Forecasting fatalities

The ViEWS team*[1, 2]

[1]Department of Peace and Conflict Research, Uppsala University
[2]Peace Research Institute Oslo (PRIO)

31 May 2022

*The report has been developed with input from Forogh Akbari, Mihai Croicu, James Dale, Tim Gåsste, Håvard Hegre, Remco Jansen, Peder Landsverk, Maxine Leis, Angelica Lindqvist-McGowan, Hannes Mueller, Malika Rakhmankulova, David Randahl, Christopher Rauh, Espen Geelmuyden Rød, and Paola Vesco.

Contents

Acknowledgements

This material has been funded by UK aid from the UK government; however the views expressed do not necessarily reflect the UK government's official policies. It builds on work funded by the European Research Council (H2020-ERC-2015-AdG 694640, ViEWS), Uppsala University, and the United Nations Economic and Social Commission for West Asia.

The forecasts were computed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

For more information about ViEWS, please see `https://viewsforecasting.org`.

1 Introduction

1.1 The political Violence Early-Warning System (ViEWS)

Knowing the dire consequences of armed conflict, preventing and containing future conflicts are high on policy-makers' agenda. Early action, however, requires early warning (World Bank Group and United Nations, 2017). Moreover, such warnings must come in a form that prevents us from forgetting about crises that already exist and persist. With a systematic and objective understanding of when, where, and for how long future conflicts will last, as well as how lethal they will be, the international community can come together to make timely and evidence-based strategic decisions to prevent or mitigate future conflicts, engage in diplomacy efforts, and allocate resources where most needed. This is what the political Violence Early-Warning System (ViEWS) (Hegre et al., 2021, `http://viewsforecasting.org`) offers.

A number of quantitative early-warning systems have recently been developed and are running with regular updates of their risk assessments, ViEWS is however by far the most comprehensive and ambitious of such systems.[1] It is an early-warning system at the frontier of research that provides monthly forecasts of impending violence 1–36 months into the future, at two levels of analysis, and for each of three different types of political violence. The system is based on well-established academic research on the causes and correlates of conflict, and consequently draws on a variety of predictors. Moreover, the system only makes use of publicly available data in order to allow for maximum transparency, further ensured by conducting – and publishing – continuous evaluations of its predictive performance (see e.g. Hegre et al., 2019; Hegre et al., 2021).[2]

[1] Prominent examples are the Early Warning Project `https://earlywarningproject.ushmm.org`; the Atrocity Forecasting project `https://politicsir.cass.anu.edu.au/research/projects/atrocity-forecasting`; the Water, Peace and Security project `https://waterpeacesecurity.org/info/methodology`).

[2] In the version of the prediction model presented here, we include data from Political Risk Services which is not publicly available. We will remove these data when finalizing the continuously updated version of the model.

ViEWS has been publishing monthly updates of its forecasts for all of Africa since July 2018 as part of a research project funded by the European Research Council and Uppsala University. In 2021, a separate instance of the system was developed with funding from UN ESCWA that expanded the geographic scope to the Middle East. As such, it covers countries that not only account for about half of the global population as of 2022, but which have been the location of about 90% of the conflict-related fatalities across the globe over the past 20 years. With funding from the GRSA fund at the UK FCDO and other sources, a third iteration of the ViEWS system was developed over 2021–2022 using the expanded scope from the ESCWA development and a new set of models that allows the system to progress from offering dichotomous (conflict/no conflict) predictions to also forecasting the number of fatalities expected in impending conflicts. Being able to predict not only whether a certain threshold of fatalities will be reached, but the number of fatalities expected, has significantly pushed the scientific envelope, and will provide policy-makers and researchers with the ability to quantify the potential impact and intensity of conflicts.

1.2 The value of – and challenge in – using a continuous prediction target

While the dichotomous conflict/no conflict forecasts that early-warning systems – including ViEWS – have relied upon to date have been a major contribution to researchers and policy-makers alike, the limitation of whether violence will exceed a given threshold of violence or not results in an unfortunate loss of valuable information. If an alert threshold is set high (e.g. at 500 fatalities per month), focus will be on high-impact cases and shift attention away from cases that are less serious but not negligible, such as recent simmering conflicts in Tunisia, Kenya, and Saudi Arabia. From a methodological point of view, a high threshold also means there are fewer cases of violence to learn from, hurting the precision of prediction models. If the threshold is set low (e.g. at 25 per year), the models have numerous cases to learn from, but the applicable cases will not distinguish between relatively minor incidents like the ones mentioned above and major conflagrations such as the Syrian civil war and the genocide in Rwanda; all conflicts with a high probability of exceeding 25 deaths in a month will receive the same level of risk alert. Moreover, the indirect impacts of wars depend not only on the presence and length of violence, but are also proportional to the number of people killed in fighting (Ghobarah, Huth, and Russett, 2004).

Refining an early-warning system to indicate whether a future conflict will cause e.g. 100, 1000, or 10,000 deaths is, however, a challenging task and mainly the answer as to why this has not been embarked upon before. This was well illustrated by the insightful contributions to the conflict prediction competition hosted by ViEWS in 2020, in which the common denominator amongst the competing research teams was a difficulty to perform better than predicting 'no change from last period' (Hegre, Vesco, and Colaresi, 2022; Vesco et al., 2022). Discussed at length in Section 2, the reason behind this predominantly concerns the distribution of the number of fatalities in past conflicts: for a large portion of the country-months over 1990-2020, the Uppsala Conflict Data Program (UCDP) recorded no violence at all; in about half of the remaining months, more than 25 fatalities were recorded; and in about 100 country-months, the more than 2,500 people were killed. On 11 occasions – in Iraq, Syria, Ethiopia and Eritrea—the conflicts took

more than 10,000 lives in a single month.[3]

Most statistical models are ill-equipped to model such distributions. Building on the research carried out by the ViEWS project and others over recent years, this has however now become feasible.

1.3 Structure of the paper

In this paper, we present the result of embarking upon the endeavor above: a production-level forecasting model that predicts the number of fatalities in impending conflict 1–36 months into the future, at two levels of analysis, and for each of three different types of political violence. We start with an overview of the dependent variable and a deeper discussion of the historic distribution of fatalities, after which we present our approach to predicting the three outcomes. We describe the constituent models informing the forecasts, the ensembling techniques that allow us to draw upon the strengths of each of the constituent models when producing the final forecasts, and we detail our calibration procedures. In Section 5, we present the true forecasts from the prediction model and offer some tools for interpretation of the results. Section 6 presents an evaluation of the predictive performance of the model.

2 The dependent variable

The outcome that the model predicts is armed conflict as defined and compiled by the Uppsala Conflict Data Program (UCDP, Gleditsch et al., 2002; Sundberg and Melander, 2013; Pettersson et al., 2021; Hegre et al., 2020). The UCDP collects data on three types of conflict (see `https://www.pcr.uu.se/research/ucdp/definitions/`):

State-based (sb) conflict The use of armed conflict over either government or territory between armed actors in which at least one is a government of a state.

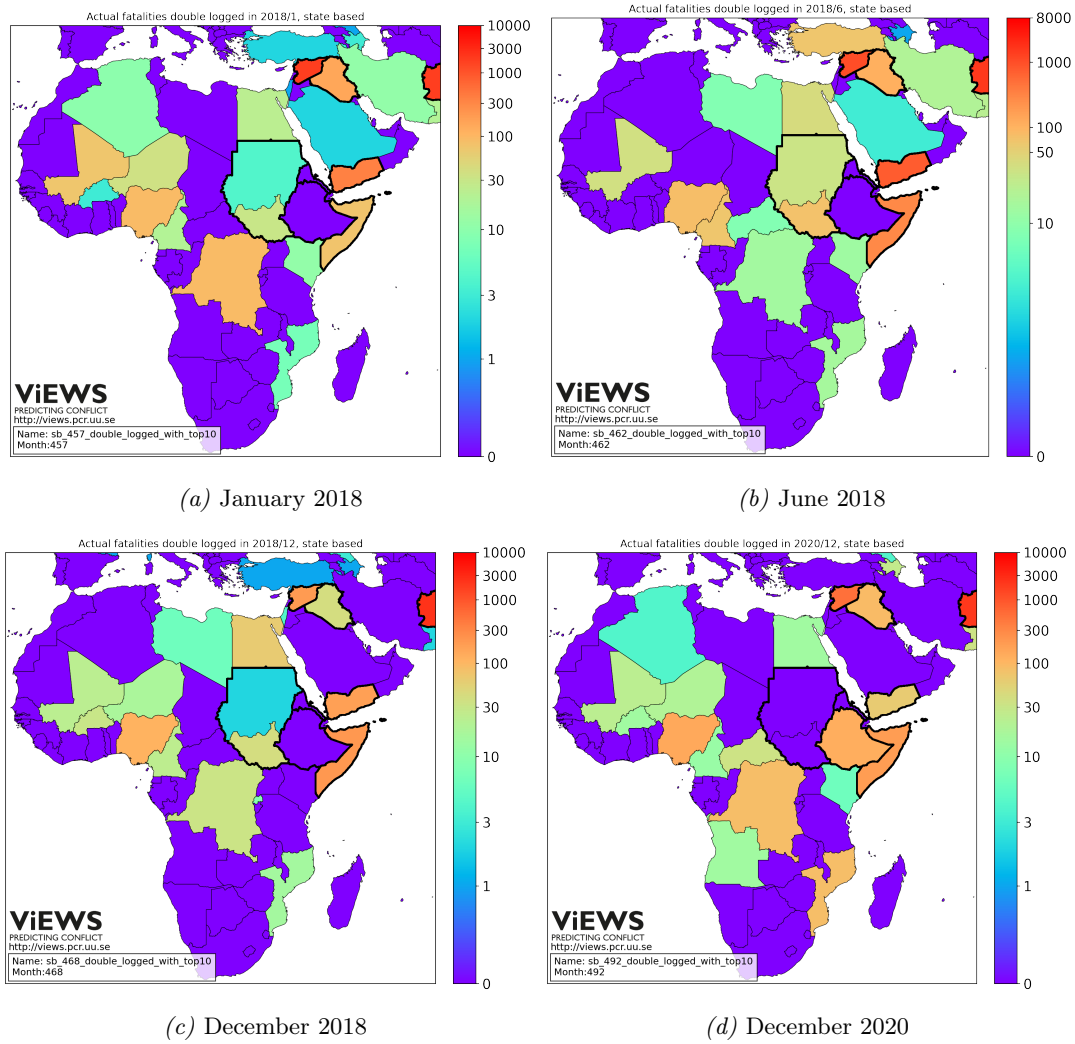
Non-state (ns) conflict The use of armed force between two or more organised armed groups, neither of which is a government of a state.

One sided (os) conflict The deliberate use of armed force by the government of a state or by a formally organised group against civilians.

The UCDP provides estimates for the number of persons killed in each of these conflict types for each of the conflict events they can document. We aggregate the fatalities across events into monthly sums, for countries and for the PRIO-GRID cell structure which divides the world into 0.5x0.5 decimal degree grids (Tollefsen, Strand, and Buhaug, 2012). We restrict forecasts to Africa and the Middle East.

[3] Had we included cases of genocide or wars before 1990 in this tabulation, we would have seen even more of these extremely violent events.

Figure 1. **Actual fatalities in Africa and Middle East, state-based conflict**



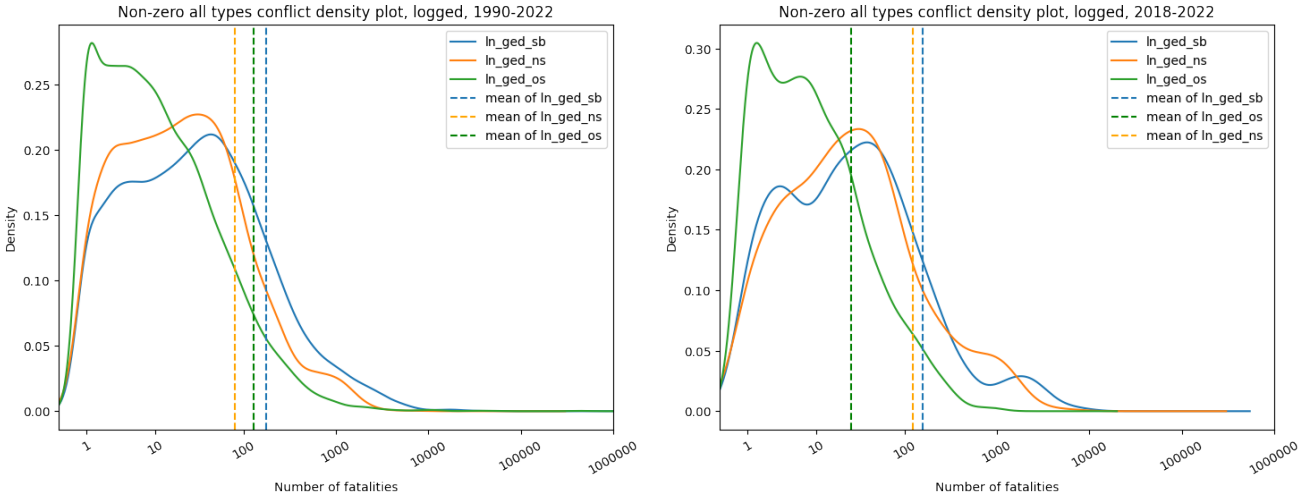
(a) January 2018 *(b)* June 2018

(c) December 2018 *(d)* December 2020

Note: The 10 countries with the most fatalities over the Jan 1990 to December 2020 period globally are marked off with thick black borders.
Source: `https://ucdp.uu.se`; Pettersson et al. (2021) and Hegre et al. (2020)

In this report, we will focus on forecasts for state-based conflict. This type of violence is the most frequent and deadly of the three UCDP categories, and the other two types frequently occur in the context of state-based violence. To that end, forecasts for state-based conflict often function as a forecast for the other two types. However, the forecasting system for state-based conflict is an excellent template for forecasting the other types, all the UCDP conflict data are available, and it is not much effort to implement similar models for the other types or even a combination of them.

Figure 2. **A challenge for forecasting models: The distribution of the outcome variables**



(a) 1990–2022 *(b)* 2018–2022

Note: Kernel density plots for all country-months with non-zero fatality counts, showing the complete period 1990–2022 as well as 2018–2022. The vertical lines show the mean (non-logged) fatality counts for the non-zero observations.
Source: UCDP GED, 2022

3 The historical distribution of fatalities

3.1 Country (cm) level

Figure 1 shows the recorded number of fatalities at the country level in map form for four selected months. The 10 countries with the most fatalities over the 1989–2020 period globally are marked off with thick black borders. Six of these are in the Africa and Middle East regions.[4]

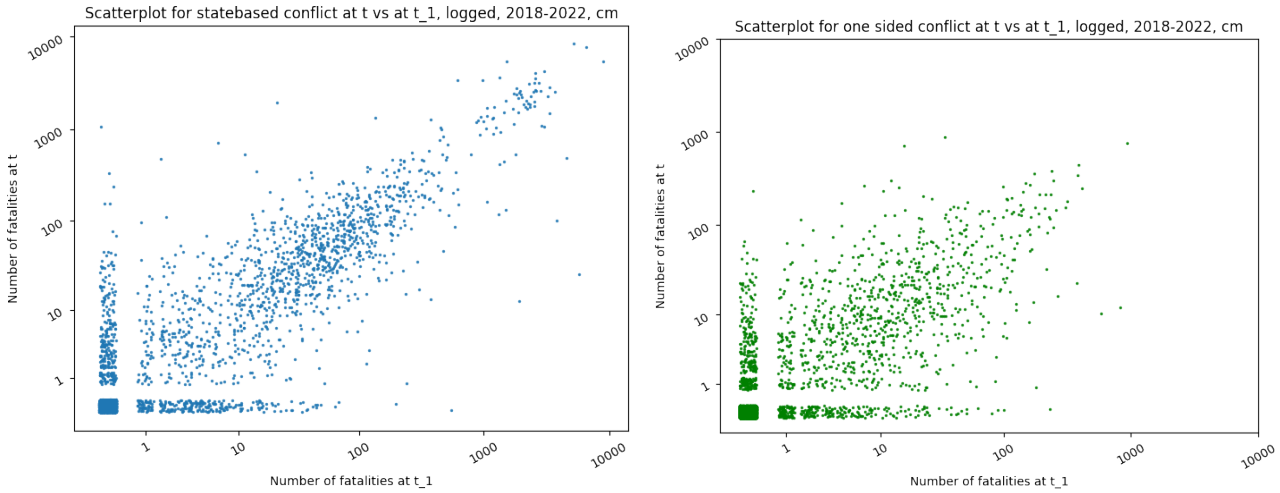
The outcome variable has a distribution that is challenging to forecast. Most observations between 1990 and 2020 are zeros (no conflict fatalities): at the country-month (cm) level, 87.5% of the observations are zeros, and at the PRIO-GRID month (pgm) level, 99.6% are zeros.

In addition to this 'zero inflation', the distribution of non-zero death counts is heavily right-skewed. Figure 2 shows a density plot of the distribution of fatality counts for all three types of violence, restricted to non-zero observations. The x axes are in log form for both sub-figures.

Over the 1990–2020 period, there were 8,700 country-months with state-based conflict. The median number of fatalities was 28 and the mean 173. We have marked off the (non-logged) means for non-zero observations with vertical dashed lines. As these descriptive statistics and the figures show, the distribution

[4]For state-based conflict, the 10 most fatal conflict countries were pre-1993 Ethiopia, post-1993 Ethiopia, pre-2011 Sudan, Iraq, Somalia, Sri Lanka, Afghanistan,Pakistan, Syria, and India. The top 10 countries for non-state violence are: Brazil, Mexico, Ethiopia, Sudan, Nigeria, Somalia, Congo (DRC), Libya, Syria, and India. The top 10 countries for one-sided violence are: Liberia, Sudan, Iraq, Nigeria, Bosnia and Herzegovina, Afghanistan, Rwanda, Congo(DRC), Syria, and India.

Figure 3. **Conflict breeds conflict: How fatality counts in one month for a country relates to fatality counts for the preceding month**



(a) State based conflict

(b) One sided conflict

Note: Scatter plot between conflict at t_1 and conflict at t for each type of conflict, 2018–2022. Observations are jittered to show the frequency of observations with similar values.
Source: UCDP GED, 2022

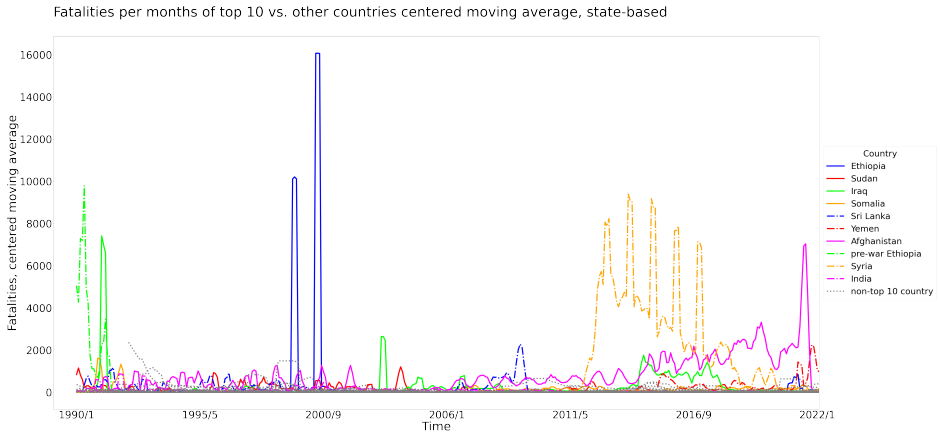
of non-zero fatality counts is heavily right-skewed. In 87 out of the 8,700 country-months, more than 2,500 people were killed in battle-related events. In one month (Ethiopia in June 2000), the UCDP recorded that more than 48,000 people died in a single month. The genocide in Rwanda is the most extreme observation in our post-1989 dataset, with close to 500,000 people killed in one-sided violence within a few weeks.[5]

The large number of zero observations as well as extremely high fatality counts is one challenging characteristic of the prediction problem. Another is that a large number of non-zero fatalities occur in the same country or location in subsequent months. Figure 3 shows how the number of deaths in one month in a country (vertical axis) relates to the number of deaths in the same country the month before (horizontal axis). For readability, the figure is restricted to the three years in the test period (2018–2020). Most non-zero observations follow another non-zero observation – a lagged dependent variable is a very strong predictor that we include in all models presented below. In a good number of country-months, however, fatality counts go from 0 to positive values, and even hundreds in the following month. Similarly, there are a good number of cases where substantial violence is followed by no deaths the month after. Predicting these spells of violence as well as when fatalities de-escalate to zero, is one of the most daunting tasks.

