Good Empirics Dominate Bad Theory: Comments on Blair and Sambanis, 2020, "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning", *JCR*

Richard K. Morgan and Andreas Beger and Michael D. Ward

04 June 2020

Introduction

Blair and Sambanis (2020; hereafter B&S) argue that theory is important for creating models that have high accuracy in forecasting civil war onset. Indeed they argue that with such theory, forecasting is more accurate than has previously been possible. Setting aside the validity of this argument, we re-examine the empirical basis for the claims made in their article. We find that these claims are false. Their theory-based escalation model does not do better than the alternatives that they examine, indeed it does worse. The reason for this reversal of their conclusion is that they have made several mistakes in their research procedure: 1) they use smoothed performance curves instead of the unsmoothed curves, 2) they mistakenly encode the dependent variable to be incidents of civil war, rather than as they claim the onset of a civil war, 3) they mistakenly calculate the predicted values for the alternative "non-theoretical" models that they compare to. In what follows we show the impact of these mistakes on the conclusions, we examine their understanding of prediction it self, and we explore the logic of their assertion about theory as well. We believe that this research by B&S is incorrect and should be retracted.

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 `remarkable accurately.'' Indeed they argue that this so-called theoretical model outperforms four other`more mechanical" alternatives. Second, they clam that the integration of structure with process are better over short forecasting windows. Third they preregisted the list of thirty countries which 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. B&S claim to be unique in assessing these forecasts. A qualitative analysis of their predictions allows them to conclude that their model is strong and a more precise test is undertaken. 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 .82 and .85 over one- and six-month windows, respectively, and AUCs as high as .92 in other specifications. Our model also consistently and sometimes dramatically outperforms the alternatives we test...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.

Review of Blair and Sambanis 2020

Blair and Sambanis (2020) (B&S for the sake of brevity hereafter) aim to examine whether theory adds value to a forecasting model when compared to non-parametric machine learning models, and specifically random forests, whose specifications are not theory-informed. They find that it does. They arrive at this conclusion by examining the problem of predicting civil war onset, and find 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 and a kitchen sink model with more than 1,000 covariates.

B&S specifically examine three questions:

- (1) How does the theoretically-driven rscalation 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 for the first half of 2016 derived from the escalation model?

To assess the first two questions B&S use global datasets 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 is civil war onset, measured using Sambanis' civil war dataset.

Both the first and second questions above rely on comparing the 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) sets.
2. Estimate the Escalation and other competing models.
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.

The corresponding results for each question are shown in B&S Tables 1 and 2, which we will replicate further below.

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, putatively drawn from a theoretical escalation model.
- Quad: ICEWS quad counts, i.e. material conflict, material cooperation, verbal conflict, verbal cooperation.
- Goldstein: -10 (conflictual) to 10 (cooperative) scores derived from the ICEWS data for interactions between the government one one side and opposition or rebel actors on the other. These are directed, thus making for 4 total covariates.
- CAMEO: counts for all CAMEO event codes, thus a total of 1,159 covariates (CAN THIS BE RIGHT?).
- Average: unweighted average of the predictions from the four models briefly described above.

The results in Table 1, aside from the core base specification results, include 8 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 as predictors has the escalation model indicators but also 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 the Escalation model to create forcasts for the first half of 2016, and in their third and final analysis, they score the forecasts accuracy using civil war onset data later observed. This is summarized in B&S Table 3.

Replication problems

While attempting to replicate B&S's results, we found several issues.

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, actual ROC curves. A reference to smoothing is made in a single sentence in the paper (p. 12):

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

This implies that the ROC curves were only smoothed in the references Figure 1, but actually all AUC-ROC calculations throughout the replication code use an option to smooth the ROC curves prior to AUC calculation.

Figure 1 shows our replication of both the smoothed ROC curves B&S report, and the actual ROC curves on the right.

ROC curves typically appear step-like in response to the distribution of positive and negative cases in the data. In this case there are also groups of cases with identical predicted probabilities, which accounts for the unusual diagonal lines. 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 a observed positive case is reached. For these ROC curves, and for that matter in the basic 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. We can only speculate, but maybe the very "step"-like nature of the ROC curves plus the presence of unusual diagonal lines accounts for the decision to use smoothed ROC curves.

