Forests of UncertainT(r)ees: Using tree-based ensembles to estimate probability distributions of future conflict[1]

Technical Report

Daniel Mittermaier[a], Tobias Bohne[a], Martin Hofer

[a]*Center for Crisis Early Warning, University of the Bundeswehr Munich*

Abstract

The high uncertainty of point predictions when forecasting conflict, especially at the subnational level, is a significant shortcoming and major obstacle to the practical application of conflict prediction systems. In our contribution to the 2023/24 ViEWS prediction challenge at the PRIO-GRID-month (*pgm*) level, we employ a quasi-hurdle combination of tree-based models to generate *pgm*-specific predictions of N=1000 samples each three to fourteen months into the future. Our strategy combines predictions from a binary classification task on the occurrence of fatalities with sample outputs from a distributional regressors trained only on non-zero targets. We address the problem of zero-inflation by interpreting the probability of the classifier as the share of non-zero predictions in the final samples drawn to represent the predicted distributions. We design a modeling pipeline to automatically tune multiple classifiers and regressors and select the best model for each prediction timestep based on tuning performance. In an effort to address a lack of data as a source of uncertainty, we additionally generate "local" model predictions for semi-automatically generated spatial clusters of violence based on *pgms* experiencing any fatalities in our training data, thus accounting for context-specific systematic differences in conflict dynamics. While all our models beat a series of benchmarks across almost all test windows and metrics, the "global"-only model (*unibw_trees_global*) and a global-local combination selected based on past performance (*unibw_trees_global-local*) scored best, with the local model (*unibw_trees_local*) only slightly worse.

[1] This paper documents a contribution to the VIEWS Prediction Challenge 2023/2024. Financial support for the Prediction Challenge was provided by the German Ministry for Foreign Affairs. For more information on the Prediction Challenge please see Hegre et al. (Forthcoming) and https://viewsforecasting.org/research/prediction-challenge-2023.

Introduction

Predicting violence at the *pgm*-level is not an easy task. The highly zero-inflated nature of the data, the coarse available data resolution of many potential determinants of conflict and, therefore, prediction features, as well as the fairly large uncertainties around recording conflict events and fatalities make it hard to beat even simple heuristics, as the findings of the first ViEWS prediction challenge showed (Vesco et al. 2022). Accounting for some of these uncertainties, the goals of this second challenge posted by the ViEWS team is not only to predict the expected number of fatalities but also to estimate uncertainty around the predictions in the form of samples from a predictive distribution (Hegre et al. 2023, Hegre et al. Forthcoming).

Despite their recent popularity and general advances around more complex approaches based on neural networks, practical applications show that often "tree-based models still outperform deep learning" (Grinsztajn et al. 2022: 1) when it comes to the tabular data primarily used in conflict research. Tree-based models also come with significant advantages in terms of interpretability, which is especially useful when communicating modeling results to policy- and decision-makers. Building on insights from previous modeling efforts, especially the strong performance of hurdle models and the potential of using ensembles (Vesco et al. 2022, Hegre et al. 2022b), we designed a modeling approach consisting of tree-based hurdle ensembles to predict violence at the *pgm*-level for our contribution to the challenge. Within this approach we incorporate distribution-specific regressors to estimate predictive distributions for each *pgm* individually. Furthermore, we address the issue of unobserved context-specific factors as a source of uncertainty in conflict models by creating a spatial ensemble consisting of multiple "local" models. With this approach we aim to account for potential systematic differences in conflict dynamics across different contexts. Lastly, we combine the best components from each of the two levels in a global-local ensemble based on past performance.

Our evaluation shows that all three approaches outperform the benchmarks across all test windows and all metrics with only very few exceptions. The global model and the global-local combined model score very similarly, while the local model performs slightly worse, with fairly marginal overall differences between the scores. The local approach therefore did not lead to an overall improvement of model performance. As the simplest of our three approaches, we prefer the global model in practice.

Methodology

Data and Modelling Setup

With our main focus on modeling strategies, we rely on the data provided by the ViEWS team in the context of the competition and use all available features in our models. The data cover Africa and the Middle East and are available monthly for each grid cell starting in 1990. As outlined in the competition call, the target is the number of fatalities from state-based armed conflict events (Hegre et al. 2023), as recorded by the Uppsala Conflict Data Program (UCDP) (Davies et al. 2023, Sundberg et al. 2013). The target is highly zero-inflated, with less than 0.4% non-zero values in the average month in our training data.

