Explainable Probabilistic Forecasting of Conflict-Related Fatalities

Oleksandr Zakotianskyi

Vrije Universiteit Amsterdam, Netherlands
o.zakotianskyi@student.vu.nl

Abstract. Can we predict wars? How certain would we be in our predictions? This research presents the first publicly available and explainable early conflict forecasting model capable of forecasting distribution of conflict-related fatalities on a country-month level. The model seeks to be maximally transparent, uses publicly available data and produces predictions up to 14 months into the future. Our model improves over 3 out of 4 benchmark years but misses the violence spikes, possibly, due to the nature of independent variables in the dataset. The presented model can complement the field of conflict-warning systems and serve as a reference against which future improvements can be evaluated.

Keywords: Interstate conflict modelling · Early Conflict Warning System · Fatalities prediction · Predicting with uncertainty

1 Introduction

A new wave of violent conflicts around the world raises concerns about the security of the whole world. According to the ACLED[1] Conflict Index 12% more conflicts occurred in 2023 compared to 2022, and the amount of conflict increased by 40% compared to 2020. An underlying assumption and motivation of conflict early warning systems is that enhanced prediction and forecasting can better inform decision-making, reduce risk, and trigger more robust prevention and response measures from international actors. Policymakers are interested in uncovering the conflict dynamics and making robust forecasts to take anticipatory actions and reduce the impact on vulnerable people. At the same time, regular citizens are looking for accurate and not biased forecasts for their own security and awareness.

A decade ago inspiring studies of Hedge et al. [1] and Chadefaux [2] marked a path towards a new generation of models - production-ready Early Conflict Warning Systems. By now at least a couple of dozen Early Conflict Warning Systems are deployed and to some degree are used for decision-making [3, 4]. Great progress in this development was thanks to the first prediction competition organized by the ViEWS[2] academic consortium in 2020. The competition focused

[1] Armed Conflict Location and Event Data
[2] The Violence & Impacts Early-Warning System

on increasing the accuracy of the state-of-the-art conflict prediction models and resulted in significantly pushing state-of-the-art. The best model for prediction on a country-month level appeared to be an ensemble of tree-based models, but also other innovative approaches were elaborated upon such as models based on topic modelling and alternative data sources.

When the improved models were deployed and feedback from stakeholders was gathered, it became apparent that additional improvements were needed. According to the feedback, the stakeholders are interested not only in the most likely outcome (a point prediction), but also in the lower-probability risk that conflicts escalate catastrophically (the tail ends of the probability distribution), and understanding how uncertain the forecasts at hand are. These preferences urge for a new generation of forecasting models that present their estimates as probability distributions.

To address this gap, the ViEWS team has announced the second prediction competition, now focusing on predicting the distribution of conflict-related fatalities instead of simple point forecasts.

In response, this paper presents an open-source model for forecasting the probability distribution of fatalities on a country-month level. This leads to the central research question of this study: **How can we develop a model that accurately predicts conflict-related fatalities while also estimating the uncertainty of these predictions?**

The model presented in this paper builds upon the Natural Gradient Boosting framework by Duan et al. [5] which allows for estimating conditional probability distribution for each separate prediction.

The accuracy of the developed model is on par with traditional gradient boosting techniques and improves over the heuristic benchmarks developed by the competition organisers for 3 out of 4 prediction years.

2 Related Work

The advancements in conflict research went in waves that started as early as the 1970s with pioneering studies of Andriole & Young[6]. However, only the most recent advancements have brought Conflict Early Warning Systems closer to a production-ready state. A great overview of currently deployed systems was done by Rod et al. [3] and Muggah & Whitlock [4].

The efforts in conflict forecasting were historically divided into two categories: theoretically supported models that utilise structural variables[3] and semantic analysis models that rely on news and political speech. Theoretically-supported models rely on econometric data and the political background of states and regions to predict the conflict propensity and future risk levels of states[7, 8]. Semantic analysis models focus on political event data as a basis for prediction by using media content and verbal statements of political actors to forecast the

[3] Examples of structural variables are economic and political features such as development indexes, GDP, education, youth bulge, life expectancy, etc.

impending escalation of conflicts in regions and states [2, 9]. Such models usually take as input news, social media feeds or official political speech.

It was shown that theoretically supported models perform well in assessing the country's risk of engaging in conflict. Structural variables can explain well an ongoing deterioration of a state or, otherwise, its successes through economic, political and social variables. Such variables explain the capabilities of the country, and its weaknesses and thus can be well mapped to an increased or decreased risk of conflict. On the other hand, structural variables are rarely updated and are very slowly moving, thus often failing to forecast conflict escalations.

Another type of models, that was made possible with the recent advancements in natural language processing, is the semantic analysis models. Their strong side is that they can spot rapid changes in people's moods as this is highly reflected in the news and social media. Such models may be insightful for estimating the temporal dimension of violence.

At the same time, both types of state-of-the-art models produce forecasts as the most likely fatality number, fatality bin or risk index. As mentioned earlier, stakeholders of the early conflict warning models wish to know not only the most likely risk but rather its probability distribution. Such probability distribution of the risk will allow stakeholders to prepare or respond depending on the changes in the distribution over time. While there are multiple ways to represent a risk of conflict, one of the most granular and easily interpretable is the amount of fatalities due to conflict in this country.

The closest model that focuses on predicting conflict-related fatalities is the model developed by Hegre et al. in the scope of the ViEWS project [10]. The ViEWS Fatalities model predicts state-based fatalities on a monthly level with global coverage. The latest release in early 2023 [11] presented an unweighted ensemble of 21 models for making forecasts up to 36 months in advance. ViEWS combines six types of models: Random forests, Gradient boosting models, Extreme gradient boosting, Light gradient boosting, Hurdle models and Markov models. While the ViEWS model is a great work and one of the most advanced publicly available models, it produces the most likely point predictions without a probability distribution.

To properly position the proposed model in this study, it is important to give an overview of the recent impactful models in the field.

In recent years, researchers proposed multiple models of predicting military conflict on national and sub-national levels with datasets consisting of economic and political variables [1, 7, 12–14] As a result, conflict literature has made significant progress in understanding which countries are at risk of suffering an armed conflict, but usually fails to accurately forecast conflict onsets in both spatial and temporal dimensions. One of the drivers that allowed not only to improve the accuracy of the models over time but also to understand the predictions is the advancement in models' explainability. The end goal of the early conflict warning system is to allow stakeholders to consult them and make decisions based on their outputs. The model needs to be explainable to validate that, for example, a high-risk prediction is not influenced by some misalignment in the input

data. For instance, an inspiring study by Baillie et al. features a model using Motre Carlo simulations to forecast conflict onset on a country-year level with only three variables in the dataset [15]. Baillie argues that an explainable and interpretable model is more likely to be used by policymakers. However, research by Ettensperger shows that complex and thus unexplainable model ensembles tend to be more accurate [16], which urges for a balance between explainability and predictive accuracy.

Additionally, to increase the accuracy of predicting conflict onsets scholars explored additional data sources. Chadefux proposed a conflict risk index that is based solely on news and is meant to indicate a risk of a conflict onset [2]. Muller & Rauh presented a latent Dirichlet allocation model for content analysis of 700 thousand English-speaking news [9] and later improved the model to include 3.5 million news and past violence data [17]. However, Bazzi et al. showed that even access to rich historical and textual data allows for accurate forecasting of spatial but not the temporal dimension of the conflict [18]. Moreover, Halkia et al. found multiple problems with textual data such as bias, duplication of some events and under-representation of others (especially in countries with severe censorship) [19].

Because of the limitations of models described above, scholars turned to alternative data sources such as night lights data [20], stock market behaviour [21], cell phone call patterns [22], climate-sensitive models [23, 24], local peacekeeping data [25] and remote sensing data [26]. Racek et al. attempted to forecast the Syrian civil war onset based on remote sensing data only [27]. Their remote sensing dataset comprised data on population, land-cover classes, nighttime lights, topography, vegetation health, crops, precipitation and temperature and was found to lead to a 1.75% predictive performance gain.

Concluding, multiple methodologies were developed in the field of conflict modelling. The state-of-the-art models still need to increase the predictive accuracy to robustly predict outbreaks of violence in previously peaceful countries and enable more informative forecasts for stakeholders, such as probability distribution forecasts. Additionally, the balance between accuracy and explainability needs to be maintained, as the stakeholders are more likely to rely on explainable and interpretable models.

3 Summary of contributions

As a first step towards predictions of fatalities distribution, this study focuses on enabling an explainable model that relies on structural variables to produce probabilistic forecasts on a country-month level.

1. We present a model for probabilistic forecasting conflict-related fatalities based on the Natural Gradient Boosting framework (Section 4.4 for framework overview).
2. We clean and improve the competition dataset by removing rows with too many missing values and adding new dimensions of data, which results in

increased performance (Section 4.3 for data preprocessing and Section 5 for performance overview).

