Predicting Violent Conflict in Africa
-
Leveraging Open Geodata and Deep Learning for Spatio-Temporal Event Detection

Master's Thesis submitted

to

Dr. Boris Thies

Prof. Dr. Thomas Nauss

University of Marburg

Department of Geography

Climatology and Environmental Modelling

Ecoinformatics

by

Darius A. Görgen

(2622343)

in partial fulfillment of the requirements

for the degree of

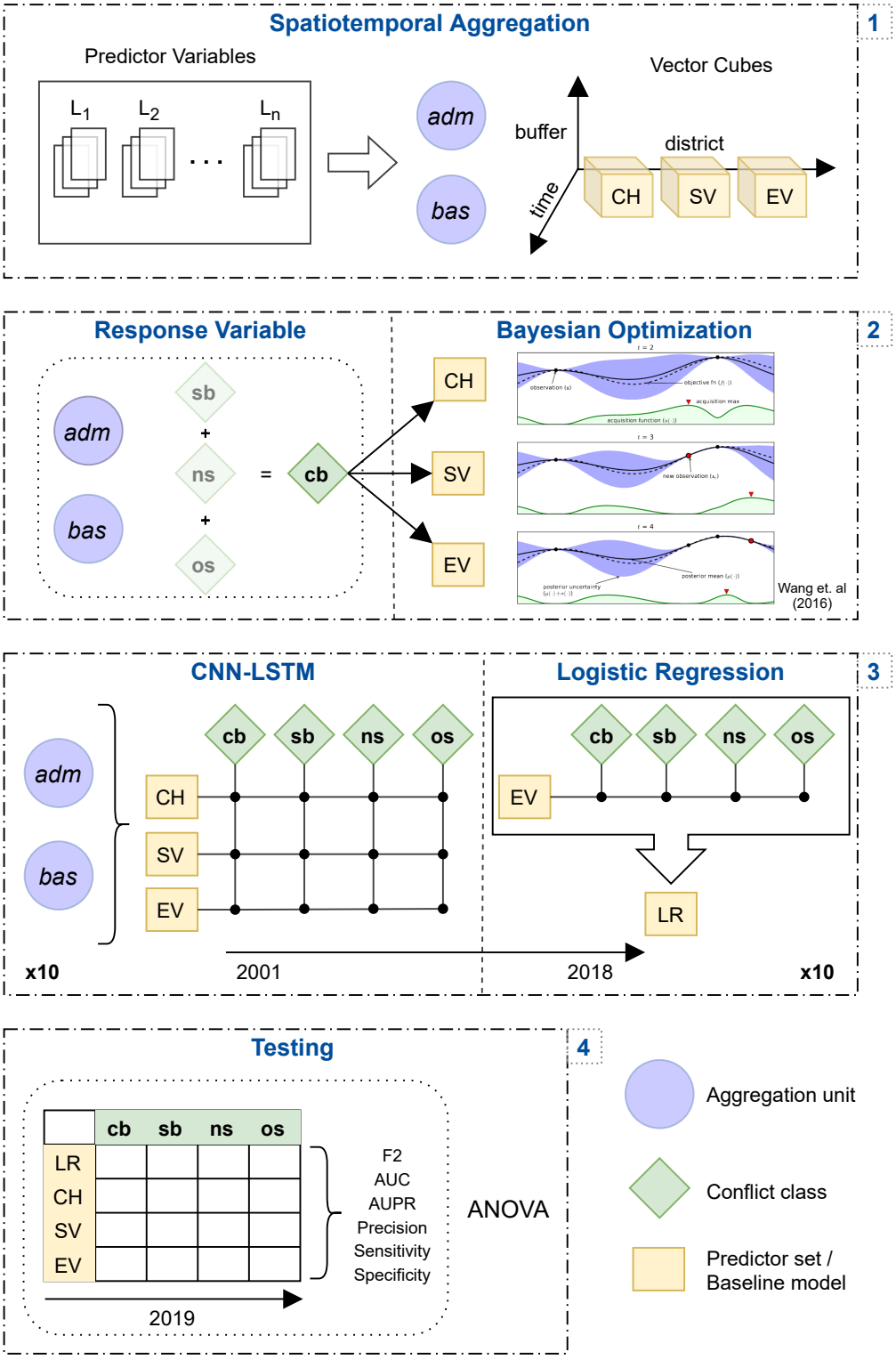
Master of Science Physical Geography

24 March, 2021

Abstract

Violent conflicts endanger human lives, the social cohesion of societies and the natural environment. While the number of intensive international conflicts has remained on a low level during the 21st century, civil wars are on the rise. Since the 1990s, research engages in predicting the outbreak of violence. However, findings on the role the natural environment plays in the emergence of violence remain mostly inconclusive. In order to contribute to the discussion this thesis sets out to compare the predictive performance of deep learning models using data from the Uppsala Conflict Data Program (UCDP) on civil conflict between 2001 to 2019. The data is simultaneously aggregated on administrative districts and sub-basin watersheds and combined with socio-economic and environmental predictors. The hyperparameters of CNN-LSTM architectures are optimized employing a Bayesian Optimization strategy. The results in terms of F_2-score suggest significant improvements for aggregating predictors on sub-basin watersheds ($+7.16, p = 3.4e^{-11}$) as well as integrating environmental predictors ($+3.98, p = 5.9e^{-05}$) for a combined conflict class. For other conflict classes, the results tend to the same direction but are not significant. Through the comparison to existing conflict prediction tools, the thesis exposes the sensitivity of prediction models to spatial scale and units of aggregation. It is argued that in order to fulfill the requirements of effective conflict prevention efforts, prediction research will have to fully integrate modern deep learning frameworks and constant data streams on different earth processes in the future.

Graphical Abstract



Contents

List of Abbreviations

Abbreviation	Explanation
adm	Sub-national administrative districts
ANOVA	Analysis of variance
AUC	Area under the Receiver Operating Characteristic curve
AUPR	Area under the precision-recall curve
bas	Sub-basin watershed districts
BO	Bayesian optimization
cb	Combined violence class
CH	Conflict history
CHIRPS	Climate Hazards Group Infrared Precipitation with Station data
CNN	Convolutional Neural Network
DL	Deep Learning
ET	Total Evapotranspiration
EV	Environmental variables
GCRI	Global Conflict Risk Index
GDP	Gross Domestic Product
GPP	Gross Primary Productivity
LR	Logistic regression baseline
LST	Land Surface Temperature
LSTM	Long Short-Term Memory
MODIS	Moderate Resolution Imaging Spectroradiometer
ns	Non-state violence class
os	State-based violence class
PET	Total Potential Evapotranspiration
ROC	Receiver Operating Characteristic
sb	One-sided violence class
SPEI	Standardized Precipitation-Evapotranspiration Index
SPI	Standardized Precipitation Index
SV	Structural variables
TRI	Terrain Ruggedness Index
UCDP	Upsala Conflict Data Program

List of Figures

List of Tables

1 Introduction

Background. While the world has seen only limited numbers of international conflicts in the 21st century, civil war is on the rise (Figure 1). These conflicts not only are associated with increasing numbers of casualties, but they also threaten the social fabric of societies through migration pressure, especially towards the urban centers (Brzoska and Fröhlich, 2016; Owain and Maslin, 2018; Reuveny, 2007), and put material assets such as infrastructure, agricultural production, and natural forests at risk (Adelaja and George, 2019; Buhaug et al., 2015; Eklund et al., 2017; Jones et al., 2017; Koren, 2018). The scientific literature has successfully revealed a multitude of pathways violent conflict, directly and indirectly, impacts socio-economic indicators. It is commonly agreed that violent conflict is a primary factor impeding sustainable development due to the vigorous impacts on people's livelihoods (Gates et al., 2012). However, the field is complex, with many interdependent relationships. Posing the question of causality, especially for the relationship between the natural environment and conflict, produced heterogeneous and even contradicting empirical results over the last decades (Ward and Bakke, 2005). Methodologically, the scientific community has been focused on using various degrees of derivatives of simple linear regression models to construct theory-based causal models to explain the occurrence, duration, and intensity of violent conflict in relation to natural and socio-economic covariates. Very often, these models proved less valuable for actually predicting violent conflict into the yet unseen future. Some researchers have attributed this shortcoming to the strategy of fitting the parameters of causal models (Colaresi and Mahmood, 2017). Most studies fit their parameters and evaluate the model *in-sample*, meaning that the complete data set is used during both stages, model building and evaluation. While this has the clear advantage of explaining why something has happened within the spatiotemporal domain of a given study, extrapolating these findings to other space-time locations or even to the future of the domain itself led to discouraging results. These shortcomings in accurately predicting the occurrence of violent conflict have not been unrecognized by the general public (Ward et al., 2010). Especially for political decision-makers, accurate predictions into the future and information on how they can act to prevent conflict are of utmost importance. This led to increasing criticism on the usability of research findings because they were neither suited to predict violence outbreaks nor did they inform decision-makers on how to act to achieve a more peaceful future.

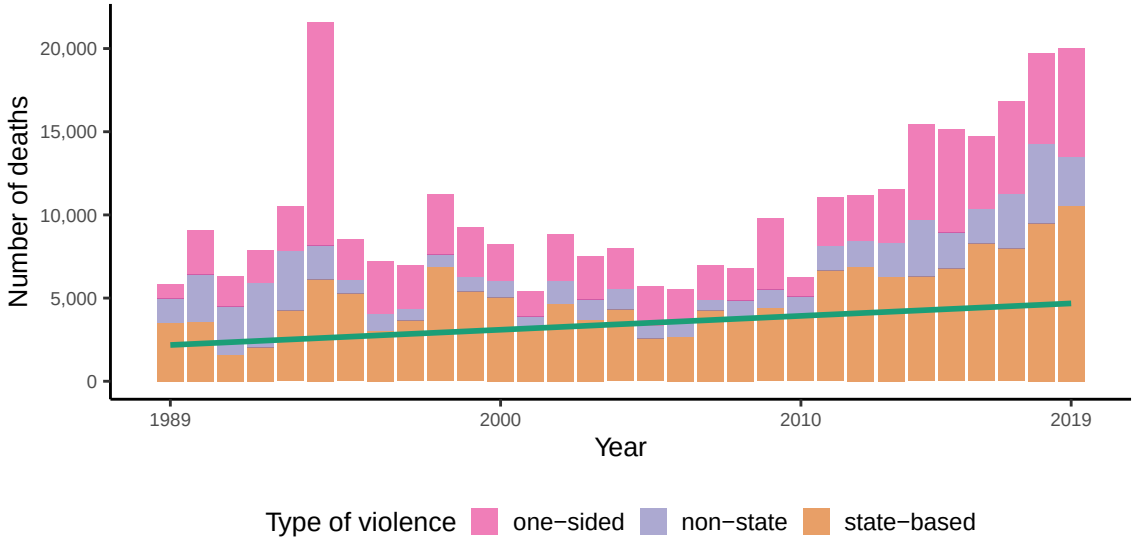


Figure 1: Total number of conflict casualties in Africa 1989-2019. The green line indicates the linear trend for the conflict classes combined.

Recent Developments. Today, massive amounts of data are collected in near-realtime. Due to recent advances in computational power it is now possible to apply computational intensive tools to analyze this data. These developments have led to a shift in the peace and conflict research community to use newly available research opportunities for more robust conflict prediction. A first adaptation to the shortcomings of causal models was a shift towards *out-of-sample* evaluation for prediction models (Ward et al., 2013; Ward et al., 2010). In short, this can be summarized in the observation that a variable is useful that accurately predicts, not necessarily that one that has a lower p-value. This approach to conflict prediction was born out of the necessity to change the community's standpoint towards causal models because they failed to predict more often than not. The question arose about the value of a conflict theory when the theoretically grounded selection of significant variables *does not* lead to useful predictions. In this context, it has been noted that establishing models for prediction itself can inform theory building (Ward et al., 2010). A model that accurately predicts will at least carry some information on the data generating process that can be integrated during the practice of theory building. The second adaptation is in the use of more sophisticated models that possibly capture non-linear relationships better. The conflict research community has seen an increase in the use of machine learning models like Random Forest (Muchlinski et al., 2016; Perry, 2013) as well as Deep Learning (DL) techniques (Beck et al., 2000; Schellens and Belyazid, 2020) to more accurately predict, both in space and time, the occurrence of violent conflict. A third shift has primarily materialized

on the left side of the equation. Recent technological advances allowed for the nearly fully automated coding of event data on violence. Examples are the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010), the Upsala Conflict Data Program (UCDP) (Pettersson and Öberg, 2020), and the Global Database of Events, Language, and Tone (GDELT) (GDELT, 2021), which automatically filter global news feeds for the occurrences of events of interest. Human intervention is mainly restricted to quality control, delivering previously unseen information richness on spatially and temporally disaggregated levels in near-realtime. Conflict research has not yet fully integrated increased data streams on the right side of the equation. Most studies still rely on highly aggregated predictors based on country-years while other research fields, such as different earth sciences, saw an tremendous increase in the availability of data mainly driven by advances of remote sensing technology. Only a few conflict prediction studies make use of disaggregated data sets on the sub-national level and higher temporal resolution. In principle, this can be explained due to the difficulties associated with a spatial-continuous mapping of socio-economic variables. The PRIO-GRID is one example to overcome this limitation using statistical methods to distribute survey data in space (Tollefsen et al., 2012). Other projects also showed promising results in mapping demographic data sets into regularly spaced grids worldwide (WorldPop, 2018). For natural resources, however, many analysis-ready data sets with high spatiotemporal resolution are already available, and more are yet to come, e.g., through the European Copernicus Program (European Union, 2021).

Environmental Causes of Conflict. Whether environmental change can be a driver of violent conflict has been asked since the 1990s. Homer-Dixon presented his famous pie metaphor differentiating between three dimensions of environmental scarcity based on several qualitative case studies in the early 1990s (Homer-Dixon, 1995, 1994, 1991). The general size of the pie from which each individual within a society gets a share is determined by the availability of natural resources. Decreasing the availability or quality of a resource will consequently decrease the overall size of the pie. Increasing population numbers or changes in the consumption pattern can lead to a reduced share available per capita. Moreover, discriminatory distributional policies can reduce the share an individual or social group receives in relation to others leading to, e.g., the marginalization of specific linguistic, religious, or ethnic populations. These dimensions of environmental scarcity can be observed in various combinations, eventually increasing the overall conflict risk.

