Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict*

Andreas Beger† Richard K. Morgan‡ Michael D. Ward§

22 September 2020

Abstract

We examine the research protocols in Blair and Sambanis (nd). We find that there are several important mistakes and research decisions that determine their results. Fixing these mistakes results in a reversal of their claim that theory based models of escalation are better at predicting onsets of civil war than other kinds of models. Their model is s outperformed by several of the ad hoc, putatively non-theoretical models they devise and examine.

1 Introduction

Blair and Sambanis (n.d., hereafter B&S) argue that theory is essential for creating models that have high accuracy in forecasting civil war onset. Indeed they assert that with such theory forecasting is more accurate than has previously been possible. We re-examine the empirical basis for the claims made in support of it. We find that these claims to be unsupported and evidence presented for them to be incorrect. Their theory-based escalation model does not do better than the alternatives that they examine. It does worse. The reason for this is that they have made several mistakes in their research procedure. Further, the performance results they report are based on smoothed performance curves, not the empirical, unsmoothed curves. This provides misleading results. In addition, two of the structural alternatives to their basic escalation model were incorrectly implemented. We also found that the scoring of their forecasts for the first half of 2016 was incorrectly performed using civil war incidence, not onset. In what follows, we show the impact of these mistakes on the conclusions.

B&S claim (page 3) to show that a model informed by procedural theories of escalation and de-escalation can predict the onset of civil wars "remarkably accurately". Indeed, B&S argue that this theoretical model outperforms four other "more mechanical" alternatives.They also claim that the integration of structure with process is better than alternatives over short forecasting windows. Third, they preregistered the list of thirty countries that have the highest risk of civil war onset. They claim that such prospective predictions are rare in the literature when, in fact, they have been routine for many years with several prominent projects.[1] B&S claim to be unique in assessing these forecasts. A qualitative analysis of their predictions allows them to conclude that their model is robust. We will return to their analysis later, after correcting the procedural mistakes we found in their research process.

Before proceeding, we quote B&S (page 24):

> Our theoretically driven model generates accurate forecasts, with base specification AUCs of 0.82 and 0.85 over one- and six-month windows, respectively, and AUCs as high as 0.92 in

*John Ahlquist, Cassy L. Dorff, and Shahryar Minhas both provided helpful feedback on this project. We are especially grateful to Paul Huth for his comments and guidance. All the code and several additional analyses can be found at our replication archive at https://github.com/andybega/Blair-Sambanis-replication.

†Predictive Heuristics, adbeger@gmail.com.

‡Varieties of Democracy Institute, University of Gothenburg, rick.morgan2@gmail.com.

§Predictive Heuristics, Duke University, University of Washington, michael.don.ward@gmail.com. Corresponding author.

[1] PITF (e.g. Goldstone et al. 2010; Beger and Ward 2017), the W-ICEWS project (Ward et al. 2013), VIEWS (Hegre et al. 2019), and others, e.g. Beger, Morgan, and Maxwell (2020).

other specifications. Our model also consistently and sometimes dramatically outperforms the alternatives we test. [...] 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.

2 Summary of Blair and Sambanis (2020)

B&S' analysis is based on the use of non-linear, non-deterministic machine learning models, and specifically random forests, one of which has a specification they argue is theory-based. Several others with more generic sets of covariates are also developed and compared. Notably, the analysis is at the country-month or country-(6 month) levels and relies in large part on indicators derived from the ICEWS event data. In short, they uphold their assumption that theory-guided empirical research produces better conflict predictions than machine learning inspired efforts that are necessarily ad hoc combinations of available variables. They arrive at this conclusion by examining the problem of predicting civil war onset. They report that a parsimonious model using a small number of covariates derived from escalation theories of conflict can forecast civil war onset better than alternative specifications based on generic covariates not specifically informed by theory, including a *kitchen sink* model with more than 1,000 covariates.

B&S specifically examine three questions:

1. How does the theoretically-driven escalation model compare in forecast performance to alternative models not informed specifically by civil war onset theories?
2. Does annual, structural information from the PITF instability forecasting model add to the escalation model's monthly and 6-month predictions?
3. How accurate were predictions using the escalation model for the first half of 2016?

To assess the first two questions, B&S use ICEWS data covering all major countries from 2001 to 2015. Two versions of the dataset are used, one at the country-month level, the other aggregated to 6-month half-years. The main outcome variable is civil war onset, measured using Sambanis' civil war dataset.

Both the first and second questions above rely on comparing their escalation model to various alternative models. The same procedure is used in both cases:

1. Split the training data into training (2001 - 2007) and test (2008 - 2015) data.
2. Estimate the escalation and other competing models with the training data.
3. Create out-of-sample (OOS) predictions from each model using the test set.
4. Calculate AUC-ROC measures for each set of OOS predictions.

To examine the first question, B&S compare the test set of the escalation model to four alternative models. The independent variables for the first set of analysis reported in Table 1 in the paper are all derived from the ICEWS event data, using domestic events between actors within a country. The models are:

- Escalation: a set of ten indicators, drawn from a theoretical escalation model, for interactions between the government on one side and opposition or rebel actors on the other.
- Quad: ICEWS quad counts, i.e. material conflict, material cooperation, verbal conflict, verbal cooperation, for interactions between the government and opposition or rebels. These are directed, thus making for four directed dyads, which with four quad categories make sixteen covariates.
- Goldstein: -10 (conflictual) to 10 (cooperative) scores derived from the ICEWS data for the same four directed dyads, for a total of four covariates.
- CAMEO: counts for all CAMEO event codes over the four actor dyads, totaling $1,160$ covariates, which are mostly zero for any country in any month.
- Average: unweighted average of the predictions from the four models briefly described above.

The corresponding results for each question are shown in B&S Tables 1 and 2, which we examine further below. We accurately replicate their Tables 1 and 2. The results in Table 1, aside from the core base specification results, include eight additional robustness tests for both the 1-month and 6-month versions.

These robustness checks vary either (1) random forecast hyperparameter values or (2) the year used to split the train/test data, or (3) alternative codings of the civil war onset dependent variable.

The second question, whether structural variables add to the escalation model, is assessed by comparing the original escalation model to four alternatives that incorporate annual, structural variables that are used in the PITF instability forecasting model:

- Escalation Only: the original basic escalation model with only ICEWS predictors.
- With PITF Predictors: a random forest that also adds the PITF annual, structural variables.
- Weighted by PITF: escalation model predictions weighted using the PITF instability model predictions.
- PITF Split Population: the training data are split into high and low risk portions based on the PITF instability model predictions, two separate escalation random forests are trained on the splits, then re-combined into a single random forest that is used to create the test set predictions.
- PITF Only: a random forest model based only on the annual, structural PITF model predictors.

The corresponding results are shown in B&S Table 2.

Finally, B&S used their escalation model to create forecasts for the first half of 2016. In their third analysis, they score the forecasts accuracy using subsequent civil war onset data. This is summarized in B&S Table 3.

3 Implementation Issues in B&S

While replicating and analyzing B&S's results, we found several issues worthy of further discussion and investigation. These are a) the use of smoothed ROC curves to draw conclusions about which model is best, b) incorrect implementations of the weighted by PITF and PITF analog split-population models, c) inconsistent test sets for the models examined, and d) incorrect scoring of the 2016 forecasts.[2]

We believe that these research decisions and issues lead B&S to incorrect conclusions. The escalation model is not the best; it actually performs worse than the atheoretical, ktchen sink model with over 1000 variables. We discuss these five issues below. We defer a complete analysis that corrects all these issues until later, as there are many possible permutations of a seriatim unfolding.