4 Methodology

This section outlines the methodologies employed to develop a predictive model for conflict-related fatalities at the country-month level. The methodology is divided into key sub-sections: an explanation of the level of analysis the model assumes and the prediction window (Section 4.1), an overview of the original ViEWS competition dataset (Section 4.2), data preprocessing pipeline (Section 4.3), the employed model (Section 4.4), scoring criteria (Section 4.5), model finetuning (Section 4.6) and overview of the heuristic competition benchmarks (Section 4.7).

4.1 Level of analysis and prediction window

We build a model that generates forecasts at the country-month level of analysis. The country-month level is useful for predictions of new conflicts and for modelling processes at the government level. The set of countries is initially defined according to Gleditsch&Ward (GW) country definitions [28], but later is mapped to Correlates of War (CoW) state membership system [29] (see Section 4.3).

The competition rules define that October should be the last data point to forecast the next calendar year. This implies that the model should make predictions for all months from January to December of the next year based on the data till October of the previous year. The proposed model uses one-step ahead modelling for each of the 12 months in the test set, where each month in the test set predicts 14 months ahead for each country. For example, if the model produces forecasts for the whole year of 2022 the test set spans from November 2020 to October 2021, and the train set includes every month until September 2022 inclusive. The dependent variable shifting is discussed in detail in Section 4.3.

4.2 Original Competition Dataset

In this study, we use the dataset published by the ViEWS team for the second prediction competition [30]. The published dataset merges multiple open-source datasets and adds spatial and temporal lags and decays for several features.

The dataset covers 382 months from Jan 1990 to Oct 2021 for 213 unique country IDs (defined according to the Gleditsch & Ward system) and comprises 128 columns. The dataset includes data on fatalities per country per month reported by the Uppsala Conflict Data Program (UCDP)[4] Georeferenced Events Dataset [31, 32], the World Development Indicators [33], data on politically excluded ethnic groups [34], demographic factors [35], protests from ACLED [36],

[4] Uppsala Conflict Data Program

data on institutions from V-Dem [37], and data on national water resources from AQUASTAT [38]. Additionally, the ViEWS team engineer temporal and spatial lag features with decays to enhance the predictive power of the dataset [39].

The dataset contains features for three types of violence coded by UCDP[40]: Uppsala Conflict Data Program distinguishes three types of violence according to the definitions by Melander et al. [40]: state-based conflict[5], one-sided violence against civilians[6], and non-state conflict[7]

As for the dependent variable, we follow the ViEWS competition guidelines and choose state-based conflict. Features for the other two types of violence features are kept in the dataset as independent variables. It was shown by Hegre et al. that a large share of one-sided violence against civilians and non-state conflict events are outcomes of state-based conflicts [1]. Much of the violence against civilians is perpetrated by governments and rebel groups to weaken opponents, and much non-state conflict is infighting between rebel groups that also conflict with the government.

The distribution of the state-based fatalities is shown in Figure 1. The dependent variable is highly skewed to the right (with a skewness of 17) and has a long tail on the right side (with a kurtosis of 7810). We see that most of the countries have near-zero numbers of fatalities with rare spikes for individual months. Interestingly, the whole dataset of 71,642 rows contains only 381 country-months with more than 100 fatalities, 199 observations with more than 1,000 fatalities and only 11 observations with more than 10,000 fatalities.

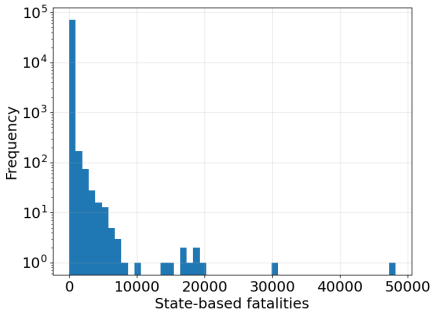


Fig. 1: Distributions of state-based fatalities. The dependent variable is highly skewed to the right with a long tail of rare violence spikes.

[5] A state-based armed conflict is a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state.

[6] The deliberate use of armed force by the government of a state or by a formally organised group against civilians.

[7] The use of armed force between two organised armed groups, neither of which is the government of a state.

Moreover, some variables are highly correlated. The correlation of more than |0.75| occurs for 78 out of 125 numeric variables, mostly for features from V-Dem, World Development Indicators and some spatial decays. The full correlation matrix is shown in the Appendix section A.1. These findings urge for data preprocessing, which is discussed in the next section.

4.3 Data preprocessing

In this section, we discuss data cleansing of the original ViEWS dataset, the creation of a dependent variable for the prediction task and the addition of the region dimension to the dataset for enhanced region dynamics modelling.

Data cleaning We find several problems with the original dataset, namely:

- There are cases where the same country has multiple country IDs in the same month and some variables for one of the instances are not populated.
- Some countries have too many missing structural variables. There are countries with almost all independent variables set to 0. This usually happens for low-populated islands, micro-states, or countries that do not exist today and lack reliable data.

To address these issues we calculate the percentage of zero values for each country-month and the mean percentage per country. To account for the fact that some features are spatial or temporal lags of the other, we exclude from the analysis all spatial and temporal lag features. Additionally, as previous violence is a good prediction of future violence and the model should be able to learn not only fluctuations of the dependent variable but also infer future violence from the country's indexes, we exclude from the analysis all features that are related to all three types of violence (see Section 4.2) and ACLED data on protests. After the column exclusions, 89 data columns are left for analysis.

The percentage of missing values per country is defined as an average proportion of zero values in a given subset of columns for this country and calculated as follows:

$$\frac{\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{N}\mathbb{1}(X_{im})}{N}\right)}{M}\times 100$$

where

- M is the amount months for an analysed country ID.
- N is the number of columns to analyse.
- $\mathbb{1}(X_{im})$ is the indicator function for the i-th (feature) column in the m-th (country-month) row, defined as:

$$\mathbb{1}(X_{im}) = \begin{cases} 1 & \text{if the value in column } i \text{ of row } m \text{ equals } 0 \\ 0 & \text{otherwise} \end{cases}$$

With the percentage of missing values calculated, the dataset is processed in two steps:

1. **Correlates-of-War mapping**: An alternative to Gleditsch&Ward State System Membership exists and was published by the Correlates of War (CoW) project. The major difference between these systems is that CoW does not include states with populations lower than 250,000, which automatically excludes all micro-states. Microstates are insignificant actors in a geopolitical landscape for which data collection is usually not done properly and major datasets do not consider these countries. By manual inspection, we found that all states from the original dataset that cannot be mapped to the CoW system are all microstates, with more than 70% of missing values. Thus, these countries are dropped from the dataset.
2. **Dropping countries with more than 20% of zero values**: Many studies find that the previous violence is a good predictor of future violence [8, 10, 18, 41, 42]. Thus, keeping countries with only populated fatality variables with their spatial and temporal lags and the majority of other structural variables being 0 would only incentivise the model to ignore the country indexes and focus on learning patterns of previous violence. To minimise the chances of such learning, we drop all countries which have a percentage of missing values of more than 20%. The calculation results of the mean missing values per country are presented in Table 1. One can see that the majority of countries with a mean zero values percentage of more than 20% are micro-states, short-lived countries or countries for which data collection was not reliable.

The resulting dataset comprises 177 unique country IDs under the Gleditsch&Ward state membership system and 169 unique country IDs under the CoW state membership system. This difference comes from the fact that for some states that experienced revolutions and changes in government structure (e.g. Sudan in July 2011, Indonesia in May 2002 or multiple instances of Russia in 1991), the Gleditsch&Ward system defines a different country ID, while CoW keeps the same.

As the last step, the country IDs are encoded using the dummy encoding scheme [43] according to the Gleditsch&Ward system.

Dependent variable shifting As explained in Section 4.1, the dependent variable is generated by shifting the state-based fatalities column by 14 months back to create a dependency suitable for forecasting. Because the Gleditsch&Ward country membership system changes country ID every time the government structure changes, the dependent variable is shifted based on CoW country ID, which does not change in such cases. As ViEWS competition suggests, the country-months dataset is merged with country-month 'actuals' for the predicted year, which creates an expected 2-month (November to January) gap between the test set and the predicted year and as a result a gap between the train set and the test set. The visualization of the split is provided in Figure 2.

