Research Article

Journal of Peace Research

The 2023/24 VIEWS Prediction challenge: Predicting the number of fatalities in armed conflict, with uncertainty

Journal of Peace Research
1–18

DOI: 10.1177/00223433241300862
journals.sagepub.com/home/jpr


Håvard Hegre[1], Paola Vesco[1], Michael Colaresi[2], Jonas Vestby[3], Alexa Timlick[3], Noorain Syed Kazmi[3], Angelica Lindqvist-McGowan[4], Friederike Becker[5], Marco Binetti[6], Tobias Bodentien[5], Tobias Bohne[6], Patrick T. Brandt[7], Thomas Chadefaux[8], Simon Drauz[5], Christoph Dworschak[9], Vito D'Orazio[10], Hannah Frank[8], Cornelius Fritz[11], Kristian Skrede Gleditsch[12], Sonja Häffner[6], Martin Hofer[6], Finn L Klebe[13], Luca Macis[14], Alexandra Malaga[15], Marius Mehrl[16], Nils W Metternich[13], Daniel Mittermaier[6], David Muchlinski[17], Hannes Mueller[18], Christian Oswald[6], Paola Pisano[14], David Randahl[4], Christopher Rauh[19], Lotta Rüter[5], Thomas Schincariol[8], Benjamin Seimon[20], Elena Siletti[14], Marco Tagliapietra[14], Chandler Thornhill[21], Johan Vegelius[22] and Julian Walterskirchen[6]

Abstract

Governmental and nongovernmental organizations have increasingly relied on early-warning systems of conflict to support their decisionmaking. Predictions of war intensity as probability distributions prove closer to what policymakers need than point estimates, as they encompass useful representations of both the most likely outcome and the lower-probability risk that conflicts escalate catastrophically. Point-estimate predictions, by contrast, fail to represent the inherent uncertainty in the distribution of conflict fatalities. Yet, current early warning systems are preponderantly focused on providing point estimates, while efforts to forecast conflict fatalities as a probability distribution remain sparse. Building on the predecessor VIEWS competition, we organize a prediction challenge to encourage endeavours in this direction. We invite researchers across multiple disciplinary fields, from conflict

[1]*Peace Research Institute Oslo (PRIO) & Department of Peace and Conflict Research, Uppsala University*
[2]*Department of Political Science, University of Pittsburgh and Peace Research Institute Oslo (PRIO)*
[3]*Peace Research Institute Oslo (PRIO)*
[4]*Department of Peace and Conflict Research, Uppsala University*
[5]*Institute of Statistics (STAT), Karlsruhe Institute of Technology (KIT)*
[6]*Center for Crisis Early Warning (CCEW), University of the Bundeswehr Munich*
[7]*School of Economic, Political, and Policy Sciences, University of Texas, Dallas*
[8]*Department of Political Science, Trinity College Dublin*
[9]*Department of Politics and International Relations, University of York*
[10]*Department of Political Science, West Virginia University*
[11]*School of Computer Science and Statistics, Trinity College Dublin*
[12]*Peace Research Institute Oslo (PRIO) & Department of Government, University of Essex*
[13]*Department of Political Science, University College London*
[14]*Department of Economics and Statistics 'Cognetti de Martiis', University of Turin*
[15]*Institute for Economic Analysis (CSIC), Barcelona*
[16]*School of Politics and International Studies, University of Leeds*
[17]*Sam Nunn School of International Affairs, Georgia Tech*
[18]*Institute for Economic Analysis (CSIC), Barcelona & Centre for Economic Policy Research (CEPR), Barcelona School of Economics*
[19]*Peace Research Institute Oslo (PRIO) & Institute for Economic Analysis (CSIC), Barcelona*
[20]*Fundació Economia Analítica*
[21]*School of Economics, Georgia Tech*
[22]*Department of Medical Sciences, Uppsala University*

Corresponding author:
Håvard Hegre.
hhegre@prio.org

studies to computer science, to forecast the number of fatalities in state-based armed conflicts, in the form of the UCDP 'best' estimates aggregated to two units of analysis (country-months and PRIO-GRID-months), with estimates of uncertainty. This article introduces the goal and motivation behind the prediction challenge, presents a set of evaluation metrics to assess the performance of the forecasting models, describes the benchmark models which the contributions are evaluated against, and summarizes the salient features of the submitted contributions.

Keywords
Armed conflict, prediction, uncertainty

Introduction

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 such as the UN and within governments. Several such organizations are developing early-warning systems, or including forecasts from systems like the Violence Early Warning System (VIEWS – Hegre et al., 2019, 2021) in their 'dashboards' to support decisionmaking. 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 interested in both the most likely outcome (a point prediction), and the lower-probability risk that conflicts escalate catastrophically (the tail ends of the probability distribution). Point predictions are arguably not very informative in the situations where early warnings are most useful, namely before large-scale violence erupts in places where there has been limited previous violence. Since it can take a long time before severe tension escalates into overt violence, point predictions tend to cluster around no violence. Forecasts in the form of probability distributions, on the other hand, can alert to a low but alarming risk of a large-scale conflagration. Users are also interested in knowing how uncertain the forecasts are. As such, predictions of war intensity as probability distributions come closer to what user groups need than point estimates or simple dichotomous predictions from classification models (Brandt et al., 2014).

Generating forecasts as probability distributions, however, is new to the field of armed conflict forecasting. Most of the existing efforts, in fact, provide forecasts as point predictions without any measures of uncertainty (e.g. Bazzi et al., 2022; Blair and Sambanis, 2020; Brandt et al., 2011; Chiba and Gleditsch, 2017; Dorff et al., 2020; Goldstone et al., 2010; Hegre et al., 2013, 2022; Mueller and Rauh, 2018; Vesco et al., 2022). To strengthen the knowledge basis for such modeling, the VIEWS project has invited research teams interested in forecasting models in general and in the prediction of armed conflict in particular, to take part in a prediction challenge where all contributors work on a common, well-defined task:

To predict the number of fatalities in armed conflict, as reported by the Uppsala Conflict Data Program (UCDP), with estimates of the uncertainty of the predictions calculated in the form of samples of forecast values.

In this article, we describe the challenge in more detail, present the task, outline the 24 contributions from 13 teams, and present the evaluation metrics that they will be evaluated on, as well as a set of benchmark models that the contributions will be scored against. By rewarding contributions that perform well both in terms of point prediction and uncertainty estimation, the challenge encourages interdisciplinary efforts to model uncertainty around armed conflict forecasts, increases our understanding of the model characteristics that are most useful to improve probabilistic forecasts, sheds light on the issues involved in evaluating such forecasts, and suggests a set of evaluation metrics that could address these issues.[1]

This article is supplemented by an interactive visualization tool available at https://predcomp.viewsforecasting.org to explore all forecasts, as well as an Online Appendix. In-depth summaries for all models can be found at https://viewsforecasting.org/research/prediction-challenge-2023/.

The challenge

The challenge builds on the predecessor VIEWS prediction competition (Hegre et al., 2022; Vesco et al., 2022), where the task 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 tools to predict

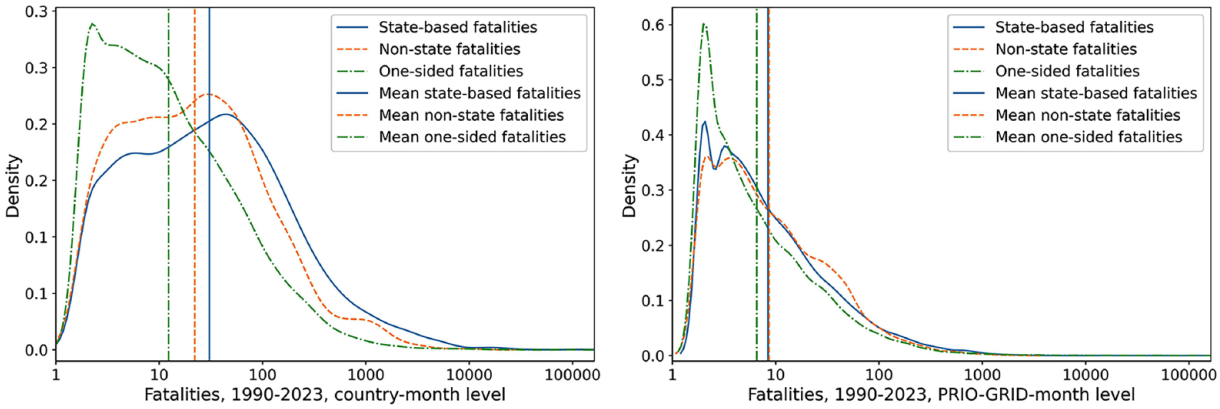


Figure 1. Distributions of observed (non-zero) fatalities for the three different types of violence reported by the UCDP-GED dataset (state-based, non-state and one-sided; Pettersson et al., 2024), 1990–2023, at our two levels of analysis: country-month (*cm*) (left) and PRIO-GRID-month (*pgm*) (right).
The density plot lines represent the natural logarithmic scale of observed fatalities for better visualization of a wide range of values. The horizontal axis is labeled to show the equivalent non-logged count of fatalities.

