Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes

Kristian Skrede Gleditsch, Department of Government, University of Essex & Peace Research Institute Oslo (PRIO)

Michael D Ward, Department of Political Science, Duke University

Abstract

Prediction is an important goal in the study of international conflict, but a large body of research has found that existing statistical models generally have disappointing predictive abilities. We show that most efforts build on models unlikely to be helpful for prediction. Many models essentially ignore the origins of conflict; Studies look either at invariant structural features believed to affect the opportunities of conflict, or factors that are believed to reduce the baseline risk of conflict, without attempting to identify the potential motivations and contentious issues over which conflicts typically arise. Researchers that have considered how contentious issues may motivate conflict and how these can be managed, using the Issues Correlates of War (ICOW) data, have not considered how these features may inform prediction. We assess the risk of dyadic interstate conflict based on the presence of specific contentious issues and conflict management events that may change the conflict potential of these contentious issues. We evaluate to what extent incorporating contentious issues and conflict management can help improve out-of-sample forecasting, as well as advance our understanding of conflict dynamics. Our results provide strong support for the idea that taking into account contentious issues can inform and improve out-of-sample forecasting.

Keywords: conflict, prediction, forecasting, issues, incompatibilities

Corresponding author: ksg@essex.ac.uk
Introduction

Research on interstate conflict is often justified by its potential to anticipate future conflicts between countries. The idea of forecasting has a long tradition in international relations scholarship, if not in actual applications, then at least in aspirations (see, among many others, Choucri & Robinson, 1978; Goldstone et al., 2010; O’Brien, 2010; Singer & Wallace, 1979). In this article, we propose a new approach to forecasting, based on contentious issues. Before turning to our proposed approach, we briefly review existing research on forecasting interstate conflict and what we see as some of the key limitations.

The most visible and best known efforts to predict political events tend to focus on either predicting outcomes in specific pre-identified situations, where the relevant actors are known and the time horizon often is short, or to predict broader forms of domestic instability rather than the risk of interstate conflict (see Bueno de Mesquita, 2002; Goldstone et al., 2010; O’Brien, 2010). Studies that do examine interstate conflict tend to be dyad specific, and often analyze variation in the outcomes of conflict (e.g., Schrodt, 1991) or continuous measures of hostile-conflictual interactions rather than conflicts as binary outcomes or larger episodes of individual conflict events (see Brandt, Freeman & Schrodt, 2011; Pevehouse & Goldstein, 1999). However, most empirical research on interstate conflict tends to consider binary conflict outcomes such as ‘dispute’ or ‘war’ on a global basis. If we believe that such global analyses are appropriate and that ‘disputes’ or ‘wars’ are meaningful theoretical categories, then we should be able to make meaningful global predictions about how likely these types of outcomes are across dyads, given the observed covariates deemed to be important.
There have been remarkably few efforts to generate global forecasts or risk profiles for interstate conflict. Moreover, the most prominent efforts to consider the predictive ability of models of interstate conflicts have based their research on models that were not actually proposed with forecasting in mind. A notable example here is Beck, King & Zeng (2000), who essentially adopt the so-called liberal peace model of Russett & Oneal (2001). Certainly, nothing akin to the Political Instability Task Force’s now annual projections (beginning with Gurr, Marshall & Khosla, 2000) exists for international conflicts. Perhaps not surprisingly, many observers are very skeptical of the ability of academic researchers to anticipate conflict between states, at least beyond very short time horizons.

Research in recent decades has seen a great deal of hypotheses generated to explain under what conditions militarized interstate conflict is more or less likely. This avenue of research has been primarily inspired by research on the so-called democratic peace, or the absence of conflict between democracies. Indeed, there are thousands of scholarly works mentioning the term militarized interstate dispute (MID), most of which use these data for some kind of empirical examination of a proposition about disputes. Yet, the evidence suggests that the ability of this body of work to forecast conflict out-of-sample is decidedly disappointing. Ward, Siverson & Cao (2007) found that most of the recent statistical studies of militarized interstate disputes in prominent political science and international relations journals were unable to predict the outbreak of a single dispute out-of-sample (see also Beck, King & Zeng, 2000).

Many researchers have sought to improve on the ability to forecast militarized interstate conflict by turning to alternative statistical methods. Beck, King & Zeng (2000), for example, find that neural networks perform marginally better than generalized linear regression models in forecasting conflict from the same input factors.¹ Changes in estimation
methods or statistical techniques *per se*, however, have at best led only to limited improvements in out-of-sample predictive ability.

Our argument is that simply identifying inappropriate methods as the key source of the problem in forecasting conflict may give us the wrong diagnosis and lead us down less productive avenues. A more fundamental problem is with models that provide a poor basis for forecasting by disregarding the motives for conflict to arise, or by only considering motives in a relatively superficial manner. Models that have been proposed for research on the democratic peace, notably the work of Russett & Oneal (2001), are primarily intended to examine whether certain characteristics of liberal institutions, such as democracy and trade, make conflict on average less likely relative to baseline risks of conflict. Although these approaches may be appropriate for testing the original propositions of interest, they essentially ignore the contentious issues that might cause states to resort to violence and instead treat these contentious issues as exogenous features, typically hidden inside a so-called ‘black box’ of the baseline risk of conflict. Our own initial foray into out-of-sample prediction for a state level model indicates that spatial information about other conflict events can help to improve forecasts (see Ward & Gleditsch, 2002). Although this allows predictions of conflict to be conditional on other observed events rather than treating each conflict as an independent observation, the approach still ignores the issues over which such conflicts may have arisen initially.

