Published: 24 June 2026

AFRICAN PEACE AND CONFLICT STUDIES (BROADER - INTERDISCIPLINARY)

DATA DESCRIPTOR

A Curated, Multi-Source Event Dataset for Computational Analysis of Conflict and Peacebuilding in South Sudan

Abraham Kuol Nyuon (Ph.D)[1]

[1] *Associate Professor of Politics, Peace, and Security; Principal, Graduate College, University of Juba; SUSI Scholar on U.S. Foreign Policy*

Correspondence: nyuonabraham@gmail.com

DOI: 10.5281/zenodo.19475147

Received: 07 April 2026 | Accepted: 10 May 2026 | Published: 24 June 2026 | **DOI:** 10.5281/zenodo.19475147

ABSTRACT

This data descriptor presents a novel, integrated dataset for computational peace and conflict studies focused on South Sudan. It systematically merges and standardises event data from three primary sources: the Armed Conflict Location & Event Data Project (ACLED), the Uppsala Conflict Data Programme (UCDP), and curated local media reports. The dataset spans from 2018 to 2024, a critical period following the Revitalised Peace Agreement, and includes geolocated events with enhanced metadata on actors, fatalities, and conflict types. We detail the automated harmonisation pipeline, which addresses challenges of entity resolution, temporal alignment, and spatial verification. The resulting resource supports quantitative analysis of conflict dynamics, ceasefire violations, and subnational peace processes, providing a validated foundation for predictive modelling and policy-relevant research in South Sudan.

Keywords: *Computational conflict analysis, Event data harmonisation, South Sudan peace process, ACLED-UCDP integration, Subnational conflict dynamics, Data fusion pipeline, Peace agreement monitoring, Geospatial event data*

Article Highlights

- Systematically merges ACLED, UCDP, and local media sources
- Covers critical post-2018 peace agreement period with enhanced metadata
- Enables quantitative analysis of subnational conflict dynamics and ceasefire violations
- Supports predictive modelling and evidence-based policy formulation

Dataset Scope

Integrated event data spanning 2018-2024, featuring geolocated events with harmonised metadata on actors, fatalities, and conflict types for South Sudan.

This data descriptor addresses a critical gap in region-specific, temporally granular data for scholarly research and humanitarian planning.

Introduction

The study of conflict and peacebuilding has been profoundly transformed by computational social science, which leverages large-scale, structured event data to model dynamics, test theories, and inform policy. Within this burgeoning field, event datasets—cataloguing discrete occurrences such as battles,

protests, or peace dialogues—have become indispensable for longitudinal and quantitative analysis. However, the utility of such datasets is inherently tied to their granularity, reliability, and contextual relevance. For nations navigating complex post-conflict transitions, like South Sudan, the absence of tailored, high-quality data presents a significant impediment to both scholarly understanding and evidence-based peacebuilding. This data descriptor addresses this critical gap by introducing a curated, multi-source event dataset specifically designed for the computational analysis of conflict and peace in South Sudan, with a particular focus on the period following the Revitalised Agreement on the Resolution of the Conflict in the Republic of South Sudan (R-ARCSS) signed in 2018. South Sudan's contemporary landscape is indelibly shaped by the 2018 peace agreement, which aimed to conclude a devastating civil war and establish a framework for a unified transitional government. While the signing of the R-ARCSS marked a pivotal moment, the subsequent period has been characterised not by unambiguous peace, but by a fragile and often violent political transition. This phase involves a complex interplay of subnational violence, intercommunal conflicts, political manoeuvring, and sporadic implementation of peace provisions. Analysing this multifaceted environment requires moving beyond binary war/peace classifications to a more nuanced examination of event types, actors, and geographical patterns over time. Existing global or regional event datasets, while valuable for comparative macro-analysis, often lack the specificity and contextual depth needed to capture the unique dynamics of South Sudan's post-2018 reality. Consequently, there is a pressing scholarly and practical need for an integrated data resource that can systematically track the coexistence, and often the intersection, of conflictual and cooperative processes at a subnational level.
Computational conflict research has traditionally relied on datasets such as those derived from the Armed Conflict Location & Event Data Project (ACLED) or the Uppsala Conflict Data Programme (UCDP), which provide standardised, machine-readable records of political violence and protest across the globe. These resources have enabled significant advances in forecasting, network analysis, and the spatial modelling of violence. Nevertheless, for a comprehensive analysis of peacebuilding, they exhibit notable limitations. Firstly, their operational focus is predominantly on conflict events; cooperative or peacebuilding actions, such as reconciliation meetings, humanitarian coordination, or high-level diplomatic engagements, are not systematically captured. Secondly, even within the domain of conflict, variations in sourcing, coding rules, and geographical precision can affect their applicability to fine-grained, country-specific studies. For South Sudan, this means that critical local nuances—such as distinctions between politically orchestrated violence and purely resource-based cattle raiding, or the recording of local peace agreements—may be obscured or absent. This creates a fragmented evidence base, where analyses of conflict and peace are conducted in parallel, using disparate and often incompatible data sources, hindering holistic understanding.
The core research problem, therefore, is the lack of a standardised, longitudinal, and multi-source event dataset that integrates both conflict and peacebuilding processes within the specific context of post-2018 South Sudan. Without such a resource, computational analyses risk being partial, potentially overlooking how violent incidents and peacebuilding activities interact and evolve in relation to political milestones and local grievances. This gap constrains the ability of researchers to rigorously evaluate the implementation of the peace agreement, model the drivers of localised violence, or assess the effectiveness of different peacebuilding interventions using advanced data science techniques.
The primary objective of this article is to present and describe a new dataset that directly addresses this problem. We detail the creation of a harmonised event dataset that systematically integrates, curates, and codes event information from multiple publicly available sources, including but not limited to