To generate predictions for the whole next year from the available training data, we chose to train separate models for each of the timesteps to predict (t+3, … t+14) and combine the resulting outputs to a full year of predictions, since we employ models which are unable to natively generate time series predictions. As the features include several lagged versions of our target variable, we are confident that this does not decrease our ability to extrapolate trends unreasonably. We strictly follow the two months time separation between features and test windows defined by the ViEWS team[2] to prevent data leakage. The training data is further limited by the time period we want to predict into the future with this approach. For example, the t+14 model generating the predictions for December 2018 can only be trained on data up to August 2016, 14 months before the October 2017 end of the training data, as this is the last month where there is sufficient future information to label the target.

In the interest of usability, we designed a modular, model-agnostic modeling pipeline in Python, which performs the tuning, training and predicting mostly automatically for the given prediction problem. We include multiple machine learning algorithms, selecting the algorithm which achieves the best performance during tuning for each timestep individually. This allowed us to integrate and test multiple different machine learning algorithms with minimal effort and means our approach can be easily reused for different prediction problems.

We perform hyperparameter tuning for each timestep only once based on the data up to October 2017 (N=4,378,740), before the first test window. Our tuning procedure is based on time series cross-validation with a 5-year sliding window through the training data. We employ the hyperopt package (Bergstra et al. 2013), which implements a Bayesian search approach with the Tree-structured Parzen Estimator

[2] For each of the six yearly test windows, the training data is artificially limited to October of the previous year. This is done to simulate the availability gap present in the data for the true future predictions, where training data is available until April 2024 to predict for the timeframe July 2024 to June 2025. Therefore, predictions need to be generated for the timesteps t+3, …, t+14.

(TPE) algorithm (Bergstra et al. 2011), to identify the best hyperparameters for each prediction timestep model. The cross-validation and scoring within the tuning procedure is implemented in the scikit-learn framework (Pedregosa et al. 2011). Where necessary, we implemented custom wrappers to integrate the different estimators and evaluation metrics into the framework. After tuning, we select the parameters with the best performance during tuning and fit the model on all available training data for a given test window before generating our predictions from the last available observations 3 months before the start of the prediction window.

Our tuning metric depends on the modeling step in the hurdle model, with further details provided below. For each chosen base metric, our tuning loss is based on the mean performance of the test splits during cross-validation. As we observed a strong tendency to overfit for some parameter combinations, we combine this with a penalty term for deviation from the mean performance of the training splits to emphasize generalizability of the models.[3]

Local Models

Much of the uncertainty of conflict prediction models likely stems from a lack of (precise enough) data on some of the factors influencing the occurrence and dynamics of violence. Consequently, these are not represented well in the available feature set, making it challenging to generate reliable predictions (Cederman & Weidmann 2017, Chadefaux 2017), with models relying heavily on past violence dynamics as a proxy for these underlying issues (Hegre et al. 2022a, Mueller & Rauh 2022).

To account for potential systematic differences across different conflict contexts, we develop a strategy to incorporate results from local models in addition to our base modeling approach. Our approach is inspired by the Geographical Random Forests (GRFs) method (Georganos et al. 2021), which we adapt to the problem of conflict prediction. While the original GRFs work by estimating a local model for each spatial unit trained on data from a spatial buffer zone, this is unsuited to the problem of conflict prediction on the *pgm* level. Either the scarcity of non-zero values would result in almost all local models not having enough training data to learn meaningful patterns of violence with a small buffer area, or the large number of grid cells in our data would make this process computationally too expensive with larger buffer sizes.

[3] The formula we use is *loss = -(mean test score - 0.5 • |mean train score - mean test score|)*, with scores where lower values are better inverted and the mean calculated across all cross-validation splits. Loss is then minimized by the TPE tuning algorithm.