Figure A-3 shows the distribution of the fatality count variables (as in Figure 2) for all country-months where there was at least one fatality the year before.

[5]These extreme observations are somewhat exaggerated due to a weakness in our current dataset: The UCDP recorded 48,000 fatalities at the border of Ethiopia and Eritrea in the period January–June 2000. They do not have sufficient source material to identify the exact date of each violent event during this war, and code the violence as distributed across these months. In our current aggregation procedure these fatalities are assigned to the last of these months. Similarly, the genocide in Rwanda is assigned to May 1994 although the violence occurred over the April and May period. We have written a revised aggregation procedure to handle this, and will update the data in the next iteration of this report.

Figure 4. **Time series for top 10 cumulative fatality countries vs all other countries, state-based violence, centered moving average**



Note: A three-month centered moving average means that the value shown for March 2016 is the average of fatalities over the three-month period February–April 2016; the value for April 2016 the average for March–May, etc.
Source: UCDP GED, 2022

In combination, these distributional aspects mean that a very large fraction of the battle-related deaths have occurred in a small number of countries. Figure 4 shows the number of fatalities per month for the 10 most deadly conflict countries over the past 30 years, as well as the total number of fatalities in all other conflicts. We identified the 10 most deadly countries by summing up all fatalities by conflict sub-type.[6]

Figure 4 shows that the global total of state-based violence over the 1990–2020 period was dominated by the Eritrean secessionist war (listed as in Ethiopia), Iraq (multiple wars from the first Gulf war and onwards), Sri Lanka, Syria, and Afghanistan.[7]

Figure 5 shows the global total number of fatalities across all countries for the 1990–2022 period.

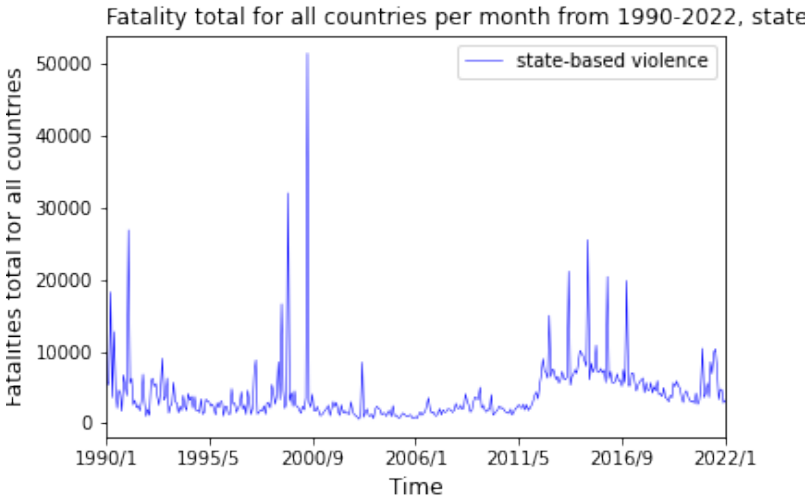
3.2 Geographical (*pgm*) level

Figure 6 shows where the UCDP recorded fatalities for four selected months in the test period (Pettersson et al., 2021; Hegre et al., 2020). The fatality counts are aggregated to the total number of deaths in each PRIO-GRID cell per month (see Tollefsen, Strand, and Buhaug, 2012, for a presentation of the PRIO grid). Figure 7 shows the distribution of fatalities over these geographical cells, restricted to PRIO-GRID months where there was at least one fatality. The distribution is even more right-skewed than for the country (*cm*) level (Figure 2). The median number of fatalities lay between 1 and 3, but a sizeable proportion exceeds 100. Even though the PRIO-GRID cells are small, about 55x55km at the equator, Rwanda only occupies seven such cells. Hence, the 1994 genocide (classified as one-sided violence by the UCDP) did not only occur in a very short time span, but also in a very condensed area. In principle, forecasting models should

[6]We aggregated counts by country ID. Following Weidmann, Kuse, and Gleditsch (2010), some countries are assigned a new distinct country IDs when its territory changes. For that reason, countries can appear multiple times in the figures.

[7]The spikes for Syria are due to an incorrect aggregation of annual data to individual months, to be corrected in the next version of the report.

Figure 5. **Time series for fatalities for all countries, state-based conflict, 1990–2022**



Source: UCDP GED, 2022

be able to make forecasts that incorporate such extreme events if possible.

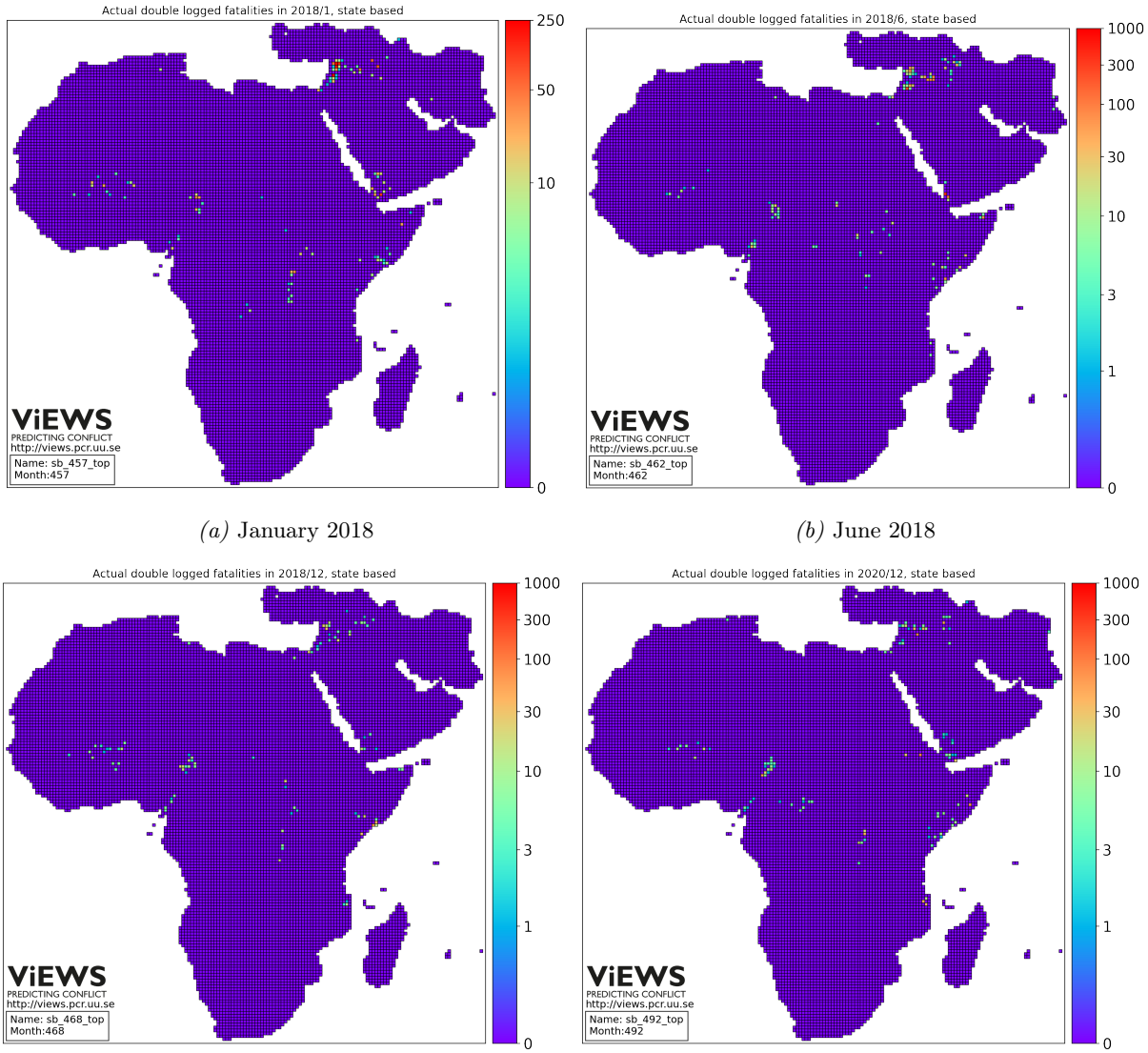
Several current conflict hotspots are similarly concentrated. The recent violence in the Tigray province occurs in only two PRIO-GRID cells, and that in Eastern DRC mainly affected a narrow, but densely populated strip along the borders to Rwanda and Uganda. Still, even when geographically concentrated, the fighting often spills over national borders, such as in the North of Nigeria and Cameroon, and in the region straddling Mali, Burkina Faso, and Niger.

3.3 Summary of review of the outcome variables

It is clear from this discussion that the outcome we develop the model for has a very challenging distribution. Most observations are zeros, and on top of that the non-zero observations are highly right-skewed. The really serious conflict occasions are fortunately quite rare. However, these rare instances are also the ones that grab most attention, and by definition affect a large number of people. Accordingly, forecasting models should be designed so that they are able to warn about these. In the section describing the *cm* models below we discuss briefly how the models succeed in capturing the distribution described here, including the rare events, to prepare for continued model development.

The descriptive statistics has also revealed some problems with the data we are currently using. These are not really errors, but are due to using a simple procedure to treat known measurement uncertainty. We have a solution to this that we will implement.

Figure 6. **Actual fatalities in Africa and Middle East, state-based conflict at PRIO-GRID (*pgm*) level**



(a) January 2018

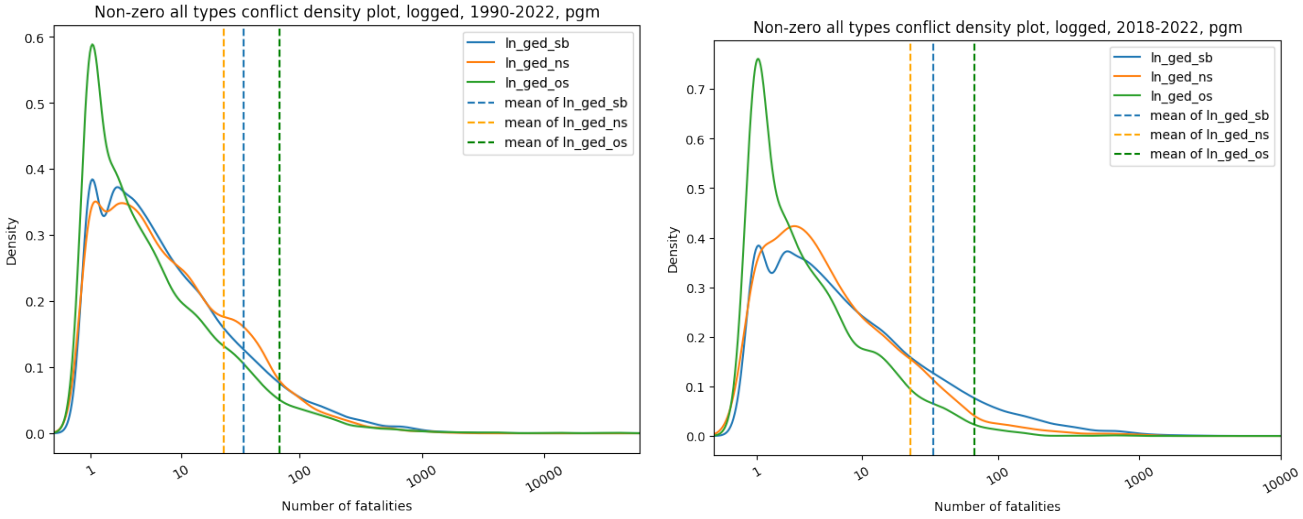
(b) June 2018

(c) December 2018

(d) December 2020

Source: UCDP GED, 2020

Figure 7. **Actual fatalities in Africa and Middle East, state-based conflict at PRIO-GRID (*pgm*) level**



(a) 1990–2020 *(b)* 2018–2020

Note: Kernel density plots for all PRIO-GRID months with non-zero fatality counts, showing the complete period 1990–2022 as well as 2018–2022, the three-year period used for evaluation of forecasts.
Source: UCDP GED and UCDP Candidate, 2022, visualized by ViEWS

4 Generating the forecasts

The forecasting system is an ensemble or collection of a set of constituent models. We first describe the constituent models, and then our approach to ensembling. The results from running these models are shown in Section 6.

4.1 Constituent models

The team has developed a set of forecasting models at the *cm* and *pgm* levels. At the *cm* level, the current setup explores 48 models. We have included more models than necessary in order to evaluate the relative usefulness of the different algorithms we specify as well as the feature sets.

The 48 models are combinations of nine feature sets and ten different machine-learning algorithms. In the figures and tables that follow, the models are labeled by their feature sets and their algorithms; e.g., the model **fat_conflicthistory_hurdle_xgb** is a fatalities model using the conflict history feature set and an XGB-based hurdle-regression algorithm.

Section 4.1.1 describes the feature sets at the country level in more detail, Section 4.1.2 the feature sets at the geographical level, and Section 4.1.3 the machine-learning algorithms. We review the relative contribution of the various algorithms and feature sets in Section 6.1.2.

Table 1. **Models at *cm* level: combinations of feature sets and algorithms**

	Algorithm										
Feature set	RF (XGB)	RF (Scikit)	GBM (scikit)	XGB	hist GBM (scikit)	LGB	LGB/LGB hurdle	RF/RF hurdle	XGB/XGB hurdle	Markov GLM	Markov RF
Baseline	X										
Conflict hist.	X	X	X	X	X	X	X	X	X		
Long conflict hist.	X			X							
V-Dem	X			X					X		
WDI	X			X					X		
Topics	X			X	X				X		
PRS	X			X					X		
Broad	X			X					X		
Greatest hits	X			X					X		
hh20	X	X	X	X	X	X	X	X	X	X	X
PCA all				X							
PCA topics				X		X					
PCA V-Dem						X					
PCA WDI						X					

Source: ViEWS, 2022

4.1.1 *cm*-level feature sets

baseline is a very simple model with only five data columns (each column representing one feature): The number of fatalities in the same country at $t-1$, three 'decay functions' of time since there was at least five fatalities in a single month, for each of the UCDP conflict types – state-based, one-sided, or non-state conflict – and log population size (Hegre et al., 2020; Pettersson et al., 2021).[8] The features in the baseline are included in all the models described below. This ensures that all models in the ensemble provides at least moderately good predictions, while guaranteeing diversity in feature sets and modelling approaches.

conflicthistory is a collection of 28 variables that together map the conflict history of a country. The features include lagged dependent variables for each conflict type as coded by the UCDP (state-based, one-sided, or non-state) for up to each of the preceding six months, 'decay functions' of time since conflict caused 5, 100, and 500 deaths in a month, for each type of violence, whether ACLED (Raleigh et al., 2010) recorded similar violence, and whether there was recent violence in any neighboring countries.

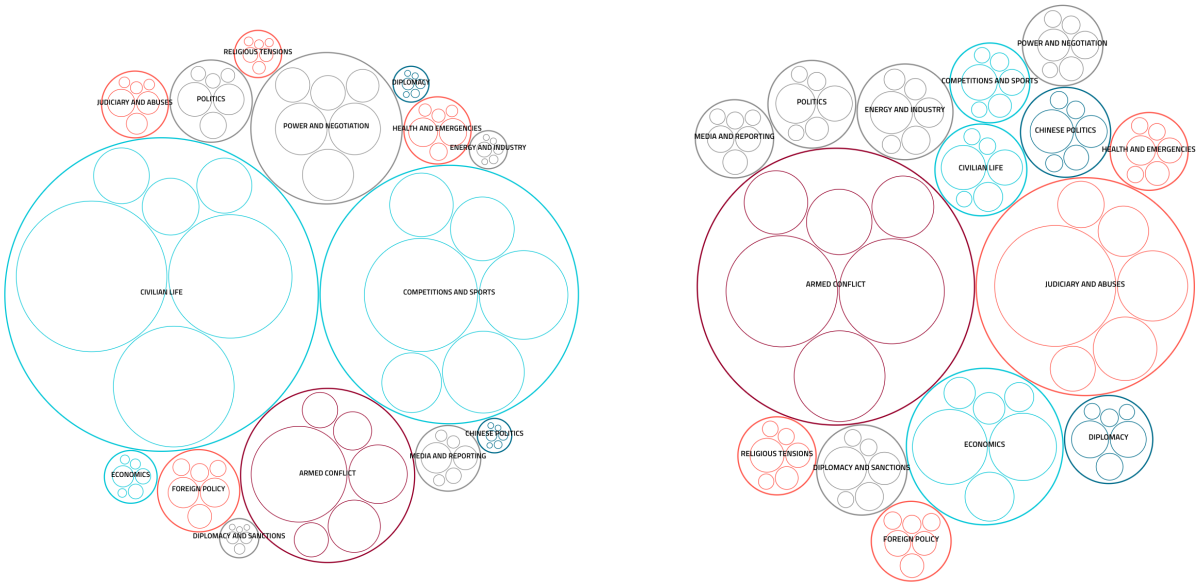
vdem includes about 60 features drawn from the Varieties of Democracy project (Coppedge et al., 2020) as well as from World Development Indicators (WDI WorldBank, 2019), as well as the baseline model features from the UCDP. The most important features are a number of conflict history lags and some WDI features (see below). Among the Varieties of Democracy features, the most important are indicators for horizontal and vertical accountability, clientilism, divided parliament control, and indicators of exclusion/discrimination.

wdi is composed of about 40 features drawn from the WDI as well as some conflict history indicators. Following the conflict history indicators, the most important predictors are indicators of migra-

[8] The 'decay function' is $e^{-tsc/\alpha}$ where tsc is the number of months since five fatalities, and α a half-life parameter.

tion and refugee population, military expenditure, international aid, total population, fertility, and population growth.

Figure 8. **The topic model of Mueller and Rauh (2020): Dominant topics in news coverage of Burkina Faso, 2017**



(a) Burkina Faso, January 2017

(b) Burkina Faso, April 2017

Source: Mueller and Rauh (2020)

topics has a number of features from Mueller and Rauh (2018), Mueller and Rauh (2020), and Mueller and Rauh (2022). The features are constructed from 3.5 million newspaper articles that are processed into 15 topics using Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003). In addition, the model contains a set of conflict history features, population size, child mortality, and some democracy features.

Figure 8 illustrates the topics for Burkina Faso. Blue topics are associated with a reduced risk of subsequent armed conflict, whereas red topics signal an increased risk. Each topic has been assigned a label in Mueller and Rauh (2020). Figure 8a shows the dominant topics in news covering the country in January 2017. 'Armed conflict' is present in the news, but do not receive nearly as much coverage as 'Civilian life' and 'Competition and sports'. In April 2017 (8b), 'Armed conflict' is the most important topic, followed by 'Judiciary and abuses' that also signals a risk of escalation.

PRS includes a number of features from the Political Risk Services (`https://www.prsgroup.com`), who make available at a monthly basis a number of political and economic indicators as well as updated risk assessments. Since the PRS data are not publicly available, they will not be part of a publicly available ViEWS system, but they have been included here to explore the extent to which they could potentially contribute to the forecasting performance (they make a discernible but very marginal contribution).

broad includes about 90 features taken from all the models described above. The most important features are, as in all models, a set of lagged conflict variables. After these, net migration, fertility rates, and a number of the Mueller & Rauh news topic features are the most important.

greatest_hits is a shortened version of the **broad** model, constructed by removing about 30 of the least important features from the former.

hh20 is a further shortened version of the broad model that retains about 40 of the most important features.

