Enhancing the regularity and granularity of conflict forecasts for policy makers

Eric Frey
Fundació d'Economia Analítica

Hannes Mueller
IAE (CSIC), BSE, CEPR

Christopher Rauh
IAE (CSIC), University of Cambridge, BSE, CEPR, PRIO

Ben Seimon
Fundació d'Economia Analítica

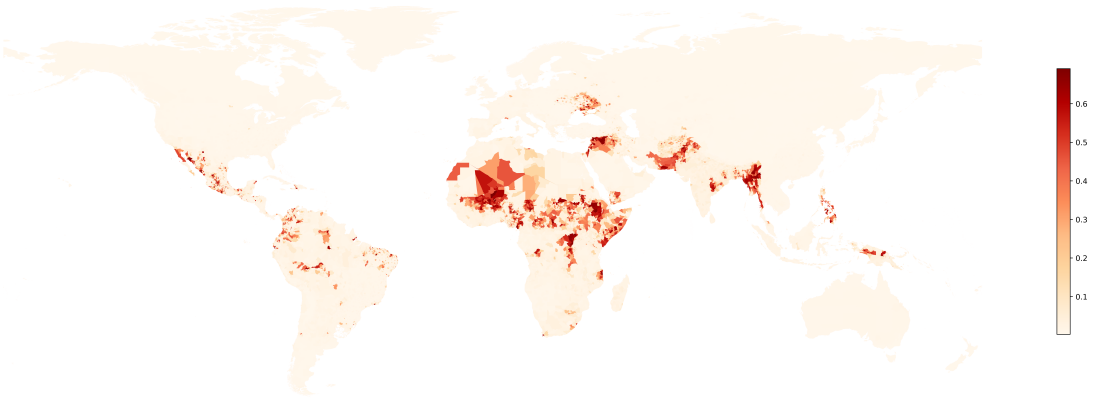
Emmett Sexton
Fundació d'Economia Analítica

March 2025

Executive summary

This report presents a major advance in conflict forecasting by integrating subnational machine learning models and newspaper text data, and to inform early-warning efforts and policy design. We develop and publicly release high-resolution forecasts of conflict risk at the ADM2 level (e.g., counties or districts), enabling more precise and actionable early-warning tools.

Global risk of experiencing any battle death within the next 12 months



Using state-of-the-art natural language processing and artificial intelligence, we identify latent indicators of instability—especially in areas with limited recent history of events—by combining structured historical event data with unstructured textual information. These forecasts cover battle-related fatalities, riots, and sexual violence, with predictions available for 3- and 12-month horizons at `conflictforecast.org`.

The report also outlines several promising directions for future work, including improved geolocation of news articles using large language models, the use of deep learning for spatial prediction, and integration with policy intervention data.

This work supports the ongoing collaboration between Conflict Forecast and the FCDO's Global Security Rapid Analysis program to improve conflict prevention and humanitarian preparedness in the face of rising global instability.

Contents

1 Introduction

Violent conflict remains one of the most pressing challenges to human security, sustainable development, and supply chains. Despite major global investments in peacebuilding, conflict continues to erupt in regions with little prior warning, displacing communities, disrupting economies, and reversing development gains. Accurate and timely forecasts can help policymakers and humanitarian actors intervene early, allocate resources more effectively, and ultimately save lives. Against this backdrop, improving the spatial precision and policy relevance of conflict forecasting systems is an urgent priority.

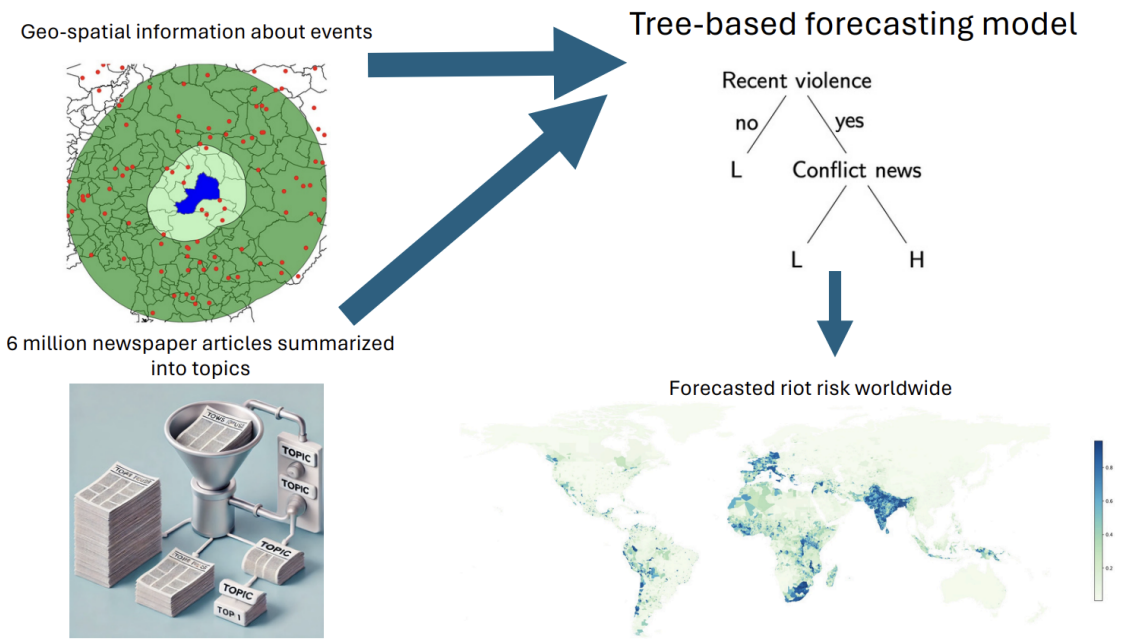
Prominent early-warning systems are made publicly available by institutions such as the Violence and Impacts Early-Warning System (VIEWS), the Armed Conflict Location and Event Data (ACLED), Conflict Alert System (CAST), and Conflict Forecast (CF). These cover a range of different conflict event data sources, spatial granularities, forecasting horizons, target definitions and evaluation metrics. At CF, our focus is on the utilization of text data to improve predictive performance for geographies with no recent history of violence. Our previous collaborations with the Global Security Rapid Analysis (GSRA) program at the Foreign Commonwealth and Development Office (FCDO) have demonstrated that our national forecasts can help to shape prevention work across geographic directorates.

Advances in artificial intelligence (AI) and machine learning offer new tools to tackle this forecasting challenge by identifying subtle and complex patterns in large volumes of data. Conflict rarely arises from a single cause; instead, it is often preceded by a confluence of signals—economic shocks, political instability, social unrest—that may only be detectable when viewed together across time and space. AI models are particularly well-suited to this task, as they can learn from diverse data sources, capture nonlinear relationships, and generalize from past patterns to new contexts. This makes them a powerful complement to traditional conflict analysis approaches, enabling earlier and more granular identification of emerging risks.

To now, our subnational forecasts have relied upon the Peace Research Institute Oslo (PRIO) grid-cells as the spatial unit for analysis, partly due to their computational convenience. The first key output of this project is an expansion to the ADM2 level across both the Upssala Conflict Data Program (UCDP) (Davies, Engström, Pettersson, and Öberg, 2024) and ACLED event datasets (Raleigh, Kishi, and Linke, 2023) in order to better align with the needs of operational and strategic policymakers. Figure 1 illustrates how we combine vast amounts of geo-spatial data and

newspaper text in a machine learning model to generate risk indicators at the ADM2 level. Predictions related to fatalities (UCDP), riots (ACLED) and sexual violence (ACLED) for both a 3 and 12 month forecasting horizon are now publicly available at `conflictforecast.org` and are updated monthly.[1]

Figure 1: Illustration of violence prediction model



A key contribution of this project is the advancement of the inclusion of newspaper text into our forecasting models at the local level. Building on recent advances in natural language processing, we develop models that leverages an extensive database of news articles to anticipate changes in conflict risk—particularly in areas with limited recent violence, where traditional time series signals are weak or absent. This approach allows us to capture latent drivers of instability, such as political unrest, repression, or environmental shocks, as reflected in the narrative structure of news coverage. By enriching structured conflict event data with unstructured textual information, we offer a complementary lens to anticipate emerging threats with improved geographic precision and lead times.

