

Forecasting Civil Wars: Theory and Structure in an Age of “Big Data” and Machine Learning

Journal of Conflict Resolution

1-31

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0022002720918923

journals.sagepub.com/home/jcr

Robert A. Blair¹  and Nicholas Sambanis²

Abstract

Does theory contribute to forecasting accuracy? We use event data to show that a parsimonious model grounded in prominent theories of conflict escalation can forecast civil war onset with high accuracy and over shorter temporal windows than has generally been possible. Our forecasting model draws on “procedural” variables, building on insights from the contentious politics literature. We show that a procedural model outperforms more inductive, atheoretical alternatives and also outperforms models based on countries’ structural characteristics, which previously dominated models of civil war onset. We find that process can substitute for structure over short forecasting windows. We also find a more direct connection between theory and forecasting than is sometimes assumed, though we suggest that future researchers treat the value-added of theory for prediction not as an assumption but rather as a hypothesis to test.

Keywords

forecasting, civil wars, event data, machine learning

¹Department of Political Science, Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA

²Department of Political Science, University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author:

Robert A. Blair, Department of Political Science, Watson Institute for International and Public Affairs, Brown University, 111 Thayer St., Providence, RI 02912, USA.

Email: robert_blair@brown.edu

Conflict forecasters have long argued that good theory is essential for accurate prediction. Chiba and Gleditsch (2017, 295) argue that “theoretical attention to relevant actors” should improve predictive performance and advocate “anchoring prediction models in theories.” Brandt, Freeman, and Schrodtt (2011, 46) aim to generate “theoretically grounded, policy-relevant forecasts,” and Goldstone et al. (2010, 194) test a variety of potential correlates of conflict that are “drawn from the theoretical literature.” Cederman and Weidmann (2017, 476) conclude their review of the literature by arguing strongly for the necessity of theory for conflict prediction.

In recent years, however, claims about the importance of theory have yielded two related tensions in conflict forecasting research. First, some of the most prominent theories of conflict have been shown to perform remarkably poorly when used to predict conflict itself. Ward, Greenhill, and Bakke (2010) find that models based on the Fearon and Laitin (2003) theory of civil war produce surprisingly inaccurate forecasts; models based on Collier and Hoeffler (2004) do not fare much better. Hill and Jones (2014) show that existing theories of state repression generally cannot predict state repression itself. Blair, Blattman, and Hartman (2017) find that widely studied causes of local violence perform poorly when used for prediction. Some interpret this as evidence that purported causes of conflict may not be causal at all. Beck, King, and Zeng (2000, 21) go so far as to argue that causal theories that fail to forecast are of “dubious validity and marginal value.”

Second, while almost all conflict forecasting models are motivated to some extent by theoretical intuitions, many are operationalized using variables that do not correspond to those intuitions in a direct way. This has become especially true with the rise of “big” event data sets like the Computational Event Data System, the Global Database of Events, Language, and Tone, and, most prominently, the World-wide Integrated Crisis Early Warning System (ICEWS). ICEWS contains data on over 300 categories of events between actor dyads, ranging from “deny responsibility” to “grant diplomatic recognition” to “engage in ethnic cleansing.” This would seem to offer many opportunities for applying theory to prediction. Yet while many conflict forecasters have adopted ICEWS, most have done so by collapsing disparate event types into broad indices that distinguish conflictive events from cooperative ones and verbal events from material ones (e.g., Beger, Dorff, and Ward 2016; Chiba and Gleditsch 2017; Montgomery, Hollenbach, and Ward 2012; Schrodtt, Woodward, and Marshall 2011). As informative as these studies have been, such broad distinctions between conflict and cooperation are too generic to capture specific theoretical insights.

One way to resolve these tensions might be to combine the best of two distinct approaches to conflict forecasting. Early conflict forecasting models were inspired by “structural” theories of civil war that emphasize the role that regime type, per capita income, natural resource endowments, rough terrain, and other slow-moving or time-invariant characteristics play in increasing or decreasing the risk of civil war onset. But with few exceptions, these structural variables change too slowly to

predict when, exactly, civil war will begin or end. Some (though still relatively few) more recent models are inspired by “procedural” theories instead, which focus on the dynamics of dissent, mobilization, and repression that precede civil war onset. But most procedural models rely on measurement strategies that are too coarse to capture the dynamics of escalation that procedural theories describe.

We attempt to address this mismatch between theory and measurement by developing a parsimonious procedural model that more directly reflects the processes of escalation and de-escalation that precede and (potentially) predict civil war onset. We make three contributions. First, we show that a model informed by procedural theories of conflict escalation can predict the onset of civil war remarkably accurately, with an area under the receiver operating characteristic (ROC) curve (AUC, the most widely used metric for evaluating predictive performance in conflict forecasting) ranging from .82 to .85 in our base specification. Equally important, we show that our theoretically informed model outperforms more mechanical alternatives including a “kitchen-sink” model that comprises all events between government, opposition, and rebel groups in the ICEWS data set—over 1,000 predictors in all. Our results suggest that in the context of rare events driven by complex causal processes, models that rely on big data and “brute force” feature selection are likely to yield disappointing results (Cederman and Weidmann 2017, 476).

Second, we assess the value-added of structural characteristics in a model otherwise composed of procedural variables only. Other studies have found that structural characteristics can be used to identify countries that are “immune” from risk in the context of split-population models (Beger, Dorff, and Ward 2016) or that the onset, duration, and recurrence of civil war can be modeled interdependently, improving predictive accuracy (Chiba and Gleditsch 2017). Somewhat surprisingly, we find that structural characteristics diminish or only marginally improve the performance of models based on process alone. Most causal models attempt to identify the structural conditions (e.g., poverty or inequality) that generate tensions that may, in turn, escalate into civil war. While these conditions may be theoretically interesting and important, they seem not to help us predict when civil war will occur. For the latter purpose, we may need to focus on the processes that produce violence closer to the moment of onset. This does not imply that conflict forecasters should abandon theory altogether; to the contrary, it simply suggests that we should be selective in the theories we choose to apply. Our findings thus complement studies that integrate structure with process (Beger, Dorff, and Ward 2016; Tikuisis, Carment, and Samy 2013) but suggest that the latter may substitute for the former over short forecasting windows.

Finally, at the beginning of 2016, we used our theoretically informed model to generate a list of 30 countries with the highest risk of civil war in the first half of the year, then preregistered that list with the Evidence in Governance and Politics (EGAP) network.¹ “True” prospective forecasting of this sort remains rare in the study of conflict. Most conflict forecasting models are evaluated retrospectively through cross-validation or retrocasts of outcomes that have already been realized.

This leaves open the possibility of “fishing” for results—that is, increasing accuracy by overfitting through multiple unreported rounds of training and testing on the same data set. The few scholars who have published true prospective forecasts have generally not returned to evaluate them later (e.g., Hegre et al. 2013).² In this article, we revisit our predictions to assess their accuracy with the benefit of hindsight. To our knowledge, this is the first effort to preregister forecasts in a formal, publicly available scholarly venue.

Theoretical Framework

Why Do Civil Wars Occur?

Broadly speaking, studies of the causes of civil war can be grouped into two approaches: those focused on the structural characteristics that generate grievances or facilitate insurgency and those focused on the dynamic processes of repression, mobilization, and accommodation that arise between the state and its opponents. Studies favoring the structural approach tend to focus on the role that grievances play in catalyzing protest and rebellion (Moore 1966; Gurr 1970; Chenoweth and Ulfelder 2017) or, alternatively, on the conditions that make it easier to rebel in the first place (Collier and Hoeffler 2004), such as the “geography” of conflict (e.g., mountainous terrain where insurgents can hide).

The grievances that motivate nonviolent claims by opposition groups are at the core of classic models of conflict (Gurr 1970; Horowitz 1985). Recent contributions to the quantitative literature on civil war also emphasize the role of grievances in rebellion (Cederman, Wimmer, and Min 2010; Cederman, Weidmann, and Gleditsch 2011; Gurr 1993; Saxton 2005; Wimmer, Cederman, and Min 2009). Other studies using the structural approach emphasize the opportunity to rebel that tends to be greater in places where the state is unable to project power nationwide due to weak administrative capacity, rough terrain, or porous borders (Fearon and Laitin 2003). Uniting these studies is a focus on the slow-moving or time-invariant structural characteristics of countries that make them vulnerable to civil war.

Yet, structural characteristics tend to make poor predictors of the timing of civil war onset (Ward et al. 2013) and often cannot distinguish between violent and nonviolent conflict. Grievances alone are not sufficient to explain the escalation of nonviolent conflicts into violent ones. This echoes Tilly (1978), who claimed that discontent is too ubiquitous to explain the timing of collective violence. Tilly’s insight has informed research on contentious politics (McAdam, Tarrow, and Tilly 2001), which has explored the processes through which protest and rebellion occur. Small-scale conflicts can escalate into large-scale civil wars as a function of patterns of interactions between the state and claim-making groups that are not captured by structural variables (Sambanis and Milanovic 2015). An important contribution of the contentious politics literature has been to draw attention to these complex and often highly contingent processes.

Process-based research on conflict escalation has focused in particular on the role that state repression or accommodation of claim-making groups can play in dynamically altering the costs and benefits of rebellion as perceived by (potential) combatants. The mechanisms through which these changes occur remain contested, and the effects of repression on escalation are likely dependent on context and may differ in the short term versus long term. On the one hand, repression may radicalize the public, widening the pool of recruits willing to use violence against the state (Kitschelt 1985). Exposure to repression generates both affective and rational motives for violence. Repression causes anger, resentment, fear, and distrust (Petersen 2002; Young 2019), which may help mobilize groups to regain their lost social status. Repression could also make violent conflict more likely by demonstrating to opposition groups the futility of nonviolence (Tarrow 1989).