Country ID	Country name	Percentage of missing values (%)	Min date	Max date	Max fatalities
187.000	Czechoslovakia	100.00	1990-01	1992-12	0
189.000	Russia (Soviet Union)	100.00	1990-01	1991-08	144
232.000	Kosovo	100.00	2008-03	2021-10	0
247.000	Yugoslavia	100.00	1991-11	1992-04	1284
248.000	Russia (Soviet Union)	100.00	1991-08	1991-08	0
250.000	Russia (Soviet Union)	100.00	1991-09	1991-09	0
252.000	Russia (Soviet Union)	100.00	1991-10	1991-10	0
253.000	Russia (Soviet Union)	100.00	1991-11	1991-11	1
254.000	Russia (Soviet Union)	100.00	1991-12	1991-12	0
188.000	Yugoslavia	98.88	1990-01	1991-11	351
227.000	Yugoslavia	97.75	1992-04	2006-06	560
26.0000	Bahamas	70.85	1990-01	2021-10	0
140.000	Brunei	60.67	1990-01	2021-10	0
27.0000	Belize	58.80	1990-01	2021-10	0
186.000	German Democratic Republic	49.44	1990-01	1990-10	0
197.000	Yemen, People's Republic	49.44	1990-01	1990-05	0
198.000	Taiwan	44.20	1990-01	2021-10	0
196.000	Yemen, Arab Republic	25.84	1990-01	1990-05	0
185.000	German Federal Republic	24.72	1990-01	1990-10	1
192.000	South Africa	19.10	1990-01	1990-03	0
230.000	Serbia	17.98	2006-06	2008-02	0
191.000	Ethiopia	14.14	1990-01	1993-05	16545
180.000	Solomon Islands	13.48	1990-01	2021-10	2
83.0000	Bosnia-Herzegovina	10.71	1992-04	2021-10	4423
....	....	....	....	....	....

Table 1: Mean percentage of missing values for Gleditsch-Ward Country IDs in the original dataset after dropping countries that could not be mapped to the CoW state membership system. All countries with more than 20% of missing values (above the separator line) are dropped.

Regions addition According to multiple studies in the peace research community, adding a region component to the dataset helps in increased accuracy or explaining variability in modelling conflict [1, 44, 45]. The intuition behind this finding is that geographical regions share similar risk factors that drive conflict and define propensities for instability. Goldstone et al. in his early research notes five regions that account for similar "regional and temporal distributions" in both the train and test datasets [8] and Boekestein reports an increase of 2-7% in prediction accuracy with added regions [45].

Hegre et al. presented a model that included regions as independent variables [1]. Contrary to Goldstone, Hegre defined nine regions according to the United Nations regional definitions. He posited that the region variable improves the quality of predictions by maximizing the explained variance in the dataset. Still, as Hegre's model produced forecasts up to 40 years ahead, he questioned

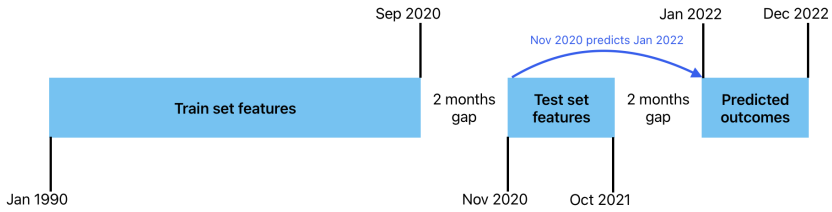


Fig. 2: Train and test datasets for 2022 prediction year. The test set spans 12 months and each month in the test set predicts 14 months ahead.

the duration of this assistance for forecasts of more than a decade forecasting horizon as the region's heterogeneity might change over time.

To incorporate the regional dynamic, we add two categorical variables: seven global and 23 smaller regions according to definitions by World Bank Development Indicators. These categorical variables are encoded using the dummy encoding scheme and added to the dataset.

Parametrization We develop logic to generate multiple versions of the dataset to test which features yield higher accuracy and test different hypotheses. Parameters that support fine-tuning and concern data preprocessing are listed in Table 2 under the 'Data processing parameters' category.

Least Important Features Drop After the country IDs and regions are encoded the number of dimensions in the resulting dataset explodes, which may lead to the model's under-performance. We define the least important features by running XGBoost regressor for each of the configurations of the dataset (see Section 4.6) and 35 of the least important features are removed from the dataset before the main algorithm is run.

4.4 Natural Gradient Boosting

In this section, we give a high-level overview of the algorithm used in this study and discuss the post-processing methodology of the predictions.

Gradient boosting is a supervised learning technique where several weak learners (or base learners) are combined in an additive ensemble [46]. The model is learnt sequentially, where the next base learner is fit against the training objective residual of the current ensemble. The output of the fitted base learner is then scaled by a learning rate and added to the ensemble.

Natural gradient boosting NGBoost is a framework for probabilistic prediction with competitive state-of-the-art performance on a variety of datasets [5]. NGBoost combines a multi-parameter boosting algorithm with the natural gradient to efficiently estimate how parameters of the presumed outcome distribution vary with the observed features. NGBoost can be used with any base learner, any

family of distributions with continuous parameters, and any scoring rule. While authors of the framework admit that point prediction will always be best with a dedicated model for that purpose, they find the loss in RMSE is not substantial if NGBoost is used to support probabilistic regression.

The NGBoost supports different base-learner models. In this study, we choose to keep a default base-learner model, which is a Decision Tree. We fine-tune the NGBoost and base learner hyperparameters in Section 4.6.

While NGBoost supports multiple distributions [47], this study tests Normal and Poisson distributions. The Poisson distribution, being a discrete probability distribution, models the dependent variable as a whole number. The Natural distribution predicts the dependent variable as a real number. Since the real number does not have a negative limit, the decent part of the predicted distribution can be negative if the prediction is around 0 but the variance is high. This requires additional handling, as negative predictions in the context of fatalities do not make sense. The handling of negative predictions is discussed in the next subsection.

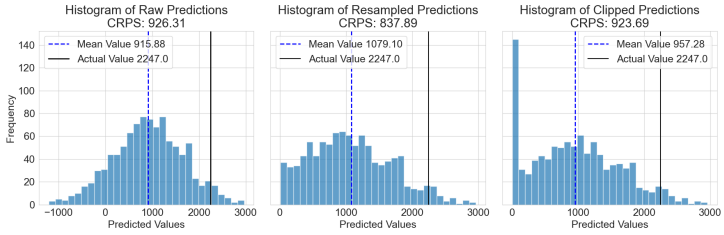
Handling Negative Predictions The gradient-boosting regression methods are unaware of feature boundaries and may make negative predictions even if the same feature was non-negative throughout the train set. While with point prediction the negative values of fatalities are simply converted to zero, in the case of distributions the handling of negative values in the distribution is dubious.

The ViEWS competition allows contestants to define their custom way of post-processing predictions but specifies default handling for negative predictions as clipping them to zero. The default clipping to zero creates a bias towards zero value, greatly changing the predicted distribution and impacting the CRPS and Mean Interval Score. We choose to resample the predicted distribution removing negative values. The processing of predictions is done in two steps:

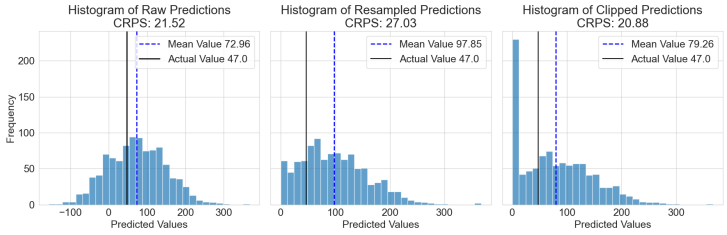
1. The model outputs 1000 predictions for each country-month in the test set.
2. All predictions are converted to integers.
3. In case some predictions have negative values, they are removed from the prediction set and the non-negative values are randomly resampled to fill in the created gaps.

The difference between clipping the negative predictions to 0 and resampling is shown in Figure 3. The resampling approach shifts the mean of the predicted distribution towards a higher value of fatalities, but, on the other hand, it does not place a disproportionate amount of confidence in the 0 value as does clipping. In Figure 3a, the model under-predicted fatalities, and resampling shifted the mean further which benefited the CRPS metric; in Figure 3b, the model over-predicted fatalities and clipping predictions to 0 benefited the CRPS metric.

Handling of removed countries The countries removed from the dataset in section 4.2 are removed from the train and test sets, implying that we do not evaluate predictions for countries with too many missing values. While this



(a) Prediction distributions for Afghanistan November 2020 (month_id: 477)



(b) Prediction distributions for Pakistan July 2020 (month_id: 473)

Fig. 3: Comparison of raw, resampled and clipped fatalities distributions with negative tails for two county-month cases.

admittedly reduces coverage of the predicted countries, on the other hand, making a model to predict fatalities based on inputs with almost every independent variable set to 0 is also unreliable.

4.5 Scoring Criteria

This section introduces three scoring metrics reported in this study and defined by the ViEWS competition.

We adhere to the main evaluation metric of the competition - the Continuous Ranked Probability Score (CRPS). The competition also defines two additional metrics that complement the CRPS: Ignorance score and Mean Interval Scores. We also report those metrics for comparability with other submissions of the competition.

Continuous Ranked Probability Score (CRPS) The CRPS is a scoring function that compares a single ground-truth value to its predicted distribution. This property makes it relevant to Bayesian machine learning, where models usually output distributional predictions rather than point-wise estimates. The CRPS is defined as follows:

The continuous rank probability score is defined as:

$$CRPS(F, y) = \int_{\mathbb{R}} (F(x) - \mathbb{1}(x - y))^2 \, dx \tag{1}$$

where $\mathbb{1}(z)$ is the indicator function, defined as

$$\mathbb{1}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This metric may be viewed as a generalization of the Brier score to infinitely small bins. Broadly, CRPS is a generalization of the mean absolute error (MAE) for any predictive distribution: if CRPS is used to compare a 'point' prediction as a cumulative distribution function with a point observation, it gives MAE.

