–years that have outflows below threshold for at least past 3 years, covering both positive (hard-onsets) and negative predictions. Panel A shows the fatalities-only model; Panel B shows the final ensemble. Bars are SHAP contributions stacked by feature type: segments to the right (left) of the vertical line increase (decrease) the forecast relative to the expected-value baseline. Country–years in bold with a + denote hard onsets; those with a - denote negative predictions.

or below the expected (baseline) value.[15] We restrict attention country–years that have remained below the displacement threshold for at least three consecutive years. These are the low-risk type of situations that generate hard to predict onsets.

In Figure 5, Panel A shows SHAP results for the fatalities model and Panel B shows the analogous for the ensemble model. Bars are stacked by feature type around zero; segments to the right increase predicted risk and segments to the left decrease it. Countries with bold labels with a + indicate cases where the prediction was followed by a hard-to-predict onsets, while countries with a - sign indicate negative cases. We display only the top twenty country–years ranked by the ensemble's predicted risk to focus on the high-risk region in Figure 4B—precisely where the precision–recall analysis for hard onsets shows the ensemble's performance advantage over the fatalities model.

Past the two top predictions in both models, the text features typically pull the prediction downward in this group of countries. This reduces the forecast and helps avert false positives. The most illustrative cases here are Kazakhstan, Israel and Turkmenistan where the text features pull the prediction down in the ensemble model. Precision is higher in the ensemble because the text features provide an additional sorting signal that helps distinguish positives from negatives. Still, the gains from text features in the outflow model are relatively modest.

3.2 Dyadic model

In Table 5 we report the relative mean absolute error (MAE) for all the base learners, i.e. a baseline model, a text-based model and a socioeconomic model, and their combined ensemble across different target flow bins. Relative MAE is defined as the ratio of the model's MAE to the naive model's MAE: hence, values below 1 indicate improvement over the naive benchmark. Keep in mind that beating the naive model, same flow as last period, is notoriously difficult to beat in yearly data. Nontheless, it fails to capture (de)escalations by construction, so performance improvements reflect the model's ability to anticipate turning points in displacement dynamics.

The results show a heterogeneous performance of the different base learners by bin. The baseline model is particularly strong in the 0 outflows bin as the model never generates high flow predictions for countries in peace with no history of outflows. As expected, across the mid-range bins (1–100, 101–1,000, and 1,001–10,000), our ensemble model consistently outperforms all standalone base learners, and reduces the MAE by around 50% compared to the naive forecast. Moreover, in the highest volume bin (10,001+), the ensemble's gains over naive strategy are smaller but still significant (18%). This highlights the challenge in predicting shifts in large flows (i.e., Ukraine's bilateral flows to Germany surge from a few hundreds in 2021 to the order of millions in 2022).

While providing point-forecast performance metrics is revealing, operational planning demands an assessment of forecast uncertainty, especially when decisions carry high hu-

[15] We compute SHAP values with a model-agnostic permutation approach for both the ensemble and fatalities model. See Appendix E for a detailed description on SHAP computation for ensemble model.

Table 5: Performance by target bin relative to naive model

	Target bin				
	0	1-100	101-1,000	1,001-10,000	10,001+
Panel A: Relative mean absolute error					
Baseline	0.11	0.64	0.50	0.55	0.83
Text	0.22	0.60	0.47	0.50	0.82
Socioeconomic	0.18	0.57	0.44	0.49	0.82
Ensemble	0.16	0.56	0.44	0.47	0.82
Panel B: Coverage					
Ensemble	0.99	0.84	0.83	0.73	0.65

Notes: Relative mean absolute error compared to naive benchmark. The naive benchmark is predicting a flow will be equal to last year's. Coverage is the share of forecasts that fall within the predicted confidence interval.

manitarian stakes. To construct prediction intervals for our dyadic model we adopt the conformal prediction framework and implement the LPCI algorithm (Batra et al., 2023). The main idea is to use the non-conformity score -in our case, residuals- in the calibration set to obtain uncertainty intervals for the test points.[16] The Quantile Regression Forest (QRF), which takes last observed residuals as the target and a history of lagged residuals and control variables as predictors, is then used to estimate the conditional quantiles of residuals for each unseen test point.[17] These quantile estimates are then directly employed to define the bounds of the prediction intervals.

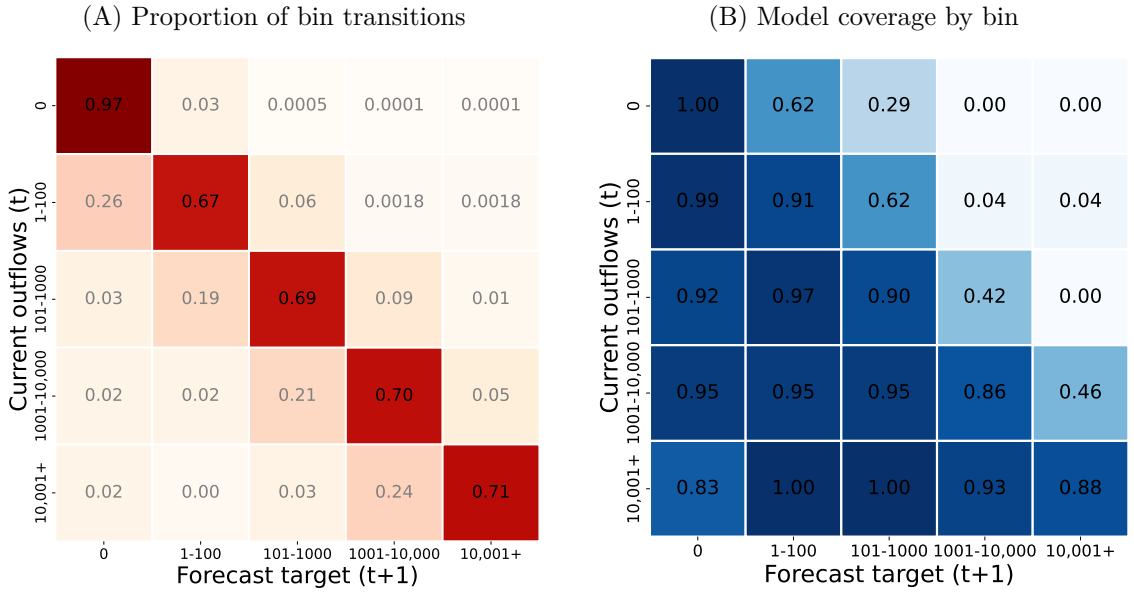
Overall coverage refers to the proportion of times the true outcome falls within its corresponding prediction interval across all observations. Overall coverage is consistently above 96%, a result largely driven by the predominance of zero flows, which almost always fall within even relatively narrow intervals. However, we are interested in whether the system delivers coverage in rare outlier cases. Coverage by bin assesses the share of times the true outcomes lie within prediction intervals across all time points for each bin. As Panel 5B shows, coverage deteriorates as bin size increases. The ensemble models achieves almost perfect coverage in the zero-flow bin, while in the mid-range flow bins (1–100, 101–1000 & 1001-10,000) coverage hovers around the 70-85% range, considerably below the confidence level of 95%. In line with the previous findings, coverage drops more for the largest bin (65%) but remains over 50% even for these extremely rare events.