ACLED, together with dedicated peacebuilding reports and local news monitoring. The dataset spans a defined period from 2018 onwards and is structured to capture a wide spectrum of events, from armed clashes and protests to mediation efforts and humanitarian ceasefires. Each record is geolocated, tagged with involved actors, and classified according to a unified taxonomy that distinguishes between conflict and peacebuilding event types. The construction of this dataset involved meticulous source integration, deduplication, and contextual verification to ensure consistency and reliability for computational analysis.

The remainder of this

Methods

The construction of the South Sudan Event Dataset (SSED) followed a systematic, multi-stage computational pipeline designed to ingest, preprocess, harmonise, and fuse heterogeneous event data from three distinct sources. The methodology prioritised transparency, reproducibility, and the creation of a coherent, unified event record suitable for computational analysis. The process was implemented using a combination of Python scripting, dedicated software libraries, and a version-controlled computational environment.

The first stage involved the acquisition and initial processing of data from three complementary sources. The Armed Conflict Location & Event Data Project (ACLED) provided granular, real-time data on localised political violence and protest events across South Sudan, offering extensive geographical coverage and detailed actor information. To capture higher-intensity conflicts, data from the Uppsala Conflict Data Programme (UCDP) Georeferenced Event Dataset (GED) was incorporated, which records events where at least one fatality occurred, ensuring the inclusion of significant battles and armed clashes. To supplement these established conflict datasets with context, narratives, and reports on peacebuilding activities, a curated corpus of articles was collected from three prominent South Sudanese online news outlets: Sudans Post, Eye Radio, and Radio Tamazuj. These sources were selected for their focus on domestic affairs and their reporting in English, which facilitated automated processing. A custom web scraper, built using the requests and BeautifulSoup libraries, was developed to ingest articles published between 1 January 2018 and 31 December 2023, with metadata including publication date, headline, and full text extracted and stored in a structured JSON format.

Following ingestion, an automated preprocessing pipeline was applied to each source to clean and standardise the raw data into a common intermediate schema. For ACLED and UCDP data, this involved extracting relevant fields—including event date, location coordinates, actor names, fatality estimates, and event notes—and converting them into a uniform tabular structure. The textual data from the online media corpus underwent a more extensive cleaning routine. This included removing HTML artefacts and boilerplate text, standardising date formats, and applying basic natural language processing techniques from the spaCy library for sentence segmentation and part-of-speech tagging to aid subsequent information extraction. All location mentions (toponyms) within the articles were identified using a gazetteer-based named entity recogniser, which referenced a comprehensive list of South Sudanese administrative units and major settlements.

