UPPSALA UNIVERSITET

ViEWS PREDICTING CONFLICT

ViEWS: Data setup and notation.

Online appendix A to ViEWS$_{2020}$: Revising and evaluating the ViEWS political Violence Early-Warning System.

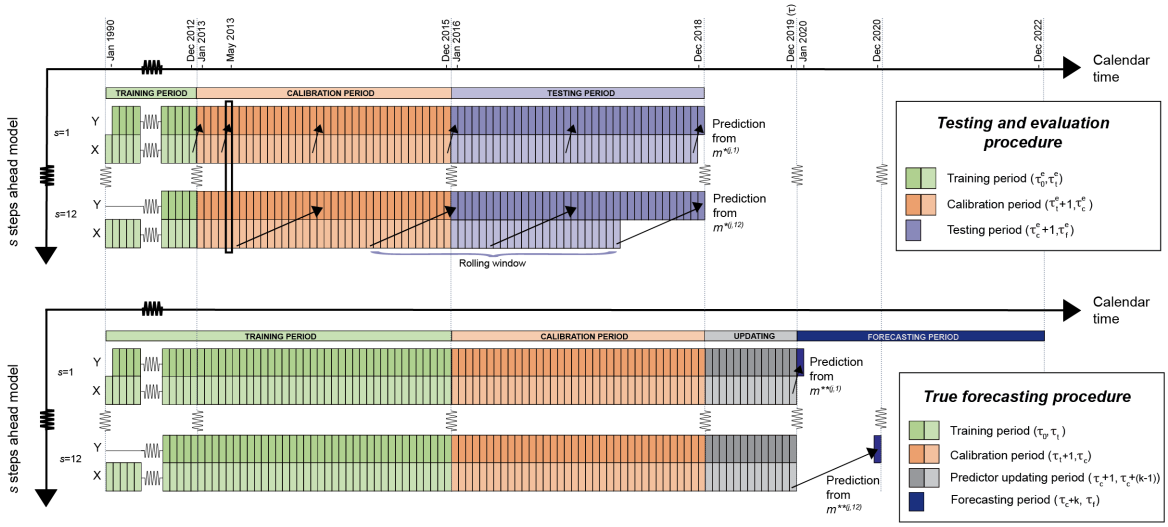
Håvard Hegre[1,2], Curtis Bell[1,3], Michael Colaresi[1,4], Mihai Croicu[1], Frederick Hoyles[1], Remco Jansen[1], Maxine Ria Leis[1], Angelica Lindqvist-McGowan[1], David Randahl[1], Espen Geelmuyden Rød[1], and Paola Vesco[1]

[1]Department of Peace and Conflict Research, Uppsala University
[2]Peace Research Institute Oslo
[3]U.S. Naval War College
[4]University of Pittsburgh

January 21, 2021

Abstract

This appendix documents the levels of analysis and dependent variables used in ViEWS, data partitioning, ensemble modeling, downsampling, handling of missing observations, and other data management procedures.



CONTENTS

A-1 LEVELS OF ANALYSIS AND DEPENDENT VARIABLES

A-1.1 Levels of analysis

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

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

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

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

A-1.2 Dependent variables

ViEWS generates predictions for the three forms of organized violence coded by the UCDP (Melander, Pettersson, and Themnér, 2016): state-based conflict (**sb**), one-sided violence against civilians (**os**), and non-state conflict (**ns**).[1] We continue to generate predictions at the country-month (*cm*) and PRIO-GRID-month (*pgm*) levels for each of these three forms of violence.

In Hegre et al., 2019, we defined the outcomes to be predicted as violence of one of these forms leading to at least one battle-related death (BRD) per month at both the PRIO-GRID and country (Tollefsen, Strand, and Buhaug, 2012) levels. This setup remains for the *pgm* level. At the *cm* level, we however now define the outcome we predict as at least 25 battle-related deaths in a given country-month. This threshold yields more relevant warnings than the single-death one. Even though every death from organized violence should be avoided, these remaining cases are more homogeneous and have much graver consequences than more isolated violent events. In the case of state-based conflict, the UCDP recorded violence with at least 25 deaths in about 11% of the country-months in Africa in the 2016–18 period. This rate is about half of the frequency of country-months with at least one death (see Table A-1).

Conflict data are primarily obtained from UCDP-GED and take the form of events (Sundberg and Melander, 2013). Historical data covering 1989–2018 are extracted from the UCDP-GED version 19.1 (Högbladh, 2019; Sundberg and Melander, 2013).[2] Newer data are provided by the new UCDP-Candidate dataset which is updated monthly (see Hegre et al., 2020, for an introduction). This allows use of conflict event data up to

[1] See Melander, Pettersson, and Themnér (2016) and `https://www.pcr.uu.se/research/ucdp/definitions/` for detailed definitions.

