journal of peace RESEARCH

Early warning signals for war in the news

Journal of Peace Research
201X, Vol XX(X) 1–14

DOI: 10.1177/0022343313507302
jpr.sagepub.com
SAGE

Thomas Chadefaux

ETH Zurich

Abstract

There have been more than 200 wars since the start of the 20th century, leading to about 35 million battle deaths. However, efforts at forecasting conflicts have so far performed poorly for lack of fine-grained and comprehensive measures of geopolitical tensions. In this article, a weekly risk index is derived by analyzing a comprehensive dataset of historical newspaper articles over the past century. News reports have the advantage of conveying information about contemporaries' interpretation of events and not having to rely on meaning inferred *a posteriori* with the benefit of hindsight. I applied this new index to a dataset of all wars within and between countries recorded since 1900, and found that the number of conflict-related news items increases dramatically prior to the onset of conflict. Using only information available at the time, the onset of a war within the next few months could be predicted with up to 85% confidence and predictions significantly improved upon existing methods both in terms of binary predictions (as measured by the area under the curve) and calibration (measured by the Brier score). Predictions also extend well before the onset of war – more than one year prior to interstate wars, and six months prior to civil wars – giving policymakers significant additional warning time.

Keywords

media, tensions, war onset, war prediction, warning signals

> There will be no European general war.... The six great powers
>
> – Germany, Austria and Italy on one side, and Russia, France and Britain on the other side – cannot afford a clash of arms .... [They] will hesitate at the last moment and endeavor to adjust matters.
>
> *Los Angeles Times,* 27 July 1914

Introduction

Up until the very outbreak of World War I on 28 July 1914, reporters took little notice of the rising tensions and the brewing conflict in Europe. In fact, in the week preceding the onset of war, worldwide newspapers mentioned 'tensions' or 'conflict' no more than at almost any time during the previous 15 years. In other words, WWI seems to have come largely as a general surprise. This is in sharp contrast with World War II, for which the rise of tensions was echoed by a steadily growing attention from the press since at least 1935.

These differences in the anticipation of war are striking and raise a number of questions. Do wars usually come unexpectedly, or is the buildup of tensions visible and the outbreak of conflict predictable? Are there systematic differences in our ability to anticipate wars – for example, are conflicts with high levels of casualties easier to anticipate than the relatively costless ones? Or perhaps the type of war – interstate or intrastate – is the determining factor? And, using only information available at the time, could we have derived earlier warning signals for war?

Unfortunately, a recurrent difficulty in predicting wars has been the absence of measures of tensions that are both fine-grained and comprehensive (Holsti, 1963; Newcombe, Newcombe & Landrus, 1974; Choucri, 1974). Historical studies of single wars abound but are hardly quantifiable, rely on hindsight, or ignore the

Corresponding author:
thomasc@ethz.ch

equally important cases in which war did not occur (Leetaru, 2011). Others have focused on the conditions that are most conducive to war, but the indicators used are typically yearly, thereby missing the escalation of tensions and the timing of the conflict outbreak (Beck, King & Zeng, 2000, 2004; De Marchi, Gelpi & Grynaviski, 2004; Gleditsch & Ward, 2011). In addition, these indicators are often poorly harmonized across countries, and their estimation (e.g. military spending) might depend on the government's goodwill or strategic interests. Finally, they cannot reliably measure the *perceived* reality of the time. Contemporaries may have been oblivious to real risk factors or, on the contrary, might have imagined them where none existed (Holsti, 1963).

To fill this gap I derived, for a large number of countries and times, a comprehensive estimate of tensions – a situation of stress and latent hostility – within and between countries. Tensions were estimated by analyzing a large dataset of historical newspaper articles. The press is an ideal source of information because it provides fast, accurate and in-depth coverage of rising tensions throughout the world. Pre-conflict escalations (Senese & Vasquez, 2010; Moul, 1988; Siverson & Tennefoss, 1984) and the process of offers and counter-offers, threats and signalling associated with it (see Powell [2002] for a review) should be partly observable by contemporaries and relayed in the media, whose reputation is based upon the provision of accurate information. A database of news reports also avoids the problem of hindsight by using only information available at the time. Finally, newspapers have an important advantage over event-base data: they can report tensions even when no actual event occurred (and hence nothing is recorded in the MID or COPDAB data). Conversely, an event might occur but not be perceived as significant by its contemporaries. In other words, news reports convey information about contemporaries' interpretation of events – or the absence thereof – and not an event description from which meaning needs to be inferred *a posteriori* with the benefit of hindsight.

The resulting dataset is a fine-grained and direct proxy for the evolution of tensions in each country. It is used here to derive an estimate of the probability of a coming war, which is then tested on existing conflict datasets – including all inter-, intra- and extrastate conflicts recorded with a starting date from January 1902 to December 2001.

Four main questions are addressed in this article. First, do newspaper articles report growing tensions, or does war usually come as a surprise? Second, can a reliable risk-index be derived from an analysis of newspapers? And would this index improve upon predictions made using only yearly variables such as military spending or regime type? Third, are different types of war – inter- or intrastate, large or small – better predicted than others? Finally, how much warning time can news provide?

The article proceeds in four steps. I first review the relevant literature on predicting conflicts. I then introduce new data on tensions collected by analyzing a large database of newspapers and discuss the conflict data on which the tensions estimates will be tested. I then present evidence that the number of reports about tensions typically rises well ahead of a conflict, and that the number of conflict-related news items is a significant predictor of conflict. Finally, I show that early warning signals can be derived from these data and used as reliable predictors of wars, using only information available at the time. I also analyze the type of war best predicted – by type and casualties – and how far ahead of time metrics about news items can provide information.

Measures of geopolitical tensions

There is a general tradeoff in existing measures of tensions between breadth – the extent of their temporal or geographic coverage – and depth – the data's coarseness (e.g. daily vs. yearly). For example, historical studies of single wars are invaluable for their depth of information and level of analysis, but are highly time-consuming and usually rely on hindsight – ignoring the dogs that did not bark. Thus, of the hundreds of books about World War I, only a few focus instead on the absence of outbreak in 1913.

At the other end of the depth–breadth scale, the international and comparative conflict literature has taken a more systematic approach to the measurement of geopolitical risk by deriving the *conditions* most conducive to war. Arms races (Glaser, 2000), longstanding territorial rivalries (Huth, 1998), large and rapid shifts in power (Powell, 2004) or rough terrain (Fearon & Laitin, 2003) are some of the factors that have been associated with an elevated risk of conflict, either internationally or domestically. This approach has the advantage of identifying the root causes of tensions in data with a very large time and geographic span, and some of these variables are incorporated in the present models. However, the indicators used are typically yearly, thereby missing important parts of the escalation and the timing of the conflict outbreak, and they might be imprecise or even manipulated (Lebovic, 1998, 1999).

Others have attempted to quantify tensions more directly. Thus the Conflict and Peace Data Bank is a 'library of daily international and domestic events or interactions' (Azar, 1980) and the World Events Interaction Survey 'a record of the flow of action and response between countries (as well as non-governmental actors, e.g. NATO) reflected in public events reported daily in the New York Times from January 1966 through December 1978' (McClelland, 1984). While these data document relevant events such as international border clashes or domestic press censorship with sufficient frequency and detail, coding is labor-intensive and the data's time coverages are, as a result, limited (see also Weidmann & Ward [2010] and Bernauer & Gleditsch [2012] for more recent efforts).[1] A more systematic coding effort is by King & Lowe (2003), but the time span is limited and only interstate wars are covered.

