Journal of Genocide Research

ISSN: 1462-3528 (Print) 1469-9494 (Online) Journal homepage: https://www.tandfonline.com/loi/cjgr20

Genocide Forecasting: Past Accuracy and New Forecasts to 2020

Benjamin E. Goldsmith & Charles Butcher

To cite this article: Benjamin E. Goldsmith & Charles Butcher (2018) Genocide Forecasting: Past Accuracy and New Forecasts to 2020, Journal of Genocide Research, 20:1, 90-107, DOI: 10.1080/14623528.2017.1379631

To link to this article: https://doi.org/10.1080/14623528.2017.1379631

Published online: 15 Oct 2017.

JOURNAL OF GENOCIDE RESEARCH, 2018
VOL. 20, NO. 1, 90–107
https://doi.org/10.1080/14623528.2017.1379631

Genocide Forecasting: Past Accuracy and New Forecasts to 2020

Benjamin E. Goldsmith [a] and Charles Butcher[b]

[a]School of Politics and International Relations, Australian National University, Acton, ACT, Australia;
[b]Department of Sociology and Political Science, Norwegian University of Science and Technology, Trondheim, Norway

ABSTRACT
We assess the accuracy of genocide forecasts made by the Atrocity Forecasting Project (AFP) for 2011–15, and present new forecasts for 2016–20. Using data from the United Nations, Genocide Watch and the Political Instability Task Force, we evaluate AFP accuracy. We compare AFP accuracy with that of forecasts from the Genocide Prevention Advisory Network. It is relatively rare in most areas of social science that researchers produce (and make public) future forecasts. It is rarer still to evaluate their accuracy once the future has arrived. AFP five-year forecasts are potentially important for genocide and politicide prevention, and have gained attention from policy makers and news media, but a systematic assessment of their accuracy has not been undertaken previously. Our evaluation of past forecast accuracy, with true-positive rates from thirty-three to fifty per cent, true-negative rates around ninety per cent, and area under the curve (AUC) statistics from .81 to .96, gives an indication of how much confidence should be placed in the 2016–20 forecasts.

ARTICLE HISTORY
Received 9 June 2017
Accepted 7 September 2017

KEYWORDS
Atrocity prevention; forecasting; genocide; politicide

Prediction is one of the most difficult and potentially most useful goals of social science. It has gained prominence recently as distinct from, and an alternative to, traditional statistical hypothesis testing.[1] Nevertheless, it is relatively rare in most areas of social science, especially outside of elections and economics, that researchers produce (and make public) future forecasts. It is rarer still to evaluate their accuracy once the future has arrived.

Prediction can have great value for policy makers if the preferred option is prevention of an event, or if a pro-active policy is more effective than a reactive one. A large number of events and policy issues conceivably fall into these categories. One that certainly does is large-scale targeted mass killing of members of ethnic or political groups, often with intended elimination of the group partially or in its entirety. For example, a 2016 United States Executive Order, which, at the time of writing, had not been rescinded by the

CONTACT Benjamin E. Goldsmith ben.goldsmith@anu.edu.au School of Politics and International Relations, Australian National University, Acton, ACT 0200, Australia

Supplemental data for this article can be accessed https://doi.org/10.1080/14623528.2017.1379631.

[1] Michael D. Ward, Brian Greenhill and Kirstin M. Bakke, "The Perils of Policy by P-value: Predicting Civil Conflicts," *Journal of Peace Research* 47, no. 4 (2010): 363–75; Philip A. Schrodt, "Seven Deadly Sins of Contemporary Quantitative Political Analysis," *Journal of Peace Research* 51, no. 2 (2014): 287–300; Adeline Lo et al., "Why Significant Variables Aren"t Automatically Good Predictors," *Proceedings of the National Academy of Sciences* 112, no. 45 (2015): 13892–7.

Trump administration, declares that "preventing mass atrocities and genocide is a core national security interest and a core moral responsibility."[2]

Predicting genocide is especially challenging because of its relative rarity, making a large number of false positives very likely. But the potential benefits are great, if a small enough list of at-risk cases can be produced to make monitoring and prevention efforts practical.[3] In this article, we assess the accuracy of one set of forecasts of genocide and politicide[4] covering the period 2011–15, produced by the Atrocity Forecasting Project (AFP). These have gained policy makers' attention in the US, Europe and elsewhere, and have been featured in major news media,[5] but a systematic assessment of their accuracy has not been undertaken previously. We find that the forecasts, while far from perfect prediction, demonstrate reasonable accuracy in standard metrics. Given this, we use a similar, updated method to produce forecasts of genocide and politicide for the period 2016–20. We also suggest avenues for further improving genocide forecasting. We emphasize that we focus our contribution exclusively on the challenging and less familiar task of assessing forecasting accuracy, rather than on theory building or attempted causal inference. This article is not about what causes genocide; it is about how to best predict it.

A Brief Primer on Forecasting Method

How might one construct a quantitative forecasting model of something like genocide? Given the relative novelty of such forecasting applications in social sciences, it is worth highlighting key aspects of general forecasting methods. There are at least three challenges that make the task of forecasting different from standard quantitative analysis. First, the information of ultimate interest is necessarily unknown. That is, the goal is to anticipate future events, while relying only on data that exist in the present. Second, because the goal is to find the most powerful combination of predictor variables that is most closely tied to the outcome of interest, in this case genocide, there is a danger that the forecaster will focus on unusual or idiosyncratic aspects of the known data that are only related to the outcome by chance or temporary circumstance. A very close fit to presently known data can be engineered through trial and error, but the variables used will probably turn out to be poor predictors in the future. This type of "overfitting" to extant data can lead to poor future forecasts based on ultimately irrelevant predictors. Third, standard ways of assessing and interpreting quantitative models in the social sciences, such as attributing "statistical significance" to regression coefficients with low standard errors or focusing on model-fit statistics such as R-squared (R^2) or Akaike's information criterion (AIC), turn out not to be the best guides to predictive accuracy.[6] Different

[2] The White House, *Comprehensive Approach to Atrocity Prevention and Response*, May 18, 2016, https://www.whitehouse.gov/the-press-office/2016/05/18/executive-order-comprehensive-approach-atrocity-prevention-and-response (accessed June 7, 2016).

[3] Sascha Nanlohy, Charles Butcher and Benjamin E. Goldsmith, "The Policy Value of Quantitative Atrocity Forecasting Models," *RUSI Journal* 162, no. 2 (2017): 24–32.

[4] In this article, we use the term "genocide" to represent both genocide and politicide. This is for ease of reading and is not meant to diminish the importance of politicide. All forecasts and analyses based on the work of AFP and Harff and Gurr include combined data for genocide and politicide.

[5] The AFP website lists coverage in the *New York Times* (March 22, 2014) and elsewhere, and presentations to the Council on Foreign Relations, the German Institute for International and Security Affairs, and the International Crisis Group. See politicsir.cass.anu.edu.au/research/projects/atrocity-forecasting (accessed September 13, 2017).

