Online Appendix

May 29, 2014

Appendix A: Coding Rules for Event Data

Our main concern was the very different data structure between GDELT and the two hand-coded data sets. GDELT codes a wide variety of events using the CAMEO coding scheme, which produces an event typology that has no easy analogue to ACLED or GED. For example, ACLED and GED both have a simple coding scheme identifying whether or not a violent event targeted civilians. However, GDELT has no clear coding for violence against civilians: for example, it has one category for "assault", containing events ranging from political repression to assassination to terrorist attacks, and another category for "fight" containing a variety of military engagements. In other words, GDELT codes by technology, while ACLED and GED code by interaction and actors.

Since our goal is to create a subset of the GDELT that maps onto the ACLED and GED as well as possible, we therefore focus on one specific type of event with clear definition across all three data sets: incidents of conventional military violence initiated by an armed group. This can include violence against civilians, but does not include riots, protests, or violence between non-military groups.

GED

The GED focuses explicitly on violent events initiated by an armed group and resulting in at least one fatality, making it the most specific in terms of event type. Only events with location identified at the second-level administrative division (equivalent to a county in the United States) or below are included.

ACLED

The ACLED also focuses on political violence, but differs from the GED in that (1) fatalities are not required for an event to make it into the data set, and (2) a variety of non-violent or non-military interactions, such as public protests or movement of headquarters, are also included in the data set.

As such, we subset by the following variables:

1. Geographic precision: only events with location identified as "a small part of a region" or below are included.

2. Actors involved: Only violent events initiated by an armed group (government forces, insurgents, or political or ethnic militias) are included. Events initiated by civilians, rioters or protestors are not included.

GDELT

The GDELT coding scheme, coupled with the 'noisy' nature of the data, means that subsetting is a more complicated process, and the size of the data set is highly sensitive to coding rules. To ensure robustness, we use three different coding schemes with various levels of 'permissiveness' in terms of information requirements and specificity.

Coding scheme 1 (moderate information requirements, presented in main paper)

1. Event type: only events in the 19 event root code ('fight') are included.

2. Event subtype: non-violent events in this category (191, 'impose blockade') are not included.

3. Geographic precision: only events with location identified at the city level are included.

4. Actor identification: only events where the initiator is identified as a militarized group (GOV, MIL, INS, or REB) are included.

Coding scheme 2 (low information requirements)

1. Event type: only events in the 190 family ('fight') are included.

2. Event subtype: non-violent events in this category (191, 'impose blockade') are not included.

3. Geographic precision: only events with location identified at the city level are included.

4. Actor identification: All actor codes, including null values, are included.

Coding scheme 3 (high information requirements)

1. Event type: only events in the 190 family ('fight') are included.
2. Event subtype: non-violent events in this category (191, 'impose blockade') are not included.
3. Geographic precision: only events with location identified at the city level are included.
4. Actor identification: only events where the initiator is identified as a militarized group (GOV, MIL, INS, or REB) and the target is identified as either a militarized group or civilians (CIV) are included.

Appendix B: Robustness Checks with Alternate GDELT Coding

This section shows figures and tables summarizing results of parallel analysis using the two alternate GDELT coding schema.

Results using coding scheme 2 (low information requirements)

Low information requirements lead to an explosion of GDELT observations, and a huge number of false positives in the data. However, all the relationships identified in the paper remain significant here as well.