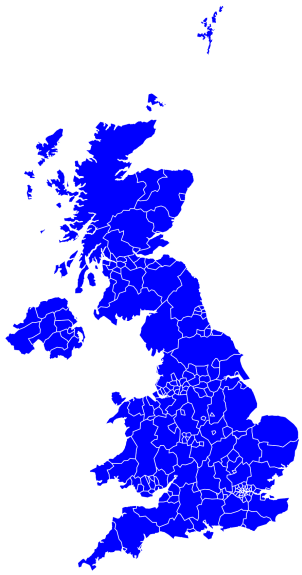
[1]See Mueller, Rauh, and Seimon (2024) for an overview of the rest of the data available on the website.

2 Data

This section details the data used for our ADM2 forecasting pipeline.[2] If applicable, we briefly describe the preprocessing steps required to aggregate/disaggregate the source data to our unit of analysis.

Unit of analysis Our unit of analysis is the ADM2/month level. We rely on the geoBoundaries Global Database of Political Administrative Boundaries Database, which is an online, open license resource of information on administrative boundaries (i.e., state, county) for every country in the world. Figure 2 demonstrates the granularity of these boundaries for the UK, where the ADM2 level corresponds to counties.

Figure 2: ADM2 divisions in the UK



In total, there are almost 50,000 administrative units worldwide across 198 countries. It should be noted that there is significant heterogeneity in the size of ADM2 units both within and across countries. For example, Brazil has over 5,500 unique ADM2 units, the most of any country. These range in size from less than $1km^2$ to over

[2] A reference for all data sources can be found in the supporting document *Data sources.xlsx*.

150,000km^2. The USA has a comparable total land size, yet has fewer than 3,500 ADM2 units which range in size from 3km^2 to over 375,000km^2.

Conflict data The Armed Conflict Location and Event Data Project (ACLED) and the Uppsala Conflict Data Program (UCDP) are the most prominent providers of conflict data. However, they differ in a number of ways, namely their spatial/temporal coverage, update frequency and data validation processes. Both are suitable for developing early-warning systems at the ADM2 level, yet their event definitions lead to different target definitions for forecasting.

- **UCDP event definition:** "An incident where armed force was used by an organized actor against another organized actor, or against civilians, resulting in at least 1 direct death at a specific location and a specific date" (Stina, 2024). They distinguish between three types of violence: state-based, one-sided and non-state events.
- **ACLED event definition:** "Involvement of designated actors – e.g. a named rebel group, a militia, or state forces (with the sole exception of Unidentified Armed Group and generic categories including Rioters, Protesters, and Civilians). They occur at a specific named location (identified by name and geographic coordinates) and on a specific day" (ACLED, 2024).

Our previous engagements with FCDO, as well as existing CF publicly available forecasts, have solely relied on UCDP as the source of "ground-truth" for the absence/presence of violence. We do not discriminate across types of violence, and instead consider total violence (i.e. the sum across state based, non-state, and one-sided conflict deaths) for any given ADM2-month.

Notice that UCDP will code an event if and only if a fatality has occurred. However, in ACLED's definition a fatality is not a necessary condition for an event. This means they have a more granular coding of event types (6 event types and 25 sub-event types), which enables the definition of target variables for conflict not related to fatalities, but instead with respect to events.[3] We choose to make forecasts for the following ACLED definitions:

[3]Importantly, these event types are hierarchical. The ACLED hierarchy is defined in descending order as battles, explosions/remote violence, violence against civilians, protests, and riots. This means that if an event could be coded in more than one event type, it is only recorded once as the highest relevant event type. For example, if a riot or protest is taking place in a city, but then it escalates to the point of bombings or grenades exploding, that event would only be coded as a 'explosion/remote violence' event type, not a riot or protest.

- **Riot**: This is its own event type. It is defined as "violent events where demonstrators or mobs engage in disruptive acts" (ACLED, 2024).
- **Sexual violence**: This is a sub-event type within 'violence against civilians'. It is defined as "individuals (regardless of gender or age) experiencing harm of a sexual nature (including, though not limited to, rape, public stripping, sexual torture, mutilation of genitals, etc.)" (ACLED, 2024).

Note that we do not exclusively use the "*sexual violence*" sub-event type coding provided by ACLED. We use that sub-event type in combination with ACLED tags. ACLED tags are additional notes that let us recover sexual violence events that take place in conjunction with event types higher in ACLED's coding hierarchy. For example, if there is reported sexual violence against civilians during a battle, this would be coded as a battle event, but its tag may indicate that sexual violence also occurred. For this work, we code an instance of sexual violence if:

- ACLED codes this a sexual violence sub-event type; or
- ACLED codes a "sexual violence" or female targetting tag against any event.

Furthermore, it is important to highlight the variation in geographical coverage over time. UCDP's maintains constant, global geographical coverage for the entire history of data provision (Jan 1989 - present). By contrast, Appendix Figure A1 shows that ACLED's data provision, for all events, begins in 1997 for 50 countries. These are exclusively located in Africa. Additional countries then enter the dataset, such that by 2018 a total of 142 countries are covered. From 2021 ACLED then has complete country coverage. This poses a problem for reliable data preprocessing and feature engineering given that we utilize a panel data structure for our forecasting pipeline. Following extensive data analysis and model experimentation, we exclude all ACLED data prior to 2010.

Finally, whilst both datasets geo-locate recorded events, significant uncertainties can arise with respect to the exact coordinates. For UCDP, we only include events coded as having geographical precision up to the ADM2 level, and for ACLED we only include events that are recorded as being exact or nearby.[4]

[4]For completeness, this corresponds to including values [1, 2, 3] for UCDP's "where_prec" variable, and [1, 2] for ACLED's "geo_precision" variable.

Population data We rely on population data curated by Liu, Cao, Li, and Jie (2024). They provide yearly data from 1990 to 2022 with a spatial granularity of $0.0083° \times 0.0083°$. Since this is not complete temporal coverage, we apply some basic pre-processing steps. We compute the total population of all raster cells whose center falls within an ADM2 unit. We then linearly interpolate to go from the ADM2/year level to ADM2/month level. Finally, ADM2 population is assumed to grow from 2022 onwards in line with the associated country level population growth rate as sourced from the World Bank.

Text data Our text data is comprised of over 6 million documents from 1989 to present. These are downloaded from Factiva and are sourced from two newspapers (the New York Times and the Economist) and three news aggregators (the Associated Press, BBC Monitor, and LatinNews).[5] Text is downloaded according to rules set in an extensive query. As a generalization, a document is downloaded if a country or capital name appears in the title or lead paragraph. One limitation relates to the inherent bias of news data, particularly in political regimes where the media is censored or restricted. The inclusion of LatinNews as a source is specifically intended to improve the text signal for Latin America since BBC Monitor generally focuses on Asia and Africa.

Standard natural language preprocessing (NLP) techniques are used, including the removal of punctuation, stop words, and lemmatization. In addition to single words (unigrams), we also consider common combinations of two or three words (bigrams and trigrams). Any token (unigram, bigram, or trigram) that appears in at least half of the documents (too frequent) or in fewer than 200 documents (too infrequent) is also removed.

