International Studies Review

International Studies Review (2013) **0**, 1–18

ANALYTICAL ESSAYS: EVALUATION, SYNTHESIS, REFLECTIONS

Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction[1]

Michael D. Ward
Duke University

Nils W. Metternich
University of College London

Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle
Duke University

Developing political forecasting models not only increases the ability of political scientists to inform public policy decisions, but is also relevant for scientific advancement. This article argues for and demonstrates the utility of creating forecasting models for predicting political conflicts in a diverse range of country settings. Apart from the benefit of making actual predictions, we argue that predictive heuristics are one gold standard of model development in the field of conflict studies. As such, they shed light on an array of important components of the political science literature on conflict dynamics. We develop and present conflict predictions that have been highly accurate for past and subsequent events, exhibiting few false-negative and false-positive categorizations. Our predictions are made at the monthly level for 6-month periods into the future, taking into account the social–spatial context of each individual country. The model has a high degree of accuracy in reproducing historical data measured monthly over the past 10 years and has approximately equal accuracy in making forecasts. Thus, forecasting in political science is increasingly accurate. At the same time, by providing a gold standard that separates model construction from model evaluation, we can defeat observational research designs and use true prediction as a way to evaluate theories. We suggest that progress in the modeling of

[1]This project was partially supported by the Defense Advanced Research Projects Agency and subsequently the Office of Naval Research (holding grants to the Lockheed Martin Corporation, Contracts FA8650-07-C-7749 and N00014-12-C-0066, respectively). For helpful insights, we thank Philippe Loustaunau, Mark Hoffman, Scott de Marchi, and Will Moore. All the bad ideas and mistakes are our own. An earlier version was presented at the 2nd International Conference on Cross-Cultural Decision Making.

Ward, Michael D., Nils W. Metternich, Cassy L. Dorff, Max Gallop, Florian M. Hollenbach, Anna Schultz, and Simon Weschle. (2013) Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction. *International Studies Review*, doi: 10.1111/misr.12072

conflict research depends on the use of prediction as a gold standard of heuristic evaluation.

Political events are frequently framed as unpredictable. Who could have predicted the Arab Spring, 9/11, or the end of the cold war? This skepticism about prediction reflects an underlying desire to forecast. Predicting political events is difficult because they result from complex social processes. However, in recent years, our capacity to collect information on social behavior and our ability to process large data have increased to degrees only foreseen in science fiction. This new ability to analyze and predict behavior confronts a demand for better political forecasts that may serve to inform and even help to structure effective policies in a world in which prediction in everyday life has become commonplace.

Only a decade ago, scholars interested in civil wars undertook their research with constrained resources, limited data, and statistical estimation capabilities that seem underdeveloped by current standards. Still, major advances did result from these efforts. Consider "Ethnicity, Insurgency and Civil War" by Fearon and Laitin (2003), one of the most venerated and cited articles about the onset of civil wars. Published in 2003, it has over 3,000 citations in scholar.google.com and almost 900 citations in the Web of Science (as of April 2013). It has been cited prominently in virtually every social science discipline in journals ranging from *Acta Sociologica* to *World Politics*; and it is *the* most downloaded article from the *American Political Science Review*.[2] This article is rightly regarded as an important, foundational piece of scholarship. However, in the summer of 2012, it was used by Jacqueline Stevens in a *New York Times* Op-Ed as evidence that political scientists are bad forecasters. That claim was wildly off the mark in that Fearon and Laitin do not focus on forecasting, and Stevens ignored other, actual forecasting efforts in political science. Stevens' point—which was taken up by the US Congress—was that government funding on quantitative approaches was being wasted on efforts that did not provide accurate policy advice. In contrast to Stevens, we argue that conflict research in political science can be substantially improved by more, not less, attention to predictions through quantitative approaches.

We argue that the increasing availability of disaggregated data and advanced estimation techniques are making forecasts of conflict more accurate and precise, thereby helping to evaluate the utility of different models and winnow the good from the bad. Forecasting also helps to prevent overfitting and reduces confirmation bias. As such, forecasting efforts can be used to help validate models, to gain greater confidence in the resulting estimates, and to ultimately present robust models that may allow us to improve the interaction with decision makers seeking greater clarity about the implications of potential actions.

First, we highlight that forecasting has been on the mind of many conflict researchers, but has never received the same attention as theoretical, descriptive, or explanatory contributions. Then to highlight the advances made in the last few years, we reassess the models Fearon and Laitin propose in regard to their forecasting ability. This serves as a benchmark to evaluate the progress made in the last decade.[3] Last, we describe a newer effort to model crises that is specifically aimed at prediction, illustrating the potential of forecasting as a means of model validation.

[2]See http://www.apsanet.org/content_30489.cfm; April 18, 2013 shows that the full text of the article has been viewed almost 7,000 times.

[3]Yet at the same time, many models have the same broad characteristics of the Fearon and Laitin approach (panel data; linear probability models) without having become canonical.

What's Prologue Is Prologue

The founding book of the quantitative study of conflict is *The Statistics of Deadly Quarrels* by Lewis Frye Richardson (1960), which began the systematic collection of data on wars with the intent of explaining and predicting their occurrence, escalation, duration, termination, and spread. Richardson's analytic and systematic exploration of war was very much in the spirit of Karl W. Deutsch's subsequent introduction to Quincy Wright's *Study of War*, when he writes that "war, to be abolished, must be understood. To be understood, it must be studied." In *The Statistics of Deadly Quarrels*, Richardson lists more than 300 wars between 1820 and 1949, generating data on each of them. Interestingly, he does not restrict his collection to a particular type of war, but simply introduces a threshold for a violent conflict to be included. In fact, the "objective" definition of war is what makes Richardson's war list the first scientific collection of conflict, rather than the list of Wright (1942) based on legal documents, or Sorokin's (1937) list that selects on "greater" nations. Richardson only has a small part of his book devoted to prediction. Presumably this is because, just as he had doubts about forecasting the weather, he did not believe that we would ever have the necessary data nor computing power to predict conflicts. However, in the last 50 years, and especially the last 20, the advances in data collection and estimation are making a new generation of forecasting models possible.

Inspired by the work of Pitirim Sorokin, Lewis Frye Richardson, and Quincy Wright, J. David Singer founded the Correlates of War Project in 1963 at the University of Michigan with the goal of systematically accumulating scientific knowledge about war. The impact of this project—which continues under the current guidance of Zeev Maoz—on the discipline cannot be overstated. It formed the gold standard of quantitative conflict research projects in the twentieth century. The collaborative effort of Melvin Small and J. David Singer collecting information on interstate and extra-state war led to the publication of *The Wages of War* in 1972. It marks the beginning of the country-year format that has dominated conflict data sets and how the discipline thinks about conflict processes.

However, the forecasts that researchers and policymakers are interested in often address much lower levels of violence than the traditional 1,000 battle-death threshold. The need to analyze civil conflicts at lower escalation levels and understand subnational dynamics was driven by an increasing number of intrastate conflicts in the early 1990s that did not fit COW's typology of intrastate wars. While leading Correlates of War (COW) scholars identified this problem early on (Sarkees, Wayman, and Singer 2003), the data project itself was slow to adjust to a changing scientific landscape. Hence, conflict data in the beginning of the twenty-first century was suddenly dominated by Scandinavians, led by the Uppsala Conflict Data Program (UCDP) (Themnér and Wallensteen 2012) and the Peace Research Institute Oslo (PRIO).

