Codebook for www.conflictforecast.org updated on

September 4, 2025

Table of Contents

This codebook provides some technical details on the provided datasets, forecast methods and guidance on how to use/interpret the models we provide on conflictforecast.org. For more details we refer the reader to the academic publications. At the time of writing, Mueller, Rauh and Seimon (2024) provides the most up to date overview of our methodology. However, please note that we are constantly refining our methodology. The latest updates are on our repository on GitHub conflictforecast. For any specific requests or queries please do not hestitate to reach out to us at conflictforecast@gmail.com.

Updates (2024.nov) Onset Model Update: Mixed Approach for Any Violence and Armed Conflict Forecasting

We have updated our approach from using the onset model exclusively to a mixed model strategy. Previously, we applied onset model predictions during both peace and conflict periods. Now, we use the onset model only during peace times, while an incidence model is applied during conflict times.

The onset model is trained using data from periods of peace, whereas the incidence model is trained using data from both peace and conflict periods. This change affects 12 of our 18 models, specifically those related to armed conflict and any violence. It does not affect the models for violence intensity.

This update aims to better capture ongoing conflict dynamics by leveraging the strengths of both models where they are most effective.

1 National predictions

In total we publish predictions from 18 Random Forest models which cover three different target variables, two different horizons and three different feature sets.

Models

Target variables

1. **Any violence:** The target variable is a binary indicator of at least one conflict-related death.
2. **Armed conflict:** The target variable is a binary indicator which meets our definition of armed conflict. We define armed conflict as a per capita measure - 0.5 deaths per 1 million inhabitants in a given month.
3. **Violence intensity:** The target variable is the number of fatalities. More specifically, we predict the log of conflict fatalities plus one.

Horizons

1. **3 months ahead:** Aggregate conflict measure over the entire next 3 months.
2. **12 months ahead:** Aggregate conflict measure over the entire next 12 months.

Feature sets

1. **Text model (text):** Relies on 15 topic generated by an LDA model that analyzes over 5 million news articles. The following table includes the descriptive names we have given to the topics.

[2]:

	Topic Description
0	Competition and Sports
1	Health and Education
2	Military Conflict
3	Politics
4	Military Technology
5	National Development
6	Political Tensions
7	Judiciary and Abuses
8	Middle East
9	Chinese Politics
10	Economics
11	Diplomacy
12	Civilian Life
13	Foreign Policy
14	Power and Negotiation

2. **Historical violence model (hist):** Relies on information about past violence of the country and neighboring countries.
3. **All model (all):** Combines the text features and historical features.

Evaluation

We use ROC-AUC to measure the performance of both any violence and armed conflict models, and the ratio of mean squared error (MSE) relative to the naive model for the intensity model. The naive model predicts the next value as the last observed value.

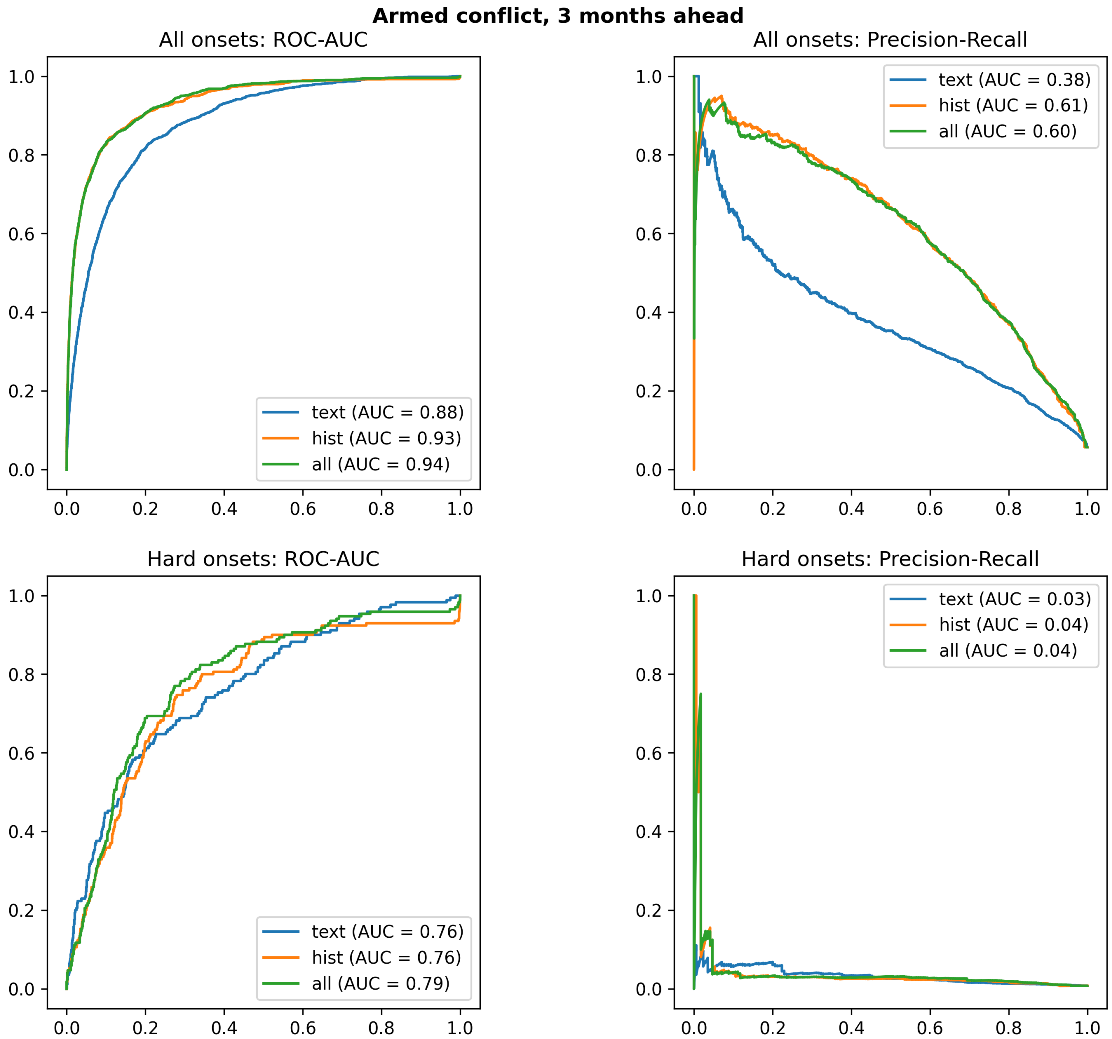
These metrics are calculated for the entire dataset (overall) and specifically for challenging cases (hard). We define challenging cases as country-months where there has been no violence in the last 60 months for the any violence model, no armed conflict in the last 60 months for the armed conflict model, and no violence in the last 12 months for the intensity model.

