Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality

HÅVARD HEGRE and HÅVARD MOKLEIV NYGÅRD
Uppsala University and Peace Research Institute Oslo (PRIO)

AND

PEDER LANDSVERK
Peace Research Institute Oslo (PRIO)

Can we predict civil war? This article sheds light on this question by evaluating 9 years of, at the time, future predictions made by Hegre et al. (2013) in 2011. We evaluate the ability of this study to predict observed conflicts in the 2010–2018 period, using multiple metrics. We also evaluate the original performance evaluation, i.e., whether the performance measures presented by Hegre et al. hold in this new 9-year window. Overall, we conclude that Hegre et al. were able to produce meaningful and reasonably accurate predictions of armed conflict. Of course, they did not always hit the mark. We find that the model has performed worse in predicting low level incidence of conflict than in predicting major armed conflict. The model also failed to predict some important broader regional shifts. These, however, represent important insights for future research and illustrate the utility in predictive models for both testing and developing theory.

Introduction

Almost 10 years ago in this journal, Hegre et al. (2013) published predictions for the probability of internal armed conflict for most countries in the world for the next 40 years. The article demonstrated that it was possible to construct predicted probabilities of armed conflict decades into the future using dynamic models and projections for key independent variables. Hegre et al. (2013) contributed to a wave of articles rekindling interest in prediction as a tool both to improve the research on armed conflict and also to make the research more policy relevant (Hegre et al. 2017). Ward, Greenhill and Bakke (2010) and Schrot (2014), in particular, argued that prediction needed to become a standard tool in evaluating models and theories on armed conflict. Such reliable long-run forecasts of armed conflict are also needed to construct projections for other critical socio-economic variables such as gross domestic product (GDP) per capita. The...
GDP per capita forecasts in Dellink et al. (2017) that underpin much climate change research, for instance, are certainly overly optimistic for all scenarios (Buhaug and Vestby 2019); however, as they completely ignore armed conflict. Joint modeling of armed conflict and such socio-economic variables is necessary to improve forecasts of both. For this to be feasible, we first need to know whether we can produce reliable forecasts of armed conflict. To that end, about a decade after these studies, the literature has reached a level of maturity that allows us to assess whether we are indeed able to make accurate and reliable predictions of civil conflict (Ryan-Mosley 2019).

These efforts have been challenged, however. As complex social phenomena, armed conflicts are claimed to be too heterogeneous and idiosyncratic to allow prediction, and what caused the conflicts of yesterday are unlikely to be useful to anticipate those of tomorrow. In a compelling critique, Bowlsby et al. (2019) review the prediction models in Goldstone et al. (2010), Hegre et al. (2013), and Chenoweth and Ulfelder (2017). They show that the predictive power varies substantially over time, in particular for the Goldstone et al. (2010) model. The predictive power peaked in the period originally used to evaluate their performance, and it performed particularly poorly in the most recent years. Cederman and Weidmann (2017) are skeptical of long-term forecasting for related reasons.

Although this skepticism is well taken, the retrospective empirical evaluation of Hegre et al. (2013) we present below suggest that there is reason to be more optimistic about long-term forecasting of conflict. The Hegre et al. (2013) study is especially useful for a reassessment. It presented global- and regional-level predictions for the expected incidence of civil conflict in the world for every year from 2010 to 2050.2 It is the only study we are aware of that has made country-level conflict forecasts publicly available.3 Included in the replication material for the study was the predicted conflict (Ryan-Mosley 2019).

We replicate the original out-of-sample evaluation done by Hegre et al. (2013) for the period 2001–2009, and then add to this a new evaluation of how well their predictions held up in the period from 2010 to 2018. More specifically, we first study how well, overall, the model performed in the 2010–2018 period in predicting both minor and major conflict. Second, we evaluate their evaluations—we examine to what extent the measures of predictive accuracy reported in the original study were indicative of the predictive performance of the model in the subsequent 9-year period. We also evaluate the forecasts in terms of precision-recall curves, a metric the original study did not consider. Third, we discuss a set of important cases that the authors either predicted correctly (such as Nigeria) or missed (such as Libya).

In the next section, we briefly explain the prediction methodology used by Hegre et al. (2013). In Section 3, we evaluate their predictions, discuss the results, and discuss some illustrative “hit” and “miss” cases. Section 4 concludes and discusses some implications of this exercise for the study of armed conflict and for the use of predictive models.

