



# Probabilistic forecasts of international bilateral migration flows

Nathan G. Welch<sup>a,1</sup> and Adrian E. Raftery<sup>a,b,1</sup>

Contributed by Adrian E. Raftery; received March 3, 2022; accepted July 15, 2022; reviewed by Peter Smith and Carl Schmertmann

We propose a method for forecasting global human migration flows. A Bayesian hierarchical model is used to make probabilistic projections of the 39,800 bilateral migration flows among the 200 most populous countries. We generate out-of-sample forecasts for all bilateral flows for the 2015 to 2020 period, using models fitted to bilateral migration flows for five 5-y periods from 1990 to 1995 through 2010 to 2015. We find that the model produces well-calibrated out-of-sample forecasts of bilateral flows, as well as total country-level inflows, outflows, and net flows. The mean absolute error decreased by 61% using our method, compared to a leading model of international migration. Out-of-sample analysis indicated that simple methods for forecasting migration flows offered accurate projections of bilateral migration flows in the near term. Our method matched or improved on the out-of-sample performance using these simple deterministic alternatives, while also accurately assessing uncertainty. We integrate the migration flow forecasting model into a fully probabilistic population projection model to generate bilateral migration flow forecasts by age and sex for all flows from 2020 to 2025 through 2040 to 2045.

bilateral migration flows | international migration | probabilistic forecasting | Bayesian hierarchical model

Recent methodological advances have made it possible to generate plausible estimates of international migration flows at a global scale. Abel (1) pioneered a method to estimate the minimum number of people who must have changed their country of residence to explain the change in migrant stocks among all countries of the world. Azose and Raftery (2) extended the minimum migration estimates to produce well-calibrated pseudo-Bayes estimates of bilateral migration flows. This was found to perform best among six methods for estimating international migration, in the sense of having the highest correlation with several common measures of migration (3).

Accurate estimates of historic migration trends and forecasts of future trends are essential to crafting effective migration policies (4), but flow forecasting method development has lagged progress in flow estimation (5). Gravity models use push factors to help explain the magnitude of out-migration from a country along with pull factors to help explain the magnitude of country-level in-migration. Work with these models has used estimates of migration flows from most of the world to a few wealthy countries and vice versa to quantify the influence of push and pull factors on the magnitudes of migration flows (6, 7). Alternatives to the gravity model approach are concerned with migration flow forecasting for a subset of countries or regions (8, 9).

We address the problem of probabilistic forecasting of international migration flows between all pairs of countries. Our approach uses a Bayesian hierarchical model that pools information across time periods and individual flows (10), helping to compensate for the small number of periods where migration flow estimates are typically available. The Bayesian approach also makes it possible to encode outside information in the model, helping to rein in implausibly large forecast variability.

This paper describes a Bayesian hierarchical model that builds on a key idea from ref. 11: Once the overall level of migration is controlled for, spatial distribution patterns are found to be remarkably stable. Rogers et al. (12) and Raymer et al. (13) used this idea to model regional or subnational migration flows. Our approach models the out-migration rate by origin country and time period. Multiplying the country-level outflow rate by the population at risk for out-migrating each period yields the number of people to allocate to each destination. Our destination model jointly allocates this migrant total to all destinations, conditional on the origin country. This conditional origin framework makes it possible to capture the variability in spatial interactions across flows from the same origin. It also yields forecasts of net migration by country that sum to zero by construction.

Out-of-sample evaluation results indicate that our model generates plausible, well-calibrated forecasts of bilateral migration flows over a short time horizon. These accommodate the possibility of major migration shocks in the future. We use our fitted model to generate probabilistic forecasts of global bilateral migration flows for quinquennial periods from the 2020 to 2025 period to 2040 to 2045.

## Significance

Accurate forecasts of migration trends are essential for effective migration policies. We develop a method for probabilistic forecasting of international migration flows between pairs of countries. Our model encodes the observation that the spatial distribution patterns of migration are stable over time. This model produces one-period-ahead probabilistic forecasts that are more accurate than a leading alternative and are well calibrated for international bilateral migration flows, as well as country-level migrant inflows, outflows, and net migration flows. The flow forecasts are sex and age specific and properly account for population change not attributable to migration.

Author affiliations: <sup>a</sup>Department of Statistics, University of Washington, Seattle, WA 98195; and <sup>b</sup>Department of Sociology, University of Washington, Seattle, WA 98195

Author contributions: N.G.W. and A.E.R. designed research, performed research, analyzed data, and wrote the paper.

Reviewers: P.S., University of Southampton; and C.S., Florida State University.

The authors declare no competing interest.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

<sup>1</sup>To whom correspondence may be addressed. Email: [nwelch@uw.edu](mailto:nwelch@uw.edu) or [raftery@uw.edu](mailto:raftery@uw.edu).

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2203822119/-DCSupplemental>.

Published August 22, 2022.

**Table 1. Out-of-sample MAE in thousands of migrants per period, MAPE, and 95% prediction interval (PI) coverage for models fitted to all 1990 to 1995 through 2010 to 2015 migration flows and tested on all 39,800 2015 to 2020 flows**

Method	MAE	MAPE	95% PI			
			Flow	In	Out	Net
Historic mean flow	1.2	139	—	—	—	—
Persistence	<b>1.0</b>	79	—	—	—	—
Gravity model	3.0	1,565	86	77	80	99
Poisson hurdle model	10.0	25,649	90	66	65	48
Bayesian flow model	1.2	<b>76</b>	<b>93</b>	<b>87</b>	<b>92</b>	<b>94</b>

A bold-face entry indicates the most accurate number for that metric.

## Results

**Out-of-Sample Validation.** We fitted five competing models of bilateral migration flows for the first five periods for which flow data were available, i.e., 1990 to 2015, using only data that would have been available at the start of the 2015 to 2020 period (2, 14). Fitted models were then used to predict the bilateral flows observed in the 2015 to 2020 period (3, 15). This approach offers some indication of how our forecasting approach might perform one period ahead. We calculated the mean absolute error (MAE) and the mean absolute percentage of error (MAPE) for all models and the prediction interval (PI) coverage for the probabilistic methods (16, 17).

The MAE measures the average absolute difference between each flow forecast and the observed flow for the 2015 to 2020 period. It summarizes prediction error on the same scale as the data for all flows. The MAPE normalizes the error by the magnitude of the observed flow, after adding 1 to both the observed and predicted flows to avoid infinite values. Normalizing errors by the magnitude of the observed flow puts the magnitude of errors for different pairs of countries on the same scale.

We also evaluated competing probabilistic methods by comparing their 95% prediction interval coverages. If a probabilistic flow model is well calibrated, then the 95% prediction intervals from a model fitted to the 1990 to 2015 flows should include about 95% of the 2015 to 2020 flows. Prediction interval coverage estimates that differ from the nominal value (95% in this case) indicate that a model may be poorly calibrated, misspecified, or both.

