ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Conflict forecasting using remote sensing data: An application to the Syrian civil war☆

Daniel Racek [a,*], Paul W. Thurner [b], Brittany I. Davidson [c], Xiao Xiang Zhu [d], Göran Kauermann [a]

[a] Institute of Statistics, Ludwig-Maximilians-University Munich, Germany
[b] Institute of Political Science, Ludwig-Maximilians-University Munich, Germany
[c] School of Management, University of Bath, United Kingdom
[d] School of Engineering and Design, Technical University of Munich, Germany

ARTICLE INFO

Keywords:
Remote sensing
Satellite imagery
Conflict prediction
Forecasting
Machine learning
Statistical modeling
Syria

ABSTRACT

Conflict research is increasingly influenced by modern computational and statistical techniques. Combined with recent advances in the collection and public availability of new data sources, this allows for more accurate forecasting models in ever more fine-grained spatial areas. This paper demonstrates the utilization of remote sensing data as a potential solution to the lack of official data sources for conflict forecasting in crisis-ridden countries. We evaluate and quantify remote sensing data's differentiated impact on forecasting accuracy across fine-grained spatial grid cells using the Syrian civil war as a use case. It can be shown that conflict, particularly its onset, can be forecasted more accurately by employing publicly available remote sensing datasets. These results are consistent across a range of established statistical and machine learning models, which raises the hope to get closer to reliable early-warning systems for conflict prediction.

1. Introduction

Conflict prediction has been considered the number one task of peace research for decades (Singer, 1973). According to Hegre, Metternich, Nygård, and Wucherpfennig (2017), early work on conflict prediction was inspired by the works of Richardson (1960), Sorokin (1962), Wright (1965) and influenced by the Correlates of War Project (Small, Singer, & Bennett, 1982), which in 1963 started systematically collecting quantitative data on war, adhering to scientific principles. During this time, the development of early-warning systems for conflict was already one of the central goals of conflict research. Shifting to the 1970s and early 1980s, interest in this space declined, where explicit prediction studies were an exception in the published literature (Hegre et al., 2017). This subsequently changed in the decades thereafter. Although there were early efforts to move away from country–year datasets (Schrodt & Gerner, 2000) such as the Correlates of War Project, in the following decades most of the prediction studies predominantly focused on large-scale country-level events, such as civil wars (Gleditsch & Ward, 2013; Goldstone et al., 2010; Harff & Gurr, 1998; King & Zeng, 2001). Only within the last decade, an increasing number of studies have moved to more fine-grained subnational levels (Bazzi et al., 2022; Hegre et al., 2019; Koren & Bagozzi, 2017), as more and more disaggregated global conflict datasets have become available (Raleigh, Linke, Hegre, & Karlsen, 2010; Sundberg & Melander, 2013). Most recently, a forecasting competition organized by ViEWS (Vesco et al., 2022) showcased the

☆ **Funding Information:** This work is supported by the Helmholtz Association, Germany under the joint research school "Munich School for Data Science - MUDS".

* Correspondence to: Institut für Statistik, Ludwigstr. 33, 80539 München, Germany.
E-mail address: daniel.racek@stat.uni-muenchen.de (D. Racek).

https://doi.org/10.1016/j.ijforecast.2023.04.001
0169-2070/

Please cite this article as: D. Racek, P.W. Thurner, B.I. Davidson et al., Conflict forecasting using remote sensing data: An application to the Syrian civil war. International Journal of Forecasting (2023), https://doi.org/10.1016/j.ijforecast.2023.04.001.

current advances in the field in both methodology and data sources.

One of the main difficulties in subnational analyses is often the lack of (reliable) structural data sources, such as population density or economic indicators commonly associated with conflict (Blattman & Miguel, 2010; Jerven, 2013). Hence, subnational studies are either conducted using the (low-level) administrative zones of those countries that have sufficient data available (see Bazzi et al., 2022), or are using the well-known 0.5 × 0.5 decimal degree PRIO grid cells, corresponding to an area of roughly 55 × 55 km (at the Equator; Tollefsen, Strand, & Buhaug, 2012). The former are highly country-dependent and thus cannot be easily compared. Administrative zones often vary greatly in size, can be irregular in shape and may change over time, which renders spatial inference more difficult (Wood & Sullivan, 2015). The latter specification via PRIO grid cells, however, is quite coarse, with a cell size of roughly 55 × 55 km, which similarly limits spatial inference. Furthermore, defining cells in decimal degrees means that their size differs by a substantial margin across the world. For example, in Africa alone, the width of the PRIO grid cells differs by up to 11 km based on this definition. But even when using one of those two spatial structures, sufficient data availability for developing and/or crisis-ridden countries are rare, which makes reliable forecasting of conflict in those countries, in which it is needed the most, particularly difficult.

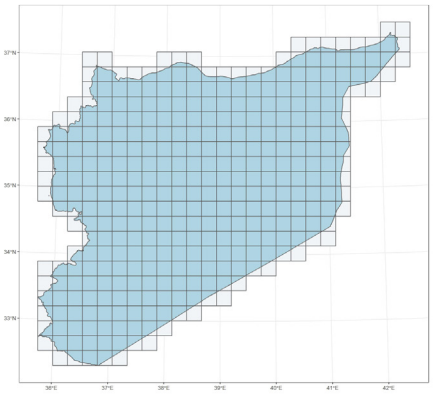
Recently, new and emerging data sources, such as news (Attinà, Carammia, & Iacus, 2022; Mueller & Rauh, 2018), social media (Zeitzoff, 2017), and remote sensing data (Avtar et al., 2021), have increasingly gained attention to solve these problems. This work examines the capabilities of remote sensing data for the task of conflict prediction. Remote sensing data are acquired by applying complex prediction pipelines on sets of high-resolution satellite images. This results in highly fine-grained datasets previously unheard of. Notably, these datasets typically have global coverage, which creates a number of opportunities for conflict research. For instance, as we show in this work, this allows for their use anywhere across the world in custom-defined spatial areas of any size or shape. In recent years, a number of new, high-quality, and high-resolution remote sensing datasets have been made publically available, such as improved global landcover maps (Buchhorn et al., 2020) and vegetation indicators (FAO, 2022). This has become possible due to long-term records of satellite imagery through satellite systems such as AVHRR and MODIS (Pedelty et al., 2007), Landsat Loveland and Dwyer (2012), and Sentinel (Berger, Moreno, Johannessen, Levelt, & Hanssen, 2012), in combination with improvements in classification techniques through, for example, deep learning (Ball, Anderson, & Chan Sr, 2017).

Our work utilizes (novel) remote sensing datasets in order to forecast conflict in self-defined, fine-grained, and regular-sized cells across Syria and tests as well as quantifies their effectiveness for this task. Syria experienced (and to some extent still experiences) one of the largest and deadliest civil wars of the past century, with more than 392,000 recorded fatalities by the end of 2020 (Pettersson et al., 2021). According to scholars, the uprising and the subsequent civil war started with mass protests in the city of Dara'a (Leenders & Heydemann, 2012, p. 142) in March 2011, which quickly escalated due to repression and the use of heavy force exerted by government security forces (Leenders & Heydemann, 2012, p. 149) and ultimately led to war. Although there was a significant decline in violence in 2020, as of today, the war is still ongoing (Human Rights Watch, 2022; Pettersson et al., 2021).

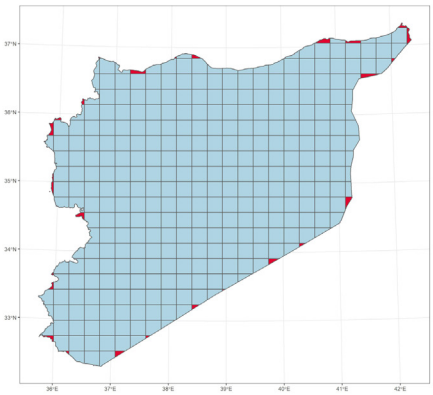
Given Syria's long history of conflict, it presents itself as an ideal use case for examining the potentials of remote sensing data, because the availability of other data sources is sparse to non-existent. To the best of our knowledge, one can likely only obtain the social demographics for the 14 Syrian governorates based on the 2004 census, as well as location polygons of selected ethnic groups. This problem of data limitation, which exists for many developing and crisis-ridden countries, can be alleviated by drawing on remote sensing data sources with global coverage, as motivated in this paper.

In this work, we systematically test the effectiveness of various remote sensing datasets for spatial forecasts of armed conflict across Syria and quantify the change in performance using each data source for this task. Only recently, remote sensing data have been identified as an essential addition in the development of early warning systems (Avtar et al., 2021). Hence, we extend this notion by systematically analyzing their effectiveness for forecasting. Additionally, we further extend current work in the field, as the use of remote sensing data allows us to conduct our forecasting in custom-defined cells. This means that we are not constrained in having to use traditional administrative zones or the PRIO grid cells in our analyses. Instead, we manually construct cells that are more fine-grained than those employed in other studies. These cells are regular and fixed in size across Syria, as they are defined in the Universal Transverse Mercator (UTM) coordinate system. It is worth noting that through this definition, our cells are indeed country-independent. Moreover, all of our employed data sources have global coverage. Hence, applications, extensions, and comparisons to other countries or even continents can easily be undertaken.

Specifically, for our analysis, we rasterize Syria into 25 × 25 km cells and match these cells with various remote sensing datasets. Then, for each cell, we construct aggregated remote sensing variables and use those alongside other traditional predictors to forecast the monthly occurrence of armed conflict. We do this through a one-step-ahead recursive window forecast using a range of established statistical and machine learning models. By repeatedly re-running our models with different specifications, in which we alter the set of included variables, we are able to quantify the effectiveness of each remote sensing dataset with respect to a classical literature-inspired baseline specification. This allows us to evaluate the gain of using remote sensing data for conflict prediction without being reliant on a specific model type. We provide details on our forecasting procedure and chosen models in our methodology section and an in-depth discussion of model selection in our discussion.



(a) The original (uncropped) cells are depicted in grey. The relevant Syrian area is marked in blue.



(b) The 34 removed cells with an area of less than 150 km^2 are depicted in red. The others, which are used in this study, are depicted in blue.

Fig. 1. Illustration of the rasterization and cropping process to construct the 322 Syrian cells used as the main observation units in the study.

We are able to show that by adding remote sensing variables to our baseline, we can consistently improve the overall forecasting performance of our models. By further differentiating conflict into onset and persistence, a distinction frequently made and discussed in the literature (Bazzi et al., 2022; Blattman & Miguel, 2010; Fearon & Laitin, 2003), we show that most of the overall performance increase stems from the former. In other words, utilizing remote sensing data primarily helps to predict new conflicts in areas not suffering from conflicts before. According to our results, population, landcover, and crop data are the most important remote sensing-based predictors of conflict. Our full specification, including variables from each dataset, performs the most consistently well across all types of models. Finally, we want to highlight the generalizability and ease of reproducibility of our study, with which we are answering a recent call by Vesco et al. (2022). Our definition of cells is country-independent and all of our employed data sources are freely available and have global coverage.

The rest of the paper is organized as follows. First, we describe all data sources utilized and how they are processed, followed by a thorough description of our methodological approaches and evaluation criteria. Next, we report the results of our study, before providing an in-depth discussion and a final conclusion.

