The 2023/24 VIEWS Prediction competition
Predicting the number of fatalities in armed conflict, with uncertainty[*]

Håvard Hegre[1, 2, 4], Paola Vesco[1, 2, 4], Michael Colaresi[2, 3, 4], and Jonas Vestby[1, 4]

[1]Peace Research Institute Oslo (PRIO)
[2]Department of Peace and Conflict Research, Uppsala University
[3]University of Pittsburgh
[4]Centre for Advanced Study, Oslo

v1.2 April 18, 2023
Source code for this document found at
`https://www.overleaf.com/project/63579fd45bcb4908c2af3632`

Abstract

This note outlines a prediction competition where the target is to forecast the number of fatalities in armed conflicts, in the form of the UCDP 'best' estimates, aggregated to the VIEWS units of analysis. The note presents the format of the contributions, the evaluation metric, and the procedures. More information on the competition, and all data referred to in this document, can be found at `https://viewsforecasting.org/prediction-competition-2/`.

[*] The research was funded by the Uppsala University, the European Research Council, project ERC-AdG 101055176 (ANTICIPATE), Riksbankens Jubileumsfond (Societies at Risk), and ..... For more information on the project see `viewsforecasting.org`.

1 Invitation and motivation

The VIEWS team invites research teams interested in forecasting models in general and in the prediction of armed conflict in particular, to take part in a prediction competition where all contributors work on a common, well-defined challenge:

- Predict the number of fatalities in armed conflict, as reported by the Uppsala Conflict Data Program (UCDP). Details on the data collection are provided in Section 2.1.
- With estimates of the uncertainty of the predictions calculated in the form of samples of forecasted values 2.3.
- Rewarding those that do well both in terms of point prediction and uncertainty, or more precisely in terms of *calibration*, *sharpness*, *focus*, *nearness*, and *propriety.*

1.1 Use case

Since the World Bank Group and United Nations (2017) report calling for 'early warning-early action' procedures, armed conflict forecasts have been in increasing demand within IGOs like the UN or within governments. Several such organizations are developing early-warning systems, or including forecasts from systems like VIEWS in their 'dashboards' supporting decision making. To the extent that the forecasts are sufficiently precise, they are used either to stimulate efforts to prevent conflict escalation, or, more realistically, to support efforts to mitigate their consequences. Typically, users are both interested in the most likely outcome (a point prediction), but also in the lower-probability risk that conflicts escalate catastrophically (the tail ends of the probability distribution). Users are also interested in knowing how uncertain the forecasts are. Setting the communication challenges aside for this context, predictions of war intensity as probability distributions comes closer to what user groups need than point estimates or simple dichotomous classification models.

1.2 Technical motivation

The competition builds on the predecessor VIEWS prediction competition (Hegre, Vesco, and Colaresi, 2022; Vesco et al., 2022), where the challenge was to predict *change* in the number of conflict fatalities. The previous competition taught us some valuable lessons on the value of forecasting conflict fatalities, while raising some important limitations and challenges. It was clear that complex models based on sophisticated algorithms and leveraging big data are the best individual tool to predict changes in fatalities – although they tend to be difficult to interpret (Vesco et al., 2022). Even very sophisticated machine learning models, however, tend to be surprised by the outbreak of conflicts in previously peaceful locations: most (but not all) of the models in fact are beaten, on common forecast evaluation scores, by a basic 'no-change' model that constantly predicts a null change in fatalities.

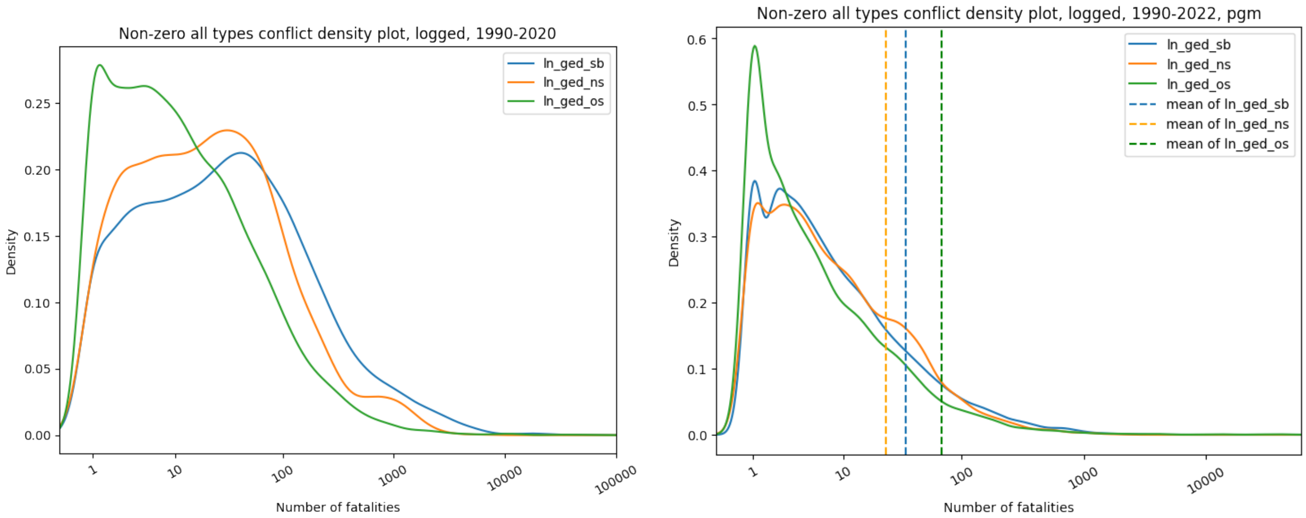

Figure 1. Distribution of observed fatalities in December 2022 at our two levels of analysis, *cm* (left) and *pgm* (right)

These findings suggest that more research is needed to improve our collective ability to forecast conflict outbreaks and (de-)escalation, but also calls for better evaluation metrics to more meaningfully assess each model's accuracy as well as the performance of combined ensembles of models (see below). Point-estimate predictions – which rely on a single measure of the predicted outcome – are difficult to evaluate in a way that creates new insights. One important disconnect between a point estimate and understanding what is missing from or well represented by a given model is the fact that point estimates, by definition, do not encode the inherent uncertainty in the distribution of conflict fatalities in a given instance. Point-estimate predictions make it more difficult to reveal mis-calibration as well as obscure the relation between the shape and location of the predictive distribution and the underlying data generation process. This divergence might be most pronounced for processes that produce skewed distributions of target observations, such as in conflict fatalities. Figure 1 demonstrates the skewness of the (non-zero) fatality counts at the two VIEWS levels of analysis – even when plotting the counts on a log axis, the distributions have a distinct skew. To add to this, 87% of all the observations at the 'country-month' (*cm*) level are zeros, and as many as 99% at the 'PRIO-GRID-month' (*pgm*) level.

