Replication Materials for:

Forecasting the Onset of Genocide and Politicide:

Annual Out-of-Sample Forecasts on a Global Dataset, 1988-2003

Replication Materials

In this replication appendix we provide further details regarding data imputation procedures, and our construction of a conditional forecasting model based on that of Harff (2003, 2012). We also provide two tables (Tables C and D) with more detail on results referred to in the article.

Data Imputation

We have imputed data for two components of our variables *across the dataset* because they were missing for cases of genocide onset. These are for 1992, 1993, and 2002 and relate to the onset years 1993 (Burundi), 1994 (Rwanda), and 2003 (Sudan). Rather than allow these observations to drop out of our models, we imputed missing values with data from the previous year. Values are imputed for all state-years with missing data, however, not just the cases of genocide onset. Values for 'executive constraints' were missing for 1992 and 1993, and we replaced missing values at year *t* with the same country's values for year *t-1*, but only in instances in which there was also a non-missing value for the overall Polity score in year *t*. Executive constraints is a component of the UnconstrainedExecutive×fdHumanDefenseBurden$_{t-1}$ and UnconstrainedExecutive×HumanDefenseBurden$_{t-1}$ variables. Only the military personnel variable was missing for 2002, so we also replaced missing observations of this variable at *t*

with its value at *t-1*. Military Personnel is also a component of UnconstrainedExecutive×fdHumanDefenseBurden$_{t-1}$ and UnconstrainedExecutive×lnHumanDefenseBurden$_{t-1}$. To avoid biasing the forecasts for the years 1993, 1994 and 2003, we made these changes for all years in the dataset. We expect this reduces the quality of our forecasts, because countries with missing data in Polity are often in transition, and imputed data generally involve some loss of precision. Nevertheless we prefer to produce some predictions for these cases, rather than none. Using past values for the same countries is a simple and realistic option for forecasting, especially given variables which tend not to exhibit great temporal variation. Practically, past observations for the same countries are likely to be available for forecasting, if present observations are missing.

We also encountered missing values for the five newly emergent states of the former Yugoslavia in 1991, but across too many variables to justify imputing data across the whole dataset. We imputed these values, but *only* for 1991 for the former Yugoslav republics. We decided to use data for Yugoslavia as a whole from 1991 for each of the states in 1991, specifically Bosnia and Herzegovina, Croatia, Macedonia, Serbia and Montenegro, and Slovenia. The variables are: the polity value, executive constraints, military personnel, total population, IGO memberships, ethnic power relations, ethnic fractionalization, infant mortality, assassinations, history of instability and genocide / politicide, and neighboring conflicts. Where these are components of other variables in the model, the other variables were also recreated for these former Yugoslav states for 1991 only. We see this as a conservative approach, given the imprecision of such imputation. But, it is important to capture the Bosnia case (Table I), and this solution uses data which were available at the time. Our imputations do not appear to positively bias the forecasts: Table III shows that performance with several imputed values for the year 1992 is among the worst for years with genocide onsets. In addition, Tables IIa and IIb present results for out-of-sample forecasts

with and without imputed data and show that imputing these data does not improve forecasting performance. Data imputation did not affect observations before 1987 which were included in the in-sample analyses, thus we do not show two sets of identical in-sample results.

Conditional Model Based on Harff (2003, 2012)

As a benchmark, we constructed a conditional forecasting model based on descriptions by Harff (2003; 2012). The results are presented in Table A, and Table B provides a comparison of forecasting performance of this adapted version of Harff's model and our model featured in the main article across several forecasting thresholds. To produce annual forecasts it was necessary to adapt her approach to an annual time-series framework. But we attempted to retain the key features of her approach, which we take to be the following: 1) the model is conditional on instability in that the 'universe of analysis consists of… instances of internal war and regime collapse… as identified by the State Failure [PITF] project' (Harff 2003, 57);[1] 2) the model is 'structural' in that it identifies underlying factors – 'a limited set of theoretically important variables' - within unstable states which increase or reduce the conditional probability of genocide onset (Harff 2003, 65). A third feature of Harff's approach, the case-control method, cannot be adapted for forecasting because it involves selecting observations which are known not to experience genocide onset, but are otherwise