4.1.2 *pgm*-level feature sets

We have specified and trained a number of *pgm* models, and include most of them in the ensemble we present below. Most models make use of the LightGBM gradient boosting model, either in standard or hurdle-regression form. LightGBM is by far the most efficient algorithm that we have explored, greatly facilitating the task of training models for 13,000 PRIO-GRID-cells across several hundred points in time, across all 356 time-steps. The comparison done at the *cm* level indicates that the algorithm in general performs at least as well as the other, much less efficient, algorithms.

baseline is a simple specification including the most important predictors of conflict in a grid-cell. First, it includes information on the spatial and temporal proximity to conflict (any type of violence coded by UCDP), as we know that conflicts tend to re-occur and are likely to cluster in space. Locations that have a legacy of violence or are neighboring violent locations, are more likely to experience conflict. In addition, we include information on population size, as the likelihood of violence is increasing in densely populated locations (Raleigh and Hegre, 2009).

conflicthistory includes the baseline features as well as a dozen additional features more closely representing the conflict history of the geographical grid cell. To better be able to capture how previous, intense violence affect the number of persons killed at a later point in time, we include decay functions of time since at least 5, 25, 100, or 500 deaths in the same cell, for each of the three types of UCDP organized violence.

natsoc includes the baseline as well as a number of natural and social geography features that are used in the current ViEWS system (Hegre et al., 2019). Social geography features include information on local poverty, distance to international borders, the capital or the nearest city, and exclusion of local ethnic group from political power. Natural geography-related features include characteristics of the local terrain (e.g. prevalence of mountains, pasture, urban, forest, etc.), distance to diamond and oil deposits. Locations with high share of ethnic exclusion, diffused poverty, disconnected from the centres of power, and close to lootable resources are more likely to experience conflict.

drought includes a number of drought and vulnerability features in addition to the baseline. The model includes the SPEI Index as a proxy for drought (Vicente-Serrano, Beguería, and López-Moreno,

2010), information on the local crops' growing season from the MIRCA dataset (Portmann, Siebert, and Döll, 2010) and data on the crops' harvest and yield from Mapspam (International Food Policy Research Institute, 2019). Local drought during the local crops' growing season may increase the risk of conflict in vulnerable communities. To account for the role of societal vulnerability in mediating the effect of climate shocks on the risk of conflict, the *drought* model also encompasses information on economic conditions, the level of development, the degree of ethnic exclusion, population size, and dependence to agriculture – all factors that have been shown to condition the climate impacts on societies and their capacity to respond to climate anomalies.

protest includes 45 features capturing the recent history of protests at the local level based on data from the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010). The model distinguishes between events based on the intensity of violence as well as the actors involved. Hence, additional four types of different protest events are included, i.e. peaceful protests, protests with low-level intervention, protests with excessive force against protesters and protests with violent behavior by demonstrators. The protest event counts are normalised by dividing them by grid cell population (Tollefsen, 2012). Multiple transformations are performed to capture the spatial and temporal variation of the different protest types. Temporal proximity to protest events taking place in first- and second-order neighbouring grid cells are the most important features to predict the count of fatalities for all steps ahead. Protest events that include excessive violence are thereby particularly influential.

confhist is a conflict history model capturing both time and space proximity to past conflict, using conflict data at the sub-national level sourced from the UCDP-GED, the geographically disaggregated version of UCDP (Croicu and Sundberg, 2015). The main features include various temporal and spatial lags of the count of fatalities by grid-cell. The moving average and sum of fatalities over the past 6 months are the most important features to predict the count of fatalities one month ahead, although their importance decreases over time. Expectedly, the moving sum and average of fatalities over the past 36 and 12 months are the most relevant features to predict further ahead into the future (*s=12,24,36*). The spatial lag of fatalities is also an important feature to predict the number of deaths from state-based conflicts, in line with findings from the empirical literature, suggesting that conflicts cluster in space (Buhaug and Gleditsch, 2008). The temporal and spatial proximity to other forms of violence (*os,ns*) is less relevant to predict *sb* fatalities, even though its importance tends to increase along the forecasting horizon.

xgb_actor_confhist is a model combining the conflict history model described above with features relating to rebel groups (actors) present in both the grid-cell and in nearby grid-cells. These features attempt to extract agency and structure of conflict processes and include the count of armed groups involved in conflicts against the government and against each other, and information on the level of conflict intensity (fatalities and number of battles) carried out by the most intense and of the average actor in the group. Measures of longevity and escalation/de-escalation for each group are also included - count of actors existing for two months in a row and surviving for a year in the grid-cell, cumulative intensities over 12 months, as well as number of actors appearing, disappearing, escalating and de-escalating in the current month. Measures for nearby actor behavior (in the queen-

contiguity spatial-lag of each cell / 3x3 convolution space) are included as well. The data is described in Croicu (2019) and has, as principal novelty, that it includes potential rebel groups as well as armed actors that never escalated (groups with very limited armed activity, below the standard 25-battle related deaths threshold). In addition, the model includes some basic structural geography measures (GDP, travel time to nearest town, agricultural density and population are included). The model is estimated using extreme gradient boosting with a gradient histogram approximation for performance (Chen and Guestrin, 2016). Hyperparameters are tuned using a simple genetic algorithm run for 20 models for 20 generations on the calibration set.

conflict_tree_lags is a conflict model employing spatial lags of the conflict variables computed using a tree-based method instead of the kernel-based convolution used in the **confhist** and **xgb_actor_confhist** models. Convolution-based lags compute sums over a finite-sized (usually 3×3 cells) kernel. The size of the kernel, and therefore the range at which one event can be modelled as influencing another, is limited by the computational effort required to compute kernel sums for every cell. The tree, in contrast, allows distance-weighted sums to be computed approximately over the whole grid. Grid cells are first placed into a hierarchy with individual cells forming the highest level, groups of four cells the level below, groups of sixteen cells the level below that, and so on. For a given cell at which the sum is to be computed, sums over nearby partner cells are computed directly, but more distant partners are aggregated into lower-level tree nodes. An approximate sum with any required distance weighting over the whole grid can then be efficiently computed for every grid cell. The **tree_lags** model employs lags of the three conflict variables computed in this fashion with distance^{-1} and distance^{-2} weightings.

conflict_sptime_dist is a model that includes a unified measure of proximity to violence in both space and time, called *spacetime*. *Spacetime* allows to assign different weights to the temporal and spatial dimension. Specifically, for every grid cell at every time-step, the *spacetime distance* s to the nearest past conflict event is computed, where for a grid cell located at (x_i, y_i) at a time-step t_i and a past conflict event at (x_e, y_e, t_e),

$$s^2 = (x_i - x_e)^2 + (y_i - y_e)^2 + \nu^2(t_i - t_e)^2 \tag{1}$$

with the constraint that $t_i - t_e \geq 0$. ν is a scaling factor with the physical dimensions of velocity which allows the time difference to be added to the spatial distance. There is no obvious 'best´ value of ν, and consequently the **sptime_dist** model includes features where space-time distances are computed with $\nu = 10$ (which has the effect of stretching the time axis and thereby privileging more recent events), 1, and 0.01 (which compresses the time axis and makes it more likely that an event in the distant past will yield the minimum value of s for a given (x_i, y_i, t_i).

broad is a model that selects the most important features from each of the models above, based on a review of feature importances for each of them. In addition to the baseline features, the model includes three features from the sptime feature set, two from the tree_lags model, distances to borders, capitals, cities, and diamond deposits, local poverty, and four features from the drought model.

4.1.3 Algorithms

We have explored a number of algorithms to relate the feature sets to the outcome we seek to predict. Most of the models we end up using are tree-based models. With the exception of the GLM Markov model, none of the generalized linear models we tried yielded good performance. We also had limited success with simple neural network models.[9]

Random forests XGBRFRegressor(n_estimators=300) or scikit's RandomForestRegressor(n_estimators=200), implementing the Random Forest (Breiman, 2001) algorithm, an ensemble of decision trees, with each tree trained on a subset of features and bootstrapped data – with the aggregate ensemble reducing.

Gradient boosting models scikit's GradientBoostingRegressor() Gradient Boosting Regressors (GBR) are another ensemble method improving decision trees sequentially by training each iteration on the residual of the past iteration. The algorithm starts by assigning equal weights to all data points. It then iteratively changes the weights by increasing the weight assigned to difficult observations that are misclassified, and lowering the weight for data points that are easy to classify or are correctly classified.

'Extreme' gradient boosting XGBRegressor(n_estimators=100, learning_rate = 0.05). The model is estimated using extreme gradient boosting (Chen and Guestrin, 2016). We use the XGBoost implementation, and performed an 'early-stopping' routine to identify the optimal number of estimators and learning rate parameters.

Light gradient boosting LightGBM (LGBMRegressor, n_estimators=100) is another gradient boosting method based on decision trees to increase the efficiency of the model and reduce memory usage. It uses novel techniques (One Side Sampling and Exclusive Feature Bundling) to overcome the limitations of histogram-based algorithm. The Light Gradient Boosting works by retaining instances that with larger gradients(those that contain more information but are under-trained) and randomly dropping data-points with small gradients. This leads to a more accurate estimation than uniformly random sampling.

Hurdle models We have also explored a set of 'hurdle models'. As indicated by the review of the prediction outcome above, most of the observations have no fatalities, and the distribution of the non-zero observations are highly right-skewed. There are reasons to think that the data-generating process that lead to whether a country or a grid cell having any fatalities at all is quite different from the one that lead to subsequent fatalities. Hurdle models take this into account by dividing the outcome into two variables, a dichotomous variable for whether there was non-zero fatalities or not, and the log count of fatalities if there was at least one fatality. The model then trains a classifier for the zero/non-zero distinction, and a regressor for the non-zero observations. At the predict stage, the predicted number of fatalities is the product of the probability of non-zero observations, and

[9] We expect more advanced long-short memory and convolutional recurrent neural network models to be effective for the *pgm* level, but we have not been able to test these as currently are lacking access to a well-specified GPU-based computer.

the expected number of fatalities given there is at least one, conditional on the predictors. We have explored a number of variants of the hurdle models, using classifier and regression versions of the tree-based algorithms described above.[10]

Markov_glm and Markov_rf Markov models are a more sophisticated formulation of the hurdle-model idea that different models do well in different situations. The models use an observed Markov modeling approach with four different latent states which produce fatalities, and where the forecast of fatalities is conditional on the likelihood of the conflict state. The four conflict states used are the same as in Randahl and Vegelius (2022), that is, 'peace', 'escalation', 'de-escalation', and 'conflict'. Transitions between states are restricted such that each state only has two possible future states. The transitions allowed are from peace to peace and to escalation, from escalation to conflict and to deescalation, from deescalation to peace and to escalation, and from conflict to conflict and to deescalation. The escalation and deescalation states are thus transient, as they do not allow transitions to themselves. Transitions between states are modelled as a binary logistic regression model, and the log number of fatalities conditional on the latent states are modelled using OLS regression. The **Markov_glm** version use logistic and linear regression models, and the **Markov_rf** models the transition between states using a random forest classifier, and the log number of fatalities conditional on the latent states using a random forest regressor. For more details on the Markov modeling approach, see Randahl and Vegelius (2022).

4.2 Ensembling – using the 'wisdom of the crowd'

No statistical or machine-learning model or algorithm can perfectly learn the patterns of behavior that link some observable predictors to subsequent observations of the number of fatalities in war. Building on a variety of theoretical and methodological perspectives – the 'wisdom of the crowd' – clearly yields the best foundation for good decisions and high-quality forecasts (Tetlock, 2005). The greater the variance of adequate models available, the better a forecasting and decision-making system performs (Page, 2007). Ensembling – grouping of diverse forecasting models – also work as a means to smooth over problems (Armstrong, Green, and Graefe, 2015). Following the approach in ViEWS (Hegre et al., 2019), we use ensembling of constituent models to aggregate insights from various models, allowing a variety of modeling algorithms and feature sets, and applying state-of-the-art model weighting algorithms (Sivanandam and Deepa, 2008; Scrucca et al., 2013; Montgomery, Hollenbach, and Ward, 2012).

The default ensemble algorithm is just the equally weighted mean. However, it is clear from the evaluation of results below that some models perform better than others, and we should be able to improve performance by giving these models more weight in our ensembles. For the *cm* level, we have developed an algorithm to learn these weights from the data. To do this, we split the data into three periods. The first period, the 'training period', include the years 1990–2012. We train the constituent models described above on data for this period, and predict for the 'calibration period'; 2013–2016. Our ensemble weighting

[10] Our hurdle model implementation is based on code developed by Geoff Hurdock: `https://geoffruddock.com/building-a-hurdle-regression-estimator-in-scikit-learn/`

and calibration model use these predictions as well as data for the true outcome for the calibration period to obtain weights and calibration parameters. We then retrain all the constituent models for the 1990–2016 period and generate predictions for the 2017–2020 period, and apply the weights and calibration parameters to produce ensemble forecasts for that period. The forecasts for the true future will use the 2017–2020 period as calibration period, and generate ensemble predictions for the 2022-2024 period.

Our model weights are obtained using a genetic algorithm (Sivanandam and Deepa, 2008; Russell and Norvig, 2020). These optimize a user-defined performance metric in the calibration data by letting a population of random model weights evolve over a large number of generations to find optimal weights. Genetic algorithms provide a fast, flexible, and intuitive way to optimize the performance metric when the inputs are high-dimensional or when there are complex restrictions on the available inputs.

The genetic ensembling algorithm, as implemented, works like this: 100 random ensembles are chosen with a random set of weights (genes), under the sole condition that the sum of those weights is between 0.5 and 3. Each of the 100 ensembles are then computed using the assigned weights and then evaluated using a mean squared error fitness function ($1/e^{mse}$) against the data in the calibration period. Pairs of the ensembles are then sampled through a weighted sampling procedure based on the fitness scores. These pairs are recombined in a simulation of genetic reproduction - a random subset of weights from one ensemble in the pair is combined with the remaining weights from the other ensemble in the pair. Then, with a probability of .2, a random subset of weights from the resulting ensemble is replaced with random weights. 100 such recombination/mutation processes are carried out, leading to 100 new organisms that form a new generation, to be put, again, through the same process as above[11]. This is repeated 500 times (generations), with the best ensembles from the last (500th) generation being used.

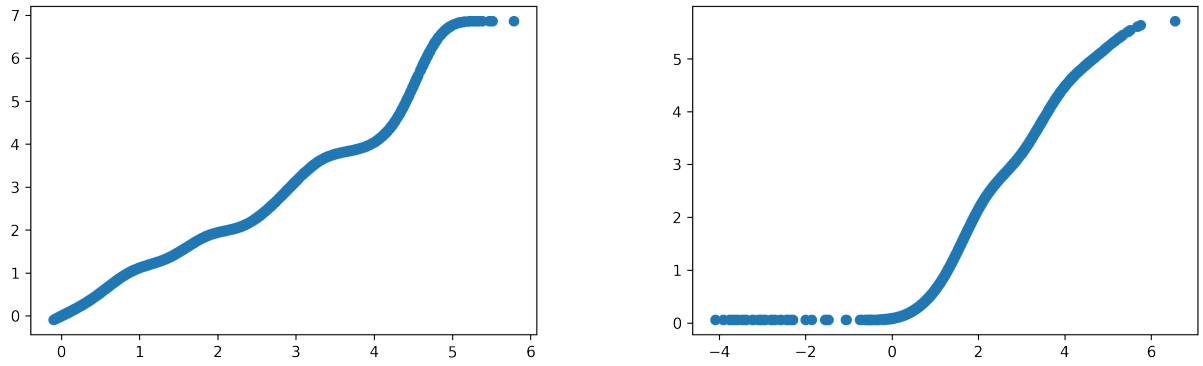
4.3 Calibration

Predictions should ideally have roughly the same distributions as the observed outcome. Given the distribution of the outcome variable (Figures 2 and 7), few algorithms would get this right without post-prediction calibration. In particular, models tend to yield distributions with smaller variance than the outcome, reducing our ability to correctly predict the total number of fatalities. Moreover, they tend to have means that are lower than the true mean, and some models produce negative values even though these do not exist in the training data.

We have explored two ways to calibrate predictions The first is to multiply the predicted number of fatalities with the constant that gives the same variance (in the calibration partition) as the outcome – in practice, this is equivalent to estimating a no-constant linear regression model with the outcome in the calibration partition as the dependent variable and the prediction as the independent variable. The estimated parameter is in most cases larger than 1, expanding the variance of the predictions and

[11] As a complication, to improve performance, at each generation, the 10 best models of the previous generations are "spared and cloned", i.e. preserved intact, without recombination and mutation for the next generation. The ten worst performers in the mutation and recombination steps are discarded, so that the population of ensembles remains 100.

Figure 9. **GAM calibration function, fat_topics_histgbm model (left) and fat_hh20_xgb model (right),** $s = 3$



Source: ViEWS, 2022

increasing the mean. We estimate this calibration model separately for each constituent model and each step, and apply the calibration parameter (the β coefficient in the OLS model) to the predictions for the test partition. We also explored including an intercept in the calibration model. This, however, in most cases yields non-zero predictions for a majority of the cases where the true outcome is zero, significantly hurting performance. Without an intercept term, however, the calibration works poorly for models like the XGBoost model that can yield negative predictions – multiplying a negative prediction with a number larger than one obviously does not help calibration.

To counter theses challenges, we have moved to use a generalized additive linear model (a GAM), using the PYGAM package (Servén D., 2020). GAM models fit the relationship between the dependent and independent variables as a very flexible function. To avoid overfitting, we constrained the model to yield calibrated predictions that are monotonically increasing in the non-calibrated predictions – if the original model ranks one case higher than another, the calibrated model also ranks it as as least as high. We set the parameters of the function so that the calibrated transformation is quite smooth, retaining most of the original prediction.

This model was estimated separately for each constituent model and for each step. Figure 9 illustrates how the function works. The calibration function works well, typically decreasing MSE by about 10% relative to the uncalibrated predictions, mostly removes zero predictions, and increases the variance.

Figure 9 shows two example calibrations. The y axis shows the calibrated predicted number of fatalities as a function of the original prediction (x axis). In the case of the fat_topics_histgbm model (left), the GAM function does not alter the original predictions much except for predictions above 4 (about 50 deaths), but pull the remaining predictions considerably upwards. In the case of fat_hh20_xgb model (right), which yields a number of large negative predictions, all negative predictions are calibrated to zero, and the remaining predictions are mostly unchanged.

The genetic algorithm can in principle yield weights that sum to less than or more than one. This, then, serves as a second calibration step.

4.3.1 Calibration at the *pgm* level

The issue of calibration is particularly acute for the pgm models. Africa and the Middle East together comprise approximately 13,000 PRIOGRID cells, which when multiplied by the 356 time – steps under consideration here, yields around five million units of analysis. In any of the conflict datasets, the vast majority of values (i.e. the numbers of fatalities) associated with these units of analysis are zero.

The imbalance between the numbers of zero and non-zero data points in problems like conflict prediction is sometimes redressed by randomly discarding a (possibly very large) fraction of the zero-valued units of analysis, so that the regression algorithms which search for patterns in the data are exposed to more equal numbers of zero and non-zero values. This is known as *downsampling* and the hope in doing this is that the algorithms are not swamped by zero values and do not, therefore, tend to simply predict zeros or very small values everywhere. Downsampling also reduces the runtime of fitting procedures, since algorithms have less data to deal with.

We examined the effect of downsampling by discarding between 70 and 98 per cent of the zero-valued data points. Very large discard fractions significantly worsened predictive performance and more moderate fractions yielded no appreciable performance improvement, while runtimes required to dispense with downsampling entirely were not prohibitive. We therefore abandoned downsampling as a data-engineering strategy.

However, with or without moderate downsampling, the predictive performance of the random forest-, LGBM- and gradient-boosting algorithms were all quite poor, as measured by examining MSE values, or more subjectively by comparing maps of predictions from the test partition with observations from the same timestep.

The clearest problem is that, while able to predict the presence or absence of conflict in roughly the correct geographic locations (although with some spread around the locations of observed conflict events), the numbers of predicted fatalities were almost always very small. This is indicative of a normalisation or calibration problem, as described in the previous section.

The best solution to this problem might be to use another evaluation metric than MSE at the *pgm* level. The pEMDiv metric (Greene et al., 2019) is the best candidate we are aware of. In future iterations of this modeling exercise, we will use a variant of this metric both to evaluate the model system and to optimize it.

Another issue is that it is desirable that the sum of predicted fatalities at the *pgm* level that make up a given country are similar to the predicted fatalities at the *cm* level. In the modeling system, we calibrate

the predictions along this line of logic (see below). The MSE scores we report, then, have a different interpretation – they should be read as ranking models in terms of the model's ability to capture the distribution of the total fatalities suggested by the *cm* ensemble.

5 True forecasts for April 2022–March 2025

Figure 10 shows the predicted number of fatalities for all countries in the Africa and Middle East regions in map form. The forecasts were produced in May 2022 based on data up to and including March 2022. Figure 11 shows the total number of predicted fatalities over the next 12 and 36 months for these countries, and 12 presents the distribution of fatalities over the most affected conflict countries.