While the country-year and conflict-year format pushed the discipline forward and allowed for a subfield that has gained great recognition within political science, forecasts should ideally be made on a daily, weekly, or monthly level, and not be restricted to yearly level. This is reflected in the increasing demand for temporally disaggregated event data (Cederman and Gleditsch 2009). Even though there has been a surge of event data in recent years, a number of early efforts were made in the seventies and eighties. Edward Azar's COPDAB (1980) and Charles McClelland's WEIS (McClelland and Hoggard 1969) can be seen as the front-runners in this trend, but many smaller and specialized efforts followed (for example, Leng's BCOW 1993). In fact, the NSF-funded Data Development in International Relations project (McGowan, Starr, Hower, Merritt, and Zinnes 1988) stated that event data had become the second most common form of data, behind the dominating non-event COW data sets.

The rise of event data in the early 1990s can be interpreted as a function of faster and cheaper computing power and dedicated scholarship. Automated event coding suddenly became more feasible, and the access to electronically available news sources increased exponentially. The Kansas Event Data System (KEDS)—now known as the Penn State Event Data Project—used automated coding of English language news reports to generate political event data focusing on the Middle East, Balkans, and West Africa. These data were used in statistical early warning models to predict political change. In addition, the proprietary VRA-Reader was developed to process the increasing volume of international news reports with more precision and extensibility than earlier sparse parsers. As part of these projects, political event coding taxonomies were developed to deal with actors, actions, and locations associated with individual events. The best known are CAMEO (Gerner, Schrodt, and Yilmaz 2009) and IDEA (Bond, Bond, Oh, Jenkins, and Taylor 2003).

However, the real success story of event data began to play out only recently. From the scientific side, there was a growing demand to analyze conflict processes on the micro-level. At the same time, government sponsors became interested in forecasts of political violence and other political events that demanded data disaggregated both temporally and geographically. In particular, the integration of geographical information systems and event data changed the way conflict data could be used and analyzed. An early leader in this effort was the ACLED project (Raleigh, Linke, Hegre, and Karlsen 2010), but by now most event data collections do disambiguation for geography. In fact, it led to an explosion of conflict-related data sets and an exciting and new way to test dynamic theories across time and space. The recently released GDELT data assembled by Kalev Leetaru is based on real-time web scraping combined with event ontologies based on the Schrodt approach known as *TABARI* (for a recent description, see Leetaru and Schrodt 2013).

In political science, prediction is typically conceptualized as a conditional exercise, in which values on a dependent variable are calculated based on some estimated, or conditional, statistical model and then compared with the actual observed values (Hildebrand, Laing, and Rosenthal 1976). But a recent trend makes political predictions about things that have not yet occurred, in the sense that the *Old Farmer's Almanac* predicts the weather for the coming year. In 1978, a volume edited by Nazli Choucri and Thomas Robinson (Choucri and Robinson 1978) provided an overview of the then current work in forecasting in international relations (IR), much of which was done in the context of policy-oriented research for the US government during the Vietnam War. There were a variety of attempts to forecast or evaluate forecasting efforts (Freeman and Job 1979; Singer and Wallace 1979; Vincent 1980), and a few efforts began to forecast internal conflict (Gurr and Lichbach 1986), but the median empirical article in political science (as well as sociology and economics) used predictions only in the sense of in-sample observational studies.[4] Doran (1999) and others provided some criticism, but most scholars avoided making predictions, perhaps because their models had enough difficulty in describing well what *had* happened.

Still there were a few scholars that continued to make predictions (yes, about the future), including Gurr and Harff (1996), Krause (1997), Davies and Gurr (1998), Pevehouse and Goldstein (1999), Schrodt and Gerner (2000), King and Zeng (2001), O'Brien (2002), Bueno de Mesquita (2002), Fearon and Laitin (2003), de Marchi, Gelpi, and Grynaviski (2004), Enders and Sandler (2005), Leblang and Satyanath (2006), Ward, Siverson and Cao (2007), Brandt,

[4]In the late 1990s, scholars of American electoral politics began making predictions of voting patterns in presidential elections (Campbell 1992).

Colaresi, and Freeman (2008), Bennett and Stam (2009), and Gleditsch and Ward (2010), among a few others.[5] However, just in the last few years, the field of conflict forecasting has expanded tremendously. The surge of prediction research in conflict and peace studies can be attributed to the new availability of spatio-temporal disaggregated data and the application of new estimation strategies. Both developments are a result of increasing computational power that allows access to large data sources and the implementation of complex statistical tools. Including our own, there are a number of political science projects that focus on conflict forecasting. For example, the Predictive Societal Indicators of Radicalism created by Murdie (2011) forecasts political violence levels 5 years into the future. A research group led by Hegre (2013) at the Peace Research Institute Oslo also forecasts the future of armed conflict. In addition, it aims at increasing our understanding of the causes of conflict by using simulations and predictions to forecast the onset and spread of conflict and democracy.

Stepping 10 Years Back

To demonstrate the huge step that has been taken by conflict researchers in the past years, we re-evaluate the foundational study by Fearon and Laitin (2003) to establish a comparison and highlight the advances in recent years. Although Fearon and Laitin's article is not about out-of-sample prediction, it uses in-sample prediction to tell its main story. In their article, they present the probability of civil war onsets over a 5-year period, conditional on ethnic homogeneity and GDP per capita, showing that the latter has a more profound effect than the former on the 5-year probability of civil war onset.[6] Given the prominence of this research, as well as the transparency of the research program it embodied, we use their framework to predict civil war onsets in the period from 2000 to 2009.[7] We illustrate how prediction can provide evidence about the validity, as well as utility, of the specific models that have been developed. We undertake a replication of Fearon and Laitin's Model (1) for the year 1999,[8] the last year in the Fearon and Laitin data set, and use the laws of probability to calculate the cumulative probability of civil wars in each country for each of the 10 subsequent years (2000–2009).[9] This approach is *cumulative*, because the probability mounts up incrementally in the absence of an onset. Table 1 shows the yearly, cumulative predicted probabilities of civil war onsets for the 10, highest-probability

[5]A summary of classified efforts was declassified and reported in Feder (1995, 2002). Especially, recent efforts include Hewitt, Wilkenfeld, and Gurr (2010) who present a yearly update of hot spots around the world; Goldstone, Bates, Epstein, Gurr, Lustik, Marshall, Ulfelder, and Woodward (2010), which contains a fairly recent global model from the US Government's Political Instability Task Force; Brandt, Freeman, and Schrodt (2011a) describing real-time efforts; and Hegre, Nygård, Strand, Urdal, and Karlsen (2013), which presents five decades of forecasts into the future. A nice overview of some of the historical efforts along with a description of current thinking about forecasting and decision support is given by O'Brien (2010), a former program manager at DARPA who has conceptualized and supported the ICEWS project under which we conducted some of the research reported in this manuscript.

