Levels of analysis and the dependent variables[*][†]

Håvard Hegre[1, 2], Sofia Nordenving[2], Mihai Croicu[2], James Dale[2], and Paola Vesco[2]

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

May 6, 2023
VIEWS Version: Fatalities002

Abstract

The VIEWS forecasting model provides monthly forecasts for the number of battle-related deaths expected in impending political violence during each of the next 36 months, as well as the probabilities that these counts will exceed a given threshold. In this paper we describe the current levels of analysis, how the outcome is defined and how historical conflict data is distributed.

[*] The research was funded by the European Research Council, project H2020-ERC-2015-AdG 694640 (ViEWS), the European Research Council Advanced Grant project *ANTICIPATE*, Riksbankens Jubileumsfond project *Societies at Risk*, Uppsala University, the Swedish Research Council project *DEMSCORE*, the UK Foreign, Commonwealth and Development Office, the United Nations High Commissioner for Refugees, and Peace Research Institute Oslo. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). For more information on the project see `viewsforecasting.org`.

[†] Forogh Akbari, Michael Colaresi, Frederick Hoyles, Remco Jansen, Angelica Lindqvist-McGowan, Maxine Ria Leis, Malika Rakhmankulova, David Randahl, and Espen Geelmuyden Rød contributed to earlier versions of this documentation document.

Contents

1 Levels of analysis

1.1 Country months

VIEWS generates forecasts at two levels of analysis: country-months (Gleditsch and Ward, 1999, abbreviated *cm* in VIEWS), and sub-national geographical location months (*pgm*). The *cm* level is particularly useful to provide predictions for entirely new conflicts where no known actors exist, and to model tensions and processes at the governmental level. The set of countries is defined by the Gleditsch-Ward country code (Gleditsch and Ward, 1999, with later updates), and the geographical extent of countries by the latest version of CShapes (Weidmann, Kuse, and Gleditsch, 2010). For the country level of analysis VIEWS provide global forecasts.

1.2 PRIO-GRID months

For the subnational forecasts, VIEWS relies on PRIO-GRID (version 2.0; Tollefsen, Strand, and Buhaug, 2012), a standardized spatial grid structure consisting of quadratic grid cells that jointly cover all areas of the world at a resolution of 0.5 x 0.5 decimal degrees. Near the equator, a side of such a cell is 55 km. This resolution is close to the precision level of the data we have for the outcomes. Investigating the spatial error of the UCDP-GED in Afghanistan, Weidmann (2014, p.1143) found that most events were "located within 50 km of where they actually occured". Given this, a finer resolution might not yield more precise forecasts. For the subnational level of analysis, we currently restrict forecasts to Africa and the Middle East.

Note that the *cm* and *pgm* definitions are not fully compatible with each other. PRIO-GRID provides a 1:1 cell-to-country correspondence by assigning the grid cell to the country taking up the largest area (Tollefsen, 2012). When PRIO-GRID cells span two or more countries, all events contained in that PRIO-GRID cell are aggregated, ignoring which country they actually took place in. In the country-month dataset, such events are assigned to the country where the event took place. Moreover, PRIO-GRID cells exist for the entire duration of the dataset, but only those months in which a country has existed in the Gleditsch and Ward (1999) country list are included in the *cm* datasets.

The grid-level structure has been retrieved directly from the PRIO-GRID API to ensure full compatibility.

2 Dependent variables

2.1 Defining armed conflict

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 three 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 (Tollefsen, Strand, and Buhaug, 2012).