We believe that greater attention to the specific reasons for why conflicts may arise and the incompatibilities that may generate the use of violence can help improve our ability to forecast conflict. Although we recognize that different models may be appropriate to evaluate particular propositions and to forecast events, in our view the enterprise of prediction has great potential for winnowing bad ideas out of theories on the causes of conflict and avoid
the problem of retrospective biases in conventional hypothesis testing on the data used to develop the hypotheses in the first place (see Ward, Greenhill & Bakke, 2010).

In fairness, much of the existing work on the statistical modeling of conflict has bypassed motivation since it is genuinely difficult to establish what states fight over and what their possible motivation for fighting might be. Nevertheless, the fact that something is difficult to evaluate does not mean that simply ignoring it is the best course of action. Another tradition in research on conflict has sought to identify incompatibilities in terms of contentious issues, such as territorial or maritime claims (Diehl, 1992; Mansbach & Vasquez, 1981). Recent efforts to examine these proposition empirically have found considerable evidence that cases where such claims exist are more likely to see militarized activities (Hensel, 2001; Hensel et al., 2008; Hensel & Mitchell, 2010). Even so, at present this line of research has primarily engaged in testing hypotheses about whether coefficient estimates are significantly different from 0 or the in-sample post-diction of conflicts, and has not yet examined if information on contentious issues may be helpful for forecasting dyadic conflict out-of-sample. Here we explicitly consider whether taking into account information on contentious issues and conflict management can help improve on forecasting interstate conflict and our understanding of conflict dynamics.

Although we focus on statistical approaches to interstate conflict in this article, many of our arguments also apply to problems in traditional theories of conflict and qualitative approaches to prediction or anticipating political events (see Tetlock, 2005). Traditional theories of interstate conflict tend to focus on structural features presumed to influence the opportunities for conflict such as the distribution of power in the international system or relative balance of power (see, e.g., Waltz, 1979). These theories display little interest in the specific incompatibilities that may motivate the use of violence. However, structural factors rarely change rapidly, but violent conflict tends to be episodic, and hence cannot be
adequately explained merely by reference to permissive conditions (see Fearon, 1995). Likewise, our core argument applies to studies of civil war, which tend to emphasize opportunities for conflict rather than motivations for conflict (see Cederman, Weidmann & Gleditsch, 2011), and where evidence for the predictive ability of existing statistical efforts seems similarly disappointing (see Ward, Greenhill & Bakke, 2010). Many political and area study experts, typically using informal methods for deriving predictions, often have strong confidence in their ability to forecast events. However, the comprehensive series of studies by Tetlock (2005), who asked experts to rate a series of outcomes which could then be compared against the historical record, provide little support for the forecasting ability of political experts.

**Contentious issues and interstate conflict**

Formal models of conflict often begin with a representation of two parties bargaining over some issue or incompatibility, where violence arises if the parties are unable to agree upon a division or an agreement that both would prefer to fighting or a costly contest (see Boulding, 1963; Morrow, 1986). Much of the existing empirical research on interstate conflict relates opportunities for peace to other features believed to systematically influence the ability of the two parties to find an agreement. These features include, for example, how the cost of conflict is greater when states have extensive economic linkages that would be destroyed in the event of a conflict, or the putatively greater transparency induced by democratic institutions of governance and political competition. Researchers will often ‘control’ for other features believed to influence the opportunities and costs of conflict, such as the geographical distance between actors and their relative power. Nevertheless, the bulk of research on interstate conflict, in particular research on the democratic peace, pays little
attention to the incompatibilities themselves and whether they are empirically identifiable. Stated differently, even if we concede that the presence of two democratic states is a sufficient condition for an absence of conflict in a dyad, then knowing that at least one state in a dyad is not democratic does not provide much leverage for predicting conflict in the very large set of dyads where at least one state is an autocracy, unless we also know if these dyads have disagreements over issues that might cause conflict.

An alternative research tradition argues for a greater focus on substantive issues in the study of conflict and highlights a multitude of potential issues that could command attention in the international arena (see Diehl, 1992; Mansbach & Vasquez, 1981). Many of these issues might be considered dangerous in the sense that they have the potential to raise the risk of the onset of militarized violence among states. Territorial claims have often been highlighted as the single most common motivation for conflict, and some researchers have argued that the bulk of military conflicts have arisen over territorial issues, both historically as well as in recent years (see Holsti, 1991; Huth and Allee, 2002; Vasquez, 1993). Moreover, the notion of enduring rivalries highlight how certain dyads tend to see much higher frequencies of disputes, both because certain specific contentious issues may generate recurrent disputes and because contentious issues may lead to deteriorating relations in general and increase the risk of the use of force (Goertz & Diehl, 1993; Diehl, 1998).

Looking only at the issues that have been deemed important in observed conflicts does not give us an adequate understanding of how much potential these issues have to generate conflict, since this alone cannot tell us how often we see territorial claims that do not lead to militarized conflicts between states. We believe that it is possible, at least in principle, to catalogue ongoing incompatibilities or contentious issues and then observe whether we see systematically higher rates of subsequent interstate conflict in dyads where states have such
contentious issues. Existing research has certainly examined whether dyads with claims see more frequent disputes, but have not considered predictive ability in-sample or whether information on claims can help improve out of sample prediction.

The fact that two states have a contentious issue could imply a higher risk of conflict but obviously does not mean that violence is inevitable. If violent conflict is costly, then states also have incentives to avoid escalation to violence through some form of peaceful settlement. A territorial claim that has previously generated the use of force may cease to create problems in dyadic relations upon the successful conclusion of a treaty. As such, we should expect the conflict potential of an issue to vary systematically with conflict management efforts and their success, in terms of whether parties can agree on proposals. Moreover, different forms of incompatibilities or contentious issues may vary in their susceptibility to conflict management efforts and their effectiveness. For example, territorial issues may better lend themselves to peaceful management through bilateral settlement efforts than do disputes arising from transnational ethnic communities and conflict within countries, where the ability of states to guarantee agreements may be more limited since the conflicts involve actors partly outside their control (Schultz, 2010).