At the core of this problem is that learning the expected mean of the fatality count distribution – what evaluation metrics such as the Mean Squared Error (MSE) favors – may not be of preeminent relevance. As argued by Tillman Gneiting, '[s]ingle-valued point forecasts... can lead to grossly misguided inferences, unless the scoring function and the forecasting task are carefully matched' (Gneiting, 2011, p. 746).[1]

For an outcome like ours, with its zero-inflated and right-skewed distribution, the mean is not

[1] Gneiting (2011) further argues that '[e]ffective point forecasting requires that the scoring function be specified a priori, or that the forecaster receives a directive in the form of a statistical functional', and that 'it is critical that the scoring function [is] consistent for [the functional], in the sense that the expected score is minimized when following the directive' (Ibid.). What this means is that if we want to use single-valued point forecasts, we have to a priori decide on an elicitable target to summarise the predictive distribution, e.g., the mean. We can then find a scoring function that is consistent with this target. In the case of the mean, that would be the square error.

sufficiently informative to balance the risks between zero and extremely high values.[2] Accordingly, we want broader criteria for our evaluation – not only a certain target representation of the predictive distribution (such as the mean), but a more general characterization of how well the predictive distribution fits observed outcomes.[3] If this is the case, then there is a need to move from single-valued point forecasts to probabilistic forecasts, where the goal is 'to maximize the sharpness of the predictive distributions subject to calibration' (Gneiting and Raftery, 2007, p. 359). We will also consider other criteria that can define useful predictive distributions relative to conflict fatalities.

For the competition to be useful for a wider range of researchers, policy-makers, and NGOs, we want to calculate different evaluation metrics for distinct use cases and depending on our goals. If we are only interested in 'one-shot forecasts' – how well the model does at predicting specific single events – we would mainly evaluate predictions at the part of the predictive distribution that is deemed the most likely outcome. If we are interested in performance over many tries, then calibration becomes more relevant. If we are only interested in extreme outcomes, then predictive performance at the tails of the predictive distribution becomes more relevant. Additionally, if the costs and benefits of prediction depend on the location of events in space and time relative to the predicted distribution, then we would want to calculate relative performance of models in that spatial-temporal context.

An important lesson of the past competition was therefore the need of going beyond point-estimate predictions, and account for the full distribution of the forecasts to better represent uncertainty, with evaluation metrics that capture different use cases, or different qualities as we call them below. This is what we seek to explore in this competition.

2 The challenge

2.1 Prediction target: Predicting the monthly number of fatalities from organized political violence

The competition consists of two parts with targets defined at two different geographical units, and contributors can choose to submit contributions at either the country level (*cm*) with global coverage, the sub-national level (PRIO-GRID-month, *pgm*) for Africa and the Middle East, or both. The temporal resolution at both the *cm* and *pgm* levels is the month. The two levels of analysis are the same as used in VIEWS (Hegre et al., 2019; Hegre et al., 2021) and in the former ViEWS prediction competition (Hegre, Vesco, and Colaresi, 2022; Vesco et al., 2022) and depicted in Figure 2. The two parts will have roughly the same structure, although some of the evaluation metrics may differ.

[2] There are infinitely many distributions with the same mean that assign different plausibilities to zero-observations and very large observations.

[3] We will define more clearly what we mean by *fit* below.

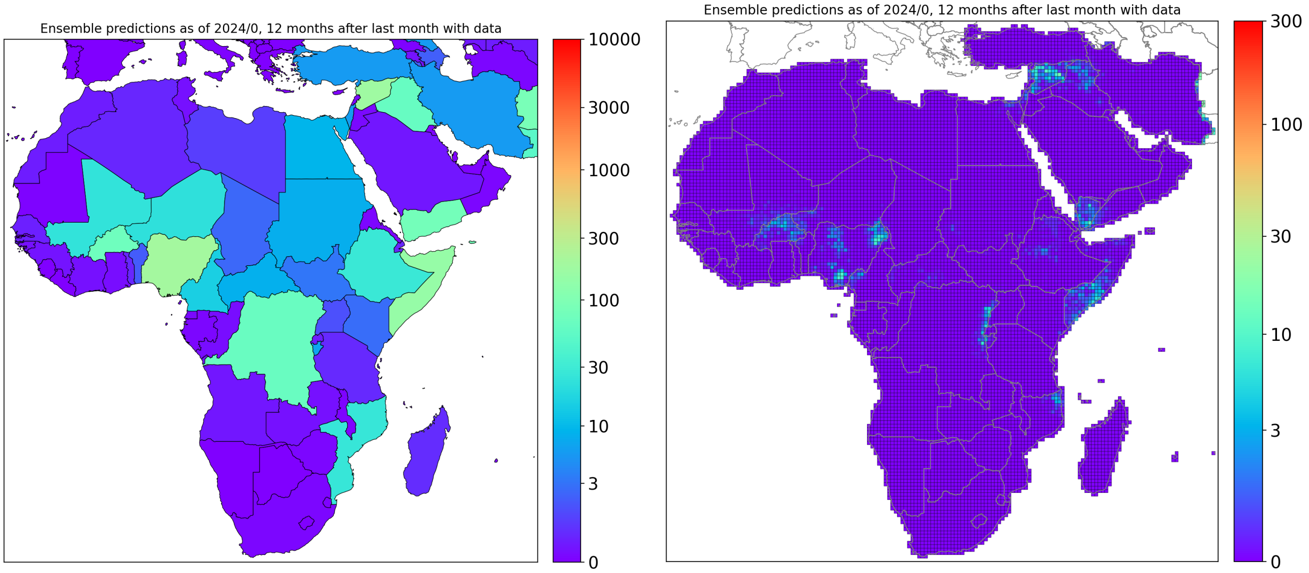
Ensemble predictions as of 2024/0, 12 months after last month with data
10000
3000
1000
300
100
30
10
3
0
Ensemble predictions as of 2024/0, 12 months after last month with data
300
100
30
10
3
0