	ACLED	GED	GDELT
ACLED	1.00	0.33	0.33
GED	0.33	1.00	0.25
GDELT	0.33	0.25	1.00

Table 1: Correlations: Event-cells by Cell-Month

	ACLED	GED	GDELT
ACLED	1.00	0.43	0.71
GED	0.43	1.00	0.39
GDELT	0.71	0.39	1.00

Table 2: Correlations: Summed Event-cells by Month

		ACLED	
		0	1
GDELT	0	530,973	3,634
	1	9,966	3,329

Table 3: Confusion Matrix: ACLED

		GED	
		0	1
GDELT	0	532,197	2,410
	1	11,203	2,002

Table 4: Confusion Matrix: GED

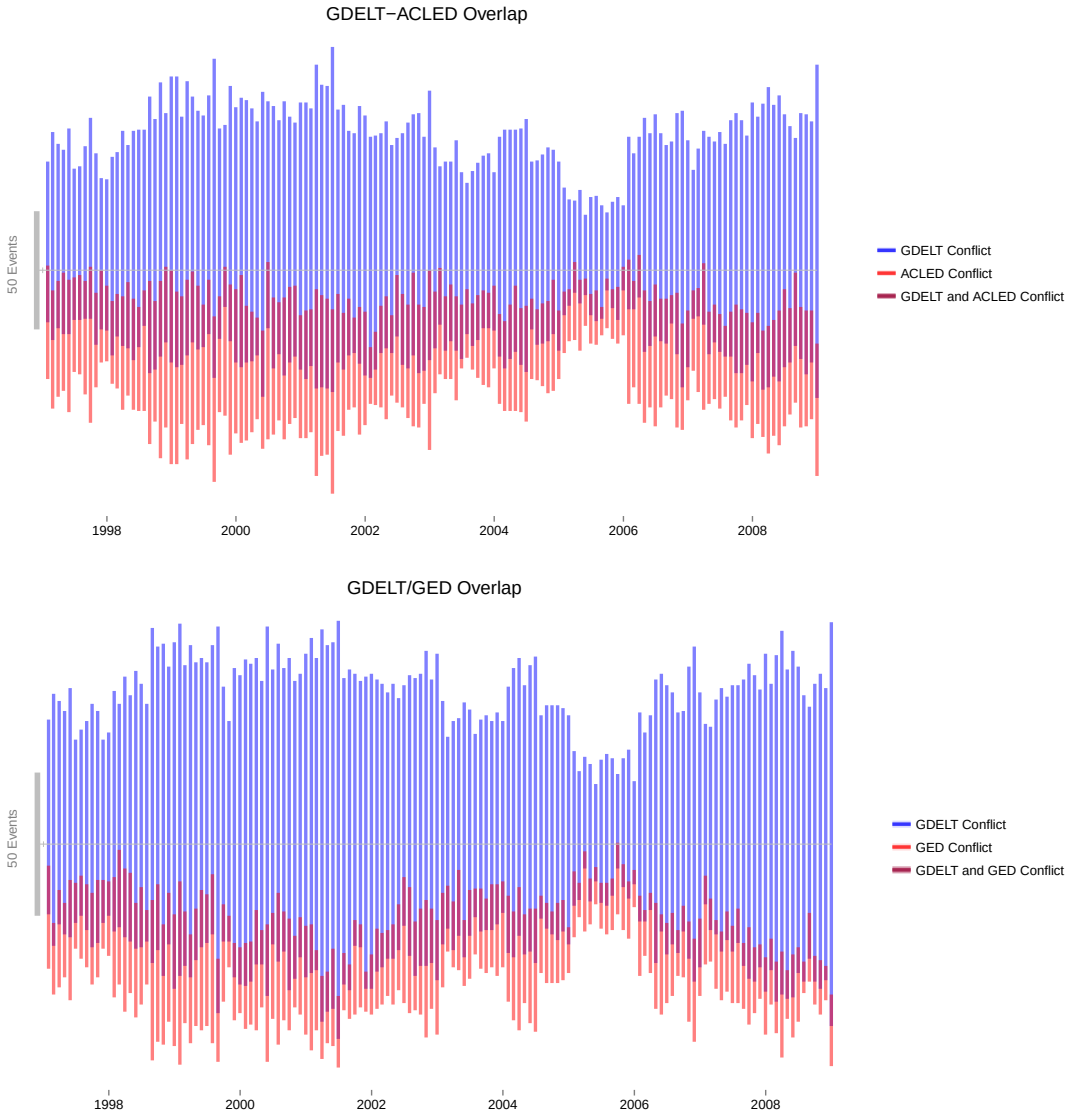

Figure 1: Grid-cell events over time

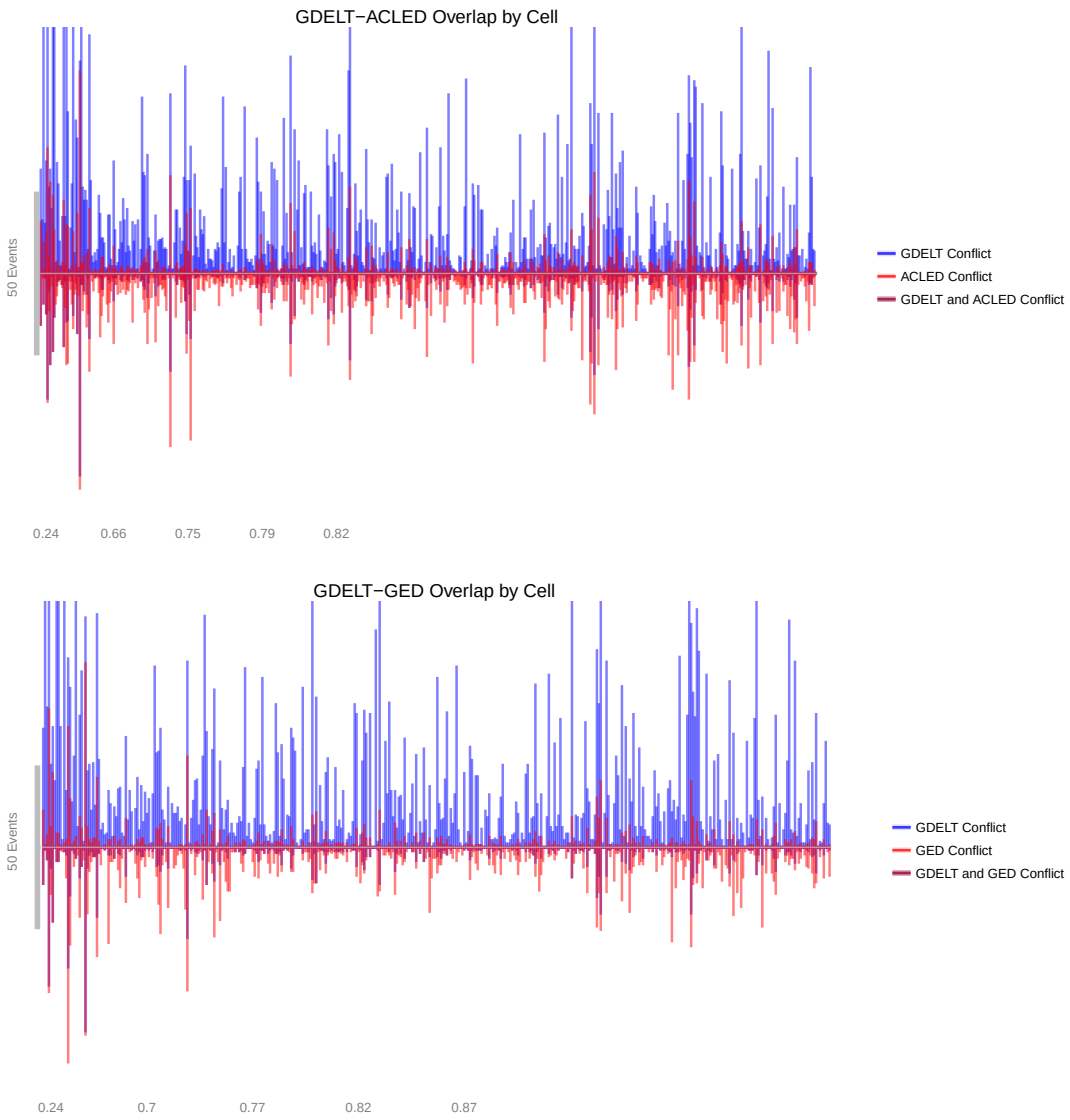

Figure 2: Grid-cell events over logged and normalized capital distance

Table 5

	GDELT-ACLED		*GDELT-GED*	
	GDELT = 1, ACLED = 0	GDELT = 0, ACLED = 1	GDELT = 1, GED = 0	GDELT = 0 GED = 1
	(1)	(2)	(3)	(4)
Population	0.68^{***}	0.61^{***}	0.70^{***}	0.68^{***}
	(0.01)	(0.02)	(0.01)	(0.02)
Capital Distance	-1.80^{***}	2.15^{***}	-1.88^{***}	3.80^{***}
	(0.11)	(0.21)	(0.11)	(0.28)
% Mountainous	0.60^{***}	0.69^{***}	0.58^{***}	0.58^{***}
	(0.04)	(0.07)	(0.04)	(0.08)
Constant	-8.78^{***}	-12.02^{***}	-9.16^{***}	-16.65^{***}
	(0.30)	(0.44)	(0.31)	(1.08)
Observations	547,812	547,812	547,812	547,812
Log Likelihood	−37,730.44	−19,204.65	−40,426.54	−13,769.00
Akaike Inf. Crit.	75,516.89	38,465.30	80,909.09	27,593.99

Note: *p<0.1; **p<0.05; ***p<0.01

Country-level fixed effects not shown.