[6]The presentation of this result can be found in Fearon and Laitin (2003) p. 83, Figure 2

[7]The replication data, plus myriad reanalyses, are available at http://www.stanford.edu/~jfearon/.

[8]Model (1) uses the following independent variables: prior war, lagged per capita income, lagged log of the population, log of the percentage of mountainous terrain, noncontiguous state, oil exporter, new state, lagged instability measure, lagged Polity IV score, ethnic fractionalization, religious fractionalization. The lags are one year.

[9]All of the analyses in this manuscript use *lme4* version 0.999999-0, and $\mathcal{R}$, version 2.15.1: platform x86_64-apple-darwin9.8.0; arch x86_64; os darwin9.8.0; system x86_64, darwin9.8.0; status; major 2; minor 15.1; year 2012; month 06; day 22; svn rev 59600; language R; version.string R version 2.15.1 (2012-06-22); nickname Roasted Marshmallows. Probabilities in year N are given by $P(\text{onset in year } N) = 1 - (1 - P(\text{onset in1999}))^{N-1999}$. If a war occurs in year *M*, between 1999 and year *N*, the exponent is set to (*N*–*M*). We had to exclude a number of countries since they did not have a predicted probability for 1999 due to missing data. This is the approach which results in the probability map found in Fearon and Laitin.

Table 1. Top 10 Cumulative Predicted Probabilities for Countries with at Least One Civil War Onset between 2000 and 2009. Probabilities in Years with Onsets Are **Bold**

	2000	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*
Nigeria	0.15	0.27	0.38	0.47	**0.54**	0.15	0.27	0.38	0.47	**0.54**
Angola	0.06	0.12	**0.17**	0.06	**0.12**	0.06	0.12	**0.17**	0.06	**0.12**
Azerbaijan	0.05	0.10	0.15	0.20	0.24	**0.28**	0.05	0.10	0.15	0.20
India	0.04	0.08	0.12	**0.16**	0.04	**0.08**	**0.04**	0.04	**0.08**	**0.04**
Niger	0.04	0.07	0.10	0.14	0.17	0.20	0.23	**0.25**	0.04	0.07
Dem. Rep. Congo	0.03	0.05	0.08	0.10	0.13	0.15	**0.17**	**0.03**	0.03	0.05
Georgia	0.02	0.05	0.07	0.09	**0.11**	0.02	0.05	0.07	**0.09**	0.02
Haiti	**0.02**	0.02	0.04	**0.07**	**0.02**	0.02	0.04	0.07	0.09	0.11
Burundi	0.02	0.04	0.07	0.09	0.11	0.13	0.15	0.17	**0.18**	0.02
Honduras	0.02	0.04	0.06	0.08	0.10	**0.12**	0.02	0.04	0.06	0.08

Table 2. Number of Correctly Predicted Onsets and False Positives for Cumulative Predicted Probabilities at Varying Probability Thresholds. Note That There Are 1,310 Out-of-Sample Cases, 40 of Which Were New Conflicts

Threshold	*Correctly Predicted Onsets*	*False Positives*
0.5	2/40	0/1270
0.3	2/40	29/1270
0.1	15/40	245/1270
0.05	24/40	495/1270

Table 3. Performance Statistics, By Year, for Cumulative Predicted Probabilities Approach for a Threshold of 0.1 (where necessary)

	2000	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*Overall*
Sensitivity	0.00	0.00	0.33	0.25	0.50	0.33	0.20	0.80	0.33	0.67	0.38
Specificity	0.99	0.95	0.90	0.87	0.83	0.78	0.72	0.71	0.67	0.63	0.81
Accuracy	0.97	0.94	0.89	0.85	0.82	0.76	0.70	0.71	0.66	0.63	0.79
Precision	0.00	0.00	0.07	0.06	0.12	0.07	0.03	0.10	0.02	0.04	0.06
Brier Score	0.02	0.02	0.02	0.03	0.04	0.05	0.04	0.04	0.04	0.03	0.03
AUC	0.59	0.43	0.72	0.56	0.74	0.62	0.55	0.75	0.71	0.68	0.65
Observed	3.00	2.00	3.00	4.00	6.00	6.00	5.00	5.00	3.00	3.00	40.00
Predicted	1.86	3.62	5.31	6.75	8.13	8.86	9.91	11.11	11.94	13.10	80.58

countries with an actual an onset.[10] Probabilities are in **bold** when an onset actually occurred. Since the probabilities increase if onset does not occur, there are also a large number of countries that have high cumulative predicted probabilities, but in which no onset occurred between 2000 and 2009. Examples include Kenya, Romania, Tajikistan, Tanzania, and Guatemala, which each have predicted cumulative probabilities at least as high as those reported in Table 1.

Table 2 gives the number of correctly predicted civil war onsets as well as the number of false positives, taking different thresholds for the cumulative predicted probability above which the model is considered to indicate a civil war onset. Using a probability threshold of 0.5, the model predicts only two out of the 40 onsets, but has no false positives. As the threshold is lowered, the number

[10]The onset and prior war variables were taken from UCDP's intrastate war database (Themnér and Wallensteen 2012).

of correct predictions goes up, but so does the number of false positives. When using 0.1 as the cutoff, 15 onsets are correctly predicted, but at the same time the model forecasts 245 onsets that did not happen. Despite the fact that many of the variables used in the Fearon and Laitin's study are statistically significant, the predictive accuracy of the model out-of-sample is not high.

Table 3 gives common performance statistics for each year, using a threshold value of 0.1 where necessary. The proportion of correctly identified both positives and negatives (sensitivity and specificity) is about 60%, but this percentage varies greatly by year. Accuracy measures the proportion of observations that are correctly classified (true positives and true negatives), which is again about 60%. Precision is the proportion of true positives out of all predicted positives. This is approximately 5%, meaning that if the model predicts an onset, it is wrong 95% of the time. Another single-number summary to evaluate predictions is the Brier score (Brier 1950), defined as the average squared deviation of the predicted probability from the true event. The Brier score is one of the few strictly proper scoring rules for predictions with binary outcomes (Gneiting and Raftery 2007). Brier scores closer to zero indicate better predictive performance. This model's Brier scores are low, because the model predicts *peace* for many countries that indeed do not have an onset of civil war. But predicting the non-onset of civil war is not the main point of the endeavor. The low AUC scores show that the model is not very accurate at predicting *conflict onsets*. The last row in Table 3 provides an estimate of the cumulative number of global onsets predicted by the model in each year (20 estimates of 0.05 should produce one onset). The model vastly overpredicts civil war onsets.