The literature on forecasting also suffers from the trade-off between breadth and depth. Predicting conflict has recently received increasing attention, whether it be for interstate wars (Beck, King & Zeng, 2000; Ward, Siverson & Cao, 2007), civil wars (Ward, Greenhill & Bakke, 2010), or other political disruptions, from state failure to political instability, genocides, human rights violations or ethnic conflicts (Schneider, Gleditsch & Carey, 2010; Bueno de Mesquita, 2009; Goldstone et al., 2010). Some work has focused on predicting the evolution of a particular conflict (Pevehouse & Goldstein, 1999; Schrodt & Gerner, 2000), sometimes using fine-grained data (Bosler & Schneider, 2011), but the time span used is limited, and the external validity of these studies is difficult to assess. Similar problems affect prediction markets (Wolfers & Zitzewitz, 2006; Berg, Nelson & Rietz, 2008). More recent projects such as Swisspeace's FAST project or the European Media Monitor project also gather large amounts of media data for the purpose of developing early warning signals, but typically focus only on the recent past and on different subjects (e.g. threats to public health). Game-theoretic approaches have also yielded encouraging results (Bueno de Mesquita, 2002; Feder, 2002), although they are usually poor at predicting the timing of a particular event. Moreover, they typically rely on detailed information from issue or area experts, so that their generalizability remains unclear.

While these problems have long been recognized (Holsti, 1963; Newcombe, Newcombe & Landrus, 1974; Choucri, 1974), there still exists no fine-grained, comprehensive data of tensions. An aggregate analysis of news articles can serve such a function.[2] The aim here is to develop an index that is both broad in time and space and fine-grained. In addition, I demonstrate the validity of the index derived by showing its ability to correctly predict wars over the past 100 years. This test of the index could validate its use as a proxy for tensions, so that it could be used to address further questions that have so far eluded researchers for lack of fine-grained and long-term data.

The data

Measuring tensions

To estimate domestic and international tensions, I relied on Google's database of newspapers, *Google News Archive*.[3] This wide collection includes a large proportion of all English-speaking newspapers, ranging from major publications such as the *New York Times*, the *Washington Post* or the *Guardian*, to more obscure local ones such as *California Oil Worker* or the *Cambridge City Tribune*.[4] In all, the database spans more than 200 years and consists of over 60 million pages. It also includes as subsets major providers of news archives such as Proquest Historical Newspapers, thereby making it the world's largest database in terms of the number of articles referenced. This comprehensiveness has the added advantage of smoothing out any particular newspaper's biases, such as those caused by their geographic location (Thai newspapers, say, might not have written as much about WWI as Germany's), their political orientation (conservative or liberal) or their substantive focus (politics, economics or art).

Within this data, the entire text of every article was searched for every week from 1902 to 2001 (data prior to 1902 in Google's database were less reliable; MID data from the CoW project stop in 2001). I then counted the number of articles mentioning a given country, together with a set of keywords typically associated with tensions. The list of keywords, generated using a

[1] 1948–78 for COPDAB, and 1966–78 for WEIS.

[2] Deutsch's idea that a careful analysis of the media could yield early warning signals for interstate conflicts (Deutsch, 1957; George, 1956) is relevant here. His insight was exploited systematically by Hunt (1997), who provides a methodology for identifying a regime's intention to launch a conflict in advance of the actual initiation using media analysis. However, his strategy also involves significant human coding and, as a result, Hunt's analysis is largely limited to positive cases – examining the pattern of editorials prior to wars or crises.

[3] See http://news.google.com/archivesearch.

[4] See http://news.google.com/newspapers for a partial list.

thesaurus to avoid any personal or linguistic bias, is the following: *tension(s), crisis, conflict, antagonism, clash, contention, discord, dissent, disunion, disunity, feud, division, fight, hostility, rupture, strife, attack, combat, shell, struggle, fighting, confrontation, impasse*. Thus, a sample search would be 'France AND tensions OR crisis OR conflict . . . ' for newspapers published between 22 July and 29 July 1914.[5] This search yielded six results, indicating that six newspaper articles mentioned at least one of the keywords in their text. This procedure was repeated for every week from 1 January 1902 to 31 December 2001, and for every country included in the Correlates of War dataset (Correlates of War Project, 2008).[6] The resulting dataset consists of 100 years worth of weekly time series for 167 countries.[7]

Examples of search results include positive hits such as: 'Sir Edward Grey seeking conference to avert a general *conflict*. *France* and Italy agree' (*New York Times*, 28 July 1914, emphasis added); 'Conditions that never before were witnessed in the foreign exchange market here resulted yesterday from the possibility of the trouble between Serbia and Austria involving *France* . . . . in such a *crisis*' (*New York Times*, 28 July 1914, emphasis added). However, the dataset remains imperfect. For example, a sentence such as 'oil revenues surge in Norway due to an increased risk of war in Iraq' would mistakenly be coded as an increase in tensions for Norway, as well as Iraq.

Moreover, this crude search does not allow us to determine who will fight with whom. We might for example infer that both France and Germany are experiencing tensions, but not whether these tensions are in relation to one another, to a third country, or simply happen to spike at the same time for altogether different reasons. In addition, while the index does improve our ability to predict all types of war, it is not a tool to predict the *type* of war that will occur. It simply tells us whether a conflict will occur at all – not its type or its participants. Moreover, although the entire text can be searched for specific keywords or sentences, legal access limitations imply that the content cannot be processed for more complex analyses (in contrast to King & Lowe [2003]). Thus, 'war will not occur' increases the estimate of tensions to the same extent as 'war will occur'.[8] Yet, a newspaper contributor writing about her belief that a conflict will *not* break out still reveals existing concerns that need to be dispelled and, as such, should be treated as a sign of tensions.

Finally, some rare mistakes were detected in Google's article-dating algorithm (only two such errors were found, both related to WWI). For example, an article reporting on World War I commemorations between France and Germany in 2000 might mention 28 July 1914 in its body, leading Google's classification algorithm to wrongly assume this was published in 1914. However, the predictions of WWI only use data up to the week prior to conflict, so that this isolated erroneous entry on 28 July 1914 does not affect the predictions for this war.

While the data could certainly be improved in the future, they are to my knowledge the most comprehensive, systematic and uniform estimate of tensions throughout countries, and I show that even the simple count used here already produces substantial results.

Conflicts

Conflicts can be broadly categorized according to their scale and the actors involved. First, conflicts range from aggressive speeches to the simple display of force to full-scale wars with thousands of deaths. Second, they can involve only states ('interstate conflicts'); one state against rebel groups (civil war or 'intrastate'); or states with non-state armed groups with no defined territorial base (extrastate wars) (Sarkees & Wayman, 2010).

I study the rise of tensions for all militarized conflicts included in the Correlates of War (CoW) or the MID data (Sarkees & Wayman, 2010; Faten, Palmer & Bremer, 2004).[9] This includes all inter-, intra- and extrastate conflicts recorded with a starting date of January 1902 to December 2001 (see Table I for a breakdown). Note

[5] More details on the search string and URL used can be found in the online appendix.

[6] The list of countries includes only those that still exist today and therefore excludes countries such as Austria-Hungary.