Data on contentious issues and interstate conflict

To what extent is it possible to identify interstate incompatibilities or contentious issues empirically with existing data? The so-called Issue Correlates of War (ICOW) project, directed by Hensel & Mitchell (2010), has developed data on various interstate claims, such as claims over disputed territory, that provide an opportunity for evaluating whether information on contentious issues can help improve conflict forecasts. We certainly do not believe that the claims catalogued by the ICOW project are the only potential
incompatibilities that may lead to the use of force. Nevertheless, the ICOW data allow for a proof-of-concept test to examine if taking explicit information on incompatibilities into account can improve on purely structural or ‘issue free’ forecasting models.

The ICOW project data attempt to identify ‘explicit evidence of contention involving official representatives of two or more nation-states over some type of issue’ (Hensel & Mitchell, 2010). The existing data consider claims about territory where two or more nation-states claim sovereignty over a specific piece of territory, the use or abuse of rivers, or maritime claims over the use of a specific maritime area. This helps bypass some of the problems in previous research that classifies issues post-hoc in observed conflicts. The ICOW project also collects data on peaceful settlement attempts, including bilateral negotiations and third party settlement efforts such as mediation and arbitration. The ICOW project also claims to collect their data explicitly without reference to militarized conflict over the claims.

In addition to the obvious advantages, there are also various limitations with the existing ICOW data that make them less than ideal for examining the potential for forecasting based on incompatibilities or contentious issues. At present, the full coverage of the data is limited to the Western Hemisphere. River claims data are also available for the Middle East and data for some claims have also been completed for Western Europe. However, we limit ourselves to the Western Hemisphere at the present, given the unclear boundaries between Western and Eastern Europe, and since we do not want to consider forecasting conflict in the Middle East from river claims alone. The existing data are only available until 2001. Still, we believe that these data can at least provide for an initial assessment of our arguments about issues, conflict management, and forecasting disputes.

**Empirical analysis**
Data

To assess forecasts of interstate conflicts out-of-sample, we first consider a Logit model estimated from the observed data on contentious claims, structural characteristics, and conflict history for the period 1900-1989. We then consider the implied predictions for the 1990-2001 period, based on the 1989 covariates.

We use this as a cross-validation experiment in which we validate our inferences from a model with a different set of data than the data we used to generate the inferences or model estimates. Although this is not a true out-of-sample prediction in the sense that we do not know the response in the forecast period at the time of the prediction, we stress that we do not actually use any information about conflict or the covariates over the forecast period in generating the predicted probabilities for the outcomes.

Our conflict data are instances of militarized interstate dispute, taken from the Maoz (2005) dyadic dispute data. We look only at the initial onset of disputes, excluding ongoing disputes. We construct a database of all annual undirected dyadic observations in the Correlates of War state list after 1900, which forms the common basis for both the MID and ICOW data. Next, we consider whether the ICOW project classifies the dyad as having a territorial, maritime, or river claim in a given year. To consider the potentially conflict dampening effects of settlement attempts, we also include an indicator of peaceful settlement attempts over the year by each specific type of issue. We return to the issue of alternative specifications and robustness tests after the discussion of the results for our core model.

To compare our results with existing work on MIDs, we also include a number of other characteristics that have been highlighted in the extant literature and that plausibly could be associated with both the presence of contentious issues and the risk of conflict. Existing work suggests that prior conflict history has a strong influence on the risk of disputes.
Accordingly, we incorporate a dichotomous variable for whether the dyad has ever previously been involved in a dispute, a count of the number of successive years that the dyad has been observed without a dispute (often called peaceyears in the literature on armed conflict), as well as the square and cube of this variable (see Beck, Katz & Tucker, 1998; Carter & Signorino, 2010). This is an important control variable, in the sense that we need to distinguish the effects of contentious issues, which may have led to disputes in the past, and the risk of recurrent disputes among previous antagonists.

Many scholars argue that democratic states are less likely to fight one another, and some have explicitly surmised that this may reflect democratic states having fewer contentious issues in the first place (see, e.g. Gartzke, 1998; Russett & Oneal, 2001). Based on the weakest link argument in democratic peace specifications, which holds that the least democratic state influences the prospects for avoiding conflict in a dyad, we consider the lowest of the two democracy scores in a dyad by using the Polity data (see Jaggers & Gurr, 1995). We set all observations with special values on the Polity scale due to transitions or interregnums to the lowest value (i.e., -10) and include estimates predicted from the Freedom House data for some small countries not included in the Polity data. To avoid problems of reverse causality or conflict causing changes in institutions, we look at the institutions in place at the start of the year (i.e., 1 January, or date of independence, if first year) for each observation.

We also include a measure of the relative power balance in a dyad. Power preponderance is often argued to deter conflict. Moreover, one might argue that the willingness to make or pursue a claim could reflect differences in power. We consider the relative power by looking at the ratio of the lowest to the highest Composite Indicator of National Capabilities (CINC) value for the two states in the dyad based on the COW
capabilities data (see Singer, 1987). Values closer to one on this ratio imply dyadic power balances closer to parity.

Finally, we also incorporate the natural log of the distance between two capital cities in a dyad since states are less likely to fight other states that are further away. This is an important control variable as some researchers have argued that the tendency for distant states to fight one another may reflect both lower opportunities as well as fewer contentious issues (Vasquez, 1995).