Ignorace Score The ignorance score, which is also called the Log score, is the log of the predictive density evaluated at the actual observation:

$$IGN(f, y) = -\log_2(f(y)) \tag{3}$$

The ignorance score is the only proper local (i.e., only on the predictive density through its value at the event that materializes) scoring rule for continuous data. The ignorance score complements the CRPS by scoring the predicted probability of the observed event, instead of the distance between the predicted and observed. Therefore it emphasizes how much belief is focused on the observed value.

Mean Interval Score The M4 competition [48] use the Mean Interval Scores (MSIS). MSIS is set up as a battle between making the prediction interval as small as possible whilst still making sure that it has a good coverage rate. It does not consider the mass of the predictive distribution within the interval, so it is not an accuracy metric like CRPS. The metric is a nice transition from a point estimate to the distribution that CRPS tests. It puts the focus on the most likely values, without narrowing to a point. The trade-off between penalizing large intervals but rewarding coverage is useful too. The scaling in MSIS is used to make the measure scale-independent as the M4 competition deals with a large set of different types of time series with varying time scales and variability. The ViEWS team has simplified this score to just the (mean) Interval Score. A similar approach is also discussed by Gneiting & Raftery [49]. The Interval Score is defined as:

$$IS_{it} = (U_{it} - L_{it}) + \frac{2}{\alpha}(L_{it} - Y_{it})\mathbb{1}(L_{it} - Y_{it}) + \frac{2}{\alpha}(Y_{it} - U_{it})\mathbb{1}(Y_{it} - U_{it}) \tag{4}$$

where U_{it} and L_{it} are the upper and lower prediction sample quantiles using the set prediction interval, $\alpha = [1 - (\text{prediction interval})]$ (e.g., for a 95% prediction interval, $\alpha = 0.05$) and $\mathbb{1}(z)$ is the indicator function as defined previously. To get the Mean Interval Score, the IS_{it} is averaged across time t and units i.

Metrics Implementation As we use the scoring code published by ViEWS the details on implementations of the scoring metrics can be found in the ViEWS prediction competition invitation [30].

4.6 Model fine-tuning

This section describes how the NGBoost model was fine-tuned and details the fine-tuned parameters.

The fine-tuning is done across all five years (from the beginning of 2018 to the end of 2022) with the mean CRPS as a target optimisation metric and as explained in Section 4.4 the removed countries in the data preprocessing state are not evaluated and thus do not influence the fine-tuning process.

We perform fine-tuning using the Optuna framework [50]. We find that fine-tuning all variables is infeasible due to the duration of the model training and thus manually select eight variables to fine-tune. Four of these variables cover dataset construction, defining which features the final dataset should include and whether to remove rows that match custom criteria. Three variables cover the hyperparameters for the NGBoost framework, and the other one defines the maximum depth of the base learner used by NGBoost. The description of all variables selected for fine-tuning, their range and the best value are summarized in Table 2.

Parameter group	Parameter name	Values tested	Best	Parameter meaning
Data processing parameters	Include country ID	[True, False]	True	Whether to encode G&W country ID
	Include month ID	[True, False]	True	Whether to encode month ID
	Drop zero rows	[0, 20, 50, 80, 99, 100]	20	Percentage of rows with zero fatalities to be dropped from the training set
	Drop the least important	[True, False]	True	Whether to drop the predefined list of the 35 least important columns
NGBoost parameters	n_estimators	[300, 400, 500]	300	Number of iterations to train
	dist	[Normal, Poisson]	Normal	Distribution to assume
	minibatch_frac	[0.5, 1]	0.5	the percentage subsample of rows to use in each boosting iteration
Base Learner Parameters	bl_max_depth	[3, 4, 5]	5	The maximum depth of the tree

Table 2: List of fine-tuned parameters and the best values chosen

4.7 Competition Benchmarks

The prediction competition develops two heuristic benchmarks that we use to validate the performance of the developed NGBoost model. As the benchmarks are based on heuristics, the methodology of their construction is described below. The benchmarks published by competition organizers cover 4 years from 2018 to 2021.

Last Historical Poisson The first benchmark model uses the last observed value as the prediction – for every month in each calendar year, the prediction is based on the observed value in October of the preceding calendar year. Because of its nature, this benchmark model is almost constant across all months. To introduce variability in the benchmark model, the ViEWS team drew 1000 samples for each country month from a Poisson distribution with mean and variance equal to the last historical value. In the majority of the cases where the last observed number of fatalities was 0, all draws are identical. Table 3a shows the evaluation of the benchmark. The upper sub-table shows the evaluation metrics aggregated by calendar year as well as the mean score across the four years. The lower sub-table shows that the scores tend to deteriorate over the months of the forecasting horizon. Naturally, the violence recorded in October in a year is a better predictor of violence in January than in December the year after.

Bootstraps from actuals The second benchmark model uses actual true fatalities. For each country, the model makes 1000 draws from the set of observed fatalities for all countries of the entire calendar year in which this month is. For instance, the prediction for any country in July 2021 is a set of random draws from the observed fatality counts for all country-month instances in 2021. Table 3b shows the evaluation of the benchmark. The expected strength of this benchmark is that it in aggregate covers all actual outcomes. Thus, the ignorance scores are low. CRPS is higher than all the other models, whereas the mean interval scores are moderately high.

5 Results

In this section, the results of the NGBoost model performance are presented and compared to the competition benchmarks published by ViEWS. Additionally, Section 5.2 discusses the model performance solely for the 2022 year, as the 2022 year is not covered by benchmarks and does not allow for direct comparison.

The evaluation results for the 2018-2022 years are presented in Table 4 which composes three subtables. The sub-table 4a shows target metrics aggregated per year and two means: one mean for 2018-2021 which is comparable to benchmarks and another one for 2018-2022, which cannot be directly compared to the benchmarks. Sub-table 4b shows the averaged per-month metrics for 2018-2021 and sub-table 4c shows the averaged per-month metrics for 2022 only.

By calendar year			
year	crps	ign	mis
2018	20.04	1.19	378.08
2019	9.64	1.04	175.83
2020	13.70	1.08	256.18
2021	37.13	1.23	722.67
Mean	20.13	1.13	383.19
By month in forecasting horizon			
month	crps	ign	mis
1	13.74	1.12	256.14
2	13.09	1.02	242.93
3	11.25	1.15	206.08
4	19.30	1.12	366.75
5	19.15	1.16	363.58
6	23.79	1.09	456.75
7	22.53	1.16	430.75
8	19.05	1.02	361.42
9	17.56	1.25	331.33
10	21.49	1.19	410.53
11	42.02	1.19	820.02
12	18.56	1.17	352.02

(a) Last historical values predictions

By calendar year			
year	crps	ign	mis
2018	23.49	1.11	453.29
2019	22.07	1.09	419.32
2020	21.27	1.09	402.35
2021	36.05	1.13	694.49
Mean	25.72	1.10	492.36
By month in forecasting horizon			
month	crps	ign	mis
1	23.25	1.12	446.28
2	20.10	1.06	383.71
3	21.35	1.13	403.33
4	25.45	1.10	485.86
5	26.83	1.15	502.16
6	31.67	1.08	613.29
7	33.88	1.15	656.47
8	24.83	1.08	473.11
9	19.19	1.17	360.46
10	22.56	1.04	431.10
11	41.30	1.09	811.85
12	18.24	1.10	340.73

(b) Bootstraps from actuals predictions

Table 3: VIEWS benchmark models evaluation for 2018-2021 years, aggregated by calendar year and by month. The last historical values benchmark tends to be more accurate which is expressed in the mean for all years, but its accuracy also deteriorates over time, contrary to the bootstraps from actuals benchmark.

5.1 General Performance (2018-2021)

This subsection evaluates NGBoost model performance against heuristic benchmarks published by ViEWS which cover 2018-2021 years and that can be directly compared to our model.

Comparing per-year accuracy we see that the model improves over both benchmarks for every year, except for only 2019, where the Last Historical Poisson yields a very low CRPS of 9.64. The mean CRPS of the model is 18.12 which improves over the Last Historical Poisson and Actuals Boostraps benchmarks by 12% and 41% respectively.

Turning to per-month prediction accuracy we see that the CRPS for the model is lower for each month in the test set compared to the corresponding months of both ViEWS benchmarks, except for month 7, where the model

slightly under-performs (CRPS of 25.60 versus 22.53 for the Last Historical Poisson).