However, as Sachs and Warner (1995) showed, the abundance of natural resources does not prevent conflicts from arising. On the contrary, based on empirical evidence, they showed that resource-rich countries were more likely to experience conflict and impeded economic development than resource-poor countries. For countries with the most available natural resources the risk of experiencing conflicts was again reduced to a low level. These findings were coined as the *resource curse* and inspired their own research line, mostly supporting the original findings (Alexeev and Conrad, 2009; Antonakakis et al., 2015; Bjorvatn et al., 2012; Boschini et al., 2013). Early on, it was noted that a high level of non-linearity is associated with the analyzed processes. Focusing explicitly on rebellions and insurgencies, Collier (1998) showed that an econometric model including the opportunity costs of rebellions could explain observed occurrences. However, they conclude that the relationship between the covariates and the conflict outcome might be non-monotonic. For example, very high rates of ethnic fractionalization do not necessarily increase the conflict risk. Instead, fractionalization into two similarly sized groups shows the highest increase. In another study, they showed that African countries show a greater baseline risk for conflict compared to other world regions due to their economic structure. But the ethnic composition primarily acts as an reduction factor of conflict risk (Collier and Hoeffler, 2002).

In the context of increasing awareness of the consequences of climate change and its challenges for sustainable development, it is surprising to see only a few attempts to link environmental change to violent conflict during the last decade. In most of the studies, natural resources are proxied by primary commodity exports (Collier and Hoeffler, 2004; Fearon and Laitin, 2003; Muchlinski et al., 2016). More recent attempts emphasize capturing non-linearities between predictor variables and the outcome, but a systematic analysis of environmental predictors is mostly absent. Hegre et al. (2019) produce an early-warning system for violence using model ensemble forecasts based on different spatial aggregation units. Natural resources in terms of different land cover classes and the distance to diamond and oil extraction sides are only included for one of the aggregation units and not systematically analyzed. Schutte (2017) presents an innovative approach based on modeling conflicts as point processes on the African continent. However, natural resource variables are restricted to accessibility and land cover data and not systematically analyzed. Witmer et al. (2017) set out to estimate the long-term dynamics of subnational conflict (2015-2065) based on different climate scenarios. While their model explicitly analyzes the impact of precipitation and temperature changes, long-term scenario simulations, due to their nature, do

not serve well as early-warning systems for decision-makers to prevent conflict. Halkia et al. (2020) presented the Global Conflict Risk Index (GCRI), a tool used by the European Union for conflict prediction. It is based on a simple logistic regression model, nonetheless achieves remarkable accuracy results. However, variables related to natural resources are included under the label of economic variables. They are limited to indices on food security and water stress as well as raw oil production. Even though not peer-reviewed, the World Resource Institute (WRI) published a technical note reporting on their efforts for a violent conflict forecasting tool (Kuzma et al., 2020). By using a Random Forest model and incorporating several environmental variables concerned with food production and water availability, they produce promising results for the African continent, the Middle East and parts of Asia. They include an analysis of variable importance which only indicates as relevant a few of the environmental predictors. Schellens and Belyazid (2020) analyze the role of natural resources in violent conflict prediction through modern machine learning techniques. They compare a standard logistic regression model versus two types of neural networks and a Random Forest model. They particularly test model performance on different sets of predictors. One of these sets includes natural resources conceptualized as agricultural production, forest area, primary commodity exports of non-renewable resources, water access, withdrawal, and available reserves, as well as resource rents. Their results indicate an improvement of the Random Forest model when natural resource variables are included. However, the analysis of the neural networks remains rather shallow because only basic network architectures were considered.

Research Question. The outlined review of the literature leads to the main research question of this thesis, namely if a modern deep learning framework and the vast availability of open geodata can positively impact the task of spatiotemporal conflict detection. Note that detection and prediction will be used interchangeably in this thesis. Continuing with a brief definition of the two basic concepts, open geodata can be defined as freely available and usable data related to some information of space. Very commonly, it is based on remote sensing imagery. Remote sensing enables the collection of spatiotemporal comprehensive measurements. In principle, it works by measuring the spectral reflectance of the Earth's surface in different bands across the electromagnetic spectrum (De Jong et al., 2007). Through physically informed transformation models, the original spectral reflectance values can be translated into measurements of physical variables, e.g., evapotranspiration from the canopy or biomass production of the vegetation cover. Also, land cover can be mapped

regularly, informing, e.g., dynamics of deforestation or the extension of agricultural farmland. There are many agencies, research institutions, and companies that provide a wide range of so-called *value-added* remote sensing products, a lot of them at zero cost (Radočaj et al., 2020). On the other hand, DL is a subcomponent of machine learning, often focusing on solving supervised classification and regression problems by using neural networks (LeCun et al., 2015). These networks consist of varying numbers and architectures of *hidden layers*, mapping a specified input to the desired output through non-linear transformations of the data values. They are supervised because the outcome of specific observations is known, and a model is tasked with optimizing its parameters towards lowering the prediction error. The model performance is evaluated based on an *out-of-sample* validation set. A suitable parametrization of the model is equivalent to its generalization potential to unseen data and consequently will yield high-performance metrics on the validation set.

In the context of conflict prediction, DL methods seem promising to deliver accurate predictions in both the temporal and the spatial domain (Emmert-Streib et al., 2020). The occurrence of conflict is a highly complex process, with many interdependent variables from the social, economic, and environmental dimensions of human reality. DL could indicate a way to yield accurate conflict predictions despite this complexity. Additionally, the availability of spatial and temporal comprehensive data sets mainly based on remote sensing technology allows gathering predictor variables suited for conflict prediction in a time and cost-efficient manner. For the presents study's design, two theoretical grounded assumptions are of particular relevance. First, it is assumed that linkages between environmental variables and conflict are not necessarily observed *in-situ* nor *instantaneously*. The first aspect of this assumption means that environmental processes in one location might impact the conflict risk at yet a distinct but related location. For example, a bad harvest, e.g., through an extensive drought, might aggravate conflict risk in a neighboring urban center due to increased food prices. The same increase in conflict risk is not necessarily observed in the rural areas where the root cause, i.e., the extensive drought, took place. The second aspect emphasizes that the problem at hand is one of a time series. For example, deforestation might not instantaneously negatively impact the livelihoods of the communities. However, ongoing deforestation in the long term can lead to increased soil erosion, undermining the basis of rural livelihoods and eventually increase conflict risk. Second, environmental change does not stop at national borders. Put differently, administrative boundaries rarely follow natural borders, e.g., in the case of a river course separating two nation-states. If environmental dynamics are the sub-

ject of analysis, this could mean that administrative districts are not necessarily the optimal choice to study these dynamics. The interdependence between environmental change and the risk of conflict might be better depicted on a different scale, e.g., by aggregating variables based on watersheds.

Guided by these assumptions, a workflow is set up to test for two distinct hypotheses. These are:

- H1: *Environmental predictors increase the performance of deep learning models for the conflict prediction task over models based solely on the conflict history and structural variables.*

- H2: *Aggregating predictor and response variables on the basis of sub-basin watersheds delivers better predictive performance than aggregating on sub-national administrative districts.*

General Methodology. To test for these hypotheses, the study's design is presented in detail in the following section. Briefly, (i) several socio-economic and environmental variables available in spatially explicit formats are collected and aggregated for sub-national administrative districts and sub-basin watersheds, (ii) spatial buffers around these districts are used to account for processes occurring in a district's larger spatial neighborhood, (iii) DL models are trained on different predictor sets consisting of the conflict history, structural, and environmental variables to solve the conflict prediction statistically formulated as a time series problem, (iv) model performances are evaluated on an *out-of-sample* validation set and cross-checked with a logistic regression baseline, (v) performance results are comprehensibly reported for all model configurations for different classes of violent conflict, allowing for an analysis of variance in the predictive performance to obtain statistical significant indications in relation to the hypotheses.

2 Data

2.1 Area of Interest

To account for the hypothesis that aggregating predictor variables on the basis of watersheds will lead to higher detection accuracy compared to administrative boundaries, two different sets of spatial units are selected. The first represents sub-national administrative boundaries (referred to as *adm* hereafter) derived from the Natural Earth project in a vector format (South, 2017). Only polygons on the African continent as well as Madagascar from the third administrative level are selected. The procedure results in a total of 847 polygons covering the study area (Figure 2).

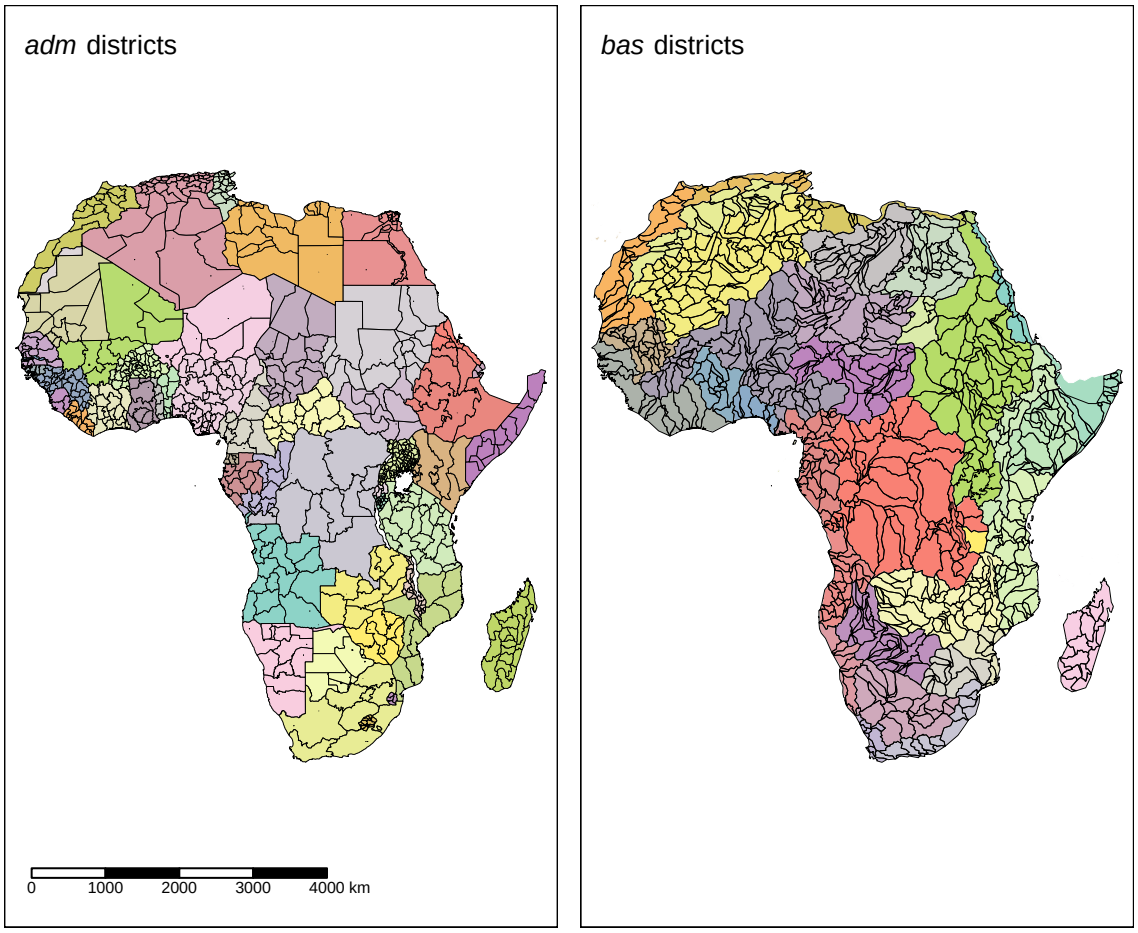


Figure 2: Overview of the administrative (left) and sub-basin districts (right) used for data aggregation.

The second set of spatial units represents sub-basin watersheds (referred to as *bas* hereafter) which are downloaded from the HydroSHEDS project (Lehner et al., 2008). To keep the

number of polygons comparable to the number of administrative units, the fifth level out of a total of 12 levels of watershed delineations is selected. The procedure results in a total number of 1013 polygons covering the African continent and Madagascar (Figure 2).

Table 1: Descriptive statistics on the area of the spatial aggregation units in km². (*adm* represents administrative units, *bas* represents sub-basin watersheds)

Spatial unit	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
adm	847	2.693	2730.437	9264.974	34837.50	37425.82	622228.3
bas	1013	0.608	8023.777	18279.980	29067.15	37924.56	313042.4

The descriptive statistics of area distribution among the two different sets reveals that the sub-basin watersheds have a lower average value of 29,067.15 km² compared to 34,837.5 km², while the administrative units have the highest maximum value of 622,228.3 km² which is nearly twice as large as the largest watershed (Table 1).

Three different spatial buffers of 50, 100, and 200 km are processed for each district to include information on the geographical neighborhood. It is assumed that processes in the larger neighborhood of a district might influence the occurrence of conflicts. This influence can take the form of spillover effects for which empirical evidence has been found, e.g., for military spending (Phillips, 2015), reducing regional trade volumes (Murdoch and Sandler, 2004) or short-term reduction in economic growth (Murdoch and Sandler, 2002). These effects can manifest in environmental changes and reduction of agricultural production, as has been the case for Syria and Iraq during the war against the so-called Islamic State (Eklund et al., 2017). For Sub-Saharan Africa, additional evidence has been presented that spillover effects are more pronounced than in the rest of the world (Carmignani and Kler, 2016). For each of the buffered areas, the same variables are extracted as for individual districts, resulting in a 3-time increase in the data set size.