In order to allocate news articles to ADM2 units, we rely on prepositions to identify locations mentioned within an article. These locations are then assigned coordinates using Nominatim, which is a geocoding software that uses OpenStreetMap data to find locations on Earth by name. Appendix Figure A2 illustrates the coverage of locations mentioned in our database.

This results in a corpus whereby documents are assigned to the ADM2/month level. Hence, we have a set of documents that act as a proxy for the news landscape for every ADM2 unit for every month between January 1989 to present.

[5]Over two-thirds of all articles are sourced from The Associated Press and BBC Monitor, with the remainder mostly sourced from New York Times or the Economist. Documents sourced from LatinNews make up a very small proportion of the total corpus (<1%).

Geographic data The geographic data we collect can be split into two categories:

1. **Static:** These relate to geographic characteristics that remain relatively stable over time, specifically elevation, water coverage, and barren land. These variables all have a spatial granularity of $1km \times 1km$.

2. **Time variant:** These relate to climatic variables that change over time, specifically temperature and rainfall. The data is sourced from NASA, is available monthly from 1980 to present, and has a spatial granularity of $0.5° \times 0.625°$.

For static data, we compute the mean value of all raster cells whose center falls within an ADM2 unit. In the case of time-variant data, we assign the mean value of all climate data cells that intersect an ADM2 unit.

Gender inequality data Finally, we collect two additional datasets that are only used to forecast sexual violence.

1. **Male Dominance Index:** This index measures the historical prevalence of different group practices that positioned males in more dominant roles than females for ethnic groups across the globe (Guarnieri and Tur-Prats, 2023). Each ethnic group is only recorded as a point on the globe. If an ADM2 contains multiple recorded ethnic groups from the database, we assign that ADM2 the average index value. If an ADM2 doesn't contain any male dominance index data within its boundaries, we assign it the index value from the nearest ethnic group.

2. **Cross-Gender Friending Ratio:** Using global Facebook data, Bailey, Johnston, Kuchler, Kumar, and Stroebel (2025) define the cross-gender friending ratio (CGFR) as "as the ratio of female friends in men's networks to the share of female friends in women's networks". This compliments the historical ethnographic Male Dominance Index by providing a normalized estimate of gender segregated social networks using more recent data (2025). We reconcile the ADM2 boundaries used in their paper with the updated boundaries used in our report by taking weighted averages of CGFRs based on proportion of land cover overlap.

3 Forecasting

This section is broken down into the following sub-sections: i) forecasting components, ii) predictors/features, and iii) evaluation. Note that this is intended as a high-

level overview of the forecasts to be updated monthly at conflictforecast.org. For more details, the reader is referred to the codebook available for download on the website. Finally, predictions pertaining to a 36 month forecasting horizon, and/or regression tasks, can be made available to GSRA/FCDO upon request to the authors.

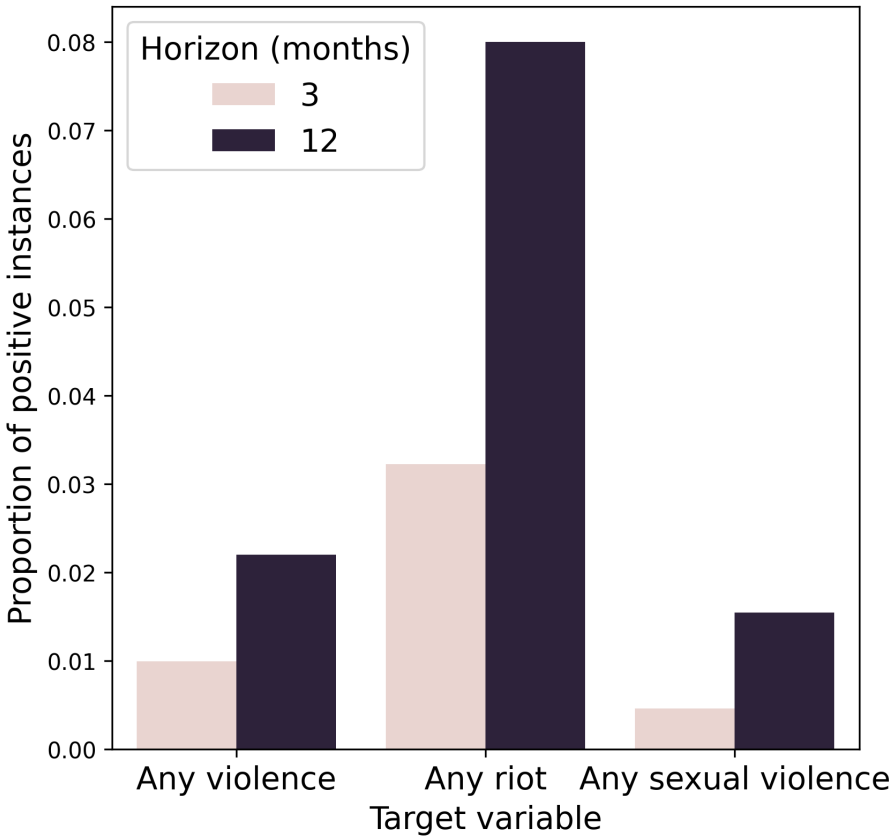
3.1 Forecasting components

Target variables Predictions with respect to 3 target variables for a forecast horizon (denoted h) of 3 months and 12 months are now publicly available and updated monthly. These are classification tasks with the following target definitions:

1. **Any violence, UCDP:** 1 if an ADM2 unit experiences at least one fatality in the next h months; 0 otherwise.
2. **Any riot, ACLED:** 1 if an ADM2 unit experiences at least one riot event in the next h months; 0 otherwise.
3. **Any sexual violence, ACLED:** 1 if an ADM2 unit experiences at least one sexual violence event in the next h months; 0 otherwise.

A measure of class imbalance is useful to understand the frequency of events. Figure 3 shows the percent of positive instances (1's) for each target variable and horizon. For example, in the case of riots, 3.22% of ADM2 regions experience at least one riot in the next 3 months.

Figure 3: Class imbalance across horizons and targets



Notes: For 'Any violence', class imbalance is computed for all ADM2 regions from January 1989 until February 2025. For 'Any riot' and 'Any sexual violence', it is computed for all ADM2 regions covered by ACLED from January 1997 until Februrary 2025.

Notice that, for any given target variable (i.e. row), the share of 1's increases as the forecasting horizon increases. Also, sexual violence events are the rarest of the event types that we forecast, with just 0.46% of ADM2 regions experiencing a sexual violence event in the next 3 months.

3.2 Predictors/features

Now that we have defined our target variable, we must select a set of generalizable features on which to train our machine-learning models and generate out-of-sample predictions. The objective is to provide the model with sufficient information to learn

complex, non-linear interactions between these predictors and the target variable.[6] They can be broadly categorized into four groups:

1. *Event history, ADM2:* Previous incidence is typically the most powerful predictor of future occurrence. Figure 4(A) shows that the likelihood of a fatality in the next 12 months is almost 70% if it has been just 1 month since the last fatality. This falls to less than 30% if it has been 12 months since the last fatality.

2. *Event history, neighbors:* There is a wealth of evidence supporting the hypothesis of spatial spillovers, particularly with respect to violence (Harari and Ferrara, 2018). To accurately capture historical event counts in neighboring regions, we rely on buffers of varying sizes as shown in Figure 4(B).

3. *Text:* We rely on the method developed by Mueller and Rauh (2017). This utilizes over 6 million geo-located newspaper articles which are passed through a Latent Dirichlet Allocation (LDA) topic model. This enables us to calculate 15 topic shares for each ADM2/month observation, where any single topic is represented by a "bag of words" as shown in Figure 4(C). These topic shares summarize how the local news landscape is changing over time in order to provide an early-warning signal for future onsets, escalations and/or de-escalations. Previous research at the national level indicates that topics such as "military conflict" predict increasing risk, whilst others such as "economics" and "civilian life" predict decreasing risk.