3.1 Smoothed ROC Curves

The most consequential issue that we found is that all AUC-ROC values reported in B&S Tables 1 and 2 are calculated using smoothed ROC curves, not the original, empirical ROC curves. The data in rare events problems like this one restrict the number of true positive rate values and lead to non-continuous ROC curves; B&S refer to smoothing only in the context of the ROC curves in their Figure 1, justified as easing interpretation. However, smoothing is also used in all AUC-ROC values they report. There is no difficulty to be overcome in using non-smoothed plots. Indeed, such plots are standard practice in conflict research and forecasting, both for visualization and when calculating AUCs. We examined all 63 references in B&S and also 15 articles in two randomly picked issues of JCR for 2020 (numbers 1 and 9), in order to survey the use of ROC curves. Of the 37 articles that used a binary outcome, 19 included either ROC curve figures or a table with AUC-ROC values. We found that 90%–17 of 19–clearly used empirical ROC curves, and 2 (10%) where we could not establish whether empirical or smoothed curves were used.[3] The survey results are not surprising since the original ideal as well as contemporaneous documentation suggest this is the gold standard. What is surprising is that anyone would choose the smoothed ROC implementation.

The next three issues we encountered all concern information in B&S Table 2, which is the key table in which B&S report AUC-ROC values to demonstrate the "Escalation" models superiority.

[2]We also note there are additional concerns arising from the question of how the random forest models were tuned by B&S, especially given the way they are used is unorthodox. We did not further investigate the latter issue as it is rendered somewhat moot by the changes in results after addressing the preceding issues.

[3]One paper reports AUC-ROC values but no figure with ROC curves, and we could not find publicly posted replication code. The other is an ambiguous case where ROC curves are manually constructed in a way that might be considered "smoothing"–and one of the two figures certainly looks smoothed–but it's not a parametric smoothing method like used in B&S. The online replication archive has more details at https://github.com/andybega/Blair-Sambanis-replication/tree/master/journal-survey. Both of the ambiguous cases include Blair as coauthor.

3.2 Incorrect "Weighted by PITF" Implementation

The "Weighted by PITF" model is described as follows in B&S, page 19:

> The [Weighted by PITF model] 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.

We believe that B&S intend that the escalation model's predictions for the test set are weighted by the PITF model predictions for the test set. However, they actually weight the **test** set predictions using the PITF model predictions for the **training** set.

As a result: (1) in the test set, all countries with names that alphabetically start after "Tongo" are dropped from the data, including Turkey, Ukraine, Yemen, etc.; and (2) the cases whose PITF training data predictions are combined with Escalation model test predictions in the weighting are nonsensical. For example, the "Weighted by PITF" model predictions for Syria from 2008 to 2015 are the Escalation test predictions for that time period weighted, respectively, by the 2002 to 2007 PITF predictions for South Africa and the 2001 to 2002 PITF predictions for Zambia.

This appears to be an easily corrected coding error on the part of B&S.

3.3 Incorrect "PITF Split Population" Implementation

The "PITF Split Population" model also appears to be incorrectly implemented, owing to a coding error. B&S describe it on page 20:

> The final approach is a random forest analog to split-population modeling. 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 [...].

This description suggests that B&S intended to run two separate random forest models, one each on the low- and high-risk training data splits. The replication code does indeed run two separate random forecasts, but they both utilize the *exact same training data*, which consist of the full training data from all other models. In short, rather than using only data for high-risk countries to train their split-sample model for high-risk cases, the data they use captures both high- and low-risk countries, similarly for the training set for their low-risk split. The model specifications are also identical; i.e., they use the same x variables and the same random forest hyper-parameter settings. The *only* difference in the models as they are implemented in the B&S replication code is due to the non-deterministic nature of the random forest model itself. If we ran both with the same random seed, they would be identical in every respect, producing two forests of identical decision trees and, thus, identical predictions.

The implementation error aside, this split-population analog model is quite odd and does not replicate the idea behind split-population modeling (Chiba, Metternich, and Ward 2015; Beger et al. 2017). Although the two RFs are trained on separate data (in our updated, fixed replication), the process of combining them actually creates a new, larger RF using both component model's underlying decision trees. Thus, while all RF models throughout (except for one of the robustness checks) are trained with 100,000 decision trees (`ntree`), the new RF model after combination does have 200,000 decision trees. Furthermore, the PITF model predictions do not impact how the combined RF model predicts at all, not even through a binary low-/high-risk split. The split-population PITF RF model is practically speaking simply another escalation model trained with N=200,000 instead of N=100,000 trees and an extra odd randomization step added to the already existing RF randomization facilities (row and column sampling for each decision tree). This does not adequately implement their split-sample modeling strategy.

3.4 Inconsistent test set N for the models in Table 2

Further, the AUC-ROC values reported in the original B&S Table 2 are calculated on slightly different numbers of underlying test set cases (see Table 2). ROC calculations for a set of predictions can only be done on the set of cases for which both non-missing predictions and non-missing outcomes are available. Those sets differ across models (columns) for each row in B&S Table 2.

Thus a difference in AUC-ROC values for two models could be because they were calculated on different sets of underlying cases, not because the models are systemically performing at a different level. In other words, the results for different models in B&S Table 2 are not comparable to one another, and conclusions drawn from such comparison are potentially incorrect.

3.5 Incorrect scoring of the 2016 forecasts

B&S present a confusion matrix to score their 2016-H1 forecasts in Table 4. Although the forecasts are for the probability of civil war *onset*, in the replication code, they are actually scored using *incidence* of civil war. By definition the incidence of civil war is much more common than onset, which is only coded for the starting year of a conflict. When forecasting rare events, it is common to have many false positives, i.e. instances where we predict onset but no onsets occur. *Ex ante*, using incidence rather than onset to score onset forecasts should decrease the number of false positives and correspondingly increase the number of true positives (good predictions) and false negatives (missed onsets).

4 Results of the updated analysis

We now turn to an examination that addresses and fixes the five issues discussed above.[4] The main results of the original analysis consist of the comparison of the Escalation to other ICEWS models (our Table 1, B&S Table 1), and a comparison of the Escalation model to models that add structural variables/information (our Table 3, B&S Table 2). We will review the substantive implications of our updated analysis below, but the bottom line is that these changes turn B&S's conclusions on their heads.

4.1 Smoothed ROC curves and AUC calculations

A reference to smoothing is made in a single sentence in B&S (p. 12):

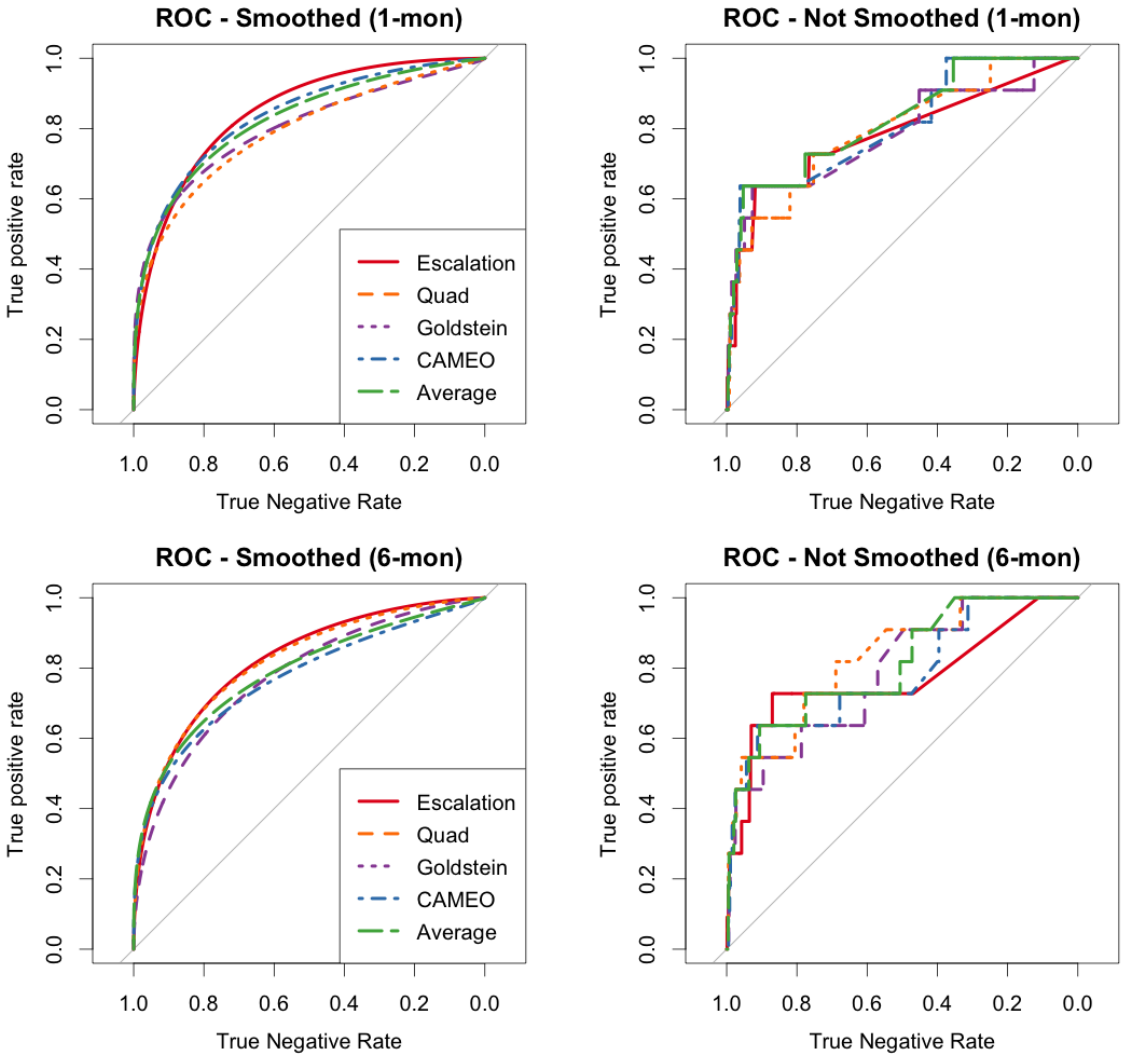
> Figure 1 displays the corresponding ROC curves, smoothed for ease of interpretation.