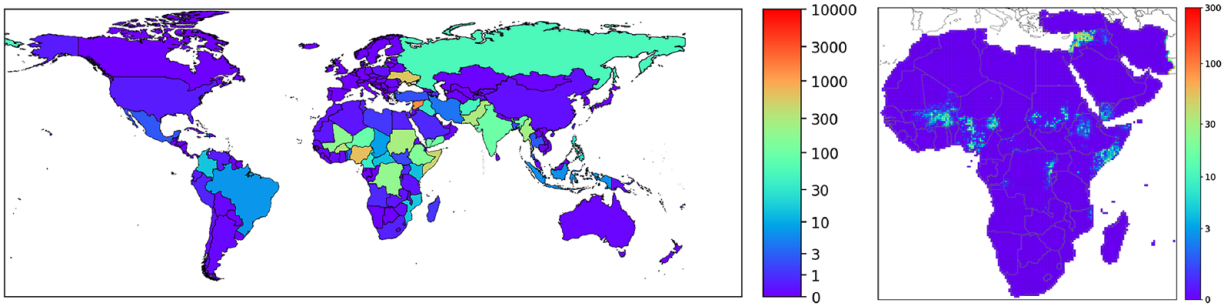
changes in fatalities – although they tend to be difficult to interpret (Vesco et al., 2022). Even very sophisticated machine learning models, however, tend to be surprised by the outbreak of conflicts in previously peaceful locations: most (but not all) of the models in fact are beaten, on common forecast evaluation metrics, by a basic 'no-change' model that constantly predicts a null change in fatalities.

These findings suggest that more research is needed to improve our collective ability to forecast conflict outbreaks and (de-)escalation, but also call for better evaluation metrics to more meaningfully assess forecast accuracy. Point-estimate predictions – which rely on a single measure of the predicted outcome – are difficult to evaluate in a way that creates new insights. Point estimates, by definition, do not encode the inherent uncertainty in the distribution of conflict fatalities for a given instance. Moreover, point-estimate predictions obscure the relation between the shape and location of the predictive distribution and the underlying data generating process – occluding important traits of forecasts, including calibration. This divergence might be most pronounced for processes that produce skewed distributions of target observations, such as in the UCDP conflict fatalities we use (Pettersson et al., 2024). Figure 1 demonstrates the skewness of the (non-zero) fatality counts at the two VIEWS levels of analysis, the 'country-month' (*cm*) and the 'PRIO-GRID-month' (*pgm*) levels. 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 *cm* level are zeros, and as many as 99% at the *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) favor – may not be of pre-eminent relevance. For an outcome like the one of interest, with its zero-inflated and right-skewed distribution/probability mass function, the expected value is not sufficiently informative to reflect the risks associated with different potential outcomes. Consider an example of two forecast distributions where the first forecast assigns a 99% chance to zero fatalities and a 1% chance to 120 fatalities, while the second places a 40% chance on zero fatalities and a 60% chance on one fatality. Both forecast distributions have a mean of 1.2, but very different practical implications in terms of crisis mitigation.[2]

If the goal is to produce an insightful characterization of the predictive distribution, which can meaningfully represent both the near-zero and the extreme cases at the tail, then there is a need to move from single-valued point forecasts to probabilistic forecasts that explicitly represent the distribution of plausible outcomes.

To explore this challenge, we distributed an invitation to contribute in April 2023. The contributors presented preliminary models to a workshop in October 2023, and their final forecasts in late June 2024.[3] In the fall of 2025, after the end of the true future forecasting window, a separate article will provide a comprehensive evaluation of the performance of the models, based on both the test window forecasts and the 12 months of 2024–2025. More details about the prediction target and windows are provided in the next section.



Country-months (*cm*). 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 et al., 2010). For the country level of analysis VIEWS data are global.

PRIO-GRID months (*pgm*), which rely on PRIO-GRID (Tollefsen et al., 2012), a standardized spatial grid structure consisting of quadratic grid cells at a resolution of 0.5×0.5 decimal degrees. Near the equator, the area of such a cell is 55×55km.

Figure 2. Levels of analysis: forecasts of (logged) armed conflict fatalities for June 2025, produced by the VIEWS ensemble model based on data up to and including June 2024, for the country-month *cm* (left) and PRIO-GRID-month *pgm* level (right).
The colors are based on a natural logarithmic scale of forecast fatalities for better visualization of a wide range of values. The legend is labeled to show the equivalent non-logged count of fatalities.

Prediction target: Predicting the monthly number of fatalities in state-based armed conflict

The prediction target is the UCDP record of the number of fatalities in state-based armed conflict, as defined in Pettersson et al. (2024). For the test prediction windows up to and including 2023, we have access to the final UCDP data as reported in that article. For the year of 2024 up to April, the UCDP Candidate data (Hegre et al., 2020) were available. Contributions were requested to present predictions as a number of draws from the predicted distribution of fatality counts.[4]

We requested contributions for two sets of prediction windows. The primary goal was to provide predictions for the true future – the 12 months from July 2024 to June 2025, based on data up to and including April 2024. The predictions were to be provided separately for each of the 12 months in this time window.[5] However, a single year of events at these aggregation levels will generate a limited amount of data for model evaluation, and the evaluation scores will vary considerably over time, as the benchmark model evaluation presented in the benchmark section below suggests. To complement the evaluation of the true future forecasts, we also requested contributions for each of the 12 months in the calendar years 2018-2023, based on data up to and including October for the preceding year. Seven sets of input data were made available to the participants, one for each of these forecasting windows.

The contributors were requested to submit forecasts in either or both of the two VIEWS levels of analysis (Hegre et al., 2019, 2021), as used in the former VIEWS prediction competition (Hegre et al., 2022; Vesco et al., 2022) and depicted in Figure 2. The subnational level forecasts were restricted to Africa and the Middle East.[6] The temporal unit for both levels of analysis is the month.

The contributions

The contributions cover a wide array of methodologies and approaches to forecast conflict fatalities with uncertainty. Several contributions aim to capture the complex spatio-temporal dynamics that characterize conflicts – which tend to cluster in space and re-occur over time – leveraging various methods, such as (empirical) quantile-based solutions (Bodentien and Rüter, 2024; Drauz and Becker, 2024), ensembles of local random forests (Mittermaier et al., 2024), Markov models (Randahl and Vegelius, 2024), sequence-based approaches and dynamic time warping (Gleditsch et al., 2024; Schincariol et al., 2024).

A few models exploit new and granular data sources, including information on early signs of tensions extracted from the news (Málaga et al., 2024), and data on the presence of and relations among actors (Gleditsch et al., 2024). Some contributions focus on the distributional challenges of the outcome, using hierarchical and two-stage models to explicitly tackle the zero inflation (Fritz et al.,

Table 1. Overview of the contributions at *cm* and *pgm* levels, sorted by team names.

Team	*Level*	*Main features*	*Reference article*
Bodentien & Rüter	*cm*	Empirically estimated quantiles assuming a negative binomial distribution.	Bodentien and Rüter (2024)
Brandt	*cm*	Bayesian density forecast allowing for the evaluation of dynamics and different distributional assumptions for the outcome (Tweedie, Negative Binomial, and Poisson).	Brandt (2024)
CCEW 'tree'	*pgm*	Combination of local and global tree-based models in a hurdle ensemble framework.	Mittermaier et al. (2024)
CCEW 'tft'	*cm, pgm*	Temporal fusion transformer models incorporating time-invariant covariates and past/future inputs.	Walterskirchen et al. (2024)
Conflict forecast	*cm, pgm*	Random forest on conflict history and 15 topics extracted from a large news corpus using Latent Dirichlet Allocation.	Málaga et al. (2024)
D'Orazio	*pgm*	Auto-machine learning systems to optimize input features, algorithms, outcome transformations and distributional assumptions.	D'Orazio (2024)
Drauz & Becker	*cm*	Equally spaced empirical quantiles sampled from conflict history, optimized for each country and forecast horizon.	Drauz and Becker (2024)
Fritz et al.	*pgm*	Three-stage hierarchical hurdle count model with classification stages at *cm* and *pgm*, and truncated count regression at *pgm* level as final stage.	Fritz et al. (2024)
Gleditsch et al.	*pgm*	Dynamic Time Warping on dyad-specific features addressing actor-related heterogeneity at the grid level.	Gleditsch et al. (2024)
Muchlinski & Thornhill	*cm*	Two-stage zero-inflated hurdle Generalized Additive Model (GAM) and ensemble machine learning.	Muchlinski and Thornhill (2024)
PaCE	*cm*	Shape-based prediction using Dynamic Time Warping to identify recurrent patterns in time-series data.	Schincariol et al. (2024)
Randahl & Vegelius	*cm*	Hidden, Observed, and Gaussian Markov models capturing different latent or observed states of conflict.	Randahl and Vegelius (2024)
Unito	*cm*	Transformer models applied to time-series conflict data, optimized through Log-Likelihood Loss function of Negative Binomial distribution.	Macis et al. (2024)

2024; Muchlinski and Thornhill, 2024), or flexibly exploring various distributional assumptions and assessing how they affect predictive performance (Brandt, 2024; D'Orazio, 2024).

Lastly, some contributions rely on complex methodological innovation and exploit the recent rise of transformer-based models, such as temporal fusion transformers (TFTs) and pre-trained attention mechanisms, allowing for greater flexibility in incorporating temporal patterns and exogenous variables into the probabilistic predictions (Macis et al., 2024; Walterskirchen et al., 2024).