As coverage would be essential for the reliability of a system like ours we analyze this further. Figure 6 juxtaposes the empirical bin-transition matrix (Panel A) with the corresponding coverage rates for each transition (Panel B). In Panel 6A, the strong concentration along the main diagonal reflects the "stickiness" or auto-correlation of bilateral flows—most dyads remain in the same volume bin year to year—while the off-diagonal entries capture the comparatively rare escalations and de-escalations. Coverage is strongest

[16]For further details on LPCI implementation, see Section 2.2 or the LPCI package documentation (`https://github.com/EconAIorg/LPCI`).

[17]Control variables from original model include dyadic stocks and flows, as well as population at origin and destination.

Figure 6: Comparison of bin transitions and escalation coverage.



Notes: In both panels the y-axis specifies the bin of current outflows and the x-axis the target bin. The darker the shade of a square the larger the number. Panel A indicates how often we observed transitions from one bin (row) to another (column). Therefore, each row sums to one. Panel B show how often our forecast interval correctly captures each transition. White squares indicate that no observations fall within that bin.

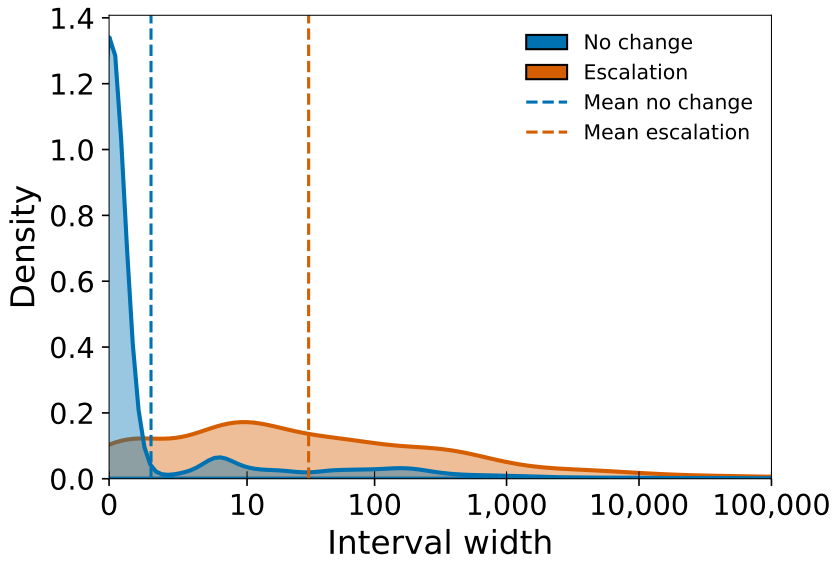
for the "no-change" bin and for de-escalations, where sufficient training samples allow the model to learn these patterns. Prediction intervals also perform well for adjacent-bin transitions, both one-step escalations and de-escalations. Coverage has a deep asymmetry. It deteriorates only for upward movements when these become rarer and more extreme. For example, among observations transitioning from 1–100 outflows (current) to 100–1,000 (target), the interval captures the true value 62% of the time, but this falls to only 4% for transitions to 10,001+ outflows.

The questions becomes whether the system indicates that "something" is coming but misses the size of the flow. In other words, the width of the prediction interval itself conveys relevant signaling information because it encodes the model's uncertainty: intervals should remain narrow when flows are predictable and expand adaptively when the model detects a heightened risk of escalation. Figure 7 shows the distribution of predicted interval widths, in logarithmic terms, for cases that experience an escalation in the forecasted period (orange) versus those with no change, defined as remaining in the same flow bin (blue).[18] Dashed lines represent the mean interval widths: 1 for no-change cases and 31 for escalations. Predictions followed by no change produce very narrow intervals for a large share of observations, reflected in the distribution's mode near zero. In contrast,

[18]For clarity, we exclude de-escalation cases. Their interval-width distribution is qualitatively similar to that of escalations, suggesting that intervals widen ahead of large changes in flows regardless of direction.

the uncertainty intervals for escalating dyads (orange) display a much more dispersed distribution with a pronounced right tail that extends into the thousands.

Figure 7: Distribution of predicted interval widths for escalations vs. no-change cases



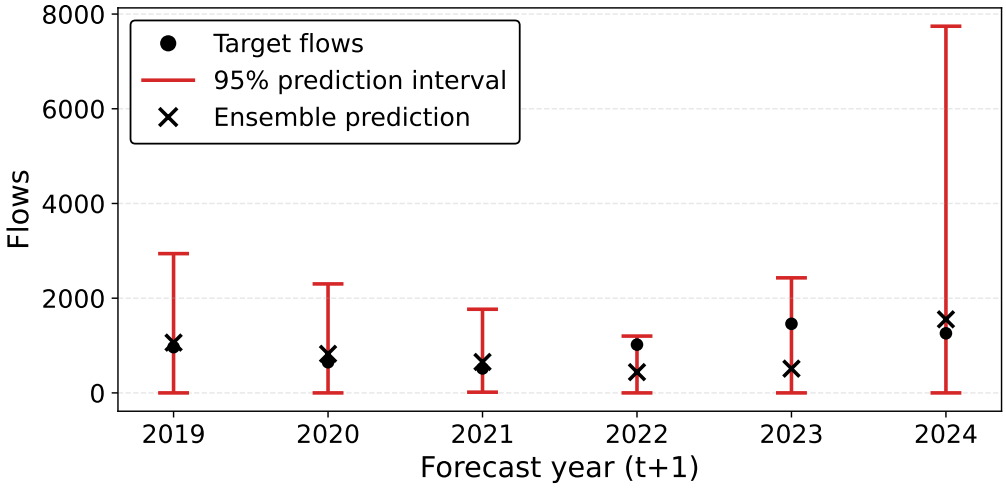
Notes: The x-axis represents the interval width for predictions followed by an escalation in the forecasted period (orange) versus those with no change (i.e., remaining in the same flow bin, blue). Dashed vertical lines indicate the respective means.