This suggest that using smoothed ROC curves will have no substantive impact on the results; however, in this smoothing determines the comparative results. Their statement also implies that the ROC curves were only smoothed in the referenced Figure 1. However, this is not the case. All AUC-ROC calculations throughout the replication code use an option to smooth the ROC curves *prior* to AUC calculation. ROC curves are constructed from the false positive and true positive rates as one moves through a set of ranked predictions, and as a result they appear step-like. Figure 1 shows our replication of both the estimated smoothed ROC curves from the B&S report (left-hand side) and the actual empirical ROC curves (right-hand side). The standard method is to compute the area under the curve (AUC) statistic on the original, empirical ROC curves that are shown on the right.

The specific predictions also include groups of cases with identical predicted probabilities, which accounts for the unusual diagonal lines seen in the panels on the right. In any case, with a sparse outcome like civil war onsets, the true positive rate on the y-axis only changes when the prediction for an observed positive case is reached. For these ROC curves, and for that matter in the fundamental train/test split used for 12 of the 18 rows/models in B&S Table 1, there are only 11 civil war onset cases in the test set. Thus, the ROC curves here are very step-like, with only 12 (11 positive cases plus 1 for TPR = 0) distinct y coordinates. Notice also that the smoothing averages the left-most almost straight line with the right-most almost straight line

[4] The code for all of our analysis undertaken for this effort may be found at https://github.com/andybega/Blair-Sambanis-replication.

Figure 1: Replication of B&S Figure 1 with both smooth and non-smoothed ROC curves



in a monotonic way. Thus, smoothing changes the ordering of alternatives. It is not a cosmetic change as suggested.

What impact did the ROC smoothing have overall on the evaluated performance of the Escalation model relative to other ICEWS models (our Table 1, B&S Table 1) and structural extensions (our Table 3, B&S Table 2)? Figure 2 shows the changes in AUC-ROC values had we used smoothed ROC curves to calculate the AUC-ROC values. Each point corresponds to the change in AUC-ROC values for one of the models in the cells in Tables 1 and 3 (the colors match those in Figure 1). The vertical bar in each plot marks the average effect of smoothing on AUC-ROC. For all alternative models, from Quad to "PITF Only", smoothing sometimes hurts and sometimes benefits, but the overall impact is negligible on average. The Escalation model *always* benefits from smoothing, with an average improvement on the order of 0.05. This is sufficient to push the Escalation model ahead of the alternative models, and accounts for the incorrect result B&S report, namely that the escalation model is generally superior.

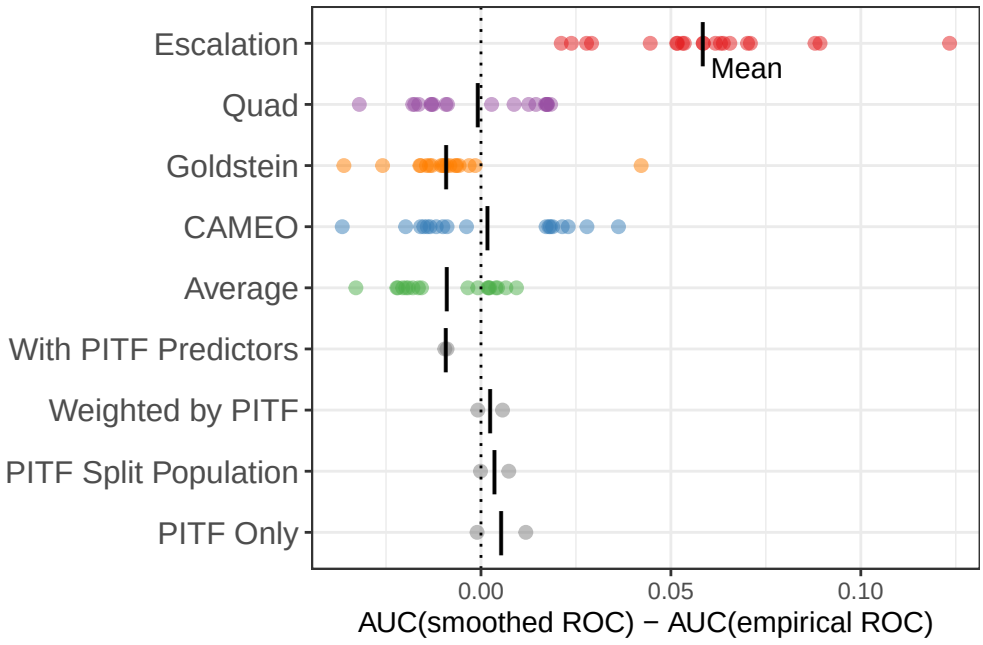


Figure 2: Gain from using smoothed ROC to calculate AUC, for each model reported in Tables 3 and 4 (B&S Tables 1 and 2). The Escalation model is the only model which consistently benefits from using smoothed ROC curves, and this advantage accounts for its apparent superiority to the other models.

In sum, it is not only the case that using smoothed ROC curves alters the results, but also that the use of smoothed ROC curves to calculate AUC-ROC values benefits **only** the escalation model. It does so consistently and by a considerable margin. All eight alternative models reported in B&S Tables 1 and 2, on average, do not gain when using smoothed ROC curves to calculate AUC. Four of the eight show a small positive bias from smoothing, while the other four display a small negative bias. The magnitudes of these are not close to the level of positive bias found in the escalation model.

4.2 Is the Escalation model superior to the alternative ICEWS models?

Table 1 (B&S Table 1) shows the comparison of the Escalation model to other alternative models based on ICEWS event data indicators. The AUC-ROC values are based on original, non-smoothed ROC curves. Both the Average and CAMEO models outperform the Escalation model in almost all instances (see also Figure A1). The Goldstein model generally outperforms the Escalation model in the 6-month version. The Quad model appears to be roughly on par with the Escalation model. Thus, the original B&S conclusion that the Escalation model is superior to the alternative models is entirely conditional on the non-standard use of smoothed ROC curves.