Prediction Methodology

The central innovation in the Hegre et al. (2013) model, is the development of a “dynamic simulation model,” estimated by means of a multinomial logit model. This model is dynamic in the sense that the prediction algorithm takes into account what has occurred at \( t - x \) when predicting the risk of armed conflict at \( t \) and, moreover, that the “realised” prediction at \( t \) is carried forward and similarly informs the probability of conflict at \( t + x \). Hegre et al. (2013) estimate the probabilities of transitioning between on of the conflict states—“no conflict,” “minor conflict,” and “major conflict”—using a multinomial logit model with the conflict level at \( t \) as the outcome variable, and the level at \( t - 1 \) as a set of dummy variables. In the multinomial model (see Greene 1997, 914–917) for the three outcomes \( j = 0: \) “no conflict,” \( j = 1: \) “minor conflict,” and \( j = 2: \) “major conflict”), the probabilities of the three outcomes are given by:

\[
p(Y_j = j) = \frac{e^{\beta_j}}{\sum_{k=0}^{2} e^{\beta_k}}
\]

Using a “dynamic multinomial logit” model allows them to capture that variables may increase the risk of conflict onset, but not its duration. This is achieved by adding interaction terms between the state at \( t - 1 \) and predictor variables. They also include information on the conflict state at earlier points in time by adding to the model a function of the number of years in each state up to \( t - 2 \).4 The same simulation framework has been used to perform long-term predictions of conflict as a function of the “shared socio-economic pathways” (Hegre et al. 2016), for studying the intensity of the conflict trap (Hegre, Nygård and Ræder 2017), and for evaluating the effectiveness of peacekeepers (Hegre, Hultman and Nygård 2019). The dynamic simulation concept has recently been extended to be used for short-term, geographically disaggregated violence early warning (Hegre et al. 2019).

Hegre et al. (2013) build an ensemble of models (Montgomery, Hollenbach and Ward 2012) using unweighted averages to combine predictions across models. Here, we only focus on the results of the combined model. The constituent models include a range of predictors that capture conflict history for both the country in question and neighboring countries; socio-economic indicators for infant mortality, oil dependence, and education rates; demographic indicators for ethnic composition, youth bulges, and population size; as well as regional markers. Models also include interactions with conflict state at \( t - 1 \) as well as the log of time in peace. For a full discussion and description of the variables, see Hegre et al. (2013).

Evaluation

We will discuss three aspects of the predictions made by Hegre et al. (2013). First, we evaluate the model’s performance for the 2010–18 period, including assessing whether the out-of-sample performance metrics reported by Hegre et al. (2013) were good estimates of the performance in the true future. Second, we look into some actual cases to explore where the prediction model did well and where it did not do so well. Third, we look into how well the

---

2The forecasts from 2012 onwards were true predictions about an unknown future, but 2009 was the last year of data available to the authors at the time of writing.

3The only exception is Hegre et al. (2016). This study covers a shorter time span, however, and is fairly similar to the one evaluated here. At the sub-national level, Wimmer et al. (2017) has made forecasts up to 2065 publicly available, and Hegre et al. (2019) monthly forecasts up to 2022.

4For full details of the model, see Hegre et al. (2013).
model has succeeded (so far) in forecasting global levels of armed conflict in the aggregate. Throughout, we compare predictions to observed data on armed conflict from the UCDP/PRIO Armed Conflict Database (Gleditsch et al. 2002; Pettersson, Högbladh and Öberg 2019).

We report the area under the receiver operator curve (AUROC) and the area under the precision-recall curve (AUPRC). The original out-of-sample performance metrics were constructed by training models on data up to and including the year 2000 and producing forecasts for 2001–2009 as if this was the future. The metrics compare these forecasts with observed conflict as reported by the UCDP. For ease of comparison, in the rows labeled “2001–2009,” we have recalculated metrics for the out-of-sample period and termination. For onset, it is clear both from Table 1 and Figure 1, dotted lines, that the out-of-sample evaluation reported in the original study suggested somewhat better performance than what achieved over the 2010–2018 period. This pattern also holds for the precision-recall curve for onset, although here the difference is very small. Table 1 indicates the same pattern holds for termination.