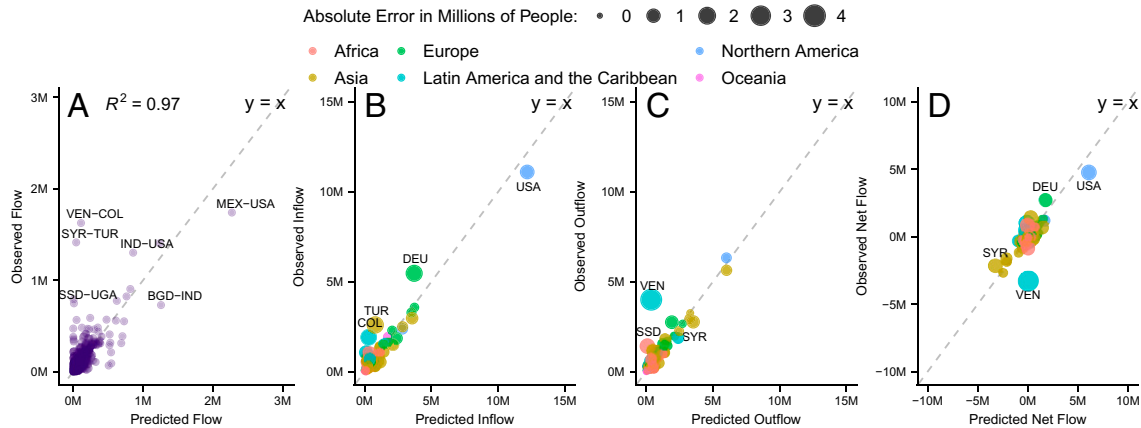
Table 1 summarizes the performance of the Bayesian flow model (BFM) alongside two simple deterministic approaches, a

standard gravity model and the Poisson hurdle model, which is a more complex gravity model (6, 7, 18). Let  $m_{i,j,t}$  denote the number of migrants from origin  $i$  to destination  $j$  during the period starting in year  $t$ . The historic mean flow model generates forecasts by projecting each bilateral flow mean over the first five periods into the 2015 to 2020 period. The historic mean flow forecast for the 2015 to 2020 period for the flow from origin  $i$  to destination  $j$  is defined as  $\bar{m}_{i,j} = \frac{1}{5} \sum_{t=1990}^{2010} m_{i,j,t}$ . The persistence flow model projects each of the most recently observed bilateral flows forward to the next period. The persistence flow forecast for the 2015 to 2020 period for the flow from origin  $i$  to destination  $j$  is defined as  $m_{i,j,2010}$ . The gravity and hurdle model specifications are detailed in *Materials and Methods*.

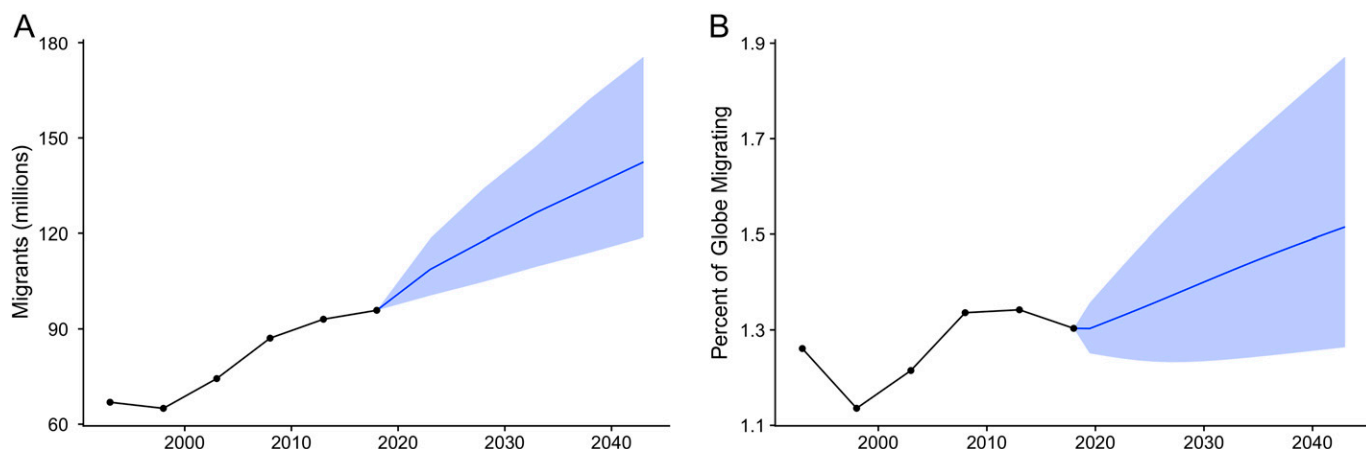
Among the probabilistic models, the BFM was the best calibrated for bilateral flows, total inflows, total outflows, and net migration flows. The BFM had the lowest MAPE among the methods considered and clearly outperformed both gravity models in terms of both MAE and MAPE. Interestingly, the simple deterministic models performed similarly to the BFM in terms of MAE and MAPE, but they fall short in that they do not produce prediction intervals.

Fig. 1*A* summarizes the distribution of observed and predicted flows for the BFM. Points that fall along the dashed line  $y = x$  indicate perfect agreement between the single best forecast and the observed flow. The estimated  $R^2$  values from the Poisson hurdle model,  $R^2 = 0.93$ , improve on the gravity model,  $R^2 = 0.83$ , and the BFM, with  $R^2 = 0.97$ , improves on both. The BFM is a good model for most flows, but there are several examples where the predicted flow leads to large errors. Examples include the flows from Venezuela (VEN) to Colombia (COL), from Syria (SYR) to Turkey (TUR), from Mexico (MEX) to the United States (USA), and from South Sudan (SSD) to Uganda (UGA). Large errors for flows originating in Venezuela and in Syria arise from major political crises in those countries (19, 20). Departures from historic norms in the Mexico to United States flow could be partly explained by actual or perceived changes in immigration policy by the Trump administration between 2017 and 2020 (21). Large errors associated with South Sudan are reasonable since the country was founded only in 2011 (22). These cases show that major migration shocks can generate observations that fall far in the tail of the forecast distribution.

Fig. 1*B–D* shows country-level forecasts of 2015 to 2020 flows compared to the best available estimates of country-level flows. Median BFM inflow forecasts for Germany (DEU) and



**Fig. 1.** (A–D) Observed 2015 to 2020 (A) flow, (B) total country inflows, (C) total country outflows, and (D) total country net flows compared to Bayesian hierarchical model median forecasts colored by United Nations Area and sized according to the absolute error in millions of people.



**Fig. 2.** (A and B) Global migration flows in millions of migrants (A) and in percentage of global population migrating (B) during 5-y periods observed from 1990 through 2020 with median forecast (solid line) and 90% prediction interval for 5-y periods from 2020 through 2045.

Turkey were smaller than the estimated values for the 2015 to 2020 period. The median BFM outflow forecast for Venezuela was smaller than the observed outflow while the median Syria outflow forecast was too large. Country-level inflow and outflow forecasts implicitly define net flow forecasts. The error in Venezuela's outflow forecast was carried over into the net flow forecast as shown in Fig. 1C: The net flow was smaller than the median forecast. This is consistent with the very large number of people that left Venezuela during the period. In summary, out-of-sample performance of bilateral flow forecasts and aggregate measures of migration indicates that the BFM was well calibrated to the best available estimates of global bilateral migration flows in the short term.