Table 1: Comparison of the escalation model to alternative ICEWS model using test set AUC-ROC, *without* smoothed ROC curves (Replication of B&S Table 1)

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.78	0.78	0.80	0.81	0.82
Terminal nodes	0.79	0.78	0.79	0.81	0.82
Sample size	0.79	0.80	0.74	0.82	0.84
Trees per forest	0.78	0.78	0.79	0.81	0.82
Training/test sets 1	0.77	0.76	0.77	0.79	0.80
Training/test sets 2	0.75	0.77	0.74	0.76	0.78
Training/test sets 3	0.71	0.79	0.69	0.72	0.74
Coding of DV 1	0.80	0.80	0.80	0.82	0.83
Coding of DV 2	0.80	0.82	0.77	0.83	0.81
Six-month forecasts					
Base specification	0.77	0.79	0.83	0.78	0.81
Terminal nodes	0.77	0.78	0.82	0.78	0.79
Sample size	0.77	0.77	0.80	0.80	0.82
Trees per forest	0.77	0.79	0.83	0.79	0.81
Training/test sets 1	0.75	0.79	0.82	0.77	0.80
Training/test sets 2	0.70	0.75	0.78	0.75	0.77
Training/test sets 3	0.85	0.72	0.84	0.72	0.81
Coding of DV 1	0.78	0.80	0.83	0.79	0.82
Coding of DV 2	0.80	0.78	0.84	0.80	0.81

The key difference between Table 1 and B&S Table 1 is whether the underlying ROC curves were original or smoothed; the test set N and coding issues did not affect this set of results.[5] B&S's original results and interpretation regarding the superiority of the B&S model over alternative ICEWS models, including the 1,000+ covariate CAMEO model, are thus entirely conditional on the use of smoothed AUC-ROC. Even the Quad model is generally as good as the escalation model in all implementations and is frequently quite a bit better.

Aside from the visual impact of ROC smoothing, the underlying motivation and methods involved in smoothing are broader and involve parametric estimation of population ROC curves (Hanley 1988, 2014). The main application is in the evaluation of medical tests (e.g. diagnostic imaging) with early methods dating back several decades and ongoing development of both parametric and non-parametric smoothing methods (e.g. Zou, Hall, and Shapiro 1997; Pulit 2016). It is not established which methods are preferable in a given application (Hanley 2014). Notably, the context in which ROC smoothing is used differs in important aspects from typical conflict research applications. Conflict data are closer to a census of the population and with well-known dependencies like spatial and temporal correlation, rather than a random sample that is approximately independently and identically distributed. It is thus neither clear that smoothing is justified nor valid. We also note that smoothed ROC plots are not widely used outside of the narrow context mentioned above.

4.3 Using Inconsistent Ns

The top portion of Table 2 shows the number of valid test set predictions that can be scored for each model in the original B&S Table 2. For the 1-month data version, the number of predictions differs by up to 500

[5]We should note that in our replication we consciously decided to not set RNG seeds, even though the random forest models are non-deterministic and vary slightly from run to run. As we changed the replication code to allow running in parallel, we cannot exactly reproduce the B&S results even with the same RNG seed. More importantly, the substantive interpretation of results should not depend on a specific RNG seed. We found that the escalation model's AUC-ROC values generally fluctuate no more than 0.01 (see https://github.com/andybega/Blair-Sambanis-replication/blob/master/rep_nosmooth/variance.md), and thus are confident that the patterns we see are robust to the initial RNG state.

Table 2: Number of valid test predictions for the escalation and structural alterantives models (B&S Table 2)

Horizon	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
Original model-specific cases					
1 month	13748	13155	13461	13748	13510
6 months	2366	2264	2317	2366	2265
Cases adjusted to common subset					
1 month	13062	13062	13062	13062	13062
6 months	2250	2250	2250	2250	2250

Table 3: Comparison of escalation model to structural extensions, using test set AUC-ROC (Replication of B&S Table 2)

Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts				
0.75	0.78	0.78	0.85	0.75
Six-month forecasts				
0.76	0.86	0.80	0.77	0.74

Note: Differences from the original B&S Table 2: (1) AUC-ROC values are computed on the common subset of cases, meaning that N is equal in each row; (2) AUC-ROC values are computed using original, non-smoothed ROC curves.

cases, and in the 6-month data version by around 100 cases. These numbers appear small enough to be not important. However, the already small number of positive cases also is affected substantially: "With PITF Predictors" and "PITF Only" models lose two or one (respectively) of 11 positive cases in the 1- and 6-month data versions.

Generally, these comparisons require the same cases. Therefore, we use predictions for a common joint subset of non-missing predictions. The resulting numbers of cases for each model are shown in the bottom portion of the table. Since the sets of incomplete cases in the original version above do not entirely overlap themselves, the common subset for all models is slightly smaller than the smallest N in the top portion of the table.

4.4 Do Structural Variables Add to the Escalation Model's Predictive Power?

Table 3 shows our replication of B&S Table 2 with (1) regular, not smoothed, AUC-ROC, (2) fixed "Weighted by PITF" and "PITF Split-Population" models, and (3) AUC-ROC values computed on the common, joint subset of tests cases for which all models have non-missing predictions. Table A2 in the appendix shows AUC-ROC values for both smoothed and non-smoothed versions, and both the original, model-varying test cases sets and our common joint subset.

B&S interpret their results as follows (page 20) and we comment on the conclusions seratim:[6]

1. B&S: "Of the approaches we test, the split-population analog is most promising . . . "

The PITF split-population analog still performs well, but the simple Escalation + PITF predictors model arguably performs better still.

2. B&S: "Adding PITF predictors improves the performance of the escalation model over six-month windows but diminishes it over one-month windows . . . "

[6]We list the "Overall, . . . " interpretation out-of-order for clarity.

Adding PITF predictors actually improves performance in both cases; the "With PITF Predictions" model strictly dominates the "Escalation Only" model by quite a margin in the 6-month data version.

3. B&S: "The weighted model performs very poorly regardless . . . "

The weighted model performs roughly on par with the Escalation Only model. One finding that remains is that the "PITF Only" model is outperformed by the "Escalation Only" model. As the former only uses annual inputs, but the data at hand are the 1-month or 6-months level, this is neither surprising, nor especially noteworthy.

4. B&S: "Overall, our results suggest that while measures of structural risk may improve predictive performance, the value they add is marginal and inconsistent. [. . .] Incorporating PITF thus significantly reduces or only slightly improves the performance of the escalation model, regardless of the approach we take. [. . .]"

The most straightforward method of incorporating the annual structural PITF variables—adding them to the predictors of the Escalation RF model—strictly outperforms the Escalation Only model. Note that the two other combination models considered are both non-standard and that the "PITF Split Population" model does not incorporate structural information at all, yet they also both do well. We thus conclude that adding structural variables improves predictive performance.

4.5 How Accurate Were the 2016 Forecasts?

B&S made forecasts for the first half of 2016 (2016-H1), and assess the accuracy of these forecasts. The forecasts themselves are probabilistic, i.e. range from 0 to 1, but for the purpose of scoring the forecasts they consider the 30 highest predictions to have been positive forecasts ("1") and the remaining forecasts to have been for no onset ("0"). They then construct confusion matrices for the forecasts, in their Table 4 and which we replicate here in the top half of our Table 4. They use two alternative codings of the outcomes for 2016-H1, one of which they call "Assuming Persistence" (this is what is in their replication data), and one of which they call "Assuming Change" (hand-coded in one of their replication scripts).

With the "Assuming Persistence" version of the outcomes, shown in the first confusion matrix there are 15 civil war onsets, of which the forecasts capture 13 (true positives) and miss 2 (false negatives). With the "Assuming Persistence" coding of outcomes, shown in the second confusion matrix, there are 16 onsets and the forecasts capture 14 of those with, still, 2 misses. Fifteen to 16 civil war onsets in the first half of 2016 is incredible, especially considering that the entire historical data from 2001 to 2015 only has 20 recorded civil war onsets. Indeed, it turns out that B&S scored their forecasts for civil war *onset* using civil war *incidence*, i.e. counting ongoing civil wars as if they were continuing onsets.