Secondary metrics, such as the Ignorance Score and Mean Interval Score, further illustrate the model's performance:

- Compared to benchmarks, the Ignorance Score of the model is lower for each year except for 2019. The mean Ignorance Score is 1.02, which is approximately 7.27% lower than the 'Bootstraps from actuals' benchmark mean of 1.10. This indicates a better probabilistic calibration of the model.
- The Mean Interval Score improves over all years of both benchmarks. The mean MIS for the model is 163.30 while the lowest of the benchmarks is 383.19, indicating more accurate prediction intervals.

These results suggest that the NGBoost model provides more accurate and reliable forecasting of conflict-related fatalities compared to the competition benchmarks.

year	crps	ign	mis
2018	15.73	0.89	181.64
2019	14.71	1.08	152.70
2020	12.82	1.07	155.55
2021	29.21	0.95	461.70
2022	72.34	1.06	1332.44
Mean[a]	14.42	1.02	163.30
Mean[b]	26.54	1.01	407.89

(a) By calendar year

month	crps	ign	mis
1	13.08	0.89	120.67
2	10.47	0.91	86.49
3	11.39	1.01	100.47
4	17.61	1.04	231.38
5	16.94	0.98	226.61
6	23.07	0.95	332.04
7	25.60	0.96	373.50
8	17.07	0.99	219.97
9	13.43	1.02	159.81
10	18.44	1.06	245.74
11	36.79	1.12	613.15
12	13.49	1.07	144.95

(b) By month in forecasting horizon 2018-2021

month	crps	ign	mis
1	12.94	0.98	137.72
2	11.50	0.96	121.59
3	68.99	1.04	1275.12
4	51.80	0.98	923.75
5	31.08	1.03	515.10
6	16.22	1.13	210.85
7	12.01	1.09	131.29
8	12.24	1.09	131.51
9	158.98	1.11	3045.86
10	22.88	1.10	330.48
11	436.38	1.17	8600.00
12	33.05	1.05	565.99

(c) By month in forecasting horizon 2022

[a] Mean for 2018-2021
[b] Mean for 2018-2022

Table 4: NGBoost model evaluation for 2018-2022 years, aggregated by calendar year and by month

5.2 Additional evaluation for the 2022 year

The ViEWS team has supplied data until October 2021, which also allows for predicting the whole year of 2022. While the results cannot be directly compared to the benchmarks as their coverage is only until 2021, the evaluation results highlight potential biases present in the model.

As can be seen in Table 4c, the prediction accuracy for the 2022 year is substantially worse than the aggregated accuracy for the 2018-2021 years. The per-year aggregations in Table 4a show that the mean CRPS for the 2018-2022 period is 26.54, which is twice as bad as the 14.42 CRPS for the 2018-2021 period.

Table 4c shows per-month metrics for 2022 and in comparison with Table 4b it is clear that the CRPS for months 3, 4, 9, and 11 is much higher than averages for previous years. The per-month distribution of predictions vs actual values is shown in Appendix section A.2.

November of 2022 stands out with a strikingly high CRPS of 436.38. Such low accuracy comes from a missed spike in fatalities in Ethiopia. In November UCDP recorded a record number of 80 thousand fatalities in Ethiopia. The instability in Tigray province has spiked a violent civil war that resulted in numerous fatalities. The model fails to predict this spike, which results in a CRPS of 8,397.3 for Ethiopia on a country level for 2022 and in CRPS of 79,519.44 for a single prediction in Ethiopia in November.

The predictions for Ethiopia and some other countries are available for visual inspection in Appendix section A.3. Interestingly, some countries seem to experience violence seasonally. For example, Algeria experienced a spike of higher magnitude at the beginning of 2021 and a lower magnitude spike at the beginning of 2022, which was correctly predicted by the model. At the same time, we see that for the Central African Republic, the model underestimates fatalities at the beginning of 2022 and predicts a slight spike which is still lower than the actual values. This behaviour brings to the attention the fact that in some cases the model heavily relies on the previous violence to model the prediction. This observation and the feature's importance are discussed in the next two sections.

5.3 Model accuracy dependency on input fatalities distribution of the month

Contrary to the Last Historical Poisson benchmark, the accuracy of the NGBoost model does not seem to decrease over time as more distant forecasts are made. The prediction accuracy for the first month is 13.08, for the ninth - 13.43 and the twelfth - 13.49 (shown in Table 4b) are roughly comparable. While this might mean that the model can generalise well, it also raises concern that some months are easier, and that's why the accuracy drastically decreases for the other months.

Figure 4 shows the distribution of the dependent variable for the 2018-2021 with the per month average CRPS for these prediction windows. It can be seen that for months that have a distribution of fatalities close to 0 and only a couple of observations with high fatalities, the CRPS is substantially lower. This proves that some months are indeed 'easier' regarding the absence of violence spikes and that the model has constantly higher accuracy for such 'easy' months.

To give another perspective on the model predictions, Figure 5 shows the per-month distribution of the predicted vs actual values for the 2021 year and the standard deviation of each prediction. It can be seen that the model tends to

underestimate fatalities. For Nov 2021 there is a single prediction for which the model predicts around 200 fatalities with high certainty, while the actual value spikes to more than 10,000.

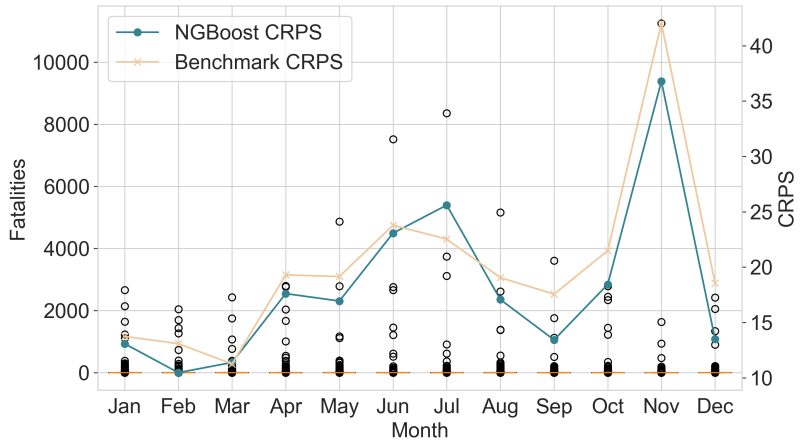


Fig. 4: Distributions of dependent variable per month for 2018-2021 and mean CRPS per month (lines with scale on the right) for the NGBoost model and the Last Historical Poisson benchmark. The CRPS score increases for the months that have a distribution of fatalities further from 0.

5.4 Feature Importance

Figure 6 shows the ten most important features and their influence on the NGBoost model output using Shapley values [51]. Previous studies indicated, that the models optimised to satisfy general criteria of the smallest prediction error of fatalities tend to rely on historical violence to make forecasts [41, 42]. As for most countries, the amount of fatalities does not fluctuate fast and spikes as rather rare - the implemented NGBoost model also finds a correlation that the previous violence is a good predictor of future violence.

Except for the previous fatalities, the important features are the percentage of the female labour force, the refugee population in a country, freedom of domestic movement, and whether the government respects civil liberties across different areas of the country.

Interestingly, the month ID does not make it to the top ten as well as the other political and economic features found statistically significant in the other conflict studies [52] (Table 2 and Table 3).

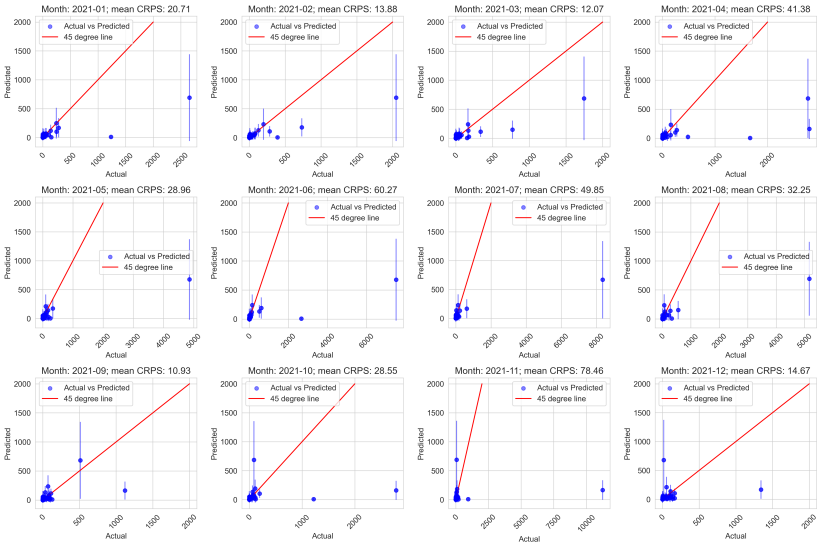


Fig. 5: Distributions of predicted values by NGBoost model vs actual values per month for 2021 prediction window along with the mean CRPS per month

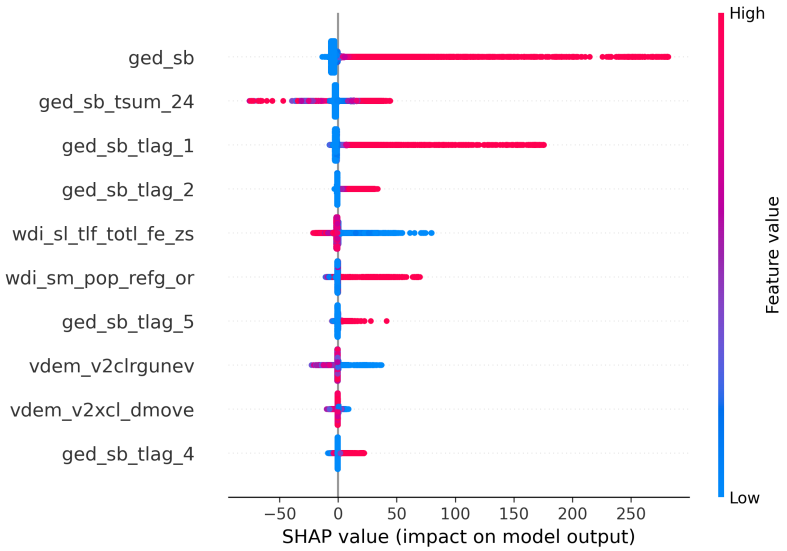


Fig. 6: SHAP tree explainer for top 10 most important features. Historical violence features mostly shape the model output, with some adjustment for structural variables.

