ISSN 1831-9424

European Commission

JRC TECHNICAL REPORT

The Global Conflict Risk Index 2022: Revised Data and Methods

Schvitz, G., Corbane, C., Van Damme, M., Galariotis, I., and Valli, I.

2022

Joint Research Centre

EUR 31330 EN

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact Information
Name: Guy Schvitz
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 267, 21027 Ispra (VA), Italy
Email: guy.schvitz@ec.europa.eu
Tel.:

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC 131326

EUR 31330 EN

PDF ISBN 978-92-76-60169-2 ISSN 1831-9424 doi:10.2760/041759 KJ-NA-31-330-EN-N

Luxembourg: Publications Office of the European Union, 2022

How to cite this report: Schvitz, G., Corbane, C., Van Damme, M., Galariotis, I. and Valli, I., *The Global Conflict Risk Index 2022: Revised Data and Methods*, Publications Office of the European Union, Luxembourg, 2022, doi:10.2760/041759, JRC 131326

Contents

Abstract

This report introduces the revised Global Conflict Risk Index (GCRI) and documents the changes and improvements made in the latest update. Our work involved a thorough revision of the GCRI's input data and methodology that resulted in the following changes: We (1) adopted a new set of conflict definitions for the outcome variable; (2) significantly reduced missing values by replacing several input data sources; (3) selected a new modelling framework for conflict intensities following a systematic comparison of 14 probability and intensity models; and (4) improved the GCRI's forecasts by incorporating short-term projections of predictor variables. We demonstrate that these revisions improved the GCRI's predictive performance. We conclude with a discussion of limitations and tasks for further research.

Acknowledgements

The European Commission's Joint Research Centre (JRC) would like to thank Céline Aucouturier from the Service for Foreign Policy Instruments (FPI) for financing this project, as well as Jonas Claes, Anja Palm, Massimo Farrugia, Marco Marzi, Pavla Danisova, and Dylan Macchiarini Crosson from the European External Action Service (EEAS) Conflict Prevention and Mediation Support Division for their unwavering support of the development of the Global Conflict Risk Index (GCRI). We also thank Sepehr Marzi (JRC.E.1), Arthur Hrast Essenfelder (JRC.E.1) and Jeremy Pal (CMCC) for their assistance in gathering and processing climate data, and thank Sarah Weiler (JRC.E.1) for editing this document.

Authors

Schvitz Guy[1], Corbane Christina[1], Van Damme Marie-Sophie[2], Galariotis Ioannis[3], Valli Igor[4]

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy.
[2] Piksel Ltd Italian Branch, Via Breda 176, 20126 Milano, Italy.
[3] Unisystems S.A, Rue Edward Steichen 26, Luxembourg, L-2540.
[4] Unisystems S.A, Via Michelangelo Buonarroti 39, 20145 Milano, Italy

1 Introduction

This report introduces the revised Global Conflict Risk Index (GCRI) and summarizes the changes made in the latest update. The GCRI is a quantitative model of conflict risk that serves as the main input into the EU's Conflict Early Warning System (EU Conflict EWS), and is funded by the Service for Foreign Policy Instruments (FPI). The model was originally developed in 2014 at the Joint Research Centre (JRC) and has since been updated and revised on a yearly basis, in close collaboration with the European External Action Service (EEAS) and the FPI.

The revised GCRI estimates the probability and intensity of violent internal conflict in 138 countries around the world over the next 4 years. These estimates are based on 22 predictor variables that represent known drivers of conflict and are grouped into 6 risk areas: Political, security, social, economy, geography & environment, and demographics. Each variable is taken from open-source datasets, which are shown in Table 1. The GCRI probability and intensity models are trained on a dataset that covers 175 countries from 1991 to the present. The final output of the GCRI consists of a regional ranking of countries based on their estimated overall conflict risk. The main steps involved in the analysis are illustrated in Figure 1.

1.1 Changes in the revised GCRI

The field of conflict forecasting is still at a relatively early stage and has evolved rapidly in the years since the GCRI's original inception. Among other developments, many new datasets have become available and increasingly sophisticated models have been adopted. In general, many new conflict forecasting projects were launched and we can benefit from new insights from the most recent wave of research. Therefore, we have conducted a thorough review of the GCRI's data sources and methodology in order to further improve the model's accuracy and reliability. This resulted in the following changes:

- **Conflict typology:** We have adopted a new conflict typology that captures a broader spectrum of violent conflict, enabling a more comprehensive understanding of internal conflict risk.
- **Data sources and variables:** The GCRI previously faced high rates of missing data. We replaced several input data sources, which helped to significantly reduce missing values and to minimize the model's reliance on imputation. In addition, we revisited the GCRI's variable selection, which led us to remove 3 variables from the model while adding 1 new variable.
- **Statistical modelling:** The previous GCRI used logistic regressions to estimate conflict probabilities and used linear models for conflict intensities. Other models exist that may be more suitable to deal with important features of the data, such as temporal dependence, cross-country heterogeneity, and the rare nature of conflict and skewed distribution of fatalities. Following a comparison of 14 models, we selected a new intensity model but decided to keep using the previous probability model, which performed comparatively well.
- **Forecasting methods:** A main challenge in forecasting is that we do not yet know the future values of predictor variables. Instead of simply extending current values into the future, we started to explore approaches to projecting the likely trajectory of predictor variables. We demonstrate that this approach is superior to the one used previously, but note that there still remains further room for improvement.

In the following sections, we discuss each of these changes in more detail. The report is organized as follows: First, we introduce the new conflict typology that defines the main outputs of the model. Second, we discuss the revised input dataset and variable selection. Third, we present the results of our model comparison exercise, based on which we selected the current set of GCRI models. Fourth, we show initial results demonstrating that projecting independent variables into the near future can improve the model's accuracy. Fifth, we evaluate the revised GCRI's performance compared to the previous version. Finally, we summarize the progress made over the previous year and conclude with a discussion of limitations and tasks for future research.

Dimension	Component	Variable	Source
Political	Regime type	Democracy	V-DEM
	Regime performance	State capacity	V-DEM
		Repression	V-DEM
		Corruption	V-DEM
Security	History of conflict	Recent internal conflict	UCDP
		Years since last conflict	UCDP
	Current conflict situation	Neighboring conflict	UCDP
		Homicide rate	IHME
Social	Social cohesion and diversity	Female empowerment	V-DEM
		Ethnic exclusion	EPR
		Transnational ethnic ties	EPR
Economy	Development and distribution	GDP per capita, log	World Bank
		Income inequality	WID
		Trade openness	World Bank
		Oil exports	World Bank
	Provisions and employment	Food security	FAO
		Unemployment	World Bank
Geography - Environment	Environment	Droughts	SPEI/CSIC
		Temperature change	FAO
Demographics	Demographics	Population, log	UN
		Youth bulge	UN
		Child mortality	World Bank

Table 1: GCRI variables and data sources

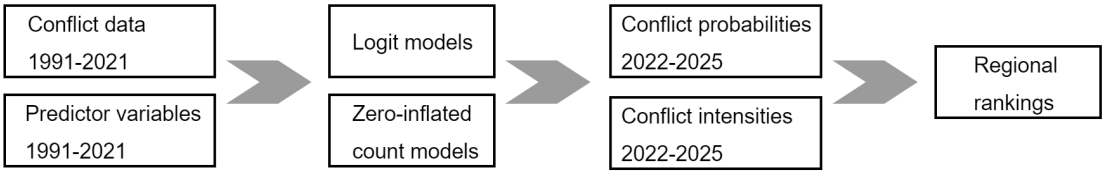

Figure 1: Overview of the GCRI workflow: From raw data to conflict risk.

2 New conflict typology

Following extensive consultations with EEAS and FPI, we have adopted a new conflict typology that covers a broader range of conflicts and enables a more comprehensive assessment of conflict risk than the one used previously: The old distinction between *"National Power"* and *"Sub-National"* conflicts was replaced by 3 new categories. In addition, the final ranking of countries is no longer done based on a single subcategory, but instead takes into account the overall risk of all 3 conflict types combined.

2.1 Moving beyond National Power vs. Sub-National conflict

Violent conflict comes in many different forms. The ongoing conflicts in Ethiopia, Myanmar or Ukraine are in many ways different from the Mexican drug war or the recent herder-farmer conflicts in the Sahel region. A key difference between these conflicts is that they revolve around different issues and are fought between different types of actors. Previous research has also shown that different types of conflict can have different causes (e.g., Buhaug and Rød, 2006), that some types of conflict last longer, while others are more fatal, and that some are more common in some regions and historical contexts than in others (Fearon, 2004, Kalyvas and Balcells, 2010, Pettersson et al., 2021).

For these reasons, conflict researchers commonly analyze different patterns of violence using separate theoretical and empirical models. For example, civil war studies often distinguish between wars fought over central government power and those fought over secession or regional autonomy. The GCRI has so far adopted the same approach, distinguishing between internal conflicts over *"National Power" (NP)* and so-called *"Sub-National" (SN)* conflicts that revolve around separatism. While this distinction is very useful to classify civil wars, it is less suitable for other types of conflict that have become increasingly common in recent years. Many recent conflicts are fought between local actors and are not related to national power struggles or separatism. As a result, they do not fit well into the NP-SN categorization.

Previously, the GCRI assigned all conflicts that do not qualify as NP to the SN category. As a result, the vast majority of conflicts were designated as SN, although many of them did not match this definition. Most importantly, the GCRI previously produced separate risk estimates for NP and SN conflicts, but generated its final ranking of countries based on the NP dimension alone. Given that SN conflicts are much more frequent (see Figure 2), this approach likely results in an incomplete understanding of conflict risk. To address these challenges, we have adopted a new typology that covers the following 3 conflict types:

- **State-based conflict (SBC)**: Armed conflict between two organized groups, one of which is a state government. The use of armed force results in at least 25 battle-related deaths per year (Gleditsch et al., 2002, p. 619).
- **Non-state conflict (NSC)**: Armed conflict between two groups, neither of which is the state. This includes conflict between rebel groups and militias, fighting between supporters of different political parties as well as conflict between social groups, usually identified along ethnic or religious lines. Fighting results in at least 25 battle-related deaths per year (Sundberg et al., 2012, p. 352).
- **One-sided violence (OSV)**: Direct and deliberate killing of civilians, perpetrated either by a state government or an armed group, resulting in at least 25 battle-related deaths per year (Eck and Hultman, 2007, p. 233).

These mutually exclusive categories and their definitions are taken from the Uppsala Conflict Dataset Program (UCDP) (Gleditsch et al., 2002, Pettersson et al., 2021). They are also widely used in the conflict research literature, which helps ensure that the GCRI's output is comparable to previous research and other forecasting models. In addition to the 3 categories, we also introduce the following overarching category:

- **Any conflict (ANY)**: This category combines the three categories SBC, NSC and OSV and is the main outcome of interest of the GCRI model.

As done previously, the revised GCRI model produces separate risk estimates for each conflict category (SBC, NSC, OSV), but now also estimates each country's overall risk of all conflicts combined (ANY). The regional ranking of countries is based on this latter category, which encompasses the 3 subcategories and therefore gives a more complete overview of conflict risk than done previously.

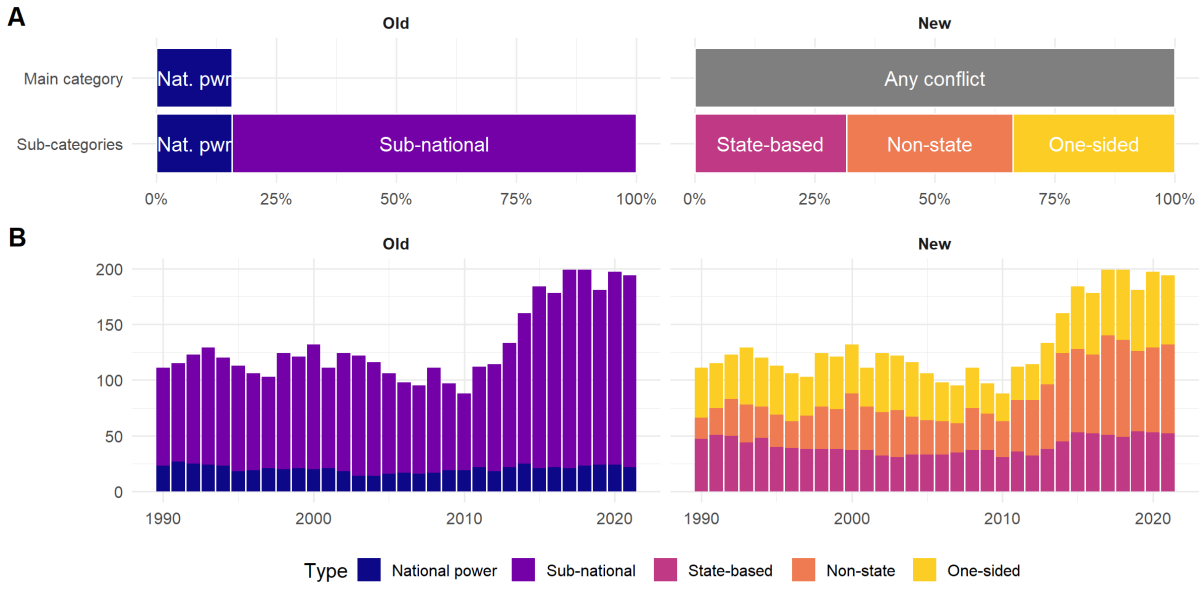

Figure 2: Distribution of UCDP conflicts by old (left) and new conflict categories (right). The old GCRI ranked countries based exclusively on their "National Power" risk, while the new model focuses on the risk of all conflicts combined ("Any conflict").