[6] Lo et al., "Why Significant Variables Aren''t Automatically Good Predictors."

ways of judging the power of individual predictors and the overall predictive performance of a model are needed.

The most common approach used by forecasters to address these challenges is called "out-of-sample" prediction. This is an intuitive procedure that allows the forecaster to simulate the future forecasting process while using only presently available data. There are numerous ways to approach it, but the fundamental elements are the division of the existing data into two samples, one for "training" a forecasting model and the other for testing it, and the use of predictive performance in the test sample as the metric for assessing forecasting accuracy.

For example, in order to build a model to forecast US presidential elections, one could gather data for predictors such as the party of the incumbent president, whether the incumbent president was running for re-election, the economic growth and unemployment rates in the election year, whether the country was at war, and the results of the most recent Congressional election. The outcome to be predicted could be the vote share of the incumbent president's party. If we are able to obtain data for the predictors and outcome in every year from, say, 1916, when Woodrow Wilson was the Democratic candidate and Charles Hughes was the Republican—the incumbent party—through 2016, this full dataset could be divided for the purpose of constructing a forecasting model using the out-of-sample approach.

A common way to split the full dataset into training and testing samples is temporally. We could define our training sample as all elections from 1916 to 1980, or roughly two-thirds of the data. That would leave elections from 1984 onwards for testing. We could use a basic quantitative method, linear regression, to fit a model using our predictors to the training data. Crucially, we will be concerned not with the statistical significance of the coefficients for the predictor variables, nor with the overall R^2 of the model, but with how close the predicted incumbent party vote share matches that in the actual election. This will be a function of the predictors included in the model, and the coefficients or relative magnitudes of their contributions to the prediction, which is what the linear regression technique estimates.

We can try to improve the model's performance within the training sample, that is, to maximize *in-sample* predictive accuracy. For example, we might decide to add a predictor that represents economic growth over the last four years, since that is the president's term in office that he might be held accountable for by voters. We might decide to include the effective tax rate of the median voter, and also discount the current economic performance in some way by that tax rate, if for example we believe that high taxes in bad times will be especially damaging to the incumbent. Of course, adding new predictors beyond our original expectations, to see if they improve in-sample prediction, risks overfitting the model to unusual characteristics of the training data. The test sample is key to guard against overfitting.

To test our model of presidential vote share and assess its out-of-sample performance, we preserve the coefficient values for each predictor that resulted from the in-sample training. We then use the same linear regression technique, with these pre-set predictor coefficients or relative magnitudes, for producing predictions for the out-of-sample elections, 1984–2016, using the out-of-sample predictor data. The model will produce predicted incumbent party vote shares for each election and these can be directly compared to the actual vote shares from the elections to assess predictive accuracy.

Statistics such as Mean Absolute Error (MAE) of the predictions can be produced, assessing how far off we were on average, above or below the actual results. If there is another forecaster with a different forecasting model, then that model's out-of-sample MAE can be compared to ours to suggest which will provide more accurate forecasts. Of course, the ultimate aim is to predict future election outcomes, but without this out-of-sample approach we would have little hope of assessing which models were likely to perform well, and which were hobbled by overfitting to existing data.

Several political scientists have produced forecasts along these lines for recent US presidential elections, and there is now a regular, quadrennial effort to assess their forecasts, discuss and improve the models.[7] The example of US presidential elections is instructive and highlights an area in which forecasting has become more common, if not always of desired accuracy. Forecasting a phenomenon like genocide has some distinct challenges, and can probably be said to be overall a more difficult task. One distinction is that genocides are discrete events, so we do not have data that measure the percentage or degree of genocide in a given country in a particular year. Rather, we typically have data that record either "yes" such an event occurred, or "no" it did not. Forecasting such binary outcomes potentially increases the difficulty because a quantitative model will nevertheless need to calculate an underlying probability or risk level for each case.

Another distinction is that genocides are (thankfully) quite rare events. In the most commonly used quantitative dataset, that developed by Barbara Harff and Ted Robert Gurr,[8] genocides erupt on average less than once per year globally for the period since 1955. This leads to a problem sometimes called "unbalanced data" in which the frequency of "no" events approaches perhaps ninety per cent or even ninety-nine per cent of all observations, while the "yes" events, the things we want to predict, occur in roughly 10 per cent or less of cases in the data. This "rare events" problem is an enduring challenge for any type of quantitative analysis, for a number of reasons. When combined with the challenges of predicting the future and uneven predictor data quality, the task of genocide forecasting can be understood to present significant challenges. Specifically, the frequency of genocide onsets is about 0.5 per cent of country-years. For comparison, using standard databases,[9] civil war onsets occur in about 1.8 per cent of country-years and coups d'état (successful or failed) in about 6.2 per cent. Nevertheless, there have been attempts, and here we focus mainly on that of the AFP, with some comparison to forecasts produced by Harff and Gurr.

The 2011–15 Forecasts

The AFP presented its forecast for 2011–15 in a report and on its website in August 2012.[10] This consisted of a list of the fifteen countries most at risk of the onset of genocide or poli-

[7] James E. Campbell, "Forecasting the 2016 American National Elections: Introduction," *PS: Political Science and Politics* 49, no. 4 (2016): 649–54; James E. Campbell, "How Accurate Were the Political Science Forecasts of the 2016 Presidential Election?" *Sabato's Crystal Ball*, University of Virginia Center for Politics, November 17, 2016, http://www.centerforpolitics.org/crystalball/articles/how-accurate-were-the-political-science-forecasts-of-the-2016-presidential-election/ (accessed June 9, 2017).

[8] Barbara Harff and Ted Robert Gurr, "Toward Empirical Theory of Genocides and Politicides: Identification and Measurement of Cases since 1945," *International Studies Quarterly* 32, no. 3 (1988): 359–71.

[9] Civil war data are from the Major Episodes of Political Violence (MEPV) dataset, and coups data are from datasets by Marshall and Marshall, and Powell and Thyne. Full reference details are found in the online supplementary materials.

[10] Atrocity Forecasting Project (AFP), *Forecasts to 2015* (2012), politicsir.cass.anu.edu.au/research/projects/atrocity-forecasting; Charles R. Butcher et al., *Understanding and Forecasting Political Instability and Genocide for Early Warning*, March

Table 1. Forecast for 2011–15: top fifteen countries at risk of the onset of genocide or politicide.

1	Central African Republic
2	Democratic Republic of the Congo
3	Chad
4	Somalia
5	Angola
6	Myanmar
7	Sri Lanka
8	Ecuador
9	Burundi
10	Afghanistan
11	Syria
12	Guinea
13	Cameroon
14	Uganda
15	Libya

ticide for the period, reproduced in Table 1. The definition of genocide/politicide developed by Barbara Harff was used, and outcome data on genocide/politicide came from the Political Instability Task Force's dataset (PITF)[11] based on Harff's coding guidelines. A generalized additive model (GAM) semiparametric statistical approach[12] was applied with data on a range of predictors to generate the forecast.