2.2 Spatiotemporal Aggregation

For the establishment of a regular data set used for the training process a number of predictor variables in the raster format need to be processed into a regular time series based on the spatial aggregation districts used for prediction (Table 3). Raw remote sensing scenes, as well as *value-added* raster products, vary greatly in the spatiotemporal resolution they are available

at (Table 3). The resolution is a function of the technical prerequisites of a given sensor or statistical method in the case of socio-economic variables and design choices by the product providers. For example, together, both satellites equipped with the Moderate Resolution Imaging Spectroradiometer (MODIS) cover every point on earth every 1 to 2 days. However, certain products such as evaporation or gross primary productivity are processed to 8-day composites. This design choice was made to assure valid measurements of certain variables that might be blocked on some day, e.g., through persistent cloud coverage, especially in the tropics (Steven Running et al., 2017). Other products, such as precipitation data or population counts are available at a monthly or yearly temporal resolution, mainly due to the statistical processes involved in their production, e.g., in the latter case by relying on yearly survey data. To account for this variability in data set characteristics, the decision is made to aggregate on a monthly district level (district-months). A pixel-based approach, which would require the transformation of all data sets to a common spatial resolution, is rejected due to the very low occurrences of the response variables and the great variability of spatial resolutions across the data sets. That way, all data sets can be processed at or close to their native resolution since the final value for a given district-month is obtained by a zonal statistic operation. To this end, the R package *gdalcubes* was used (Appel and Pebesma, 2019). This package was specifically designed for the efficient harmonization of remote sensing products with varying spatiotemporal resolutions. It programmatically conducts the transformations needed to bring a given raster data set to the desired spatio-temporal resolution and provides efficient routines to calculate zonal statistics for the intersection areas between rasters and polygons.

For data sets that are available at a yearly time scale or static over the complete time-series a last-observation-carried-forward approach is0 chosen to obtain a monthly data set. For data sets that are available at a finer resolution, a suitable aggregation operation to the measured variable is chosen (Table 3). In summary, a vector data cube consisting of 847 (1013) districts for *adm* (*bas*) for 228 individual months (2001 to 2019) and 176 ((40 predictors + 4 responses) x 4 buffers) variables summing up to a total of 33,988,416 (40,649,664) individual data points is constructed. The individual variables selected for training are explained in detail in the following sections.

2.3 Response Variable

Data from UCDP is used for the response variable (Pettersson and Öberg, 2020). The data is an event database, meaning that every reported violent conflict with at least 25 fatalities per year is included as a single event. Each event is associated with information on involved actors, time, location, and estimated fatalities by party involved. Since the data is mainly collected automatically, inaccuracies concerning the exact location or the number of deaths are frequently observed. However, the information on the reliability for each logged event is included. The latest version of the database as of the time of writing this thesis for the African continent is used. Only events occurring between 2001 and 2019 are included. The data is further filtered to contain only events for which the quality of the geographic localization is assured to be accurate on the sub-national level. Also, only events are included with timestamps which are reported to be accurate on a monthly scale. The data is provided distinguishing between three different classes of conflict. These are state-based violence (referred to **sb** hereafter), non-state violence (**ns**), and one-sided violence (**os**) as defined in Gleditsch et al. (2002), Sundberg et al. (2012), and Eck and Hultman (2007), respectively. UCDP summarizes these definitions as "violence between two organized actors of which at least one is the government of a state, violence between actors of which neither party is the government of a state, and lastly, violence against unarmed civilians perpetrated by organized non-state groups or governments" (Allansson, 2021). For each class, a spatiotemporal aggregation based on the different analysis units on a monthly time scale is applied. Additionally, by a summation of casulties a fourth class representing a combination of the three base classes is introduced (referred to as **cb**). These aggregations represent the basic unit of analysis for this project and are used as the response objects for the detection algorithm. Because the final prediction problem is conceptualized as a binary classification task (peace vs. conflict) any district-month with fatalities greater than 0 is transformed to a value of 1, representing conflict, while district-months with no fatalities are assigned a value of 0, representing peace.

Table 2: Percentage of conflict district-months for different training data sets across aggregation units and classes of conflict.

Dataset	**cb**	**sb**	**ns**	**os**
adm				
Training	3.102	1.752	0.775	1.130
Validation	4.801	2.607	1.830	1.702
Test	5.239	2.986	1.682	2.096
bas				
Training	2.347	1.282	0.650	0.912
Validation	4.187	2.229	1.571	1.456
Test	4.660	2.694	1.493	1.917

General: Values are given in percent.

For the *adm* representation of the data, the total number of district-months is 193,116 (847 districts x 228 months). For the *bas* representation it is 230,964 (1,013 districts x 228 months). To summarize, for each district, a time-series worth of 228 data points for four different response variables is available. Table 2 summarizes the occurrence of conflict district-months per outcome variable for both aggregation units. The definition of training, validation, and testing data set will be further discussed in Section 3.4.

2.4 Predictor Variables

The search of predictor variables was guided by a literature review of quantitative studies in the field of conflict research and the selected variables are presented in Table 3. Variables that have been shown to correlate with different forms of conflict and proved useful predictors were of primary interest. However, it became evident that most studies were concerned with predicting conflict work on a country-year basis (Halkia et al., 2020; Hegre et al., 2019; Schellens and Belyazid, 2020). In cases the units of analysis were more fine-grained, they are most commonly restricted to sub-national administrative boundaries (Kuzma et al., 2020). For this kind of problem formulation, several socio-economic variables that are collected on an administrative level are easily adaptable. For example, indicators collected on a national level such as a democracy index, literacy rates, or income inequality can be disaggregated to the sub-national level or held constant for all districts within a nation. However, natural boundaries, such as watersheds, rarely follow political boundaries but they often

intersect with each other. Thus, variables available as national aggregates can not easily be disaggregated at the watershed level. The use of such administrative-bound variables was not rendered feasible for this study.

In a more specific context, this meant that only predictor variables which could be aggregated on both the levels of *adm* and *bas* are included. Additionally, a district's spatial neighborhood was theorized to serve as a predictor. These neighborhood predictors were conceptualized as spatial buffers of 50, 100, and 200 km around each district. To enable this line of problem formulation, the search for predictors was limited to data sets in the raster format. In the current project, raw satellite imagery would not prove very useful to model the outcome variable. Instead, the search was concentrated on so-called *value-added* products, e.g. products for which transformations into measurements of physical variables already have been applied.

In the following section, the selected data sets used to extract predictor variables are explained in detail. The structure mirrors the experimental design of this study in the sense that predictors are grouped into three distinct sets of variables. The most basic predictor set only contains information on the past conflict history (referred to as CH). Structural predictors (SV) contain the conflict history and selected structural variables for which a measurement is available at least once every year. The environmental predictor set (EV) contains all variables of the two preceding variable sets and selected environmental variables for which a data measurement is available every month. Thus, the predictor sets not only incorporate different aspects of a district in terms of its structural and environmental characteristics, but they also differ in the time-scale observations for given predictor variables are recorded.

Table 3: Spatio-temporal properties of predictor variables.

Name	Spatial Resolution	Temporal Resolution	Unit	Aggregation
Baseline				
Conflict history	-	monthly	binary	-
Structural				
Terrain Ruggedness Index (log)	0.0008°	static	m	mean
Travel time (log)	0.008°	static	minutes	mean
Livestock (log)	0.08°	static	2010 heads	sum
Population (log)	0.008°	yearly	persons	sum
Youth bulge	0.008°	yearly	%	sum
Dependency ratio	0.008°	yearly	%	mean
GDP (log)	0.08°	yearly	2011 USD	mean
Cropland	0.005°	yearly	%	sum
Forest cover	0.005°	yearly	%	sum
Builtup area	0.005°	yearly	%	sum
Grassland	0.005°	yearly	%	sum
Shrubland	0.005°	yearly	%	sum
Barren land	0.005°	yearly	%	sum
Water bodies	0.005°	yearly	%	sum
Environmental				
Precipitation	0.05°	monthly	mm	mean
Precipitation anomaly	0.05°	monthly	mm	mean
SPI	0.05°	monthly	-	mean
SPEI	0.05°	monthly	-	mean
Land Surface Temperature	0.05°	monthly	K	mean
Evapotranspiration	0.01°	monthly	kg/m^2	mean
Gross Primary Productivity	0.01°	monthly	$kg\ C/m^2$	mean
Precipitatation agr.	0.05°	monthly	mm	mean
Precipitation anonmaly agr.	0.05	monthly	mm	mean
SPI agr.	0.05°	monthly	-	mean
SPEI agr.	0.05°	monthly	-	mean
Land Surface Temperature agr.	0.005°	monthly	K	mean
Evapotranspiration agr.	0.005°	monthly	kg/m^2	mean
Gross Primary Productivity agr.	0.005°	monthly	$kg\ C/m^2$	mean

General: Variables denoted with *agr.* were calculated by a multiplicative interaction with a binary cropland mask.

2.4.1 Conflict History (CH)

The most basic predictor set consists only of the past conflict history for every district. Depending on which outcome variable is predicted (e.g. **cb**, **sb**, **ns**, or **os**), four different predictors are available. These are the conflict history for the unit itself and its three spatial buffers of 50, 100, and 200 km representing the conflict history in the spatial neighborhood. The value of measurement is a binary encoding whether a given district-month experienced conflict or not.

2.4.2 Structural Variables (SV)

Terrain Ruggedness Index (TRI). The TRI is used as an indicator for the landscape organization of a given district. Several studies have found significant correlations between mountainous areas and conflict. However, a clear definition of the parameter is hardly found in any of these publications (Collier and Hoeffler, 2004; Fearon and Laitin, 2003; Hegre et al., 2019; Muchlinski et al., 2016). Here, the TRI is used as an indicator of the ruggedness of the terrain based on the calculation of a digital elevation model (DEM). The TRI estimates the landscape heterogeneity in a grid cell's neighborhood, calculating the sum of change in elevation to all eight neighbors of a cell (Riley et al., 1999). The DEM from the Shuttle Radar Topography Mission (SRTM) is used (Jarvis et al., 2008). This DEM comes at a resolution of 0.0008° and it is processed at its native resolution. The unit of measurement is in meters. This variable represents a static value and does not change within the time series. There are no missing values for land areas. The mean of all pixels within a district is extracted as a zonal statistic for the district level aggregation. Due to the great variance of the TRI between the districts, this predictor's natural logarithm is used during training. The distribution of values between *adm* and *bas* are quite similar with the administrative districts showing slightly higher average and maximal values (Table 4).

Table 4: Descriptive statistics of the Terrain Ruggedness Index based on different spatial aggregation units. (Unit of measurment: meters)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
adm	0.839	2.101	3.584	5.175	6.993	28.539
bas	0.842	1.999	2.853	3.913	4.801	21.346

Travel Time. This predictor is used as an indicator of the centrality or remoteness of a district. The data set used indicates the travel time from a given cell to the closest city of 50,000 or more inhabitants (Nelson, 2008). It is calculated based on a cost-distance model incorporating various variables contributing to the accessibility of a location such as roads, railways, land-cover, slope, and others (Nelson, 2008). This indicator has been found a valuable predictor for the intensity of violence (Schutte, 2017). The native resolution of the data set is 0.008°. It is static and does not change within the time series. The unit of measurement is in minutes of travel time to the nearest city. There are no missing values on land areas. For the aggregation on the district level, the mean travel time is calculated. Due to the great variance of travel time between the districts, this predictor's natural logarithm is used during training. The *bas* districts show an average travel time nearly twice as high compared to the *adm* districts (Table 5).

Table 5: Descriptive statistics of travel time to cities $\geq 50,000$ inhabitants based on different spatial aggregation units. (Unit of measurement: minutes)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
adm	4.312	161.133	264.22	353.441	416.073	2571.178
bas	45.8	290.209	459.661	690.55	848.293	5004.453

Livestock. Numbers on livestock are used as an indicator for the characteristics of the agricultural sector in a district. A data set containing the global distribution for cattle, buffaloes, horses, sheep, goats, pigs, chicken, and ducks representing the condition in the year 2010 is used (Gilbert et al., 2018). The data set comes at a spatial resolution of 0.08°. There are some missing values for districts in the Sahara desert. The numbers for different species are summed on a pixel basis, and the total sum of livestock heads is calculated for each district. The literature review revealed that this indicator has not been used in previous studies on conflict prediction. Due to the great variance of livestock numbers between the districts, the natural logarithm of this predictor is used during training. The aggregation of *adm* and *bas* districts generally shows an equal distribution with *adm* being characterized by slightly higher livestock counts (Table 6).

Table 6: Descriptive statistics of total livestock numbers based on different spatial aggregation units. (Unit of measurement: 2010 heads)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	0	439279	1197960	3009761	3411511	128835648	1680
bas	0	73380	442958	2510632	1937566	104541794	1440

Population. The total number of the population is frequently used as a predictor across different conflict studies (Collier and Hoeffler, 2004; Fearon and Laitin, 2003; Halkia et al., 2020; Hegre et al., 2019). A spatially explicit data set produced by the WorldPOP project is used in this study (WorldPop, 2018). This data set is produced following the methodology reported in Pezzulo et al. (2017). Based on over 6,000 subnational data records, a spatially continuous high-resolution gridded data set of female and male population counts by age cohorts was produced at a resolution of 0.008°. The unit of measurement is the total number of persons irrespective of the sex. Per district, the sum of the total population is calculated as a zonal statistic. There is one district for the *bas* representation with missing data. During training, the population count's natural logarithm was used because the scale of the population counts varies greatly between districts. The values are generally comparable between the *adm* and *bas* districts though *adm* shows a higher average value while *bas* shows the highest maximum value (Table 7).

Table 7: Descriptive statistics of total population numbers based on different spatial aggregation units. (Unit of measurement: persons)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	586	210917	481571	1192672	1256860	38961168	-
bas	0	11689	141719	998214	787904	49778275	240

Youth Bulge. Halkia et al. (2020) and Schellens and Belyazid (2020) recently used this predictor to characterize the age structure of a country. They define the youth bulge as the percentage of the population aged between 15 and 24 divided by the population older than 25. In this thesis, it is calculated using data from the WorldPop project (WorldPop, 2018). For every pixel, the total number of people for the respective age groups is calculated. Then, the sum of persons per age group is extracted for every district to finally calculate the

youth bulge percentage. There are no missing values. The distribution of values between *adm* and *bas* is relatively similar (Table 8) ranging from 0 % for *bas* districts and a minimum value of 19% for *adm* districts to maximum values of 104 % and 115 %, respectively. The average value is comparable between the spatial units between 54.7 % to 57 %.

