# ViEWS: A political Violence Early Warning System

Håvard Hegre[1], Marie Allansson[1], Michael Colaresi[1,2], Mihai Croicu[1], Hanne Fjelde[1], Frederick Hoyles[1], Lisa Hultman[1], Stina Högbladh[1], Remco Jansen[1], Naima Mouhleb[1], Sayyed Auwn Muhammad[1], Desiree Nilsson[1], Håvard Mokleiv Nygård[1,3], Gudlaug Olafsdottir[1], Kristina Petrova[1], David Randahl[1], Espen Geelmuyden Rød[1], Nina von Uexkull[1], and Jonas Vestby[3]

[1]Department of Peace and Conflict Research, Uppsala University
[2]University of Pittsburgh
[3]Peace Research Institute Oslo

June 25, 2018

**Abstract**

The article presents ViEWS – a political Violence Early Warning System. The system seeks to be maximally transparent, be publicly available, and have uniform coverage. The article argues that given reasonable performance, ViEWS can be a useful complement to non-public early warning systems. It presents and evaluates the ViEWS forecasts for conflict in Africa at the grid-cell month level for several forms of conflict. The forecasts indicate a very strong persistence of conflict in regions in Africa with a recent history of political violence, but also alert to relatively new conflicts such as in Southern Cameroon. The article also summarizes the methodology employed to generate these forecasts. The system breaks the forecasting problem into several constituent parts, analyzing the risk of three separate types of conflict at two levels of analysis, and uses model averaging to combine numerous modeling approaches for each.

# 1 Introducing ViEWS

Armed conflicts and other forms of large-scale political violence continue to kill and maim thousands of people every month across the globe. For every person killed, hundreds are forced to relocate within countries and across borders. Armed conflicts have disastrous economic consequences where they occur (Gates et al., 2012). They invariably undermine the functioning of political systems and of the public services policy-makers seek to provide, and prevent a range of countries in the developing world from escaping poverty (Collier et al., 2003). Conflicts hinder international actors in providing humanitarian assistance where needed.

The challenges of preventing, mitigating, and adapting to large-scale political violence are particularly daunting when it escalates in locations and at times where it is not expected. Policy-makers and first responders would benefit greatly from a system that systematically monitors all locations at risk of conflict and assesses the probability of conflict escalation, continuation, and diffusion. This article presents ViEWS– a political Violence Early-Warning System which seeks to address this need.[1]

We present the framework of ViEWS as well as its first public forecasts. ViEWS assesses the probability of political violence events in the future – events that may be continuation of ongoing conflict, entirely new conflicts, or diffusion of ongoing conflicts to new places.

In order to be useful, ViEWS aims for maximal transparency, uniform coverage, and public availability. Transparency requires that the risk assessments can be traced back to a fully specified argument and accessible information. ViEWS is exclusively based on publicly available data. Its results, input data, and procedures are available to researchers and the international community. Such transparency is essential to allow readers and potential users to evaluate what lies behind a given set of forecasts. Uniform coverage of the regions at risk helps alerting observers to locations that receive less attention than the conflict potential requires. In principle, ViEWS seeks to be able to issue a warning with equal probability for any location within this area, independent of its geo-strategic importance, past conflict history, or current humanitarian situation. Public availability of the results is useful for domestic actors and small international NGOs, but is also essential to ensure complete transparency regarding decisions made on the basis of specific warnings.

ViEWS also seeks methodological innovation. Some of that innovation is documented in the current system. For instance, we present methods to account for the dynamic escalatory processes typical of conflict. Events in Tunisia and in Syria initially looked quite similar, but only Syria escalated to a full-blown civil war over the year that followed the initial demonstrations. Forecasting all events in advance would be very difficult, but a model should be able to provide a probability distribution over various escalation trajectories.

ViEWS provides forecasts for three forms of political violence: armed conflict involving states and rebel groups, armed conflict between non-state actors, and violence against civilians (Melander, Pettersson, and Themnér, 2016). Forecasts apply to specific sub-national geographical units, and countries. To be useful as an early-warning system, it will be running continuously, generating forecasts that are updated as soon as new data become available. This is made possible by the close links between

[1] A detailed description of the project and the project participants is found at `http://www.pcr.uu.se/research/views/`.

ViEWS and the Uppsala Conflict Data Program, who have been publishing on armed conflict and other forms of political violence for more than two decades (Gleditsch et al., 2002; Sundberg and Melander, 2013).

In its present form, ViEWS primarily concentrates on theoretical and methodological development, and on exploring how accurately it is possible to predict political violence. For now, ViEWS is limited in scope to Africa and updated monthly, based on frequent data releases provided by the UCDP (Hegre et al., 2018).[2]

In this article, we present the ViEWS forecasts as of 1 June 2018, summarize the evaluation of their predictive performance, and outline the methodological framework behind the project and the data used. More details on the methods used are provided in a set of background papers available at the ViEWS web page. These are referred to where appropriate below.

The project is currently collecting and adapting data to expand the system, such as data on elections past and future, military coups, economic indicators, and ethnic composition, and is developing an expert survey. As new data are collected over the course of project, ViEWS forecasts will be evaluated and our models replaced, reformulated, and retrained accordingly. This allows us to gradually improve the forecasting performance of the system by comparing new combinations of risk factors, statistical model formulations, and projection algorithms with the performance of the previous versions.

To document this development, we will follow up with a series of annual short articles to present the future development of ViEWS. In these, we will present new methodological, theoretical, and data-related innovations in ViEWS relative to the previous articles, and show a brief extract of the most recent forecasts for the year following the previous publication. We will also present a formalized comparison between the forecasts published the year before and the events that actually happened the year after. As such, the series format would ensure that the forecasts are evaluated not only out-of-sample, but as true forecasts of events that were unavailable to the researchers at the time their models were trained.

This article first presents the forecasts from the current specification of the system as well as an evaluation of its predictive performance. We then present the data we use and summarize the methods applied.

## 2 Current forecasts

The ViEWS forecasts are assessments of the probability of three types of conflict events. At the end of each month, we publish forecasts for the coming month and for the subsequent 36 months at the project website (`http://www.pcr.uu.se/research/views`).[3]

The conflict data come from the newly released ViEWS Outcomes dataset (Hegre et al., 2018; Croicu and Hegre, 2018). This, in turn, is based on the UCDP-GED event dataset (Sundberg and Melander, 2013), with additions and aggregations from the new UCDP-Candidate dataset (Hegre et al., 2018).

[2]Given sufficient funding to cover the required data-collection needs, these ambitions will be scaled up to a wider geographic scale.

[3]The procedure in practice also involves creating a 'now-cast' for one month to accommodate for time it takes the UCDP to finalize its monthly coding and for ViEWS to prepare other input variables. What we report here, involves a now-cast for May 2018.

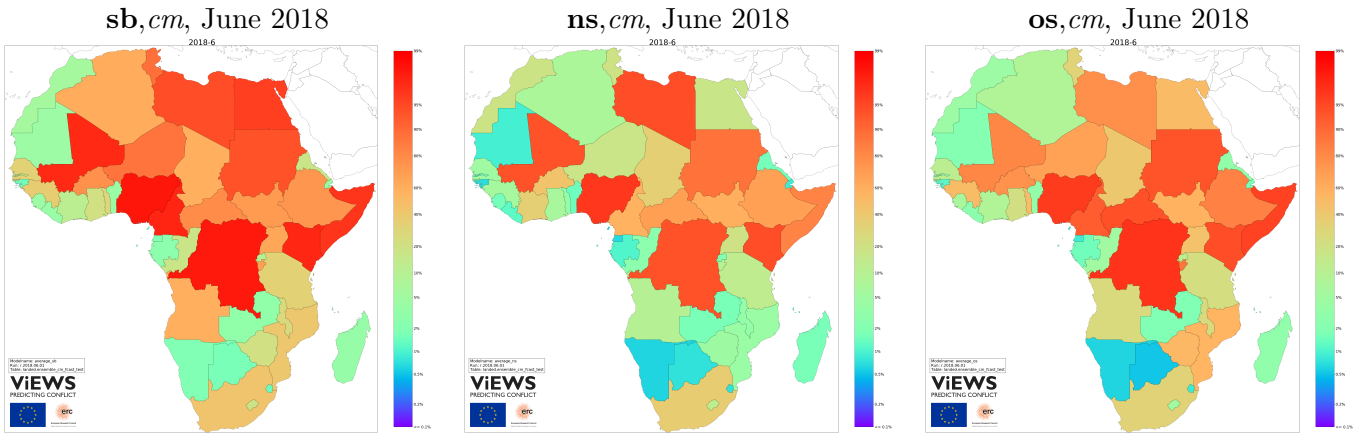



Figure 1. Country-level ViEWS forecasts, state-based conflict (left), non-state conflict (centre), one-sided (right). Predicted probabilities of at least one UCDP-GED event for June 2018, based on the ViEWS system as of 1 June 2018.

The UCDP-GED and UCDP-Candidate cover three conflict types: **sb** – state-based armed conflict which involves fighting between a government and organized rebel actors as well as inter-state war (Gleditsch et al., 2002); **os** – one-sided violence which includes the deliberate and direct targeting of unarmed civilians (Eck and Hultman, 2007); and **ns** – non-state conflict which involves inter-communal violence, fighting between rebel organizations, and fighting between political parties (Sundberg, Eck, and Kreutz, 2012).

In this article, we present results at two levels of analysis – for countries as well as sub-national geographical locations.[4] All ViEWS models are designed at a monthly temporal scale. Throughout, the dependent variable is whether there was at least one UCDP-GED event with at least one fatality within the unit of analysis.[5] Complete forecasts are available at `http://www.pcr.uu.se/research/views/data/downloads/`.

Our forecasts for the country-month (called *cm* hereafter) are presented in Figure 1.[6] This level is particularly useful to provide predictions for entirely new conflicts where no known actors exist, and to model tensions and processes at the government level. ViEWS maintains global coverage for the *cm* level although we collect more data for Africa than for the rest of the world.

The plots in Figure 1 show the ViEWS **sb**,*cm* forecasts for the immediate future – what will happen in June 2018? We show the probability of at least one event in June 2018 based on data up to and including April in the same year. Countries with red color have forecast probabilities close to 1, whereas blue countries have forecasts at less than 0.01. When the forecasts indicate that no events is as likely as at least one event, countries are drawn with an orange color.

Our models yield results in line with mainstream studies of conflict at the country level.[7] For

[4] See Section 4.5 for details. ViEWS will add an actor level at a later stage.

[5] For more details, see Section 5.1. The aggregation procedures transforming the UCDP-GED events into the dependent variables used in ViEWS are described in Hegre et al. (2018).

[6] We use the cShapes dataset (Weidmann, Kuse, and Gleditsch, 2010) and the Gleditsch-Ward country code (Gleditsch and Ward, 1999) to identify countries.

[7] The interpretations outlined below are based on the parameter estimates and imporance scores of the models in Table 3. Detailed results for all models are available in the document 'ViEWS monthly estimate tables, June 2018' at

instance, we forecast a high probability of **sb** conflict in countries with large populations (Fearon and Laitin, 2003; Raleigh and Hegre, 2009), in non-democracies and countries with recent regime change (Hegre et al., 2001; Cederman, Hug, and Krebs, 2010), with low or negative growth rates (Collier and Hoeffler, 2004), and with low education levels (Thyne, 2006) or other indicators of low socio-economic development.

We forecast a high probability of conflict in countries that have a recent history of conflict or with recent protest events. In Mali, Nigeria, and DR Congo conflict is almost certain. We also forecast a high probability of state-based conflict (**sb**) in Cameroon, driven by recent events. Tensions and violence have escalated since separatists symbolically declared independence of 'Ambazonia' in October 2017. The separatist violence, involving several groups, continued throughout the spring in 2018 (International Crisis Group, 2018). There have also been clashes between government forces and IS (often referred to as Boko Haram) in the northern part of the country.[8] In Kenya, clashes between the government and Al-Shabaab have been reported in every month up to April 2018, and these are likely to continue.[9]

The forecast maps for non-state conflict (**ns**) and one-sided violence (**os**) follow partly the same patterns as **sb**. Our forecasts for **ns** depends on the same factors as for **sb**, but seems less depressed by democratic institutions and socio-economic development than **sb** events. More importantly, the patterns of past events differ across conflict types (see Figure 2). Cameroon and Egypt, for instance, have not had much **ns** conflict, wheras Libya and Sudan have seen a lot. We forecast a high probability also of **ns** in Kenya due to recent confrontations between cattle rustlers and herders. Furthermore, actors with unclear affiliation carried out attacks against civilians (*Kenyan minister declares curfew in restive Mt Elgon region*).

The forecasts for **os** respond to about about the same factors, but are less clearly related to protests and regime change. They also occur more frequently in newly independent countries. Kenya, again, will see continued one-sided violence, most likely perpetrated by the Al-Shabaab.[10]

To complement the country-level forecasts, Figure 2 presents forecasts at fine-grained sub-national geographical locations, focusing primarily on *where* events happen. We refer to this level as *pgm*. Here, ViEWS relies on the PRIO-GRID (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. Around equator, a side of such a cell is around 55 km. This resolution is close to the precision level of the data we have for the outcomes.[11]

The upper row of maps in Figure 2 shows ViEWS forecasts for the *pgm* level for each of the three outcomes. The color mapping is the same as for the *cm* forecasts. The densest risk clusters for state-based conflict are in northern Nigeria, the North Kivu province in DRC, Somalia, and Darfur. All of these regions have been ravaged with violence for years. These maps reflect that countries' recent conflict history is the strongest predictor of future violence. To demonstrate this, we plot maps

`https://www.pcr.uu.se/research/views/publications/appendices/`.

[8] See `http://ucdp.uu.se/#/statebased/12422`.