Table 1 reports metrics of predictive accuracy. It is comparable to table 2 in Hegre et al. (2013), but adds metrics for 2010–18 as well as the area under the precision-recall curve (AUPRC). The original out-of-sample performance metrics were constructed by training models on data up to and including the year 2000 and producing forecasts for 2001–2009 as if this was the future. The metrics compare these forecasts with observed conflict as reported by the UCDP. For ease of comparison, in the rows labeled “2001–2009,” we have recalculated metrics for the out-of-sample period in the original study. In the rows labeled “2010–2018,” we report the same metrics assessing how the predictions of Hegre et al. (2013) compare to observed conflicts over the 2010–2018 period. As in the original study, we report performance across the entire period (2001–2009, compared to 2010–2018) and the 3-year period at the end of the evaluation window (2007–2009, compared to 2016–2018).

We report the area under the receiver operator curve (AUROC) and the AUPRC for incidence (whether there was conflict or not in a given country year) as well as onset (new conflict that year) and termination (conflict ended that year). Moreover, we report true and false positive rates (respectively, TPR and FPR) at two cut-off points for categorizing a probability as a positive prediction, both with respect to incidence of conflict.

We first compare the predictive performance of the true future (2010–18) with the out-of-sample performance originally reported. For incidence, AUROC is 0.919 for 2010–2018, as compared to 0.944 in the original out-of-sample evaluation. This drop in performance is noticeable, but not very large—the drop is within the 95 percent confidence interval for the original metric, and much smaller than the drop reported in Bowlsby et al. (2019) when reevaluating the study of Goldstone et al. (2010). This is also borne out in Figure 1 which shows the receiver operator curves (left) and the precision-recall curves (right) for the original 2001–2009 and the new 2010–2018 period. In both cases, the two solid lines for incidence of conflict for the two periods are close to each other, indicating that the model performs just as well in the two periods.

The differences are somewhat larger when we turn to onset and termination. For onset, it is clear both from Table 1 and Figure 1, dotted lines, that the out-of-sample evaluation reported in the original study suggested somewhat better performance than what achieved over the 2010–2018 period. This pattern also holds for the precision-recall curve for onset, although here the difference is very small. Table 1 indicates the same pattern holds for termination.

Similarly visible, but not dramatic differences appear when we look at the TPR, especially for the higher, 0.50 cut-off rule. The TPR for the $p > 0.50$ cut-off was 0.662 for the 2001–2009 window vs 0.657 for the 2010–2018 window. For the $p > 0.50$ cut-off, performance declined, respectively, from 0.369 to 0.275. This small decline in performance in terms of the TPR is, however, partially offset by a reduction, that is, an improvement, in the rate at which the model produces false positives.

Why is the performance of the model somewhat poorer than originally claimed? Bowlsby et al. (2019, 12) suggest that the "drivers of political instability are time variant." The

Table 1. Performance metrics, AUROC, AUPRC, TPR, and FPR, for combined model evaluated on data for 2001–2009 as well as 2010–2018

<table>
<thead>
<tr>
<th></th>
<th>AUROC</th>
<th>AUPRC</th>
<th>TPR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>DeLong 95 percent CI</td>
<td>Score</td>
<td></td>
</tr>
<tr>
<td>2001–2009</td>
<td>Incidence</td>
<td>0.944 (0.929, 0.959)</td>
<td>0.797</td>
<td>0.369</td>
</tr>
<tr>
<td></td>
<td>Onset</td>
<td>0.863 (0.813, 0.912)</td>
<td>0.153</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Termination</td>
<td>0.813 (0.765, 0.861)</td>
<td>0.071</td>
<td>–</td>
</tr>
<tr>
<td>2007–2009</td>
<td>Incidence</td>
<td>0.948 (0.928, 0.969)</td>
<td>0.790</td>
<td>0.192</td>
</tr>
<tr>
<td></td>
<td>Onset</td>
<td>0.887 (0.832, 0.941)</td>
<td>0.154</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Termination</td>
<td>0.821 (0.748, 0.894)</td>
<td>0.063</td>
<td>–</td>
</tr>
<tr>
<td>2010–2018</td>
<td>Incidence</td>
<td>0.919 (0.899, 0.939)</td>
<td>0.774</td>
<td>0.275</td>
</tr>
<tr>
<td></td>
<td>Onset</td>
<td>0.824 (0.759, 0.888)</td>
<td>0.140</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Termination</td>
<td>0.779 (0.716, 0.842)</td>
<td>0.057</td>
<td>–</td>
</tr>
<tr>
<td>2016–2018</td>
<td>Incidence</td>
<td>0.907 (0.872, 0.942)</td>
<td>0.741</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>Onset</td>
<td>0.873 (0.787, 0.959)</td>
<td>0.264</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Termination</td>
<td>0.744 (0.624, 0.865)</td>
<td>0.077</td>
<td>–</td>
</tr>
</tbody>
</table>