[7] Undoubtedly, the list remains ad hoc, and another set of keywords may better measure tensions or predict conflict. To ensure the results' robustness, I have therefore collected counts of news items for other sets of words (but only for a limited set of countries, given the computational challenges involved). Overall, no significant qualitative differences were found in the results. For example, using only *tensions* as a keyword led to a lower total number of news items (since it excludes all news articles mentioning for example *conflict* but not *tensions*), but to a time series highly correlated with ours.

[8] This limitation could be circumvented by relying on full-text databases such as Proquest Historical News, but their scope is far more limited than Google News Archive.

[9] See Intrastate war data v4.1, interstate war data v4.0 and version 3.0 of the Extra-State War data set; MID v3.0.

Table I. Frequencies of country-events (1902–2001) by conflict type and battle death count

	Battle deaths			
Type	*<1,000 ('minor')*	*(1-10,000) ('small')*	*10,000+ ('major')*	*Total*
Interstate wars	4,045	145	30	4,220
Intrastate wars	0	201	33	234
Extrastate wars	3	21	5	29
Total	4,048	367	68	4,483

A country-event is coded as 1 for every week in which a conflict breaks out (but not for an ongoing conflict), 0 otherwise.

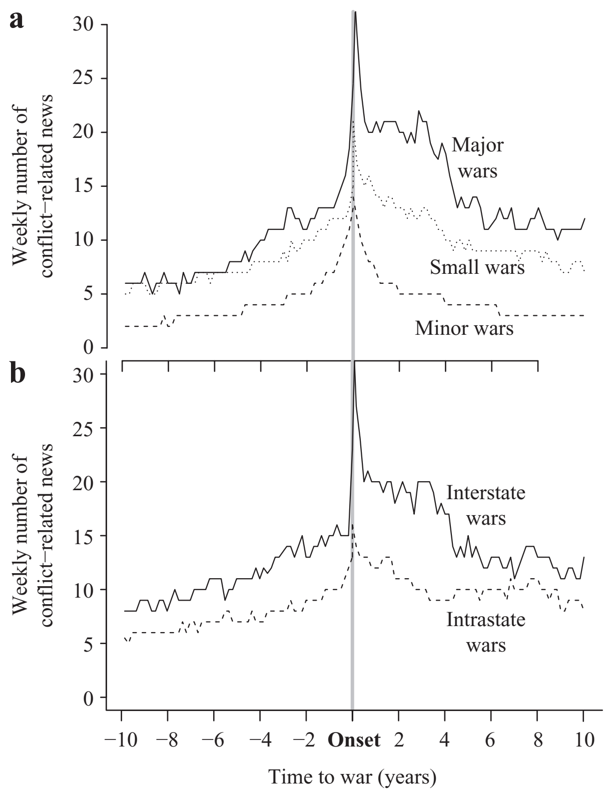

Figure 1. Median weekly number of conflict-related news items as a function of time to conflict
a. Evolution by war size. Major wars have at least 10,000 battle deaths; small wars have [1000, 10000); minor wars have less than 1,000.
b. Evolution by war type: interstate and intrastate wars with at least 1,000 battle deaths.

also that country-conflicts are used, which implies that World War I includes an entry for all 15 participants listed in the CoW data. In total, this means that the dependent variable is composed of 4,952 wars, broken down by type and casualty levels as reported in Table I. This variety will test the ability of the present measure of tension to announce not only large-scale interstate wars, but also bloodless domestic conflicts.

Conflict-related news items signal geopolitical risk

Bivariate relationship

The average weekly number of conflict-related news items (mode = 0, mean = 28.28, median = 6, s.d. = 126.92) varies considerably in time – for example increasing in the United States from an average of 28 in the 1900s to 209 in the 1990s – and space – ranging from 0.12 for Suriname to 376 for Georgia. However, conflict-related news counts dramatically increase in the months and years preceding conflict, and rapidly recede thereafter (see Figure 1).

This trend applies independently of war size (casualties) or type (inter- or intrastate). The pattern emerges remarkably early: a visible upward trend appears at least three to five years before large wars, and two to four years before minor wars. Unsurprisingly, I also find that the number of conflict-related news items is much higher within the year that precedes the outbreak of war than at other times, for wars of any scale or type (see Table II).

Yet, the number of conflict-related news items may simply reflect changes in other variables (e.g. military spending) and hence would not carry additional information. Moreover, it does not inform us about the evolution of the number of conflict-related news items in cases where war does *not* occur. I therefore tested the specific explanatory power of news with a logistic regression model to which, in addition to conflict-related news counts, I added potentially confounding variables.

Multivariate analysis

The following standard logit model was fitted:[10]

[10] Note that a standard logit is used here, without correcting for biases associated with rare events, because the dependent variable – the occurrence of a conflict within the next three months – actually occurs in about 13% of cases (though it is rarer for wars involving a larger number of casualties).

Table II. Comparison of the median number of conflict-related news items one year prior to war and its overall median

Conflict type	W	$\overline{W}$	U	n_1	n_2	*p (one-tailed)*
All events	7	4	1.790×10^{10}	97,686	328,008	< 0.001
Interstate wars	15	6	2.166×10^{9}	6,809	503,587	< 0.001
Intrastate wars	11	5	3.534×10^{9}	12,357	477,625	< 0.001
Large-scale wars	13	6	1.408×10^{9}	4,676	496,601	< 0.001
Small wars	11	5	4.594×10^{9}	16,491	467,973	< 0.001
Minor wars	7	5	2.075×10^{10}	104,697	354,004	< 0.001

Mann-Whitney U statistic comparing the median number of conflict-related news items within the year preceding war (W) to its median during weeks not within one year of a conflict ($\overline{W}$), for various categories of conflict. We use the Mann-Whitney U statistics instead of the typical t-test because of the skewed distribution of the number of conflict-related news items. However, similar results hold with the t-test.

$$\mathbb{E}[Y_{it}|X_{it}] = 1/(1 + e^{-\beta X_{it}}),$$

where $Y_{it} \in \{0, 1\}$ is the occurrence or not in country i and at time t of a war within the next three months, $\beta = [\beta_0, \beta_1, \dots, \beta_k]$ is a vector of coefficients, and $\mathbf{X}_{it}$ is a matrix of explanatory variables which, in addition to conflict-related news, includes other variables that are likely to affect a country's odds of experiencing a conflict: a variable measuring a country's national material capabilities using the Composite Index of National Capability (*Cinc*) from the Correlates of War (Singer, Bremer & Stuckey, 1972).[11] A variable measuring the yearly change in a country's *Cinc* was also included to account for the explanation that large and rapid shifts in power can lead to war (Organski & Kugler, 1981; Chadefaux, 2011). The effect of political regime types (e.g. democracy vs. autocracy) on the probability of war has also been the subject of much research (Ray, 1998). A *Polity* variable (Marshall, Jaggers & Gurr, 2010) was therefore also added.

I also included the total number of news items in the world (*World news*) and its interaction with news counts at the state level to control for the fact that increases in conflict-related news in one country do not necessarily indicate an increase in tensions, but might simply reflect a general trend (for example because the total number of news articles is skewed toward more recent years).

Finally, I included a variable measuring the number of weeks since the last conflict (*Peace weeks*) as a measure of temporal dependence, as well as the square and cube of this variable (see Beck, Katz & Tucker [1998] and Carter & Signorino [2010]). This variable controls for the possibility that conflict is more likely to erupt after previous disputes than after a long period of peace.[12]

Several models were tested. Model 1 is the baseline model, including only a constant and the 'time since conflict' variables. Model 2 adds the weekly number of conflict-related news items to Model 1, but no control variables. Model 3 is the 'structural' model, including variables that have been found to be important predictors of conflict in the literature (but not conflict-related news). Model 4 includes all control and interaction variables described above, together with conflict-related news.