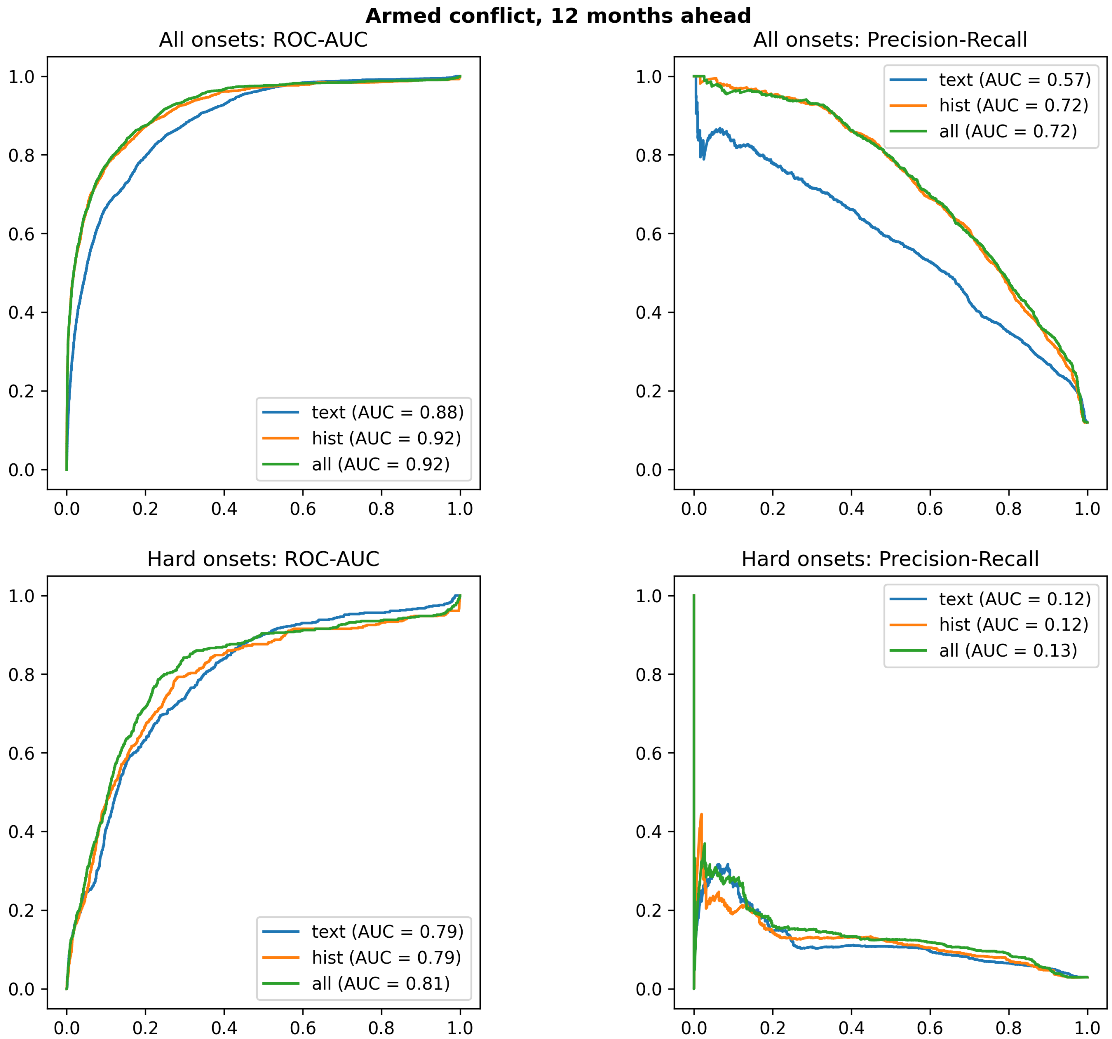
The table presents metrics for each target across three feature sets (all, hist, text) for the current month update using the pipeline on our repository on GitHub conflictforecast. Notice that the models retain forecasting power in the challenging cases. Furthermore, the inclusion of text features improve the performance for hard cases for both the any violence and armed conflict models.