[9] See `http://ucdp.uu.se/#/statebased/10589`.

[10] See `http://ucdp.uu.se/#/onesided/1071`.

[11] Investigating the spatial error of the UCDP-GED in Afghanistan, Weidmann (2014, p.1143) found that most events where "located within 50 km of where they actually occured". Given this, a finer resolution would be unlikely to yield more precise forecasts.

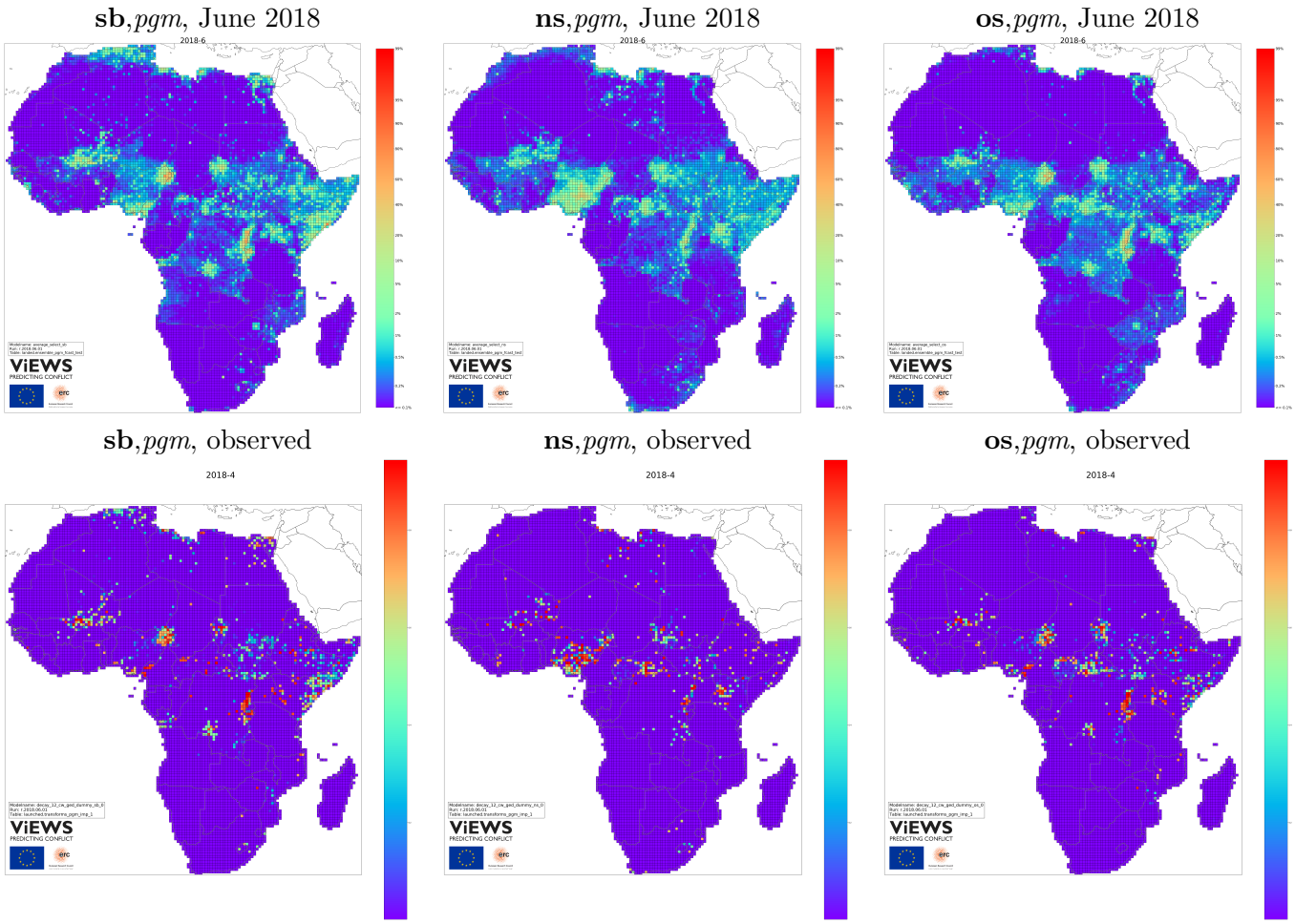



Figure 2. Upper row: Forecasts June 2018–June 2021, *pgm* level, ensemble model. Predicted probabilities of at least one UCDP-GED event for June 2018, based on the ViEWS system as of 1 June 2018. Lower row: Decay function of time since most recent event, halflife 12 months, as recorded in Hegre et al. (2018).

depicting the recent history of violence in each PRIO-GRID cell in the low row of the figure. Red cells had conflict in April 2018, and purple ones have not seen conflict in many years.

Geographical features are also important – the low population concentration in Sahara translates into a low risk of conflict, and conflicts are more likely in border regions than close to countries' capitals. The maps also show how country-level risk assessments influence the geographical forecasts. Zambia, Botswana, and Tanzania, for instance, have markedly lower probability of future conflict than their neighboring countries.

The forecasts for non-state conflict and one-sided violence depend on the same factors although with somewhat different implications. For **ns**, we forecast main clusters in central Nigeria, Central African Republic, North Kivu, Darfur and the Kenyan Rift Valley. For **os**, northern Nigeria, Darfur, North Kivu, and Burundi are the primary hotspots.

All forecasts shown so far have been for June 2018, the second month after the most recent data available. Figure 3 indicates how the forecasts change over time. The color mapping is roughly the

same as above, but here correspond to the forecasted proportion of PRIO-GRID cells in **sb** conflict for each country. In Burundi, for instance, we expect about 18% of the cells to have conflict in each month. In Ethiopia, the forecast is 1.2%.

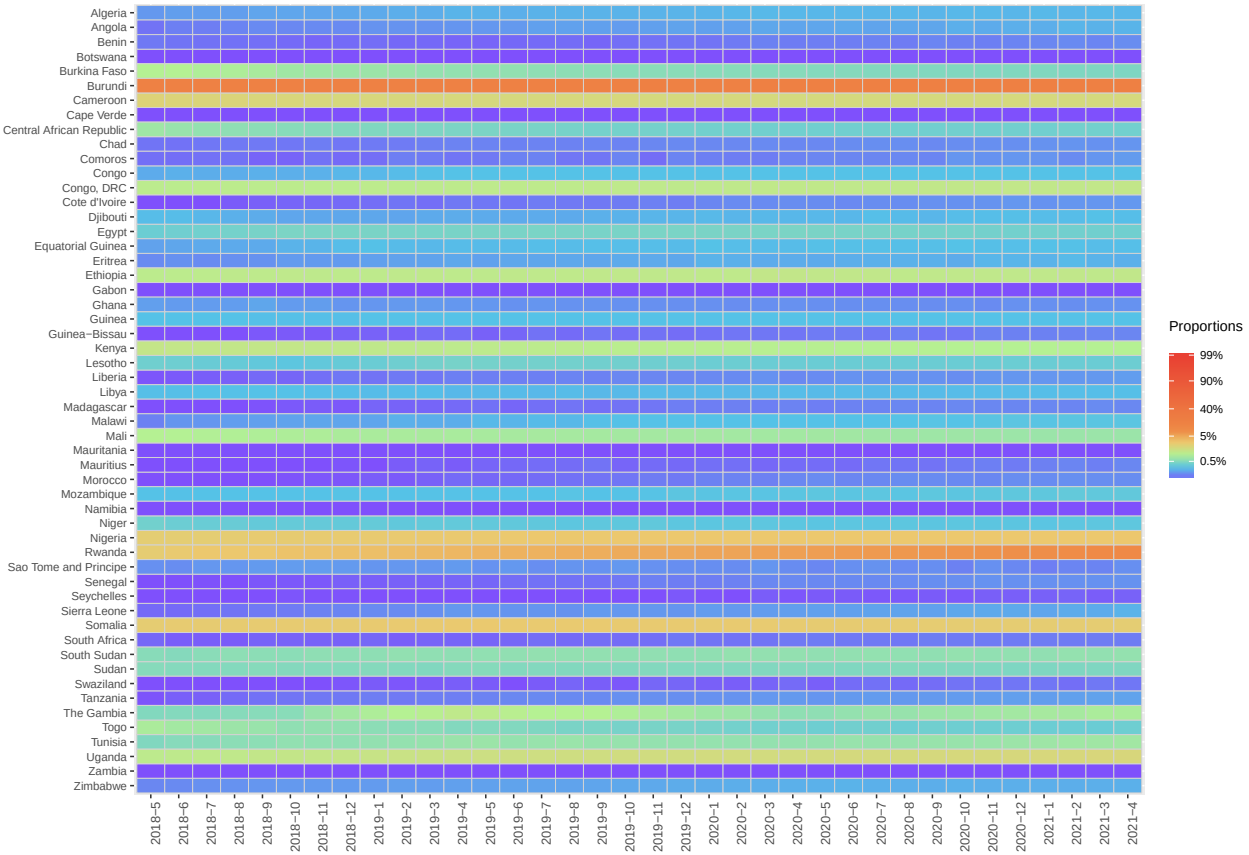



Figure 3. Forecasted proportion of *pgm* cells with conflict events, by country, June 2018–June 2021

Since our projections for the exogenous variables are relatively stable, our forecasting models yield about constant average risks of conflict over time – most countries have the same color throughout the period. There are a couple of exceptions. We forecast a decreasing danger of conflict in Togo, for instance, as the impact of isolated violent events in in 2017 recedes (*'Three killed' in Togo opposition clashes*). On the other hand, there is an increasing danger of conflict in DRC and Burundi spilling over into Rwanda. The forecasted proportion of *pgms* with conflict in Rwanda increases from 0.025 in June 2018 to 0.085 in mid-2021.

Figure 4 takes a closer look at Western Central Africa over time. There is a sizeable orange cluster, with probabilities around 30–50% in each *pgm* in the Eastern DRC and Burundi for June 2018. This cluster is slowly expanding over the forecasting period, and spills into Rwanda.

Our models reflect that forecasted violence in these clusters change little over time. The ViEWS models contain information about conflict events many years into the past, and the underlying estimates indicate that African conflicts are very persistent. Most of the variation in Figure 3 is between countries, not across time. Some tendency of a 'regression to the mean' can be seen in Figures 3 and 4, however. Since we can be less certain the further into the future we forecast, we expect differences across countries

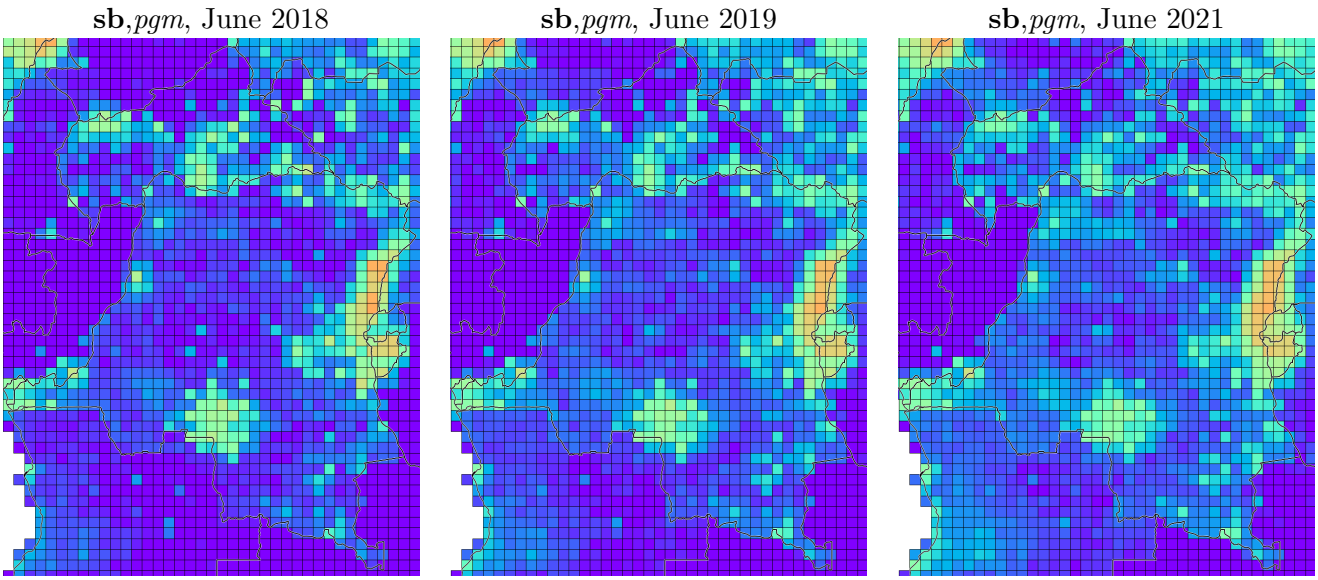



Figure 4. Forecasts June 2018 – June 2021, state-based conflict, DRC, CAR, Rwanda, Burundi

and locations to become less marked.

We return to other factors that determine our forecasts when reviewing the methods used below (section 4).

## 3 How well do we predict? Evaluation of models

To evaluate the out-of-sample predictive performance of the ViEWS forecasting system, we partition our observed measures of conflict into three disjoint periods, one for estimating statistical models (training period), one for calibrating predicted probabilities from the models (calibration period) and one for forecasting (testing/forecasting period). When we evaluate the system, we have information about conflict in the testing/forecasting period that we compare predictions to. This is not the case when producing forecasts into the future (beyond April 2018). Here, the testing/forecasting period must await observations before we can evaluate how well we predict. Table 1 shows how we partitioned the data. We used the 'forecast' periodization for the forecasts presented above, and use the 'evaluation' periodization for the evaluations discussed in this section.