Table 1 presents a short overview of the models – the references in the table contain links to the more extensive presentations of the models written by each contributor team, available at https://predcomp.viewsforecasting.org. The Online Appendix also presents very short summaries outlining each team's modeling strategies.

Evaluation and metrics

Scoring committee

The evaluation of the contributions will be done by a scoring committee consisting of members of the forecasting expert community as well of the user community: Philip Schrodt (Parus Analytics, former Pennsylvania State University), Céline Cunen (Norwegian Computing Center and University of Oslo), Thomas Mayer (Preview, German Ministry of Foreign Affairs), and Seth Caldwell (formerly United Nations Office for the Coordination of Humanitarian Affairs – UN OCHA). The scoring committee will provide an independent evaluation of the models, based primarily on the quantitative scoring outlined below (which, for technical reasons, will be computed by the VIEWS team), but also on the short summaries of the contributions, as well as any additional criteria the committee itself may deem relevant.

The role of the scoring committee is thus to provide a comprehensive assessment of the contributions that goes beyond the quantitative evaluation, and account for additional aspects of the models – such as creative innovations or impressive replicability – that may represent a valuable contribution despite not being necessarily rewarded by the quantitative scoring. We will place approximately equal weight on the joint performance across the test period predictions and the predictions for 2024, the true future.

Evaluation guiding principles

For the challenge to be useful for a wide range of researchers and users, we need to incorporate several evaluation metrics and principles for distinct use cases, accounting for different aspects of the predictive distributions, as different facets of forecasts will potentially serve distinctive purposes. To this end, we begin by defining the traits that we would like to see in our probabilistic forecasts. For example, two related goals of probabilistic forecast evaluation are to reward the 'sharpness of the distribution, subject to calibration' (Gneiting and Raftery, 2007: 359). Sharpness is a priority because it represents higher plausibility on values and more certainty, all else equal. Calibration in turn seeks to reward forecasts that match the longer-run, average propensity for events of a certain size to occur with the forecast probability of those events. We also seek to prioritize forecasts that provide as much information on the event that actually materializes as possible. This is known as focus or locality and is distinct from sharpness because it depends on the observed value.

In addition to sharpness, calibration, and focus, we also want to induce propriety – the honest reporting of probabilities over outcomes from competitors where possible. We discuss these criteria and how they jointly lead us to selecting corresponding evaluation metrics below.

What to reward: Evaluation criteria and metrics

Here, we present the main metrics used to evaluate the contributions and 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:

- $f_i^{(m)}(x)$: The forecast distribution/probability mass function (pmf) from model m for instance i, dropping the t subscript for clarity, over possible outcome values x.
- $F_i^{(m)}(x)$: The forecast cumulative density function (CDF) associated with $f_i^{(m)}(x)$.
- y_i: The observed value for instance i.

Table 2. Beneficial qualities of probabilistic forecasting systems and how they are assessed by core evaluation metrics.

Rule	*Desirable qualities*			
	Calibration	*Sharpness*	*Focus*	*Propriety*
CRPS	X	X	-	X
ab-Log Score	-	X	X	x
MIS	X	X	-	X

Large *X* means the metric is highly useful for assessing the respective quality. Small *x* means the metric captures the quality partially or with conditions.

As noted above, we will do the evaluation in terms of y_i as the non-logged count of fatalities, in place of the logged fatalities ($ln(y_i+1)$) used in Hegre et al. (2022).

The metrics are designed to reward the four qualities of probabilistic forecasting systems mentioned above – calibration, sharpness, focus, and propriety. Table 2 summarizes these four traits and links them to the main metrics that we will be using.

Calibration: A model is well calibrated when the predicted frequency of $f_i^{(m)}(x)$-values corresponds to the observed frequency of $y_i = x$ in new data. To judge calibration we need to jointly analyze and compare $f_i^{(m)}(x)$ and y_i. For instance, if a well-calibrated model predicts a 30% probability of 100 deaths and we then receive 100 new observations, the new actual data should record approximately 30 observations of 100 deaths.

Sharpness: Concentrated predictive distributions are preferred as they encode more certainty, as defined by Shannon (1948), across all possible values of y_i. Unlike calibration, sharpness is necessarily a function of $f_i^{(m)}(x)$ and not y_i. A model that predicts with 90% probability that the true value is 0 and 10% that it is 101 is sharper as compared to a forecast that specifies a 50% chance of 0 and a 50% chance of 120. The forecasting community generally, although not unanimously, agrees that an ideal forecast should maximize sharpness subject to calibration (Du, 2021; Gneiting et al., 2007; Smith et al., 2015).

Focus: refers to the aim that useful predictive distributions should provide high plausibility at the exact value that materializes. This quality is a function of both $f_i^{(m)}(x)$ and y_i. Focus is sometimes referred to as locality (Du, 2021). To see the unique value of focus as

compared to sharpness, imagine the forecast distribution as a flashlight pointed at a nearly infinite ruler that runs from 0 to some very large number. The ruler has a mark at some value that represents y_i, the actual value. As the plausibility of specific values for the observation (from the point of the view of the forecast) increases, more and more light is thrown on those values on the ruler. Unlike sharpness, which would simply measure the highest intensity of light regardless of the value marked on the ruler, focus rewards how much light is cast exactly on or close to the marked value on the ruler. Disregarding focus would risk privileging models that assign much less plausibility to the actual values solely because they are better calibrated in areas far away from the realized outcome (Du, 2021). Focus corresponds to the argument 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: 689, cited in Gneiting and Raftery, 2007: 365–366). Focus is also a concept that contrasts with calibration, since calibration would analyze the pattern of the intensity of light across the whole ruler and how it matched the frequency of actual marks/observations.

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 (Czado et al., 2009; Gneiting et al., 2007). In contrast, an improper score might reward increased certainty or a shifted mode for the distribution to hedge relative to the true underlying beliefs.[7] While propriety might seem like it should always and everywhere apply, as with each of the other traits, there are trade-offs. For example, in the current domain of the challenge, directly measuring focus with a proper scoring rule like the the raw log score (defined on the full count sample space and not just a coarser range) is not possible without computing infinite penalties regularly, as we discuss below.

Metrics

The scoring committee will consider the metrics below when evaluating the contributions. The main scoring and ranking of the contributions will be done in terms of the continuous rank probability score (CRPS). The other metrics will be used for secondary scoring and rankings, to facilitate a richer discussion of model performance. The code implementing the evaluation, including all the detailed adaptations reviewed below, is found in https://github.com/prio-data/prediction_competition_2023.

Main metric: Continuous rank probability score (CRPS)

CRPS values sharpness subject to calibration, and is an assessment of the full forecast distribution given the outcome. It is also a proper scoring function (Gneiting and Raftery, 2007). The CRPS for forecast from model m for an individual observation i is defined as:

$$CRPS\left(F_i^{(m)}, y_i\right) = \int_{\mathbb{R}} \left(F_i^{(m)}(x) - \mathbb{1}(x \geq y_i)\right)^2 dx$$

where $\mathbb{1}(z)$ is the indicator function defined as:

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

The individual CRPS scores are averaged for each model across the evaluation observations in a given set.

A few examples will help clarify how the CRPS is computed. First, we imagine two actual observations that will occur in the future, taking the value $y_1 = 0$ and $y_2 = 101$. These observations are representative of our data, characterized by frequent zeros as well as rare non-zero values that are relatively far from zero. For each of the two observations, we introduce two forecasts. For simplicity, the forecasts vary across models but not across observations. The probability that a forecast from model m for observation i takes on the value x is $Pr^{(m)}(y_i = x)$, i.e. the pmf of the forecast for that observation over all x values. We define

$$Pr^{(1)}(y_i = 0) = .9, Pr^{(1)}(y_i = 101) = .1,$$
$$Pr^{(1)}(y_i = \tilde{x}) = 0\ \forall \tilde{x} \notin \{0, 101\}$$

and

$$Pr^{(2)}(y_i = 0) = .5, Pr^{(2)}(y_i = 120) = .5,$$
$$Pr^{(2)}(y_i = \tilde{x}) = 0\ \forall \tilde{x} \notin \{0, 120\}$$

As noted, the CRPS represents the forecasts by their associated CDF, so we calculate $Pr^{(m)}(y_i \leq x)$ for each forecast-observation pair, yielding $F_i^{(m)}(x)$. Each individual observation contributes to the averaged CRPS for model m, as shown in the subplots in Figure 3. The