[1] 'The dependent variable represents the conditional probability that a genocide or politicide will begin one year later in a country already experiencing failure. This avoids the problem of comparing the risks of genocide in Rwanda and Sudan with, say, the negligible risks in France and Canada. Instead, the objective is to examine countries experiencing episodes of internal wars and regime collapse and determine why geno-/politicide occurred during such events in Rwanda and Sudan but not, say, in Liberia or Nigeria' (Harff 2003, 65).

similar to observations which are known to have experienced genocide onset (Harff 2003, 65). However, this approach is plausible for assessing which variables should be included in the model, which Harff has already established and thus we construct a conditional forecasting model based on the variables she has used.

As noted in the main article, there are six variables in her model. We have received updated versions of two of these from PITF.[2] The remainder we have reconstructed based on descriptions by Harff (2003; 2012). We had to diverge from her description for the variable "Prior Genocide" because this dropped out of the models due to collinearity. Instead, we substituted the cubic polynomial (linear, squared, and cubed terms) for years since the previous genocide. Data sources are PITF data on genocide/politicide and instability (Marshall, Gurr, and Harff, 2010), the Polity IV dataset (Marshall, Jaggers and Gurr, 2010), the Expanded Trade and GDP Data (Gleditsch 2002), and the MAR dataset (Marshall, 2010). In addition, Harff (2012) includes an indicator for the risk of future instability over three years based on a model by Hewitt in the same volume. We have also attempted to reproduce this model as accurately as possible. Hewitt's (2012, 9) model forecasts instability onset over three years, and includes 'Institutional Consistency,' 'Partial Democracy,' 'Economic

[2] Email communication with Monty Marshall on 2 August 2012. The data for Elite Ethnicity and Elite Ideology were provided on the condition that we do not share them beyond our research group.

Openness,' 'Infant Mortality Rates,' 'Militarization,' and 'Neighborhood Security'.[3] One adjustment we made, which *increases* the ultimate out-of-sample forecasting performance of the adapted Harff model, is to use probit regression instead of logit. While the AUC statistics for the Hewitt-type model in Table A are considerably lower than those for our instability model (Table IIa), substituting our forecasted probability of instability into the Genocide model in Table A affected its forecasting performance only moderately. The in-sample AUC is lower, .9091, and the out-of-sample AUC is somewhat higher at .7883.

Adding the 15-year rolling average of total instability years, as Harff's (2003) description implies, actually worsens the out-of-sample forecasting of the model to AUC < .68, whether it is added in combination with the probability of instability onset, or replaces the probability of instability onset.

References

Abouharb, M. Rodwan and Anessa L. Kimball (2007) A New Dataset on Infant Mortality Rates, 1816-2002. *Journal of Peace Research* 44(6): 743-754

Gleditsch, Kristian S. (2002) Expanded Trade and GDP Data. *Journal of Conflict Resolution* 46(5): 712-24.

Harff, Barbara (2003) No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder since 1955. *American Political Science Review* 97(1): 57-73.

Harff, Barbara (2012) Assessing Risks of Genocide and Politicide: A Global Watch List for 2012. In *Peace and Conflict 2012*, edited by J. Joseph Hewitt, Johnathan Wilkenfeld and Ted Robert Gurr. College Park, MD: University of Maryland, 2012, 53-56.

[3] These variables are for the most part identical to variables used in our or the adapted Harff model: $MixedRegime_{t-1}$, $TradeGDPshare_{t-1}$, $InfantMortalityRate_{t-1}$, $HumanDefenseBurden_{t-1}$. $NeighborSecurity_{t-1}$ is similar to $NeighborConflict_{t-1}$, with the exception that as per Hewitt's model it is coded '1' if there is any conflict in neighboring countries, rather than counting the number of neighboring countries with conflict.