As it turns out, this decision has a dramatic impact on the AUC-ROC calculations. As we will show in the results further below, the difference between the smoothed and original ROC AUC values is up to 0.12–a huge difference given that row-wise, the models being compared typically differ by only 0.05 or less–and differentially impacts the models that are being compared. In fact, B&S's original results and interpretation are entirely conditional on the use of smoothed AUC-ROC.

To our knowledge, the norm is to calculate AUC-ROC values using original, not smoothed ROC curves. We are in fact not aware of other work that uses smoothed ROC curves for AUC calculations. It might be that there are theoretical reasons justifying the use of smoothed ROC curves over the original ROC curves; but given that this decision dramatically impacts the interpretation of the B&S results, it minimally would have warranted an explicit discussion in the paper. This is not the case, and it is only mentioned in the sentence we quote above.

The next three issues we encountered all concern information in Table 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.
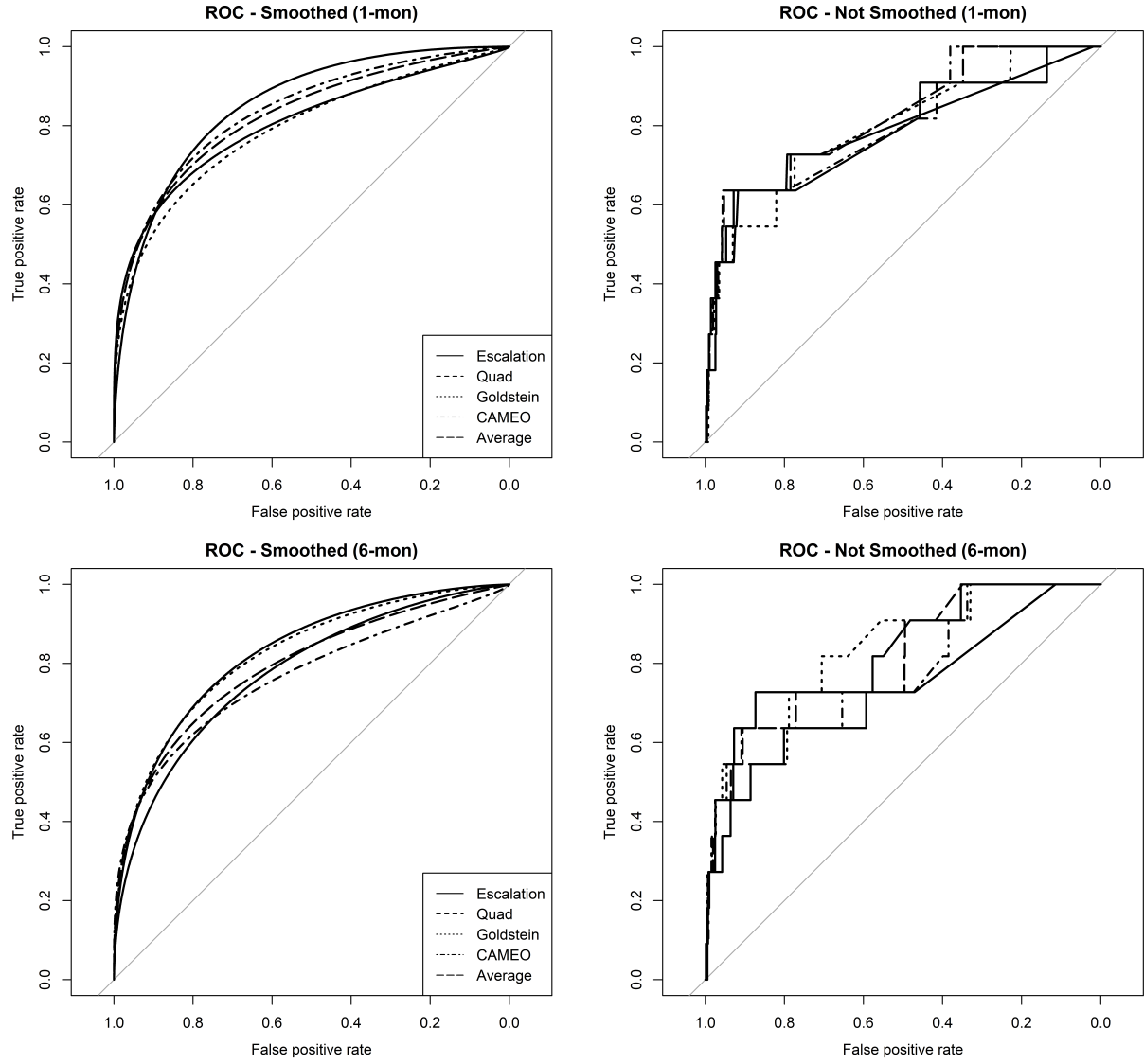
Figure 1: Replication of B&S Figure 1 with both smooth and non-smoothed ROC curves.