To evaluate the performance of the models, we first estimated models using data in the training period (January 1990 – December 2010) and obtained predicted probabilities for the calibration period (January 2011 – December 2013). Second, we estimated the same models using data from the training and calibration period (January 1990 – December 2013) and obtained predicted probabilities for testing/forecasting period (January 2014 – December 2016). Finally, the predicted probabilities for the testing/forecasting period were calibrated using information about the actual occurrence and predicted probabilities in the calibration period.

| | Periodization | |
|---|---|---|
| | Evaluation | Forecast |
| Training period | January 1990 – December 2010 | January 1990 – April 2015 |
| Calibration period | January 2011 – December 2013 | May 2015 – April 2018 |
| Testing/forecasting period | January 2014 – December 2016 | [May]June 2018 – June 2021 |

Table 1. Partitioning of data for forecasting, evaluation, and estimating model weights

**Overall performance**

Table 2 shows summary statistics of the predictive performance of the ViEWS forecasts or the test period (2014–16). As we elaborate on in Section 4, the ViEWS forecasts are based on combinations of multiple constituent models. These are labelled as *ensembles* in Table 2. However, since the performance of the ensembles cannot be interpreted in and by themselves, we define a simple set of baseline models for comparison. These are varying-intercept models without any features beyond which country (for *cm* models) or PRIO-GRID cell (for *pgm*) they refer to. The baseline models, then, simply predicts the probability of conflict in each group in the test period to equal the mean probability of conflict in the same group in the training period.

Performance is assessed based on the *Area Under the Receiver Operator Curve (AUROC)*, *Area Under the Precision-Recall curve (AUPR)*, and *Brier score*.[12] AUROC summarizes performance as a relative measure of the true positive rate and the false positive rate of predictions. The goal is to maximize true positives relative to false positives. In other words, the measure rewards models for increasing detection of actual conflict (true positives) relative to "false alarms" (false positives). AUPR is a relative measure of how precisely a model predicts true positives and the true positive rate.

Precision is measured as the proportion of predicted conflict onsets that are correct. This means that the AUPR measure rewards models for getting conflicts correct once a model predicts them. Since only a small percentage of observations experience conflict, it is more difficult to get predictions of conflicts correct than it is to get predictions of the absence of conflict correct. AUPR is therefore a more demanding measure than AUROC.

The Brier score measures the accuracy of probabilistic predictions. It favors sharp, accurate probabilistic predictions (near 0 or 1), which is different to the relative ordering of the forecasts that is needed for the computation of the AUPR and AUROC.

The table shows that the ensembles increase the predictive performance considerably in comparison to these baseline models. For **sb** at the *cm* level, for example, there is an increase in AUROC from 0.585 to 0.930, a decrease in Brier score from 0.175 to 0.092, and an increase in AUPR from 0.345 to 0.797. There are similar increases for the **os** and **ns** ensembles as well. At the *pgm* level, there are also large differences between the baseline models and the ensembles. For **sb**, AUROC increases from 0.536 to 0.896 and AUPR from 0.002 to 0.049. There is a smaller change in Brier score from 0.00632 to 0.00589. The discrepancy between improvements in AUROC, AUPR and Brier score highlights that the ensembles can successfully sort observations into higher and lower probability and thus improve AUROC and AUPR. At the same time, the modest improvement in the Brier score shows that the value of the ensemble predicted probabilities for true positives is still quite far from 1. In other words,

[12]These metrics are discussed in more detail in Section 4.3.

| Model | AUROC | Brier Score | AUPR |
|---|---|---|---|
| *cm*, **sb**, baseline | 0.5947 | 0.1754 | 0.3451 |
| *cm*, **sb**, ensemble | 0.9307 | 0.0925 | 0.7974 |
| *cm*, **os**, baseline | 0.6230 | 0.1359 | 0.3191 |
| *cm*, **os**, ensemble | 0.9234 | 0.0758 | 0.7852 |
| *cm*, **ns**, baseline | 0.6575 | 0.0970 | 0.3265 |
| *cm*, **ns**, ensemble | 0.9233 | 0.0616 | 0.6797 |
| *pgm*, **sb**, baseline | 0.5337 | 0.00632 | 0.0051 |
| *pgm*, **sb**, ensemble | 0.9520 | 0.00589 | 0.2466 |
| *pgm*, **os**, baseline | 0.5655 | 0.00523 | 0.0047 |
| *pgm*, **os**, ensemble | 0.9548 | 0.00503 | 0.1990 |
| *pgm*, **ns**, baseline | 0.5369 | 0.00394 | 0.0025 |
| *pgm*, **ns**, ensemble | 0.8965 | 0.00393 | 0.0493 |

Table 2. Summary statistics for predictive performance averaged across all months in test window (see Table 1).

the ensemble is not making sharp predictions that clearly separate the classes on the probability scale. The pattern for **sb** is consistent with **ns** and **os**.

These evaluations show that the current system does very well relative to the baseline models. These models are intuitive, but obviously underspecified. They serve as a point of reference in the absence of any established baselines. As ViEWS moves forward, the metrics reported here for the ensemble models will constitute the baselines for future comparisons. These numbers also constitute a new frame of reference for others that need to gauge the performance of their models applied to the forecasting problem defined here.

### Performance over time

When discussing Figure 3, we assumed that the ViEWS forecasts for 2021 are more uncertain than those for 2018. Figure 5 investigates formally how the predictive performance of the ViEWS forecasts change depending on how far into the future we move. These evaluations enable us to gauge the feasibility of forecasting up to 36 months into the future, and also how reliable forecasts are in the long term relative to the short term. The figure shows predictive performance for each outcome-level combination. The top row of plots shows performance for the conflict outcomes at the *cm* level, the bottom row the same for *pgm*. In the plots in the left column, the $y$ axis shows AUROC and the right column AUPR. Since the predictive performance differs between the models, the $y$ axis varies from plot to plot. The $x$ axis shows the month of the forecast, moving from 1 to 36 months into the future. The colors indicate the conflict type, blue for **sb**, green for **os**, and red for **ns**. The lines are smoothed using a loess function.[13]

At the *cm* level, both the AUROC and AUPR declines over time for **sb**, **os**, and **ns** in figure 5. This is what we would expect: as we move further into the future, it becomes increasingly difficult to obtain accurate predictions. The deterioration is substantial, especially for AUPR. In the top left, AUROC decreases from close to 1 in the beginning of the forecasting period to between 0.85 and 0.9 at the end. In the top right, we see a large decrease in AUPR from around 1 to between 0.65 and 0.75.

More strikingly, AUROC and AUPR at the *pgm* level does not change much over time for **sb**. As

[13]If the figures were plotted without smoothing, we would see a zig-zag pattern reflecting that we use short time windows in the evaluation (month) and that the incidence of conflict in each month fluctuates more than our forecasted risk.

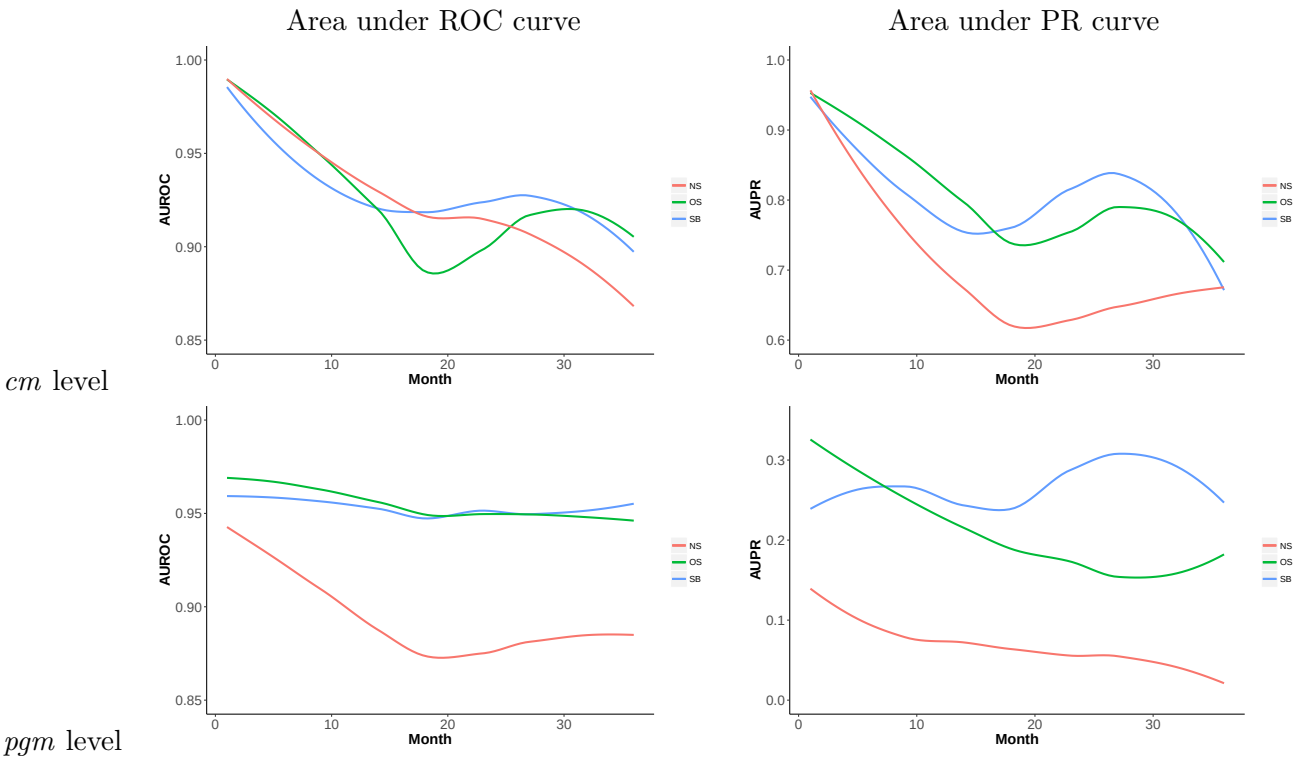



Figure 5. Performance over time, *cm* (top) and *pgm* (bottom). AUROC (left) and AUPR (right), by month in forecasting window. The lines are smoothed with a loess function. Note that the y-axis differs in each plot.

we saw in Figure 2, the geographical patterns of sb conflict have been quite stable over the past years, and our forecasts appear to be doing quite well in capturing that stability. For **os**, there is no change for AUROC but a clear downward slope for AUPR – from above 0.3 in the beginning of the period, to around 0.2 at the end. For **ns**, there is a drop in both AUPR and AUROC over time. The drop is also fairly large, especially for AUPR, which changes from around 0.15 at the highest to almost 0 at the lowest.

In Figure A-2 in Appendix D we show that the mean predicted probabilities across Africa are consistent with the actual mean probabilities, although our model is not able to forecast the significant variation over time in conflict levels over the 2014–2016 period.

These results consolidate our expectation that the ViEWS forecasts perform better in the near than the distant future, but that forecasting over a three-year horizon is fully within reach. By looking at changes over time, we can quantify the expected loss in predictive performance of the forecasts in specific time intervals, which is important when communicating the forecasts. Again, a major function of this evaluation is to establish a baseline to compare future models with.

# 4 Methodology

ViEWS applies a 'divide and conquer' strategy to the forecasting problem. In addition to analyzing separately the three outcomes at each of the three levels of analysis, we combine multiple modeling

strategies in model ensembles. We outline here the major features of the system in its current form, and sketch a few areas where we are working to improve it.

As we proceed with developing the system, we will follow the guidelines of Colaresi and Mahmood (2017), who suggest an iterative loop, whereby model representations are built from domain knowledge, their parameters computed, their performance critiqued, and then the successes and particularly the failures of the previous models inform a new generation of model representations. Crucial to this machine learning-inspired workflow are visual tools, such as model criticism and biseparation plots for researchers to inspect patterns that are captured by some models and ensembles but missed by others. We too, will expand on these tools, looking at mistakes in the geographical context.