The correct confusion matrices that use civil war onsets to score the forecasts are shown in the second part of Table 4. In the "Assuming Persistence" coding, there are no civil war onsets in the data for 2016-H1. The alternative "Assuming Persistence" coding records 2 civil war onsets in 2016-H1. As a result, if we compare the precision and recall of the forecasts with the correct scoring to the incorrect scoring in B&S, almost all of the forecasts are false positives. This is the norm when predicting rare events, and thus not unusual at all. The forecasts do correctly capture the 2 civil war onsets in the "Assuming Change" version of the outcome coding. This is, at face value, a credit to the model's accuracy.

5 Discussion

B&S advocate the use of theory to guide prediction. But "theory" is not defined by B&S and may be an ambiguous and undefined concept. "Theory" has many interpretations in social science. For some—the EITM movement, for example (Aldrich, Alt, and Lupia 2008)—theory is a system of internally consistent, logically connected propositions from which additional statements may be deduced. The additional statements can be subjected to empirical scrutiny. An example might be William Riker's theory of coalitions (Riker 1962) which posits propositions in game theoretic form from which the size principal of minimal winning coalitions may be deduced and further examined empirically. A more recent example of strong theory is found in *The Political Logic of Survival* (Bueno de Mesquita et al. 2005). For the purposes of discussion we call this strong theory.

Table 4: Replication of B&S Table 4: Confusion Matrices for Six-month Escalation Model forecasts for the first half of 2016.

Original confusion matrices with civil war *incidence*

Assuming Persistence

	$P=0$	$P=1$
$Y=0$	132	17
$Y=1$	2	13

Assuming Change

	$P=0$	$P=1$
$Y=0$	132	16
$Y=1$	2	14

Fixed confusion matrices with civil war *onset*

Assuming Persistence

	$P=0$	$P=1$
$Y=0$	134	30
$Y=1$	0	0

Assuming Change

	$P=0$	$P=1$
$Y=0$	134	28
$Y=1$	0	2

However, for many, a theory is a falsifiable story or narrative that makes causal claims about the relationships among variables. Typically it is fused together from a meta-analysis of prior work on the topic under consideration. We call these types of theory weak theory. They are weak because they do not include two characteristics of strong theory, though they may include one:

1. A theory is separate from the case or cases being examined. It is intended to be general.
2. A theory generates new statements that in principle can be empirically falsified, not proven.

Most social science research uses a weak form of theory. In the case of B&S they base their theoretical thinking on research that argues that conflict is not the product of structural factors, but rather comes from an interaction of state repression and the changes in costs and benefits to citizens for engaging in anti-government conflict. Following a wide swath of scholars, they argue that the escalation and de-escalations of interactions between citizens and government are what typically lead up to the onset of civil wars. Operationally, this boils down to four types of concepts: demands, accommodations, nonviolent repression, and 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.
- 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.
- 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.

These variables, taken from ICEWS, are the main components of the escalation model in the B&S article. The escalation theory is that the forces behind these variables interact such that as each of them increases, they independently increase the probability of civil conflict. The actual mechanisms of the dynamism among these variables are not specified.

Indeed most empirical social science research follows this approach: 1) delineation of an argument, 2) selection of variables that represent a plausible or compelling representation of the concepts, followed by 3) utilization of these variables in a statistical model with a dependent variable that measures the main phenomena of

investigation. It is worth pointing out that this is simply one set of variables that are selected to represent the concepts in the theoretical story. A strong theory gives more guidance on compelling choices than does a weak theory. Accordingly, while the choice is plausible is not necessarily compelling. Another choice might be equally plausible. In the ICEWS example, we might include variables on intragovernmental demands (from the military toward the central government), or further look a the dynamics among opposition groups. In practical terms there is a wide array of choices that are available to instantiate any particular theory. A second point is that putting them in a linear regression may be a particularly crude instantiation of complicated dynamics. One general principal that is on point in the B&S article and is more widely relevant is that dynamical phenomenon like civil wars need to have explanatory models that change rapidly over time—and by implication that are represented by data that change fairly quickly.

At this level of detail the theoretical model is more about choosing the correct variables and this is the contrast between theoretical prediction models and atheoretical machine learning that B&S focus on. But this dichotomy is misguided. The data that are used in a predictive model have to come from somewhere. Even if the inclusion of some factors is not explicitly justified, it is hardly the case that one assembles data without the expectation that it will be useful. It is not a coincidence that nobody uses 1,180 variable strawman models whose variable vectors predominantly consist of 0's, even if such a model coincidentally ends up performing well. Rather, we posit that most of the cases B&S might characterize as atheoretical enter the "theoretical", with some awareness of existing arguments and work that is used to select data sources for inclusion in a dataset.[7]

Furthermore, if what we mean by the concept "theory" is such that it suggests a large range of variables that could be related to an outcome, what do we really gain from having this theory in our predictive model? We can construct a model designed to predict well while avoiding overfitting, and then go back and try to couch it in theoretical language—which, given the malleability of extant theory, we probably could in many cases—but are we then better off? Certainly that's not what B&S are trying to suggest since this would be a problematic way to "evaluate" a theory. But when theory is extensive or malleable so that it suggests a broad range of variables and specifications, and thus a broad range of specifications could in fact be tied to a theory, it is hard to demonstrate that this theory is in fact not the result of predictive modeling's equivalent to p-hacking, namely developing a model specification and associated "theory" with any eye towards out-of-sample predictive accuracy (cf. Colaresi and Mahmood 2017 for how to avoid this problem while leveraging out-of-sample predictions). Incidentally, this is also a recipe for creating overfit models whose accuracy does not generalize and whose underlying theories may by extension not be externally valid (for a more detailed treatment of this issue, see Fariss and Jones 2018).

Conversely, not all factors that are "theoretically" important or suggested by the literature are also important for prediction, and some factors that turn out to be good predictors do not originate from prior theory. For example, both Ward, Greenhill, and Bakke (2010) and Hill and Jones (2014) examine a variety of variables that are considered "theoretically" important for explaining civil war onset and state repression, respectively, and find that there is substantial variation in how useful they are for for actually predicting those outcomes. And it is entirely possible that a focus on building accurate predictive models could uncover factors overlooked in the literature, e.g. the strong association between infant mortality and a broad variety of conflict outcomes like political instability (Goldstone et al. 2010), irregular leadership changes (Beger et al. 2017), and coups.[8]

Aside from selection of the variables that go into a model, another way in which we could contrast theoretical from atheoretical prediction models is in terms of their flexibility in regard to functional form. Theory should not only tell us what factors matter, but also how they relate to each other. We think that few would be strongly committed to the idea that the linear additive equations that dominate published regression analyses, maybe with the occasional polynomial or multiplicative interaction terms, are *the* correct functional specification. Is the danger posed by low-level violence the same when tensions are high, with high demands and high non-violent repression, as when tensions are low? Are those relationships linear in the log-odds? Some machine learning models, random forests included, are flexible in this regard and to some extent capable of learning non-linear and interactive relationships. B&S implicitly concede on this point since they use

[7] While we do not undertake this, there are a variety of procedures that one might use for variable choices, ranging from staring at lists of variables to machine learning approaches.

[8] Andreas Beger and Michael D. Ward, 2020, "Coup forecasts for 2020", https://www.predictiveheuristics.com/forecasts

random forests for both their Escalation model and the atheoretical alternatives. But in the absence of strong a priori theoretical expectations about functional form, why not use machine learning tools to try to uncover any such non-linear relationships in need of explanation, and use this to improve our theory (see also Fariss and Jones 2018)?

Rather than trying to frame theoretical modeling and machine learning as competing endeavors, a more useful perspective is to consider the ways in which they can mutually reinforce each other. Even if the analysis in B&S fails to support the point, we concede that predicting on the basis of a strong theoretical model is preferable to inductive prediction. Beyond accurate predictions, such models would allow us to make causal statements that are harder to make on the basis ad hoc models. But this is hard to do in practice. The problem is that many extant theoretical models are in fact so malleable to interpretation that they essentially are ad hoc. They are useful for identifying variables that could be useful for prediction, and this is in fact already taking place.

Most importantly, a single study is unlikely to be convincing that theory-based models are better (or worse) than ad hoc models based mainly in machine learning. Even if B&S were correct that in their study the theory based escalation model performed better, it would simply be a case study of a single case where it was preferable. It establishes no general tendency, let alone law-like evidence for general theory. Nor, against it.