The model suggests conflict will remain extremely violent in Yemen, with up to about 1000 deaths from state-based violence every month over the next three years. Conflict will also remain intense in Nigeria, Somalia, and Syria, and to a somewhat lesser extent in Mali, Burkina Faso, and DR Congo.

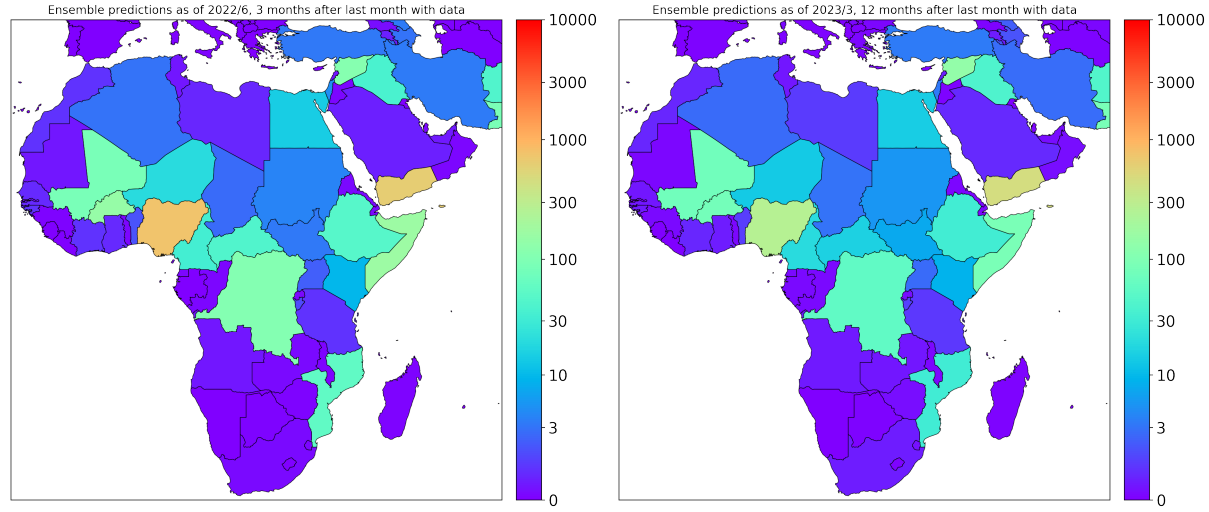
Figure 13d shows what drives these forecasts. Most importantly, past conflict history drives the forecasts for the most violent countries listed above (Figure 13a and 13b). This is accentuated in many countries by a combination of poverty and non-democratic institutions (in particular, in Central African Republic and Somalia). Other countries with relatively high risk of fatalities despite having limited amounts of recent violence are Guinea, Sierra Leone, South Sudan, Zambia, and Zimbabwe.

5.1 Surrogate models to understand the key drivers of armed conflict

To interpret the ensemble model predictions we have developed a set of 'surrogate models' (Molnar, 2021). These models link the ensemble predictions to a small number of input features by means of a simple model. In our implementation, we have used generalized additive models (GAMs) to aid interpretation. These are estimated using the ensemble predictions for the test partition for a given step as dependent variable, and the input variable appropriately time-shifted as independent variable. Figure 14 shows the results for two of these models, for two steps. Figure 14a shows how the (log) number of fatalities predicted six months into the future relate to the (log) number of deaths in **sb** conflict six months earlier, i.e. the most recent month of data informing the surrogate models presented here. This variable on its own is sufficient to explain about 91% of the prediction. Figure 14b shows the same for the case where we seek to predict 36 months into the future. Still, this variable accounts for about 87% of the variance in the ensemble predictions.

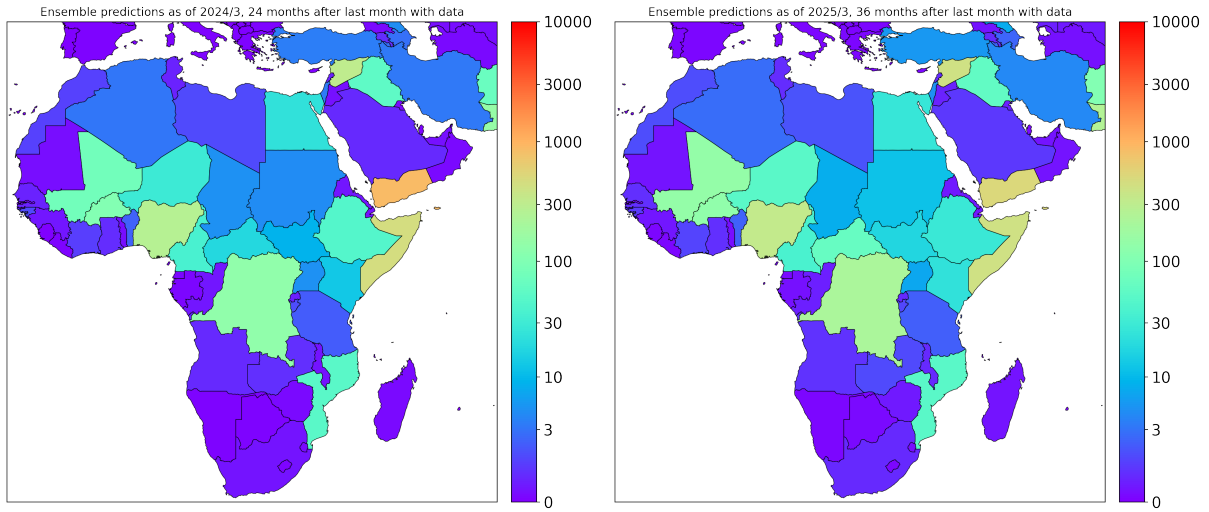
Figures 14c and 14d shows the relationship between the ensemble predictions 6 and 36 months into the future, as functions of the 'liberal democracy' score from V-Dem (Coppedge et al., 2020) and infant mortality rate from the World Development indicators (WorldBank, 2019), which we see as a proxy for the extent of poverty and under-development in the country. This model ignores conflict history, and is

Figure 10. **Prediction maps for the future, based on data up to and including March 2022**



(a) June 2022 ($s = 3$)

(b) April 2023 ($s = 12$)

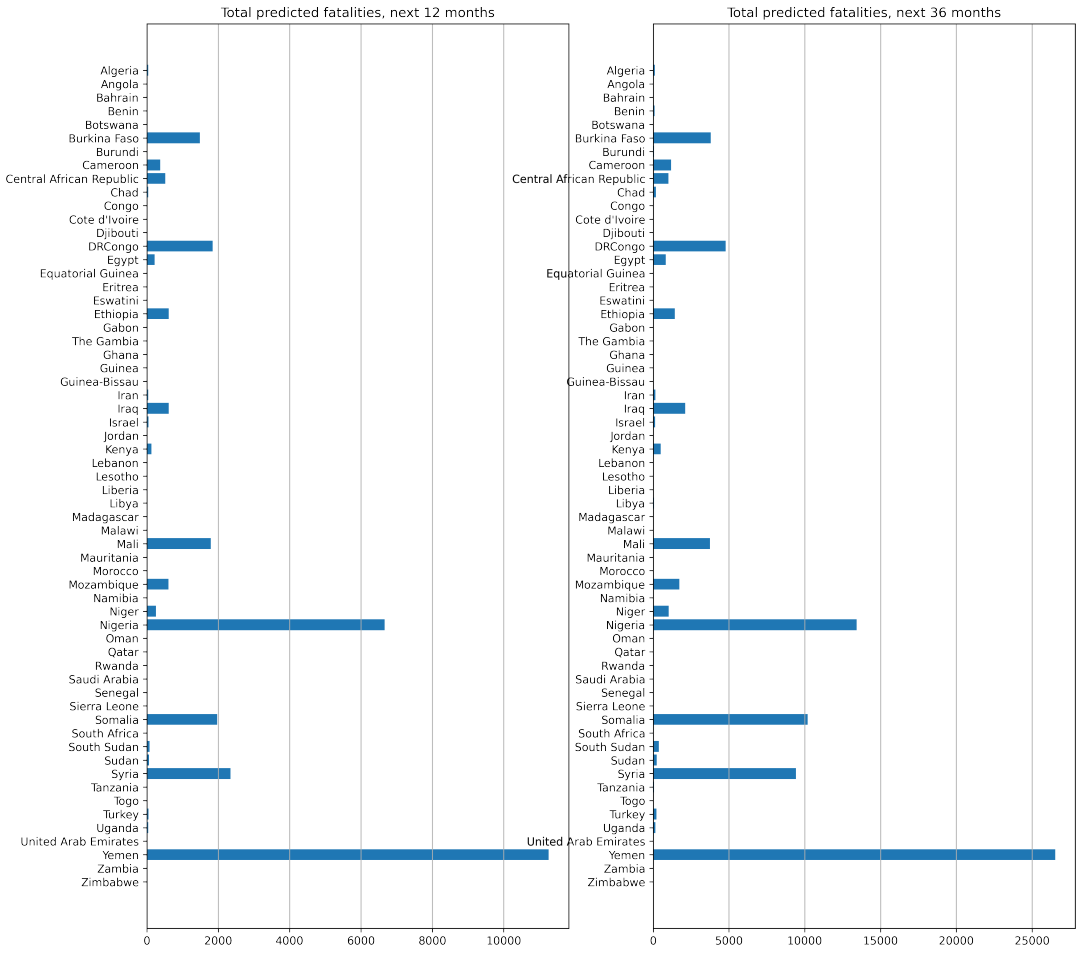


(c) March 2024 ($s = 24$)

(d) March 2025 ($s = 36$)

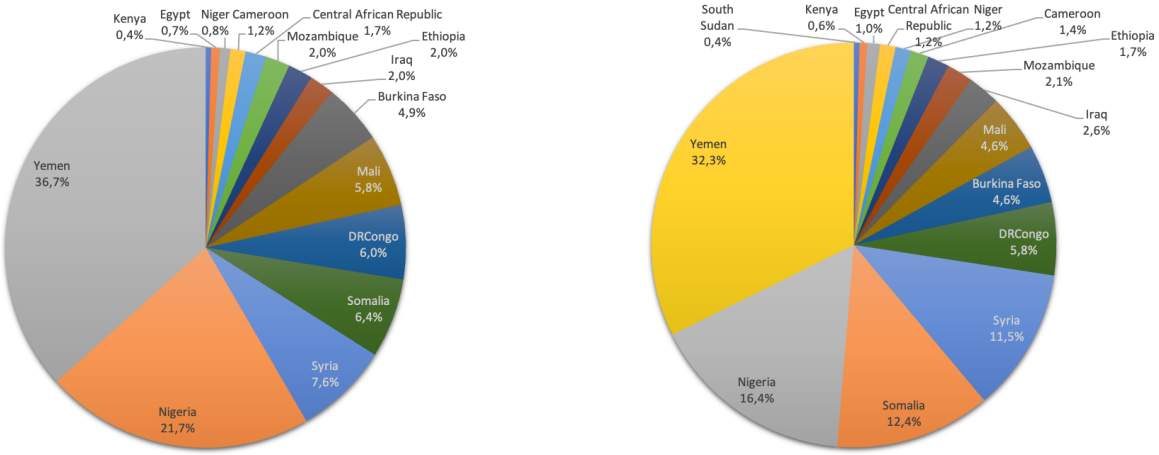
Note: Forecasts for 3, 12, 24, and 36 months into the future relative to the last month of data, March 2022.
Source: ViEWS, 2022

Figure 11. **Predicted total fatalities over the next 12/36 months, all countries**



Note: Predicted total fatalities over the next 12 and 36 months relative to the last month of data, March 2022.
Source: ViEWS, 2022

Figure 12. **Predicted share of total fatalities over the next 12/36 months amongst the most affected conflict countries**



Note: Predicted share of total predicted fatalities in the most important countries over the next 12/36 months, relative to the last month of data (March 2022).
Source: ViEWS, 2022

then able to account for about 17% of the variance in the predicted number of fatalities. The relationships are in line with what has been found in earlier studies (see Hegre, 2014; Hegre, 2018, for reviews): The risk of fatalities is increasing in infant mortality rate, and has an inverted-U relationship to the level of democracy. The predicted number of fatalities is highest for political systems that are rated with 0.2 on the liberal democracy score, the current level in Ethiopia, for instance, and lower for more or less democratic political systems. Countries that are both very poor and are partial democracies are the most at risk of intense political violence, according to the forecasting model.

6 Evaluation of predictive performance

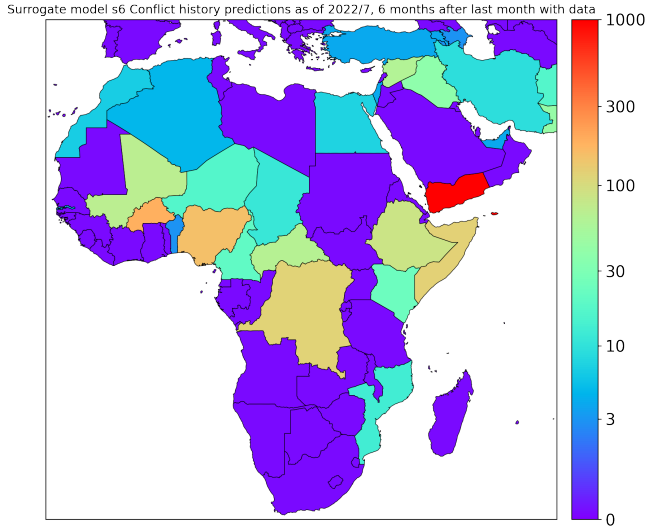
In this section, we review the predictive performance of the models at the *cm* and *pgm* level, including comparing the relative contribution of the various methodological alternatives explored in the project.

6.1 *cm* level

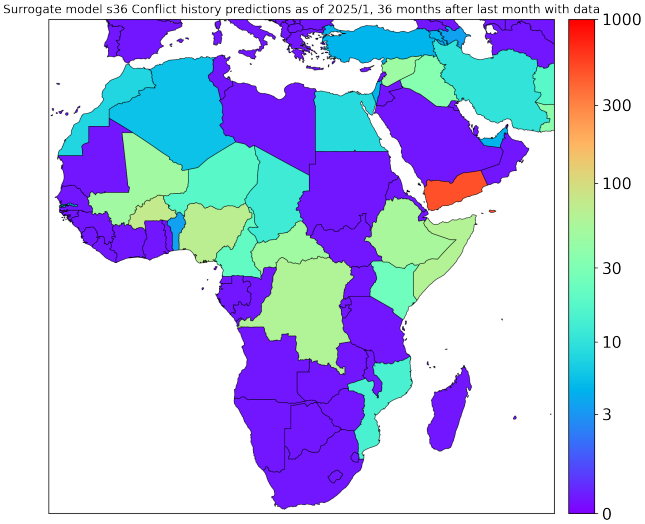
6.1.1 Constituent models

Figure 15 shows Mean Squared Error (MSE) for the entire range of constituent models explored for the project, as well as for the unweighted ensemble, as a heatmap, for all steps, for the calibration period 2013–2016. Models are sorted according to the feature sets that go into them, with the ensemble at the bottom. Blue color corresponds to low MSE and good performance, red color to high MSE and poor

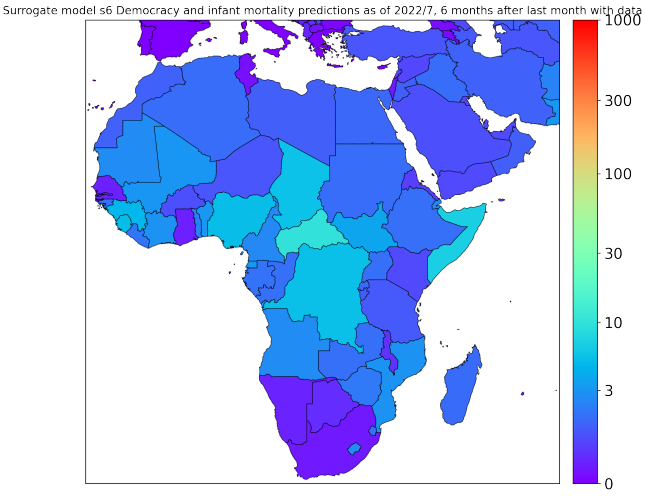
Figure 13. **Surrogate model prediction maps for the future, based on data up to and including January 2022**



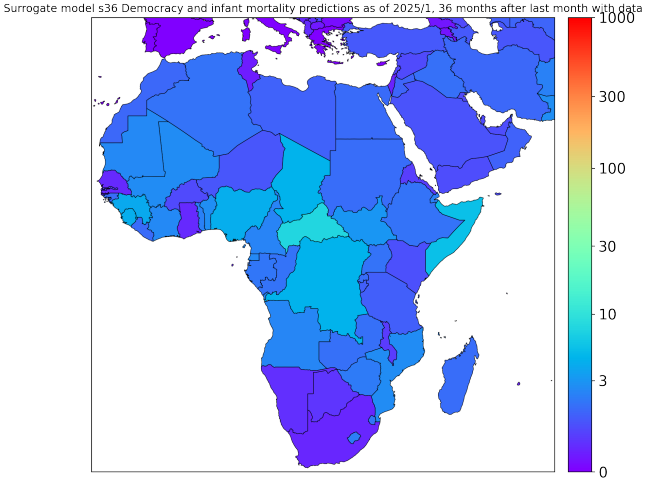
(a) Conflict history, July 2022 ($s = 6$)



(b) Conflict history, January 2025 ($s = 36$)



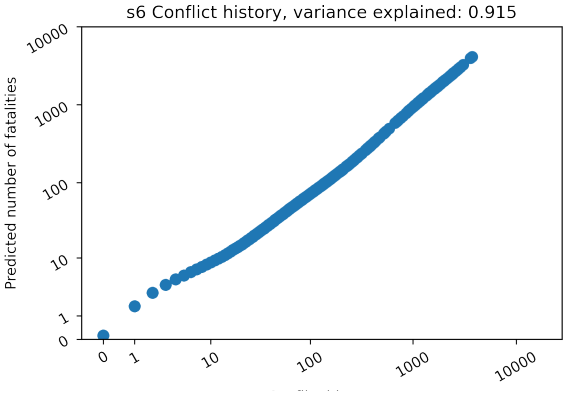
(c) Democracy and infant mortality, July 2022 ($s = 6$)



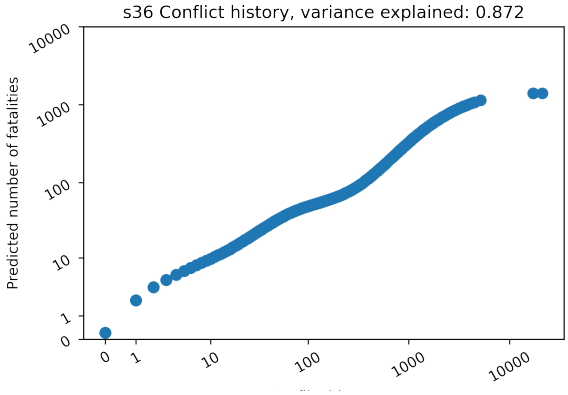
(d) Democracy and infant mortality, January 2025 ($s = 36$)

Source: ViEWS, 2022

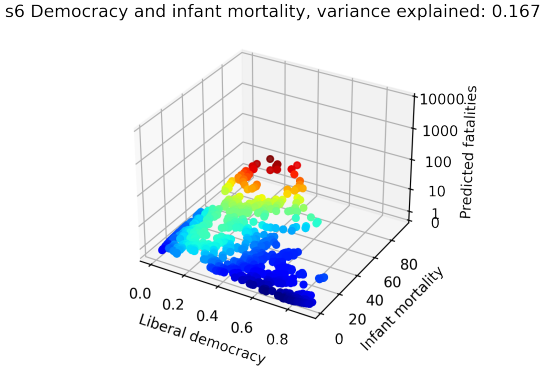
Figure 14. **Surrogate models, test partition, all countries globally**



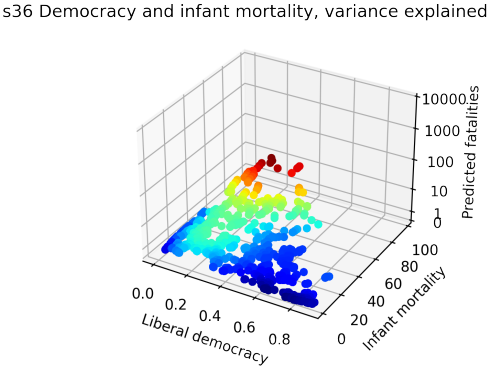
(a) Conflict history, $s = 6$



(b) Conflict history, $s = 36$



(c) Democracy and infant mortality rate, $s = 6$

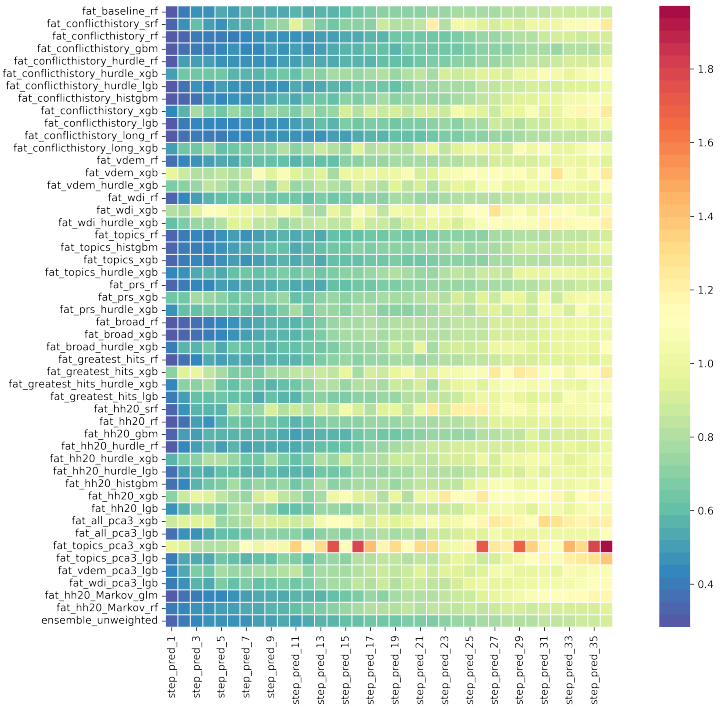


(d) Democracy and infant mortality rate, $s = 36$

Source: ViEWS, 2022

performance. Obviously, MSE increases the further into the future we seek to forecast: Mean squared error is typically 2–3 times higher at step 36 than at step 1.