Figure 2. ViEWS ensemble point predictions for January 2024, *cm* and *pgm* level

1+4 prediction windows

We seek contributions for two sets of prediction windows. The primary goal is to provide predictions for the true future – the full year of 2024, based on data up to and including October 2023. The predictions should be made separately for each of the twelve months in 2024.[4]

However, a single year of events at these aggregation levels will generate a limited amount of data for model evaluation. To complement the evaluation of the forecasts, we will also request contributions for each of the calendar years of 2018, 2019, 2020, and 2021, based on data up to and including October for the preceding year.

Five sets of input data will be made available to the participants, one for each of these forecasting windows.

State-based conflict

The prediction target will be the Uppsala Conflict Data Program (UCDP)'s coding of the number of fatalities in state-based armed conflict, as defined in Davies, Pettersson, and Öberg (2022). For the test prediction windows up to and including 2021, we have access to the final UCDP data as reported in that article. For the year of 2024, the UCDP Candidate data will be available (Hegre et al., 2020). Contributions should present predictions as fatality counts.[5]

[4] Contributors may decide whether they want to submit identical predictions for each month, or fine-tune predictions separately for each of them.

[5] In the Hegre, Vesco, and Colaresi (2022) competition, the target was specified in log form. There are some compelling reasons for making forecasts in the count form: Adding 1 to y to allow for (log) zeros is an arbitrary choice. Since the vast majority of cases are zero, the choice we apply will make a difference. We believe that evaluating models on the original, non-logged scale most likely helps rewarding models that are willing to inch up

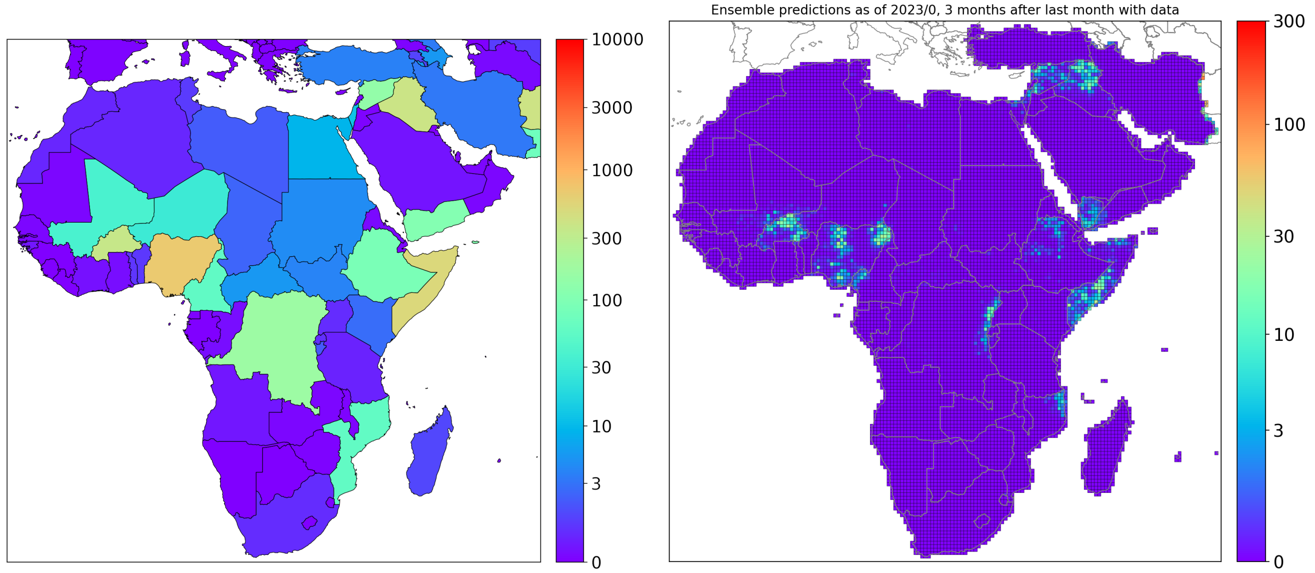
10000
3000
1000
300
100
30
10
3
0
Ensemble predictions as of 2023/0, 3 months after last month with data
300
100
30
10
3
0

Figure 3. Predicted fatalities in December 2022 at our two levels of analysis

The VIEWS levels of analysis

Forecasts can be made at one of the two VIEWS levels of analysis or both:

Country-months (Gleditsch and Ward, 1999, abbreviated *cm* in VIEWS), and sub-national geographical location months (*pgm*). The set of countries is defined by the Gleditsch-Ward country code (Gleditsch and Ward, 1999, with later updates), and the geographical extent of countries by the latest version of CShapes (Weidmann, Kuse, and Gleditsch, 2010). For the country level of analysis VIEWS data are global.

PRIO-GRID months (*pgm*), which rely 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. For the subnational level of analysis, we currently restrict forecasts to Africa and the Middle East. [6]

away from 0. We may still calculate some auxiliary evaluation metrics based on a log-transformed version.

[6] 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.