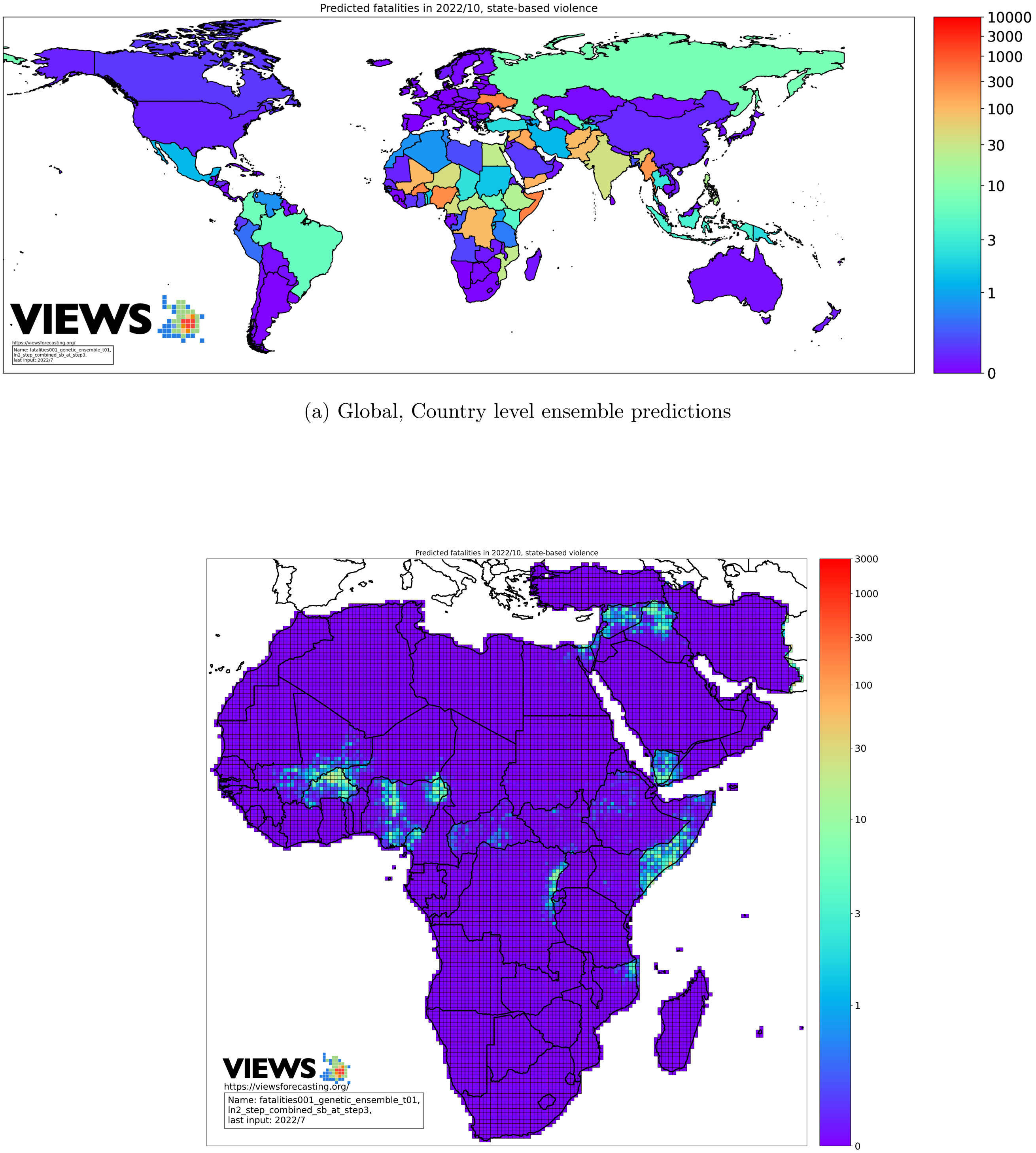
Historical data covering 1989–2021 are extracted from the UCDP-GED version 22.1 (Högbladh, 2022; Sundberg and Melander, 2013).[1] Newer data are provided by the UCDP-Candidate dataset which is updated monthly (Hegre et al., 2020). This allows use of conflict event data up to one month before the forecasting window. Since the candidate data are coded using a smaller set of sources than the final UCDP-GED data, there are some discrepancies between the two (Hegre et al., 2020). The candidate data are replaced with final UCDP-GED data as they come available through each annual release of GED.

2.2 Dichotomous version

With previous VIEWS models (Hegre et al., 2021), we defined the outcomes to be the predicted probability of violence leading to at least 25 battle-related deaths in a given country-month and at least one battle-related death (BRD) per month at the subnational level of analysis.

Datasets from the previous probability model are still available through the VIEWS API. These datasets have the prefix $r\_$ (see the list of available datasets).

[1] The UCDP-GED raw data are publicly available through the UCDP-GED API (Croicu and Sundberg, 2013). VIEWS automatically retrieves these data from the API each month and aggregates to the VIEWS units of analysis. Usage of the API is described at `http://ucdp.uu.se/apidocs/`; the data are available as version 22.1 (1989–2021).