Genocide and politicide are defined by Harff[13] as:

> … the promotion, execution, and/or implied consent of sustained policies by governing elites or their agents—or, in the case of civil war, either of the contending authorities—that are intended to destroy, in whole or part, a communal, political, or politicized ethnic group. In genocides the victimized groups are defined by their perpetrators primarily in terms of their communal characteristics. In politicides, in contrast, groups are defined primarily in terms of their political opposition to the regime and dominant groups.

The PITF data for genocides and politicides record events from 1955 to 2015. Figure 1 shows both the total number of events ongoing in each year, and new event onsets. The AFP attempts to forecast these onsets of new instances of genocide and politicide.

AFP predictor variables can be divided into structural, slow-changing factors such as ethnic divisions, infant mortality rates and political institutions, and factors with greater temporal variance such as elections, conflicts in neighbouring states and the use of guerrilla war tactics. The AFP claims that inclusion of such time-sensitive predictors, along with use of an unconditional model producing forecasts for all states in the international system, are among the strengths of its approach contributing to improved forecasting

2012, politicsir.cass.anu.edu.au/research/projects/atrocity-forecasting/publications (accessed September 13, 2017). The AFP 2016–20 forecasts are also available at http://politicsir.cass.anu.edu.au/research/projects/atrocity-forecasting/forecasts (accessed September 13, 2017).

[11] Monty G. Marshall, Ted Robert Gurr and Barbara Harff, *PITF-State Failure Problem Set: Internal Wars and Failures of Governance, 1955–2014*, May 6, 2015, http://www.systemicpeace.org/inscrdata.html (accessed August 7, 2017). Harff and Gurr have clarified for us that they have not been involved with PITF since 2010, so coding of cases was done independently of them from that point onwards (personal communication, June 24, 2016).

[12] Trevor Hastie and Robert Tibshirani, *Generalized Additive Models* (London: Chapman & Hall/CRC, 1990); Nathaniel Beck and Simon Jackman, "Beyond Linearity by Default: Generalized Additive Models," *American Journal of Political Science* 42, no. 2 (1998): 596–627.

[13] Harff and Gurr, "Toward Empirical Theory," 360. The authors have clarified that the definition presented in this article was developed solely by Harff (personal communication, June 24, 2016).

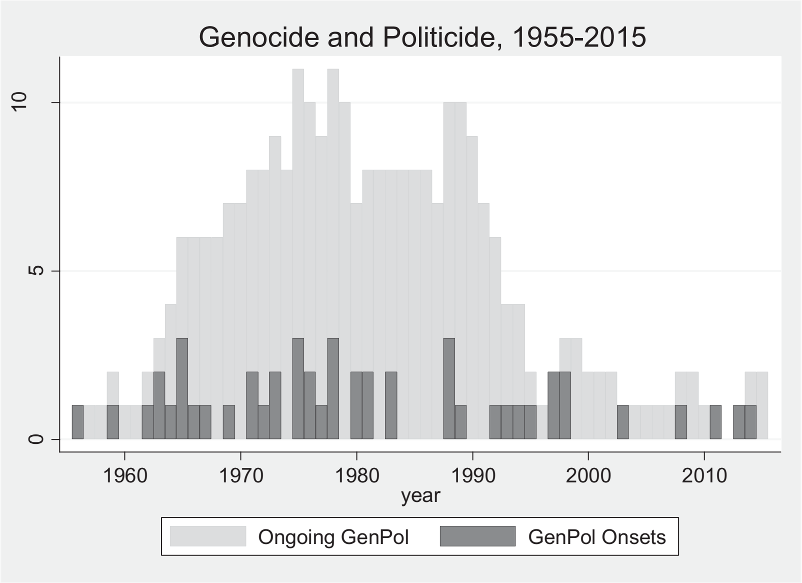

Figure 1. Onsets and Ongoing Cases of Genocide and Politicide, 1955–2015.

accuracy.[14] Nineteen predictors were measured, 1974–2010, and forecasts were produced for the subsequent five-year period. The AFP reports a procedure for developing forecasts involving three basic steps: (1) assemble a set of predictors based on existing scholarly literature and original theorizing; (2) train the model on a sample of earlier data to achieve very good in-sample fit or prediction; and then (3) test the model on a sample of later data to assess its out-of-sample predictive performance for events not used to develop the model. For example, their original model was trained on data for 1955–87, and tested on data for 1988–2003.[15] An important point is that any model such as the AFP's with the primary aim of maximizing forecast accuracy may diverge from models specified for causal inference or other forms of theoretical analysis, and should not be interpreted in those ways.[16]

Since the forecast period has now passed, we are in a position to assess the AFP's actual forecasting performance. In the 2011–15 forecasts shown in Table 1, countries are listed in descending order of risk, although the AFP presents the fifteen as a group to be most at risk, relative to all others for which forecasts were produced, ranked 16–142 (see the supplementary material).[17]

Of the fifteen cases in Table 1, we believe that the Central African Republic (CAR), Libya, Syria and Myanmar came closest to actual genocide or politicide onset over the period. CAR and Myanmar, we believe, seemed particularly counter-intuitive in 2011 and 2012,

[14] Benjamin E. Goldsmith et al., "Forecasting the Onset of Genocide and Politicide: Annual Out-of-Sample Forecasts on a Global Dataset, 1988–2003," *Journal of Peace Research* 50, no. 4 (2013): 437–52.

[15] Butcher et al., *Understanding and Forecasting Political Instability*; Goldsmith et al., "Forecasting the Onset of Genocide and Politicide."

[16] Lo et al., "Why Significant Variables Aren"t Automatically Good Predictors."

[17] Our initial assessments were produced in February 2016, and are very similar to those presented here. We have expanded and refined the assessment and also included new PITF data which only became available in mid 2017.

but proved to be at considerable risk. Anecdotally, these cases point to the potential value of the list, and in general to lists developed using rigorous, systematic quantitative approaches, rather than qualitative judgement. For example, it was not until after a destabilizing coup in 2013 that the International Crisis Group[18] began to signal serious concern about CAR. The AFP approach "saw" the risk in this case based only on data up to 2010, placing it at the top of the list.