Table 6: Logit results: event mismatch by cell-month

	Dependent variable:		
	ACLED Conflict	GED Conflict	GDELT Conflict
	(5)	(6)	(7)
$\text{Conflict}_{(t-1)}$	2.60^{***}	2.63^{***}	2.51^{***}
	(0.04)	(0.05)	(0.02)
$\text{Spatial lag}_{(t-1)}$	0.06^{***}	0.13^{***}	0.01
	(0.003)	(0.01)	(0.003)
Distance to capital	0.14	1.18^{***}	-1.41^{***}
	(0.15)	(0.18)	(0.11)
Population	0.60^{***}	0.65^{***}	0.62^{***}
	(0.01)	(0.02)	(0.01)
% Mountainous	0.44^{***}	0.31^{***}	0.48^{***}
	(0.06)	(0.07)	(0.04)
Constant	-10.24^{***}	-12.05^{***}	-8.73^{***}
	(0.36)	(0.44)	(0.30)
Observations	547,812	547,812	547,812
Log Likelihood	−26,121.86	−19,047.40	−39,372.64
Akaike Inf. Crit.	52,303.71	38,154.80	78,805.29
Note:			*p<0.1; **p<0.05; ***p<0.01
			Country-level fixed effects not shown.

Table 7: Logistic regression results. Dependent variable: Occurrence of violence in cell/month.

Results using coding scheme 3 (high information requirements)

High information requirements dramatically lower the number of 'valid' GDELT observations. While all regression results hold substantively, the balance of false positives and false negatives shifts significantly.

