*Journal of Forecasting*

WILEY

RESEARCH ARTICLE OPEN ACCESS

# Temporal Patterns in Migration Flows Evidence from South Sudan

Thomas Schincariol | Thomas Chadefaux

Department of Political Science, Trinity College Dublin, Dublin, Ireland

**Correspondence:** Thomas Schincariol (schincat@tcd.ie)

**Received:** 14 May 2024 | **Revised:** 28 August 2024 | **Accepted:** 23 October 2024

**Funding:** This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 101002240).

**Keywords:** forecasting | migration dynamics | temporal patterns

## ABSTRACT

What explains the variation in migration flows over time and space? Existing work has contributed to a rich understanding of the factors that affect why and when people leave. What is less understood are the dynamics of migration flows over time. Existing work typically focuses on static variables at the country-year level and ignores the temporal dynamics. Are there recurring temporal patterns in migration flows? And can we use these patterns to improve our forecasts of the number of migrants? Here, we introduce new methods to uncover temporal sequences—motifs—in the number of migrants over time and use these motifs for forecasting. By developing a multivariable shape similarity-based model, we show that temporal patterns do exist. Moreover, using these patterns results in better out-of-sample forecasts than a benchmark of statistical and neural networks models. We apply the new method to the case of South Sudan.

## 1 | Introduction

What explains the variation in migration flows over time and space? Existing work has contributed to a rich understanding of the factors that determine what pushes people away from their home: war, poverty, or oppression—push factors, and what attracts them elsewhere: economic opportunities or freedom—pull factors. What is less understood is the dynamics of migration flows over time. Existing work typically focuses on static variables at the country-year level. The analysis may involve more fine-grained spatial units, but rarely is the temporal dynamic investigated.

In contrast to this predominantly static approach, our research aims to improve the accuracy of migration flow predictions by identifying dynamic temporal patterns in migration flows. This study investigates whether the sequence of events, rather than just individual events, can provide additional insights into migration patterns. By identifying these temporal patterns, we aim to enhance the accuracy of forecasts regarding migrant numbers and address the shortcomings of current predictive models.

Our novel methodology combines dynamic time warping (DTW) with pattern recognition algorithms. This approach enables us to discern not just the presence of migration motifs but also their specific characteristics and temporal variations. By applying these methods to fine-grained migration data, we aim to more accurately forecast future migration trends and understand the underlying dynamics driving these flows.

Additionally, our approach benefits from relying on a limited number of covariates rather than employing complex machine learning models with extensive features. This strategy offers significant advantages: it reduces the risk of overfitting, simplifies interpretation, and requires less computational power and data. Complex machine learning models, while powerful, often need large datasets to perform effectively and face issues related to reproducibility and interpretability due to their nondeterministic

nature. In contrast, our streamlined approach provides a practical and robust solution, making it more accessible and applicable to various contexts where data may be limited.

To validate our motifs, we take advantage of a unique dataset that is temporally fine-grained. We use data from the International Organization for Migration's Displacement Tracking Matrix (IOM-DTM), focusing specifically on forced displacement. This includes internal displacement within South Sudan and, where applicable, cross-border refugee movements. The DTM provides fine-grained temporal and spatial data on out-flows from locations within South Sudan, allowing us to analyze short-term fluctuations and localized patterns in forced displacement.

We show that our method, focusing on temporal shapes—distinctive patterns in migration data over time—yields better results than traditional benchmarks set by models like regression models or neural networks. We apply this approach to the case of South Sudan, a context marked by a complex interplay of food prices and conflict fatalities. Our findings offer more accurate predictive models and provide insights that could inform migration policy and aid in understanding similar migration dynamics in other regions.

Our findings are important for a number of reasons. First, migration studies suffer from a lack of fine-grained data. Data are usually coarse and unavailable at the subnational level. However, we show here that even with a small temporal coverage, we can extract important temporal patterns. These results also have important implications for policymakers and actors in the field such as NGOs. Predicting migrant flows matters, and although those on the ground may develop over time an intuition for the patterns we uncover here, the approach here has the advantage of extracting patterns that may have been missed in practice.

## 2 | Patterns in Migration Flows

### 2.1 | The Covariates of Migration

The question of why and when people leave their homes has received a substantial amount of attention in the literature. The individual decisions leading each person or family to leave are complex. Whether a person flees may depend on their means, their tolerance for risk (Engel and Ibáñez 2007), or their ability to secure a better alternative elsewhere. These and other factors may in turn vary by age (Kassar and Dourgnon 2014), education (Van Dalen, Groenewold, and Schoorl 2005), gender (Lauby and Stark 1988), or personality (Arcand and Mbaye 2013).

Beyond individual motives, a number of factors correlate with migration. Most can be classified either as push or pull factors. Conflict (Davenport, Moore, and Poe 2003), human rights violations (Moore and Shellman 2004), corruption (Dimant, Krieger, and Meierrieks 2013), or climate change (Dasgupta et al. 2016), for example, tend to "push" people away from their home. On the other hand, pull factors may attract migrants to other countries: easily accessible neighbors (Turkoglu and Chadefaux 2019), well-functioning institutions (Ariu, Docquier, and Squicciarini 2016), or income opportunities (Ortega and Peri 2009).

While much of the literature focuses on push and pull factors, recent research has begun to explore patterns within migration flows themselves. Some studies at the country-year level have provided initial insights into temporal variations in migration (Melander and Öberg 2006; Fearon and Shaver 2020). However, these analyses often miss finer grained temporal dynamics.

In short, we know that migration—whether forced or not—is affected by a multitude of factors and their interactions (De Haas 2011). What is less well understood are the dynamic processes that lead to migration. Indeed, it is likely that it is not solely the raw value of these variables that matters but also their evolution over time. In other words, it may not be enough to say that the level of a particular variable matters; we also need to understand how its variation over time may affect the dynamic of migration itself.

### 2.2 | Why Patterns?

Why would migration flows follow regular temporal patterns? While it is well understood that migration is not random but is driven by various covariates, the idea that there may be recurring temporal patterns in refugee flows has, to our knowledge, not been explored. Of course, existing models do incorporate some elements of temporal dependence, such as lags or first differences. However, these models cannot account (i) for the variety of possible patterns; (ii) for their possible variations over time—the coefficients that are true today may no longer be tomorrow; and (iii) for the fact that these patterns may be warped and stretched versions of each other, such that a one-to-one matching approach will not be able to uncover them.

First, patterns may emerge out of the complex interplay of variables. Patterns emerge spontaneously in natural (Malchow, Petrovskii, and Venturino 2007), animal (Baurmann, Gross, and Feudel 2007), and social phenomena (Chadefaux 2021). Indeed, a highly simplistic model alone is enough to generate temporal patterns. Suppose, for example, that the number of migrants and the number of violent armed attacks are linked by a system of two differential equations.

$$\frac{dp}{dt} = \alpha p - \beta p v, \tag{1}$$

$$\frac{dv}{dt} = \delta p v - \gamma v, \tag{2}$$

where $p$ denotes the population left in the area (and hence the number of migrants) and $v$ denotes the number of violent events in the same area. $\frac{dv}{dt}$ and $\frac{dx}{dt}$ therefore represent the growth rates of migrants and violent events at a particular time $t$. The underlying idea is that violent events are more likely to take place when the population remains in the area—when there are people to victimize; but people are more likely to flee when there is more violence. This is clearly not a realistic model of migration and has no intention to represent the underlying dynamic or causes of migration. However, it illustrates that even simple rules of interaction can lead to recurring nonlinear temporal patterns (Figure 1).