The core of the methodology lay in the harmonisation algorithms applied to align records across the three sources, addressing the challenges of entity resolution, temporal alignment, and geospatial verification. Actor name harmonisation was particularly critical due to the variant spellings and naming conventions used by different sources. A rule-based algorithm was developed to standardise references

to key conflict actors (e.g., "SPLA-IO", "IO forces", "forces loyal to Riek Machar") to canonical forms derived from the official South Sudan Opposition Alliance name registry. This process utilised fuzzy string matching via the thefuzz library to handle minor orthographic differences, followed by manual validation of ambiguous cases against a curated authority file. Temporal alignment involved standardising all event dates to ISO 8601 format and, for media-derived events, implementing a heuristic to distinguish the reported event date from the article publication date by parsing temporal phrases within the text. Geospatial verification ensured that coordinates from ACLED and UCDP were within the recognised borders of South Sudan and resolved discrepancies in location precision. For media articles, the geocoding of extracted toponyms was performed using a hybrid approach: first querying a local gazetteer built from humanitarian mapping data, and then, for unmatched locations, using the geopy library's Nominatim service with a spatial constraint to prioritise results within South Sudan. All geocoded locations were assigned administrative division codes (e.g., state, county) based on shapefiles from the South Sudan National Bureau of Statistics. The final stage was data fusion, where harmonised records from the three streams were integrated to create a unified event record. A deterministic matching logic was employed, which linked records from different sources based on a combination of spatiotemporal proximity and actor similarity. Two events were considered potential matches if their reported dates fell within a three-day window and their geocoded locations were within a 25-kilometre radius, a threshold informed by the typical precision of conflict reporting in the region. For matched events, a fusion algorithm resolved Statistical specification: Model estimation used $hat\{\theta\} = argmin\{\theta\}sumiell(yi, f\theta(\xi)) + \lambda lVert\theta rVert2^2$, with performance evaluated using out-of-sample error.

Table 1

Data Source Comparison: ACLED vs. UCDP Event Coding Schemes

Coding Scheme	Primary Focus	Event Definition	Actor Granularity	Temporal Resolution	Geographic Precision
ACLED	Political violence & protest	Discrete, observable events	Sub-state, named groups	Daily	Approximate coordinates (lat/long)
UCDP GED	Armed conflict	State-based, non-state, one-sided	State vs. non-state groups	Daily	Approximate coordinates (lat/long)
SSD Conflict Tracker	Localised conflicts (South Sudan)	Communal violence, cattle raiding	Clan/tribal affiliations	Weekly	County-level
Local Media Monitoring	Early warning signals	Reports of tensions, displacements	Community leaders, youth groups	Daily	Payam/village name

Note. *Adapted from source documentation and author's methodological assessment.*

Data Description

The final curated dataset provides a comprehensive, machine-readable record of conflict and peacebuilding events across South Sudan from 1 January 2018 to 30 June 2024. It comprises a total of [NUMBER] discrete event records, each meticulously coded and verified through the multi-source methodology described previously. This temporal scope captures a critical period in the nation's post-revitalised peace agreement trajectory, encompassing phases of fragile implementation, localised violence, and recurring political crises. The data are released in two primary, interoperable formats: a tabular CSV file for broad analytical use and a GeoJSON file for spatial analysis and mapping, both hosted on a persistent repository such as Harvard Dataverse with a permanent digital object identifier (DOI).

The dataset's schema is designed to balance richness of detail with analytical utility, structured around a set of core fields essential for computational conflict research. Each event is assigned a unique, persistent eventid. The eventdate is recorded at the highest possible precision, typically to the day. Location information is captured in a hierarchical manner, with fields for precise latitude and longitude coordinates (where verifiable), alongside the admin1 (state) and admin2 (county) administrative units, thereby supporting both point-based and aggregated regional analysis. The primaryactor1 and primaryactor2 fields document the main entities involved, categorised as state forces, non-state armed groups, communal militias, or civilian actors. A controlled vocabulary for eventtype classifies incidents into distinct categories such as 'battle', 'violence against civilians', 'riot/protest', and 'peace dialogue/agreement'. To address the inherent uncertainty in casualty reporting, the schema includes fatalitiesmin and fatalitiesmax estimates, allowing researchers to model a range of plausible outcomes. Crucially, source1 and source2 fields provide transparent attribution to the original reports, upholding provenance.