2.2 Timeline

We expect the following timeline for the project:

- Early March 2023: Distribute formal invitation to participate
- 1 May 2023: data, benchmarks, and code available to participants
- **Deadline 1**: 1 June 2023: Participants submit abstracts to organizers
- 1 July 2023: Decide which contributions to include in competition
- **Deadline 2**: 10-11 October 2023: Workshop for selected contributors. Contributors submit by 25 September:
    1. Preliminary forecasts
    2. Updated 200-word summary
    3. 1500-word summary papers presenting methods
- About 1 October 2023: Working version of webpage to monitor results
- 1 December: Organizers provide all participants with updated data up to and including 31 October 2023
- **Deadline 3**: 10 December: Contributors submit the final predictions for test window (2019–2022) and true future (January 2024 ($s = 3$)–December 2024 ($s = 14$))
- 20 January 2023: All predictions available on dedicated webpage
- 1 January 2024: Start of prediction window
- 31 December 2024: End of forecasting window
- 1 March 2025: Final draft of evaluation article, based on full forecasting window
- 1 May 2025: Online publication of evaluation article

2.3 What to contribute

The teams will submit the following at three deadlines:

- Abstract of contribution (Deadline 1)
- Short summary of model/contribution, basis for introduction article (Deadline 2)
- 3–5 page write-up of model/contribution, for workshop and as working paper (Deadline 2)

- Preliminary forecasts for the four test windows 2018–2021, using the corresponding training data I (Deadline 2)
- Final forecasts for test window 2018–2021 and forecasts for true future, using the training data being made available in late November (Deadline 3)
- Code for their models

Teams can submit predictions in one of two formats. For each test-set observation, a team can provide either (or both) of the following:

1. for each observation, indexed by either

   $$month\_id, country\_id, draw$$

   or

   $$priogrid\_id, month\_id, draw$$

   supply between 15 and 1000 samples (int32) from their predictive distribution for Y_{it} (where i is either *country_id* or *priogrid_id* and t is *month_id*). We will take these samples and calculate the necessary functions for each of our evaluation metrics. Contributions in the form of samples are convenient, as they will allow the scoring team to evaluate properly the tails of the distributions, and it is also practically easier to fit empirical CDFs/pdfs with samples when the contributed models can potentially generate arbitrary distributions of predictions (i.e. do not follow a specific distribution). See the benchmark models provided at `https://viewsforecasting.org/prediction-competition-2/` for examples.

2. for each observation i, t supply a point prediction for Y_{it}. We will generate a set of samples using this point predictions by assuming that the predictive distribution has a Poisson distribution with the point prediction provided i, t as mean and variance, and evaluate using the same functions as if we had received samples directly from the team. We offer this contribution format to prospective participants that choose to focus on other aspects of the modeling than the shape of the predictive distribution. The Poisson model probably understates the uncertainty of the predictions – contributors concerned with this should submit predictions as draws from their predictive distribution.

Contributors will submit their predictions as .parquet files[7] indexed by country or PRIO-GRID id, month id, and draw id. We will provide information regarding the technical submission solution at a later point in time.

Below, we specify the evaluation scores we will use, and, for the most important ones, the packages we will use.

[7] See Apache Parquet project, `https://parquet.apache.org`. There are packages and functions to read and write parquet files in all major open source computing languages commonly used for data science and computational social science work, including python, Julia, and R.

3 Data to provide to the contributors

The VIEWS team will make the following data available to the contributors, for the *cm* and *pgm* level.

By May 1 2023:

- Training data I, for the 1990–2016 period. About 100 predictors each at *cm* level and *pgm*
- Prediction data for the test period – the data needed to produce forecasts for 1–12 months forward, for all months in the period 2018–2021 (or 2019–2022)
- Benchmark models as probability distributions

By May 15 2023:

- Possibly: VIEWS constituent models (50 draws from each) for contributors interested in using them for ensemble models, for the 2014–2017 period

By 1 December 2023:

- Prediction data up to and including October 2023 – the data needed to produce forecasts for 1–12 months forward, for all months in 2024

4 Evaluation and metrics

The evaluation of the contributions will be done by a scoring committee consisting of 3–4 members from the forecasting expert community as well one or two representatives from the user community. The scoring committee will have technical assistance from the VIEWS team but will perform their evaluation independently. We will place approximately equal weight on the joint performance across the test period predictions and the predictions for 2024, the true future.

Here, we present the main metrics used to evaluate the contributions, our motivations for their relative importance, and discuss some practicalities of evaluation and power/data sparseness concerns.

We will use the following notation in what follows:

- $f(x)$: The forecast distribution/mass function

	Desirable qualities				
Rule	Calibration	Sharpness	Focus	Nearness	Propriety
CRPS	X	X	-	x	X
IGN	-	X	X	-	x
MIS	X	X	-	-	X
WaRPS	X	x	-	X	-

Table 1. Beneficial qualities of probabilistic forecasting systems and how they are assessed by core evaluation metrics. Large X-s mean the metric is highly useful for assessing the respective quality. Small x-s mean the metric captures the quality partially.[a]

[a] CRPS gets a little x for distance-sensitivity because it is sensitive to misses within observation bins (across values for an observation), but not across observations. WaRPS gets a little x for sharpness because of the fact that the observations arrive as heaps of mass means that sharp distributions are preferred (as it is based on RPS which is related to CRPS).

- $F(x)$: The forecast cumulative distribution function
- y: The observed value

As noted above, we will do the evaluation in terms of y as the non-logged count of fatalities, rather based on $ln(y+1)$ as in Hegre, Vesco, and Colaresi (2022). As shown in Figure 1, the distribution of the target variable is challenging, being zero-inflated and with heavy tails. The evaluation metrics we propose below should be well equipped to evaluate predictive distributions seeking to match this target.

Note that we will evaluate the predictions for 2024 against the VIEWS aggregations of the UCDP-Candidate data (Hegre et al., 2020), since these are the only data that will be available at that time. For the other test periods, we will evaluate against VIEWS aggregates of the final UCDP-GED data (Sundberg and Melander, 2013).

4.1 What to reward

The metrics are designed to reward five qualities of probabilistic forecasting systems – calibration, sharpness, focus, nearness, and propriety.

Table 1 summarizes these five and link them to the main metrics that we will be using.