2. Data

2.1. Constructing the dataset

For our analysis, we rasterize Syria into fine-grained evenly sized 25 $\times$ 25 km (625 km^2) grid cells, in which we forecast the monthly occurrence of armed conflict from 2011 to 2020.[1] We chose this time period because 2011 marks the year in which the Syrian civil war broke out, and 2020 is the last year for which the Uppsala Conflict Data Program (UCDP) Georeferenced Event Dataset (GED) provides conflict data. The rasterization takes place in accordance with the UTM coordinate system (zone 36N), a projected coordinate system, in order to avoid cell size distortions. Such distortions would otherwise occur when using any geographic coordinate system (e.g. the World Geodetic System (WGS) 1984 with latitude/longitude coordinates). Additionally, we crop any cells crossing the Syrian border, because for this case study we are only interested in conflict taking place within Syria. Last but not least, we remove any cells with an area of less than 150 km^2 after cropping (roughly a quarter of the full cell size), as those cells almost never experience conflict and hence might skew the results. Nonetheless, we additionally report our results without this filtering in Appendix I. This process results in a total of 356 cells, of which 34 are removed after cropping. The whole process is illustrated in Fig. 1. The resulting 322 cells are our primary observations of interest in this study.

For the time period between 2011 and 2020, we match all subsequently described (remote sensing) datasets based on time and geolocation to our cells of interest and motivate their potential use based on findings from the conflict literature. We emphasize that all data used in this study are freely available and can be directly accessed and downloaded from the web addresses provided in Table A.6 in Appendix A. All our code and the processed datasets can be found on the Open Science Framework (OSF) here.

For our remote sensing data sources, perfect spatial matching is not always possible, as we are matching a raster of cells (consisting of remote sensing data) to our raster of Syrian cells. Hence, in some instances, the cells of the former will overlap with two or more Syrian cells. To solve this issue, we relatively distribute their contribution based on the percentage of covered area of the respective Syrian cells. An illustrative example of this procedure is shown in Fig. 2.

[1] We exclude the Quneitra Governorate, which mainly consists of the Israeli-occupied Golan Heights and the United Nations Disengagement Observer Force (UNDOF) buffer zone.

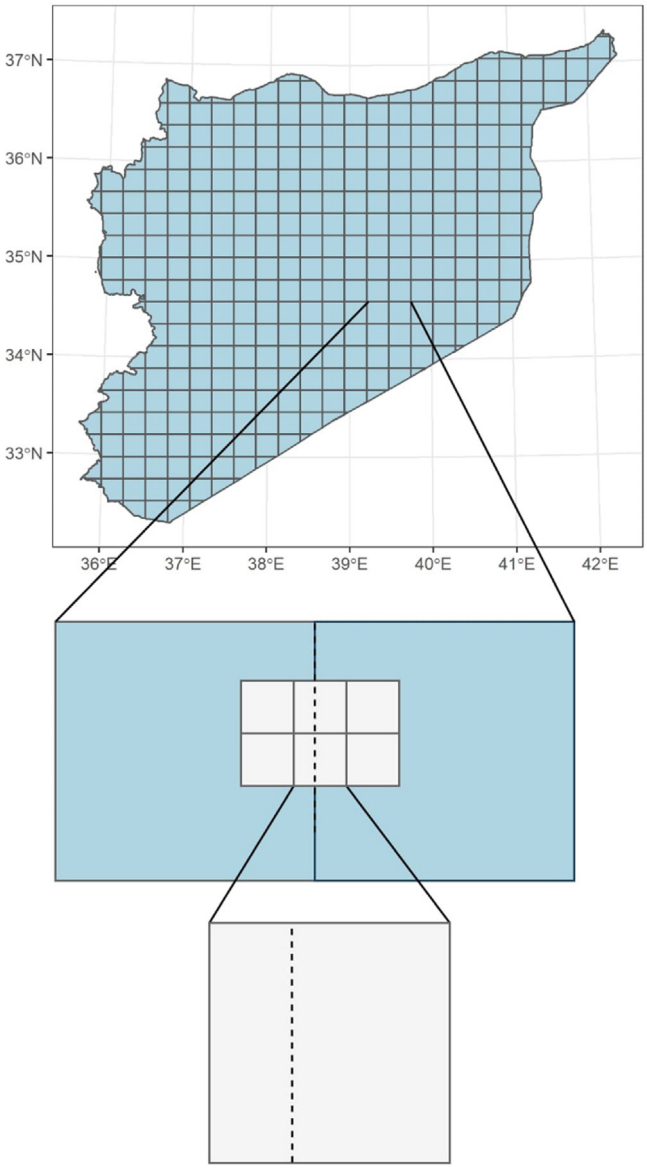


Fig. 2. Illustration of the matching process for overlapping remote sensing data grid cells. The top shows all the resulting Syrian grid cells (in blue) after the rasterization process described in Section 2.1. In the middle, we zoom in on two of those constructed cells and plot six remote sensing data cells (in grey; e.g. consisting of population numbers) onto them. These remote sensing cells (and their data values) are matched to the Syrian grid cells in the matching routine. The four cells on the sides (two on the left, and two on the right) are entirely covered by a single Syrian grid cell and thus matched in their entirety to the respective cell. The two remote sensing cells in the middle are partially covered by both of the Syrian grid cells. Hence, a relative distribution of their data values is necessary. In the bottom, we zoom in on one of those remote sensing data cells. From the dashed line we can infer that one-third of the cell is covered by the left Syrian grid cell, and the remaining two-thirds by the right cell. Hence, in this instance, we would allocate one-third of the data value (e.g. the amount of population) to the left cell and two-thirds to the right cell. The same strategy is applied for all remote sensing datasets.

2.2. Conflict & ethnic data

Data on armed conflict were drawn from the widely known UCDP Georeferenced Event Dataset (GED) (Sundberg & Melander, 2013). The dataset reports events of organized violence, resulting in at least one estimated direct death through armed force, across the world from 1989 to 2020. The data were systematically collected and coded by experienced researchers using national and international news reports, as well as data from NGOs and international organizations. Each event is (among others) assigned a specific date, place, type of violence, and estimated number of resulting fatalities. This study focuses on forecasting battle-related fatalities, i.e. deaths resulting from either state-based or non-state conflict between organized parties. It additionally uses (lagged) one-sided violence, i.e. violence against civilians, as an explanatory variable in order to account for preceding escalatory processes. For a small portion of events, the exact location and/or time point is unknown. We discard these events

in our analysis.[2] As a result, around 15% of the Syrian events from 2011 onwards are discarded. Naturally, this is a limitation in our study. All remaining events are matched based on location and time to the respective cell in the respective month of the study period. Hence, for each cell we have aggregated monthly information on the prevalence of conflict from the beginning of 2011 to the end of 2020.

As noted by scholars (Abosedra, Fakih, & Haimoun, 2021; Ismail, 2011), ethnicities play a central role in the Syrian civil war. Hence, we gather information on the location of "politically relevant ethnic groups" from the Geo-referencing Ethnic Power Relations (GeoEPR) 2021 dataset (Vogt et al., 2015), which is part of the Geographical Research On War, Unified Platform ($GROW^{up}$) (Girardin, Hunziker, Cederman, Bormann, & Vogt, 2015). It assigns every ethnic group to settlement patterns and provides polygons of their location globally from 1946 to 2020. For Syria, the identified (and largest) ethnic groups are the Sunni Arabs, Alawis, Christians, Kurds, and Druze. For each of those groups we create an indicator that describes whether at least 5% of the respective cell area is covered by the respective settlement polygon.

2.3. Remote sensing data

Population data come from the WorldPop project (Tatem, 2017). The project estimates population numbers across the world in 100 m resolution grid cells, with the help of census data and detailed geospatial datasets through a semi-automated dasymetric modeling technique using random forests. In order to ensure that these estimates are as close to reported real-world population numbers, we chose the estimates adjusted to match national UN numbers. Furthermore, because the Syrian population estimates are based on the 2004 census, and the civil war reportedly led to large-scale migration across the country (Kelley, Mohtadi, Cane, Seager, & Kushnir, 2015), we only employ the population numbers stemming from the year 2010 in this study. Additional information on, for example, age and sex structures are only available on a country-wide level and thus are not considered. We aggregate these estimated population numbers for each of the 322 Syrian cells to obtain total amounts. In accordance with earlier studies (Raleigh & Hegre, 2009), we expect areas with higher population numbers to be more likely to experience conflict and thus this information to be highly relevant.

Landcover information is drawn from the Copernicus Global landcover map collection 3 (Buchhorn et al., 2020), which classifies the entire world into 23 different landcover classes at 100 m resolution through an elaborate prediction pipeline (supervised classification, expert rules, temporal cleaning via break detection, etc.) on a yearly basis from 2015 to 2019. We employ some of the "level 1" classes, such as cropland, forest, and permanent water, for which the authors report an average accuracy of 80.6%. Data from each year of the landcover map are matched to the cells accordingly, and average shares for each class (per cell) are derived. These landcover classes contain structural information that might be relevant for the prediction of conflict, for instance, the amount of urban area (related to population, see above; and/or economic activity, see below), crop area (see below), bare area (e.g. desert), tree-covered area (with potential hideouts), or the existence of rivers. The latter might be related to strategically important locations, as rivers might allow for the transportation of weapons and food, and can be used for energy production.

Topography data are collected from (Amatulli et al., 2018). The authors calculate a variety of elevation-based topographic variables such as slope, roughness, and terrain ruggedness for the entire globe using the digital elevation model products of the 250 m Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) (Danielson & Gesch, 2011), which was released in 2010 by the United States Geological Survey and the National Geospatial-Intelligence Agency. They aggregate each variable to a resolution of 1, 5, 10, 50, and 100 km using different spatial aggregation methods. In this study, we employ several topographic variables from the most fine-grained 1 km resolution dataset and derive the respective median of each variable (as also used by the authors for the original aggregate calculation) for each cell. Topography, specifically ruggedness or rough terrain, has been associated with conflict (Collier & Hoeffler, 2004; Fearon & Laitin, 2003), as such terrain arguably provides protection and opportunities to hide for rebels.

Satellite-observed nighttime lights are drawn from the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) version 4 nighttime stable lights, which discards ephemeral events and background noise, for the years 2010 to 2013 (Baugh, Elvidge, Ghosh, & Ziskin, 2010) with a 1 km resolution. In earlier studies, nighttime illumination has been shown to be a good indicator of built infrastructure, and thus of economic activity, on a country-wide level (see Elvidge, Hsu, Baugh, & Ghosh, 2014). A more recent study has demonstrated that nighttime lights are also a good predictor of economic wealth at local (within-country) levels (Weidmann & Schutte, 2017). We derive the total amount of nighttime lights for each of our cells and, similar to other studies (Bazzi et al., 2022; Weidmann & Schutte, 2017), calculate a logged per capita value. Generally, countries with persistent conflict are associated with lower per capita gross domestic product (GDP) (Collier, 2004; Pinstrup-Andersen & Shimokawa, 2008). In our specific setting, nighttime lights might point to industry-heavy and wealthier areas, which might again constitute strategically important locations in conflicts and thus improve predictive performance.