[3]:

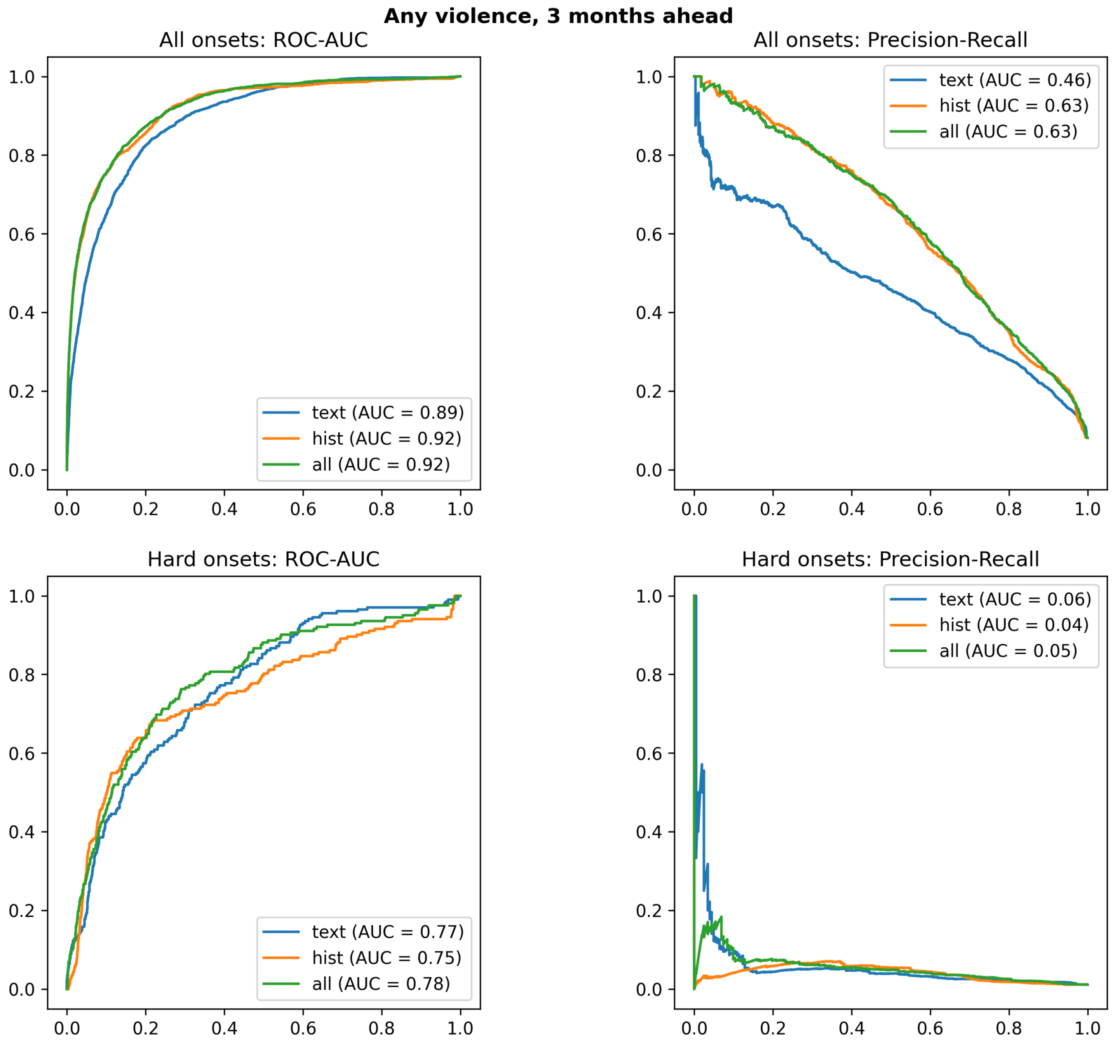
		all	hist	text
target	cases			
ons_armedconf_12	overall	0.92	0.92	0.88
	hard	0.81	0.79	0.79
ons_armedconf_03	overall	0.94	0.93	0.88
	hard	0.79	0.76	0.76
ons_anyviolence_12	overall	0.90	0.90	0.88
	hard	0.79	0.76	0.78
ons_anyviolence_03	overall	0.92	0.92	0.89
	hard	0.78	0.75	0.77
int_lnbest_12	overall	0.81	0.82	1.82
	hard	0.87	0.88	1.93
int_lnbest_03	overall	0.75	0.75	1.46
	hard	0.97	0.99	2.57