Repression may also enable escalation by reducing individuals' identification with the state and making ethno-regional identities more salient, which should make it easier for elites to mobilize opposition to the state and recruit rebels (Nair and Sambanis 2019). Repression that takes the form of political exclusion of ethnic groups has been shown to increase the risk of ethnic war onset (Cederman, Wimmer, and Min 2010). Recent research by Lindemann and Wimmer (2018) investigates the conditions under which groups with grievances are likely to rebel and finds evidence that state repression increases the likelihood of escalation (as do other factors including having access to sanctuaries beyond the control of the state). Nonviolent repression and restrictions of group rights can delegitimize the state and help mobilize support for the pursuit of autonomy, self-determination, or reform (Wimmer, Cederman, and Min 2009; Hechter 2001).

On the other hand, repression may convince citizens that the costs of resistance are too high, especially when the potential benefits are low (Shadmehr 2014). Thus, participation in violent protest could decrease if the threat of a state crackdown is high (Nair and Sambanis 2019), and claim-making groups could be appeased by state concessions of regional autonomy or political inclusion (Cederman et al. 2015). Such concessions and accommodations are generally believed to help prevent conflict escalation (Lichbach 1987)—an idea that has also received empirical support in the literature on ethnic civil wars (Cederman, Wimmer, and Min 2010; Cederman et al. 2015). However, the conflict resolution potential of concessions is also contingent on context. Concessions to groups challenging the state could fuel conflict if they are perceived as signs of state weakness (Walter 2006) or if they empower local elites to make ever-increasing claims (Brancati 2006).

These ambiguous and sometimes conflicting results on the effects of repression on conflict escalation are partly due to lack of attention to the strategic nature of escalation processes (Pierskalla 2010) but are also due in part to the multiple interactions and contingencies that determine the net effect of repression on civil war onset. While a few studies have highlighted some important interactions, many remain unexplored since the state's decision to repress some groups and accommodate others is multifaceted as are the determinants of groups' reactions to state

policy. While capturing these dynamics empirically is a challenge, we attempt to model some of this complexity by applying inherently interactive machine learning techniques that are well suited to estimate contingent relationships without prespecifying all potentially relevant interactions.

As some recent studies have noted, structural and procedural accounts are not mutually exclusive and can be incorporated into more unified frameworks in which levels of repression are determined by structural sources of grievance and vice versa (e.g., Shadmehr 2014). The debate between structure- and process-based theories corresponds in the forecasting literature to a divide between models focused on “fundamentals”—country-year variables that change slowly, if at all—and those focused on “behaviors,” typically operationalized using event data at the subannual and/or subnational level. Similarly, attempts to reconcile these two sides of the theoretical debate correspond in the forecasting literature to models that incorporate both structural and procedural variables simultaneously, for example, in the context of split-population estimators (Bagozzi et al. 2015; Beger, Dorff, and Ward 2016). Our analysis draws on these approaches.

Building Our Forecasting Model of Civil War

Attempts to leverage structural theories for purposes of forecasting have generally yielded disappointing results (Hill and Jones 2014; Ward, Greenhill, and Bakke 2010). Our analysis therefore integrates insights from the structural literature in civil war studies with the procedural literature on contentious politics using event data to model the dynamics of escalation and de-escalation. While we are not the first to use event data to forecast conflict, previous studies have relied on proxies that do not closely correspond to theoretical priors in the literature, creating a gap between theory and measurement that we seek to close. With just one exception that we are aware of (Ward et al. 2013), existing models rely on either “Quad” counts or “Goldstein” scales (e.g., Beger, Dorff, and Ward 2016; Chiba and Gleditsch 2017; Montgomery, Hollenbach, and Ward 2012; Schrodtt, Woodward, and Marshall 2011). Quad counts aggregate disparate events into four general categories—verbal cooperation, material cooperation, verbal conflict, and material conflict (Schrodtt, Woodward, and Marshall 2011). Goldstein scales are similar, except that they further aggregate events into scales, with each event weighted according to how conflictive or cooperative it is deemed to be.³ These approaches are useful for distilling complex data sets into simple variables but are too coarse to capture the dynamics of escalation and de-escalation that procedural theories propose.

Of course, it is possible that event data sets are simply too noisy to model escalation more precisely, in which case Quad counts and Goldstein scales may be the best we can do. It is also possible that theory itself is unnecessary for forecasting, in which case the degree of correspondence between theory and measurement should not matter. Muchlinski et al. (2015, 88), for example, note that “quite accurate predictions” can sometimes be made even with a “relatively poor

understanding of the underlying causal processes” or through the use of causal models that are known, *ex ante*, to be incorrect (p. 88). Ward (2017) argues that conflict scholars should abandon “worship of theory” in favor of prediction. Some have gone further, suggesting that the combination of big data, sophisticated machine learning algorithms, and virtually limitless computing power has rendered theory “obsolete” (Anderson 2008). But to our knowledge, these ideas have never been tested, at least not within political science.

We use ICEWS event data to build models that incorporate proxies for conflict escalation and de-escalation processes, following the theoretical framework outlined above. We begin by coding events of peaceful claim-making by opposition or (potential) rebel groups. We then code the government’s response, either repression or accommodation or some combination of the two. Here, we follow Lichbach (1987), who proposes that the effects of repression could be conditional on other strategies used by the state. Lichbach shows that if repression is used with accommodation as a mixed strategy, it will reduce mobilization among groups that are not being repressed and that accommodations might prevent escalation when the costs of violent rebellion are high. Finally, we code incidents of low-level violence (e.g., assassinations and violent protests) that may precede and predict escalation to civil war.

In sum, our approach aims to capture group-based demands and state reactions—both repressive and accommodative—in a quantitative model that can be tested using event data. We then compare the predictive performance of our model to alternatives composed of Quad counts or Goldstein scales and to a more mechanical, atheoretical alternative composed of all possible events between governments, rebels, and opposition groups in the ICEWS data set—over 1,000 predictors in all. The better (or worse) our model performs relative to these alternatives, the more (or less) useful we consider the theoretical intuitions underlying it to have been. Finally, we compare our procedural model to an alternative that combines process-based variables with structural characteristics generally believed to raise or lower the risk of civil war. This allows us to test whether structural variables improve predictive performance or whether process might substitute for structure, at least over short forecasting windows.

Data

Operationalizing Civil War

Our dependent variable is derived from the Sambanis (2004) data set, updated through the end of 2015.⁴ We use the Sambanis data set for two reasons. First, it constitutes the most careful attempt to code not just the year but also the month in which civil wars begin, with explicit coding rules and detailed documentation for each case contained in a 256-page online supplement. Second and more significantly, the Sambanis data set conceptualizes civil war as a continuous process: most conflicts are coded as spanning the entire period from onset to termination without

interruption and thus do not start and stop from one year to the next depending on whether they happen to reach a particular threshold of battle-related deaths.⁵ We believe the former approach is of much greater practical and theoretical interest than the latter. We justify our choice of this data set in further detail in the Online Appendix and test the robustness of our results to variations on the Sambanis coding rules. We code our outcome as 1 for each country-month in which a civil war begins, 0 for each country-month of peace, and missing for each country-month of ongoing conflict.

Operationalizing Escalation

We use the ICEWS data set to construct four sets of predictors capturing the dynamics of escalation and de-escalation described above: demands, accommodations, nonviolent repression, and low-level violence. Each predictor is operationalized as an event count at the country-month level. All are coded dyadically and directionally and are lagged by either one month (for our one-month models) or six months (for our six-month models). The variable *demands* captures the first step in the escalation process, when opposition or (potential) rebel groups appeal to the government for reform.⁶ The government may respond to these demands with *non-violent repression*, *accommodation*, or some combination of the two.

These actions on the part of the government shape the costs and benefits of rebellion from the perspective of potential combatants by radicalizing recruits, inducing fear and resentment, or informing rational assessments of the costs and benefits of different claim-making strategies. Depending on the context and the mix of strategies used by the state, claim-making groups may decide that further escalation is either unnecessary (if they are appeased with concessions or accommodation), futile (if they fear more extreme violent repression), or necessary (if nonviolent claims are ineffective and the costs of violence are not prohibitively high). If accommodations fail to satisfy the opposition group, or if nonviolent repression fails to quash them, they may begin to engage in low-level violence against the state—protests, skirmishes, and selective killings. While most incidents of this sort defuse relatively quickly, some escalate into civil war. Small, isolated outbreaks of violence are an early warning sign for larger conflict escalation that we capture in our model using proxies for *low-level violence*.

Demands include events in which either opposition or rebels is the source and government is the target and typically involve appeals for political or institutional reform, policy change, or extension of rights. Conflict and Mediation Event Observations (CAMEO) codes used to define actors and events are listed in the Online Appendix. *Nonviolent repression* and *accommodation* include events in which government is the source and either opposition or rebels is the target. Nonviolent repression includes threats, sanctions, restrictions on freedom of movement, or rejections of group demands. Accommodations involve efforts to ease these restrictions or to accede to group demands. *Low-level violence* includes all events in which

(1) government is the source and opposition is the target, (2) opposition is the source and government is the target, (3) government is the source and rebels are the target, or (4) rebels are the source and government is the target. Most of these events are violent protests, but they also include property destruction, assassinations, and other forms of isolated, small-scale conflict and violence.⁷ We limit our analysis to events within (rather than between) countries and to the period beginning January 1, 2001,⁸ and ending December 31, 2015.

Several caveats are warranted. First, CAMEO codes correspond closely but not perfectly to the concepts we wish to measure, and one could experiment with different aggregations than the four we propose. Exhaustively comparing different permutations of these variables may prove worthwhile but is beyond the scope of this article. Second, there is enough overlap between categories that some events could justifiably be included as components of two or more variables simultaneously.⁹ Again, for compactness and tractability, we do not explore these variations here. Third, some event types occur infrequently or not at all in the ICEWS data set—at least not in the years or among the actors we consider.¹⁰ This is probably the result of a fourth concern about idiosyncrasies in the text-scraping algorithm used to populate the ICEWS data set, which may misclassify actors, event types, or both.