**Forecast Evaluation.** We produced forecasts for all 39,800 flows from the 2020 to 2025 period through the 2040 to 2045 period using the most updated estimates of bilateral migration flows from period 1990 to 1995 through period 2015 to 2020 (3). A summary of each bilateral flow forecast is included in *SI Appendix*.

One appeal of well-calibrated probabilistic migration flow forecasts is that aggregate quantities or functions of flows lead to valid approximations of the statistical distribution associated with the function applied to individual flow forecast trajectories. We use this fact to evaluate the BFM forecasts of country-level inflows, outflows, and net flows and the percentage of the globe migrating. Global flow estimates are available for only six periods, making multiple-period-ahead out-of-sample evaluation infeasible. We use aggregate measures of migration to assess the plausibility of longer-term forecasts from the BFM.

Fig. 2 shows the estimated and forecasted number of people migrating globally in each period. The median forecast is that the number of people migrating in 2040 to 2045 will be nearly 50% larger than the number of people migrating in the 2015 to 2020 period. However, much of this increase is due to the projected increase in world population. After accounting for global population growth, the percentage of world population migrating increases from about 1.3% in the 2015 to 2020 period to 1.5% in the 2040 to 2045 period—an increase of only about 16%.

Table 2 summarizes the estimated increase in the total number of people migrating around the world for the last period where data are available (2015 to 2020) and the last period in the forecast (2040 to 2045). Growth in global migration will be driven first by the increasing global population and to a lesser extent by growth in the out-migration rates in a few large countries.

**Case study: Germany.** Fig. 3 shows the BFM forecasts for Germany in millions of migrants per 5-y period from 2020 through 2045. Fig. 3A shows the historic net migration rate in Germany from 1950 through 2020 along with the United Nations (UN) projection of net migration from the 2019 Revision of World Population Prospects (WPP) (15). The UN net migration projection falls within the 90% prediction interval constructed from joint forecasts of bilateral flows into and out of Germany. Even though the median forecast and prediction interval suggest more positive net flows into Germany using the BFM, our net migration forecast interval contains the UN's net migration projection for all forecast periods. The BFM forecast of net flows also appears plausible given the range of past net flows; however, the 90% prediction intervals from the BFM indicate that net outflows from Germany are possible over the coming years. This net outflow could occur if for example many of the migrants who fled humanitarian crises over the last decade return to their home countries once the humanitarian situation ends.

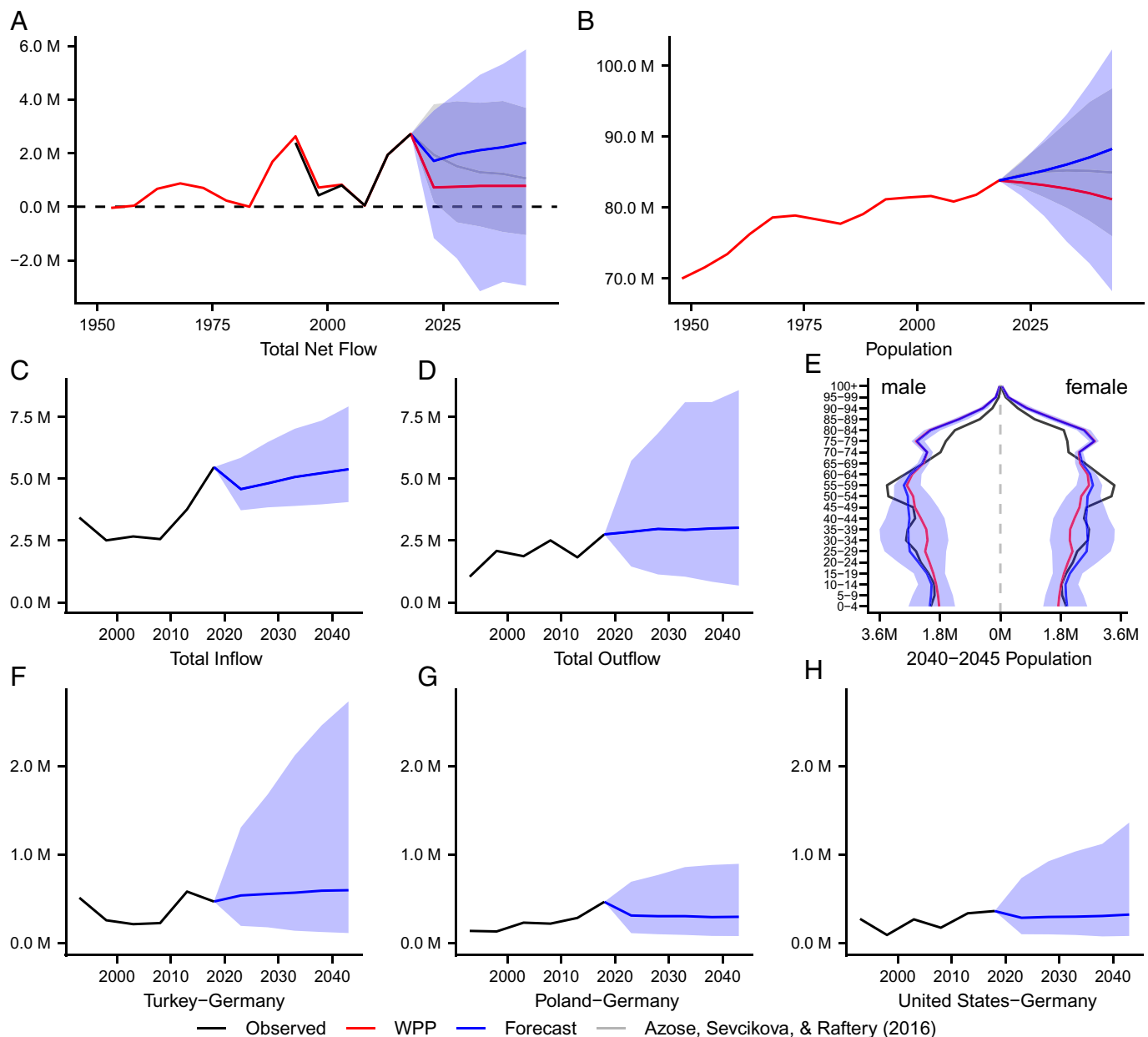
Fig. 3B shows the range of uncertainty about the total population of Germany generated by the net migration flow forecast through 2045. The median population forecast using the BFM is higher than the WPP 2019 projection and shows large uncertainty due to uncertainty about total outflows from Germany and possible large inflows from a few countries.