What can we learn from this? The most problematic aspect of the Fearon and Laitin model is that many country-years with a high predicted probability of conflict actually have no onset. It does well at predicting *peace*, but faces challenges predicting *conflict*. Consider the following medical analogy: A test has been developed to detect some disease that about 35 out of 1,000 people in the population have. If everybody is tested, for the 35 who have the disease, the test will tell 16 of them that they are infected, while the other 19 are told they are healthy. Out of the 965 without the disease, 193 are diagnosed and receive treatment. This means that only about 8% of the treated are actually sick and 55% of the sick are left untreated. These are the performance statistics of the Fearon and Laitin model.[11] Thus, a model with a number of statistically significant variables may nevertheless be poorly equipped to tell us what we (and policymakers) are really interested in: When and where do civil wars occur? The fact that, for example, the coefficient for oil exporter is positive and significant in a regression does *not* necessarily mean that focusing on this variable (among others) will equip us better to explain why civil war broke out (or will break out) in a certain country, but not in another. The poor predictive performance is not an indictment of Fearon and Laitin's contribution, nor is it evidence that prediction is too treacherous to attempt. Rather, it points to an opening for social scientists and to the benefits of embracing prediction as a concept. First, it establishes a framework for rigorous and ongoing cross-validation of our models. This cross-validation offers us the opportunity to test our theories, their scope, and their portability, which can provide valuable input in the theory-building process. Finally, generating predictions makes the implications of our research more accessible to the policy community and the general public. Specifically, it underscores the opportunity for developing better models of civil war onsets.

Although cross-validation techniques are well known in political science, they are not frequently used. Cross-validation is useful to minimize overfitting and maximize predictive power. The prediction framework incorporates two types of

[11]Using the out-of-sample approach with a threshold of 0.1.

cross-validation. Initially, it subsets the data into a training set and a test set. Coefficients for independent variables are estimated on the training set and then along with available data are used to predict outcomes in the test data. This provides an opportunity to evaluate whether the model and the data on which it is based are convolved in such a way that the model is only valid for the given data, or whether the model can provide information about the patterns in data that were isolated from the model development phase. Second, prediction invites the collection of new data and encourages the re-evaluation of the model. Though the data were partitioned for in-sample versus out-of-sample predictions, there is a certain risk of what one might want to call "second-order overfitting," in which models are optimized to make good out-of-sample predictions on a particular test set. This kind of overfitting is especially likely in models with a small number of cases, a large number of variables, and short time series. Hence, we believe model evaluation by out-of-sample prediction is particularly powerful if the out-of-sample data were not available to the researcher during the modeling phase. This is especially important in an arena wherein most analyses are done with observational, as opposed to experimental, data.

An oft-heard claim is that political science is interested in *explanation* rather than *prediction*. But given what we have shown above, we should be wary of saying that we can *explain* what causes civil war, just because we have a number of statistically significant variables. Assessing the quality of statistical models using prediction can tell us how well we understand social phenomena and help us improve our explanations of where and when they occur. Building theory-driven forecast models, with the concomitant continual availability of new out-of-sample data for testing, can lead social science research toward better theory and explanations, not just better predictions. In short, good explanations based on good social science theory should be able to generate accurate predictions, even if they will be probabilistic.

Prediction also creates particular incentives that may be more broadly useful. Because true out-of-sample prediction could involve a shift in context (for example, time, location, level), it requires integrated theories to be successful. When conducting all analyses *ex post*, the researcher will choose the theory she thinks is most likely to apply, often focusing on the novel or unusual. For example, parts of the Arab Spring were organized by social media, and hence, many commentators focused on this novel tool as an explanation for the success of the revolution. However, there were a host of other factors that made revolutions in the Arab world likely. High unemployment, low growth rates, aging dictators, and religious divisions are long-standing explanations for popular uprisings that were all present in this context. When creating predictions, it becomes much harder to inadvertently select advantageous models for a particular context. The choices between competing models have to be "endogenized" in order to have a durable and portable tool for prediction. This requirement—that we delimit the context in which our theory applies—makes our understanding and use of theory more precise. Making models portable across time also brings our focus to what societies have in common, leading us toward a more systematic understanding of social processes.

Developing theoretically motivated, cross-validated models can make our findings more accessible to a wider audience. Sharing the knowledge generated by our research is an important part of the enterprise. In addition, the practice of generating these predictions, for researchers, gives us a reminder of and an answer to the often embarrassing question, "so what?" Predictions bind our independent variables to outcomes in a concrete way, bringing clarity, both to ourselves and to others, as to the mechanisms in our models. Current events make predictions of civil conflict even more desirable. The Arab Spring and its aftermath, continuing violence in Afghanistan, and sudden agreements between

Georgia and Russia always raise the same question: Could we have predicted these events to prepare for or even influence their emergence? Our profession struggles to do so. There are, of course, many who are interested in generating a predictive, analytical social science in the policy realm. During the Vietnam War, Jeffrey S. Milstein, a young Ph.D. from Stanford University, conducted quantitative and simulation studies that were among the first ever, real predictions in the discipline of IR (Milstein and Mitchell 1968; Milstein 1974). These were aimed at elucidating the dynamics of the then ongoing conflict in South-East Asia. Robert McNamara, the US Secretary of Defense, was promoting a systems analytic perspective, which suggested that a war of attrition would allow the United States to outlast the Viet Cong. Milstein's analysis was the first (outside of the Department of Defense) to illustrate that the dynamics were likely to turn out differently. And in fact, Milstein, not McNamara, was correct. Since Milstein's early work, there have been a variety of efforts within the policy community to craft accurate predictive models that can inform, if not necessarily guide, policy.[12]

This article is a reminder that despite unique features in every crisis, we, as a discipline, should strive for the main prize: the identification of general mechanisms that allow us to make predictions about future events. In fact, the ability to predict future crises can be understood as the gold standard to scientifically advance the study of conflict, peace, and crises. The goal is to have something as theoretically sound as the Fearon and Laitin models that can *also* explain the data that have not yet been collected, namely the future.

Stepping into the Future: New Civil War Model

We develop and present a predictive model of conflict that incorporates behavioral and institutional variables to predict the escalation of conflict.[13] Our model for predicting conflict has an approach somewhat different than forecasting projects that produce highly aggregated forecasts. For example, Hegre et al. (2013) make country-year predictions 50 years into the future, based largely on annual, demographic dynamics inherent in the UN population projections. The Political Instability Task Force produces probabilities of conflict in countries for a 2-year period into the future (Goldstone et al. 2010; Hewitt et al. 2010) and similarly Murdie (2011) provides forecasts that are based on 5-year projection intervals. In a sense, these forecasting models can be seen as complementary to our approach, which is temporally disaggregated well below the annual level. In the future, it is likely that these aggregated predictions can be combined with our disaggregated predictions via a principled statistical aggregator (Montgomery, Hollenbach, and Ward 2012). However, the point of this article is not to conduct a horse race (Brandt, Freeman, and Schrodt 2011b), but rather to suggest that there is a principled way forward to predict conflicts with highly disaggregated data that capture the dynamics of civil conflict in a way that can be improved by the use of cross-validation and forecast evaluation. Thus, this article is about a future that is not dominated by data that have been collected for different purposes at an annual aggregation, but rather that is gathered explicitly to focus on the disaggregated ebb and flow of conflict within and across societies.

Another difference to existing models is our methodological approach. Commonly, scholars estimate logit regression models assuming that all countries have

[12]Many of these are in the public domain and include recent efforts. A recent report by the Army Environmental Policy Institute (2011) lists no fewer that twelve ongoing projects that touch on some aspect of forecasting in the environmental realm (somewhat loosely conceptualized).