We infer that the intent is that the Escalation model's predictions for the test set are weighted by the PITF model predictions for the test set. The implementation actually weighs the test set predictions using the PITF model predictions for the training set.[1] This appears to be a coding error.

Incorrect "PITF Split Population" implementation

Similarly, the "PITF Split Population" model appears to be incorrectly implemented. B&S describe it as, on page 20:

> The final approach is a random forest analog to split-population modeling. first compute the average PITF predicted probability for each country across all years in our training set. We define those that fall in the bottom quartile as "low risk" and the rest as "high risk." We then run our escalation model on the high-risk and low-risk subsets separately, combining the results into a single random forest (column 4).

The intention clearly was 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 are both run on the exact same training data, which consists of the full training data all other models are run on. The models are also identical otherwise, 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 RNG seed, they would be identical in every respect, producing identical predictions.[2]

The implementation error aside, this split-population analog model is actually quite odd and does not actually replicate an analogue of the idea behind split-population modeling. Although the two RFs are trained on separate data (in our updated, fixed replication), the process of combining them actually just creates a new, larger RF using both component model's underlying decision (regression[3]) 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 indeed have 200,000 decision trees. Furthermore, the PITF model predictions do not impact the way 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 just 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).

Inconsistent test set N for the models in Table 2

Lastly, the AUC-ROC values reported in the original B&S Table 2 are calculated on the basis of slightly different numbers of underlying test set cases (see Table 8). 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 Table 2. Thus a difference in AUC-ROC values for two models could be due to the fact that 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 actually not comparable to the other, and any conclusions drawn from such comparison are potentially incorrect.

We fix this issue by only using predictions for common joint subset of cases that all models have non-missing predictions for. The original B&S Table 2 1-month have N=11,806–12,495 versus a common joint subset of 9,811, and for the 6-month row N=2,070–2,233 versus N=1,915 for the common joint subset.

---

[1]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.

[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. We have documented details at https://github.com/rickmorgan2/Blair-Sambanis-replication/issues/5.

[3]See further below. Although the RF models are used for a binary decision problem, the actual implementation uses regression RFs for continuous outcomes.

Table 1: 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	0.83	0.82
Terminal nodes	0.86	0.79	0.78	0.83	0.82
Sample size	0.85	0.81	0.70	0.86	0.84
Trees per forest	0.84	0.80	0.78	0.83	0.82
Training/test sets 1	0.86	0.78	0.75	0.81	0.80
Training/test sets 2	0.81	0.79	0.72	0.78	0.78
Training/test sets 3	0.79	0.80	0.69	0.75	0.75
Coding of DV 1	0.86	0.81	0.80	0.84	0.83
Coding of DV 2	0.92	0.81	0.81	0.82	0.81
Six-month forecasts					
Base specification	0.82	0.78	0.82	0.77	0.79
Terminal nodes	0.80	0.75	0.81	0.75	0.77
Sample size	0.83	0.78	0.78	0.78	0.79
Trees per forest	0.82	0.78	0.82	0.77	0.79
Training/test sets 1	0.79	0.77	0.81	0.75	0.78
Training/test sets 2	0.72	0.73	0.77	0.74	0.75
Training/test sets 3	0.88	0.70	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	0.82	0.79	0.79

Incorrect scoring of the 2016 forecasts

B&S show a confusiong 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 the much more common incidence of civil war, i.e. including ongoing civil wars as "1"'s.

The relevant variables in the data are "incidence_civil_ns" and "incidence_civil_ns_plus1", which appears to be a 1-period lead version of the DV that is used in the actual prediction models. The incidence DV 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 in how these are coded.

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 back to incidence.

Results of the updated replication

Basic interpretation of Table 1

Table 1 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.