Crop production statistics are drawn from MapSPAM (Yu et al., 2020). MapSPAM estimates detailed patterns

[2] This includes all events for which the UCDP variable where_prec is larger than two (i.e. only the second-order administrative division for the location is known) and all events for which date_prec is larger than four (i.e. a day range which is longer than 30 days is reported). All remaining events have a geo-precision of <= 25 km and a reported time period of <= 30 days and thus can be unambiguously assigned to the most likely month/cell combination. Note that there is still some uncertainty left using this assignment, as it is not always guaranteed to be correct.

of crop statistics for 42 different crops in 10 km grid cells across the world. This is achieved through an estimation procedure that combines information on crops (from the lowest available administrative units), land-cover classes, and climate and soil conditions—all three derived from satellite imagery. We employ the most recent (2010) statistics and calculate the total amount of production for all types of crops for each of our cells. According to reports (Eng & Martinez, 2014), rebel areas with agricultural crops were specifically targeted by the Syrian army as a form of punishment by the government. More generally, it is common that conflict parties deliberately destroy infrastructure and resources for food production (Messer & Cohen, 2015) or try to seize cropland in order to secure and guarantee access to food for sustenance (Koren & Bagozzi, 2017).

Daytime temperature recordings are collected from the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature/Emissivity Daily (MOD11A1) version 6.1 (Wan, Hook, & Hulley, 2021). It retrieves daily temperature levels at 1 km resolution across the world from 2000 onwards, using the MODIS thermal infrared channel received by satellite sensors, and is validated by accurate ground-based measurements. In this work, we calculate the average monthly temperature for each cell over the entire study period. Drought (and thus to an extent temperature) has been discussed as a potential contributing factor to the outbreak of the Syrian civil war (Kelley et al., 2015) and has been related to an increase in violence against civilians (Bagozzi, Koren, & Mukherjee, 2017) as well as conflict in general (Von Uexkull, Croicu, Fjelde, & Buhaug, 2016), as it threatens food security. Moreover, we can see drought as a form of local income shock, which can trigger violent mobilization in the case of existing grievances such as ethnic political cleavage (Buhaug, Croicu, Fjelde, & Uexkull, 2021).

Monthly precipitation data come from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). CHIRPS estimates rainfall maps at 5 km resolution going all the way back to 1981 (Funk et al., 2015) using climatology models, satellite imagery, and local station data. We derive the total amount of precipitation for each cell in a given month over the study period. Like temperature, precipitation is closely related to drought.

Last but not least, we use monthly data on vegetation health from the Food and Agriculture Organization (FAO) of the UN (FAO, 2022). More specifically, we draw on the Vegetation Health Index (VHI) developed by Kogan (1997), which is a composite indicator constructed from the Vegetation Condition Index (VCI) and the Temperature Condition Index (TCI), both derived from Advanced Very High Resolution Radiometer (AVHRR) satellite imagery. The VHI has been used in numerous studies to identify droughts (see for example Kogan, Yang, Wei, Zhiyuan, & Xianfeng, 2005 and Rojas, Vrieling, & Rembold, 2011) and is reported for 10-day periods from 1984 onwards. We derive the monthly average VHI for each cell over the entire study period. We provide an overview on the spatial and (employed) temporal resolution of each remote sensing data source in Table A.7 in Appendix A.

3. Methodology & models

3.1. Variables & model specifications

For our forecasting analysis, we define our target variable as a binary measure, indicating whether there was at least one direct fatality as a result of armed force through either state-based or non-state conflict (battles) in a given cell in a given month. With 322 grid cells, from the beginning of our study period in 2011 to the end of 2020 (=120 months), this adds up to a total of 38,640 observations, of which 6698 (∼17.3%) experienced conflict according to our definition.

Our explanatory variables (predictors) are constructed from the data sources listed in Section 2. All time-varying explanatory variables are lagged by one month or year, in order to reflect a real-world setting. For example, we only have temperature data available for the current month, and not for the month ahead for which we want to forecast conflict. We construct our "zero-model" using only temporal and spatial covariates. The zero-model accounts for all effects relating to the spatial and temporal structure of the civil war and thus should already capture some of the main effects. We obtain our baseline specification by extending the zero-model with a set of covariates capturing both civilian and battle-related fatalities of the past 12 months, in addition to ethnic indicators. The former are generally included in most well-performing forecasting routines (see for example Bazzi et al., 2022; Fritz, Mehrl, Thurner, & Kauermann, 2022; Hegre et al., 2019), whereas the latter are specifically included because of the Syrian use case, as explained above. Hence, our baseline model constitutes a specification a researcher would typically employ when not including any type of special or novel dataset and thus should allow for a fair performance comparison against more complex specifications.

In order to test the effectiveness of different remote sensing data sources for our task at hand, we extend our baseline specification by separately adding variables from each source. Doing this for all of the remote sensing datasets reported in Section 2.3, we obtain a total of eight additional specifications. Finally, we construct a full specification, which adds variables from all of the remote sensing datasets jointly to the baseline. An overview over all 11 tested specifications is provided in Table 1. A detailed table with the exact variables used in each specification is reported in Table C.8 in Appendix C. All specifications are run and their performance evaluated with the forecasting procedure described in the subsequent section.

3.2. Forecasting setup

Forecasting is conducted through a monthly one-step-ahead recursive (expanding) window classification, in which for each month t, we forecast conflict in $t+1$. This means our models are trained on all historical data from the first month t_0 up to the current month t. This is a commonly employed evaluation strategy for time series data, as explained e.g. in Petropoulos et al. (2022). We leave multiple-step-ahead forecasts as a potential future work (forecasting conflict for $t+2$, $t+3$, etc., while

Table 1
Model specifications.

#	Specification	Remote sensing data source
1	Zero-Model	–
2	Baseline	–
3	Baseline + Population	Tatem (2017)
4	Baseline + Landcover Classes	Buchhorn et al. (2020)
5	Baseline + Nighttime Lights	Baugh et al. (2010)
6	Baseline + Topography	Amatulli et al. (2018)
7	Baseline + Vegetation Health	FAO (2022)
8	Baseline + Crops	Yu et al. (2020)
9	Baseline + Precipitation	Funk et al. (2015)
10	Baseline + Temperature	Wan et al. (2021)
11	Baseline + All	All of the above

Notes: Definition of the different specifications tested in this study, with a reference to the respective remote sensing data source used.

being in month t), as this would require training separate models for each step ahead and thus substantially increase computational complexity. With a time period from 2011 to 2020, we have $t = 1, 2, \ldots, 120$. In order to ensure we have enough data to train our models, we skip the first year and start our forecasting procedure in $t = 12$. Because most of the employed models require the specification of hyperparameters, hyperparameter tuning is carried out every 12 steps of the forecast, by training the models on data from t_0 up to $t - 1$ and forecasting for t, optimizing the AUROC (the area under the receiver operator characteristic curve; see Section 3.4 for an exact definition). We decided against tuning our models at each step in order to reduce computational complexity. Moreover, the goal of this work is to analyze and highlight the capabilities of remote sensing data for conflict prediction, rather than optimizing performance scores to the very last digit. As a result, we have a total of nine tuning runs ($108/12 = 9$) per model per specification over the entire study period.

Specifically, the procedure at each step $t = 12, 13, \ldots, 119$ is as follows:

1. Assign all observations from t_0 up to t to the training dataset. Assign all observations in $t + 1$ to the test dataset.
2. If $t \; mod \; 12 = 0$: conduct hyperparameter tuning by repeatedly training the model on data from t_0 up to $t-1$ and evaluating the performance for t. Save the best performing hyperparameters for the model.[3]
3. Load the last hyperparameter specification for the model.
4. Train the model with those hyperparameters on the entire training dataset.
5. Predict (forecast) conflict for all observations in the test dataset.

This procedure is repeated for each of the specifications reported in Table 1 and each of the models listed hereafter.

[3] We define *mod* as the modulo operation, which returns the remainder of the division.

3.3. Models, packages, & hyperparameters

In our study, we employ a range of established statistical and machine learning models in and outside the field. For an in-depth discussion of the selection process, see Section 5. In the following, we list each of the chosen models and name the packages employed to conduct the analysis. Moreover, we report the selected hyperparameters. Hyperparameter optimization is carried out for those parameters that are generally understood to be most vital to the performance and set to default values for the remaining ones, in an attempt to reduce computational complexity. This limitation is necessary, as we are running 11 specifications with nine tuning runs each per model, resulting in a total of 99 tuning runs per model.

The following models are employed in our study:

1. Least absolute shrinkage and selection operator (LASSO): We make use of the R package *glmnet* to fit the logistic LASSO regression. We optimize the only hyperparameter λ, specifying the weight of the penalty.
2. Generalized additive model (GAM): We use the R package *mgcv* for fitting, add selection penalties to our model, and make use of both thin plate splines and P-Splines in our models. See Table C.8 in Appendix C for a complete model description. The selection penalties are optimized during training. Additional hyperparameters, such as the number of knots, are not tuned and are left at their default values.
3. Random forests (RF): We employ the highly optimized R implementation *ranger*. Since RFs rarely overfit when increasing the number of fitted trees, we set this number to be sufficiently large, at 500. We also tested larger numbers of trees, but they did not lead to any relevant changes in performance. We optimize the maximum depth of the trees and the minimum node size. The remaining hyperparameters are left at their default values.
4. Gradient boosting (GB): We employ the commonly used extreme gradient boosting (XGBoost) algorithm and its R implementation *xgboost*. We optimize the number of rounds, the learning rate (eta), and the minimum child weight, whereas the remaining hyperparameters are left at their default values.

See Appendix B for more information about the models, including references for more details.

3.4. Performance evaluation

Our main evaluation criteria are the area under the receiver operator characteristic (ROC) curve (AUROC) and the area under the precision–recall curve (AUPRC), which are described subsequently.

The ROC curve plots the true positive rate (TPR), also known as the recall and defined as the ratio of correctly identified positives ($\frac{TP}{P}$), against the false positive rate (FPR), defined as the ratio of false positives to negatives ($\frac{FP}{N}$). The curve describes the tradeoff between the two

when choosing different classification thresholds for a trained model. This means that the AUROC is between 0 and 1, where 0.5 describes a random classifier and 1 describes a perfect classifier. The advantage of the AUROC is its invariance to the classification threshold. That is, it alleviates the difficult decision of choosing a threshold or thresholds for the performance analysis. Hence, it is often used in general classification scenarios and more specifically in the conflict literature (Bazzi et al., 2022; Hegre et al., 2019; Hegre, Nygård, & Landsverk, 2021).

In imbalanced classification settings, the AUROC can sometimes be misleading, as the focus does not lie in the prediction of the minority (positive) class (see Cranmer and Desmarais (2017) for a discussion on this topic in empirical political research). In our study, we face such an issue, as only ∼16% of our observations experience conflict and thus are assigned to the positive class. In such cases, the area under the precision–recall curve (AUPRC), which again does not require thresholding, is typically used in addition. The precision–recall curve describes the tradeoff between precision, defined as the ratio of true positives over all positive predictions ($\frac{TP}{TP+FP}$), and recall when using different classification thresholds. The AUPRC is similarly between 0 and 1, where 1 describes a classifier that is perfectly able to identify the positive class, and 0 the opposite.