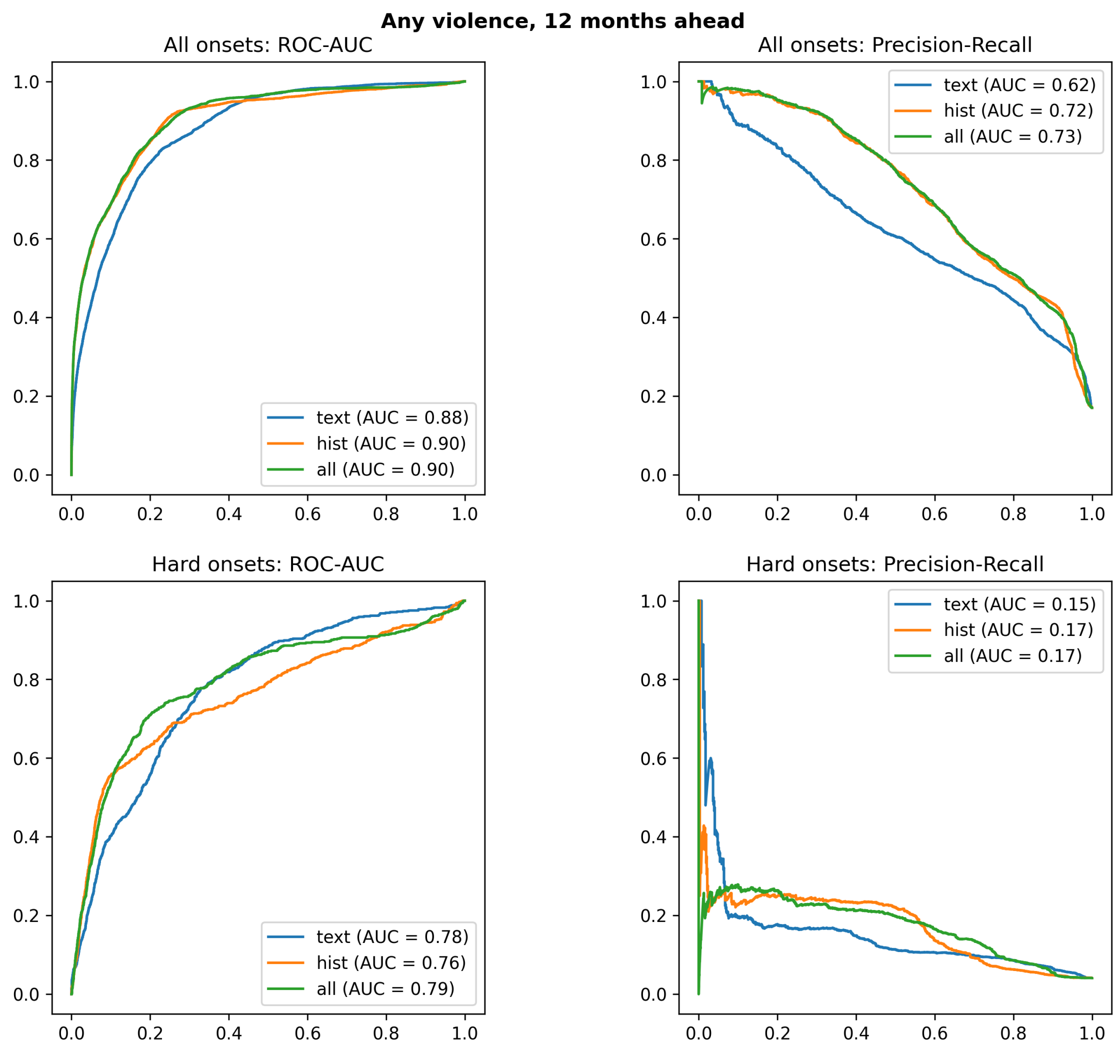
Below we show the performance of the any violence and armed conflict models with respect to ROC-AUC and precision-recall curves, and the intensity model with respect to the mean squared error (MSE). Note that for the violence intensity model plots, the bins are based on recorded UCDP fatalities at time t. The "onset" bin shows performance conditional on 0 fatalities at time t and >0 fatalities in the prediction window.