Table 2: Replication of B&S Table 1 *without* smoothed ROC curves; test set AUC-ROC for various models

Model	Escalation	Quad	Goldstein	CAMEO	Average
One-month forecasts					
Base specification	0.78	0.78	0.79	0.80	0.82
Terminal nodes	0.79	0.78	0.78	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	0.78	0.76	0.76	0.79	0.80
Training/test sets 2	0.75	0.77	0.73	0.76	0.78
Training/test sets 3	0.70	0.79	0.69	0.73	0.74
Coding of DV 1	0.80	0.79	0.80	0.82	0.83
Coding of DV 2	0.80	0.82	0.78	0.83	0.81
Six-month forecasts					
Base specification	0.77	0.78	0.83	0.79	0.81
Terminal nodes	0.77	0.78	0.82	0.78	0.79
Sample size	0.78	0.77	0.80	0.80	0.82
Trees per forest	0.77	0.79	0.83	0.78	0.81
Training/test sets 1	0.75	0.79	0.82	0.78	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	0.71	0.81
Coding of DV 1	0.77	0.80	0.83	0.79	0.82
Coding of DV 2	0.80	0.78	0.84	0.80	0.81

Table 2 shows a version of B&S Table 1 with the conventional non-smoothed ROC curves. The Average model outperforms the Escalation model in 17 out of 18 cases, and the CAMEO model outperforms in 16 of 18 cases, with one tie. 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 completely conditional on the non-standard use of smoothed ROC curves, and overturns when using traditional AUC-ROC calculations.

Do structural variables add to the Escalation model?

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 that all models have non-missing predictions for. Table 9 further below 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 the results as follows, on page 20:[4]

> Of the approaches we test, the split-population analog is most promising

This is no the case anymore. It outperforms in the 1-month version and underperforms the Escalation model in the 6-month version.

> Adding PITF predictors improves the performance of the escalation model over six-month windows but diminishes it over one-month windows.

Adding PITF predictors actually improves performance in both cases; the "With PITF Predictions" model strictly dominates the "Escalation Only" model.

> The weighted model performs very poorly regardless.

It performs roughly on par with the Escalation Only model.

---

[4]We list the "Overall, . . . " interpretation out of order, last, for clarity.

Table 3: Replication of B&S Table 2: Test set AUC-ROC for Escalation model with and without structural PITF contribution

Escalation Only	With PITF Predictors	Weighted by PITF	PITF Split Population	PITF Only
One-month forecasts				
0.69	0.71	0.69	0.81	0.67
Six-month forecasts				
0.79	0.87	0.80	0.69	0.71

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.

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 not surprising.

> 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 in fact actually incorporate structural information at all. We thus conclude that adding structural variables actually clearly improves predictive performance.

The effect of using smoothed ROC curves

What impact did the ROC smoothing have overall on the results reported in B&S Tables 1 and 2? Tables 4 and 5 show the increase in AUC-ROC when using smoothed ROC curves, compared to the standard non-smoothed AUC-ROC. Positive values indicate that smoothing increased a model's apparent performance. The Escalation model is the only model that consistently had a benefit from smoothing. For all 8 other models, smoothing sometimes gave a benefit, sometimes not.

A more dramatic difference stands out when we consider the overall average impact of smoothing across all flavors of a model reported in the columns of the tables. The Average, Goldstein, Quad, PITF Split Population, and Weighted by PITF models are slightly hurt by smoothing, but on the order of less than 0.01 in absolute magnitude. The "With PITF Predictors" model is hurt on the order of 0.02, i.e. it appears substantially worse with smoothing. The CAMEO and PITF Only models benefit slightly, on the order of 0.01 or less. The Escalation Only model on the other hand has an average boost of 0.058 to it's AUC-ROC from smoothing. Considering the spread of AUC-ROC values if we compare across rows in B&S Tables 1 and 2, that boost is substantial.

The use of smoothed ROC curves to calculate AUC-ROC values only clearly benefits the Escalation model, it does so consistently, and by a considerable margin. All 8 alternative models reported in Tables 1 and 2 do not on average gain when using smoothed ROC curves to calculate AUC.