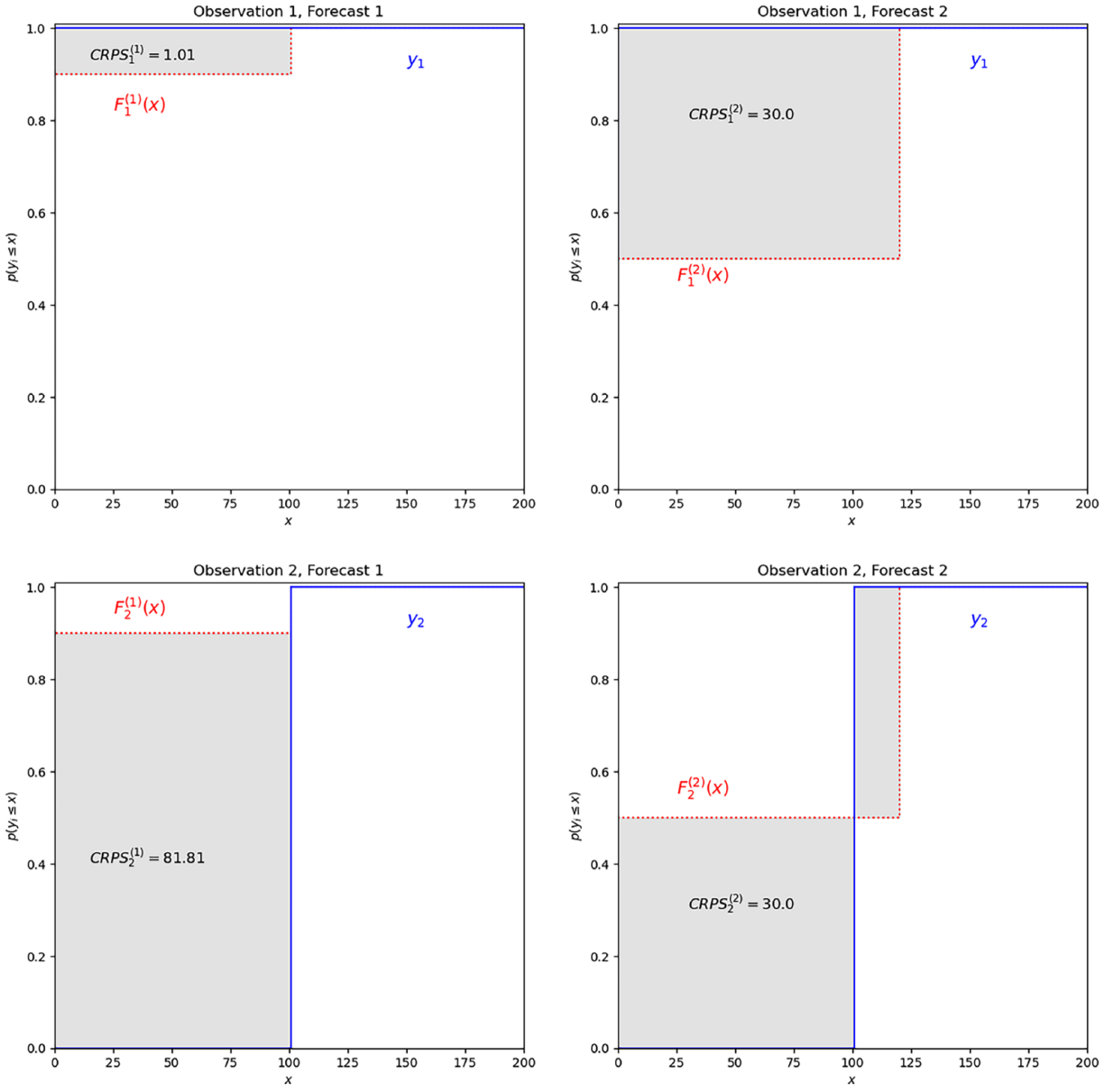


Figure 3. Illustration of the CRPS, inspired by Bracher et al. (2021).
Two exemplary forecasts (in columns) are used to predict two exemplary observations (in rows). Their performance is evaluated by their individual CRPS values, represented by the grey area. The dotted red line in each plot is the CDF of the forecast for that observation. The solid blue line is the step function representing the empirical CDF of the actual observed value. The gray area is the (sum of) squared differences between these two functions and hence constitutes the CRPS value. Smaller differences lead to less area and thus smaller CRPS values are preferred.

dotted red line in each plot is the CDF of the forecast for that observation; the solid blue line is the step function representing the actual value, while the gray area is the difference between these two functions and is squared to compute the individual value. In the top left panel for model 1 and observation 1, the solid blue line jumps up to the actual value of $y_1 = 0$ at $x = 0$ and the CDF $F_1^{(1)}(x)$, jumps to .9 at zero and then 1 at 101. The CRPS measures the square of the area in gray that is the difference between the actual step-function and the forecast CDF. We can quickly see that this area (and thus the individual contribution to the CRPS) is minimized when the forecast CDF is the matching step function that assigns all probability to the actual value.

If we move to the top right panel, we see the forecast from the second model for the first observation, $F_1^{(2)}(x)$

in dotted red, where the CDF jumps only to .5 at $x = 0$ and does not rise to 1 until $x = 120$. Notice that the actual value $y_1 = 0$ (in solid blue) is the same in both plots in the top row, because these are forecasts for the same observation. The area is greater in the right than in the left plot, and the individual CRPS value is smaller (better) for the first forecast on the first observation ($CRPS_1^{(1)} = 1.01$) as compared to the second forecast for that first observation ($CRPS_1^{(2)} = 30$). CRPS rewards the first model on the first observation because it is both sharp (has a large jump in cumulative probability over a narrow range of x) and better calibrated (the shape of the CDF matches more closely the actual value).

The situation is reversed when we move to the second row. The second observed value is $y_2 = 101$ while the forecasts remain the same for the two models. The step function for the actual value (in solid blue) stays at 0 until 101 while the first forecast jumps up to .9 at $x = 0$ (bottom left panel). The large difference between the solid blue and dotted red lines leads to a larger $CRPS_2^{(1)} = 81.81$. The second forecast for the second observation (bottom, right panel) stays lower at .1 from $x = 0$ to $x = 101$, where the actual value occurs. Even though there is an additional area to the right of the actual value here from $x = 101$ to $x = 120$, it is smaller than the area for forecast 1, observation 2. Therefore, we have $CRPS_2^{(2)} = 30$. This result highlights that CRPS does not simply value sharpness. The first forecast is still sharper than the second, but relative to the second observation, the calibration is much worse.

Averaging the two individual contributions to CRPS for each model leads to $CRPS^{(1)} = 41.41$ and $CRPS^{(2)} = 30$. Since the lower value is better, model 2 would be preferred in this stylized example.[8] These examples illustrate a key facet of the CRPS: probability mass that is far from the actual value influences the score. The CDF representation stretches the probability mass across x so that a non-zero probability of a value far from the actual observed value is carried across that distance, creating an area that increases the CRPS average. Thus, the CRPS depends crucially on the entire pmf, through the CDF, not just on the values of the forecast that are close to the actual value. Du (2021) argues that this fact is 'unfortunate' and suggests supplementing any analysis based on CRPS with a local scoring rule, such as the Ignorance/Log score – which we turn to in *Secondary metric I* because it values focus (see also Smith et al., 2015).

Implementation: We compute the measure using the `properscoring.crps_ensemble()` function in Python, as implemented in `xskillscore.crps_ensemble()`. We weigh each sample forecast equally. This approach uses the Empirical CDF to elicit probabilities (see Krüger et al. (2021).

Secondary metric I: adjusted, binned-Log score

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

$$Log\ Score\left(f_i^{(m)}, y_i\right) = -log_2\left(f_i^{(m)}\left(y_i\right)\right)$$

The Log Score evaluates how much belief (probability) the forecast assigns to the actual observed outcome and is also a 'proper' score as it encourages honest forecasting. In the Online Appendix, we walk through the same examples from the CRPS example for the Log Score and the adjusted, binned Log Score (*ab-Log Score*) we introduce here. Here, it suffices to note that the $-log\left(0\right) = \infty$ and thus the Log Score returns an infinite penalty when any probability assigned to any actual value is zero. As we detail further below and in the Online Appendix, we solve this problem by first binning the fatality counts into a coarser resolution a priori, with sets of values mapped to bins in b, and then adding an a priori adjustment value $\omega > 0$ to each bin such that there is a finite maximum penalty. The (*ab-Log Score*) is calculated as the negative of the log of the adjusted-binned probability within the bin that the actual value occurs within:

$$ab\text{-}Log\ Score\left(f_i^{(m)}, y_i\right) = -log\left(\frac{\left(\sum_{z \in b_{w_i}} f_i^{(m)}\left(x = z\right)\right) \times n^{(samps)} + \omega}{n^{(samps)} + n^{(bins)}\omega}\right)$$

where w_i is the bin within which y_i falls. The value $n^{(samps)}$ is the number of samples used to calculate $f_i^{(m)}$ and $n^{(bins)} > 1$ is the number of bins.[9] Our adjusted, binned-Log Score uses a domain specific definition of focus and locality through the definition of the bin width (defined in the implementation details below). In addition, while the *ab-Log Score* is not a proper scoring rule on the original fine-grained sample space of fatality counts, an un-adjusted version is proper on the sample space of the bin-probabilities. In future work, we can experiment with using the probability directly as a score (i.e. linear).

Implementation: It is not always straightforward to practically represent $f_i^{(m)}\left(y_i = x\right)$ for all possible values of x. We utilize samples to represent forecasts, but these are necessarily finite, and thus will not cover

every possible (or even realized) value. Our binning and then adjustment for the *ab-Log Score* solves this problem without requiring competitors to submit forecasts in a completely different format. We simply use $\sum_{z \in b_{w_i}} f_i^{(m)}(x = z)$, where $f_i^{(m)}(x = z)$ is the proportion of samples that take on the value z.[10]

Binning allows us to utilize domain knowledge to calibrate what is close/local across the sample space of counts, without having probability mass at extreme values far away from the actual observation warp the score as in the CRPS. We view zero as a discrete value from the non-zero values (i.e. it receives its own bin) and then use increasing bin lengths to represent the dual facts that right-tail probabilities generally are smaller (i.e. we need to bin more values there) and that perceptions of what are similar types of fatality magnitudes are relatively coarser the larger the count. This leads to our binning scheme:

$$\{[0,0],[1,2],[3,5],[6,10],[11,25],[26,50],$$
$$[51,100],[101,250],[251,500],[501,1000],[1001,\infty)\}$$

In addition, we set $\omega = 1$, which adds one pseudo-sample to each bin. In future work, different sets of bins and ω values can be investigated with the forecasts the competition will generate.