Predicted fatalities in 2022/10, state-based violence
10000
3000
1000
300
100
30
10
3
1
0
VIEWS
https://viewsforecasting.org/
Name: fatalities001_genetic_ensemble_t01, ln2_step_combined_sb_at_step3, last input: 2022/7
Predicted fatalities in 2022/10, state-based violence
3000
1000
300
100
30
10
3
1
0
VIEWS
https://viewsforecasting.org/
Name: fatalities001_genetic_ensemble_t01, ln2_step_combined_sb_at_step3, last input: 2022/7

(a) Global, Country level ensemble predictions

(b) Africa and the Middle East, PRIO-GRID level

Figure 1. Geographical scope

2.3 Fatalities count version

The outcome from the current fatalities model (Hegre et al., 2022) is defined as predicted number of fatalities per country-month and PRIO-GRID month.

In the VIEWS API these datasets are available under the model name $fatalities$. The current version is 002 (see the list of available prediction datasets).

The outcome variables in the API follow the same structure. For instance, state-based violence has the variable name *sc_cm_sb_main*, where *sc* stands for step combined, *cm* for country-month, *sb* for state-based and finally, *main* indicate that the output is from the main ensemble model. The outcome is defined as predicted fatalities per country-month in impending state-based conflict, expressed in natural logged form plus 1 ($ln(fatalities + 1)$).

The fatalities model version likewise produce a dichotomous probability variable, indicating the probability that the fatalities count will exceed the thresholds described above. To generate these predictions we exponentiate the outcome to get the real number of predicted fatalities and replace them with 0 if the prediction is below 25 BRDs for the country level and 1 BRD for the PRIO-GRID level and 1 otherwise. Second, we run a logistic-regression model to predict this new dichotomous outcome, using the main ensemble model predictions at each step. These probabilities constitute our dichotomous predictions.

The naming convention for these variables follow the same structure, with an addition to indicate that it is the dichotomous version. For instance, *sc_cm_sb_dich_main* provides dichotomous predictions for the probability of at least 25 BRDs per country-month in impending state-based conflict.

3 The historical distribution of fatalities – A prediction problem

The dependent variables: descriptive statistics

For most country-months, the UCDP records no violence at all. Most of the remaining months have a low number of fatalities, but a significant and politically important proportion sees extreme levels of violence (Hegre et al., 2022, p.7). Most statistical models are ill-equipped to model such distributions. Building on the basic research carried out by the VIEWS project and others over the past years, this has however now become feasible. Table 1 shows descriptive statistics for the various dependent variables we use.

3.1 Country (*cm*) level

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.

The outcome variable has a distribution that is challenging to forecast. Most observations between 1990 and 2022 are zeros (no conflict fatalities): at the country-month (*cm*) level, 83.03% of the observations

Variable	State-based	One-sided	Non-State
Country month			
$\geq$ 1 BRDs	0.129	0.096	0.049
$\geq$ 25 BRDs	0.067	0.027	0.022
Mean BRDs, all country months	24.093	11.980	4.114
St. Dev.	452.214	1644.646	56.656
Median BRDs, all country months	0.000	0.000	0.000
PRIO-Grid month			
$\geq$ 1 BRDs	0.0046	0.0023	0.0014
$\geq$ 25 BRDs	0.0008	0.0003	0.0003
Mean BRDs, all grid months	0.1674	0.0316	0.1507
St. Dev.	44.5456	2.8541	81.4813
Median BRDs, all grid months	0.0000	0.0000	0.0000

Table 1. Descriptive statistics of dependent variables 1990 – 2022.
BRDs: Battle-Related Deaths

are zeros, and at the PRIO-GRID month (*pgm*) level, 99.29% are zeros.