| Model name | Dynamic mode | Sampling strategy | Stat. model | Features | Temporal coverage | Weight in ensemble | AUROC (sb) | Brier (sb) | AUPR (sb) |
|---|---|---|---|---|---|---|---|---|---|
| **Country-level** (*cm*) | | | | | | | | | |
| ds_cm_canon_base* | DS | All obs., global | Logit | *cm* core | 1990–2016 | 0.1667 | 0.94419 | 0.08766 | 0.81801 |
| ds_cm_acled_base* | DS | All obs., Africa | Logit | *cm* core + ACLED | 1997–2016 | 0.1667 | 0.93900 | 0.09025 | 0.80143 |
| osa_cm_acled_base_logit* | OSA | All obs., Africa | Logit | *cm* core + ACLED | 1997–2016 | 0.1667 | 0.91981 | 0.09851 | 0.77482 |
| osa_cm_acled_base_rf* | OSA | All obs., Africa | RF | *cm* core + ACLED | 1997–2016 | 0.1667 | 0.90150 | 0.09934 | 0.76869 |
| osa_cm_canon_base_logit* | OSA | All obs., global | Logit | *cm* core | 1990–2016 | 0.1667 | 0.91855 | 0.09959 | 0.78842 |
| osa_cm_canon_base_rf* | OSA | All obs., global | RF | *cm* core | 1990–2016 | 0.1667 | 0.89669 | 0.09771 | 0.76452 |
| **PRIOGRID-level** (*pgm*) | | | | | | | | | |
| ds_pgm_canon_nocm* | DS | All obs. | Logit | *pgm* core | 1990–2016 | 0.0833 | 0.91103 | 0.00611 | 0.20776 |
| ds_pgm_canon_wcm* | DS | All obs. | Logit | *pgm* core + *cm* vars. | 1990–2016 | 0.0833 | 0.91756 | 0.00600 | 0.21749 |
| ds_pgm_acled_wcm* | DS | All obs. | Logit | *pgm* core + ACLED + *cm* vars. | 1997–2016 | 0.0833 | 0.91856 | 0.00607 | 0.19714 |
| osa_pgm_acled_nocm_rf* | OSA | Downs. | RF | *pgm* core + ACLED | 1997–2016 | 0.0833 | 0.94875 | 0.00605 | 0.20373 |
| osa_pgm_acled_wcm_rf* | OSA | Downs. | RF | *pgm* core + ACLED + *cm* vars. | 1997–2016 | 0.0833 | 0.94845 | 0.00601 | 0.20507 |
| osa_pgm_canon_nocm_logit* | OSA | All obs. | Logit | *pgm* core | 1990–2016 | 0.0833 | 0.92290 | 0.00596 | 0.21989 |
| osa_pgm_canon_nocm_rf* | OSA | Downs. | RF | *pgm* core | 1990–2016 | 0.0833 | 0.94506 | 0.00606 | 0.19487 |
| osa_pgm_canon_wcm_rf* | OSA | Downs. | RF | *pgm* core + *cm* vars. | 1990–2016 | 0.0833 | 0.94858 | 0.00602 | 0.19662 |
| cl_ds_pgm_canon_nocm* | DS | All obs. | Logit | *pgm* core + *cm* core | 1990–2016 | 0.0833 | 0.91469 | 0.00598 | 0.20110 |
| cl_ds_pgm_acled_nocm* | DS | All obs. | Logit | *pgm* core + ACLED | 1997–2016 | 0.0833 | 0.91765 | 0.00603 | 0.19961 |
| cl_osa_pgm_canon_nocm_logit* | OSA | All obs. | Logit | *pgm* core + *cm* core | 1990–2016 | 0.0833 | 0.91082 | 0.00593 | 0.21472 |
| cl_osa_pgm_acled_nocm_rf* | DS | All obs. | Logit | *pgm* core + ACLED + *cm* core | 1997–2016 | 0.0833 | 0.94201 | 0.00608 | 0.20097 |

Table 3. Models in the ViEWS ensemble. See Appendix C for the features in the core and ACLED models. Full estimation results for all models are provided in the 'ViEWS monthly estimate tables, June 2018' appendix (`https://www.pcr.uu.se/research/views/publications/appendices/`).

### 4.1 The individual statistical models in ViEWS

The conflict research community has laid the ground for an early-warning system through careful isolation of theoretically manageable sub-components of complex phenomena, and concomitant systematic, disaggregated data-collection efforts. ViEWS integrates these research efforts into a theoretically and methodologically consistent forecasting system by means of an ensemble of statistical models. The current collection of models is listed in Table 3. The first column in this table shows the name the model currently has in our database. We will discuss the contents of the remaining columns as we proceed.

We selected these models with the aim of including various combinations of the individual modeling strategies we describe below regarding dynamic mode, sampling strategy, statistical model, and feature selection. It would not be feasible to include all possible combinations, so we excluded a number of models that yielded very similar predictions to those included. The first aspect of variation is the features included in the models. We currently have defined two sets of features at each level of analysis, 'core' and 'acled'. The 'features' column above shows which model are based on which feature set. The same set of features is used across all models and all outcomes, but estimated separately for each. The '*cm* core' set is based on standard country-level models of armed conflict such as Hegre et al. (2001), Fearon and Laitin (2003), and Cederman, Wimmer, and Min (2010), and include lagged dependent variables, demographic and economic features, political institutions and ethnic exclusion.

At the *pgm* level, there are three groups of variables. We include a set of variables that represent the conflict history of each cell. We include information on the history of all three conflict outcomes in all models. There is also a set of static geographical features drawn from PRIO-GRID, such as the size of the population in a grid cell, the type of terrain at the location, or distance to the nearest oil field. Finally, in some models we also include some *cm* core variabels copied to each PRIO-GRID cell within each country. This set includes the presence of *cm*-level conflict.

The 'ACLED' models (at both the *cm* and *pgm* levels) contain all the core features in addition to the most recent data on protests from ACLED (Raleigh et al., 2010).[14]

Most variables in all of these models are treated as exogenous, except for conflict history variables that are simulated historically (see Section 4.8). The variables in the model are listed in Appendix C. A Detailed descriptions of each variable and complete estimation results for each model are found at `http://pcr.uu.se/research/views/publications/appendices/`.

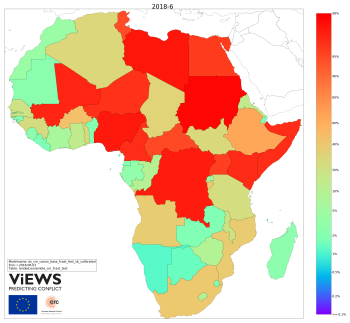
The second element of model variation is estimation strategy. Some of the models were estimated using logit, others using random forests. All random forest models are estimated on an asymmetrically downsampled dataset (see Appendix D).

The third aspect is how we handle the dynamic aspect of the forecasts. ViEWS employs two alternative strategies to compute forecasts several months into the future from data with monthly resolution. We call these Dynamic simulation (DS in Table 3) and One step ahead (osa). We discuss this in more detail in Section 4.6.

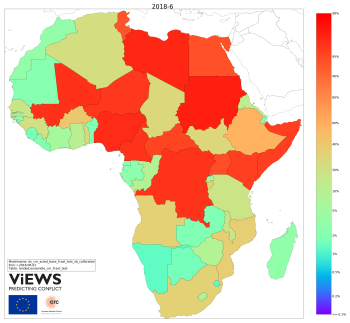
The final element we vary is how we incorporate information from the cm level in the pgm models. We have models that incorporate no information from the *cm* level (nocm), with *cm* predictors (wcm),

[14]Since ACLED has limited coverage outside Africa, we estimate these models for Africa only at the *cm* level.

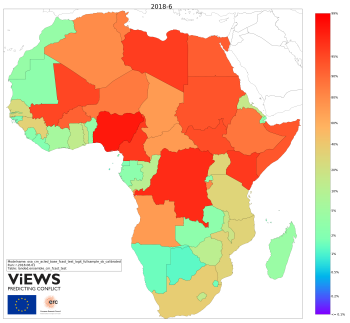
and the product of the *cm* and *pgm* predicted probabilities (cl). We return to this in Section 4.5.



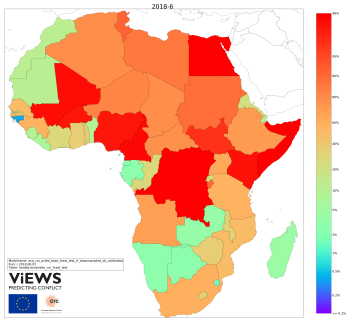
ds_cm_canon_base
Model 1: Core features, logit, dynamic simulation



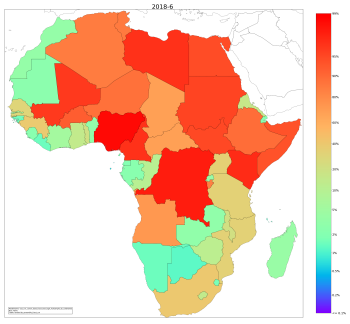
ds_cm_acled_base
Model 2: Core + protests, logit, dynamic simulation



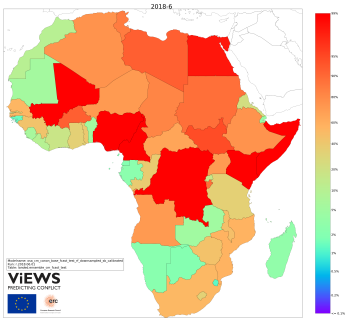
osa_cm_acled_base_logit
Model 3: Core + protests, logit, one-step-ahead



osa_cm_acled_base_rf
Model 4: Core + protests, random forest, one-step-ahead



osa_cm_canon_base_logit
Model 5: Core features, logit , one-step-ahead



osa_cm_canon_base_rf
Model 6: Core features, random forest, one-step-ahead

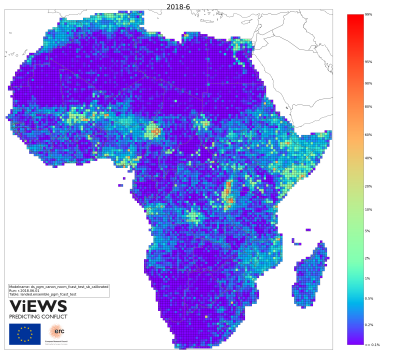
Figure 6. Country-level **sb** forecasts for June 2018, constituent models

Figure 6 shows the predicted probabilities for **sb** conflict June 2018 for each of the constituent *cm* models. At the cm level, the various modeling strategies we choose have only marginal impact on the results. Table 3 also report the predictive performance metrics for each model for the **sb** outcome. Complete evaluation metrics for the models are reported in Appendix A.1.

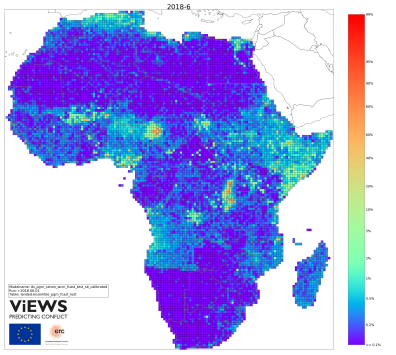
Figure 7 shows the predicted probabilities for **sb** conflict June 2018 for each of the constituent models at the *pgm* level.

## 4.2 Ensembles

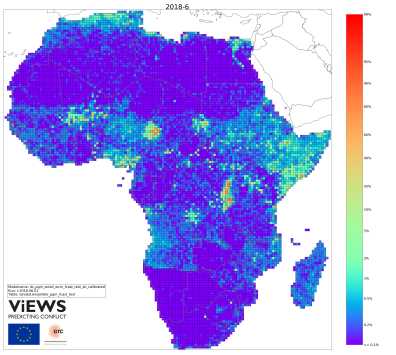
The ViEWS forecasts are combinations of the constituent models in Table 3, commonly referred to as *ensembles*. The combination of multiple constituent models in ensembles have been shown to produce more robust forecasts that often have higher predictive performance than even the most accurate of the component models that go into them (Armstrong, 2001). We have explored two strategies for assigning weights to the constituent models in the ensembles. First, we simply used the unweighted average prediction from all models. Second, we estimated weights using Ensemble Bayesian Model Averaging (EBMA, for technical details see Montgomery, Hollenbach, and Ward, 2012; Beger, Dorff, and Ward, 2014; Raftery and Lewis, 1992; Raftery et al., 2005). EBMA weights the constituent models in the ensemble based on their performance in the calibration period (see Table 1). In short, EBMA uses an expectation maximization algorithm to assign these weights based on overall predictive performance



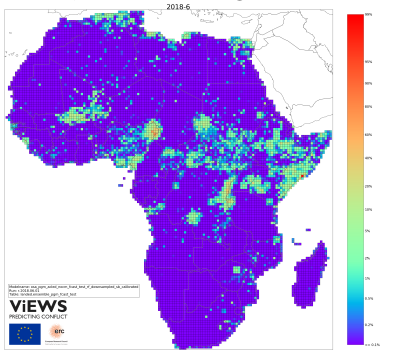
ds_pgm_canon_nocm
Model 1: Core features, logit, dynamic simulation, no *cm* modeling



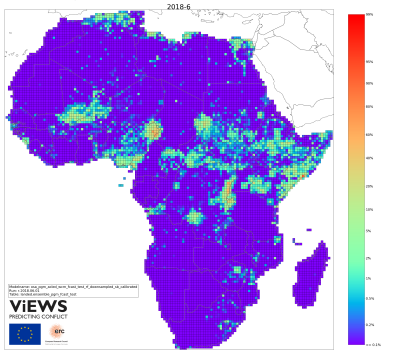
ds_pgm_canon_wcm
Model 2: Core features, logit, dynamic simulation, *cm* features



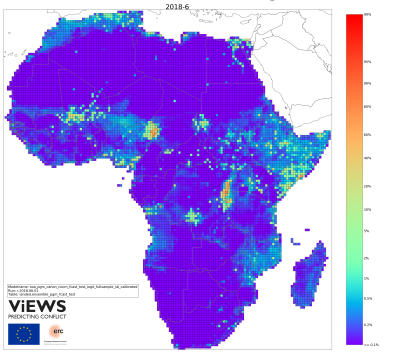
ds_pgm_acled_wcm
Model 3: Core+protests, logit, dynamic simulation, no *cm* modeling



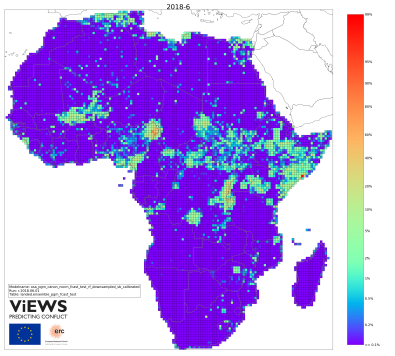
osa_pgm_acled_nocm_
Model 4: Core + protests, random forests, one-step-ahead, no *cm* modeling



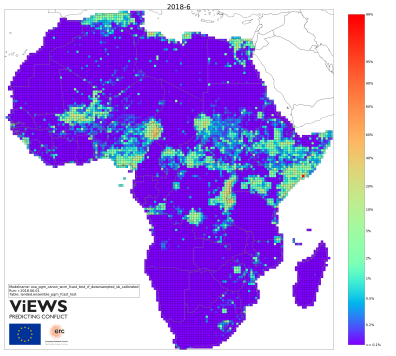
osa_pgm_acled_wcm_
Model 5: Core + protests, random forests, one-step-ahead, *cm* features