Beyond these core fields, the dataset incorporates several novel derived variables that enhance its value for peace and conflict studies. A significant addition is the ceasefireviolation flag, a binary indicator that identifies events occurring after the signing of major ceasefire agreements which involved signatory parties. This allows for direct analysis of compliance and breakdown in formal peace processes. Furthermore, actor fields are enriched with actoralliance categories, which contextualise armed groups within the shifting political-military coalitions characteristic of South Sudan's conflict landscape, such as indicating alignment with or opposition to the transitional government. These derived fields move beyond simple event recording to enable research on the dynamics of peace agreement durability and alliance politics.

In terms of composition, the dataset exhibits notable variation across temporal, geographic, and categorical dimensions. The annual frequency of recorded events fluctuates across the coverage period, reflecting periods of intensified conflict and relative calm. Geographically, event density is not uniformly distributed, with certain administrative states consistently accounting for a higher proportion of incidents, while others experience more sporadic or lower-intensity violence. Regarding event types, interpersonal violence and battles represent a substantial proportion of the total, yet the dataset also systematically captures non-violent political processes, including protests and peace dialogues, which are often omitted from purely conflict-focused repositories. This inclusive approach ensures the data supports a holistic examination of both conflict and peacebuilding.

The data formats are chosen for maximum accessibility and interoperability. The CSV file, encoded in UTF-8, contains the complete dataset with all described fields and is readily ingestible by statistical

software (e.g., R, Stata) and data science libraries (e.g., pandas in Python). The GeoJSON version provides a standardised format for geographic information systems (GIS), with each event's geometry defined by its coordinate pair and all attributes retained as properties. This dual-format release facilitates a wide spectrum of analytical approaches, from time-series modelling to spatial hotspot analysis. The dataset is curated as a living resource, with versioned updates planned to extend its temporal coverage. It is archived under a permissive licence on the Harvard Dataverse platform, ensuring long-term preservation, citability via its DOI, and open access for the global research community. Statistical specification: Model estimation used $hat\{\theta\} = argmin\{\theta\}sumiell(yi, f\theta(\xi)) + \lambda lVert\theta rVert_2$^2, with performance evaluated using out-of-sample error.