[13]More information about our modeling approach can be found at mdwardlab.com which describes the Crisis Prediction (CRISP) modeling activities and provides links to a variety of our forecasting activities.

the same baseline risk and are affected similarly by a set of covariates. In contrast, consonant with modern statistical theory, we estimate hierarchical models, a type of mixed effects model. These have the ability to provide a general framework for understanding a phenomenon, without requiring that the coefficients be exactly the same for each and every case. This makes sense in a world in which there are lots of unmodeled aspects and in which there is substantial heterogeneity to the objects studied. These kinds of models have facets that operate with groupings, typically at different levels (for example, country), and they also keep track of the variation between the groupings. In addition, we are not using annual aggregates that lead to a country-year format, but rather focus on smaller temporal aggregations. This approach allows us (i) to learn about processes that may vary from one place or time to another; (ii) use all the data while compromising between within-group estimates that are highly uncertain because they are based on averages, and use the more precise individual estimates that plausibly ignore influences that occur at the level of a group; as well as (iii) keep track of the uncertainty and covariation across the different levels. As an example, it may well be that accumulated inequalities tend to be associated with rebellious onsets in a fairly predictable way, but that this relationship is perhaps slightly different for autocracies than it is in democracies. One simple way to model this is with an interaction term, but that ignores the variation that may occur among the groupings (dictators may get a lot of foreign aid) and the individual effects within each country.[14] We could model civil war separately in each country and then average over all the coefficients we obtained. Hierarchical models are a compromise between this pooled approach and one in which there are groups of countries that are modeled together.

We briefly detail the substantive foundations of our modeling approach. Strikingly, most civil conflict predictions are based on structural factors such as political institutions and economic indices. Behavior is largely absent, or rather, behavior is not explicitly modeled in many empirical models. We argue that political institutions matter both in how they condition the behavior of their constituents and in the institutional capacity to respond to constituent behavior. More precisely, we stress that whether a civil conflict escalates to civil war depends on how the government addresses citizen demands and grievances. If the government has institutional restraints that prevent accommodation or repression of civil unrest, violence is likely to escalate and turn into a civil war. This argument is similar in spirit to that presented in Bueno de Mesquita and Smith (2009), which portrays a game among the regime, its supporters, and those in opposition. As a result, we focus on including the events that occur among the government, its supporters, and its opponents.

At the same time, structural factors help to condition the impact of these events, and we focus on the regime characteristics of the government, along with the exclusive character of the state to garner information about the most salient of these factors. In addition, we examine the local, meso-scale, environment for evidence that violence occurs in neighboring locations. All of these kinds of variables are important, but the inclusion of behavior is especially salient. The addition of behavior to theories that predict conflict is important for two reasons. First, behavior is implicit in the institutional and structural variables in typical conflict prediction models. Second, theories of conflict are theories of behavior. However, structural and institutional variables do not always behave the way that we expect. The Laitin and Fearon Random Narratives project,[15] for example, conducts a series of analyses of individual false predictions their model makes.

[14]See Gelman and Hill (2007) and Pinheiro and Bates (2009) for a more complete statement of the benefits of this approach.

[15]Described in Laitin and Fearon (2009).

In the case of Burkina Faso, they conclude that the behavioral mechanisms they had proposed do take place, but conflict did not escalate. They take this as a validation of their theory, although the prediction was incorrect. We contend that the difference between discontent and civil war is substantially larger than the effort political scientists have dedicated to explaining it. While individual explorations of the behavioral elements of conflict are valuable for theory building, an explicit theoretical integration of behavior makes it possible to trace the causal channels for civil war—from discontent through escalation—in a systematic way.

In the most general case, we can model civil war (and other crisis events) using hierarchical models in which both the intercept and slope vary. Simply stated, this means that we group the data along an indicator, such as level of executive constraints, creating a different intercept for each group. Thus, the varying intercepts correspond to group indicators, and the varying slopes represent an interaction between predictor variables x and the group indicators:

$$\Pr(y_{it} = 1) = \text{logit}^{-1}(\alpha_j + \beta_j X_{it} + \gamma Z_{it})$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma\right)$$

where i denotes the countries, t the month, and j the grouping variable, α_j are the grouping variable's random intercepts. x_{it} is a predictor variable; β_j is the associated random coefficient; γ is a vector of fixed effects associated with Z_{it}.

In this article, we only utilize varying intercepts for each country. In addition to dealing with country-specific effects, we model civil conflict as a function of behavior as represented by lagged event data. These data are gleaned from natural language processing of a continuously updated harvest of news stories, primarily taken from Factiva, a proprietary repository of news stories from over 200 sources around the world. The baseline event coder is called JABARI, a java variant of TABARI (Text Analysis By Augmented Replacement Instructions), the former developed by Philip Schrodt and colleagues.[16] From this event data, we construct explanatory variables that indicate whether high-intensity events (for example, protests, fighting, killings) or low-intensity events (for example, demands or threats) are taking place between the government and opposition groups.[17]

These data are augmented with a variety of other attribute and network data. We use country attributes, coded on a monthly or yearly basis from the Polity (Democracy and Autocracy scores), World Bank (Child Mortality rate), and Excluded Population (Cederman, Wimmer and Min 2010) databases.[18] In addition, we use information about relations between the countries, including geography. We use this data to predict occurrences of civil conflict based on UCDP data. The results from our empirical model, measured at the monthly level from 1997 to 2011, are provided in Table 4.

Both high- and low-intensity conflictual events are associated with higher probabilities of civil war, but high-intensity events are about twice as powerful on average as those classified as low intensity. At the same time, civil wars are more likely

[16]See http://eventdata.psu.edu/.

[17]Quincy Wright's *Study of War* argued that if up-to-date and comprehensive predictions based on indices were available they, "should have a value for statesmen, similar to that of weather maps for farmers or of business indices for businessmen … Such indices could be used not only for studying the probability of war between particular pairs of states but also for ascertaining the changes in the general tension level within a state or throughout the world" (p. 1270ff).

[18]Polity began as an outgrowth of research by Gurr (1974), was institutionalized in efforts of several scholars (Marshall, Jaggers, and Gurr 2009), and is updated on a frequent basis in http://www.systemicpeace.org/polity/polity4.htm, as part of the PITF's ongoing activities. GDP data were taken from the World Bank (annual).

TABLE 4. Hierarchical Model Estimates for Conflict Occurrence in UCDP Data

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	Z
(Intercept)	−9.40	1.65	−5.69
High-intensity Conflictual Events	1.74	0.14	12.83
Low-intensity Conflictual Events	1.00	0.19	5.21
Excluded Population	7.62	1.20	6.33
Excluded Population^2	−9.12	1.50	−6.09
Log Child Mortality	0.52	0.36	1.43
Democracy	−0.28	0.05	−5.20
Autocracy	−0.11	0.06	−1.90
Spatial Low-Intensity Conflictual Events	0.93	0.43	2.16

(a) In-Sample Separation Plot

(b) Out-of-Sample Separation Plot

FIG. 1. Separation Plots for CRISP Model Prediction of UCDP Data. (a) In-Sample Separation Plot. (b) Out-of-Sample Separation Plot

to exist in societies in which large slices of the population are excluded from political access to the state. This relationship is non-linear, which is in line with previous findings in the literature. Also in line with existing research, richer countries (as represented by a lower rate of infant mortality) are less likely to have civil wars, and both very democratic and very autocratic countries have a reduced risk of civil war. Finally, conflicts seem to spread, in the sense that conflictual events in neighboring countries are positively associated with conflict occurrence.