2.2 Conflict data and operationalization

To measure internal armed conflict we use data from UCDP, which provides data on SBC, NSC and OSV in 176 countries since 1989.[5] The data records yearly information on individual conflicts, which we aggregate to the country level. For each type of conflict, we construct the following two variables:

— **Conflict incidence:** This variable takes on the value 1 if there is at least one ongoing conflict with more than 25 battle-related deaths in a country-year and remains 0 otherwise, and is used to model and predict conflict probabilities.

— **Conflict intensity:** This variable counts the total number of battle-related deaths in each country-year, and is used to model and predict conflict intensities.

The GCRI previously discretized conflict intensities by dividing fatality estimates into 12 intervals prior to the analysis. We no longer take this additional step, in order to avoid discarding potentially important variation in fatalities. Instead, we fit intensity models directly on the raw fatality estimates, but log-transform and rescale our final predictions to a 0-10 conflict intensity scale after the analysis.[6]

While UCDP publishes yearly updates of all its datasets, the data are made available 1-2 months later than the GCRI's yearly updates. As a result, we lack definitive data on conflicts in the previous year,[7] which is needed to estimate conflict risk in the next 1-4 years. Previously, the GCRI relied on data from the Heidelberg Institute

[5] UCDP data coverage for state-based conflict starts in 1946, but we limit the analysis of all conflicts to the post-1991 period for consistency reasons and to avoid data limitations for other variables. UCDP also covers a small number of inter-state conflicts, which we remove from the dataset.

[6] The GCRI previously coded conflict intensity based on a combination of total fatalities and the number of conflicts. This is potentially misleading, since a country with 3 low-fatality conflicts could receive a higher score than a country with a single high-fatality conflict. For this reason, we decided to code conflict intensities based on the estimated fatalities alone.

[7] Each yearly UCDP release covers conflicts up to and including the previous year.

for International Conflict Research (HIIK) (HIIK, 2022) to fill this gap in data coverage. The HIIK data is released before the yearly GCRI updates, but its main disadvantage is that it uses a different approach to defining and measuring conflict, which is not fully compatible with UCDP's definitions.

Fortunately, UCDP recently introduced a new *Candidate Events Dataset*, which is designed specifically for near real-time conflict monitoring and forecasting (Hegre et al., 2020). The dataset is released on a monthly basis and covers all recent events that match UCDP's conflict typology according to their initial data collection. We rely on the candidate events data to code conflict occurrence and intensity at the country-level in the preceding year, using the most recent information available.

It should be noted that the candidate events data is by definition preliminary, and some events are added, removed or modified as part of an annual vetting process (Hegre et al., 2020). As a result, there are some inevitable discrepancies between the preliminary and final UCDP data that can affect the final output of the GCRI. However, we also found that any discrepancies between the preliminary and final UCDP data are much smaller than the differences between UCDP and HIIK data. Therefore, we decided to rely on the candidate events dataset instead of HIIK for the most recent conflict data, but plan to continue exploring ways to reduce the GCRI's sensitivity to possible coding errors in the preliminary UCDP data.

3 Revision of variables and data sources

The most important revisions in this year's update concern the input data and variables. We replaced the data sources of 9 variables and took additional steps to minimize the proportion of missing values. In addition, we revisited the GCRI's variable selection according to 5 criteria, based on which we removed 3 variables from the model while adding a new one. These changes are described in more detail below.

3.1 Addressing missing data problems

The GCRI relies on open-source data on known conflict drivers to estimate future conflict risk. Any statistical model is only as good as the data used to fit the model, and as such the requirements in terms of data quality and coverage are quite high. Unfortunately, not all of the original GCRI source datasets were satisfactory in that regard: Some datasets had insufficient temporal coverage, while others contained incomplete data for many countries or excluded several countries altogether. The result was a training dataset with high rates of missing data.

The problems of missing data are twofold: First, data is rarely missing at random, which can result in biased model estimates unless properly accounted for (Honaker and King, 2010). Second, we can only estimate a country's conflict risk if we have complete data for all predictor variables in that particular year. A common response to missing data problems is *imputation*, which refers to statistical methods to estimate missing data values from the overall distribution of the observed data. Imputation works well in datasets with relatively small percentages of missing data, but becomes problematic if the overall gaps in data coverage are large.

In the previous GCRI input dataset, 13 out of 24 input variables had missing value rates well above 20%. Moreover, many variables excluded several countries altogether and omitted many consecutive years for the entire sample. This makes imputation especially difficult, as we lack the information needed to estimate the unobserved values. As a general rule, it is best to minimize the need for imputation in the first place by relying on complete data wherever possible. We took two steps to achieve this: First, we standardized the list of countries to avoid including countries excluded from multiple datasets. Second, we replaced many of the original data sources with newer datasets that provide much better coverage. Both steps are discussed in the following sections.

3.1.1 Standardized list of countries

Missing data in part results from the fact that not all data sources cover the same set of countries. While most cross-country datasets aim to provide global coverage, they do not all cover the same list of countries. If we combine multiple datasets based on alternative country lists, this inevitably results in missing data in some cases. Therefore, a first step to addressing missing data problems is to start with a standardized list of states. Within the social sciences, two widely used lists are the Correlates of War list (Russett et al., 1968) and a second list of states introduced by Gleditsch and Ward, 1999.

We chose to rely on the second list, which is fully compatible with the UCDP conflict database and with most other cross-national datasets in social science.[8] The Gleditsch and Ward (GW) list includes countries with a population greater than 250'000 that have autonomous control over their territory and that are recognized as independent by other regional actors. The list covers a total of 176 countries in the post-cold war period. Although the EU Conflict EWS focuses on a smaller subset of 138 countries, we train the GCRI model on the full set of countries in order to make optimal use of all available data. After adopting the GW list, the revised GCRI dataset no longer covers the following 18 countries:

> *Antigua & Barbuda, Dominica, Grenada, Kiribati, Liechtenstein, Marshall Islands, Micronesia (Federated States of), Nauru, Palau, Samoa, São Tomé & Príncipe, Seychelles, St. Kitts & Nevis, St. Lucia, St. Vincent & Grenadines, Tonga, Tuvalu, and Vanuatu.*

[8] It should be noted that in the post-1991 period the Gleditsch and Ward and Correlates of War state lists are nearly identical.

Each of these countries are micro-states without a recent history of violent conflict and are not of major concern to the EU Conflict EWS. Removing these countries helps improve data coverage as most of them do not appear in multiple cross-country datasets. In addition, 4 countries were added to the data since 1991, including Czechoslovakia and Yugoslavia, which no longer exist today but nonetheless serve as important inputs into the training data.

We made one additional modification by adding Palestine to the revised list of states. Although Palestine is not universally recognized as an independent entity and is not part of the GW list, it plays an important role in the EU conflict EWS as a potential conflict zone. Moreover, the fact that most datasets record information on both Israel and Palestine allows us to model the two territories separately, which should contribute to a better understanding of conflict risk in both cases.[9] However, this distinction is made purely for analytical purposes, is done without prejudice to the individual positions of Member States and shall not be construed as recognition of a State of Palestine.

3.1.2 Replacing data sources

When the GCRI was originally developed in 2014, open-source data on socio-economic indicators was still relatively scarce. This explains why some of the previous GCRI predictor variables have incomplete coverage. Fortunately, a growing number of datasets have become available each year, making it much easier to address existing gaps in data coverage. Taking advantage of this trend, we reviewed all GCRI data sources to assess whether alternative sources exist with improved coverage. We established the following criteria: First, the dataset should contain valid measures of the variable of interest. Second, it should provide global coverage from 1991 to the present. Third, the data should be open source and fourth, the data should be updated regularly, ideally on a yearly basis.

For most variables, we found alternative data sources that allowed us to significantly reduce missing values. Figure 3A demonstrates the improvements by comparing missing value rates between the old and new GCRI input data. Predictor variables in the old dataset contained 19% missing data on average, compared to just 5% in the revised data. Coverage over time has also improved considerably, as shown in Figure 3B, which summarizes the yearly percentage of missing values for each variable. Contrary to the old dataset, most variables in the new data have near-complete coverage over time. Most missings in the new dataset are concentrated in 2021, which is due to the fact that many input datsets were not yet updated at the time of the analysis.[10]

Table 2 shows a side-by side comparison of data sources used in the old and new GCRI input data. In the new dataset, we rely on alternative data providers for 9 of the variables. Indicators for "Democracy", "State capacity", "Repression" and "Corruption" are now taken from the Varieties of Democracy (V-Dem) dataset (Coppedge et al., 2021). For "Homicide rates", we rely on the Global Burden of Disease dataset from the Institute for Health Metrics and Evaluation (Lopez and Murray, 1998). Data on "Transnational ethnic ties" is taken from the Ethnic Power Relations (EPR) Dataset Family (Vogt et al., 2015).[11] "Income inequality" is measured using the World Inequality Database (WID) (Alvaredo, 2018). To measure exposure to extreme droughts ("Droughts"), the new GCRI uses the CSIC Standardized Precipitation Evapotranspiration Index (Vicente-Serrano et al., 2010). "Temperature change" data is taken from the FAOSTAT Temperature change dataset (FAO, 2022).

For three variables we use the same data providers as before, but construct alternative indicators that capture the same concepts but contain much fewer missing values. To measure "Ethnic exclusion", we use a variable from the EPR dataset that estimates the percentage of a country's population that belongs to politically excluded ethnic groups.[12] For "Oil exports", we use an indicator of oil rents as a percentage of GDP from the

[9] It should be noted that UCDP's conflict datasets do not distinguish between Israel and Palestine. To address this, we use location data from the Georeferenced Event Dataset to assign past conflict events to the respective territories (Sundberg and Melander, 2013).

[10] Note that in this comparison the old GCRI dataset does not include the year 2021 and we are therefore slightly underestimating missing value rates compared to the new data.

[11] Previously, the GCRI used two distinct "Ethnic power" variables for NP and SN conflicts. We no longer make such a distinction and instead use the same "Ethnic exclusion" variable across all conflict types.

[12] This new variable also corresponds closely to the concept of "Ethnic exclusion" as a driver of conflict (Cederman et al., 2013).

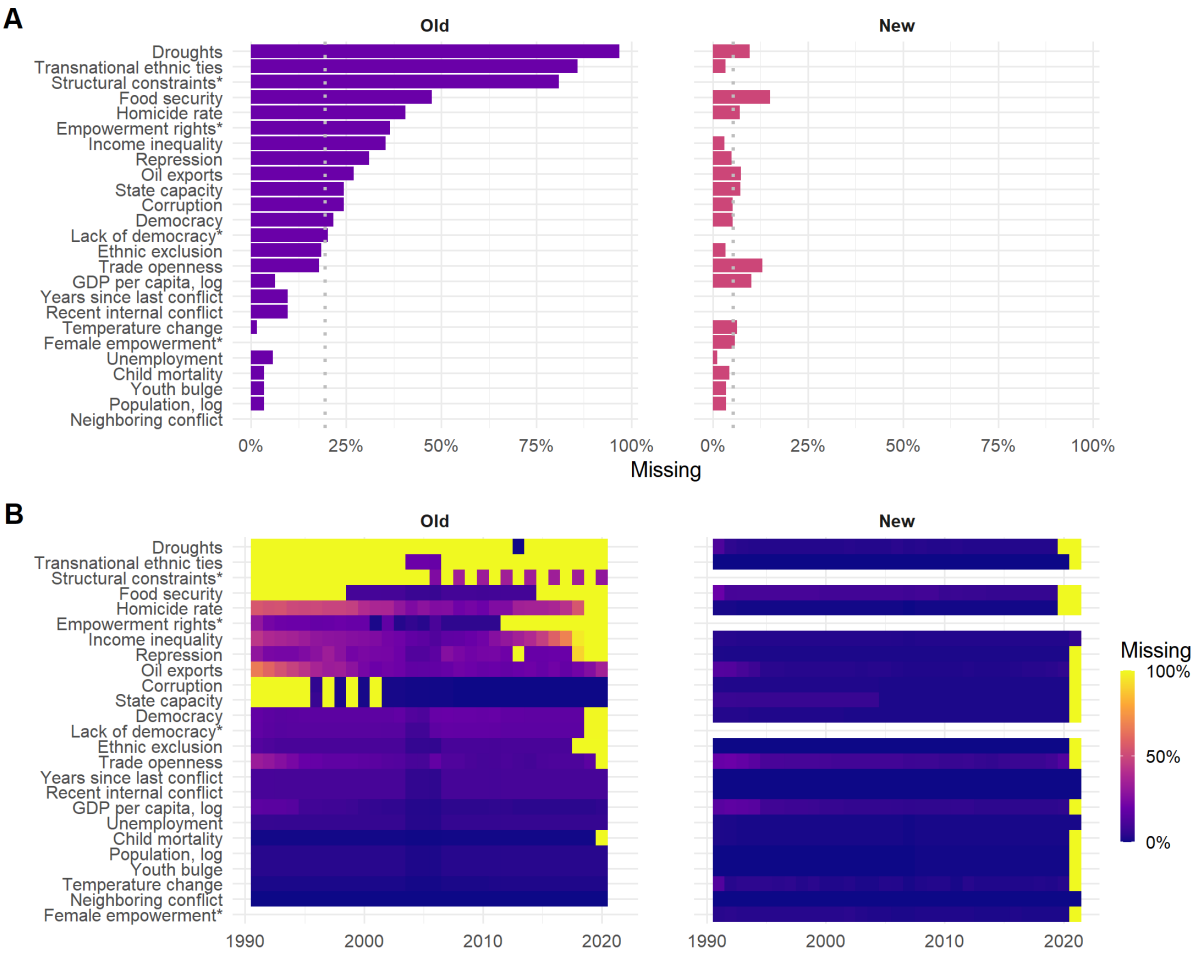

Figure 3: Missing values in the old and new GCRI input data. Panel A: Total % missing. Panel B: % Missing by year. Variables not included in both datasets are marked with an asterisk.