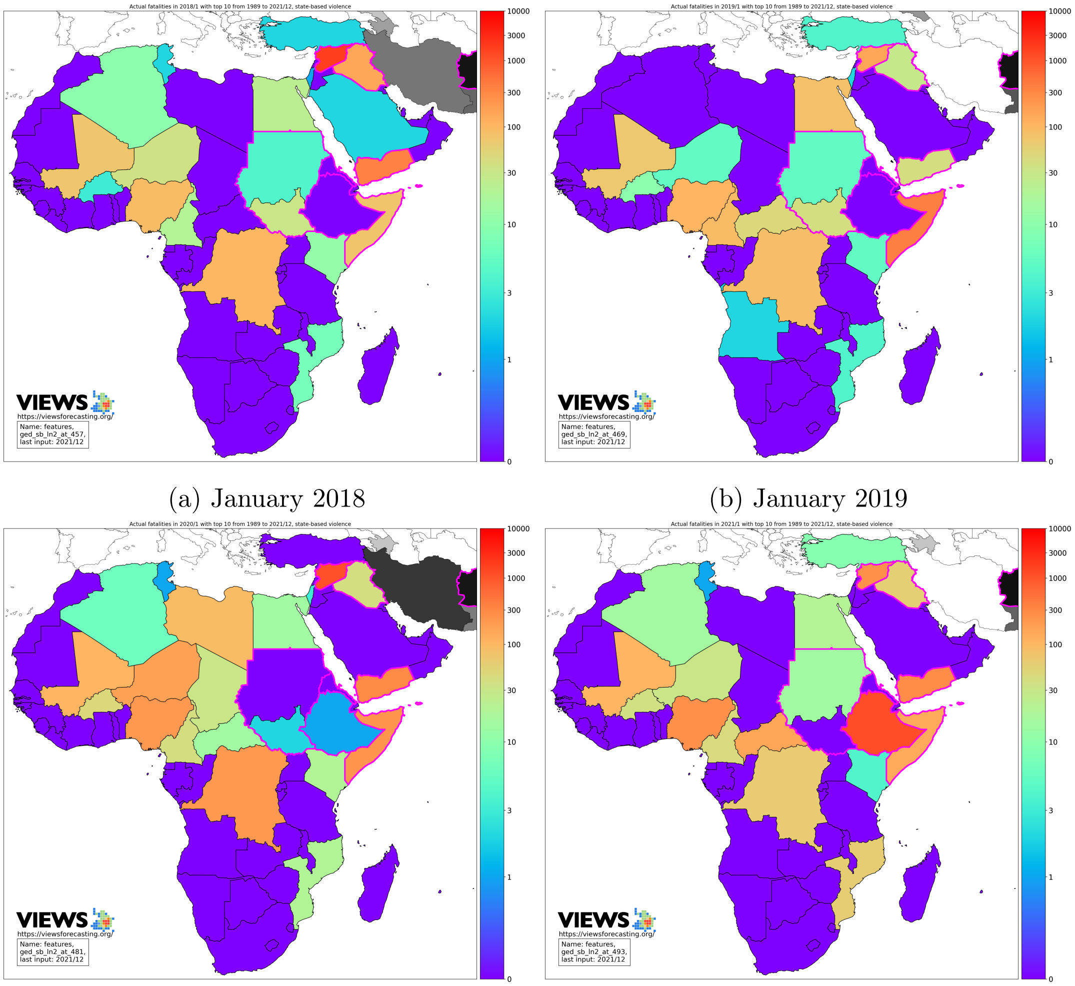
In addition to this 'zero inflation', the distribution of non-zero death counts is heavily right-skewed. Figure 3 shows density plots of the distribution of fatality counts for all three types of violence and at both levels of analysis, restricted to non-zero observations. The x axes are in log form for both sub-figures. For PRIO-GRID months (figure b) the distribution is even more right-skewed than at the country level.

Over the 1990–2022 period, there were 9,587 country-months with state-based conflict. The median number of fatalities was 27 and the mean 187. 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 of non-zero fatality counts is heavily right-skewed. In 110 out of the 9,587 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.[2]

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

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 4 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 (2018–2022). Most non-zero observations follow another non-zero observation – a lagged dependent variable is a very strong

[2] 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.


Actual fatalities in 2018/1 with top 10 from 1989 to 2021/12, state-based violence
Actual fatalities in 2019/1 with top 10 from 1989 to 2021/12, state-based violence
Actual fatalities in 2020/1 with top 10 from 1989 to 2021/12, state-based violence
Actual fatalities in 2021/1 with top 10 from 1989 to 2021/12, state-based violence
10000
3000
1000
300
100
30
10
3
1
0
VIEWS
https://viewsforecasting.org/
Name: features, ged_sb_ln2_at_457, last input: 2021/12
Name: features, ged_sb_ln2_at_469, last input: 2021/12
Name: features, ged_sb_ln2_at_481, last input: 2021/12
Name: features, ged_sb_ln2_at_493, last input: 2021/12