4. *Climate:* We include features related to precipitation and temperature. We hypothesise that deviations from historical climatic conditions can be a precursor to violence and/or social unrest, where there exists some supporting evidence from the literature Hsiang, Burke, and Miguel (2013), though so far the predictive power is limited (Mach, Kraan, Adger, Buhaug, Burke, Fearon, Field, Hendrix, Maystadt, O'Loughlin, et al., 2019).

Finally, we introduce a novel election feature. We have manually curated a dataset containing the dates of presidential and/or parliamentary elections on a monthly basis from 1989 to 2030 for 194 countries.[7] Figure 4(D) shows how the average riot

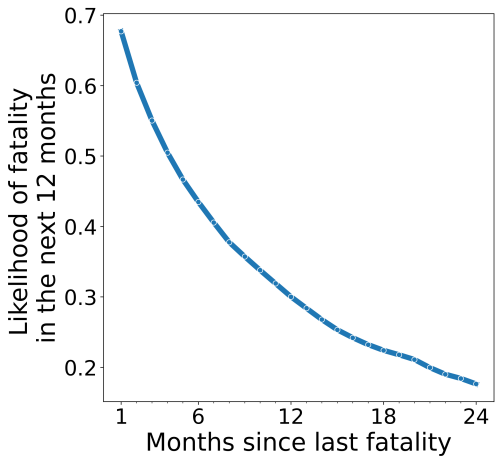
[6] It's important to note that predictors are not necessarily causal in nature. In fact, it is quite common that variables which are causally related to the target are not good predictors.

[7] We source this information from `electionguide.org` and Wikipedia. Elections from 2025 to 2030 are hard-coded and subject to change.

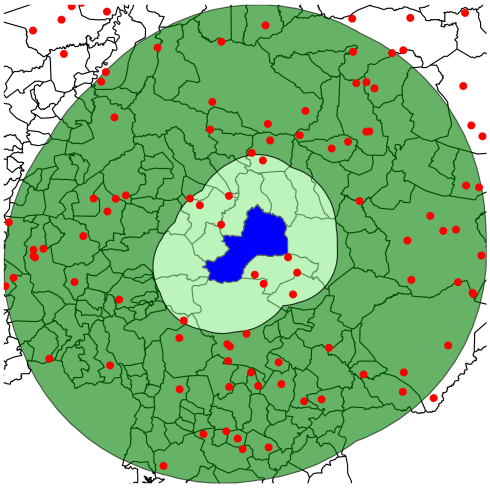
count increases as an election approaches, peaks at the month of an election, and quickly tails off in the months after an election.

Figure 4: Predictors/features

(A) Event history, ADM2

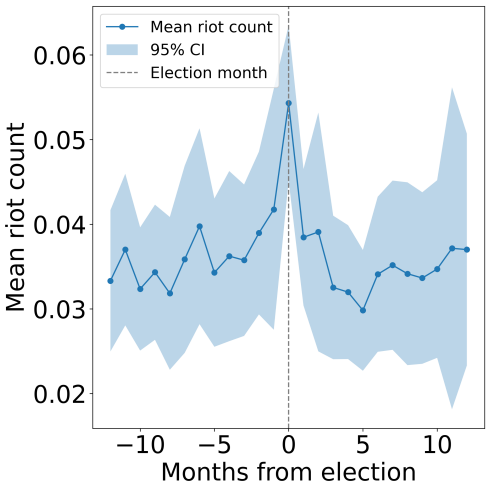


(B) Event history, neighbors



(C) Text

(D) Elections



NOTES: Panel A shows the likelihood of any battle-related death (UCDP) in the next 12 months, conditional on the number of months since the last fatality. Panel B demonstrates the use of buffers to capture neighboring events. The ADM2 unit of reference is the blue highlighted unit. The red dots indicate events. Red dots within the light-shaded green region would be classified as falling within a 25km buffer, whilst those in the dark-shaded green region would be classified as falling within a 25-100km buffer. Panel C shows the top 100 words associated with the military conflict topic as at Februrary 2025. Panel D shows the average riot count, with the 95% confidence interval, in the months before, during and after an election.

Subset methodology The final key component of our forecasting pipeline is the segmentation of observations into 6 distinct groups. We hypothesise that the predictors of risk at the ADM2 level depend on: i) event history, and ii) spatial spillovers. As a result, we train 6 different, specialist models on each of the sub-samples shown in Figure 5.

Figure 5: Subset matrix

		ADM2 event history		
		No recent (>24 months)	*Recent (<=24 months)*	*Ongoing*
Neighboring event	*No event within 25km*	20,571,167 (96.1%)	436,021 (2.0%)	50,013 (0.23%)
	Event within 25km	180,407 (0.8%)	125,047 (0.6%)	54,377 (0.25%)

NOTES: The table shows the number (share) of total observations according to our subset rules for UCDP any violence from January 1989 to February 2025.

This enables us to develop custom feature sets depending on the subset, and will also serve as a useful evaluation tool. For example, the subset in the top left contains ADM2 units which have not experienced an event (e.g. a battle related fatality or a riot) for more than 2 years. There are also no ongoing events within a 25km buffer of the selected unit. Hence, these are situations where event history is unlikely to provide a useful forecasting signal, and instead our model might rely on text or climatic features to capture future risk. By contrast, the subset in the bottom right represents a very different context. It is defined as all ADM2 units where there is both an ongoing event in the selected unit, and an event has occurred within a 25km buffer. In these cases, event history is a critical feature that must be included to optimize forecasting performance.

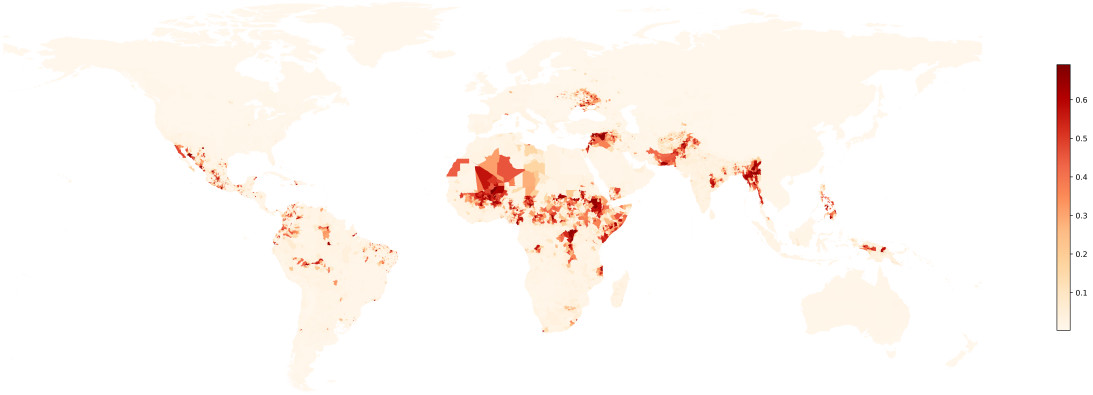
Tree-based machine learning algorithms Tree-based algorithms are robust, widely used ensemble machine learning algorithms that excel in predictive modelling tasks for tabular data. Their strength lies in handling high-dimensional data and capturing complex non-linear relationships. Their starting point is a decision tree, which generates predictions by repeatedly splitting the data based on feature values, much like a flowchart. Whilst a single decision tree is simple and interpretable, it is highly susceptible to over-fitting on the training data. To mitigate this, bagging and

boosting—two forms of ensemble learning—construct multiple trees and combine their outputs. In bagging, each tree is trained on a different random subset of the dataset, and their predictions are aggregated, reducing variance and stabilizing results. Boosting, on the other hand, builds trees sequentially, giving extra emphasis to instances misclassified by earlier trees to systematically refine the model. By blending the outcomes from numerous trees, these methods enhance generalization and yield more reliable predictions.