5 Note that Hegre et al. (2013) only report the AUROC. Since the publication of that study, however, many of the weaknesses of the AUROC, especially when confronted with heavily skewed data such as armed conflict, has become apparent (Cranmer and Desmarais 2017). Reporting also the AUPRC has now become routine in the literature.

6 The confidence interval for AUROC is calculated using the method of DeLong, DeLong and Clarke-Pearson (1988). The method exploits the fact that the empirical AUROC is equal to the Mann–Whitney statistic, so that the variance-covariance matrix can be calculated according to the general theory of U-statistics.

7 The replications in Bowlsby et al. (2019) suggest that the performance of the Hegre et al. (2013) model is much more stable over time than the one in Goldstone et al. (2010).

8 The curves for termination are not included in Figure 1 for legibility reasons.

9 Note that a 0.50 cut-off threshold is a very high bar for classifying something as a positive prediction. Civil war is a rare event so predicted probabilities of conflict well below 0.50 should be seen as important signals. Indeed, in recent work on violence early warning, Hegre et al. (2019) find that an optimal cut-off threshold probably is closer to 0.10.
are good reasons to consider the forecasts of conflict incidence. To assess possible humanitarian consequences, there is the fact that it still goes on 10 years after the onset of war in Syria in 2011 was a shocking disaster, but the length of a conflict is of crucial importance. The incidence model must also do well in terms of onset. More than 10 percent of the country years. It does well even in terms of precision-recall which is the most appropriate metrics set for unbalanced data as is the case here. AUPRCs range from 0.74 to 0.80. To gain some intuition for what these figures mean, we can inspect some points in the precision-recall curves (Figure 1). If the target is a recall of 0.8, estimated precision is 60 percent. In other words, if we require that the model correctly predicts as conflict 80 percent of actual conflict years, 60 percent of these alerts will be correct. For onset and termination, performance is much less impressive. At an 80 percent recall target, the model predicted conflict in a given country in a given year correctly in only 10 percent of the country years.

For policy purposes, predicting onset correctly may seem much more important than predicting incidence. However, when forecasting years and decades into the future, the difference is less important. The majority of conflicts that will be active in 2030 are unlikely to have started yet, so a good incidence model must also do well in terms of onset. Moreover, the duration of a conflict is of crucial importance. The onset of war in Syria in 2011 was a shocking disaster, but the tragedy is completed by the fact that it still goes on 10 years after. To assess possible humanitarian consequences, there are good reasons to consider the forecasts of conflict incidence in addition to onset.

How could the Hegre et al. (2013) model be improved? A closer look at where the Hegre et al. (2013) study hits the spot and where it misses is instructive.

The AUROC and AUPRC metrics respond to the model’s ability to rank countries correctly in terms of risk. In Appendix Table A-2 we list the fifty countries with the highest predicted probabilities. The table shows that 26 of the 35 countries with conflict in 2018 are in the top-30 list of predicted probabilities. Three of the four major conflicts are in the same list. Figure 2 indicates where the models were least precise. The colors represent the difference between the average predicted probability of conflict and the proportion of years in conflict over the 2010-2018 period. Countries plotted with red color had much more conflict than predicted, and those in blue had considerably less. The top map shows the comparison for all UCDP conflicts, the bottom for major conflicts only.