	ACLED	GED	GDELT
ACLED	1.00	0.33	0.20
GED	0.33	1.00	0.15
GDELT	0.20	0.15	1.00

Table 8: Correlations: Events by Cell-Month

	ACLED	GED	GDELT
ACLED	1.00	0.43	0.54
GED	0.43	1.00	0.36
GDELT	0.54	0.36	1.00

Table 9: Correlations: Events by Cell-Month

		ACLED	
	0	1	
GDELT	0	540,029	6,184
	1	910	689

Table 10: Confusion Matrix: ACLED

		GED	
	0	1	
GDELT	0	542,220	3,993
	1	1,180	419

Table 11: Confusion Matrix: GED

Figure 3: Grid-cell events over time

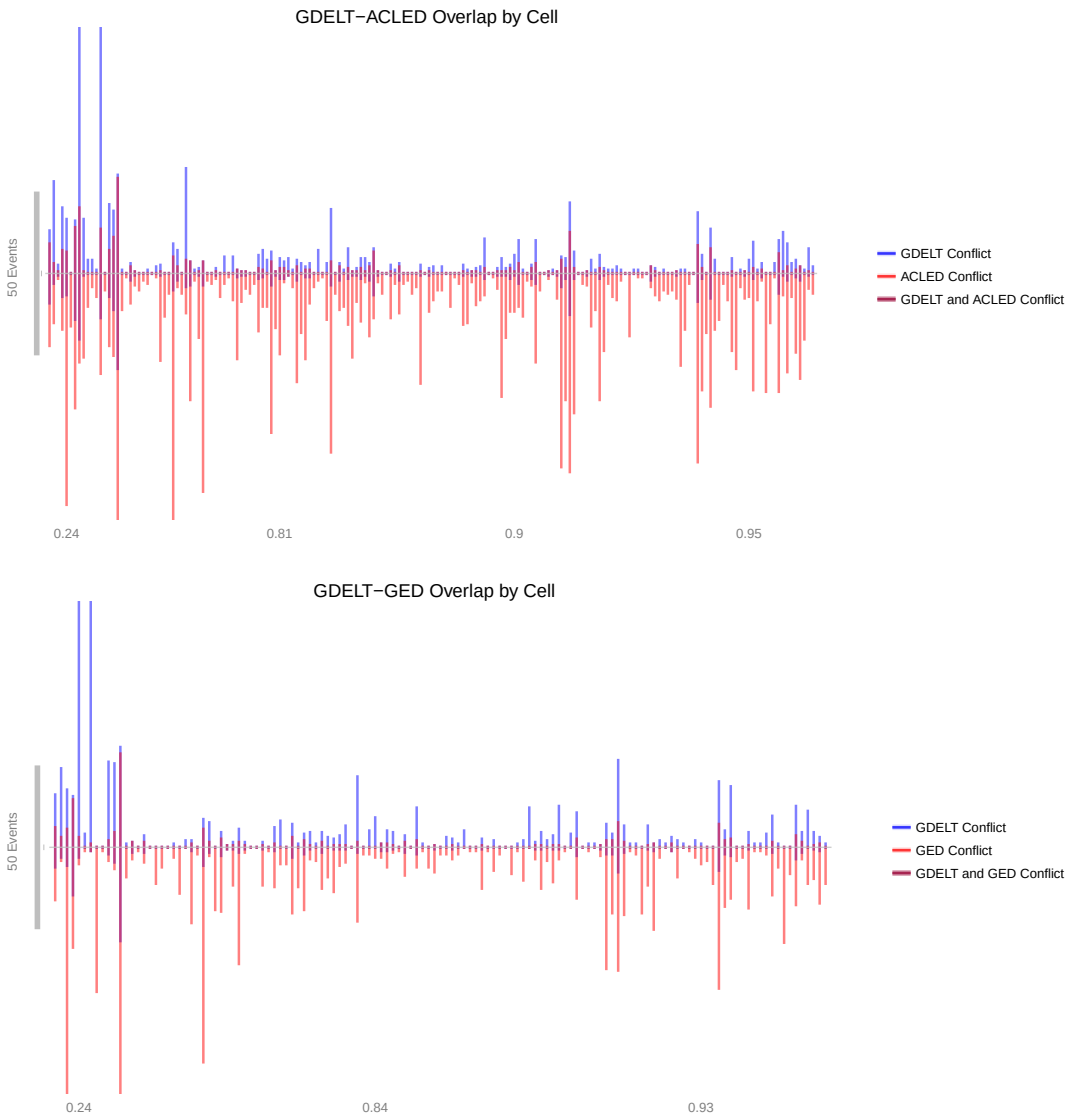

Figure 4: Grid-cell events over logged and normalized capital distance

Table 12

	Dependent variable:			
	GDELT = 1, ACLED = 0	GDELT = 0, ACLED = 1	GDELT = 1, GED = 0	GDELT = 0 GED = 1
Population	0.62^{***}	0.69^{***}	0.70^{***}	0.73^{***}
	(0.04)	(0.01)	(0.03)	(0.02)
Capital Distance	-5.31^{***}	0.47^{***}	-5.18^{***}	0.87^{***}
	(0.30)	(0.15)	(0.26)	(0.18)
% Mountainous	0.89^{***}	0.62^{***}	0.75^{***}	0.51^{***}
	(0.16)	(0.05)	(0.14)	(0.06)
Constant	-7.14^{***}	-11.02^{***}	-9.63^{***}	-12.61^{***}
	(0.79)	(0.35)	(1.14)	(0.47)
Observations	547,812	547,812	547,812	547,812
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13: Logit results: event mismatch by cell-month

	Dependent variable:		
	ACLED Conflict	GED Conflict	GDELT Conflict
	(5)	(6)	(7)
$\text{Conflict}_{(t-1)}$	2.60^{***}	2.63^{***}	2.80^{***}
	(0.04)	(0.05)	(0.08)
$\text{Spatial lag}_{(t-1)}$	0.06^{***}	0.13^{***}	-0.004
	(0.003)	(0.01)	(0.01)
Distance to capital	0.14	1.18^{***}	-3.46^{***}
	(0.15)	(0.18)	(0.25)
Population	0.60^{***}	0.65^{***}	0.67^{***}
	(0.01)	(0.02)	(0.03)
% Mountainous	0.44^{***}	0.31^{***}	0.59^{***}
	(0.06)	(0.07)	(0.13)
Constant	-10.24^{***}	-12.05^{***}	-9.36^{***}
	(0.36)	(0.44)	(0.70)
Observations	547,812	547,812	547,812
Log Likelihood	−26,121.86	−19,047.40	−7,128.79
Akaike Inf. Crit.	52,303.71	38,154.80	14,317.58
Note:			*p<0.1; **p<0.05; ***p<0.01
			Country-level fixed effects not shown.

Table 14: Logistic regression results. Dependent variable: Occurrence of violence in cell/month.

Appendix C: Robustness Checks using Count Data

The dependent variable in these models is the number of events occurring in a given cell-month, not the general presence or absence of any violent events. In this case, we find a slight but significant capital-distance bias for both ACLED and GED; however, the pattern of over-reporting near the capital is much stronger for GDELT in all cases. We are unsurprised by the fact that more event *reports* occur closer to the capital, as even for hand-coded data sets with a wide variety of sources,

the capital city is likely to be covered by media and other sources much more closely. The fact that the tendency for GDELT to over-report events close to the capital even after binarizing the data tells us that this coverage issue affects not only the number of events reported, but the geographic pattern of reported events. As such, these results further reinforce our basic finding: that GDELT has a severe bias in reporting events closer to the capital city.