4. Results

Our overall results for the entire study period for our four models and 11 different specifications are reported in Table 2 and Table 3, respectively. The former reports the results using the AUROC, the latter using the AUPRC. In both tables, for each model and given each specification, we report the absolute number of the respective evaluation criteria. Additionally, in brackets, we report the relative difference (in percentage) compared to our baseline specification (using the same model), to describe the performance gain (or loss) of the each specification with respect to our baseline. The ROC curves and PR curves are reported in Appendix I. Bootstrapped confidence intervals for the performance of three of our specifications (zero-model, baseline, and full specification) are also available there.

We start the analysis of our results with the AUROC, i.e. Table 2. Our zero-model specification, which only includes spatial and temporal variables, is arguably already performing well, as the AUROC is between 0.77 (LASSO) and 0.923 (RF). This can be explained by two factors. First, a significant part of Syria is covered by desert (95 out of 322 cells are covered by more than 80% "bare" area according to our landcover map; see Appendix I for an illustration) with little to no inhabitants (mean population of 2224 vs. 82,845 in the remaining cells), in which naturally almost no conflict took place over the study period (1.7% vs. 26.7% of observations). This can be largely captured through our spatial variables. Second, civil wars are generally characterized by location-specific battle lines that might slowly change over time (Raleigh & Hegre, 2009), which can be (partly) captured through a combination of both spatial and temporal variables. We can also see that both GB and RF, which can freely model non-linear effects, are much better at capturing the two aforementioned patterns, compared to our GAM, which can only partially model them, and LASSO, which can only model linear effects.

Moving to our baseline specification, which adds both lagged fatality information as well as ethnic indicators, we can see performance increases compared to the zero-model specification for all models. These increases are rather small for our non-linear models (RF and GB) and significantly larger for the other two (GAM and LASSO), as also confirmed by our bootstrapped confidence intervals (see Appendix I). This means that both RF and GB can model the spatial and temporal structure of the civil war so well that adding additional information on past fatalities and ethnicities only leads to minor improvements in overall performance, whereas both GAM and LASSO profit much more from the addition. In our baseline specification, all models perform well (AUROC $\geq$ 0.91) and the differences between the models are not substantial.

Next, we look into the performance gains when including remote sensing data, our main question of interest in this study. We can see (small) performance increases across the board (compared to our baseline) for the individual specifications using landcover classes, population, and crops, whereas the results for the other remote sensing data sources are more mixed and model-dependent. For both nighttime lights and temperature, we can see (minor) performance gains for GAM, RF, and GB, whereas the LASSO performance is on par with the baseline. In other cases, for instance for topography, the inclusion leads to either negligible performance increases or even small decreases, depending on which model performance we look at. For topography, the highest increase is achieved by GB with only +0.26%, and we can spot a decrease for RF (−0.21%). Patterns such as this can occur if the dataset itself does not contribute a lot of relevant information to our task at hand, i.e. forecasting conflict. Then, the inclusion can possibly lead to overfitting and performance will fall off, as our models will pick up random signals (noise) from the respective dataset during training. Although both GB and GAM show minor performance gains for topography, we can arguably still conclude that our chosen topography variables contain little to no relevant information to predict conflict, at least without combining it with additional remote sensing information (from other data sources). Similar arguments can be made for vegetation health and precipitation. Finally, our full specification (Baseline + All), including variables from all remote sensing data sources, performs consistently better than our baseline (0.37%–1.75%) but (for GB and RF) slightly worse than some of the individual specifications, again hinting at the fact that some remote sensing data sources (or the combination of them) are unnecessarily included. Overall, we can conclude that the inclusion of remote sensing data (marginally) increases our predictive performance in terms of the AUROC.

As noted above, the AUROC can be misleading, particularly if a researcher is interested in identifying the minority class, the prevalence of conflict in our setting.

Table 2
AUROC performance (all observations).

Specification	AUROC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.850 (−6.68%)	0.767 (−16.37%)	0.923 (−0.62%)	0.917 (−0.7%)
Baseline	0.910	0.917	0.929	0.923
Baseline + Landcover Classes	0.912 (+0.23%)	0.922 (+0.53%)	0.933 (+0.43%)	0.927 (+0.39%)
Baseline + Population	0.915 (+0.49%)	0.919 (+0.26%)	0.933 (+0.51%)	0.929 (+0.67%)
Baseline + Nighttime Lights	0.919 (+0.99%)	0.917 (+0%)	0.931 (+0.22%)	0.926 (+0.25%)
Baseline + Topography	0.911 (+0.06%)	0.917 (+0%)	0.927 (−0.21%)	0.926 (+0.26%)
Baseline + Vegetation Health	0.912 (+0.19%)	0.917 (−0.02%)	0.929 (+0.04%)	0.926 (+0.35%)
Baseline + Crops	0.916 (+0.63%)	0.919 (+0.23%)	**0.933 (+0.46%)**	**0.930 (+0.74%)**
Baseline + Precipitation	0.916 (+0.58%)	0.917 (+0.02%)	0.928 (−0.05%)	0.924 (+0.12%)
Baseline + Temperature	0.917 (+0.73%)	0.917 (+0%)	0.929 (+0.02%)	0.925 (+0.19%)
Baseline + All	**0.926 (+1.75%)**	**0.923 (+0.73%)**	0.932 (+0.37%)	0.929 (+0.59%)

Notes: Average area under the receicer operator characteristics curve (AUROC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline specification itself. The best performing specification for each model is highlighted in bold.

Table 3
AUPRC performance (all observations).

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.657 (−16.84%)	0.499 (−35.76%)	0.784 (−3.11%)	0.779 (−2.23%)
Baseline	0.790	0.777	0.810	0.796
Baseline + Landcover Classes	0.794 (+0.49%)	0.786 (+1.11%)	0.821 (+1.43%)	0.805 (+1.12%)
Baseline + Population	0.796 (+0.68%)	0.781 (+0.5%)	**0.822 (+1.5%)**	0.809 (+1.64%)
Baseline + Nighttime Lights	0.798 (+1%)	0.777 (+0%)	0.815 (+0.64%)	0.802 (+0.74%)
Baseline + Topography	0.792 (+0.24%)	0.774 (−0.46%)	0.802 (-1%)	0.802 (+0.65%)
Baseline + Vegetation Health	0.794 (+0.51%)	0.773 (−0.52%)	0.815 (+0.72%)	0.807 (+1.35%)
Baseline + Crops	0.797 (+0.9%)	0.781 (+0.47%)	0.822 (+1.49%)	**0.813 (+2.08%)**
Baseline + Precipitation	0.797 (+0.88%)	0.774 (−0.46%)	0.813 (+0.47%)	0.799 (+0.34%)
Baseline + Temperature	0.797 (+0.85%)	0.777 (−0.02%)	0.814 (+0.5%)	0.804 (+0.95%)
Baseline + All	**0.805 (+1.91%)**	**0.789 (+1.51%)**	0.817 (+0.95%)	0.812 (+1.93%)

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline specification itself. The best performing specification for each model is highlighted in bold.

Hence, Table 3 reports the AUPRC of the different specifications for our forecasting task. First of all, we can see that the difference between the zero-model and the baseline is significantly larger than before. This makes sense in that, for our positive-class (conflict) observations, knowing and modeling the location of the Syrian Desert will reduce the number of false positives (FPs) and hence increase the precision, but it will not increase the number of true positives (TPs) and thus the recall. For the latter, knowing past conflict through information on lagged fatalities will most certainly have a positive effect, hence the larger performance difference between the zero-model and the baseline. Notably, this difference is not as large for both RF and GB, as both models can more easily capture the dynamics of the civil war through non-linear combinations of both spatial and temporal variables.

Moving to the remote sensing specifications, we see similar results as above. We report increases for landcover classes, population, crops, and nighttime lights. Notably, the relative increases are around double in percentage points compared to earlier and range up to 2.08% (GB with crops). For vegetation health, precipitation, and temperature we similarly report performance gains, with the exception for LASSO, whereas topography is more mixed, as it again shows decreases for both LASSO and RF. Our full specification performs consistently well (0.95%–1.93%) and is only marginally outperformed by some individual specifications for RF and GB.

By setting a probability threshold, we can analyze the performance increase of our full specification compared to the baseline in individual observation numbers. For reasons of simplicity, we do not tune this threshold and instead set it to the standard value of 0.5. Doing this, out of 34,776 conflict observations in the forecasting test sample, the GAM is able to predict 31,266 correctly (+89 compared to the baseline), and LASSO, RF, and GB correctly predict 31,147 (+20), 31,452 (+53), and 31,330 (+78), respectively. Note that this does not necessarily mean that we are able to correctly forecast the same instances as with our baseline. Moreover, we can very likely achieve better performance by tuning the probability threshold.

Next, we explore where some of these performance increases stem from. By differentiating our observations

Table 4
AUPRC performance for conflict onset observations.

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.241 (−26.79%)	0.161 (−48.49%)	0.357 (−1.49%)	0.323 (−1.9%)
Baseline	0.329	0.312	0.362	0.330
Baseline + Landcover Classes	0.336 (+1.94%)	0.324 (+4%)	0.384 (+6.03%)	0.341 (+3.61%)
Baseline + Population	0.337 (+2.5%)	0.318 (+2.11%)	0.38 (+4.88%)	0.352 (+6.84%)
Baseline + Nighttime Lights	0.331 (+0.39%)	0.312 (+0.02%)	0.37 (+2.07%)	0.336 (+1.82%)
Baseline + Topography	0.335 (+1.59%)	0.308 (−1.13%)	0.365 (+0.7%)	0.346 (+5.1%)
Baseline + Vegetation Health	0.336 (+1.9%)	0.309 (−0.99%)	0.364 (+0.53%)	0.338 (+2.59%)
Baseline + Crops	0.336 (+2.05%)	0.317 (+1.77%)	0.378 (+4.37%)	**0.359 (+8.83%)**
Baseline + Precipitation	0.338 (+2.61%)	0.308 (−1.06%)	0.361 (−0.33%)	0.334 (+1.28%)
Baseline + Temperature	0.334 (+1.55%)	0.312 (+0%)	0.362 (+0.01%)	0.342 (+3.66%)
Baseline + All	**0.347 (+5.42%)**	**0.327 (+4.9%)**	**0.386 (+6.66%)**	0.356 (+7.91%)

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The sample is limited to all conflict onset observations as described in the text. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline itself. The best performing specification for each model is highlighted in bold.

into two categories, we can analyze how well we are able to predict conflict onset vs. conflict persistence. We define our conflict onset observations as observations with no conflict in the current month (t). Hence, analyzing the AUPRC and thus the (minority) conflict class in the next month ($t + 1$) means that we are analyzing how well we are able to predict the outbreak of conflict from one month to the next. This category includes 28,141 observations, of which 1988 experience such an outbreak. We define our conflict persistence observations as those observations that experience conflict in the current month (t). Similarly, by analyzing the AUPRC, we are analyzing how well we are able to predict the persistence of conflict from one month to the next. This category comprises the remaining 6635 observations, of which 4652 experience persistent conflict according to our definition. Based on this distinction, we recalculate the AUPRC from Table 3 for all specifications and report the results in Table 4 (onset) and Table 5 (persistence). By comparing the two tables, we can immediately spot the performance difference in forecasting conflict onset vs. persistence. Predicting "new" conflict is generally more difficult (AUPRC $\sim$ 0.35) than predicting the continuance of it (AUPRC $\sim$ 0.9).