World Development Indicators (World Bank, 2020). To measure "Food security", we rely on the total daily average kcal available to each person according to the FAO Food Balance Sheets (FAO, 2001). Each of these indicators has fewer missing values than the previous ones, as shown in Figure 3. Tables A2 and A3 in the Appendix contain more detailed information on how each variable was constructed.

Variable	Old source	New source
Democracy	Polity IV	V-DEM
Lack of democracy	Polity IV	—
State capacity	World Bank	V-DEM
Repression	PTS	V-DEM
Corruption	World Bank	V-DEM
Empowerment rights	CIRI	—
Recent internal conflict	UCDP	UCDP
Years since last conflict	UCDP	UCDP
Neighboring conflict	UCDP	UCDP
Homicide rate	World Bank	IHME
Female empowerment	—	V-DEM
Ethnic exclusion	EPR	EPR
Transnational ethnic ties	MAR	EPR
GDP per capita, log	World Bank	World Bank
Income inequality	SWIID	WID
Trade openness	World Bank	World Bank
Oil exports	World Bank	World Bank
Food security	FAO	FAO
Unemployment	World Bank	World Bank
Droughts	WRI	SPEI/CSIC
Temperature change	SPEI/CSIC	FAO
Structural constraints	BTI	—
Population, log	UN	UN
Youth bulge	UN	UN
Child mortality	World Bank	World Bank

Table 2: Comparison of old and new data sources

3.1.3 Filling gaps in data coverage

By using a standardized list of states and replacing many input data sources, we were able to minimize value rates. However, the new GCRI dataset still contains around 5% missings overall, as shown in Figure 3. To fill these remaining gaps in data coverage, we took the following approach:

— **Patching**: For some variables, alternative datasets exist that cover most of the missing values. In particular, the Maddison project provides GDP and population estimates for some of the countries not covered by our current data sources (Bolt and Van Zanden, 2020). We therefore used these datasets to patch these gaps in data coverage.

— **Projection**: Many input datasets did not yet contain values for 2020 and 2021 at the time of the analysis. To address this, we adopted a new approach to projecting trends from the last observed data points into the near future. This approach, which is described in more detail in Section 5, is also applied to the years 2022-2025.

— **Imputation**: For the remaining missing values, we relied on multiple imputation. We used an imputation method designed for time-series-cross-section data, which takes into account that each country is

observed repeatedly over time (Honaker and King, 2010).[13]

The final data contains no missing values.

3.2 Variable selection

As part of the revisions of input data and variables, we also reviewed the GCRI variable selection. This led us to remove 3 variables that were deemed problematic. In addition, we also added a new female empowerment index to the GCRI, following a review of potential new variables and data sources. The total number of variables was therefore reduced from 24 to 22.

3.2.1 Selection criteria

Variable selection is a challenging task, which involves balancing multiple competing criteria. On the one hand, policy makers that rely on quantitative risk models want to ensure that they incorporate all relevant risk factors, and are less likely to trust models that may seem too simplistic or incomplete. On the other hand, the literature on predictive modelling has shown that sparser models tend to outperform overly complex ones. Overloading a model with too many predictor variables can create two main problems: First, it increases the likelihood that some variables are correlated, which can lead to unreliable estimates due to multicollinearity (Mansfield and Helms, 1982). Second, increasing the number of variables also increases the risk of overfitting. In this case, the model may perform exceedingly well on the training data, but loses predictive capacity when confronted with new data (James et al., 2013, Ying, 2019).

The authors of the original GCRI carefully selected 24 variables based on an in-depth review of the literature and extensive consultations with experts and policy makers (De Groeve et al., 2014). As a result, the selection of variables is generally backed up by theory and empirical evidence. At the same time, our knowledge on the causes of conflict keeps evolving, and we still know relatively little about which variables are needed to predict conflict. This is because most previous research has prioritized explanation over prediction, focusing mainly on finding statistically significant relationships between variables and conflict. However, many robustly significant variables have been shown to perform poorly in predictive models (Ward et al., 2010, Schrodt, 2014).

While statistical significance indicates whether or not two variables are related, it does not say much about the magnitude of the relationship. Significant effects may be very small or stem from just a handful of influential observations in the data (Lo et al., 2015). Conversely, some variables that are not causally linked to conflict might still serve as useful predictors. For example, child mortality is known to perform quite well in prediction models, although it is not usually seen a cause of conflict by itself (Hegre et al., 2017a). In short, identifying the main conflict predictors remains a challenging task in its own right.

To tackle these issues, prediction models commonly employ automatic variable selection procedures that identify the optimal number and combination of variables. Such analyses usually start with a large list of candidate variables, of which only a subset is retained that maximizes predictive performance. One potential drawback is that the resulting models may contain just a small number of variables, or the selected variables may not be very intuitive and may thus be difficult to interpret. For example, a recent model by Baillie et al., 2021 consists of only 3 predictor variables, whereas D'Orazio and Lin, 2022 develop a model that performs very well, but relies on variables that are difficult to grasp intuitively.[14] There is therefore a potential trade-off between theoretical intuition, parsimony and predictive accuracy.

For the purposes of the GCRI, relying entirely on automatic variable selection is not a suitable option. Besides improving model performance, we also aim to preserve the model's explainability and the theoretical intuition behind the current variables, while avoiding drastic changes in the methodology that could undermine policy

[13] We average over 25 imputations using the Amelia II package in R: `https://www.rdocumentation.org/packages/Amelia/versions/1.8.0`.

[14] Key variables in the D'Orazio and Lin, 2022 model include various spatial lags, temporal decay functions as well as the country name and month of observation.

makers' confidence in the system. Keeping these different objectives in mind, we reviewed the existing variables according to the following 5 criteria:

- **Theoretical justification:** There should be a clear theoretical argument supporting a relationship between the variable and conflict.
- **Empirical support:** There needs to be empirical evidence that the variable influences, or is otherwise indicative of conflict risk.
- **Data coverage:** There should be sufficient data available for all 176 countries in the post-Cold War period, and the dataset should be open-source and regularly updated (ideally once every year).
- **No overlap:** The variable should not be highly correlated with other variables in the model. Including the variable should not create high levels of multicollinearity.
- **Predictive performance:** Adding the variable should reduce, or at least not increase prediction errors.

3.2.2 Theoretical support

The GCRI's original variable selection was based on a review of over 200 quantitative conflict research publications and interviews with country experts and practitioners at the EEAS (De Groeve et al., 2014). Not surprisingly, therefore, each of the 24 risk factors in the original GCRI has strong theoretical support. Another strength of the GCRI is that its variables were chosen to reflect 6 broad risk areas: Political, Security, Social, Economy, Geography & Environment and Demographics.[15] This gives policy-makers an intuitive understanding of each variable and ensures that the model reflects a wide range of risk areas, each of which may be equally important in theory.[16]

One challenge with the original selection of variables is that there appears to be some conceptual overlap between some variables, as their definitions already suggest: For example, "Regime type" appears to measure very similar concepts as the "Lack of democracy" and "Empowerment rights" variables, and "Structural constraints" may overlap with "State capacity" and variables related to economic development. Such redundancies should be avoided, as they could result in high correlations and potentially in multicollinearity. We examine these issues in more detail further below.

3.2.3 Empirical support

Previous GCRI reports offer detailed reviews of the empirical evidence supporting each variable (De Groeve et al., 2014, Halkia et al., 2017). However, since research on the causes of conflict continues to evolve, we briefly revisited the current state of knowledge on each variable. Overall, our conclusions are in line with the assessments made previously, but there were some exceptions: For some variables, we found no strong evidence that the variable either influences or predicts conflict risk. This group includes "Trade openness", "Income inequality", "Unemployment" and the demographic "Youth bulge".

Although international trade is known to affect the risk of interstate conflict (Gartzke, 2007, Schneider and Gleditsch, 2010), its effects on domestic conflict remain less certain (Dixon, 2009). Income inequality has been shown to affect the risk of domestic conflict, but this applies primarily to inequality between ethnic groups and not to individual-level inequality (Cederman et al., 2011b, Stewart, 2011). Previous research has failed to find a systematic relationship between unemployment and conflict risk (Berman et al., 2011), although there is broader evidence that declines in economic opportunity can exacerbate ongoing conflicts (Dube and Vargas, 2013). Lastly, although the idea of a destabilizing youth bulge has received much attention, several studies have failed to find any support for it (see Fearon, 2011).

[15] The GCRI initially combined Geography and Demographics, but these were later divided into separate categories.

[16] The classification of variables into risk areas is also used to construct a composite indicator, which consists of standardized risk scores on each of the 6 dimensions.

Other variables are supported by some evidence, but the overall empirical record remains mixed. This group includes "Corruption", "Oil exports", "Transnational ethnic ties", "Food security", "Droughts" and "Temperature change". Mixed evidence can stem from difficulties in isolating the causal effects of certain variables, or it may be that some variables affect conflict risk only under specific conditions. For example, there is evidence that corruption reduces conflict risk in oil-rich states, but may increase instability in other contexts (Fjelde, 2009). Oil wealth itself may increase conflict risk if oil is located in regions home to marginalized ethnic groups, but not necessarily in other cases (Ross, 2015). The conflict-inducing effect of transnational ethnic ties depends in part on the group's access to power or on previous border changes (Goemans and Schultz, 2017, Cederman et al., 2022). Research on the link between food insecurity and conflict remains inconclusive, although there is evidence that food price shocks can provoke domestic unrest (Martin-Shields and Stojetz, 2019). Finally, there is widespread agreement that climate variability can increase conflict risk, but research thus far has found the effects to be small and contingent on the political and economic context of affected regions (Mach et al., 2019, IPCC, 2022).

To be sure, an absence of evidence does not equal evidence of absence. Moreover, even if a variable does not *cause* conflict, it may still help to *predict* conflict risk. A useful analogy here is the canary in the coal mine, which can play a crucial role in detecting toxic gas leakages, despite not having any causal impact on the outcome (Hegre et al., 2017a). In short, while evidence on causal effects is certainly useful, it should not be viewed a deal-breaker in variable selection. Instead, it is important to also pay closer attention to other criteria, including predictive performance.

3.2.4 Data coverage

As discussed in Section 3.1.2, we were able to minimize missing data for most variables by replacing several input data sources. However, there were still two variables with high proportions of missing values in the revised dataset:

First, the original GCRI included a **"Structural constraints"** indicator, which is taken from the Bertelsmann Transformation Index (BTI) and reflects *"structural difficulties [that] constrain the political leadership's governance capacity"* (Bertelsmann Foundation, 2022, p. 37). These constraints can stem from poverty, a lack of education, geographic obstacles and infrastructural deficiencies. The BTI itself is not well suited for a global analysis as it only covers a subset of "countries in transition",[17] does not provide yearly coverage and records no data prior to 2006. As a result, over 80% of observations are missing, and we were unable to find suitable replacement data for this indicator.

A second variable captures **"Empowerment rights"**, based on an additive index of indicators of freedom of speech, freedom of assembly, freedom of movement and related indicators taken from the CIRI human rights dataset (Cingranelli et al., 2014). The dataset has not been updated since 2014 and records no data beyond 2011. As we lack data for at least 10 consecutive years, imputation is not feasible. Although there are other replacement datasets that measure similar concepts, we also found that this variable overlaps to a considerable degree with other variables in the model and is therefore considered redundant (see next section).

3.2.5 Overlap and multicollinearity

Considering the large number of variables in the GCRI, it is likely that some are going to be highly correlated. Correlations can arise if multiple variables tap into the same concepts or if they are causally related, which can be problematic if it introduces multicollinearity. In this case, one independent variable may be a near-perfect predictor of another variable. This makes it especially difficult to estimate the independent effect of either variable, and our model estimates may become unreliable as a result (Mansfield and Helms, 1982).

To assess the degree of overlap, we first examined pairwise correlations between the 24 original GCRI variables. The results are shown in Figure 4. Following Dormann et al., 2013, we use 0.7 as the threshold for problematic correlations. Only a few variable pairs exceed this threshold: "Democracy" correlates highly with a

[17] This category excludes 40 countries in the GCRI training dataset.