Calibration A model is well calibrated when the predicted frequency of y-values corresponds to the observed frequency of y in new data (this is a joint function of $f(x)$ and y). For instance, if the model predicts a 30% probability of at least 100 deaths in 10 observations, is it the case in the observed data that approximately 3 out of those 10 suffered at least 100 deaths?

Sharpness Concentrated predictive distributions are preferred as they encode more information, as defined by Shannon (1948), across all possible values of y (this is a function of $f(x)$ only). A model that predicts with 80% probability that the true value is between 20 and 50 is preferred to one that specifies the 80% prediction interval to be between 10 and 100, everything else equal.

Focus refers to the aim that predictive distributions should provide peak(s)/high density regions that hold information on new data as it arrives – that is, high probability density for the events that materialize. This quality is a function of $f(x)$ evaluated at y. The whole predictive distribution is not as important, according to focus, which is also similar to locality, as what information is contained within the predictive distribution around the actual value observed. Some argue that evaluation of the full predictive distribution is not warranted, as 'when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value should be taken into account' (Bernardo 1979 p. 689, cited in Gneiting and Raftery, 2007, 365f.).

Nearness Predictive distributions should array the plausibility of future values near, based on a usefully applied geometry, to actual values. In our context, 'near' means close in time and space – predicting violence three months too early is better than predicting twelve months too early, and predicting violence 100 km from where it happens is better than 1,000 km off. This quality is called 'sensitivity to distance' by Gneiting and Raftery (2007) (but they mean it in a limited bin-by-bin sense) and is offered as a partial rejoinder to only relying on the log score, which ignores the 'nearness' of all the plausibility in the distribution that is not located at the observed value. (this is a function of $f(x)$, y, and an underlying geometry of distance in our conceptualization).

Propriety encourages the reporting of predictive distributions that represent the honest beliefs of the forecaster or model. Proper scoring rules accomplish this by ensuring that the maximization of the expected reward for the forecaster occurs when reporting their underlying beliefs and not bending the shape of those beliefs in a particular direction (Gneiting and Raftery., 2007; Czado, Gneiting, and Held, 2009). In contrast, an improper score might reward increased certainty or a shifted mode for the distribution to hedge relative to the true underlying beliefs. A survey of the use of proper scoring functions across different scientific domains can be found in (Carvalho, 2016), and a discussion with specific reference to count data is contained in Czado, Gneiting, and Held (2009).

4.2 Metrics

The scoring committee will consider the metrics below when evaluating the contributions. The main scoring will be done in terms of the CRPS. The other metrics will be used for secondary scoring and facilitate a richer discussion of model performance.

Main metric: Continuous Rank Probability Score (CRPS)

The continuous rank probability score is defined as:

$$CRPS(F,y) = \int_{\mathbb{R}} (F(x) - \mathbb{1}(x-y))^2 dx$$

where $\mathbb{1}(z)$ is the indicator function (or Heaviside step function), defined as

$$\mathbb{1}(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

One can think of the metric as a generalization of the Brier score to infinitely small bins. Broadly, CRPS is a generalization of the MAE for any predictive distribution: if CRPS is used to compare a 'point' prediction as a CDF with a point observation, it gives MAE.[8] Unlike the Ignorance Score (see below), CRPS takes the distance between the predicted and observed value into account.

Implementation: We will compute the measure using the `properscoring.crps_ensemble()` function/package as implemented in `xskillscore.crps_ensemble()`. We will weigh each sample forecast equally.

Secondary metric I: Log score/ignorance score

The log score (also called ignorance score) is the log of the predictive density evaluated at the actual observation:

$$IGN(f,y) = -log_2(f(y))$$

The ignorance score is the only proper local (i.e., only on the predictive density through its value at the event that materializes) scoring rule for continuous data. The ignorance score complements the CRPS by scoring the predicted probability of the observed event, instead of the distance between the predicted and observed. Therefore it emphasizes how much belief is focused at the observed value (see implementation details below however).

[8] For finite count data, the CRPS is really a Ranked Probability score (Czado, Gneiting, and Held, 2009; Kolassa, 2016). We will follow convention and refer to the metric as CRPS here as the integral does apply to the CDF of the counts as described above.

Implementation: The Ignorance Score can in practice be difficult to work with. The reason is that it not always straight-forward to elicit a predicted probability for any outcome. Since we base our computations on samples and not full probability distributions, we need some way to translate samples into probabilities. There are many ways to do this, however. Our preferred way is to bin the predicted and observed outcomes, and then add each of the possible binned outcomes to each forecast sample. E.g., if the binned forecast sample is [0,0,0,2], which represents three samples that take on the value of 0 and one sample that takes on the value of 2, and we have binned using three different categories (and not just the two that appear in this specific set of samples), the final forecast sample that we evaluate with is [0,0,0,2,0,1,2], where we have added 1 sample from categories 0, 1, and 2. This way, we will guarantee being able to elicit non-zero probabilities for each outcome. This approach is similar to assuming a naive uniform prior distribution. For this scheme to be fair across submissions, the number of submitted samples must be equal for all models and observations. We will therefore up-sample any submissions with less than 1000 samples before calculating the Ignorance Score. By having a maximum number of samples, we also limit how certain any forecast can be. E.g., with 11 bins, we can at most be right 1001 out of 1011 times (0.99%, or Ignorance score of 0.014).

We will use the following binning scheme: 0, 1–2, 3–5, 6–10, 11–25, 26–50, 51–100, 101–250, 251–500, 501–1000, 1001–.