These caveats notwithstanding, ICEWS is the best and most thoroughly studied source of event data on interactions between governments, rebels, and opposition groups worldwide and is likely to continue attracting users as text-scraping improves. Moreover, the noise in ICEWS should privilege mechanical over theoretically grounded approaches to prediction since mechanical approaches can leverage any information in the data for purposes of forecasting whether or not that information maps onto relevant theoretical intuitions. We thus view ICEWS as a hard test for the value-added of theory for prediction.¹¹ In the Online Appendix, we explore ICEWS data quality in detail through a case study of Iraq. For further discussion of known issues with ICEWS, see Schrodtt and Van Brackle (2012), and for a discussion of reporting bias in media-based event data sets more generally, see Weidmann (2016).

Empirical Strategy

Models

We test five models:

1. The *escalation* model, composed of country-month event counts for (1) demands, (2) accommodations, (3) nonviolent repression, and (4) low-level violence as defined above. These variables are motivated by existing procedural theories of civil war and are designed to capture the violent and nonviolent interactions between governments, rebels, and opposition groups that should precede—and potentially predict—civil war onset.

2. The *Quad* model, composed of country-month Quad counts for (1) material conflict, (2) material cooperation, (3) verbal conflict, and (4) verbal cooperation (Leetaru and Schrodtt 2014).
3. The *Goldstein* (1992) model, composed of country-month Goldstein scales, which take values between -10 and 10 where -10 is the most conflictive and 10 is the most cooperative.
4. The *CAMEO* model, composed of country-month counts for all events in the CAMEO codebook—1,159 predictors in all.
5. The *average* model, composed of the unweighted average of predicted probabilities from the four models above.

For the *Quad*, *Goldstein*, and *CAMEO* models, we include all events in which (1) government is the source and opposition is the target, (2) opposition is the source and government is the target, (3) government is the source and rebels are the target, or (4) rebels are the source and government is the target. We include the *average* model because ensembles often outperform individual predictions (Montgomery, Hollenbach, and Ward 2012); though in this case, we are more interested in the comparison between models whose features correspond more and less closely to the underlying theories. We also explore various approaches to combining these ICEWS-based predictors with structural characteristics.

Estimation

The variety of techniques for forecasting rare events is immense and growing. We use ensembles of classification trees called random forests. Classification trees divide a data set into increasingly homogeneous subgroups through recursive partitioning. The process begins by identifying a predictor that most accurately distinguishes positive cases (onsets of civil war) from negative ones (peace). This first split partitions the data into two subsamples called nodes. Additional predictors are then selected to further improve the purity of each node (i.e., the extent to which it is populated entirely by positives or entirely by negatives). Each observation is passed down the tree until it reaches a terminal node at which point a prediction is made. Random forests iterate this process over many random subsamples of the data and many random subsets of predictors, increasing stability and obviating the need for cross-validation (Siroky 2009).

Random forests have a number of especially attractive properties for our purposes. First, unlike more familiar maximum likelihood estimators (e.g., logit or probit), random forests generate predictions based on random subsamples of variables and observations and so can accommodate hundreds of highly collinear predictors simultaneously. This is especially important for our *CAMEO* model that includes over 1,000 features. Second, and also unlike maximum likelihood estimators, random forests can ignore variables that do not improve predictive performance. This, too, is especially important for our *CAMEO*

model. Third, unlike regression shrinkage and selection techniques (e.g., ridge regression or least absolute shrinkage and selection operator), random forests are inherently interactive, as the role of each predictor depends on all the others at higher nodes. Structural models of civil war onset often rely on stringent functional form assumptions, some of which are never made explicit (Fearon and Laitin 2003; Collier and Hoeffler 2004); our approach is nonparametric, allowing us to avoid this problem. Procedural theories of escalation also suggest that the relationship between demands, accommodations, nonviolent repression, and low-level violence may be highly interactive and contingent. Random forests allow for these complexities.

Fourth, random forests can easily be adjusted to accommodate extreme class imbalances.¹² This is important because civil war onsets are rare events, especially when operationalized at the country-month level. Fifth and related, random forests are far less susceptible to overfitting than linear and maximum likelihood estimators and also many other machine learning algorithms (Breiman 2001; Siroky 2009). This is important for all of our models, especially the *CAMEO* model. (We further mitigate the risk of overfitting by restricting the size of each tree in the forest as discussed below.) Sixth, random forests are well established and widely used in the machine learning literature and have recently begun to emerge in economics and political science as well (Bazzi et al. 2019; Blair, Blattman, and Hartman 2017; Hill and Jones 2014; Muchlinski et al. 2015). Indeed, one recent study shows that random forests outperform maximum likelihood estimators when forecasting civil war onset using structural characteristics alone (Muchlinski et al. 2015). Finally, unlike other inherently nonlinear models (e.g., neural networks), random forests are intuitive and thus relatively easy to understand even for nonspecialists.

In our base specification, we implement random forests with five terminal nodes and 100 observations per tree (sampled without replacement) and 100,000 trees per forest. We train each model on a subset of data beginning January 1, 2001, and ending December 31, 2007, then test the models on the remaining data, forecasting over intervals of either one or six months. Because performance can be sensitive to these parameter choices, we also consider models with 10 (rather than five) terminal nodes, 500 (rather than 100) observations per tree, and 1,000,000 (rather than 100,000) trees per forest. We also consider three alternate start dates for the test set (January 1, 2009, 2010, or 2011) and two alternate coding rules for the dependent variable described in detail in the Online Appendix.¹³

Evaluating Model Performance

There are a variety of metrics available for assessing predictive performance. In accordance with our pre-analysis plan (PAP), we focus on the area under the receiver operating characteristic (ROC) curve (AUC). The AUC estimates the likelihood that a model assigns a higher predicted probability to a randomly chosen positive case than a randomly chosen negative one. The AUC is probably the most widely used

metric in the forecasting literature and has a number of advantages for our purposes: it is intuitive and easy to interpret; it incorporates true and false positives (Type I errors) and true and false negatives (Type II errors) simultaneously; it does not require specifying an arbitrary “discrimination threshold” above which we predict civil war will occur;¹⁴ it offers a single summary metric to facilitate comparison across models; and it is sensitive enough to detect variation in predictive performance, even with a rare outcome. This latter point is especially important: while we are interested in the absolute performance of our models, we are particularly interested in the relative performance of more and less theoretical alternatives. The AUC provides an easy way to assess this.

To complement the AUC, we also provide precision-recall curves and separation plots for the base specifications of our five models. Precision is the ratio of true positives to the sum of true and false positives at a particular discrimination threshold; recall is the ratio of true positives to the sum of true positives and false negatives. To construct separation plots (Greenhill, Ward, and Sacks 2011), we first arrange the predicted probabilities from each model in ascending order, then overlay bars representing actual civil war onsets at their corresponding points on the predicted probability distribution. While separation plots cannot be distilled to a single summary metric (like the AUC), they are informative and easy to understand and provide a useful visual representation of predictive performance. As additional complements to the AUC, in the Online Appendix, we also report Brier and F1 scores for our base specifications.

Results

Model Performance

Table 1 reports AUCs for our five models over one-month (top panel) and six-month (bottom panel) forecasting windows. Figure 1 displays the corresponding ROC curves, smoothed for ease of interpretation. ROC curves plot the trade-off between true and false positive rates across many discrimination thresholds. The higher the AUC, and the closer the ROC curve to a right angle, the better the model. Results from our base specification are reported in the top row of each panel in Table 1; the remaining rows report results for the alternate specifications described above. Figure 1 displays ROC curves for the base specification alone.

The *escalation* model outperforms the alternatives over both one-month and six-month forecasting windows. Over one-month windows, the AUC for the *escalation* model ranges from a low of .79 to a high of .92, with a mean and median of .85 across the nine specifications we test. (While there is no universal rule of thumb to apply here, AUCs in the .60s are generally considered poor, .70s average, .80s good, and .90s excellent.) In contrast, the AUC ranges from .78 to .81 for the *Quad* model, .69 to .81 for the *Goldstein* model, .75 to .86 for the *CAMEO* model, and .76 to .84 for the *average* model. The *escalation* model’s AUCs are almost all good to excellent; the others are good in some specifications but average in others.

Table 1. Out-of-sample Area under the Receiver Operating Characteristic Curves for One-month and Six-month Forecasts.

Model	Escalation	Quad	Goldstein	CAMEO	Avg.
One-month forecasts					
Base specification	.85	.80	.79	.82	.82
Terminal nodes	.85	.80	.78	.83	.82
Sample size	.85	.81	.71	.86	.84
Trees per forest	.85	.80	.78	.83	.82
Training/test sets 1	.86	.78	.76	.81	.80
Training/test sets 2	.81	.79	.73	.77	.78
Training/test sets 3	.79	.81	.69	.75	.76
Coding of DV 1	.86	.81	.79	.84	.83
Coding of DV 2	.92	.80	.81	.81	.81
Six-month forecasts					
Base specification	.82	.78	.82	.76	.79
Terminal nodes	.80	.76	.81	.76	.78
Sample size	.83	.78	.78	.79	.79
Trees per forest	.82	.78	.82	.77	.79
Training/test sets 1	.79	.78	.81	.76	.78
Training/test sets 2	.73	.73	.76	.73	.75
Training/test sets 3	.88	.71	.81	.68	.79
Coding of dependent variable 1	.83	.78	.82	.78	.80
Coding of dependent variable 2	.83	.77	.83	.78	.79

Note: AUCs for our five random forests models. The top row in each panel reports AUCs for the base specification. We also report results with 10 rather than five terminal nodes (second row); 500 rather than 100 observations per tree (third row); 1,000,000 rather than 100,000 trees per forest (fourth row); a test set that begins January 1, 2009, January 1, 2010, or January 1, 2011 (fifth, sixth, and seventh rows, respectively); and alternate codings of the dependent variable (eighth and ninth rows) as described in the Online Appendix.