	GDELT	ACLED	GED
GDELT	1.00	0.31	0.34
ACLED	0.31	1.00	0.48
GED	0.34	0.48	1.00

Table 15: Correlations: Number of Events by Cell-Month

	Dependent variable:		
	ACLED Conflict	GED Conflict	GDELT Conflict
	(5)	(6)	(7)
$\text{Conflict}_{(t-1)}$	0.48^{***}	0.51^{***}	0.52^{***}
	(0.001)	(0.001)	(0.001)
$\text{Spatial lag}_{(t-1)}$	0.01^{***}	0.01^{***}	-0.01^{***}
	(0.0004)	(0.0004)	(0.004)
capdist_norm	-0.10^{***}	-0.07^{***}	-0.33^{***}
	(0.01)	(0.004)	(0.01)
pop_2000	0.004^{***}	0.002^{***}	0.0003
	(0.0004)	(0.0002)	(0.0005)
mnt	0.01^{***}	0.01^{***}	-0.01^{**}
	(0.003)	(0.001)	(0.003)
Constant	0.16^{***}	0.06^{***}	0.33^{***}
	(0.03)	(0.01)	(0.03)
Observations	547,812	547,812	547,812
R^2	0.28	0.29	0.28
Adjusted R^2	0.28	0.29	0.28
Note:			*p<0.1; **p<0.05; ***p<0.01
			Country-level fixed effects not shown.

Table 16: Logistic regression results. Dependent variable: Occurrence of violence in cell/month.

Appendix D: Robustness Checks with Alternate ACLED/GED Coding

This section shows figures and tables summarizing results of parallel analysis using a subset of sources for ACLED and GED. Here we subset all observations to only include those gleaned from the four main sources used to create GDELT: Agence France-Presse, Associated Press, Xinhua, and BBC Monitoring.

Results using coding scheme 2 (low information requirements)

Subsetting the sources used to create ACLED/GED produces a smaller pair of data sets that oddly enough, correlate worse overall with GDELT.