Secondary metric II: Mean Interval Scores (MIS)

The M4 competition (Makridakis, Spiliotis, and Assimakopoulos, 2018) use the Mean Interval Scores (MSIS). MSIS is set up as a battle between making the prediction interval as small as possible whilst still making sure that we have a good coverage rate. It does not consider the mass of the predictive distribution within the interval, so it is not an accuracy metric like CRPS. The metric is a nice transition from a point estimate to the distribution that CRPS tests. It puts the focus on the most likely values, without narrowing to a point. The trade-off between penalizing large intervals but rewarding coverage is useful too. The scaling in MSIS is used to make the measure scale-independent as the M4 competition deals with a large set of different types of time-series with varying time-scales and variability. Since this is not needed here, we have simplified this score to just the (mean) Interval Score, which is also discussed in (Gneiting and Raftery, 2007). The Interval Score is defined as:

$$IS_{it} = (U_{it} - L_{it}) + \frac{2}{a}(L_{it} - Y_{it})\mathbb{1}(L_{it} - Y_{it}) + \frac{2}{a}(Y_{it} - U_{it})\mathbb{1}(Y_{it} - U_{it})$$

where U_{it} and L_{it} are the upper and lower prediction sample quantiles using the set prediction interval, $a = [1 - (\text{prediction interval})]$ (e.g. for a 95% prediction interval, $a = 0.05$) and $\mathbb{1}(z)$ is the indicator function as defined previously. To get the Mean Interval Score, the IS_{it} is averaged across time t and units i.

Implementation: It is not straight-forward to calculate quantiles from a sample (Hyndman and Fan, 1996). We use the linear (Gumbel) method for interpolation, which is the default approach in NumPy. We estimate the MIS for the 90% prediction interval (a = 0.1). We consider this mainly a calibration metric, where we want to measure the ability to issue forecasts that most of the time includes the observed outcome. Since CRPS is already accounting for the error-distance aspect, we opted for as wide prediction intervals as we deemed possible to elicit. Through simulation, we found that the ability of forecasts with 1000 samples to accurately provide estimates of quantiles for overdispersed distributions outside the 90% prediction interval tapers off quickly, which is why we did not settle for a wider interval. It is also at the tail end of the distribution that the choice of the quantile interpolation method matters the most. By setting this to 90%, we are attempting to reduce the effect of the computation and number of samples – which would have a potentially large effect on wider intervals – to learn about the range of the most likely values that extend towards the tails of the count distribution.

Secondary metric III: mu-pEMDiv

An additional metric is a new multiplex pseudo-Earth Mover Divergence Score (mu-pEMDiv). This metric is an expansion of pseudo-Earth Mover Divergence (pEMDiv) (Hegre, Vesco, and Colaresi, 2022). While pEMDiv calculated the cost of moving excess prediction mass across space and/or time (or any network with defined node-to-node distances), mu-pEMDiv also allows the transportation of excess prediction mass across binned-values of the target variable (eg fatality count bins). The new measure first bins the count of fatalities into J bins. For simplicity, think of 5 bins: 0 but less than 1 fatalities, 1 to less than 10 fatalities, 10 to less than 100 fatalities, 100 to less than 1000 fatalities, over 1000 fatalities. The samples are used to create a probability for each observation i from forecast model m that bin j will occur. The lowest bin is kept as a reference distribution. Call the forecast $\hat{f}_{(m)(x)}$ and the actual $f_y(x)$ as an indicator value that takes on the value of 1 for the category observed and 0 everywhere else. We then calculate $\hat{f}^{(m)}(x) - f_y(x)$, yielding surpluses and deficits at particular bins of the target variable.

Instead of running CRPS or another metric on these ordered categories, however, mu-pEMDiv uses the pEMDiv algorithm to run on a directed graph (where a grid or time series chain is a special case) that takes into account space, time, and bin adjacency simultaneously. This is akin to each bin (after the lowest bin) representing a new layer in a multiplex network. The costs to transport mass are calculated across edges for space, time, and bins. Specifically, a bin j at unit/observation i, t is connected to bin $j+1$ for units/observation i, t until the max category J. This is the extension of pEMDiv into a multiplex network. Similar to pEMDiv, each layer in this multiplex connects nodes i, t that are contiguous and weights those edges by a defined distance/cost. Thus, there are costs to all these movements and a scale has to be created such that the movement of a particle across adjacent bins is equivalent to the cost of moving it across some space/time distance. This can be set to 1, same as moving adjacent edges to test out. The cost of creating mass or destroying mass at the last bin J must be greater than the half the maximum distance across the multiplex. This ensures mass is moved instead of teleported by the algorithm. However, in the mu-pEMDiv, mass can not move across bins, where the cost is that mass creation/destruction cost divided by one minus the number of categories.

pEMDiv is then run with each node in the directed graph now being an observation-bin combo (so if there are N observations and J bins there are $N \times (J-1)$ nodes with a potential surplus or a deficit) You can move each surplus particle across space and time for example, and also across bins. There is particle mass creation and distribution just like in normal pEMDiv, additionally. For only two bins, mu-pEMDiv reduces to pEMDiv.

Implementation: mu-pEMDiv is implemented by setting up a network with space, time, and bins (not including the lowest bin) defining nodes. This can be accomplished with the original pEMDiv package in Python (Hegre, Vesco, and Colaresi, 2022). More implementation details and code will be available at a later date.

4.3 Other metrics

Threshold-weighted CRPS: twCRPS

Evaluating and exploring tail behavior is not an easy task (Lerch et al., 2017). It is also a real question whether we have enough power to evaluate tail behavior, particularly when only eliciting 1000 samples from the predictive distribution. Gneiting and Ranjan (2011) propose a threshold-weighted CRPS that we will consider. It is implemented here: `https://github.com/properscoring/properscoring/blob/master/properscoring/_brier.py`.

Average coverage difference

The M4-competition also uses the average coverage difference as a supplementary measure to explore the calibration of models. The ACD is the difference in (1-how often the future observed values are outside the predicted X% interval) - X). We would need to specify X (e.g., 95%).