**Empirical results**

We consider three different model specifications for the risk of MIDs. Our first model, which we will refer to as the Issues based model, includes only the presence of territorial, maritime, and river claims; whether the outstanding claims see at least one settlement attempt for each type of claims, and the indicators of conflict history discussed above.

Our second model, which we will label the Structural characteristics model, considers various ‘conventional’ or ‘structural’ characteristics of countries or dyads believed to influence conflict, independently of the motivation or issues that may lead to conflict. This model includes the minimum democracy level of the two parties, the relative power balance, and distance in addition to the dyadic conflict history terms.

Lastly, our third model, which we will refer to as the Combined model, incorporates all the covariates in the first and second models.

**In-sample results**

Before turning to discuss the regression estimates, we briefly provide some summary statistics of dispute occurrences, conditional on claims and prior disputes. There are a total
of 239 dyad-years with dispute onset in our sample, out of a total of 24,792 dyad years. Out of these 239 disputes, 143 involved a dyad with a territorial claim. Other forms of claims are relatively less common among the dispute cases, but 175 of the 239 disputes take place in a dyad with one or more of the three forms of claims. Obviously, dyads with ongoing claims do not see militarized disputes all the time; only about 8.75% of the years with a territorial claims actually see an onset of a dispute. This share, however, is much higher than the share of onset years in dyads without territorial claims, which is less than one-half of one percent. There is some overlap between the various claim indicators, but not so much as to create problems of collinearity. We provide full descriptive statistics as well as a correlation matrix for all the regressors in the supplementary appendix.

Table I reports the estimated coefficients for the three models in the training sample. Our main interest here is to examine whether taking into account contentious issues and conflict management on contentious claims can improve on the ability to forecast conflict compared to conventional or structural models that disregard issues altogether, only consider whether a dyad has previously seen a dispute, or only consider the amount of time since previous conflicts. As such, the fit for the estimation sample is less relevant than the differences in the ability to predict to the subsequent time period, but we first briefly comment on the estimation sample results. The coefficient estimates reported in Table I yield strong support for the argument that we are more likely to see disputes in the presence of claims, as all the coefficients for contentious issues in the Issues based model have positive estimated coefficients. This finding holds even when we control for prior conflict history. As such, the results cannot be dismissed as an artifact of dyads with previous disputes simply being more likely to have new disputes.
These results, however, do not fully support the arguments about how conflict management should influence contentious issues. Although we find negative coefficients for peaceful settlement attempts for river based claims and maritime claims, only the river based claim settlement term is large enough to outweigh the positive influence of the presence of claims over rivers. The coefficient for peaceful settlement attempts for maritime claims is very close to 0. Finally, the coefficient for peaceful settlement attempts and territorial claims is large and positive, suggesting that peaceful settlement attempts over territorial claims imply a higher risk of conflict than just the claim itself. It is possible that this may pick up on the timing of settlement attempts during crises and situations with tension where conflict may be more likely and that the risk of conflict already is elevated, especially if settlement attempts are unsuccessful. We also experienced with a longer window of previous settlement attempts, to see if there was any evidence of a lower risk of conflict in the more extended aftermath of settlement attempts. However, we found little evidence for this. One possible interpretation of the relatively weak results for peaceful settlement attempts in decreasing the risk of disputes (especially for territorial claims) is that settlement attempts primarily affect the risk of conflict when successful, and if so, their effect on the risk of is largely captured in the termination of the claim itself.

With regards to the features emphasized in conventional structural models, we note that the evidence for democracy and dyadic power balances to discriminate with respect to the likelihood of disputes seems very limited in this particular sample. Neither of the two coefficients are significantly different from 0 by conventional standards in the Structural
characteristics model or the Combined model. Finally, it is clear that the log-likelihood for the Combined model is considerably smaller than Structural characteristics model, suggesting a better fit to the data. Moreover, when we compare these to the log-likelihood for theIssues based model constrained to the same sample as the other two models (-875.54), we find only a relatively small difference. This strongly suggests that the information on contentious claim does much more of the heavy lifting and discriminates conflict better than the traditional structure features, at least in this particular estimation sample.

Beyond the issue of the sign and significance of particular coefficients, does taking into consideration contentious issues improve our ability to account for disputes in-sample? MIDs are obviously relatively rare events, and thus there are likely to be few instances where the expected risk of a dispute would exceed the likelihood of a ‘no dispute’ outcome in a period as short as a single year. However, a good model with discriminating ability in-sample should systematically assign higher values to the observations where we see disputes and lower to those where we do not. Table II summarizes the results at a prediction threshold of $\hat{p} > 0.25$, restricted to the common set of observations across the three models. As can be seen, we find that the Structural characteristics model identifies 8 actual disputes, with 219 missed disputes and 29 false positives, while the Issues based model identifies 43 actual disputes at the same threshold, with 184 missed disputes and 72 false positives. The Combined model identifies 44 disputes but also generates 9 more false positives than the Issues based model. This again supports the notion that considering contentious claims can notably increase the ability to discriminate between the observations with and without disputes.

Table II in here
Any single threshold for considering an event as ‘predicted’ could be seen as an arbitrary description of the continuous distribution of the probabilities of conflict. A receiver operating characteristic or ROC curve provides a way to evaluate the discriminating ability of models without choosing a specific threshold. The ROC curve compares the share of correctly identified 0s and 1s for different prediction thresholds. A 50-50 guess would give points along the diagonal 45 degree. A perfect prediction with no false negatives and no false positives would fall in the upper left corner. As the ROC for a corresponding model is further above the y = x line, the better the discriminative ability of a model. Figure 1 shows that a model based on contentious issues consistently outperforms a conventional model in the relevant ranges for the in-sample estimates. The so-called Area Under the Curve (AUC) figure for the Issues based model is 0.91 and 0.92 for the Combined model, compared to 0.89 for the conventional Structural characteristics model. Although the Combined model does somewhat better than the Issues based model without the additional conventional structural features, the degree of improvement is relatively limited.