By comparing the performance for conflict onset across our specifications (Table 4), we can see the clear positive impact of including remote sensing data sources. For example, the inclusion of landcover classes leads to performance increases of 1.94% to 6.03%. We can see mixed results for the individual specifications regarding topography, vegetation health and precipitation, and (substantial) performance increases for landcover classes, population, nighttime lights, crops, and temperature. The combination of all remote sensing data sources leads to the highest increases with respect to the baseline for the GAM (5.42%), LASSO (4.9%), and RF (6.66%), and to the second-highest increase for GB (7.91%). Hence, when it comes to the prediction of conflict onset, the inclusion of remote sensing data increases the performance considerably, and a combination of different data sources seems to works well.

Moving to Table 5 and thus the results for conflict persistence, the performance gains when adding remote sensing data to our baseline are much more moderate. Similar data sources seem to perform well and not so well. The full model leads to consistent performance increases, but they are much smaller compared to conflict onset (0.62%–1.61%).

5. Discussion

In this work, we set out to test the effectiveness of various remote sensing datasets for conflict prediction. A number of key findings can be inferred from our results. First and foremost, our results confirm that remote sensing data help to increase overall predictive performance according to both the AUROC (up to 1.75% for the full specification) and AUPRC (1.93%). The overall increases seem rather small at first glance, but this was to be expected due to three reasons. First, our baseline performed well from the outset (AUROC $\geq$ 0.91; AUPRC $\geq$ 0.78), which is a common finding in the literature that is reinforced for our Syrian case study (e.g. see Bazzi et al., 2022; Hegre et al., 2019. Hence, large increases in performance are typically difficult to achieve and much less expected. Second, our marginal performance gains are in line with (and sometimes above) those reported in studies across the literature, all of which extend a baseline specification consisting of lagged fatality information by additional predictors (AUROC increase of $\sim$1.6% in Bazzi et al., 2022; $\sim$1.2% increase in AUROC in Hegre et al., 2019; $\sim$0.8% decrease in MSE (=performance increase) in Mueller & Rauh, 2022). Third, our baseline is actually a richer model than those employed as a baseline in the cited studies, as we additionally include spatial, temporal, and ethnic variables. Therefore, our results being consistent with current literature, despite employing this richer baseline, demonstrates the additional value of remote sensing data impressively. Fourth, even when removing contextual information such as lagged fatalities and ethnicities from the model (=zero-model), remote sensing

Table 5
AUPRC performance for conflict persistence observations.

Specification	AUPRC for Model			
	GAM	LASSO	RF	GB
Zero-Model	0.869 (−4.39%)	0.825 (−7.16%)	0.894 (−1.97%)	0.886 (−1.89%)
Baseline	0.909	0.889	0.912	0.903
Baseline + Landcover Classes	0.911 (+0.3%)	0.895 (+0.67%)	0.923 (+1.19%)	0.912 (+0.96%)
Baseline + Population	0.912 (+0.34%)	0.891 (+0.24%)	0.924 (+1.31%)	0.914 (+1.25%)
Baseline + Nighttime Lights	0.912 (+0.41%)	0.889 (+0%)	0.917 (+0.56%)	0.909 (+0.69%)
Baseline + Topography	0.910 (+0.09%)	0.886 (−0.29%)	0.904 (−0.84%)	0.906 (+0.34%)
Baseline + Vegetation Health	0.911 (+0.24%)	0.886 (−0.33%)	0.919 (+0.82%)	0.915 (+1.31%)
Baseline + Crops	0.913 (+0.42%)	0.891 (+0.22%)	**0.924 (+1.32%)**	0.918 (+1.58%)
Baseline + Precipitation	0.912 (+0.38%)	0.886 (−0.28%)	0.917 (+0.59%)	0.907 (+0.41%)
Baseline + Temperature	0.912 (+0.33%)	0.889 (−0.02%)	0.917 (+0.6%)	0.911 (+0.9%)
Baseline + All	**0.915 (+0.64%)**	**0.898 (+1.05%)**	0.918 (+0.62%)	**0.918 (+1.61%)**

Notes: Average area under the precision–recall curve (AUPRC) performance for one-step ahead forecasts over the entire forecasting horizon of the different model specifications and types. The sample is limited to all conflict persistence observations as described in the text. For an explanation of the AUPRC, see Section 3.4. Each row reports the performance of one specification (details in Section 3.2), each column of one type of model (details in Section 3.3). In brackets we report the relative performance difference to our baseline specification (of the same model), except for the baseline itself. The best performing specification for each model is highlighted in bold.

data provide clear performance gains when they are included (see Appendix D). Overall, we can conclude that remote sensing data indeed provide an additional source of information relevant for the prediction of conflict.

Although we did not initially set out to differentiate between the onset and persistence of conflict, during our performance analysis we identified that remote sensing data are particularly important for correctly predicting the onset of conflict. Generally, predicting conflict onset is considered a much more difficult task in the conflict literature, as similar studies (Hegre et al., 2021; Mueller & Rauh, 2022), as well as our results (lower AUPRC), show. Providing additional information through remote sensing data sources turns out to be particularly important for this challenging task. Notably, our full specification, which includes all remote sensing data sources, performs the best with one exception (GB; but in second place here), which supports our argument. By inspecting the onset results more closely (see Appendix F), we can identify that the prediction task (expectedly) becomes more and more difficult as the number of months since last conflict increases (lower AUPRC). At the same time, the relative performance increase of the full specification becomes larger and larger. Hence, according to our results, the importance of remote sensing data increases, as the time that has passed since last conflict becomes longer.

This leads to our third observation. Depending on the model and setting (onset vs. persistence), particular specifications perform better or worse, respectively. There are a few implications to draw from this finding. First, these results point to the fact that training and using individual models for onset and persistence, respectively (possibly with different data sources or variables included), might improve forecasting performance. To the best of our knowledge, this has not been considered in the literature yet. Second, including all remote sensing data sources into the models might not necessarily be the best choice. For our non-linear models (RF and GB), individual specifications (with only one remote sensing data source included) at times outperform the full specification (with all of them included). Notably, these patterns are not consistent across models. This makes sense in that regard, insofar as different models vary in their ability to extract relevant information for the conflict prediction task. Both RF and GB are in theory able to extract (highly) non-linear relationships between the target variable (conflict) and our explanatory variables, including possible interactions, whereas both the GAM and LASSO are only partially able to or are unable to, respectively. Hence, for the former two models, the same variables offer more possibly relevant information (e.g. reflected in the higher performance of the baseline), which makes it more likely that some of the remote sensing data sources are rendered redundant, and thus that their inclusion leads to overfitting and a performance decrease. As a consequence, researchers not only need to be careful what variables or data sources they consider, but also need to take into account the model they intend to use when making any decisions on variable inclusions. Future works could consider automating this data-driven process to achieve the best possible performances. For example, one could pursue a forward selection process, in which data sources and/or specific variables are continually added to the model in a systematic fashion (starting from the baseline) as long as the performances are increasing. A researcher could even go so far as to perform this forward selection and thus adjust the included variables for each step of the study period (i.e. every time a new model is trained) in order to achieve optimal results across the entire forecasting horizon. Note that such a selection approach would need to be treated in a similar fashion as hyperparameter tuning; i.e. we need to guarantee a true out-of-sample performance evaluation.

Next, we investigated the performance of our models over time (see Appendix E for the results and a more thorough discussion). We conclude that both the GAM and to some extent LASSO profit considerably from the inclusion of remote sensing data early into the study period, where less training data are available. Closer to the end of the study period, the AUPRC performance starts to drop for all models and specifications, as conflict events are thinning out and the task of correctly identifying conflict becomes

considerably more difficult. Nonetheless, the relative performance gains from including remote sensing data are largely steady throughout, strengthening the confidence of our findings.

Moreover, we want to contribute to the ongoing discussion in the literature on the tradeoff between explanation and prediction (Hegre et al., 2017) by highlighting that our model performances do not substantially differ. Although there is a clear ranking in terms of model performance (RF > GB > GAM > LASSO), once we arrive at our baseline specification, the performance difference in both AUROC (up to 1.3%) and AUPRC (4.2%) between the models for any specification is much lower than one might expect. Notably, for our full specification, these differences further decrease. Hence, a researcher could easily fall back on using one of the inherently interpretable models, such as LASSO or the GAM, without giving up or forgoing substantial performance gains. Note that in-depth analyses in this setting remain a difficult endeavor, even with interpretable models, as we are re-training our models for each step of the study period (e.g. here, 108 different GAMs for one specification only). Moreover, identifying causal chains of effects additionally requires distinct variable setups and theoretical considerations. Nonetheless, the small performance gap reported here seems promising.

This brings us to our sixth point of discussion: our selection of models. Drawing from the latest ViEWS conflict prediction competition (Vesco et al., 2022), and similar studies such as (Bazzi et al., 2022; Hegre et al., 2019), we decided on a versatile set of commonly employed models, ranging from classical statistical to common machine learning models, in order to ensure that our results are consistent across a variety of different models. All chosen models (GAM, LASSO, RF, and GB) have proven themselves to perform well in the conflict forecasting domain (see the cited studies) as well as across a variety of other fields (Bastin et al., 2019; Chaudhary, Richardson, Schoeman, & Costello, 2021; Fabbri et al., 2020; Fife & D'Onofrio, 2022; Greener, Kandathil, Moffat, & Jones, 2022; Rustam et al., 2020; Schroeders, Schmidt, & Gnambs, 2022; Xie & Zhu, 2020). Our set of chosen models consists of two "simpler" models with the GAM and LASSO, which have a limited capacity to model non-linear and interaction effects but remain easier to analyze and interpret. On the other hand, with RF and GB, we chose two models that can freely model non-linear and interaction effects, but are consequently much more difficult to analyze and interpret (and hence oftentimes considered black-box models), with GB additionally being more prone to overfitting. We refrained from including neural networks, as they require large amounts of data to train and are generally outperformed by classical machine learning models on tabular datasets such as the one here (Borisov et al., 2021). Notably, redefining our observations in the form of images and implicitly taking into account the spatial structure of each cell and their surroundings—and thus taking advantage of the high resolution the remote sensing datasets offer, and moving away from the tabular structure—through convolutional neural networks (CNNs) might be a viable future path. Moreover, we decided against including an ensemble of different individual models, as for example pursued in Bazzi et al. (2022) and in the ViEWS forecasting competition (Vesco et al., 2022), since such an ensemble is simply a weighted combination of individual models. Hence, performance increases in the individual models (as reported here across the board) are very likely similarly reflected in the performance of the ensemble.