When colors are light, the difference between prediction and reality was slight. The map and Appendix Table A-2 indicates the Hegre et al. (2013) model performed well in predicting the continuation of a series of both minor and major armed conflicts, such as the long-running conflicts in for instance India, Afghanistan, and Somalia. The model also does well in predicting the escalation of previously low-intensity conflicts that have since become higher-intensity ones, such as for instance the conflict in Nigeria, Niger, or the Central African Republic. Even Burkina Faso, a country that had not seen state-based conflict since 1987 according to UCDP when the forecasts were made, is in the top quarter of the risk rankings.

However, the model under-predicted severely for Syria, Libya, Yemen, and Ukraine, for both major and minor conflict. The model ranked Yemen highly (Appendix Table A-2), with an annual probability of either conflict above 25 percent, but, as any other forecasts made in 2010, missed the other three.

The model over-predicted conflict for Peru, Angola, Sri Lanka, and Ethiopia. All of these countries were in conflict in 2009, the last year with data in the model. Among countries with ongoing conflict, Peru and Sri Lanka were assessed as considerably more likely to end the conflicts than continue them, though. For major conflict, it assigned a
too high probability to India and Russia. Again, the predicted probabilities of major conflict were relatively low (see Appendix Table A-2). On the other hand, the model suggested a 88-percent chance of conflict in Ethiopia, given its long-standing conflict. In 2018, that prediction was a false positive, but in 2020 this case returned to a true positive, despite the 2019 Nobel peace prize to Abiy Ahmed Ali.

**Ability to Predict Correctly in the Aggregate**

The relatively poor performance of the fixed-threshold metrics TPR and FPR (Table 1) as compared to the rank-based metrics, suggests that the model, and the predicted probabilities it produces, are not properly calibrated. The metrics indicate that Hegre et al. (2013) under-predict conflict. This is clear in Figure 3 which plots the observed incidence (measured in terms of percent of countries) of conflicts in the world from 1970 to 2018 (solid line) along with predicted incidence for the 2010–2050 period (dashed line). The two vertical lines mark 2009 and 2018, the final years of the two prediction periods we examine. The model performs well in predicting the proportion of countries in major armed conflict. The model did not anticipate the strong increase from 2013 to 2014, but after the decline of major conflict from 2017 to 2018 (Pettersson, Högbladh and Öberg 2019), the actual proportion of countries dropped to the forecasted level.  

Now consider the line for “Either”—minor or major conflict incidence combined. Here, Hegre et al. (2013) predicted a pronounced and continuous decline in the number of conflicts from the 2009 level. In contrast to this, the world in this period saw a substantial increase in the number of armed conflicts. Indeed, in absolute numbers, there were more active armed conflicts in 2015 than registered at any time since 1946 (Allansson, Melander and Themnér 2017). From the above, we know that the model does well at predicting major conflict, implying that this divergence is driven by a failure of the model to predict more low-level, minor, conflicts. We discuss potential reasons for this below.

**Regional Performance**

Some regions, in particular the Middle East and North Africa (MENA) region, has seen especially pronounced differences in their conflict trajectories across the old and the new evaluation window. Indeed, the old evaluation window ended right before the onset of the “Arab spring.” To delve deeper into this and to see to what extent the decrease in performance is region-specific, Table 2 breaks down the

---

11 The shaded area represents the 80 percent confidence band. The reported uncertainty was under-estimated in the original. See Hegre et al. (2016) for improved estimates.
Can We Predict Armed Conflict?

Figure 3. Observed (1970–2018, solid line) and predicted (2010–2048, dashed line) global incidence of minor or major (top line) or major only (bottom line) armed conflict.

Table 2. AUROC and AUPRC, by region, new and original evaluation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Europe, North America and Oceania</td>
<td>0.853</td>
<td>0.014</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>South and Central America and the Carribbean</td>
<td>0.983</td>
<td>0.860</td>
<td>0.995</td>
<td>0.884</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>0.978</td>
<td>0.922</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MENA</td>
<td>0.974</td>
<td>0.939</td>
<td>0.844</td>
<td>0.833</td>
</tr>
<tr>
<td>West Africa</td>
<td>0.614</td>
<td>0.195</td>
<td>0.886</td>
<td>0.587</td>
</tr>
<tr>
<td>East and Central Africa</td>
<td>0.862</td>
<td>0.704</td>
<td>0.864</td>
<td>0.739</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>0.974</td>
<td>0.605</td>
<td>0.911</td>
<td>0.171</td>
</tr>
<tr>
<td>South and Central Asia</td>
<td>0.935</td>
<td>0.889</td>
<td>0.936</td>
<td>0.901</td>
</tr>
<tr>
<td>East and South-East Asia</td>
<td>0.939</td>
<td>0.859</td>
<td>0.983</td>
<td>0.976</td>
</tr>
</tbody>
</table>