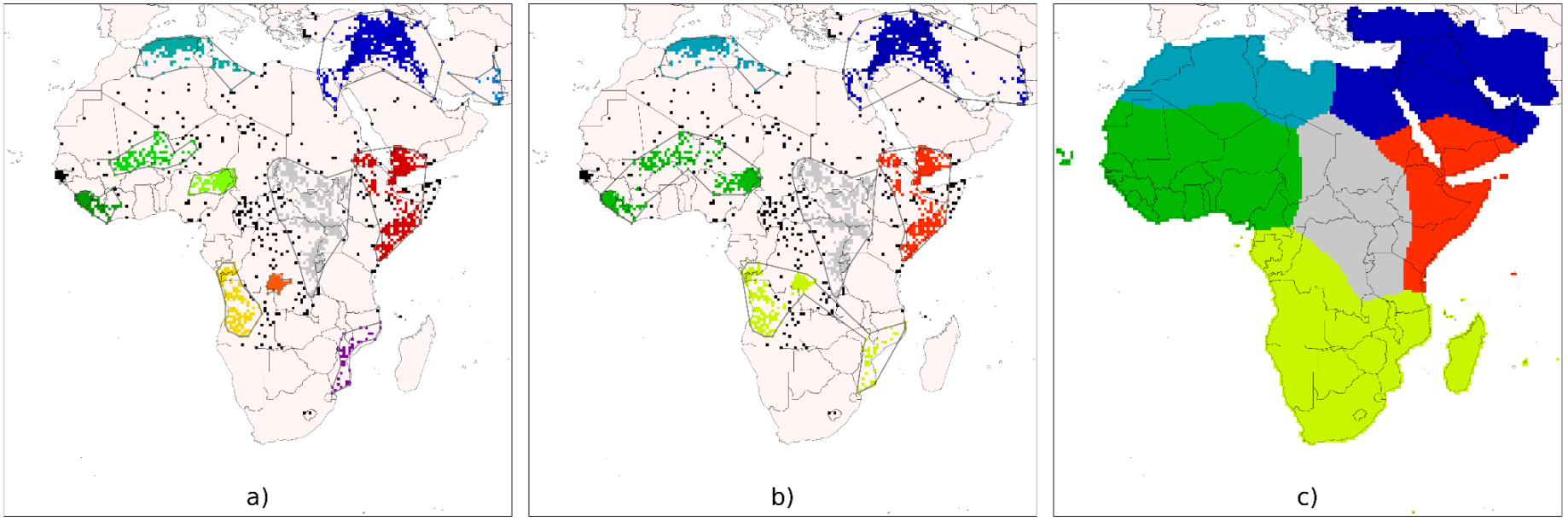

Figure 1: Visualization of the creation process of the clusters for local models. a) clusters created by manually tuned HDBSCAN algorithm with corresponding polygons; b) clusters after merging of smaller clusters with updated polygons; c) final clusters with grid-cells without violence assigned. Grid cells shown in a) and b) are those experiencing any violence in the training data (1990-2017). Black grid cells displayed in a) and b) are not assigned to any clusters by HDBSCAN initially.

Instead, we create custom contiguous geographic clusters based on the spatial distribution of cells with any recorded fatalities in the training data (1990-2017) using the HDBSCAN clustering algorithm (Campello et al. 2013)[4]. We performed an initial manual tuning of the clustering algorithm until inspections of the results yielded groupings, which plausibly corresponded to visually discernible patterns, resulting in eleven clusters. To ensure sufficient non-zero training data in each cluster for the hurdle regression models, we further reduce this down to six clusters by iterating over the clusters and combining smaller clusters with their nearest neighbors based on centroid distance of polygons drawn around each cluster, requiring a minimum of 1000 *pgms* with non-zero fatalities. To subsequently assign any cells not containing conflict, we first draw new polygons around the combined grid cells of each cluster. Cells remaining outside these polygons were assigned to the cluster with the nearest boundary to the cell center. The procedure and resulting clusters are visualized in Figure 1. We subsequently train separate "local" models for each of the clusters following the same procedure as with the "global" models, described in the following section. Each grid cell is therefore assigned not only to exactly one cluster but also to one corresponding local set of models. Combining the predictions from all local models yields predictions for the whole geographical area of interest.

Hurdle Approach

Our overarching modeling approach is a variation on the principle of hurdle models. Hurdle models are a combination of two modeling steps: a first step consisting of a classifier determining whether the hurdle is reached, trained on all available training data, and a second step consisting of a regressor determining the predicted value, trained on only the subsection of the training data where the target has reached the hurdle. The hurdle technique relies on the assumption that different ranges of the data follow different distributions. This is often the case when the data is zero-inflated, as the distinction between zero and non-zero may follow a different logic than that determining the non-zero data value. This is a reasonable assumption for conflict data as well, on the one hand due to the nature of conflict event data generation, and on the other hand since the determinants of the intensity of conflict may differ from the determinants of the occurrence of conflict. Moreover, hurdle models have been demonstrated to work comparatively well in predicting conflict intensity (see e.g. Hegre et al. 2022b). We perform tuning and prediction for each of these two steps separately to allow for custom combinations of global and local predictions as further discussed below.