Last but not least, we want to briefly reflect on the performance and thus importance of different remote sensing datasets. According to our individual results, landcover classes, population, and crop data provide the highest and most consistent overall performance increases, whereas the remaining datasets seem less important. Nonetheless, once we look at the results for conflict onset, most, if not all, of the remote sensing datasets seem to provide relevant information for the prediction of conflict. This is confirmed by the fact that our full specification performed the best in three out of four models. Moreover, by looking at the variables selected by LASSO, we can see that in our full specification, each remote sensing dataset is selected in more than half of the model fits (see Table G.14 in Appendix G), with landcover classes, population, and crop data being selected most often. Finally, by inspecting the feature importance scores for RF (full specification, see Fig. H.3 in Appendix H), we can corroborate this pattern. While our lagged battle-related features seem to be most important for the prediction, both landcover classes and population are not too far behind. On average, each remote sensing dataset contributes to the model performance according to these importance scores. Hence, we conclude that each remote sensing data source seems to provide relevant information for the prediction of conflict, but some of them are more important than others.

6. Conclusion

We tested the effectiveness and capabilities of remote sensing data for conflict prediction in the context of the Syrian civil war. Using remote sensing data enabled us to conduct our study in self-defined, fine-grained, and evenly sized spatial cells across Syria. Our results confirmed that including a variety of remote sensing datasets consistently improved forecasting performance compared to a rich baseline independent of the chosen prediction model. As our analysis showed, a large portion of this performance gain came from correctly identifying the onset of conflict. We conclude that remote sensing data can and indeed should be used to forecast conflict in countries with a lack of reliable official data sources.

Future work could try to take advantage of the fine-grained spatial structure of the remote sensing datasets through specific modeling techniques. Moreover, as more and more high-quality datasets are published, identifying causal effects for conflict on a subnational level might prove to be possible. Finally, evaluating other emerging data sources such as news or social media, and finding ways to combine disparate and emerging datasets into a joint model, may further improve forecasting performance. This provides an interesting set of data sources, methods, and approaches to further advance conflict research globally.

CRediT authorship contribution statement

D.R., P.W.T., X.X.Z. and G.K. conceived the research. P.W.T., B.I.D., X.X.Z. and G.K. supervised the research. D.R., P.W.T. and G.K. designed the methodology. D.R. collected & processed the data, conducted the study and analyzed the results. D.R. and B.I.D. created the visualizations. D.R., P.W.T. and G.K. wrote the first draft. D.R., P.W.T., B.I.D. and G.K. edited and all authors approved the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Data sources

See Tables A.6 and A.7.

Appendix B. Model descriptions

Here, we provide concise descriptions of the models employed in this study. For details we refer the reader to the respective publications.

1. Least absolute shrinkage and selection operator (LASSO): Using LASSO (Tibshirani, 1996), we fit a simple (logistic) regression model in conjunction with an L_1-penalty on the coefficients. This shrinks the model coefficients compared to a standard regression. Additionally, some of the coefficients (those for "less important" variables) are set to exactly 0, which resembles a feature selection. Generally, the penalty reduces the generalization error of the fitted model. Using LASSO, we can only capture linear effects of the included covariates. Hence, any non-linear effects or interactions of covariates need to be explicitly included in the model through transformations, as done in standard regression models.
2. Generalized additive model (GAM): A GAM (Hastie, 2017) is a generalized linear model (in our case a logistic regression) with an additional set of linearly included unknown smooth functions of (some of the) explanatory variables. The set of smooth functions is chosen by the user, in terms of both the included variables and the types of functions. Commonly chosen smooth functions are, for example, thin plate regression splines (Wood, 2003), cubic regression splines (Durrleman & Simon, 1989), and P-splines (Eilers & Marx, 1996). By including these smooth terms into the model, the GAM is able to model non-linear effects of the chosen variables. Similar to LASSO, interactions need to be explicitly included in the model.
3. Random forests (RF): RF (Breiman, 2001) is an ensemble consisting of multiple (decision) trees. The overall prediction is the average of the individual predictions over all trees. RFs typically employ bagging and a random selection of the features in order

Table A.6
Data sources with web addresses for download.

Data	Source	Downloaded from:
Conflict	Sundberg and Melander (2013)	https://ucdp.uu.se/downloads/
Ethnicity	Vogt et al. (2015)	https://icr.ethz.ch/data/epr/geoepr/
Population	Tatem (2017)	https://hub.worldpop.org/project/categories?id=3
Landcover Classes	Buchhorn et al. (2020)	https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_Landcover_100m_Proba-V-C3_Global
Nighttime Lights	Baugh et al. (2010)	https://developers.google.com/earth-engine/datasets/catalog/NOAA_DMSP-OLS_NIGHTTIME_LIGHTS
Topography	Amatulli et al. (2018)	http://www.earthenv.org/topography
Vegetation Health	FAO (2022)	https://data.apps.fao.org/map/catalog/srv/eng/catalog.search#/metadata/84e27651-0bb4-4a26-8b4a-2b10bbccb7e0
Crops	Yu et al. (2020)	https://www.mapspam.info/data/
Precipitation	Funk et al. (2015)	https://developers.google.com/earth-engine/datasets/catalog/UCSB-CHG_CHIRPS_DAILY
Temperature	Wan et al. (2021)	https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MOD11A1

Table A.7
Remote sensing data sources temporal & spatial resolution.

Data	Available temporal resolution	Temporal resolution & time period employed	Spatial resolution
Population	Yearly	Fixed (2010)	100 m
Landcover Classes	Yearly	Yearly from 2015–2019, fixed before & after	100 m
Nighttime Lights	Yearly	Yearly from 2010–2013	1 km
Topography	None	Fixed	1 km
Vegetation Health	10 days	Monthly from 2010–2020	1 km
Crops	10 years	Fixed (2010)	10 km
Precipitation	Daily	Monthly from 2010–2020	5 km
Temperature	Daily	Monthly from 2010–2020	1 km

to reduce the correlation between the trees and thus the variance of the ensemble. RF is able to freely model non-linear effects and the interactions between the variables without the requirement of including any of them explicitly.

4. Gradient boosting (GB): In boosting, we construct an ensemble of weak prediction models, typically (decision) trees, and iteratively apply the model to modified versions of the training data. After each iteration, misclassified inputs are assigned higher weights such that they are focused on in the next training iteration. GB (Friedman, 2001) mimics this process by viewing the boosting algorithm as iterative functional gradient descent and "nudging" the model prediction function step by step, closer to the real data points. An efficient and scalable implementation of this technique is XGBoost (Chen & Guestrin, 2016), which is employed in this study. GB can freely model non-linear effects and the interactions between the variables without including any of them explicitly.

Appendix C. Specifications

See Table C.8.

Appendix D. Additional model specification results

The following two tables expand on our main results by reporting the performance of an additional specification that extends the zero-model with all remote sensing data sources (Zero-Model + All). Hence, the results of Table D.9 correspond to Table 2 and Table D.10 corresponds to Table 3 in the main text. Notably, we only compare the performance between these two specifications. The results show that even without including information on lagged fatalities and ethnicities, remote sensing data increase predictive performance in both AUROC and AUPRC. The performance increases are particularly large for the "simpler" models (GAM and LASSO) and smaller for the other two (RF and GB). Arguably, both RF and GB already learn most of the relevant remote sensing information implicitly through a (non-linear) combination of spatial and temporal variables (e.g. more populous areas experience more conflict during certain time periods). Nonetheless, even without additional contextual information (lagged fatalities, ethnicities), we see consistent performance increases when adding remote sensing data to the respective model across all four models.

Appendix E. Performance over time

Here, we provide additional insights into the performance of our different models and specifications over the study period. Table E.11 splits up the AUROC performance reported in Table 2 in the main text into separate scores for each year for both the baseline and the full specification. Table E.12 does the same for the AUPRC performance reported in Table 3.

Table E.11 shows that both the GAM and LASSO (and to a smaller extent, RF) considerably benefit from the inclusion of remote sensing data in the first year of the analysis period, as indicated by the large performance gains in AUROC. Apparently, the "simpler" models particularly struggle with the lack of historical information when only using baseline features, but most of this performance gap can be made up for by the inclusion of remote sensing data. In the remaining years, the performance increases from baseline to full specification are mostly steady (with some smaller fluctuations). Notably, we see a small drop in performance from 2016 onwards. We attribute this drop to the changing circumstances (Russian support in form of airstrikes, pushing back of the Islamic State by both the Kurds and the Assad regime, and Turkish offensives in northern Syria) that are not immediately picked up by the models. Nonetheless, we want to highlight that our performance gains from remote sensing data are (mostly) consistent even throughout this period.

Similarly, Table E.12 shows considerable AUPRC performance gains for the GAM with the inclusion of remote sensing in the first year. Again, we can identify a drop in performance in 2016, from which the performance continues to fall off. While we can attribute some of this performance drop to the changing circumstances described above, after 2018 especially, the task of correctly identifying conflict becomes considerably more difficult, as the civil war moves more towards the border regions next to Turkey and Iraq, and fewer cells experience conflict. In 2018, only 13% of the observations suffer from conflict (see the second column in the table), compared to 25.2% in 2015. In 2020 we are down to only 9.3%. The increased difficulty in the prediction task that results from this is distinctly reflected across all models and specifications, with a considerably lower AUPRC of around 0.55. Nonetheless, on average, the performance increase from the use of remote sensing continues to hold even throughout these years.

Appendix F. Detailed onset performance

The following table offers a more detailed view on the performance results for conflict onset, by differentiating observations by the number (#) of months since the last conflict took place in the respective cell. Overall, we can see that the prediction task becomes more and more difficult as the number of months increase (i.e. AUPRC decreases as we move down the rows), across all models. At the same time, the relative performance increase of the full specification increases substantially. Even when leaving out the first row (conflict persistence), the performance increase rises by 7 (GB) to 16 (RF) percentage points when moving down to the last row (months since last conflict $\geq$ 6). We can observe this pattern consistently across all models. Notably, for GB we can spot a performance decrease with respect to the baseline twice (for 3 and 5 months). We attribute this effect to potential overfitting, since the number of observations is quite low (1259 and 773 vs. 23,202 for $\geq$ 6 months), GB is particularly prone to overfitting and we cannot observe the same pattern for any of the other models. Hence, overall, we can conclude that remote sensing data become more and more important for the prediction of conflict as the time since last conflict increases (see Table F.13).

Table C.8
Specifications with covariates.

#	Specification	Included covariates
1	Zero-Model	- Time Trend: integer GAM: Included using a P-Spline - Monthly dummies (11): binary - Cell size in km^2: numeric - Latitude: numeric GAM: Inclusion see Longitude - Longitude: numeric GAM: Included with Latitude using a thin plate spline - Distance to capital in km: numeric
2	Baseline	- All from specification 1) - Alawi dummy (share $>$ 0.05): binary - Christians dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Druze dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Kurds dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - Sunni Arabs dummy (share > 0.05): binary LASSO: Included as interaction with Time Trend GAM: Included as interaction with Time Trend using a P-Spline - # of fatalities through battle last month: integer - # of fatalities through battle last 12 months: integer - # of month since last fatality through battle: integer - # of civilian fatalities last month: integer - # of civilian fatalities last 12 months: integer - # of month since last civilian fatality: integer
3	Baseline + Population	- All from specification 2) - Total amount of population (logged): numeric
4	Baseline + Landcover Classes	- All from specification 2 - Share crop area: numeric - Share bare area: numeric - Share built area: numeric - Share grass & shrub area: numeric - Permanent water dummy (share $>$ 0.01): binary - Tree-covered area dummy (share $>$ 0.01): binary
5	Baseline + Nighttime Lights	- All from specification 2) - Total amount of stable lights per person (logged): numeric
6	Baseline + Topography	- All from specification 2) - Elevation (median): numeric - Slope (median): numeric - Vector ruggedness measure (median): numeric
7	Baseline + Vegetation Health	- All from specification 2) - Vegetation health index (average): numeric
8	Baseline + Crops	- All from specification 2) - Total amount of food crops (logged): numeric - Total amount of non-food crops (logged): numeric
9	Baseline + Precipitation	- All from specification 2) - Total amount of precipitation: numeric
10	Baseline + Temperature	- All from specification 2) - Day temperature (average): numeric
11	Baseline + All	- All variables

Notes: If not noted otherwise, all variables are included linearly without any interaction in both LASSO and the GAM.