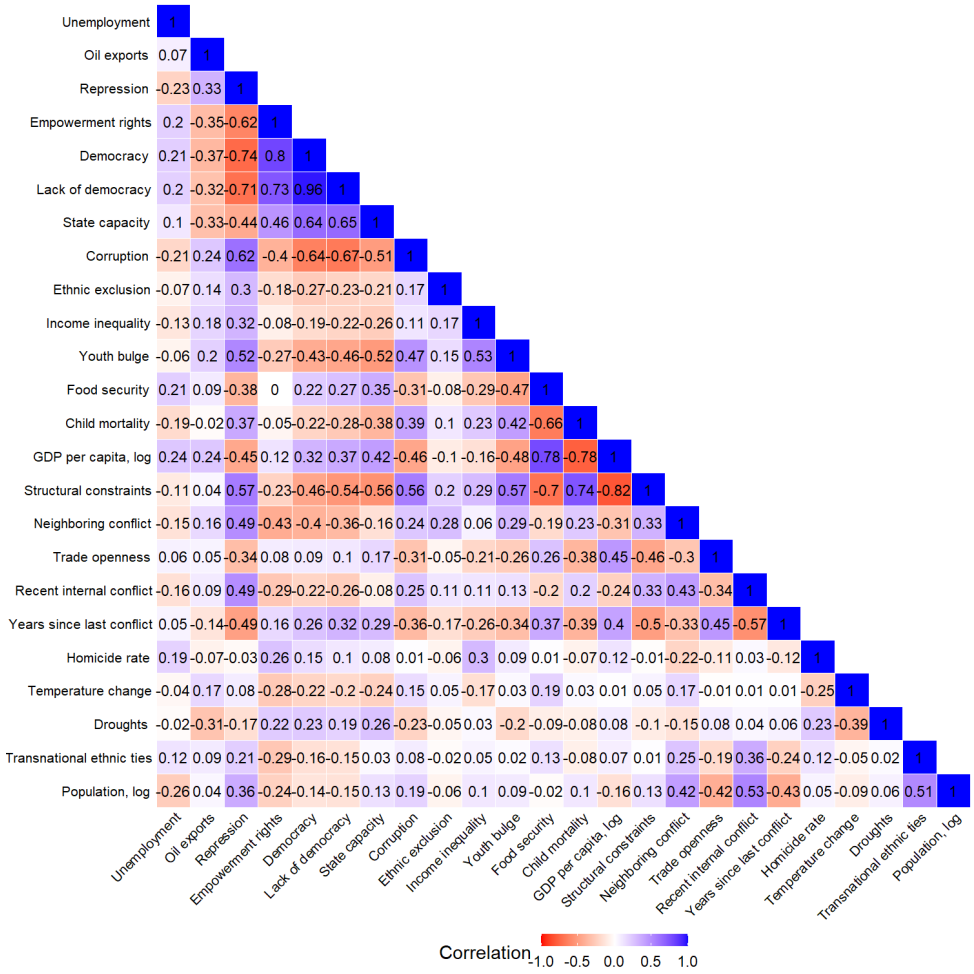

Figure 4: Pairwise correlations between variables

"Lack of democracy" and "Empowerment rights", "GDP per capita" is highly correlated with "Food security", and "Structural constraints" is highly correlated with "Child mortality".

The high correlations between the democracy variables and "Empowerment rights" are not surprising: The former two variables are based on VDEM's liberal democracy index, which covers many of the same concepts as the empowerment rights variable.[18] Likewise, the "Structural constraints" variable overlaps with 3 other variables that are also related to poverty and poor government performance. Given that both "Empowerment rights" and "Structural constraints" also have insufficient data coverage, we considered them as the most problematic and decided to remove them from the GCRI.

In a second step, we estimated regression models with the remaining variables and computed the Variance Inflation Factor (VIF), which is a standard measure of multicollinearity. This was done for the GCRI probability and intensity models, as shown in Figure 5.[19] Following James et al., 2013, we used a VIF value of 10 as the threshold for problematic multicollinearity, but also treated any values above 5 as potentially concerning.

A first set of models with all the remaining variables has VIF values well above the problematic threshold. This is mostly driven by the high correlation between the "Democracy" and "Lack of democracy" variables: We decided to remove the latter variable, as we considered it slightly less intuitive and it performed less well than the democracy index. As a result, overall levels of multicollinearity dropped below the problematic threshold. "GDP per capita" still has a relatively high VIF, which is most likely due to its overlap with "Child mortality" and "Food security". However, given that the highest VIF values are only slightly above 5, we do not view the remaining

[18] V-Dem's liberal democracy index *"emphasizes the importance of protecting individual and minority rights against the tyranny of the state and the tyranny of the majority. […]. This is achieved by constitutionally protected civil liberties, strong rule of law, an independent judiciary, and effective checks and balances that, together, limit the exercise of executive power. To make this a measure of liberal democracy, the index also takes the level of electoral democracy into account."* (Coppedge et al., 2021).

[19] For the purpose of this analysis, we only estimated models for the "Any conflict" category.

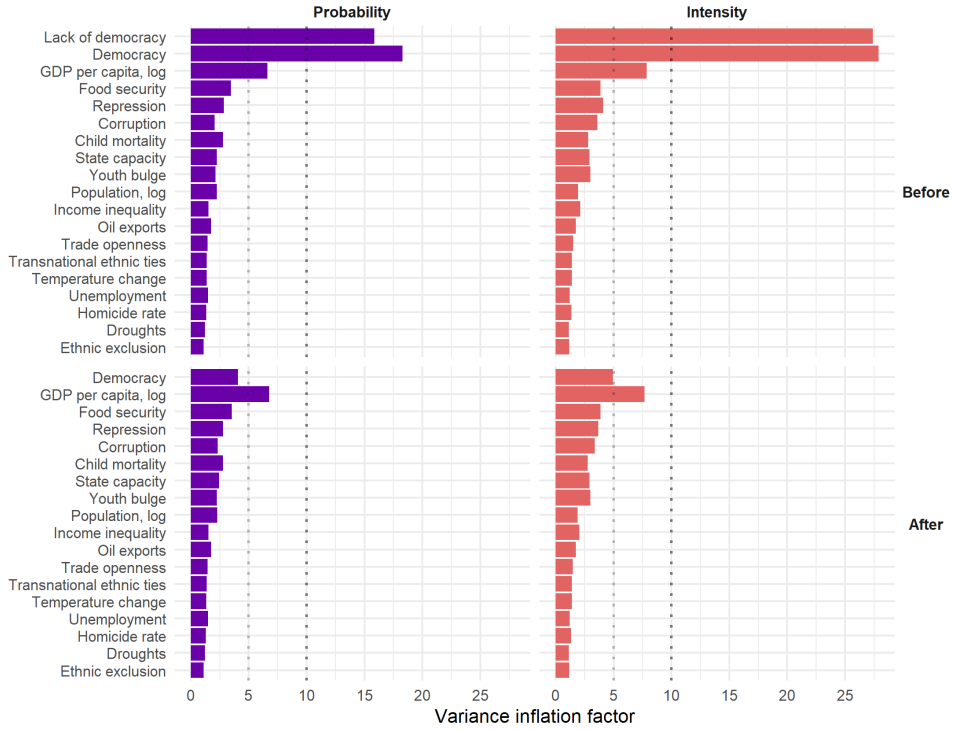

Figure 5: Multicollinearity before and after removing the "Lack of democracy" variable. Dotted lines indicate thresholds of potentially problematic (left) and concerning (right) levels of multicollinearity

multicollinearity as especially concerning.

3.2.6 Predictive performance

A key variable selection criterion is predictive performance: Adding a variable to the model should improve, or at least not reduce the accuracy of our predictions. To assess this, we estimated a series of empty probability models that contain just one predictor variable at a time, and assessed each model's performance to get a general idea of each variable's importance. Each model was trained on a random partition of the revised data consisting of 8 consecutive years between 1991 and 2021, and tested by making a prediction one year into the future. This procedure was repeated 10 times. Figure 6 illustrates the split between training and testing data in a simplified example.

Each prediction was compared against observed outcomes in the testing data to assess its accuracy. As a summary measure of model performance, we use the Area under the Precision-Recall Curve (AUPRC), which indicates how many instances of violent conflict were correctly identified as such.[20] Each model was trained and tested 10 times on different subsets of data. Figure 7 summarizes the results for each of the variables and across all conflict types, ranking the variables from most (top) to least important (bottom).

The results show considerable differences between variables. By far the most influential variables relate to past and current levels of conflict. This aligns with previous research that has identified previous conflict and neighboring conflict as some of the most important risk factors (Gleditsch, 2007, Salehyan, 2011, Hegre et al., 2017b). Other important variables include "Population size", "Trade openness", "Corruption", "Child mortality", "GDP per capita", "Ethnic exclusion" and "Food security". At the bottom of the list, we find "Unemployment", "Homicide rates", "Income inequality", "Droughts" and "Temperature changes".

While individual comparisons help to identify the most important variables, they are not sufficient. In a multivariate model, the effect of some variables may be absorbed by others, or may depend on other variables

[20] We prefer the AUPRC over alternative metrics such as the Area under the ROC curve or Brier scores because the AUPRC ignores the "true negatives" (i.e. instances of no conflict), which constitute the vast majority of observations and are therefore easiest to predict.

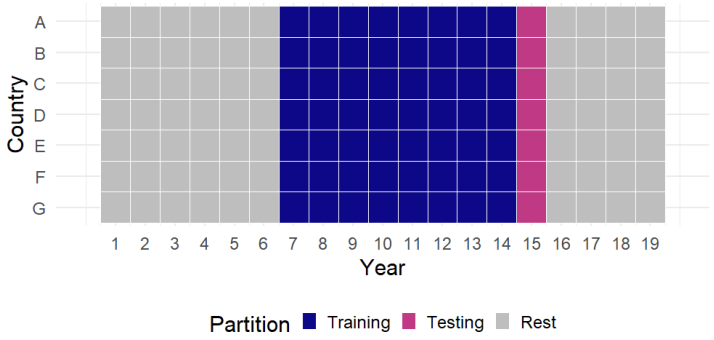

Figure 6: Illustration: Partitioning data into a training and testing set. Training data is used to fit the model and testing data to make and validate predictions.

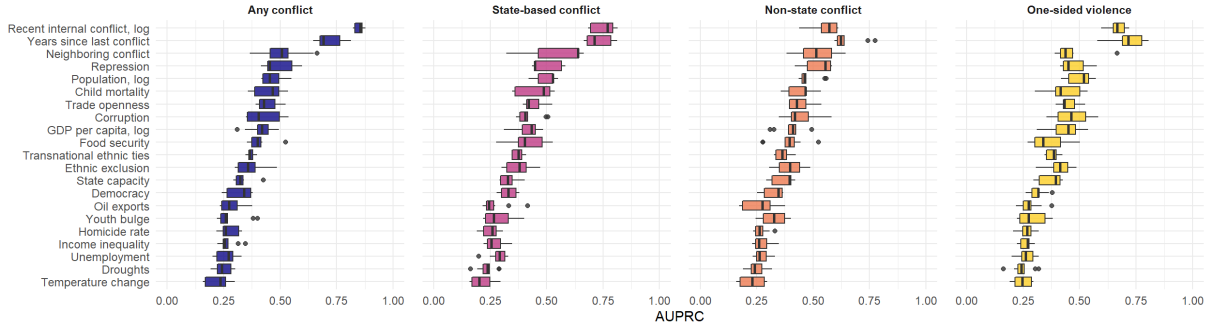

Figure 7: Variable importance. Variables are sorted from most to least important based on their predictive performance compared to an empty model. Box plots show distribution of AUPRC scores over 10 cross-validations.

in the model. Therefore, we also examined variable importance using automated variable selection. Although we do not use this approach to select the final set of variables, it is still useful to consider which combination of variables drives the results. We used stepwise forward selection, which starts with an empty model and then gradually adds the most to least important variables until the best model fit on the training data is reached.[21] In this case, the training data covers all years from 1991 to 2016. The resulting model contains only 11 of the remaining variables:

> *Recent internal conflict, Years since last conflict, Population (log), Repression, State capacity, Ethnic exclusion, Democracy, Neighboring conflict, Corruption, Income inequality and Unemployment.*

Note that this list of variables does not exactly match the 11 most important variables according to the previous analysis: Some influential variables such as "Trade openness", "Child mortality" and "GDP per capita" are missing, while other relatively unimportant variables such as "Income inequality" and "Unemployment" are included. This highlights that selecting the right combination of variables can be as important as selecting the right variables individually.

Since our variable selection method only considered data from 1991 to 2016, we also compare the predictive performance of the full and reduced model on a testing dataset that covers 2017 to 2020. In addition to the AUPRC score, we also consider other common metrics, such as the Area under the Receiver Operator Curve (AUROC), the Expected Proportion of Correct Predictions (EPCP) and the Brier score (see Greenhill et al., 2011 for a discussion). The results are shown in Table 3, highlighting the best-performing model in bold.

In the testing data, the performance of the full model is nearly identical to the reduced model with 11 variables. The former achieves a slightly better performance than the latter, but the differences are so small they are hardly relevant. For example, the difference in AUPRC values is only < 0.0004. In short, this comparison suggests that a subset of 11 variables drives most of the GCRI's results, but it also shows that adding the

[21] Using stepwise backward selection leads to identical results.

remaining 10 variables does no harm to the model's predictive performance. We therefore see no good reason to remove further variables from the GCRI based on predictive performance alone.

Model	AUPRC	AUROC	EPCP	Brier
Reduced model (11 vars)	0.8840	0.9479	0.8698	0.0656
Full model (21 vars)	**0.8844**	**0.9487**	**0.8706**	**0.0652**

Table 3: Predictive performance: Full vs. reduced model. Best performance highlighted in bold.

Lastly, a remaining concern given the large number of variables is overfitting: If our model becomes too specific to the training data, it can lose its performance on new datasets. Halkia et al., 2020 already assessed this for the previous GCRI model by comparing its in-sample and out-of-sample performance, which suggested there was no overfitting. We replicated this analysis with the revised GCRI, comparing performance of the full model on a training dataset (1999-2016) and testing data (2017-2020), as shown in Table 4.