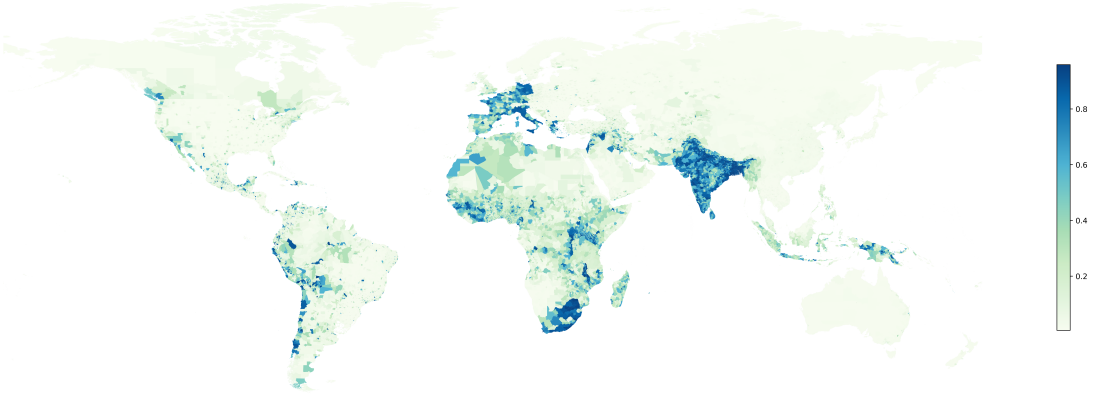
Predictions In Figure 6 we show our predictions for the next 12 months.

Figure 6: Global risk

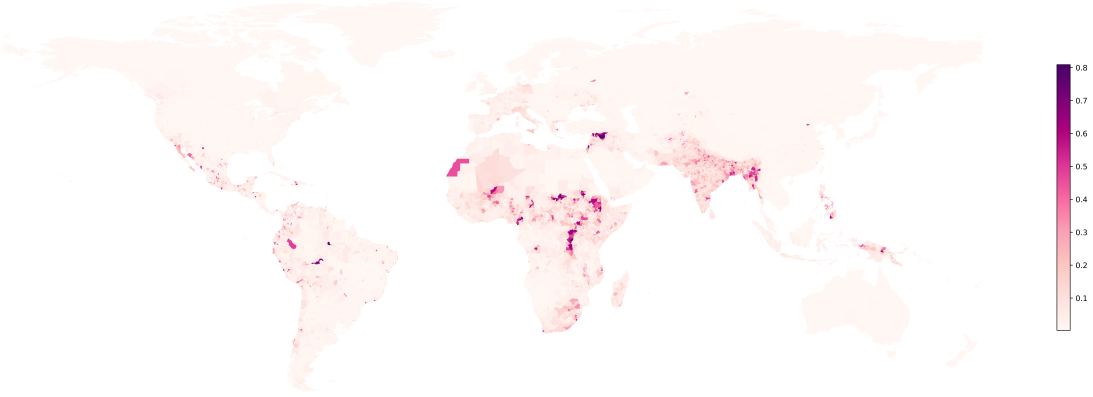
a) Any violence risk



b) Riot risk



c) Sexual violence risk



3.3 Evaluation

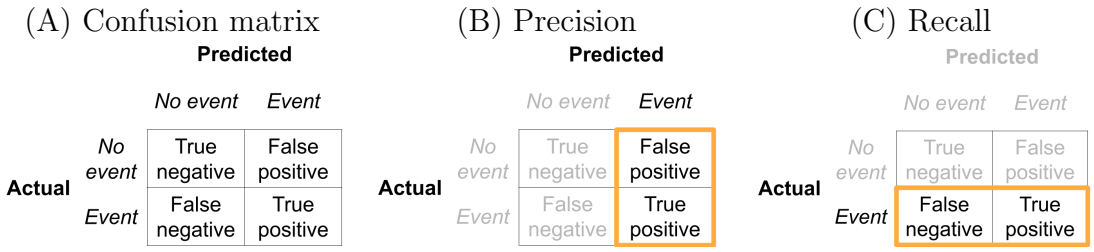
In Appendix A.2 we outline the pseudo-out-of sample procedure. In the following, we explain how we evaluate our predictions.

Metrics Our key metrics for evaluation are precision and recall:

- **Precision**: Of the predictions that we make, how many are correct?
- **Recall**: Of the realized events, how many did we predict?

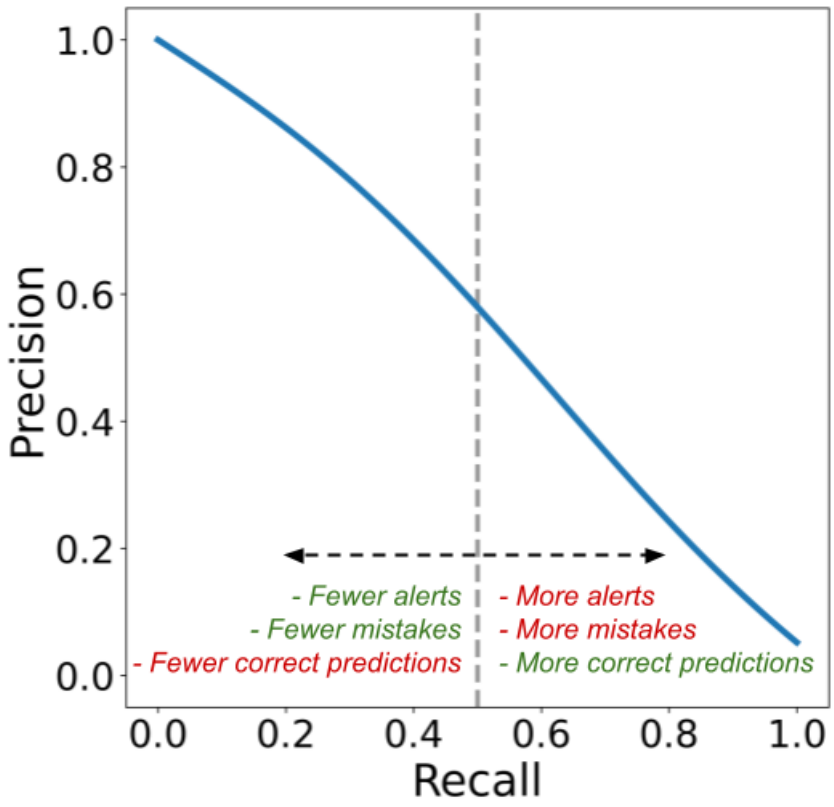
More concretely, their computations can be visualized for a binary classification task using a confusion matrix as in Figure 7. Precision is defined as $\frac{\text{True positives}}{\text{False positives} + \text{True positives}}$ and recall is defined as $\frac{\text{True positives}}{\text{False negatives} + \text{True positives}}$.

Figure 7: Confusion matrix: precision and recall



There is a direct trade-off between these two metrics, as captured by the precision-recall curve. In a classification setting, machine learning models output a probability between 0 and 1. To compute precision and recall, it is necessary to set a threshold to binarize our predictions. For example, setting the threshold at 95% would mean we predict a 1 for all observations where the predicted probability exceeds 95%, and we would predict a 0 for the remaining observations. In this case, we would be in a "high-precision" mindset, since we would be assigning 1's to only the cases where our model is very confident that an event *will* materialise. The converse is true if we set the threshold at 5% for a "high recall" mindset. We would only be assigning 0's to cases where our model is very confident that an event *will not* materialise. By computing precision and recall for all thresholds between 0 and 1, we derive a precision-recall curve, and the trade-offs for decision-makers when setting a threshold, as described in Figure 8.