The results are starker when we compare across models but within specifications. Over one-month windows, the *escalation* model outperforms the *Quad* model by as many as 13 points on the AUC and underperforms in just one specification (by just two points), outperforms the *Goldstein* model by as many as 14 points and never underperforms, outperforms the *CAMEO* model by as many as 11 points and underperforms in just one specification (by just one point), and outperforms the *average* model by as many as 11 points and never underperforms.

A comparison to other recent conflict forecasting exercises suggests these are substantial differences. Muchlinski et al. (2015, 94), for example, describe differences in AUC ranging from .09 to .14 as “sizable margins”; Ward, Greenhill, and Bakke (2010, 367) describe a .10 difference as “quite substantial”; Weidmann and Ward (2010, 891) describe a seven-point increase as evidence of “clearly visible improvement”; Blair, Blattman, and Hartman (2017, 307) describe six- to nine-point increases as “significant”; and Gleditsch and Ward (2013) describe differences of

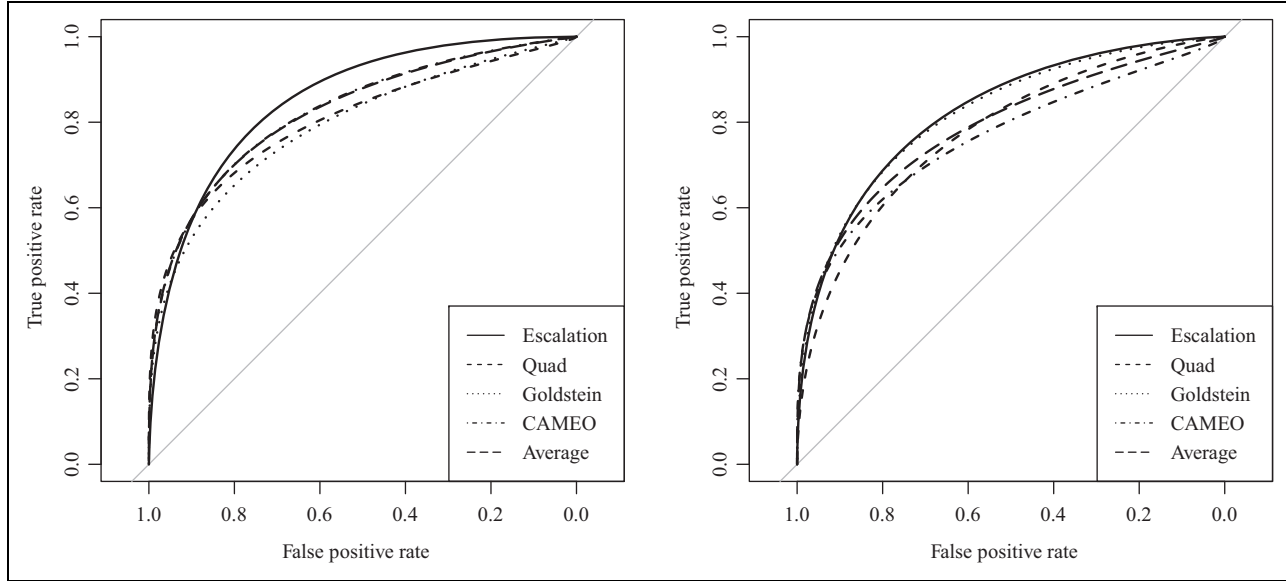


Figure 1. Out-of-sample receiver operating characteristic (ROC) curves for the base specification of our five random forests models. The left and right plots show results from one-month and six-month forecasts, respectively.

just two to three points as evidence of “consistent” improvement. Indeed, given the rarity of our dependent variable, gaining even a few points in AUC is important.

Our results are more mixed over six-month windows, but the *escalation* model continues to dominate, with a minimum AUC of .73 and a maximum of .88. While the *Goldstein* and *average* models outperform at the bottom of this distribution (with lows of .76 and .75, respectively), they underperform at the top, with highs of .83 and .80, respectively. Comparing across models but within specifications, the *escalation* model outperforms the *Quad* model by as many as 17 points and never underperforms, outperforms the *Goldstein* model by as many as seven points but performs equally well in three specifications and underperforms in three (by one to three points), outperforms the *CAMEO* model by as many as 20 points and never underperforms, and outperforms the *average* model by as many as nine points and underperforms in just one specification by two points. While *Goldstein* is a close competitor to the *escalation* model over six-month windows, it underperforms dramatically over one-month windows.

The alternatives are also much less consistent in their performance relative to one another. While the *Goldstein* model performs as well or better than the *quad*, *CAMEO*, and *average* models in all but one specification over six-month windows, it performs worse than the others in all but one specification over one-month windows. While the *CAMEO* model performs as well as or better than the *Quad* model in all but two specifications over one-month windows, the *Quad* model performs as well as or better than the *CAMEO* model in all but two specifications over six-month windows (though the margins are generally small). And while the *average* model outperforms both the *Quad* and *CAMEO* models in most specifications, it still underperforms the *escalation* model in all but one specification. In other words, while the *escalation* model outperforms regardless of the specification or forecasting window we choose, the relative performance of the other models is idiosyncratic.

As an alternative metric, Figure 2 plots precision-recall curves for the base specification of the one-month and six-month models. The figure suggests that the *escalation* model outperforms at low and high recall in particular, underperforming (but not by much) at mid recall. Figures 3 and 4 also display separation plots for the base specification of our one-month and six-month models, respectively. The more concentrated the vertical bars on the right-hand side of the plot, the better the model. While the rarity of civil war onset makes some of these plots difficult to compare visually, the *escalation* model clearly outperforms the alternatives. Figures 3 and 4 also clarify the source of this superiority. Over both one-month and six-month windows, there are three onsets in particular that *escalation* struggles to detect—Syria, Mali, and Libya, all in 2011—and even these are clustered toward the right-hand side of the figure, at least in the one-month model. The other models are more idiosyncratic, with more onsets scattered along the left-hand side of the plots, and fewer concentrated on the right-hand side.

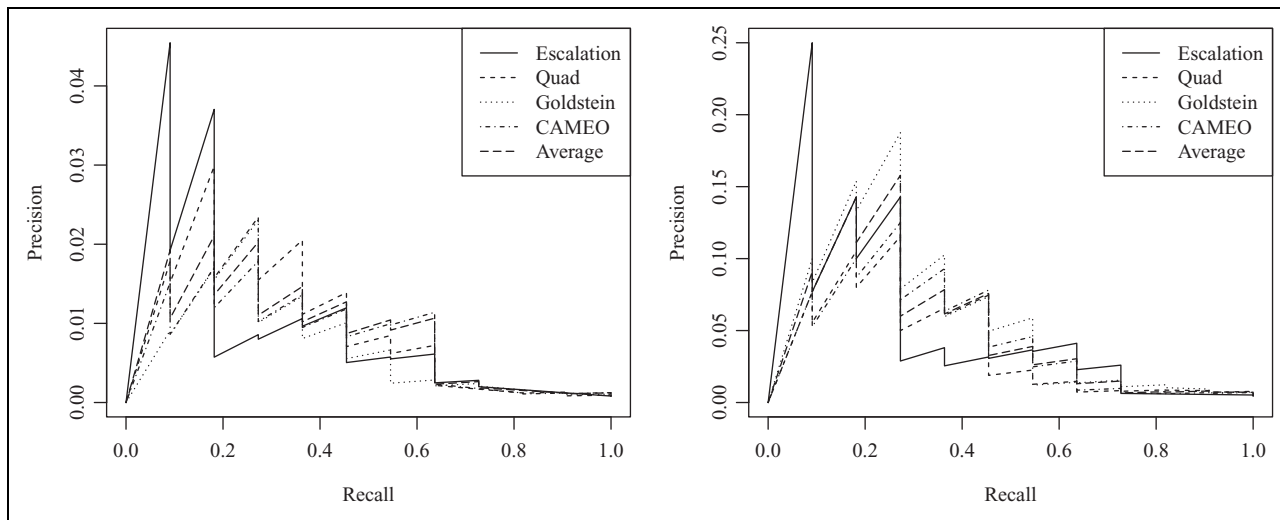


Figure 2. Out-of-sample precision-recall curves for the base specification of our five random forests models. The left and right plots show results from one-month and six-month forecasts, respectively.