	ACLED	GED	GDELT
ACLED	1.00	0.27	0.21
GED	0.27	1.00	0.17
GDELT	0.21	0.17	1.00

Table 17: Correlations: Event-cells by Cell-Month

	ACLED	GED	GDELT
ACLED	1.00	0.46	0.46
GED	0.48	1.00	0.43
GDELT	0.46	0.43	1.00

Table 18: Correlations: Summed Event-cells by Month

		ACLED	
		0	1
GDELT	0	541,537	2,546
	1	2,988	741

Table 19: Confusion Matrix: ACLED

		GED	
		0	1
GDELT	0	541,941	2,142
	1	3,172	557

Table 20: Confusion Matrix: GED

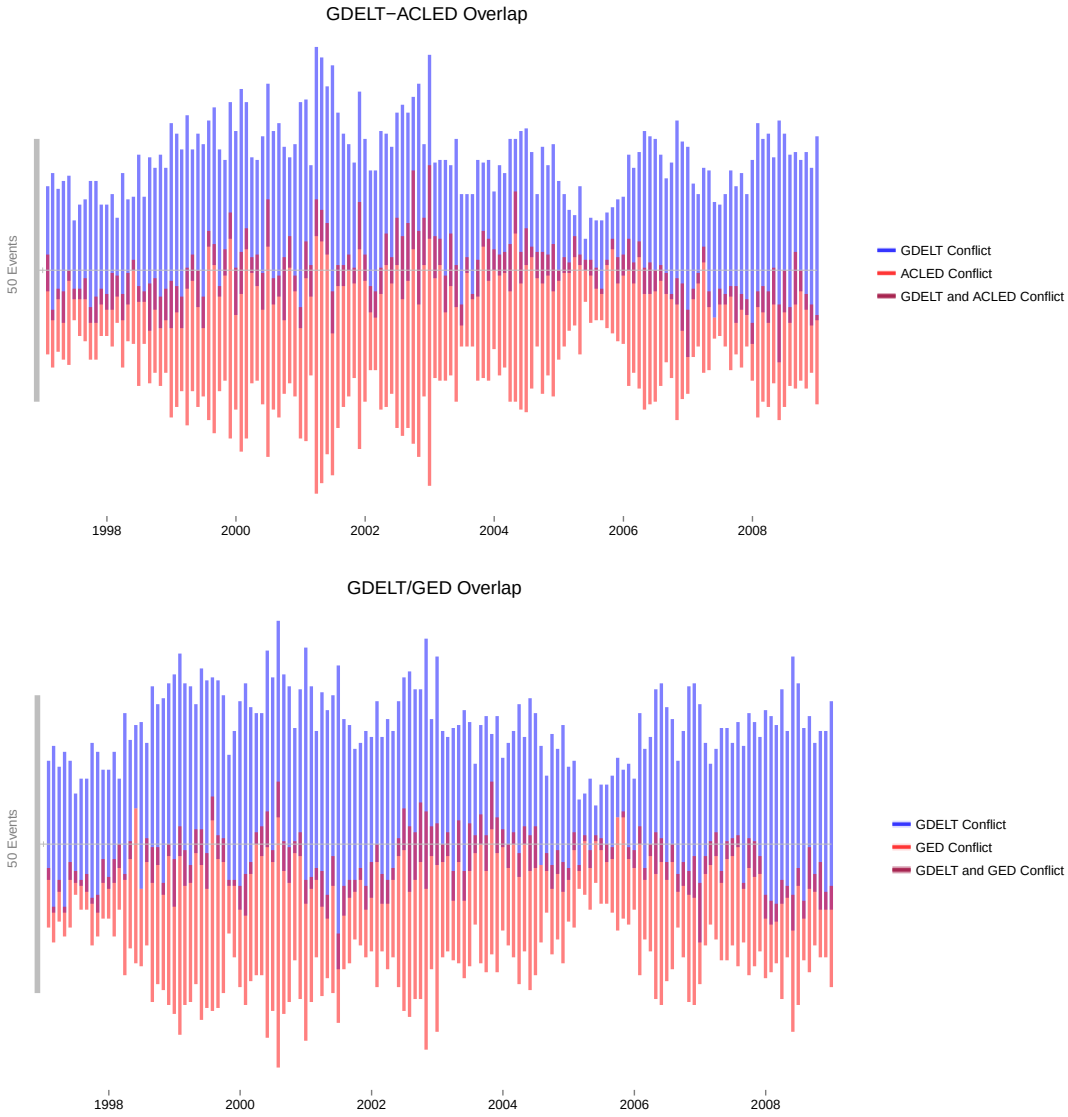

Figure 5: Grid-cell events over time

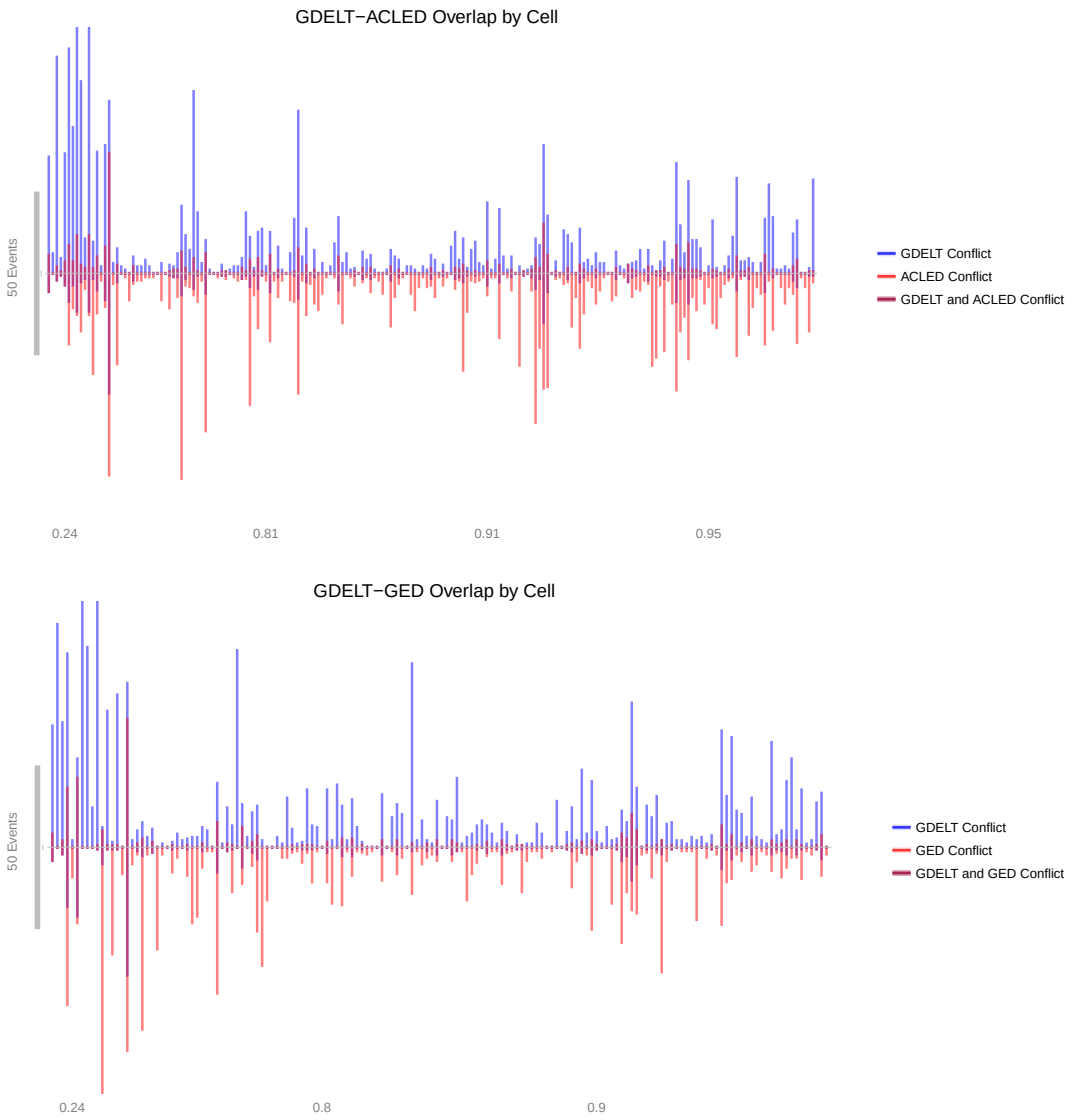

Figure 6: Grid-cell events over logged and normalized capital distance

Table 21

	GDELT-ACLED		*GDELT-GED*	
	GDELT = 1, ACLED = 0	GDELT = 0, ACLED = 1	GDELT = 1, GED = 0	GDELT = 0 GED = 1
	(1)	(2)	(3)	(4)
Population	8.32***	7.46***	8.32***	8.13***
	(0.26)	(0.25)	(0.25)	(0.25)
Capital Distance	−7.86***	3.03***	−7.65***	−0.48
	(0.28)	(0.45)	(0.27)	(0.42)
% Mountainous	0.67***	0.53***	0.70***	0.26***
	(0.08)	(0.08)	(0.08)	(0.08)
Constant	−17.51***	−25.39***	−18.00***	−24.58***
	(0.90)	(0.99)	(0.91)	(1.08)
Observations	547,812	547,812	547,812	547,812