Second, migration may be driven by variables that themselves follow repeating temporal patterns. Migration and

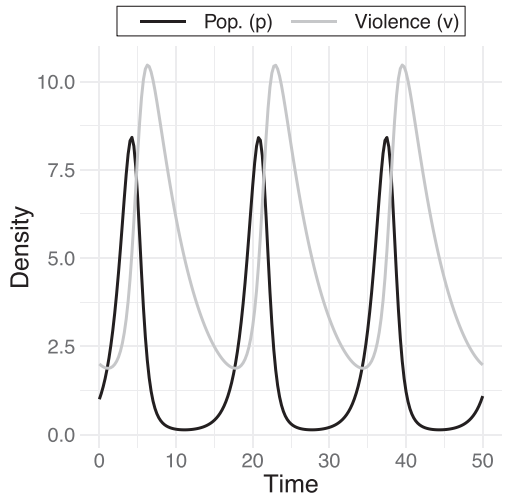



**FIGURE 1** | Temporal patterns can emerge from simple differential equations.

refugee flows, for example, are frequently associated with conflict (Schon 2019), which itself can follow repeating temporal patterns (Chadefaux 2021). Climate-related factors have also been linked to migration (Kaczan and Orgill-Meyer 2020), and some of these patterns are obviously at least seasonal. Such seasonality is easily captured by an ARIMA model. However, others may be more a complex and harder to define a priori. It may be the case, for example, that a recurrent pattern of droughts over months or years drive people away. Such complex and nonlinear patterns will be difficult or impossible to uncover for regular approaches relying on lags or differencing.

Another cause for the emergence of migratory patterns is that people are likely to react to real or observed patterns in their environment. Migrants may, for example, respond to ongoing violence in the country. While the level of violence itself matters, it is likely that they will also base their decision to move on the expected evolution of future trends. These expectations will be based on past trends or mental models. If so, migration itself will mimic the underlying patterns of these variables.

Indeed, humans excel at detecting patterns in their environment—conflict, climate, technology, or their networks. A long literature in psychology and neuroscience shows the ability of various organisms to form complex temporal representations of events (Gallistel 1990).[1] People construct "mental models of temporal sequences of events" (Schaeken, Johnson-Laird, and D'Ydewalle 1996). "Humans build discrete mental simulations of possibilities—mental models—when they reason." "Reasoners do not represent all of the time points across which an event might endure. Instead, they construct discrete tokens that stand in place of the beginnings and endings of those events" (Kelly and Khemlani 2019). In other words, humans are able to adjust their mental models to events of different duration, as long as these events form a similar pattern delineated by key inflection points. For example, while a single episode of drought may not prompt widespread migration, a series of such events significantly increases the likelihood of people relocating. This pattern suggests that not only the occurrence of adverse events but their frequency and perceived future trend influence migration decisions. Similarly, people will observe the unfolding of conflict events in their regions over time. As a result, their behavior itself is likely to reflect these patterns—or a combination of them. Looking at aggregate number of events or deaths in a given month may therefore not be sufficient to capture the potentially complex patterns of violence that lead people to flee.

**TABLE 1** | Two temporal sequences with varying lengths and speed.

| Time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *a* | 5 | 3 | 1 | 3 | 5 | 3 | | | | | |
| *b* | 5 | 4 | 3 | 2 | 1 | 2 | 3 | 4 | 5 | 4 | 3 |

Finally, patterns may arise due to factors unique to the migration context. A crucial element of these dynamics is the influence of migration networks and remittances in shaping migration trends over time. Migration networks, established by earlier migrants, can facilitate subsequent migration by reducing costs and risks for new migrants (Massey et al. 1993). These networks can lead to cumulative causation, where initial migration creates social structures that increase the likelihood of further migration (Massey, Goldring, and Durand 1994). Similarly, remittances sent by migrants to their home communities can transform the economic landscape, potentially encouraging or discouraging further migration depending on their usage (Taylor 1999). These factors can create self-reinforcing patterns in migration flows that evolve over time, underscoring the importance of examining not only the drivers of migration but also the temporal patterns within the flows themselves.

### 2.3 | A Methodological Problem

There are good reasons, however, for why temporal dynamics have received limited attention. This is due to two limitations of the tools we have worked with so far. First, data are typically limited to national and yearly levels, which does not offer the level of granularity needed to understand the dynamics of migration over time, but rather only lets us compare levels over long periods of time. Some work has started to look at more granular levels (Mayen, Wood, and Frazier 2022), but the data remains scarce.

A second limitation is the difficulty of existing methods in the social sciences to account for temporal dynamics. The methods we use—typically regression—are ill-suited to incorporate the possibly complex ebbs and flows of a complex process such as migration. Existing methods rely on correlation, which compares the raw values of the covariates, and does not allow for possible warping in time and space. Correlation (and therefore regression) relies on the comparison of more or less static covariates, and only allows limited comparisons of time sequences. As a simple example, consider the two sequences presented in Table 1.

Both are clearly similar in that they follow a "down, up, down" pattern. Yet, *a* unfolds over 6 time periods, but *b* over 11. Calculating the correlation between the two series is impossible due to their different lengths. Even if we were to limit our attention to only six observations of *b*—as a regression would—the correlation between the two is nearly zero. Existing methods that rely on correlation patterns, such as ARIMA or distributed

lag models can therefore not adequately account for these patterns. While ARIMA models, for example, can model complex and even nonlinear sequences, they are poor at recognizing similar sequences that may unfold at different speeds. The same is true of standard regression models.

To be sure, there are alternatives to statistical approaches that can model time sequences effectively. Systems dynamics models and agent-based models (ABMs) offer detailed simulations of complex systems, capturing interactions and feedback loops. However, they require significant development time and extensive domain knowledge, making them resource intensive. Deep learning models for time series can also uncover complex patterns in data but often require large datasets to perform effectively and face issues related to reproducibility and interpretability due to their nondeterministic nature. Additionally, machine learning models can be finicky, often needing careful tuning and large amounts of data to yield reliable results. Furthermore, relying on a limited number of covariates, as we have done, offers significant advantages over complex machine learning models with extensive features. This approach reduces the risk of overfitting, simplifies interpretation, and requires less computational power and data, making it more practical for many applications.

This inability to account for variations in length and speed is unfortunate, because events may unfold at different paces. Episodes of violence and their associated trajectories and shape may take place over days in some place, months in others. Ethnic cleansing may unfold over weeks, as was the case of Rwanda, or years, as in Darfur. The resulting patterns of migration may as a consequence be either fast or slow.

However, new data and methods now allow us to take into account not only static factors, such as the number of conflict events or food prices, but also the dynamics—the patterns—of these factors. We take advantage of methods derived mainly from speech recognition to account for the possibility of patterns at different temporal levels.

As an illustration of our approach (detailed below), consider the sequences presented in Figure 2. They show the number of migrants (left, blue), the food price (center, green) from the Melut region in South Sudan in 2022 (top) and Guit region in 2021 (bottom). The dynamics of monthly migration (blue) in Melut are visually similar to those in Guit. The same applies to food prices (green). Both situations lead to similar migration outcomes (red). Using information from Guit could help forecast events in Melut. Below, we show how to identify similarities across multiple dimensions and find close matches that improve future predictions.

Our approach offers several significant advantages over using complex machine learning models with extensive features. First, by focusing on a limited number of covariates, we significantly reduce the risk of overfitting. Overfitting occurs when a model is too closely aligned with the training data, capturing noise rather than the underlying pattern. Models with extensive features are particularly prone to this issue because they can fit the training data very well but fail to generalize to new, unseen data. In contrast, our approach with fewer covariates helps ensure that the model captures the essential patterns that are likely to be present in future data, improving its robustness and predictive power.

Second, a simplified model with fewer covariates enhances interpretability. In migration studies, it is crucial for policymakers and practitioners to understand the factors driving migration flows. Complex machine learning models, while powerful, often operate as "black boxes," making it difficult to extract meaningful insights about the relationship between covariates and migration patterns. Our approach, by limiting the number of covariates, makes it easier to interpret the model's outputs and understand the impact of each factor on migration flows. This transparency is valuable for developing targeted interventions and informing policy decisions.

Third, models with a limited number of covariates require less computational power and data, making them more practical for many applications. Complex machine learning models with extensive features often demand substantial computational resources


**FIGURE 2** | Human migration (blue) and food price (in green) for Melut region in South Sudan (top) and Guit (bottom). The red section represents the migration figures for the 3 months following the depicted period (blue) for both regions.

and large datasets to perform effectively. In many real-world scenarios, especially in regions with limited data availability, such requirements can be prohibitive. Our approach is more efficient and accessible, allowing for effective migration forecasting even with smaller datasets and limited computational infrastructure.