6 Conclusion

We encountered several issues in the code underlying the B&S analysis. The problems we encountered are not subjective modeling choices. They bear no implications on the question of whether theory based model are preferable. However, when we fix these issues and perform an updated analysis, the B&S conclusions are all essentially overturned. In other words, B&S findings are based on a faulty analysis, invalid, and should be disregarded as evidence in favor of theory-based models.

In contradistinction to the conclusions offered by B&S, we find that when correctly specified and implemented:

- The theory-driven escalation model is outperformed both by the low-effort 1,160 predictor all-CAMEO model and the ensemble average model.
- The Average ensemble model and the CAMEO models outperform the escalation model in all instances.
- Adding structural variables substantially improves the escalation model's performance.

More importantly, trying to contrast machine learning and theory may be misguided. Even if a "theory" section is missing, predictive models are informed by previous work and arguments. If we had strong general theories of something like civil war onset, we could replace more inductive, ad hoc, methods with cleary specified models. But we do not. What we have instead are weak theories that could be interpreted to suggest any range of equations for predicting an outcome. Instead of trying to find an instantiation of such a "theoretical" model that can outperform a more ad hoc alternative, we could use the flexibility of machine learning methods and insights that predictive modeling brings to improve our weak theories.

Blair & Sambanis have focused attention on comparing forecasting models and forecasts in the civil war domain. As social science becomes more adept at predictive analysis, this will doubtless be of increasing importance. However, these comparisons must be made carefully to ensure those correct inferences are drawn. We continue to think that weak theory is overrated (Ward 2016), and that machine learning and big data will allow us to learn new things. It will be interesting to see how the evidence for these claims is adjudicated with additional usage and careful evaluation. Social science has leaned on theory for (at least) seventy years, but it might be time to find useful ways to incorporate some of the recently developed methods in the arena of machine learning and modern statistics as well.

7 References

Aldrich, John H., James E. Alt, and Arthur Lupia. 2008. "The Eitm Approach: Origins and Interpretations." In *The Oxford Handbook of Political Methodology*.

Beger, Andreas, Daniel W. Hill, Jr., Nils. W. Metternich, Shahryar Minhas, and Michael D. Ward. 2017. "Splitting It Up: The `spduration` Split-Population Duration Regression Package for Time-Varying Covariates." *The $\mathcal{R}$ Journal* 9 (2): 474–86. https://journal.r-project.org/archive/2017/RJ-2017-056/index.html.

Beger, Andreas, Richard K. Morgan, and Laura Maxwell. 2020. "The Democratic Spaces Barometer: Global Forecasts of Autocratization and Democratization." https://www.v-dem.net/en/analysis/DemSpace/.

Beger, Andreas, and Michael D. Ward. 2017. "Lessons from Near Real-Time Forecasting of Irregular Leadership Changes." *Journal of Peace Research* 54 (2): 141–56.

Blair, Robert A., and Nicholas Sambanis. n.d. "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning." *Journal of Conflict Resolution* TBD (firstview): 0022002720918923.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1). Springer: 5–32.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson, and James D. Morrow. 2005. *The Logic of Political Survival*. Cambridge, Mass: MIT Press.

Cederman, Lars-Erik, and Nils B. Weidmann. 2017. "Predicting Armed Conflict: Time to Adjust Our Expectations?" *Science* 355 (6324): 474–76.

Chiba, Daina, Nils W. Metternich, and Michael D. Ward. 2015. "Every Story Has a Beginning, Middle, and an End (but Not Always in That Order): Predicting Duration Dynamics in a Unified Framework." *Political Science Research and Methods* 3 (3): 515–41.

Colaresi, Michael, and Zuhaib Mahmood. 2017. "Do the Robot: Lessons from Machine Learning to Improve Conflict Forecasting." *Journal of Peace Research* 54 (2): 193–214.

Fariss, Christopher J., and Zachary M. Jones. 2018. "Enhancing Validity in Observational Settings When Replication Is Not Possible." *Political Science Research and Methods* 6 (2). Cambridge University Press: 365–80.

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). Wiley Online Library: 190–208.

Hanley, James A. 1988. "The Robustness of the "Binormal" Assumptions Used in Fitting ROC Curves." *Medical Decision Making* 8 (3): 197–203. https://doi.org/10.1177/0272989X8800800308.

———. 2014. "Receiver Operating Characteristic (ROC) Curves." In *Wiley Statsref: Statistics Reference Online*. American Cancer Society. https://doi.org/10.1002/9781118445112.stat05255.

Hegre, Håvard, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, et al. 2019. "ViEWS: A Political Violence Early-Warning System." *Journal of Peace Research* 56 (2): 155–74.

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108 (03). Cambridge University Press: 661–87.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Malley, James D., Jochen Kruppa, Anirban DasGupta, Karen G. Malley, and Andreas Ziegler. 2012. "Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines." *Methods of Information in Medicine* 51 (1): 74–81. https://www.thieme-connect.de/products/ejournals/abstract/10.3414/ME00-01-0052.

Pulit, Michał. 2016. "A New Method of Kernel-Smoothing Estimation of the Roc Curve." *Metrika* 79 (5): 603–34.

Riker, William H. 1962. *The Theory of Political Coalitions*. Yale University Press.

Ward, Michael D. 2016. "Can We Predict Politics? Toward What End?" *Journal of Global Security Studies* 1 (1): 80–91.

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* 16 (4): 473–644.

Zou, Kelly H., William J. Hall, and David E. Shapiro. 1997. "Smooth Non-Parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests." *Statistics in Medicine* 16 (19): 2143–56.

A Appendix

A.1 Code references for coding errors

A.1.1 Section 3.2: Incorrect "Weighted by PITF" Implementation

We believe the intention with the "Weighted by PITF" model is to combine the PITF test predictions with the Escalation test predictions, i.e. weight the latter with the former. Due to an error, the PITF *training* set predictions are used, not the *test* set predictions.

See `1mo_run_escalation_weighted_PITF.R` line 4, where the PITF predictions are taken from the training data set (`train$pred_prob_plus1`). The next line is a hack extending the shorter `weight` vector with missing values to avoid a R warning when it is multiplied with the longer vector of escalation model test set predictions. Similarly in the 6-month version of this file.

Uncovering which cases in the PITF training predictions are matched to the cases that the Escalation model test set predictions correspond to requires tracing back several more data objects. We have documented details in issue #3 in the GitHub repo.

A.1.2 Section 3.3: Incorrect "PITF Split Population" Implementation

The intention is to train seperate random forests on data split into high and low-risk groups based on PITF predictions, and then combine them back into one random forest. Actually both random forests are trained on identical data, essentially production a random forest that just has N_trees * 2.

Disentangling this coding error is not straightforward as it occurs over several R scripts and requires (or at least is easier to verify by) running partway through the actual replication until the objects holding the training data for the models are instantiated and can be examined. Details are documented in issue #5 in the GitHub repo.

A.1.3 Section 3.5: Incorrect scoring of the 2016 forecasts

The relevant variables in the data are "incidence_civil_ns" and "incidence_civil_ns_plus1", which appears to be a 1-period lead (ie., t_+1) version of the dependent variable that is used in the actual prediction models. The incidence dependent variable contains both 0/1 and missing values. By examining the pattern of missing values, it seems clear that this was originally an incidence variable indicating whether a country was at civil war in a given year or not, and which was converted to an onset version so that onsets retain the value of 1 but continuing civil war years are coded as missing. This reflects common practice.

However, by examining the code used to generated Table 4, we were able to confirm that the onset forecasts are assessed using incidence, not onset. In the file `6mo_make_confusion_matrix.do` on line 52, missing values in "incidence_civil_ns" are recoded to 1, thus reverting the onset coding of this variable to incidence.