The results of the multivariate model confirm the results of the simple correlation described above. The risk of war significantly increases with the number of conflict-related news items, even after controlling for the structural variables described above. This result applies whether conflict is defined as events with more than 10,000 battle deaths; those with deaths between 1,000 and 10,000; or those with less than 1,000 deaths. It also holds regardless of war type, for interstate, intrastate and extrastate wars alike (see Tables III and IV).

Out-of-sample predictive power of news

While an increased number of conflict-related news items signals a higher risk of war, forecasting wars remains a needle in a haystack problem. A better test of the value added

[11] Studies have suggested that neural networks might yield better predictions than simple logits (Beck, King & Zeng, 2000). However, I chose the standard logit here because it is more transparent and easier to interpret (De Marchi, Gelpi & Grynaviski, 2004). The goal is not to generate the most accurate prediction but rather to show in the simplest manner that the introduction of a simple estimate of tensions yields highly improved predictive results, when compared to a model without such an index. I expect this improvement to apply to more complex models, which would actually be able to make better use of the additional information offered by the 'news' variable. Finally, the method is most likely irrelevant, as all three models are estimated using the very same method.

[12] Other variables (e.g. GDP per capita, population size or ethnic cleavages) may be important predictors for conflict, but they are rarely available prior to 1945. In addition, some of the variables used here (e.g. *Cinc*) include variables that can act as proxies for these variables (e.g. proportion of urban population as a proxy for GDP per capita).

Table III. Logit models of the onset of conflict within the next three months

	Model 1 (*'Baseline'*)		*Model 2*		*Model 3* (*'Structural'*)		*Model 4* (*'News'*)	
Const.	−1.332***	(0.006)	−1.720***	(0.011)	−1.683***	(0.009)	−2.400***	(0.033)
Time since conflict	−0.001***	(0.000)	−0.001***	(0.0)	−0.001***	(0.0)	−0.001***	(0.0)
Time since conflict^2	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)
Time since conflict^3	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)
News (lag)			0.158***	(0.003)			**0.486*****	(0.011)
War ongoing					0.327***	(0.011)	0.230***	(0.012)
Cinc (lag)					8.356***	(0.115)	6.153***	(0.126)
Power Shift					−7.732***	(1.135)	−5.958***	(1.118)
Polity (lag)					−0.006***	(0.001)	−0.007***	(0.001)
World news (lag)							0.227***	(0.017)
News × world news							−0.162***	(0.005)
Nagelkerke R-sq.	0.114		0.123		0.141		0.152	
Likelihood-ratio	27336		29585		33980		36772	
BIC	291568		289333		284977		282224	

Logit models regressing the onset of conflict of any size or type within the next three months against various variable combinations. The model with conflict-related news reported in the main text is Model 4. The 'structural' model is Model 3. The baseline model uses only a 'days since the last conflict' variable (Model 1). Standard errors are reported in parentheses; $N = 475,715$. $^*p < 0.5$, $^{**}p < 0.01$.

Table IV. Logit models of the onset of different types of conflict within the next three months

	β_{News}	*BIC Baseline*	*BIC Structural*	*BIC News*	*LR test*
Any size/type	.486** (.011)	291,566	284,977	282,223	< .001
Deaths ∈[1k,10 k)	.350** (.027)	53,365	52,918	52,380	< .001
Deaths > 10k	.707** (.070)	16,059	15,867	15,644	< .001
Interstate (> 1k deaths)	.741** (.052)	20,013	19,808	18,972	< .001
Intrastate (> 1k deaths)	.322** (.033)	38,387	37,892	37,775	< .001
Extrastate (> 1k deaths)	.692** (.180)	4,669	4,698	4,667	< .001

Results of logit models using different dependent variables. The first column reports the coefficient for the 'News' (logged) variable. Thus, the top left number (.446) corresponds to the bolded cell in Table III. $^*p < 0.5$, $^{**}p < 0.01$. The next three columns report the AIC statistic for the three main models (Models 1, 3 and 4) in Table III. The last column reports the likelihood ratio tests of the model including conflict-related news (Model 4 in Table III) against the same model without conflict-related news. The likelihood ratio test is a method for hypothesis testing by which the fit of a dataset to a more complex model is compared with its fit to a simpler model using the likelihood ratio statistic (twice the ratio of the likelihoods of the two models). The improvement in fit is evaluated using a χ^2 distribution.

of the number of conflict-related news items is its ability to improve predictions. That is, can the probability of a coming war be better estimated using the measure of tensions derived from newspapers than without?

To evaluate the predictive power of the model, I computed a large number of out-of-sample predictions.[13] Using a randomly selected time t as a reference point, all prior data (i.e. $[1902, t)$) were used as a 'learning' set from which the model's coefficients were derived. I then used these coefficients to derive out-of-sample predictions about the *following* year (i.e. $[t+1, t+2]$.[14]

[13] Although a cross-validation approach (e.g. K-fold cross-validation) yields similar results here (see online appendix, Figure 1), it is not reproduced here because cross-validation can be problematic when the data are not independent (as is the case for time series), since leaving out an observation (or a group of them) does not remove all the associated information due to the correlations with other observations. Even worse in the case of time series, cross-validation implies using future information to predict the past.

[14] Predictions start at $t + 1$ – that is, I use a 'buffer' period of one year – to avoid any contamination from past data. Structural data being yearly, they are the same in, say, March 1995 as in December 1995. In other words, yearly data (e.g. military spending) from March 1995 already incorporate some of the information that will only become available later. This means that without this buffer period (i.e. if predictions were made from 15 March 1995 to 15 March 1996), I would actually be using future information. In practice, this makes little difference for the results.

Table V. Sample predictions in the week preceding interstate wars, 1980–2001

	Base	*Structural*	*News*
Iraq, 1980-09-11	34.7	51.8	69.8
Iran, 1980-09-12	34.7	28.6	90
Syria, 1982-04-09	90	93	77.8
Chad, 1986-11-06	20.7	32.4	76.4
Lybia, 1986-11-07	20.7	33.3	80.3
China, 1986-12-25	89	99	100
Iraq, 1990-07-19	87	91	83.4
Kuwait, 1990-07-21	17.6	41.5	52.9
Yugoslavia, 1992-03-26	49.2	50.2	79.7
Armenia, 1993-01-28	14.4	7	32.4
Azerbaijan, 1993-01-28	14.4	15.4	37.4
Eritrea, 1998-04-23	5	23.5	10.5
Afghanistan, 2001-09-29	1.8	24.4	7.8
Average	36.9	45.5	61.4

Sample predictions for the week preceding interstate wars with at least 1,000 battle deaths over the period 1980–2001. Each cell is color-coded according to the estimated risk quantile in which this prediction falls (white = [0,33) quantile; light grey = [33,66); dark grey = [66,100]. For most cases, the estimated risk is more often correctly estimated to be higher by the model with news than by either the structural model or the baseline model. However, war came as a surprise in two cases (Eritrea and Afghanistan).