## 3 | Data

Migration data are often limited to the country-year level. Most datasets list the flow of migrants from one country to the next in any given year. While this level of analysis is adequate for a static analysis of the factors leading to migration, they are not sufficient to understand the temporal dynamics at play. Yearly, national-level data fail to capture the subtleties of migration trends that occur on a finer temporal scale.

To address this gap, we take advantage of a unique dataset from the IOM-DTM flow monitoring surveys. The dataset offers a detailed view of migration at the subnational level, providing monthly data on the origins of migrants within South Sudan from January 2020 to October 2022. The IOM-DTM captures displacement, including both internal displacement within South Sudan and cross-border refugee movements. These data encompass individuals fleeing due to conflict, violence, human rights violations, economic, healthcare, and natural or human-made disasters.[2]

The IOM-DTM conducts surveys at 31 critical transit points across South Sudan and its borders, providing a level of temporal detail that annual data cannot achieve. The focus of our analysis is the monthly count of migrants leaving each of the 76 provinces (Administrative Region 2) in South Sudan.[3] For each of these regions, we obtain a time series of 34 data points, each representing the number of migrants leaving each region monthly from January 2020 to October 2022.

Additionally, we incorporate two key exogenous variables into our analysis: food prices and conflict data. The inclusion of these variables is motivated by their significance as major push factors in migration, with both datasets offering the necessary granularity for our analysis at a monthly and regional level. Food price data are derived from the World Bank's dataset, focusing on the monthly prices of beans in 29 markets across South Sudan (Andrée 2021). These prices are then associated with the nearest regions to each market, based on the assumption that populations primarily access their nearest markets. Food prices can serve as a valuable proxy for assessing both climatic and economic conditions. Extreme events such as droughts or floods have a direct and significant impact on food prices. Therefore, fluctuations in food prices give insights on the environmental and economic dynamics within a region. The conflict data are sourced from the Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) (Sundberg and Melander 2013; Davies, Pettersson, and Öberg 2022), which provides us with monthly data on conflict events for each region in the country over the entire period.

## 4 | Methods

Our approach to uncover and use temporal patterns in migration data is twofold: First, it involves the identification of recurring temporal motifs or patterns in migration data—"motifs." These motifs are not predefined but are instead extracted directly from the data, allowing for the discovery of patterns that traditional static models may overlook. Second, we use these extracted motifs to forecast migration at the regional level.

This methodological innovation hinges on the application of DTW, a technique that allows for the flexible comparison of time series data, even when these series unfold at different speeds. This flexibility is crucial for accurately matching temporal patterns across different contexts and time frames. The following sections detail the specific steps of this approach, including the extraction of motifs from the data, the application of DTW for pattern matching, and the integration of these elements into a predictive model for migration flows.

### 4.1 | Searching for Patterns

Identifying patterns in migration data is challenging due to their inherent complexity and variability. There are generally two approaches to uncover these patterns. The first would involve predefining patterns based on existing theories. For example, an "up–down–up" pattern in migration could be hypothesized to reflect specific sequences of events leading to increased migrant flow. However, this approach has limitations. First, existing theories do not discuss temporal patterns, but rather focus on static covariates like push and pull factors. As a result, we do not have a good theoretical starting point to hypothesize the existence of particular shapes. Furthermore, relying solely on predefined patterns risks missing out on discovering novel, potentially more significant patterns that existing theories do not anticipate.

Our approach is more flexible. Instead of testing predefined patterns, we extract relevant patterns directly from the data. This method involves analyzing historical data to find matching sequences and using these patterns to forecast future events. This approach has two advantages. First, it significantly reduces the risk of overfitting, ensuring the patterns we identify reflect genuine trends rather than random noise. Second, by validating our model's predictions against actual future events, we can differentiate between true underlying phenomena and mere coincidences in the dataset. Our approach to uncovering migration patterns thus involves two main steps: first, measuring the similarity between time series and then using these similarity scores to create forecasts. We detail these steps below.

### 4.2 | Matching Subsequences

Traditional correlation analysis can fail to recognize similarities between time series. For example, time series $a$ and $b$ in Figure A1 look alike, yet their correlation cannot be calculated over the entire range, and is near zero regardless of how we truncate $b$. We therefore rely on *DTW*, a method that allows us to match time series with more flexibility on the time dimension. DTW is a method for measuring the similarity between two time series by examining how well the overall patterns of the time series match, even if they do not align perfectly in time. This is different from traditional methods that only focus on the differences in values at each specific time point.

The essence of DTW lies in its ability to adjust time indices between two series. This involves strategically modifying the time dimension of the series through stretching or compressing, and possibly duplicating or shifting data points, to achieve alignment of similar patterns. Such adjustments allow for a more natural comparison beyond mere point-to-point differences, focusing on the overarching shape or pattern similarity between the series.[4] DTW allows us to find similarities between time series, even when these series unfold at different speeds. The advantage is that we can uncover time series with similar shapes, in contrast to existing approaches, which focus on the raw values of the covariates.

We apply DTW to find resemblances between time series representing migration and covariates (food prices and conflict) even when they unfold at different rates. By focusing on shape similarities, we can better anticipate future trends in migration. The distance between two region-months (our unit of analysis) $i$ and $j$ is defined as the sum of the DTW distances across these variables:

$$d_{i,j} = \sum_{v} \mathrm{DTW}(seq_{i,v}, seq_{j,v}), \tag{3}$$

where $v$ represents each variable of interest—food prices, conflict fatalities, and migration.[5] Based on the distance $d_{i,j}$, the observation is classified as a match if the distance is smaller than a threshold $\theta$.[6] The overall process is described in Figure 3.

### 4.3 | Forecasts

To forecast the number of migrants originating from a particular region in a given month, we follow three simple steps. First, the most recent time series of region $i$ is extracted. Then, time series with a suitably small distance $d_{i,j}$ are extracted in the historical data. To identify these matches, we consider all regions of South Sudan in the past. Finally, the subsequent $h \in \{1,2,3\}$ values of these matches—what we call past futures—are collected (squared sequences in Figure 3). The predicted values are then the mean values of these past futures. Thus, the prediction for region $i$ and time $t+h$ is the mean of matching units $j$ at time $s+h$, where $s$ is the time at which matching unit $j$ was observed.[7] Note that forecasts are made for all $h \in \{1,2,3\}$ simultaneously at time $t$. As a result, the quality of predictions is expected to decrease as $h$ increases.

### 4.4 | Evaluation

We evaluate our forecasts by comparing them to two benchmarks. The first, autoregressive integrated moving average (ARIMA), is a widely used model for forecasting time series. The model is defined by the following equation:

$$\Delta^d Y_t = \mu + \phi_1 \Delta^d Y_{t-1} + \ \dots \ + \phi_p \Delta^d Y_{t-p} + \theta_1 \epsilon_{t-1} + \ \dots \ + \theta_q \epsilon_{t-q}, \tag{4}$$

where $\Delta^d Y_t$ represents the $d$th-order differenced series, $\mu$ is the intercept $\phi_1, ..., \phi_p$ are the autoregressive coefficients, $\theta_1, ..., \theta_q$ are the moving average coefficients.

The ARIMA model can handle a variety of time series patterns, including trends and seasonality, by incorporating differencing, which makes it robust in dealing with nonstationary data. The model's capability to integrate autoregressive and moving average components allows it to capture both short-term and long-term dependencies in the data.

Our second benchmark, long short-term memory (LSTM), is a derivative of recurrent neural networks (RNNs), with the added advantage of being able to handle long-term dependencies in sequential data. This makes them ideal for complex time series forecasting tasks where context and history significantly influence future predictions (see Appendix B for details).