[2] The UCDP-GED raw data are publicly available through the UCDP-GED API (Croicu and Sundberg, 2013). ViEWS automatically retrieves these data from the API each month and aggregates to the ViEWS units of analysis. Usage of the API is described at `http://ucdp.uu.se/apidocs/`; the data are available as version 19.1 (1989–2018).

one month before the forecasting window. Here, we use UCDP-GED data up until and including December 2018, and UCDP-Candidate data up until December 2019. We aggregate all UCDP events up to our two levels of analysis.

The dependent variables: descriptive statistics

The evaluation metrics we use are dependent on the data they are evaluated for. In particular, they are all in varying ways dependent on class balance. Table A-1 shows descriptive statistics for the various dependent variables we use.

Table A-1. Descriptive statistics of dependent variables 2013–15 and 2016–18..

Variable	Mean, 2013–15, *cm*	Mean, 2016–18, *cm*	Mean, 2013–15, *pgm*	Mean, 2016–18, *pgm*
State-based (**sb**), ≥ 1 BRDs	0.209	0.240	0.000643	0.000572
State-based (**sb**), ≥ 25 BRDs	0.096	0.110		
One-sided (**os**), ≥ 1 BRDs	0.151	0.222	0.000528	0.000232
One-sided (**os**), ≥ 25 BRDs	0.062	0.061		
Non-state (**ns**), ≥ 1 BRDs	0.105	0.146	0.000349	0.000356
Non-state (**ns**), ≥ 25 BRDs	0.061	0.073		

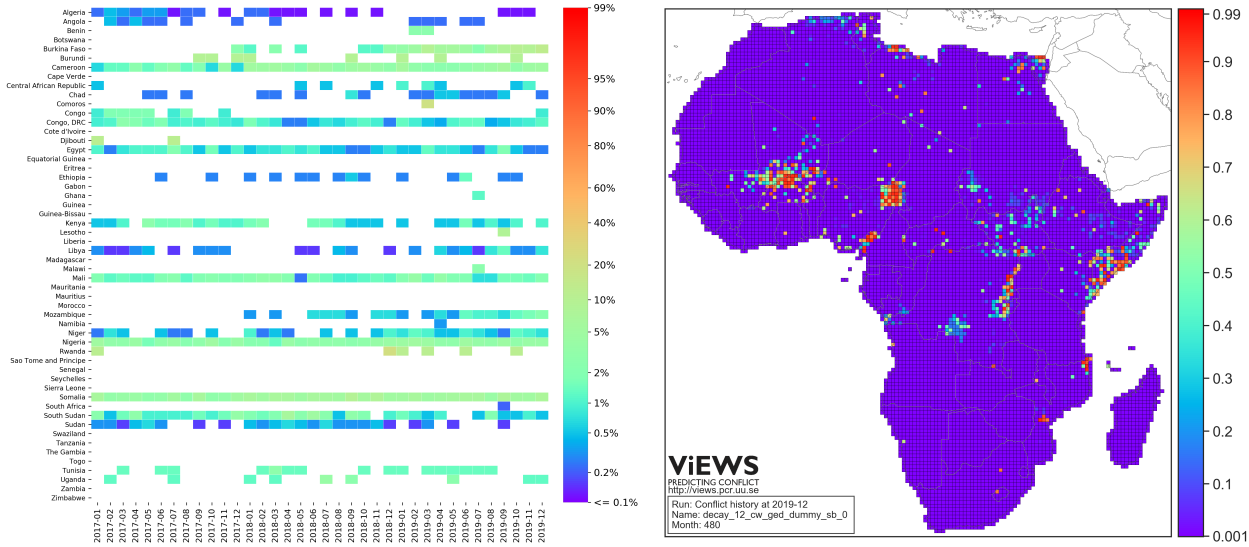
The UCDP records violence in a sizeable proportion of the country months (*cm*). When we use at least 1 BRD per country-month as the threshold, as many as 20.9% of the country-months had state-based conflict in 2013–15, and 24% in 2016–18. The UCDP records violence somewhat less frequently for the one-sided (**os**) and non-state (**ns**) political violence categories. From 2020, we have changed our setup to use of at least 25 battle-related deaths as our threshold at the *cm* level. Seen from Table A-1, this reduces the proportion of country-months where we observe conflict to just under half of the figures for the single-death threshold.

At the *pgm* level, our dataset has a very strong class imbalance. About 0.06% of PRIO-GRID-months had state-based conflict events in the two periods, between 0.02 and 0.05% had one-sided violence, and about 0.035% had non-state conflict events. We therefore continue to use the single-death threshold for the much smaller PRIO-GRID unit of analysis, and also include models trained on the lower threshold in our model ensemble.