Hewitt, J. Joseph (2012) The Peace and Conflict Instability Ledger: Ranking States on Future Risks. In *Peace and Conflict 2012*, edited by J. Joseph Hewitt, Johnathan Wilkenfeld and Ted Robert Gurr. College Park, MD: University of Maryland, 2012, 5-24.

Marshall, Monty G (2010) *Major Episodes of Political Violence (MEPV) and Conflict Regions, 1946-2008*. (http://www.systemicpeace.org/inscr/inscr.htm).

Marshall, Monty G; Ted Robert Gurr, and Barbara Harff (2010) *PITF – State Failure Problem Set: Internal Wars and Failures of Governance, 1955-2009. Dataset and Coding Guidelines*. (http://www.systemicpeace.org/inscr/inscr.htm).

Marshall, Monty G; Keith Jaggers and Ted Robert Gurr (2010) *PolityIV Project: Political Regime Characteristics and Transitions, 1800-2010, Dataset Users' Manual*, Centre for Systemic Peace. (http://www.systemicpeace.org/polity/polity4.htm).

Minorities at Risk Project (2009) *Minorities at Risk Dataset*. College Park, MD: Center for International Development and Conflict Management. Accessed 17 January 2012 at: http://www.cidcm.umd.edu/mar/.

Tables

Table A – Adaptation of Harff's Conditional Genocide /Politicide Onset Model

Risks of Future Instability		*1974-1987* β	*P*	*Genocide / Politicide Onset*		*1974-1987* β	*P*
Infant mortality	$InfantMortalityRate_{t-1}$	0.003	0.127	Ethnic character of the ruling elite	$EliteEthnicity_{t-1}$	0.367	0.285
Economic openness	$(Trade/GDP)_{t-1}$	-0.004	0.221	Ideological orientation of the ruling elite	$EliteIdeology_{t-1}$	1.073	0.003
Militarization	$HumanDefenseBurden_{t-1}$	15.170	0.087	Trade openness	$(Trade/GDP)_{t-1}$	0.001	0.943
Regime consistency	$MixedRegime_{t-1}$	0.003	0.375	Regime type: Authoritarian	$FullAuthoritarian_{t-1}$	2.774	0.001
Partial democracy	$PartialDemocracy_{t-1}$	0.858	0.007	Regime type: Partial authoritarian	$PartialAuthoritarian_{t-1}$	3.327	0.000
Neighborhood war	$NeighborConflictDum_{t-1}$	0.086	0.592	Regime type: Partial democracy	$PartialDemocracy_{t-1}$	1.319	0.417
	Constant	-2.143	0.000	Active ethnic discrimination	$StateLedDiscrimination_{t-1}$	0.199	0.739
				Prior genocide	$YearsSinceGenocide_{t-1}$	-0.600	0.025
					$YearsSinceGenocide^2_{t-1}$	0.049	0.006
					$YearsSinceGenocide^3_{t-1}$	-0.001	0.004
				Risks of Future Instability	$Hewitt_PrInst_{t123}$	18.685	0.045
					Constant	-8.813	0.000
	n	1869			n	230	
	Wald $\chi2(6)$	27.41	0.000		Wald $\chi2(11)$	589.12	0.000
AUC (1974-1987): 0.7001, n = 1869				AUC (1974-1987): 0.9302, n = 230			
AUC (1988-2003), imputed data: 0.6641, n = 2374				AUC (1988-2003), imputed data: 0.7478, n = 437			
AUC (1988-2003), without imputed data: 0.6647, n = 2333				AUC (1988-2003), without imputed data: 0.7492, n = 420			

Notes: Probit models with robust standard errors.

Table B - 1988-2003 Out-of-Sample Forecast Comparison

True Positives	True-positive Rate	False Positives	True-negative Rate
Unconditional Model as in Table IIb, n = 2219			
7	0.636	212	0.898
8	0.727	244	0.882
9	0.818	358	0.827
10	0.909	432	0.792
Conditional Adapted Harff Model, n = 493			
3	0.429	57	0.883
4	0.571	122	0.749
5	0.714	130	0.733
6	0.857	236	0.514

Note: sample sizes for all forecast state-years including ongoing genocides.