___________

Figure 1 in here

___________

Separation plots provide a different way of displaying these results (see Greenhill, Ward & Sacks, 2011). Separation plots for binary data arrange observations in increasing order by their predicted values from left to right and plots dark and light panels depending on whether observations are events or non-events, where a solid line indicates the actual
predicted probabilities for each observation. A good model would produce a plot where all the events (with dark panels) are clustered at the right-hand side of the plot and all the non-events (with light panels) at the left-hand side. The plot allows us to graphically ascertain the discriminative ability of a model: the greater scattering of dark panels reflects a weaker relationship between predicted probabilities and the observed outcomes. We present the in-sample plots for three models in Figure 2. Although the bulk of the actual disputes fall within the highest levels of the predicted probabilities for all the three models, we can see that the highest predicted probabilities are generally much larger for the Issues based model in panel (a) and the Combined model in panel (c). By contrast, the Structural characteristics model in panel (b) assigns lower predicted probabilities for the observations with the highest predicted risks, and we have a larger number of observations with actual disputes assigned low predicted probabilities.

Figure 2 in here

Out-of-sample results

We now turn to examine whether contentious issues and conflict management can provide any help in predicting disputes out-of-sample. There are only 19 dyads that have disputes in the Western Hemisphere over the period 1990-2001, out of a total of 595 dyads, and thus we have a relatively limited set of conflicts in the out-of-sample forecasting test set.

Simple descriptive statistics reveal that 15 out of these 19 MIDs involved dyads with
contentious claims active in 1989. Out of 15 territorial claims, 10 saw disputes over the period 1990-2001. In the same period, 2 out 3 river claims saw disputes, while among the maritime claims 12 out 26 — or slightly less than half — actually experiences disputes. By comparison, although most dyads that saw disputes had previously seen disputes, the vast majority of the 65 dyads that have previously experienced disputes did not see new conflicts, suggesting that conflict history alone greatly over-predicts the actual incidence of subsequent disputes.

We now turn to predictions from the models based on the 1989 observed covariates and assess how these compare to the actual record of disputes over the period 1990-2001. One way to evaluate the accuracy of the predicted probabilities over this interval is to consider the aggregated predicted probabilities from the annual disputes probabilities for 1989 over the test interval — i.e., \( \hat{p}^* = 1 - (1 - \hat{p})^{11} \), since the risk of conflict over the period overall is the complement of the probability that none of the 11 years in the period will see conflict. We consider conflict cases predicted if the aggregated predicted probabilities exceed 0.5 for the 11 year interval, i.e., \( \hat{p}^* > 0.5 \), so that the model suggests that conflict would be more likely than peace, and we then compare these predictions to the actual disputes. These results are summarized in Table III. In this instance, we find a somewhat similar pattern of results for the predicted and observed outcomes as in the in-sample data, but the better performance of the Issues based model is considerably greater out-of-sample. The Issues based model identifies 10 out of the 19 disputes with only 3 false positives. By contrast, the Structural characteristics model identifies 8 disputes but also has a substantially higher number of false positives. The Combined model identifies 11 of the 19 disputes but also has an additional false positive (i.e., 4), compared to the Issues based model. Although the number of disputes is relatively small in this sample, the differences between the models and the improvement
made by considering contentious issues are quite compelling, especially for such a simple model. The results also suggest that more complex models are not necessarily better and yield the well-known statistical result that more complex models, which run the risk of over-fitting in-sample, often do notably worse out-of-sample.

__________

Table III in here

__________

Given the low number of disputes in the prediction set, the out-of-sample ROCs look distinctively jagged (see Figure 3). It is generally the case, however, that the Issues based model performs better than a conventional Structural characteristics model. The area under the curve (AUC) values are 0.90 for the Issues based model and the Combined model and 0.88 for the Structural characteristics model. The separation plots for the three models shown in Figure 4 show that the bulk of events are concentrated at the highest levels of predicted probabilities. Still, we can note a smaller number of false negatives for the Issues based model in panels (a) and (c) and see more missed conflicts with lower predicted probabilities for the Structural characteristics model in panel (b).

__________

Figure 3 in here

__________

Beyond examining how the overall predicted probabilities compare with the actual outcomes, it is instructive to look at the record of actual disputes to see in what instances the
models discriminate relatively better, whether the models plausibly predict events for the ‘right’ reasons, and whether the disputes are something that these models realistically could anticipate as likely events. We review the narratives provided by the MID data for the most recent updates to evaluate if the actual disputes that take place are incidents that one should expect this type of model to identify as well as if other types or information could help inform forecasting. The lower right cell of Table IV lists the dyads that the Issues based model correctly predicts to experience disputes over the period and the specific dispute numbers. By comparison, the Structural characteristics model, without any information on claims, misses the actual disputes between USA and Canada, Nicaragua and Colombia, and Guyana and Suriname; all of which have outstanding territorial and maritime claims (as well as river claims in the case of the USA and Canada). The Structural characteristics model, however, does identify Haiti and the Dominican Republic as a high risk dyad, based on a recent dispute between the two, while this is not identified as a high risk dyad by the Issues based model since the two states have no outstanding claims. This dyadic dispute, however, was part of a multilateral dispute where the Dominican Republic joins a US-led naval blockade to restore the rule of former Haitian President Aristide after he was deposed in a coup, rather than a bilateral conflict initiated by the Dominican Republic. Thus, the apparently successful prediction of the Structural characteristics model is perhaps better seen as a fluke.