In other words, our conformal intervals tend to widen when an escalation is on the horizon. This suggests that, despite the high incidence of zeros and the yearly granularity of the data, the model captures informative second-moment dynamics in addition to point predictions. Such uncertainty signals could be valuable for designing loss functions for risk-averse policymakers or in settings with asymmetric costs of over- and under-prediction.

To visualize this behavior, we present one specific dyad—refugee flows from the Republic of Congo to the France. France has been one of the primary destinations for outflows from the Republic of the Congo during our 2019–2024 test period, and in 2023 received roughly 43% of all newly displaced Congolese. Figure 8 illustrates the forecast dynamics for the COG-FRA dyad in the test set. The x-axis is indexed by the forecasted year (the year being predicted). Black dots denote the realized flows, black crosses the corresponding ensemble point predictions, and the red vertical bars the 95% prediction intervals.

Overall coverage for the dyad is 100%, as the realized flow falls within the predicted 95% interval in every forecasted year. An instructive pattern emerges in the 2023 forecast. The 2022 residual is larger than in preceding years but still small in magnitude. However, the control variables included in the QRF resemble configurations that historically co-occurred with escalations across dyads. As a result, the LPCI framework raises the upper conditional quantile and the prediction interval widens even in the absence of large recent residuals. Consequently, even though the ensemble point prediction underpredicts the escalation, the prediction interval still covers the realized outcome, illustrating how interval

Figure 8: Predicted flows from the Republic of Congo to France



Notes: Black dots denote the realized displacement flows from the Republic of the Congo to France in the forecasted year (t+1). Black crosses indicate the ensemble point predictions, and red vertical bars show the corresponding 95% conformal prediction intervals.

widening can signal elevated risk even when point forecasts do not fully anticipate the surge.

The 2024 forecast reflects a different mechanism. Here, the large residual observed in 2023, together with feature configurations that historically co-occurred with escalations across dyads, leads the QRF calibration step to raise the upper conditional quantile more aggressively. As a result, the prediction interval widens substantially, signaling elevated uncertainty. The ensemble point prediction also shifts upward relative to the previous forecast, tracking the persistence of high flows, even though the realized outcome does not escalate further relative to 2023. Taken together, these two episodes illustrate the broader pattern shown in Figure 7: prediction intervals tend to widen in the presence of escalation risk, with the extent of the widening reflecting the combined influence of recent residuals and escalation-related feature patterns..

4 What is driving the predictions?

One advantage of the tree-based design is that text features can capture additional push and pull dynamics that go beyond the standard gravity models. To examine this explicitly, we study how Google Trends features contribute to our dyadic regression forecast for the final out-of-sample forecasted year, 2025, using predictions issued in December 2024. Hence, in this exercise both the predictions and Google Trends topics are generated in December 2024 to forecast 2025 flows. This model, trained on the maximum amount of data, gives us an opportunity to understand how the model uses the text features in prediction.

To assess predictive capability beyond conventional features, we begin by computing

SHAP values for the final ensemble using a model-agnostic permutation framework.[19] This approach enables us to quantify each feature's marginal contribution to the 2024 out-of-sample forecasts. We first identify and rank the most influential dyadic predictors, highlighting the role of Google Trends–based origin topics and migration-related search activity at the destination. We then turn to the Google Trends–derived indicators themselves—origin-topic embeddings, the GTI (country-name searches at origin referencing a potential destination), and destination-side search volumes—and examine their aggregate contribution to predicted flows across all dyads.

Top dyadic predictors

We compute SHAP values on the original scale of the outcome (number of people), allowing each feature's influence to be interpreted directly: its SHAP value represents the change in the predicted number of individuals attributable to that feature, relative to the model's baseline prediction. Permutation SHAP also ensures that interactions among features are handled consistently, preventing shared effects from being counted more than once.

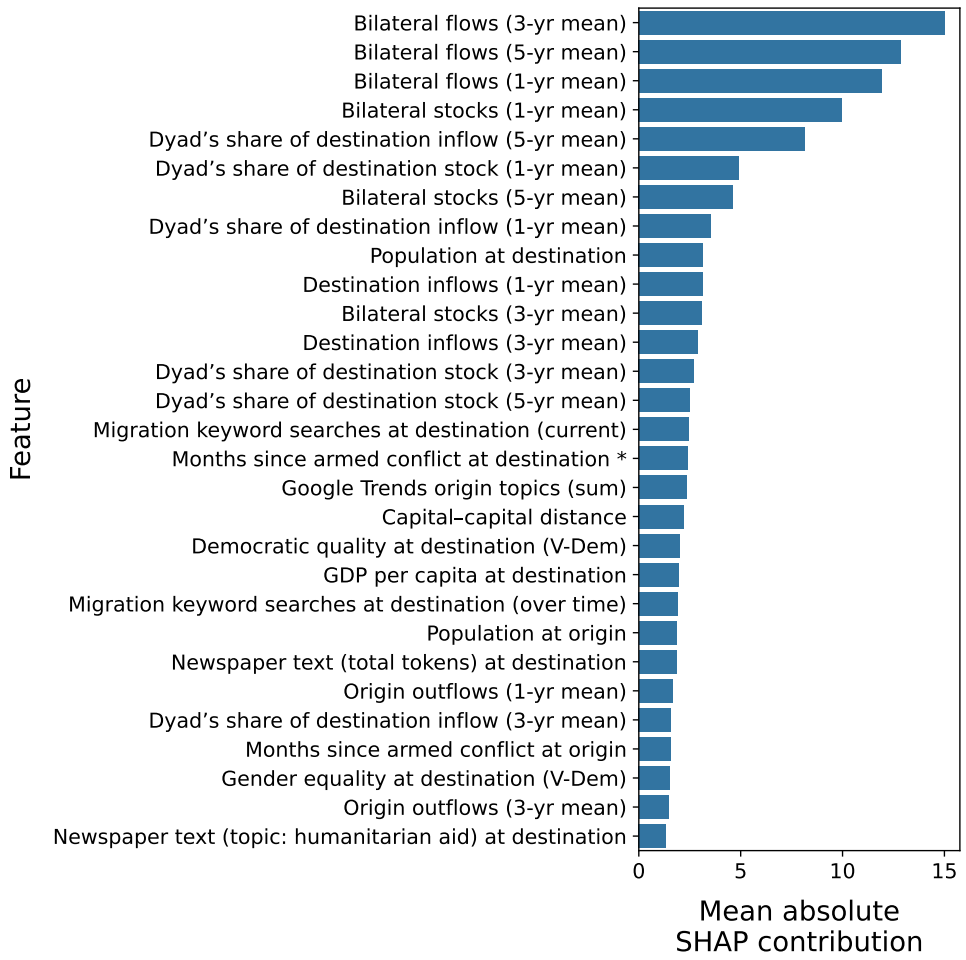
Figure 9 presents the features with the largest predictive contributions in the 2025 out-of-sample forecast. Out of nearly 150 candidate predictors included in the model, these stand out as the variables that most strongly shape predicted flows. The x-axis corresponds to the mean absolute marginal contribution of each feature reported in original scale, number of displaced people. Hence, the values represent the average (across dyads) absolute change in predicted flows (in people) attributable to each feature. Four of the top five predictors correspond to past bilateral flows or stocks at different time horizons, both recent (1-year means) and longer-term (3-year and 5-year means), underscoring the persistence of displacement, whereby historical flows and existing migrant stocks remain the strongest predictors of future movements. Moreover, dyadic indicators capturing the share of inflows from an origin at the destination and the share of destination stocks already settled from that origin, further reflect how existing communities at host countries shape future displacement patterns.