Table 4: 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	0.07	0.02	0.00	0.02	0.00
Terminal nodes	0.06	0.01	-0.01	0.02	0.00
Sample size	0.06	0.01	-0.03	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	0.00	-0.01	0.02	0.00
Training/test sets 3	0.09	0.01	-0.01	0.02	0.01
Coding of DV 1	0.06	0.02	0.00	0.03	0.00
Coding of DV 2	0.12	-0.01	0.04	-0.01	0.01
Six-month forecasts					
Base specification	0.05	-0.01	-0.01	-0.02	-0.02
Terminal nodes	0.02	-0.02	-0.01	-0.02	-0.02
Sample size	0.05	0.00	-0.02	-0.01	-0.03
Trees per forest	0.05	-0.01	-0.01	-0.01	-0.02
Training/test sets 1	0.04	-0.01	-0.01	-0.02	-0.02
Training/test sets 2	0.03	-0.01	-0.02	-0.01	-0.02
Training/test sets 3	0.03	-0.02	-0.03	-0.02	-0.02
Coding of DV 1	0.05	-0.03	-0.01	-0.01	-0.02
Coding of DV 2	0.03	-0.01	-0.01	-0.01	-0.02

Table 5: 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	0.05	-0.03	0	-0.02	0.02
Six-month forecasts					
Base specification	0.06	-0.01	0	0.01	0.00

How accurate were the 2016 forecasts?

Additional concerns

Random forest hyperparameters

What initially sparked our interest in the paper was the unusual choice of hyperparameter settings for the random forest models estimated. Table 6 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 predictived 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