Fig. 3A and B also shows the 90% prediction interval using the population projection methods in Azose et al. (23) fitted to WPP 2019 data. The 90% BFM prediction intervals for net migration and population counts are wider than the probabilistic population projections using the net migration model in ref. 23. Wider prediction intervals using the BFM are reasonable as quinquennial bilateral global flow data are available in fewer than half the

**Table 2. Total global migration in millions of migrants per 5-y period and percentage of the population migrating**

	2015 to 2020	2040 to 2045 forecast		
		5%	50%	95%
Sum of global flows, millions	96	119	142	176
Percentage of population migrating	1.3	1.3	1.5	1.9

The columns correspond to the 5th, 50th (median), and 95th percentiles of the predictive distribution.



**Fig. 3.** (A–H) Observations and 90% prediction interval forecasts in millions of people per 5-y period for (A) total net flow, (B) population, (C) total inflow, (D) total outflow, (E) population by age and sex (black denotes 2015 to 2020 period), and (F–H) bilateral flows with Germany as origin in descending order by historic magnitude.

number of periods where country-level net migration data are available. Probabilistic net migration model prediction intervals can be wider than BFM intervals if, for example, variation in the net migration rate is much larger in periods prior to 1990 to 2020.

Fig. 3 C and D shows the total estimated flows into and out of Germany every 5 y from 1990 to 2020 and forecasts of flows from 2020 through 2045 implied by the BFM. The outflow forecast effectively continues the historic pattern with relatively wide uncertainty. The median inflow forecast moderates from the historic peak in the 2015 to 2020 period before growing again through 2045. Inflow forecast patterns like this reflect a return to long-term historic destination preferences encoded in the destination component of the model and/or a fall in the total outflow from one or more countries from one period to the next. Inflow forecast counts tend to increase with the increasing populations in sending country populations.

The cumulative impacts of migration uncertainty on the age profile of the German population by the end of the forecast period in 2040 to 2045 are shown in the probabilistic population pyramid in Fig. 3E. The wide intervals for the 20- to 50-y age groups reflect uncertainty about migration and the fact that international migration tends to be largely concentrated in the 20- to 35-y age group. The median forecast of the German age profile is similar to the current profile, but with wide uncertainty.

Fig. 3 F–H shows estimates of the largest flows into Germany from 1990 to 2020 and the BFM forecasts for those flows from 2020 to 2045. The median forecasts are approximately equal to the mean of the historic flows. Flow forecasts from Turkey to Germany contribute the largest number of migrants to the median flow forecast along with the largest degree of uncertainty in future inflows. See *SI Appendix* for all country-level forecast summary plots.

## Discussion

Bayesian probabilistic forecasting methods for demographic processes have become increasingly widespread in the last 15 y (8, 23–26), and here we extend this approach to forecasting bilateral international migration flows. Our approach builds on the multiplicative components model (13), arguments for separating the outflow rate and spatial interaction process (12), and the relative stability of the spatial distribution of global flows over the past three decades (11, 27). The mean departure rate functional form is inspired by the net migration flow model of ref. 28 and reflects the fact that the net migration rate can be decomposed into inflow and outflow components.

We model out-migration and spatial interaction conditionally on the origin and population at risk for out-migration. This approach separates the magnitude of a flow from the spatial interaction as in refs. 11–13; however, conditioning on the origin and population at risk is a departure from these methods. We also model spatial interactions jointly, conditional on the origin and magnitude of outflows. This approach leads to implicit estimates of country-level inflows and net flows. It also ensures that the magnitude of global outflows equals that of global inflows.

Spatial interaction is modeled with a centered logratio model (29). Our spatial interaction model does not use covariates, but could accommodate covariates in estimating the spatial interactions. Others have suggested the logratio model for the multinomial likelihood model of migration flows (30). We evaluated that formulation as well, but chose the centered logratio formulation to remove the influence of the choice of a baseline country on forecasts of destination preferences of all other countries.

In principle, our approach could yield nonzero probabilities of implausibly dense populations in some countries. In practice, we did not observe this to any great extent, but if desired it could be addressed by incorporating country-level inflow limits in the model; we did not do so here. We did, however, put thresholds on the fraction of the population leaving in any one period, as described in *Materials and Methods*.

It is too early to quantify the impact of the COVID-19 pandemic (31) on global migration flows over the 2020 to 2025 period. Early indications from some parts of the world suggest that global migration may have fallen during the 2020 to 2021 period due to strict border controls (32). In other places, flows may be on track to hit historic highs despite the pandemic (33).

The flow estimates that we used to fit the BFM rely on migrant stock estimates that are compiled every 5 y (3). Data generated from social media platforms, search engine inquiries, and other digital trace data might offer alternative and more timely sources of migrant stock estimates (34–41). Bias-corrected migrant stock estimates derived from big data have the potential to improve the time resolution of migration flow estimates, especially for regional and subnational contexts where platform adaptation among the population is more widespread. However, early applications of digital trace data used to study demographic and public health trends suggest that additional work may be needed to resolve the signals present in big data (37, 42, 43).

Large outflows from India, China, Indonesia, and several African countries have little impact on the overall population of those countries through 2045. However, uncertainty about the magnitude of outflows from these countries generates substantial uncertainty in the population age distribution in both sending and receiving countries by the end of the forecasting period. Also, very large outflows from India, China, Indonesia, and populous countries in Africa are possible and could generate very large inflows for a few popular destinations, unless destination countries constrain the flow of migrants into their countries.

Our choice of a time-invariant destination model reflects the fact that the share of migrants leaving one region for another has been relatively stable over time and that we expect this pattern to persist. An origin's outflow rate and population size are the main factors influencing the magnitude of outflows in the BFM. Even if departure rates continue to follow historic patterns, the absolute number of migrants leaving high-fertility countries or ones with currently young populations will increase.

## Materials and Methods

**Data.** BFM out-of-sample results and forecasts are based on the pseudo-Bayes estimates of flows during 5-y periods starting in 1990 and running through 2020 (2, 3). The flow matrix for  $C$  countries in the period starting at time  $t$  is

$$M_t := \begin{pmatrix} 0 & m_{1,2,t} & \dots & m_{1,C,t} \\ m_{2,1,t} & 0 & \dots & m_{2,C,t} \\ \vdots & \vdots & \ddots & \vdots \\ m_{C,1,t} & m_{C,2,t} & \dots & 0 \end{pmatrix}, \quad [1]$$

where  $m_{i,j,t}$  is the flow of individuals from country  $i$  to country  $j$  during the period starting at time  $t$ . The off-diagonal entries show the total number of movers whose place of residence at the end of the period was different from the one they had at the beginning of the period. The diagonal entries are set to zero as we are interested only in modeling the magnitude of migration flows.

The sum of the entries in row  $i$ ,  $m_{i,+t} = \sum_j m_{i,j,t}$ , is the total number of people whose residence was in country  $i$  at the start of period  $t$  but was some other place at the end of period  $t$ . This sum approximates the outflows from origin  $i$  during the period. Similarly, the sum of the entries in column  $j$ ,  $m_{+,j,t} = \sum_i m_{i,j,t}$ , is the total number of people whose residence was in country  $j$  at the end of the period but was somewhere else at the beginning of the period. This is the total of the inflows to destination  $j$  during period  $t$ . The net flow for country  $c$  is  $r_{c,t} = m_{+,c,t} - m_{c,+t}$ .