Table C [non-onset years]

Table C – Forecasts for Years Without Genocide/Politicide Onset

Onsets and Forecasts for 1990

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
Indonesia	ongoing	0.191	1	0.84%
Uganda	0	0.172	2	1.68%
Rwanda	0	0.092	3	2.52%
El Salvador	0	0.076	4	3.36%
DRC	0	0.063	5	4.20%
Chad	0	0.046	6	5.04%
Afghanistan	ongoing	0.043	7	5.88%
Ethiopia	0	0.043	8	6.72%
Angola	ongoing	0.034	9	7.56%
Algeria	0	0.032	10	8.40%
n (1974-1989)		n (1990)		
1768		119		

Onsets and Forecasts for 1991

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
Rwanda	0	0.243	1	0.85%
Indonesia	ongoing	0.152	2	1.71%
Uganda	0	0.145	3	2.56%
Chad	0	0.073	4	3.42%
DRC	0	0.069	5	4.27%
Kenya	0	0.053	6	5.13%
Ethiopia	0	0.052	7	5.98%
Afghanistan	ongoing	0.044	8	6.84%
Angola	ongoing	0.029	9	7.69%
El Salvador	0	0.021	10	8.55%
n (1974-1990)		n (1991)		
1878		117		

Onsets and Forecasts for 1995

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
Rwanda	0	0.667	1	0.75%
Burundi	0	0.605	2	1.49%
DRC	0	0.482	3	2.24%
Algeria	0	0.324	4	2.99%
Tajikistan	0	0.216	5	3.73%
Guatemala	0	0.132	6	4.48%
Cambodia	0	0.117	7	5.22%
Afghanistan	0	0.086	8	5.97%
Uganda	0	0.080	9	6.72%
Angola	0	0.066	10	7.46%
n (1974-1994)		n (1995)		
2334		134		

Onsets and Forecasts for 1996

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
Algeria	0	0.297	1	0.74%
Burundi	0	0.261	2	1.47%
Cambodia	0	0.221	3	2.21%
Guatemala	0	0.105	4	2.94%
Rwanda	0	0.096	5	3.68%
DRC	0	0.089	6	4.41%
Azerbaijan	0	0.065	7	5.15%
Bangladesh	0	0.059	8	5.88%
Georgia	0	0.057	9	6.62%
Tajikistan	0	0.055	10	7.35%
n (1974-1995)		n (1996)		
2467		136		

Table C (continued)

Onsets and Forecasts for 1999

country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
DRC	0	0.079	1	0.71%
Angola	ongoing	0.063	2	1.42%
Uganda	0	0.049	3	2.13%
Afghanistan	0	0.035	4	2.84%
Algeria	0	0.032	5	3.55%
Nigeria	0	0.026	6	4.26%
Zambia	0	0.025	7	4.96%
Sudan	ongoing	0.022	8	5.67%
Indonesia	0	0.019	9	6.38%
Congo	0	0.016	10	7.09%
n (1974-1998)		n (1999)		
2883		141		

Onsets and Forecasts for 2000

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
DRC	0	0.070	1	0.71%
Afghanistan	0	0.064	2	1.42%
Cambodia	0	0.056	3	2.13%
Algeria	0	0.047	4	2.84%
Iran	0	0.032	5	3.55%
Uganda	0	0.032	6	4.26%
Angola	ongoing	0.031	7	4.96%
Zambia	0	0.024	8	5.67%
Sudan	ongoing	0.022	9	6.38%
Burundi	0	0.020	10	7.09%
n (1974-1999)		n (2000)		
3022		141		

Onsets and Forecasts for 2001

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
Afghanistan	0	0.081	1	0.71%
Cambodia	0	0.059	2	1.43%
DRC	0	0.058	3	2.14%
Pakistan	0	0.032	4	2.86%
Uganda	0	0.028	5	3.57%
Angola	ongoing	0.028	6	4.29%
Algeria	0	0.022	7	5.00%
Indonesia	0	0.022	8	5.71%
Sudan	ongoing	0.021	9	6.43%
Burundi	0	0.017	10	7.14%
n (1974-2000)		n (2001)		
3161		140		