The dependent variables: recent history

Figure A-1 summarizes the most recent observations for the **sb** outcome variables. Figure A-1a shows the proportion of PRIO-GRID cells with at least one event for each month in the 2017–2019 period. Figure A-1b depicts the recent history of violence in each PRIO-GRID cell. Red cells had conflict in November 2019, and purple ones have not seen conflict in many years. More precisely, the variable that is mapped is $z = e^{-m/h}$ where m is the number of months since last month with at least one fatality and h a halflife parameter—the number of months it takes before z 'decays' to half the value. Figures A-2 and A-3 show the same for the other two outcome variables.

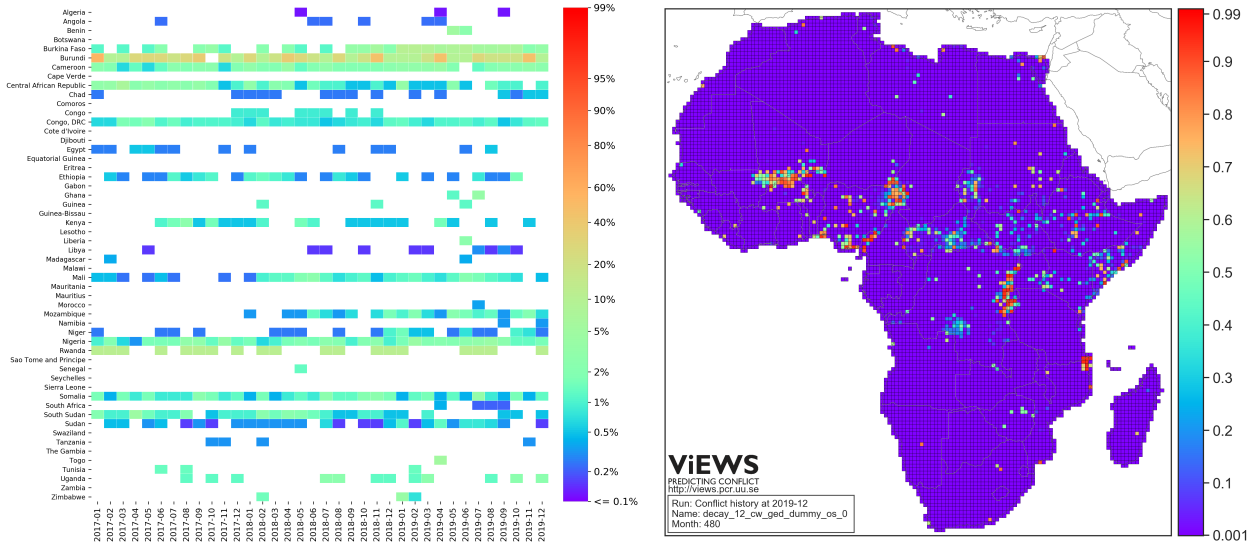
Figure A-1. State-based conflict (**sb**), *pgm* level, as recorded in Sundberg and Melander (2013) and Hegre et al. (2020).



(a) Observed proportion of PRIO-GRID cells with conflict, by country and month, 2017–2019

(b) Decay function of time since most recent event up to December 2019, halflife 12 months

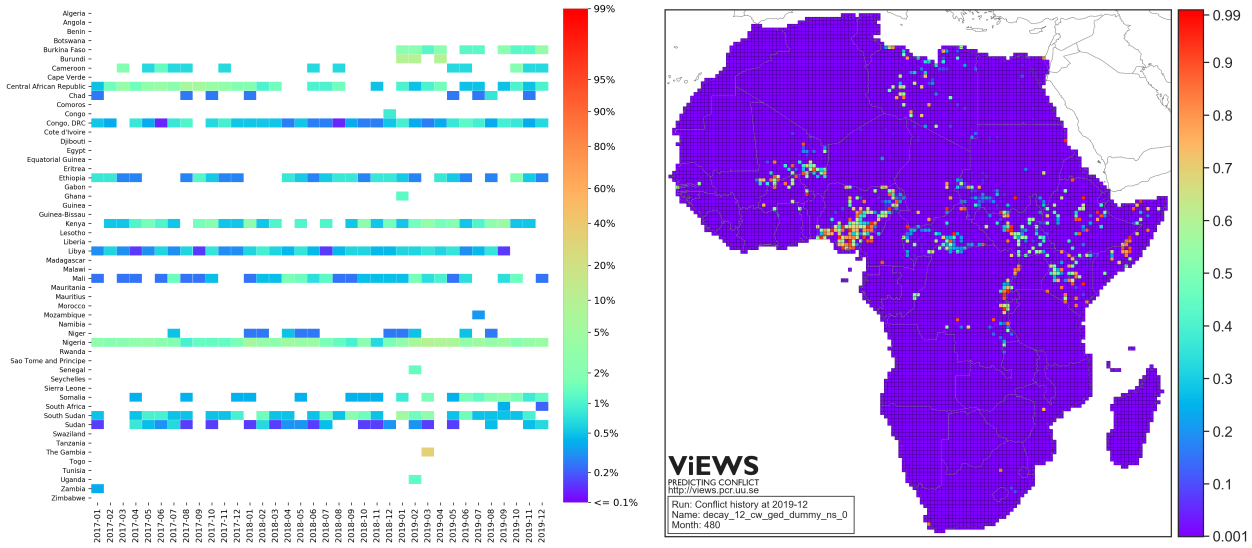
Figure A-2. One-sided violence (**os**), *pgm* level, as recorded in Sundberg and Melander (2013) and Hegre et al. (2020).