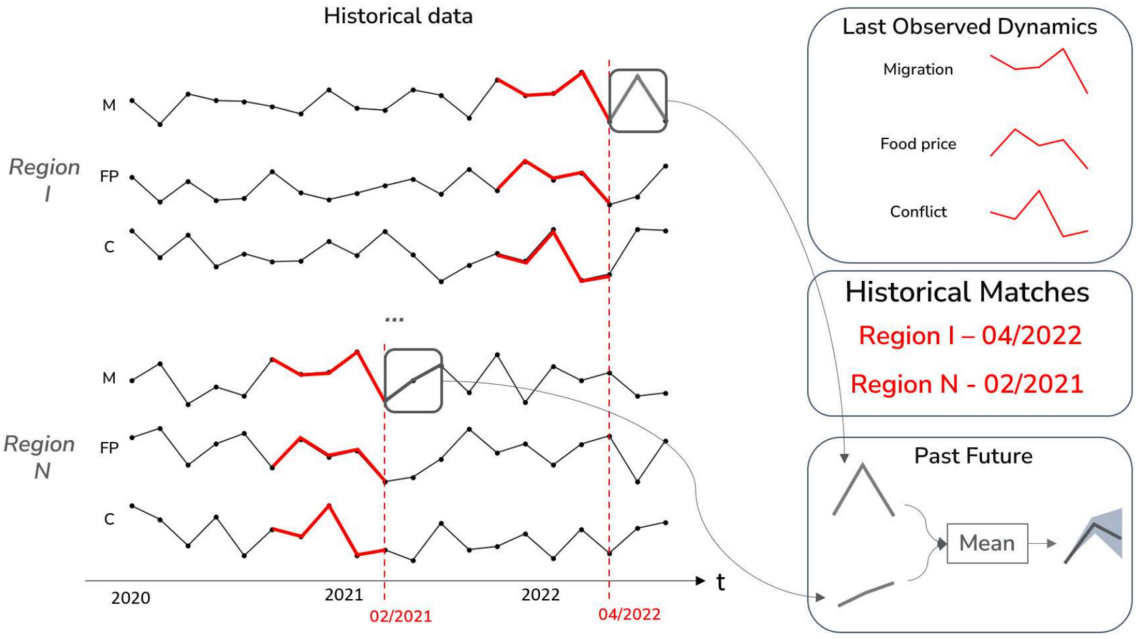



**FIGURE 3** | Illustrative representation of the shape similarity model's match-finding and forecasting methodology. Displayed in the top-right quadrant are the most recent shape observations for each variable within the specified region. A rolling window approach is applied to each region to identify instances of similarity. Upon recognizing a similar case (e.g., April 2022 in Region I), the migration values of the ensuing month are recorded and categorized as "Past Future." The predicted value is then computed as the average of these "Past Future" sequences.

For each of the three models, the division of data into training and test sets follows a 90/10 ratio. This allocation translates to a test set duration of 3 months. The decision to opt for this distribution is based on the limited amount of data we have at our disposal, yet still leaves us with 228 observations (three observations for each of the 76 regions)—enough to derive statistically significant conclusions.

## 5 | Results

### 5.1 | Patterns

Figure 4 displays sample matches from the data, illustrating specific recurring instances of patterns. These observed patterns are likely the result of the complex interplay between various factors influencing migration, such as economic conditions, conflict, climate patterns, and human behavior as individuals respond to their changing environment.

However, sample repetitions are not enough to reach any conclusions, as these repetitions may occur by chance. After all, even a suitably long white noise series would include repeating patterns. Consequently, a pivotal question is whether these patterns are meaningful. Investigating the outcomes that follow such patterns can provide insights. If patterns do carry information, then we would expect similar patterns to lead to similar outcomes—something we would not expect from white noise. Suppose, for example, that pattern *A* is followed by a wide range of outcomes—many migrants sometimes, few others—whereas pattern *B* is consistently followed by a large number of migrants. Clearly, pattern *B* in this case is more informative.

To put this concept into practice, we gathered the standard deviation of the events following each identified pattern. A substantial standard deviation would imply that the observed patterns are predominantly stochastic and, hence, offer negligible predictive utility. In contrast, a minimal standard deviation signals that the outcomes subsequent to the patterns exhibit a high degree of uniformity. In Figure 5, the *y* axis denotes the standard deviation of the "postpatterns"—the sequence of three observations that immediately follows the pattern—while the *x* axis sorts patterns by their proximity. A clear inverse relationship is observed: Similar patterns are followed by similar outcomes, whereas dissimilar patterns exhibit larger standard deviations. This trend suggests that closer patterns lead to more consistent and predictable migration flows, suggesting that temporal motifs in migration data do carry important information.

Figure C1 illustrates instances where patterns of human migration, food prices, and conflict fatalities are combined, along with the visual representation of their corresponding following migration values. The combination of patterns associated with the highest standard deviation (top) and the lowest SD (bottom) of the following values are displayed.

### 5.2 | Forecasts

We now validate the idea that these patterns carry valuable information by using them to forecast future migration values. To compare the performance of our proposed model against benchmarks, we rely on the log ratio of mean squared error (MSE) as our primary metric. The choice of MSE is motivated by its widespread use and ease of interpretation, making it an ideal


**FIGURE 4** | Illustrative matches of temporal migration patterns (red), food prices (green), and conflict (yellow) from various regions and time periods in South Sudan, identified through dynamic time warping analysis. Each row showcases the data from a distinct region, with time plotted on the *x* axis and the number of migrants on the *y* axis. The columns represent these three variables for different regions or time periods. The intensity of the grey in the background series indicates the degree of similarity to the colored reference series (darker = more similar).

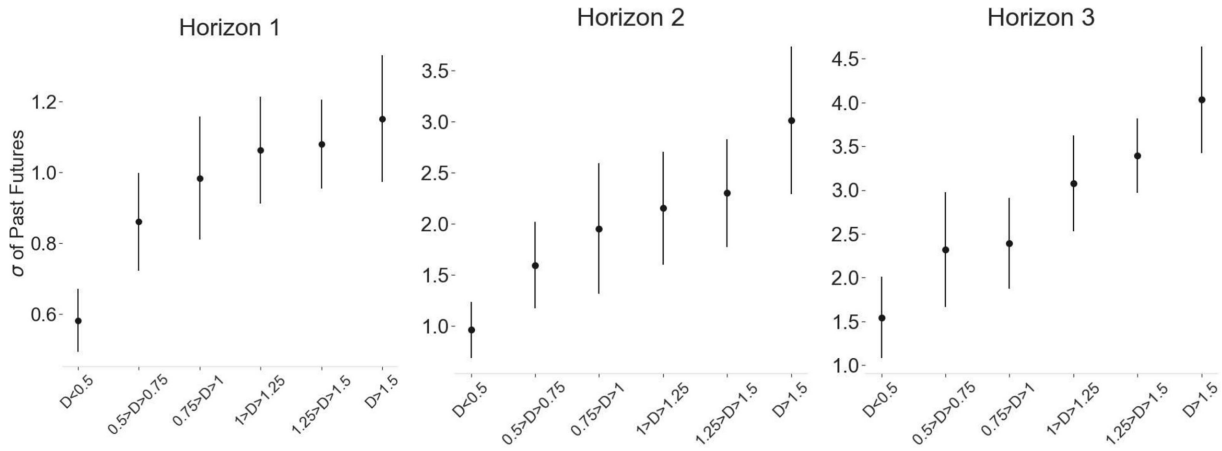



**FIGURE 5** | Pattern dissimilarity ($D$) and postpattern migration consistency. The graph illustrates the inverse relationship between the degree of similarity in migration patterns ($x$ axis) and the standard deviation ($\sigma$) of subsequent migration flows ($y$ axis). Lower standard deviations following postpatterns (the three observations that immediately a sequence) indicate more predictable migration outcomes, underscoring the potential of using pattern similarity as a forecasting tool. The data suggest that as the proximity of patterns increases, the predictability of migration trends also increases, indicating that patterns carry important information.

standard for assessing and contrasting the efficacy of the different models in our study.

In Figure 6, we present the log ratio of the MSEs from our predictions.[8] Each point on the plot represents the average squared errors across all time periods for a particular region.[9] A positive ratio indicates that the benchmark has larger errors than our model—that is, its performance is worse. We find that the ratios for both benchmarks are significantly above zero (1.92 and 0.89 for LSTM and ARIMA, respectively, with $t$ tests significant at the $p < 0.001$ level). These results hold for a variety of metrics (Table 2). These results hold for all $h \in \{1,2,3\}$ but not beyond. This is to be expected, as outperforming a model that more or less predicts the immediate past (ARIMA) becomes exceedingly difficult without incorporating additional information.