5.5 Analysis of country forecasts

One of the most beneficial features of the NGBoost model is that it produces a separate forecast per month for the whole forecasting window, which allows for per-country analysis and comparisons to benchmarks. In this section, we look into the feature importance of predictions for Ethiopia, Algeria, and Turkey utilising the benefits of the NGBoost framework explainability.

Figure 7 shows three country-month predictions for 2022 with the SHAP force plots for each. The force plot for the November spike in Ethiopia is shown in Figure 7a, which highlights that the model had no clue about the upcoming surge of violence. The model predicts 90 fatalities based on historical fluctuations with the actual value for this month being 79,609. We see that percentage of female labour feature, which was found important in Section 5.4, is relatively high (46.5%) and decrease the estimated number of fatalities.

On the other hand, in Figure 7b we see that the model overestimated fatalities for Algeria in March 2022. The forces that push the estimate above the actual value are three protests that happened that month according to the ACLED database and a low percentage of female labour.

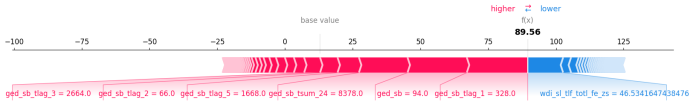
Figure 7c shows a prediction for Turkey in January 2022, where again the model overestimates actual fatalities. Interestingly, the model adds a country bias for all predictions in 2022, which pushes all estimates by about 7 fatalities. All the other features that mainly shape this forecast are previous violence features with a small positive influence from the moderate value of the freedom of domestic movement index.

Overall, the force plot analysis indicates that the model relies on historical violence and country bias to build its predictions with political and economic features only slightly adjusting the forecasts. The implications of the findings above are discussed further in Section 6 along with possible improvements in Section 7.

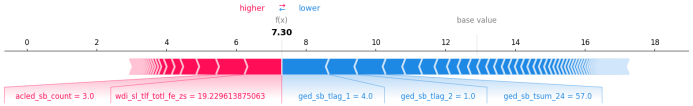
6 Discussion

The model presented in this paper is the first attempt to build an explainable probabilistic regression model to forecast state-based fatalities on a country-month level with global coverage. While the model improves over the heuristic benchmarks published by the competition organizers, it possesses several limitations that do not allow it to generalise well. The model heavily relies on historical fatalities to make future forecasts, and only moderately considers other economic and political features that theoretically are drivers for the conflict.

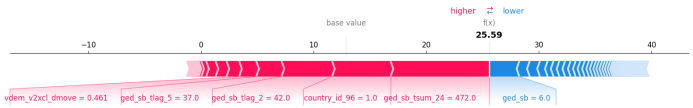
On the other hand, the economic and political features are mostly static and do not change swiftly even when the conflict bursts. One can argue that economic and political features may explain the risk (likelihood) of the conflict, but cannot indicate an outburst of a conflict, as they do not capture sudden changes in people's mood and views or situational context. An outburst of a Ukrainian war in February 2022 serves as a good example in this case. The



(a) Prediction for Ethiopia on Nov 2022, with actual of 79,609



(b) Prediction for Algeria on March 2022, with actual of 3



(c) Prediction for Turkey on Jan 2022, with actual of 3

Fig. 7: SHAP force plot for three cases showcasing how the features are weighted for separate predictions.

economic and political features present in the dataset do not indicate the Russian troops gathering on a Ukrainian border three months before the invasion, which makes it questionable whether such conflict can be predicted with the present dataset.

Analysing results closely, we see the model does not capture the concept of war and the fact that if two countries are at war then they both are more likely to incur high fatalities. There are two reasons for the lack of this understanding: the absence of geopolitical features in the dataset that may explain relationships between countries and that the spatiotemporal features present in the dataset simply indicate that there was some violence in the neighbouring countries but do not explain the relationship of the analysed country to the neighbours' fatalities.

Another limitation of the current model is that it tests only two distributions. While the normal distribution was found to be the most accurate, arguably, the country's risks of violence are not always distributed normally. For conflict research, the tails of the probability distribution are of the most interest as they might warn on a small probability of the worst-case scenario and a high probability of the most likely forecast.

Moreover, the CRPS as a target metric used to optimise the model, is not ideal as it tries to satisfy a general case (which for most countries means prolonged peaceful periods with rare outbursts). While generic scoring might be suitable for cases where the spikes happen frequently, predicting rare spikes re-

quires metrics that allow for learning these patterns and do not severely punish temporal errors as CRPS does. The possible target metrics might be multiplex pseudo-Earth Mover Divergence Score [53], which estimates the cost of moving excess prediction mass across space and time. This metric works with binned forecasts, which is not exactly suitable for predicting the exact number of fatalities, but the idea is that it still favours the predicted spike slightly shifted in time. Additionally, reporting the model evaluation scores only for the periods with changes in fatalities as was done by Muller & Rauh [54] might be more informative. Arguably, conflict forecasting models are valuable when they forecast escalation and de-escalation correctly and do not simply extrapolate history.

7 Future Work

A promising step forward is the addition of high-frequency data such as news and political speech. Utilising a recently published dataset for topic analysis of news by Muller et al. [55] will add a much-needed high-frequency component to the feature set. Future work should also focus on working with the dependent variable, as due to the high skewness to the right, the model struggles to output predictions with more than 500 fatalities. A possible solution that should be tested is to perform a log transformation of the dependent variable, which will better fit a normal distribution.

Another promising improvement is to employ a weighted ensemble of models. This approach worked greatly for the ViEWS team for their weighted ensemble optimised for point prediction. The ViEWS lab shows that implementing multi-horizon forecasting reduces the error rate for the first month by 2-3 times compared to the 36th month[10].

Further, experimenting with deep learning techniques such as Long Short-Term Memory (LSTM) networks and Transformer architectures can be beneficial. While Ettensperger shows that the best LSTM network could only reach and not surpass the performance of Random Forest trained on the same data [42], the dataset used in that study comprised only 851 country-years, which is arguably too small for deep learning. With a country-month dataset of 70 thousand observations, deep learning may already be feasible. LSTM networks can capture long-term dependencies in sequential data suitable for the conflict modelling task. Transformer architecture, leveraging the self-attention mechanism, can capture complex patterns and dependencies, and may be able to improve the prediction accuracy of rare spikes in violence.

Moreover, experimenting with other representations that support modelling relations between countries may allow a model to better capture geopolitical dynamics. Future work may focus on developing a framework for representing countries as nodes in a Spatio-Temporal Attention Graph Neural Network [56, 57] and their relations as edges. But as the number of countries changes over time and node edges might also change this approach brings challenges that need to be solved before it can be implemented.

Additionally, implementing a data pipeline for updates of the model's data will allow for real-time monthly forecasts. While this step only makes sense when the accuracy is in place, it will be important in the future to deliver a fully-fledged conflict prediction model.

8 Replication data

The replication scripts are freely available on the GitHub page of the project.

References

[1] Håvard Hegre et al. "Predicting Armed Conflict, 2010–2050". In: *International Studies Quarterly* 57.2 (June 2013), pp. 250–270. ISSN: 0020-8833. DOI: 10.1111/isqu.12007. eprint: https://academic.oup.com/isq/article-pdf/57/2/250/5056073/57-2-250.pdf. URL: https://doi.org/10.1111/isqu.12007.

[2] Thomas Chadefaux. "Early warning signals for war in the news". In: *Journal of Peace Research* 51.1 (2014), pp. 5–18. DOI: 10.1177/0022343313507302.

[3] Espen Geelmuyden Rød, Tim Gåsste, and Håvard Hegre. "A review and comparison of conflict early warning systems". In: *International Journal of Forecasting* 40.1 (Jan. 2024), pp. 96–112. DOI: 10.1016/j.ijforecast.2023.01.001.