(a) Observed proportion of PRIO-GRID cells with conflict, by country and month, 2017–2019

(b) Decay function of time since most recent event up to December 2019, halflife 12 months

Figure A-3. Non-state conflict (**ns**), *pgm* level, as recorded in Sundberg and Melander (2013) and Hegre et al. (2020).



(a) Observed proportion of PRIO-GRID cells with conflict, by country and month, 2017–2019

(b) Decay function of time since most recent event up to December 2019, halflife 12 months

A-2 DATA PARTITIONING AND EVALUATION

In this section, we expand on the presentation in the main article of our procedures for partitioning data and for calibration, model weighting, and out-of-sample evaluation.

A-2.1 Data partitioning

As in Hegre et al., 2019, we split all of the available data into three partitions. In our notation, the time periods for these partitions are defined based on the time stamps for the observed outcomes Y. Table A-2 summarizes how we currently split our data. Internally and in online documentation, we refer to the month December 2019 as month 480, and will use the numeric month id in examples below.[3]

τ refers to calendar time,[4] but we add subscripts to identify when the partitions start and end. Because the partitions differ between evaluation and true forecasting, we have also added the superscript ^{e} to all notations of our evaluation partitions (see Table A-2).

The first relevant partition is the *training period* from τ_0^e, the first month with data, to τ_t^e (τ_0 to τ_t in the forecasting periodization). We use the notation $Y_{\tau_0:\tau_t^e}$ to define the beginning and ending time-points for the labeled instances in the training data.

The second set of observations is defined as the *calibration period* and is bounded by the cut-points τ_t^e+1 and τ_c^e (τ_t+1 and τ_c in the forecasting periodization), and is represented as $Y_{(\tau_t^e+1):\tau_c^e}$.

A third set is added only to the forecasting periodization. The UCDP releases final, carefully vetted events data annually—the most recent version covers 1989–2018 (Pettersson, Högbladh, and Öberg, 2019). To ensure maximal consistency, we use only final events data for training, weighting, and calibration, so we set the calibration to the 2016–2018 period. However, the UCDP has provided 'candidate data' for 2019 for Africa at a monthly update schedule (Hegre et al., 2020). We make use of these updated data when computing forecasts, in order to allow data input up until the month before the true forecasting period. We call this the *predictor updating period*. It starts immediately upon the end of the calibration period, at τ_c+1, and runs up

[3]The ViEWS month id is a counter that started on 1 in January 1980.

[4]This is the reference point from which the project is operating. We are assuming we can observe values less than τ, but not values greater than τ, when computing models.

until the first month of true forecasts at $s = 1$, to $\tau_c + (k - 1)$. $k - 1$ thus represents the months from which we rely on UCDP-Candidate data as opposed to UCDP-GED data.

The last set is the *testing/forecasting period*, extending from $\tau_c^e + 1$ to τ_f^e in the evaluation periodization, and from $\tau_c + k$ to τ_f in the forecasting periodization.

Table A-2. Partitioning of data for estimating model weights, hyper-parameter tuning, evaluation, and forecasting

	Periodization	
	Evaluation	Forecast
Training period	$\tau_0^e = 121$ (January 1990) $\tau_t^e = 396$ (December 2012)	$\tau_0 = 121$ (January 1990) $\tau_t = 432$ (December 2015)
Calibration period	$\tau_t^e + 1 = 397$ (January 2013) $\tau_c^e = 432$ (December 2015)	$\tau_t + 1 = 433$ (January 2016) $\tau_c = 468$ (December 2018)
Predictor updating period	n/a n/a	$\tau_c + 1 = 469$ (January 2019) $\tau_c + (k - 1) = 480$ (December 2019)
Testing/forecasting period	$\tau_c^e + 1 = 433$ (January 2016) $\tau_f^e = 468$ (December 2018)	$\tau_c + k = 481$ (January 2020) $\tau_f = 516$ (December 2022)

The 'forecast' periodization is for actual forecasting, the 'evaluation' periodization for testing models and ensembles. We use the training periods to train models and the calibration periods for hyper-parameter tuning and estimating model weights. After calibration, EBMA, and hyper-parameter tuning, we retrain our models using both the training and calibration partitions.

A-2.2 Procedures for calibration, model weighting, and out-of-sample evaluation

All ViEWS models are subjected to a careful out-of-sample evaluation of predictive performance. We have developed a setup to maximize the precision of this evaluation, using available data as efficiently as possible.