(a) January 2018

(b) January 2019

(c) January 2020

(d) January 2021

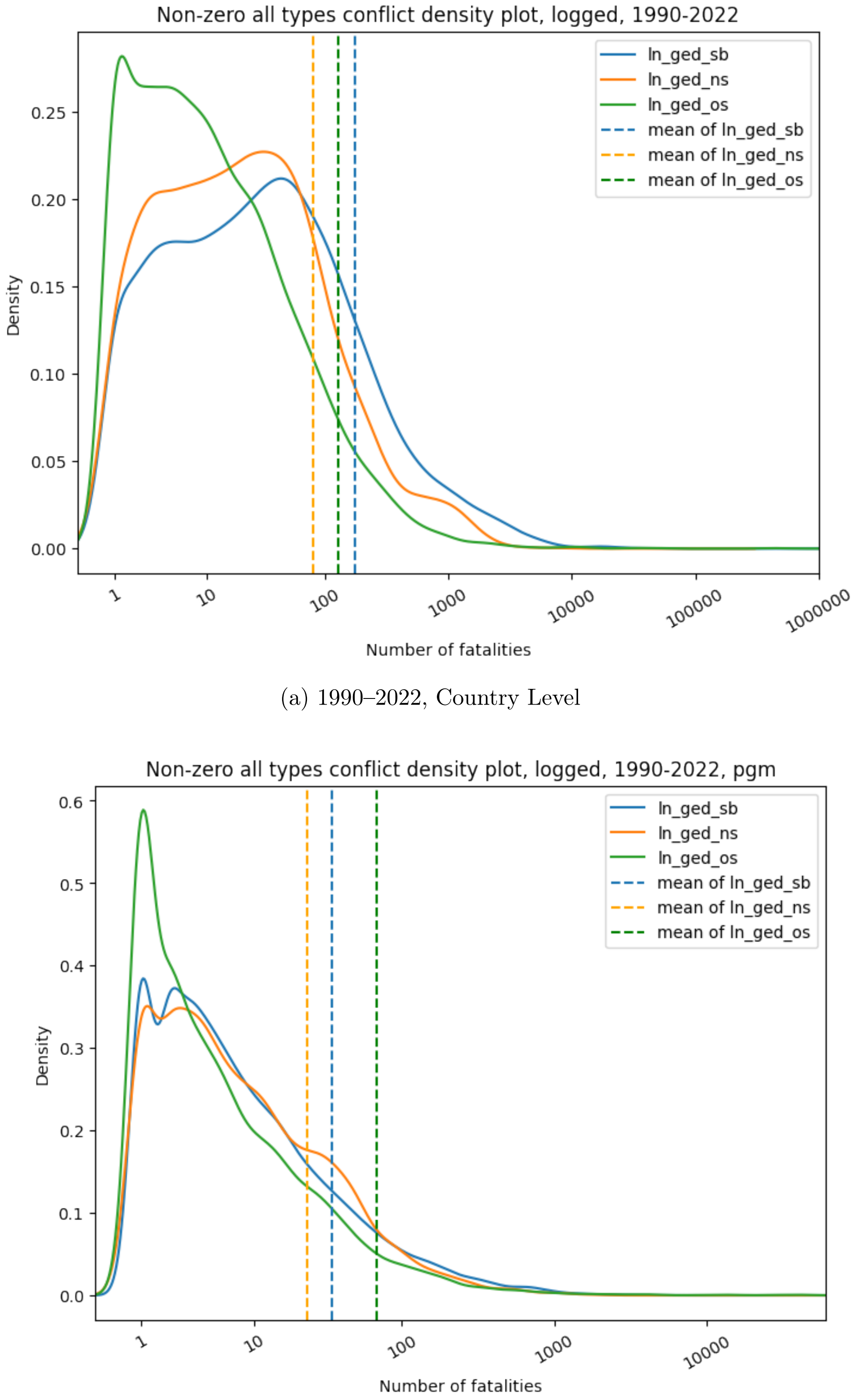
Figure 2. Actual fatalities in Africa and Middle East, state-based conflict. The 10 countries with the most fatalities over the 1989–2021 period globally are marked off with thick pink borders. Seven out of the ten countries are in Africa and the Middle East.

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.

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 5a shows the number of fatalities per month for the 10 most deadly conflict countries over the past 30 years. It shows that the global total of state-based violence over the 1990–2022 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.[3] We identified the 10 most deadly countries by summing up all fatalities by conflict sub-type.[4] Figure 5b shows the global total number of fatalities across all countries for the 1990–2022 period.

[3] 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.