[4] Robert Muggah and Mark Whitlock. "Reflections on the Evolution of Conflict Early Warning". In: *Stability: International Journal of Security and Development* 10 (2022). URL: https://link.gale.com/apps/doc/A698837336/AONE.

[5] Tony Duan et al. *NGBoost: Natural Gradient Boosting for Probabilistic Prediction*. 2020. arXiv: 1910.03225 [cs.LG].

[6] Stephen J. Andriole and Robert A. Young. "Toward the Development of an Integrated Crisis Warning System". In: *International Studies Quarterly* 21.1 (1977), p. 107. DOI: 10.2307/2600149.

[7] George W Williford and Douglas B Atkinson. "A Bayesian forecasting model of international conflict". In: *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 17.3 (Feb. 2019), pp. 235–242. DOI: 10.1177/1548512919827659.

[8] Jack A. Goldstone et al. "A Global Model for Forecasting Political Instability". In: *American Journal of Political Science* 54.1 (2009), pp. 190–208. DOI: 10.1111/j.1540-5907.2009.00426.x.

[9] Hannes Mueller and Christopher Rauh. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". In: *American Political Science Review* 112.2 (2017), pp. 358–375. DOI: 10.1017/s0003055417000570.

[10] Hegre Håvard et al. "ViEWS: Forecasting Fatalities v2". In: *DiVA* (2022). Accessed: 2024-06-12. URL: https://www.diva-portal.org/smash/get/diva2:1667048/FULLTEXT01.pdf.

[11] Håvard Hegre et al. *Forecasting fatalities (002) - ViEWS*. Accessed: 2024-06-12. 2023. URL: https://viewsforecasting.org/early-warning-system/models/fatalities002/.

[12] Håvard Hegre, Håvard Mokleiv Nygård, and Ranveig Flaten Ræder. "Evaluating the scope and intensity of the conflict trap: A dynamic simulation approach". In: *Journal of Peace Research* 54.2 (2017), pp. 243–261. ISSN: 00223433, 14603578. URL: http://www.jstor.org/stable/44511209 (visited on 06/15/2024).

[13] M Halkia et al. *The Global Conflict Risk Index – Artificial intelligence for conflict prevention*. Publications Office, 2019. DOI: doi/10.2760/004232.

[14] G Schvitz et al. *The Global Conflict Risk Index 2022 – Revised data and methods*. Publications Office of the European Union, 2022. DOI: doi/10.2760/041759.

[15] Emma Baillie et al. "Explainable models for forecasting the emergence of political instability". In: *PLOS ONE* 16.7 (July 2021), pp. 1–18. DOI: 10.1371/journal.pone.0254350. URL: https://doi.org/10.1371/journal.pone.0254350.

[16] Felix Ettensperger. "Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging with multiple tree-based algorithms and variance promoting data configurations". In: *International Interactions* 48.4 (2021), pp. 555–578. DOI: 10.1080/03050629.2022.1993209.

[17] Hannes Mueller and Christopher Rauh. "Using past violence and current news to predict changes in violence". In: *International Interactions* 48.4 (May 2022), pp. 579–596. DOI: 10.1080/03050629.2022.2063853.

[18] Samuel Bazzi et al. "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia". In: *The Review of Economics and Statistics* 104.4 (2022), pp. 764–779. DOI: 10.1162/rest_a_01016.

[19] S Halkia et al. *Dynamic Global Conflict Risk Index*. Publications Office, 2019. DOI: doi/10.2760/846412.

[20] Nils B Weidmann and Sebastian Schutte. "Using night light emissions for the prediction of local wealth". In: *Journal of Peace Research* 54.2 (2016), pp. 125–140. DOI: 10.1177/0022343316630359.

[21] Thomas Chadefaux. "Market anticipations of conflict onsets". In: *Journal of Peace Research* 54.2 (2017), pp. 313–327. DOI: 10.1177/0022343316687615.

[22] Daniel Berger, Shankar Kalyanaraman, and Sera Linardi. "Violence and Cell Phone Communication: Behavior and Prediction in Cote DDIvoire". In: *SSRN Electronic Journal* (2014). DOI: 10.2139/ssrn.2526336.

[23] Quansheng Ge et al. "Modelling armed conflict risk under climate change with machine learning and time-series data". In: *Nature Communications* 13.1 (May 2022). DOI: 10.1038/s41467-022-30356-x.

[24] Frank DW Witmer et al. "Subnational violent conflict forecasts for sub-Saharan Africa, 2015—65, using climate-sensitive models". In: *Journal of Peace Research* 54.2 (2017), pp. 175–192. ISSN: 00223433, 14603578. URL: http://www.jstor.org/stable/44511205 (visited on 06/15/2024).

[25] Lisa Hultman, Maxine Leis, and Desirée Nilsson. "Employing local peacekeeping data to forecast changes in violence". In: *International Interactions* 48.4 (2022), pp. 823–840. DOI: 10.1080/03050629.2022.2055010.

[26] Ram Avtar et al. "Remote Sensing for International Peace and Security: Its Role and Implications". In: *Remote Sensing* 13.3 (2021), p. 439. DOI: 10.3390/rs13030439.

[27] Daniel Racek et al. "Conflict forecasting using remote sensing data: An application to the Syrian civil war". In: *International Journal of Forecasting* 40.1 (2024), pp. 373–391. DOI: 10.1016/j.ijforecast.2023.04.001.

[28] Kristian Skrede Gleditsch and Michael Ward. "A revised list of wars between and within independent states, 1816-2002". In: *Copyright © Taylor* 30 (Jan. 2004), pp. 231–262.

[29] Correlates of War Project. *State System Membership List, v2016*. http://correlatesofwar.org. 2017.

[30] Håvard Hegre et al. "Prediction Challenge 2023/2024". In: (2023). Accessed: 2024-06-10. URL: https://viewsforecasting.org/research/prediction-challenge-2023/.

[31] Shawn Davies, Therése Pettersson, and Magnus Öberg. "Organized violence 1989–2022, and the return of conflict between states". In: *Journal of Peace Research* 60.4 (2023), pp. 691–708. DOI: 10.1177/00223433231185169.

[32] Ralph Sundberg and Erik Melander. "Introducing the UCDP Georeferenced Event Dataset". In: *Journal of Peace Research* 50.4 (2013), pp. 523–532. DOI: 10.1177/0022343313484347.

[33] *The World Bank Annual Report 2015*. The World Bank, 2015. DOI: 10.1596/978-1-4648-0574-5.

[34] Lars-Erik Cederman, Andreas Wimmer, and Brian Min. "Why Do Ethnic Groups Rebel? New Data and Analysis". In: *World Politics* 62.1 (2009), pp. 87–119. DOI: 10.1017/s0043887109990219.

[35] Wolfgang Lutz. "Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000". In: *Vienna Yearbook of Population Research* 2007 (2007), pp. 193–235. DOI: 10.1553/populationyearbook2007s193.

[36] Clionadh Raleigh et al. "Introducing ACLED: An Armed Conflict Location and Event Dataset". In: *Journal of Peace Research* 47.5 (2010), pp. 651–660. DOI: 10.1177/0022343310378914.

[37] Michael Coppedge et al. "Conceptualizing and Measuring Democracy: A New Approach". In: *Perspectives on Politics* 9.2 (2011), pp. 247–267. DOI: 10.1017/s1537592711000880.

[38] Food and Agriculture Organization. *AQUASTAT Glossary*. https://www.fao.org/aquastat/en/. FAO website (accessed on 17 June 2024). 2019.

[39] Håvard Hegre et al. *The 2023/24 VIEWS Prediction competition: Predicting the number of fatalities in armed conflict, with uncertainty*. Accessed on 2024-07-03. 2023. URL: https://viewsforecasting.org/wp-content/uploads/VIEWS_2023.24_Prediction_Competition_Invitation.pdf.

[40] Marie Allansson, Erik Melander, and Lotta Themnér. "Organized violence, 1989–2016". In: *Journal of Peace Research* 54.4 (2017), pp. 574–587. DOI: 10.1177/0022343317718773. eprint: https://doi.org/10.1177/0022343317718773. URL: https://doi.org/10.1177/0022343317718773.

[41] Michael D. Ward et al. "Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction". In: *International Studies Review* 15.4 (2013), pp. 473–490. ISSN: 15219488, 14682486. URL: http://www.jstor.org/stable/24032984 (visited on 06/19/2024).

[42] Felix Ettensperger. "Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field". In: *Quality amp; Quantity* 54.2 (May 2019), pp. 567–601. DOI: 10.1007/s11135-019-00882-w.

[43] scikit-learn developers. *OneHotEncoder*. Accessed: 2024-07-04. 2024. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[44] Sarah Neumann, Darryl Ahner, and Raymond R. Hill. "Forecasting country conflict using statistical learning methods". In: *Journal of Defense Analytics and Logistics* 6.1 (2022), pp. 59–72. DOI: 10.1108/jdal-10-2021-0014.