The "Ground Truth"

More generally, forecasts can be evaluated against the "ground truth" of actual events, once the period to which the forecasts applied has passed. We selected three indicators of onsets of genocide or politicide, the ground truth for the AFP forecasts. Because these are rare events, occurring on average less than once per year or in about 0.5 per cent of country-years, and coding is uncertain for some cases, we sought not only post-2010 events coded by PITF, but also other indicators against which to test forecast accuracy. Because PITF had coded genocide and politicide events for the period only up to 2014 until mid 2017, we also wanted to find data that could be used in a more timely way to continually assess genocide forecasting. Two further sources seemed reasonably appropriate: warnings issued by the United Nations Special Advisers for the Prevention of Genocide and the Responsibility to Protect[19] and the categorizations of the non-governmental organization Genocide Watch.[20] We coded all UN warnings ("statements") identifying a particular country, or part of a country, as at risk of violence relevant for genocide prevention or invoking the responsibility to protect a population at risk (R2P). When there were two or more warnings per country per year, we counted this as one warning. We excluded warnings that did not specify a country or countries as being at risk, for example general warnings against hate speech by religious figures, and we excluded other statements that did not give specific warnings such as those recognizing investigations or commemorations of historical atrocities. We coded Genocide Watch cases as onsets when a country reached Stage 9 ("extermination") in their framework from a lower stage in a given year. Table 2 shows the events we coded across all three indicators of ground truth.

Evaluations

Basic criteria for evaluating forecasting performance are relatively straightforward, although a wide range of approaches exist depending on types of forecasts and ground truth events. For applications in which the outcome of interest is binary, as in the case of genocide, measures such as Mean Absolute Error are not applicable, as noted. Rather, we can classify categories of correct and incorrect predictions. Fundamentally, we can assess whether something happens that was predicted, and whether nothing happens

[18] International Crisis Group (ICG), *Central African Republic: Priorities of the Transition. Africa Report No. 203*, June 11, 2013, http://www.crisisgroup.org/en/regions/africa/central-africa/central-african-republic/203-central-african-republic-priorities-of-the-transition.aspx.

[19] See http://www.un.org/en/preventgenocide/adviser/ (accessed February 7, 2016). The website was reconfigured in 2017 and the new relevant URL is: http://www.un.org/en/genocideprevention/public-statements.html. Indeed, Harff was involved in developing the UN"s criteria to identify cases (personal communication, June 24, 2016).

[20] See http://genocidewatch.net/ (accessed February 7, 2016). We emphasize that neither the UN warnings nor the Genocide Watch cases were identified using statistical analysis or other forecasting techniques, but by ex post qualitative assessment of recent or current events, and therefore are appropriate to treat as "ground truth" in our exercise.

Table 2. Ground truth events, 2011–15.

Country	UN Warnings	Genocide Watch Onsets	PITF Onsets
Burundi	2015		
Myanmar	2015		
Syria	2015		
Yemen	2015		
Central African Republic	2014		
Iraq	2014	2014	2014
Israel	2014		
Nigeria		2014	
South Sudan	2014	2014	
Syria	2014		
Central African Republic	2013		2013
Egypt	2013		
Mali	2013		
Myanmar	2013		
South Sudan	2013		
Syria	2013		
Syria	2012		
Democratic Republic of the Congo		2011	
Ivory Coast	2011		
Libya	2011	2011	
Sudan	2011	2011	
Syria	2011	2011	
Uganda		2011	
Yemen		2011	

Table 3. Contingency table of forecasting assessment categories.

		Ground Truth Observation	
		No Onset	Genocide Onset
Forecast	No Onset	(true negatives)	(false negatives)
	Genocide Onset	(false positives)	(true positives)

when nothing was predicted. In other words, we look for true and false positive predictions, and true and false negative predictions. These can be arranged in a contingency table (confusion matrix) as in Table 3, providing four categories for assessing accuracy.

A major challenge with rare-events (unbalanced) data is to predict a good number of "yes" (genocide/true positive) outcomes correctly, while not also capturing a very high number of the much more common "no" (non-genocide/false positive) events. When fewer than 1 in 100 of the observations in the data are actual genocides, developing a model that will reliably place those cases near the top of a list of at-risk countries for each year in out-of-sample testing is a major challenge.

In what follows, we use three approaches to assess forecasting performance, briefly explaining the rationale behind each approach as we go. First, we do a simple assessment of all four possible categories for the 142 states on the AFP list, with a focus on true-positive and true-negative rates, considering the fifteen most highly ranked, listed in Table 1, as positive forecasts, and the remaining 127 as negative forecasts. Second, we compare false positives and true positives for our forecasts and those made public by Harff and Gurr for several years.[21] Third, we use so-called Receiver-Operating-Characteristic (ROC)

[21] Genocide Prevention Advisory Network, *Barbara Harff's Risk Assessments* (2015), http://www.gpanet.org/content/barbara-harffs-risk-assessments (accessed February 7, 2016).

analysis to assess the accuracy of AFP forecasts across all 142 states to which they assigned a risk score.

The AFP's forecasting outcomes for true and false positives and true and false negatives across the three outcome indicators are presented in Table 4.

While assessing each of the four categories separately can tell us that the AFP did not achieve perfect prediction, i.e. zero false positives and zero false negatives, it is hard to assess accuracy from these four categories alone. Somewhat more informative are the true-positive rate (also called recall or sensitivity) and the true-negative rate (specificity). The true-positive rate is the proportion of correctly predicted positives. For the AFP's prediction of Genocide Watch onsets, this would be four out of nine, or forty-four per cent, for example. The true-negative rate would be 122 out of 133, or ninety-two per cent. The corresponding rates for UN warnings are thirty-three per cent and ninety-two percent, and for PITF onsets, fifty per cent and ninety per cent. These are included in the lower rows of Table 4, and tell us roughly that the AFP list anticipated from a third to half of the relevant events, and also anticipated nine out of ten non-events.[22]

Given the difficulty of the task, we find the true-positive and true-negative rates encouraging. But the frequency of false positives and false negatives highlights the need to consider any such forecasts as indicative, not definitive. Nevertheless, the value of identifying otherwise non-obvious cases is high, and a distinct advantage of global, quantitative approaches such as the AFP's.

It is also important to assess forecasts in comparative perspective, to understand how good or useful they are relative to other available sources of early warning. In the next section, we undertake a limited comparison of AFP forecasts with the only other quantitative forecasts of genocide and politicide of which we are aware.[23]

Comparison

There is no other set of genocide forecasts that makes exactly the same type of predictions as the AFP: genocide/politicide onsets over a five-year period. Nevertheless, it is important to attempt comparison because policy makers and others seeking to prevent genocide should be interested in the relative accuracy of existing lists of at-risk countries.

The forecasts of Harff and Gurr posted to the Genocide Prevention Advocacy Network website[24] use the same outcome variable, onsets of genocide or politicide, but produce annual forecasts. While the AFP produces a list of fifteen at-risk countries over the five-year period, Harff and Gurr produce a list of seventeen to twenty countries at risk over each one-year period.

[22] No genocide risk was estimated for South Sudan, which became independent in 2011 while AFP forecasts for 2011–15 were based on data from 2010.

[23] Other forecasting efforts focus on substantially different outcomes, such as "state-led mass killing" (see the review: Ernesto Verdeja, "Predicting Genocide and Mass Atrocities," *Genocide Studies and Prevention* 9, no. 3 [2016]: 13–32). Perhaps the most prominent of such efforts is the Early Warning Project (https://www.earlywarningproject.org/definitions), which focuses on all mass killing events with 1,000 or more non-combatant civilian deaths.