Figure 15. **MSE, calibration partition**



Source: ViEWS, 2022

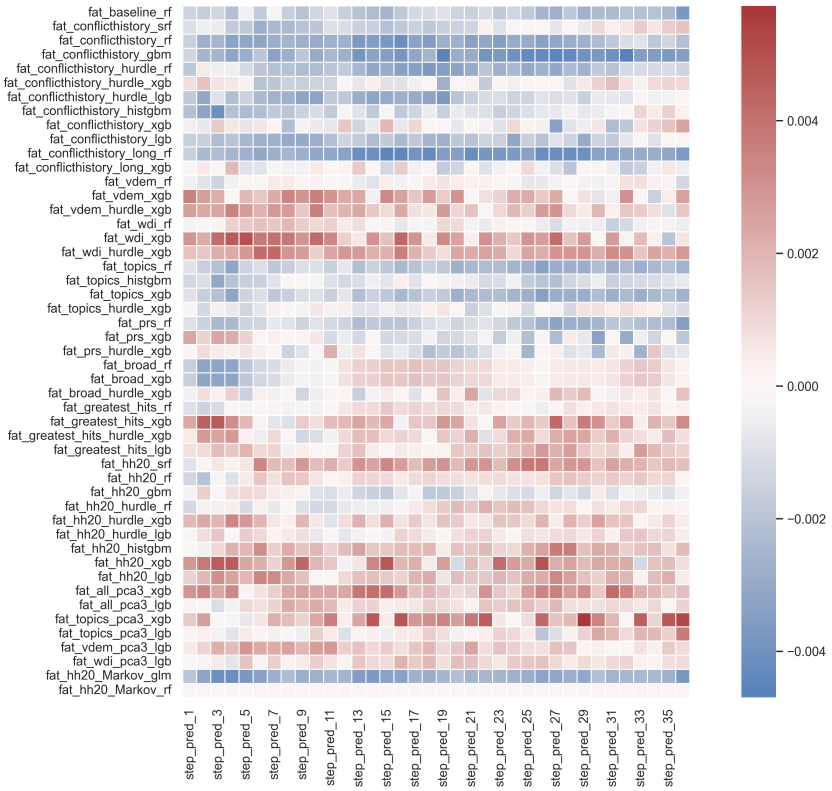
Table 2 shows MSE for the test partition in table form for selected steps.

Figure A-1 shows MSE for the fraction of the test partition that had actual zeros (top) as well as actual non-zeros (bottom), to further explore the relative strengths and weaknesses of the models. For the most part, these plots confirm the picture from MSE overall, with some interesting exceptions. The random-forest based Markov model yields a high MSE for the zero cases, but is very precise for predicting the non-zeros, for instance.

Figure 16 looks at model performance from a different angle: It shows how much the MSE for an unweighted ensemble changes if the model is removed from this ensemble. Cells have blue color when the ensemble deteriorates if the model is removed for that step, and red color if the ensemble improves. Seven of the conflict history models contribute positively the ensemble for all steps, three of the topics models, one of the PRS models. Only one of the 'hh20' models is contributing. Models that are exclusively based on V-Dem and WDI features mostly hurt the ensemble, as do most of the ´broad' and 'greatest hits' models.

Finally, Table 3 shows MSE for the cases that actually had non-zero fatality counts in the 2018–20 period.

Figure 16. **Ablation MSE: Changes to the MSE of an unweighted ensemble, calibration partition, if the model is removed from the ensemble.**



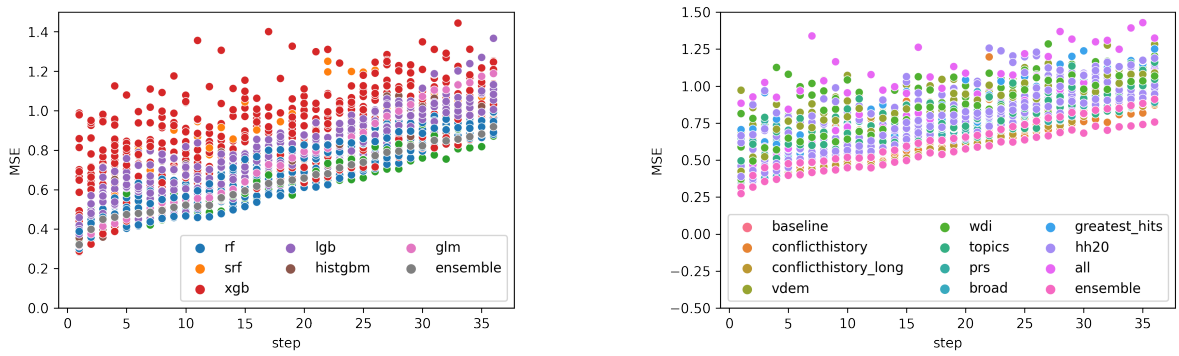
Note: Blue color/negative values means the MSE of the ensemble is lower/better if the model is kept in the ensemble, red color/positive values that the MSE of the ensemble is higher if it is retained.
Source: ViEWS, 2022

Table 2. **MSE for constituent models and ensembles, test partition, calibrated models**

	step_pred_1	step_pred_3	step_pred_6	step_pred_12	step_pred_36
fat_baseline_rf	0.312350	0.348635	0.482028	0.762247	0.615954
fat_conflicthistory_srf	0.271888	0.388347	0.397856	0.567990	0.795355
fat_conflicthistory_rf	0.243884	0.309597	0.438257	0.550901	0.619640
fat_conflicthistory_gbm	0.239752	0.346730	0.339781	0.383258	0.646735
fat_conflicthistory_hurdle_rf	0.239677	0.314409	0.334517	0.424023	0.697962
fat_conflicthistory_hurdle_xgb	0.348079	0.391089	0.445749	0.478384	0.781236
fat_conflicthistory_hurdle_lgb	0.253937	0.330650	0.337711	0.398766	0.624152
fat_conflicthistory_histgbm	0.244727	0.323168	0.366490	0.433894	0.723607
fat_conflicthistory_xgb	0.400629	0.648744	0.636688	0.538982	0.757464
fat_conflicthistory_lgb	0.257139	0.315741	0.362989	0.387724	0.708629
fat_conflicthistory_long_rf	0.234813	0.315027	0.355938	0.445409	0.593512
fat_conflicthistory_long_xgb	0.526209	0.716729	0.643766	0.710156	0.794975
fat_vdem_rf	0.297589	0.397761	0.737763	1.862160	0.828352
fat_vdem_xgb	0.518204	0.616339	0.573585	0.597758	1.069021
fat_vdem_hurdle_xgb	0.386985	0.504486	0.585525	0.576335	0.755514
fat_wdi_rf	0.277754	0.355100	0.653053	0.497221	0.656402
fat_wdi_xgb	0.553252	0.670171	0.590272	0.625111	0.840372
fat_wdi_hurdle_xgb	0.422284	0.514358	0.540469	0.503373	0.798807
fat_topics_rf	0.269212	0.377123	0.418048	0.427984	0.638771
fat_topics_histgbm	0.269763	0.365392	0.423085	0.421580	0.667724
fat_topics_xgb	0.269212	0.377123	0.418048	0.427984	0.638771
fat_topics_hurdle_xgb	0.378810	0.508782	0.483442	0.494552	0.705131
fat_prs_rf	0.267412	0.328939	0.430272	0.569960	0.780147
fat_prs_xgb	0.397621	0.513776	0.527073	0.564223	0.767457
fat_prs_hurdle_xgb	0.399098	0.450055	0.504209	0.577334	0.709295
fat_broad_rf	0.238619	0.302460	0.410058	0.579081	0.768517
fat_broad_xgb	0.238619	0.302460	0.410058	0.579081	0.768517
fat_broad_hurdle_xgb	0.346410	0.417060	0.437902	0.481655	0.727715
fat_greatest_hits_rf	0.242592	0.384008	0.523712	0.533727	0.725425
fat_greatest_hits_xgb	0.457396	0.503204	0.568745	0.624928	0.843833
fat_greatest_hits_hurdle_xgb	0.338278	0.467331	0.449010	0.474543	0.758770
fat_greatest_hits_lgb	0.294072	0.390850	0.395873	0.499952	0.673218
fat_hh20_srf	0.304743	0.655106	0.573686	0.771910	0.786554
fat_hh20_rf	0.242222	0.430797	0.982336	0.494854	0.718563
fat_hh20_gbm	0.264341	0.341445	0.382068	0.457804	0.674115
fat_hh20_hurdle_rf	0.245551	0.340191	0.401259	0.470854	0.694925
fat_hh20_hurdle_xgb	0.367649	0.548605	0.530213	0.529297	0.748032
fat_hh20_hurdle_lgb	0.250011	0.374491	0.429056	0.460599	0.650877
fat_hh20_histgbm	0.257422	0.354729	0.419473	0.517115	0.793001
fat_hh20_xgb	0.416823	0.573836	0.587328	0.626272	1.015177
fat_hh20_lgb	0.290986	0.421984	0.499616	0.490089	0.702540
fat_all_pca3_xgb	0.689024	0.743891	0.747092	0.794703	0.970532
fat_all_pca3_lgb	0.313916	0.406369	0.499983	0.599801	0.659584
fat_topics_pca3_xgb	0.835653	0.771567	0.772422	0.955923	1.673361
fat_topics_pca3_lgb	0.341363	0.493407	0.523855	0.676104	1.545612
fat_vdem_pca3_lgb	0.331461	0.429191	0.512842	0.549633	0.769146
fat_wdi_pca3_lgb	0.336640	0.501225	0.577609	0.621943	0.735220
fat_hh20_Markov_glm	0.266196	0.340344	0.440420	0.553014	1.288579
fat_hh20_Markov_rf	0.327788	0.398252	0.442994	0.503946	0.741810
ensemble_unweighted	0.234541	0.297511	0.318613	0.361841	0.547852

Source: ViEWS, 2022

Figure 17. **MSE as function of steps into the future, algorithms, and feature sets**



(a) Steps and algorithm

(b) Steps and feature sets

Source: ViEWS, 2022

For $s = 1$ month in to the future, models have mean squared errors just over 1, which corresponds to missing the true fatality with about a factor of 3. For $s = 24$, models miss by about 2 units on a log scale, or about a factor of 8. Again, the topics models perform well, but other models are not far behind.[12] The **vdem_short**, **wdi_broad**, and **broad_short** models also do well. The Markov models are somewhat weaker than the best-performing models.

6.1.2 Comparing MSEs across algorithms and feature sets

Some feature sets and algorithms perform better than others. Table 18 shows how mean performance across models and steps vary with algorithms used, feature sets underlying them, and other modeling characteristics.[13] Figure 17 shows the same information in visual form.

This analysis reflects that mean squared error increases as the forecasting horizon is extended, more precisely by 0.015 per step on average.

The coefficients in Table 18 shows how a given model characteristic change the MSE compared to the unweighted ensemble model. The results suggest that the hurdle models do considerably better than the ensemble, and the Markov model construction also clearly improves performance. The PCA models perform as well as the ordinary models, indicating that they could add something useful to the ensemble as their predictions are quite distinct from the others.

As for feature sets, models based on the two conflict history sets and the news topics models in general

[12] In the next stage of this project, we will use boot-strapping techniques to assess the uncertainty of the differences in MSEs between models.

[13] The table is the result from estimating an OLS regression using MSEs of models as the dependent variable and the characteristics as independent ones. Estimated standard errors should not be used for any hypothesis testing, but the estimated coefficients are good indications of a characteristic's contribution to relative performance.

Table 3. **MSE for constituent models, only observations with actual non-zeros, test partition, calibrated models**

	step_pred_1	step_pred_3	step_pred_6	step_pred_12	step_pred_36
fat_baseline_rf	1.503345	1.524623	2.336305	4.141495	3.085684
fat_conflicthistory_srf	1.343787	2.024064	1.895271	2.821754	3.179284
fat_conflicthistory_rf	1.164621	1.476284	2.116914	2.821917	2.230528
fat_conflicthistory_gbm	1.153559	1.482670	1.343545	1.879099	2.531581
fat_conflicthistory_hurdle_rf	1.159798	1.391232	1.468947	1.683432	2.285727
fat_conflicthistory_hurdle_xgb	1.732104	1.735561	2.109369	2.123059	3.400394
fat_conflicthistory_hurdle_lgb	1.177005	1.534752	1.447181	1.659308	2.496271
fat_conflicthistory_histgbm	1.150633	1.464606	1.450148	1.941896	2.671167
fat_conflicthistory_xgb	2.165823	2.448412	2.322757	2.321712	3.450760
fat_conflicthistory_lgb	1.186864	1.476629	1.453354	1.644653	2.704289
fat_conflicthistory_long_rf	1.121134	1.452637	1.544338	2.036023	2.470182
fat_conflicthistory_long_xgb	1.827653	3.046872	2.823402	2.966699	3.130239
fat_vdem_rf	1.454486	1.972611	3.918311	9.025943	4.116031
fat_vdem_xgb	2.214157	2.457624	2.839655	2.520093	5.256171
fat_vdem_hurdle_xgb	1.725652	2.462229	2.748871	2.235513	3.777703
fat_wdi_rf	1.271859	1.572185	3.101892	2.254350	3.460709
fat_wdi_xgb	2.199253	2.896549	2.948675	3.186534	4.249318
fat_wdi_hurdle_xgb	1.909017	2.369795	2.200511	2.115526	4.145712
fat_topics_rf	1.275998	1.807971	2.079043	1.953468	2.913266
fat_topics_histgbm	1.341753	1.837880	2.209088	2.062236	3.020464
fat_topics_xgb	1.275998	1.807971	2.079043	1.953468	2.913266
fat_topics_hurdle_xgb	1.992092	2.609829	2.480245	2.447909	3.150093
fat_prs_rf	1.200323	1.487704	2.074943	2.901292	3.377009
fat_prs_xgb	2.029237	2.396553	2.446589	2.786389	3.772877
fat_prs_hurdle_xgb	2.046677	2.020685	2.316531	2.415765	3.391370
fat_broad_rf	1.143640	1.402724	1.707521	2.326928	3.166042
fat_broad_xgb	1.143640	1.402724	1.707521	2.326928	3.166042
fat_broad_hurdle_xgb	1.853530	2.070326	1.935385	2.172782	3.525120
fat_greatest_hits_rf	1.182829	1.921499	2.234830	2.206989	3.331416
fat_greatest_hits_xgb	1.986365	2.330007	2.735423	3.065358	3.918275
fat_greatest_hits_hurdle_xgb	1.765032	2.132586	2.172283	2.043721	3.856693
fat_greatest_hits_lgb	1.373099	1.738249	1.645467	2.060743	3.245598
fat_hh20_srf	1.492986	2.156872	2.076605	3.915923	3.604609
fat_hh20_rf	1.162554	2.106171	5.573140	2.394193	3.154381
fat_hh20_gbm	1.155355	1.427249	1.706866	1.874625	2.894999
fat_hh20_hurdle_rf	1.184339	1.593804	1.653093	1.955720	2.888666
fat_hh20_hurdle_xgb	1.679095	2.372503	2.420406	2.357178	3.322519
fat_hh20_hurdle_lgb	1.189455	1.600164	1.833084	2.030360	2.798055
fat_hh20_histgbm	1.246684	1.695281	1.637910	2.095908	3.283830
fat_hh20_xgb	1.910751	2.713801	2.292699	2.806152	4.153993
fat_hh20_lgb	1.287638	1.722793	2.048325	2.037719	2.900494
fat_all_pca3_xgb	2.557523	3.014174	3.622915	3.292943	4.498346
fat_all_pca3_lgb	1.379224	1.955464	2.331547	2.646223	2.938416
fat_topics_pca3_xgb	3.499885	4.482024	3.959814	4.605906	10.443977
fat_topics_pca3_lgb	1.709011	2.515193	2.389118	2.384080	5.502493
fat_vdem_pca3_lgb	1.542693	1.828547	2.090385	2.064645	3.565875
fat_wdi_pca3_lgb	1.523892	2.457583	2.830616	2.570071	3.717958
fat_hh20_Markov_glm	1.192426	1.524279	1.974117	2.451173	5.719302
fat_hh20_Markov_rf	1.236295	1.506650	1.723119	1.926638	3.014789
ensemble_unweighted	1.077267	1.341794	1.344365	1.551380	2.472969

Source: ViEWS, 2022

Figure 18. **How MSE of models vary with algorithms and feature sets, calibration partition. Ensemble models as reference category.**

OLS Regression Results

Dep. Variable:	MSE	R-squared:	0.781
Model:	OLS	Adj. R-squared:	0.779
Method:	Least Squares	F-statistic:	318.1
Date:	Tue, 31 May 2022	Prob (F-statistic):	0.00
Time:	10:53:34	Log-Likelihood:	1568.4
No. Observations:	1800	AIC:	-3095.
Df Residuals:	1779	BIC:	-2979.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3623	0.018	20.698	0.000	0.328	0.397
step	0.0154	0.000	66.675	0.000	0.015	0.016
Hurdle	-0.0691	0.007	-10.567	0.000	-0.082	-0.056
Markov	-0.0265	0.019	-1.416	0.157	-0.063	0.010
PCA	-2.109e-17	2.32e-17	-0.910	0.363	-6.66e-17	2.44e-17
Featureset_all	0.1248	0.014	9.075	0.000	0.098	0.152
Featureset_baseline	0.0136	0.018	0.756	0.450	-0.022	0.049
Featureset_broad	-0.0238	0.012	-2.014	0.044	-0.047	-0.001
Featureset_conflicthistory	0.0006	0.009	0.061	0.951	-0.017	0.018
Featureset_conflicthistory_long	-0.0368	0.014	-2.709	0.007	-0.063	-0.010
Featureset_greatest_hits	0.0648	0.011	6.015	0.000	0.044	0.086
Featureset_hh20	0.1032	0.009	11.408	0.000	0.085	0.121
Featureset_prs	-0.0236	0.012	-1.999	0.046	-0.047	-0.000
Featureset_topics	0.0358	0.010	3.678	0.000	0.017	0.055
Featureset_vdem	0.1114	0.011	10.342	0.000	0.090	0.133
Featureset_wdi	0.1132	0.011	10.501	0.000	0.092	0.134
Algorithm__gbm	-0.0587	0.016	-3.653	0.000	-0.090	-0.027
Algorithm_glm	0.0334	0.025	1.317	0.188	-0.016	0.083
Algorithm_histgbm	0.0472	0.014	3.270	0.001	0.019	0.076
Algorithm_lgb	0.0727	0.012	6.054	0.000	0.049	0.096
Algorithm_rf	0.0087	0.011	0.760	0.447	-0.014	0.031
Algorithm_srf	0.1646	0.016	10.250	0.000	0.133	0.196
Algorithm_xgb	0.2151	0.012	18.644	0.000	0.193	0.238

Omnibus:	931.675	Durbin-Watson:	0.641
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17377.487
Skew:	1.991	Prob(JB):	0.00
Kurtosis:	17.691	Cond. No.	1.10e+18

Source: ViEWS, 2022

perform well.[14] Models based on the Varieties of Democracy (vdem) and World Development Indicators (wdi), on the other hand, do not do that well. Some of the models that make use of the feature sets combining indicators from these sets do well, however, in particular the **broad** feature set. The Political Risk Services features also contribute positively to the mix.