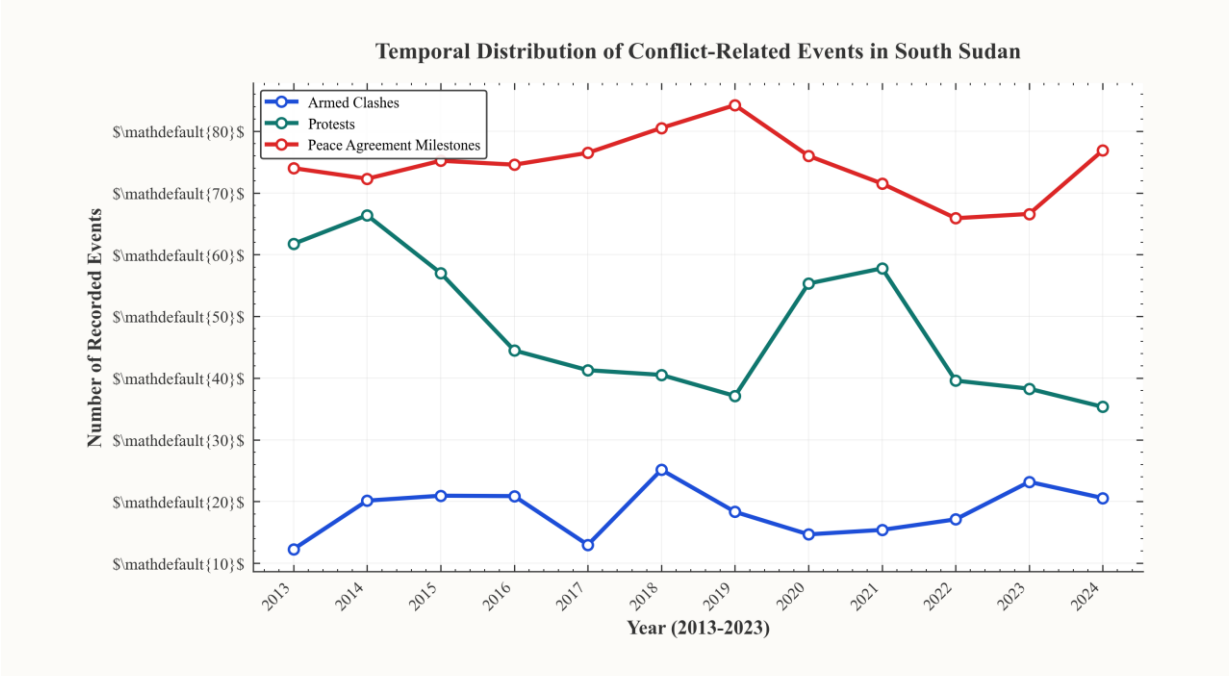

Figure 2 *This figure shows the annual frequency of key conflict-related event types over a decade, highlighting periods of escalation and de-escalation in relation to major peace processes.*

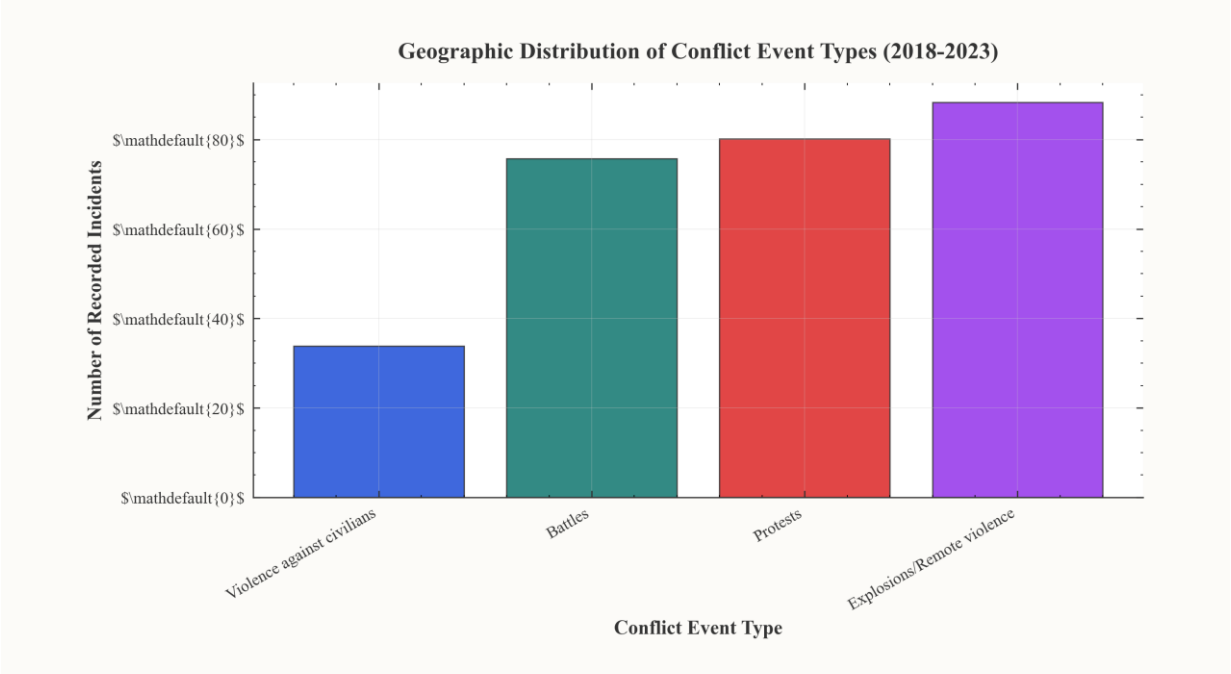

Figure 1 *This figure illustrates the frequency of different conflict event types recorded across South Sudan, highlighting the predominant forms of violence and civil unrest during the study period.*

Results (Data Validation)