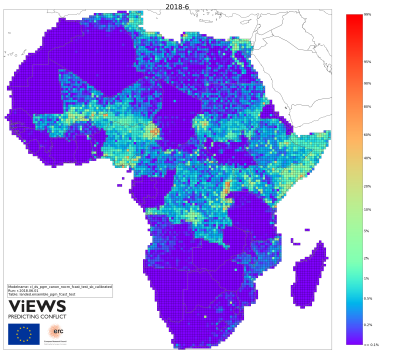
osa_pgm_canon_nocm
Model 6: Core features, logit, one-step-ahead, no *cm* modeling



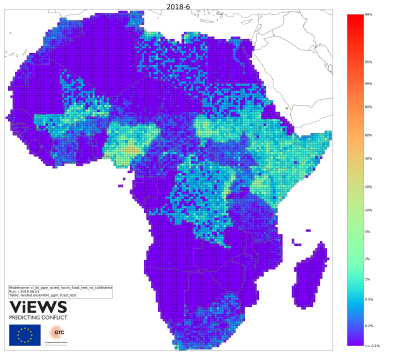
osa_pgm_canon_nocm
Model 7: Core features, random forests, one-step-ahead, no *cm* modeling



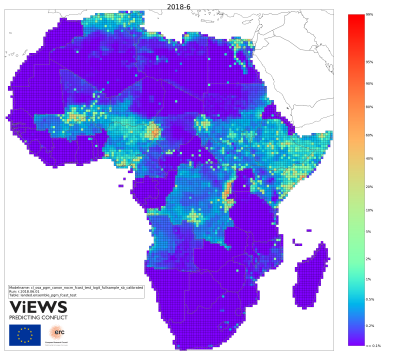
osa_pgm_canon_wcm
Model 8: Core features, random forests, one-step-ahead, *cm* features



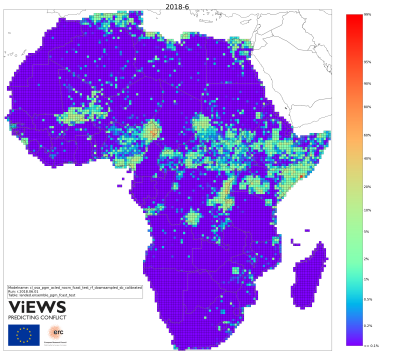
cl_ds_pgm_canon
Model 9: Core features, logit, dynamic simulation, products of *cm* and *pgm* probabilities



cl_ds_pgm_acled
Model 10: Core + protests, logit, dynamic simulation, products of *cm* and *pgm* probabilities



cl_osa_pgm_canon
Model 11: Core features, logit, one-step-ahead, products of *cm* and *pgm* probabilities



cl_osa_pgm_acled
Model 12: Core + protests, random forests, one-step-ahead, products of *cm* and *pgm* probabilities

Figure 7. PRIO-GRID level forecasts June 2018, constituent models

and the degree to which constituent models pick up unique information.

Overall, we have found that the two approaches yield very similar results (for similar conclusions in a different context, see Graefe et al., 2015). The estimated EBMA weights are reported in Appendix A.3.

The most likely reason that the weighed approach of EBMA does not outperform the unweighted approach is that the relative performance of the constituent models varies over time. The EBMA calibrates weights based on the performance of constituent models in the calibration period (Table 1). However, these weights may not be suitable for other time periods (the testing/forecasting period). This can be the case if conflicts in the calibration period are driven by certain factors, for example ethnic cleavages or coup d'etats, and if these factors are not as important in the testing/forecasting period. Such clustering of conflict with similar causes in time (and space) is not unusual, consider for example the numerous conflicts in the immediate aftermath of the breakdown of the Soviet Union or the Arab Spring uprisings. The unweighted average ensembles are not affected by such temporal fluctuations.

Since the results are similar, and EBMA involves considerable complexity compared to the unweighted average ensembles, the ViEWS forecasts are currently based on the latter. However, we will continue to combine models using both approaches to accumulate more evidence on the relative performance of these two ensemble types. In particular, we expect future implementations in ViEWS to impact the utility of EBMA. In particular, we plan to generate ViEWS forecasts with ensembles of more distinct models (for example themed models such as "conflict history", "geography", "political institutions"). Given the estimation procedures of EBMA, it may perform better if the constituent models are more distinct since they are more likely to produce unique forecasts.

### 4.3 The principles and metrics we apply to select and weight models

We currently evaluate our models using the AUPR, AUROC, and Brier scores. Having a suite of performance metrics is useful since model performance is multidimensional. In many model comparisons, one model outperforms another in terms of all three measures, so we can safely conclude that the one is an improvement over another using these metrics. Yet, in other situations, the picture is less consistent. For example, one model might make fewer false positive predictions, but more false negatives, than a second model. In a situation like ours, where the presence of an event is the rarer class, this can lead to a high AUROC for the first model, and a higher AUPR for the second.

Similarly, the Brier score favors sharp, accurate probabilistic predictions (near 0 or 1) while only the relative ordering of the forecasts are needed for the computation of the AUPR and AUROC. Each performance metric includes an implicit relative cost for different types of mistakes (false positives, false negatives) and relative rewards for different types of successes (true positives, true negatives). Since we are more interested in predicting instances of political violence than the absence of such, we give priority to the AUPR over the AUROC, as the former rewards models more for accurately predicting a one, as compared to a zero. We also look to Brier scores when the AUPR and AUROC scores are close or inconsistent. It is also useful to inspect the confusion matrices for various prediction thresholds. We report a number of these in Appendix A.2.

Another important evaluation metric is the set of model weights obtained in the EBMA estimations. While there is not a direct equivalence between the Brier score in the calibration window and EBMA model weights, their relative ranking of models should be similar. The point estimates for the model weights are approximations of the mode for the posterior distribution of the probability of the model. This set of weights is the most useful computed representation of the constituent models, given the data, any priors and the set of models in the ensemble. A model that produces forecasts in the calibration window that are close to the actual values will both have a low Brier score and high model weight, relative to other constituent models.

ViEWS will also work to develop and adapt a number of other performance metrics. We will develop a domain-specific evaluation measure based on differential classification rewards and misclassification costs. We will adapt the concept of 'Earth Mover Distance' (EMD) as a score with which to compare models (Pele and Werman, 2008; Pele and Werman, 2009). EMD calculates the minimum amount of work necessary to move one distribution of values arrayed in n-dimensional space, such as our predictions across space and time, to a target distribution, such as the actual observations of conflict or its absence across pgms.

ViEWS will also disaggregate performance metrics by whether they are calculated on 'new' cases of conflict or whether the unit continued to experience conflict in the time period under investigation. A model that maximizes AUPR for PRIO-GRID cells that have never seen conflict before should go into our ensemble even if it performs poorly for continued conflict, for example.

### 4.4 Combining three types of political violence outcomes

As shown in Figures 1 and 2, the forecasts for the three outcomes are quite similar. This partly reflects that the various forms of organized violence occur through similar processes and are constrained by the same factors. In addition, there is considerable spillover between the forms of violence. One-sided violence, for instance, are most frequently perpetrated in the context of a state-based conflict. We still believe there is ample scope for improving the system by modeling more carefully how the three outcomes affect each other and how they are distinct.

Combining these three outcomes in a single system brings several advantages. First, they together constitute a reasonable definition of political violence that subsumes conflicts such as the ongoing war in Syria, the 1994 genocide in Rwanda, and drug-related organized violence in Mexico. The system allows them to be modeled separately since they follow different dynamics and involve different types of actors. At the same time, VIEWS allows the various types of violence to serve as early warning indicators for each other.

### 4.5 Combining three levels of analysis

The risk of conflict in a given location is influenced by local factors as well as country-level factors. Hence, ViEWS is working to make the two levels of analysis inform each other. Our *pgm* model ensemble currently combines three approaches. The first (models labeled *_nocm_* in Table 3) approach is to ignore *cm* factors. The second (*_wcm_*) is to add a few core *cm* variables to the *pgm* model specification. This approach may be suboptimal, however, since it tends to 'smear' the country-level

risk evenly out across the country's territory. This may lead the model ensemble to over-predict in low-risk locations even when our *pgm* models are able to differentiate between the local risk levels. To counter this, we also include some models (*_cl_*) that are based on the product of the predicted probability at the *cm* and *pgm* level.

Evaluation of the predictive performance of the system indicates that the combination of these three approaches yield better results than each of them in separation.

From 2019, the project will expand the system to also include actors as units, specifying *who* are involved in events. Here, ViEWS will focus on all relevant pairings of actors identified by the UCDP as participants in the political violence events they record, as well as all governments and selected other actors, such as protest movements (Weidmann and Rød, Forthcoming). This level allows tracing actor-related escalation, termination, recurrence, transformation, and external involvement in conflicts, complementing the geographically defined system. An early attempt to set up actor-level predictions appears in Croicu and Hegre (2018).

### 4.6 Handling dynamics

The second column in Table 3 specifies the 'dynamic mode' – how ViEWS specifies models that see monthly data up to the most recent month $t$ into forecasts $k$ months into the future.

#### Dynamic simulation

The first 'dynamic mode' is labeled 'dynamic simulation'. This procedure builds on Hegre et al. (2013) and Hegre et al. (2016) and is discussed at length there.[15] In short, the procedure involves simulating the model parameters based on the estimated coefficients and the variance-covariance matrix of the estimates from the model. In addition, we compute the predicted probabilities for the outcomes for the first month $t$, draw outcomes, recalculate the history variables so that the input feature matrix $X_{t+1}$ at $t+1$ reflects that draw. This is repeated for each month for the forecasting window, and for each simulated set of parameters.

If we are interested in forecasting 2 months into the forecasting period, we first train the constituent models, estimate the weights and produce our ensemble one-month ahead forecast. To produce forecasts for the next month, we need the input feature matrices $X_{t+1}^{(k)}$. For many constituent models, these input features will themselves be functions of actual conflict (e.g., lagged conflict indicators, time since last conflict, spatial distance to nearest conflict). Since these do not exist for next month (after the training window), we use the prediction as the probability of an unobserved feature, for example for conflict at time $t+1$, when forecasting conflict at $t+2$. A simulated value is drawn from this probability, and recorded within a new simulated set of predictors $\widetilde{X}_{t+1}^{(k)}$.

The predictions for the three outcomes are obtained simultaneously within each time step. For each of these, we compute the predicted probability at $t+1$ as a function of information available at $t$,

[15]The first author's original software 'PRIOsim' was rewritten in C# and C++ by Joakim Karlsen for use in these publications. The Python routines underlying the current projects is based on the 'Dynasim' reimplementation of this, written by Jonas Vestby and Frederick Hoyles (Hegre, Hoyles, and Vestby, 2018).

including the status for the other two outcomes. This procedure repeats for every month to the end of the forecasting window.

#### 'One-step-ahead' modeling

In its current implementation, the dynamic simulation mode is restricted to a narrow selection of parametric models (OLS and logit), but the 'one-step-ahead' mode can make use of any model implemented in scikit-learn (Pedregosa et al., 2011).

In this mode, we predict each step into the future ($t+1$, (...), $t+12$, (...) $t+36$) independently, as opposed to dynamic simulation which moves forward through the time sequentially. We do this by estimating a set of models of the form $f_s(X_{t-s})$ where $s$ denotes the number of months into the future to forecast. In regression notation these take the form

$$y_{t+s} = \beta_s X_t$$

The ViEWS 'One-step-ahead' mode does this by time-shifting the right hand side variables with respect to the outcome before models are trained, thus making the model a link function between the future ($y_{t+s}$) and the present ($X_t$).

### 4.7 Handling of missing data

The methods used by ViEWS require that the input data used for the simulations and predictions are complete. Dropping observations with missing values would make it impossible to make predictions for those observations. Additionally, this would create bias in the estimation for the models (Allison, 1999).

We perform multiple imputation using the Amelia II package in R (Honaker et al., 2011). Each missing value is replaced multiple times with appropriate values, creating 5 different complete datasets. For Dynasim, all five datasets are used simultaneously and the results are calculated using the Rubin Rules. Due to technical issues in the One-step-ahead forecast procedure, only one imputed dataset is used for this forecast. Using one dataset instead of five does not bias the results, but reduces statistical efficiency (Buuren, 2012). We provide more details on the issue of missingness in the ViEWS project in an appendix at `pcr.uu.se/research/publications/appendices`. For a more comprehensive test of missing data methods see Randahl (2016).

### 4.8 Projections

In order to provide forecasts for the future, the system requires that the input feature matrix $X_t^{(k)}$ is defined for all the timesteps $t$ over the forecast window. ViEWS will make use of three strategies to project these input features:

The first strategy is provided by our dynamic forecasting system (Section 4.6). The forecasts we generate for state-based conflict or any other endogenous variable at $t$ are used as projected inputs in each relevant equation at $t+1$. A similar approach can be used for other events – these are forecasted using similar methods although only used as input features to the system. For instance, ACLED protest

events (Raleigh et al., 2010) are used in a couple of the ensemble models. Here, we generate projections for these events using the Dynasim approach (see Section 4.6) with ACLED protests in a given *pgm* as the dependent variable.[16]

The second is to use information from external sources to project events. For some features, this is quite straightforward: most countries, for instance, have scheduled dates for elections over the next few years. ViEWS will also search for projections for other features such as droughts in a given location, or expected growth rates for a given country.

The third is to define very simple projections such that an assumption that a feature is unchanged over the forecasting window. This approach is the one we use for most predictors in the system.

# 5 Data

## 5.1 Conflict data

Data on conflict are primarily obtained from UCDP-GED and take the form of events (individual incidents of conflict and organized violence resulting in fatalities and taking place at a given time and place Sundberg and Melander, 2013). Historical data covering 1989–2016 are extracted from the UCDP GED version 17.2 (Croicu and Sundberg, 2013; Allansson, Melander, and Themnér, 2017).[17]

Newer data are provided by the new UCDP-Candidate dataset that provides data in a close to real-time fashion with monthly updates (see Hegre et al., 2018, for an introduction). This allows producing forecasts using input data that extend up to one month before the forecasting window. Here, we use data up to and including April 2018, aggregated according to the procedures explained in Hegre et al. (2018).