[4] We tested two additional versions of generating clusters: One with clusters created via an alternative clustering algorithm, DBSCAN, and a similar manual tuning of clustering parameters, and one with clusters corresponding to the United Nations Statistics Division sub-regions, with grid cells assigned to countries based on a majority rule. Both performed slightly worse in testing.

Classifier

Our classifiers are trained on a dummy variable indicating whether or not any fatalities occurred in a given *pgm*. We include two different classifiers in our modeling pipeline, random forests (Breiman 2001) and eXtreme gradient boosting (Chen & Guestrin 2016) models[5].

- **Random Forests (RF)** estimate probability by aggregating the results of a multitude of decision trees. Each tree in the forest is built from a different sample of data, using a technique called bootstrap aggregating, or bagging. Additionally, Random Forests employ random feature selection, where each split in a tree considers only a random subset of features.

- **eXtreme Gradient Boosting (XGB)**, employs a regularized learning objective that balances model complexity and predictive accuracy. The system utilizes gradient tree boosting, where the model is trained in an additive manner, incrementally improving the predictions by minimizing a loss function using second-order gradient statistics. Models are built sequentially, correcting the errors of previous models.

Our tuning metric of choice for classification is the average precision score, which summarizes the precision-recall-curve. This metric is well suited for zero-inflated classification tasks as it does not take into account whether zeroes are predicted correctly, which we argue is of little interest given the relative scarcity of violence (also see Saito & Rehmsmeier 2015). Tuning performance for both model types is fairly similar, with RF performing slightly better in all global models, while XGB is favored for most local models.

Regressor

To estimate uncertainty around predictions, we rely on regressors designed to output distributions directly rather than trying to estimate distributions around point predictions ex post. We include three different tree-based distributional regressors in our modeling pipeline, quantile regression forests (Meinshausen 2006), distributional random forests (Cevid et al. 2022) and natural gradient boosting for probabilistic regression (Duan et al. 2019):

- **Quantile Regression Forests (QRF)** extend the RF methodology to estimate conditional quantiles. Like RF, the QRF algorithm involves growing an ensemble of decision trees using a randomized node and split point selection process. Unlike traditional RFs, which only retain the mean response in each leaf, QRF retains all observed responses, enabling the estimation of the entire conditional distribution. The algorithm calculates the conditional quantile by averaging the

[5] We also tested logit models as a "simple" alternative approach, which performed worse by a factor of 2-3 on average.

weighted distribution of observed responses, with weights derived from the original RF methodology. We use evenly spaced quantile steps to generate the samples for our predictions with this algorithm.

- **Distributional Random Forests (DRF)** are another extension of the traditional RF framework used to estimate the entire conditional distribution of univariate or multivariate responses. The methodology involves constructing trees that split data points based on a novel criterion derived from the Maximum Mean Discrepancy (MMD) statistic, which measures differences in distributions rather than just differences in means. This splitting criterion is applied recursively to ensure that the distributions in the resulting child nodes are as homogeneous as possible. Each tree in the forest is grown to optimize this distributional metric. The final forest model uses a weighted combination of trees to estimate the full conditional distribution of the response variables. This approach allows DRF to adaptively weight training data points based on their relevance to the prediction, providing a robust and flexible method for modeling complex dependencies.

- **Natural Gradient Boosting (NGB)** for probabilistic regression extends gradient boosting to the estimation of probability distributions. This involves boosting the parameters of a specified parametric distribution using a natural gradient, which corrects the training dynamics for more stable and efficient learning. The algorithm integrates three modular components: a simple base learner, a parametric probability distribution, and a proper scoring rule. The natural gradient is employed to optimize the parameters of the conditional distribution, ensuring that the updates are invariant to reparameterization and efficiently exploit the curvature of the score in distributional space. We use decision trees as the base learner, log-normal probability distributions and the CRPS as the scoring rule.

Following the principle of hurdle models, we train our models only on *pgms* with non-zero targets while still generating predictions for all *pgms*. As our tuning metric, we use the competition's main metric, the Continuous Ranked Probability Score (CRPS) (see Hegre et al. 2023). In line with the maximum number of samples allowed by the prediction challenge, we set our regression models to output 1000 samples of the predicted distribution. This results in a wider range of possible values, ensuring the inclusion of low-probability outcomes. NGB performed best during tuning, being chosen in 75% of global models and 80% of local models, with the other two regressors only chosen occasionally.