In Figure 7, the geographical distribution of results is illustrated. Each region is color coded based on the best model for that region. This gives us a sense not only of how much the MSE improves, but how often. Out of the three models, our model, labeled as "ShapeFinder" (dark blue), has the lowest MSE in 70% of the regions. ARIMA, indicated by the light blue color, and LSTM, in purple, are best in significantly fewer regions—15 and 8, respectively. Additionally, in the three regions with the highest count of migration values (Morobo, Juba, and Rubkona), our model has the best performance.

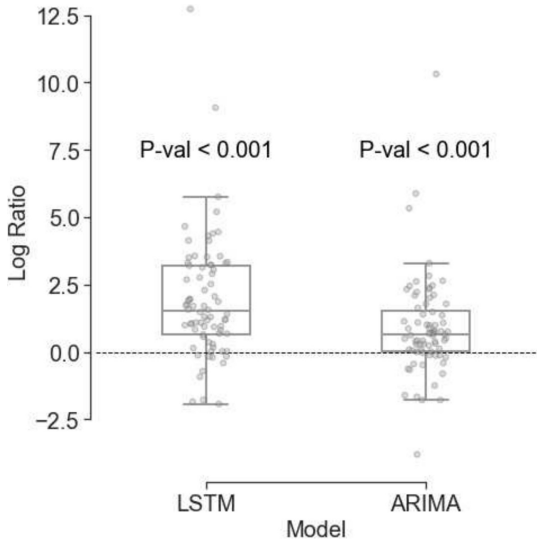



**FIGURE 6** | Log ratio of the mean squared error (MSE) for benchmark models (LSTM and ARIMA) compared to the shape similarity model. Each point represents a region of South Sudan. Points above the zero line indicate regions where the benchmark model has a higher (i.e., worse) MSE than the shape similarity model.

## 5.3 | Covariates

Furthermore, regional variations in food prices and conflict fatalities provide further insights. Figure 8b highlights the North-East regions with higher food prices. In these regions, which exhibit the highest mean value in food prices, ARIMA or LSTM models perform best in 80% of the cases. This is further supported by the regression results shown in Figure 8a. The negative regression coefficient implies that as food prices increase, model performance deteriorates. This observation is consistent with the findings of Smith and Wesselbaum (2022), where a linear relationship between food insecurity and migration flows was identified. Given this linearity, simpler models like ARIMA perform well, suggesting that our more complex model may not be necessary.

Conversely, Figure 9b shows regions with the highest sum of fatalities due to conflict. In these regions, 80% of the highest fatalities are best modeled using the ShapeFinder model. In Figure 9a, the quadratic regression between conflict fatalities and the MSE log ratio is displayed for the two models, indicating that our model performs better in conflict-affected regions, but its effectiveness decreases as the number of fatalities rises. These findings underscore the importance of considering regional characteristics when selecting and evaluating models. The consistent performance of different models in varying contexts highlights the robustness of the proposed modeling framework, despite changes in the weight configurations.

**TABLE 2** | Mean and 95% confidence intervals for various prediction metrics across different models.

| Metric | LSTM | ARIMA | ShapeFinder |
|---|---|---|---|
| RMSE | $260 \pm 159$ | $149 \pm 107$ | $120 \pm 86$ |
| Med. Abs. Err. | $238 \pm 148$ | $121 \pm 88$ | $81 \pm 58$ |
| MAPE | $4.5 \times 10^{16} \pm 5.9 \times 10^{16}$ | $9.8 \times 10^{15} \pm 1.1 \times 10^{16}$ | $1.8 \times 10^{15} \pm 1.3 \times 10^{15}$ |
| MAE | $248 \pm 155$ | $134 \pm 100$ | $103 \pm 75$ |

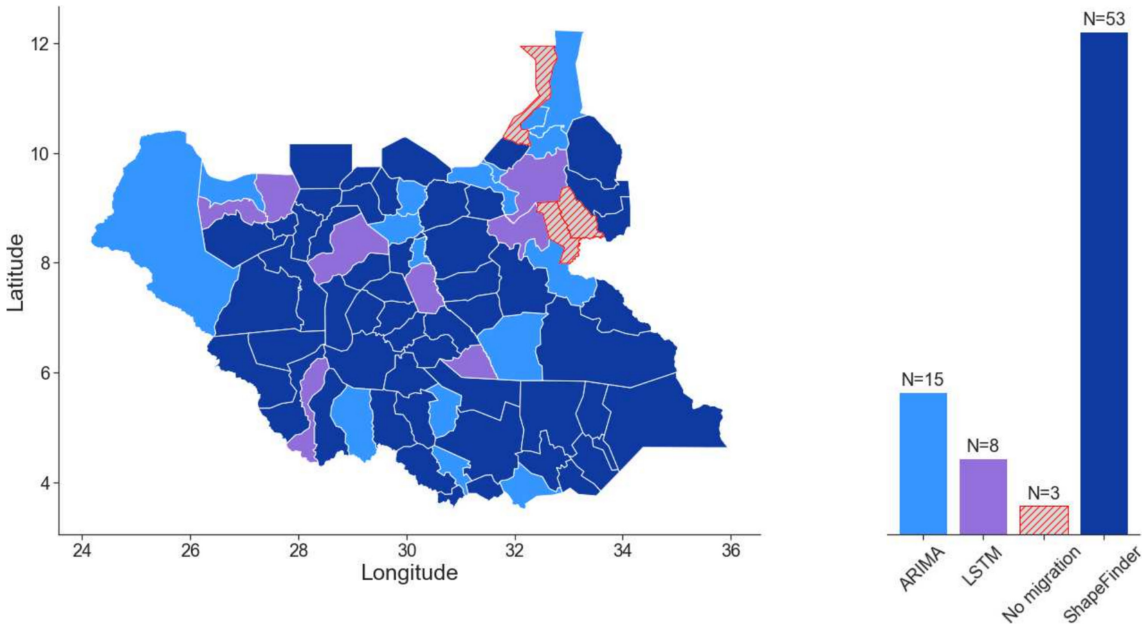



**FIGURE 7** | Map of the best-performing model for each region in South Sudan. Different colors indicate the model with the lowest mean squared error (MSE) in each respective region. The accompanying bar plot shows the number of regions where each model achieved the best performance.

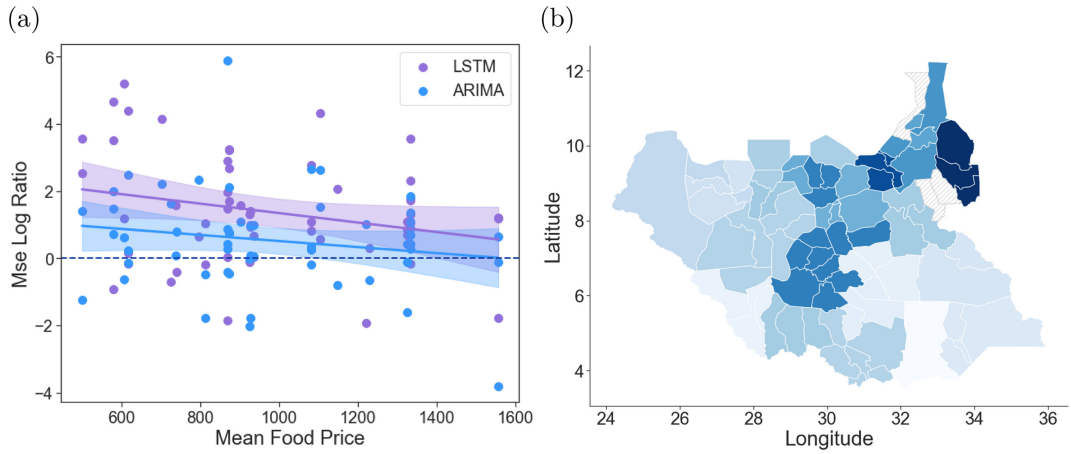



**FIGURE 8** | (a) MSE log ratio (in $y$) per mean food price (in $x$). One point represents one model/region compared to our model. A positive point shows better performance for our model, contrary to negative points. The line represents the linear regression curves for LSTM (purple) and ARIMA (blue). The shaded area is the 95% confidence interval. (b) Average food price per region before the forecasting evaluation period (February 2022–July 2022). The darker blue represents higher values.