AUROC and AUPRC by region, again comparing the regional performance now to what was reported in the original study.\(^{12}\) At the regional level, we do see a decrease in both metrics for the MENA region. The performance for 2010–2018 for West Africa is actually better than in the original out-of-sample evaluation, whereas Southern Africa is worse. Beyond these regions, differences are small or negligible.

What West Africa and MENA have in common is important, and points to three important limitations of the model. First, no authoritative projections for democratic institutions exist and Hegre et al. (2013) were therefore unable to include democracy, changes to democratic institutions, or other regime characteristics in their model.\(^{13}\) Apart from conflict history, the most important driver of conflict in the model is lack of socio-economic development. Since Syria and Libya are (or were) middle-income countries, the model did not assign a high risk to them. Studies indicate that democratization processes are associated with a heightened risk of internal armed conflict (Hegre et al. 2001). The lack of features covering political institutions, especially democracy, in the model meant that it would be hard pressed to predict at least the timing of the “Arab spring” dynamics with a high degree of accuracy. As we elaborate on below, developing forecasts of democratization processes is of high importance to improve conflict forecasts.

Second, conflict history is among the most important predictors in the Hegre et al. (2013) model. West Africa and MENA both saw substantial regional-level changes in their conflict trajectories between the different windows. West Africa had seen a number of conflicts throughout the 1990s and early 2000s. This, however, changed substantially as the wars in Liberia and Sierra Leone ended in 2002 and 2003. Consequently, West Africa was much more peaceful in the evaluation period of the original study than it had been over the last decade. Similarly, the MENA region had seen relatively few armed conflicts in the decade before the Hegre et al. (2013) prediction period started in 2010, something which then changed dramatically with the onset of armed conflicts in Yemen, Libya, and Syria. To some extent, these conflict history variables appear to make the model, at a regional level, somewhat too static—both in locking larger regions in conflict and in peace—and very cautious about predicting broader, regional level, changes to conflict incidence.

Third, and related to the last point, the Hegre et al. (2013) model does include regional-level and neighborhood factors as predictors, but these appear to be unable to account for what is arguably regionally correlated changes, such as those seen in West Africa and the MENA. Very likely, the error terms exhibit regional-level patterns of between-country correlation. If the model fails to predict one country then this prediction error increases the likelihood of a mis-prediction for another country in the same

---

\(^{12}\) Scores for two regions could not be computed for 2010–2018 as there was no observed conflict in Western Europe, North America and Oceania, and since the model separated perfectly for Eastern Europe.

\(^{13}\) Hegre et al. (2013) do include a number of socio-economic factors, including infant mortality and education rates, and these variables are of course highly correlated with regime type.
region—prediction errors are correlated. Because of this, the model is partly blind to regional conflict dynamics that can lead to broader shifts both in a more peaceful and more violent direction, such as those seen in, respectively, West Africa in the early 2000s and in MENA in the early 2010s. We discuss some possible solutions to these issues below.

**Long-Term Forecasting of Armed Conflict: the Way Forward**

Have attempts at predicting armed conflict been excessively ambitious? Cederman and Weidmann (2017) remind us that especially for long-run predictions it is important to take into account “massive historical complexity and contingency in human systems” as well the likelihood of systemic shocks that may both alter the relationship between conflict and key predictors as well as what the key predictors are themselves. One crucially important approach to learn about the role and importance of such dynamics, however, is to build and assess theoretically motivated long-term forecasting models based on insights that are widely accepted (Colaresi and Mahmood 2017). Only then will we be able to say something meaningful about the role and importance of “historical complexity” and systemic shocks. To cite the great H. G. Wells: “Science is not an account of facts but a criticism and analysis of facts ... It is therefore not merely legitimate ... to attempt forecasts of general trend of events, but to not attempt them is a frank confession of the futility of the science” (Wells 1925, 1). Here we have evaluated the first 9 years of such an attempt to produce 40 years of such predictions. The general conclusion, we argue, is that creating meaningful, useful, and reasonable predictions does appear to be possible. In contrast to Bowlsby et al. (2019) we also find that the predictive accuracy is quite stable over time and that predictive performance has not declined much in the ten years since the study we focus on here was published.