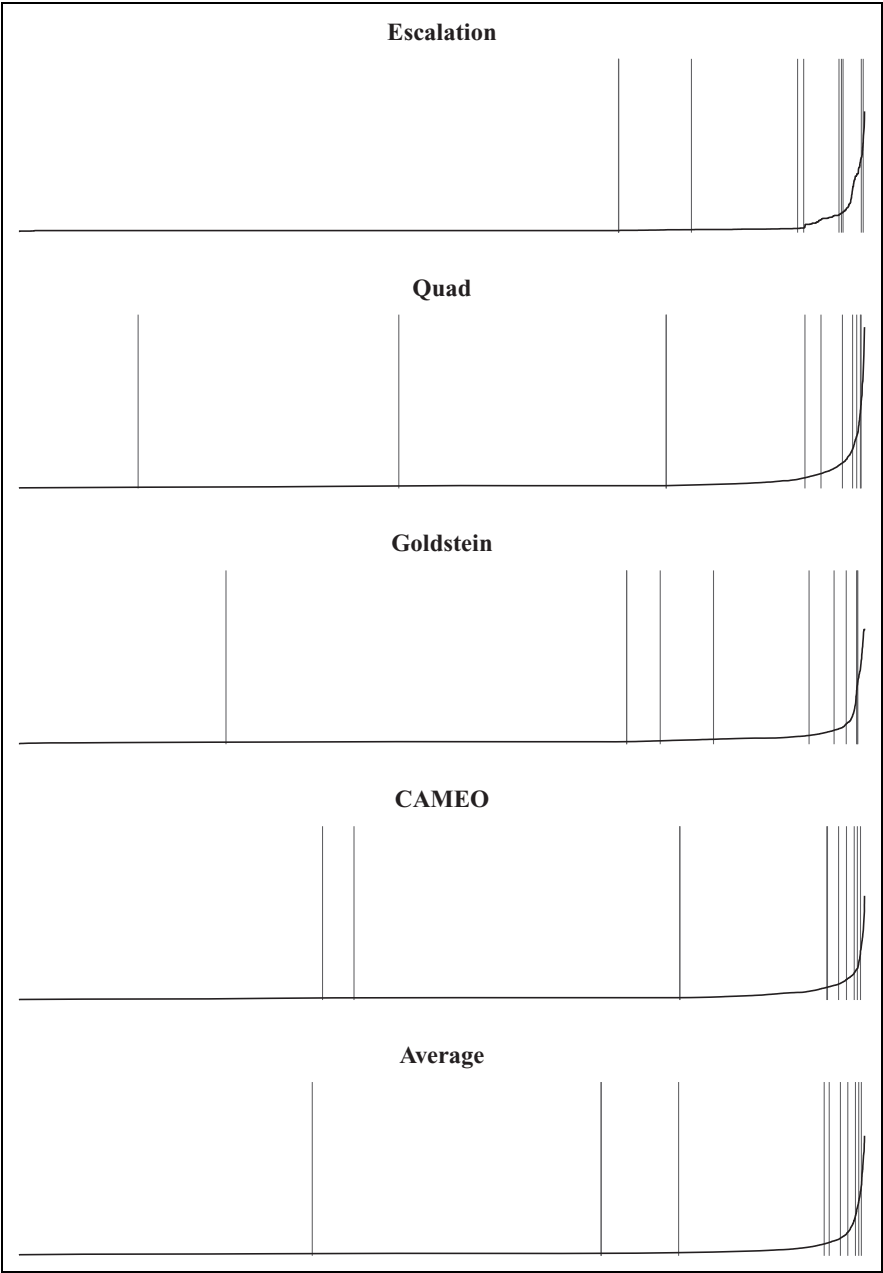


Figure 3. Out-of-sample separation plots for the base specification of our five random forests models over one-month forecasting windows. The vertical bars denote civil war onsets.

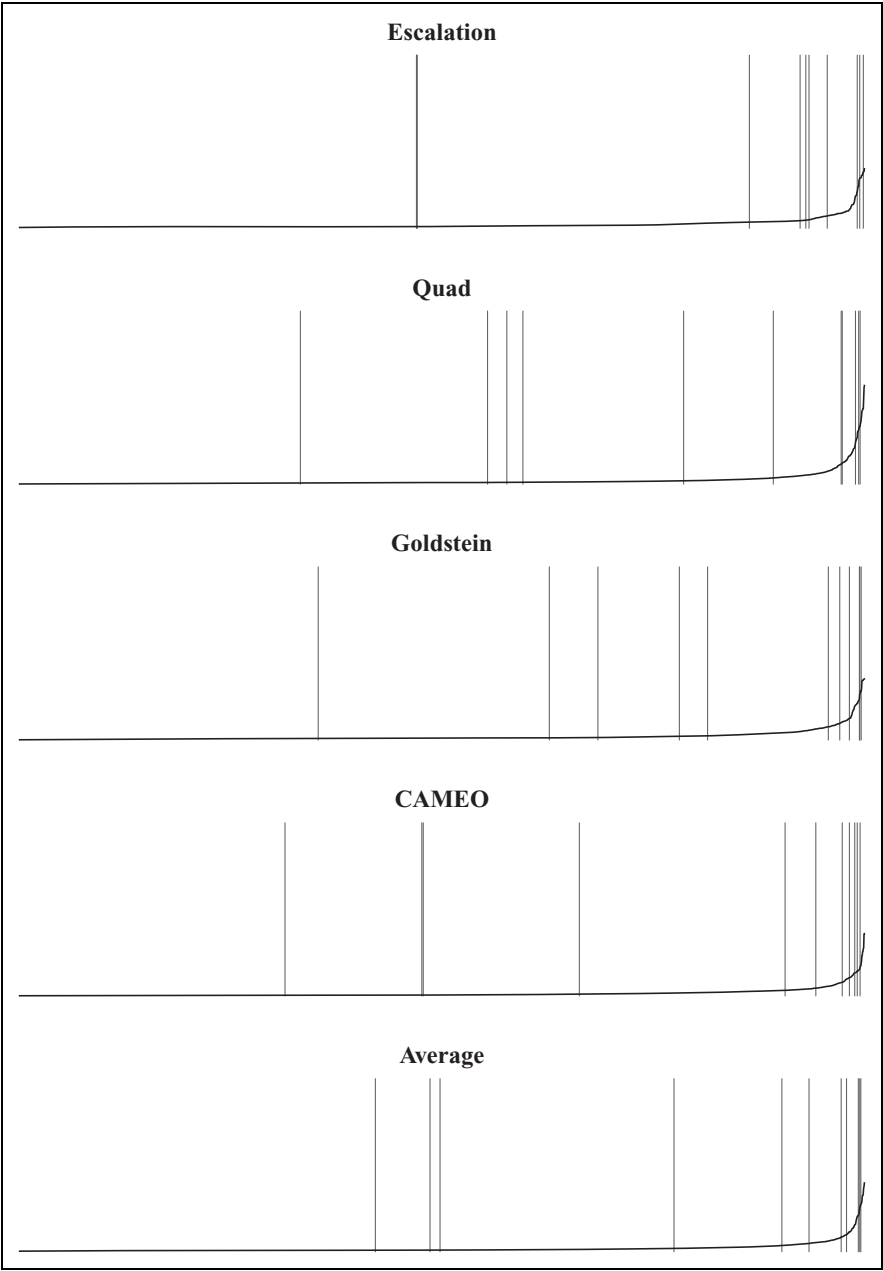


Figure 4. Out-of-sample separation plots for the base specification of our five random forests models over six-month forecasting windows. The vertical bars denote civil war onsets.

Table 2. Out-of-sample Area under the Receiver Operating Characteristic Curves for Models Including PITF.

	<i>Escalation Only</i>	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts	.85	.78	.53	.84	.76
Six-month forecasts	.82	.86	.52	.83	.74

Note: AUCs for the *escalation* model incorporating Political Instability Task Force (PITF) predictors. The top and bottom panels report results from one-month and six-month forecasts, respectively.

The separation plots also indicate that the *escalation* model's performance in terms of AUC is not simply a result of correctly predicting spells of peace; the model also outperforms when predicting onsets of civil war. Nor is the performance of the *escalation* model an artifact of parsimony alone. As forecasters have long recognized, simple models often outperform complex ones in out-of-sample tests. This may explain why the *escalation* model outperforms the *CAMEO* model, since the latter is composed of over 1,000 predictors, the former only 10. Parsimony cannot, however, explain why *escalation* outperforms *Quad* and *Goldstein*, which are also parsimonious. (Indeed, *Goldstein* is the most parsimonious of the models we test, with just four features.) This suggests that predictive power improves not just with a more parsimonious selection of variables but with a closer match between theory and measurement as well.

Combining Process with Structure

Our predictions thus far do not incorporate the structural characteristics typically included in models of civil war onset. While structural variables change too slowly to predict *when* conflict will occur (at least subannually), they may help predict *where* it will occur, potentially in ways that ICEWS cannot. Opposition groups, for example, may make as many demands of government in high-risk countries as in low-risk ones, but low-risk countries may have structural features that make accommodation more likely. Identifying these countries *ex ante* may improve the performance of our models.

To estimate structural risk, we rely primarily on the Political Instability Task Force's (PITF) "Internal Conflict Watch List" that provides annual predicted probabilities of onset for all countries in the world. Predictions are derived from the model in Goldstone et al. (2010), which is composed of four structural variables: regime type, infant mortality, prevalence of civil war in neighboring countries, and state-led discrimination.¹⁵ In Table 2, we test three approaches to combining process (ICEWS) with structure (PITF), focusing on the *escalation* model. The first (column 2) simply adds PITF predictors alongside the ICEWS-based ones. The second (column 3) uses PITF predicted probabilities to weight the results of the *escalation* model, ensuring that high-risk and low-risk countries that happen to take similar

values on ICEWS-based predictors are nonetheless assigned different predicted probabilities in most months.

The final approach is a random forest analog to split-population modeling.¹⁶ We first compute the average PITF predicted probability for each country across all years in our training set. We define those that fall in the bottom quartile as “low risk” and the rest as “high risk.” We then run our *escalation* model on the high-risk and low-risk subsets separately, combining the results into a single random forest (column 4). For purposes of comparison, Table 2 also reproduces the *escalation* model results from Table 1 (column 1) and includes results from a purely structural model composed of the four PITF predictors alone (column 5).

Overall, our results suggest that while measures of structural risk may improve predictive performance, the value they add is marginal and inconsistent. Of the approaches we test, the split-population analog is most promising, with AUCs of .84 and .83 over one-month and six-month windows, respectively. These are very similar to the AUCs of the *escalation* model. Adding PITF predictors improves the performance of the *escalation* model over six-month windows but diminishes it over one-month windows. The weighted model performs very poorly regardless. Incorporating PITF thus significantly reduces or only slightly improves the performance of the *escalation* model, regardless of the approach we take. This is counterintuitive but may reflect the relatively short windows over which we are forecasting. The longer the window, the more informative slow-moving structural variables should be. Over shorter windows, process may substitute for structure, making the latter superfluous for prediction.