Table 8: Descriptive statistics of the youth bulge based on different spatial aggregation units. (Unit of measurement: percent)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
adm	19.142	51.194	58.057	56.977	64.523	115.117
bas	0	49.142	56.041	54.66	62.041	103.923

Dependency Ratio. This variable describes the ratio between dependents to the working-age persons in a population. It is used as an official indicator by the World Bank, and it was chosen here in addition to the youth bulge predictor to include some information on the elderly of a given population. Dependents are defined as being younger than 15 or older than 64. People between these age classes are considered as the working-age population. The procedure to calculate this variable is the same as for the youth bulge variable, except that different age groups were chosen. There are no missing values. The distribution of values between *adm* and *bas* is relatively similar (Table 9) ranging from 0 % for *bas* districts and 28 % for *adm* districts to a maximum value of 157 % for *adm* districts and 150 % for *bas* districts.

Table 9: Descriptive statistics of the dependency ratio based on different spatial aggregation units. (Unit of measurement: percent)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
adm	27.865	73.846	96.74	91.647	108.656	157.408
bas	0	70.011	94.09	88.617	107.447	150.318

Gross Domestic Product (GDP). In most studies concerned with conflict prediction, GDP is found as a valuable predictor (Collier and Hoeffler, 2004; Halkia et al., 2020; Hegre et al., 2019; Muchlinski et al., 2016; Rost et al., 2009; Schellens and Belyazid, 2020; Schutte, 2017; Ward and Bakke, 2005). The variable is included in different forms, e.g., the natural logarithm, GDP per capita, or the growth of GDP between

time steps. In this project, a yearly available gridded data set spanning the years 1990 to 2015 at a spatial resolution of 0.08° is used (Kummu et al., 2018). Missing values from 2016 to 2019 are interpolated by a simple linear interpolation on a pixel basis. There are some missing values, mainly for very small districts located at the coastal zones across the whole continent where the raster data set showed no values. The unit of measurement is 2011 international US Dollars. The average GDP value is calculated as a zonal statistic. Because the values vary greatly between districts, the natural logarithm of GDP is used as a predictor during training. Both *adm* and *bas* show very similar descriptive statistics ranging from about 280 USD to about 44,800 USD (Table 10).

Table 10: Descriptive statistics of the Gross Domestic Product based on different spatial aggregation units. (Unit of measurement: 2011 USD)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	284.04	1228.172	1889.668	4439.684	5569.510	44808.06	1680
bas	278.97	1501.878	3206.946	5855.363	9346.887	44808.06	960

Land Cover. In order to include information on the spatial structure of a district, several land cover classes were included as predictors in the training process. Different forms of land cover have been included in prior studies of conflict prediction. Hegre et al. (2019) include areal statistics on agriculture, barren land, shrubland, pasture, urban areas, and forest cover in their prediction model. Schellens and Belyazid (2020) include the percentage of arable land and forest cover in their analysis, while Schutte (2017) decided to include a remote sensing based proxy for green vegetation. In this thesis, several broad classes of land cover are included. These are cropland, forest cover, built-up area, grassland, shrubland, barren land, and water bodies. All of these predictors are delineated using the MCD12Q1 product generated yearly from the Terra and Aqua MODIS satellites (Friedl and Sulla-Menashe, Damien, 2019). The product is delivered with five different land cover classification schemes, of which only two are selected due to the suitability of their classification scheme (Sulla-Menashe and Friedl, 2018). These are the classifications of the International Geosphere-Biosphere Programme (IGBP) and the University of Maryland (UMD). For a pixel to be assigned one of the target classes listed above, both the IGBP and UMD classification need to correspond to the same class. The data set is processed at a spatial resolution of 0.005°. For each district, the count of pixels for each class is extracted. The percentage of coverage

is calculated, with 0 representing a valid value for cases when the respective class is not present. The data is available every year. Missing values occur for districts, where neither of the target land cover classes are present. Concerning the distribution of values, very low and high coverages are observed (Table 11). The average values reveal some differences between *adm* and *bas* districts, for example, that the coverage with barren land is almost three times as high for *bas* districts compared to *adm*. Also, higher rates of cropland coverage are observed with *adm* districts.

Table 11: Descriptive statistics of different land cover classes based on different spatial aggregation units. (Unit of measurement: percent)

Spatial Unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Barren land							
adm	0	0.00000	0.00000	10.6738625	0.14146	99.99545	10164
bas	0	0.00000	0.02318	32.2734785	98.59107	100.00000	12156
Cropland							
adm	0	0.04878	2.97109	20.2042559	35.22953	99.68764	10164
bas	0	0.00000	0.02545	6.0829911	2.57099	89.23104	12156
Forest							
adm	0	0.00000	0.07553	7.5630393	5.19282	99.30686	10164
bas	0	0.00000	0.00000	6.8620305	2.32165	99.93229	12156
Grassland							
adm	0	1.27006	12.92306	29.6158713	55.24245	100.00000	10164
bas	0	0.08412	12.74015	29.3516104	57.14479	100.00000	12156
Shrubland							
adm	0	1.44282	15.51877	27.7411470	47.93473	99.73342	10164
bas	0	0.00992	6.88072	24.1363455	46.26372	100.00000	12156
Built-up							
adm	0	0.01929	0.08562	2.2936831	0.43554	100.00000	10164
bas	0	0.00000	0.02036	0.2668843	0.11210	35.74467	12156
Water							
adm	0	0.00000	0.09216	1.9081408	1.19737	82.05121	10164
bas	0	0.00000	0.00000	1.0266599	0.34510	78.57147	12156

2.4.3 Environmental Variables (EV)

Precipitation. A global data set with monthly precipitation amount is provided by the Climate Hazards Group Infrared Precipitation with Station data (CHIRPS) (Funk et al., 2015). Precipitation has been used previously as an indicator for long-term conflict risk (Witmer et al., 2017). The data set does not solely rely on the in-situ measurement of rainfall, instead available station measurements are extrapolated to unknown areas using different sources of remotely sensed imagery (Funk et al., 2015). It consists of quasi-global rainfall estimates starting from 1981 to the near-present at a spatial resolution of 0.05° and a temporal resoulution of one month. The unit of measurement is *mm*, and the data is processed at its native resolution. The average rainfall estimate is extracted per district. Due to the data set's statistical generation process, there are no missing values for areas on land. Some coastal districts do not align with the raster's spatial extent, resulting in a few missing values. *adm* districts show higher average and maximum rainfall rates compared to *bas* districts (Table 12). Maximum values above 1000 mm within a month are observed for both aggregation units.

Table 12: Descriptive statistics of precipitation amounts based on different spatial aggregation units. (Unit of measurement: mm)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	0	5.951	43.221	82.598	131.588	1563.218	720
bas	0	1.250	10.364	53.394	77.110	1124.251	720

Precipitation Anomalies. This variable describes the observed anomalies in precipitation in relation to a long-term observed monthly average precipitation at a specific location. It is calculated using the CHIRPS data set (Funk et al., 2015). Because rainfall estimates are available starting in 1981 (see above), the average rainfall can be calculated on a pixel basis for a 30-year period (1981 - 2010) for every month. This averaged value is then subtracted from each pixel for the period of interest (2001 - 2019). The unit of measurement is *mm*, and the data is processed at its native resolution. As a zonal statistic, the mean of all pixels within a district is extracted. Similar to the precipitation variable, for some coastal districts missing data is present. The average anomaly for *adm* districts is with 1.4 mm more than twice as high as for *bas* districts with 0.6 mm (Table 13).

Table 13: Descriptive statistics of precipitation anomalies based on different spatial aggregation units. (Unit of measurement: mm)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	-363.784	-9.176	-0.062	1.404	8.01	1214.561	720
bas	-274.300	-3.447	-0.021	0.635	1.34	852.608	720

Standardized Precipitation Index (SPI). This index was developed by McKee et al. (1993) and is widely used in research as an indicator for droughts. For its calculation solely data on precipitation is needed. That is why it is sometimes referred to as an indicator for meteorological drought (Hayes et al., 2011). For a time-series of a location, different sets of averaging periods of variable length are calculated. For each month in the desired output sequence, a Gamma function is fitted on the averaging period to capture the probability of precipitation (McKee et al., 1993). This probability is then used to calculate deviation of the observed precipitation from the normally distributed probability density. McKee et al. (1993) use arbitrary but regular reference periods of up to 48 months, which they interpret as representing short- to long-term precipitation deficits and surpluses. However, because increasing the length of the reference period means fewer data points at the beginning of the time series, here the SPI is calculated for 1, 3, 6, and 12 month reference periods only using the R package SPEI (Begueria and Vicente-Serrano, 2017). The calculation is based on the native resolution of the CHIRPS data set. Even though the value of measurement is dimensionless, because it is standardized comparisons across time and spatial districts are possible. The mean of all pixels within a district is extracted for each reference period as a zonal statistic. Missing values are increasing from the 1 month to 12 month reference period because of the cut-off behavior of the SPI calculation (Table 14). The distribution behaves similar between *adm* and *bas* districts, with the lowest values present for SPI1 between -6 and -7. Maximum values are also quite similar reaching values between 6 and 8, while the average SPI values are only slightly above 0.

Table 14: Descriptive statistics of the Standardized Precipitation Index (SPI) based on different spatial aggregation units. (Unit of measurement: dimensionless)

Spatial Unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
SPI1							
adm	-6.021	-0.579	-0.051	0.040	0.607	8.045	1100
bas	-6.863	-0.589	-0.138	0.031	0.536	8.067	2571
SPI3							
adm	-5.241	-0.580	0.000	0.045	0.640	7.984	2654
bas	-4.987	-0.592	-0.073	0.038	0.586	8.142	3331
SPI6							
adm	-4.360	-0.577	0.024	0.040	0.647	6.977	5180
bas	-4.490	-0.599	-0.018	0.037	0.621	8.210	6005
SPI12							
adm	-5.726	-0.570	0.025	0.044	0.652	5.829	10233
bas	-4.917	-0.591	0.003	0.047	0.644	6.565	12059

Standardized Precipitation Evaporation Index (SPEI). The SPEI can be understood as an extension of the SPI in the sense that, in addition to precipitation, Potential Evaportranspiration (PET) is included in the calculation. Consequently, changes in the water balance due to varying temperatures, a major issue in the context of global warming, can be accounted for (Vicente-Serrano et al., 2010). For the calculation, the PET is subtracted from the precipitation values. PET data are obtained using the MxD16A2 products, which provide 8-Day estimates of PET (Steve Running et al., 2017a, 2017b). To construct a monthly data set, the 8-day composites are firstly divided by 8 to retrieve a specific value for each day of the year. Then, the daily data is aggregated to a monthly scale by taking the sum of daily values. Even though the data comes at a higher spatial resolution, it is processed on the resolution at which precipitation data from CHIRPS is available (i.e., 0.05°). After the subtraction of PET from precipitation, SPEI is calculated for 1, 3, 6, and 12 month reference periods as explained in the section above. PET is not available for large parts of the Sahara because the algorithm used to calculate it is not specified for desert areas. As a zonal statistic, the mean of all pixels within a district is extracted. The SPEI distribution between *adm* and *bas* districts is very similar, with value ranges between approximately -5 to 6 and mean values slightly above 0 (Table 15).

Table 15: Descriptive statistics of the Standardized Precipitation-Evapotranspiration Index (SPEI) based on different spatial aggregation units. (Unit of measurement: dimensionless)

Spatial Unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
SPEI1							
adm	-4.850	-0.601	0.007	0.033	0.648	5.938	973
bas	-4.566	-0.594	-0.012	0.047	0.652	6.388	2230
SPEI3							
adm	-4.218	-0.614	0.016	0.036	0.672	5.006	2651
bas	-3.841	-0.600	0.005	0.048	0.673	5.133	3310
SPEI6							
adm	-5.158	-0.616	0.020	0.033	0.668	4.426	5183
bas	-4.092	-0.614	0.016	0.044	0.684	5.131	6005
SPEI12							
adm	-3.607	-0.616	0.026	0.034	0.687	3.815	10246
bas	-3.446	-0.614	0.026	0.050	0.705	4.122	12059

Evapotranspiration (ET). This variable is extracted from the MxD16A2 product, as explained in the section above. However, in contrast to PET, the data is processed at a resolution of 0.01°. Evapotranspiration measures the amount of water that evaporates from the soil and plants to the atmosphere and is an essential variable in analyzing plant productivity, water consumption, and drought conditions (Senay et al., 2020). ET is not available for large parts of the Sahara because the algorithm used to calculate ET is not specified for desert areas. For every pixel, the total sum of ET is calculated per month. The average ET is extracted per district. The unit of measurement is kg/m^2. *bas* districts show a lower average ET compared to *adm* districts resulting in a difference of about 100 mm (Table 16).

Table 16: Descriptive statistics of evapotranspiration (ET) based on different spatial aggregation units. (Unit of measurement: kg/m^2)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	13	151.834	441.748	504.116	831.892	1724.832	644
bas	13	63.408	248.074	396.118	705.162	1751.375	39598

Land Surface Temperature (LST). Several studies have analyzed the effect of temperature on conflicts, e.g. for long-term projections (Witmer et al., 2017). Temperature is an essential climatic component that substantially influences the interaction between other measurable components such as ET and PET (Vicente-Serrano et al., 2010). LST is obtained by the MxD11CV monthly LST data sets, which provides temperature estimates for land areas during day and nighttime (Wan et al., 2015a, 2015b). Here, only daytime LST is used as a predictor. The data set is processed at a spatial resolution of 0.05°. It irregularly contains missing data concentrated at tropical locations due to persistent cloud coverage within an observation window. The unit of measurement is Kelvin. As a zonal statistic, the mean of all pixels within a district is extracted. The distribution between *bas* and *adm* districts is quite similar, with *bas* districts showing slightly higher daytime LST on average (Table 17).

Table 17: Descriptive statistics of land surface temperature (LST) based on different spatial aggregation units. (Unit of measurement: Kelvin)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	281.086	299.803	303.002	304.329	308.934	327.465	1597
bas	269.420	300.942	306.416	307.223	312.981	329.598	1766

Gross Primary Productivity (GPP). As a proxy for biomass production, GPP is included in this study. It measures the amount of carbon produced within in a grid cell. It is extracted using the MxD17A2H Gross Primary Productivity 8-Day data sets (Running et al., 2015a, 2015b). The same procedure to obtain a monthly set as with the ET variable explained above is conducted. For GPP, the monthly sum is calculated. Similar to PET and ET, there are missing data in the Sahara region because the production algorithm is not specified to calculate GPP over desert areas. The data is processed at a spatial resolution of 0.01° and is available at a monthly time-scale. The unit of measurement is kg C/m^2. On average, *adm* districts are characterized with higher GPP compared to *bas* districts (Table 18).