Appendix G. Lasso model selection

See Table G.14.

Appendix H. Feature importance

See Fig. H.3.

Appendix I. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.04.001. There, we 1) provide further information on the Syrian desert cells, 2) provide the ROC curves & PR curves, 3) report the bootstrapped performance and 4) report performance results without the cell filtering.

Table D.9
AUROC performance (all observations), expansion.

Specification	AUROC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.850	0.767	0.923	0.917
Zero-Model + All	0.892 (+5.04%)	0.876 (+14.21%)	0.926 (+0.34%)	0.925 (+0.92%)

Notes: Expansion of the performance results of Table 2 in the main results section. Here, we compare the results of our zero-model, with a specification that adds all remote sensing data sources to the zero-model. Cell filtering takes place. Relative performance differences are calculated with respect to the zero-model.

Table D.10
AUPRC performance (all observations), expansion.

Specification	AUPRC for model			
	GAM	LASSO	RF	GB
Zero-Model	0.657	0.499	0.784	0.779
Zero-Model + All	0.725 (+10.38%)	0.665 (+33.18%)	0.786 (+0.19%)	0.793 (+1.91%)

Notes: Expansion of the performance results of Table 3 in the main results section. Here, we compare the results of our zero-model, with a specification that adds all remote sensing data sources to the zero-model. Cell filtering takes place. Relative performance differences are calculated with respect to the zero-model.

Table E.11
AUROC Performance Yearly (All Observations).

Year	Share Pos.	GAM		LASSO		RF		GB	
		Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
2012	0.182	0.818	0.923 (+12.83%)	0.924	0.952 (+2.95%)	0.931	0.948 (+1.83%)	0.942	0.951 (+0.93%)
2013	0.250	0.945	0.947 (+0.19%)	0.947	0.951 (+0.39%)	0.950	0.953 (+0.35%)	0.948	0.95 (+0.26%)
2014	0.258	0.929	0.933 (+0.42%)	0.920	0.920 (+0%)	0.933	0.938 (+0.47%)	0.929	0.93 (+0.05%)
2015	0.252	0.942	0.943 (+0.11%)	0.941	0.944 (+0.38%)	0.945	0.948 (+0.31%)	0.940	0.944 (+0.38%)
2016	0.236	0.914	0.912 (−0.19%)	0.912	0.914 (+0.19%)	0.914	0.916 (+0.25%)	0.897	0.913 (+1.81%)
2017	0.209	0.899	0.901 (+0.24%)	0.889	0.901 (+1.31%)	0.904	0.905 (+0.11%)	0.898	0.903 (+0.59%)
2018	0.130	0.912	0.911 (−0.12%)	0.910	0.915 (+0.54%)	0.911	0.917 (+0.66%)	0.900	0.912 (+1.38%)
2019	0.108	0.915	0.918 (+0.33%)	0.903	0.891 (−1.31%)	0.916	0.922 (+0.63%)	0.910	0.905 (−0.53%)
2020	0.093	0.895	0.899 (+0.38%)	0.891	0.898 (+0.77%)	0.902	0.902 (+0.01%)	0.900	0.895 (−0.56%)

Notes: Detailed AUROC performance results for all observations differentiated by the forecasting year for the two main specifications (Baseline, Baseline + All). The year 2011 is not included, as it is only used for training. The second column (Share Pos.) reports the share of observations experiencing conflict in the respective year. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

Table E.12
AUPRC performance yearly (all observations).

Year	Share Pos.	GAM		LASSO		RF		GB	
		Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
2012	0.182	0.703	0.82 (+16.59%)	0.850	0.872 (+2.64%)	0.859	0.872 (+1.55%)	0.861	0.868 (+0.89%)
2013	0.250	0.885	0.888 (+0.34%)	0.889	0.891 (+0.26%)	0.887	0.900 (+1.53%)	0.889	0.894 (+0.59%)
2014	0.258	0.839	0.848 (+1.01%)	0.826	0.826 (+0%)	0.854	0.860 (+0.71%)	0.841	0.843 (+0.18%)
2015	0.252	0.864	0.861 (−0.24%)	0.862	0.863 (+0.11%)	0.870	0.876 (+0.75%)	0.858	0.867 (+1.1%)
2016	0.236	0.803	0.799 (−0.47%)	0.808	0.811 (+0.45%)	0.807	0.810 (+0.34%)	0.767	0.809 (+5.48%)
2017	0.209	0.742	0.744 (+0.26%)	0.730	0.747 (+2.35%)	0.747	0.751 (+0.56%)	0.728	0.746 (+2.38%)
2018	0.130	0.703	0.696 (−0.93%)	0.693	0.711 (+2.63%)	0.721	0.730 (+1.35%)	0.696	0.713 (+2.53%)
2019	0.108	0.695	0.700 (+0.68%)	0.692	0.673 (−2.72%)	0.707	0.713 (+0.91%)	0.691	0.694 (+0.46%)
2020	0.093	0.525	0.534 (+1.59%)	0.520	0.531 (+2.28%)	0.568	0.570 (+0.39%)	0.557	0.559 (+0.36%)

Notes: Detailed AUPRC performance results for all observations differentiated by the forecasting year for the two main specifications (Baseline, Baseline + All). The year 2011 is not included, as it is only used for training. The second column (Share Pos.) reports the share of observations experiencing conflict in the respective year. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

Table F.13
AUPRC performance since last conflict (all observations).

# Months last conflict	GAM		LASSO		RF		GB	
	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All	Baseline	Baseline + All
1	0.909	0.915 (+0.64%)	0.889	0.898 (+1.05%)	0.912	0.918 (+0.62%)	0.903	0.918 (+1.61%)
2	0.513	0.532 (+3.73%)	0.486	0.503 (+3.49%)	0.535	0.576 (+7.55%)	0.502	0.550 (+9.51%)
3	0.378	0.398 (+5.38%)	0.360	0.380 (+5.38%)	0.378	0.397 (+5.18%)	0.377	0.354 (−5.94%)
4	0.276	0.277 (+0.43%)	0.268	0.276 (+3%)	0.297	0.333 (+12.16%)	0.242	0.261 (+7.86%)
5	0.208	0.212 (+2.28%)	0.181	0.212 (+17.47%)	0.210	0.247 (+17.37%)	0.203	0.183 (−10.05%)
≥6	0.094	0.105 (+12.07%)	0.076	0.091 (+19.27%)	0.088	0.103 (+16.16%)	0.084	0.098 (+16.64%)

Notes: Detailed AUPRC performance results for all observations differentiated by the number (#) of months since last conflict for the two main specifications (Baseline, Baseline + All). The results in the first row (1 month since last conflict) are equivalent to the results reported in Table 5 in the main text, as they represent conflict persistence. The five remaining rows offer a more detailed view on the performance for conflict onset. As in the main text, the relative performance differences (%) are calculated with respect to the baseline specification of the same model type.

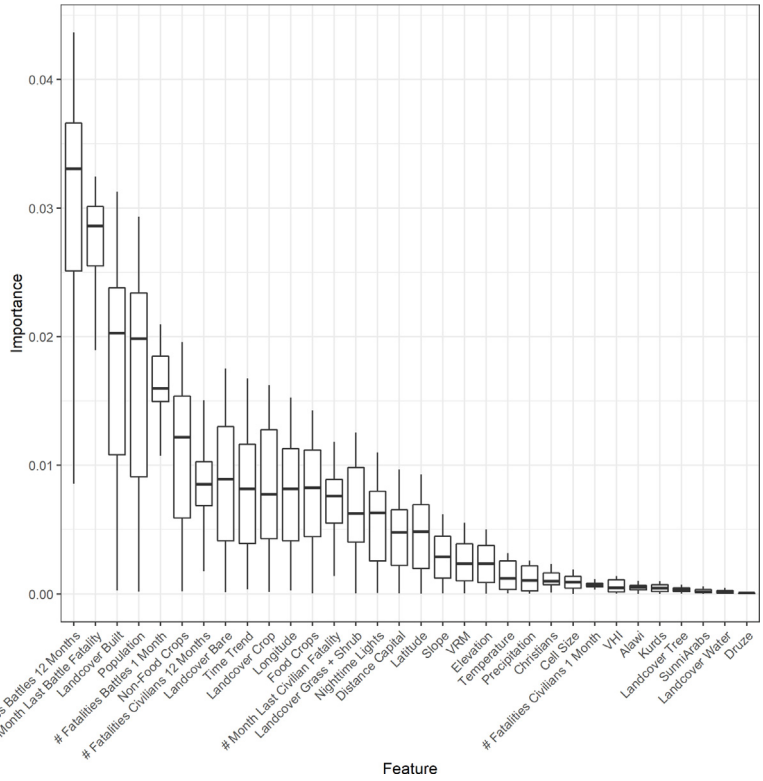


Fig. H.3. Boxplot of the permutation feature importance scores for the full specification (Baseline + All) using RF. We derived the importance scores for each of the 108 models (trained over the entire forecasting horizon) by calculating the average increase in classification error on the out-of-bag data sample for all trees. The features are ordered with respect to their median importance. We excluded the monthly dummies from our feature set, as they require a more elaborate permutation strategy. The boxplots for all other specifications are available on the OSF.

Table G.14
Remote sensing dataset inclusions by LASSO.

Remote sensing dataset	% Included
Landcover classes	77.8%
Population	76.9%
Nighttime Lights	55.6%
Topography	60.2%
Vegetation Health	66.7%
Crops	73.1%
Precipitation	73.1%
Temperature	50.9%

Notes: % of times a remote sensing dataset was included by LASSO in the full specification over the entire forecasting horizon (108 model fits). Each time one variable of a remote sensing dataset is included, we count this as an inclusion of the respective dataset.

References

Abosedra, S., Fakih, A., & Haimoun, N. (2021). *Ethnic divisions and the onset of civil wars in Syria*: *Technical Report*, GLO Discussion Paper.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., et al. (2018). A suite of global cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*, 1–15. http://dx.doi.org/10.1038/sdata.2018.40.

Attinà, F., Carammia, M., & Iacus, S. M. (2022). Forecasting change in conflict fatalities with dynamic elastic net. arXiv preprint arXiv: 2205.14073.

Avtar, R., Kouser, A., Kumar, A., Singh, D., Misra, P., Gupta, A., et al. (2021). Remote sensing for international peace and security: Its role and implications. *Remote Sensing*, *13*(439).