Quasi-Hurdle Ensemble

Hurdle model point predictions are usually generated via a simple multiplication of the output of the classifier and the predicted value of the regressor. Given that we work

with N=1000 samples instead, a multiplication of the classification probability with each of the samples would result in non-integer predictions, which is not in line with the nature of fatality counts. At the same time, our tree-based regressors trained only on non-zero targets never produce zero predictions[6] and a multiplication would therefore likely overestimate the probability of violence occurring. While both issues could be partially addressed with rounding, we opt to instead interpret the classification probability as the percentage of the ensemble sample taken from the non-zero predictions, with the remaining share of the 1000 samples filled with zero values. In testing, this also performed better than the multiplicative approach.[7] We combine the global and local classification and regression predictions separately to generate one purely global and one purely local set of final predictions.

Global-Local Ensemble

While the GRF algorithm inspiring our local approach combines local and global predictions via weighted means, we found this not be beneficial during testing, with performance always falling between the two "pure" predictions. However, our approach allows us to selectively pick and choose the components for the hurdle ensemble from either the global or the local model, creating a third set of predictions combining global and local model outputs. We do so by testing the combined performance of the hurdle ensemble for each of the four possible combinations of classification and regression predictions on a cluster-by-cluster basis, selecting the global-local combination for each cluster which performed best across the three years[8] prior to a given prediction window.

Evaluation on Test Windows

To judge the performance of our three ensemble predictions[9], we compare them to several benchmarks provided by the ViEWS team across the six yearly test windows. Those are two naive benchmarks and three "conflictology" benchmarks based on medium- to long-term conflict history. The naive benchmarks are samples drawn from a Poisson distribution centered around the last observed values for each grid cell (Hegre

[6] The only exception here is NGB which does also predict zeroes. For consistency with the other algorithms we replace all zeros with ones in the NGB predictions.

[7] We also tested selecting either all-zero samples or the full non-zero sample based on the predicted probability from the classifier and a threshold, which performed significantly worse.

[8] We also tested this using only one or two years of prior data, which resulted in worse performance of the combined prediction. This is likely connected to a fairly high volatility in performance across years, with prior performance not correlated enough to future performance. An increase to five years of prior data did also not lead to meaningful improvements.

[9] For simplicity, we subsequently use the term "model" in this context to refer to the respective specification used to generate our global-only, local-only, or global-local ensemble prediction.

et al. 2023) and predictions with only zero values, which we initially included based on the extremely high proportion of zeros in the target and which have since also been added to the suite of benchmarks by the ViEWS team. The "conflictology" benchmarks all treat historic fatality counts as draws from the predictive distribution to generate forecasts. The first benchmark ("conflictology") uses fatality counts from a specific grid cell during the previous 12 months, for the respective prediction window (12 draws). The second benchmark ("conflictology neighbors") follows the same principle, but uses the combined conflict history of the grid cell and its immediate neighbors (108 draws). The third benchmark ("bootstrap 240") draws 1000 random samples from the grid cell's conflict history of the last 240 months. All also adhere to the two-months gap between training and test data. For instance, the "conflictology" samples for all months in 2018 are based on the observed fatalities from November 2016 to October 2017.[10]

Model	Year	CRPS	IGN	MIS	MSE	MAE
Global	**2018**	**0.1300**	**0.0568**	**3.4406**	60.554	0.3106
Local	**2018**	0.1376	0.0688	4.4840	**56.312**	0.4892
Global-local	**2018**	0.1324	0.0578	3.5510	61.223	0.3337
All-zero	**2018**	0.1444	0.0763	5.7765	65.815	**0.1444**
Poisson (last)	**2018**	0.3860	0.1001	13.879	169.82	0.4048
Conflictology	**2018**	0.1919	0.7887	4.6422	91.442	0.3510
Conf. neighbors	**2018**	0.1473	0.1559	5.2755	69.367	0.3651
Bootstrap 240	**2018**	0.1443	0.0770	5.7765	67.003	0.3108
Global	**2019**	**0.1010**	**0.0550**	**2.5629**	17.806	0.2381
Local	**2019**	0.1040	0.0668	3.5783	**16.915**	0.4141
Global-local	**2019**	0.1011	0.0553	2.6446	21.628	0.2732
All-zero	**2019**	0.1154	0.0777	4.6178	17.239	**0.1154**
Poisson (last)	**2019**	0.1442	0.0870	5.0394	18.237	0.1532
Conflictology	**2019**	0.1184	0.7860	3.1259	32.223	0.2180
Conf. neighbors	**2019**	0.1068	0.1539	3.2557	18.191	0.2320
Bootstrap 240	**2019**	0.1154	0.0783	4.6178	18.499	0.2861
Global	**2020**	**0.1180**	0.0622	3.1919	**16.187**	0.2065
Local	**2020**	0.1221	0.0730	4.1652	19.272	0.4123
Global-local	**2020**	0.1181	**0.0621**	**3.1773**	18.404	0.2269

[10] Description based on the source code at https://github.com/prio-data/prediction_competition_2023/blob/a45796ce8d1ffdd82e879e05c46d90c58b460a66/benchmark.py.