Visual evaluation

Probability integral transforms and verification rank histograms could be used to explore calibration more visually. We will use such tools to further probe the performance in the competition, but it will not be used to rank the contributions.

year	crps	ign	mis
2018	20.03	1.19	738.02
2019	9.62	1.04	338.35
2020	13.68	1.09	497.83
2021	37.11	1.23	1429.31
Mean	20.11	1.14	750.88

(a) By calendar year

month	crps	ign	mis
1	13.71	1.10	496.17
2	13.08	1.00	470.85
3	11.23	1.16	397.35
4	19.29	1.11	718.10
5	19.14	1.16	712.02
6	23.76	1.09	897.36
7	22.51	1.16	846.35
8	19.02	1.02	707.32
9	17.54	1.27	646.34
10	21.48	1.20	806.07
11	42.01	1.19	1624.98
12	18.54	1.17	687.63

(b) By month in forecasting horizon

Table 2. Benchmark model evaluation, *cm* level, VIEWS/UCDP last historical values predictions with uncertainty estimated assuming Poisson distribution.

5 Benchmark models

5.1 *cm* level

Table 2 shows the evaluation of a model that uses the last observed value as the prediction – for every month in each calendar year, the prediction is based on the observed value in October the preceding calendar year. To introduce variability in the benchmark model, we drew 1000 samples for each country month from a Poisson distribution with mean and variance equal to the last historical value. In the majority of the cases where the last observed number of fatalities was 0, all draws are identical, but in the cases were the last historical value was non-zero, there is some variation.

Sub-table 2a shows the evaluation metrics aggregated by calendar year as well as the mean score across the four years. On average across the four years 2018–2021, the CRPS for this model is 20.11, the ignorance score is 1.14, and the mean interval score is 750.82. For all years, the scores for 2021 are considerably higher than for the other years. This is mostly due to Ethiopia, where the number of deaths recorded by the UCDP increased from 0 to 1445 from October 2021 to the month after, and then decreased to 160 in December, as violence escalated in the Tigray province.[9] Sub-table 2b shows that the scores tend to deteriorate over the months of the forecasting horizon – naturally, the violence recorded in October in a year is a better predictor of violence in January than in December the year after.

Table 3 shows the evaluation of another heuristic model – each prediction in this model for a given

[9] In their 2022 update, the UCDP has adjusted considerably the fatality count estimate upwards for this conflict.

year	crps	ign	mis
2018	23.44	1.11	828.71
2019	22.02	1.09	765.92
2020	21.22	1.08	715.05
2021	35.99	1.12	1292.25
Mean	25.67	1.10	900.48

(a) By calendar year

month	crps	ign	mis
1	23.20	1.11	800.44
2	20.04	1.05	693.18
3	21.29	1.12	719.12
4	25.40	1.10	879.51
5	26.79	1.15	902.93
6	31.61	1.08	1144.24
7	33.82	1.14	1228.21
8	24.79	1.08	864.01
9	19.13	1.16	638.84
10	22.52	1.04	794.11
11	41.25	1.08	1542.28
12	18.18	1.09	598.94

(b) By month in forecasting horizon

Table 3. Benchmark model evaluation, *cm* level, bootstraps from actuals. Left: by calendar year, right: by month in forecasting horizon.

country month is a draw from the set of actuals for the entire calendar year containing that month. In other words, the prediction for Ethiopia for November 2021 is a set of 1,000 random draws from the observed fatality counts for all country months in 2021. The expected strength of the model is that it in aggregate covers all actual outcomes. Consequently, the ignorance scores are low. CRPS is higher than all the other models, whereas the mean interval scores are moderately high.

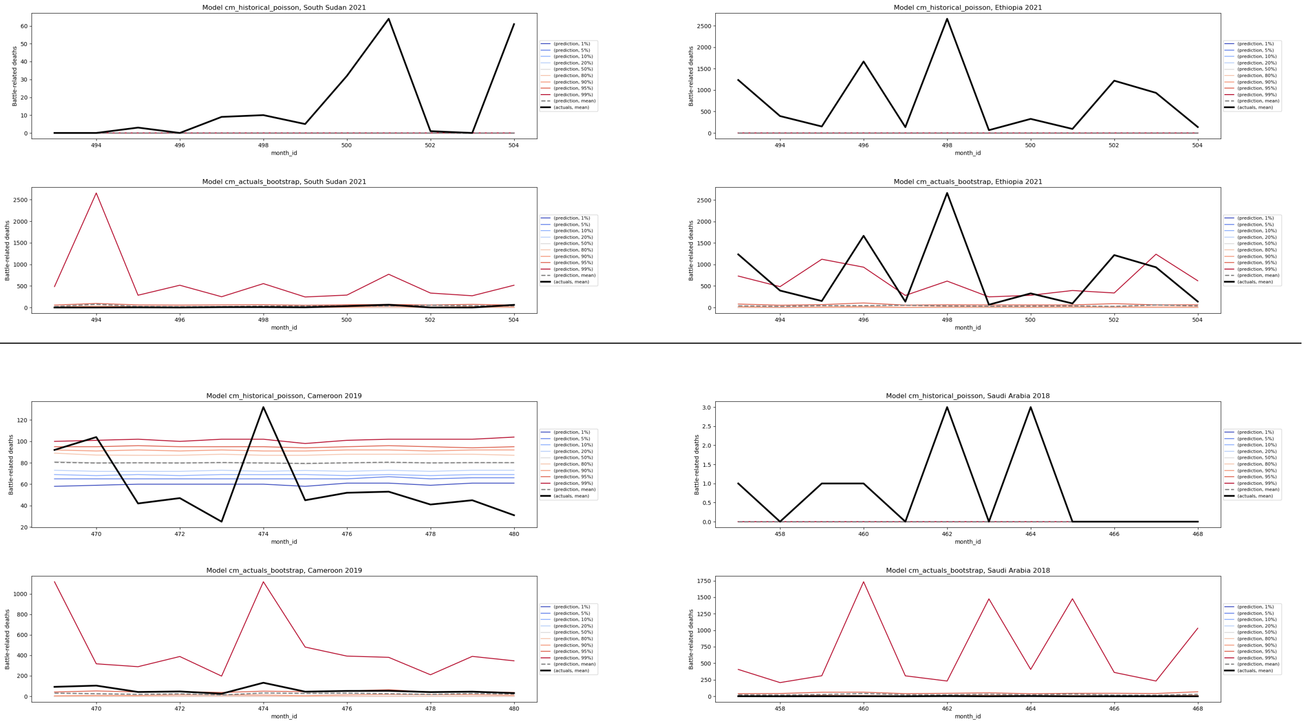

Figure 4. Predictions and actuals, benchmark models, South Sudan 2021, Ethiopia 2021, Cameroon 2019, Saudi Arabia 2018