Log Likelihood	−13,765.25	−13,172.47	−14,391.47	−11,371.54
Akaike Inf. Crit.	27,586.51	26,400.93	28,838.93	22,799.07

Note: *p<0.1; **p<0.05; ***p<0.01
Country-level fixed effects not shown.

Table 22: Logit results: event mismatch by cell-month

	Dependent variable:		
	ACLED Conflict	GED Conflict	GDELT Conflict
	(5)	(6)	(7)
$\text{Conflict}_{(t-1)}$	2.62^{***}	2.63^{***}	2.84^{***}
	(0.06)	(0.06)	(0.05)
$\text{Spatial lag}_{(t-1)}$	0.10^{***}	0.16^{***}	0.01
	(0.01)	(0.01)	(0.004)
Capital Distance	-0.19	-0.59	-5.29^{***}
	(0.36)	(0.39)	(0.29)
Population	6.96^{***}	7.33^{***}	7.60^{***}
	(0.24)	(0.24)	(0.24)
% Mountainous	0.35^{***}	0.14^{*}	0.54^{***}
	(0.08)	(0.08)	(0.08)
Constant	-21.29^{***}	-22.10^{***}	-18.38^{***}
	(0.90)	(0.96)	(0.88)
Observations	547,812	547,812	547,812
Log Likelihood	$-14{,}267.98$	$-12{,}172.38$	$-14{,}472.43$
Akaike Inf. Crit.	28,595.96	24,404.77	29,004.86
Note:			*p<0.1; **p<0.05; ***p<0.01
			Country-level fixed effects not shown.

Table 23: Logistic regression results. Dependent variable: Occurrence of violence in cell/month.