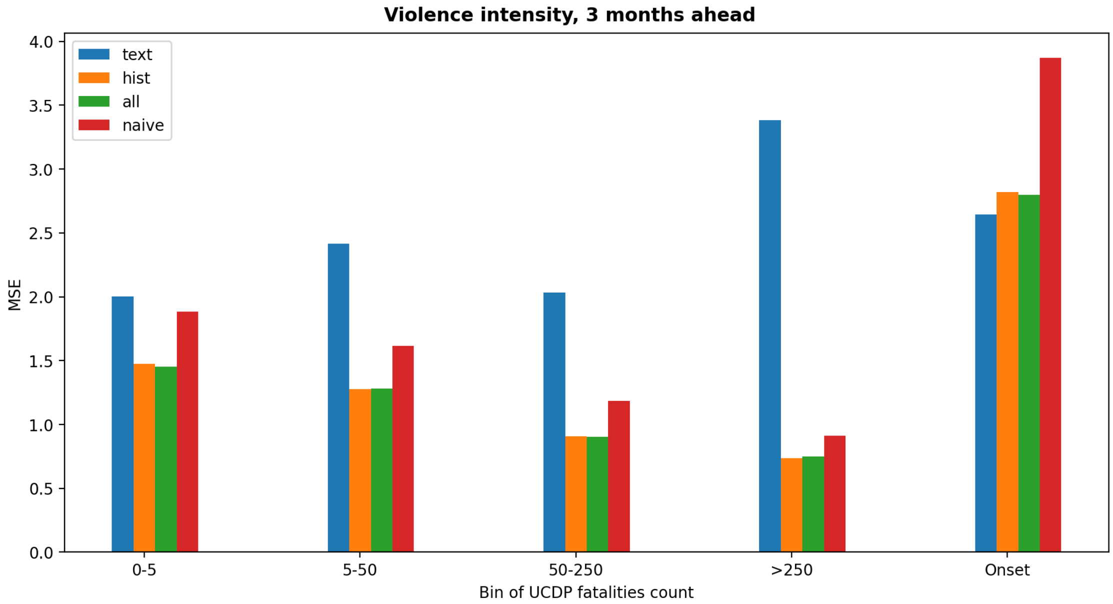

Armed conflict, 3 months ahead
All onsets: ROC-AUC
text (AUC = 0.88)
hist (AUC = 0.93)
all (AUC = 0.94)
All onsets: Precision-Recall
text (AUC = 0.38)
hist (AUC = 0.61)
all (AUC = 0.60)
Hard onsets: ROC-AUC
text (AUC = 0.76)
hist (AUC = 0.76)
all (AUC = 0.79)
Hard onsets: Precision-Recall
text (AUC = 0.03)
hist (AUC = 0.04)
all (AUC = 0.04)

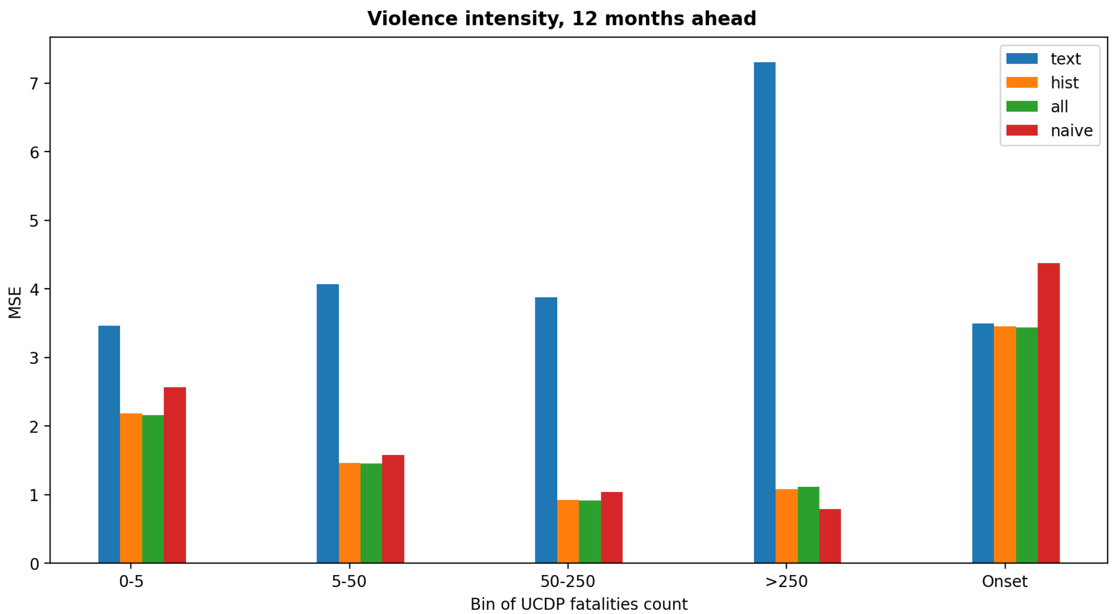


Armed conflict, 12 months ahead
All onsets: ROC-AUC
text (AUC = 0.88)
hist (AUC = 0.92)
all (AUC = 0.92)
All onsets: Precision-Recall
text (AUC = 0.57)
hist (AUC = 0.72)
all (AUC = 0.72)
Hard onsets: ROC-AUC
text (AUC = 0.79)
hist (AUC = 0.79)
all (AUC = 0.81)
Hard onsets: Precision-Recall
text (AUC = 0.12)
hist (AUC = 0.12)
all (AUC = 0.13)
0.0
0.2
0.4
0.6
0.8
1.0