In the following, we will refer to a specification as a 'model' m^j. When we use input data up to December 2015, we generate forecasts for each of the 36 months from January 2016 to December 2018. We refer to these as steps $s \in [1, 36]$. We train each of these models specifically for each s. Model $m^{(j,1)}$ is trained to predict $s = 1$ month into the future, $m^{(j,6)}$ $s = 6$ months forward, and so on. Our procedure can be summarized as follows:

For each model $m^{(j)}$ and step $s \in [1, 36]$ in the evaluation periodization we do the following:

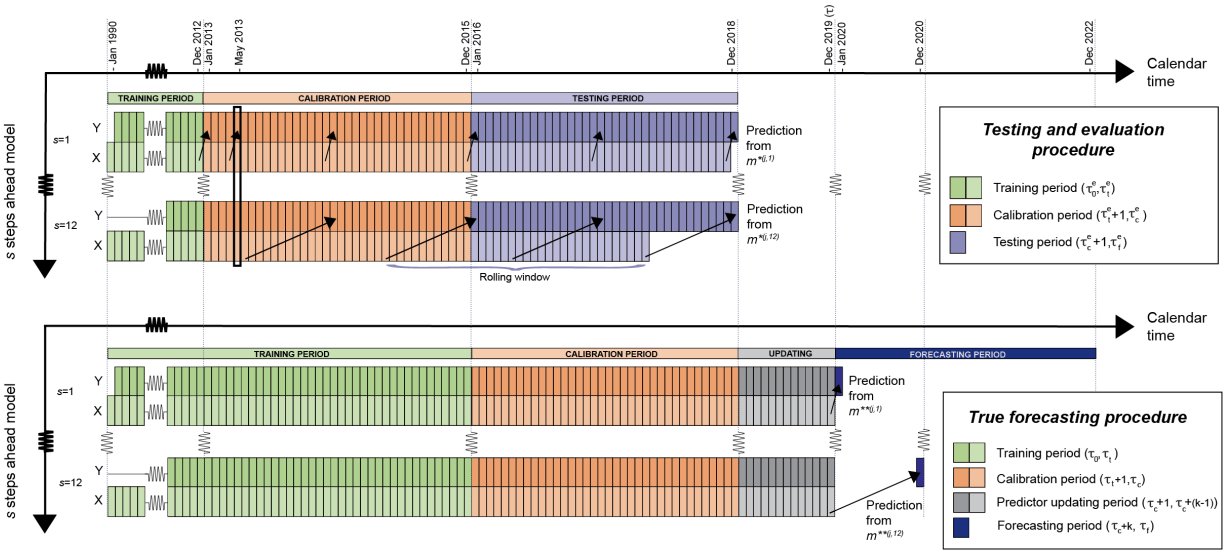
1. Train model $m^{(j,s)}$ on monthly data from τ_0^e to τ_t^e
2. Generate predictions from $m^{(j,s)}$ trained in (1) for all months i in the calibration period ($\tau_t^e + 1, \tau_c^e$), using data up to s months before i
3. Calibrate models, obtain ensemble weights, and tune hyper-parameters using the predictions from (2) along with the actuals for all months in the calibration period
4. Retrain model $m^{(j,s)}$ using both the training and calibration periods (τ_0^e, τ_c^e)
5. Generate predictions for the testing/forecasting period ($\tau_c^e + 1, \tau_f^e$) from the new model $m^{(j,s)}$ retrained in (4)

In step 5, the procedure is different when we evaluate models than when we generate true forecasts. The two variants of the procedure are summarized in Figure A-4. The horizontal bar labeled $s = 1$ indicates how we produce forecasts one month into the future. The bars are divided into three blocks, each of which consists of a number of rectangles representing months. The start and end months τ of these periods, only numerically referenced in the figure, are specified in Table A-2.

Each rectangle in Figure A-4 is divided into data for the outcome (Y) and for predictors (X). The green block represents the training data partition, stretching from month τ_0 to month τ_t. The arrows show how our models use input data X available at a given month to learn the relationship to the outcome Y for the subsequent s-month(s) ahead.[5]

[5]Note that the first months i with observed outcomes Y cannot be related to any features since we do not have data before τ_0, illustrated by the missing rectangles in the green blocks in Figure A-4.

Figure A-4. Timeshifting and periodization in the current pipeline for testing (top) and true forecasting (bottom).



The gray sections in the lower figure display the gap between the calibration and forecasting windows in the true forecasting procedure due to missing UCDP-GED data post-2018. UCDP-GED data are compiled on an annual basis, giving rise to this gap. The first dark blue rectangle appearing in the forecasting set of the bottom figure is month $\tau + 1$, here January 2020 (481).