Model	Year	CRPS	IGN	MIS	MSE	MAE
All-zero	**2020**	0.1319	0.0900	5.2749	17.218	**0.1319**
Poisson (last)	**2020**	0.1646	0.0975	5.7744	18.785	0.1749
Conflictology	**2020**	0.1275	0.7893	3.6535	18.586	0.1895
Conf. neighbors	**2020**	0.1230	0.1595	3.5811	16.653	0.1994
Bootstrap 240	**2020**	0.1317	0.0898	5.2749	18.132	0.2872
Global	**2021**	0.9246	**0.0690**	35.296	81843.2	1.0130
Local	**2021**	0.9293	0.0795	36.308	81844.3	1.2107
Global-local	**2021**	**0.9243**	**0.0690**	35.233	81843.1	1.0204
All-zero	**2021**	0.9398	0.0979	37.592	81844.9	**0.9398**
Poisson (last)	**2021**	0.9703	0.1061	37.924	81843.2	0.9813
Conflictology	**2021**	0.9302	0.7930	**35.048**	**81841.8**	0.9991
Conf. neighbors	**2021**	0.9279	0.1653	35.454	81843.0	1.0189
Bootstrap 240	**2021**	0.9396	0.0972	37.592	81844.9	1.0794
Global	**2022**	1.1274	0.0676	43.974	98532.3	1.2461
Local	**2022**	1.1289	0.0775	44.748	98528.7	1.4561
Global-local	**2022**	**1.1263**	**0.0674**	43.818	98519.6	1.2767
All-zero	**2022**	1.1375	0.0965	45.499	98560.1	**1.1375**
Poisson (last)	**2022**	1.4565	0.1180	56.598	98771.2	1.4768
Conflictology	**2022**	1.1419	0.7939	**43.386**	**98183.2**	1.2833
Conf. neighbors	**2022**	1.1311	0.1645	44.069	98440.0	1.2966
Bootstrap 240	**2022**	1.1373	0.0954	45.499	98560.0	1.2842
Global	**2023**	**0.2147**	**0.0720**	**7.3453**	189.74	0.3590
Local	**2023**	0.2207	0.0821	8.7784	212.79	1.1777
Global-local	**2023**	0.2175	0.0721	8.1044	239.99	0.6551
All-zero	**2023**	0.2236	0.0996	8.9446	**163.33**	**0.2236**
Poisson (last)	**2023**	9.7500	0.1268	387.10	1134057	9.7827
Conflictology	**2023**	0.5237	0.7958	17.030	34704.4	1.9570
Conf. neighbors	**2023**	0.2499	0.1677	12.133	4027.1	1.9647
Bootstrap 240	**2023**	0.2234	0.0985	8.9446	171.23	0.4475

Table 1: Overview of naive model and benchmark metrics. Best results for each year are marked in bold. Lower scores signify better performance. Note that the CRPS is always equal to the MAE in the case of all-zero predictions.

The evaluation results for our models and the benchmarks are reported in Table 1. We base our evaluation on the challenge's main metric CRPS and the two additional

metrics specified in the prediction challenge (Hegre et al. 2023), the Ignorance Score (IGN) and the Mean Interval Score (MIS), all of which evaluate samples from a distribution. As the results show, our three models are reliably able to beat the benchmarks, with very few exceptions. Our global model performs slightly better than the local model, while the global-local combination performs about the same as the global model. Depending on the metric, it even pulls ahead in two to three out of the six test windows. Despite this, we **prefer the global model as our main submission** for the prediction challenge, as it is the simpler approach. We also provide the other sets of predictions for comparison. While this means we are unable to exploit systematic differences across contexts with our approach to improve predictive performance, the good performance of the local models still opens interesting new avenues for further research, e.g. through the inclusion of only locally available data. However, differences in performance between all three models as well as improvements over most benchmarks are miniscule, which makes us hesitant to draw systematic conclusions based on the characteristics of the individual scores.