Any violence, 3 months ahead
All onsets: ROC-AUC
text (AUC = 0.89)
hist (AUC = 0.92)
all (AUC = 0.92)
All onsets: Precision-Recall
text (AUC = 0.46)
hist (AUC = 0.63)
all (AUC = 0.63)
Hard onsets: ROC-AUC
text (AUC = 0.77)
hist (AUC = 0.75)
all (AUC = 0.78)
Hard onsets: Precision-Recall
text (AUC = 0.06)
hist (AUC = 0.04)
all (AUC = 0.05)
0.0
0.2
0.4
0.6
0.8
1.0


Any violence, 12 months ahead
All onsets: ROC-AUC
text (AUC = 0.88)
hist (AUC = 0.90)
all (AUC = 0.90)
All onsets: Precision-Recall
text (AUC = 0.62)
hist (AUC = 0.72)
all (AUC = 0.73)
Hard onsets: ROC-AUC
text (AUC = 0.78)
hist (AUC = 0.76)
all (AUC = 0.79)
Hard onsets: Precision-Recall
text (AUC = 0.15)
hist (AUC = 0.17)
all (AUC = 0.17)





2 Which national model should you use?

For countries with extended periods of peace, the any violence and armed conflict forecasts provide an indication of the possible risk of an outbreak. However, this is less valuable for countries currently in violence or stuck in the conflict trap. In these situations, the violence intensity forecast can provide an indication of future escalations/de-escalations.

Our default model is the model using all features, for armed conflict, 12 months ahead. This model predicts the outbreak of violence that exceeds 0.5 fatalities per 1 million inhabitants of a country in a given month. These are serious outbreaks of internal political violence for the respective country relative to their population. For countries like India or China this risk is lower due to their large population size.

Any violence is any fatality due to political conflict and here population does not play a role in the estimation of the risk. As the threshold here is lower, risk estimates are much higher for this type of conflict. The absolute threshold also makes more populous countries relatively riskier.

The 12 months ahead model uses a longer period to train the random forest. This means the model can pick up subtler signals but it also means that risks are more disperse. The 3 months ahead model will pick up more escalations of lower level violence like riots and activity by armed actors.

The question on which model to use therefore depends on the goal of using the forecast. For a measure of general political risks that might destabilize an entire country use the armed conflict forecast, 12 months ahead. For forecasts of violence itself use the any violence forecasts.

The all model is the better forecast overall as it uses conflict history in its forecast. However, this also means that the forecast is dominated by conflict dynamics and history. Countries which suffered recent, intense violence will be particularly risky. This means that some new developments will be forecasted slightly less precisely. The text model is trained with text alone to predict and it will, in many circumstances, even beat the best model in picking up new developments in a country.

We recommend to never use the text model when predicting intensity. The best model performs much better in most circumstances in this model. We do not recommend the use of the intensity model in countries which have been at peace for several years.

3 Subnational predictions

Models

The subnational level forecasts are still in their beta version and are not derived from published academic research. In their current form, the prediction models are an extension of the national predictions and were developed as part of a project for the Foreign, Commonwealth and Development Office of the UK.

In terms of target variables and horizons, we only provide an any violence forecast for a horizon period of 12 months. That is, we predict the likelihood of any battle death in a grid cell within the next 12 months.

Evaluation

We will soon add an overview of performance at the subnational level.

4 Method details

Although we are continuously experimenting with different models and ensembles, we typically rely on a Random Forest and implement a rolling forecast methodology. The purpose of the rolling forecast is to replicate the information set that would be available to a decision maker. In other words, they would observe features until period and make a forecast for the aggregate window

< + where is equal to 3 or 12. In other words, when the window is 3 months, then we are predicting, for instance, the likelihood of any battle death over the entirety of the 3 months. We do not predict outcomes for each of the 3 months separately. Similarly, when predicting the 12 month window, we consider the aggregate outcome over the next 12 months, and not for each of the next 12 months separately. For example, a armed conflict forecast in 2015m1 for a window of 12 months is predicting the likelihood of an armed conflict outbreak in any of the following 12 months. The same applies to the intensity forecast - for a forecasting window of 12 months, we are predicting the average number of fatalities per month over the next 12 months.

When training on past data for performance measurement we do not use information on conflicts that have not broken out but will break out. For example, when predicting conflict from a standpoint of information in June 2020 we do not use the fact that we know a conflict broke out in July 2020. We do this by setting existing positive training observations to missing in the training sample when the conflict has not broken out yet. This means that the forecast in June 2020 cannot rely on the observation in May 2020 for that country. We also set observations to missing for which it is not clear, yet, whether a conflict will break out. We are also not evaluating our model on these observations. For example, in May 2021 we will not evaluate performance on observations in the period April 2020 to May 2021 when forecasting 12 months ahead. We will not evaluate performance on observations in the period February 2021 to May 2021.

We train models to learn a functional form using all data from 1989m1 to 2009m12. With the resulting model, we then produce out of sample predictions on a rolling basis from 2010m1 onwards. For any violence and armed conflict, hyperparameters are chosen by maximising the area under the curve (AUC) of the receiver operating characteristics (ROCs) curve. This applies to both the national and subnational models. In the case of violence intensity, we seek to minimise the mean squared error (MSE).