When *evaluating* the models, we generate predictions for each month in the testing period, just as in step 2. We then match $s = 2$ forecasts for January 2016 (based on input data up to November 2015) with what actually happened in January 2016, $s = 2$ forecasts for February 2016 with actuals for February 2016, etc, for all s. This means that a model $m^{(j,2)}$ targeting $s = 2$ is evaluated against all 36 months in our calibration or testing period.

When we generate *true forecasts*, however, we only make forecasts based on the most recent input data. For the forecasts presented below, we have data up to December 2019. We make one set of forecasts at $s = 1$ for January 2020, one at $s = 2$ for February 2020, etc.

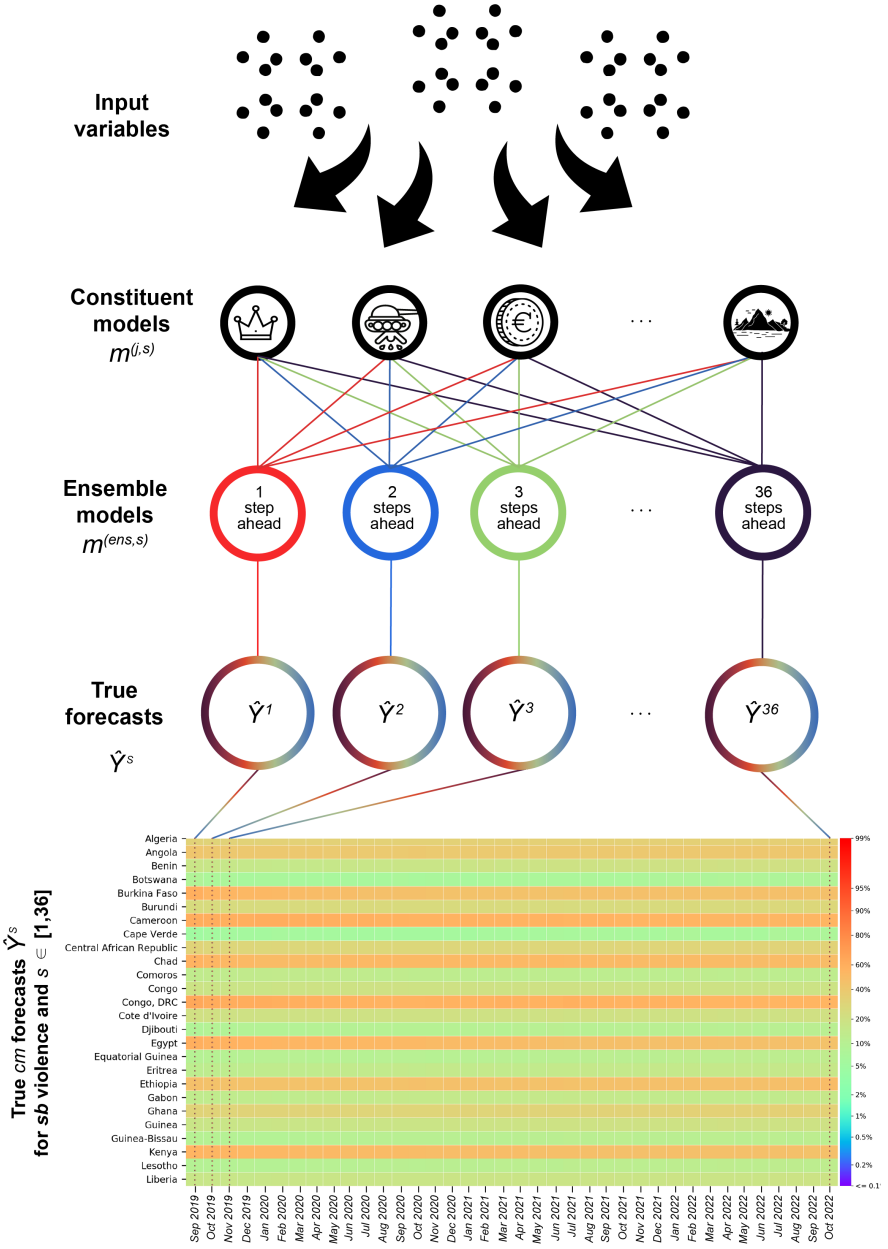
In Hegre et al. (2019), we used the procedure we now use for forecasts also when evaluating, calibrating, and estimating model weights. The new setup provides us with much more data for testing and calibration as we are able to reuse multiple times the set of actuals Y for a given month, e.g. month 401 (May 2013), marked off in Figure A-4.[6]

The changes do not affect training, but have other important benefits. Most importantly, using data for actual conflicts for all months in the calibration period for each step s means we are now able to estimate ensemble weights specifically for each s. We also have more data for hyper-parameter tuning, allowing us to introduce new algorithms. In addition, our evaluation of individual models in the ensemble yields more precise results, as we can allow similar model specifications to perform differently for different s. We may now capture that some models are more important for forecasting the immediate future and others for the more distant ones.