Forecasts for First Half of 2016

Finally, Table 3 reports “true” forecasts for the first half of 2016, using predictors measured through December 31, 2015. These forecasts were preregistered with the EGAP network to mitigate concerns about “fishing” for results (e.g., by retroactively adjusting our model parameters to improve performance).¹⁷ For compactness and ease of exposition, we report results for the 30 highest risk countries only, based on six-month forecasts from the *escalation* model. In raw terms, the risk of civil war in these countries ranges from a high of seven percent to a low of one percent. The predicted probabilities of onset become almost indistinguishably small lower down the list, making the top 30 an appropriate place to focus our attention.

Importantly, we make no assumptions about the persistence of civil war between the second half of 2015 and the first half of 2016, meaning that all countries—even those with ongoing conflicts as of December 31, 2015—are included in the 2016 test set. As a result, 14 of the 30 countries on the list (and nine of the top 10) are cases of ongoing conflict. Equally important, four of the countries on the list (India, Afghanistan, Russia, and Colombia) are coded as ongoing conflicts for the duration of our training set as well. These constitute de facto out-of-sample tests since our model was not trained on them at all. The fact that our model nonetheless identifies them as

Table 3. 2016 Prediction Rankings for Six-month *Escalation* Model.

Country	Predicted probability	Country	Predicted probability	Country	Predicted probability
Nigeria	.074	Burundi	.041	Thailand	.016
India	.072	Egypt	.039	Iran	.016
Iraq	.072	Yemen	.036	Myanmar	.015
Somalia	.071	Colombia	.027	Montenegro	.013
Syria	.065	Mali	.022	Bangladesh	.012
Pakistan	.065	China	.021	Niger	.012
Philippines	.065	Indonesia	.019	El Salvador	.011
Turkey	.064	Ukraine	.019	France	.010
Afghanistan	.064	Sudan	.017	Ghana	.010
Russia	.047	Lebanon	.016	Tajikistan	.010

Note: Ordinal ranking of the top 30 countries at highest risk of civil war in the first half of 2016 based on predicted probabilities from the *escalation* model using predictors measured through December 31, 2015.

high risk in the test set suggests that our model parameters are portable not just over time but across space as well.

Other results are similarly encouraging. The fact that the list includes countries suffering severe, geographically dispersed violence (such as Afghanistan and Syria) as well as those suffering milder, more geographically isolated violence (such as India and Colombia) suggests that our predictors are capable of capturing a wide range of conflict dynamics. The overlap between our list and the PITF Watch List for 2016 is encouraging as well. The PITF identifies 18 of the countries on our list as being “in crisis,” including 16 of our top 20 and all of our top seven.¹⁸ Of the remaining 12, the PITF classifies three (Indonesia, Iran, and Bangladesh) as “most” at risk, five (Turkey, Lebanon, Niger, Tajikistan, and France) as “significant” risk, and three (Burundi, El Salvador, and Ghana) as “some” risk. (The PITF classifies Montenegro as low risk.) These parallels are especially remarkable given that ICEWS is newer and noisier than the data sets on which the PITF relies (e.g., Minorities at Risk, which began in 1986).

Our model also identifies escalating risk in countries that the PITF model misses. Neither Turkey nor Burundi was coded as an ongoing civil war in the second half of 2015, but both experienced episodes of instability in the first half of 2016 that would arguably qualify as onsets according to Sambanis’s definition. In Turkey, the government reached a ceasefire with the Kurdistan Workers’ Party (PKK) in 2014, but conflict reignited in July 2015 with waves of violence between state security forces and PKK-backed militias and radical splinter groups. But violence did not reach levels that would warrant inclusion as a civil war until 2016. In July 2016, one month after the end of our test set, a faction of the Turkish Armed Forces attempted a coup d’état, further escalating the crisis and initiating a prolonged period of purges and violent state repression.

In Burundi, President Pierre Nkurunziza incited popular protests and an attempted coup d’état in May 2015 after announcing his intention to run for a third

term in office, in defiance of constitutional limits. The death toll remained low through the second half of 2015 but spiked with a series of grenade attacks and targeted killings of military and government officials in the first three months of 2016.¹⁹ By March, the European Union had responded by suspending direct budgetary support to the country. Observers in the media were similarly alarmed. Instability in Burundi was significant enough that the country ranked fifth on *Business Insider's* list of "16 countries with the most civil unrest" in 2016 (Martin 2016) and was also included on *Foreign Policy's* list of "10 conflicts to watch in 2016" (Guéhenno 2016). Turkey was included on the latter list as well. (*Business Insider's* list is based on research by the risk consultancy firm Verisk Maplecroft.)

There was evidence of unrest in some of our apparent false positives as well. Egypt has witnessed frequent violence since the Arab Spring, including a series of attacks by the terrorist group Sinai Province between January and June 2016. Security in Lebanon eroded through the first half of 2016 as well with bombings in Beirut and the Bekaa Valley and an unresolved political impasse surrounding the presidential election of April 2014 spurring protests and boycotts. In Indonesia, Islamic State claimed responsibility for a series of bombings in January 2016, and peaceful protests in Papua and West Papua the following month were accompanied by violent attacks by the Free Papua Movement and by the arrest of hundreds of activists.

Perhaps the most obvious false positive on our list is France. But our model is not unique in this respect. France ranked surprisingly high on the PITF Watch List as well and was classified as being at "significant" risk of civil war; France was also ranked #16 on *Business Insider's* list of 16 countries with the most civil unrest in 2016 and was deemed the developed country at highest risk of conflict by the firm Verisk Maplecroft (Martin 2016). Whatever sources of risk our model is detecting, other models seem to be detecting them too—even if they are false alarms.²⁰

There are some false negatives as well. The two most conspicuous are the Democratic Republic of the Congo (DRC) and Libya—countries with ongoing civil wars in the first half of 2016. While we cannot know for certain why our model failed to detect these cases, a possible explanation may lie in the declining severity of both conflicts over the course of 2015. The Social Conflict Analysis Database (SCAD), for example, records 125 incidents of violence in Libya in 2015, down from 183 the year before. Moreover, SCAD records a substantial reduction in the frequency of incidents between the first and second halves of 2015. The Armed Conflict Location and Event Dataset (ACLED) reveals a similar (though less dramatic) trend. Both SCAD and ACLED suggest that the frequency of incidents in DRC declined over the course of 2015 as well. Other potential omissions from our list include the Central African Republic, where the Seleka alliance overthrew the government in 2013, and Venezuela, where protests erupted in 2016 following the Supreme Court's controversial decision to seize control of the legislature.

Overall, however, our model appears to produce relatively few obvious false positives or negatives. As a more formal test, Table 4 presents confusion matrices for our

Table 4. 2016 Confusion Matrices for Six-month *Escalation* Model.**Assuming Persistence**

		Predicted	
		0	1
Observed	0	132	17
	1	2	13

Assuming Change

		Predicted	
		0	1
Observed	0	132	16
	1	2	14

Note: Confusion matrices based on predicted probabilities from the six-month *escalation* model. We code as 1 the top 30 countries at highest risk of civil war in the first half of 2016. The top panel assumes that all ongoing civil wars as of December 31, 2015, continued in the first half of 2016 and that no new civil wars began. The bottom panel codes the ongoing civil war in Colombia ending in the first half of 2016 and codes new civil wars beginning in Turkey and Burundi.

2016 forecasts using the cutoff implicit in Table 3 (i.e., the predicted probability threshold that separates the 30 countries at highest risk of onset from the rest of the test set). Since the Sambanis data set was not updated through 2016 at the time of our analysis, there is inevitably some discretion in the way we code onsets and terminations in the first half of 2016. We therefore present two variations on this exercise. The top panel of Table 4 presents a confusion matrix assuming no new onsets or terminations. In other words, this panel assumes that all countries that were in conflict in the second half of 2015 continued to be in conflict in the first half of 2016 and that all countries that were at peace in the second half of 2015 continued to be at peace.

Following the discussion above, the bottom panel instead assumes that new civil wars began in Turkey and Burundi in the first half of 2016. It also assumes that the civil war in Colombia ended by the first half of 2016. A peace process with the country's largest rebel group, the Revolutionary Armed Forces of Colombia (FARC), began with exploratory meetings in 2011. Perhaps the most important moment occurred in September 2015, when the government and FARC agreed on terms for transitional justice. But negotiations continued into 2016, and the agreement was not finalized until June 23 of that year. Irrespective of these coding decisions, the formal results in Table 4 confirm our more informal analysis above. Our model correctly predicts 13 to 14 cases of civil war and 132 cases of peace. It generates only two false negatives (DRC and Libya) and 16 or 17 false positives, for a true positive to false positive ratio of .76 or .88, depending on the coding choices we use. These are reasonable given the rarity of the dependent variable.

Conclusion

We assess whether a theoretically driven procedural model based on ICEWS event data can forecast the onset of civil war better than more mechanical alternatives. We also test whether slow-moving structural characteristics can improve the performance of models based on fast-moving “process” variables alone. While all conflict forecasting models are motivated to some extent by theory, we improve on recent studies using event data by closing the gap between the predictors we measure and the theories from which they are derived. Following the lead of experimental social scientists, we also retrospectively evaluate the accuracy of “true” forecasts for the first half of 2016, which we preregistered with the EGAP network in order to mitigate concerns about “fishing” for results.

Our theoretically driven model generates accurate forecasts, with base specification AUCs of .82 and .85 over one- and six-month windows, respectively, and AUCs as high as .92 in other specifications. Our model also consistently and sometimes dramatically outperforms the alternatives we test. Structural characteristics may improve the performance of procedural models, but only marginally, suggesting that process can substitute for structure over short forecasting windows and can generate better predictions (and potentially more useful policy prescriptions) than the “standard” models of civil war (Ward, Greenhill, and Bakke 2010). Our forecasts for the first half of 2016 also identified a number of at-risk countries (and two potential civil war onsets) that models based on structural characteristics alone appear to have missed.