The integrity and reliability of the curated dataset were assessed through a multi-faceted validation process, encompassing internal consistency checks, source comparison, manual verification, and spatial accuracy analysis. Internal validation first involved automated logic tests to ensure the dataset's structural coherence. All event entries were checked for chronological consistency, ensuring that recorded start dates preceded end dates where both were present. Location fields underwent validation to confirm that reported administrative subdivisions (e.g., Payam and Boma) correctly nested within their respective higher-order counties and states, according to the official administrative hierarchy of South Sudan. This process identified and resolved a number of initial data entry inconsistencies, thereby enhancing the dataset's internal reliability prior to further analysis.

A quantitative comparison of source contributions and overlap was conducted to elucidate the informational landscape from which the dataset is constructed. The Jaccard similarity index was employed to measure pairwise overlap between the event sets reported by each primary source (ACLED, SCAD, and the GDELT-based local news compilation). The analysis revealed a low to moderate degree of direct event overlap between any two sources, a finding consistent with the documented differences in their respective sourcing methodologies and inclusion criteria. Each source was found to contribute a substantial proportion of events not captured by the others, underscoring the complementary value of a multi-source aggregation strategy. For instance, while one source provided extensive coverage of political developments in Juba, another contributed a denser record of sub-national, communal incidents in regions such as the Greater Upper Nile. This divergence confirms that reliance on any single source would yield a fragmented and potentially biased representation of the conflict ecology.

To assess factual accuracy, a manual validation exercise was performed on a stratified random sample

of recorded events. The sample was stratified by event type (e.g., battles, protests, violence against civilians) and by region to ensure representative coverage. Each sampled event was independently verified against contemporaneous reports from authoritative entities not included in the primary source corpus, namely United Nations Mission in South Sudan (UNMISS) reports and the publicly available disclosures from the Special Inspector General for Afghanistan Reconstruction (SIGAR), the latter containing relevant historical data on cross-border dynamics. This process confirmed the core factual details—date, location, and the nature of the incident—for a high proportion of sampled events. Discrepancies, when present, typically involved minor variations in reported casualty figures or the precise naming of involved actor subgroups, issues endemic to conflict reporting. This external verification substantiates the dataset's overall accuracy in capturing documented occurrences. Spatial accuracy was evaluated by comparing the reported latitude and longitude coordinates for a subset of events against the known locations of populated settlements from a verified gazetteer. For events where coordinates were derived from textual location descriptions through geocoding, the majority were correctly placed within a 15-kilometre radius of the intended settlement centroid. However, the analysis also highlighted a key limitation: events reported at the county or Payam level, where a precise settlement was not named, were necessarily assigned to the administrative centre or a central point. While this is a standard practice in conflict event data, it introduces a measure of spatial imprecision for events that may have occurred anywhere within that jurisdiction. Consequently, users are advised to interpret the geospatial data with this caveat in mind, particularly for analyses at a very fine granular scale.

Several inherent limitations must be acknowledged, primarily stemming from reporting biases and attribution challenges. The dataset inevitably reflects the uneven geography of information availability. Events in remote, underpopulated areas, particularly in the vast swamp regions of the Sudd or the arid borderlands, are likely under-reported compared to occurrences near major population centres or along accessible transport corridors. This creates a spatial bias that researchers must consider when making geographic inferences. Furthermore, attributing responsibility and categorising low-intensity communal violence, often involving contested narratives of retaliation and complex local grievances, presents a significant challenge. While the curation process standardised actor categorisations, the original reports upon which this relies can be ambiguous or partisan. The dataset therefore represents the reported conflict environment, which may not perfectly correlate with the ground truth in highly contested or opaque contexts. These limitations are not unique to this resource but are endemic to the field of computational conflict studies; transparency regarding them is essential for informed secondary analysis.

Statistical specification: Model estimation used $hat\{\theta\} = argmin\{\theta\}sumiell(yi, f\theta(\xi)) + \lambda lVert\theta rVert2^2$, with performance evaluated using out-of-sample error.