The results confirm that there is indeed no overfitting. Instead, the model even performs better on the testing data than on the training dataset in this particular case. This improvement may be due to "luck", as the period 2017-2020 might contain more conflicts that are easy to predict for the GCRI than other periods. In any case, this alleviates concerns of overfitting as a result of too many variables.

Model	AUPRC	AUROC	EPCP	Brier
Training data (1991-2016)	0.884	0.949	0.871	0.065
Testing data (2017-2020)	**0.921**	**0.956**	**0.891**	**0.058**

Table 4: Predictive performance: Training vs testing data (full model). Best performance highlighted in bold.

3.3 Summary: Variables no longer in the GCRI

To recap, we conducted a systematic review of all GCRI variables based on 5 separate criteria, and found that there was sufficient theoretical support for the inclusion of each variable, while empirical support existed in most, but not all cases. Data coverage is sufficient for most variables, except for "Structural constraints" and "Empowerment rights", which still contained high rates of missing values in the revised dataset. In addition, these two variables also overlapped with other variables in the model which leads to multicollinearity. Another variable that introduced problematic levels of multicollinearity is the "Lack of democracy" variable, which is a near-perfect predictor of the "Democracy" indicator. Based on these findings, we decided to remove these three variables from the GCRI going forward.

Our analysis also generated useful insights into the importance of each variable: A small number of variables related to current and past conflict have an outsize effect on the results, while many others appear to be less relevant. Using automated variable selection, we found that 11 out of 21 variables did most of the heavy lifting, while the remaining 10 variables appear to have very little impact on the GCRI predictions. However, we also found that using the full set of variables does not reduce model performance or result in overfitting, which justifies keeping them in the model.

3.4 New GCRI variable: Female empowerment

As part of our review of variables, we also conducted research on potential new variables and data sources. Following this initial research and a series of tests, we decided to add an index of female empowerment to the current list of predictor variables. Previous research has found a strong correlation between gender inequality and conflict risk (Caprioli, 2005), while the broader involvement of women in societal decision-making is viewed as an important ingredient for peace (GIWPS and PRIO, 2019, Crespo-Sancho, 2017). Even though the causal

link between gender inequality and conflict remains difficult to establish, it can still be seen as a useful indicator for inequality and repression more generally.

To measure this concept, we rely the *Women Political Empowerment Index* from the V-Dem dataset. The index includes 3 equally-weighted dimensions: (i) fundamental civil liberties, (ii) women's open discussion of political issues and participation in civil society organizations, and (iii) the descriptive representation of women in formal political positions (Coppedge et al., 2021).

Adding the "Female empowerment" variable did not have a major impact on the GCRI as a whole: The variable's correlations with regime type and repression are relatively high, but adding the variable only slightly increases multicollinearity (see Figures A2 and A3 in the Appendix). Although the variable itself ranks in the middle in terms of importance (Figure A4), we found that adding it to the existing 21 variables has no discernible impact on predictive performance. Despite these limitations, we concluded that "Female empowerment" still constitutes a useful addition from a theoretical perspective, as it complements "Ethnic exclusion" and "Transnational ethnic ties" as two other relevant attributes on the social dimension.

In future iterations of the GCRI, we plan to further explore other variables that could be added to the current selection. Potentially useful variables include the proximity of elections and the risk of droughts for agriculture. Going forward, we plan to base any decision on adding or removing variables on the 5 criteria used in the present analysis.

4 Revisiting the GCRI modelling framework

The original GCRI used logistic regressions to model and predict conflict probabilities, and relied on linear regressions for conflict fatalities. These two approaches are widely used in the literature but also have known limitations. For example, both models treat each observation in the data as independent, which likely does not apply if the data contains repeated observations for the same units over time. If a model fails to account for such important characteristics of the data, this could result in biased parameter estimates and decreased predictive performance. As part of revising the GCRI, we therefore reviewed and tested alternative statistical models. We identified the following challenges that the current models do not sufficiently address:

Temporal dependence

The GCRI relies on a typical *Time-Series Cross-Sectional* dataset, in which a set of countries is observed over multiple years. In such contexts, the independence assumption usually does not hold: A country's current conflict risk likely depends on its history of conflict. Ignoring temporal dependencies would also mean that we assume a constant risk over time, which is usually inaccurate (Carter and Signorino, 2010). The GCRI previously accounted for time dependence to some extent by incorporating conflict history variables, but this may not be sufficient. As a possible alternative, we explored time-series models, which are commonly used to forecast observations based on historical data. In this analysis, we estimated probability and intensity models with autoregressive integrated moving average (ARIMA) errors (Shumway and Stoffer, 2017a). In contrast to most regression models that pool observations from many countries, this approach entails estimating separate models for each country.

Unobserved heterogeneity

Beyond the observed country-specific risk factors, there may also be unobserved country characteristics that drive conflict risk. Some countries may be more prone to conflict to begin with for reasons not fully reflected by the predictor variables. This in turn might cause us to over- or underestimate conflict risk in certain cases. To account for this, we estimated a series of mixed models,[22] which are designed to deal with clustered data structures[23] (Kreft and De Leeuw, 1998, Gelman, 2006). This allowed us to estimate different baseline risks for each country. We estimated a simple generalized mixed model with random intercepts for the probabilities, and estimated 4 types of linear mixed models for the intensities. In the latter case, each model used a different approach to dealing with variation over time per country.

Overdispersion

The linear model used previously assumes that conflict intensities are normally distributed. While many real-world phenomena such as height, intelligence or shoe sizes fit a normal distribution, this is not the case for violent conflict. Instead, conflict fatalities follow a power-law distribution similar to the magnitude of earth-quakes (Cederman et al., 2011a, Cirillo and Taleb, 2016). To deal with this skewed distribution of conflict intensities, we estimated a series of count models, including negative binomial and Poisson models for conflict intensities (Hilbe, 2011).

Rare events

Violent conflict is rare, making it difficult to estimate relationships between a large set of variables and few non-zero outcomes. This can lead us to underestimate the probability of rare events and could also make the results more sensitive to measurement error in the input data (King and Zeng, 2001). To deal with excess zeros in the probability models, we estimated a rare events logistic regression model (King and Zeng, 2001), and replicated a

[22] Mixed models are also commonly referred to as hierarchical or multi-level models.
[23] In our case the data are clustered across time and countries.

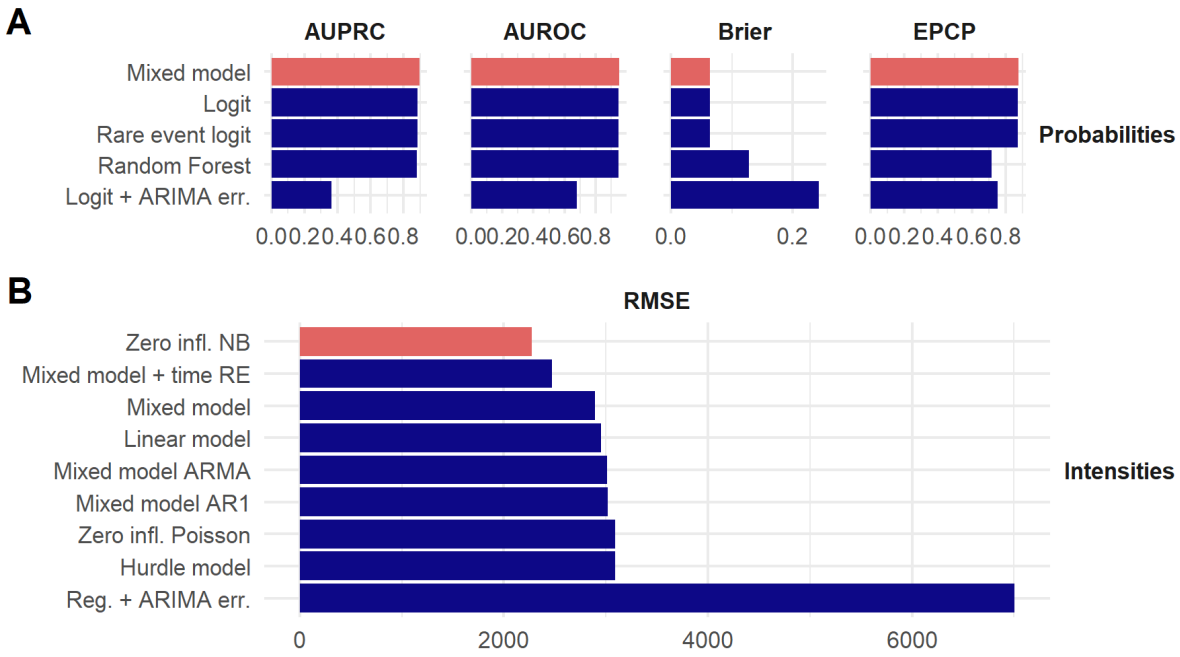

Figure 8: Results of model comparisons: (A) Probability models and (B) Intensity models. Best performance scores highlighted in red.

random forest model by Muchlinski et al., 2016.[24]. For the intensity models, we estimated zero-inflated negative binomial and poisson regression as well as a hurdle model. The main advantage of these approaches is that they estimate non-conflict outcomes separately from conflict fatalities, making it less likely that we underestimate the outcome when dealing with rare events.

4.1 Results of model comparisons

Our final comparison included 5 different probability models and 9 intensity models, each using the same set of predictor variables. Each model was trained on a dataset ranging from 1991 to 2016 and validated on the years 2017 to 2020. Figure 8 summarizes the results. Among the probability models, the generalized mixed model performs best, followed by the standard logistic regression model used previously, the rare events logit and random forest model and finally the logistic model with ARIMA errors. However, it should be noted that differences between the 3 top-performing models are so small they are hardly relevant. For example, the difference in AUPRC values between the mixed model and standard logistic model is only < 0.004.

Differences among intensity models are much clearer, as shown in Panel B. The comparison here focuses on the Root Mean Square Error (RMSE), a standard metric for continuous outcome models. The zero-inflated negative binomial model achieves the smallest overall prediction error, followed by a linear mixed model with time random effects, a simple linear mixed model, the standard linear model used previously, two mixed models with auto-regressive errors and two other count models. Lastly, the linear model with ARIMA errors generates the largest prediction errors by a wide margin.

We used the results of this model comparison as the basis for our final selection of models. The choice was clear for the intensity models, as the zero-inflated negative binomial evidently outperformed the other approaches. We therefore adopted this approach for the intensity models instead of the linear model used previously. For the probability models, the decision was less straightforward. The generalized mixed model performed best in this comparison, but the improvement relative to the logistic model hardly matters in substantive terms. In addition, we were unable to estimate the mixed models on some portions of the data.

Mixed models often fail to converge when attempting to estimate too many parameters from limited amounts

[24] The authors found that their random forest model outperformed standard logistic regression models in predicting rare outcomes.

of data, or if the outcome itself is a rare event.[25] If a model fails to converge, the estimates and predictions may be unreliable. To avoid such issues, we decided to keep the standard logistic regression model for the time being, which proved to perform nearly as well as the mixed model but faced much fewer convergence issues.

[25] In this particular case, we aim to estimate 176 country-specific intercepts and coefficients for 22 variables from a dataset that contains 176 countries, no more than 30 observations per country and relatively few instances of conflict.

5 Predicting the predictor variables

The main idea behind predictive modelling is that we can estimate unobserved outcomes based on (1) an observed set of predictor variables and (2) a model of the relationship between predictor variables and the outcome.[26] If a model performs well enough on previously observed data, we can feed in new values for the predictor variables to estimate unobserved outcomes. This is straightforward if the new input values are already known, but is more challenging if they have not yet been observed.

This of course applies to any attempt to predict future outcomes, where both the outcome and the distribution of predictor variables are unknown. In this case, we need to make certain assumptions about the future values of predictors or estimate them from the available data. The simplest approach is to extend the last observed values of the model's predictors into the near future. This approach is often referred to as *Last Observation Carried Forward (LOCF)*, and was also used so far in the GCRI, which produced a single prediction for the next four years based on the most recent available data. In doing so, we implicitly assumed that the values of all predictor variables would remain constant over the coming 4 years. One problem with this approach is that predictor variables can and often do change in the near term, which may also influence conflict risk.

For example, as the fallout of COVID-19 and the Ukraine war demonstrate, regional crises can drive down GDP and food security in vulnerable countries, which in turn can lead to political instability. A country's security situation might also change due to a sudden outbreak of conflict in a neighboring country, while an increased risk of conflict in the next year also has implications for all subsequent years. As the "conflict trap" phenomenon illustrates, ongoing conflicts may become self-sustaining and previous conflicts often recur, in part because they can undermine socio-economic development and political institutions.

Indeed, the challenge of "predicting the predictor variables" remains an important but understudied issue. Some studies have demonstrated the potential of simulation methods to deal with compounding risks and spatial-temporal dependencies (Hegre et al., 2017b, Weidmann and Ward, 2010), while some have called for better prediction models of important variables such as democracy or oil dependence (Hegre et al., 2021). To be sure, either approach constitutes a major challenge in its own right and is certainly beyond the scope of the current analysis. In addition, the problem is arguably less relevant for the GCRI, which has a time horizon of only 4 years, compared to models that aim to predict conflicts many decades into the future.