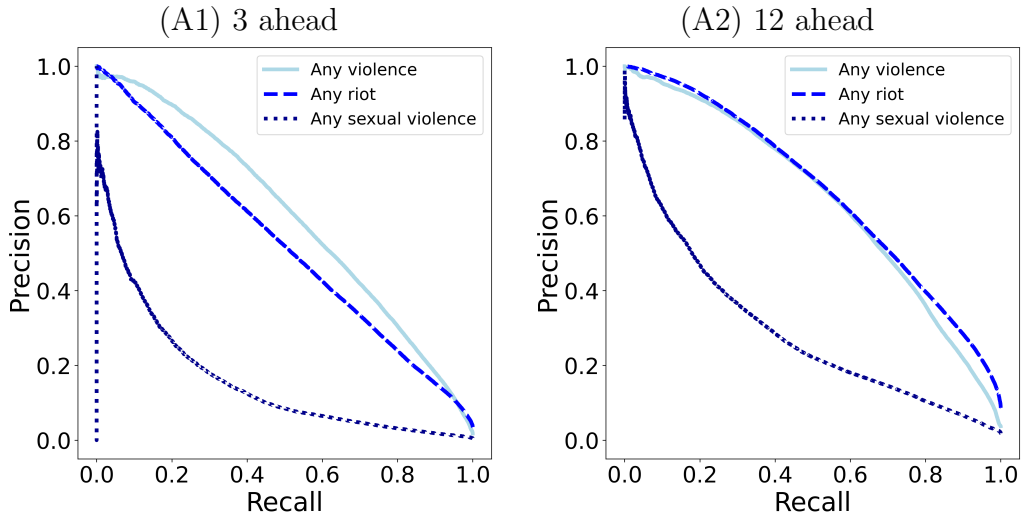
Figure 8: Precision-recall trade-off



Notes: The figure shows the trade-off between precision-recall starting from a base case of setting a threshold that results in 50% recall (grey dashed line). The blue line is derived using dummy predictions and is for demonstration purposes only.

Model performance Figures 9 and A3 show the precision-recall curves for all targets and for each of the 3 and 12 month forecasting horizons. The rows refer to the target and the columns refer to the forecasting horizon. Figure 9 shows performance in the aggregate, whilst Appendix Figure A3 evaluates the models with respect to each subset as described in Section 3.1. The corresponding legend for the latter is shown in Appendix Figure A4.

Figure 9: Model performance: precision-recall curves



Note that performance tends to be stronger in the 12 ahead than in the 3 month ahead case.

In Appendix Figure A3 we present the results broken down by subset. There are two key takeaways for our model performance:

1. **ADM2 event history**: Irrespective of target and horizon, performance decreases as we move from the "ongoing" to the "recent" to the "no recent" event subsets. The "no recent" subset is defined as the break out of an event in situations where there has been at least 2 years of no event. We typically also call these "hard onsets" because these are the hardest to predict. This is reflected in our forecasting performance.

2. **Spatial spillovers**: Performance improves in cases where there are ongoing events in neighboring units. This clearly demonstrates that effectively capturing spatial spillovers as part of feature engineering plays a crucial role in forecasting at the ADM2 level.

Finally, performance for sexual violence is the worst across the three targets. We posit that this may be partially due to low data quality, where our analysis suggests there are likely to be a significant number of false negatives coded into the dataset -

i.e. an ADM2-month has been coded as having no event in the source data, when in fact an event occurs. Additionally, there is a lack of regularly updated data sources that could provide a useful signal for possible future instances of sexual violence. The additional datasets we have included to try and improve performance are cross-sectional, which means the model cannot learn from changes over time. There is substantial work to be done to further improve data quality, availability and feature engineering in order to make these predictions more useful for policymakers.

4 Conclusion

In this report, we have outlined the forecasting methodology and performance of forecasts at the ADM2 level across three different targets (violence, riots and sexual violence) and multiple horizons (3 and 12 months). This represents a significant expansion of regularly updated forecasts which policymakers can use to support both strategic and operational policy decision-making.

Work undertaken for this project has brought to light a range of possible future research areas which we believe could be of benefit to GSRA/FCDO and the wider community. These include:

1. **Article geolocation:** This project highlighted that we have significant sparsity of text data at the ADM2 level. This is driven by the challenging task of allocating news articles to ADM2 units. The recent explosion of open-source Large Language Models (LLMs) provides a fruitful avenue to explore, particularly given their capabilities to understand context. Importantly, we believe that improved text coverage across time and space could have significant benefits for performance with respect to onsets.

2. **Model architectures:** For the moment, we have defaulted to classic tabular machine-learning algorithms (i.e. tree-based methods) to generate predictions. Training deep-learning models, such as Graph Neural Networks, could yield performance improvements, but requires a substantial amount of both human and computational resource that was not available in this project.

3. **Policy analysis:** Finally, there are an increasing number of geo-located policy datasets. For example, the recently published Geocoded Official Development Assistance Dataset has thousands of projects from European countries, United States, China, India, and the World Bank (Bomprezzi, 2025). This opens the possibility of combining these with our ADM2 risk estimates to analyze the causal effect of policy instruments on reductions in conflict risk.

References

ACLED (2024): "Armed Conflict Location and Event Data Codebook," Available at https://acleddata.com/knowledge-base/codebook/.

Bailey, Michael and Johnston, Drew and Kuchler, Theresa and Kumar, Ayush and Stroebel, Johannes (2025): "Cross-Gender Social Ties Around the World," American Economic Association Papers Proceedings.

Bomprezzi, Pietro; Dreher, Axel; Fuchs, Andreas; Hailer, Teresa; Kammerlander, Andreas; Kaplan, Lennart; Marchesi, Silvia; Masi, Tania; Robert, Charlotte; Unfried, Kerstin (2025): "Wedded to Prosperity? Informal Influence and Regional Favoritism," CEPR Discussion Paper, 18878 (v.2). Available at https://godad.uni-goettingen.de/home/.

Davies, Shawn and Engström, Garoun and Pettersson, Therese and Öberg, Magnus (2024): "Organized violence 1989–2023, and the prevalence of organized crime groups," Journal of Peace Research, 61(4), 673–693.

Guarnieri, Eleonora and Tur-Prats, Ana (2023): "Cultural Distance and Conflict-Related Sexual Violence," The Quarterly Journal of Economics.

Mariaflavia Harari and Eliana La Ferrara (2018): "Conflict, Climate, and Cells: A Disaggregated Analysis," The Review of Economics and Statistics, MIT Press, 100(4), pages 594-608.

Hsiang, Solomon M and Burke, Marshall and Miguel, Edward (2013): "Quantifying the influence of climate on human conflict," Science, 341(6151), 1235367.

Liu, Luling and Cao, Xin and Li, Shijie and Jie, Na (2024): "A 31-year (1990–2020) global gridded population dataset generated by cluster analysis and statistical learning," Scientific Data, 11(1), 124.

Mach, Katharine J and Kraan, Caroline M and Adger, W Neil and Buhaug, Halvard and Burke, Marshall and Fearon, James D and Field, Christopher B and Hendrix, Cullen S and Maystadt, Jean-Francois and O'Loughlin, John and others (2019): "Climate as a risk factor for armed conflict," Nature, 571(7764), 193–197.

Mueller, Hannes and Rauh, Christopher (2017): "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text," American Political Science Review, 112, 1–18.