The UCDP-Candidate data are in the form of 'candidate events'. Many UCDP definitions (see Gleditsch et al., 2002) are applicable only on a per calendar-year basis, and the final UCDP-GED dataset can only be compiled after the end of the year. The UCDP-Candidate events dataset consequently relaxes the UCDP requirement of a 25-battle related deaths threshold in order for a conflict to be recorded, as well as the requirement of a 'stated goal of incompatibility'. The violence in Togo referred to above, for instance, did not make it to the published version of UCDP-GED.[18] Further, due to data-collection constraints,[19] the very strict requirements in terms of known and clear parties to the conflict are also relaxed as long as there are sufficiently strong indications that such events have a high likelihood of inclusion in the final UCDP datasets at the end of the year.[20] UCDP and ViEWS

[16] We do not include these forecasts in our main presentations, though.

[17] 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 aggregate to the units of analysis described above as described in Hegre et al. (2018). Usage of the API is described in `http://ucdp.uu.se/apidocs/`; the data are available as version 17.2 (1989–2016).

[18] The protest violence in Togo 2017 was not included in the UCDP state-based category since the protesters did not fulfill the organizational criterion. The violence could not be included in the one-sided violence category either since some protesters were armed.

[19] Examples of data-collection constraints are unavailability of certain in-depth sources so close in time to the actual violence.

[20] The decision whether to incluce a report in the UCDP-Candidate data is taken by the UCDP coder and project manager in charge of data collection. In the UCDP-GED system, these relaxed-criteria events are clearly marked with a code status distinct from that of clear events that will definitely be included in the final UCDP-GED dataset.

have developed a coding procedure for such events with a goal of making the monthly candidate event sample as close in content to the final dataset as possible.[21]

UCDP-GED includes high-resolution temporal and geographical references that are easy to match to predictor data. In about 15% of the cases, however, UCDP coders have been unable to identify the location more precisely than for instance a given second-order administrative region. In such cases, the UCDP assigns the center point of the region as a place-holder location and marks the event with a precision score. For prediction purposes, the place-holder solution is not optimal. Hence, ViEWS has developed a method to multiply impute the location of this and other types of imprecise codings, as documented in Croicu and Hegre (2018). All the forecasts reported by ViEWS are based on a set of 5 imputed location datasets. Croicu and Hegre (2018) show that this improves the predictive performance of the system considerably.

### 5.2 Protests

Data on protests are extracted from ACLED (Raleigh et al., 2010), covering the entire period of data availability up to the last completely coded month (April 2018). Data is automatically retrieved from the ACLED API each month and aggregated to the resolution required for analysis.

### 5.3 PRIO-GRID

We make use of PRIO-GRID version 2.0 (Tollefsen, Strand, and Buhaug, 2012) for our grid-level structure, as retrieved from its API, with the geographic structure downloaded directly from the website. As PRIO-GRID data are currently only available on a yearly basis, data for each year was applied to each month in the given calendar year.

Where predictor data are only available at specific intervals (such as population data or per-cell GDP data, where data are available every 5- or 10-years), data were interpolated linearly between two observations. Similarly, where data were not available for the entire period under analysis, observations were linearly extrapolated from the first/last two observations in both directions. The non-interpolated data was also kept with the goal of creating custom imputations in the future to better account for the lack of such data.

### 5.4 Country-level data

The models include a range of *cm*-level data. Data on political institutions are based on V-Dem (Coppedge et al., 2011). Economic data come from World Bank (2017), and data on ethnic exclusion from EPR (Vogt et al., 2015). Demographic data come from the IIASA (IIASA, 2014). Complete sources are given in the 'ViEWS Independent Variables' online appendix (`https://www.pcr.uu.se/research/views/publications/appendices/`).

[21] Further improvements to this procedure are being rolled out, including clear markers for each type of relaxed assumptions made, in order to improve sample consistency and uncertainty management across the entire product. Approximately 30% of all coded events for the first three months of 2017 are "relaxed-criteria" events (137/467), the total sample of events for these months is consistent with the total sample available within UCDP GED for the same months in the past 10 years (being within one standard deviation from the mean).

# 6 Conclusions

This article has presented initial results from the ViEWS forecasting system and summarized the methodology behind these forecasts. We have accounted for the evaluation procedures of ViEWS and established a frame of reference for this forecasting problem. The evaluation indicates that the system generates very accurate forecasts for conflict-prone regions in Africa.

ViEWS is being developed according to four guiding principles: public availability, uniform coverage, transparency, and methodological innovation. Public availability is ensured by the complete release of all data and procedures through the ViEWS website. An implication of this principle is that ViEWS is restricted to using data that allows us to make the data available, even if predictive performance could be improved by the inclusion of such data. ViEWS safeguards uniform coverage by relying on the UCDP suite of datasets, which consistently applies a clearly articulated definition and procedures that minimize the risk of overseeing conflict.

Transparency and replicability are essential objectives for ViEWS. They are also challenging given the complexity of the system. The system is documented through this paper, a number of auxilliary papers available at our website, and the website itself. We are making the source code available upon request and are working toward making it publicly available on Github.

There is ample room for improving ViEWS. An important function of this paper is to document the performance of the current system so that we have a baseline against which we can assess the gains we make from further modeling. As transparency is a core value of this project, we will continue to publicly report our performance and document our system moving forward. Since there exists no universally valid measure of predictive performance, improvement on multi-dimensional metrics compared to an established, problem-specific baseline is necessary to track progress and setbacks. ViEWS welcomes any interested party to take our replication material and compare their own forecasting approaches against the baseline and evaluation routines presented here.

The forecasts we have presented are mostly driven by past conflict events as coded by the UCDP. 'Early warning' of new conflicts is a major challenge that requires collecting data we currently do not have. Such data collection is a top priority for the project now that the infrastructure of the system is ready.

How useful is the current system? Even though ViEWS currently has limited ability to forecast entirely new conflicts, two important results can be discerned. First, the system models geographical diffusion quite well. Because of its vicinity to the various conflicts in Nigeria, the recent tensions in Cameroon are reflected as a high predicted probability of further violence. Secondly, the models show how persistent organized violence is in Africa. Our results indicate that the major conflict clusters in Africa will continue to be very violent over the coming three-year period.

Can the system be misused? A government that sees our risk assessment for a location inside territory under its rule might conceivably be led to preempt the conflict, possibly through violent means. However, we do not believe this is a great danger. First, local governments have much better information about what is going on in their own countries than any system based on open-source data can deliver. Second, even if ViEWS should be perceived to be able to reveal unknown information, being alerted to a risk of this form does not necessarily have to lead to undesirable responses. Governments

may take steps to reduce tensions peacefully. More importantly, outside observers that lack the inside information of local governments may use high-quality forecasts to take action. NGOs can apply pressure on conflict actors or prepare for humanitarian assistance. Large organizations such as the UN may use the forecasts when they decide on whether to deploy peacekeeping operations (Hegre, Hultman, and Nygård, 2019).

Moreover, many governments have their own intelligence systems upon which they act in response to threats of organized violence. Such intelligence is never publicly available. If an open-source early-warning system such as ViEWS can be sufficiently accurate, critical voices may in some cases use this to challenge the assessments government actions are based on.

We envision our work on ViEWS as a step towards a future where high-resolution forecasts of conflict at practically useful spatial and temporal scales are publicly available. While any such system for violence will necessarily be less precise than the modeling of most non-human physical systems, the goal is to improve outcomes relative to a world where these forecasts do not exist. Even an imperfect future system that builds on the current ViEWS architecture has the potential to inform the placement of peacekeepers, the deployment of NGO resources, and even the decisions of private citizens; potentially saving lives. Attaining this goal will take a community of researchers collaborating across domain specialties to identify mistakes, suggest innovations, and incorporate successful new ideas into a computational infrastructure. We believe ViEWS is a necessary start towards bringing this vision to fruition.

## Replication data and source code

Replication data, source code, and datasets with detailed predictions are available at `http://pcr.uu.se/research/views/data/replication-data/`.

## Funding

ViEWS receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 694640). ViEWS computations are performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

## References


Agence France Presse. *'Three killed' in Togo opposition clashes.* (Visited on 10/19/2017).

Allansson, Marie, Erik Melander, and Lotta Themnér (2017). "Organized violence, 1989–2016". In: *Journal of Peace Research* 54.4, pp. 574–587.

Allison, Paul D. (1999). *Missing data.* Thousand Oaks, CA: Sage.

Armstrong, J Scott (2001). "Combining forecasts". In: *Principles of forecasting.* Springer, pp. 417–439.

BBC Monitoring Africa. *Kenyan minister declares curfew in restive Mt Elgon region.* (Visited on 03/06/2018).

Beger, Andreas, Cassy L Dorff, and Michael D Ward (2014). "Ensemble forecasting of irregular leadership change". In: *Research & Politics* 1.3.

Buuren, Stef van (2012). *Flexible imputation of missing data.* CRC press.

Cederman, Lars-Erik, Simon Hug, and Lutz F. Krebs (2010). "Democratization and civil war: Empirical evidence". In: *Journal of Peace Research* 47.4, pp. 377–394.

Cederman, Lars-Erik, Andreas Wimmer, and Brian Min (2010). "Why do ethnic groups rebel? New data and analysis". In: *World Politics* 62.1, pp. 87–119.

Chao Chen, Andy Liaw and Leo Breiman (2004). "Using Random Forests to Learn Imbalanced Data". In: *University of California-Berkley Tech Report 666.*

Colaresi, Michael and Zuhaib Mahmood (2017). "Do the Robot: Lessons from Machine Learning to Improve Conflict Forecasting". In: *Journal of Peace Research* 54.2, pp. 193–214.

Collier, Paul and Anke Hoeffler (2004). "Greed and Grievance in Civil War". In: *Oxford Economic Papers* 56.4, pp. 563–595.

Collier, Paul, Lani Elliot, Håvard Hegre, Anke Hoeffler, Marta Reynal-Querol, and Nicholas Sambanis (2003). *Breaking the Conflict Trap. Civil War and Development Policy.* Oxford: Oxford University Press.

Coppedge, Michael, John Gerring, David Altman, Michael Bernhard, Steven Fish, Allen Hicken, Matthew Kroenig, Staffan I. Lindberg, Kelly McMann, Pamela Paxton, Holli A. Semetko, Svend-Erik Skaaning, Jeffrey Staton, and Jan Teorell (2011). "Conceptualizing and Measuring Democracy: A New Approach". In: *Perspectives on Politics* 9.02, pp. 247–267.

Croicu, Mihai and Håvard Hegre (2018). *A Fast Spatial Multiple Imputation Procedure for Imprecise Armed Conflict Events.* Typescript, Uppsala University/ViEWS. `http://www.pcr.uu.se/research/views`.

Croicu, Mihai and Ralph Sundberg (2013). *UCDP Georeferenced Event Dataset Codebook Version 4.0.* Typescript, Uppsala Conflict Data Program. URL: `http://www.pcr.uu.se/research/ucdp/datasets/ucdp_ged/`.

Eck, Kristine and Lisa Hultman (2007). "One-Sided Violence against Civilians in War: Insights from New Fatality Data". In: *Journal of Peace Research* 44.2, pp. 233–246.

Fearon, James D. and David D. Laitin (2003). "Ethnicity, Insurgency, and Civil War". In: *American Political Science Review* 97.1, pp. 75–90.

Gates, Scott, Håvard Hegre, Håvard Mokleiv Nygård, and Håvard Strand (2012). "Development Consequences of Armed Conflict". In: *World Development* 40.9, pp. 1713–1722.

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, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002). "Armed Conflict 1946–2001: A New Dataset". In: *Journal of Peace Research* 39.5, pp. 615–637.

Graefe, Andreas, Helmut Küchenhoff, Veronika Stierle, and Bernhard Riedl (2015). "Limitations of Ensemble Bayesian Model Averaging for forecasting social science problems". In: *International Journal of Forecasting* 31.3, pp. 943–951.

Hegre, Håvard, Frederick Hoyles, and Jonas Vestby (2018). *Dynasim – A fast and simple simulator of endogenous (simultaneous) panel data regression models*. Typescript, Uppsala University.

Hegre, Håvard, Lisa Hultman, and Håvard Mokleiv Nygård (2019). "Evaluating the conflict-reducing effect of UN peacekeeping operations". In: *Journal of Politics* in press.

Hegre, Håvard, Tanja Ellingsen, Scott Gates, and Nils Petter Gleditsch (2001). "Toward a Democratic Civil Peace? Democracy, Political Change, and Civil War, 1816–1992". In: *American Political Science Review* 95.1, pp. 33–48.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal (2013). "Predicting Armed Conflict 2010–2050". In: *International Studies Quarterly* 55.2, pp. 250–270. DOI: 10.1111/isqu.12007.

Hegre, Håvard, Halvard Buhaug, Katherine V. Calvin, Jonas Nordkvelle, Stephanie T. Waldhoff, and Elisabeth Gilmore (2016). "Forecasting civil conflict along the shared socioeconomic pathways". In: *Environmental Research Letters* 11.5, p. 054002. DOI: 10.188/1748-9326/11/5/054002.

Hegre, Håvard, Mihai Croicu, Kristine Eck, and Stina Högbladh (2018). *UCDP-monthly. Monthly updated organized violence data in event and aggregated forms*. Typescript, Uppsala University. http://www.pcr.uu.se/research/views/.