The flow matrix can underestimate the total number of people who migrated during each 5-y period. Some people will start and finish the period with residence  $i$  even though they established multiple residences other than  $i$  throughout the period. A person might also start the period with residence  $i$ , establish several residences throughout the period, and reside in country  $j$  at the end of the period. The flow estimate would capture only the change from  $i$  to  $j$ .

Flow estimates are available for  $C = 200$  countries during the six 5-y periods starting in years  $t \in \{1990, 1995, 2000, 2005, 2010, 2015\}$ . This translates to 39,800 bilateral migration flows observed during six periods starting in 1990 and ending in 2020. We do not address uncertainty in historic migration flow estimates; however, future work might account for such uncertainties in the historic flow estimates to reflect an additional source of forecast uncertainty.

**Bayesian Flow Model.** We fitted a Bayesian hierarchical model to all available flow estimates. The model takes advantage of three properties of the data-generating process: 1) Every individual must start from one origin, 2) every individual can choose just one of  $(C - 1)$  possible destinations, and 3) the spatial distribution of migration flows remains relatively constant over time. We exploit these three properties in the specification of the hierarchical model, which is as follows:

$$\begin{aligned} \text{Observations} & \left\{ m_{i,j,t} \mid \pi_{i,j,t}, \delta_{i,t} \stackrel{\text{ind}}{\sim} \text{Multinomial}(N_{i,t}, \pi_{i,j,t}) \right. \\ \text{Outflow} & \left\{ \begin{aligned} N_{i,t} &= \lfloor \delta_{i,t} P_{i,t} + 1/2 \rfloor \\ \log \delta_{i,t} &\stackrel{\text{ind}}{\sim} \text{Normal}((1 - \phi) \mu_i + \phi \log \delta_{i,t-1}, \sigma_i^2) \\ \phi &\sim \text{Uniform}(0, 1) \\ \mu_i &\stackrel{\text{iid}}{\sim} \text{Normal}(\nu, \tau_0^2) \\ \nu &\sim \text{Normal}(\mu_0, 100^2) \\ \sigma_i &\stackrel{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \end{aligned} \right. \end{aligned}$$



$$\text{Inflow} \begin{cases} \pi_{i,j,t} = \exp \eta_{i,j,t} / \sum_{j \neq i} \exp \eta_{i,j,t} \\ \eta_{i,j,t} \stackrel{\text{ind}}{\sim} \text{Normal}(\kappa_{i,j}, \psi_{i,j}^2) \\ \kappa_{i,j} \stackrel{\text{iid}}{\sim} \text{Normal}(0, 10^2) \\ \psi_{i,j} \stackrel{\text{iid}}{\sim} \text{Beta}(p_0, q_0), \end{cases}$$

where “ind” means independently distributed, “iid” means identically and independently distributed, and  $\lfloor x \rfloor$  denotes the floor of  $x$ , i.e., the largest integer smaller than  $x$ . Table 3 defines each term of the model.

Under this specification, the expected flow from origin  $i$  to destination  $j$  during period  $t$  is

$$\mathbb{E}[m_{i,j,t} | \pi_{i,j,t}, \delta_{i,t}] = \pi_{i,j,t} N_{i,t} \approx \pi_{i,j,t} \delta_{i,t} P_{i,t}. \quad [2]$$

This expected value encodes the three defining features of the generative process. Each flow is composed of individuals who belonged to the population of the origin country at the start of the period. Only a subset of the country's population migrates, and that fraction for origin country  $i$  in period  $t$  is  $\delta_{i,t}$ . Every individual leaving origin  $i$  in period  $t$  migrates to only one destination, and the distribution of destinations is relatively stable over time. Destination preference stability is encoded in the parameterization of the  $\pi_{i,\cdot,t}$  component of the model.

For each origin, the out-migration rate is modeled on the logarithmic scale as a weighted average of two quantities: the logarithm of the last departure rate,  $\log \delta_{i,t-1}$ , and the long-term mean rate for country  $i$ ,  $\mu_i$ . Stochastic variation around  $\log$  out-migration rates is captured by  $\sigma_i^2$ . This model ensures that departure rates never drop below zero.

The mean of  $\pi_{i,j,t}$  over  $t$  represents the long-term relative tendency of migrants from origin  $i$  to move to destination  $j$ . Variables in this model are time constant to reflect the assumption that destination preferences encoded by  $\pi_{i,\cdot,t}$  are stable from period to period. Variation about the long-term tendencies from period to

period is represented by a random intercept mixed-effects model. Spatial distribution tendencies ( $\pi_{i,\cdot,t}$ ) are encoded by an overparameterized model equivalent to the centered logratio (clr) transformation (29); namely,

$$\eta_{i,j,t} = \text{clr}(\pi_{i,j,t}) = \log \frac{\pi_{i,j,t}}{\left(\prod_{j=1}^{C-1} \pi_{i,j,t}\right)^{1/(C-1)}}. \quad [3]$$

This model provides a flexible framework that could be extended to take origin- and/or destination-specific variables into account, e.g., colonial relationships, shared language, and economic differences creating push/pull forces among country pairs. A sum-to-zero constraint on  $\kappa_{i,j}$  makes this model identifiable.

The model is implemented by Markov chain Monte Carlo (MCMC) using the R NIMBLE software package (44–46). NIMBLE is designed to sample from the posterior distribution of a Bayesian hierarchical model and is optimized for computational efficiency; however, sampling the model remained computationally slow. We overcame the computational challenges by approximating the model and splitting it into an outflow component and an inflow component.

The outflow component includes all parts of the model involving  $\delta_{i,t}$ . Country-level departure rates,  $\delta_{i,t}$ , are hidden variables in our model; however, the total number of people departing origin  $i$  in period  $t$  is only important to the multinomial likelihood in that it yields  $N_{i,t} = \sum_j m_{i,j,t}$ . This makes it possible to sample the posteriors for  $\mu_i$ ,  $\phi$ , and  $\sigma_i$  by approximating  $\delta_{i,t}$  with the observed out-migration rate for origin  $i$ ,  $d_{i,t} = \sum_j \frac{m_{i,j,t}}{P_{i,t}}$ . Our model assumes  $\delta_{i,t} > 0$ ; all observed  $d_{i,t}$  are positive.

The inflow component includes all parts of the model involving  $\pi_{i,j,t}$ . Information in the data shared across different origin countries is fully contained in the priors for  $\kappa_{i,j}$  and  $\psi_{i,j}$ . This makes it possible to parallelize this portion of the MCMC sampler. A simulation in [SI Appendix](#) showed that inferences from the full model and the approximate parallel implementation were very similar.

The  $\delta_{i,t}$  approximation and parallelized inflow implementation reduced processing time by several orders of magnitude. Parallelized performance gains, however, make it necessary to specify a number of hyperparameters that might otherwise be estimated simultaneously with the rest of the parameters in the model.