With these settings our calculation of the *ab-Log Score* simplifies to

$$ab\text{-}Log\ Score_i^{(m)} = -log\left(\frac{\left(\sum_{z \in b_{w_i}} f_i^{(m)}(x = z)\right) \times 1000 + 1}{1011}\right)$$

Returning to the problematic infinite value in our simple example where $f_2^{(2)}(y_2 = 101) = 0$ from the raw Log Score, we can now see that the *ab-Log Score* value is simply $-log(501/1011) = .702$ because the .5 probability assigned by the forecast to the value $x = 120$ is within the same bin (101–250) as the actual value $y_2 = 101$, with all other values in that bin equal to zero. If the actual values were revealed to be in a different bin, to which the second forecast assigned a zero value (e.g. 26–50), then there would still be a finite, but large value, equal to $-log(1/1011) = 6.92$ given $\omega = 1$ and our other settings.[11]

Secondary metric II: Mean interval scores (MIS)

The mean interval score (MIS) evaluates both the calibration and the sharpness of the forecasts, and is set up to reward forecasts that can calculate prediction intervals (based on a lower and upper quantile) as narrow as possible while still ensuring that they cover the actual value with sufficient frequency. The score does not consider any mass of the predictive distribution outside of the interval, making it distinct from the CDF-based CRPS.

$$IS_i^{(m)} = \left(U_i^{(m)} - L_i^{(m)}\right) + \frac{2}{a}\left(L_i^{(m)} - y_i\right) \times \mathbb{1}\left(L_i^{(m)} > y_i\right) + \frac{2}{a}\left(y_i - U_i^{(m)}\right)\mathbb{1}\left(y_i > U_i^{(m)}\right)$$

where $U_i^{(m)}$ denotes the (upper) $1 - \frac{a}{2}$ prediction sample quantile and $L_i^{(m)}$ represents the (lower) $\frac{a}{2}$ prediction sample quantile, given that we consider the $(1-a) \times 100\%$ prediction interval (i.e. for a 95% prediction interval, $a = 0.05$). $\mathbb{1}(z)$ is the indicator function as defined in the evaluation and metrics section.

To compute the Mean Interval Score, the IS_i is averaged across instances for each model. The first term measures the interval width – which rewards sharpness and is only based on the prediction interval itself. The next set of terms measures coverage (or lack thereof) and penalizes forecasts by the weighted distance between the nearest interval limit and the actual value when the observed value is not covered by the interval. For example, to form the 90% interval for our first forecast on our first observation, where $f_1^{(1)}(0) = .9$, $f_1^{(1)}(101) = .1$, and $f_1^{(1)}(x) = 0 \forall x \notin 0,101$, we have $L_1^{(1)} = 0$ and $U_1^{(1)} = 101$ using the .05 and .95 quantiles. Because $y_1 = 0$, which is included in the interval, $IS_1^{(1)} = 101 - 0 = 101$. In the Online Appendix, we include a more detailed discussion of how the interval score is computed on the example forecasts and observations referenced above.

Implementation: It is not straightforward to calculate quantiles from a sample (Hyndman and Fan, 1996). We use the linear (Gumbel) method for interpolation, which is the default approach in the Python package NumPy, and a 90% prediction interval (a = 0.1). Since CRPS is already accounting for an aspect of the distance between the cumulative prediction mass and the observed value, we opted for a relatively wide prediction interval definition (90%). Through simulation, we found that the ability of forecasts with 1,000 samples to accurately provide estimates of quantiles for overdispersed distributions outside the 90% prediction interval tapers off quickly, which is why we did not push for an even wider interval definition (e.g. 99%). It is also at the tail end of the distribution that the choice of the quantile interpolation method matters the most. By setting the prediction interval to 90%, we aim to minimize the impact of the choice of interpolation method and

number of samples – which would have a potentially large effect on wider intervals – while still learning about the range of the most likely values that extend towards the tails of the count distribution – but not too far into the tails, where the estimates become less dependable.

Benchmark models and evaluation

Since the beginning of the challenge, we made available a couple of simple benchmark models to provide some context for the evaluation metrics and a baseline for comparison during the model development phase. Following the recommendations by the scoring committee, we added a few benchmark models later on as a complement to the initial benchmarks. However, we do not use the latter benchmarks – marked as ph ('post-hoc') – as scoring rule as they were not disclosed to the participants from the start. The post-hoc benchmarks are still useful for informational purposes, as a simple heuristic that facilitates the comparison across models. Importantly, we note that the contribution of Drauz and Becker (2024) has been based on the same logic of the *conflictology* benchmarks since the opening of the prediction challenge, before we decided to add this benchmark.[12]

The VIEWS_bm_exactly_zero benchmark (at both *cm* and *pgm* level) forecasts zero fatalities for all country-months or PRIO-GRID months. The VIEWS_bm_last_historical (*cm/pgm*) uses the last observed value as the point prediction for a given unit of analysis and for each of the months in the forecasting window, and generates a probability distribution by drawing from the Poisson distribution with the point prediction as mean and variance. The VIEWS_bm_ph_conflictology_country12 (*cm/pgm)* uses the set of values from each of the past 12 months for the same country as the forecast distribution for that country. Using empirical probability distributions as a benchmark has long been a common practice in meteorology (Murphy and Winkler, 1984). Similarly, the VIEWS_bm_ph_conflictology_neighbors12 (*pgm*) takes the values observed over the past 12 months in the cell as well as its first order neighbors as the 'draw' from the prediction distribution. The VIEWS_bm_conflictology_bootstrap240 (*cm/pgm*) randomly draws a set of values from the past 240 months (20 years) from any country in the world as the forecast for a given country.[13]

Table 3 shows evaluation scores for the four benchmark models described above, for each of the years 2018–2023 for which we have historical data, and each of the three metrics under consideration. We also show the average scores across the six years in the row labeled 'overall'. Table 3a shows the scores for the benchmark model VIEWS_bm_exactly_zero where all units are forecast as zero fatalities. Mean scores across all six years are 56.84 for CRPS, 1.59 for *ab-Log Score*, and 1,136.80 for MIS. From 2014 to 2019, state-based fatalities were declining, as reflected by the lower VIEWS_bm_exactly_zero CRPS scores in 2018–2019. The higher CRPS values for the VIEWS_bm_exact_zero model in 2021 are due to a decisive increase in fatalities recorded by the UCDP for that year: 2021 was the deadliest year since the Rwandan genocide, driven by the conflicts in Ethiopia and Ukraine (Pettersson et al., 2024). Relative to 2021, the UCDP reported lower deaths in 2022, driven by the de-escalation in Afghanistan and Yemen. The VIEWS_bm_exactly_zero model inherently fails to capture these variations in fatalities over time, and this is reflected in its CRPS scores. Nine out of the ten highest CRPS values are attributable to Ethiopia and Ukraine, and six of the nine occur in 2022. Yemen in November 2021 rounds out the top ten highest CRPS values. Moreover, although state-based fatalities have increased four-fold since 2020, the majority of country-months still do not record any fatalities. The sudden peaks in fatalities across country-months further contribute to the decline of the CRPS score for the VIEWS_bm_last_historical CRPS observed in 2021–2023.

Table 3b shows the scores for the VIEWS_bm_last_historical model that predicts that the violence observed in the last month (with available data) will continue unchanged (with some added uncertainty). For the first three years, this model does better than the exactly zero model, but for 2021–2023 it is even more surprised by the new wars than the exactly zero fatalities model. This supports examining the last six historical years for the models' evaluation, in order to smooth the influence of uneven pulses of events.

VIEWS_bm_ph_conflictology_country12 (Table 3c) is the strongest of the benchmarks – using the historical observations for the last 12 months as the prediction distribution, it performs much better on all metrics. The scores equal to 49.36 for CRPS, 0.65 for *ab-Log Score*, and 873.53 for MIS.

Table 3d scores the final *cm* benchmark, the VIEWS_bm_conflictology_bootstrap240 model that predicts that all country-months have the same probability distribution as the global record back to the late 1990s. Being just as uncertain and pessimistic as the exactly zero benchmark is confidently optimistic, it does considerably better. In terms of CRPS, it is actually the best model for 2022.

Table 4 scores the five benchmark models at the *pgm* level.

Table 3. Benchmark model evaluation, *cm* level, four benchmark models, 2018–2023.