Mueller, Hannes and Rauh, Christopher and Seimon, Ben (2024): "Introducing a global dataset on conflict forecasts and news topics," Data & Policy, 6, e17.

Raleigh, Clionadh and Kishi, Roudabeh and Linke, Andrew (2023): "Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices," Humanities and Social Sciences Communications, 10(1), 1–17.

Högbladh Stina (2024): "UCDP Georeferenced Event Dataset Codebook Version 24.1," Department of Peace and Conflict Research, Uppsala University.

A Appendix

A.1 Additional figures

Figure A1: ACLED country coverage over time

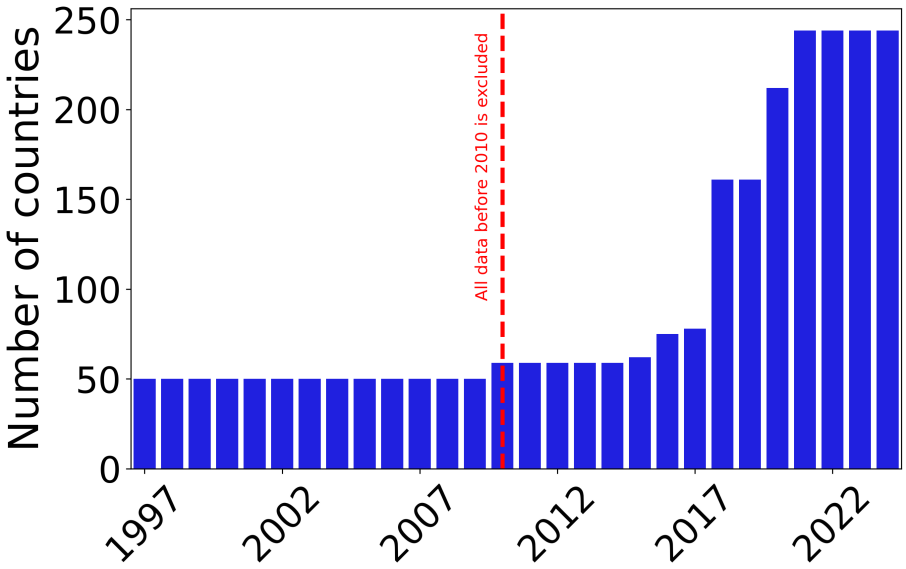


Figure A2: Locations mentioned in newspaper database

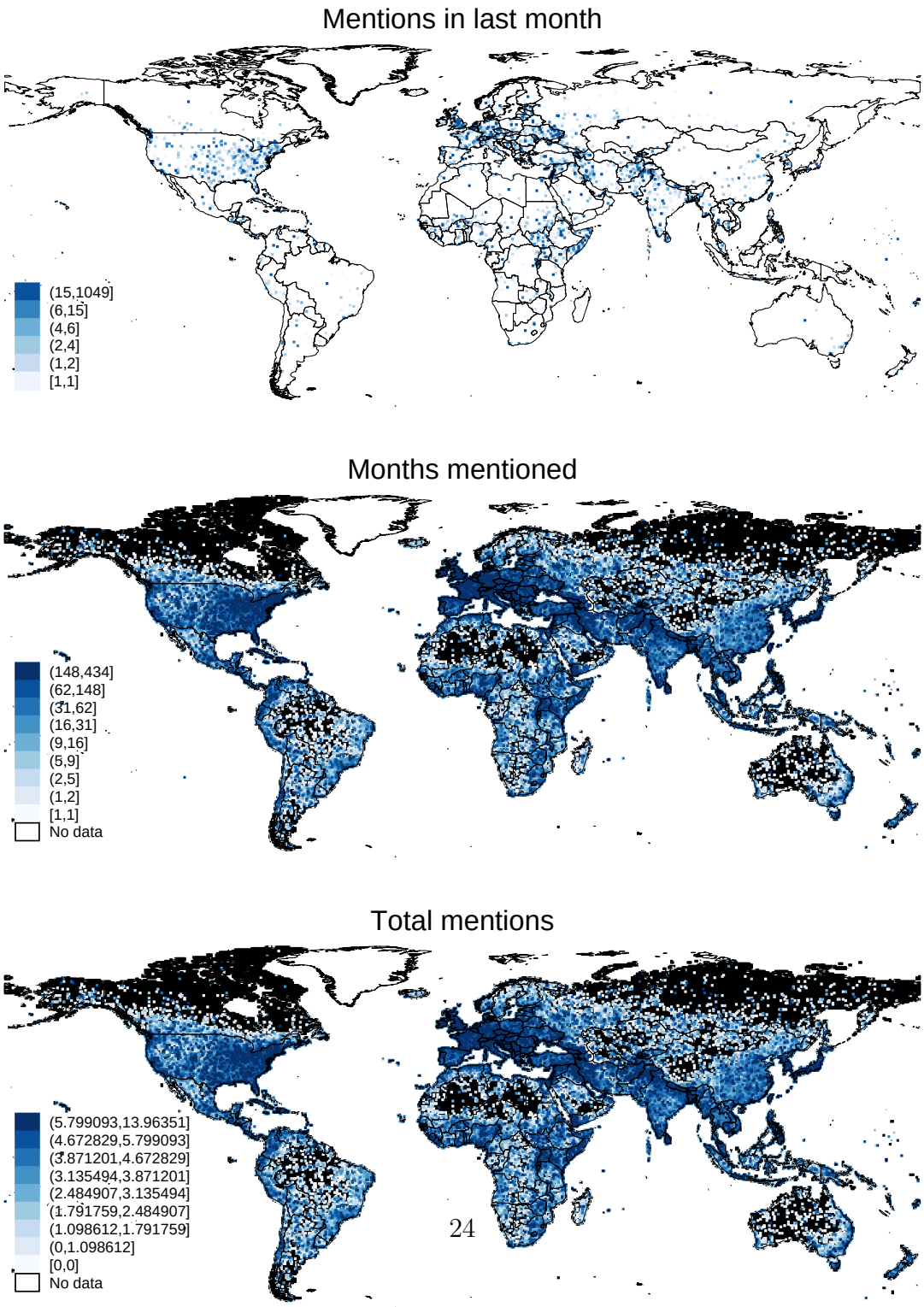
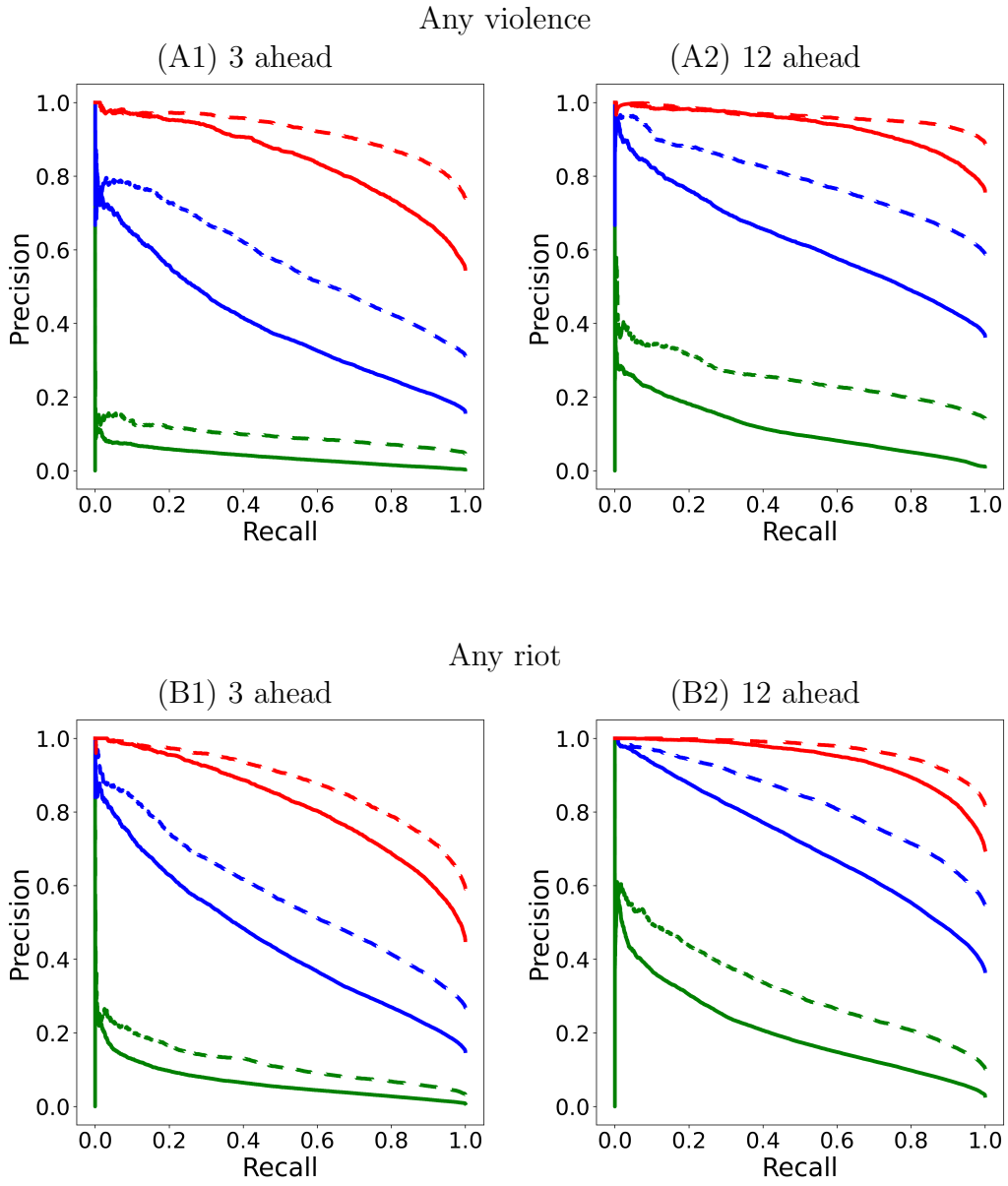