Honaker, James, Gary King, Matthew Blackwell, and others (2011). "Amelia II: A program for missing data". In: *Journal of statistical software* 45.7, pp. 1–47. URL: http://artax.karlin.mff.cuni.cz/~hans/src/doc/r-cran-amelia/amelia.pdf (visited on 05/15/2016).

IIASA (2014). *SSP Database (version 0.93)*. URL: https://secure.iiasa.ac.at/web-apps/ene/SspDb/dsd?Action=htmlpage\&page=about.

International Crisis Group (2018). *Crisis watch: Global overview, May 2018: Cameroon*. https://www.crisisgroup.org/ (Visited on 06/20/2018).

Melander, Erik, Therése Pettersson, and Lotta Themnér (2016). "Organized violence, 1989–2015". In: *Journal of Peace Research* 53.5, pp. 727–742.

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.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pele, Ofir and Michael Werman (2008). "A linear time histogram metric for improved sift matching". In: *Computer Vision–ECCV 2008*. Springer, pp. 495–508.

— (2009). "Fast and robust earth mover's distances". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE, pp. 460–467.

Raftery, Adrian E and Steven M Lewis (1992). "Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo". In: *Statistical Science* 7.4, pp. 493–497.

Raftery, Adrian E., Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski (2005). "Using Bayesian Model Averaging to Calibrate Forecast Ensembles". In: *American Meteorological Society* 133.1, pp. 1155–1173.

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, Hå vard Hegre, Joakim Karlsen, and Andrew Linke (2010). "Introducing ACLED: An Armed Conflict Location and Event Dataset". In: *Journal of Peace Research* 47.5, pp. 651–660.

Randahl, David (2016). *Raoul: An R-Package for Handling Missing Data*. URL: `http://www.diva-portal.org/smash/get/diva2:940656/FULLTEXT01.pdf`.

Ricardo Barandela Jose Salvador Sanchez, Vicente GarcÄ±*aAndEdgarRangel* (2003). "Strategies for learning in class imbalance problems". In: *Pattern Recognition* 36.3, pp. 849–851.

Sundberg, Ralph, Kristine Eck, and Joakim Kreutz (2012). "Introducing the UCDP Non-State Conflict Dataset". In: *Journal of Peace Research* 49, pp. 351–362.

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`.

Thyne, Clayton (2006). "ABC's, 123's, and the Golden Rule: The Pacifying Effect of Education on Civil War, 1980–1999". In: *International Studies Quarterly* 50.4, pp. 733–754.

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`.

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin (2015). "Integrating Data on Ethnicity, Geography, and Conflict The Ethnic Power Relations Data Set Family". In: *Journal of Conflict Resolution* 59.7, pp. 1327–42.

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.

Weidmann, Nils B. and Espen Geelmuyden Rød (Forthcoming). *The Internet and Political Protest in Autocracies*. Oxford University Press.

World Bank (2017). *World Development Indicators*. Washington DC: The World Bank.

# Appendices

# A Evaluation results, test January 2014–December 2016

## A.1 Results, evaluation

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_cm_canon_base_eval_test_sb_calibrated | 0.94419 | 0.08766 | 0.81801 |
| ds_cm_acled_base_eval_test_sb_calibrated | 0.93900 | 0.09025 | 0.80143 |
| osa_cm_acled_base_eval_test_logit_fullsample_sb_calibrated | 0.91981 | 0.09851 | 0.77482 |
| osa_cm_acled_base_eval_test_rf_downsampled_sb_calibrated | 0.90150 | 0.09934 | 0.76869 |
| osa_cm_canon_base_eval_test_logit_fullsample_sb_calibrated | 0.91855 | 0.09959 | 0.78842 |
| osa_cm_canon_base_eval_test_rf_downsampled_sb_calibrated | 0.89669 | 0.09771 | 0.76452 |
| average_sb | 0.93066 | 0.09248 | 0.79742 |
| ebma_sb | 0.93022 | 0.09364 | 0.79845 |

Table A-1. SB constituent models and ensembles, 36 months, CM level

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_cm_canon_base_eval_test_os_calibrated | 0.92258 | 0.08116 | 0.77910 |
| ds_cm_acled_base_eval_test_os_calibrated | 0.92475 | 0.07886 | 0.75480 |
| osa_cm_acled_base_eval_test_logit_fullsample_os_calibrated | 0.90482 | 0.08229 | 0.71093 |
| osa_cm_acled_base_eval_test_rf_downsampled_os_calibrated | 0.91562 | 0.07980 | 0.76279 |
| osa_cm_canon_base_eval_test_logit_fullsample_os_calibrated | 0.90546 | 0.07783 | 0.76821 |
| osa_cm_canon_base_eval_test_rf_downsampled_os_calibrated | 0.91019 | 0.07747 | 0.75544 |
| average_os | 0.92341 | 0.07582 | 0.78521 |
| ebma_os | 0.92292 | 0.07742 | 0.78454 |

Table A-2. OS constituent models and ensembles, 36 months, CM level

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_cm_canon_base_eval_test_ns_calibrated | 0.90912 | 0.06488 | 0.64501 |
| ds_cm_acled_base_eval_test_ns_calibrated | 0.87874 | 0.06876 | 0.62372 |
| osa_cm_acled_base_eval_test_logit_fullsample_ns_calibrated | 0.88705 | 0.06605 | 0.60998 |
| osa_cm_acled_base_eval_test_rf_downsampled_ns_calibrated | 0.93445 | 0.06030 | 0.65482 |
| osa_cm_canon_base_eval_test_logit_fullsample_ns_calibrated | 0.92166 | 0.06127 | 0.69039 |
| osa_cm_canon_base_eval_test_rf_downsampled_ns_calibrated | 0.91697 | 0.06228 | 0.67825 |
| average_ns | 0.92338 | 0.06157 | 0.67972 |
| ebma_ns | 0.92323 | 0.06278 | 0.67855 |

Table A-3. NS constituent models and ensembles, 36 months, CM level

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_pgm_canon_nocm_eval_test_sb_calibrated | 0.91103 | 0.00611 | 0.20776 |
| ds_pgm_canon_wcm_eval_test_sb_calibrated | 0.91756 | 0.00600 | 0.21749 |
| ds_pgm_acled_wcm_eval_test_sb_calibrated | 0.91856 | 0.00607 | 0.19714 |
| osa_pgm_acled_nocm_eval_test_rf_downsampled_sb_calibrated | 0.94875 | 0.00605 | 0.20373 |
| osa_pgm_acled_wcm_eval_test_rf_downsampled_sb_calibrated | 0.94845 | 0.00601 | 0.20507 |
| osa_pgm_canon_nocm_eval_test_logit_fullsample_sb_calibrated | 0.92290 | 0.00596 | 0.21989 |
| osa_pgm_canon_nocm_eval_test_rf_downsampled_sb_calibrated | 0.94506 | 0.00606 | 0.19487 |
| osa_pgm_canon_wcm_eval_test_rf_downsampled_sb_calibrated | 0.94858 | 0.00602 | 0.19662 |
| CL_ds_pgm_canon_nocm_eval_test_sb_calibrated | 0.91469 | 0.00598 | 0.20110 |
| CL_ds_pgm_acled_nocm_eval_test_sb_calibrated | 0.91765 | 0.00603 | 0.19961 |
| CL_osa_pgm_canon_nocm_eval_test_logit_fullsample_sb_calibrated | 0.91082 | 0.00593 | 0.21472 |
| CL_osa_pgm_acled_nocm_eval_test_rf_downsampled_sb_calibrated | 0.94201 | 0.00608 | 0.20097 |
| average_select_sb | 0.95198 | 0.00589 | 0.24656 |
| ebma | 0.95196 | 0.00590 | 0.24419 |

Table A-4. SB constituent models and ensembles, 36 months, PGM level

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_pgm_canon_nocm_eval_test_os_calibrated | 0.89694 | 0.00529 | 0.17001 |
| ds_pgm_canon_wcm_eval_test_os_calibrated | 0.91438 | 0.00518 | 0.17426 |
| ds_pgm_acled_wcm_eval_test_os_calibrated | 0.91934 | 0.00536 | 0.10382 |
| osa_pgm_acled_nocm_eval_test_rf_downsampled_os_calibrated | 0.94970 | 0.00516 | 0.15320 |
| osa_pgm_acled_wcm_eval_test_rf_downsampled_os_calibrated | 0.95608 | 0.00513 | 0.15819 |
| osa_pgm_canon_nocm_eval_test_logit_fullsample_os_calibrated | 0.90232 | 0.00513 | 0.14426 |
| osa_pgm_canon_nocm_eval_test_rf_downsampled_os_calibrated | 0.94807 | 0.00516 | 0.15784 |
| osa_pgm_canon_wcm_eval_test_rf_downsampled_os_calibrated | 0.95523 | 0.00512 | 0.15487 |
| CL_ds_pgm_canon_nocm_eval_test_os_calibrated | 0.92453 | 0.00536 | 0.18576 |
| CL_ds_pgm_acled_nocm_eval_test_os_calibrated | 0.92757 | 0.00545 | 0.17025 |
| CL_osa_pgm_canon_nocm_eval_test_logit_fullsample_os_calibrated | 0.91544 | 0.00508 | 0.16602 |
| CL_osa_pgm_acled_nocm_eval_test_rf_downsampled_os_calibrated | 0.95202 | 0.00513 | 0.16038 |
| average_select_os | 0.95484 | 0.00503 | 0.19899 |
| ebma | 0.95446 | 0.00509 | 0.19282 |

Table A-5. OS constituent models and ensembles, 36 months, PGM level

| | ROC_AUC | Brier_Score | PR_AUC |
|---|---|---|---|
| ds_pgm_canon_nocm_eval_test_ns_calibrated | 0.75801 | 0.00393 | 0.03170 |
| ds_pgm_canon_wcm_eval_test_ns_calibrated | 0.75018 | 0.00393 | 0.02962 |
| ds_pgm_acled_wcm_eval_test_ns_calibrated | 0.75136 | 0.00398 | 0.02164 |
| osa_pgm_acled_nocm_eval_test_rf_downsampled_ns_calibrated | 0.91142 | 0.00393 | 0.05368 |
| osa_pgm_acled_wcm_eval_test_rf_downsampled_ns_calibrated | 0.92565 | 0.00393 | 0.04966 |
| osa_pgm_canon_nocm_eval_test_logit_fullsample_ns_calibrated | 0.77797 | 0.00394 | 0.03900 |
| osa_pgm_canon_nocm_eval_test_rf_downsampled_ns_calibrated | 0.90118 | 0.00393 | 0.04877 |
| osa_pgm_canon_wcm_eval_test_rf_downsampled_ns_calibrated | 0.91914 | 0.00392 | 0.05301 |
| CL_ds_pgm_canon_nocm_eval_test_ns_calibrated | 0.80563 | 0.00393 | 0.03604 |
| CL_ds_pgm_acled_nocm_eval_test_ns_calibrated | 0.79437 | 0.00393 | 0.02626 |
| CL_osa_pgm_canon_nocm_eval_test_logit_fullsample_ns_calibrated | 0.82810 | 0.00393 | 0.05063 |
| CL_osa_pgm_acled_nocm_eval_test_rf_downsampled_ns_calibrated | 0.92229 | 0.00393 | 0.06356 |
| average_select_ns | 0.89649 | 0.00393 | 0.04929 |
| ebma | 0.89356 | 0.00393 | 0.04895 |

Table A-6. NS constituent models and ensembles, 36 months, PGM level

## A.2 Confusion matrices

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 1194 | 36 |
| Pr 0.1-0.25 | 135 | 80 |
| Pr 0.25-0.5 | 86 | 17 |
| Pr 0.5-1 | 95 | 301 |

Table A-7. Confusion matrix, SB, CM level

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 1314 | 53 |
| Pr 0.1-0.25 | 139 | 39 |
| Pr 0.25-0.5 | 72 | 17 |
| Pr 0.5-1 | 90 | 220 |

Table A-8. Confusion matrix, OS, CM level

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 1441 | 45 |
| Pr 0.1-0.25 | 162 | 21 |
| Pr 0.25-0.5 | 46 | 23 |
| Pr 0.5-1 | 77 | 129 |

Table A-9. Confusion matrix, NS, CM level

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 380982 | 1026 |
| Pr 0.1-0.25 | 1332 | 302 |
| Pr 0.25-0.5 | 322 | 203 |
| Pr 0.5-1 | 24 | 73 |

Table A-10. Confusion matrix, SB, PGM level

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 381914 | 807 |
| Pr 0.1-0.25 | 823 | 192 |
| Pr 0.25-0.5 | 350 | 116 |
| Pr 0.5-1 | 10 | 52 |

Table A-11. Confusion matrix, OS, PGM level

| | 0 | 1 |
|---|---|---|
| Pr 0-0.1 | 383570 | 597 |
| Pr 0.1-0.25 | 83 | 14 |

Table A-12. Confusion matrix, NS, PGM level

### A.3 EBMA weights, evaluation

| | Weight SB | Weight OS | Weight NS |
|---|---|---|---|
| ds_cm_canon_base_eval_test_sb | 0.1782 | 0.0919 | 0.0883 |
| ds_cm_acled_base_eval_test_sb | 0.1429 | 0.2703 | 0.4220 |
| osa_cm_acled_base_eval_test_logit_fullsample_sb | 0.1948 | 0.2242 | 0.0835 |
| osa_cm_acled_base_eval_test_rf_downsampled_sb | 0.1733 | 0.1523 | 0.2076 |
| osa_cm_canon_base_eval_test_logit_fullsample_sb | 0.2038 | 0.0855 | 0.0617 |
| osa_cm_canon_base_eval_test_rf_downsampled_sb | 0.1069 | 0.1758 | 0.1368 |

Table A-13. Model weights, CM level:SB, OS, NS

| | Weight SB | Weight OS | Weight NS |
|---|---|---|---|
| ds_pgm_canon_nocm_eval_test_sb | 0.0826 | 0.0833 | 0.0833 |
| ds_pgm_canon_wcm_eval_test_sb | 0.0827 | 0.0833 | 0.0833 |
| ds_pgm_acled_wcm_eval_test_sb | 0.0829 | 0.0833 | 0.0833 |
| osa_pgm_acled_nocm_eval_test_rf_downsampled_sb | 0.0841 | 0.0833 | 0.0834 |
| osa_pgm_acled_wcm_eval_test_rf_downsampled_sb | 0.0838 | 0.0834 | 0.0833 |
| osa_pgm_canon_nocm_eval_test_logit_fullsample_sb | 0.0829 | 0.0833 | 0.0833 |
| osa_pgm_canon_nocm_eval_test_rf_downsampled_sb | 0.0837 | 0.0833 | 0.0834 |
| osa_pgm_canon_wcm_eval_test_rf_downsampled_sb | 0.0836 | 0.0833 | 0.0834 |
| CL_ds_pgm_canon_nocm_eval_test_sb | 0.0830 | 0.0834 | 0.0833 |
| CL_ds_pgm_acled_nocm_eval_test_sb | 0.0834 | 0.0834 | 0.0833 |
| CL_osa_pgm_canon_nocm_eval_test_logit_fullsample_sb | 0.0831 | 0.0833 | 0.0833 |
| CL_osa_pgm_acled_nocm_eval_test_rf_downsampled_sb | 0.0842 | 0.0834 | 0.0834 |

Table A-14. Model weights, PGM level: SB, OS, NS

## B How sets of features are combined to models

This appendix describes the combination of features into models. The core sets of variables for pgm are presented in table A-15.