Onsets and Forecasts for 2002

Country	*Onset of Genocide / Politicide*	*Forecast*	*Rank Order*	*Percent of "n"*
DRC	0	0.170	1	0.72%
Congo	0	0.053	2	1.44%
Sudan	ongoing	0.036	3	2.16%
Zimbabwe	0	0.027	4	2.88%
Angola	ongoing	0.025	5	3.60%
Nigeria	0	0.023	6	4.32%
Tajikistan	0	0.023	7	5.04%
Pakistan	0	0.019	8	5.76%
Indonesia	0	0.017	9	6.47%
Uzbekistan	0	0.016	10	7.19%
n (1974-2001)		n (2002)		
3299		139		

Table D - Predictor Variable Values for Genocide Onsets, 1988-2003

Forecast %	*Onset year*	*State*	*StateLedDiscrimination$_{t-1}$*	*PreviousGenocides$_{t-1}$*	*ElectionPeriod$_t$*	*Instability$_{t-1}$*	*PrInst$_t$*	*UnconstrainedExecutive×HumanDefenseBurden$_{t-1}$*	*YearsSinceGenocide$_{t-1}$*	*EthnicPowerRelations$_{t-1}$*	*EthnicFractionalization$_{t-1}$*	*Polity$_{t-1}$*	*PolityChange3$_{t-1}$*	*UnconstrainedExecutive×fdHumanDefenseBurden$_{t-1}$*	*NumberIGOs$_{t-1}$*	*Assassination$_{t-1}$*	*HumanDefenseBurden$_{t-1}$*	*Assassination$_{t-2}$*
1.45%	2003	Sudan	1	2	0	1	0.99	6724.51	18	2	0.71	-6	1	0.00	72	0	6724.51	0
2.13%	1997	DRC	1	2	0	1	0.99	2833.80	18	2	0.93	0	0	563.93	58	0	2833.80	0
2.56%	1994	Rwanda	1	1	0	1	0.91	1924.25	29	1	0.18	-6	1	9.99	44	1	1924.25	0
3.31%	1988	Burundi	1	1	0	0	0.01	646.66	21	2	0.33	-7	0	-0.81	39	0	646.66	0
3.52%	1998	Angola	1	1	0	1	0.99	0.00	21	2	0.76	-3	-2	0.00	45	0	5799.59	0
4.96%	1988	Iraq	1	1	1	1	0.98	54058.20	23	2	0.55	-9	0	6006.47	55	0	54058.20	0
13.22%	1988	Somalia	1	0	0	0	0.04	3190.44	27	0	0.81	-7	0	0.00	46	0	3190.44	0
26.72%	1993	Burundi	0	2	1	1	0.91	833.84	3	1	0.33	-3	4	63.30	42	0	833.84	0
40.00%	1989	Sri Lanka	1	0	1	1	0.99	0.00	40	1	0.43	5	0	0.00	58	0	1321.42	1
42.25%	1998	Yugoslavia	1	0	0	0	0.02	7258.42	49	1	0.57	-6	1	636.63	38	1	7258.42	0
42.62%	1992	Bos. & Herz.	0	0	0	0	0.00	0.00	31	1	0.57	-5	0	0.00	53	1	10666.73	0
		Mean	0.25	0.19	0.54	0.17	0.22	1524.02	35.88	0.92	0.48	1.76	0.76	-28.38	55.12	0.12	7498.19	0.11
		SD	0.43	0.51	0.50	0.37	0.37	6671.37	11.15	0.51	0.26	7.14	3.37	1773.45	22.04	0.32	18108.33	0.32
		Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-10.00	-15.00	-54594.44	1.00	0.00	0.00	0.00
		Max	1.00	3.00	1.00	1.00	1.00	82935.81	54.00	2.00	1.00	10.00	17.00	31579.44	129.00	1.00	207158.70	1.00