[24] Accessed 7 February 2016 at http://www.gpanet.org/content/barbara-harffs-risk-assessments; further details provided in the supplementary materials.

Table 4. AFP forecasting performance across three outcome indicators.

Forecast for 2011–15: Top 15 Countries at Risk of the Onset of Genocide or Politicide		UN Warnings	Genocide Watch Onsets	PITF
1	Central African Republic	2013, 2014		2013
2	Democratic Republic of the Congo		2011	
3	Chad			
4	Somalia			
5	Angola			
6	Myanmar	2013, 2015		
7	Sri Lanka			
8	Ecuador			
9	Burundi			
10	Afghanistan			
11	Syria	2011–2015	2011	
12	Guinea			
13	Cameroon			
14	Uganda		2011	
15	Libya	2011	2011	
	AFP True Positives	4	4	1
	AFP False Positives	11	11	14
	AFP False Negatives	8	5	1
	AFP True Negatives	119	122	126
	True-positive rate	33%	44%	50%
	True-negative rate	92%	92%	90%

Notes: UN warnings false negatives include South Sudan.

We limit our comparisons to true positives and false positives, because the negative-case lists are not available for the Harff and Gurr forecasts.[25] We discuss two types of comparisons: annual comparisons treating the AFP's list as distinct annual forecasts with the same fifteen countries for each year, and five-year forecasts treating Harff and Gurr's 2011 list as covering 2011–15.

Harff and Gurr produced forecasts for 2011, 2013 and 2015. We take each in turn to compare annual true and false positives, comparing results based on the UN warnings, since that gives the largest number of cases to work with. We also briefly note performance for the Genocide Watch and PITF data.

The two cases for which UN warnings were produced in 2011 that also appeared on the Harff and Gurr list were Sudan and Syria.[26] Since there were twenty countries in their list, there are eighteen false positives. The AFP list identified Syria and Libya, which were the subject of UN warnings, so the AFP also had two true positives. Since the AFP list contains fifteen countries, the number of false positives was thirteen. In 2013, Harff and Gurr had two true positives, Myanmar and Syria, and fifteen false positives. The AFP had three true positives, Myanmar, Syria and the Central African Republic, and twelve false positives.

[25] We contacted them but they could not provide the full lists on which the publicly available forecasts were based. This limits the types of comparisons we can make. False negatives in this context are costly, so comparing what proportion of false positives are required by each list to achieve a desired true-positive rate might be a good indicator of the usefulness of competing forecasts (see Ryan Kennedy, "Making Useful Conflict Predictions: Methods for Addressing Skewed Classes and Implementing Cost-Sensitive Learning in the Study of State Failure," *Journal of Peace Research* 52, no. 5 [2015]: 649–64). Without the full rankings from Harff and Gurr’’s forecasts, we cannot do this here. Harff and Gurr did provide recently created full lists of forecasts using Cox hazard models. The resulting annual lists of at-risk countries, 2011–15, were substantially different from the publicly available forecasts at GAPNet, however.

[26] We take the Harff and Gurr forecasts as pertaining to 2011 because the title of the list is "Country Risks of Genocide and Politicide in 2011." However, the subtitle of the chapter containing the list is "A Global Watch List for 2012," and it is not clear from the text which year to choose. If we choose 2012, then the comparative results are very similar: both Harff and Gurr and AFP have one true positive, Syria.

Table 5. Comparisons of annual and five-year forecasts for UN warnings.

	2011	2013	2015	2011–15
Harff & Gurr				
True positives	2	2	3	4
False positives	18	15	17	16
AFP				
True positives	2	3	3	5
False positives	13	12	12	10

Notes: 2011–15 forecast uses Harff and Gurr's 2011 list.

For 2015, Harff and Gurr had three true positives, Myanmar, Syria and Yemen, and seventeen false positives, while the AFP had three true positives, Myanmar, Syria and Burundi, and twelve false positives. These are summarized in Table 5.

Overall, then, for these three years, Harff and Gurr have seven true positives and forty false positives. Of all their at-risk countries, 14.9 per cent received UN warnings.[27] The AFP has overall eight true positives for the three years and thirty-seven false positives, giving 17.8 per cent of cases on the AFP list as true positives (counted as a new list in each year).

If we just focus on true positives for the UN warnings for these three years, the true-positive rate as described in the previous section is seven of fourteen, or fifty per cent, for Harff and Gurr, and eight of fourteen, or fifty-seven per cent, for the AFP.

Looking briefly at the Genocide Watch and PITF ground truth data, Harff and Gurr's 2011 list has twenty per cent true positives for Genocide Watch in 2011 (zero per cent for other years) and five per cent true positives for PITF's 2013 CAR case (zero per cent if their 2013 list is used). The AFP's list has twenty-seven per cent true positives for Genocide Watch onsets and seven per cent true positives for PITF genocide/politicide onsets. Thus, the AFP shows marginally but consistently higher accuracy across all annual indicators of ground truth.[28]

However, AFP forecasts are meant to cover a five-year period. If we take the 2011 Harff and Gurr forecasts as also applicable to the entire period, 2011–15, we can compare based on this standard (final column of Table 5). Of the thirteen countries for which there was at least one UN warning, 2011–15, Harff and Gurr's 2011 list identifies four (Myanmar, Syria, Central African Republic and Sudan) while the AFP list identifies five (Myanmar, Syria, Burundi, Central African Republic and Libya). This equates to true-positive rates of twenty per cent for Harff and Gurr and thirty-three per cent for the AFP.

While the first (annual) comparison method in principle favours the annual Harff and Gurr results, the second favours the five-year AFP results. The AFP does marginally better in either comparison, although there is no statistically significant difference.[29] What does seem clear is that the AFP's approach of producing forecasts for a five-year period, and focusing on a shorter list of at-risk countries, does not yield fewer true positives. That is, if the goal is to provide as short a list as possible without sacrificing accuracy, the AFP's approach achieves this better than that of Harff and Gurr. Because two true

[27] This statistic is sometimes called precision or positive predictive value. It is calculated as: True Positives / (True Positives + False Positives).

[28] We note that the higher percentages for AFP forecasts are due to the lower false positives. Because the AFP list is shorter, the forecasts have greater "precision." Numbers of true positives are equal: four for Genocide Watch and one for PITF. But see also note 30.

[29] We used chi-squared and Fisher''s exact tests to assess this. See supplementary materials.

positives (CAR and Uganda) occur outside the top fifteen in Harff and Gurr's 2011 list, while in no case do their lists have more true positives than the AFP, the AFP approach appears more suited to producing a short list.[30] Practically, the five-year forecasting method also has the advantage of giving a longer time-span in which to attempt prevention, and identify and monitor at-risk cases.