Figure A3: Model performance by subset: precision-recall curves



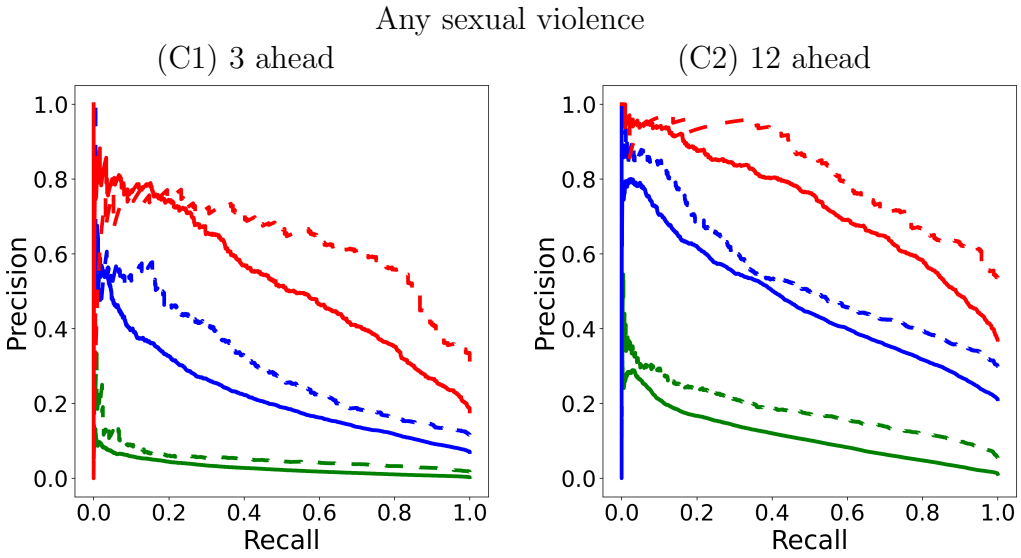
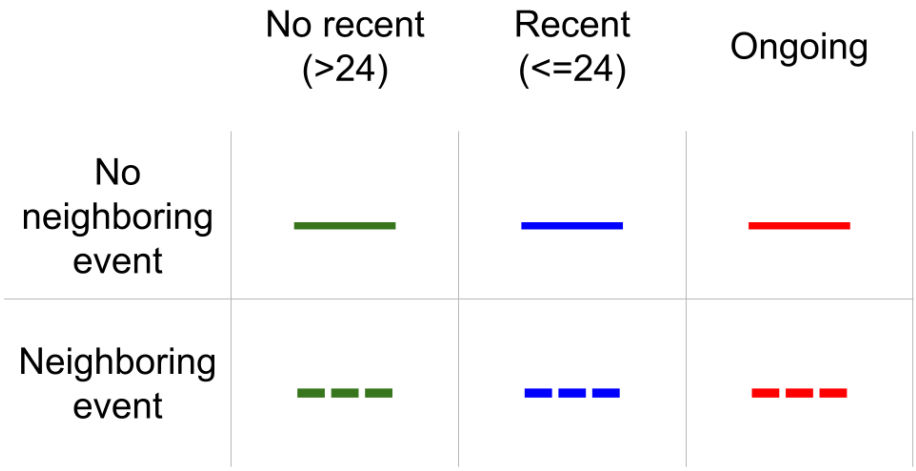


Figure A4: Model performance legend



A.2 Pseudo-out-of-sample procedure

Having provided an overview of the predictors and target variables, we now provide a description of our forecasting methodology. We define forecasting as using all available information at a given time t to make a prediction about a specific outcome in the future. Critically, we utilize pseudo-out-of-sample forecasting as illustrated in Figure A5. In simple terms, this iteratively splits the available data into training and test (i.e. out-of-sample or "unseen" data) sets along the time dimension. We are not interested in performance of the system within the training sample. Modern machine learning methods can easily fit to cases that they see - known as overfitting - but this does not guarantee that observations outside the training sample can be predicted. Instead, the fundamental principle for evaluating how "good" your forecasts are is to check performance on the out-of-sample data.

To elucidate this further, assume we are forecasting any violence in the next 12 months for a particular country. At the end of May 2020, we train our model on data up to and including May 2020 and forecast forward. These predictions can then be held against the *realized* outcome i.e. whether any violence was observed between June 2020 through May 2020. We then do a time-step forward, end of June 2020, pretend we knew only what could have been known by that time, retrain the system, forecast forward and evaluate our performance. We continue this process iteratively up and until the most recent month. This way of training and testing mimics the situation faced by a policymaker at any time and gives us a realistic view of how well it will perform in production.

Figure A5: Visualization of an expanding window forecast. Image credit: Taken from a kaggle competition notebook that credits an Uber blog that has meanwhile been removed.

This is a simplified description of our forecasting methodology. In practice, there is an additional step in this process: hyperparameter tuning. Furthermore, the choice of horizon adds nuance to how we can split the data along the time dimension for training, tuning, testing and evaluating. We start by training from the first available data point. [8] Our test set runs from 2020m1 to present, for which we can only evaluate performance up until the target variable is known. We use the previous four years before the test set to fine-tune the model's hyperparameters.[9]

[8]Note that due to computational constraints, we start feature engineering from 1989m1 for UCDP, but the training set starts in 2010m1. For ACLED we start feature engineering in 2010m1 and the training set starts in 2012m1.

[9]Average precision is the scoring metric chosen for optimizing hyperparameters.