## 6 | Discussion

By expanding beyond static value-based models, this study enhances our understanding of migration trends in several ways. Static models typically rely on fixed data points, offering a snapshot view that may overlook underlying trends and patterns. In contrast, a dynamic approach analyzes changes over time, revealing how migration patterns—not just raw values—influence overall migration. Additionally, the inclusion of sequences associated with food prices and conflict dynamics allows us to model more complex temporal interactions and understand how they affect migration.

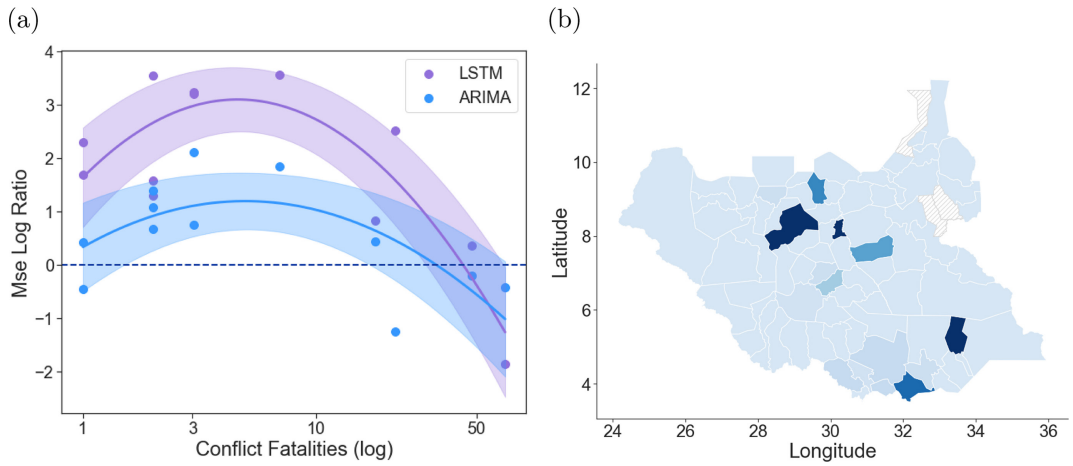



**FIGURE 9** | (a) MSE log ratio (in $y$) per conflict fatalities (log) (in $x$). One point represent one model/region comparison to our model. A positive point shows better performance for our model, contrary to negative points. The line represents the quadratic regression curves for LSTM (purple) and ARIMA (blue). The shaded area is the 95% confidence interval. (b) Sum of conflict fatalities per region before the forecasting evaluation period (February 2022–July 2022). The darker blue represents higher values.

However, the study also faces significant data challenges. Continuous, fine-resolution data on migration are scarce, with most available datasets being limited in temporal and geographical scope. While projects like those led by the International Organization for Migration (IOM) in Iraq offer better granularity, they are hindered by data discontinuities, which pose a challenge for time series analysis and pattern extraction. The cost and logistical complexity of gathering migration data, typically through interviews at key migration points, contribute to the rarity of long-term data projects. Emerging techniques using nightlight satellite imagery or social media data show promise but are still developing and currently lack reliability.

Another limitation is that while our model excels at identifying changes, it cannot predict shifts from long periods of inactivity to activity. This limitation is not specific to our approach but is common in autoregressive models. Indeed, most statistical models often miss the onset of new trends following prolonged stability.

On the other hand, a significant advantage of the methodology presented here is its efficacy in extracting meaningful patterns and making predictions, even with small datasets. This capability is particularly crucial in fields like migration studies, where data can be scarce or irregular. The model's ability to extract and use underlying patterns in limited datasets sets it apart from traditional models that often require large, comprehensive datasets to yield accurate forecasts. Furthermore, by relying on a limited number of covariates, our approach reduces the risk of overfitting, simplifies interpretation, and requires less computational power and data, making it more practical for many applications.

The methodology presented thus has broader applications beyond migration studies and could contribute to forecasting in fields such as conflict fatalities, or epidemiology, where data can be scarce. By focusing on extracting and using temporal patterns, our approach provides a robust framework for improving predictive accuracy and understanding complex temporal dynamics across various domains.


### Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 101002240). Open access funding provided by IReL.


### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The data that support the findings of this study are available in South_Sudan_Migra at https://github.com/ThomasSchinca/South_Sudan_Migra, Reference Number a45d99a. These data were derived from the following resources available in the public domain:

- Market food price (WorldBank): https://microdata.worldbank.org/index.php/catalog/4505
- Conflict fatalities (UCDP): https://ucdp.uu.se/downloads/
- Migration flows IOM dataset: https://dtm.iom.int/datasets?f%5B0%5D=dataset_component%3A4&f%5B1%5D=dataset_country%3A81

### Endnotes

[1] Multiple theories account for our ability to recognize temporal patterns in our environments. Feature detection theory, for example, suggests that we process information by sorting and filtering incoming stimuli. Our feature detectors are individual or groups of neurons that encode specific features. More complex feature detectors can detect more elaborate and specific features. We can identify meaningful sequences of events thanks to our feature detection system (Reed 2013).

[2] Note that the definition of "migration" in this context also include voluntary or economic migration, such as seasonal labor movements or long-term economic relocation.

[3] Three additional regions (Luakpiny, Manyo, and Ulang) were excluded due to the absence of migrant data during the study period.

[4] For a more detailed explanation of the DTW algorithm and its implementation, please refer to Appendix A and Müller (2007).

[5] Note that using a simple sum of these distances may not be the most effective approach. A more refined method could involve a weighted average, where the weights are derived from the data itself, potentially

leading to improved accuracy. Nevertheless, our current focus is on maintaining simplicity in our methodology, and as such, we opt for the straightforward summative approach at this stage. Results for different weights do not substantially differ from the ones presented here and are reported in Figure D2.

[6] The model uses an observation window of 6 months for all variables. The threshold to classify two cases as similar is set to 3, using normalized (min–max from 0 to 1) sequences. Several threshold values were evaluated, however, the highest scoring threshold value of 3 was found to be the most effective, achieving a valid balance between specificity and generalization. The individual flexibility for each variable is approximately 1 under this threshold configuration.

[7] Note that all subsequences are normalized with min–max scaling from 0 to 1 to facilitate the comparison of shapes rather than magnitudes. The "past futures" are scaled based on their corresponding subsequences, and the forecasted values are determined from normalized mean values. Afterward, the forecasted values undergo a reverse normalization transformation to obtain the final prediction, ensuring its adjustment to the original scale of the data.

[8] Using ratios directly can be misleading due to their asymmetry: Decreases fall between zero and one (e.g., 0.5 means half), while increases can be any number above. So simple ratios are not symmetric around the point of no change (1.0). By applying the logarithm to ratios, we achieve symmetry: no change corresponds to zero, increases are positive, and decreases are negative, making the log of 2.0 and 0.5 symmetric around zero.

[9] In other words, each point corresponds to $\log\left(\frac{MSE_{benchmark}+1}{MSE_{ShapeFinder}+1}\right)$. The addition of one is a common technique to ensure that the log function is defined even when MSE values are zero.

## References


Andrée, B. P. J. 2021. "Monthly Food Price Estimates by Product and Market." WLD_2021_RTFP_v02_M; Version 12–02.

Arcand, J.-L., and L. Mbaye. 2013. "Braving the Waves: The Role of Time and Risk Preferences in Illegal Migration From Senegal."

Ariu, A., F. Docquier, and M. P. Squicciarini. 2016. "Governance Quality and Net Migration Flows." *Regional Science and Urban Economics* 60: 238–248.

Baurmann, M., T. Gross, and U. Feudel. 2007. "Instabilities in Spatially Extended Predator–Prey Systems: Spatio-Temporal Patterns in the Neighborhood of Turing–Hopf Bifurcations." *Journal of Theoretical Biology* 245, no. 2: 220–229.

Chadefaux, T. 2021. "A Shape-Based Approach to Conflict Forecasting." *International Interactions* 48: 633–648.