Figure 4 in here

The lower left cell of Table IV lists the actual disputes that took place over the period
that the Issues based model missed. As alluded to above, many of the missed dyadic dispute cases originate from MID 4016, or incidents generated by the efforts to restore the rule of Haitian President Aristide. Clearly, this is not a dispute related to any contentious territorial, maritime, or river claim, so it is unreasonable to expect that the Issues based model should identify this since this does not include any information on regime claims. Nevertheless, one could at least in principle envision collecting information on incompatibilities of this type. Indeed, Tures (2000) has classified regime claims empirically and includes Haiti as a case of this, although his data covers only up to 1992 for the Western Hemisphere and hence are less useful for our out-of-sample test. Belize and Guatemala is an interesting case, illustrating how territorial claims may be dormant for long periods, but become activated by new events and developments. Although Guatemala has territorial claims on Belize, and did not recognize the country until 1990, no actual militarized interstate disputes are recorded up through 1989. According to the MID narratives, however, the previously dormant situation was activated following the 1993 announcement that the UK would close its last military base and after the government suspended a territorial accord signed by its predecessor before ratification. Although these events happened after the end of our estimation period (i.e., after 1989), one could at least in principle imagine that forecasts could be improved conditionally on information about issue evolution or relevant events. Disputes 4153 between El Salvador and Nicaragua, 4261 between the USA and Venezuela, and 4149, 4154, and 4155 between Venezuela and Trinidad and Tobago are fishing disputes that follow long periods without MIDs, and hence are unlikely based on the model. There is no information available on MID 3550 in the MID narratives.
The upper right cell of Table IV lists the false positives, i.e., the cases where the model predicted disputes but no disputes actually took place. Although the US and Nicaragua had a long history of conflict over the Sandinista government in the 1980s, bilateral relations improved following the democratic elections in 1990 won by the opposition, which essentially removed the original contentious issue. Again, although our model does not include claims over regimes, one could imagine that such claims could be identified empirically, and information on elections and leadership or coalition turnover could be used to update forecasts. USA-Panama is another example where the political dynamics fundamentally changed following the 1989 US invasion, which was quickly followed by competitive elections and a new leadership. The third dyad, Argentina-Chile, is a good example of how contentious issues can be settled peacefully. Although Argentina and Chile have a long history of border claims and disputes, bilateral relations improved notably in the 1990s and most of the outstanding border issues were settled bilaterally. The last remaining border issue, the disputed El Laguna del Desierto, was decided in favor of Argentina by international arbitration in 1994. Again, more information on conflict management and the specific nature of settlement attempts could help provide information on declining risk of conflict to update forecasts.

**Discussion**

The ICOW data contain information on a number of other characteristics beyond the
claims and peaceful settlement attempts that we have considered here. As such, one might ask whether incorporating such features might help further inform out of sample forecasting. We have considered a number of different specifications. One might imagine that the risk that contentious claims should escalate to conflict would depend on the salience of the claim. The ICOW data include a 12 point salience index of a claim to the two parties in a dyad, based on features such as whether a territory involves strategic resources or a deemed to be a homeland. However, although we believe assessing issue salience in principle may be helpful, we found that although including this improved the fit in-sample, it actually decreased the predictive ability out-of-sample. As such, there is no support for issue salience improving predictive ability out-of-sample in this specific sample. This may be due to idiosyncracies in this particular sample. However, there are also more fundamental reasons that may account for the out-of-sample in-sample discrepancies. The so-called issue of overfitting alerts us to how models that have better in-sample fit can often have worse predictive ability out-of-sample. Moreover, this particular salience index scale builds on a number of assumptions regarding the weights on the different components that may not be optimal. In light of this, it may be helpful to do additional validation on how to assess issue salience, and we do not pursue the issue of weighting claims by salience here.

The example of the 1994 Argentina and Chile agreement cited above suggests that successful agreements could change the risk of conflict in dyad. We experimented by introducing post agreement windows to see if these were associated with decreases in risk, however, we found no support for this. Although we believe that agreements over contentious claims are likely to change long term risk, it appears as if most of the information is contained in the presence or absence of the claim itself. It is of course possible that events and progress on issues other than the claims contained in the ICOW data may lead to
structural shifts in the risk of conflict (consider, for example, visits by head of states such as Brandt’s Ostpolitik), but we are unable to consider this issue.

We have considered predictions over a 11 year window (1991-2001), using the information available at the start of the forecast window (i.e., 1990). This emulates an out-of-sample forecast, in the sense that out-of-sample forecasts cannot assume knowledge of the covariates and how they will evolve over time. However, one could imagine many other types of forecast. For example, one could imagine a one step ahead forecast, where the original predictions made at the outset could be revised as new information on the covariates becomes available. This may be helpful in certain circumstances, but entails prediction over a much shorter time horizon and is hence not comparable with an out of sample forecast over a long window. Moreover, one could imagine forecasts based on forecasting the values of the relevant covariates, such as the likelihood that claims will terminate based on the prior interaction, which could be informed by more information on the interaction of the parties and conflict management activities. We have focused on a simple application here, assuming that we do not know how covariates may move in the future. However, forecasts based on forecasts of the covariates is certainly an interesting avenue for future. Finally, it may be helpful to consider a range of forecasts under different scenarios, for example on the possible impact political shocks, such as democratization and shifts in political coalitions.