ROC Analysis

Another way to assess forecasting accuracy is to consider all predictions (positive and negative) across all known outcomes. We believe this is preferable, but we only have access to the full list of AFP risk scores, so we do not undertake a comparison. Harff and Gurr produce their forecasts for a conditional sample limited to countries experiencing political instability (in 2011), or produce separate assessments with separate rankings for those with and without instability (2013, 2015), which precludes a balanced comparison.[31] We briefly discuss comparisons using the Harff and Gurr conditional sample at the end of this section.

Receiver-Operating-Characteristic (ROC) analysis assesses all possible prediction thresholds by plotting the corresponding true-positive (sensitivity) and false-positive (1-specificity) rates.[32] ROC analysis is especially helpful because it provides a relevant and intuitive metric that can be compared across models and often across applications (although other metrics exist and have their advantages). The area under the curve (AUC) measures the portion of the graph captured under the ROC curve. An AUC of 0.5 indicates prediction no better than chance; an AUC of 1.0 indicates perfect prediction. The AFP's previous forecasts achieved AUC statistics for out-of-sample forecasting of .8878 (for the period 1988–2003[33]) and .9218 (for the period 1990–2010[34]).

Here we present ROC results based on AFP risk scores produced via a GAM for all 142 countries in the AFP dataset for the 2011–15 forecasts. Outcomes are the UN warnings, Genocide Watch onsets and PITF onsets shown in Table 4, scored 1 for onsets 2011–15, 0 otherwise. AUC scores are .8557 for UN warnings (Figure 2), .8088 for Genocide Watch onsets (Figure 3) and .9643 for PITF genocide/politicide onsets (Figure 4).[35]

That the AUC statistic is somewhat lower for the ground truth data that are related to, but not the same as, the genocide/politicide data the AFP model was trained on, is not surprising. It is moderately encouraging that the AFP forecasting model yields an AUC

[30] While retrospectively adjusting a forecast based on known outcomes is not a legitimate way to evaluate that forecast, it is nevertheless informative to consider how Harff and Gurr''s lists perform when limited to their top fifteen cases, as AFP forecasts are. Some true positives would be lost. For UN warnings, Harff and Gurr''s 2011 forecast ranks CAR 16th, so the 2011–15 forecast would lose one true positive. For Genocide Watch, Harff and Gurr''s 2011 forecast ranks Uganda 17th, so one true positive is lost for both the 2011 and 2011–15 forecasts. For PITF, Harff and Gurr would lose their one (and only) true positive for 2011–15 (again due to CAR''s 16th ranking for 2011).

[31] See note 25. For attempts at such comparisons, see Goldsmith et al., "Forecasting the Onset of Genocide and Politicide."

[32] On ROC, see David W. Hosmer and Stanley Lemeshow, *Applied Logistic Regression*, 2nd ed. (New York: Wiley, 2000), chap. 5.

[33] Goldsmith et al., "Forecasting the Onset of Genocide and Politicide."

[34] Butcher et al., *Understanding and Forecasting Political Instability*.

[35] We encounter a complication with the data at this stage. When we re-run the GAM script used by the AFP to produce its 2011–15 forecasts, we are not able to precisely replicate their forecasts. Consultation with AFP team members who wrote and ran the script leads us to believe that this is due to software updates in the R platform and packages used. The results are similar, but we provide further assessment of the comparability of our replication to the AFP forecasts in Table A2 in the supplementary materials. The original list of forecasts with risk scores for each country was not saved by AFP, so only the R scripts, data files and ordinal top-fifteen ranking remain.

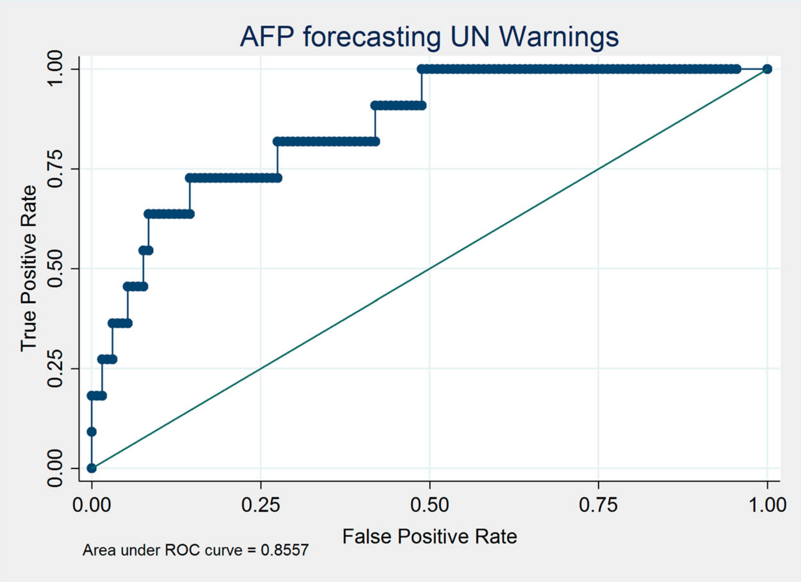

Figure 2. ROC AUC for AFP Forecasts of UN Warnings.

above 0.80 for cases receiving UN warnings and also cases from Genocide Watch. Of course, the AUC for forecasting the PITF data is quite high, and this is also encouraging.

We can also compare the performance of the AFP forecasts with forecasts supplied to us by Harff and Gurr. While these are substantially different from their publicly available forecasts, they rely on an updated method, Cox hazard models, which Harff and Gurr prefer (see note 25 and supplementary materials). For the 2011 forecast, both the old

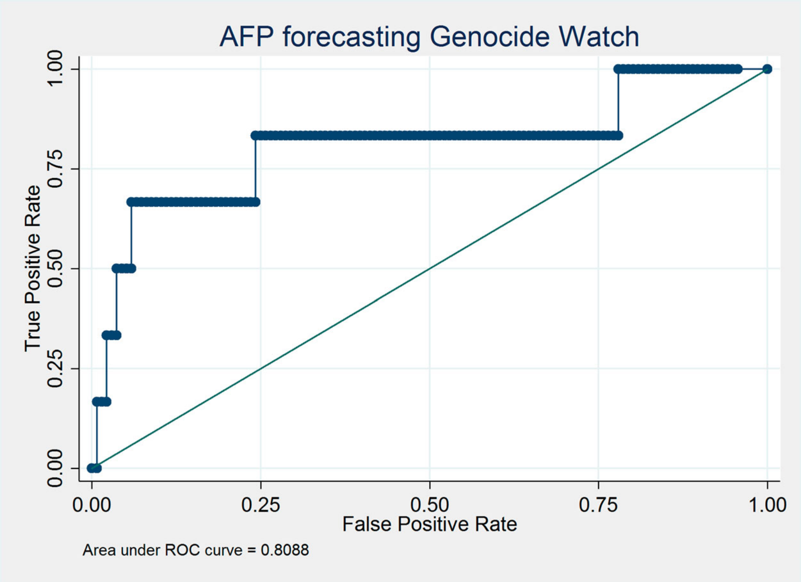

Figure 3. ROC AUC for AFP Forecasts of Genocide Watch Onsets.