Table 6: 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

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 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:

```
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.[5] 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.

AUC-ROC sensitivity

RNG and variation

Pre-registration

- The escalation model specification, prior to looking at data!
- The hyper-parameter tuning strategy.

---

[5]The **ranger** package.

Conclusion

B&S has a clear misunderstanding of the role that out-of-sample prediction can plan in analysis. On the one hand it can be used for simple forecasting while on the other it can be used to evaluate the performance of models, specifically to overfitting and bias. They further misrepresent cross-validation which is not about fishing for results, as claimed by B&S. The current state-of-the-art uses prediction and cross-validation most frequently to provide supportive evidence that is independent of the estimation procedure and the in-sample data. It is incorrect to paint this procedure as atheoretical since many studies will have some explanation for how the model was constructed. These procedures may simply be used to provide evidence for a theoretical argument. Though that argument may be unconvincing and/or atheoretical.

The B&S approach that is advocated is to use theory to guide prediction. But theory is a quite ambiguous and undefined concept. It is not a procedure. What they actually do is to create a model with four right hand side variables that is supposed to capture a complicated repression-dissent dynamic. The dynamic is probably not linear, but their model is. There is a wide-ranging literature on this dynamic that they do not rely on to construct their model. As such their baseline comparison is hardly a standard bearer for strong theory.

Further, they complete misunderstand the use of ICEWS event data in current research. They claim that most uses to date have focused on the quad categories, but this ignores a wide swath of literature (Steinert-Threlkeld ApsR 2017) and Metternich et al (AJPX 2013) that uses a sepcific action—such as protest—defined in the CAMEO ontology. In the AJPS article we hand coded, for example, every actor in Thailand and focused on an analysis of how those have been interacting.

We encountered several issues in the code underlying the B&S analysis.

The issues we encountered are not subjective modeling choices.

When we fix these issue and perform an updated analysis, the conclusions B&S draw all essentially overturn. In other words, B&S findings are based on a faulty analysis, and invalid.

Using the same analysis B&S intend to use, we in fact find that:

- the theory-driven Escalation model is outperformed both by the low-effort 1,160 predictor all CAMEO model and the Average ensemble model
- structural variables substantially improve the Escalation model's performance when added to the pool of predictors the underlying random forest model draws on

Bueno de Mesquita (2011) , Ward (2016) , Montgomery, Hollenbach, and Ward (2012a), Chiba, Metternich, and Ward (2015), Brandt (2015) Montgomery, Hollenbach, and Ward (2012b),Ward et al. (2013),Chadefaux (2015),Weidmann and Ward (2010),Pilster and Böhmelt (2014),P. T. Brandt, Freeman, and Schrodt (2011a),Gleditsch (2015),Bennett and Stam (2009),Shellman, Levey, and Young (2013),Hegre, Hultman, and Nygård (2011),Ward, Greenhill, and Bakke (2010),Freeman and Job (1979),Gneiting and Raftery (2005) Choucri (1974),Choucri and Robinson (1979),Goldstone et al. (2010),tetlock:etal:2017, Rost, Schneider, and Kleibl (2009),
Hopple and Andriole (1984), Andriole and Young (1977), P. T. Brandt, Freeman, and Schrodt (2011b), Brandt, Freeman, and Schrodt (2014), Mellers et al. (2015), Tetlock and Gardner (2015), Hegre et al. (2019), Weidmann and Ward (2010), Beger and Ward (2017), Schrodt (1991), O'Brien (2010), O'Brien (2002), Hegre et al. (2019), Hegre (2017), Hegre et al. (2016), Chiba and Gleditsch (2017), Buhaug, Cederman, and Gleditsch (2014), Colaresi and Mahmood (2017), Ramakrishnan et al. (2014), Schleussner et al. (2016), Kupilik and Witmer (2018), Witmer et al. (2017) O'Loughlin et al. (2012) Hegre et al. (2013) Buhaug and Gleditsch (2008)

AB: add a table comparing B&S original claims and updated results.

Table 7: Replication of B&S Table 4: 2016 Confusion Matrices for Six-month Escalation Model.

header	Observed	Predicted0	Predicted1
Original, scored with civil war incidence			
Assuming Persistence	0	132	17
	1	2	13
Assuming Change	0	132	16
	1	2	14
Fixed, scored with civil war onset			
Assuming Persistence	0	134	30
	1	0	0
Assuming Change	0	134	28
	1	0	2

Table 8: Number of valid test predictions for each cell in 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	9811	9811	9811	9811	9811
6 months	1915	1915	1915	1915	1915

References

Andriole, Stephen J., and Robert A. Young. 1977. "Toward the Development of an Integrated Crisis Warning System." *International Studies Quarterly* 21 (1): 107–50.

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.

Bennett, D. Scott, and Allan C. Stam. 2009. "Revisiting Predictions of War Duration." *Conflict Management and Peace Science* 26 (3): 256–67.

Blair, Robert A., and Nicholas Sambanis. 2020. "Forecasting Civil Wars: Theory and Structure in an Age of 'Big Data' and Machine Learning." *Journal of Conflict Resolution.*

Brandt, Patrick. 2015. "Forecasting Conflicts: Long and Short Term Predictions Based on Different Training Set Considerations." University of Texas Dallas; Peace Research Institute Oslo.

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

———. 2011b. "Real Time, Time Series Forecasting of Inter-and Intra-State Political Conflict." *Conflict Management and Peace Science* 28 (1). Sage Publications: 41–64.

———. 2014. "Evaluating Forecasts of Political Conflict Dynamics." *International Journal of Forecasting* 30 (4). Elsevier: 944–62.

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

Table 9: 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	0.76	0.80	0.81	0.77	