Besides dyadic features, we observe that most explanatory power comes from the pull-side i.e. destination features. Salient predictors include the number of months since armed conflict at the destination, which signals recent instability, and classic gravity determinants such as capital-to-capital distance and GDP per capita at destination.[20] Socio-political V-Dem indicators also emerge as meaningful predictors, most notably polyarchy (electoral democratic quality) and the gender-equality measure, capturing at the destination dimensions of women's empowerment, civil liberties, and political participation, as well as the extent to which power is distributed equitably across social groups. On the push side, the most notable origin predictor is past-year outflows, which capture the most recent displacement dynamics and the inertia of ongoing movements. Additionally, origin

[19] See Appendix E for a description on how to compute SHAP values for the final ensemble regressor with a permutation approach.

[20] We define armed conflict as cases where monthly fatalities surpass 50 per million inhabitants.

Figure 9: Ranking of predictors for dyadic regressor by mean absolute SHAP contribution



Notes: SHAP values are computed for the $T = 2024$ model using a model-agnostic permutation approach applied to the final ensemble model. Bars report, for each feature, the mean absolute SHAP value across all dyads in original scale, and thus summarize the magnitude—but not the sign—of its contribution to predicted flows. (∗)Armed conflict is defined as cases in which monthly fatalities exceed 50 per million inhabitants.

population is also predictive of bilateral flows between dyads.

Text and Google Trends signals also contribute to explaining variation in predicted flows across dyads. On first sight, the gains from text may seem marginal. However, the fact that previous flows explain so much of the variation is analog to past conflict dominating the predictive power in the conflict literature (Mueller and Rauh 2018). Any additional gain in predictive power is very hard to come by and becomes extremely valuable for cases with no previous flows. Newspaper features capture how host countries are portrayed in the media: the overall token volume reflects the intensity of newspaper coverage on a destination, while the share of newspaper coverage focused on humanitarian aid may signal the degree to which the destination is represented as favorable toward offering assistance. Additionally, migration-related searches *at destination* are likewise

informative. Consistently positive SHAP values for destination-side searches indicate a positive association with predicted flows, which is consistent with these searches being generated by refugees and asylum seekers already settled in the destination who seek migration-relevant information (e.g., procedures, rights, family reunification). This is not unrealistic given recent findings by Ponticelli et al. (2024) who show that presence of migrants can be proxied through Google searches at destination. Finally, the sum of origin-specific Google Trends topics also helps account for variation in predicted flows. Taken together, the Google Trends data capture both push and pull dynamics, and thereby help explain a meaningful share of the model's predictive signal. We turn towards a detailed analysis of these factors for the true out-of-sample model next.

Google Trends features in the true out-of-sample

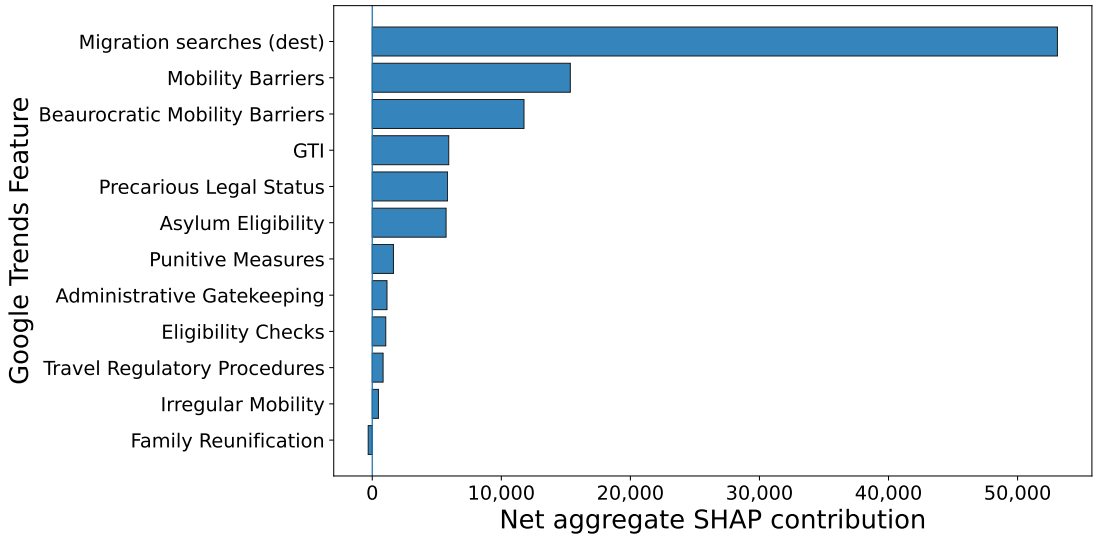
Google Trends predictors account for a meaningful component of the model's out-of-sample predictive signal. Figure 10 displays the net permutation SHAP contributions of the Google Trends predictors for the 2025 out-of-sample prediction. In this setting, the x-axis reports the net aggregate SHAP value, defined as the sum of SHAP contributions (positive and negative) across all dyads. In other words, the length of each bar reflects how much a feature pushes predicted flows up or down from base value accounting for interactions with other variables. The y-axis shows the disaggregated LDA-derived origin topics, the GTI and migration-related Google Trends search volume at the destination.[21]

Figure 10 shows that the destination-side migration searches dominate the Google Trends block, with a large positive net contribution of over 50,000 people across all dyads. This indicates that higher destination search intensity is associated with sizeable upward shifts in predicted flows. These net positive SHAP contributions are consistent with migration-relevant searches carried out by refugees and asylum seekers already settled at destination, which would naturally correlate with corridors experiencing higher realized or prospective flows. More broadly, since recent displacement dynamics—captured by lagged flows and stocks—are the strongest predictors of future movement, destination-side searches can be interpreted as a higher-frequency, near–real time signal that tracks active or emerging displacement corridors as conditions evolve, thereby complementing (and partially updating) inherently retrospective flow and stock measures.