Table 18: Descriptive statistics of gross primary productivity (GPP) based on different spatial aggregation units. (Unit of measurement: kg C/m^2)

Spatial unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
adm	0	287.909	906.974	971.592	1573.883	3628.234	480
bas	0	166.240	543.187	777.121	1334.267	3474.603	39360

Interaction Variables. For environmental variables available at a monthly time scale, interaction variables with a binary mask for cropland are created. This interaction is expected to represent better productivity changes in the agricultural sector than the averaged variables for the entire district. Capturing these variables over time might translate into an accurate description of the agricultural sector's development in terms of productivity in- or decreases, crop failure, and drought stress. It should be noted that the cropland mask is relatively coarse in resolution. Small-scale agriculture or agroforestry systems are not likely to be captured due to spectral mixture with other land cover classes at these locations. For reasons of brevity, the descriptive statistics of these interaction variables are reported in the Appendix (Table A1).

3 Methods

3.1 Model Specifications

In the following section, the model specifications and the training and validation procedures are outlined. The core model of this thesis consists of a Convolutional-Long-Short-Term-Memory Neural Network (CNN-LSTM). However, as a baseline reference to asses the CNN-LSTM performance, a standard logistic regression model (LR) is also constructed. Starting from the LR model, relevant concepts for the establishment of the training architecture are introduced.

3.1.1 Logistic Regression

LR models have been the standard in conflict research for some time. One advantage of regression models is that the model outcome is easy to interpret by humans, which is of primary importance when research results are used to inform policy decisions. Recently, Halkia et al. (2020) published the methodology for the GCRI developed for the European Union Conflict Early Warning System which is based on LR. Their model outperforms or achieves comparable accuracy metrics compared to several conflict prediction tools based on more complex modeling procedures. This indicates that a LR model constitutes a viable choice to compare it with the results of more advanced modeling techniques such as neural networks. A binomial LR predicts the probability of an observation belonging to either one of two categories of a dichotomous dependent variable based on a number of independent variables following Equation (1).

$$P(Y) = \frac{e^{\gamma}}{1+e^{\gamma}}; \gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{1}$$

In this form, $P(Y)$ will take a value between 0 and 1 and represents the probability of a conflict occurrence. $x_1 ... x_n$ represent the predictor variables while $\beta_0 ... \beta_n$ represent the model coefficients to be fitted. From the equation, it becomes evident that logistic regressions assume a linear relationship between the predictors and the response variable. Secondly, LR is not intrinsically designed to handle a time axis, even though one could include time as an additional independent variable. It also means that LR can only produce one output at a time for each observation. If one wants to predict conflict for several months into the future, a specific model must be trained for each month in the prediction horizon.

Additionally, LR models are highly sensitive to class imbalances in the training data set.

Conflict prediction is a classification task with a very high class imbalance (Table 2). Various techniques exist to cope with class imbalance (Ali et al., 2015). One among them which has been previously used in conflict research is downsampling (Hegre et al., 2019). Here, before fitting the model, the majority class, which is represented by non-conflict district-months, is randomly downsampled to match the size of the minority class. This balanced subset is then used to fit the model parameters. Metric evaluations can be applied on a hold-out test data set, characterized by the original imbalanced distribution between conflict and non-conflict district months. The procedure to map the probabilistic model output to a binary classification follows Equation (2):

$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y) \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here, λ is a threshold value usually selected to be 0.5. However, specifically in modeling tasks with a high class imbalance such as conflict prediction, an optimal threshold might be searched for a given accuracy metric - a process referred to as threshold tuning (Zou et al., 2016).

3.1.2 CNN-LSTM

Basic Concepts of Neural Networks. Before explaining the fundamentals of CNN-LSTMs, a first step is to define the modeling task which might guide the reader through the descriptions to follow. In the present case, the modeling task is one of a multivariate time series prediction with a multi-step prediction horizon. The available training data set can be formally described as in Equation (3),

$$X_t^L = x_t^1, x_{t+1}^1 ..., x_{t-N+1}^1, ... \ ..., x_t^L, ..., x_{t-N+1}^L \tag{3}$$

where X_t^L is the predictor matrix with L predictors and t available timesteps. The training process then comprises the search for a function $f(X_t^L)$ which maps the input features to a multi-step prediction vector following Equation (4),

$$\hat{Y} = f(X_t^L) = \hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+h} \tag{4}$$

where $\hat{Y}$ is the predicted outcome vector of h time steps into the future. Commonly, the modeling function is trained on equal length inputs, such as the last 24 hours for energy price forecasting (Cordoni, 2020), the last week in the case of particle matter concentration forecasting (Li et al., 2020), or even several years worth of data in the case of solar irradiance

and photovoltaic power prediction (Rajagukguk et al., 2020). However, sophisticated deep learning models are also able to model variable-length inputs to variable-length outputs, i.e., in the case of language translation models (Yang et al., 2020).

The most fundamental concept in neural networks is that of a single neuron. This neuron receives some inputs X, learns a weight matrix W which then is used for a multiplicative interaction with the inputs, and an additional bias term b is added. The final output $\hat{Y}$ is obtained by passing the results of this interaction through a possibly non-linear activation function, also called activation function, here represented by a σ-function (5):

$$\hat{Y} = \sigma\Big(WX + b\Big). \tag{5}$$

These predicted outcomes can be used to calculate the error in comparison to the observed outcomes based on a specific loss function also referred to as cost function C. In a general notation, a given loss function calculates the loss between predicted and observed values which in itself is a function of the weights, the bias, the activation function and the inputs and observed values based on Equation (6):

$$Loss = C(Y - \hat{Y}) = C(Wb\sigma XY). \tag{6}$$

The loss is used to adjust the weight matrix W and the bias term b to more precisely match the expected outputs. This is achieved through a process called backpropagation. It is the process governing how a network *learns* to optimize the model function from Equation (4). To apply backpropagation, partial derivatives of the loss function in relation to its components are calculated from the output towards the inputs. This way, the gradient of change in the loss conditioned by the single components can be estimated. The gradients' calculation is not included in detail for brevity. However, a comprehensive explanation of the process can be found in Parr and Howard (2018).

In relation to the weight matrix W, the gradients are defined by the partial derivatives of the loss in relation to the single weights according to Equation (7):

$$\nabla C_W = \begin{bmatrix} \frac{\partial C}{\partial w_1} \\ \frac{\partial C}{\partial w_2} \\ \vdots \\ \frac{\partial C}{\partial w_n} \end{bmatrix} \tag{7}$$

The gradients are used to update the weight matrix in the direction the loss function is suspected to decrease most rapidly following Equation (8)

$$W^* = W - \eta \nabla C_W, \tag{8}$$

where η is a constant steering how much the weight matrix is adjusted in the gradients' direction referred to as learning rate, and W^* is the adjusted weight matrix. The same principle applies to the bias term. It should be noted that the process of backpropagation described here is simplified to a single-layer network. For multi-layer networks, the error terms have to be backpropagated from one layer to the next, in the direction from the output layer towards the input layer. However, the general concept remains the same. In this context, it is essential to differentiate between different training strategies which differ from each other mainly by the fact when gradients are calculated, and weights are adjusted during training. The first is called batch gradient descent. Here the gradients are calculated on based on all available observations. Once all training samples have been passed through the network, the cost function evaluates the loss and the errors are then backpropagated to adjust the weight matrix and bias terms. The second strategy is referred to as stochastic gradient descent or online learning, where a backpropagation takes place for every single sample. The last one is called mini-batch gradient descent and is the most widely used strategy. Based on a pre-defined batch size, the training data set is separated into equally sized batches. Backpropagation is then applied once a batch has been passed through the network. Additionally, several functions governing the adaptation process exist. They are referred to as optimizers and they mainly differ in the way the learning rate is adapted during the training process. An analysis of these differences is out of this thesis' scope, and the reader is referred to Sun et al. (2019) for a comprehensive overview.

Convolutional Neural Networks. In contrast to the simple neural network structure explained in the section above, CNNs work by applying a convolution kernel to the inputs. This operation effectively summarizes the input values based on a specific number of different kernels with a shared kernel width (Figure 3). This behavior of CNNs made them most attractive to tasks involving 2D-data such as image classification (Krizhevsky et al., 2017) or the analysis of remote sensing imagery (Song et al., 2019). Despite this traditional usage, CNNs successfully have been employed with 1D data structures with a time-component such as audio signals (Lee et al., 2009), activity detection and heart failure (Zheng et al., 2014) or stock price forecasting (Mehtab et al., 2020).

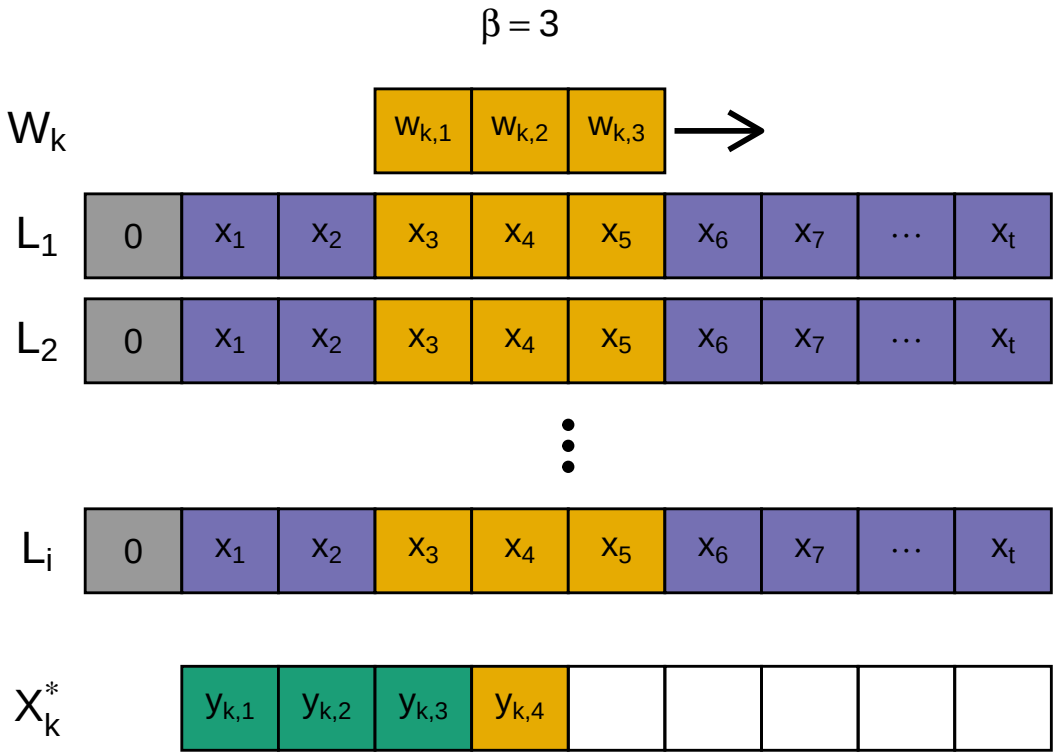


Figure 3: Scheme of a 1D convolution operation for a specific kernel k with width $\beta = 3$. Yellow squares indicate the current convolution at $t = 4$.

Figure 3 is an exemplary convolution operation for a specific kernel k with the kernel width $\beta = 3$. The input matrix X consists of L_i individual predictors with t time steps. The computation of the output involves sliding the kernel window W_k along the time axis. At the edges of the time series, there is not enough data for the convolution operation. One way to overcome this limitation is called zero-padding. A value of 0 is assumed for unavailable data, indicated by the gray boxes on the left. Another method, which would eventually alter

the size of the time axis, is to simply drop the observations for which no calculation with the kernel width β is possible. This behavior is leveraged in many applications to effectively reduce the size of the data sequence at higher abstraction levels within the network, e.g., in applications of CNNs in image recognition (Krizhevsky et al., 2017). The individual outputs $y_{k,t}$ are obtained by summing up the products of the inputs and the weights within the kernel window, indicated by the yellow boxes. The weights of W_k remain the same for one kernel but not across different kernels. Note that the result of the operation was denoted X_k^* to indicate it is an intermediate output, so one does not confuse it with the final network output $\hat{Y}$. Convolutional layers rarely represent the final layer in a network. It is prevalent to stack multiple convolutional layers on top of each other so that the outputs of the first are used as the inputs to the next (Rawat and Wang, 2017). For this reason, a given 1D convolutional layer at position l within the network and an input X^{l-1} is defined by Equation (9),

$$X_\beta^l = \sigma\Big(\sum_{i=1}^{L} X_i^{l-1} \cdot k_{i\beta}^l + b_{i\beta}^l\Big) \tag{9}$$

where k is the number of kernels, β indicates the size of the filter kernels, L is the number of input features in X^{l-1}, the bias is represented with b, σ is an activation function and $(\cdot)$ represents the convolutional operation explained in Figure 3. In practice, convolutional layers are often combined with so-called pooling layers, which combine the advantages of further reducing the dimensionality of the inputs and extracting latent patterns in the data (Rawat and Wang, 2017). The pooling layers exist in two favors, namely average and maximum pooling. It requires the specification of a pooling window, like the kernel window in the convolutional layer, which will then be applied over the input data to generate the average or maximum value of the observation window. A difference to the convolution kernel is that the pooling will be applied to each kernel individually and that most commonly, the pooling windows will be non-overlapping. Again, by using a zero-padding strategy, the shortening of the time series can be averted.