This, of course, does not at all imply that the predictions published by Hegre et al. (2013) always hit the mark. Far from it. Three weaknesses in general appear. First, in terms of forecasting the aggregate, global, level of conflict, the model performs worst for low-intensity, “minor,” conflict. This suggests that the model is not optimally calibrated, in terms of predictors, for such low-level conflict. To some extent this may be because low-intensity conflict are more stochastic than higher-intensity major conflicts. The shock of the “Arab spring” in combination with the failure to mount effective international response to incipient violence, led to a number of conflicts that the model failed to predict. Three of them, Syria, Yemen, and Libya, descended into major conflagrations. The rest of these cases (e.g., Tunisia and Jordan) remained minor. Possibly, the structural variables that dominate the Hegre et al. (2013) model largely constrain countries from major violence but not from minor incidences.

The model seems to be quite good at predicting escalation from low to high levels of violence, but considerably less impressive at predicting escalation from no conflict to minor conflict—the model does not perform particularly well for the onset of armed conflict. The latter is intrinsically a more difficult problem, and requires a different set of predictors than the ones included here.14 However, such “early-warning signals” that change quickly and alert to alarming recent events are almost by definition not available when seeking to forecast decades into the future. A model might do better by including such predictors for the first few years only, and use structural variables for the longer-term forecasts. Such a construction, however, would add complexity and lead to some losses in interpretability. It might be better to treat short- and long-term prediction problems separately, and assess studies like Hegre et al. (2013) in terms of their performance for the long-term incidence of conflict rather than for onset in the short term.

Second, not surprisingly given the lack of variables measuring aspects of the political system, the model fails to predict dynamics that are driven by changes to political systems. The prediction failures related to the Arab spring that we have discussed are cases in point. The termination of conflict in Ethiopia is also related to a positive political development the model was not set up to incorporate. Over the long run, political changes would be partially covered by aspects of socio-economic development included in the model, since socio-economic development is a powerful predictor of political development (Przeworski et al. 2000). Over shorter periods, however, the model will miss such dynamics, and the turbulence in the transition phase will not be well modeled. A more comprehensive solution is to forecast changes to political systems and the onset and termination of armed conflict jointly.

Third, the model appears to be too conservative in the face of broader regional-level shifts. The model clearly underestimated the extent to which conflict spreads from one country to another. This indicates that the model should have better accounted for the plausibly correlated error terms between countries in a region. In the case of the Arab spring, the unrest spread not only through geographical adjacency, but also across the Arab cultural sphere. The model would do better if it had modeled spread of conflict within the same language or religion group in addition to through geographical neighbors. With a better diffusion model, the spike in 2014–2018 in Figure 3 might have been covered by the prediction interval. Alternatively, models that capture systemic shocks to the relations between variables (such as Mitchell, Gates and Hegre 1999; Cunen, Hjort and Nygård 2020) might be necessary.

These are important technical learning points as the discipline continues its attempts to build conflict-forecasting models. But they also represent theoretical insights. The difference in predictive accuracy between minor and major conflict indicates that there is more heterogeneity between these types of conflict then we typically account for in models of civil war. This should be addressed in future research. The regional clustering also represents an avenue for future theorizing. Of course, a large literature does exist that examines spatial diffusion of conflict, but our exercise indicates that more work is needed.

Predicting conflict 40 years into the future obviously entails pushing a model to its extreme. However, it may be the most effective way to learn something enduring about the state of theory and our understanding of civil war dynamics. If we had been completely unable to predict armed conflict 10 years into the future based on variables that the discipline essentially agree are critical, how much could we really say that we knew about civil war? Luckily, this exercise shows that we know a lot—and that we still have a lot to learn.