As expected, the GTI also contributes positively to predicted flows, consistent with the idea that more frequent origin searches for a specific destination country name reflect heightened attention to—and potentially greater orientation toward—that destination. Among the origin-side topic embeddings, a small set of themes accounts for most of the positive contribution: *mobility barriers* (top words: screening, sponsor, discrimination, restriction) and *bureaucratic mobility barriers* (top words: restrict, customs, unauthorized, advisor, seeker, visa free) stand out as the strongest topics, suggesting people are searching for information about obstacles to movement, whether administrative requirements such

[21] See Section 2.1 for a full characterization of the ten LDA-derived Google Trends origin topics, including their defining top words.

Figure 10: Net SHAP value across all dyads per Google Trends origin topics, Google Trends Index, and migration keyword searches at destination



Notes: The horizontal axis reports the summed SHAP values across all dyads in their original scale (people), although the underlying plot uses a logarithmic scale for visualization purposes. The vertical axis lists origin-specific Google Trends topics, the Google Trends Index (GTI), and the destination Google Trends search volume. All values are from the $T = 2024$ model.

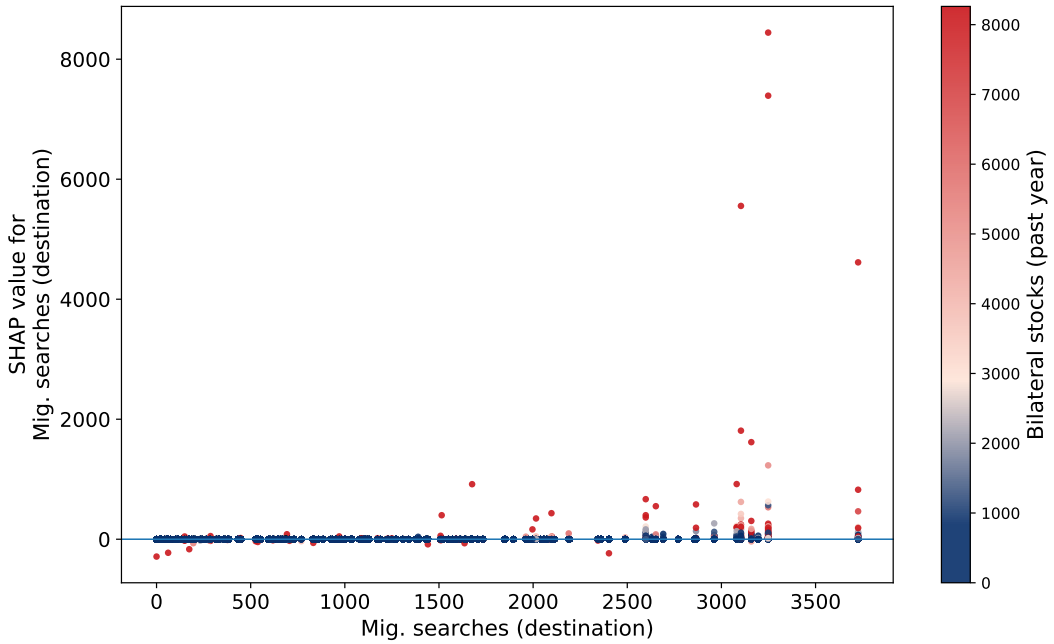
as customs and visa regulations, or broader constraints such as screening and discrimination.[22] Topics related to *Precarious Legal Status* (top words: undocumented, persecute, immigrate) and *Asylum Eligibility* (top words: nationality, eligible, asylum) also contribute positively, pointing to elevated information demand around legal standing and eligibility conditions.

In Figure 11 we return to the idea that searches at destination are used by the model to identify displacement corridors. The Figure presents a SHAP dependence plot for destination-side migration searches at the dyad level. Each point is an origin–destination dyad: the horizontal axis reports the level of migration-related Google Trends searches observed at the destination, and the vertical axis reports the corresponding SHAP value (in people). When stocks are higher (lighter/red points), the same increase in destination searches is much more often linked to large positive SHAP values, indicating that the model amplifies search signals for established corridors. Notably, high stocks alone do not generate positive contributions: with high stocks but low search intensity, SHAP values remain near zero (and can even turn negative). The model is able to exploit a combination of high bilateral displacement stocks and high migration-related searches at destination to predict displacement flows on the dyad.

A useful lens for interpreting the predictive role of destination-side Google searches is the classic migration-networks literature. Munshi (2003) shows that migrant networks

[22] See Appendix C for word cloud visualizations of the mobility barriers and bureaucratic mobility barriers topics.

Figure 11: SHAP dependence plot of destination-side migration searches colored by lagged bilateral stocks



Notes: Each point corresponds to an origin–destination dyad. The horizontal axis reports the level of migration-related Google Trends searches at the destination. The vertical axis reports the SHAP value (in people) associated with destination-side searches, i.e., the marginal contribution of this feature to the model prediction relative to the baseline. Point color indicates level of lagged bilateral migrant stocks (past year). All values are from the $T = 2024$ model.

facilitate location choice by lowering information frictions and smoothing access to opportunities at destination, implying that corridors with stronger networks should exhibit more responsive, higher-frequency signals of mobility intent. Complementarily, McKenzie and Rapoport (2010) documents that networks shape not only migration levels but also who migrates, consistent with networks operating as an information and cost-reduction technology that changes the effective barriers to moving along specific routes. In our dyadic forecasting model destination-side search intensity becomes most informative precisely where bilateral stocks are already high, suggesting that searches by established communities at destination act as a real-time proxy for corridor salience. Searches at destination can therefore complement the variables in the baseline model and improve short-horizon updating of dyadic forecasts even when the outcome is only observed annually.

5 Conclusion

This paper introduces a global Early Warning System (EWS) that uses machine learning techniques and Natural Language Processing (NLP) methods to forecast forced displacement flows at scale, capturing their likelihood, magnitude and destinations. Our study demonstrates the value of augmenting predictor sets with text-based features, showing

benefits in early warning performance. Feeding Google Trends real-time searches and a corpus of millions of media articles into our forecasts enhances our capacity to foresee forced displacement crises, from offering early indications of "silent" country-level onsets that structural indicators fail to signal, to producing corridor-specific volume forecasts that guide resource planning at the origin and destination. The outflows classifier delivers an actionable early-warning alert when risk is still latent, while the dyadic model translates that alert into quantified, origin-to-destination flow projections and their confidence bounds.