Dasgupta, S., M. D. Moqbul Hossain, M. Huq, and D. Wheeler. 2016. "Facing the Hungry Tide: Climate Change, Livelihood Threats, and Household Responses in Coastal Bangladesh." *Climate Change Economics* 7, no. 03: 1650007.

Davenport, C., W. Moore, and S. Poe. 2003. "Sometimes You Just Have to Leave: Domestic Threats and Forced Migration, 1964-1989." *International Interactions* 29, no. 1: 27–55.

Davies, S., T. Pettersson, and M. Öberg. 2022. "Organized Violence 1989–2021 and Drone Warfare." *Journal of Peace Research* 59, no. 4: 593–610.

De Haas, H. 2011. *The Determinants of International Migration: Conceptualising Policy, Origin and Destination Effects*. IMI Working Paper Series 32. Oxford, UK: International Migration Institute, University of Oxford.

Dimant, E., T. Krieger, and D. Meierrieks. 2013. "The Effect of Corruption on Migration, 1985–2000." *Applied Economics Letters* 20, no. 13: 1270–1274.

Engel, S., and A. M. Ibáñez. 2007. "Displacement due to Violence in Colombia: A Household-Level Analysis." *Economic development and cultural change* 55, no. 2: 335–365.

Fearon, J. D., and A. Shaver. 2020. *Civil War Violence and Refugee Outflows*. Stanford, CA: Stanford University. Unpublished paper.

Gallistel, C. R. 1990. *The Organization of Learning*. Cambridge, MA: MIT Press.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9, no. 8: 1735–1780.

Kaczan, D. J., and J. Orgill-Meyer. 2020. "The Impact of Climate Change on Migration: A Synthesis of Recent Empirical Insights." *Climatic Change* 158, no. 3-4: 281–300.

Kassar, H., and P. Dourgnon. 2014. "The Big Crossing: Illegal Boat Migrants in the Mediterranean." *European Journal of Public Health* 24, no. suppl_1: 11–15.

Kelly, L. J., and S. Khemlani. 2019. "The Consistency of Durative Relations." In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, edited by A. Goel, C. Seifert, and C. Freksa, 1998–2004. Austin, TX: Cognitive Science Society.

Lauby, J., and O. Stark. 1988. "Individual Migration as a Family Strategy: Young Women in the Philippines." *Population Studies* 42, no. 3: 473–486.

Malchow, H., S. V. Petrovskii, and E. Venturino. 2007. *Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation*. Boca Raton, FL: Chapman and Hall/CRC.

Massey, D. S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor. 1993. "Theories of International Migration: A Review and Appraisal." *Population and Development Review* 19: 431–466.

Massey, D. S., L. Goldring, and J. Durand. 1994. "Continuities in Transnational Migration: An Analysis of Nineteen Mexican Communities." *American Journal of Sociology* 99, no. 6: 1492–1533.

Mayen, J. V., E. Wood, and T. Frazier. 2022. "Practical Flood Risk Reduction Strategies in South Sudan." *Journal of Emergency Management* 20, no. 8: 123–136.

Melander, E., and M. Öberg. 2006. "Time to Go? Duration Dependence in Forced Migration." *International Interactions* 32, no. 2: 129–152.

Moore, W. H., and S. M. Shellman. 2004. "Fear of Persecution: Forced Migration, 1952-1995." *Journal of Conflict Resolution* 48, no. 5: 723–745.

Müller, M. 2007. "Dynamic Time Warping." *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer, pp. 69–84.

Ortega, F., and G. Peri. 2009. "The Causes and Effects of International Migrations: Evidence From OECD Countries 1980-2005." Cambridge, MA: National Bureau of Economic Research Technical Report National Bureau of Economic Research Technical Report

Reed, S. K. 2013. *Psychological Processes in Pattern Recognition*. Cambridge, MA: Academic Press.

Schaeken, W., P. N. Johnson-Laird, and G. D'Ydewalle. 1996. "Mental Models and Temporal Reasoning." *Cognition* 60, no. 3: 205–234.

Schon, J. 2019. "Motivation and Opportunity for Conflict-Induced Migration: An Analysis of Syrian Migration Timing." *Journal of Peace Research* 56, no. 1: 12–27.

Smith, M. D., and D. Wesselbaum. 2022. "Food Insecurity and International Migration Flows." *International Migration Review* 56, no. 2: 615–635.

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50, no. 4: 523–532.

Taylor, J. E. 1999. "The New Economics of Labour Migration and the Role of Remittances in the Migration Process." *International Migration* 37, no. 1: 63–88.

Turkoglu, O., and T. Chadefaux. 2019. "Nowhere to Go? Why Do Some Civil Wars Generate More Refugees Than Others?." *International Interactions* 45, no. 2: 401–420.

Van Dalen, H. P., G. Groenewold, and J. J. Schoorl. 2005. "Out of Africa: What Drives the Pressure to Emigrate?." *Journal of Population Economics* 18, no. 4: 741–778.

## Appendix A

### Dynamic Time Warping (DTW): Technical Details

DTW is an algorithm used for measuring the similarity between two temporal sequences which may vary in speed. Originally developed for speech recognition, it has since found applications in a wide range of fields including data mining and time series analysis.

The fundamental principle of DTW is to compare two time series by aligning their indices in a way that minimizes the cumulative distance between them. This comparison is not restricted to matching equivalent time points, allowing DTW to effectively handle sequences that unfold at different rates.

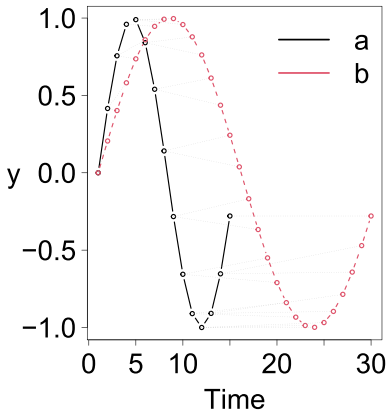



**FIGURE A1** | Dynamic time warping.

Consider two time series $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$. DTW aims to find an alignment between these two series that minimizes the total distance. The DTW distance between $X$ and $Y$ is defined as

$$DTW(X,Y) = \min \sqrt{\sum_{k=1}^{K} d(w_k)^2}, \tag{A1}$$

where $w_k$ are the elements of the warping path $W$, a set of pairs that define a mapping between $X$ and $Y$, and $d(w_k)$ is the distance between the paired points in $X$ and $Y$.

The warping path $W$ is typically computed using dynamic programming. The key idea is to build a matrix $D$ where each element $D[i,j]$ represents the distance between $x_i$ and $y_j$ plus the minimum cumulative distance to reach $D[i,j]$ from the starting point (1,1). The optimal path is then traced back from $D[n,m]$ to $D[1,1]$.

The distance matrix $D$ is defined as

$$D[i,j] = d(x_i,y_j) + \min(D[i-1,j], D[i,j-1], D[i-1,j-1]), \tag{A2}$$

where $d(x_i,y_j)$ is the distance between $x_i$ and $y_j$, often calculated using the Euclidean distance.

## Appendix B

### Long Short-Term Memory (LSTM) Network: Detailed Overview

LSTM networks are a type of recurrent neural network (RNN) particularly suited for learning from sequences of data. They are widely used for their ability to capture long-term dependencies in sequence data, which traditional RNNs often fail to do.