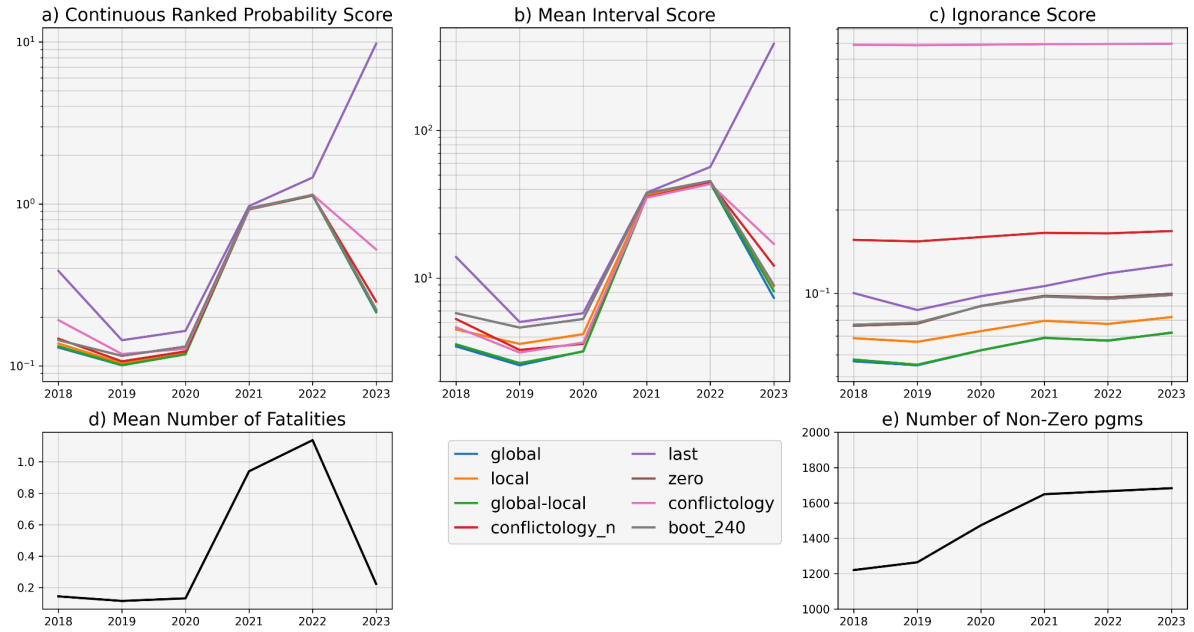

Figure 2: Performance of models and benchmark for all test windows over time (a-c) on a log scale. As the comparison shows, CRPS and MIS are highly correlated with the mean number of fatalities (d) while the IGN is strongly correlated to the number of non-zero observations (e).

The CRPS and MIS are almost perfectly correlated with the number of fatalities (> 0.99), while the mean IGN score strongly correlates to the count of non-zero *pgms* (> 0.96) for most models and benchmarks (Figure 2).[11] Comparison across years is therefore likely not appropriate. Following this correlation the performance on the CRPS and MIS for all models drops significantly in 2021 and 2022, two years with a

[11] This holds true even for random predictions in a cursory examination of this relationship, leading us to believe this is a characteristic of the metrics in relation to the highly zero-inflated observed actuals.

comparatively high number of (mean) fatalities in the data. This is largely driven by single outliers in the data with very high fatality counts[12], while the number of non-zero observations and therefore also the IGN stays more consistent across all years. 2021 and 2022 are also the only test windows where our models fall behind the benchmarks on the MIS, although this is not necessarily related.

While somewhat outside the scope of this prediction challeng, we complement this analysis with the mean squared error (MSE) and the mean absolute error (MAE) evaluated against the sample mean to better understand where the mean predictions of our samples fall. The MSE reacts more strongly to severe misses, while these barely impact the MAE. Given the zero-inflated nature of the data, the MAE strongly favors the all-zero benchmark, with no clear winner emerging from the comparison of the other models. Our best model tends to be slightly ahead of the benchmarks in the MSE, indicating a fairly well-centered distribution, until the score breaks down in 2021 and 2022 due to the large outliers. In 2023 we do not suffer the same penalty as the recent history based benchmarks when the mean number of fatalities returns to its prior "normal".

Github

The complete code of our contribution to the 2023/24 ViEWS prediction challenge including our predictions is available at https://github.com/DaMitti/views_competition_unibw_trees.