A.2 Additional replication tables

Table A1 is our replication of B&S Table 1 with smoothed AUC-ROC. The results differ slightly from the original B&S Table 1, typically by no more than 0.01, due to the non-deterministic nature of the RF models. It is the case that B&S set the RNG seed in their replication code, which should theoretically allow exact reproduction, but (1) there was a change in more recent versions of R that affected the RNG seeding process, and (2) we refactored the replication script to allow one to run the models in parallel. In any case, the interpretation of results should not be sensitive to random variation, i.e. it should not depend on using a specific RNG seed. On the basis of these results, B&S conclude that the escalation model is generally superior to the alternatives, and we can replicate that interpretation when using smoothed ROC curves.

A.3 Model to model comparison plots

Figures A1 and A2 replicate the information in Tables 1 and 3 in a way makes the comparison of the escalation model to the alternative models easier. Each facet shows the escalation AUC-ROC for all model settings

Table A1: Replication of B&S Table 1 with smoothed ROC curves; test set AUC-ROC for various models

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.85	0.80	0.79	0.84	0.83
Terminal nodes	0.86	0.79	0.77	0.83	0.82
Sample size	0.85	0.81	0.71	0.86	0.84
Trees per forest	0.85	0.80	0.78	0.83	0.82
Training/test sets 1	0.86	0.78	0.76	0.81	0.80
Training/test sets 2	0.82	0.79	0.72	0.77	0.78
Training/test sets 3	0.80	0.80	0.69	0.74	0.75
Coding of DV 1	0.86	0.81	0.80	0.84	0.83
Coding of DV 2	0.92	0.80	0.81	0.81	0.81
Six-month forecasts					
Base specification	0.82	0.78	0.82	0.77	0.79
Terminal nodes	0.79	0.76	0.81	0.77	0.77
Sample size	0.83	0.78	0.78	0.79	0.79
Trees per forest	0.82	0.77	0.82	0.77	0.79
Training/test sets 1	0.80	0.78	0.81	0.76	0.78
Training/test sets 2	0.72	0.74	0.77	0.73	0.75
Training/test sets 3	0.88	0.71	0.81	0.68	0.79
Coding of DV 1	0.83	0.77	0.82	0.79	0.80
Coding of DV 2	0.83	0.77	0.83	0.78	0.79

(the rows in Table A1 and base specification for all models in Table 3) on the left, and the AUC-ROC for an alternative model on the right, with a connecting line. If the lines slope up to the right, the alternative model is better.

A.4 Random forest hyper-parameter selection

What initially sparked our interest in the paper was the unusual choice of hyperparameter settings for the random forest models estimated. Table A5 shows the default values used by the implementation of random forest that B&S use (from the **randomForest** R package), in contrast to the basic settings used by B&S for most the models reported in the paper.

As the outcome is a binary indicator of civil war onset, one would typically use a classification random forest that predicts 0 or 1 labels directly. The implementation of random forests that B&S use ((Liaw and Wiener 2002)) is based on the original Breiman (2001) implementation and calculates predictive probabilities by averaging over the "0" or "1" votes from all constituent decision trees. The conventional wisdom regarding the number of trees in a random forest is that it needs to be large enough to stabilize performance, but without any additional gain or harm in accuracy beyond a certain number. From the other default settings, which are generally not uninformed choices, one can see that the basic logic is to grow a forest with a relatively small number of trees, but where each tree is fairly extensive, and operates on a large bootstrapped sample of the original training data. These are of course only heuristics and it is usual to attempt to find better hyper-parameter methods through some form of tuning.

B&S in contrast fit very large forests with 100,000 trees in the basic model form, but where each tree only operates on a very small sub-sample (N=100 or 500), drawn without replacement, of the available training data. This approach only works due to the choice to use regression, not classification, trees. Trying to use classification trees with the other parameter settings does not work at all because it is almost guaranteed that a sample of 100 from the ~11,000 training data rows with 9 positive cases will only include 0 (negative) outcomes in the sample. As it is, using regression with a 0 or 1 outcome produces warnings when estimating the models:

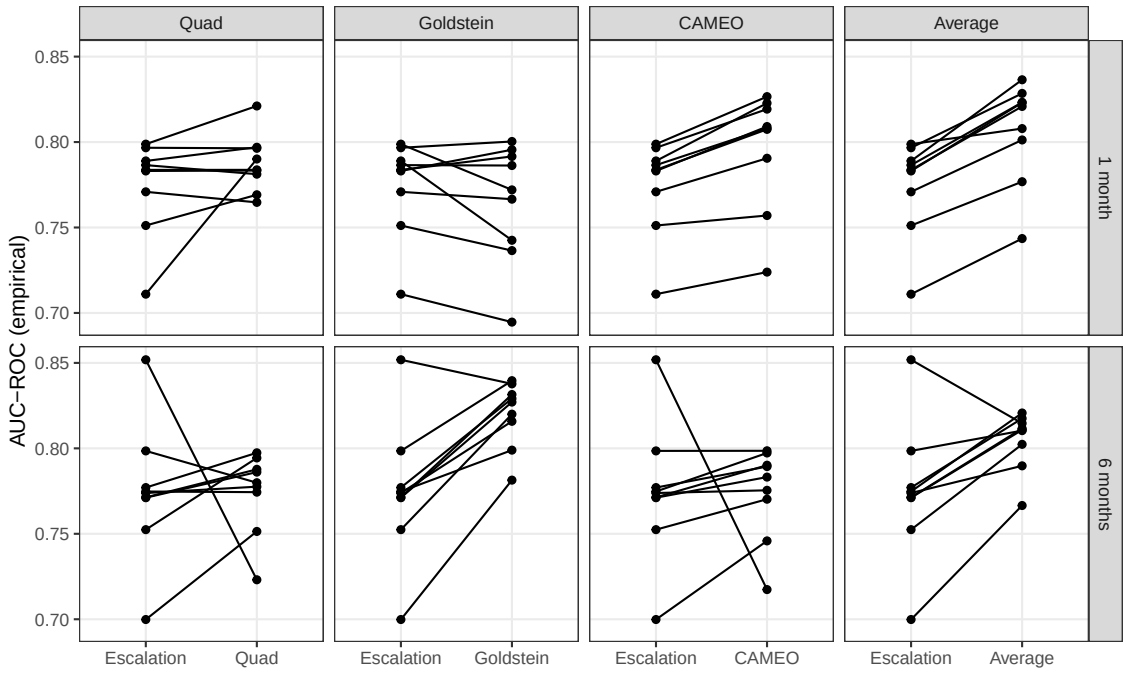


Figure A1: Escalation to alternative comparisons for the ICEWS models (B&S Table 1)

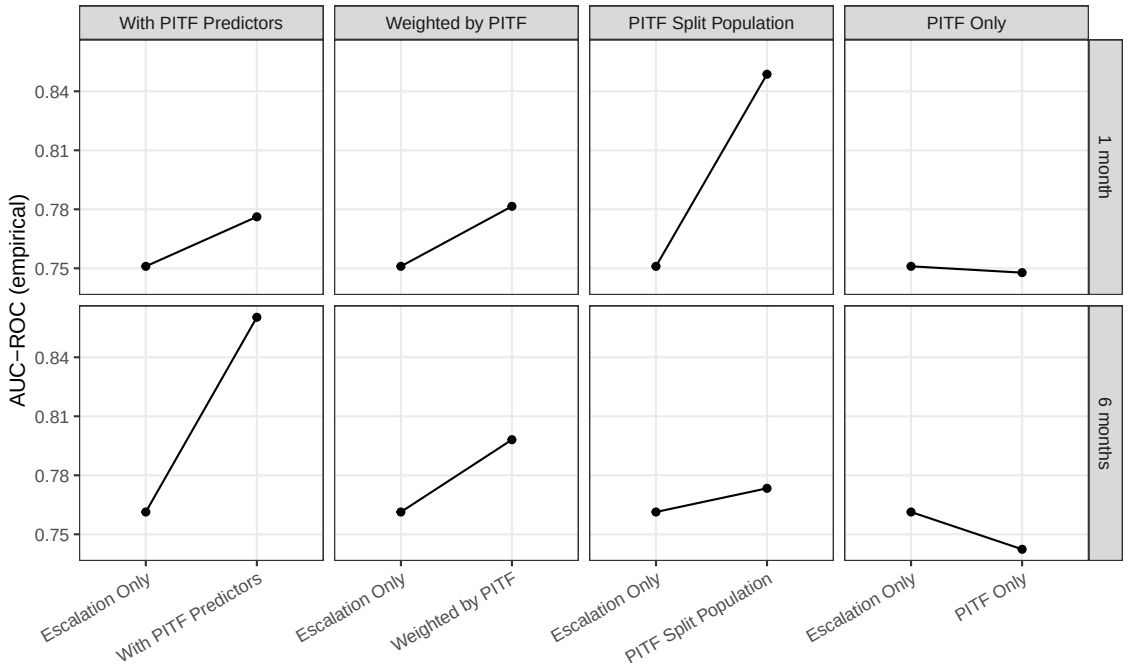


Figure A2: Escalation to alternative comparisons for the structural models (B&S Table 2)