(a) VIEWS_bm_exactly_zero

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	24.13	1.56	482.61
2019	23.02	1.56	460.38
2020	32.04	1.55	640.81
2021	87.34	1.61	1,746.78
2022	120.97	1.63	2,419.36
2023	53.54	1.61	1,070.86
Overall	56.84	1.59	1,136.80

(b) VIEWS_bm_last_historical

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	20.17	1.20	380.62
2019	9.48	1.05	172.69
2020	23.70	1.11	455.81
2021	85.61	1.23	1,690.71
2022	131.02	1.12	2,599.28
2023	678.96	1.12	13,523.46
Overall	158.16	1.14	3,137.09

(c) VIEWS_bm_ph_conflictology_country12

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	14.48	0.64	186.55
2019	9.15	0.61	89.06
2020	21.34	0.57	344.96
2021	76.85	0.69	1,435.55
2022	124.00	0.69	2,142.13
2023	50.36	0.68	1,042.92
Overall	49.36	0.65	873.53

(d) VIEWS_bm_conflictology_bootstrap240

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	23.58	1.12	454.09
2019	22.46	1.11	426.01
2020	31.42	1.12	606.00
2021	86.63	1.15	1,708.30
2022	120.25	1.15	2,380.74
2023	52.72	1.15	1,030.99
Overall	56.17	1.14	1,101.02

Table 4. Benchmark model evaluation, *pgm* level, five benchmark many distributions with the same mean that assign different plausibilities models, 2018–2023.

(a) VIEWS_bm_exactly_zero

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	0.14	0.09	2.89
2019	0.12	0.09	2.31
2020	0.13	0.11	2.64
2021	0.94	0.12	18.80
2022	1.14	0.12	22.75
2023	0.22	0.12	4.47
Overall	0.45	0.11	8.98

(b) VIEWS_bm_last_historical

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	0.39	0.12	7.15
2019	0.14	0.11	2.62
2020	0.16	0.12	2.99
2021	0.97	0.13	19.08
2022	1.46	0.15	28.53
2023	9.75	0.15	193.97
Overall	2.15	0.13	42.39

(c) VIEWS_bm_ph_conflictol_country12

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	0.19	0.08	2.83
2019	0.12	0.08	1.89
2020	0.13	0.08	2.07
2021	0.93	0.09	17.87
2022	1.14	0.10	22.28
2023	0.52	0.10	13.22
Overall	0.51	0.09	10.03

(d) VIEWS_bm_ph_conflictol_neighbors12

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	0.15	0.08	3.06
2019	0.11	0.08	1.88
2020	0.12	0.08	2.12
2021	0.93	0.10	18.11
2022	1.13	0.10	22.48
2023	0.25	0.10	4.03
Overall	0.45	0.09	8.61

(e) VIEWS_bm_conflictology_bootstrap240

Year	*CRPS*	*ab-Log Score*	*MIS*
2018	0.14	0.09	2.89
2019	0.12	0.10	2.31
2020	0.13	0.11	2.64
2021	0.94	0.12	18.80
2022	1.14	0.12	22.75
2023	0.22	0.12	4.47
Overall	0.45	0.11	8.98

The VIEWS_bm_exactly_zero model (Table 4a) has average scores of 0.45 for CRPS, 0.11 for *ab-Log Score*, and 8.98 for MIS. Just as at the country level, the scores are much worse for the years 2021–2022 than for the first three years. The extreme zero-inflation of the fatality count at the *pgm* level makes the exactly-zero model hard to beat. The VIEWS_bm_last_historical obtains worse scores across all metrics. The VIEWS_bm_ph_conflictology_country12 also performs worse than the optimistic VIEWS_bm_exactly_zero model for both CRPS and MIS, but slightly better for *ab-Log Score*. The variant of the conflictology benchmark model that also includes the observations for the immediate neighbors of each grid cell, on the other hand, does better than the exactly zero model. The fundamentally uncertain VIEWS_bm_conflictology_bootstrap240 model performs very similarly to the exactly zero model – since 99.5% of the historical values are zero, the predictions are almost identical.

Finding new collective wisdom: Ensembling the probabilistic forecasts

One of the benefits of the competition format is aligning the efforts of experts on a common set of tasks. This opens up the possibility of gleaning new collective insights from the overlap and joint contributions of the information the teams provide. The practice of combining discrete but related forecasts together is known as ensembling (Hegre et al., 2023) or prediction synthesis. We will explore how to usefully combine the probabilistic forecasts into ensembles in the follow-up evaluation article. There are both theoretical (Page, 2018) and practical (Vesco et al., 2022) reasons to expect that combinations of diverse perspectives can produce knowledge in complicated environments – of which political violence surely qualifies. However, combining probabilistic forecasts is a distinct computational task in comparison to averaging expected values across point forecasts as has been done in the past. In fact, we are not calculating a new expected value from the expected values, but instead computing a new, mixed probabilistic forecast distribution from individual constituent forecast distributions.

While a full discussion is beyond the scope of the current work, we plan to follow a two-pronged strategy. First, we will use the inverse of the competitors' CRPS values for the test sets, rescaled to sum to 1, as weights for sampling probabilities to draw forecast simulations across the competitors' contributions for each observation. This will provide a new ensemble of simulations/samples that represents the wisdom of the competitors. We will also explore truncating the forecasts included in the probabilistic ensemble by a minimum CRPS threshold, as well as by including only the models that finish in the top five ranking in any of the three main metrics (i.e. CRPS, *ab-Log Score*, or MIS). Second, and more aspirationally, we will build probabilistic versions of random forest and boosting methods. These will output samples from forecast distributions instead of point values as is conventional.[14] The new probabilistic algorithms will be used to dynamically combine the input forecasts and simulations from the competitors as a form of probabilistic stacking. We can then additionally assess models based on their unique or distinct contribution relative to the full set of all models, or their diversity relative to the collective contribution as measured by the ensemble, as well as compare when and where certain models contribute more or less to the ensemble of all forecasts.

Conclusions

This article presents the structure of the VIEWS 2023/24 prediction challenge, motivates the need to go beyond point predictions, summarizes the main models that have been presented to the challenge, and illustrates the evaluation principles and metrics used to score the models.

The forecasts for the true future have been posted, together with a pre-print of this paper.[1] The evaluation of the models will be reported in a second article from the challenge, to be finalized in the fall of 2025, after the end of the true future forecasting window.

Our experience from the last prediction competition (Hegre et al., 2022) was that fielding these sets of coordinated tasks generated new knowledge not only from an analysis of the forecasts themselves – focusing talented research teams on a small set of similar problems – but also by highlighting new, open questions related to evaluation and infrastructure. In the case of the current challenge, the time is ripe to improve probabilistic forecasts in peace science as well as the systems that can generate them in near-real-time. Specifically, Kolassa (2016) and Kolassa (2020) highlight that evaluation of forecasts depends on the context within which the data are collected and used. Our competition can illuminate insights on how our chosen evaluation metrics discriminate between models that are tasked with forecasting political violence at distinct spatial resolutions, while the shortcomings and surprises that are revealed are avenues for new innovations (Colaresi and Mahmood, 2017). In addition, this joint challenge provides ongoing research

with avenues to usefully and practically represent probabilistic forecasts and communicate uncertainty in policy-relevant contexts. To sustain the challenge, the organizers have been supplying the first shared infrastructure of its kind to the peace science community – moving beyond point predictions (and even intervals) to the presentation of forecast distributions. With this investment, the research teams interested in these topics can integrate our open-source tools into their workflows, as well as learn what works and does not work to accelerate their own discoveries into the future. This effort provides the research community with new baseline performance to compete against, evaluation criteria to measure with, data to collect and utilize, model representations to deploy and fit, and insights about the future to reduce practical harm and increase scientific understanding.

Data availability statement

The data that were made available to the contributors can be accessed at https://viewsforecasting.org/research/prediction-challenge-2023/, with a codebook and instructions to the participants. All the submitted forecasts are available at the same location. The codes used to collect, clean, and evaluate the forecasts, and to generate the benchmark model forecasts, are available at https://github.com/prio-data/prediction_competition_2023. The data, codebook and supplementary information are also available via https://www.prio.org/jpr/datasets/

Acknowledgements

Great thanks to the PREVIEW team and especially to Ida Bauer and Thomas Mayer for their help in organizing the workshop in Berlin.

Funding

The VIEWS Prediction Challenge 2023/2024 received financial support from the German Ministry for Foreign Affairs. The VIEWS team was also funded by the European Research Council under Horizon Europe (Grant agreement no. 101055176, ANTICIPATE) and the Research Council of Norway (Grant agreement no. 334977, UFFAC), and the Center for Advanced Study (CAS) at The Norwegian Academy of Science and Letters.

ORCID iDs

Håvard Hegre https://orcid.org/0000-0002-5076-0994
Paola Vesco https://orcid.org/0000-0002-0368-0633
Alexa Timlick https://orcid.org/0009-0005-6562-6279
AngelicaLindqvist-McGowan https://orcid.org/0000-0002-8132-3551
Tobias Bohne https://orcid.org/0009-0009-6478-432X
Christoph Dworschak https://orcid.org/0000-0003-0196-9545
Kristian Skrede Gleditsch https://orcid.org/0000-0003-4149-3211
Sonja Häffner https://orcid.org/0000-0001-6914-9709
Luca Macis https://orcid.org/0000-0003-0001-304X
Alexandra Malaga https://orcid.org/0009-0003-9439-1443
Marius Mehrl https://orcid.org/0000-0002-5825-9256
Nils W Metternich https://orcid.org/0000-0001-8757-0409
Daniel Mittermaier https://orcid.org/0009-0003-6284-5017
David Muchlinski https://orcid.org/0000-0002-2195-1943
Christian Oswald https://orcid.org/0000-0001-8291-9544
David Randahl https://orcid.org/0000-0003-1069-6067
Benjamin Seimon https://orcid.org/0009-0006-2655-4097
Julian Walterskirchen https://orcid.org/0000-0002-5587-7075

Notes

1. A pre-print of this paper was published at ArXiv.org: https://arxiv.org/abs/2407.11045v1 and at https://predcomp.viewsforecasting.org at the start of the true future prediction window. The models are subject to continuous evaluation at https://predcomp.viewsforecasting.org as soon as the data on actual conflicts are released by the UCDP.
2. This is because there are infinitely many distributions with the same mean that assign different plausibilities to zero-observations and very large observations.
3. For the details of the submission procedures, see the Online Appendix as well as https://viewsforecasting.org/research/prediction-challenge-2023/.
4. In the Hegre et al. (2022) competition, the target was specified in log form. Adding 1 to the predicted number of fatalities 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 is likely to – and usefully so – reward models that are willing to move up above 0. We may still calculate some auxiliary evaluation metrics based on a log-transformed version in the future.