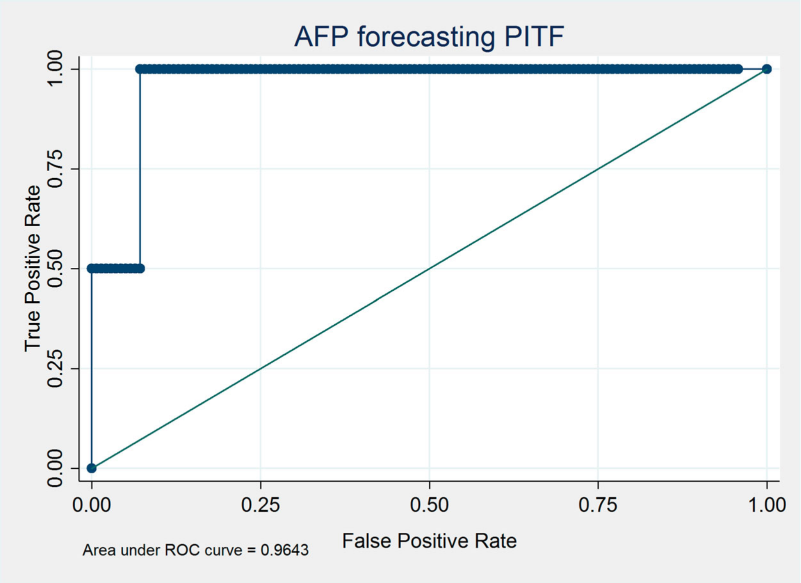

Figure 4. ROC AUC for AFP Forecasts of PITF Onsets.

and new lists contain twenty countries, but there is an overlap of only eight. Using only this conditional sample of twenty cases, Harff and Gurr do a modestly better job forecasting UN warnings, 2011–15, with an AUC of .7292 to the AFP's .6771, and a slightly better job forecasting the Genocide Watch data (.4267 versus .4000). But the AFP do much better with PITF events (.9444 versus .5278). Given the small numbers involved, we are hesitant to place too much stock in these ROC comparisons, but they do show that even in a conditional sample defined by Harff and Gurr's criteria, the AFP's forecasts perform about as well or better.

With these reasonably good initial indications of AFP forecasting performance, we can have a better idea of how much confidence to place in AFP forecasts moving forward. With this in mind, we produced forecasts with the same basic approach for the next five-year period.

New Forecasts, 2016–20

Our new forecasts use the AFP method and updated data (Table 6 and Figure 5). Given reasonably good performance for 2011–15, we expect similar performance from these forecasts. We might even see better performance, because while using the same GAM approach, our data are more complete and we have made some adjustments regarding new states and other data issues to exploit the predictor data more fully. We use a generalized additive model with a logit link. We train the model on data for the period 1955–2014, which is an improvement in that we extend the training period two decades deeper into the past, relative to the AFP's previous forecast. We then use data for predictors in 2015 as the basis for our 2016–20 forecast. In the supplementary materials, we provide full methodological and data details, while here we present only the model structure including all predictors, and the forecast it produces.

Table 6. Forecast for 2016–20: top fifteen countries at risk of the onset of genocide or politicide.

1	South Sudan
2	Sudan
3	Iraq
4	Nigeria
5	Yemen
6	Syria
7	Afghanistan
8	Somalia
9	Russia
10	Libya
11	Mali
12	Central African Republic
13	Pakistan
14	Egypt
15	Algeria

The forecasting model can be written as:

$$\begin{aligned}
&\mathrm{GAM}_{\mathrm{logit}}(\mathrm{GPonset}_{t+1\ \mathrm{through}\ t+5}) = \beta_0 + f_1(\mathrm{Regime}) + f_2(\mathrm{Population}_{\mathrm{ln}}) + f_3(\mathrm{InfantMortality})\\
&\quad + \beta_4(\mathrm{EthnicFractionalization}) + f_5(\mathrm{EthnicFractionalization}) + \beta_6(\mathrm{StateLedDiscrimination})\\
&\quad + \beta_7(\mathrm{MENA}) + \beta_8(\mathrm{CS\_Asia}) + \beta_9(\mathrm{Instability}) + f_{10}(\mathrm{RegimeChange}) + \beta_{11}(\mathrm{NeighborConflict})\\
&\quad + \beta_{12}(\mathrm{GuerrillaTactics}) + \beta_{13}(\mathrm{InternationalizedCivilWar}) + \beta_{14}(\mathrm{InterstateWar})\\
&\quad + \beta_{15}(\mathrm{Election}_{t\ \mathrm{through}\ t+2}) + \beta_{16}(\mathrm{Election}_{t+1\mathrm{through}\ t+3}) + \beta_{17}(\mathrm{Election}_{t+2\ \mathrm{through}\ t+4})\\
&\quad + f_{18}(\mathrm{noGPyears}) + f_{19}(\mathrm{Time})
\end{aligned}$$

Concluding Comments

We have assessed the accuracy of the AFP's 2011–15 forecasts of genocide and politicide, and found reasonably good performance predicting three distinct measures of ground truth, and against performance of forecasts in the same period by Harff and Gurr.

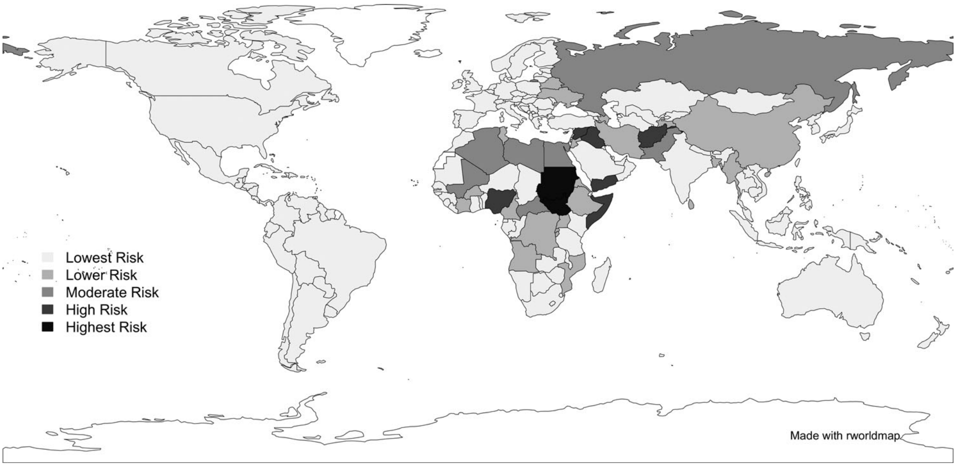

Figure 5. Map of predicted genocide/politicide risk, 2016–20.

Note: The risk categories correspond to the following: Highest Risk = top one per cent of states ranked by risk; High Risk = top five per cent of states ranked by risk; Moderate Risk = top ten per cent of states ranked by risk; Low Risk = top twenty-five per cent of states ranked by risk; Lowest Risk = outside of the top twenty-five per cent ranked by risk.