This procedure was repeated with 200 random cutoff points t.[15]

To gain some intuition for the results, I report in Table V predictions for the week prior to every interstate war over the 1980–2001 period (those for which I have complete data). Note that for all but two country-wars (Eritrea and Afghanistan), the predictions of the model with news indicate a higher risk than those of the structural model. The results are qualitatively similar for all types of war or if the time frame is extended to the entire 1920–2001 period.[16]

The predictive power of the resulting risk index can be evaluated more systematically along two main dimensions: first, its discriminating power – the ability to assign a higher probability to outcomes that occur than to those that do not. Second, the model's calibration shows its ability to assign subjective probabilities to outcomes that correspond to their objective probability – events with an estimated predicted probability of 20% should occur about 20% of the time. It is important to use both measures, as a model may have strong calibration but weak discrimination, or vice versa (see Steyerberg et al. [2010] for a review of the concepts of calibration and discrimination).[17] I review both measures in turn.

Improved binary predictions

There are two main ways to evaluate the quality – the discrimination power – of binary predictions $\hat{Y} \in \{0,1\}$. First, we might ask about the probability that a warning is issued given a forthcoming war, $P(\hat{Y}=1|Y=1)$, or the absence thereof, $P(\hat{Y}=1|Y=0)$. Answers to these questions can be addressed using the Receiver Operating Characteristic (ROC) curve. Conversely, the probability that a war occurs given a prior warning, $P(Y=1|\hat{Y}=1)$, is also of interest, and is addressed in the Precision-Recall curve.

Receiver operating characteristic curve. How often are warnings issued prior to wars, and how frequent are false warnings? Consider for example a threshold at the 99th percentile (i.e. an alarm is raised when $\hat{Y}$ is greater than this value). In this case, the true positive rate ($TPR = \frac{\text{NbTruePositives}}{\text{NbPositives}}$) for the model with news is 11.8% and the false positive rate ($FPR = \frac{\text{NbFalsePositives}}{\text{NbNegatives}}$) is 0.97%. This is a large improvement over the predictions made by the model without news (TPR = 6% only for the same FPR).

Clearly, the true positive rate and false positive rate depend on the cutoff chosen – the threshold above which a prediction $\hat{Y} \in \{0,1\}$ is labeled as a warning. I therefore calculated TPR and FPR for all possible values of the threshold, and plotted the results in an ROC curve (see Figure 2a). Visual inspection of the curve already reveals that the inclusion of the number of conflict-related news items in a model of conflict significantly improves binary predictions.

The improvement can be quantified by calculating the area under the ROC curve (AUC) (see Figure 2b) – a larger area implies a curve that is closer to the top left corner, and hence better binary predictions. In the case of interstate wars, for example, the area under the curve

[15] The cutoff sampling was done with replacement, though this has little effect on the results.

[16] A similar intuition can be gained from separation plots (Greenhill, Ward & Sacks, 2011), reported in the online appendix, Figure 2.

[17] For example, a model which estimates the risk to be 49% prior to all peace events and 51% prior to all war events has perfect discrimination but poor calibration. On the contrary, a model that assigns to all events a probability equal to the prevalence of the outcome (2% here) has perfect calibration but no discrimination.

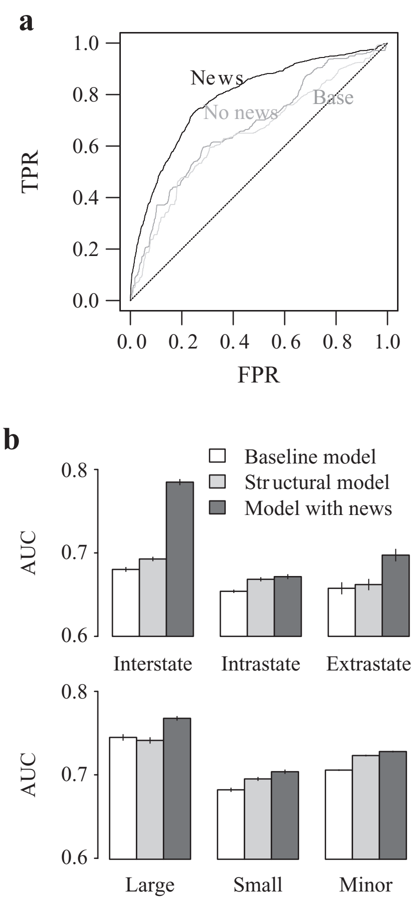

Figure 2. Receiver-operating characteristics (ROC) curves and area thereunder
a. Receiver operating characteristic (ROC) curves displaying the trade-off between the true positive rate ($\text{TPR} \equiv P(\hat{Y} = 1|Y = 1)$) and the false positive rate ($\text{FPR} \equiv P(\hat{Y} = 1|Y = 0)$) for interstate wars only (Sing et al., 2005). The curves show the trade-offs in the choice of a threshold for the detection of war, between the detection of a large proportion of wars (high TPR) and false warnings (high FPR). The best possible prediction method would yield a point in the upper left corner of the ROC space, representing no false negatives and no false positives. Curves that tend toward this ideal are better than those closer to the 45 degree line, which corresponds to random predictions.
b. Area under the curve (AUC) of the three models for different classes of war. The error bars represent the standard errors of the AUC, obtained by bootstrapping.

for the model with news is 0.785, as opposed to 0.693 for the structural model and 0.68 for the base model, which illustrates the considerable improvement offered by the inclusion of news in the model. This finding is not limited to interstate wars. The inclusion of conflict-related news significantly improves the predictions over a model with only structural variables for all types of wars, although the improvement is more modest for small conflicts and intrastate wars (see Figure 2b).[18]

Precision-recall curves. Because the ROC curve is conditioned on the actual occurrence of war ($P(\hat{Y} = 1|Y = 1)$), it has the advantage of being independent of the prevalence of conflict $P(Y = 1)$. Yet an almost perfect ROC curve can also be misleading, as it does not inform us about the perhaps more relevant question: how often are alarms actually followed by conflict, that is, $P(Y = 1|\hat{Y} = 1)$ – a statistic referred to as 'precision'?[19]

Figure 3a plots the typical precision-recall curve, often used in information retrieval (Manning & Schütze, 1999), which has been cited as an alternative to ROC curves for skewed data (Bockhorst & Craven, 2005). Here we see that for any recall level, the precision of the model with news is far higher than in the model without news. As in the ROC curve, this improvement can be quantified by calculating the area under the precision-recall curve. Figure 3b shows how much improvement is gained by the inclusion of news in the model. Again, we find a dramatic improvement for interstate wars, and a much more modest improvement for intrastate wars.

[18] The reader might wonder about the discrepancy between the AUC for interstate wars reported here and those obtained by, for example, Beck, King & Zeng (2004). The different results are explained first by different dependent variables (dyad-wars vs. country-wars), and by differences in the coarseness of our respective time series (yearly for Beck, King & Zeng [2004], weekly here). I show in ongoing work that adding information about conflict-related news items to the model in Beck, King & Zeng (2004) also leads to significant improvements in the predictions.

[19] Obviously, the two probabilities are related by Bayes' theorem, but it is nevertheless informative to report this result – often called 'precision' – in particular in the context of rare events, as it informs us about the reliability of a prediction. To see this, consider the following simple table:

		Y = 0	Y = 1
$\hat{Y}$	0	95	0
	1	4	1

Here wars are always correctly anticipated when they occur (i.e. $P(\hat{Y} = 1|Y = 1) = 1$), and false warnings are rare ($P(\hat{Y} = 1|Y = 0) = 4/99 \approx 0.04$). But these results can be misleading, as it is also the case that a warning is actually followed by a war in only 20% of cases ($P(Y = 1|\hat{Y} = 1) = 0.2\%$). It is this 'precision' that I now report.