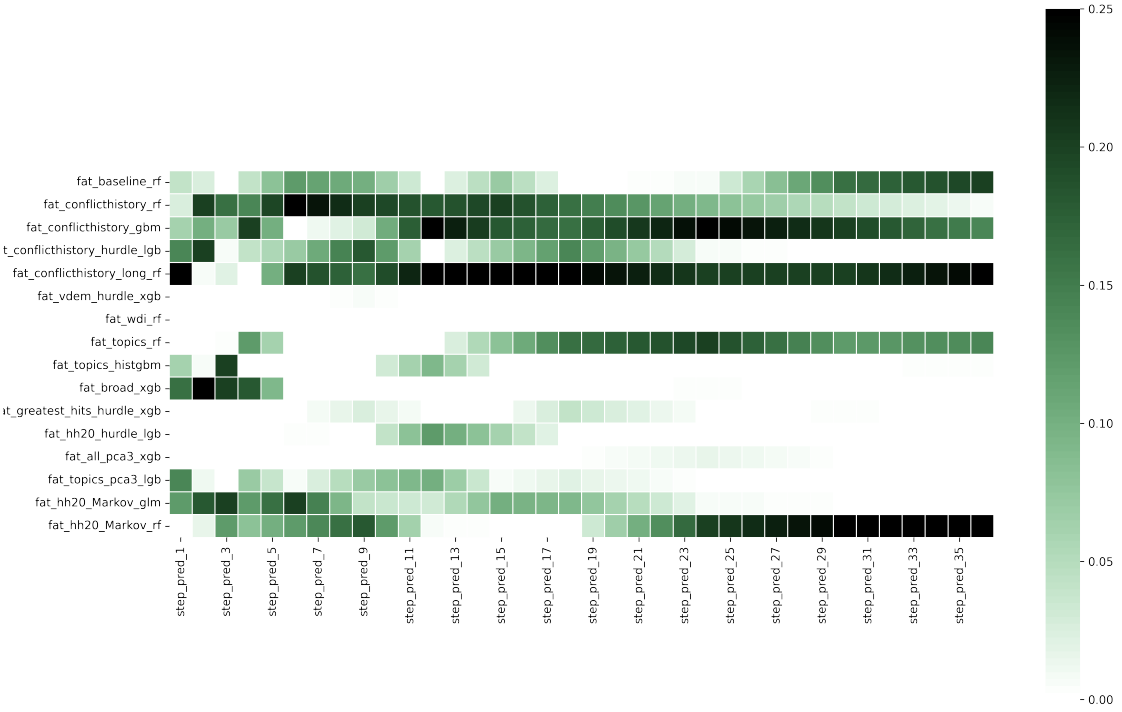
In terms of algorithms, the results suggest that the SciKit gradient boosting model performs very well, and the XGBoost-based random forest model. The other gradient boosting models are less performant.

6.1.3 Ensembles: Performance and estimated weights

It is clear from Figure 15 that the unweighted ensemble performs better than any of the constituent models, at all steps. Tables 2 and 3 show that this holds also for the test partition.

[14] The coefficients in Table 18 for these feature sets are negative or close to zero, i.e. better or comparable to the performance of the ensembles.

Figure 19. **Weights from the genetic algorithm, *cm* level**



Source: ViEWS, 2022

To define a *cm* model ensemble that is more practical for production purposes, and that could be weighted efficiently using the procedure described below, we selected 16 models that all were good performers according to the metrics showed above, that were maximally diverse (Page, 2007), that included all data sources ViEWS will be maintaining at least in one model, and excluded data sources that ViEWS will not be maintaining. The models and the weights they obtain by the genetic weighting procedure are shown in Figure 19.

Figure 19 shows the weights obtained using the genetic algorithm described above. The genetic algorithm selected values from a list of discrete values:

```
[0, 0.001, 0.002, 0.003, 0.005, 0.007, 0.010, 0.015, 0.020, 0.025, 0.030, 0.04, 0.05,
0.06, 0.07, 0.08, 0.09, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.25]
```

These values were chosen so that there were more weight options at the lower end of the scale to allow for low-weight models that pick up on detailed aspects of the prediction problem. The highest weight allowed was 0.30 – no single model was allowed to account for more than 30% of the predictions to avoid

over-fitting to the calibration partition.

The algorithm optimized weights for 11 of the 36 steps:

```
[1,2,4,6,9,12,15,18,24,30,36]
```

For the remaining steps, we calculated the linear interpolation between the most proximate steps. From Figure 19, we see that just over half of the 48 models are assigned weights. Since the predictions from several models are highly correlated, the weighting algorithm in some cases are uncertain what weight to give each of these. For instance, the two Markov models and the two PRS models take turns in being important for some step ranges.

Highly weighted models are, as expected, the same as the models that perform well according to MSE (Figure 15) as well as the ablation score (Figure 16): in particular, two of the conflict history models obtain weights above 0.20 for several steps, as does one of the topics models for a few steps, and the random forest Markov model. Except for the Markov models, most highly performant models make use of relatively sparse, thematic feature sets – in particular conflict history, but also the topics and PRS feature sets. This is not to say that the other features are ignored, since features from V-Dem and WDI are important for all cross-thematic feature sets as well as the Markov model.

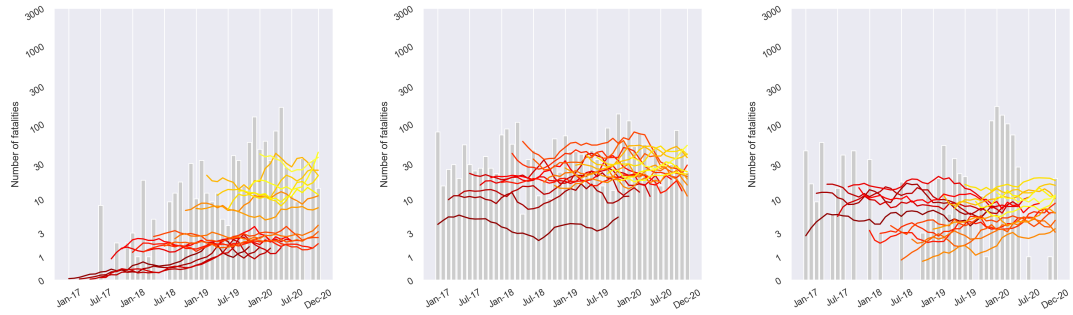
Tables 2 and 3 show that the weighted ensemble outperforms the equally-weighted ensemble for the test partition by a wide margin. MSE for $s = 1$ is 0.233 for the equally weighted ensemble, and 0.227 for the weighted one. To set this in context, MSE for the best model at $s = 1$ is 0.235, at par with the equally-weighted ensemble. At $s = 36$, MSE for the weighted ensemble is 0.460, as compared to 0.541 for the equally-weighted one, and 0.593 for the best individual model. It is clear that the ensemble weighting strategy yields good results, and is more valuable the harder is the prediction problem. Moreover, the weighted ensemble clearly outperforms also the unweighted ensemble.

6.1.4 Detailed predictions

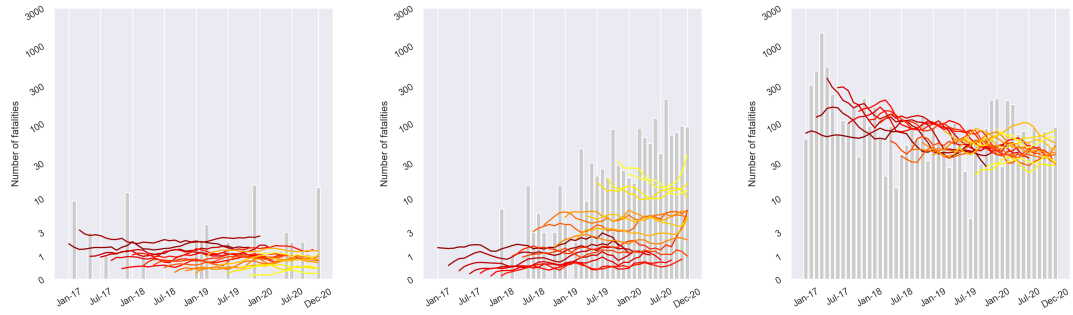
Figure 20 shows the forecasts for the weighted ensemble model for a selection of countries. The bars shows the number of fatalities recorded by the UCDP in each month from January 2017 to December 2020, on a log axis. The lines show predictions for the 36 months starting at the month of the left-hand side . To visualize how model predictions change over time we include such plots for a number of starting points, one line for every second month.

The points forming these lines are the predictions for 1,2,3, etc. months ahead given this set of input data available at that month. The lines are colored so that they gradually change toward yellow as the time of forecast generation moves forward in time. We include only every second month of forecast generation in the figures to increase legibility. For instance, the dark red lines starting in January 2017 are based on data up to and including December 2016 – let us call this the 'time of forecast generation'. For

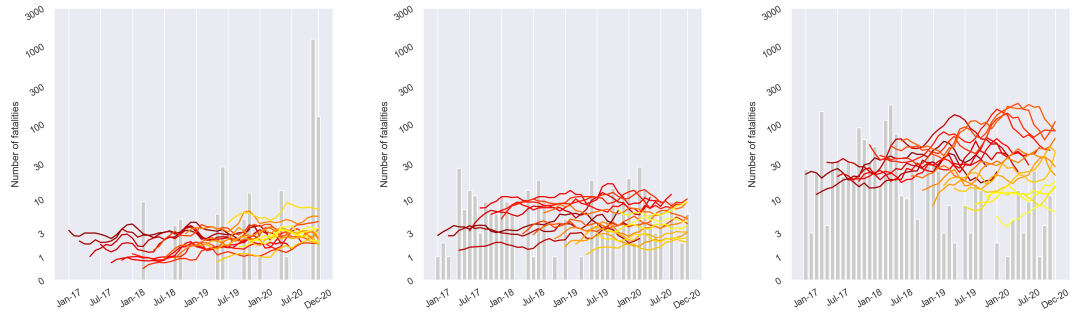
Figure 20. **How forecasts change over time, 12-month periods, January 2017–December 2020, weighted ensemble model forecasts, selected countries**



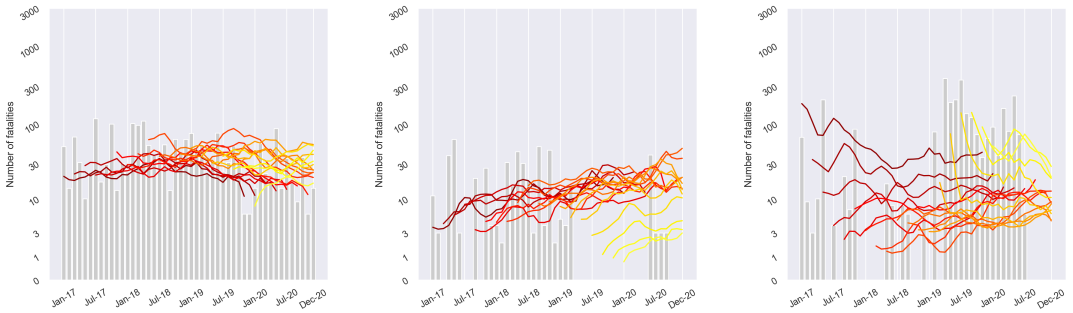
(a) **Ensemble predictions**: Burkina Faso (left), Mali (middle), Niger (right)



(b) **Ensemble predictions**: Angola (left), Mozambique (middle), DR Congo (right)



(c) **Ensemble predictions**: Ethiopia (left), Kenya (middle), South Sudan (right)



(d) **Ensemble predictions**: Egypt (left), Sudan (middle), Libya (right)

Source: ViEWS, 2022

instance, for Burkina Faso in the upper left plot, the dark red line show that the model as of December 2016 predicted close to zero fatalities per month in the first months, but increasing to 1–3 deaths every month by mid-2020, three years into the future (for a view of the UCDP data going into the forecasts, see `https://ucdp.uu.se/country/439`). This predicted future trend of increasing risk reflects the fact that in January 2017, Burkina Faso did not have any preceding state-based violence, but many other risk factors were present, including some preceding violence (In January 2016, the AQIM killed 25 civilians in Ouagadougou city). In July 2017 (month 451), the UCDP recorded 8 fatalities. The month after (light red line), the model predicted more fatalities will follow. As intermittent violence was observed in the subsequent months, the model adjusts its 36-month forecasts upward. By the end of 2018, the model predicts about 10 deaths in each of the following months.

Figure 20 shows how the forecasts change over time for a variety of countries. Some of these saw an escalation of violence from low levels in 2017–2018. Just as in Burkina Faso, Mozambique had been quite calm up to 2016. Mozambique, however, saw some fighting between the government and RENAMO in the central part of the country, so the model suggested some deaths from political violence in the first forecasts in 2017. In November-December 2017, three civilians were killed in by *Ansar al-Sunnah* in the Cabo Delgado province, setting off the following conflict with the government of Mozambique. After a few months of initial violence, the model consistently predict continued violence.

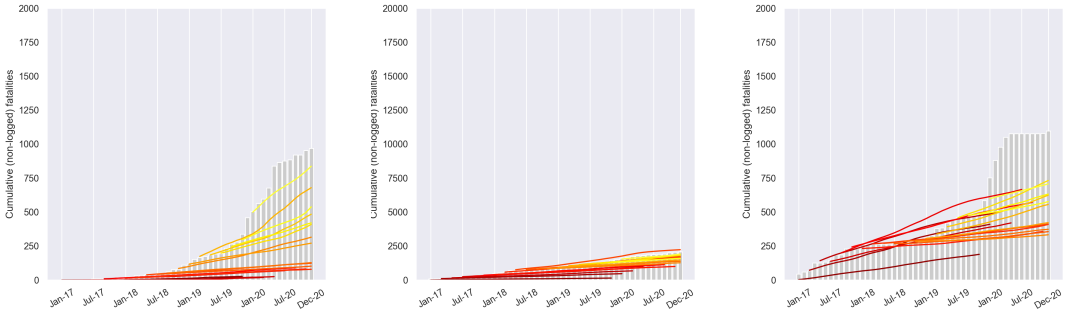
Other countries had established regular fighting before the 2017–2020 we are looking at. In Mali, for instance, state-based armed conflict had been going on for some years at the end of 2016 (see `https://ucdp.uu.se/country/432`). Seen from December 2016 (dark red line), the model predicts relatively low levels of violence, since violence had been relatively muted in the three preceding years. As soon as violence re-escalated from early 2017, the model adjusts forecasts and predicts scores of deaths every month from mid-2017 onwards. Other countries had more intermittent violence during the 2017–2020 period. Angola, for instance, saw some flareups of violence in the Cabinda province. Since this conflict had been going on for a few years but at a very low intensity level, the predictions from early 2017 expected 1–3 deaths per year, but later reduced the risk assessments.

The model correctly predicted de-escalation in DR Congo. In other countries, e.g. Sudan and South Sudan, the model predicted continued escalation, and only slowly adjusted predictions during the peaceful period in 2019 in both countries.

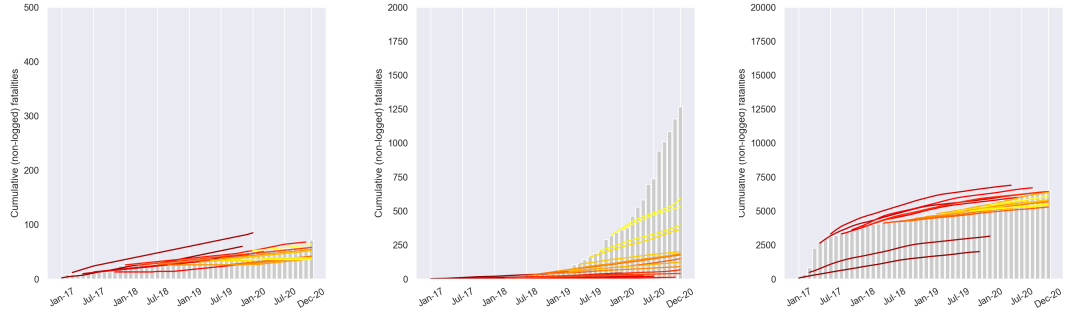
Figure 21 shows the same forecasts for the same countries, but accumulating deaths from January 2017 onwards.

The evaluation of the *cm* level will be invaluable when we continue to improve our modeling. The random forest models varying input features provides lots of information about which features are important, and some lessons regarding the optimal sizes of feature sets for such models. Conflict history is obviously extremely important, and we will continue to improve how we model this. We have in preparation better ways to incorporate spatial and temporal lags, for instance.