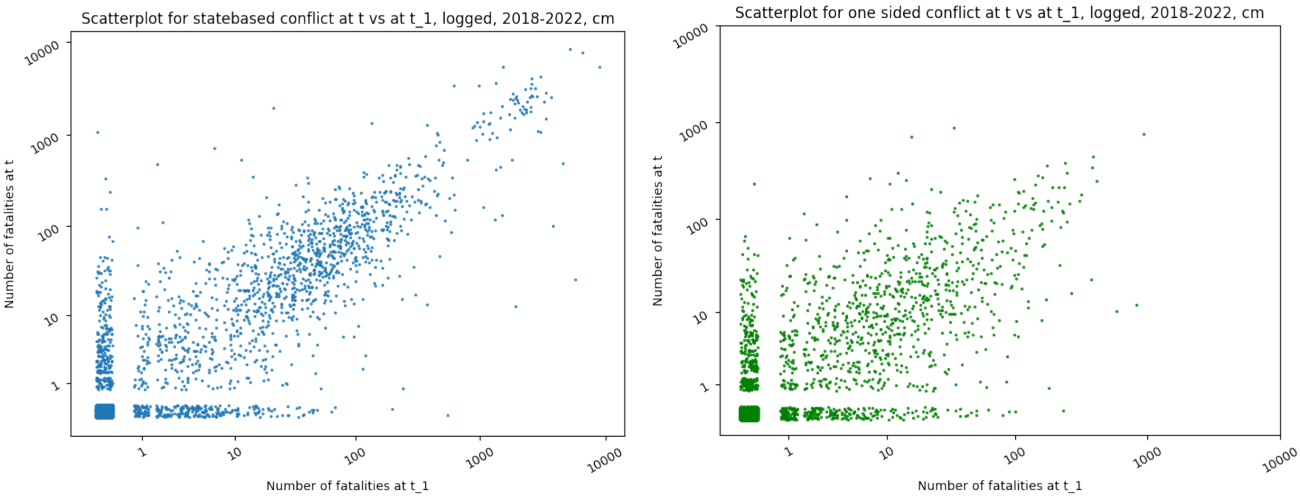
[4] We aggregated counts by country ID. Following Weidmann, Kuse, and Gleditsch (2010), some countries are assigned a new distinct country IDs when its territory changes. For that reason, countries can appear multiple times in the figures.



(a) 1990–2022, Country Level

(b) 1990–2022, PRIO-GRID level

Figure 3. Kernel density plots for all country-months/PRIO-GRID-months with non-zero fatality counts 1990–2022. The vertical lines show the mean (non-logged) fatality counts for the non-zero observations. *Source:* UCDP GED, 2022