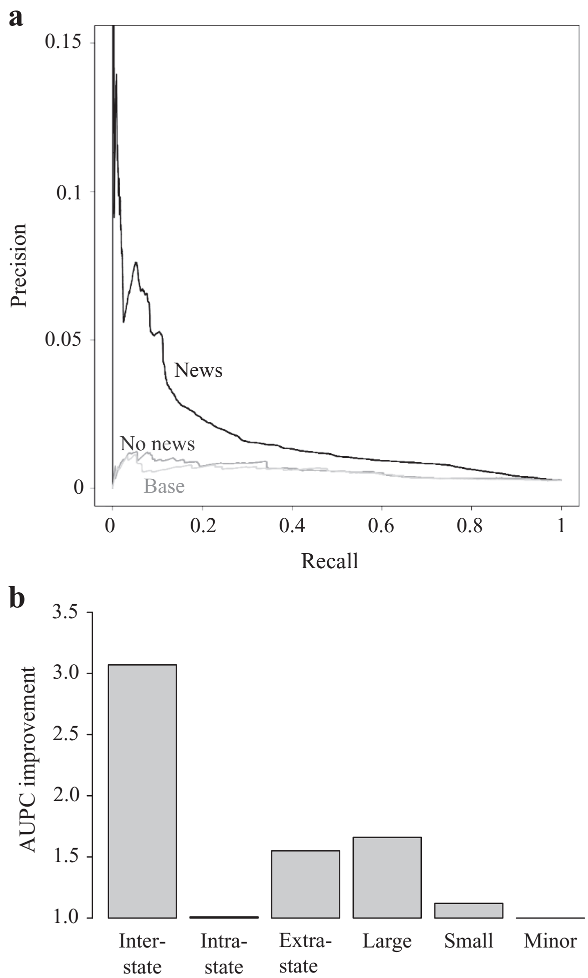

Figure 3. Precision-recall curves and area thereunder
a. Precision-recall (PR) curve for the onset of interstate wars ($> 1,000$ battle deaths) within three months. 'Recall' refers to the probability that a given war is correctly predicted ($P(\hat{Y} = 1 | Y = 1)$), while precision is the probability that a war actually happens given that a positive warning was issued ($P(Y = 1 | \hat{Y} = 1)$).
b. For each type of war and model, the area under the precision-recall curve (AUPC) was computed and compared across models. We measure the improvement between the structural model and the model with news as $\frac{AUPC_{news}}{AUPC_{struct.}}$.

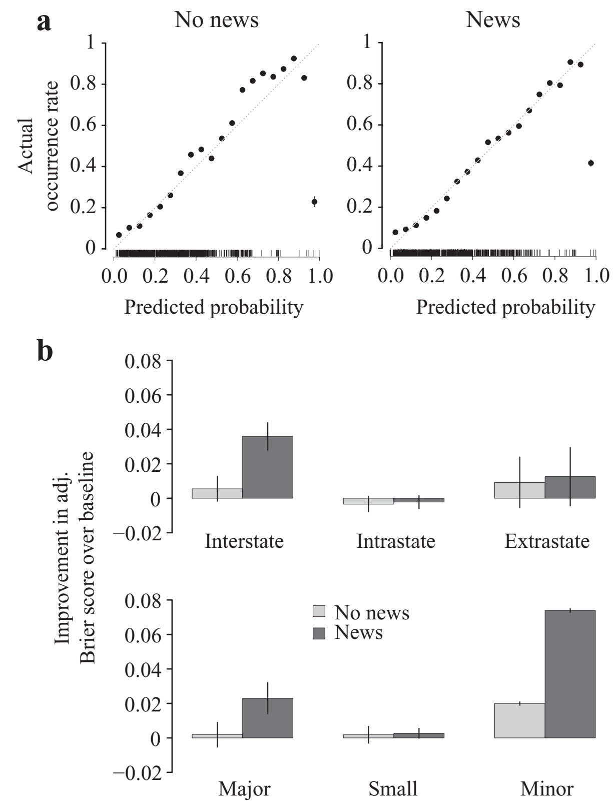

Figure 4. Calibration plots of the risk index against actual occurrence rate of conflict, 1920–2001 and adjusted Brier scores
a. The plot is a visual estimate of how closely the predicted probabilities agree with the actual outcomes (all wars within one year). In the ideal case, conflict should happen in x% of cases when the prediction is x (i.e. $P(Y = 1 | \hat{Y} = x) = x$). A perfect model would make predictions along the 45 degree line. (Note that the drop in actual occurrence rate when the predicted probability is close to 1 is caused by the low number of observations in that bin. A coarser binning removes this effect.)
b. Adjusted Brier score – a measure of calibration – for different types of war. The plot displays the improvement in the adjusted Brier score, compared to the baseline model (improvement$= adj.BS_{base}/adj.BS_{model} - 1$).

Improved probabilistic predictions

Policymakers are not only interested in binary predictions – 'Will war occur?' – but also estimates of the probability of an event – 'What is the probability of a war onset within the next few months?'. I now show that, in addition to improving binary predictions, the tension-based index also improves estimates of the *probability* of war. I calculated for given ranges of predictions (i.e. $\hat{Y} \in [0, 0.05)$, $\hat{Y} \in [0.05, 0.1)$, etc.) the prevalence of conflict in the next three months. That is, for all cases in which the probability of conflict was estimated between, say, 20% and 25%, did war actually occur in 20% to 25% of these cases?

Calibration can first be assessed visually – using a calibration plot as in Figure 4a, which shows the agreement between the estimated risk derived from the model and the actual risk of war. Better calibrations are those that (i) follow the 45-degree line and (ii) predict a larger range of values. I note that the occurrence of a conflict within three months can be forecasted with up to 85% confidence, meaning that wars occurred within three months of about 85% of cases in which an 85% risk was predicted. Conversely, peace tended to prevail when the index forecasted a low risk of conflict. Visually, we also see the better fit of the model with conflict-related news than the structural model, which departs more from the 45 degree line.

The improvement in calibration as compared to the structural model can be quantified using the Brier Score (BS) – a metric often used in meteorology (Brier, 1950; see Wilks, 2011: 331 for a good introduction to the Brier score). For each of N predictions, the Brier score measures the mean squared difference between the predicted probability $\hat{Y}_n \in [0,1]$ and the actual outcome $Y_n \in \{1,0\}$ – the occurrence or not of a conflict within the next three months. A lower Brier score indicates better calibration of predictions

$$BS = \frac{1}{N}\sum_{n \in N}(\hat{Y}_n - Y_n)^2.$$

Unfortunately, the Brier score is ill suited for the estimation of calibration in the case of rare events. Indeed an excellent (i.e. low) Brier score can be obtained by simply predicting probabilities equal to the overall probability of conflict in the population. For example, predicting $\hat{Y} = 0$ is likely to yield a very good (low) Brier score, since most cases are indeed 0, whereas a prediction that detects rare events but also issues a few false positives will be punished despite its ability to find needles in the haystack.

To address this problem, I therefore propose an alternative score in which the punishment received for a mistake is weighed by the overall probability of a given outcome:

$$BS_{adj.} = \frac{1}{N}\sum_{n \in N}(\hat{Y}_n - Y_n)^2 \times (1 - P(Y_n = y_n)),$$

where $y_n \in \{0,1\}$. Consider for example a sample size N = 1,000 with 999 negatives and 1 positive. Then for $\hat{Y}_n = 0$ when $Y_n = 1$, the increase (i.e. loss) in the adjusted Brier score is 0.999.[20] If, however, $\hat{Y}_n = 1$ when $Y_n = 0$, the loss is 0.001 – a smaller loss reflecting the idea that missing the rare event should be punished more severely than deviating from the almost certain prediction value.[21] Figure 4b plots the relative improvement in the adjusted Brier score by type and magnitude of war.