Long Short-Term Memory Network. The basic building block of LSTMs is recurrency. In deep learning, this is implemented by a simple recurrent cell which is defined by Equation (10).

$$Y_t = h_t; h_t = \sigma\Big(W_h h_{t-1} + W_x X_t + b\Big) \tag{10}$$

Here, the recurrent information, also referred to as hidden state, h_t is defined by the previous hidden state h_{t-1}, the current input X_t and the associated learnable weights W_h, W_x and the bias b. σ is an activation function. That way, recurrent networks have information from earlier time steps available when X_t is processed. However, long-term dependencies in the input sequence are not very well captured by this simple recurrent cell because of the exploding or vanishing gradient problem (Yu et al., 2019). To capture long-term dependencies in the data, Hochreiter and Schmidhuber (1997) proposed an extension to the simple recurrent cell referred to as Long Short-Term Memory cell, which was later modified to today's most common LSTM architecture by Gers et al. (2000).

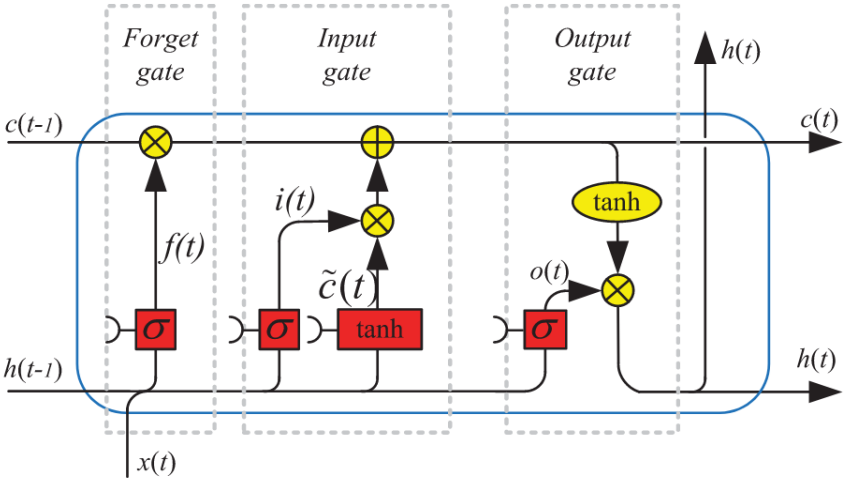


Figure 4: Scheme of a Long Short-Term Memory cell. (Source: Yu et al. (2019))

Figure 4 depicts the inner structure of a LSTM cell. The input data flows from left to right. There are two inputs to the cell, namely the inputs x_t as well as h_{t-1}, similar to the basic recurrent cell. The difference is found *within* the LSTM cell, where red boxes represent so-called gates which are functions with trainable parameters controlling the information flow inside the cell. The cell state c_{t-1} moves from left to right on the top of the box. At the first intersection, the cell state is updated by a point-wise multiplication with the result of $\sigma(h_{t-1}x_t)$, which is either 0 or 1 for specific locations. This gate governs which parts of the cell state are set to 0, which is why one refers to it as the forget gate $f(t)$. The second

gate is slightly more complex. First, another variant with individual weights of $\sigma(h_{t-1}x_t)$ is calculated. Then a point-wise multiplication with $\widetilde{c}(x_t)$ determines which new information is added linearly to the cell state. Because new information is added to the cell state, this gate is referred to as the input gate. At the final gate, referred to as the output gate $o(t)$, another multiplicative interaction with the updated cell state and the input determines the cell output h_t. Additionally, the new cell state c_t is an output, both of which will be used in the next step of an unfolded LSTM to process the input at x_{t+1}. Mathematically, such an LSTM cell is defined by the following equations:

$$\begin{aligned} f(t) &= \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f), \\ i(t) &= \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i), \\ \widetilde{c}(t) &= tanh(W_{\widetilde{c}h}h_{t-1} + W_{\widetilde{c}x}x_t + b_{\widetilde{c}}), \\ c(t) &= f(t) \cdot c_{t-1} + i(t) \cdot \widetilde{c}_t, \\ o(t) &= \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_0), \\ h_t &= o_t \cdot tanh(c_t). \end{aligned} \tag{11}$$

A common regularization strategy to reduce the tendency to overfit the training data is to include dropout layers after an LSTM layer. The dropout layer randomly silences a specific percentage of the neurons in a LSTM, thus directing each neuron's learning process towards learning more general features in the input sequence. Also, multiple LSTM layers can be stacked on top of each other, similar to CNNs, so that the first layer's output will be used as the input to the next. LSTMs have been reported to achieve considerable results in various problem fields for their capacity to capture both long and short-term dependencies. Researchers from Google achieved a high accuracy by using a LSTM in a sequence-to-sequence problem in machine translation (Wu et al., 2016). Other use cases are the prediction of stream flows in rivers (Hu et al., 2020), estimates of monthly rainfall (Chhetri et al., 2020), or predicting the occurrences of armed conflict in India (Hao et al., 2020).

Model architecture. The proposed model leverages both CNNs and LSTMs by combining them in a multi-input model. CNNs are used at the top of the model in order to yield representative feature map encodings of the high-dimensional conditions of a district and its spatial neighborhood. There are four parallel branches in the network for the model to pick up differences between a district and its buffer zones. Each of these branches processes the available predictors for the respective zone, i.e., the district or its buffers. Figure 5 shows one such branch as an example for the network architecture. Note that all four branches follow the same concept but that the specific architecture, i.e., the number of layers and neurons is determined during a hyperparameter optimization explained in the following section.

The first component of a branch consists of a 1D-CNN based on zero-padding that a second CNN layer can follow before the signal goes through a local pooling layer that averages or maximizes a sequence based on a temporal window determined by the pool size. Note that because the network is fed with variable-length input sequences, explained in detail in Section 3.4, a global pooling operation is necessary before the LSTM layers. For equal length input sequences, a flattening layer is typically used to flatten the time sequence to one dimension. With variable-length inputs, this would result in different output sizes, which the LSTM layers could not process. Thus, the global pooling layer reduces the input sequences to the number of kernels in the previous layer. This reduction is then repeated h times according to the desired prediction horizon and fed into a sequence of a maximum of three LSTM layers coupled with individual dropout layers.

Because this general network structure is applied to a total of four different inputs, the network will produce four distinct sequences. These sequences represent what the network has learned from the input data for each of the buffer zones. They are concatenated and put through a small, fully connected model with three layers to retrieve a single output. Each of these layers shares the same activation function and number of neurons. Since each neuron is fully connected to every neuron in the next layer, this part of the model is referred to as a fully connected model. Figure 6 shows the model structure from the concatenation of the four input sequences to the final output. The final layer only has one neuron but is time distributed so that it outputs a single value for every time step in the prediction horizon, which is set here to 12 months. For its activation function, it must output values between 0 and 1 so that it can be interpreted as a probability for the occurrence of conflict, which can be mapped to a specific prediction following Equation (2).

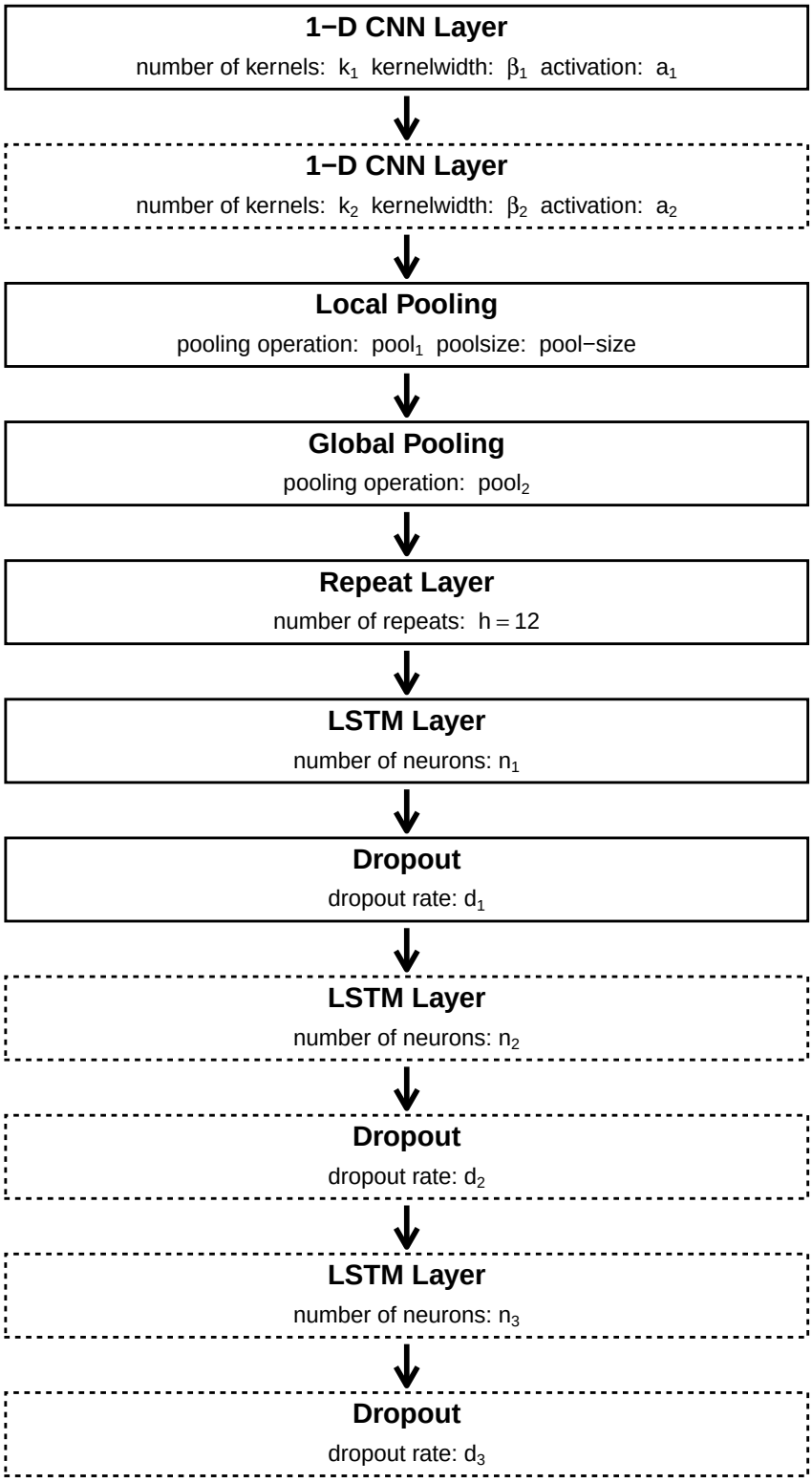


Figure 5: Proposed architecture of a single CNN-LSTM branch. Bold lines indicate mandatory layers, dashed lines indicate potential layers which are determined together with other parameters during a hyperparamter optimization process.

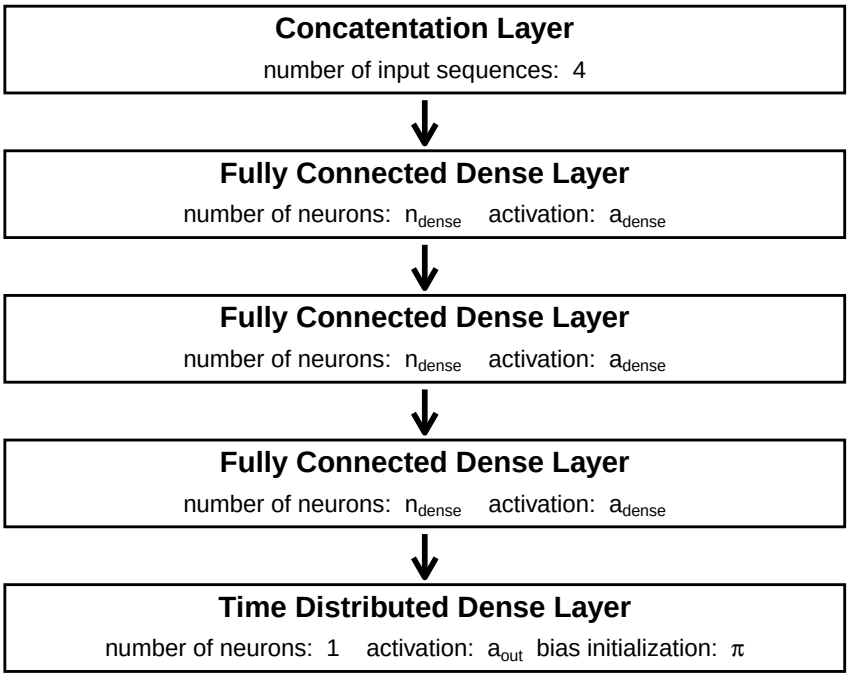


Figure 6: Proposed architechture of the fully connected output model.

As mentioned before, the outcome variable is characterized by a very high class imbalance (Table 2). In traditional approaches, researchers have counterbalanced this fact by artificially altering the distribution during training (Halkia et al., 2020; Schellens and Belyazid, 2020). Using downsampling approaches, however, means that there is a reduction in available training data. Because of the relatively short available sequences and a low number of observations, downsampling was not conducted when training the neural network. Instead, a specialized loss function was used to differentiate between easy-to-learn examples, referred to as the background class, and hard-to-classify examples also called foreground class. This function is called focal loss and was initially designed to detect rare objects in image segmentation tasks (Lin et al., 2018). As indicated, the focal loss down-weights the contribution of easy-to-classify examples to the overall loss, thus directing the training process towards optimizing for hard-to-classify examples. It is defined mathematically by Equation (12)

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{12}$$

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \, log(p_t),$$

where p is the probability estimation for an observation outputted by the model, α is a weighting factor that attributes different weights to the background and foreground class and

γ is a parameter governing the magnitude with which easy examples are down-weighted. The original authors state that this loss function has two beneficial properties for contexts with high class imbalance. The first is that for an observation wrongly classified as the background class, while p_t is small, the loss remains nearly unaffected because the modulating factor $(1-p_t)^\gamma$ tends towards 1. However, when $p_t \to 1$, the term goes towards 0, which means that the contribution of well-classified examples is weighted down. The second property is that the focusing parameter γ leans itself to adjusting the rate at which easy examples are down-weighted based on the problem at hand. When $\gamma = 0$, the loss is equal to cross-entropy loss, i.e., all examples contribute equally to the overall loss. Both α and γ are parameters that are optimized during the hyperparameters optimization stage. Additionally, the authors initiate the final output layer with a small value π, effectively reducing the probability the network will estimate the occurrence of the foreground class during the early stages of training. They report that this has positive impacts on training stability in high class imbalance scenarios (Lin et al., 2018). This initialization bias is also determined during the optimization stage.