(a) State based conflict

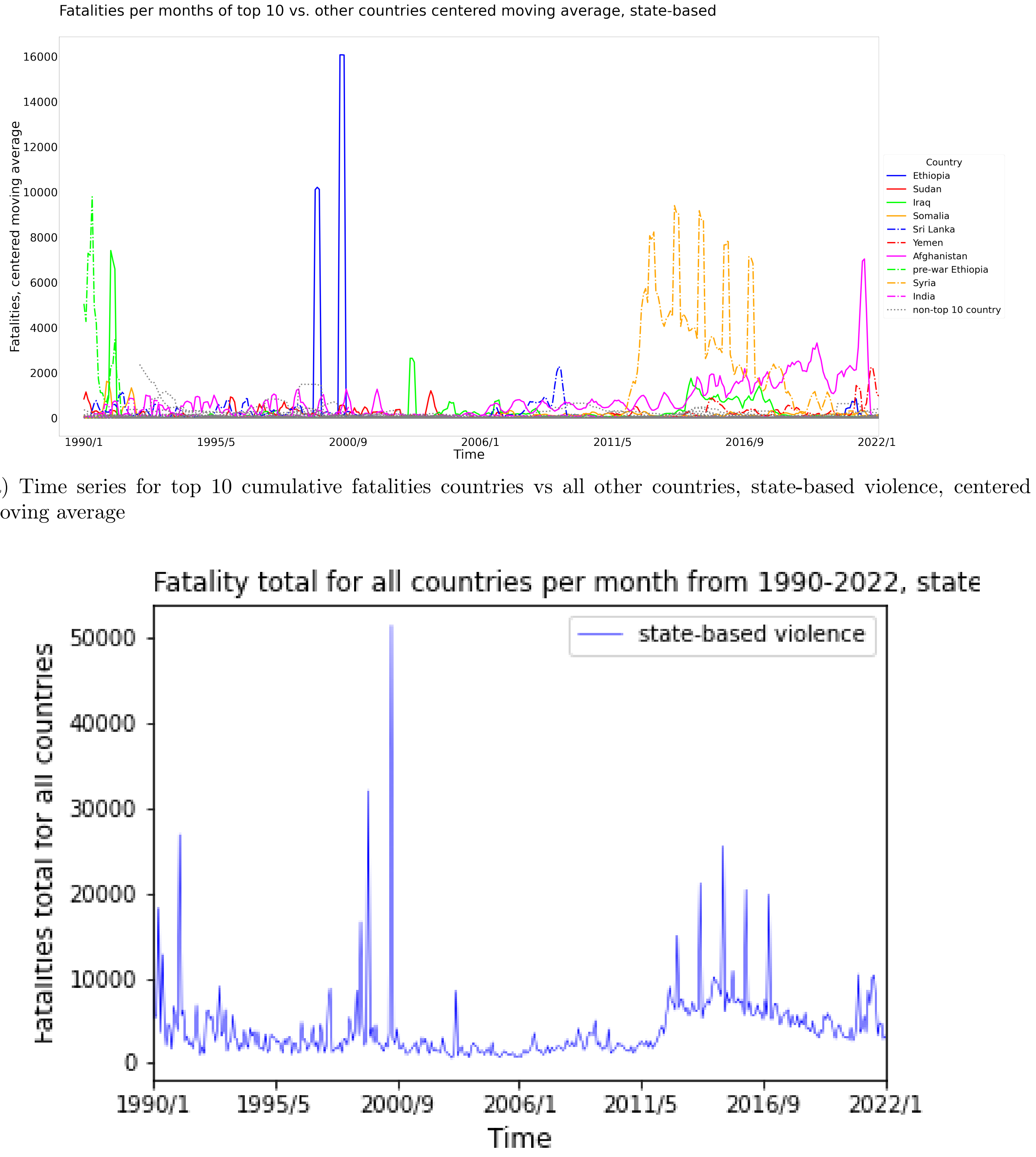
(b) One sided conflict

Figure 4. 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

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). The median number of fatalities lay between 4 and 6, 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 be able to make forecasts that incorporate such extreme events if possible.

Several current conflict hotspots are similarly concentrated. The 2020 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.