Despite the heavy reliance on gravity indicators (distance & past flows), Google Trends add a real-time lens on mobility intentions. Origin-level topic themes (from LDA) capture push intensity at the origin, while pull factors are reflected in both the destination-focused GTI and the destination migration-related Google Trends searches. Overall, these search signals improve the model's out-of-sample predictions beyond gravity-based determinants.

Looking ahead, several avenues can deepen and broaden our forecasting toolkit. First, more work should go into building models that exploit search by compatriots at destination. Our findings suggest that these searches could lend precision to dyadic models. Second, more work should go into building dedicated spike-detection models that can identify and quantify those extreme tail-risk surges. Finally, researchers need to be aware of the ethical constraints to model development. By relying on global UNHCR displacement data, our model shifts the focus away from an exclusive emphasis on migration flows into developed countries and toward the broader dynamics of forced displacement. Our hope is that this can contribute to the massive humanitarian effort necessary to best help refugees.

References

Batra, D., Mercuri, S., and Khraishi, R. (2023). Conformal predictions for longitudinal data. *arXiv preprint arXiv:2310.02863*.

Böhme, M. H., Gröger, A., and Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142:102347.

Boss, K., Groeger, A., Heidland, T., Krueger, F., and Zheng, C. (2023). Forecasting bilateral refugee flows with high-dimensional data and machine learning techniques. *Journal of Economic Geography*, 25:3–19.

Carammia, M., Iacus, S. M., and Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, 12(1):1457.

Central Intelligence Agency (2024). The world factbook: Languages. `https://www.cia.gov/the-world-factbook/`. Accessed: 05-2025.

DRC (2023). Foresight project, technical note. `https://pro.drc.ngo/media/l4qj4ltf/200821_foresight_technical_note.pdf`.

Kurz, C. F., Maier, W., and Rink, C. (2020). A greedy stacking algorithm for model ensembling and domain weighting. *BMC research notes*, 13:1–6.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Martineau, J. S. (2010). Red flags: A model for the early warning of refugee outflows. *Journal of Immigrant & Refugee Studies*, 8(2):135–157.

Masaki, T. and Madson, B. (2023). Data gaps in microdata in the context of forced displacement. Policy Research Working Paper 10631, The World Bank, Washington, DC. License: CC BY 3.0 IGO.

McKenzie, D. and Rapoport, H. (2010). Self-selection patterns in mexico–u.s. migration: The role of migration networks. *The Review of Economics and Statistics*, 92(4):811–821.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Moraga, J. F.-H. and López Molina, G. (2024). Gravity predictions of international migration flows.

Mueller, H. and Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2):358–375.

Mueller, H. and Rauh, C. (2022). The hard problem of prediction for conflict prevention. *Journal of the European Economic Association*, 20(6):2440–2467.

Mueller, H., Rauh, C., and Seimon, B. (2023). Introducing a global dataset on conflict forecasts and news topics. *Data Policy*, 5:e12.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *The Quarterly Journal of Economics*, 118(2):549–599.

Ponticelli, J., Tesei, A., and Manacorda, M. (2024). The international transmission of democratic values: Evidence from the african diaspora. Working Paper, March 20,

2024.

Prokhorenkova, N., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features.

SerpApi (2024). Google trends api. https://serpapi.com/google-trends-api.

Suleimenova, D., Bell, D., and Groen, D. (2017). A generalized simulation development approach for predicting refugee destinations. *Scientific reports*, 7(1):13377.

Suleimenova, D. et al. (2021). Forecasting forced migration by coupling an agent-based simulation approach with weather data. In *EGU General Assembly 2021*, pages EGU21–16086.

UNHCR (2023). Nowcasting of refugee and asylum-seeker statistics. https://www.unhcr.org/refugee-statistics/insights/explainers/nowcasting-refugees-asylum-seekers.html. Accessed: 05-2025.

UNHCR (2024). Forced displacement flow dataset. https://www.unhcr.org/refugee-statistics/insights/explainers/forcibly-displaced-flow-data.html. Accesed: 05-2025.

Welch, N. G. and Raftery, A. E. (2022). Probabilistic forecasts of international bilateral migration flows. *Proceedings of the National Academy of Sciences*, 119(35):e2203822119.

Appendix

A Google Trends

Table A1: Aggregate descriptive statistics of Google Trends search terms (outflows)

	Outflows model			Dyadic model		
Search term	Mean	SD	Share zeros	Mean	SD	Share zeros
advisor advisors adviser advisers	12.21	17.87	0.47	133.61	387.68	0.62
applicant applicants	5.12	12.61	0.64	69.71	294.59	0.71
apply application applications	24.84	24.40	0.25	215.52	544.36	0.57
arrival arrivals	14.29	19.43	0.43	144.13	407.57	0.61
asylum	7.67	14.07	0.57	88.88	286.87	0.66
asylum seeker asylum seekers	1.58	6.39	0.81	24.85	153.70	0.82
border control border controls	1.03	4.82	0.83	15.91	106.49	0.84
camp camps	11.01	16.04	0.41	115.95	354.64	0.60
certificate certificates	14.58	18.65	0.35	145.36	405.12	0.58
checkpoint checkpoints	7.33	15.64	0.68	77.35	294.97	0.74
citizen citizens citizenship citizenships	16.11	19.24	0.33	154.98	429.19	0.58
consulate consulates	8.56	14.02	0.48	86.23	287.74	0.63
crisis crises	10.57	14.48	0.37	101.90	299.42	0.59
customs	12.13	17.66	0.43	123.69	385.24	0.62
deportation deportations deported	2.98	7.83	0.71	35.44	150.27	0.75
detain detained	7.24	15.49	0.65	80.88	298.29	0.72
detention	5.20	12.95	0.70	74.40	298.37	0.74
discriminate discriminatory discrimination	9.67	15.68	0.51	110.97	339.98	0.63
displace displaces displaced	6.63	13.55	0.63	77.57	281.23	0.70
document documents documentation	19.87	21.37	0.29	187.83	490.20	0.57
dual citizenship dual nationality	2.11	7.31	0.78	32.20	163.39	0.80
eligible eligibility	9.03	16.08	0.54	100.03	344.37	0.67
embassy embassies	12.18	15.89	0.31	96.03	306.44	0.58
emigrant emigrants	4.59	9.68	0.65	53.26	194.96	0.72
emigrate emigrated emigration	8.46	14.43	0.56	93.71	300.99	0.67
evacuee evacuees evacuate evacuation	5.46	12.02	0.64	63.77	244.83	0.70
exile exiles	4.48	9.29	0.64	49.88	173.44	0.70
flee flees fled fleeing	9.91	16.10	0.52	108.95	331.07	0.64
foreigner foreigners	12.35	17.26	0.46	121.56	342.79	0.62
genocide	4.35	9.41	0.64	57.61	211.49	0.71
home homeland	25.93	22.57	0.22	211.27	519.39	0.56
host country	1.14	5.58	0.83	17.52	128.23	0.84
humanitarian	5.03	10.88	0.62	56.95	223.83	0.70
identity card identity cards	5.64	12.61	0.67	70.59	264.58	0.73
illegal illegally	10.00	17.40	0.57	120.84	386.53	0.67
immigrant immigrants	6.90	12.67	0.57	81.43	272.77	0.67
immigrate immigrated	4.93	11.39	0.67	57.46	242.37	0.74
immigration	8.31	13.50	0.46	82.06	279.06	0.64