LSTMs have been successfully applied in various domains such as natural language processing, time series forecasting, and sequence

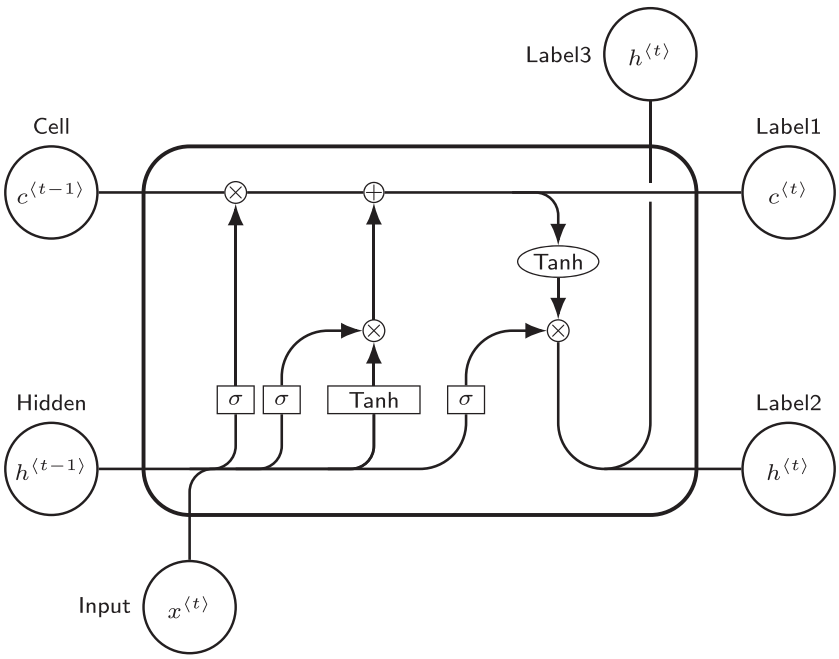



**FIGURE B1** | Schematic representation of an LSTM cell. The cell state $C_t$ is updated based on the inputs from the forget gate, input gate, and output gate, influenced by the previous cell state $C_{t-1}$ and hidden state $h_{t-1}$.

generation, showcasing their versatility and effectiveness in handling sequential data with complex temporal dependencies. We present here a very short introduction to this type of neural network. For a more comprehensive review of LSTMs, see Hochreiter and Schmidhuber (1997).

The LSTM network consists of a unique structure with memory cells, as presented in Figure B1, and multiple gates controlling the flow of information. The mathematical equations defining the LSTM architecture are as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \text{ (forget gate activation)} \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \text{ (input gate activation)} \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \text{ (cell input activation)} \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \text{ (cell state)} \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \text{ (output gate activation)} \\
h_t &= o_t \cdot \tanh(C_t) \text{ (hidden state)}
\end{aligned}
$$

Here, $f_t$, $i_t$, and $o_t$ represent the activations of the forget gate, input gate, and output gate, respectively. $\tilde{C}_t$ is the cell input activation, $C_t$ is the cell state, and $h_t$ is the hidden state. $W_f$, $W_i$, $W_C$, $W_o$, $b_f$, $b_i$, $b_C$, and $b_o$ are the respective weights and biases of the LSTM. The functions $\sigma$ and tanh denote the sigmoid activation function and the hyperbolic tangent activation function, respectively.

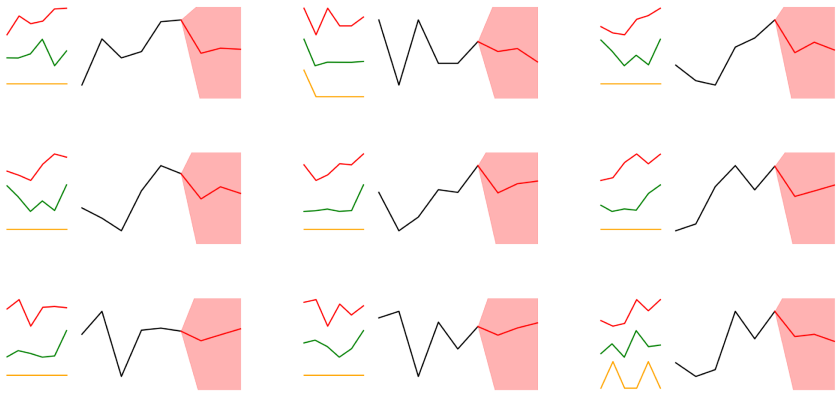
The LSTM makes decisions about what to store, delete, and output at each step in the sequence through its gates:

- *Forget gate*: Decides what information is discarded from the cell state.
- *Input gate*: Updates the cell state with new information.
- *Output gate*: Determines the next hidden state, influencing the output at the current step based on the cell state.

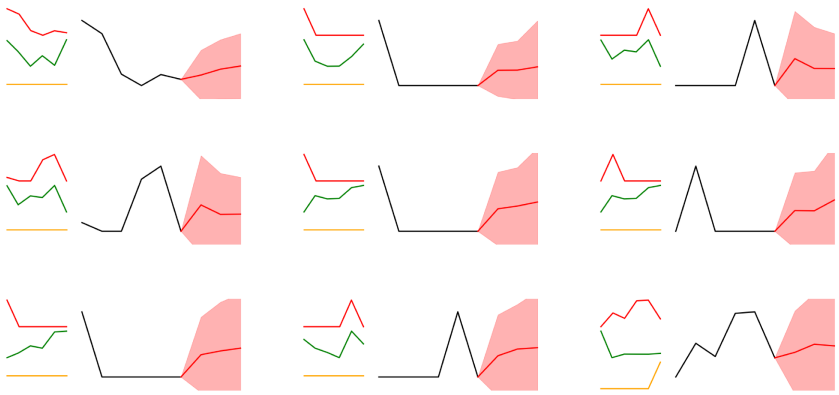
In our model, the LSTM architecture consists of two layers, each with 10 cells. We utilize an Adam optimizer for training, which is known for its efficiency in handling sparse gradients and adaptive learning rates. The loss function employed is based on the mean squared error (MSE) of the fit. Although validation data are commonly used to prevent overfitting, in our case, it did not yield improved results due to the limited size of the validation dataset.

## Appendix C

### Sample Matches of Identified Patterns



(a) Pattern with the highest standard deviation



(b) Pattern with the lowest standard deviation

**FIGURE C1** | Illustrative patterns of temporal migration patterns (red), food prices (green), and conflict (yellow) from the nine dynamic combinations with the highest and lowest mean standard deviation of the past futures, identified through dynamic time warping analysis. Each cell shows the combination of patterns on the left, with migration dynamic (in black) and the past future (in red) with the standard deviation as a confidence interval on the right.

## Appendix D

## Additional Results

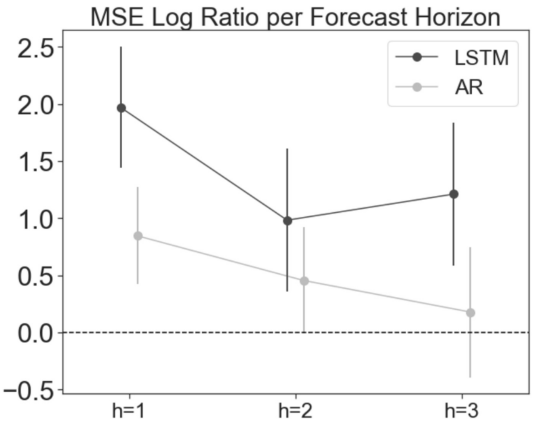



**FIGURE D1** | MSE log ratio for forecast horizons $h \in [1,2,3]$. Higher values indicate that the model performs worse than the ShapeFinder.

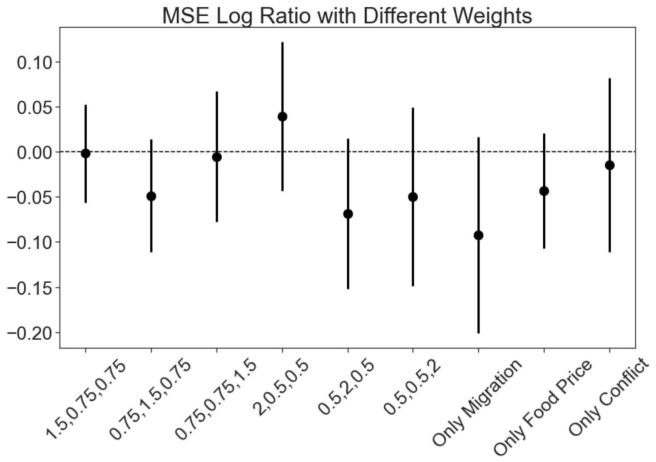



**FIGURE D2** | MSE log ratio per model weight configuration compared to the equal weight model. It illustrates the relationship between model weights and the performance outcomes. Notably, variations in model weights do not statistically alter the model's results. The error bars across different weight configurations overlap, indicating that changes in weights do not lead to statistically significant differences in the model's performance.