5. Contributors may decide whether they want to submit identical predictions for each month, or fine-tune predictions separately for each of them.
6. This geographic scope is dictated by the VIEWS infrastructure which is currently available for Africa and Middle-East only at the PRIO-GRID level. An expansion of the system to global forecasts at the PRIO-GRID level is ongoing and will likely be in place in 2025. 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 et al., 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.
7. 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 et al. (2009).
8. For simplicity we include two observations here and constant forecasts. However, to be clear there are many more zeros than non-zeros in the fatalities data, as noted, and the forecasts we score, outside of a few baseline models, vary across observations.
9. In the Online Appendix we also discuss further why we did not choose to only assign ω to a bin that holds the actual value and where the forecast has zero probability.
10. There are other ways to calculate this probability of a value from samples, but they are less simple and often less useful. For example, Krüger et al. (2021) provides evidence that kernel density estimation is an unstable approach for the Log Score, and this is particularly true for zero-inflated data such as in our context. Our approach does necessitate that we have the same number of samples for each forecast. In the rare case where we do not receive forecasts in a format that yields 1,000 samples, we up-sample the forecasts to generate 1,000 simulations. The use of a consistent number of samples and a value of ω set a priori ensures that the the addition of each binned outcome is fair in the sense of Siegert et al. (2019).
11. In the rare case where we receive a forecast that is formatted in such a way that we do not have or cannot generate 1,000 samples with the rules denoted above, we utilize scipy.signal.resample to generate 1,000 samples.
12. As such, we highlight that they were the first participants to apply this idea in this challenge and in our field. If this benchmark turns out to do well, their model is also likely to do well, and will be duly credited for their innovation.
13. The forecasted distribution is consequently similar for all countries. The model should be well calibrated at a global level but perform very poorly for individual country-months.
14. We have already developed prototype versions of probabilistic random forest models for this purpose.

References

Bazzi S, Blair RA, Blattman C, et al. (2022) The promise and pitfalls of conflict prediction: Evidence from Colombia and Indonesia. *Review of Economics and Statistics* 104(4): 764–779.

Blair RA and Sambanis N (2020) Forecasting civil wars: Theory and structure in an age of 'big data' and machine learning. *Journal of Conflict Resolution* 64(10): 1885–1915.

Bodentien T and Rüter L (2024) Forecasting monthly fatalities via a negative binomial distribution and comparison with a hurdle model and neural networks. Available at: https://viewsforecasting.org/wp-content/uploads/BodentienVIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Bracher J, Ray EL, Gneiting T, et al. (2021) Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology* 17(2): 1–15.

Brandt PT (2024) Bayesian density forecasts for VIEWS. Available at: https://viewsforecasting.org/wp-content/uploads/Brandt_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Brandt PT, Freeman JR and Schrodt PA (2011) Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science* 28(1): 41–64.

Brandt PT, Schrodt PA and Freeman JR (2014) Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting* 30(4): 944–962.

Carvalho A (2016) An overview of applications of proper scoring rules. *Decision Analysis* 13(4): 223–242.

Chiba D and Gleditsch KS (2017) The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data. *Journal of Peace Research* 54(2): 275–297.

Colaresi M and Mahmood Z (2017) Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research* 54(2): 193–214.

Czado C, Gneiting T and Held L (2009) Predictive model assessment for count data. *Biometrics* 65(4): 1254–1261.

D'Orazio V (2024) Probabilistic conflict forecasting with automated machine learning. Available at: https://viewsforecasting.org/wp-content/uploads/Dorazio_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Dorff C, Gallop M and Minhas S (2020) Networks of violence: Predicting conflict in Nigeria. *Journal of Politics* 82(2): 476–493.

Drauz S and Becker F (2024) Probabilistic forecasting of conflict fatalities: Historical quantiles vs. Bayesian penalized regression. Available at: https://viewsforecasting.org/wp-content/uploads/Drauz_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Du H (2021) Beyond strictly proper scoring rules: The importance of being local. *Weather and Forecasting* 36(2): 457–468.

Fritz C, Dworschak C and Mehrl M (2024) Predicting uncertainty in stages: Using a semiparametric hierarchical hurdle model for predicting distributions of conflict fatalities. Available at: https://viewsforecasting.org/wp-content/uploads/Fritz_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Gleditsch KS and Ward MD (1999) A revised list of independent states since the Congress of Vienna. *International Interactions* 25(4): 393–413.

Gleditsch KS, Klebe F and Metternich NW (2024) Random forest predictions with dyad features. Available at: https://viewsforecasting.org/wp-content/uploads/Gleditsch_2nd_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Gneiting T and Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477): 359–378.

Gneiting T, Balabdaoui F and Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2): 243–68.

Goldstone JA, Bates RH, Epstein DL, et al. (2010) A global model for forecasting political instability. *American Journal of Political Science* 54(1): 190–208.

Hegre H, Allansson M, Basedau M, et al. (2019) ViEWS: A political Violence Early Warning System. *Journal of Peace Research* 56(2): 155–174.

Hegre H, Bell C, Colaresi M, et al. (2021) ViEWS2020: Revising and evaluating the ViEWS political Violence Early-Warning System. *Journal of Peace Research* 58(3): 599–611.

Hegre H, Colaresi M, Nordenving S, et al. (2023) Ensembling and calibration in VIEWS. *VIEWS documentation paper series, Fatalities002*. Available at: https://viewsforecasting.org/wp-content/uploads/VIEWS_documentation_Ensembling_Fatalities002.pdf (accessed 29 October 2024).

Hegre H, Croicu M, Eck K, et al. (2020) Introducing the UCDP Candidate Events Dataset. *Research & Politics* 7(3): 2053168020935257.

Hegre H, Karlsen J, Nygård HM, et al. (2013) Predicting armed conflict 2010–2050. *International Studies Quarterly* 57(2): 250–270.

Hegre H, Vesco P and Colaresi M (2022) Lessons from an escalation prediction competition. *International Interactions* 48(4): 521–554.

Hyndman RJ and Fan Y (1996) Sample quantiles in statistical packages. *The American Statistician* 50(4): 361–365.

Kolassa S (2016) Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting* 32(3): 788–803.

Kolassa S (2020) Why the 'best' point forecast depends on the error or accuracy measure. *International Journal of Forecasting* 36(1): M4 Competition, 208–211.

Krüger F, Lerch S, Thorarinsdottir TL, et al. (2021) Predictive inference based on Markov chain Monte Carlo output. *International Statistical Review* 89(2): 274–301.

Macis L, Tagliapietra M, Siletti E, et al. (2024) Predicting fatalities with pre-trained temporal transformers: A time series regression approach. Available at: https://viewsforecasting.org/wp-content/uploads/Unito_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Málaga A, Mueller H, Rauh C, et al. (2024) Predicting fatalities using newspaper text. Available at: https://viewsforecasting.org/wp-content/uploads/ConflictForecast_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Mittermaier D, Bohne T and Hofer M (2024) Forests of uncertaint(r)ees: Using tree-based ensembles to estimate probability distributions of future conflict. Available at: https://viewsforecasting.org/wp-content/uploads/Mittermaier_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Muchlinski D and Thornhill C (2024) A zero-inflated Poisson generalized additive model for forecasting conflict fatalities. Available at: https://viewsforecasting.org/wp-content/uploads/Muchlinski_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Mueller H and Rauh C (2018) Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review* 112(2): 358–375.

Murphy AH and Winkler RL (1984) Probability forecasting in meterology. *Journal of the American Statistical Association* 79(387): 489.

Page SE (2018) *The Model Thinker: What You Need to Know to Make Data Work for You*, 1st edn. New York, NY: Basic Books.

Pettersson T, Davies S, Engström G, et al. (2024) Organized violence 1989–2023, and the prevalence of organized crime groups. *Journal of Peace Research* 61(4): 673–693.

Randahl D and Vegelius J (2024) Forecasting fatalities from state based conflicts using Markov models. Available at: https://viewsforecasting.org/wp-content/uploads/Randahl_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Schincariol T, Frank H and Chadefaux T (2024) Temporal patterns in conflict prediction: An improved shape-based approach. Available at: https://viewsforecasting.org/wp-content/uploads/Pace_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27(3): 379–423.

Siegert S, Ferro CAT, Stephenson DB, et al. (2019) The ensemble-adjusted ignorance score for forecasts issued as normal distributions. *Quarterly Journal of the Royal Meteorological Society* 145(February): 129–139.