Continued on next page

Table A1 (continued)

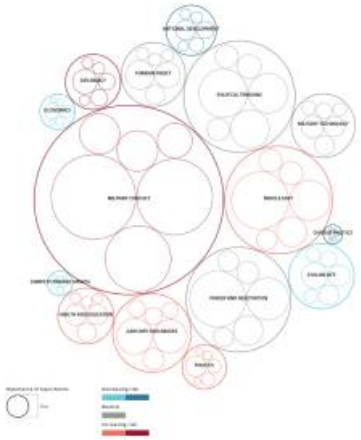
	Outflows model			Dyadic model		
Search term	Mean	SD	Share zeros	Mean	SD	Share zeros
marriage	16.33	18.17	0.32	137.21	365.16	0.59
migrant migrants	6.33	11.70	0.55	71.68	242.89	0.65
migrate migration	11.42	16.62	0.43	123.68	365.22	0.61
militant militants	3.86	9.92	0.70	43.58	185.83	0.74
militia	4.79	11.45	0.70	54.91	214.71	0.74
nationality nationalities	13.34	18.78	0.43	138.30	401.88	0.61
nationalization nationalisation	2.08	7.42	0.77	22.02	127.68	0.80
naturalization naturalisation	3.50	10.36	0.74	46.36	225.42	0.78
passport passports	13.40	17.41	0.38	126.20	357.38	0.59
permit permits	15.00	20.20	0.42	152.93	425.28	0.60
persecute persecutes persecuted	5.72	13.40	0.69	75.15	290.59	0.73
persecution	5.32	12.47	0.69	72.75	272.61	0.73
political asylum	1.34	6.18	0.84	18.57	129.70	0.85
political refugee	0.63	3.85	0.87	7.86	70.68	0.87
protection	14.33	17.92	0.36	135.73	382.68	0.59
quota quotas	9.78	16.53	0.53	111.86	351.14	0.64
refugee refugees	3.37	7.45	0.58	35.00	133.35	0.67
repatriate repatriates repatriated	3.26	9.43	0.72	33.68	169.96	0.77
repatriation	2.65	8.69	0.77	38.69	190.37	0.79
required document required documents	4.07	11.31	0.75	50.91	239.62	0.78
resettle resettles resettled	6.39	16.08	0.72	71.30	310.72	0.78
resettlement	3.37	9.77	0.77	44.30	202.27	0.79
restrict restricts restricted	10.81	18.45	0.55	134.26	416.05	0.65
restriction restrictions	7.71	14.30	0.51	82.54	286.84	0.64
sanctions	6.17	12.28	0.57	60.17	210.13	0.66
schengen	6.26	12.19	0.58	76.25	268.52	0.67
screening	9.46	17.08	0.58	115.16	377.76	0.67
seeker seekers	8.48	14.57	0.53	92.22	305.13	0.65
shelter	6.78	13.67	0.63	85.39	292.26	0.70
smuggler smugglers smuggling	6.53	12.82	0.60	75.22	256.96	0.68
social security	7.86	14.05	0.55	86.64	292.59	0.66
sponsor	7.65	14.06	0.60	99.33	320.61	0.68
spouse spouses	13.53	20.01	0.49	141.27	414.25	0.63
stateless	1.55	6.85	0.82	20.99	143.48	0.83
trafficked trafficking	9.39	16.05	0.55	103.24	330.95	0.66
unauthorised unauthorized	4.26	11.69	0.75	62.74	272.02	0.77
undocumented	1.90	8.03	0.83	25.78	162.07	0.84
uproot uproots uprooted	3.97	10.88	0.74	48.17	220.56	0.78
verification	10.53	19.04	0.57	129.26	420.41	0.67
visa free	3.32	10.35	0.76	42.67	218.77	0.79
visa visas	20.53	20.69	0.25	182.14	474.23	0.57
waiver waivers	8.68	16.26	0.59	111.65	364.01	0.67

Notes: The first three columns report summary statistics for the outflows model's Google Trends search terms, pooled over all origin–year observations. The last three columns report the analogous statistics for the dyadic model, i.e. the same key terms interacted with each destination country name, pooled over all origin–year-destination observations. Mean and SD are the sample average and standard deviation of the raw Google Trends index, while Share zeros is the fraction of observations with no detectable search volume

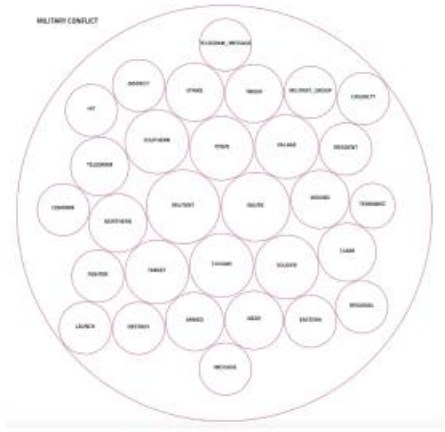
B Newspaper text

Figure B1: Examples of topics and keywords for Syria in November 2024

(A) Topics generated by LDA



(B) Keywords associated to military conflict



Notes: Screenshots from `conflictforecast.org`. See Mueller et al. (2023) for explanation of the data.

Figure A shows the different topic shares generate by the LDA in Syria in November 2024, where "Military Conflict" is shown alongside related topics such as "Political Tensions", "Middle East", "Judiciary and Abuses", and "Civilian Life." The color coding indicates the risk level associated with each topic, with "Military Conflict" and related themes showing increasing risk (indicated by red hues). The size of the circles reflects the importance of each topic, with larger circles representing more prominent themes in the text corpus for that period.