References

Bergstra, James; Rémi Bardenet, Yoshua Bengio & Balázs Kégl (2011) Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems* 24: 1–9.

Bergstra, James; Daniel Yamins & David Cox (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *International Conference on Machine Learning*: 115–123.

Breiman, Leo (2001) Random Forests. *Machine Learning* 45(1): 5–32.

Campello, Ricardo J. G. B.; Davoud Moulavi & Joerg Sander (2013) Density-Based Clustering Based on Hierarchical Density Estimates. In: Jian Pei (ed.) Advances in knowledge discovery and data mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold

[12] Dropping the largest outlying *pgm* from the 2022 data, which has more than 100.000 fatalities, results in a roughly 3-4x improvement of the metric values; e.g. the CRPS for our global model changes from 1.1274 to 0.3527 (3.6x reduction) and the MIS changes from 43.974 to 13.001 (3.4x reduction).

Coast, Australia, April 14-17, 2013 proceedings. LNCS sublibrary. SL 7, Artificial intelligence 7818-7819. Berlin, New York: Springer, 160–172.

Cevid, Domagoj; Loris Michel, Jeffrey Näf, Peter Bühlmann & Nicolai Meinshausen (2022) Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression. *Journal of Machine Learning Research* 23: 1–79.

Chadefaux, Thomas (2017) Conflict forecasting and its limits. *Data Science* 1(1-2): 7–17.

Chen, Tianqi & Carlos Guestrin (2016) XGBoost: A Scalable Tree Boosting System (https://arxiv.org/pdf/1603.02754).

Davies, Shawn, Therese Pettersson & Magnus Öberg (2023). Organized violence 1989-2022 and the return of conflicts between states?. *Journal of Peace Research* 60(4).

Duan, Tony; Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Y. Ng & Alejandro Schuler (2019) NGBoost: Natural Gradient Boosting for Probabilistic Prediction (http://arxiv.org/pdf/1910.03225).

Pedregosa, Fabian; Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85): 2825–2830.

Georganos, Stefanos; Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff & Stamatis Kalogirou (2021) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International* 36(2): 121–136.

Grinsztajn, Leo; Edouard Oyallon & Gael Varoquaux (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems* 35: 507–520.

Hegre, Håvard; Angelica Lindqvist-McGowan, Paola Vesco, James Dale, Mihai Croicu & David Randahl (2022) Forecasting fatalities in armed conflict: Forecasts for April 2022–March 2025 (https://www.prio.org/publications/13338).

Hegre, Håvard; Forogh Akbari, Mihai Croicu, James Dale, Tim Gåsste, Remco Jansen, Peder Landsverk, Maxine Leis, Angelica Lindqvist-McGowan, Hannes Mueller, Malika Rakhmankulova, David Randahl, Christopher Rauh, Espen Geelmuyden Rød & Paola Vesco (2022). Forecasting fatalities (http://uu.diva-portal.org/smash/record.jsf?dswid=939&pid=diva2%3A1667048).

Hegre, Håvard; Paola Vesco, Michael Colaresi & Jonas Vestby (2023) The 2023/24 VIEWS Prediction competition: Predicting the number of fatalities in armed conflict, with uncertainty (https://viewsforecasting.org/wp-content/uploads/VIEWS_2023.24_Prediction_Competition_Invitation.pdf).

Hegre, Håvard et al. (Forthcoming) The 2023/24 VIEWS prediction competition. *Journal of Peace Research*, XXX.

Johnson, Reid A. (2024) quantile-forest: A Python Package for Quantile Regression Forests. *Journal of Open Source Software* 9(93): 5976.

Meinshausen, Nicolai (2006) Quantile Regression Forests. *Journal of Machine Learning Research* 7: 983–999.

Mueller, Hannes & Christopher Rauh (2022) The Hard Problem of Prediction for Conflict Prevention. *Journal of the European Economic Association* 20(6): 2440–2467.

Näf, Jeffrey (2022) DRF: A Random Forest for (almost) everything - Towards Data Science (https://towardsdatascience.com/drf-a-random-forest-for-almost-everything-625fa5c3bcb8).

Saito, Takaya & Marc Rehmsmeier (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10(3): e0118432.

Sundberg, Ralph and Erik Melander (2013) Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research* 50(4).

Vesco, Paola; Håvard Hegre, Michael Colaresi, Remco Bastiaan Jansen, Adeline Lo, Gregor Reisch & Nils B. Weidmann (2022) United They Stand: Findings from an Escalation Prediction Competition. *International Interactions*: 1–37.