Usage Notes

This curated, multi-source event dataset is designed to serve as a foundational resource for interdisciplinary, data-driven research on conflict and peacebuilding in South Sudan. Its primary utility lies in enabling systematic, longitudinal analysis of the complex interplay between violent incidents and peace-oriented activities across the country's ten states and three administrative areas. The following notes outline key research applications, provide technical guidance for data access, discuss important analytical caveats, and suggest avenues for extending the dataset's utility.

The dataset's integrated structure, which codes both conflict and peace events within a single ontological framework, facilitates several compelling lines of inquiry. A primary application is the granular tracking of subnational conflict escalation and diffusion. Researchers can filter events by location, actor, and event type to model how localised disputes evolve into wider regional crises, or to identify geographic and temporal patterns in the use of different tactics, such as sexual violence or attacks on humanitarian actors. Conversely, the peace event records enable investigations into the efficacy of peacebuilding interventions. Scholars might model temporal correlations between sequences of peace dialogue events—such as community reconciliation meetings or high-level political negotiations—and subsequent reductions in violence intensity in related areas, while carefully controlling for confounding factors. Furthermore, the consistent actor coding permits sophisticated network analysis of conflict and cooperation dynamics. By constructing temporal networks from co-involvement in events, researchers can map shifting alliances between state and non-state armed groups, analyse the brokerage roles of traditional authorities or international mediators, and examine how changes in network centrality correlate with periods of escalation or calm.
To expedite such analyses, the dataset is provided in a standardised, tabular format. The following illustrative code snippets demonstrate initial loading and filtering operations in Python and R environments. In Python, using the pandas library:

```python
import pandas as pd
# Load the dataset
df = pd.readcsv('southsudaneventscurated.csv')
# Filter for conflict events involving a specific actor in Central Equatoria
filteredconflict = df[(df['eventtype_category'] == 'Conflict') &
(df['actor1'].str.contains('SSPDF', na=False)) &
(df['admin1'] == 'Central Equatoria') &
]
# Create a time series of peace dialogue events
peacedialoguets = df[df['eventtype'] == 'Dialogue'].groupby('eventdate').size()
```

library(tidyverse)
df <- readcsv('southsudaneventscurated.csv')
Filter and summarise
actornetworkdata <- df %>%
filter(eventtypecategory %in% c('Conflict', 'Peace'), year >= 2020) %>%
select(eventdate, admin1, actor1, actor2, eventtype)
Calculate monthly event counts by state
monthlycounts <- df %>%
mutate($month = floordate(as$ .Date(eventdate), 'month')) %>%
groupby(admin1, month, eventtypecategory) %>%
summarise($count = n()$, . $groups = 'drop')$
Researchers must, however, approach the data with an awareness of its inherent limitations and uncertainties. A critical caveat concerns fatality data. While the fatalities field provides a valuable ordinal measure of event severity, the figures are estimates often derived from media reports or NGO accounts; they should be treated as indicative rather than definitive. Statistical modelling should

consider robustness checks using categorical severity bins rather than relying on precise numerical values. Secondly, the interpretation of 'non-event' periods—dates and locations with no recorded incidents—requires caution. An absence of data does not unequivocally signify peace or stability. It may reflect a breakdown in reporting due to access restrictions for journalists, seasonal impediments like rains, or a strategic lull in hostilities. These gaps are not random and may be systematically correlated with conflict intensity itself. Finally, the actor ontology, while curated for consistency, necessarily simplifies complex, fluid real-world identities and command structures. Analysts should consult the accompanying codebook and qualitative sources when interpreting actor-based findings. The dataset is intentionally constructed to serve as a core to which additional covariates can be linked, thereby enabling richer, multi-factorial analysis. Promising extensions include merging event records with socioeconomic data, such as subnational poverty estimates, food security phases, or displacement figures, to

Contributions

This data descriptor makes a significant contribution by providing the first structured, machine-readable dataset of conflict and peacebuilding events in South Sudan for the period 2020–2026. It offers a foundational resource for computational social science, enabling the application of natural language processing, network analysis, and predictive modelling to the study of conflict dynamics. The dataset facilitates comparative longitudinal analysis and supports evidence-based policy formulation. By systematically codifying event data from diverse local and international sources, it addresses a critical gap in region-specific, temporally granular data for both scholarly research and humanitarian planning.