Figure 1 illustrates the fit of the model both in- and out-of-sample using separation plots. These plots provide a summary of the fit for each model by demonstrating the range and degree of variation among the predicted probabilities and the degree to which predicted probabilities correspond to actual instances of the event. They are created by ordering all country-years according to their predicted probability of civil war onset, from lowest on the left to highest on the right. The black line through the center of the plot represents this probability. Countries where a civil war actually occurred are red, while those where no civil war occurred are white (Greenhill, Ward, and Sacks 2011). Red lines on the left show false negatives, while white events on the right show false positives. A good fit would be visualized with more red panels (events occurring) stacked at the right end of the plot. These plots show that: (i) the actual civil wars are among those cases with higher predicted probabilities, and (ii) as expected, the out-of-sample fit of the model is slightly worse than the in-sample fit. However, even out-of-sample, the model fits the data very well.

Table 5 reports the standard performance statistics for fit in binary models (assuming a classification cutoff of 0.5). Again, the out-of-sample performance is

Table 5. Performance Statistics, In-Sample and Out-of-Sample for Civil War Model

	In-Sample	*Out-of-Sample*
Sensitivity	0.84	0.70
Specificity	0.98	0.98
Accuracy	0.97	0.94
Precision	0.88	0.86
Brier Score	0.03	0.04
AUC	0.99	0.98
N. Civil Wars	2468.00	286.00
Predicted N. Civil Wars	2458.37	253.10

Table 6. Number of Correctly Predicted Conflicts and False Positives for CRISP Models on UCDP Data

Threshold	*Correctly Predicted*	*False Positives*
0.5	199/286	33/1781
0.3	232/286	61/1781
0.1	245/286	117/1781
0.05	261/286	163/1781

Table 7. Predicted Probability for Out-of-Sample Onsets in UCDP Data. With the Exception of Libya, These Are Considered to Be Minor Conflicts by UCDP, Rather than Civil Wars

	Start of Conflict	*Probability of CW*	*Prior Month*	*Next Month*
Senegal	December 2011	0.26	0.25	
Nigeria	March 2011	0.11	0.1	0.1
Syria	October 2011	0.00037	0.00038	0.00039
Libya	March 2011	0.00044	0.00043	0.00044

a bit worse than in-sample, on all of these single-number estimates, but the out-of-sample evaluations fare very well, except that they underpredict the actual number of civil wars. The in-sample performance is pretty much exactly on target, in that all the events occur in cases which have the highest predicted probabilities, and the countries with the lowest predicted probabilities do not exhibit any onsets of new civil wars.

Table 6 demonstrates that as the classification threshold is reduced, the proportion of correctly predicted civil wars goes *up*, but the number of false positives also increases.[19] Table 7 lists the four actual civil war onsets during the out-of-sample period, the month during which the conflict is deemed to have begun, and the predicted probability as given by the model for that month, as well as for the immediate prior and subsequent months. This suggests that the model in this particular out-sample period is not performing as well as suggested by the in-sample results and the statistical results.

[19]The difference between Tables 5 and 6 is explained by the fact that in Table 5 we are presenting the total number of wars that would be predicted from the sum of all the probabilities. Thus, 10 probabilities of 0.10 would result in a prediction of 1 civil war. Table 6 shows exactly how many are individually predicted to be above the given threshold: here 199 have predicted probabilities greater than 0.50.

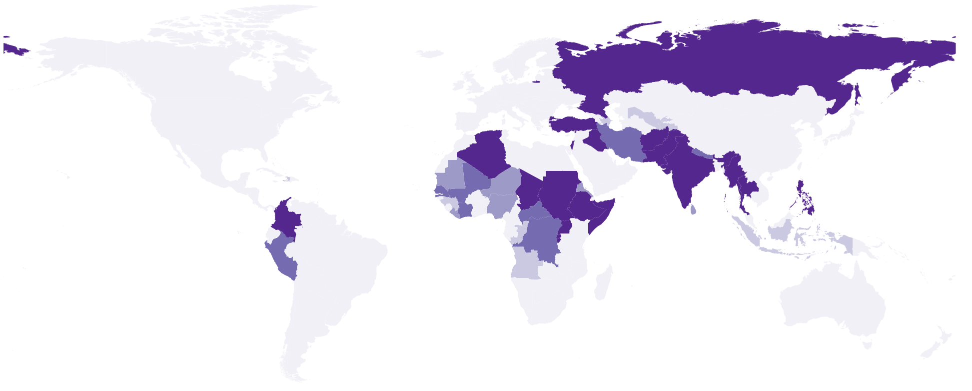

Fig. 2. Map for CRISP Model Out-of-Sample Predictions of UCDP Armed Conflicts (December 2011), with Darker Colors Presenting Higher Probabilities

Finally, in Figure 2, we provide a heat map illustrating where in the world we probabilistically expect to find civil wars in December 2011, the last out-of-sample period. This may seem unsurprising to some, but it is nonetheless developed by using theoretically based statistical models that have been annealed with cross-validation and are subject to actually being wrong. For that alone, it represents a step toward better understanding. Clearly we have a way to go with creating better models of civil war onset. Yet without actually undertaking those predictions, we might just congratulate ourselves about the statistical significance of our empirical efforts.[20]

Conclusion

In this article, we have demonstrated the utility of creating forecasting models for predicting political conflicts in a diverse range of country settings. We have shown that statistical models of civil war can be highly accurate, generating few false negative and positive predictions, both in- and out-of-sample. These models can shed light on an array of important components of the political science literature on conflict dynamics. Our model is far from perfect and misses some important events, such as the violence that transformed Libya. But it should be clear that one benefit of forecasting is that we actually know what we got wrong.

One frequent, and quickly surfaced, criticism of predictions in the social sciences is that social phenomena such as international crises are simply too complicated to predict by any means. Precisely because political conflicts are quite complicated, we should expand reasoning into mechanisms that can support the complications. Indeed, complex systems involve a wide variety of mechanisms and phenomena that are not easily described, let alone understood in isolation. A good example is meteorology, wherein we receive a variety of forecasts every day. These forecasts are typically generated by combining a large number of predictions that are based on meteorological models of weather, which in turn rely on the physics and chemistry of what is governing the various interacting systems. These systems each use a vast amount of measured data on the stocks and flows of various physical characteristics and allow for heterogeneity, so that predictions are not the same everywhere. Also, they permit an increasingly accurate scale of prediction.

[20]We avoid comparison against a simple model of lagged dependent variables because we do not believe that such a model can *in principle* be thought of as a theoretical, or explanatory model for onsets of rare events.