In our provision of risk statistics, we do not distinguish between countries which are in conflict from those that are not. And therefore countries with the highest risk estimates will mostly already suffer fatalities from armed conflict according to the UCDP.

Also note that we update our model twice a month. On the 1st of each month, we aim to collect the text information as quickly as possible from the previous month and update the _text model. For the _all and _hist models, we need a valid conflict history of fatalities to provide an update. This information on fatalities is provided by UCDP around the 20th of next month. For example, on the 20th of February 2021 the UCDP released its estimates of fatalities during January 2021. Therefore, we aim to publish the second update around the 22nd of each month, using the full information set on conflict history and text up until the end of the previous month.

5 Code disclaimers

Sep 2025: The September update, which incorporates news articles from August 2025, was produced without new articles from Latin News, as access to this source was temporarily unavailable.

Jan 2025: Starting with the January 2025 update, new New York Times articles are no longer included in our text corpus, following changes in data availability from our provider. Articles included prior to this date remain part of the corpus.

Feb 2024: Please note we underwent a pipeline transition in February 2024 for the national models. Table 1 documents changes to the variable naming conventions. Prior to Feb 2024 download files will correspond with “old_name”, and after that they will correspond with “name”.

Dec 2022: We completely overhauled our method and switched text data source from LexisNexis to Factiva. We reinsured that forecasts are robust to this and performance is extremely similar in both the best and text model.

May 2021: We spotted a minor coding error in our previous training and evaluation method – this made previous estimates and evaluations less valid for out-of-sample performance.

6 Academic publications

Mueller, H., Rauh, C., Seimon, B. (2024) Introducing a Global Dataset on Conflict Forecasts and News Topics. Forthcoming Data & Policy. https://www.econ.cam.ac.uk/research/cwpe-abstracts?cwpe=2404

Mueller, H., & Rauh, C. (2022). The hard problem of prediction for conflict prevention. Journal of the European Economic Association, 20(6), 2440-2467. https://academic.oup.com/jeea/article/20/6/2440/6574413

Mueller, H., & Rauh, C. (2022). Using past violence and current news to predict changes in violence. International Interactions, 48(4), 579-596. https://www.tandfonline.com/doi/full/10.1080/03050629.2022.2063853?casa_token=h5hIkuCJq7UAAAAA%3AI mQoYDo6BAkU1jmIE4Afq0JG-1geXeJLIZ-x-w

Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review, 112(2), 358-375. https://www.cambridge.org/core/journals/american-political-science-review/article/abs/reading-between-the-lines-prediction-of-political-violence-using-newspaper-text/4EABB473AFE18F157EEDE4339F34ABB0

7 Appendix

Table 1: National predictions dataset descriptor

name	old_name	description
isocode	isocode	Three-letter country codes defined in ISO 3166-1.
period	year-month	Forecast date represented as YYYYMM, based on news articles available until the last day of the same YYYYMM.
fatalities_ucdp	best	Best estimate of fatalities by UCDP. Includes state-based, non-state and one-sided battle-related deaths.

name	old_name	description
population	populationwb	Annual population data sourced from the World Bank, with monthly values interpolated and extrapolated from July 2022 onward.
anyviolence	anyviolence	1 if there were any fatalities in the country during the month, and 0 otherwise.
ons_anyviolence_03_target	ons_anyviolence3	1 if a country experiences an increase in fatalities from zero to any positive number in the next three months, 0 if it maintains zero fatalities, and missing otherwise.
ons_anyviolence_03_text	text_model	Estimated probability of any violence in the next three months using only population and text-related variables.
ons_anyviolence_03_hist	nan	Estimated probability of any violence in the next three months using only population and conflict dynamic variables.
ons_anyviolence_03_all	best_model	Estimated probability of any violence in the next three months using population, text-related variables and conflict dynamic variables.
ons_anyviolence_03_naive	nan	1 if there were any fatalities in the past three months, and 0 otherwise.
ons_anyviolence_12_target	ons_anyviolence12	1 if a country experiences an increase in fatalities from zero to any positive number in the next twelve months, 0 if it maintains zero fatalities, and missing otherwise.
ons_anyviolence_12_text	text_model	Estimated probability of any violence in the next twelve months using only population and text-related variables.
ons_anyviolence_12_hist	nan	Estimated probability of any violence in the next twelve months using only population and conflict dynamic variables.