**Prior Specification.** There are several user-specified parameters in the BFM. We set the prior parameters using empirical aggregate metrics across all flows. This preserves the information-sharing benefits of hierarchical modeling while improving computation time by several orders of magnitude as follows:

- $a_0, b_0$ : These define the prior distribution of the  $\sigma_i$ . More than 97% of the SDs of  $\log d_{i,t}$  by origin are smaller than 1. Values of  $\sigma_i$  near 0 can lead to implausibly narrow intervals for some small countries. Values of  $\sigma_i$  larger than 1 can yield a few highly variable outflows, leading to implausibly large proportions of the population out-migrating. We therefore set  $a_0$  and  $b_0$  so that values of  $\sigma_i$  close to zero are unlikely and values above 1 are impossible. We did this by minimizing the sum of the differences of the 2.5 and 97.5% quantiles from  $\text{Beta}(a_0, b_0)$  and (0.15, 0.99).
- $\tau_0$ : This parameter represents the amount of variation around the long-term outflow mean for each origin. We set this value equal to approximately three times the SD of the mean log origin outflow rate over  $t$ .
- $\mu_0$ : This parameter represents the mean of long-term means of the log outflows for each origin. We set it equal to the mean over  $i$  of the log origin outflow rate means over  $t$  within each origin  $i$ .
- $p_0, q_0$ : We found  $p_0$  and  $q_0$  by minimizing the sum of the differences between the 2.5 and 97.5% quantiles of the  $\text{Beta}(p_0, q_0)$  distribution and the quantiles of means by origin of SDs of  $\text{clr}(\pi_{i,j,t})$  over  $t$  for each destination  $j$  for all positive flows. We calibrated this prior to means to avoid degenerate specifications that can oversample values near zero. This approach implies that realizations of  $\psi_{i,j}$  are sampled around the distribution of the SD grand mean.

**Forecasting.** Migration flow data used to fit the BFM are not disaggregated by age or sex; however, forecasts must be disaggregated by age and sex. Age-specific and sex-specific forecasts have implications for fertility and mortality in the sending and receiving countries. We use the Rogers–Castro migration age

**Table 3. Bayesian flow model notation**

Parameter	Definition
$m_{i,j,t}$	Integer-valued flow (data) from origin $i$ to destination $j$ during period starting in year $t$
$N_{i,t}$	Number of migrants departing origin $i$ for period starting in year $t$ after rounding to the nearest integer value
$P_{i,t}$	Number of person years in origin $i$ and period $t$ (000s)
$\delta_{i,t}$	Out-migration rate in migrants per 1,000 person years for origin $i$ for period starting in year $t$
$\pi_{i,j,t}$	Conditional probability of moving to $j$ given the origin is $i$ in period $t$ with $\pi_{i,i,t} = 0$ and $\sum_j \pi_{i,j,t} = 1$ for all $i$ and $t$
$\pi_{i,\cdot,t}$	Conditional probability vector of migrants from origin $i$ moving to destinations $j \neq i$ in period $t$ with $\pi_{i,i,t} = 0$ and $\sum_j \pi_{i,j,t} = 1$
$\phi$	Global weight on mean departure rate function
$\mu_i$	Long-term mean out-migration rate for origin $i$
$\sigma_i^2$	Temporal variation around departure rate from origin $i$
$\eta_{i,j,t}$	Destination weight at time $t$ for destination $j$ among migrants from origin $i$
$\kappa_{i,j}$	Mean destination weight for $j$ among migrants from origin $i$
$\psi_{i,j}^2$	Variance of destination weights $j$ from origin $i$
$\nu$	Grand mean of long-term departure rates
$\mu_0$	User-specified prior for long-term departure rate grand mean
$\tau_0^2$	User-specified variance about the long-term departure rate grand mean
$a_0, b_0$	User-specified parameters for temporal variation around departure rates
$p_0, q_0$	User-specified parameters for destination probability variation for origin $i$
$C$	Number of countries

schedule used in the United Nations projections for many countries including China in the 2020 to 2025 period (15, 47). This schedule was used for all countries and forecast periods. Sex-specific forecasts are necessary to generate accurate impacts of fertility changes for the sending and receiving countries. We modeled the sex composition of outflows by splitting outflow forecasts by sex in proportion to the empirical distribution of the origin population.

After drawing  $S$  samples from the posterior distribution of the parameters of the BFM, we generated probabilistic forecasts as follows:

For  $s = 1, \dots, S$ ,

- 1) Sample a departure rate for origin  $i$  during period  $t$  from the posterior predictive distribution for  $\delta_{i,t}^{(s)}$  and convert it to an integer  $N_{i,t}^{(s)} = \text{round}(\delta_{i,t}^{(s)} P_{i,t})$ .
- 2) Sample an allocation vector from the posterior predictive distribution of  $\pi_i^{(s)}$ .
- 3) Sample a flow vector from the density  $m_{i,t}^{(s)} | N_{i,t}^{(s)}, \pi_i^{(s)} \sim \text{Multinomial}(N_{i,t}^{(s)}, \pi_i^{(s)})$ .
- 4) Distribute outflow counts by age group according to a Rogers-Castro migration schedule (47) and sex according to the observed proportions present in the population.
- 5) Increment  $t$ ; update  $P_{i,t}$  for all  $i$  with population changes due to migration, fertility, and mortality; and repeat until reaching the last period of the forecast.

For  $C$  countries,  $T$  periods, and  $S$  samples, this procedure leads to a  $C \times C \times T \times S$ -dimensional array. Each flow matrix sampled using this procedure is guaranteed to yield zero total net global migration, i.e., to ensure that the sum of inflows and the sum of outflows across countries are equal. Furthermore, country inflows, country net flows, regional flows, and global flows are all implicitly defined by the model.

Population projections were generated from a custom-adapted version of the bayesPop R package (48). Our implementation makes it possible to trace inflows back to specific origins by age and sex for each period. Models of mortality and fertility implemented in bayesPop make forecasts generated from our implementation bona fide population projections based on bilateral migration flows rather than country-level net migration.

We put thresholds on the fraction of the population leaving in any one period to avoid unprecedented decreases in origin country populations. Thresholds were empirically calculated using the historic flow and population data. Migration forecasts among Gulf Cooperation Council (GCC) member countries (United Arab Emirates, Bahrain, Kuwait, Qatar, Oman, Saudi Arabia) and labor-supplying countries (Bangladesh, Egypt, India, Indonesia, Pakistan, Philippines) require special treatment (23). We calculated GCC and GCC-labor bounds as the maximum observed percentage of the population departing these countries from 1990 to 2020. GCC member countries were constrained so that no more than 42% of the population departs in any single period. GCC labor-supplying countries were constrained so that no more than 3% of the population departs in any one period. The non-GCC bound was set to the 99th percentile of historic mean departure rates among non-GCC and non-GCC labor countries. This bound constrains departure rates for all flows unrelated to GCC countries so that no more than 16% of the population departs in any single period.