Indeed, the first attempt at weather prediction comes from Richardson (1922) over a century ago, when he used his mathematical approach to predict (retrospectively) the weather for May 20, 1910, by hand, using data from that date to predict the weather 6 hours subsequently. When corrected by modern smoothing techniques, Richardson's predictions were quite accurate, although he did not perceive them to be adequate at the time (Lynch 2006). To Richardson, a global model of weather forecasts would have taken tens of thousands of human calculators, which from his perspective seemed impossible. Most current global weather models are based on Richardson's equations.

Interestingly, Richardson turned away from weather predictions and wrote one of the first books on the statistical analysis of war: *Statistics of Deadly Quarrels*. After his experiences in the First World War, he thus focused on another complicated phenomenon to predict. Given advancements in theory, data collection, statistics, and computational power, we might be at an important point to push the boundaries of predicting political phenomenon beyond what we believed was possible only a few years ago. To preemptively declare defeat at the forecasting task seems foolish.

Finally, though some see the main benefit of prediction as creating a kind of social radar,[21] the real benefit of prediction may actually be as a heuristic allowing further probing of the empirical validity of specific models. Political science—especially where samples and experiments are not feasible—has an enormous vulnerability to over-reliance on the available data. In a statistical sense, this is often seen as overfitting: We use all the data to generate models that are dependent on all the data. That no longer seems like a very good research design. Being able to use our models to describe data we haven't seen before should be one gold standard criterion for model evaluation. In the face of a torrent of new data about the world, we can do this in almost real time. This permits the possibility of generating predictions about the future that may be useful beyond the validation of our theories.

The sabermetric revolution in sports portrayed in the movie *Moneyball* pitted geeky statistical modelers against wizened scouts. Some will see the same dynamics between those with "Big Data" and "Big Algorithms" and the subject matter experts with even more detailed knowledge of single cases. Tetlock has shown that some subject matter experts are especially bad forecasters (Tetlock 2005). This is the conclusion that was reached by Stevens, drawing on Tetlock's earlier efforts. Currently, Tetlock is involved in a large, government-funded project (The Good Judgement Project) to improve forecasting, in part by aggregating different forecasts. Some of these forecasts include statistical models that are quite sophisticated. Tetlock has shown that his forecasting group substantially outperforms internal government prediction markets (http://www.edge.org/conversation/how-to-win-at-forecasting). The same is true of some models, but without using prediction as a heuristic, we won't necessarily find this out. Even among those who are convinced the old scouts are the best scouts, there is renewed attention to how to evaluate political judgment such that it uses more data, keeps better track of mistakes, and integrates multiple estimates of outcomes.[22] Many others have begun researching ways to bring together better ideas in the generation of forecasts, better ideas that incorporate accountability, and multiple methodologies for making predictions (Bueno de Mesquita 2010; Tetlock 2010; Clauset and Woodard 2013).

[21]There is, among the policy community, a great optimism that these kinds of models will provide a unified way of observing as well as predicting social behavior; see Maybury (2010).

[22]See the recent articles published by the National Academy on this topic (Fischhoff and Chauvin 2011).

References

Army Environmental Policy Institute. (2011) *Environmental Change and Fragile States: Early Warning Needs, Opportunities, & Intervention.* Fort Belvoir, VA: AEPI Report.

Azar, Edward. (1980) The Conflict and Peace Data Bank (COPDAB) Project. *Journal of Conflict Resolution* 24: 143–152.

Bennett, D. Scott, and Allan C. Stam. (2009) Revisiting Predictions of War Duration. *Conflict Management and Peace Science* 26 (3): 256.

Bond, Doug, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles L. Taylor. (2003) Integrated Data for Events Analysis (IDEA): An Event Typology for Automated Events Data Development. *Journal of Peace Research* 40 (6): 733–745.

Brandt, Patrick T., Michael Colaresi, and John R. Freeman. (2008) The Dynamics of Reciprocity, Accountability, and Credibility. *Journal of Conflict Resolution* 52 (3): 343–374.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. (2011a) Real Time, Time Series Forecasting of Political Conflict. *Conflict Management and Peace Science* 28 (1): 41–64.

Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. (2011b) Racing Horses: Constructing and Evaluating Forecasts in Political Science. Presented at the Society for Political Methodology 28th Annual Summer Meeting. Princeton, NJ.

Bremer, Stuart A., and Thomas R. Cusack. (1995) *The Process of War: Advancing the Scientific Study of War.* Luxembourg and Philadelphia, PA: Gordon and Breach.

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. (2010) Judging Judgment. *Critical Review* 22 (4): 355–388.

Bueno de Mesquita, Bruce, and Alastair M. Smith. (2009) Political Survival and Endogenous Institutional Change. *Comparative Political Studies* 42 (2): 167–197.

Campbell, James E. (1992) Forecasting the Presidential Vote in the States. *American Journal of Political Science* 36 (2): 386–407.

Cederman, Lars-Erik, and Kristian S. Gleditsch. (2009) Introduction to Special Issue on "Disaggregating Civil War." *Journal of Conflict Resolution* 53 (4): 487–495.

Cederman, Lars-Erik, Andreas Wimmer, and Brian Min. (2010) Why Do Ethnic Groups Rebel? New Data and Analysis. *World Politics* 62 (1): 87–119.

Choucri, Nazli, and Thomas W. Robinson, Eds. (1978) *Forecasting in International Relations: Theory, Methods, Problems, Prospects.* San Francisco: W.H. Freeman.

Clauset, Aaron, and Ryan Woodard. (2013) Estimating the Historical and Future Probabilities of Large Terrorist Events. *Annals of Applied Statistics* (Forthcoming).

Davies, John L., and Ted R. Gurr. (1998) *Preventive Measures: Building Risk Assessment and Crisis Early Warning Systems.* Lanham, MD: Rowman & Littlefield Publishers.

de Marchi, Scott, Christopher Gelpi, and Jeffrey D. Grynaviski. (2004) Untangling Neural Nets. *American Political Science Review* 98 (2): 371–378.

Doran, Charles F. (1999) Why Forecasts Fail: The Limits and Potential of Forecasting in International Relations and Economics. *International Studies Review* 1 (2): 11–41.

Enders, Walter, and Todd Sandler. (2005) After 9/11: Is It All Different Now? *Journal of Conflict Resolution* 49 (2): 259–277.

Fearon, James D., and David D. Laitin. (2003) Ethnicity, Insurgency, and Civil War. *American Political Science Review* 97 (1): 75–90.

Feder, Stanley. (1995) Factions and Policon: New Ways to Analyze Politics. In *Inside CIA's Private World: Declassified Articles from the Agency's Internal Journal, 1955–1992*, edited by H. B. Westerfield. New Haven: Yale University Press.

Feder, Stanley A., Ed. (2002) Forecasting for Policy Making in the Post-Cold War Period. *Annual Review of Political Science* 5: 111–125.

Fischhoff, Baruch, and Cherie Chauvin, Eds. (2011) *Intelligence Analysis: Behavioral and Social Scientific Foundations.* Washington, DC: National Academies Press.

Freeman, John R., and Brian L. Job. (1979) Scientific Forecasts in International Relations: Problems of Definition and Epistemology. *International Studies Quarterly* 23 (1): 113–143.

Gelman, Andrew, and Jennifer Hill. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Analytical Methods for Social Research Cambridge, UK and NYC.