Figure A-5 summarizes the data handling procedures for the forecasting application.

[6] E.g., when evaluating, a $s = 1$ model prediction based on data up to and including April 2013 can be compared to the the actuals of May 2013, as can a step-12 forecast based on data up to May 2012.

Figure A-5. Simplified illustration of ViEWS' true forecasting procedure, exemplified for **sb** violence at the *cm* level.



The solid color connector lines (red, blue, green and purple) between the constituent and ensemble models in the figure above show how each step-specific ensemble model makes use of step-specific features from relevant constituent models in order to produce its forecasts. What constituent models are relevant for a given ensemble depends on the combination of type of violence (**sb, ns, os**) and unit of analysis (*pgm, cm*) that it is to forecast. The constituent models are in turn informed by a selection of themed input variables, relating e.g. to REIGN data (`https://www.oefresearch.org/datasets/reign`), conflict history, GDP, or geography (as pictured above). At the bottom of the figure, we see an example of predictions as compiled for a subset of countries for the period September 2019 to October 2022, as predicted by our old ensemble in September 2019. This particular heatmap thus displays the predicted probabilities of at least 1 battle-related death occurring as a result of **sb** violence at the *cm* level for all steps $s \in [1, 36]$.

A-3 COMBINING THREE TYPES OF POLITICAL VIOLENCE OUTCOMES

The figures depicting our forecasts (Figure 7 in the main article, and additional figures in Appendix F), are relatively similar—locations with a high predicted probability of one type of violence also have a high risk of the other two. This partly reflects that the various forms of organized violence occur through related processes and are constrained by the same factors. In addition, there is considerable spillover between the forms of violence. One-sided violence, for instance, is most frequently perpetrated in the context of a state-based conflict. We 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.

A-4 DOWNSAMPLING AND CALIBRATION

A-4.1 Downsampling

A majority of the models in ViEWS were trained on all available observations. Some of our random forests, however, were trained on a downsampled dataset. When downsampling, we keep all conflict events and randomly sample a proportion of the remaining observations. This serves two purposes.

First, it reduces the computational burden. 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 the computationally intensive random forest models using this data, 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.

Second, downsampling is a way of inducing an asymmetric cost-function into our computation. Assuming that 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, even at the cost of over-estimating some violence in peaceful circumstances. By reducing the proportion of non-events, we also reduce 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, then 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).

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

A-4.2 Calibration

Forecasting requires that each model is well calibrated: 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 a downsampled dataset do not have this property, and require calibration. The same applies to the models that are constructed as the product of *cm* and *pgm* probabilities. We therefore calibrate the constituent models before entering them into the ensemble.

We use the calibration partition to calibrate the models. We obtain re-centering and re-scaling parameters γ_{0i}, γ_{1i} by estimating logistic regression models for each constituent model on the calibration period:

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

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

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

where $\hat{p}_{cal}(Y_v^c = 1)$ is our calibrated predicted probability of conflict.

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 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 under-predicts or over-predicts. We show such calibration plots for our new and old *cm* ensembles in Figure D-7 in Appendix D.

A-5 HANDLING MISSING OR INCOMPLETE DATA IN VIEWS

A-5.1 Dependent variables

UCDP-GED includes high-resolution temporal and geographical references. In about 15% of the cases, however, UCDP has 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 for multiple imputation of their locations, as documented in Croicu and Hegre (2018). This method employs the locations of precisely known events within the same conflict and within close temporal proximity to determine an empirical spatial probability distribution of latent conflict propensity for each uncertain event. We then sample this empirical probability distribution multiple times. All the forecasts reported by ViEWS are based on a set of five imputed location datasets. Croicu and Hegre (2018) show that this improves the predictive performance of the system considerably.

A-5.2 Predictors

Some variables in the predictor data used by the ViEWS project contains a certain amount of missing data. This is problematic for several reason. Firstly, making predictions require the data to be complete, i.e. all values for the variables we use to make predictions must be known. Secondly, not using appropriate methods for handling the missing data in the training models risk creating biased parameter estimates and/or standard errors (Allison, 2009) which may have an adverse affect on the predictive capabilities of the models.

Depending on the mechanisms for how the missingness appear, different missing data methods have different advantages. The most common method for handling missing data is to simply remove all obser-

vations which are not complete in *listwise deletion* (Lall, 2016). Listwise deletion does, however, assume that the data are missing completely at random, i.e. that the reason for the missingness is independent both of the value of the variable itself *and* independent of the values of the other variables in the model. In social sciences this is most often not a reasonable assumption. If it does not hold, the results from the analysis will be biased (Allison, 2009; Graham, 2009; Schafer and Graham, 2002).