Still, it is likely that the GCRI would benefit from an improved modelling of predictor variables in the near future. To investigate this, we have explored time-series forecasting models to estimate the likely trajectory of several predictor variables. This analysis was limited to 12 continuous variables:

> *Homicide rate, GDP per capita (log), Income inequality, Trade openness, Oil exports, Food security, Unemployment, Droughts, Temperature change, Population (log), Youth bulge, and Child mortality.*

We did not generate projections for 10 variables that were either discrete / categorical or derived from such variables, as these are more difficult to predict. In these cases, we simply carried the last observation forward, as done previously:

> *Democracy, State capacity, Repression, Corruption, Recent internal conflict, Years since last conflict, Neighboring conflict, Female empowerment, Ethnic exclusion, and Transnational ethnic ties.*

For each country and continuous variable, we constructed a training dataset that covers the years 1991 to 2016, and a testing dataset covering 2017 to 2020. We then estimated two sets of time-series models that were validated by comparing predicted with observed values: First, we estimated linear regression models with ARIMA errors (Shumway and Stoffer, 2017a). Second, we estimated ETS models that use exponential smoothing to predict short-term time trends (Harrison, 1967). We then compared the performance of both models within the testing data, and chose the best-performing model for each country and variable to generate projections for the period 2022 to 2025. The basic procedure is illustrated in Figure 9, which shows a comparison of projected

[26] In the literature the term "prediction" does not only refer to estimating future outcomes, but decribes the task of estimating unobserved data more generally.

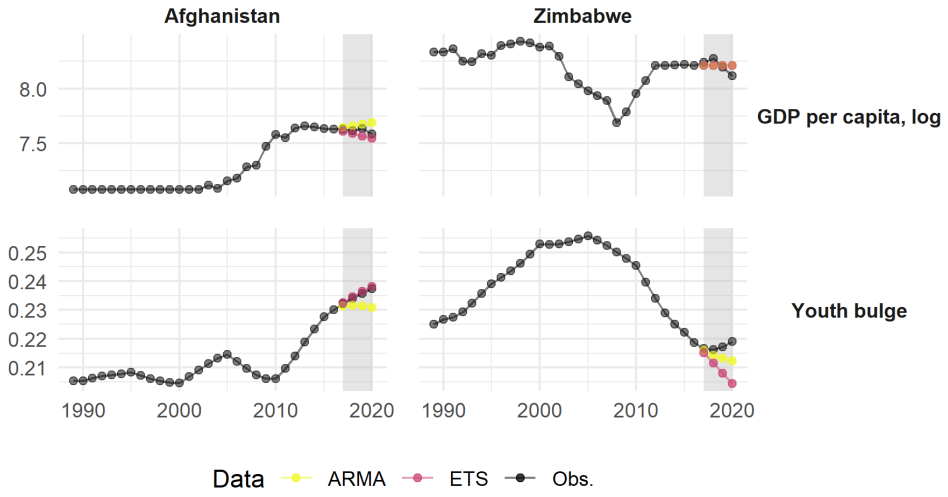

Figure 9: Comparison of projected and observed "GDP per capita" and "Youth bulge" variables for 2017-2020 in 2 countries. Forecast period highlighted in grey.

and observed values for the "GDP per capita" and "Youth bulge" variables in Afghanistan and Zimbabwe. This illustrates that in some cases ARMA yields better results, in others ETS is superior, while in other cases both methods produce identical results.

The most important question of course is whether our new approach helps improve the GCRI's performance. To assess this, we trained a probability model on a dataset from 1991 to 2016 and generated predictions for 2017 to 2020 based on two sets of input data. In the first case, we extended the values from 2016 to all subsequent years, while in the second case we included projected data for 12 continuous variables. The results in Table 5 confirm that using projections indeed helps improve predictive performance. For example, the AUPRC value improved by around 3%, which is not major but still notable.

Input data	AUPRC	AUROC	EPCP	Brier
LOCF	0.900	0.950	0.883	0.065
Projections	**0.924**	**0.959**	**0.892**	**0.056**

Table 5: Predictive performance: LOCF vs. projections. Bes performance highlighted in bold

In sum, we have introduced a new prediction framework that incorporates projections for all continuous predictor variables over the next 4 years. For each country and variable, we selected one out of two time-series forecasting models that achieved the best performance, and used this model to generate projections from the last observed values. While these initial results are encouraging, there still remains room for improvement.

Most importantly, we have not generated any projections for discrete or categorical variables,[27] as this proved to be more challenging. However, it is conceivable that improved projections of such variables could have an even bigger impact on the results. For example, if we foresee an increased risk of conflict onset in country A over the next 4 years, this is likely to influence conflict risk in neighboring countries B and C as well, given that conflicts often spread across borders. Therefore, improving our forecasts of variables such as "Recent conflict" or "Neighboring conflict" remains an important task for future research.

[27] This also includes variables that are not categorical per se, but which were derived from categorical variables, such as the "Time since last conflict", the "Corruption" index or "Transnational ethnic ties."

6 Conclusions

The goal of our work has been to further enhance and improve the GCRI, building on the most recent developments in conflict research and the broader forecasting literature. This has led to the following improvements:

First, we adopted a new, more comprehensive conflict typology that is also more compatible with previous research than the one used previously. In addition, the regional rankings are no longer based on a conflict subcategory but instead reflect the combined risk of all types of internal conflict, which ensures that different types of conflict are treated equally at the outset.

Second, we significantly reduced missing values in the GCRI input dataset by standardizing the list of states and replacing the data sources for 9 variables. By using data sources with better coverage, we also minimized the need for imputation and therefore reduced the overall noise in the input data. This is arguably the most important improvement, as the quality of the input data contributes significantly to the quality of the output.

While revising the input data, we also re-assessed the GCRI's variable selection in terms of theoretical and empirical support, data availability, overlap and multicollinearity, as well as each variable's contribution to model performance. Following this assessment, we excluded 3 problematic variables from the model and added a new indicator of female empowerment. We recommend using the same selection criteria for any other candidate variables going forward.

Third, we conducted a systematic comparison of 14 modelling frameworks, based on which we selected the updated probability and intensity models. Among the probability models, the best model performed only marginally better than the logistic regression model used previously, and we decided to keep the latter approach for practical reasons. Among the intensity models, we found that the zero inflated negative binomial model clearly outperformed the linear model used previously, and decided to use this new approach in the updated GCRI model. Overall, the results of this analysis suggest that the choice of modelling framework may be less relevant than the variable selection or the quality of the input data. Still, it is possible that further improvements can be made by exploring alternative modelling approaches, or by using an ensemble that combines several models with different strengths and weaknesses. We plan to explore this in the next GCRI update.

Fourth and finally, we took a first step towards projecting predictor variables into the near future. Instead of simply carrying the last observed values over to the next 4 years, we generated projections based on recent trends for 12 continuous predictor variables, which helped to improve the model's predictive accuracy. These initial results are promising, and suggest that investing more into the challenge of predicting the predictor variables may further improve predictive performance.

To assess the combined effect of these improvements on the GCRI as a whole, we compared the performance of the previous and revised GCRI probability models, as shown in Table 6. This shows a clear improvement: The new GCRI has an AUPRC model around 5% higher than the old model.

Model	AUPRC	AUROC	EPCP	Brier
Old GCRI	0.878	0.950	0.879	0.073
New GCRI	**0.924**	**0.959**	**0.892**	**0.056**

Table 6: Predictive performance: Old vs. new GCRI model. Best-performing models in bold

6.1 Limitations

Despite the improvements in the revised GCRI, the system as a whole still has several limitations. First, the GCRI's conflict risk assessments are largely based on "slow-moving", structural variables that do not change much over time. This is illustrated in Figure A1, which depicts the distribution of each variable across time. This reliance on structural risk factors makes it difficult to anticipate sudden changes in a conflict risk, for example due to economic shocks or other short-term events. Similarly, because the GCRI focuses on country-level conflict risk, it

does not reveal where in a country the risk of conflict may be highest. Such information may be especially useful in the case of large countries with many potential conflict hotspots in different regions. A third limitation is that the GCRI is only updated once every year, whereas policy makers often require more frequent and up-to-date risk assessments.

Lastly, one important limitation is that the GCRI exclusively deals with internal conflict, and does not consider the risk of armed conflict between states. This is not an exception; the conflict literature has long neglected interstate conflicts, as these were generally considered too unlikely in the post-1945 period. In contrast, Russia's invasion of Ukraine, resurging tensions between other post-Soviet states and growing concerns over a potential Chinese invasion of Taiwan demonstrate that there remains a real risk of interstate war, which the GCRI currently does not capture. Modelling and predicting interstate conflict is challenging, as it involves dealing with bilateral relationships and networks between states, which certainly goes beyond the scope of the GCRI. Instead, it might be more feasible to gain further insights into interstate conflict by an improved monitoring of existing rivalries, for example through news reports. At the very least, it is important to keep in mind that the GCRI is limited to domestic conflict, and will therefore fail to warn of conflicts fought primarily between states.

6.2 Future research

Our work over the next year will focus on addressing some of the GCRI's remaining limitations. First, we aim to further develop and improve the input data and modelling frameworks. The main task here is to improve the projections of predictor variables. If we are better able to estimate the trajectory of important variables such as "Neighboring conflict", "GDP per capita" or "Food security" over the next 4 years, this is likely to improve the accuracy of our conflict risk estimates. For some variables, we may be able to use existing projections by other research teams, while for others we could consider more advanced projection methods such as dynamic simulation (see Hegre et al., 2017b).

In addition, we plan to dedicate most of our time into developing a shorter-term prediction model. We have completed a pilot study on a *Dynamic Conflict Risk Model (DCRM)* that assesses conflict risk at the sub-national level in Africa over the next 1-6 months, and are currently working to further develop this model and expand it to other regions. Ultimately, the goal of this project is to complement the existing system: While the GCRI continues to give a yearly assessment of structural risks at the country level, the DCRM will give more regular updates and detailed insights into short-term increases in conflict risk (by allowing to deal with sudden changes in conflict risk due to disasters, economic shocks, or other short-term events) and the exact location of potential conflict hot spots within countries.

References

Alvaredo, F., 'World inequality report 2018'. In 'World Inequality Report 2018', Harvard University Press, 2018.

Baillie, E., Howe, P. D., Perfors, A., Miller, T., Kashima, Y. and Beger, A., 'Explainable models for forecasting the emergence of political instability', *Plos one*, Vol. 16, No 7, 2021, p. e0254350.

Beck, N., Katz, J. N. and Tucker, R., 'Taking time seriously: Time-series-cross-section analysis with a binary dependent variable', *American Journal of Political Science*, Vol. 42, No 4, 1998, pp. 1260–1288.

Berman, E., Callen, M., Felter, J. H. and Shapiro, J. N., 'Do working men rebel? insurgency and unemployment in afghanistan, iraq, and the philippines', *Journal of Conflict Resolution*, Vol. 55, No 4, 2011, pp. 496–528.

Bertelsmann Foundation, 'Bertelsmann Transformation Index 2022', *Politische Gestaltung im internationalen Vergleich. Gütersloh, Verlag Bertelsmann Stiftung*, 2022.

Bolt, J. and Van Zanden, J. L., 'Maddison style estimates of the evolution of the world economy. a new 2020 update', *Maddison-Project Working Paper WP-15, University of Groningen, Groningen, The Netherlands*, 2020.

Brandt, P. T., D'Orazio, V., Khan, L., Li, Y.-F., Osorio, J. and Sianan, M., 'Conflict forecasting with event data and spatio-temporal graph convolutional networks', *International Interactions*, 2022, pp. 1–23.

Buhaug, H., Cederman, L.-E. and Gleditsch, K. S., 'Square pegs in round holes: Inequalities, grievances, and civil war', *International Studies Quarterly*, Vol. 58, No 2, 2014, pp. 418–431.

Buhaug, H. and Rød, J. K., 'Local determinants of african civil wars, 1970–2001', *Political geography*, Vol. 25, No 3, 2006, pp. 315–335.

Caprioli, M., 'Primed for Violence: The Role of Gender Inequality in Predicting Internal Conflict', *International Studies Quarterly*, Vol. 49, No 2, 2005, pp. 161–178. ISSN 00208833, 14682478. URL `http://www.jstor.org/stable/3693510`.

Carter, D. B. and Signorino, C. S., 'Back to the future: Modeling time dependence in binary data', *Political Analysis*, Vol. 18, No 3, 2010, pp. 271–292.

Cederman, L.-E., Gleditsch, K. S. and Buhaug, H., 'Inequality, grievances, and civil war', Cambridge University Press, 2013.

Cederman, L.-E., Rüegger, S. and Schvitz, G., 'Redemption through rebellion: Border change, lost unity, and nationalist conflict', *American Journal of Political Science*, Vol. 66, No 1, 2022, pp. 24–42.

Cederman, L.-E., Warren, T. C. and Sornette, D., 'Testing clausewitz: Nationalism, mass mobilization, and the severity of war', *International Organization*, Vol. 65, No 4, 2011a, pp. 605–638.

Cederman, L.-E., Weidmann, N. B. and Gleditsch, K. S., 'Horizontal inequalities and ethnonationalist civil war: A global comparison', *American political science review*, Vol. 105, No 3, 2011b, pp. 478–495.

Cingranelli, D. L., Richards, D. L. and Clay, K. C., 'The ciri human rights dataset', 2014.

Cirillo, P. and Taleb, N. N., 'On the statistical properties and tail risk of violent conflicts', *Physica A: Statistical Mechanics and its Applications*, Vol. 452, 2016, pp. 29–45.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Alizada, N., Altman, D., Bernhard, M., Cornell, A., Fish, M. S. et al., 'V-dem [country–year/country–date] dataset v11. 1', *Varieties of Democracy Project. Available at: https://doi. org/10.23696/vdemds21*, 2021.