Gerner, Deborah J., Philip A. Schrodt, and Ömür Yilmaz. (2009) Conflict and Mediation Event Observations (CAMEO) Codebook. Available at: http://eventdata.psu.edu/data.dir/cameo.html.

Gleditsch, Kristian S., and Michael D. Ward. (2010) Contentious Issues and Forecasting Interstate Disputes. Presented to the 2010 Annual Meeting of the International Studies Association.

Gneiting, Tilmann, and Adrian E. Raftery. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102 (477): 359–378.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted R. Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. (2010) A Global Model for Forecasting Political Instability. *American Journal of Political Science* 54 (1): 190–208.

Greenhill, Brian D., Michael D. Ward, and Audrey Sacks. (2011) The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models. *American Journal of Political Science* 55 (4): 991–1003.

Gurr, Ted R. (1974) Persistence and Change in Political Systems, 1800–1971. *American Political Science Review* 74 (4): 1482–1504.

Gurr, Ted R., and Barbara Harff. (1996) *Early Warning of Communal Conflict and Humanitarian Crisis*. Tokyo: United Nations University Press, Monograph Series on Governance and Conflict Resolution.

Gurr, Ted R., and Mark I. Lichbach. (1986) Forecasting Internal Conflict: A Competitive Evaluation of Empirical Theories. *Comparative Political Studies* 19 (1): 3–38.

Hegre, Håvard, Håvard Nygård, Håvard Strand, Henrik Urdal, and Joakim Karlsen. (2013) Predicting Armed Conflict, 2011–2050. *International Studies Quarterly* (Forthcoming).

Hewitt, J. Joseph, Jonathan Wilkenfeld, and Ted R. Gurr. (2010) *Peace and Conflict 2010*. Boulder, CO: Paradigm Publishers.

Hildebrand, David K., James D. Laing, and Howard Rosenthal. (1976) Prediction Analysis in Political Research. *American Political Science Review* 70 (2): 509–535.

King, Gary, and Langche Zeng. (2001) Improving Forecasts of State Failure. *World Politics* 53 (4): 623–658.

Krause, George A. (1997) Voters, Information Heterogeneity, and the Dynamics of Aggregate Economic Expectations. *American Journal of Political Science* 41 (4): 1170–1200.

Laitin, David D., and James D. Fearon. (2009) Integrating Qualitative and Quantitative Methods. In *Oxford Handbook of Political Science*, edited by Robert E. Goodin. New York: Oxford University Press.

Leblang, David, and Shanker Satyanath. (2006) Institutions, Expectations, and Currency Crises. *International Organization* 60 (01): 245–262.

Leetaru, Kalev, and Philip A. Schrodt. (2013) GDELT Global Data on Events, Location, and Tone, 1979–2012. Presented at the Annual Meetings of the International Studies Association, April 2013, San Francisco, CA.

Leng, Russell J. (1993) *Interstate Crisis Behavior, 1816–1980: Realism versus Reciprocity*. Cambridge and New York: Cambridge University Press.

Lynch, Peter. (2006) *The Emergence of Numerical Weather Prediction: Richardson's Dream*. Cambridge: Cambridge University Press.

Marshall, Monty G., Keith Jaggers, and Ted R. Gurr. (2009) *Polity IV Project: Political Regime Characteristics and Transition 1800–2007*. College Park, MD: CIDCM, University of Maryland.

Maybury, Mark. (2010) Social Radar for Smart Power. Technical Papers 10-0745 Mitre Corporation Bedford, MA.

McClelland, Charles A., and Gary Hoggard. (1969) Conflict Patterns in the Interactions among Nations. In *International Politics and Foreign Policy*, edited by James N. Rosenau. New York: The Free Press.

McGowan, Pat, Harvey Starr, Gretchen Hower, Richard L. Merritt, and Dina A. Zinnes. (1988) *International Data as a National Resource*. Vol. 14 of International Interactions: Special Issue. New York: Taylor and Francis.

Milstein, Jeffery S. (1974) *Dynamics of the Vietnam War*. Columbus, OH, Ohio State University Press.

Milstein, Jeffrey S., and William C. Mitchell. (1968) Dynamics of the Vietnam Conflict. *Peace Research Society (International) Papers* 10 163–213.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. (2012) Improving Predictions Using Ensemble Bayesian Model Averaging. *Political Analysis* 20 (3): 271–291.

Murdie, Amanda. (2011) : A Risk Assessment Model of the Determinants of Political Violence. Technical report Council on Foreign Relations Washington, DC.

O'Brien, Sean P. (2002) Anticipating the Good, the Bad, and the Ugly: An Early Warning Approach to Conflict and Instability Analysis. *Journal of Conflict Resolution* 46 (6): 791–811.

O'Brien, Sean P. (2010) Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review* 12 (1): 87–104.

Pevehouse, Jon C., and Joshua S. Goldstein. (1999) Serbian Compliance or Defiance in Kosovo? Statistical Analysis and Real-Time Predictions. *Journal of Conflict Resolution* 43 (4): 538–546.

Pinheiro, J. C., and D. M. Bates. (2009) *Mixed-Effects Models in S and S-Plus*. Statistics and Computing. New York: Springer.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. (2010) Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. *Journal of Peace Research* 47 (5): 651–660.

Richardson, Lewis F. (1922) *Weather Prediction by Numerical Process*. Cambridge: Cambridge University Press.

Richardson, Lewis F. (1960) *Statistics of Deadly Quarrels*. Chicago and Pittsburgh, PA: Quadrangle/Boxwood.

Sarkees, Meredith R., Frank W. Wayman, and David J. Singer. (2003) Inter-State, Intra-State, and Extra-State Wars: A Comprehensive Look at Their Distribution Over Time, 1816–1997. *International Studies Quarterly* 47 (1): 47–70.

Schrodt, Philip A., and 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.

Singer, J. D., and Michael D. Wallace, Eds. (1979) *To Augur Well: Early Warning Indicators in World Politics*. Beverly Hills, CA: SAGE.

Sorokin, Pitirim A. (1937) *Social and Cultural Dynamics*. London: Owen.

Stevens, Jacqueline. (2012) Political Scientists Are Lousy Forecasters. *New York Times* June 23.

Tetlock, Philip. (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, Philip E. (2010) Second Thoughts about *Expert Political Judgment*: Reply to the Symposium. *Critical Review* 22 (4): 467–488.

Themnér, Lotta, and Peter Wallensteen. (2012) Armed Conflict, 1946–2011. *Journal of Peace Research* 49 (4): 565–575.

Vincent, Jack E. (1980) Scientific Prediction versus Crystal Ball Gazing: Can the Unknown Be Known? *International Studies Quarterly* 24 (3): 450–454.

Ward, Michael D., Randolph M. Siverson, and Xun Cao. (2007) Disputes, Democracies, and Dependencies: A Reexamination of the Kantian Peace. *American Journal of Political Science* 51 (3): 583–601.

World Bank. (Annual) World Development Indicators. URL: http://data.worldbank.org/data-catalog/world-development-indicators.

Wright, Quincy. (1942) *A Study of War*, Reprinted 1957 ed. Chicago: University of Chicago Press.