**Gravity Model.** We evaluated the Bayesian hierarchical model by comparing out-of-sample performance to a gravity model of migration (7). After removing flows with zero migrants, we used ordinary least squares to fit the gravity model:

$$\begin{aligned} \log m_{i,j,t} = & \beta_0 + \beta_1 \log P_{i,t} + \beta_2 \log P_{j,t} + \beta_3 \log D_{i,j} \\ & + \beta_4 \log PSR_{i,t} + \beta_5 \log PSR_{j,t} \\ & + \beta_6 \log IMR_{i,t} + \beta_7 \log IMR_{j,t} \\ & + \beta_8 \log urban_{i,t} + \beta_9 \log urban_{j,t} \\ & + \beta_{10} \log LA_i + \beta_{11} \log LA_j \\ & + \beta_{12} LL_i + \beta_{13} LL_j \\ & + \beta_{14} LB_{i,j} + \beta_{15} OL_{i,j} + \beta_{16} COL_{i,j} \\ & + \beta_{17} (t - 2000) + \beta_{18} (t - 2000)^2 + \varepsilon_{i,j,t}. \end{aligned}$$

\*Destination model parameters are time constant; hence, posterior predictive distribution realizations  $\pi_{i,t}^{(s)}$  are denoted  $\pi_i^{(s)}$ .

We used UN estimates of population, distance between capitals, potential support ratio, and infant mortality ratios. All other variables were obtained from the CEPII database (49) or were manually coded. Table 4 gives the definition for each covariate in the gravity model.

**Hurdle Model.** More than half of the historic flows are zero, but the gravity model approach (6, 7) does not account for these migration flows explicitly. To deal with this, we also fitted a gravity-like hurdle model using a two-stage method that removes the need to censor the data.

A hurdle model is a two-stage alternative to a zero-inflated mixture model. Hurdle models explicitly model the generative process, leading to counts equaling zero. If an observation is greater than zero, then the hurdle is crossed and the second stage of the model is fitted to the positive counts (18). The hurdle gravity model with a Poisson count component and a binomial zero component is as follows:

$$\begin{aligned} m_{i,j,t} | m_{i,j,t} > 0 & \stackrel{\text{ind}}{\sim} \text{Poisson}(\lambda_{i,j,t}), \\ \mathbb{1}_{m_{i,j,t} > 0} & \stackrel{\text{ind}}{\sim} \text{Binomial}(1, \omega_{i,j,t}), \end{aligned}$$

with positive count covariate matrix  $X^{[1+]}$  and zero covariate matrix  $X^{[0]}$ ,

$$\begin{aligned} \log \lambda_{i,j,t} & = X_{i,j,t}^{[1+]} \beta, \\ \text{logit } \omega_{i,j,t} & = X_{i,j,t}^{[0]} \gamma. \end{aligned}$$

The mean for the positive component,  $\lambda_{i,j,t}$ , uses the same regressors as the gravity model, yielding estimates for the parameter vector  $\beta = (\beta_0, \beta_1, \dots, \beta_{18})$ .

The mean for the zero component includes the populations  $P_{i,t}$  and  $P_{j,t}$ , the distance between the capitals  $D_{i,j}$ , a shared land border indicator  $LB_{i,j}$ , and a shared official language indicator  $OL_{i,j}$ . Hence, the parameter vector for the zero component is  $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_5)$ . Table 1 shows that the out-of-sample coverage for the truncated Poisson hurdle model was quite good, but that the prediction errors for both the gravity model and the hurdle model were much larger on average than for the other models considered.

**Mean Absolute Percentage of Error.** For a total of  $F$  flows in the flow matrix  $M$  and the flow matrix estimate  $\tilde{M}$ , the mean absolute percentage of error is defined as

$$\text{MAPE}(M, \tilde{M}) = \frac{100}{F} \sum_{i \neq j} \frac{|m_{i,j} - \tilde{m}_{i,j}|}{m_{i,j} + 1}. \quad [4]$$

Normalizing the errors by the magnitude of the observed flow puts the magnitude of errors into context. This means that the errors are measured relative to the size of the underlying flows. We use  $m_{i,j} + 1$  instead of  $m_{i,j}$  in the denominator to avoid degeneracies that arise when flows are equal to zero.

**Table 4. Gravity model covariates**

Parameter	Definition
$P_{i,t}$	Population of country $i$ at the start of period $t$
$D_{i,j}$	Distance between capitals of country $i$ and country $j$
$PSR_{i,t}$	Potential support ratio for country $i$ at the start of period $t$ , i.e., the number of people aged 15 to 64 y per person aged 65+ y
$IMR_{i,t}$	Infant mortality ratio for country $i$ at the start of period $t$
$urban_{i,t}$	Percentage of the population living in urban settings for country $i$ at the start of period $t$
$LA_i$	Land area of country $i$
$LL_i$	Landlocked indicator for country $i$
$LB_{i,j}$	Indicator of shared land border between countries $i$ and $j$
$OL_{i,j}$	Indicator of shared official language in countries $i$ and $j$
$COL_{i,j}$	Indicator of colonial relationship between countries $i$ and $j$
$t$	First year of period
$\varepsilon_{i,j,t}$	Variation not explained by the model

**Data, Materials, and Software Availability.** Data and code have been deposited in <https://github.com/ngwelch/bayesFlow> (50). All other study data are included in the article and/or *SI Appendix*.

**ACKNOWLEDGMENTS.** This work was supported by NIH Grant R01 HD070936. We thank Hana Ševčíková for helpful discussions and technical assistance and the reviewers for helpful comments.