Increased warning time

Finally, forecasting wars is most useful if it can give policymakers sufficient warning time to react, and ideally avert the disaster. More precisely, I consider successful a warning that occurs at any time within a period $[t - 3mo - \delta, t - \delta]$, where t denotes the onset of war, and δ denotes a warning time. So far I have taken δ to be one week (i.e. a warning was deemed correct if it occurred within three months of the conflict onset, but no later than one week prior to war). How do results change for larger values of δ?

The area under the curve was computed for increasingly larger values of δ and the present risk index was found to significantly outperform other models even with large warning times (see Figure 5). In other words, the additional information offered by a count of conflict-related news items extends well before the onset of war – more than one year earlier in the case of interstate wars, and about six months earlier for intrastate wars – giving policymakers significant additional warning time.

Conclusion

The prediction of wars has received relatively little attention in the literature, in sharp contrast to fields such as finance and geology. One important difference is the availability of data: whereas financial data are readily available in fine-grained time series, information about military spending, diplomatic agreements and other international events is far more difficult to collect, harmonize and analyze.

In this context, three main contributions were made. First, I collected a new dataset on the weekly occurrence by country of certain conflict-related terms. While imperfect – the list is ad hoc and is a simple count, not an in-depth analysis of the content of each article – these data offer some rare advantages: their frequency (weekly) is superior to most existing data; their time span is also long, going back to the beginning of the 20th century; and they are largely independent of issues of harmonization, reliability or manipulation.

A second contribution is to show that these data are a strong predictor of conflict. The number of conflict-related news items increases dramatically prior to conflicts, and therefore we can conjecture that contemporaries do witness and notice the rise of tensions. Wars rarely emerge out of nowhere, though more research will be needed on the interesting cases in which journalists failed to pick up relevant clues, and hence where war came as a surprise.

Finally, I showed the ability of the measure of tensions based on conflict-related news to function as a

[20] $1^2 \times (1 - P(Y_n = 1)) = 1 \times (1 - 0.001) = 0.999$.

[21] $1^2 \times (1 - P(Y_n = 0)) = 1 \times (1 - 0.999) = 0.001$.

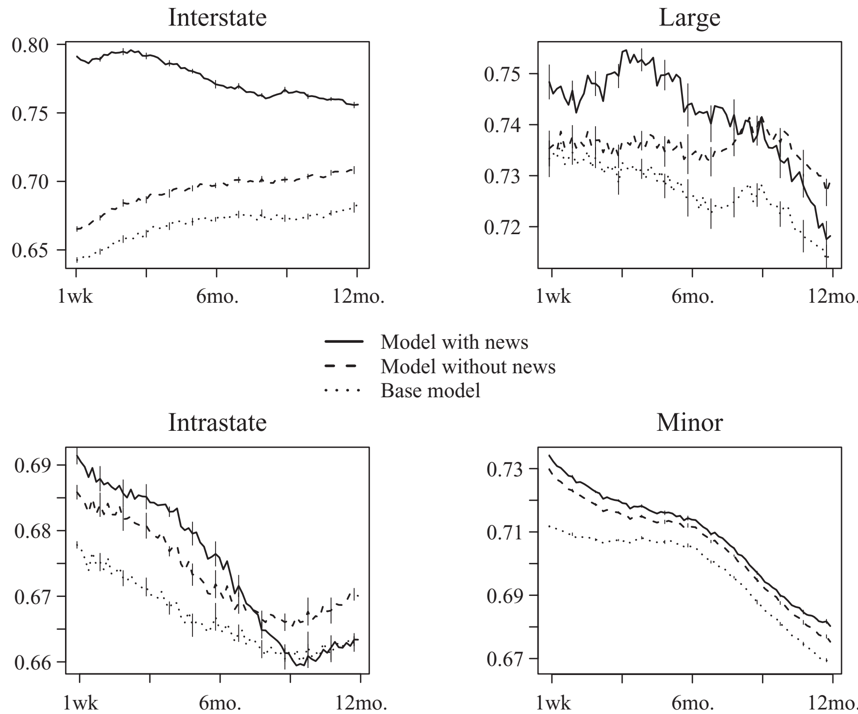

Figure 5. Predictive power (AUC) of various models for different warning times
Predictive power of the three models as warning time increases, as measured by the Area Under the Curve. 'Interstate' ('intrastate') wars with at least 1,000 battle deaths; 'large' ('minor') wars are those with at least 10, 000 (less than 1, 000) battle deaths. The vertical bars represent the standard error of the AUC, obtained by bootstrap.

reliable early warning signal, using only information available at the time. In particular, it improves for every type of war (inter- or intrastate, large or small) the precision with which we can answer questions such as 'Will a war occur next year?', 'What is the probability of a war happening next year?' and 'Will a war happen in exactly one year from today?'.

However, it is important to emphasize that the aim here was not simply to maximize predictive power. More complex models may perform better in that regard, using expert opinions, game theory, or more elaborate statistical techniques (Obrien, 2010). Rather, the goal was mainly to assess the value-added of an index of conflict-related news over structural variables. More complex models would have improved the predictive value of both models, but would not have given us additional information about either the value of the index derived here, or the ability of contemporaries to predict wars. In turn, the validation of the present index is crucial for future research, as it establishes it as a reliable historical metric of international tensions, and hence shows that it can be used to study important questions that have eluded scholars for lack of sufficiently fine-grained and comprehensive historical data. For example, what are the *immediate* geopolitical effects of changes such as democratization or increases in military budgets? How do alliance networks affect the spreading of tensions over time and space? And do democracies just not fight each other, or do they simply keep their escalation short of war. In that sense, prediction here was merely a way to validate the index of tensions as a proxy for geopolitical tensions. I hope that scholars will use this index to re-examine in more depth some of the central questions of international relations.

Replication data

All analyses were conducted using R 2.15.3. The dataset, replication R files and an output log for the empirical analysis can be found at http://www.prio.no/jpr/datasets.

Acknowledgements

I wish to thank two anonymous reviewers and the editor at the *Journal of Peace Research* for their particularly thoughtful feedback. Bob Axelrod, Robert ten Brincke, Karsten Donnay, Hein Goemans, Dirk Helbing, Rob Hyndman, Michael Mäs, Jim Morrow, Amin Mazloumian, Jonson Porteux, Sebastian Schutte, and participants of the SOMS internal seminar also contributed valuable comments.

References

Azar, Edward E (1980) The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution* 24(1): 143–152.

Beck, Nathaniel; Jonathan N Katz & Richard Tucker (1998) Taking time seriously: Time-series–cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42(4): 1260–1288.

Beck, Nathaniel; Gary King & Langche Zeng (2000) Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* 94(1): 21–36.

Beck, Nathaniel; Gary King & Langche Zeng (2004) Theory and evidence in international conflict: A response to de Marchi, Gelpi, and Grynaviski. *American Political Science Review* 98(2): 379–389.

Berg, Joyce E; Forrest D Nelson & Thomas A Rietz (2008) Prediction market accuracy in the long run. *International Journal of Forecasting* 24(2): 285–300.