3.2 Bayesian Hyperparameter Optimization

Hyperparameters are not directly involved in predicting a particular output, so they are often contrasted with model parameters. They substantially influence the overall training process and the accuracy of the predictions (Albahli et al., 2020). Table 19 summaries the notation of hyperparameters and the associated value ranges. Note that the branch-specific hyperparameters will be optimized for the four different branches in the network corresponding to the districts and the three buffer zones. Various strategies to apply hyperparameter optimization exist. Among the most widely used are grid search, random search, Bayesian Optimization (BO), and more recently the training of machine-learning models to predict the accuracy of different model configurations, also called meta-learning (Baik et al., 2020; Yu and Zhu, 2020).

Table 19: Overview of model hyperparameters.

Name	Description	Value Ranges
Branch specific hyperparameters		
$lstm\_layers$	Number of LSTM Layers	$1-3$
$double\_cnn$	Use of a second CNN layer	Yes, No
a_{cnn}	Activation function for CNN layers	$sigmoid, hard\_sigmoid, softmax, softplus, softsign$
k_{cnn}	Number of kernels in CNN layers	$12-128$
β_{cnn}	Kernel width for CNN layers	$3-24$
$pool_1$	Pooling operation for local pooling	$maximum, average$
$pool\_size$	Pool size for local pooling	$3-24$
$pool_2$	Pooling operation for global pooling	$maximum, average$
n_1	Number of neurons in first LSTM layer	$12-128$
d_1	Rate of dropout in first LSTM layer	$0-0.5$
n_2	Number of neurons in second LSTM layer	$12-128$
d_2	Rate of dropout in second LSTM layer	$0-0.5$
n_3	Number of neurons in third LSTM layer	$12-128$
d_3	Rate of dropout in third LSTM layer	$0-0.5$
Global hyperparameters		
a_{dense}	Activation of the dense model	$sigmoid, hard\_sigmoid, softmax, softplus, softsign, relu, elu, selu, tanh$
n_{dense}	Neurons per layer in the dense model	$12-128$
a_{out}	Activation of the output layer	$sigmoid, hard\_sigmoid, softmax$
π	Value of bias initialization of output layer	$0-1$
α	Alpha parameter of focal loss function	$0-1$
γ	Gamma parameter of focal loss function	$0-10$
$opti$	Optimizer function	$rmsprop, adam, adadelta, adagrad, adamax, sgd$
lr	Learning rate of the optimizer function	$1^{-10}-1$

General: Branch specific parameters are optimized individually for the 0/50/100/200 km input branches. Global parameters are optimized once per model.

Different approaches have in common what Shahriari et al. (2016) called "taking the human out of the loop." Hyperparameter tuning in this way can be understood as a process of reducing the impact of a researcher's subjectivity on the model construction towards the machine and the data controlling the training process. In its most extreme form, this thought leads towards machines being able to determine how they can learn a specific problem by themselves (Yao et al., 2019). The approaches, however, differ in complexity and the way they cope with the problem to balance between search time and accuracy. For example, grid search might yield equally high accuracies but the training time needed to achieve these can be very high compared with BO (Wu et al., 2019). While the meta-learning approach certainly is beyond this thesis's scope, an optimization strategy was searched with a reasonable balance between computing efficiency in terms of time and high accuracy. The choice was made to apply BO because it explicitly leverages prior information on the performance of hyperparameters to determine the next set to be explored and has a proven record of delivering robust results since its original publication in the 1970s (Mockus et al., 2014). In essence, BO works by iteratively updating the beliefs on the distribution of an accuracy metric for an objective function f based on the knowledge of prior samples of parameter x. The goal is to find the global maximum for x^+ within a pre-defined search space A following Equation (13)

$$x^+ = arg \max_{x \in A} f(x). \tag{13}$$

This is achieved by updating the prior probability $P(f(x))$ given data D to get the posterior probability $P(f(x)|D)$, which is referred to as the Bayes' theorem (Bayes and Price, 1763). Assume that we have accumulated a dataset $D_{1:t-1} = [(x_1, y_1) \dots (x_{t-1}, y_{t-1})]$ where x_1 is the value of a hyperparameter at trial $t = 1$ and y_1 is the result of the objective function $y_1 = f(x_1)$ which, in the case at hand, is the performance of the proposed model measured by a specific accuracy metric. This knowledge data can be queried by an acquisition function u to retrieve the next promising candidate x_t following Equation (14)

$$x_t = arg \max_{x} u(x|D_{1:t-1}) \tag{14}$$

With this candidate at hand the model is retrained to obtain an additional measurement on the performance. The knowledge data set is updated by $D_{1:t} = \{D_{1:t-1}, (x_t, y_t)\}$ and a new posterior probability for $f(x)$ can be estimated. BO works by calculating the probability density function for a given parameter x based on a Gaussian process. The details of these calculations are beyond this thesis's scope, but the reader is referred to Rasmussen and Williams (2006) for a comprehensive analysis of its application in machine learning. As

an acquisition function, the Upper Confidence Bound (UCB) function was used to calculate the next candidate parameter x_t. For a comprehensive overview of a BO process using UCB be referred to (Srinivas et al., 2012). It should be noted that BO, in its essence, is a sequential problem because the results of one iteration will impact the next. Even though there have been efforts to parallelize BO (Nomura, 2020), these approaches do not alter the basic sequential characteristic of the algorithm and come with higher management costs for the researcher.

3.3 Performance Metrics

The performance of a model is validated on a specific set of accuracy metrics. However, performance is often a multi-dimensional problem, which is why several metrics are used in this thesis. These were selected to represent different dimensions of a model's performance to differentiate between the occurrence of conflicts versus peace.

In its essence, the problem of this thesis is one of a binary classification. The core component for the performance assessment of a binary classification problem is a simple two-class confusion matrix depicted in Table 20 (Tharwat, 2020).

Table 20: Concept of a binary confusion matrix.

	Observation	
	Positives	*Negatives*
Prediction		
Positives	True Positives (TP)	False Positives (FP)
Negatives	False Negatives (FN)	True Negatives (TN)

From the table above, it is evident that there are two types of errors. False Positives (FP), also referred to as Type I error, are observations that were falsely classified as the positive class while in reality they belong to the negative class. False Negatives (FN), referred to as the Type II error, are observations that were predicted as the negative class, however, they belong to the positive class (Tharwat, 2020). The most widely used accuracy metric derived from a confusion matrix is Overall Accuracy (OA). It is simply calculated as the rate

of correctly classified observations (Equation (15))

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \ . \tag{15}$$

However, OA is not very well suited for classification problems with high class imbalance (Tharwat, 2020). A model for a classification problem where the positive class only covers 1 % of the observations can achieve an OA of 99 % only by always predicting the negative class. In the context of class imbalance, other metrics derived from the confusion matrix help to paint a more concise picture of a model's capability to predict a particular outcome. The rate at which positive examples are correctly classified (True Positive Rate), also called sensitivity or recall, sheds light on a models capability to correctly identify positive observations (Equation (16))

$$Sensitivity = TPR = \frac{TP}{TP + FN} \ . \tag{16}$$

The False Positive Rate (FPR) contains information on the rate a model falsely predicts a positive outcome in relation to all observations with a negative outcome thus indicating the rate negative observations are treated as positives (Equation (17))

$$FPR = \frac{FP}{FP + TN} \ . \tag{17}$$

Shifting the focus towards the negative class, specificity also called True Negative Rate (TNR) describes the rate negative observations are correctly identified (Equation (18))

$$Specificity = TNR = \frac{TN}{TN + FP} \ . \tag{18}$$

The False Negative Rate (FNR) then describes the rate that positive observations are falsely classified as the negative class (Equation (19))

$$FNR = \frac{FN}{FN + TP} \ . \tag{19}$$

In addition to these metrics, the precision of a model, also referred to as Positive Predictive Value (PPV) captures the rate examples classified as the positive class actually represent positive observations (Equation (20))

$$Precision = PPV = \frac{TP}{TP + FP} \ . \tag{20}$$

Note that sensitivity and precision always are in a fragile balance with each other. While higher sensitivity values can be achieved by decreasing the number of false negatives, there naturally will be a higher number of false positives, and the precision is decreased. The same

holds if one tries to increase the precision, which will lead to lower sensitivity values. To harmonize this relationship into a single metric the F_β-score is used (Equation (21))

$$F_\beta = (1 + \beta^2)\frac{Precision * Sensitivity}{\beta^2 * Precision + Sensitivity} \ . \tag{21}$$

The most widely used is $\beta = 1$, which will result in a harmonic mean between precision and sensitivity, also referred to as F_1-score (Tharwat, 2020). With increasing values of β, sensitivity is emphasized over precision. $\beta = 2$ will put the double weight of sensitivity compared to precision. This metric, referred to as F_2-score, is the central metric for this thesis. It was chosen because a model's capability to not miss out on the occurrence of conflict district-months was considered more important than a model's tendency towards so-called false alarms, i.e., the prediction of conflict when in reality, there was peace. If this model was used in conflict prevention or crisis early warning systems, missing out on a potential conflict can lead to the loss of lives. It is therefore considered advantageous when a model can correctly detect positive examples at a high rate. However, using the F_2-score as the central optimization metric ensures that the precision of a model still influences its performance evaluation so that frequently predicting the positive class itself would not lead to very high performance scores.

Two additional metrics which can not directly be derived from the confusion matrix are included as well. These are the Area Under the Receiver Operating Characteristic (AUROC - AUC for short) as well as the Area Under the Precision-Recall Curve (AUPRC - AUPR for short) (Fawcett, 2006). These metrics represent the relative tradeoffs between sensitivity and FPR, and precision and sensitivity, respectively. To get an intuition about the calculation, one can imagine that the threshold value to map a model's output to a specific class prediction steadily increases from 0 to 1. In the former case, for each threshold, the values for sensitivity and FPR are recorded. In the latter case, precision and sensitivity are the metrics of interest. In reality, a more efficient algorithm is used to calculate these metrics (Fawcett, 2006). Both metrics are then plotted against each other. The area under theses curves generalizes the performance of a classifier into a single metric. The generated plots of both the ROC and the PRC, can be used to compare the performance of different classifiers visually (Fawcett, 2006).

3.4 Training & Validation Process

Recall that the training data is available as a sequence of L predictors of t time steps in the form of X_t^L from Equation (3). In total, there are $t = 228$ time steps available from January 2001 to December 2019. The total number of predictors is $L = 176$, but only a subset is fed to the respective models, except for the regression baseline and the environmental models which are trained on the complete data set. The prediction horizon is set to 12 months so that for each x_t there is an associated outcome vector of the form $y = [y_{t+1}, y_{t+2}, ..., y_{t+12}]$.

In order to be able to evaluate the potential of a model to generalize, out-of-sample evaluation data sets are needed. The available data is split into three different sets, which are used for training, validation, and testing according to Table 21.

Table 21: Split of the available data to training, validation and testing data sets.

Name	Purpose	Range
Training	Fit model parameters	Jan. 2001 - Dec. 2016
Validation	Fit model hyperparameters	Jan. 2017 - Dec. 2018
Test	Performance estimation	Jan. 2019 - Dec. 2019

For the LR baseline, all available predictors are used. Because no hyperparameter tuning is involved in fitting the model, the training and validation data sets are combined to estimate the model parameters. Undersampling is performed so that an equal amount of observations with conflict and peace district-months is included. While all conflict observations are included, peace observations are randomly sampled. This process is repeated ten times, each with different selected peace observations. For each month in the prediction horizon, models are trained individually. Time is not included as a predictor variable, so that the model predictions e.g. for six month into the future are based solely on the predictors at time step t: $\hat{y}_{t+6} = f(x_t)$. The predictor variables are normalized based on the distribution in the combined training and validation set. District-months with missing data are dropped for the logistic regression model. Threshold tuning on the model's probability predictions is applied in the sense that the threshold is chosen which delivers the best F_2-score on the training data. With this threshold, the performance metrics are evaluated on the test set and averaged across the ten repeats. The average of the monthly performance metrics is calculated and referred to as the global performance.

Handling the data for training the neural networks differs considerably compared to the LR baseline. For all neural network models, the data is handed over as a time series with increasing length and a minimum length of 48 months. Batch gradient descent was chosen as a training strategy, meaning that the first step in an epoch of training a neural network consists of 48 months worth of data for all the districts. Then gradients are calculated, and the learnable model parameters are updated. The next step in an epoch then consists of training data worth of $t = 48\,months + step\_size$ time steps for all available districts. During hyperparameter optimization, a $step\_size = 4$ was chosen, effectively reducing the size of the training data set to $\frac{1}{4}$. During the models' final training, $step\_size = 1$ was chosen so that the complete data set is used to estimate the final model parameters. One epoch of training is completed once all time steps have been presented to the model. Performance on the validation set is calculated at the end of an epoch. A global termination criterion for training is implemented for all models, once 200 epochs have been trained. This is combined with different early stopping policies for the optimization and the final training stage, which will lead to varying numbers of epochs from one model to another. Similar to the logistic regression model, the predictor variables are normalized based on the distribution in the training data set. Missing values, in contrast, are not dropped but imputed with a value of -1. Due to normalization, valid values are in the range of $[0, 1]$, so replacing missing data with a constant value represents a valid training strategy for deep learning frameworks (Ipsen et al., 2020). Also, missing values mainly occurred systematically in the present data set. For example, values of evapotranspiration are systematically missing for the Sahara region. Other imputation strategies, such as mean imputation of interpolation between the last and next observed value in a time series, cannot deliver robust results in such cases. Also, a comparable pattern of missingness is expected in the test set, and missing values are expected to occur not in relation to the outcome variable but due to the characteristics governing their respective generating process.