An alternative method for handling missing data which does not require this assumption is imputation. Here, the missing data are replaced (imputed) with some plausible values. Several possible imputation methods exist. Among the more naive are to use either the mean or the predicted values from a linear regression of the variable on all other variables. Mean and regression imputation do, however, also bias the results unless some strong assumptions are fulfilled (Buuren, 2012). The solution to these problems is to use multiple imputation.[7]

The ViEWS project uses Multiple Imputation with the Amelia II package in R to replace the missing data. Amelia II uses bootstrapping and the expectation-maximization (EM) algorithm to impute each missing value m times to create m complete datasets. The number of imputations, m, affects a number of statistical quantities, including power and efficiency. To achieve reasonable statistical efficiency, as few as five complete imputed datasets are needed. A higher number of imputations do, however, lead to both higher efficiency and higher power and precision (see for instance Graham, Olchowski, and Gilreath, 2007; White, Royston, and Wood, 2011). As imputation is computationally intensive in large datasets, and the ViEWS project due to its focus on prediction is less concerned with statistical power, five imputations are currently used. Our dynamic simulation procedure uses all five datasets simultaneously and the results are aggregated using the Rubin Rules (Rubin, 1987). In the one-step-ahead forecasts, only one imputed dataset is used at this time. Using one dataset instead of five will not bias the results, but will reduce the statistical efficiency (Buuren, 2012).

The ViEWS project will conduct further tests on how different missing data techniques affect prediction and simulation, in order to create a best practice for missing data in predictive studies.

A-6 DATA MANAGEMENT

Data are currently stored in a large, highly normalized (3NF) Postgres database to avoid redundancy and ensure consistency. Each unit of analysis (*pgm, cm* and corresponding units for countries, country years, PRIO-GRID cells and PRIO-GRID years) 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 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.

[7] For a more comprehensive test of missing data methods, see Randahl (2016).

REFERENCES

Allison, Paul D. (2009). "Missing Data". In: *The SAGE handbook of quantitative methods in psychology*. Ed. by Roger E. Millsap and Alberto Maydeu-Olivares. Sage Publications.

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

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

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/.

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.

Graham, John W (2009). "Missing data analysis: Making it work in the real world". In: *Annual review of psychology* 60, pp. 549–576.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath (2007). "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory". In: *Prevention Science* 8.3, pp. 206–213.

Hegre, Håvard, Marie Allansson, Matthias Basedau, Mike Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayeed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby (2019). "ViEWS: A political Violence Early Warning System". In: *Journal of Peace Research* 56.2, pp. 155–174. URL: https://doi.org/10.1177/0022343319823860.

Hegre, Håvard, Mihai Croicu, Kristine Eck, and Stina Högbladh (2020). "Introducing the UCDP Candidate Events Dataset". In: *Research and Politics* in press.

Högbladh, Stina (2019). *UCDP GED Codebook version 19.1*. Department of Peace and Conflict Research, Uppsala University.

Lall, Ranjit (2016). "How Multiple Imputation Makes a Difference". en. In: *Political Analysis* 24.4, pp. 414–433. DOI: 10.1093/pan/mpw020.

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

Pettersson, Therése, Stina Högbladh, and Magnus Öberg (2019). "Organized violence, 1989–2018 and peace agreements". In: *Journal of Peace Research* 56.4, pp. 589–603. DOI: 10.1177/0022343319856046. URL: https://doi.org/10.1177/0022343319856046.

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 Garcıa and Edgar Rangel (2003). "Strategies for learning in class imbalance problems". In: *Pattern Recognition* 36.3, pp. 849–851.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, Joseph L. and John W. Graham (2002). "Missing data: Our view of the state of the art." In: *Psychological Methods* 7.2, pp. 147–177. DOI: 10.1037//1082-989X.7.2.147. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.7.2.147 (visited on 04/27/2015).

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.

Tollefsen, Andreas Forø (2012). *PRIO-GRID Codebook*. Typescript, PRIO. URL: http://file.prio.no/ReplicationData/PRIO-GRID/PRIO-GRID_codebook_v1_01.pdf.

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.

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.

White, Ian R., Patrick Royston, and Angela M. Wood (Feb. 2011). "Multiple imputation using chained equations: Issues and guidance for practice". en. In: *Statistics in Medicine* 30.4, pp. 377–399. DOI: 10.1002/sim.4067. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4067 (visited on 10/08/2018).

FUNDING

ViEWS is funded by the European Research Council, grant number 694640 and Uppsala University.

European Research Council
Established by the European Commission

COLLABORATIONS

ViEWS has an active interaction with other projects, including CLIMSEC, CAVE and CROP at PRIO (https://prio.org/), the MISTRA Geopolitics project, and most importantly the Uppsala Conflict Data Program (https://ucdp.uu.se/) at Uppsala University.

CODEBASE & PUBLICATIONS

ViEWS' codebase is available at:

https://github.com/UppsalaConflictDataProgram/OpenViEWS2

The full pool of publications are accessible at:

https://pcr.uu.se/research/views/publications/