[45] Benjamin C. Boekestein. "A Predictive Logistic Regression Model of World Conflict Using Open Source Data". Theses and Dissertations, 101. MS thesis. Air Force Institute of Technology, 2015. URL: https://scholar.afit.edu/etd/101.

[46] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[47] Stanford ML Group. *Usage of NGBoost for Regression Distributions*. Accessed: 2024-06-18. URL: https://stanfordmlgroup.github.io/ngboost/1-useage.html#Regression-Distributions.

[48] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PLOS ONE* 13.3 (2018), e0194889. DOI: 10.1371/journal.pone.0194889.

[49] Tilmann Gneiting and Adrian E Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378. DOI: 10.1198/016214506000001437.

[50] Takuya Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. arXiv: 1907.10902 [cs.LG].

[51] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[52] Benjamin Leiby and Darryl Ahner. "Datasets and Models for Globally Predicting Country Conflict and Peace: A Survey". In: *Military Operations Research* 28.3 (2023), pp. 87–112. ISSN: 10825983, 21632758. URL: https://www.jstor.org/stable/27254917 (visited on 06/08/2024).

[53] Paola Vesco Håvard Hegre and Michael Colaresi. "Lessons from an escalation prediction competition". In: *International Interactions* 48.4 (2022), pp. 521–554. DOI: 10.1080/03050629.2022.2070745. eprint: https://doi.org/10.1080/03050629.2022.2070745. URL: https://doi.org/10.1080/03050629.2022.2070745.

[54] Hannes Mueller and Christopher Rauh. "The Hard Problem of Prediction for Conflict Prevention". In: *Journal of the European Economic Association* 20.6 (Apr. 2022), pp. 2440–2467. ISSN: 1542-4766. DOI: 10.1093/jeea/jvac025. eprint: https://academic.oup.com/jeea/article-pdf/20/6/2440/48333706/jvac025.pdf. URL: https://doi.org/10.1093/jeea/jvac025.

[55] H Mueller, C Rauh, and B Seimon. "Introducing a global dataset on conflict forecasts and news topics". In: (2024). DOI: 10.17863/CAM.106231. URL: https://www.repository.cam.ac.uk/handle/1810/364649.

[56] Petar Veličković et al. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML].

[57] Zahraa Al Sahili and Mariette Awad. *Spatio-Temporal Graph Neural Networks: A Survey*. 2023. arXiv: 2301.10569 [cs.LG].

A Appendix

A.1 Correlation Matrix

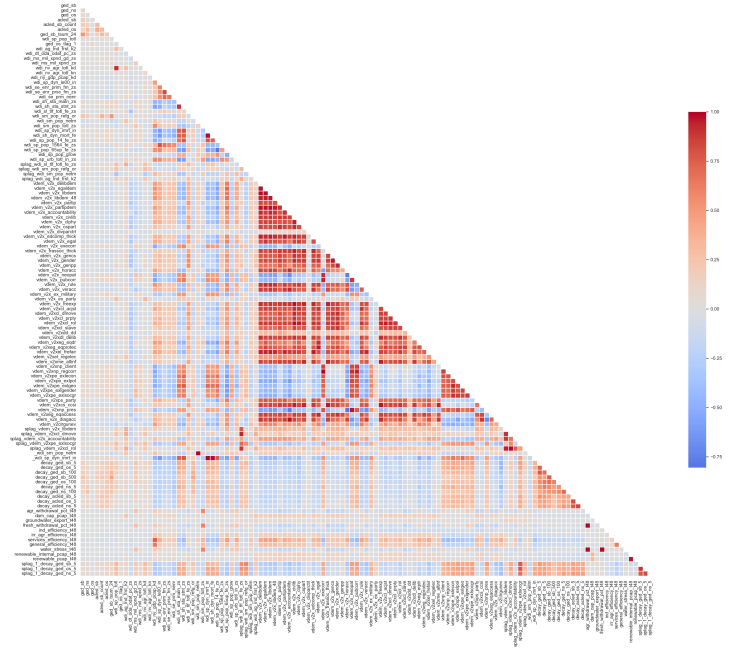


Fig. 8: Correlation matrix for original ViEWS competition dataset

A.2 Predicted vs Actual values for 2022

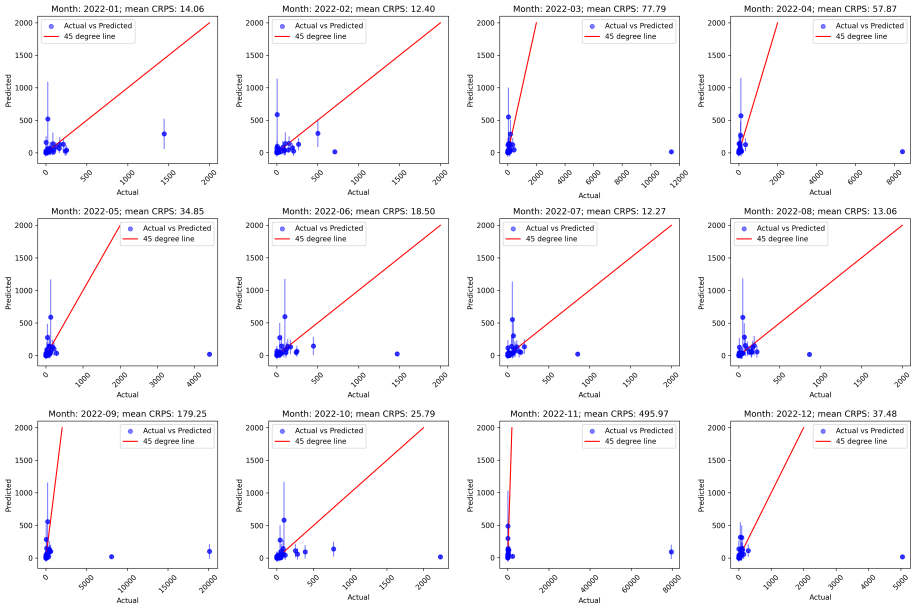


Fig. 9: Distributions of predicted values by NGBoost model vs actual values per month for 2022 prediction window along with the mean CRPS per month

A.3 Predictions for some countries in 2022 prediction window

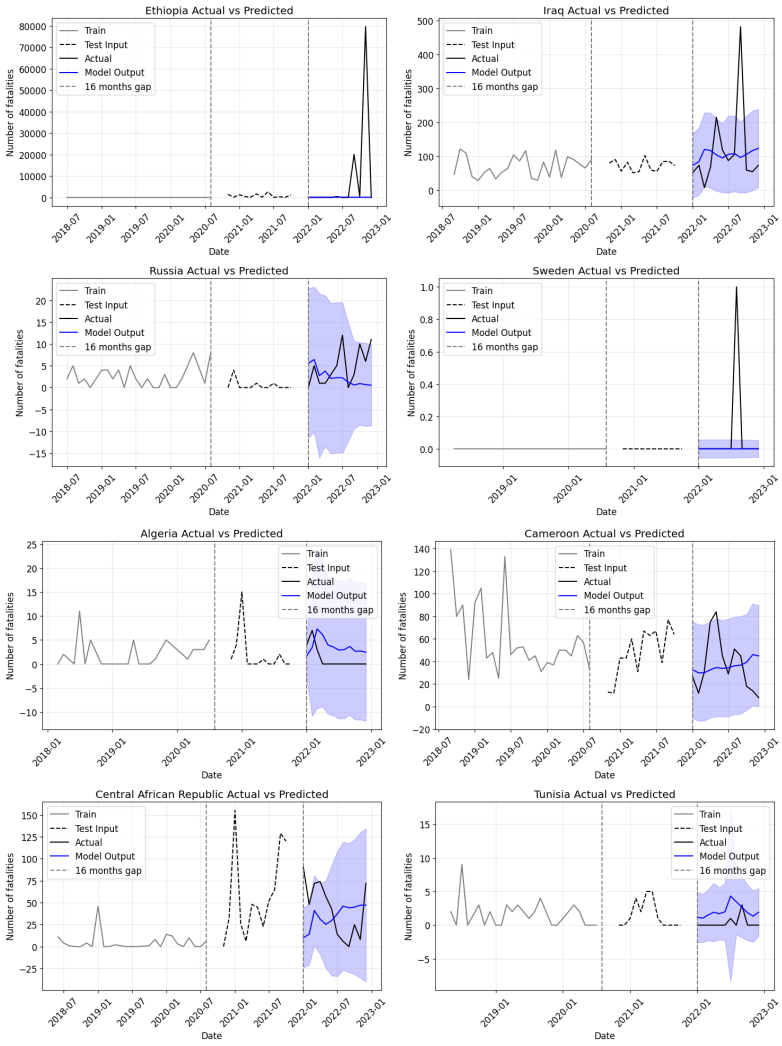


Fig. 10: Predictions for 2022. The highlight is Ethiopia, for which the model misses the spike in violence. For the other countries, the predictions are moderately accurate.