Crespo-Sancho, C., 'The Role of Gender in the Prevention of Violent Conflict. Background paper for the United Nations-World Bank Flagship Study, Pathways for Peace: Inclusive Approaches to Preventing Violent Conflict', *World Bank, Washington, DC*, 2017.

Davies, S., Pettersson, T. and Öberg, M., 'Organized violence 1989–2021 and drone warfare', *Journal of Peace Research*, 2022, p. 00223433221108428.

De Groeve, T., Mandrella, S., Daniel Pardo Serrano, Hachemer, P. and Vernaccini, L., 'Global conflict risk index-handbook for data and statistical analysis. data management, imputation methods and code documentation (r package)', 2015.

De Groeve, T., Vernaccini, L. and Hachemer, P., 'The global conflict risk index (gcri): A quantitative model. concept and methodology', 2014.

Dixon, J., 'What Causes Civil Wars? Integrating Quantitative Research Findings', *International Studies Review*, Vol. 11, No 4, 2009, pp. 707–735. ISSN 15219488, 14682486. URL `http://www.jstor.org/stable/40389163`.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al., 'Collinearity: a review of methods to deal with it and a simulation study evaluating their performance', *Ecography*, Vol. 36, No 1, 2013, pp. 27–46.

Dube, O. and Vargas, J. F., 'Commodity price shocks and civil conflict: Evidence from colombia', *The review of economic studies*, Vol. 80, No 4, 2013, pp. 1384–1421.

D'Orazio, V. and Lin, Y., 'Forecasting conflict in africa with automated machine learning systems', *International Interactions*, 2022, pp. 1–24.

Eck, K. and Hultman, L., 'One-sided violence against civilians in war: Insights from new fatality data', *Journal of Peace Research*, Vol. 44, No 2, 2007, pp. 233–246.

FAO, 'Food balance sheets: A handbook'. 2001.

FAO, 'Temperature change statistics 1961-2021', Tech. rep., 2022. URL `https://www.fao.org/3/cb9051en/cb9051en.pdf`.

Fearon, J. D., 'Why do some civil wars last so much longer than others?', *Journal of peace research*, Vol. 41, No 3, 2004, pp. 275–301.

Fearon, J. D., 'Governance and civil war onset', 2011.

Fjelde, H., 'Buying peace? oil wealth, corruption and civil war, 1985—99', *Journal of peace research*, Vol. 46, No 2, 2009, pp. 199–218.

Gartzke, E., 'The capitalist peace', *American journal of political science*, Vol. 51, No 1, 2007, pp. 166–191.

Gelman, A., 'Multilevel (hierarchical) modeling: what it can and cannot do', *Technometrics*, Vol. 48, No 3, 2006, pp. 432–435.

GIWPS and PRIO, 'Women, Peace and Security Index 2019/20: Tracking sustainable peace through inclusion, justice, and security for women', Washington, DC: GIWPS and PRIO, 2019.

Gleditsch, K. S., 'Transnational dimensions of civil war', *Journal of peace research*, Vol. 44, No 3, 2007, pp. 293–309.

Gleditsch, K. S. and Ward, M. D., 'A revised list of independent states since the congress of vienna', *International Interactions*, Vol. 25, No 4, 1999, pp. 393–413.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M. and Strand, H., 'Armed conflict 1946-2001: A new dataset', *Journal of Peace Research*, Vol. 39, No 5, 2002, pp. 615–637.

Goemans, H. E. and Schultz, K. A., 'The politics of territorial claims: A geospatial approach applied to africa', *International Organization*, Vol. 71, No 1, 2017, pp. 31–64.

Greenhill, B., Ward, M. D. and Sacks, A., 'The separation plot: A new visual method for evaluating the fit of binary models', *American Journal of Political Science*, Vol. 55, No 4, 2011, pp. 991–1002.

Halkia, M., Ferri, S., Joubert-Boitat, I. and Saporiti, F., 'Conflict risk indicators: Significance and data management in the gcri', *Luxembourg: Publications Office of the European Union*, 2017, p. 12.

Halkia, M., Ferri, S., Schellens, M. K., Papazoglou, M. and Thomakos, D., 'The global conflict risk index: A quantitative tool for policy support on conflict prevention', *Progress in Disaster Science*, Vol. 6, 2020, p. 100069.

Harrison, P., 'Exponential smoothing and short-term sales forecasting', *Management Science*, Vol. 13, No 11, 1967, pp. 821–842.

Hastie, T., Tibshirani, R. and Friedman, J., 'The elements of statistical learning: Data mining, inference, and prediction', Springer, New York, 2 edn., 2001.

Hegre, H., 'Democracy and armed conflict', *Journal of Peace Research*, Vol. 51, No 2, 2014, pp. 159–172.

Hegre, H., Croicu, M., Eck, K. and Högbladh, S., 'Introducing the ucdp candidate events dataset', *Research & Politics*, Vol. 7, No 3, 2020, p. 2053168020935257.

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H. and Urdal, H., 'Predicting armed conflict, 2010–2050', *International Studies Quarterly*, Vol. 57, No 2, 2013, pp. 250–270.

Hegre, H., Metternich, N. W., Nygård, H. M. and Wucherpfennig, J., 'Introduction: Forecasting in peace research'. 2017a.

Hegre, H., Nygård, H. M. and Landsverk, P., 'Can we predict armed conflict? how the first 9 years of published forecasts stand up to reality', *International Studies Quarterly*, Vol. 65, No 3, 2021, pp. 660–668.

Hegre, H., Nygård, H. M. and Ræder, R. F., 'Evaluating the scope and intensity of the conflict trap: A dynamic simulation approach', *Journal of Peace Research*, Vol. 54, No 2, 2017b, pp. 243–261.

HIIK, 'Conflict barometer 2021', *Disputes, Non-Violent Crises, Violent Crises, Limited Wars Wars*, 2022.

Hilbe, J. M., 'Negative binomial regression', Cambridge University Press, 2011.

Hoch, J., de Bruin, S. P., Buhaug, H., von Uexkull, N., van Beek, R. and Wanders, N., 'Projecting armed conflict risk in africa towards 2050 along the shared socio-economic pathways: a machine learning approach', 2021.

Honaker, J. and King, G., 'What to do about missing values in time-series cross-section data', *American journal of political science*, Vol. 54, No 2, 2010, pp. 561–581.

Hyndman, R., 'The ARIMAX model muddle'. `https://robjhyndman.com/hyndsight/arimax`. Last accessed on 2022-02-02.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. and Yasmeen, F., *forecast: Forecasting functions for time series and linear models*, 2022. URL `https://pkg.robjhyndman.com/forecast/`. R package version 8.16.

Hyndman, R. J. and Khandakar, Y., 'Automatic time series forecasting: the forecast package for R', *Journal of Statistical Software*, Vol. 26, No 3, 2008, pp. 1–22. .

IPCC, 'Climate Change 2022: Impacts, Adaptation and Vulnerability', Geneva: Intergovernmental Panel on Climate Change (www.ipcc.ch), 2022.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 'An introduction to statistical learning', Vol. 112. Springer, 2013.

Kalyvas, S. N. and Balcells, L., 'International system and technologies of rebellion: How the end of the cold war shaped internal conflict', *American Political Science Review*, Vol. 104, No 3, 2010, pp. 415–429.

King, G. and Zeng, L., 'Logistic regression in rare events data', *Political analysis*, Vol. 9, No 2, 2001, pp. 137–163.

Kreft, I. G. and De Leeuw, J., 'Introducing multilevel modeling', Sage, 1998.

Lo, A., Chernoff, H., Zheng, T. and Lo, S.-H., 'Why significant variables aren't automatically good predictors', *Proceedings of the National Academy of Sciences*, Vol. 112, No 45, 2015, pp. 13892–13897.

Lopez, A. D. and Murray, C. C., 'The global burden of disease, 1990–2020', *Nature medicine*, Vol. 4, No 11, 1998, pp. 1241–1243.

Mach, K. J., Kraan, C. M., Adger, W. N., Buhaug, H., Burke, M., Fearon, J. D., Field, C. B., Hendrix, C. S., Maystadt, J.-F., O'Loughlin, J. et al., 'Climate as a risk factor for armed conflict', *Nature*, Vol. 571, No 7764, 2019, pp. 193–197.

Mansfield, E. R. and Helms, B. P., 'Detecting multicollinearity', *The American Statistician*, Vol. 36, No 3a, 1982, pp. 158–160.

Martin, P., Mayer, T. and Thoenig, M., 'Make Trade Not War?', *Review of Economic Studies*, Vol. 75, 2008, p. 865–900.

Martin-Shields, C. P. and Stojetz, W., 'Food security and conflict: Empirical challenges and future opportunities for research and policy making on food security and conflict', *World Development*, Vol. 119, 2019, pp. 150–164.

Menard, S., 'Applied logistic regression analysis', 106. Sage, 2002.

Muchlinski, D., Siroky, D., He, J. and Kocher, M., 'Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data', *Political Analysis*, Vol. 24, No 1, 2016, pp. 87–103.

Pettersson, T., Davies, S., Deniz, A., Engström, G., Hawach, N., Högbladh, S. and Öberg, M. S. M., 'Organized violence 1989–2020, with a special emphasis on syria', *Journal of Peace Research*, Vol. 58, No 4, 2021, pp. 809–825.

Ross, M. L., 'What have we learned about the resource curse?', *Annual review of political science*, Vol. 18, 2015, pp. 239–259.

Russett, B., Singer, J. D. and Small, M., 'A standardized list of national political entities in the twentieth century', *American Political Science Review*, Vol. 62, No 2, 1968, pp. 932–51.

Salehyan, I., 'Rebels without borders: transnational insurgencies in world politics', Cornell University Press, 2011.

Schneider, G. and Gleditsch, N. P., 'The capitalist peace: The origins and prospects of a liberal idea', *International Interactions*, Vol. 36, No 2, 2010, pp. 107–114.

Schrodt, P. A., 'Seven deadly sins of contemporary quantitative political analysis', *Journal of peace research*, Vol. 51, No 2, 2014, pp. 287–300.

Shumway, R. H. and Stoffer, D. S., 'Arima models'. In 'Time series analysis and its applications', Springer, 2017a. pp. 75–163.

Shumway, R. H. and Stoffer, D. S., 'Time series analysis and its applications', Springer, New York, 4 edn., 2017b.

Steenbergen, M. R. and Jones, B. S., 'Modeling multilevel data structures', *american Journal of political Science*, 2002, pp. 218–237.

Stewart, F., 'Horizontal inequalities as a cause of conflict: A review of crise findings', 2011.

Sundberg, R., Eck, K. and Kreutz, J., 'Introducing the UCDP non-state conflict dataset', *Journal of Peace Research*, Vol. 49, No 2, 2012, pp. 351–362.

Sundberg, R. and Melander, E., 'Introducing the ucdp georeferenced event dataset', *Journal of Peace Research*, Vol. 50, No 4, 2013, pp. 523–532.

Vicente-Serrano, S. M., Beguería, S. and López-Moreno, J. I., 'A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index', *Journal of climate*, Vol. 23, No 7, 2010, pp. 1696–1718.

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P. and Girardin, L., 'Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family', *Journal of Conflict Resolution*, Vol. 59, No 7, 2015, pp. 1327–1342.

Ward, M. D., Greenhill, B. D. and Bakke, K. M., 'The perils of policy by p-value: Predicting civil conflicts', *Journal of peace research*, Vol. 47, No 4, 2010, pp. 363–375.

Weidmann, N. B. and Ward, M. D., 'Predicting conflict in space and time', *Journal of Conflict Resolution*, Vol. 54, No 6, 2010, pp. 883–901.

World Bank, 'Databank: world development indicators', 2020.

Ying, X., 'An overview of overfitting and its solutions', In 'Journal of physics: Conference series', Vol. 1168. IOP Publishing, p. 022022.