---

14 Promising efforts make use of news sources (Chadeaux 2014; Mueller and Rauh 2018), events data (Ward and Beger 2017; Chiba and Gleditsch 2017), or predictors such as killing of journalists (Gohdes and Carey 2017).
weaknesses of the study discussed here suggest a set of key lessons for future research:

Forecasting the predictors of conflict. Human systems are complex and contingent (Cederman and Weidmann 2017). The Hegre et al. (2013) model failed to predict the war in Syria in part because it ignored the importance of political systems. To incorporate such variables in long forecasts, it is necessary to include forecasts of changes to political institutions in the model. Some attempts to do so now exist (Morgan, Beger and Glynn 2019; Andrijevic et al. 2020) although these still do not have the status of ‘authoritative’. The same applies to ‘oil dependence’ and any other factor that scholars see as essential to predict conflict.

Model conflict and other predictors jointly. The model in Hegre et al. (2013) relies on the projections for population and infant mortality that were available in 2010. Since then, several new projections have become available that build on the “Shared Socioeconomic Pathways” (O’Neill et al. 2014). These are much improved as they represent a set of future scenarios that come close to spanning the vast uncertainties regarding the remainder of this century. Armed conflict, however, is largely ignored in these projections. The GDP per capita forecasts in Dellink et al. (2017), for instance, that underpin much climate change research, are certainly overly optimistic for all scenarios (Buhaug and Vestby 2019). The governance projections in Andrijevic et al. (2020) suffer from the same problem. Joint modeling of armed conflict and predictors is necessary to improve forecasts of both. Possibly, the over-optimism in Hegre et al. (2013) is partly due to an over-optimism in the projections for exogenous drivers of conflict that it relies on. The population projections in United Nations (2007), for instance, are vague on the basis for the projected reduction in infant mortality, and may well be over-optimistic with respect to that particular indicator. Some projects do joint modeling (e.g., the International Futures project; Hughes 2019), but much more effort is required to solve this challenging problem.

Improved modeling. Logistic regression models still belong in any forecaster’s toolbox, but should now be complemented by new approaches from the machine-learning literature. There are two constraints that restrict the menu of methods, however. Since conflict data at the country-year level are very sparse, methods such as random forests and penalized regression models that safeguard against overfitting are most promising. Moreover, since the forecasting horizon—for instance, to the end of the 21st century—is longer than the period for which we have high-quality data, models need to be fit into some sort of dynamic simulation framework to be useful.

Combine levels of analysis. Modeling conflict spillover is another shortcoming of Hegre et al. (2013). This is partly due to a lack of data—conflict is fortunately quite rare at the country level, and spillovers even more so. In addition, spillovers are often local—conflicts in the Caucasus are unlikely to affect the far East of Russia, for instance. Models pitched at detailed geographical locations are much more likely to succeed (Wittern et al. 2017). However, since several crucial processes work at the government level (e.g. contests over the nature of political systems), effective long-term modeling must combine insights from multiple levels of analysis. Some ideas for how to achieve this are suggested in Hegre et al. (2019). Since data at more fine-grained geographical levels are less sparse, data-hungry models such as various recurrent neural network models as well as boosted tree models can be applied effectively.

Evaluation. Considerably more conceptual work is needed on how to evaluate to most effectively evaluate forecasts. This includes foundational issues such as what metrics are most appropriate and what windows should be considered. We have shown that the addition of AUPRC qualifies the conclusions to be drawn from the original study. The mismatch between performance at the country-year level and the global aggregate also suggests that evaluation should look at performance from multiple angles. This has implications for how we train models—most algorithms optimize on only one evaluation criterion, but good models should weigh different metrics against each other.

Maintain running forecasting systems. A strength of the Hegre et al. (2013) study is that they produced and made openly available a long timer series of future forecasts that could be evaluated. To move this field forward, we need more systems that are set up to produce regular forecasts, forecasts that then over time can be evaluated across different metrics. One such system is the Violence Early Warning System (Hegre et al. 2019).

Better data. Finally, richer, more precise data with greater temporal and spatial coverage is most important of all. Over the past decade, the UCDP-GED (Sundberg and Melander 2013) time series have become 30 percent longer and the V-Dem project has established the most comprehensive democracy data ever produced (Coppedge et al. 2020). New efforts to improve on the forecasts reviews here can start from a much richer selection and data, and substantial gains in precision are to be expected.

Funding

The authors acknowledge funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 694640, ViEWS), and the Research Council of Norway (Grant No. 275400).

Supplementary Information

Supplementary information is available at the International Studies Quarterly data archive.

References