- G. J. Abel, Estimating global migration flow tables using place of birth data. *Demogr. Res.* **28**, 505–546 (2013).
- J. J. Azose, A. E. Raftery, Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 116–122 (2019).
- G. J. Abel, J. E. Cohen, Bilateral international migration flow estimates for 200 countries. *Sci. Data* **6**, 82 (2019).
- F. Willekens, Towards causal forecasting of international migration. *Vienna Yearb. Popul. Res.* **16**, 199–218 (2018).
- J. Bijak, "Forecasting international migration: Selected theories, models, and methods." Warsaw, Poland: Central European Forum for Migration Research, 2006. [http://www.cefmr.pan.pl/docs/cefmr\\_wp\\_2006-04.pdf](http://www.cefmr.pan.pl/docs/cefmr_wp_2006-04.pdf). Accessed 17 October 2019.
- J. E. Cohen, M. Roig, D. C. Reuman, C. GoGwilt, International migration beyond gravity: A statistical model for use in population projections. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15269–15274 (2008).
- K. Kim, J. E. Cohen, Determinants of international migration flows to and from industrialized countries: A panel data approach beyond gravity. *Int. Migr. Rev.* **44**, 899–932 (2010).
- A. Wiśniowski, P. W. Smith, J. Bijak, J. Raymer, J. J. Forster, Bayesian population forecasting: Extending the Lee-Carter method. *Demography* **52**, 1035–1059 (2015).
- J. Raymer, A. Wiśniowski, Applying and testing a forecasting model for age and sex patterns of immigration and emigration. *Popul. Stud. (Camb.)* **72**, 339–355 (2018).
- A. Gelman, J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2006).
- F. Willekens, N. Baydar, "Population structures and models" in *Developments in Spatial Demography*, R. Woods, P. Rees (Eds.), (Allen & Unwin, 1986), pp. 203–244.
- A. Rogers, F. Willekens, J. Little, J. Raymer, Describing migration spatial structure. *Pap. Reg. Sci.* **81**, 29–48 (2002).
- J. Raymer, X. Bai, P. W. Smith, *Developments in Demographic Forecasting*, S. Mazzucco, N. Keilman, Eds. (Springer Nature, 2020).
- United Nations, Department of Economic and Social Affairs, *World Population Prospects: The 2015 Revision, Methodology of the United Nations Population Estimates and Projections* (United Nations, New York, NY, 2015).
- United Nations, Department of Economic and Social Affairs, *World Population Prospects: The 2019 Revision, Methodology of the United Nations Population Estimates and Projections* (United Nations, New York, NY, 2019).
- J. Bijak *et al.*, Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *J. Forecast.* **38**, 470–487 (2019).
- C. Czado, T. Gneiting, L. Held, Predictive model assessment for count data. *Biometrics* **65**, 1254–1261 (2009).
- S. Jackman, *Bayesian Analysis for the Social Sciences* (John Wiley & Sons, 2009), vol. 846.
- D. Bahar, T. Piccone, H. Trinkunas, "Venezuela: A Path out of Misery." Policy Brief Series on the New Geopolitics. [http://www.brookings.edu/wp-content/uploads/2018/10/FP\\_20181023\\_venezuela.pdf](http://www.brookings.edu/wp-content/uploads/2018/10/FP_20181023_venezuela.pdf). Accessed 19 May 2021.
- B. Haddad, The political economy of Syria: Realities and challenges. *Middle East Policy* **18**, 46 (2011).
- M. Hoekstra, S. Orozco-Aleman, Illegal Immigration: The Trump Effect. No. w28909. National Bureau of Economic Research, 2021. [http://www.nber.org/system/files/working\\_papers/w28909/w28909.pdf](http://www.nber.org/system/files/working_papers/w28909/w28909.pdf). Accessed June 2021.
- P. Martell, After independence, what next for South Sudan? *Afr. Renew.* **25**, 3–5 (2011).
- J. J. Azose, H. Ševčíková, A. E. Raftery, Probabilistic population projections with migration uncertainty. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6460–6465 (2016).
- J. Bijak, A. Wiśniowski, *Forecasting International Migration in Europe: A Bayesian View* (The Springer Series on Demographic Methods and Population Analysis, Springer Netherlands, 2010).
- L. Alkema *et al.*, Probabilistic projections of the total fertility rate for all countries. *Demography* **48**, 815–839 (2011).
- A. E. Raftery, J. L. Chunn, P. Gerland, H. Ševčíková, Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50**, 777–801 (2013).
- G. J. Abel, J. DeWaard, J. T. Ha, Z. W. Almqvist, The form and evolution of international migration networks, 1990–2015. *Popul. Space Place* **27**, e2432 (2021).
- J. J. Azose, A. E. Raftery, Bayesian probabilistic projection of international migration. *Demography* **52**, 1627–1650 (2015).
- J. Aitchison, *The Statistical Analysis of Compositional Data* (Chapman and Hall, Ltd., 1986).
- J. Raymer, F. Willekens, *International Migration in Europe: Data, Models and Estimates* (Wiley, 2008).
- A. Fontanet *et al.*, SARS-CoV-2 variants and ending the COVID-19 pandemic. *Lancet* **397**, 952–954 (2021).
- European Commission – Press release, Migration Statistics Update: The Impact of COVID-19 (2021). [http://www.ec.europa.eu/commission/presscorner/detail/en/ip\\_21\\_232](http://www.ec.europa.eu/commission/presscorner/detail/en/ip_21_232). Accessed 4 May 2021.
- M. Hackman, A. A. Caldwell, Surge of Migrants at U.S. Southern Border: Biden's Plan and What You Need to Know (2021). <http://www.wsj.com/articles/surge-of-migrants-at-u-s-southern-border-bidens-plan-and-what-you-need-to-know-11616788003>. Accessed 4 May 2021.
- E. Zagheni, I. Weber, "You are where you E-mail: Using E-mail data to estimate international migration rates" in *Proceedings of the 4th Annual ACM Web Science Conference, WebSci '12* (Association for Computing Machinery, New York, NY, 2012), pp. 348–351.
- E. Zagheni, V. R. K. Garimella, I. Weber, B. State, "Inferring international and internal migration patterns from Twitter data" in *Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion* (Association for Computing Machinery, New York, NY, 2014), pp. 439–444.
- E. Zagheni, I. Weber, K. Gummadi, Leveraging Facebook's advertising platform to monitor stocks of migrants. *Popul. Dev. Rev.* **43**, 721–734 (2017).
- N. Cesare, H. Lee, T. McCormick, E. Spiro, E. Zagheni, Promises and pitfalls of using digital traces for demographic research. *Demography* **55**, 1979–1999 (2018).
- L. Fiorio *et al.*, Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography* **58**, 51–74 (2021).
- F. Billari, E. Zagheni, "Big data and population processes: A revolution?" in *Proceedings of the Conference of the Italian Statistical Society, A. Petrucci, R. Verde*, Eds. (Firenze University Press, Firenze, Italy, 2017), pp. 167–178.
- D. Lazer, J. Radford, Data ex machina: Introduction to big data. *Annu. Rev. Sociol.* **43**, 19–39 (2017).
- M. Alexander, K. Polimis, E. Zagheni, Combining social media and survey data to nowcast migrant stocks in the United States. *Popul. Res. Policy Rev.* **41**, 1–28 (2022).
- D. Ruths, J. Pfeffer, Social sciences. Social media for large studies of behavior. *Science* **346**, 1063–1064 (2014).
- D. Lazer, R. Kennedy, G. King, A. Vespignani, Big data. The parable of Google Flu: Traps in big data analysis. *Science* **343**, 1203–1205 (2014).
- P. de Valpine *et al.*, Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Stat.* **26**, 403–413 (2017).
- P. de Valpine *et al.*, NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling. R package version 0.12.1 (2021). <https://cran.r-project.org/web/packages/nimble/index.html>. Accessed 24 October 2021.
- R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2021).
- A. Rogers, L. J. Castro, "Model migration schedules" (Tech. Rep. RR-81-030, International Institute for Applied System Analysis, Laxenburg, Austria, 1981).
- H. Ševčíková, A. E. Raftery, bayesPop: Probabilistic population projections. *J. Stat. Softw.* **75**, 1–29 (2016).
- T. Mayer, S. Zignago, Notes on CEPII's Distance Measures: The GeoDist database, (Centre d'Études Prospectives et d'Informations Internationales, Working Papers 2011-25 (2011). Retrieved 15 July 2018, from [http://www.cepii.fr/PDF\\_PUB/wp/2011/wp2011-25.pdf](http://www.cepii.fr/PDF_PUB/wp/2011/wp2011-25.pdf).
- N. G. Welch *et al.*, Data from "Probabilistic forecasts of international bilateral migration flows." GitHub. <https://github.com/ngwelch/bayesFlow>. Deposited 14 July 2022.