The fact that our most mechanical model (*CAMEO*) performs so poorly suggests that in the context of a prediction problem involving rare events and complex causal processes, a purely inductive approach to feature selection cannot adjudicate relevant from irrelevant predictors and so does not generate accurate forecasts. In a recent review, Cederman and Weidmann (2017, 476) argue that “the hope that big data will somehow yield valid forecasts through theory-free ‘brute force’ is misplaced in the area of political violence.” Our results lend some credence to this claim.

Our analysis is not without limitations. One is our inability to link sequences of actions and reactions between actor dyads. This limitation arises from the structure of the ICEWS data set, which does not record connections between related events, and thus cannot capture the tit-for-tat dynamics that typically characterize conflict escalation. Recording these connections would almost certainly require more intimate involvement of human coders at various stages of the data collection process. This may be a worthwhile investment for countries viewed as high priorities for monitoring and intervention. In the meantime, our analysis suggests that ICEWS can be used to forecast civil wars with high levels of accuracy, especially when informed by theories of conflict escalation. Several recent studies have introduced innovative methods for modeling dependencies between multiple sets of actors in civil wars and other contexts (Dorff, Gallop, and Minhas 2020; Minhas, Hoff, and Ward 2019);

these approaches could plausibly be applied to future iterations of ICEWS, though not in its current form.

Given the poor predictive performance of many apparently theoretically important variables, some forecasters have questioned whether theory is necessary for prediction at all, especially as vast computing power and sophisticated machine learning techniques become increasingly accessible. To our knowledge, ours is the first study to treat the value-added of theory not as an assumption but rather as a hypothesis to test. Our results argue for a closer connection between theory and prediction than is sometimes assumed, at least within the study of conflict. Conflict forecasting models are most informative when compared to some benchmark (Cederman and Weidmann 2017; Cranmer and Desmarais 2017); we urge future researchers to consider using more atheoretical models as benchmarks for assessing the value-added of more theoretical ones.²¹ Exercises of this sort will only become more important as computing power continues to increase and as “big” data sets like ICEWS continue to proliferate. We view the combination of existing theories with these relatively new data sets as one of the literature’s most interesting and promising frontiers.

Authors’ Note

The views expressed herein are the principal investigators’ alone and do not represent the views of the US government.

Acknowledgments

We thank Michael Colaresi, Benjamin Fisher, Philip Schrodt, Jonas Vesby, participants in the Political Instability Task Force semiannual conferences of summer 2014 and winter 2015, participants in the 2018 American Political Science Association conference, and three anonymous reviewers.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research described in this project was sponsored by the Political Instability Task Force (PITF). The PITF is funded by the Central Intelligence Agency.

ORCID iD

Robert A. Blair  <https://orcid.org/0000-0001-6054-1421>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Our pre-analysis plan (PAP) is available at <https://egap.org/registration/1760>.
2. The closest exception is Beger, Dorff, and Ward (2016), who retrospectively evaluate forecasts of “irregular leadership change” in 2014 (though as best we can tell these forecasts were not published in a publicly accessible venue). Some conflict forecasters—notably Jay Ulfelder, Andreas Beger, Håvard Hegre, and Michael Ward—have published their predictions on their websites or as blog posts. We view this as a welcome trend though not a substitute for formal preregistration.
3. Originally, these weights were defined by a panel of faculty members from the University of Southern California; see Goldstein (1992).
4. Because South Sudan did not exist as a country until after the end of our training set, we exclude it from our analysis. We do, however, include Sudan.
5. Specifically, Sambanis (2004, 829-31) defines civil war as an armed conflict (1) that takes place within the territory of an internationally recognized state with a population of at least 500,000; (2) in which parties are organized along political and military lines, in pursuit of explicit, public political goals; (3) in which the government is a combatant; (4) in which rebels recruit primarily from within the country where the armed conflict is waged; (5) that is characterized by “sustained” violence, meaning there is no three-year period during which the death toll falls below 500; and (6) in which the weaker party is able to mount “effective resistance” against the stronger party, meaning that it inflicts at least 100 deaths. Importantly, onset is coded as the first year in which the armed conflict causes at least 500 deaths; if it does not cause 500 deaths in its first year, it is coded as starting in that year only if it causes a total of at least 1,000 deaths over the next three years. Termination is coded when a peace agreement produces at least six months of peace or when a victory by one side produces either six months of peace or a regime change.
6. This variable would capture protests and other nonviolent claims made at the outset of a conflict as in recent studies of conflict escalation by Chenoweth and Stephan (2017) and Cunningham et al. (2017).
7. Of course, from a theoretical perspective, it is not especially surprising or interesting to posit that low-level violent protests might escalate into higher-level civil war. But as Brubaker and Laitin (1998, 426) noted in an early review, the relationship is not mechanical. Low-level violent protests are far more common than civil wars, and there is little empirical evidence demonstrating that higher levels of conflict mechanically predict higher levels of violence or that higher levels of violence mechanically predict the onset of civil war. In any event, low-level violence is just one of four predictors in our theoretically motivated model.
8. We opt not to use earlier Integrated Crisis Early Warning System (ICEWS) data based on advice from the Political Instability Task Force (PITF) as the ICEWS actor and event dictionaries were overhauled in 2000.
9. For example, Conflict and Mediation Event Observations (CAMEO) code 124 (“Refuse to yield, not specified below”) might be classified as nonviolent repression when the source is government and the target opposition or perhaps as a demand when the source is opposition and the target is government.

10. For example, in the version of the data set we use, there is no case of rebel groups violently protesting for leadership change (CAMEO code 1451) and only one case of government easing a ban on opposition political parties (CAMEO code 0812).
11. Of course, one could argue that our decision to use event data is itself a theoretical one, as is our decision to focus on events between governments and opposition groups or rebels (rather than, say, between governments and civil society organizations). Even our more mechanical model is therefore not entirely devoid of theoretical motivation. Nonetheless, it is far less theoretically grounded than our preferred alternative, as it assumes no *ex ante* difference in the predictive power of disparate event types, from “make pessimistic comment” (CAMEO code 12) to “consider policy option” (CAMEO code 14) to “bring lawsuit against” (CAMEO code 115) to “kill by physical assault” (CAMEO code 1822).
12. There are a variety of adjustments one can make: “down-sampling” the majority class, “up-sampling” the minority class, generating synthetic data for the minority class, or simply increasing the number of trees in the forest to ensure adequate minority class representation. We use the latter approach. For a recent example of down-sampling, see Muchlinski et al. (2015).
13. Again, there are infinitely many permutations of these parameters that we might test. We aim to demonstrate that the relative performance of our models is consistent across a range of equally plausible options.
14. The disadvantage of metrics such as accuracy, sensitivity (true positive rate), and specificity (true negative rate) is that they can only be estimated at specific discrimination thresholds. The choice of threshold is inevitably arbitrary, and the results may not be comparable across models. Goldstone et al. (2010), for example, report predictive performance at the discrimination threshold that equalizes accuracy, sensitivity, and specificity. In our case, however, the rarity of the dependent variable ensures that these three metrics are almost never exactly equalized, complicating comparison across models. Blair, Blattman, and Hartman (2017) instead use the threshold that maximizes sensitivity while maintaining an accuracy rate of at least 50 percent, but again, disparities in the maximum sensitivity rate achieved by our models preclude easy comparison.
15. The latter is operationalized as a score of 4 or higher on a discrimination index from the Minorities at Risk data set.
16. Split-population models consist of two nested likelihood functions. The first estimates the probability that a given country is “at risk” of civil war, and the second estimates the probability that a civil war will actually occur in that country. Our third model is similar in spirit to this approach, except that we replace the first likelihood function with predicted probabilities gleaned from PITF.
17. Due to delays in the release of the last tranche of ICEWS data for 2015, the predictions were not officially registered until February 22, 2016. Nonetheless, they were prospective in that they used data through 2015 only and in that they were preregistered well before the end of the six-month test window. The list in Table 3 also diverges slightly from the one we preregistered in order to ensure replicability after updates to ICEWS. The disparities are small and inconsequential. In our PAP, Iran was ranked above Thailand, rather than vice versa; Niger was ranked above Bangladesh, rather than vice versa; and

Ghana and Tajikistan were ranked above France, rather than vice versa. For all of these pairs of cases, the predicted probabilities of onset are identical to three or more decimal places.

18. The PITF uses a five-category classification system to characterize risk: (1) in crisis, (2) most risk, (3) significant risk, (4) some risk, and (5) low risk.
19. See the International Crisis Group's "Crisis Watch" project available at <https://www.crisisgroup.org/crisiswatch/database>.
20. One possible explanation is media overreporting of low-level violence associated with perceived religious conflicts. ICEWS records an abnormally high incidence of low-level violence in France, especially relative to other rich industrialized nations.
21. Cranmer and Desmarais (2017) make a similar point but argue for the use of "state-of-the-literature" models or models based on lagged values of the dependent variable.