Figure B shows the key terms most associated with the theme "military conflict". Prominent terms like "strike", "troop", "militant", "casualty", and "village" highlight the focus on military operations, armed groups, and civilian impacts within the context of conflict.

C Google Trends topics

The word clouds below present two illustrative examples of migration-related Google Trends topics extracted using LDA.

Figure C1: Examples of topics generated by LDA for migration queries

(A) Topic: mobility barriers

(B) Topic: bureaucratic mobility barriers

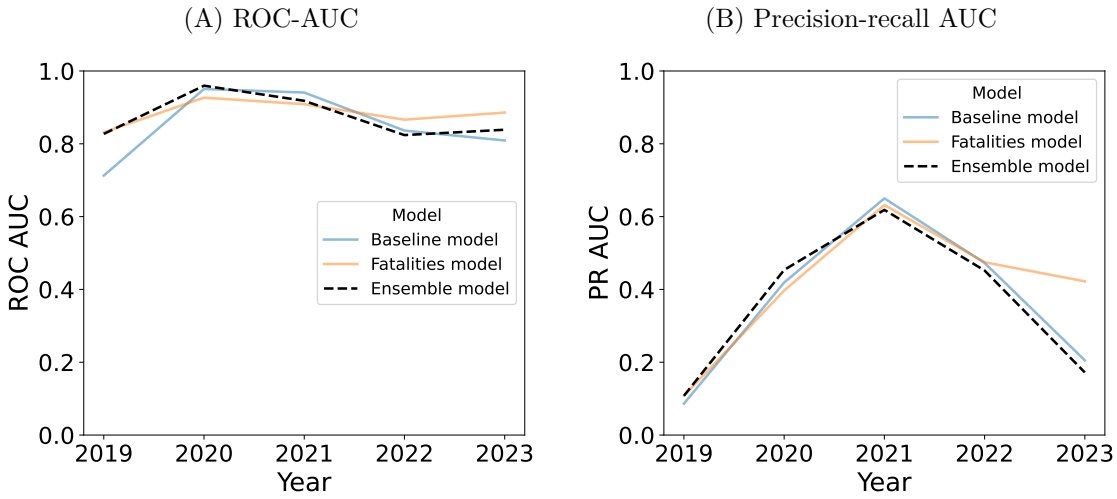
Notes: The panels depict the top keywords of two of the ten LDA topics. The size of the words is proportional to their importance. The location conveys no information.

The first topic, shown in Figure C1A, centers on mobility barriers. Prominent terms such as "screening", "sponsor", "discrimination" and "restriction" highlight obstacles to movement that are not strictly administrative but rather intangible or process-related. These words suggest searches related to eligibility hurdles, discriminatory practices, and procedural checks that migrants may face when attempting to move across borders.

The second topic, displayed in Figure C1B, reflects bureaucratic mobility barriers. Its vocabulary—featuring "customs", "visa", "unauthorized", "quota", "detention" and "advisor"—underscores the formal, state-imposed controls that regulate cross-border mobility. Here, the focus is on visas, quotas, customs procedures, and other institutional mechanisms that shape legal access to host countries.

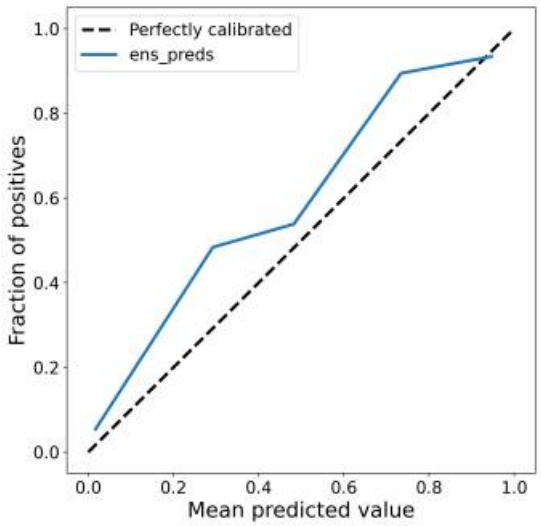
D Additional results

Figure D1: Evolution of evaluation metrics for onset events in test years



Notes: As 2018 is a non-turbulent year with a single onset, we cannot compute onset performance for this test year.

Figure D2: Calibration plot for outflows model



Notes: This plot divides predictions into bins and, for each bin, shows the model's average forecasted probability on the x-axis against the actual fraction of positive events on the y-axis. The dashed 45° diagonal represents perfect calibration, where predicted and observed rates coincide.

E Permutation SHAP for the ensemble regressor

To explain the final displacement forecasts of the ensemble model, we use permutation SHAP values (Lundberg and Lee, 2017). In our setting, the ensemble combines $K = 3$ CatBoost base learners, each trained on a different feature subset. Let $f_k(\mathbf{x})$ denote the prediction of base model k in log-transformed space, where each model outputs

$$f_k(\mathbf{x}) \;=\; \log\big(1 + \tilde{y}_k(\mathbf{x})\big),$$

with $\tilde{y}_k(\mathbf{x})$ representing the flow prediction of model k in original scale (people). After finding optimal stacking weights, the ensemble forms a convex combination in log space,

$$f_{\text{ens}}(\mathbf{x}) \;=\; \sum_{k=1}^{K} w_k\, f_k(\mathbf{x}), \qquad \text{with} \quad \sum_{k=1}^{K} w_k = 1,$$

and the final prediction in the original outcome scale (people) is obtained via the inverse transformation

$$\hat{y}(\mathbf{x}) \;=\; g\big(f_{\text{ens}}(\mathbf{x})\big) \;=\; \exp\big(f_{\text{ens}}(\mathbf{x})\big) - 1,$$

For a given evaluation year t, we treat $\hat{y}(\mathbf{x})$ as a black-box regressor and compute permutation SHAP values on the test dyads for that year. The background data for year t consist of all dyads with years strictly earlier than t, from which we draw a subsample to define an *independent* masker. We then call the permutation-based SHAP explainer with the identity link, so that SHAP values are computed directly in the original outcome scale. This yields, for each observation j and feature i,

$$\hat{y}^{(j)} \;\approx\; \phi_0 \;+\; \sum_{i=1}^{p} \phi_i^{(j)},$$

where ϕ_0 is the expected prediction under the background distribution (the SHAP base value) and $\phi_i^{(j)}$ is the SHAP value of feature i for observation j, measured in people. In this representation, each $\phi_i^{(j)}$ has a straightforward interpretation: it indicates how many individuals feature i contributes to shifting the prediction for observation j above or below the baseline level ϕ_0. Because permutation SHAP is based on Shapley values, it allocates the contribution of interaction effects across features in a way that avoids double-counting, while preserving local additivity in the original space.