No	0.78	0.77	0.80	0.80	0.75	
6 months						
Yes	0.82	0.86	0.81	0.79	0.74	
No	0.77	0.86	0.81	0.82	0.74	
Cases adjusted to common subset						
1 month						
Yes	0.74	0.68	0.69	0.80	0.69	
No	0.69	0.71	0.69	0.81	0.67	
6 months						
Yes	0.85	0.86	0.79	0.70	0.71	
No	0.79	0.87	0.80	0.69	0.71	

Bueno de Mesquita, Bruce. 2011. "A New Model for Predicting Policy Choices: Preliminary Tests." *Conflict Management and Peace Science* 28 (1): 65–85.

Buhaug, Halvard, Lars-Erik Cederman, and Kristian Skrede Gleditsch. 2014. "Square Pegs in Round Holes: Inequalities, Grievances, and Civil War." *International Studies Quarterly* 58 (2). Wiley: 418–31.

Buhaug, Halvard, and Kristian Skrede Gleditsch. 2008. "Contagion or Confusion? Why Conflicts Cluster in Space." *International Studies Quarterly* 52 (2): 215–33.

Chadefaux, Thomas. 2015. "Predictably Unpredictable: The Limits of Conflict Forecasting."

Chiba, Daina, and Kristian Skrede Gleditsch. 2017. "The Shape of Things to Come? Expanding the Inequality and Grievance Model for Civil War Forecasts with Event Data." *Journal of Peace Research* 54 (2): 275–97.

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.

Choucri, Nazli. 1974. "Forecasting in International Relations: Problems and Prospects." *International Interactions* 1 (2): 63–68.

Choucri, Nazli, and Thomas W. Robinson, eds. 1979. *Forecasting in International Relations: Theory, Methods, Problems, and Prospects.* San Francisco, CA: W.H. Freeman.

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.

Freeman, John R., and Brian L. Job. 1979. "Scientific Forecasts in International Relations: Problems of Definition and Epistemology." *International Studies Quarterly* 23 (1). Blackwell Publishing on behalf of The International Studies Association: 113–43.

Gleditsch, Kristian Skrede. 2015. "Predicting Ucdp/Prio Civil Wars: Expanding the Inequality and Grievance

Model with Event Data." University of Essex; Peace Research Institute Oslo.

Gneiting, Tilman, and Adrian E. Raftery. 2005. "Weather Forecasting with Ensemble Methods." *Science* 310: 248–49.

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.

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.

Hegre, Håvard, Halvard Buhaug, Katherine V.Calvin, Jonas Nordkvelle, Stephanie T. Waldhoff, and Elisabeth Gilmore. 2016. "Forecasting Civil Conflict Along the Shared Socioeconomic Pathways." *Environmental Research Letters* 11 (5): 054002.

Hegre, Håvard, Lisa Hultman, and Håvard Mokleiv Nygård. 2011. "Simulating the Effect of Peacekeeping Operations 2010–2035." In *Social Computing, Behavioral-Cultural Modeling and Prediction*, edited by John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai., 325–32. Lecture Notes in Computer Science 6589. Springer.

Hegre, Håvard, Joakim Karlsen, Håvard Mokleiv Nygård, Håvard Strand, and Henrik Urdal. 2013. "Predicting Armed Conflict, 2010–2050." *International Studies Quarterly* 57 (2). Wiley: 250–70.

Hegre, Nils W Metternich, Håvard. 2017. "Introduction: Forecasting in Peace Research." *Journal of Peace Research* 54 (2): 113–24.

Hopple, G.W., and S.J. Andriole. 1984. *National Security Crisis Forecasting and Management.* Westview Press, Boulder, Colo.

Kupilik, Matthew, and Frank Witmer. 2018. "Spatio-Temporal Violent Event Prediction Using Gaussian Process Regression." *Journal of Computational Social Science* 1 (2): 437–51.

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, J.D., J. Kruppa, A. Dasgupta, K.G. Malley, and A. 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.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, et al. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267–81.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012a. "Ensemble Predictions of the 2012 Us Presidential Election." *PS: Political Science & Politics* 45: 651–54.

Montgomery, Jacob M., Florian Hollenbach, and Michael D. Ward. 2012b. "Improving Predictions Using Ensemble Bayesian Model Averaging." *Political Analysis* 20 (3): 271–91.

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

———. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12 (1): 87–104.

O'Loughlin, John, Frank D. W. Witmer, Andrew M. Linke, Arlene Laing, Andrew Getterman, and Jimy Dudhia. 2012. "Climate Variability and Conflict Risk in East-Africa, 1990-2009." *Proceedings of the National Academy of Sciences* 109 (45): 18344–9.

Pilster, Ulrich, and Tobias Böhmelt. 2014. "Predicting the Duration of the Syrian Insurgency." *Research & Politics* 1 (2).

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, et al. 2014. "'Beating the News' with Embers: Forecasting Civil Unrest Using Open Source Indicators." In *Proceedings of the 20th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 1799–1808.

Rost, N., G. Schneider, and J. Kleibl. 2009. "A Global Risk Assessment Model for Civil Wars." *Social Science Research* 38 (4). Elsevier: 921–33.

Schleussner, Carl-Friedrich, Jonathan F. Donges, Reik V. Donner, and Hans Joachim Schellnhuber. 2016. "Armed-Conflict Risks Enhanced by Climate-Related Disasters in Ethnically Fractionalized Countries." *Proceedings of the National Academy of Sciences of the United States of America* 113 (33). National Academy of Sciences: 9216–21.

Schrodt, Philip A. 1991. "Prediction of Interstate Conflict Outcomes Using a Neural Network." *Social Science Computer Review* 9 (3): 359–80.

Shellman, Stephen M., Brian P. Levey, and Joseph K. Young. 2013. "Shifting Sands: Explaining and Predicting Phase Shifts by Dissident Organizations." *Journal of Peace Research* 50 (3). SAGE Publications: 319–36.

Tetlock, Philip, and Dan Gardner. 2015. *Superforecasting: The Art and Science of Prediction.* New York, NY: Crown Publishers.

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.

Weidmann, N.B., and M.D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6). SAGE Publications: 883–901.

Witmer, Frank D. W., Andrew M. Linke, John O'Loughlin, Andrew Gettelman, and Arlene Laing. 2017. "Subnational Violent Conflict Forecasts for Sub-Saharan Africa, 2015â65, Using Climate-Sensitive Models." *Journal of Peace Research* 54 (2): 175–92.