Whether onsets of genocides and politicides were measured by UN warning statements, Genocide Watch ratings, or coding by the Political Instability Task Force, the AFP's short list of fifteen at-risk countries captured between a third and half of the onsets over the period, while minimizing false positives. There is certainly room for improvement, but on almost every available measure, the performance was superior to the existing alternative. Importantly, a forecasting model like the AFP's using global rather than conditional data is more likely to identify counter-intuitive or unexpected at-risk cases, a distinct advantage over qualitative and country-expert early warning systems.

A further advantage of the AFP approach is that it produces a short list of only fifteen countries at risk over a relatively long period of time. This five-year perspective helps to reduce the rare-events nature of the modelling, which likely increases model accuracy. If the countries at risk actually changed dramatically year by year, such an approach might be problematic. But this does not seem to be the case for genocide and politicide, with all indicators of ground truth as well as Harff and Gurr's lists exhibiting year-on-year overlap among the "likely suspects." The brevity of the AFP at-risk list can reduce the false-positive rate, an important practical feature if decisions must be made about the allocation of scarce government, UN or non-governmental organization (NGO) resources to monitoring or preventing genocide onsets among a handful of the most dangerous cases. The longer time-frame has similar benefits, allowing diplomacy, advocacy, raising public awareness, military planning and similar activities greater time to be organized, and potentially to have an impact.

In the course of working with the AFP models and the existing data on genocide and politicide in particular, we have developed ideas for further improving genocide forecasting. Perhaps most importantly, the coding of cases of genocide and politicide should be made more transparent and rigorous, and ideally more timely. It is often not clear why some cases are included and others are not, and why start or end dates are assigned as they are. Other scholars have identified problems in the quantitative and qualitative study of genocide, including lack of definitional consensus, incomplete documentation and non-reproducible codings.[36]

One promising direction is to abandon the tendency to use exclusively binary coding that forces coder decisions along one arbitrary threshold based on numerous qualitatively assessed event characteristics. A severity scale of targeted ethnic or political mass killing could yield an ordinal measure that would at least make the decision to label an event genocide when it reaches a certain threshold more transparent, and thus more replicable. Forecasting accuracy depends crucially on the quality of the training dataset. If outcomes are coded inconsistently, forecasting performance will necessarily suffer.

The AFP report that the GAM approach performed better in out-of-sample tests than attempts to capitalize on the strengths of a range of techniques using Bayesian ensemble methods.[37] But this approach has shown promise in related types of mass-atrocity forecasting.[38] More recent ensemble options such as error-correcting output coding,

[36] Ernesto Verdeja, "The Political Science of Genocide: An Emerging Research Agenda," *Perspectives on Politics* 10, no. 2 (2012): 307–21; Jay Ulfelder and Benjamin Valentino, *Assessing the Risks of State-Sponsored Mass Killing* (2008), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1703426 (accessed September 28, 2017).

[37] Butcher et al., *Understanding and Forecasting Political Instability*.

[38] Ulfelder and Valentino, *Assessing the Risks of State-Sponsored Mass Killing*. See also more recent unpublished work of the Early Warning Project (https://www.earlywarningproject.org/).

bagging and boosting are potentially promising.[39] Galar et al.[40] review the state of the art in ensemble techniques for the relevant type of data, unbalanced datasets with binary outcomes, proposing a taxonomy to address rare-event problems. While improving data quality, and perhaps finding new powerful predictors, have much to promise, there is potential to combine these with a more powerful computational approach.

There is growing promise for prediction of armed conflict and political violence,[41] but prediction of rare events like genocide is among the most challenging tasks in social science. While there is much work to do, we believe that the 2016–20 forecasts in this article provide an important source of early warning for genocide and politicide events in coming months and years. *We are not aware of a genocide forecasting approach with a record of higher accuracy or reliability.* Most basically, we urge intensive monitoring of these fifteen countries as first-priority, highest-risk cases. If genocidal killing is going to happen in the period up to 2020, chances are high that it will happen in one or more of these countries. We urge satellite monitoring such as that done by the Sentinel project,[42] as well as attention from risk analysts such as the International Crisis Group, and of course by intelligence agencies in the US, Europe, Australia and other governments concerned with preventing mass atrocities.

ORCID

Benjamin E. Goldsmith http://orcid.org/0000-0002-3247-3174

Acknowledgements

We thank Barbara Harff and Ted Robert Gurr for feedback on an earlier draft and for generously sharing their recent forecasting data with us. Earlier versions of this article were presented at the Australian Institute for International Affairs, the 2014 International Studies Association working group “Forecasting International Events,” the University of Dar es Salaam and the University of Pretoria. We thank all discussion participants and anonymous reviewers for very helpful feedback.

Disclosure statement

Benjamin E. Goldsmith is a Chief Investigator and Charles Butcher is a Research Associate of the Atrocity Forecasting Project.

Funding

We gratefully acknowledge support for this research from an Australian Research Council Discovery Project grant [grant no. DP1601101122].

[39] Thomas G. Dietterich, “Ensemble Methods in Machine Learning,” in *Multiple Classifier Systems*, ed. Josef Kittler and Fabio Roli (Berlin: Springer, 2000), 1–15; Guoqiang Zhong and Chen-Lin Liu, “Error-Correcting Output Codes Based Ensemble Feature Extraction,” *Pattern Recognition* 46, no. 4 (2013): 1091–100.

[40] Mikel Galar et al., “A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. Systems, Man, and Cybernetics Part C: Applications and Reviews,” *IEEE Transactions* 42, no. 4 (2012): 463–84.

[41] Lars-Erik Cederman and Nils B. Weidmann, “Predicting Armed Conflict,” *Science* 355, no. 6324 (2017): 474–6.

[42] See https://thesentinelproject.org/?s=satellite (accessed September 6, 2017).

Notes on contributors

Benjamin E. Goldsmith is Professor and Australian Research Council Future Fellow in the School of Politics and International Relations at the Australian National University. His research and teaching are in the areas of international relations, comparative foreign policy, and atrocity forecasting. He is the author of the book *Imitation in International Relations: Observational Learning, Analogies, and Foreign Policy in Russia and Ukraine* (Palgrave Macmillan 2005), as well as articles in leading academic journals including *Comparative Political Studies*, *European Journal of International Relations*, *Journal of Conflict Resolution*, *Journal of Peace Research*, *Journal of Politics*, *Quarterly Journal of Political Science* and *World Politics*.

Charles Butcher is Associate Professor of Political Science in the Department of Sociology and Political Science at the Norwegian University of Science and Technology. His work has been published or is forthcoming in *Journal of Peace Research*, *Journal of Conflict Resolution*, *International Studies Quarterly*, *International Interactions*, *Comparative Political Studies*, *The Review of International Studies* and *Third World Quarterly*.