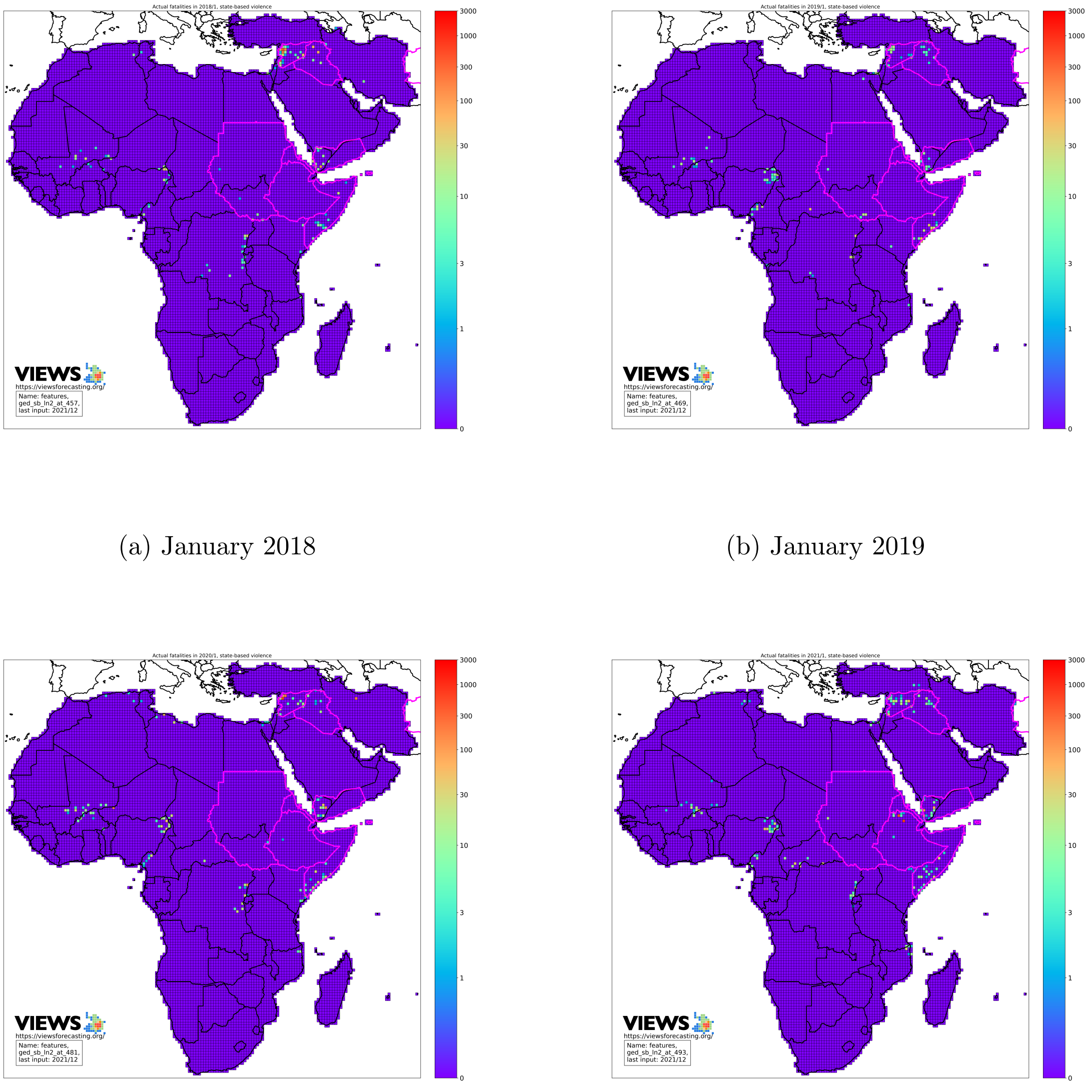


(a) Time series for top 10 cumulative fatalities countries vs all other countries, state-based violence, centered moving average

(b) Time series for fatalities for all countries, state-based violence, 1990–2022

Figure 5. State-based fatalities over time.

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


Actual fatalities in 2018/1, state-based violence
Actual fatalities in 2019/1, state-based violence
Actual fatalities in 2020/1, state-based violence
Actual fatalities in 2021/1, state-based violence
3000
1000
300
100
30
10
3
1
0
VIEWS
https://viewsforecasting.org/
Name: features, ged_sb_ln2_at_457, last input: 2021/12
Name: features, ged_sb_ln2_at_469, last input: 2021/12
Name: features, ged_sb_ln2_at_481, last input: 2021/12
Name: features, ged_sb_ln2_at_493, last input: 2021/12

(a) January 2018

(b) January 2019

(c) January 2020

(d) January 2021

Figure 6. Actual fatalities in Africa and Middle East, state-based conflict at PRIO-GRID (*pgm*) level. The 10 countries with the most fatalities over the 1989–2021 period globally are marked off with thick pink borders. *Source:* UCDP GED, 2022

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 paper describing our models we discuss briefly how they 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.

4 Change history

4.1 Fatalities002

No change to units of analysis and outcome definitions.

4.2 Fatalities001

The fatalities count model was introduced with the Fatalities001 version, thanks to funding from the UK FCDO. The extension was first presented in Hegre et al. (2022). The dichotomous (conflict/no conflict) forecasts are still made available every month and complement the estimated battle-related deaths predictions. However, refining an early-warning system so it indicates whether a future conflict will cause 100, 1000, or 10,000 deaths has significantly pushed the scientific envelope and provide policy-makers and researchers with the ability to quantify the potential impact and intensity of conflicts, and promote opportunities for early preventative action.[5] From February 2022, VIEWS has published monthly updates of the fatalities model through its API, but for state-based conflict only.

4.3 ViEWS-ESCWA

The VIEWS system was expanded to cover the Middle East (including Turkey and Iran) thanks to funding from the UN ESCWA (Theisen et al., 2021).

[5] 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. 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 and major conflagrations. 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).

4.4 ViEWS2020

From ViEWS2020 (Hegre et al., 2021), we changed the definition of the dichotomous outcome at the *cm* level from at least 1 death per month to at least 25 deaths per month.

4.5 ViEWS2018

The first version of the ViEWS early warning system, the 'ViEWS2018' version launched in July 2018 (Hegre et al., 2019) only provided forecasts in the dichotomous form.

References

Croicu, Mihai and Ralph Sundberg (2013). *UCDP Georeferenced Event Dataset Codebook Version 4.0.* Typescript, Uppsala Conflict Data Program.

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, Kristian S. and Michael D. Ward (1999). "A Revised List of Independent States since the Congress of Vienna". In: *International Interactions* 25.4, pp. 393–413.

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

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 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.

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. (2022). *Forecasting Fatalities.* Uppsala: working paper.

Högbladh, Stina (2022). *UCDP GED Codebook version 22.1.* Department of Peace and Conflict Research, Uppsala University.

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.

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.

Theisen, Ole Magnus et al. (2021). *Understanding the Potential Linkages between Climate Change and Conflict in the Arab Region.* E/ESCWA/CL6.GCP/2021/TP.9. UN ESCWA, Beirut, Lebanon.

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.

Weidmann, Nils B. (2014). "On the Accuracy of Media-based Conflict Event Data". In: *Journal of Conflict Resolution* Online first, DOI: 10.1177/0022002714530431, pp. 1–21.

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.