name	old_name	description
ons_anyviolence_12_all	best_model	Estimated probability of any violence in the next twelve months using population, text-related variables and conflict dynamic variables.
ons_anyviolence_12_naive	nan	1 if there were any fatalities in the past twelve months, and 0 otherwise.
armedconf	armedconf	1 if the fatalities per million inhabitants in the country during the month exceed 0.5, and 0 otherwise.
ons_armedconf_03_target	ons_armedconf3	1 if a country transitions from no armed conflict to armed conflict in the next three months, 0 if it remains armed conflict-free, and missing otherwise.
ons_armedconf_03_text	text_model	Estimated probability of armed conflict in the next three months using only population and text-related variables.
ons_armedconf_03_hist	nan	Estimated probability of armed conflict in the next three months using only population and conflict dynamic variables.
ons_armedconf_03_all	best_model	Estimated probability of armed conflict in the next three months using population, text-related variables and conflict dynamic variables.
ons_armedconf_03_naive	nan	1 if there was no armed conflict in the past three months, and 0 otherwise.
ons_armedconf_12_target	ons_armedconf12	1 if a country transitions from no armed conflict to armed conflict in the next twelve months, 0 if it remains armed conflict-free, and missing otherwise.

name	old_name	description
ons_armedconf_12_text	text_model	Estimated probability of armed conflict in the next twelve months using only population and text-related variables.
ons_armedconf_12_hist	nan	Estimated probability of armed conflict in the next twelve months using only population and conflict dynamic variables.
ons_armedconf_12_all	best_model	Estimated probability of armed conflict in the next twelve months using population, text-related variables and conflict dynamic variables.
ons_armedconf_12_naive	nan	1 if there was no armed conflict in the past twelve months, and 0 otherwise.
lnbest	nan	Best estimate of fatailites by UCDP, transformed using log(fatalities + 1).
int_lnbest_03_target	nan	Sum of fatalities for the next three months, in log(x+1).
int_lnbest_03_text	nan	Estimated sum of fatalities for the next three months using only population and text-related variables, in log(x+1).
int_lnbest_03_hist	nan	Estimated sum of fatalities for the next three months using only population and conflict dynamic variables, in log(x+1).
int_lnbest_03_all	nan	Estimated sum of fatalities for the next three months using population, text-related variables and conflict dynamic variables, in log(x+1).
int_lnbest_03_naive	nan	Sum of fatalities for the past three months, in log(x+1).
int_lnbest_12	nan	Sum of fatalities for the next twelve months, in log(x+1).

name	old_name	description
int_lnbest_12_text	nan	Estimated sum of fatalities for the next twelve months using only population and text-related variables, in log(x+1).
int_lnbest_12_hist	nan	Estimated sum of fatalities for the next twelve months using only population and conflict dynamic variables, in log(x+1).
int_lnbest_12_all	nan	Estimated sum of fatalities for the next twelve months using population, text-related variables and conflict dynamic variables, in log(x+1).
int_lnbest_12_naive	nan	Sum of fatalities for the past twelve months, in log(x+1).
tokens	nan	Total number of unique tokens across all news articles within a given month for a specific country.
stock_tokens	nan	Cumulative sum of tokens, with a monthly decay rate of 0.8.
stock_topic_0	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_1	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_2	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_3	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.

name	old_name	description
stock_topic_4	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_5	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_6	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_7	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_8	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_9	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_10	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_11	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_12	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.

name	old_name	description
stock_topic_13	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
stock_topic_14	nan	Cumulative sum of average topic share across all news articles withing a given month for a specific country, with a monthly decay rate of 0.8.
discounted_best	nan	Cumulative sum of fatalities, with a monthly decay rate of 0.95.
discounted_anyviolence	nan	Cumulative sum of anyviolence variable, with a monthly decay rate of 0.95.
discounted_armedconf	nan	Cumulative sum of armedconf variable, with a monthly decay rate of 0.95.
neighbors_best	nan	Average number of fatalities among the neighboring countries for the current month.
neighbors_anyviolence	nan	Share of neighboring countries experiencing any violence in the current month.
neighbors_armedconf	nan	Share of neighboring countries experiencing armed conflict in the current month.
ongoing_anyviolence	nan	Number of months of ongoing violence.
ongoing_armedconf	nan	Number of months of ongoing armed conflict.
ongoing_civilwar	nan	Number of months of ongoing civil war.
since_anyviolence	anyviolence_dp	Number of months since the last month with any violence.
since_armedconf	armedconf_dp	Number of months since the last month with armed conflict.
since_civilwar	nan	Number of months since the last month with civil war.
past_bestpc_6	nan	Sum of fatalities per 1000 inhabitants over the past 6 months.

name	old_name	description
past_bestpc_12	nan	Sum of fatalities per 1000 inhabitants over the past 12 months.
past_bestpc_60	nan	Sum of fatalities per 1000 inhabitants over the past 60 months.
past_bestpc_120	nan	Sum of fatalities per 1000 inhabitants over the past 120 months.

nan indicates this variable was not previously provided

Table 2: National model feature sets

We will soon add a table describing the specific features used for the respective national models.

Table 3: Subnational predictions dataset descriptor

Variable	Definition
gid	Unique PRIO grid cell identifier.
isocode	Three-letter country codes defined in ISO 3166-1.
countryname	Full country name.
year	Observation year.
month	Observation month.
risk	Estimated probability of any violence in the next twelve months using population, geographic, text-related and conflict dynamic variables.
Longitude	East-west position on Earth's surface.
Latitiude	North-south position on Earth's surface.

Table 4: Subnational model feature sets

We will soon add a table describing the specific features used for the subnational model.