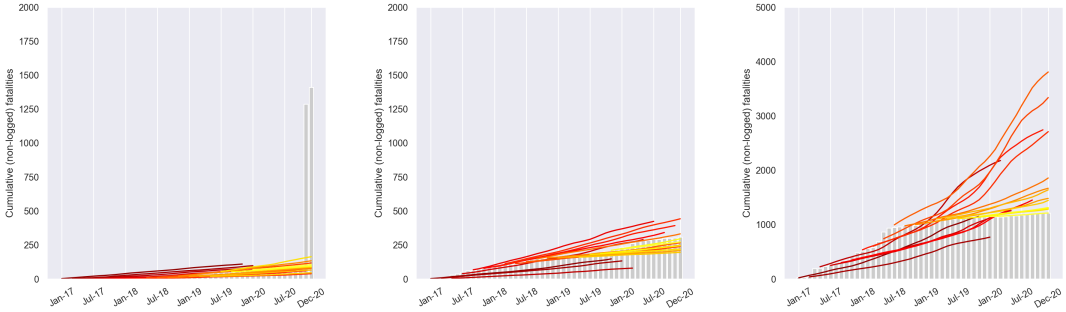
Figure 21. **Cumulative forecasts over time, 12-month periods, January 2017–December 2020**



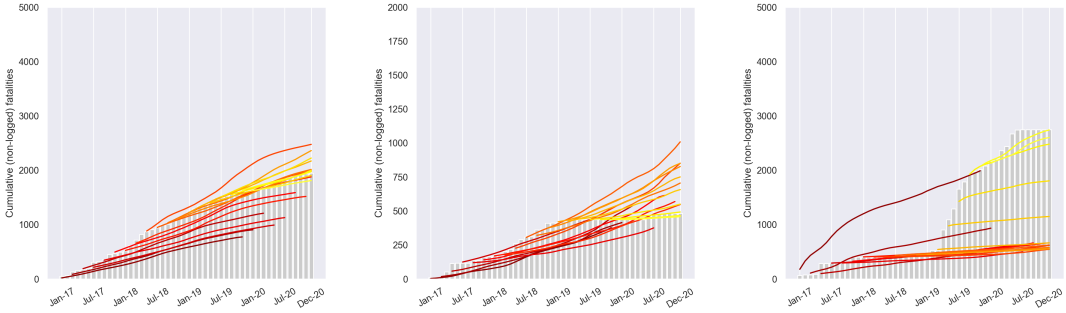
(a) **Ensemble predictions**: Burkina Faso (left), Mali (middle), Niger (right)



(b) **Ensemble predictions**: Angola (left), Mozambique (middle), DR Congo (right)



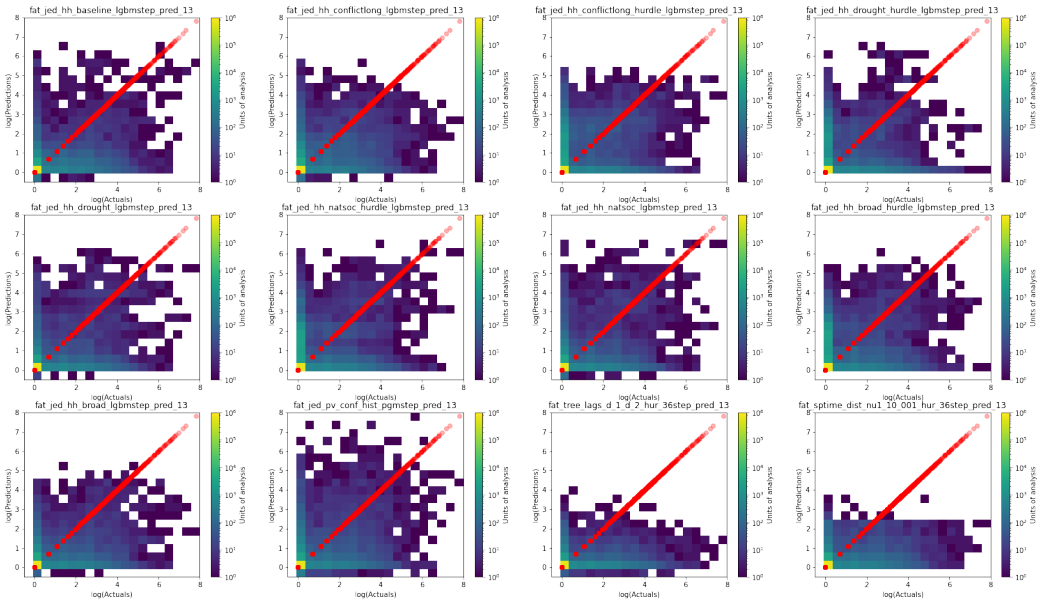
(c) **Ensemble predictions**: Ethiopia (left), Kenya (middle), South Sudan (right)



(d) **Ensemble predictions**: Egypt (left), Sudan (middle), Libya (right)

Source: ViEWS, 2022

Figure 22. **Two dimensional histograms examining the correlation between predicted and observed outcomes in the test partition for the *pgm* constituent models**



Source: ViEWS, 2022

6.2 *pgm* level

6.2.1 Constituent models

We begin by comparing the predicted outcomes and observed outcomes in the test partition, which is most easily done visually by plotting these quantities against each other. The large number of PRIOGRID cells makes the use of scatter plots for this purpose impractical and we therefore construct images of two–dimensional histograms, where each block of colour represents a histogram bin and the colour itself denotes how many units of analysis (datapoints) fall into that bin. A perfect result would result in all the units of analysis falling into the bins along the diagonal of the plot, which represents (predicted outcome = observed outcome). Figure 22 shows such a plot for all twelve pgm constituent models. The red dots indicate the location of the diagonal.

With the exception of the **pv_conf_hist** model, all the constituent models notably underpredict the outcomes in the test partition, since most of the filled histogram bins fall below the red line. All models generate significant numbers of false positives where the observations are zero but the predictions are not and, to a somewhat lesser extent, all models generate false negatives where the predictions are zero but the observations are not. These points in principle represent a severe problem for calibration, which in general involves multiplying data values by normalising factors. The only normalising factor which completely removes a false positive is zero, and no multiplying factor can repair a false negative. Attempting to calibrate the constituent models individually is thus unlikely to yield significantly improved results.

Instead, we will use the principle of ensembling. In an ensemble, the results of constituent models are added together. This should help to ease the problem of false negatives, since it is unlikely that all twelve models will yield a false negative in a given cell. In principle, ensembling may help remove false positives as well, if some models generate *negative* predictions in cells where others produce false positives. Some of the constituent models do indeed yield a small number of negative predictions, as shown in Figure 22, but this is in general undesirable and should not be relied upon to remove false positives.

For reasons that will be discussed below, we have constructed an *unweighted* ensemble of the pgm constituent models. This raises the possibility that, if there is poor consensus between the models (i.e. if, for a given cell, only a few models predict non–zero outcomes), ensembling will simply result in washing out the predictions, yielding predicted outcomes that are everywhere small or zero. In Figure 23, we show two–dimensional histograms denoting the correlations between all possible pairs of pgm constituent models.

The consensus between constituent models is indeed generally rather poor, meaning that models in general do not agree on the outcome in given cells, and the 'washing out´ discussed above in fact eventuates.

6.2.2 Ensembles

We constructed a simple unweighted ensemble model twelve ten of the constituent *pgm* models. As with the constituent models, we construct a 2D histogram in the test partition comparing the outcomes predicted by the ensemble to the true outcomes, shown in Figure 24.

The outcomes predicted by the ensemble are all extremely low and, at least subjectively, its performance must be judged extremely poor.

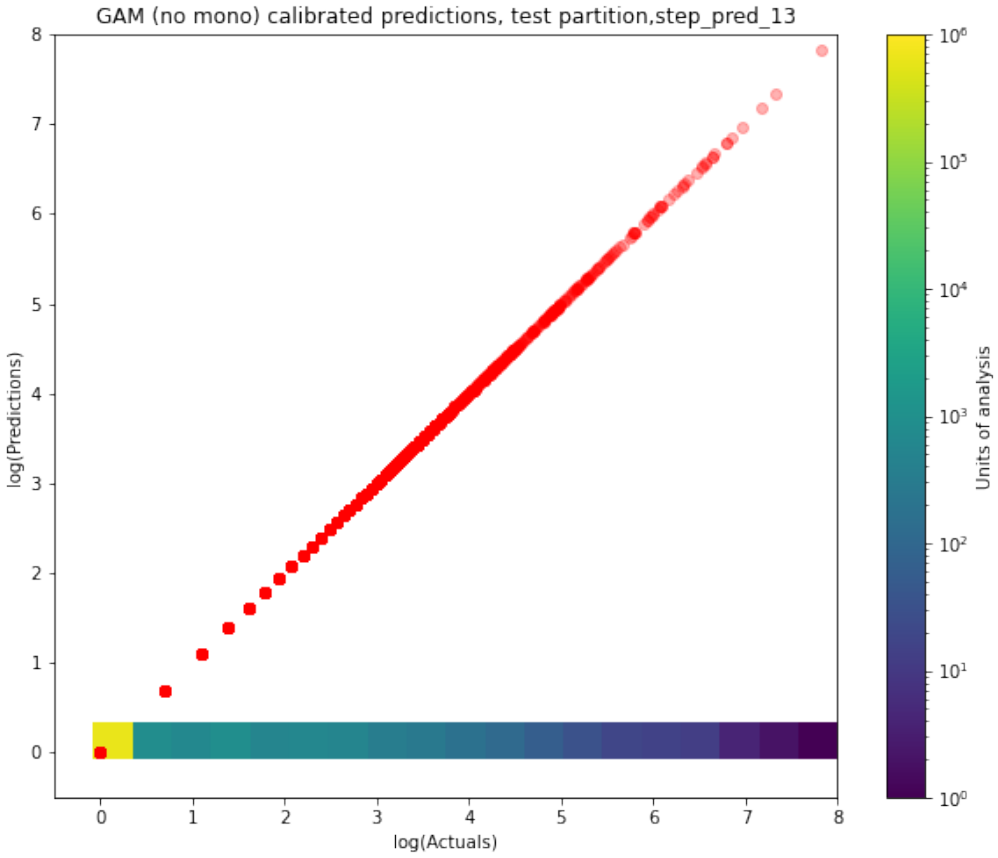
This issue may be alleviated by calibrating the ensemble (as opposed to the constituent models). We employed the same GAM–based calibration technique used for the cm–level predictions but found that the improvement was extremely modest, with predicted numbers of fatalities uniformly much lower than those observed. We find that attempting a GAM fit to the pgm–level data, using the constraint that the fitted values must increase monotonically, yields a fit function, shown in red in Figure 25 which fails to match the dynamic range of the predictions to that of the outcomes, so that the calibrated predictions retain values substantially smaller. The failure of the calibration function to explore the larger values of the observed outcome is likely due to the large numbers of false positives present in the ensemble predictions, which drag the fit function down to lower values. The fit function is then attempting to fulfil two mutually contradictory tasks: suppressing the values of the false positives towards zero, and inflating the values of the true positives towards the desired higher values. The result is then a compromise which achieves neither task particularly well.

The performance of the GAM–calibrated ensemble model, shown in Figure 26, is substantially better than that of the uncalibrated ensemble, but the GAM–calibrated model still notably underpredicts the

Figure 23. **Two dimensional histograms examining the correlation between predicted outcomes in the test partition for all pairs of *pgm* constituent models**

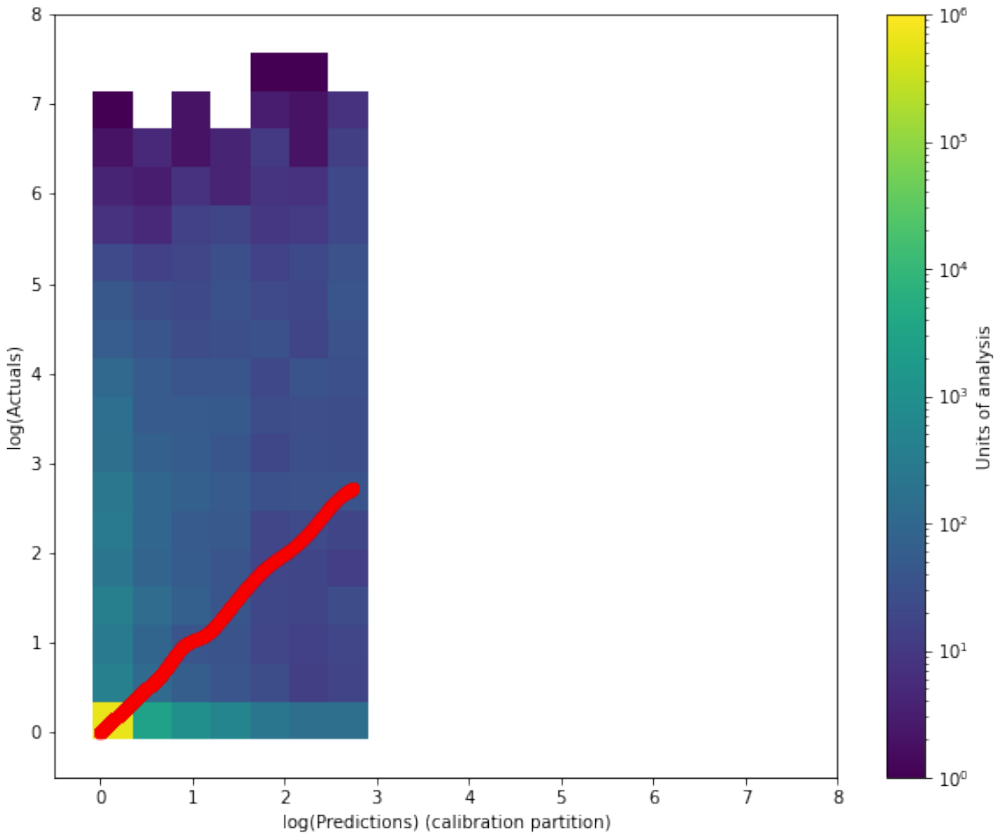
Source: ViEWS, 2022

Figure 24. **Two dimensional histogram examining the performance of the uncalibrated, unweighted pgm ensemble model.**



Source: ViEWS, 2022

Figure 25. **Two dimensional histogram examining showing the observations plotted against the *pgm* ensemble predictions in the calibration partition as a histogram with the result fit shown in red**

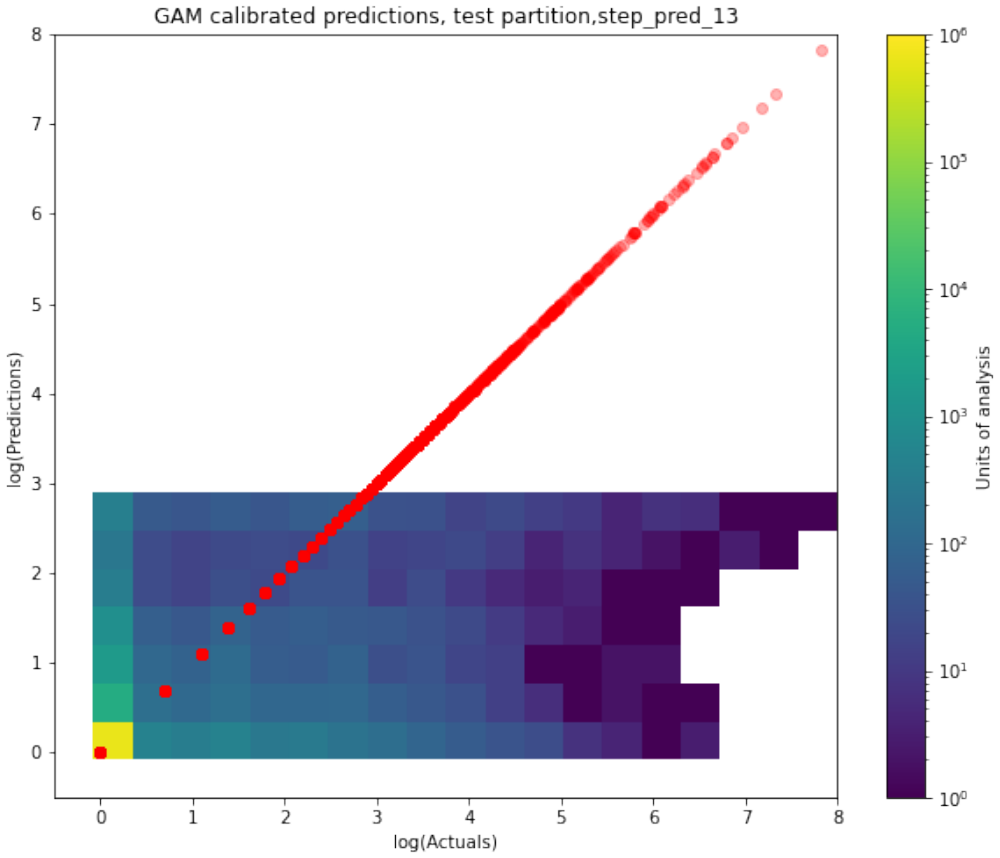


Source: ViEWS, 2022

outcomes, as a result of the inability of the GAM–generated calibration function to access the larger values of the observed outcome. The MSE values derived from the GAM-calibrated predictions were in fact *slightly worse* than those from the uncalibrated predictions (Fig 28).

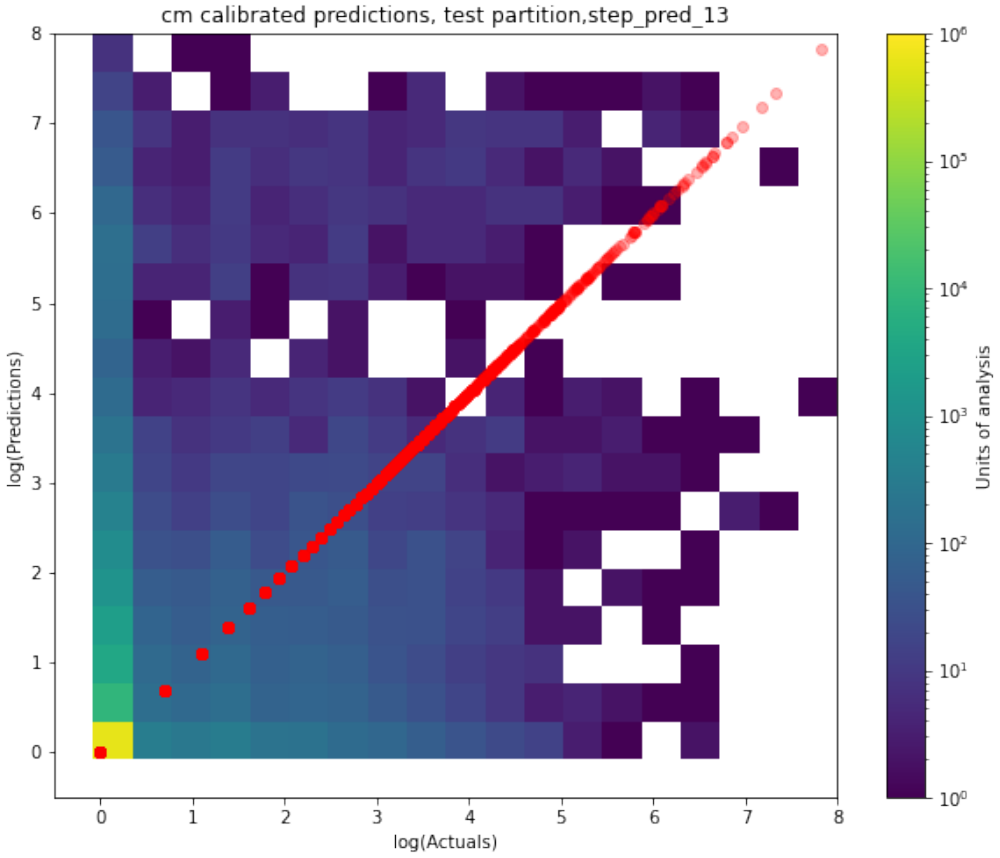
At present, the constituent models and ensembles we have constructed at the pgm level are apparently not able, of themselves, to provide a solution to the normalisation problem. We therefore elected to employ a different calibration method, using the *cm-level predictions* as a baseline. The method is very straightforward: for a given month, for a given country, we identify the group of PRIOGRID cells belonging to that country at that time, sum their predicted fatalities, and then multiply those fatalities by the ratio of the predicted fatalities at the cm–level to the sum. This ensures that the predicted fatalities at pgm–level per country, and the predicted fatalities at cm–level are the same. We justify this technique on the grounds that, since countries are generally much larger geographical units than individual PRIOGRID cells, country–level predictions are *by definition* more reliable.

Figure 26. **Two dimensional histogram examining the performance of the GAM–calibrated, unweighted pgm ensemble model**



Source: ViEWS, 2022

Figure 27. **Two dimensional histogram examining the performance of the cm–calibrated, unweighted *pgm* ensemble model**



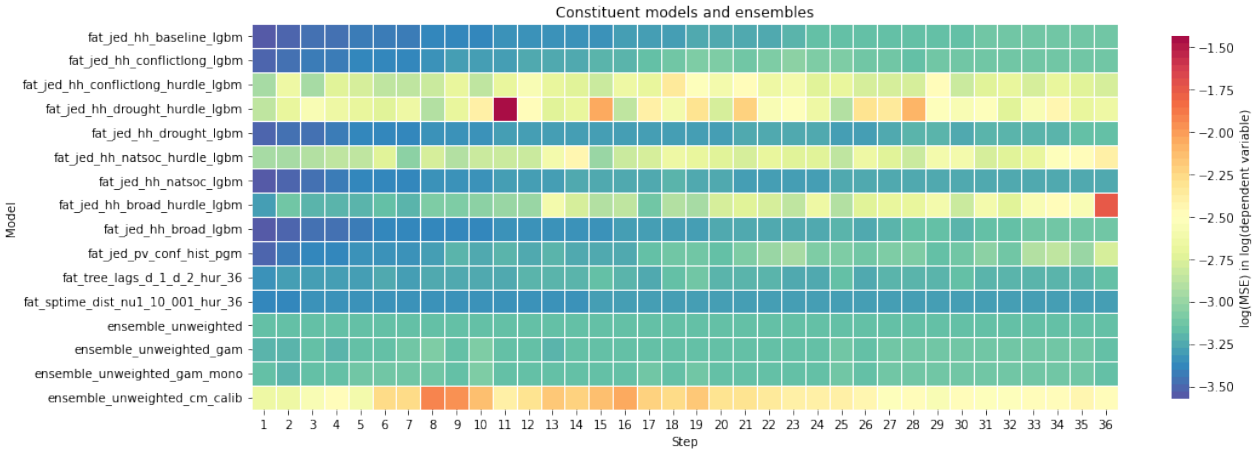
Source: ViEWS, 2022

The process effectively turns the pgm-level predictions into a two-stage process, where the regression analysis *at the PRIO-GRID level* is used to predict in which cells fatalities will occur and the *relative numbers of fatalities between cells*, while independent predictions *at the country level* are used to renormalize these relative numbers of fatalities, yielding absolute numbers which agree across the two levels of analysis (as they should, in any case). Figure 27 shows the result of applying this procedure to the unweighted pgm ensemble, using the weighted cm ensemble as a calibrator.