**Conclusion**

In the concluding section, we first summarize what we have shown in this article, provide more general comments on the broader implications for forecasting, and discuss other
promising avenues for future research. We have highlighted how many of the models that have been considered for predicting interstate disputes are unlikely to provide a suitable basis for forecasting interstate conflict since they afford no or only limited information on incompatibilities or the possible motives for conflict. We do not claim to provide new theoretical contributions to our insights on how claims may generate violence between states, but see the specific empirical analyses and results presented in this article primarily as a proof-of-concept, suggesting that forecasting can be improved by greater attention to theories of conflict. In our view, however, the results are very encouraging and strongly suggest that the prospects for forecasting may not be as bleak as commonly assumed.

There are many obvious limitations to the specific application presented here. We are dealing with a small number of cases and limited data in this simple experiment. Still, we believe that our results provide ample support for the idea that our ability to forecast conflict can be improved by greater attention to theories of conflict and emphasizing the contentious issues over which conflict may arise, and suggest promising avenues for future research. Classifying and tracking contentious issues and claims that have the potential to motivate violence can help us identify possible dyads that actually have a risk of dispute better than more structural opportunity models or purely historically driven trend extrapolations. Although the data here are quite limited in spatial and temporal scope, this is something that in principle could be done across a broad range of issues and types of conflict. For example, we have highlighted how issues such as regime claims in principle could be ex ante observable for predicting interstate disputes.

It would certainly be possible to improve on the dynamic element of the simple models that we have examined here. One might argue that contentious issues by themselves are, to some extent, close to enduring structure in forecasting over a short window. Although these
are ex ante observable, looking at claims alone does not tell us anything about the dynamics of conflict or the specific events in evolution of bilateral relations such as the timing of crisis and conflict management efforts. Although we were unable to find much evidence of clear declines in risk following conflict settlement attempts in the ICOW data, more data on the form and content of settlement attempts could, at least in principle, help us understand shifts in relations between actors and create better conditional forecasts. Combining an issues based approach to identifying risks between dyads with event data, or other information on the evolution of issues, may aid to further take into account the escalatory and de-escalatory moves that are likely to precipitate the use of violence, or its avoidance.

Finally, forecasting can play a potentially important role in theory evaluation. Most theory evaluation is currently done by testing hypotheses through statistical models for observed data, with statistical significance tests for coefficients as the most common criterion, but statistical significance tells us little about substantive significance or ‘oomph’ (see, for example, McCloskey & Ziliak, 1996). Instead, assessing the ability of models to generate large differences in predicted values is often a better device for theory evaluation. Moreover, out-of-sample forecasts are a more adequate form of theory testing than sample tests that are often conducted based on the events that inspired the theories in the first place. Lastly, researchers commonly ‘overfit’ by selecting an overly complex model that picks up on idiosyncrasies in the estimation sample but performs poorly out-of-sample. In this sense, a turn to out-of-sample forecasting bears promise of correcting some of the known problems in conventional hypothesis testing and providing better insights on understanding interstate conflict. At the same time we have tried to focus the out-of-sample testing in an area in which there are concrete hypotheses about what specific claims and issues may have caused the conflict to erupt.
References


Choucri, Nazli & Thomas W. Robinson, eds. 1978. *Forecasting in International Relations:*


**Replication data**

All results in this article were generated using R (2.13.2). The dataset used in the empirical analysis can be found at http://www.prio.no/jpr/datasets.

**Acknowledgements**

We thank Tobias Böhmelt, Lars-Erik Cederman, Govinda Clayton Cassy Dorff, Carsten Donnay, Karin Dyrstad, Paul Hensel, Nils Petter Gleditsch, Halvor Mehlum, Nils Metternich, Ulrich Pilster, Andrea Ruggeri and Phil Schrodt as well as participants at ETH Zurich and the 2012 Norwegian National Political Science Conference for helpful discussions and comments. The authors are listed in alphabetical order, equal authorship implied. Gleditsch would like to acknowledge the support of the Research Council of Norway (180441/V10). Ward thanks the Defense Advanced Research Projects Agency for supporting his work on predicting conflict events.
### Tables and figures

#### Table I. Risk of disputes, logit estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Issues based model</th>
<th>Structural characteristics model</th>
<th>Combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>SE($\hat{\beta}$)</td>
<td>$\hat{\beta}$</td>
</tr>
<tr>
<td>Intercept</td>
<td>-4.557</td>
<td>0.165</td>
<td>-2.706</td>
</tr>
<tr>
<td>Previous MID</td>
<td>1.732</td>
<td>0.184</td>
<td>1.607</td>
</tr>
<tr>
<td>$p_y$</td>
<td>0.004</td>
<td>0.001</td>
<td>0.005</td>
</tr>
<tr>
<td>$p_y^2$</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Territorial claim</td>
<td>1.163</td>
<td>0.189</td>
<td>1.041</td>
</tr>
<tr>
<td>River claim</td>
<td>0.804</td>
<td>0.308</td>
<td>0.649</td>
</tr>
<tr>
<td>Maritime claim</td>
<td>0.528</td>
<td>0.204</td>
<td>0.489</td>
</tr>
<tr>
<td>Peaceful settlement attempt - terr.</td>
<td>1.132</td>
<td>0.269</td>
<td>1.129</td>
</tr>
<tr>
<td>Peaceful settlement attempt - river</td>
<td>-1.464</td>
<td>0.544</td>
<td>-1.278</td>
</tr>
<tr>
<td>Peaceful settlement attempt - mar.</td>
<td>-0.034</td>
<td>0.364</td>
<td>-0.072</td>
</tr>
<tr>
<td>Lower democracy score</td>
<td></td>
<td></td>
<td>0.007</td>
</tr>
<tr>
<td>Balance ratio</td>
<td></td>
<td></td>
<td>-0.058</td>
</tr>
<tr>
<td>ln(distance)</td>
<td></td>
<td></td>
<td>-0.312</td>
</tr>
<tr>
<td>Observations</td>
<td>24,792</td>
<td>22,230</td>
<td>22,230</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-954.25</td>
<td>-925.42</td>
<td>-871.62</td>
</tr>
<tr>
<td>LR-$\chi^2$</td>
<td>785.99 (df=11)</td>
<td>843.63 (df=8)</td>
<td>951.24 (df=14)</td>
</tr>
</tbody>
</table>
Table II. Actual by predicted disputes, in-sample