Bernauer, Thomas & Nils P Gleditsch (2012) New event data in conflict research. *International Interactions* 38(4): 375–381.

Bockhorst, Joseph & Mark Craven (2005) Markov networks for detecting overlapping elements in sequence data. In: *Advances in Neural Information Processing Systems 17* (http://books.nips.cc/nips17.html).

Bosler, Naomi & Gerald Schneider (2011) The oracle or the crowd? Experts versus the stock market in forecasting ceasefire success in the Levant. Paper presented at the annual meeting of the International Studies Association, Montreal, 16 March (http://citation.allacademic.com/meta/p501371_index.html).

Brier, Glenn W (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78(1): 1–3.

Bueno de Mesquita, Bruce (2002) *Predicting Politics*. Columbus, OH: Ohio State University Press.

Bueno de Mesquita, Bruce (2009) *The Predictioneers Game*. New York: Random House.

Carter, David B & Curtis S Signorino (2010) Back to the future: Modeling time dependence in binary data. *Political Analysis* 18(3): 271–292.

Chadefaux, Thomas (2011) Bargaining over power: When do shifts in power lead to war? *International Theory* 3(2): 228–253.

Choucri, Nazli (1974) Forecasting in international relations: Problems and prospects. *International Interactions* 1(2): 63–86.

Correlates of War Project (2008) State System Membership List, v2008.1 (http://correlatesofwar.org).

De Marchi, Scott; Christopher Gelpi & Jeffrey D Grynaviski (2004) Untangling neural nets. *American Political Science Review* 98(2): 371–378.

Deutsch, Karl (1957) Mass communications and the loss of freedom in national decision-making: A possible research approach to interstate conflicts. *Conflict Resolution* 1(2): 200–211.

Faten, Ghosn; Glenn Palmer & Stuart Bremer (2004) The MID3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science* 21(2): 133–154.

Fearon, James D & David D Laitin (2003) Ethnicity, insurgency, and civil war. *American Political Science Review* 97(1): 75–90.

Feder, Stanley A (2002) Forecasting for policy making in the post-Cold War period. *Annual Review of Political Science* 5(1): 111–125.

George, Alexander L (1956) Prediction of political action by means of propaganda analysis. *Public Opinion Quarterly* 20(1): 334–345.

Glaser, Charles L (2000) The causes and consequences of arms races. *Annual Review of Political Science* 3(1): 251–276.

Gleditsch, Kristian S & Michael D Ward (2011) Forecasting is difficult, especially the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research* 50(1): 17–31.

Goldstone, Jack A; Robert H Bates, David L Epstein, Ted R Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder & Mark Woodward (2010) A global model for forecasting political instability. *American Journal of Political Science* 54(1): 190–208.

Greenhill, Brian; Michael D Ward & Audrey Sacks (2011) The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science* 55(4): 991–1002.

Holsti, Ole R (1963) The value of international tension measurement. *Journal of Conflict Resolution* 7(3): 608–617.

Hunt, William B (1997) *Getting to War: Predicting International Conflict with Mass Media Indicators*. Ann Arbor, MI: University of Michigan Press.

Huth, Paul K (1998) *Standing Your Ground: Territorial Disputes and International Conflict*. Ann Arbor, MI: University of Michigan Press.

King, Gary & Will Lowe (2003) An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization* 57(3): 617–642.

Lebovic, James H (1998) Consider the source: Organizational bias in estimates of foreign military spending. *International Studies Quarterly* 42(1): 161–174.

Lebovic, James H (1999) Using military spending data: The complexity of simple inference. *Journal of Peace Research* 36(6): 681–697.

Leetaru, Kalev H (2011) Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday* 16(9) (http://firstmonday.org/ojs/index.php/fm/article/view/3663).

Manning, Christopher D & Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marshall, Monty G; Keith Jaggers & Ted R Gurr (2010) Polity IV Project: Political regime characteristics and transitions, 1800–2010. Version p4v2010. Center for Systemic Peace (http//:www.systemicpeace.org/polity/polity4.htm).

McClelland, Charles A (1984) *World Event/Interaction Survey (WEIS) Project, 1966–1978*. Inter-University Consortium for Political and Social Research.

Moul, William Brian (1988) Balances of power and the escalation to war of serious disputes among the European great powers, 1815–1939: Some evidence. *American Journal of Political Science* 32(2): 241–275.

Newcombe, Alan G; Nora S Newcombe & Gary D Landrus (1974) The development of an inter-nation tensiometer. *International Interactions* 1(1): 3–18.

Obrien, Sean P (2010) Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review* 12(1): 87–104.

Organski, Abramo FK & Jacek Kugler (1981) *The War Ledger*. Chicago, IL: University of Chicago Press.

Pevehouse, John C & Joshua S Goldstein (1999) Serbian compliance or defiance in Kosovo? Statistical analysis and real-time predictions. *Journal of Conflict Resolution* 43(4): 538–546.

Powell, Robert (2002) Bargaining theory and international conflict. *Annual Review of Political Science* 5(1): 1–30.

Powell, Robert (2004) The inefficient use of power: Costly conflict with complete information. *American Political Science Review* 98(2): 231–241.

Ray, James L (1998) Does democracy cause peace? *Annual Review of Political Science* 1(1): 27–46.

Sarkees, Meredith R & Frank Wayman (2010) *Resort to War, 1816–2007*. Beverly Hills, CA: Sage.

Schneider, Gerald; Nils P Gleditsch & Sabine C Carey (2010) Exploring the past, anticipating the future: A symposium. *International Studies Review* 12(1): 1–7.

Schrodt, Philip A & Deborah J Gerner (2000) Using cluster analysis to derive early warning indicators for political change in the Middle East, 1979–1996. *American Political Science Review* 94(4): 803–818.

Senese, Paul D & John A Vasquez (2010) *The Steps to War: An Empirical Study*. Princeton, NJ: Princeton University Press.

Sing, Tobias; Oliver Sander, Niko Beerenwinkel & Thomas Lengauer (2005) ROCR: Visualizing classifier performance in R. *Bioinformatics* 21(20): 3940–3941.

Singer, David J; Stuart Bremer & John Stuckey (1972) Capability distribution, uncertainty, and major power war, 1820–1965. In: Bruce Russett (ed.) *Peace, War, and Numbers*. Beverly Hills, CA: Sage, 19–48.

Siverson, Randolph M & Michael R Tennefoss (1984) Power, alliance, and the escalation of international conflict, 1815–1965. *American Political Science Review* 78(4): 1057–1069.

Steyerberg, Ewout W; Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina & Michael W Kat-tan (2010) Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 21(1): 128.

Ward, Michael D; Randolph M Siverson & Xun Cao (2007) Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science* 51(3): 583–601.

Ward, Michael D; Brian D Greenhill & Kristin M Bakke (2010) The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research* 47(4): 363–375.

Weidmann, Nils B & Michael D Ward (2010) Predicting conflict in space and time. *Journal of Conflict Resolution* 54(6): 883–901.

Wilks, Daniel S (2011) *Statistical Methods in the Atmospheric Sciences*. Oxford: Academic.

Wolfers, Justin & Eric Zitzewitz (2006) Prediction markets in theory and practice. NBER working paper no. 12083.

THOMAS CHADEFAUX, b. 1980, PhD in Political Science (University of Michigan, 2009); Visiting Assistant Professor, University of Rochester (2009–10); Postdoctoral Researcher, Swiss Federal Institute of Technology, Zurich (2010–).