The cm–calibrated ensemble appears to be superior to the GAM–calibrated model, at least in the sense that the observations and predictions have very similar dynamic ranges. However, as expected, there is still clearly an issue with false positives, which this calibration scheme cannot remedy.

The evaluations of the constituent and ensemble models discussed thus far, based on visual examination of two–dimensional histograms are arguably somewhat subjective and certainly not quantitative. In Figure 28, we show the log of the MSE in the logged dependent variable for all 36 steps for all twelve constituent

Figure 28. **MSE scores for twelve constituent *pgm* models (top 12 rows) and four unweighted ensembles derived from these models (bottom fours rows), for steps 1–36.**



Note: From top to bottom, the ensembles are uncalibrated; GAM–calibrated; GAM–calibrated imposing the monotonic constraint; calibrated using the country–month ensemble, so that the total fatalities in the pg cells belonging to a given country–month are equal to the predicted fatalities for that country–month from the cm–level ensemble.
Source: ViEWS, 2022

models, the uncalibrated ensemble, the GAM–calibrated model discussed above, a second GAM–calibrated model in which the constraint that the fit function be monotonically–increasing is dropped (which yields a very similar model), and the cm–calibrated model.

It is noteworthy that, of the four ensembles, the uncalibrated one is objectively the worst, but has (by a slim margin) the best MSE scores. Additionally, the predictions resulting from the hybrid cm calibration procedure, while evidently yielding much more plausible numbers of fatalities than the either the uncalibrated or GAM–calibrated predictions, *yield MSE scores which are substantially worse*. The reason for this apparent discrepancy appears to be quite straightforward. Observed fatalities in a given month at pg–level generally consist of a rather small number of discrete (i.e. very geographically restricted) events with strongly varying levels of severity, surrounded by non–events (i.e. cells with no fatalities in that month). By contrast, the predicted fatalities for a given month tend to be much more smoothly distributed in space, predicting non–zero fatalities in many cells where the observed values are zero, as well as (in most cases) correctly predicting fatalities where they are indeed observed.

Under these circumstances, two things are likely to be true: (i) the normalisation of the total numbers of predicted fatalities is likely to be too low, requiring calibration to generate plausible numbers of deaths (as detailed above); (ii) the increases in predicted fatalities resulting from renormalising/recalibrating will mainly occur in cells where the *observed* fatalities are zero. This in turn almost guarantees that a better–calibrated model will have a worse MSE score. MSE scores, at least used in the standard fashion, are thus of no help in ranking such models, and may in fact lead one to discard a better model in favour of a worse one.

7 Conclusions

We have presented an initial set of conflict fatality forecasting models and explored their ability to predict the outcome – monthly counts of the number of direct battle-related deaths in state-based armed conflict. At the same time, we have presented a number of tools to evaluate the forecasts and inspect the predictions. We have also developed methods to calibrate and aggregate predictions from constituent models that work well.

References

Armstrong, J. Scott, Kesten C. Green, and Andreas Graefe (2015). "Golden rule of forecasting: Be conservative". In: *Journal of Business Research* 68.8. Special Issue on Simple Versus Complex Forecasting, pp. 1717–1731. DOI: https://doi.org/10.1016/j.jbusres.2015.03.031.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3, pp. 993–1022.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Buhaug, Halvard and Kristian Skrede Gleditsch (2008). "Contagion or Confusion? Why Conflicts Cluster in Space". In: *International Studies Quarterly* 52, pp. 215–233.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Coppedge, Michael et al. (2020). *V-Dem Codebook v10*. Varieties of Democracy (V-Dem) Project.

Croicu, Mihai (2019). "Introducing a large, global, disaggregated, near-real-time conflict actor dataset". In: *the 60th Annual Meeting of the International Studies Association, Toronto, Ontario, Canada*.

Croicu, Mihai and Ralph Sundberg (2015). "UCDP Georeferenced Event Dataset Codebook Version 4.0". In: *Journal of Peace Research* 50.4, pp. 523–532.

Ghobarah, Hazam Adam, Paul K. Huth, and Paul Russett (2004). "The Post-War Public Health Effects of Civil Conflict". In: *Social Science and Medicine* 59, pp. 869–884.

Gleditsch, Nils Petter et al. (2002). "Armed conflict 1946-2001: A new dataset". In: *Journal of peace research* 39.5, pp. 615–637.

Greene, Kevin et al. (2019). *Move It or Lose It: Introducing Pseudo-Earth Mover Divergence as a Context-sensitive Metric for Evaluating and Improving Forecasting and Prediction Systems*. Presented to the 2019 Barcelona GSE Summer Forum, workshop on Forecasting political and economic crisis: Social science meets machine learning'.

Hegre, Håvard (2014). "Democracy and armed conflict". In: *Journal of Peace Research* 51.2, pp. 159–172. DOI: 10.1177/0022343313512852.

Hegre, Håvard (2018). "Civil Conflict and Development". In: *Oxford University Press Handbook on the Politics of Development*. Ed. by Nicholas van de Walle & Carol Lancaster. Oxford: Oxford University Press.

Hegre, Håvard et al. (2019). "ViEWS: A political Violence Early Warning System". In: *Journal of Peace Research* 56.2, pp. 155–174. DOI: 10.1177/0022343319823860.

Hegre, Håvard, Paola Vesco, and Michael Colaresi (2022). "Lessons from an Escalation Prediction Competition". In: *International Interactions* 48.x, pp. 000–000.

Hegre, Håvard et al. (2020). "Introducing the UCDP Candidate Events Dataset". In: *Research & Politics* 7.3, p. 2053168020935257. DOI: 10.1177/2053168020935257. eprint: https://doi.org/10.1177/2053168020935257.

Hegre, Håvard et al. (2021). "ViEWS$_{2020}$: Revising and evaluating the ViEWS political Violence Early-Warning System". In: *Journal of Peace Research* 58.3, pp. 599–611. DOI: 10.1177/0022343320962157. eprint: https://doi.org/10.1177/0022343320962157.

International Food Policy Research Institute (2019). *Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version 1.1*. type: dataset. DOI: 10.7910/DVN/PRFF8V.

Molnar, Christoph (2021). *Interpretable Machine Learning. A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/.

Montgomery, Jacob M, Florian M Hollenbach, and Michael D Ward (2012). "Improving predictions using ensemble Bayesian model averaging". In: *Political Analysis* 20.3, pp. 271–291.

Mueller, Hannes and Christopher Rauh (2018). "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text". In: *American Political Science Review* 112.2, pp. 358–375. DOI: 10.1017/S0003055417000570.

— (2020). *The Hard Problem of Prediction for Conflict Prevention*. Cambridge Working Papers in Economics 2015, Faculty of Economics, University of Cambridge.

— (2022). "Using past violence and current news to predict changes in violence". In: *International Interactions* 48.x, pp. 000–000.

Page, Scott E. (2007). *The difference: how the power of diversity creaetes better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Pettersson, Therése et al. (2021). "Organized violence 1989–2020, with a special emphasis on Syria". In: *Journal of Peace Research* 58.4, pp. 809–825. DOI: 10.1177/00223433211026126. eprint: https://doi.org/10.1177/00223433211026126.

Portmann, Felix T., Stefan Siebert, and Petra Döll (Mar. 2010). "MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling: MONTHLY IRRIGATED AND RAINFED CROP AREAS". In: *Global Biogeochemical Cycles* 24.1, n/a–n/a. DOI: 10.1029/2008GB003435.

Raleigh, Clionadh and Håvard Hegre (2009). "Population, Size, and Civil War. A Geographically Disaggregated Analysis". In: *Political Geography* 28.4, pp. 224–238.

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

Randahl, David and Johan Vegelius (2022). "Predicting escalating and de-escalating violence in Africa using Markov Models". In: *International Interactions* 48.x, pp. 000–000.

Russell, Stuart and Peter Norvig (2020). *Artificial Intelligence: A Modern Approach*. 4th ed. Prentice Hall.

Scrucca, Luca et al. (2013). "GA: a package for genetic algorithms in R". In: *Journal of Statistical Software* 53.4, pp. 1–37.

Servén D., Brummitt C. (2020). *pyGAM: Generalized Additive Models in Python*. Zenodo. DOI: 10.5281/zenodo.1208723.

Sivanandam, SN and SN Deepa (2008). "Genetic algorithms". In: *Introduction to genetic algorithms*. Springer, pp. 15–37.

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

Tetlock, Philip E. (2005). *Expert Political Judgment: How good is it? How can we know?* Princeton: Princeton University Press.

Tollefsen, Andreas Forø (2012). *PRIO-GRID Codebook*. Typescript, PRIO.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug (2012). "PRIO-GRID: A unified spatial data structure". In: *Journal of Peace Research* 49.2, pp. 363–374. DOI: 10.1177/0022343311431287. eprint: http://jpr.sagepub.com/content/49/2/363.full.pdf+html.

Vesco, Paola et al. (2022). "United They Stand: Findings from an Escalation Prediction Competition". In: *International Interactions* 0.0, pp. 1–37. DOI: 10.1080/03050629.2022.2029856. eprint: https://doi.org/10.1080/03050629.2022.2029856.

Vicente-Serrano, Sergio M., Santiago Beguería, and Juan I. López-Moreno (2010). "A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index". In: *Journal of Climate* 23.7, pp. 1696–1718. DOI: 10.1175/2009JCLI2909.1. eprint: https://doi.org/10.1175/2009JCLI2909.1.

Weidmann, Nils B., Doreen Kuse, and Kristian Skrede Gleditsch (2010). "The geography of the international system: The CShapes dataset". In: *International Interactions* 36.1, pp. 86–106.

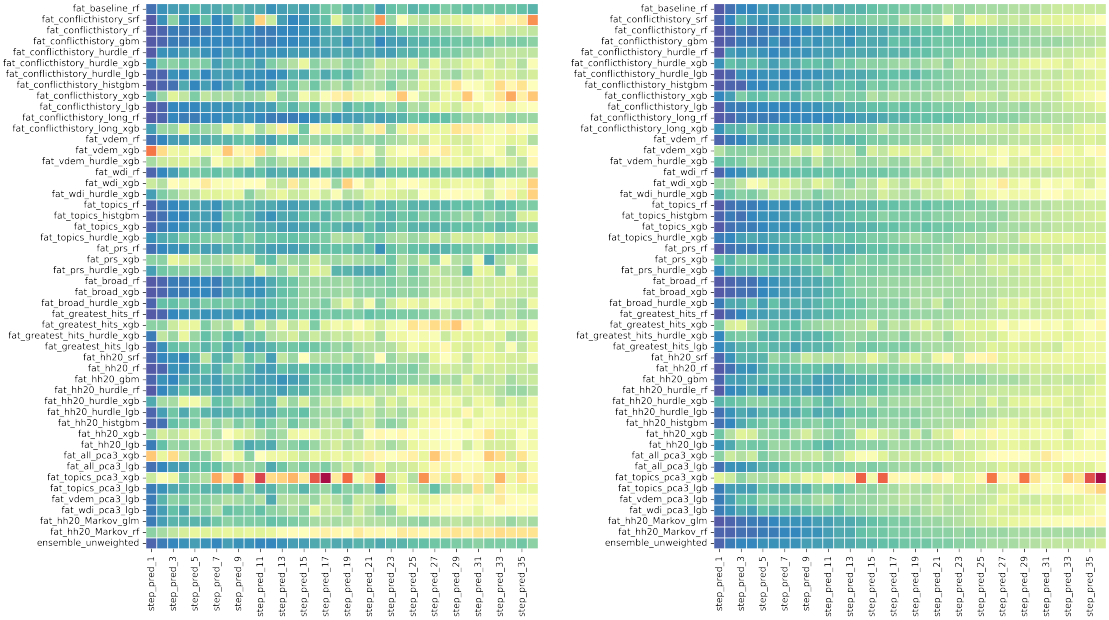
World Bank Group and United Nations (2017). *Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict. Main Messages and Emerging Policy Directions*. International Bank for Reconstruction and Development/The World Bank.

WorldBank (2019). *World Development Indicators*. Washington DC: World Bank.

A-1 APPENDIX

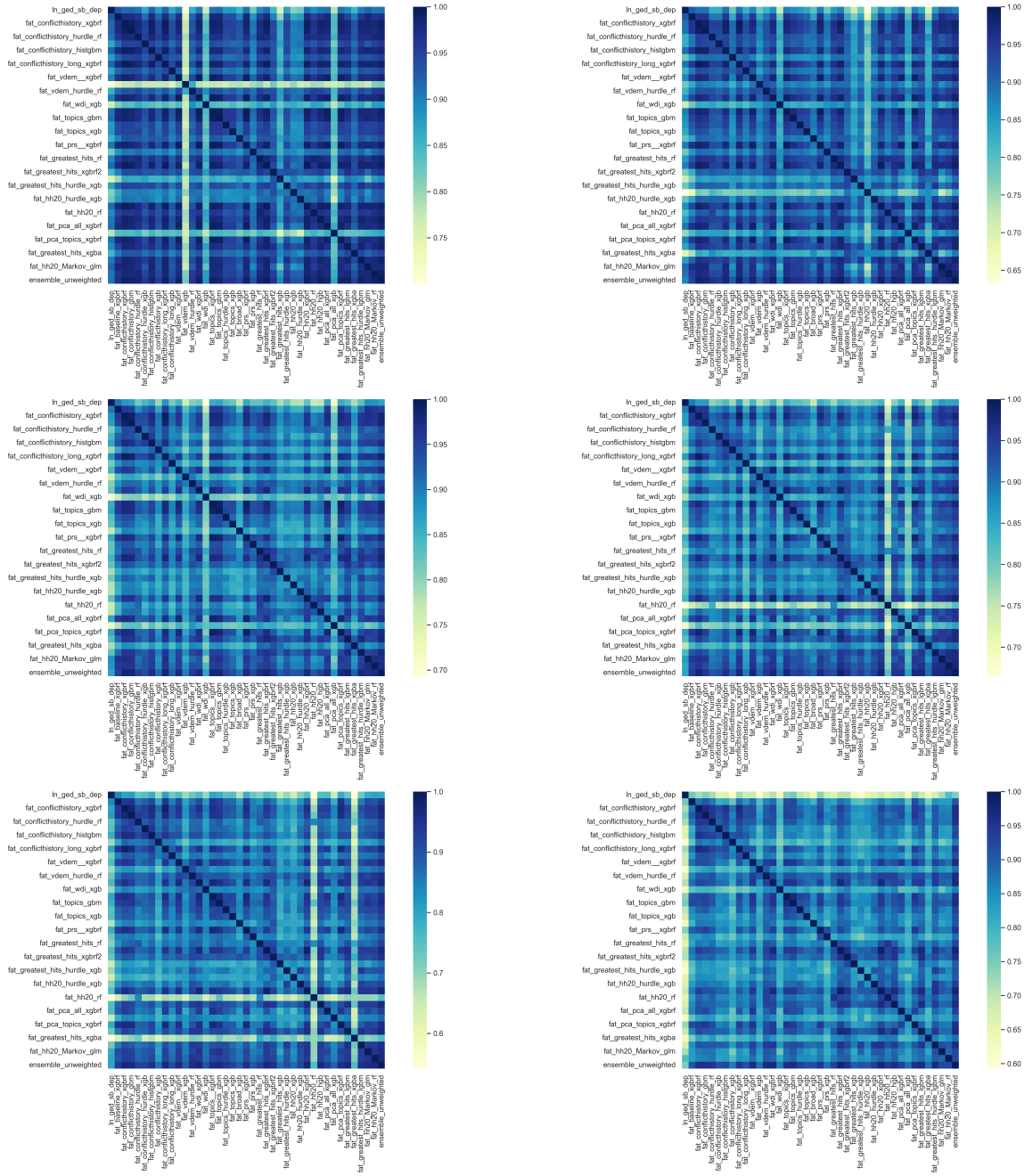
A-1.1 Model selection plots

Figure A-1. **MSEs calibration partition, zero observations (left), non-zeros (right)**



Source: ViEWS, 2022

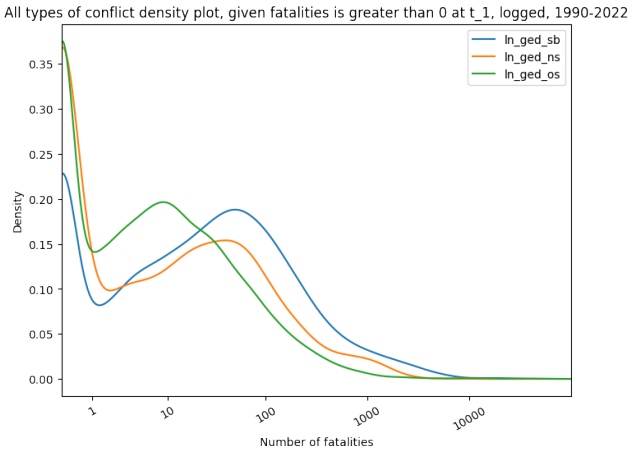
Figure A-2. **Correlation between predictions, calibration partition**



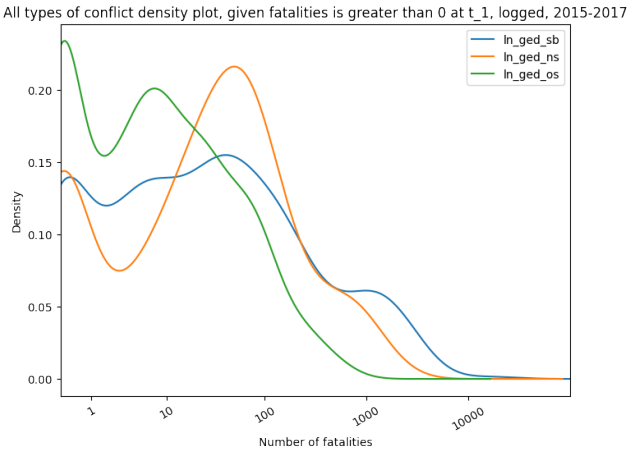
Source: ViEWS, 2022

A-1.2 Figures and descriptives for the outcomes, including non-state and one-sided violence

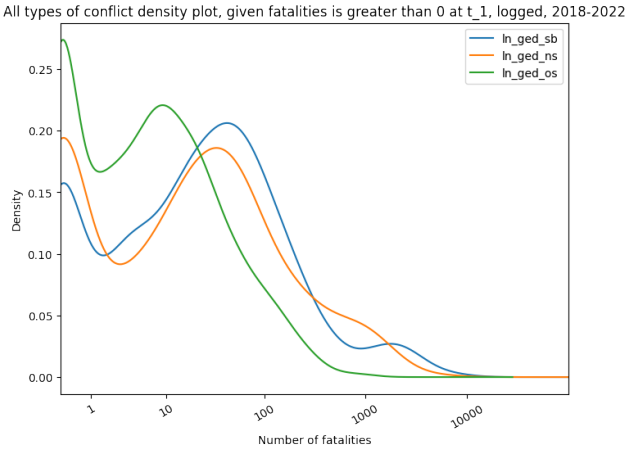
Figure A-3. **Probability of conflict given that the number of fatalities is greater than 0 at t_1 for 1990–2020, 2015–2017 and 2018–2020**



(a) 1990-2020



(b) 2015-2017



(c) 2018-2020

Source: ViEWS, 2022

A-1.2.1 Descriptive tables, cm level

Table A-1. **Summary statistics, cm level for 1990-2020**

Conflict type	Proportions of zeros	Mean	Standard Deviation	Mean (logged)	Standard Deviation (logged)
State based conflict	0.875366	21.622541	333.093440	3.118947	5.811421
Non-state conflict	0.908392	12.229980	1667.839386	1.563573	7.419884
One sided conflict	0.953048	3.775856	53.268666	2.582485	3.993947

Table A-2. **Summary statistics, cm level for 1990-2020**

Conflict type	Proportions of zeros	Mean	Standard Deviation	Mean (logged)	Standard Deviation (logged)
State based conflict	0.849913	22.728185	192.048178	3.166664	5.262940
Non-state conflict	0.882781	10.124636	114.810763	2.409162	3.113305
One sided conflict	0.932519	2.697789	21.495275	1.307735	4.751958