Bagozzi, B. E., Koren, O., & Mukherjee, B. (2017). Droughts land appropriation, and rebel violence in the developing world. *The Journal of Politics*, *79*, 1057–1072.

Ball, J. E., Anderson, D. T., & Chan Sr, C. S. (2017). Comprehensive survey of deep learning in remote sensing: Theories tools, and challenges for the community. *Journal of Applied Remote Sensing*, *11*, Article 042609.

Bastin, J. F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., et al. (2019). The global tree restoration potential. *Science*, *365*, 76–79.

Baugh, K., Elvidge, C. D., Ghosh, T., & Ziskin, D. (2010). Development of a 2009 stable lights product using DMSP-OLS data. *Proceedings of the Asia-Pacific Advanced Network*, *30*(114).

Bazzi, S., Blair, R. A., Blattman, C., Dube, O., Gudgeon, M., & Peck, R. (2022). The promise and pitfalls of conflict prediction: Evidence from Colombia and Indonesia. *The Review of Economics and Statistics*, *104*, 764–779.

Berger, M., Moreno, J., Johannessen, J. A., Levelt, P. F., & Hanssen, R. F. (2012). ESA's sentinel missions in support of Earth system science. *Remote Sensing of Environment*, *120*, 84–90.

Blattman, C., & Miguel, E. (2010). Civil war. *Journal of Economic Literature*, *48*, 3–57.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Buchhorn, M., Smets, B., Bertels, L., Roo, B. De., Lesiv, M., Tsendbazar, N. E., et al. (2020). Copernicus global land service: land cover 100 m: version 3 globe 2015-2019. http://dx.doi.org/10.5281/Zenodo.3939050.

Buhaug, H., Croicu, M., Fjelde, H., & Uexkull, N. von. (2021). A conditional model of local income shock and civil conflict. *The Journal of Politics*, *83*, 354–366.

Chaudhary, C., Richardson, A. J., Schoeman, D. S., & Costello, M. J. (2021). Global warming is causing a more pronounced dip in marine species richness around the equator. *Proceedings of the National Academy of Sciences*, *118*, Article e2015094118.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining* (pp. 785–794).

Collier, P. (2004). *Development and conflict* (pp. 1–12). Centre for the Study of African Economies.

Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford Economic Papers*, *56*, 563–595.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*, 145–166.

Danielson, J. J., & Gesch, D. B. (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010)*. US Department of the Interior, US Geological Survey Washington, DC, USA.

Durrleman, S., & Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, *8*, 551–561.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*, 89–121.

Elvidge, C. D., Hsu, F. C., Baugh, K. E., & Ghosh, T. (2014). National trends in satellite-observed lighting. In *Global Urban Monitoring and Assessment Through Earth Observation, vol. 23* (pp. 97–118).

Eng, B., & Martinez, J. (2014). Starvation, submission and survival the Syrian War through the prism of food. Middle East Report 44.

Fabbri, C., Kasper, S., Kautzky, A., Zohar, J., Souery, D., Montgomery, S., et al. (2020). A polygenic predictor of treatment-resistant depression using whole exome sequencing and genome-wide genotyping. *Translational Psychiatry*, *10*, 1–12.

FAO (2022). Agricultural stress index system (ASIS). dataset identifier: 84e27651-0bb4-4a26-8b4a-2b10bbccb7e0 http://www.fao.org/giews/earthobservation/. (Accessed 15 March 2022).

Fearon, J. D., & Laitin, D. D. (2003). Ethnicity insurgency, and civil war. *American Political Science Review*, *97*, 75–90.

Fife, D. A., & D'Onofrio, J. (2022). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Fritz, C., Mehrl, M., Thurner, P. W., & Kauermann, G. (2022). The role of governmental weapons procurements in forecasting monthly fatalities in intrastate conflicts: a semiparametric hierarchical hurdle model. *International Interactions*, 1–22.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes version 2.0. *Scientific Data*, *2*, 1–21. http://dx.doi.org/10.1038/sdata.2015.66.

Girardin, L., Hunziker, P., Cederman, L. E., Bormann, N. C., & Vogt, M. (2015). Growup-geographical research on war, unified platform. ETH Zurich. http://growup.ethz.ch.

Gleditsch, K. S., & Ward, M. D. (2013). Forecasting is difficult especially about the future: using contentious issues to forecast interstate disputes. *Journal of Peace Research*, *50*, 17–31.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., et al. (2010). A global model for forecasting political instability. *American Journal of Political Science*, *54*, 190–208.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*, 40–55.

Harff, B., & Gurr, T. R. (1998). Systematic early warning of humanitarian emergencies. *Journal of Peace Research*, *35*, 551–579.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S. Routledge* (pp. 249–307).

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., et al. (2019). Views: A political violence early-warning system. *Journal of Peace Research*, *56*, 155–174.

Hegre, H., Metternich, N. W., Nygård, H. M., & Wucherpfennig, J. (2017). Introduction: forecasting in peace research.

Hegre, H., Nygård, H. M., & Landsverk, P. (2021). Can we predict armed conflict? How the first 9 years of published forecasts stand up to reality. *International Studies Quarterly*, *65*, 660–668.

Human Rights Watch (2022). Syria: Events of 2021. https://www.hrw.org/world-report/2022/country-chapters/syria. (Accessed 11 April 2022).

Ismail, S. (2011). The Syrian uprising: Imagining and performing the nation. *Studies in Ethnicity and Nationalism*, *11*, 538–549.

Jerven, M. (2013). *Poor numbers: How we are misled by african development statistics and what to do about it*. Cornell University Press.

Kelley, C. P., Mohtadi, S., Cane, M. A., Seager, R., & Kushnir, Y. (2015). Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proceedings of the National Academy of Sciences*, *112*, 3241–3246.
King, G., & Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, *53*, 623–658.
Kogan, F. N. (1997). Global drought watch from space. *Bulletin of the American Meteorological Society*, *78*, 621–636.
Kogan, F., Yang, B., Wei, G., Zhiyuan, P., & Xianfeng, J. (2005). Modelling corn production in China using AVHRR-based vegetation health indices. *International Journal of Remote Sensing*, *26*, 2325–2336.
Koren, O., & Bagozzi, B. E. (2017). Living off the land: The connection between cropland food security, and violence against civilians. *Journal of Peace Research*, *54*, 351–364.
Leenders, R., & Heydemann, S. (2012). Popular mobilization in Syria: Opportunity and threat and the social networks of the early risers. *Mediterranean Politics*, *17*, 139–159.
Loveland, T. R., & Dwyer, J. L. (2012). Landsat: Building a strong future. *Remote Sensing of Environment*, *122*, 22–29.
Messer, E., & Cohen, M. J. (2015). Breaking the links between conflict and hunger redux. *World Medical & Health Policy*, *7*, 211–233.
Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, *112*, 358–375.
Mueller, H., & Rauh, C. (2022). Using past violence and current news to predict changes in violence. *International Interactions*, *48*, 579–596.
Pedelty, J., Devadiga, S., Masuoka, E., Brown, M., Pinzon, J., Tucker, C., et al. (2007). Generating a long-term land data record from the AVHRR and MODIS instruments. In *2007 IEEE International geoscience and remote sensing symposium* (pp. 1021–1025). IEEE.
Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*, 705–871.
Pettersson, T., Davies, S., Deniz, A., Engström, G., Hawach, N., Högbladh, S., et al. (2021). Organized violence 1989–2020 with a special emphasis on Syria. *Journal of Peace Research*, *58*, 809–825.
Pinstrup-Andersen, P., & Shimokawa, S. (2008). Do poverty and poor health and nutrition increase the risk of armed conflict onset? *Food Policy*, *33*, 513–520.
Raleigh, C., & Hegre, H. (2009). Population size concentration, and civil war. A Geographically Disaggregated Analysis. *Political Geography*, *28*, 224–238.
Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, *47*, 651–660.
Richardson, L. F. (1960). *Statistics of deadly quarrels. vol. 10*. Boxwood Press.
Rojas, O., Vrieling, A., & Rembold, F. (2011). Assessing drought probability for agricultural areas in Africa with coarse resolution remote sensing imagery. *Remote Sensing of Environment*, *115*, 343–352.
Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., et al. (2020). Covid-19 future forecasting using supervised machine learning models. *IEEE Access*, *8*, 101489–101499.
Schrodt, P. A., & Gerner, D. J. (2000). Cluster-based early warning indicators for political change in the contemporary Levant. *American Political Science Review*, *94*, 803–817.
Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*, 29–56.
Singer, J. D. (1973). The peace researcher and foreign policy prediction. *Peace Science Society (International)*, *21*, 1–13.
Small, M., Singer, J. D., & Bennett, R. (1982). *Resort to arms: International and civil wars 1816-1980*. SAGE Publications, Incorporated.
Sorokin, P. A. (1962). *Social and Cultural Dynamics: Fluctuation of social relationships, war, and revolution. volume 3*. Bedminster Press.
Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset Version 21.1. *Journal of Peace Research*, *50*, 523–532.
Tatem, A. J. (2017). Worldpop open data for spatial demography. *Scientific Data*, *4*, 1–4. http://dx.doi.org/10.5258/SOTON/WP00660.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*, 267–288.
Tollefsen, A. F., Strand, H., & Buhaug, H. (2012). PRIO-GRID: A unified spatial data structure. *Journal of Peace Research*, *49*, 363–374.
Vesco, P., Hegre, H., Colaresi, M., Jansen, R. B., Lo, A., Reisch, G., et al. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, *48*, 860–896.
Vogt, M., Bormann, N. C., Rüegger, S., Cederman, L. E., Hunziker, P., & Girardin, L. (2015). Integrating data on ethnicity geography, and conflict: The ethnic power relations data set family, version 2021. *Journal of Conflict Resolution*, *59*, 1327–1342. http://dx.doi.org/10.1177/0022002715591215.
Von Uexkull, N., Croicu, M., Fjelde, H., & Buhaug, H. (2016). Civil conflict sensitivity to growing-season drought. *Proceedings of the National Academy of Sciences*, *113*, 12391–12396.
Wan, Z., Hook, S., & Hulley, G. (2021). MOD11A1 MODIS/terra land surface temperature/emissivity daily L3 global 1km SIN Grid, Version 6.1. http://dx.doi.org/10.5067/MODIS/MOD11A1.061, NASA EOSDIS Land Processes DAAC, 2015.
Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, *54*, 125–140.
Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *65*, 95–114.
Wood, R. M., & Sullivan, C. (2015). Doing harm by doing good? The negative externalities of humanitarian aid provision during civil conflict. *The Journal of Politics*, *77*, 736–748.
Wright, Q. (1965). *A study of war* (2nd edition). Chicago, IL: University of Chicago Press.
Xie, J., & Zhu, Y. (2020). Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment*, *724*, Article 138201.
Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K., Fritz, S., et al. (2020). A cultivated planet in 2010–Part 2: The global gridded agricultural-production maps. *Earth System Science Data*, *12*, 3545–3572. http://dx.doi.org/10.5194/essd-12-3545-2020.
Zeitzoff, T. (2017). How social media is changing conflict. *Journal of Conflict Resolution*, *61*, 1970–1991.