5.2 *pgm* level

year	crps	ign	mis
2018	53.66	2.45	1990.39
2019	21.96	2.71	763.93
2020	19.80	3.13	668.73
2021	23.10	2.45	804.89
Mean	29.63	2.68	1056.99

(a) By calendar year

month	crps	ign	mis
1	29.60	2.68	1056.24
2	29.58	2.68	1055.35
3	29.56	2.68	1054.17
4	29.63	2.68	1057.32
5	29.61	2.68	1056.05
6	29.64	2.69	1057.60
7	29.62	2.68	1056.67
8	29.62	2.68	1057.09
9	29.63	2.69	1056.73
10	29.69	2.69	1059.18
11	29.69	2.69	1059.96
12	29.64	2.69	1057.48

(b) By month in forecasting horizon

Table 4. Benchmark model evaluation, *pgm* level, VIEWS/UDCP last historical values with uncertainty estimated assuming Poisson distribution.

We have prepared a benchmark model at the *pgm* level of analysis – additional models will be prepared and made available to prospective participants at `https://viewsforecasting.org/prediction-competition-2/`.

Table 4 shows the evaluation of a *pgm* model that uses the last observed value as the prediction – for every month in each calendar year, the prediction is based on the observed value in October the preceding calendar year. Similarly to the benchmark model at the *cm* level, we drew 1000 samples for each PRIO-GRID month from a Poisson distribution with mean and variance equal to the last historical value. In the majority of the cases where the last observed number of fatalities was 0, all draws are identical, but in the cases were the last historical value was non-zero, there is some variation.

Sub-table 4a shows the evaluation metrics aggregated by calendar year as well as the mean score across the four years. On average across the four years 2018–2021, the CRPS for this model is 29.63, the ignorance score is 2.68, and the mean interval score is 1056.99. For the CRPS and the MIS, the scores for 2018 are considerably higher than for the other years (but the ignorance score is low). Sub-table 4b shows that the scores tend to deteriorate very slightly over the months of the forecasting horizon.

5.3 Ensembling

We will generate an ensemble from the contributed models, with models weighted by some function of the CRPS. Building on the ensemble, we can assess models based on their unique or distinct contribution relative to the set of all models, or their diversity relative to the average contribution, as well as evaluate the joint contribution of all models.

6 Publication strategy

The VIEWS team will set up a new section in `https://viewsforecasting.org` where we present the contributions. We will seek to provide monthly updates of the contents throughout the true future forecasting window:

- All predictions
- Actuals and prediction errors
- Evaluation scores

We will also publish the contributors' summary papers as items in the VIEWS working papers series (under development).

The team is in dialogue with a good, peer-reviewed journal about publishing the following:

- An introduction article at the start of the prediction window, with all selected contributors as authors. The article would include the core information in this concept note and summarise each contribution
- A summary article published just after the end of the true future prediction window (early 2025), where we evaluate the contributions and the scoring committee ranking of the contributions according to the different metrics

References

Carvalho, Arthur (2016). "An Overview of Applications of Proper Scoring Rules". In: *Decision Analysis* 13.4, pp. 223–242. DOI: 10.1287/deca.2016.0337.

Czado, Claudia, Tilmann Gneiting, and Leonhard Held (2009). "Predictive Model Assessment for Count Data". In: *Biometrics* 65.4, pp. 1254–1261. DOI: https://doi.org/10.1111/j.1541-0420.2009.01191.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2009.01191.x.

Davies, Shawn, Therése Pettersson, and Magnus Öberg (2022). "Organized violence 1989–2021 and drone warfare". In: *Journal of Peace Research* 59.4, pp. 593–610. DOI: 10.1177/00223433221108428. eprint: https://doi.org/10.1177/00223433221108428.

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.

Gneiting Tilmann, Fadoua Balabdaoui and Adrian E. Raftery. (2007). “Probabilistic Forecasts, Calibration and Sharpness”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2, 243—68. DOI: 10.1111/j.1467-9868.2007.00587.x.

Gneiting, Tilmann (2011). “Making and Evaluating Point Forecasts”. In: *Journal of the American Statistical Association* 106.494, pp. 746–762. DOI: 10.1198/jasa.2011.r10138. eprint: https://doi.org/10.1198/jasa.2011.r10138.

Gneiting, Tilmann and Adrian E Raftery (2007). “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Gneiting, Tilmann and Roopesh Ranjan (2011). “Comparing density forecasts using threshold-and quantile-weighted scoring rules”. In: *Journal of Business & Economic Statistics* 29.3, pp. 411–422. DOI: doi.org/10.1198/jbes.2010.08110.

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

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

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

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

Hyndman, Rob J. and Yanan Fan (1996). “Sample Quantiles in Statistical Packages”. In: *The American Statistician* 50.4, pp. 361–365.

Kolassa, Stephan (2016). “Evaluating predictive count data distributions in retail sales forecasting”. In: *International Journal of Forecasting* 32.3, pp. 788–803. DOI: https://doi.org/10.1016/j.ijforecast.2015.12.004.

Lerch, Sebastian et al. (2017). “Forecaster’s dilemma: extreme events and forecast evaluation”. In: *Statistical Science*, pp. 106–127.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2018). “The M4 Competition: Results, findings, conclusion and way forward”. In: *International Journal of Forecasting* 34.4, pp. 802–808. DOI: 0.1016/j.ijforecast.2018.06.001.

Shannon, Claude Elwood (Oct. 1948). “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27, pp. 379–423.

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.

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

Vesco, Paola et al. (2022). “United They Stand: Findings from an Escalation Prediction Competition”. In: *International Interactions* 48.4, pp. 1–37. DOI: 10.1080/03050629.2022.2029856. eprint: https://doi.org/10.1080/03050629.2022.2029856.

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

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