Table A2: Replication of B&S Table 2 with smoothed/original ROC and with original varying N cases or adjusting for common case set with constant N

	Smoothed ROC	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
Original model-specific cases						
1 month						
	Yes	0.85	0.77	0.80	0.81	0.76
	No	0.79	0.78	0.80	0.79	0.75
6 months						
	Yes	0.82	0.85	0.81	0.78	0.74
	No	0.77	0.86	0.81	0.78	0.74
Cases adjusted to common subset						
1 month						
	Yes	0.81	0.77	0.79	0.86	0.76
	No	0.75	0.78	0.78	0.85	0.75
6 months						
	Yes	0.82	0.85	0.80	0.77	0.74
	No	0.76	0.86	0.80	0.77	0.74

```
Warning message:
In randomForest.default(y = as.integer(train_df$incidence_civil_ns_plus1 ==  :
  The response has five or fewer unique values.  Are you sure you want to do regression?
```

As it turns out, using regression random forests for this kind of binary classification problem in order to obtain probability estimates matches the probability random forest approach suggested and positively evaluated in Malley et al. (2012), and which is used in another prominent R implementation of random forests.[9] It is not clear whether this is intentional, as the Malley paper is not cited in B&S.

In any case, B&S's random forest approach appears to work really well. We tried to construct classification random forests tuned via cross-validation on the training data set partition, i.e. without touching the test data, but were unable to develop models that consistently match the B&S random forest method in both cross-validated out-of-sample training predictions and test set predictions.

Given that they are relatively unorthodox, yet appear to work very well, we wonder how the hyper-parameter values were determined. Two specific concerns are that this was not done with an eye towards test set accuracy, which would invalidate the independence of the out-of-sample test set, and whether the specific hyper-parameter values are optimized for only one model, or were optimized and found to work well for all models. There is no discussion of the random forest tuning strategy or how the specific hyper-parameter methods were determined in the paper.

Given the dramatic changes in results as a result of the preceding issues, we did not further investigate these potential concerns.

A.5 Different smoothing methods

The **pROC** package `smooth()` function includes several different smoothing methods. The default, which BS use in their code, is binormal smoothing. Here is a replication of the smooth benefits shown in Figure 2 with

[9] The **ranger** package.

Table A3: Smoothing advantage for B&S Table 1: the gain in AUC-ROC when calculated using smoothed ROC curves

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.06	0.02	-0.01	0.03	0.00
Terminal nodes	0.07	0.01	-0.01	0.02	0.00
Sample size	0.06	0.01	-0.04	0.04	0.00
Trees per forest	0.06	0.02	-0.01	0.02	0.00
Training/test sets 1	0.09	0.02	-0.01	0.02	0.00
Training/test sets 2	0.07	0.02	-0.01	0.02	0.00
Training/test sets 3	0.09	0.01	0.00	0.02	0.01
Coding of DV 1	0.07	0.02	0.00	0.02	0.00
Coding of DV 2	0.12	-0.02	0.04	-0.01	0.01
Six-month forecasts					
Base specification	0.05	-0.01	-0.01	-0.01	-0.02
Terminal nodes	0.02	-0.02	-0.01	-0.01	-0.02
Sample size	0.05	0.00	-0.02	-0.01	-0.03
Trees per forest	0.05	-0.01	-0.01	-0.02	-0.02
Training/test sets 1	0.04	-0.01	-0.01	-0.02	-0.02
Training/test sets 2	0.02	-0.01	-0.02	-0.01	-0.02
Training/test sets 3	0.03	-0.02	-0.03	-0.04	-0.02
Coding of DV 1	0.05	-0.03	-0.01	0.00	-0.02
Coding of DV 2	0.03	-0.01	-0.01	-0.02	-0.02

Table A4: Smoothing advantage for B&S Table 2: the gain in AUC-ROC when calculated using smoothed ROC curves

Model	Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts					
Base specification	0.06	-0.01	0.01	0.01	0.01
Six-month forecasts					
Base specification	0.06	-0.01	0.00	0.00	0.00

two other smoothing methods. (We could not get "logcondens" and "logcondens.smooth" to work, and they are based on another external package that has not been updated since 2016.)

Only the "binormal" smoothing method produces a pattern of AUC-ROCs that clearly elevates the Escalation model above the alternatives. The "density" method does not on average change the AUC-ROC values very much from their empirical ROC curve equivalents. The "fitdistr" method tends to increase all AUC-ROC values, although maybe slightly more so for the Escalation model than the main alternatives.

```
## Warning: package 'yardstick' was built under R version 3.6.2
```

Table A5: Random forest (`randomForest()`) default versus B&S hyperparameters

Hyperparameter	Default heuristic	Default values (Escalation)	B&S value
type		classification	regression
ntree		500	100,000 or 1e6
mtry	`floor(sqrt(ncol(x)))`	3	3
replace		true	false
sampsize	`nrow(x)` if replace, else `ceiling(.632*nrow(x))`	11,869	100 or 500
nodesize	1 for classification	1	1
maxnodes		null	5 or 10

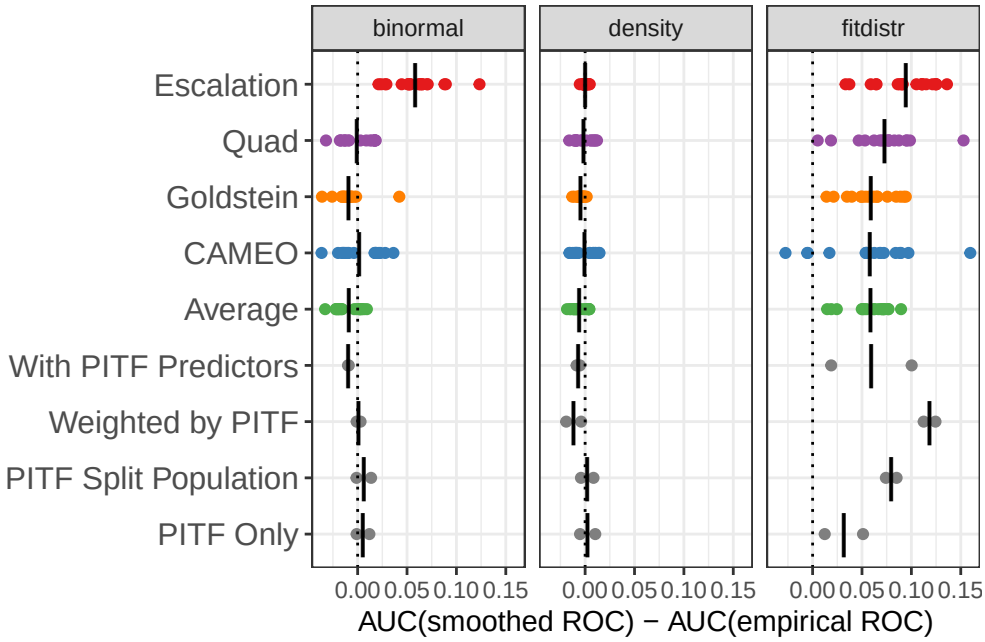


Figure A3: Gain from using smoothed ROC to calculate AUC, for each model reported in Tables 3 and 4 (B&S Tables 1 and 2), with 3 different smoothing methods (binormal is the default method and used in B&S)