List of abbreviations and definitions

ACLED Armed Conflict Location and Event Dataset

ARIMA Autoregressive Integrated Moving Average

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

COW Correlates of War Dataset

DCRM Dynamic Conflict Risk Model

EEAS European External Action Service

EPR Ethnic Power Relations Dataset

EWS Early Warning System

FPI Service for Foreign Policy Instruments

GCRI Global Conflict Risk Index

GED Georeferenced Event Dataset

GW Gleditsch and Ward Country List

JRC Joint Research Centre

LOCF Last Observation Carried Forward

NSC Non-State Conflict

NP National Power

OSV One-Sided Violence

SBC State-Based Conflict

SN Sub-National

RMSE Root Mean Squared Error

SPEI Standardized Precipitation Evapotranspiration Index

UCDP Uppsala Conflict Data Program

V-DEM Varieties of Democracy Dataset

ViEWS Violence Early-Warning System

VIF Variance Inflation Factor

WID World Inequality Database

List of figures

List of tables

Annexes

Variable	Min.	Max.	Mean	Median	Std. Dev
Democracy	0	0.89	0.4	0.36	0.27
State capacity	-3.04	2.93	0.92	0.9	1.22
Repression	-3.38	3.09	-1.05	-1.25	1.51
Corruption	0	0.97	0.51	0.56	0.3
Recent internal conflict, log	0	12.99	1.54	0	2.77
Years since last conflict	0	32	10.23	7	10.24
Neighboring conflict	0	7	1.12	1	1.38
Homicide rate	0.05	106.82	7.81	4.13	10.43
Female empowerment	0.04	0.97	0.7	0.75	0.2
Ethnic exclusion	0	1	0.16	0.08	0.22
Transnational ethnic ties	0	19	2.7	2	2.79
GDP per capita, log	6.08	11.7	9.13	9.21	1.21
Income inequality	0.32	0.84	0.57	0.58	0.08
Trade openness	0.02	437.33	82.98	71.68	51.36
Oil exports	0	66.71	3.78	0.02	9.34
Food security, log	7.32	8.28	7.91	7.93	0.18
Unemployment	0.1	38.8	8.13	6.33	6.28
Droughts	-7.2	2.97	-0.16	-0.1	0.89
Temperature change	-1.28	3.7	1.02	0.98	0.64
Population, log	5.26	14.18	9.1	9.15	1.65
Youth bulge	0.09	0.28	0.19	0.2	0.04
Child mortality	1.44	341.2	44.92	24.23	49.25

Table A1: Summary statistics, GCRI predictor variables, after imputation (1991-2021)

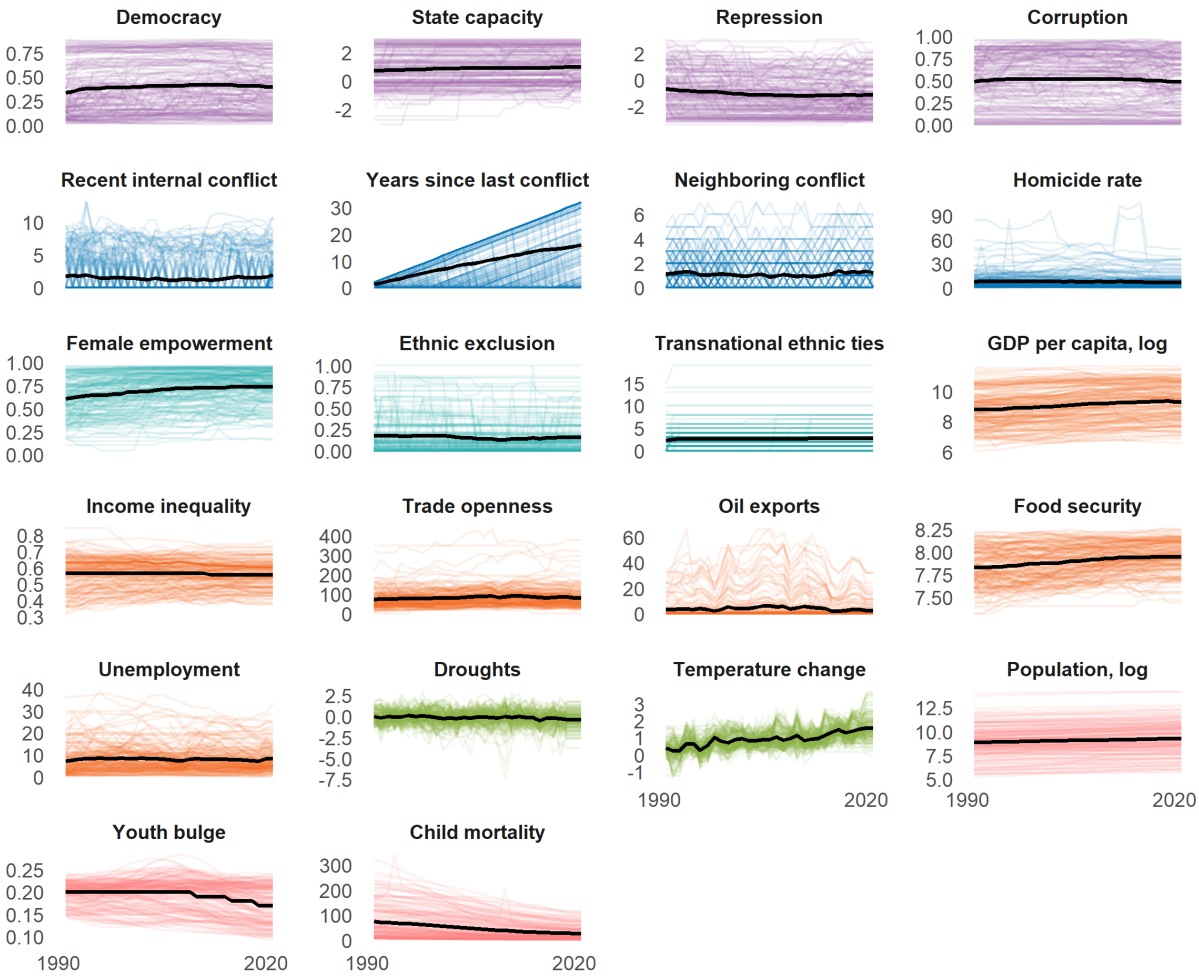

Figure A1: Variable distribution after imputation. Colored lines represent individual country values, black lines indicate the dataset average.

	Variable	GCRI label	Source	Source variable	Source label	Comment
Political	Democracy	DEMOCRACY	V-DEM	Liberal democracy index	v2x_libdem	
	State capacity	STATE_CAPACITY	V-DEM	State fiscal source of revenue	v2stfisccap	
	Repression	REPRESSION	V-DEM	Freedom from political killings	v2clkill	
	Corruption	CORRUPTION	V-DEM	Political corruption index	v2x_corr	
Security	Recent internal conflict	LN_CON_[xxx]_HI_EST	UCDP			UCDP data aggregated to country level.Log of battle deaths in T-1
	Years since last conflict	CON_[xxx]_YRS	UCDP			UCDP data aggregated to country level.
	Neighboring conflict	CON_[xxx]_NB_CT	UCDP			UCDP data aggregated to country level.
	Homicide rate	HOMIC	IHME	Homicide deaths per 100000 people		
Social	Female empowerment	FEMALE_EMPOWER	V-DEM	Women political empowerment index	v2x_gender	
	Ethnic exclusion	ETHNIC_EXCL	EPR	% of population that belongs to politically excluded groups	lexclpop	
	Transnational ethnic ties	TEK_COUNT	EPR	Number of ethnic groups with kin groups across the border		Variable derived from EPR-TEK and CShapes data on country borders.

Table A2: Construction of GCRI variables, overview.

	Variable	GCRI label	Source	Source variable	Source label	Comment
Economy	GDP per capita, log	LN_GDP_PC	World Bank	Trade (% of GDP)	NE.TRD.GNFS.ZS	
	Income inequality	INEQUALITY	WID	Gini coefficient, pre-tax national income, adults (equal-split)	gptinc992j	
	Trade openness	TRADE_DEP	World Bank	Unemployment, total (% of total labor force) (modeled ILO estimate)	SL.UEM.TOTL.ZS	
	Oil exports	OIL_DEP	World Bank	Oil rents (% of GDP)	NY.GDP.PETR.RT.ZS	
	Food security	LN_FOOD_KCAL_PC	FAO	Temperature change	°C	
	Unemployment	UNEMPLOYMENT	World Bank	GDP per capita, PPP (constant 2017 international $)	NY.GDP.PCAP.PP.KD	
Geography - Environment	Droughts	SPEI_AVG	SPEI/CSIC	Standardized Precipitation-Evapotranspiration Index		Variable derived from SPEI and CShapes data on country borders.
	Temperature change	TEMP_CHG	FAO	Grand Total Food supply	kcal/capita/day	Natureal log of kcal/capita/day
Demographics	Population, log	LN_POP	UN	Total population		Sum across all age groups, logged
	Youth bulge	YOUTH_BULGE	UN			% of Population aged 15-25
	Child mortality	CHILD_MORT	World Bank	Mortality rate, under-5 (per 1,0 live births)	SH.DYN.MORT	

Table A3: Construction of GCRI variables, overview.

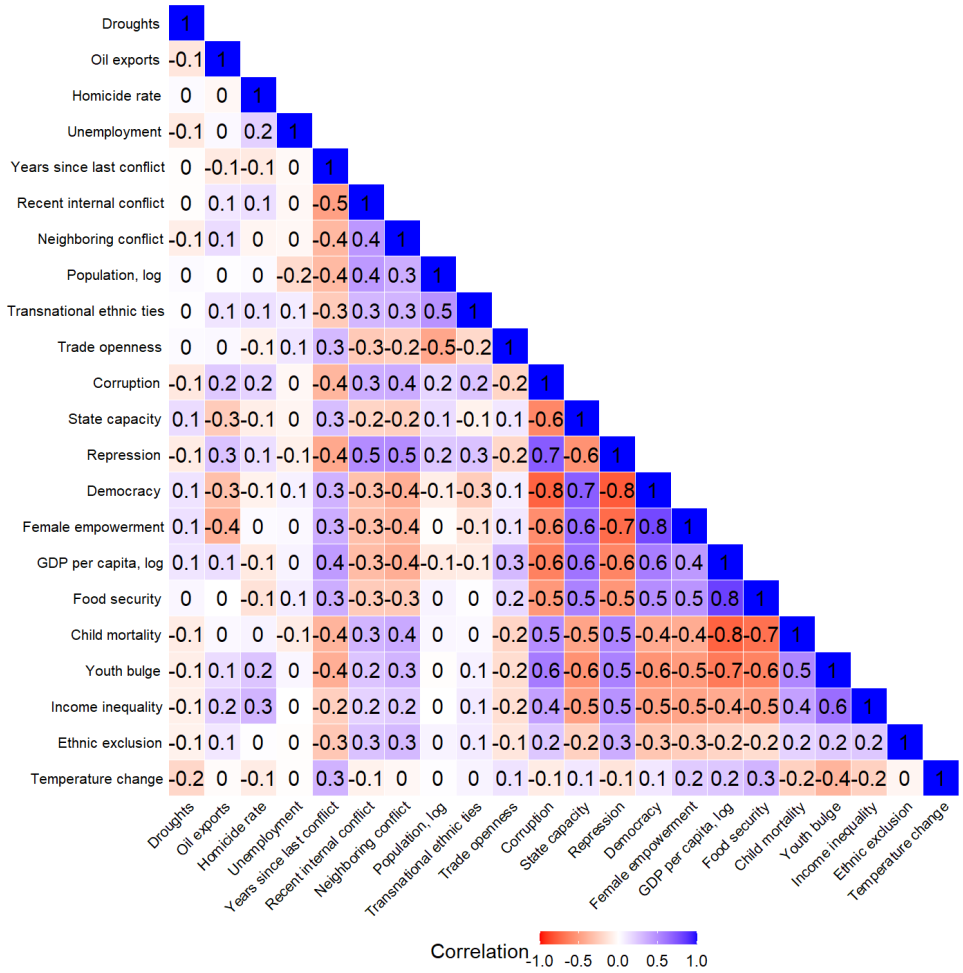

Figure A2: Pairwise correlations between final set of 22 variables

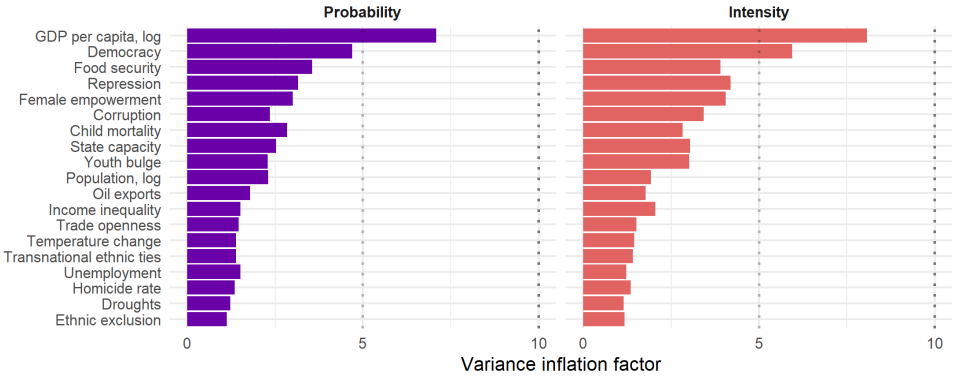

Figure A3: Multicollinearity for the final set of 22 variables. Dotted lines indicate thresholds of potentially problematic (left) and concerning (right) levels of multicollinearity

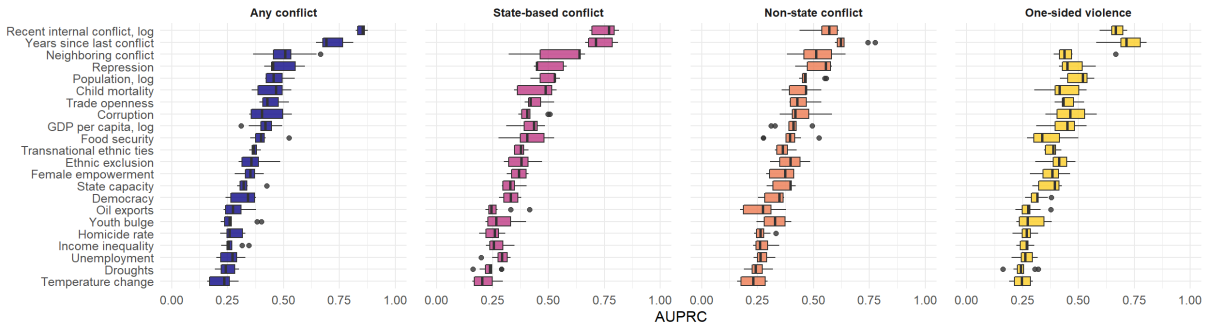

Figure A4: Variable importance, final set of 22 variables. Variables are sorted from most to least important based on their predictive performance compared to an empty model. Box plots show distribution of AUPRC scores over 10 cross-validations.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

EU Science

Publications Office
of the European Union