- pgm models denoted with wcm are "with country month predictors". This means they include the cm core set of features at the pgm level.
- endog common are the endogenous (lagged, decayed, spatially lagged) dependent variables that can feed information across the models. They are included in every model.
- endog specific are those derived variables of the outcome which are only included in the outcomes model itself, they are not cross-referenced by other models.
- Models that include ACLED protests do so in a manner identical to how the other outcomes are included.

| pgm_core | cm core | endog common | endog specific | |
|---|---|---|---|---|
| ln_bdist3 | fvp_lngdpcap_nonoilrent | l1_ged_dummy_sb | l2_outcome | |
| ln_ttime | fvp_lngdpcap_oilrent | l1_ged_dummy_ns | l3_outcome | |
| ln_capdist | ln_fvp_population200 | l1_ged_dummy_os | l4_outcome | |
| ln_pop | fvp_grgdpcap_oilrent | decay_12_cw_ged_dummy_sb_0 | l5_outcome | |
| ln_dist_diamsec | fvp_grgdpcap_nonoilrent | decay_12_cw_ged_dummy_ns_0 | l6_outcome | |
| ln_dist_petroleum | ln_fvp_timeindep | decay_12_cw_ged_dummy_os_0 | l7_outcome | |
| gcp_li_mer | ln_fvp_timesincepreindepwar | q_1_1_l1_sb | l8_outcome | |
| imr_mean | ln_fvp_timesinceregimechange | q_1_1_l1_ns | l9_outcome | |
| mountains_mean | fvp_demo | q_1_1_l1_os | l10_outcome | |
| urban_ih_li | fvp_semi | | l11_outcome | |
| excluded_dummy_li | ssp2_edu_sec_15_24_prop | | l12_outcome | |
| agri_ih_li | fvp_prop_excluded | | q_1_1_l2_outcome | |
| barren_ih_li | ssp2_urban_share_iias | | q_1_1_l3_outcome | |
| forest_ih_li | ssp2_urban_share_iias | | | |
| savanna_ih_li | | | | |
| shrub_ih_li | | | | |
| pasture_ih_li | | | | |

Table A-15. Feature sets, endog common are shared endogenous variables between models for all outcomes. endog specific are endogenous variables specific to each models' particular outcome, so a model predicted state-based conflict has 12 lags of state-based violence but only one lag of non-state violence.

# C Descriptive statistics of dependent variables

The *pgm* dataset has very strong class imbalance.0.3% of pgms had state based conflict events, 0.1% had non state conflict events, and 0.2% had one sided violence.

| | |
|---|---|
| avg_sb | 0.003227 |
| avg_ns | 0.001123 |
| avg_os | 0.002112 |
| avg_decay_sb | 0.044849 |
| avg_decay_ns | 0.033029 |
| avg_decay_os | 0.039404 |
| stdev_decay_sb | 0.129578 |
| stdev_decay_ns | 0.099688 |
| stdev_decay_os | 0.116185 |

Table A-16. Descriptive statistics of dependent variables 1990-2018. Also includes decay functions of those dependent variables

# D Downsampling and calibration

## D.1 Downsampling

A majority of the models in ViEWS were trained on all available observations. All our random forests, however, were trained on a down-sampled dataset. This serves two purposes. The *pgm* unit of analysis consists of about 11,000 units for Africa only. Sampled monthly over the 1990–2014 period for the training dataset, this amounts to a dataset with about 3.21M rows. Only 13,739 of these – 0.4% – contain observations of UCDP-GED events. To facilitate the estimation of computationally intensive

models, we trained them on a dataset containing all *pgm* units with at least one UCDP-GED event and a random sample of 10% of the remaining observations.

Downsampling is also another way of inducing an asymmetric cost-function into our computation. If incorrectly predicting peace when there is violence, is more costly than predicting violence when there is actually peace, then we would like our forecasts to hue more closely to predicting the events of violence, when possible, even at the cost of over-estimating some violence in peaceful circumstances. By downsampling very frequent non-events, we are reducing their influence on the fitted models relative to the instances where events occurred. If observations that result in events and non-events are weighted equally, than any unique signal in the rare minority class may be lost. For example, a distinct, but rare, data-generation process might lead to a higher probability of an event as compared to most observations, which will have a lower probability of an event. In this case, downsampling will help our algorithms learn the patterns in cases where violent events occurred, instead of those patterns being treated as random, rare, noise deviating from the more frequent non-events. This should produce higher precision, for example, as compared to non-weighted training where events are highly infrequent because the model has been trained to work harder to predict events, as compared to non-downsampled cases (Ricardo Barandela and Rangel, 2003; Chao Chen and Breiman, 2004).

Trivially, downsampling frequent events leads to more predictions of events simply by artificially shifting the mean upwards. Our calibration procedure transform these predicted probabilities so that they in aggregate yield a predicted conflict intensity that is as close to the actual as possible.

Out-of-sample evaluation of various models indicate that models estimated on down-sampled data predict about as well as those trained on the entire dataset. Hence, the efficiency gain come at negligible performance loss.

### D.2 Calibration

Forecasting requires that each model is well calibrated – that is, that the average predicted outcome probabilities for a set of cases is similar to the actual relative frequency for that set. Models that were trained on an asymmetrically downsampled dataset do not have this property, and requires calibration. The same applies to the models that are constructed as the product of *cm* and *pgm* probabilities. To ensure that our ensemble forecasts are well calibrated, we calibrated each constituent model before entering them into the ensemble.

We use the calibration partition to calibrate the models. We obtain recentering and rescaling parameters $\gamma_{0i}, \gamma_{1i}$ by estimating logistic regression models for each constituent model on the calibration period (see Table 1):

$$logit(p(Y_v^c = 1)) = \hat{\gamma}_{0i} + \hat{\gamma}_{1i} z_{iv}^c$$

where $z_{iv}^c$ is the log odds of conflict for model $i$ on conflict type $v$. The rescaling parameters $\hat{\gamma}_{0i}, \hat{\gamma}_{1i}$ are then used to shift and strengthen the probabilities in the forecasting period (Table 1).

If a model is well calibrated, then an event occurs approximately $x$ percent of the time when the model suggests that there is an $x$ percent chance of an event occurring. This can be gauged visually with calibration plots. In calibration plots, the predicted probabilities are binned on the x-axis

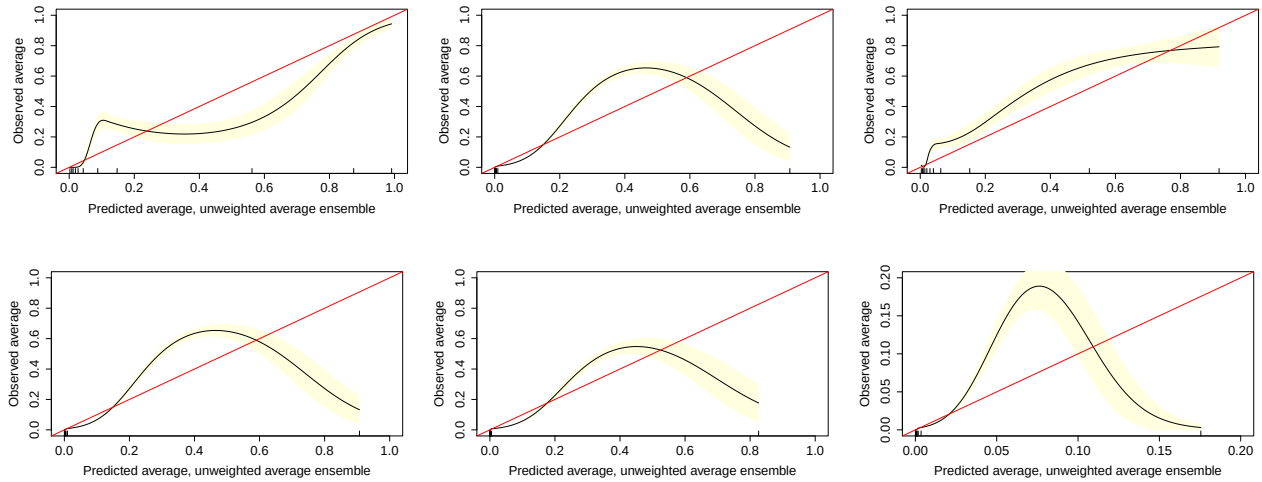



Figure A-1. Calibration plots, *cm* (top) and *pgm* (bottom) level. Left: **sb**. Centre: **os**. Right: **ns**. Note that the x-axis and y-axis is different in the bottom right plot.

and the frequency of actual events within the observations in each bin is plotted on the y-axis. A perfectly calibrated model follows a 45 degree angle. Deviations indicate that the model underpredicts or overpredicts. We show calibration plots for our six ensembles in Figure A-1. The top panel plots the *cm* ensembles. Overall, the *cm* ensembles assign both too low and too high probabilities. On the left hand side of each plot, we can see that the predicted probabilities are lower than the actual probability. In the middle of the plots, however, the predicted probability is too high. Overall, the *cm*-**ns** ensemble (top right) is better calibrated than the *cm*-**sb** (top left) and *cm*-**os** (top center) ensembles. The bottom panel plots the *pgm* ensembles. Here, we can also see that all three ensembles assign both too low and too high probabilities. In particular, the actual probability is much lower than the predicted probability when the predicted probability is higher than 0.5 (*pgm*-**sb** and *pgm*-**os** ensembles) and 0.1 (*pgm*-**ns** ensemble).

We can also evaluate how the calibration of models change over time. In Figure A-2, we display the mean actual/predicted probability of conflict on the y-axis, and months in the testing/forecasting period on the x-axis. The colors indicate the conflict type, blue for **sb**, green for **os**, and red for **ns**. Moreover, solid lines are the observed relative frequencies and the dotted lines the predicted probabilities from the unweighted average ensembles. The plots show ...

# E Data Management

Data are currently stored in a large, highly normalized (3NF) Postgres database for redundancy and consistency. Each unit of analysis (*pgm, py, pg, cm, cy, c, am, ay, a*) thus has its own individual table and storespace; each piece of data is stored once and only once; each individual relation is unique across the entire 90 GB database. These measures eliminate errors stemming from data duplication across various datasets as well as mitigate potential human errors.

Quantities of interest are computed and stored back in the database automatically through an

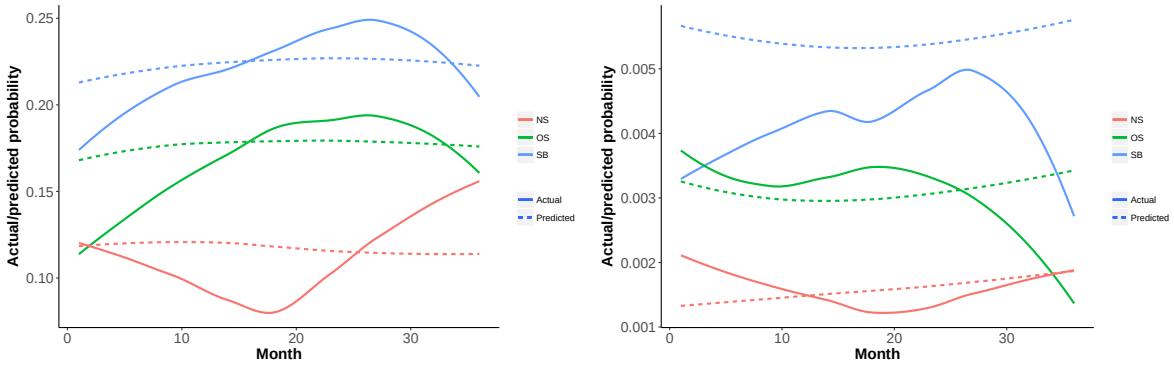



Figure A-2. Calibration over time, *cm* (left) and *pgm* (right). The solid lines are smoothed with a loess function. Note that the y-axis differs in each plot.

organized data ingestion process with each monthly update, as are imputations and monthly estimation results.

The database is completely versioned; new versions are automatically produced through a custom-built backup-and-store mechanism every week as well as after each data update cycle. For security, a unique hash of each backup is created with each backup and stored separately from the very beginning, preventing data tampering or accidental corruption.