<table>
<thead>
<tr>
<th></th>
<th>Issues based model</th>
<th>Structural characteristics model</th>
<th>Combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{p} &lt; 0.25 )</td>
<td>( \hat{p} &gt; 0.25 )</td>
<td>( \hat{p} &lt; 0.25 )</td>
</tr>
<tr>
<td>No dispute</td>
<td>21,931</td>
<td>72</td>
<td>21,974</td>
</tr>
<tr>
<td>Dispute</td>
<td>184</td>
<td>43</td>
<td>219</td>
</tr>
</tbody>
</table>

No dispute: 21,931 disputes, 184 predicted to be disputes, 72 actual disputes. Dispute: 21,974 disputes, 219 predicted to be disputes, 29 actual disputes. Combined model: 21,922 disputes, 183 predicted to be disputes, 81 actual disputes.
Table III. Actual by predicted disputes, out-of-sample

<table>
<thead>
<tr>
<th></th>
<th>Issues based model $\hat{p}^* &lt; 0.5$</th>
<th>Issues based model $\hat{p}^* &gt; 0.5$</th>
<th>Structural characteristics model $\hat{p}^* &lt; 0.5$</th>
<th>Structural characteristics model $\hat{p}^* &gt; 0.5$</th>
<th>Combined model $\hat{p}^* &lt; 0.5$</th>
<th>Combined model $\hat{p}^* &gt; 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No dispute</td>
<td>573</td>
<td>3</td>
<td>379</td>
<td>8</td>
<td>383</td>
<td>4</td>
</tr>
<tr>
<td>Dispute</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>8</td>
<td>11</td>
</tr>
</tbody>
</table>
Table IV. Actual versus predicted dispute dyads, out-of-sample

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\rho}^*&lt;0.5)</th>
<th>(\hat{\rho}^*&gt;0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No dispute</td>
<td>(573 dyads)</td>
<td>USA-Nicaragua</td>
</tr>
<tr>
<td></td>
<td></td>
<td>USA-Panama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chile-Argentina</td>
</tr>
<tr>
<td>Dispute</td>
<td>Canada-Haiti (4016)</td>
<td>USA-Canada (3972, 4183)</td>
</tr>
<tr>
<td></td>
<td>USA-Haiti (4016)</td>
<td>USA-Cuba (3950, 4196)</td>
</tr>
<tr>
<td></td>
<td>Haiti-Dominican Republic (4016)</td>
<td>Honduras-El Salvador (4010)</td>
</tr>
<tr>
<td></td>
<td>Haiti-Argentina (4016)</td>
<td>Honduras-Nicaragua (3988, 4011-2, 4140-3, 4171, 4259, 4327)</td>
</tr>
<tr>
<td></td>
<td>Belize-Guatemala (4014-5, 4150-2)</td>
<td>Nicaragua-Costa Rica (4146, 4147)</td>
</tr>
<tr>
<td></td>
<td>El Salvador-Nicaragua (4153)</td>
<td>Nicaragua-Colombia (4145, 4263)</td>
</tr>
<tr>
<td></td>
<td>Trinidad-Venezuela (4149, 4154-5)</td>
<td>Colombia-Venezuela (4009, 4172, 4219, 4262)</td>
</tr>
<tr>
<td></td>
<td>USA-Venezuela (4261)</td>
<td>Venezuela-Guyana (4260)</td>
</tr>
<tr>
<td></td>
<td>USA-Peru (3550)</td>
<td>Guyana-Suriname (4156, 4157)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ecuador-Peru (3987, 4013, 4143-4, 4189)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses indicate MID dispute numbers.
Figure 1. In-sample ROC curve
Figure 2: In-sample separation plots
Figure 3. Out-of-sample ROC curve
Figure 4. Out-of-sample Separation Plots
Biographical statements

KRISTIAN SKREDE GLEDITSCH, b. 1971, PhD in Political Science (University of Colorado, Boulder, 1999); Professor, Department of Government, University of Essex (2005– ); Research Associate, Centre for the Study of Civil War, PRIO (2003– ). Author of All International Politics is Local: The Diffusion of Conflict, Integration, and Democratization (University of Michigan Press, 2002) and recent articles in American Political Science Review, International Organization, Journal of conflict Resolution, and World Politics.

MICHAEL D WARD, b. 1948, PhD in Political Science (Northwestern University, 1977); Professor, Department of Political Science, Duke University (2009– ). Current main interest: dynamic analysis of international networks.
Endnotes

1 However, even this marginal improvement is disputed (see in particular de Marchi, Gelpi & Grynaviski, 2004).

2 The ICOW data also consider militarized disputes as violent settlement attempts, but we only consider the peaceful settlement attempts in our analyses here.

3 Since MIDs are rare events, one might ask whether the conventional Logit MLE estimates may display rare events bias. We have re-estimated all the models using the King & Zeng (2001) rare events Logit estimator. However, we note that King & Zeng state that this primarily is an issue when the N is low (less than 200), and in this case the results are near identical, and most of the difference in the coefficient estimates are extremely small. We report these results in the supplementary appendix.