Smith LA, Suckling EB, Thompson EL, et al. (2015) Towards improving the framework for probabilistic forecast evaluation. *Climatic Change* 132: 31–45.

Tollefsen AF, Strand H and Buhaug H (2012) PRIO-GRID: A unified spatial data structure. *Journal of Peace Research* 49(2): 363–374.

Vesco P, Hegre H, Colaresi M, et al. (2022) United they stand: Findings from an escalation prediction competition. *International Interactions* 48(4): 1–37.

Walterskirchen J, Häffner S, Oswald C, et al. (2024) Taking time seriously: Predicting conflict fatalities using temporal fusion transformers. Available at: https://viewsforecasting.org/wp-content/uploads/tft_ccew_VIEWSPredictionChallenge2023.pdf (accessed 29 October 2024).

Weidmann NB, Kuse D and Gleditsch KS (2010) The geography of the international system: The CShapes dataset. *International Interactions* 36(1): 86–106.

World Bank Group and United Nations (2017) *Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict. Main Messages and Emerging Policy Directions*. Washington, DC: World Bank.

HÅVARD HEGRE, b. 1964, Dr Philos in Political Science (University of Oslo, 2004); Research Professor, Peace Research Institute Oslo (2005–present) and Department of Peace and Conflict Research, Uppsala University (2013–present).

PAOLA VESCO, b. 1990, PhD in Science and Management of Climate Change (Ca' Foscari University, 2020); Senior Researcher, Peace Research Institute Oslo (2024–present); Researcher, Department of Peace and Conflict Research, Uppsala University (2020–present).

MICHAEL COLARESI, b. 1976, PhD in Political Science (Indiana University, 2002); William S. Dietrich II Chair of Political Science, University of Pittsburgh (2017–present); External Researcher, Peace Research Institute Oslo (PRIO) (2024–present).

JONAS VESTBY, b. 1984, PhD in Political Science (University of Oslo, 2018); Senior Researcher, Peace Research Institute Oslo (PRIO) (2018–present).

ALEXA TIMLICK, b. 1998, MSSc in Peace and Conflict Studies (Uppsala University, 2024); Research Assistant, Peace Research Institute Oslo (PRIO) (2024–present).

NOORAIN SYED KAZMI, b. 1988, MSSc in Industrial IT and Automation (University of South-Eastern Norway, 2023); Research Assistant, Peace Research Institute Oslo (PRIO) (2023–24).

ANGELICA LINDQVIST-MCGOWAN, b. 1993, MA in Holocaust and Genocide Studies (Uppsala University, 2019); Operations and Outreach Manager, Department of Peace and Conflict Research, Uppsala University (2019–present).

FRIEDERIKE BECKER, b. 1996, MSc in Applied Statistics (Göttingen University, 2022); Institute of Statistics, Karlsruhe Institute of Technology (2022–present).

MARCO NICOLA BINETTI, b. 1992, PhD in Government (University of Essex, 2021); Postdoctoral Researcher, University of Bremen (2024–present).

TOBIAS BODENTIEN, b. 1999, BSc Industrial Engineering (KIT Karlsruhe, 2023); PhD student, Institute of Statistics, Karlsruhe Institute of Technology.

TOBIAS BOHNE, b. 1992, MA International Relations and Development Policy (University of Duisburg-Essen, 2022); Research Associate, Center for Crisis Early Warning, University of the Bundeswehr Munich (2022–present).

PATRICK T BRANDT, b. 1972, PhD in Political Science (Indiana University, Bloomington, 2001); Professor of Political Science, University of Texas at Dallas (2015–present).

THOMAS CHADEFAUX, b. 1980, PhD in Political Science (University of Michigan, 2009); Professor, Department of Political Science, Trinity College Dublin (2021–present).

SIMON DRAUZ, b. 2000, BSc in Industrial Engineering and Management (Karlsruhe Institute of Technology, 2024), MSc in Data Science (LMU Munich, 2024–present).

CHRISTOPH DWORSCHAK, b. 1993, PhD in Government (University of Essex, 2020); Lecturer in Quantitative Political Science, University of York (2022–present).

VITO D'ORAZIO, b. 1985, PhD in Political Science (Pennsylvania State University, 2013); Woodburn Associate Professor of Political Science and Data Science, West Virginia University (2022–present); Research Fellow, Modern War Institute at West Point (2024–25).

HANNAH FRANK, b. 1994, MSSc in Peace and Conflict Studies (Uppsala University, 2021); PhD candidate, Department of Political Science, Trinity College Dublin (2021–present).

CORNELIUS FRITZ, b. 1993, Dr. rer. nat. (LMU Munich, 2022); Assistant Professor in Statistics, Trinity College Dublin (2024–present).

KRISTIAN SKREDE GLEDITSCH, b. 1971, PhD in Political Science (University of Colorado, Boulder, 1999);

Regius Professor of Political Science, University of Essex (2005–present); Research Associate, Peace Research Institute Oslo (2003–present).

SONJA HÄFFNER, b. 1995, MSc in Economics (University of Regensburg, 2021); Junior Machine Learning Developer, Peace Research Institute Oslo (2025–present); Research Associate, Center for Crisis Early Warning, University of the Bundeswehr (2021–23).

MARTIN HOFER, b. 1989, PhD in Mathematics (LMU Munich, 2018); Team Lead Data Science and Software Development, Center for Crisis Early Warning, University of the Bundeswehr Munich (2021–23).

FINN L KLEBE, b. 1996, PhD candidate, University College London (2021–present).

LUCA MACIS, b. 1992, Master's degree in Mathematics (Stochastics and Data Science, University of Turin, 2022); PhD candidate in Economics and Statistics, University of Turin (2023–present).

ALEXANDRA MALAGA, b. 1991, MSc in Data Science and Decision Making (Barcelona School of Economics, 2022), MSc in Economics (Universidad del Pacífico - Perú, 2015); Data Scientist, Institute for Economic Analysis CSIC, Barcelona (2022–present).

MARIUS MEHRL, b. 1993, PhD in Government (University of Essex, 2020); Lecturer in Quantitative International Relations, University of Leeds (2022–present).

NILS W METTERNICH, b. 1979, PhD in Government (University of Essex, 2011); Professor of Political Science, University College London (2021–present).

DANIEL MITTERMAIER, b. 1990, MA in Politics and Public Administration (University of Konstanz, 2018); Research Associate, Center for Crisis Early Warning, University of the Bundeswehr Munich (2021–present).

DAVID MUCHLINSKI, b. 1984, PhD in Political Science (Arizona State University, 2013); Assistant Professor of International Affairs, Sam Nunn School of International Affairs, Georgia Institute of Technology (2018–present).

HANNES MUELLER, b. 1978, PhD in Economics (London School of Economics, 2008); Tenured Scientist, Institute for Economic Analysis CSIC, Barcelona (2008–present); Program Director for the Data Science and Decision Making MSc, Barcelona School of Economics (2008–present); Centre for Economic Policy Research (2015–present).

CHRISTIAN OSWALD, b. 1986, PhD in Political Science (Trinity College Dublin, 2023); Research Associate, Center for Crisis Early Warning, University of the Bundeswehr Munich (2022–present).

PAOLA PISANO, b. 1977, PhD in Innovation (University of Turin, 2007); Professor of Economic and Innovation Management, University of Turin (2007–present).

DAVID RANDAHL, b. 1990, PhD in Peace and Conflict Research (Uppsala University, 2022); Researcher, Department of Peace and Conflict Research, Uppsala University (2022–present).

CHRISTOPHER RAUH, b. 1983, PhD in Economics (Universitat Autònoma de Barcelona); Institute for Economic Analysis CSIC (2024–present); Associate Senior Researcher, Peace Research Institute Oslo (2023–present); Professor of Economics and Data Science, University of Cambridge (2022–24).

LOTTA RÜTER, b. 1995, MSc in Economathematics (Karlsruhe Institute of Technology, 2021); PhD student in Statistical Methods and Econometrics, Karlsruhe Institute of Technology (2021–present).

THOMAS SCHINCARIOL, b. 1997, MSc in Control and Computer Engineering (University of Technology of Troyes, 2021); PhD candidate, Department of Political Science, Trinity College Dublin (2021–present).

BENJAMIN SEIMON, b. 1997, MSc in Data Science and Decision Making (Barcelona School of Economics, 2022); Data Scientist, Fundació Economia Analítica (2022–present).

ELENA SILETTI, b. 1972, PhD in Statistics (University of Milan-Bicocca, 2008); Assistant Professor, University of Turin, (2022–present).

MARCO TAGLIAPIETRA, b. 1997, Master's degree in Mathematics (Stochastics and Data Science, University of Turin, 2022); PhD candidate in Economics and Statistics at the University of Turin (2023–present).

CHANDLER THORNHILL, b. 1994, MSc in International Affairs (Georgia Institute of Technology, 2021); PhD student in International Affairs, Science and Technology, Georgia Institute of Technology (2021–present).

JOHAN VEGELIUS, b. 1981, PhD in Statistics (Uppsala University 2022), PhD in physics (2011); Postdoc researcher, Department of Medical Sciences, Uppsala University (2022–present).

JULIAN WALTERSKIRCHEN, b. 1992, MSc in International Relations (University of Essex, 2017); PhD student, University of Gothenburg (2024–present); Research Associate, Center for Crisis Early Warning, University of the Bundeswehr Munich (2021–24).