References

- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *WIRED*, June 23. Accessed April 6, 2020. <https://www.wired.com/2008/06/pb-theory/>.
- Bagozzi, Benjamin E., Daniel W. Hill, Will H. Moore, and Bumba Mukherjee. 2015. "Modeling Two Types of Peace: The Zero-inflated Ordered Probit (ZiOP) Model in Conflict Research." *Journal of Conflict Resolution* 59 (4): 728-52.
- Bazzi, Samuel, Robert Al. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon, and Ricard Peck. 2019. "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia." NBER Working Paper No. 25980. Accessed April 6, 2020. <https://www.nber.org/papers/w25980>.
- Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *The American Political Science Review* 94 (1): 21-35.
- Beger, Andreas, Cassy L. Dorff, and Michael D. Ward. 2016. "Irregular Leadership Changes in 2014: Forecasts Using Ensemble, Split-population Duration Models." *International Journal of Forecasting* 32 (1): 98-111.
- Blair, Robert A., Christopher Blattman, and Alexandra Hartman. 2017. "Predicting Local Violence: Evidence from a Panel Survey in Liberia." *Journal of Peace Research* 54 (2): 298-312.
- Brancati, Dawn. 2006. "Decentralization: Fueling the Fire or Dampening the Flames of Ethnic Conflict." *International Organization* 60:651-85.
- Brandt, Patrick T., John R. Freeman, and Philip A. Schrodt. 2011. "Real Time, Time Series Forecasting of Inter- and Intra-state Political Conflict." *Conflict Management and Peace Science* 28 (1): 41-64.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5-32.
- Brubaker, Rogers, and David Laitin. 1998. "Ethnic and Nationalist Violence." *Annual Review of Sociology* 24:423-52.
- Cederman, Lars-Erik, Simon Hug, Andreas Schädel, and Julian Wucherpfennig. 2015. "Territorial Autonomy in the Shadow of Conflict: Too Little, Too Late?" *American Political Science Review* 109 (02): 354-70.

- Cederman, Lars-Erik, and Nils B. Weidmann. 2017. "Predicting Armed Conflict: Time to Adjust Our Expectations?" *Science* 355 (6324): 474-76.
- Cederman, Lars-Erik, Nils B. Weidmann, and Kristian Skrede Gleditsch. 2011. "Horizontal Inequalities and Ethnonationalist Civil War: A Global Comparison." *American Political Science Review* 105 (3): 478-95.
- Cederman, Lars-Erik, Andreas Wimmer, and Brian Min. 2010. "Why Do Ethnic Groups Rebel? New Data and Analysis." *World Politics* 62 (01): 87-119.
- Chenoweth, Erica, and Maria J. Stephan. 2017. *Why Civil Resistance Works*. New York: Columbia University Press.
- Chenoweth, Erica, and Jay Ulfelder. 2017. "Can Structural Conditions Explain the Onset of Nonviolent Uprisings?" *Journal of Conflict Resolution* 61 (2): 298-324.
- Chiba, Daina, and Kristian Skrede Gleditsch. 2017. "The Shape of Things to Come? Expanding the Inequality and Grievance Model for Civil War Forecasts with Event Data." *Journal of Peace Research* 54 (2): 275-97.
- Collier, Paul, and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56 (4): 563-95.
- Cranmer, Skyler J., and Bruce A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25 (2): 145-66.
- Cunningham, David E., Kristian Skrede Gleditsch, Belén González, Dragana Vidović, and Peter B. White. 2017. "Words and Deeds: From Incompatibilities to Outcomes in Anti-government Disputes." *Journal of Peace Research* 54 (4): 468-83.
- Dorff, Cassy, Max Gallop, and Shahryar Minhas. 2020. "Networks of Violence: Predicting Conflict in Nigeria." *Journal of Politics*. doi:10.1086/706459.
- Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (1): 75-90.
- Gleditsch, Kristian Skrede, and Michael D. Ward. 2013. "Forecasting Is Difficult, Especially about the Future: Using Contentious Issues to Forecast Interstate Disputes." *Journal of Peace Research* 50 (1): 17-31.
- Goldstein, Joshua S. 1992. "A Conflict-Cooperation Scale for WEIS Events Data." *Journal of Conflict Resolution* 36 (2): 369-85.
- Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert 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, 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): 991-1002.
- Guéhenno, Jean-Marie. 2016. "10 Conflicts to Watch in 2016." *Foreign Policy*, January 3. Accessed April 6, 2020. <http://foreignpolicy.com/2016/01/03/10-conflicts-to-watch-in-2016/>.
- Gurr, Ted Robert. 1970. *Why Men Rebel*. Princeton, NJ: Princeton University Press.
- Gurr, Ted Robert. 1993. "Why Minorities Rebel: A Global Analysis of Communal Mobilization and Conflict Since 1945." *International Political Science Review* 14:161-201.

- Hechter, Michael. 2001. *Containing Nationalism*. Oxford, UK: Oxford University Press.
- Hegre, Håvard, Joakim Karlsen, Håvard Moksleiv Nygård, Håvard Strand, and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010–2050." *International Studies Quarterly* 57 (2): 250–70.
- Hill, Daniel W., Jr., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108 (3): 661–87.
- Horowitz, Donald. 1985. *Ethnic Groups in Conflict*. Los Angeles: University of California Press.
- Kitschelt, Herbert. 1985. "New Social Movements in West Germany and the United States." *Political Power and Social Theory* 5:273–324.
- Leetaru, Kalev, and Philip A. Schrodt. 2014. "GDELT: Global Data on Events, Location and Tone, 1979–2012." Paper presented at the International Studies Association Annual Convention, San Francisco, CA. Accessed April 6, 2020. <http://data.gdelproject.org/documentation/ISA.2013.GDELT.pdf>.
- Lichbach, Mark. 1987. "Deterrence or Escalation? The Puzzle of Aggregate Studies of Repression and Dissent." *Journal of Conflict Resolution* 31 (2): 266–97.
- Lindemann, Stefan, and Andreas Wimmer. 2018. "Repression and Refuge: Why Only Some Politically Excluded Ethnic Groups Rebel." *Journal of Peace Research* 55 (3): 305–19.
- Martin, Will. 2016. "The 16 Countries with the Most Civil Unrest." *Business Insider*, August 3. Accessed April 6, 2020. <http://www.businessinsider.com/countries-with-the-highest-risk-of-civil-unrest-worldwide-2016-8>.
- McAdam, Doug, Sidney G. Tarrow, and Charles Tilly. 2001. *Dynamics of Contention*. Cambridge, MA: Cambridge University Press.
- Minhas, Shahryar, Peter D. Hoff, and Michael D. Ward. 2019. "Inferential Approaches for Network Analysis: AMEN for Latent Factor Models." *Political Analysis* 27 (2): 208–22.
- Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20 (3): 271–91.
- Moore, Barrington, Jr. 1966. *Social Origins of Dictatorship and Democracy*. Boston, MA: Beacon Press.
- Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. 2015. "Comparing Random Forest with Logistic Regression for Predicting Class-imbalanced Civil War Onset Data." *Political Analysis* 24 (1): 87–103.
- Nair, Kashmir, and Nicholas Sambanis. 2019. "Violence Exposure and Ethnic Identification: Evidence from Kashmir." *International Organization* 73:329–63.
- Petersen, Roger. 2002. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-century Eastern Europe*. Cambridge, UK: Cambridge University Press.
- Pierskalla, Jan Henryk. 2010. "Protest, Deterrence, and Escalation: The Strategic Calculus of Government Repression." *Journal of Conflict Resolution* 54:117–45.
- Sambanis, Nicholas. 2004. "What Is Civil War? Conceptual and Empirical Complexities of an Operational Definition." *Journal of Conflict Resolution* 48 (6): 814–58.
- Sambanis, Nicholas, and Branko Milanovic. 2015. "Explaining Regional Autonomy Differences in Decentralized Countries." *Comparative Political Studies* 47:1830–55.

- Saxton, Gregory D. 2005. "Repression, Grievances, Mobilization, and Rebellion: A New Test of Gurr's Model of Ethnopolitical Rebellion." *International Interactions* 31 (1): 87-116.
- Schrodt, Philip A., and David Van Brackle. 2012. "Automated Coding of Political Event Data." In *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian, 23-49. New York: Springer.
- Schrodt, Philip A., Mark Woodward, and Monty G. Marshall. 2011. "Forecasting Political Conflict in Asia Using Latent Dirichlet Allocation Models." Paper presented at the European Political Science Association Annual Meeting, Dublin. Accessed April 6, 2020. <http://parusanalytics.com/eventdata/papers.dir/Schrodt.EPSA.2011.final.pdf>.
- Shadmehr, Mehdi. 2014. "Mobilization, Repression, and Revolution: Grievances and Opportunities in Contentious Politics." *The Journal of Politics* 76 (3): 621-35.
- Siroky, David S. 2009. "Navigating Random Forests and Related Advances in Algorithmic Modeling." *Statistics Surveys* 3:147-63.
- Tarrow, Sidney. 1989. *Democracy and Disorder: Social Conflict, Political Protest, and Democracy in Italy, 1965-1975*. New York: Oxford University Press.
- Tikuisis, Peter, David Carment, and Yiagadeesen Samy. 2013. "Prediction of Intrastate Conflict Using State Structural Factors and Events Data." *Journal of Conflict Resolution* 57 (3): 410-44.
- Tilly, Charles. 1978. *From Mobilization to Revolution*. Reading, MA: Addison-Wesley.
- Walter, Barbara F. 2006. "Information, Uncertainty, and the Decision to Secede." *International Organization* 60 (1): 105-35.
- Ward, Michael D. 2017. "Do We Have Too Much Theory in International Relations or Do We Need Less? Waltz Was Wrong, Tetlock Was Right." *Oxford Research Encyclopedia of Politics*, July. Accessed April 6, 2020. <http://politics.oxfordre.com/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-301>.
- Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. "The Perils of Policy by P-value: Predicting Civil Conflicts." *Journal of Peace Research* 47 (4): 363-75.
- 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* 15 (4): 473-90.
- Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1): 206-18.
- Weidmann, Nils B., and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6): 883-901.
- Wimmer, Andreas, Lars-Erik Cederman, and Brian Min. 2009. "Ethnic Politics and Armed Conflict: A Configurational Analysis of a New Global Dataset." *American Sociological Review* 74 (2): 316-37.
- Young, Lauren. 2019. "The Psychology of State Repression." *American Political Science Review* 103:645-68.