Hyperparameter optimization for the neural networks is performed only once with the outcome variable **cb** to minimize computation time. Even though it can be expected that the optimal configurations of the network architecture might differ for the outcome variables, this simplification drastically reduces training time. Additionally, the internal structure of the data does not change with the outcome variable, so that it can be expected that this one-to-serve-all approach will lead to satisfactory performance with other outcome variables. For both aggregation units, *adm* and *bas*, and the different predictor sets (CH, SV, EV) opti-

mization is conducted individually, resulting in a total number of six optimization processes. For one optimization process, the first 100 trials are conducted with random initialization of the hyperparameters. These first trials serve as the knowledge base to estimate the performance of yet-to-be-tested parameters. Another iteration of 100 trials are then used to explore and optimize the parameter space. During this stage, the training set is used to optimize model parameters. The F_2-score is evaluated on the validation data set at the end of each epoch. During training, threshold tuning is not applied, which is why $\lambda = 0.5$ is used as the cut-off value to calculate the F_2-score. Early stopping was included when the F_2-score on the validation data set does not improve for ten epochs above a threshold of 0.001.

The final training process is conducted for each predictor set on each of the four outcome variables for both aggregation units. The model is set up with the optimal hyperparameters found during the optimization stage. There are two steps involved in the final training stage. The first step consists of training the model parameters on the training set with a slightly changed early stopping policy on evaluating the validation set. Here, training is stopped after 20 epochs of no improvement in the F_2-score above a threshold of 0.0001. The policy is slightly changed because firstly, more training data is available, and secondly, the hyperparameters have been optimized for the response variable **cb** only. To account for possible variations during training, the stopping criterion is relaxed so that a higher number of training epochs are possible. Once training has been completed, threshold tuning is applied based on the validation data set to find the threshold which optimizes the F_2-score. During the second step, the model parameters are trained on the validation set, which has been held out during the first step. Because there is no other independent data set to validate this training process against, the early stopping criterion is set to a decrease in the overall loss smaller than 0.0001 within 10 epochs. After the second training step, the model performance is evaluated on the test set. For this, the model's estimation for the probability of conflict is predicted, then the optimized threshold from the first training step is applied, and the performance metrics are recorded. Because the weights of different layers are randomly initialized, the prediction results can substantially differ when training is repeated. To account for this variability, training is repeated 10 times on each predictor set for both aggregation units and all outcome variables, resulting in a total number of 240 distinct DL models (10 repeats x 4 outcome classes x 3 predictor sets x 2 aggregation units).

3.5 Analysis of Variance

Leveraging the availability of ten repeats per outcome variable and predictor set a two-way analysis of variance (ANOVA) is used to derive statistical indications to answer hypothesis H1 and H2. ANOVA is used to analyze the impact of different treatment factors on the outcome of an experiment. The four different model and predictor themes (LR, CH, SV, and EV) in combination with the spatial representation (*adm* and *bas*) are conceptualized as two different levels of treatment. The outcome of the experiment is measured by the achieved global F_2-score. Three assumptions must hold for the classical Fisher's ANOVA (Fisher, 1921). The first is the independence of observations between and within groups of treatment. This independence is given since the results of one predictor set do not impact the results of another. The ten repeats of the DL models for a given predictor-unit combination are independent from each other because instead of a cross-validation the total available data set is used for each repeat (Raschka, 2020). For the LR model, downsampling is conducted randomly so that for each repeat peace observations have the same probability to be included during training. The second assumption is that residuals are normally distributed and thirdly homogeneity of variance is assumed. Visualizations of the residuals are used to assess if the assumptions are violated. While the assumption of normal distribution holds (Figure A20), the assumptions on homogeneous variance is slightly violated (Figure A21). For that reason, instead of the classical ANOVA, the Welch-James test is applied (James, 1951; Welch, 1951) which tests for H_0 that no statistical significant difference in the mean of the outcome is observed between treatments. It is a non-parametric test and thus can be applied for cases where variance is not homogeneous. When H_0 can be refused, usually post-hoc tests are applied to investigate the differences between groups. For cases where two or more treatment groups are tested against, the test is specified by terms describing the individual group's contributions (main effect) as well as the interaction terms between groups (interaction effect). When interaction effects are significant, main effects are excluded from further analysis. The Games-Howell test is applied as a post-hoc test. It can be applied when six or more observations for each group are present (Lee and Lee, 2018). It delivers an estimate of the absolute difference between treatment groups and reports on the statistical significance of these differences. The results thus allow a detailed comparison of the achieved performances for different predictor-unit combinations based on statistical significance and thus play a central role in the confirmation of hypothesis H1 and H2.

4 Results and Discussion

4.1 Hyperparameter Tuning

The results of the BO process are reported in Table 22. For each of the predictor sets, an optimization was conducted for both the *adm* and *bas* units. Except for the SV-*bas* model, a double-layered CNN was found to deliver the best results. Concerning the activation function a_{cnn}, the picture is less clear, but *softmax* was selected in $\frac{8}{24}$ of the cases followed by the *hard_sigmoid* activation function ($\frac{7}{24}$). The kernel numbers k_{cnn} are selected only twice above a value of 120, approximating the maximum possible value of 128. The size of the network generally seems sufficient to capture the data structure. The selected kernel widths β_{cnn} for the CH-*bas* model are at the maximum of 24 months for both the 50 and 100 km branch, suggesting that for this variable configuration, there could be a performance improvement with bigger kernel widths. For the remaining configurations, the kernel widths are mainly between five to 22 months. For parameter $pool_1$, in about half the cases maximum pooling is selected with *pool_size* between 3 and 24 and only in $\frac{6}{24}$ cases the pool size is below 12. Generally, information beyond a period of 12 months seems to represent the signal for conflicts better. As pooling operation $pool_2$, maximum pooling is selected in $\frac{17}{24}$ cases, indicating that maximizing the values through the time series is beneficial for the conflict prediction task. It is of interest that average pooling was only chosen for 100 and 200 km inputs. The signal from these inputs seems to be better represented by an average layer.

For most variable configurations, the selection of two LSTM-layers yielded the best results, except for the CH-*bas* and EV-*adm* configurations. Similar to the CNN part of the model, the results suggest that the sizes of $n_{1:3}$ generally seem to be sufficient for the data complexity. Only in $\frac{7}{72}$ cases more than 120 neurons are selected for a LSTM layer. Instead, the number of neurons is frequently below 100, indicating that smaller network sizes are able to learn a pattern from the data. There is no clear pattern concerning the number of neurons across variable configurations or between layers. Also, dropout rates $d_{1:3}$ vary substantially, with the lowest rate of 1 % to the highest of 50 % dropout. As the activation function a_{dense} for the fully connected model, *softplus* and *relu* both are selected twice. The number of neurons n_{dense} is between 21 and 97 neurons. The maximum of 128 is not approximated by any of the model configurations.

Table 22: Results of the Bayesian hyperparameter optimization.

	Conflict History (CH)		**Structural Variables (SV)**		**Environmental Variables (EV)**	
	adm	*bas*	*adm*	*bas*	*adm*	*bas*
double_cnn	*Yes/Yes/Yes/Yes*	*Yes/Yes/Yes/Yes*	*Yes/Yes/Yes/Yes*	*No/No/No/No*	*Yes/Yes/Yes/Yes*	*Yes/Yes/Yes/Yes*
a_{cnn}	*softsign/hard_sigmoid*	*softsign/hard_sigmoid*	*softmax/softmax*	*sigmoid/softplus*	*hard_sigmoid/softmax*	*softmax/softmax*
	softsign/softmax	*sigmoid/softsign*	*hard_sigmoid/softmax*	*softplus/softmax*	*hard_sigmoid/hard_sigmoid*	*hard_sigmoid/softmax*
k_{cnn}	106/125/45/78	86/99/39/128	95/70/63/42	41/102/67/43	19/41/94/99	92/76/66/42
β_{cnn}	5/10/22/8	19/24/24/6	10/9/9/16	22/12/14/5	12/6/13/14	8/10/10/17
$pool_1$	*max/max/avg./max*	*avg./max/max/max*	*max/avg./avg./max*	*avg./max/avg./avg.*	*max/avg./max/avg.*	*max/avg./avg./max*
pool_size	15/23/3/14	24/19/10/21	12/18/17/18	16/6/10/9	12/16/9/22	12/18/17/16
$pool_2$	*max/max/max/avg.*	*max/max/avg./avg.*	*max/max/max/avg.*	*max/max/avg./max*	*max/max/max/avg.*	*max/max/max/avg.*
lstm_layers	2/2/2/2	3/3/3/3	2/2/2/2	2/2/2/2	3/3/3/3	2/2/2/2
n_1	114/83/114/106	97/12/12/78	109/128/43/23	23/59/38/92	88/28/85/122	107/127/38/18
d_1	0.22/0.08/0.14/0.49	0.23/0.2/0.38/0.25	0.02/0.09/0.07/0.06	0.22/0.14/0.02/0.2	0.2/0.1/0.09/0.31	0.02/0.07/0.05/0.06
n_2	79/53/62/123	128/34/84/65	83/82/128/21	37/114/83/56	111/19/13/104	76/80/128/21
d_2	0.05/0.28/0.37/0.08	0.5/0.24/0.05/0.16	0.34/0.03/0/0.29	0.26/0.12/0.46/0.47	0.23/0.04/0.09/0.49	0.34/0.02/0.01/0.31
n_3	-/-/-/-	85/92/96/57	-/-/-/-	-/-/-/-	54/109/106/13	-/-/-/-
d_3	-/-/-/-	0.437/0.231/0.255/0.172	-/-/-/-	-/-/-/-	0.202/0.477/0.284/0.285	-/-/-/-
a_{dense}	*softplus*	*elu*	*relu*	*softplus*	*selu*	*relu*
n_{dense}	21	97	32	38	95	29
a_{out}	*sigmoid*	*sigmoid*	*sigmoid*	*sigmoid*	*sigmoid*	*hard_sigmoid*
π	0.4529	0.6534	0.2404	0.3856	0.5697	0.275
α	0.9245	0.7928	0.8056	0.9087	0.722	0.8544
γ	6.2896	5.8844	7.8011	1.1974	4.8053	8.1588
opti	*adagrad*	*adam*	*adamax*	*adagrad*	*adamax*	*adamax*
lr	0.0244	0.0116	0.0246	0.0258	0.0117	0.0226

General: Multiple values indicate the results for the input branch of buffer size 0/50/100/200 km, respecitvley.

For the output layer activation a_{out}, *sigmoid* is selected in $\frac{5}{6}$ cases, *hard_sigmoid* is selected only once. This can be expected since *sigmoid* is usually the choice for probabilistic model output whereas *hard_sigmoid* is faster to compute but also less accurate. The results of the bias initialization show that the lowest initialization value selected is $\pi = 0.2404$. This result seems to be counter-intuitive since the high class imbalance in the data set the layer should be initialized with relatively small values (Lin et al., 2018). However, during training, the cut-off threshold to differentiate between conflict and peace district-months was held constant at a value of 0.5, thus $\frac{4}{6}$ models are still initialized with a bias towards the background class. The weight factor α is comparable between model configurations between 0.722 to 0.925, close to the highest possible value of 1. It thus effectively downweighs the contribution of the frequent peace observations during the calculation of the focal loss. The parameter γ takes comparable values between 4.8053 and 8.1588 except for the variable configuration SV-*bas*. Here, with $\gamma = 1.1974$, an exceptionally low value is selected, leading to higher contributions of correctly identified observations to the overall loss compared to the other models. As an optimizer, *adam* or derivation of it are selected. In $\frac{3}{6}$ cases, it is the *adamax* optimizer. This indicates that optimizers that adapt the learning rate during a training process and keep information on prior gradient calculations, called momentum, are beneficial for the problem at hand. Initiating the learning rate between values of 0.0116 to 0.0258 delivers the best results across all model configurations. Overall, BO seems to be able to determine well performing model configurations for different representations of the data. It should be kept in mind that there are some limitations in the selected approach because BO was only conducted once on the outcome variable **cb**. However, especially the parameters of the focal loss function that substantially influence a model's ability to learn from the data are expected to be sensitive to the specific class distribution per outcome class. It can be expected that repeating the BO for each outcome class would lead to better performance.

BO and hyper-parameter tuning in general are not frequently conducted in conflict prediction studies. The main reason is that linear models seldom require hyper-parameters to be tuned. But even for more complex model algorithms, tuning is often omitted. Hegre et al. (2019) do not optimize the hyper-parameters of their Random Forest models but hold them static at values found empirically that produce accurate predictions. This is also true for the model of Kuzma et al. (2020), where the selection of hyper-parameters for their Random Forest model was guided by manual search. Schellens and Belyazid (2020) report for both their neural network and the Random Forest model that hyper-parameters were delineated using

"rules-of-thumb" (Schellens and Belyazid, 2020, p. 4), i.e., the parameters were set in advance based on the size of the training data set, and no search was implemented. For Random Forest, the tuneable parameters are only a few, comprising the number of trees and the number of variables to be considered at node splits (Breiman, 2001). It should be noted that Random Forest can deliver robust results with low sensitivity to its hyper-parameters and that e.g., increasing the number of trees is not always beneficial, especially for classification tasks (Probst and Boulesteix, 2018). For DL models, the number of tuneable parameters is substantially higher. Experimental studies from other domains have shown that model performance is very sensitive to their initialization rendering hyper-parameter optimization a necessity (Cooney et al., 2020; Zhang and Wallace, 2016).

4.2 Global Performance

The following considerations are focused on the F_2-score, the AUPR metric, precision and sensitivity for reasons of brevity. Visualizations of additional accuracy metrics as well as a comprehensive table comparing these metrics for model specifications are listed in the Appendix (Figure A9 and Table A2). The F_2-score serves as the primary performance metric in this thesis. From Figure 7, it becomes evident that framing the detection problem as a time series and applying a DL framework substantially increases the F_2-score compared to the LR baseline. However, the LR baseline based on the *bas* districts achieves higher scores for all conflict classes compared to *adm*-LR, showing that aggregation on the basis *bas* of districts is beneficial for the prediction. The score gains from the LR baseline to the CH theme are considerable for both aggregation units. For the outcome classes **sb** and **ns**, both aggregation units show an increased variance for the 10 repeats in the CH set, but for *bas* districts this variance is substantially larger compared to the *adm* districts. The results suggest that the history of recent conflicts as the sole predictor already is able to achieve better performance compared to the fully specified LR baseline. However, the performance seems not to be very stable but dependent on the random initialization of model parameters, especially for *bas* models. The performance on the outcome classes **cb** and **sb** increases with more complex predictor sets for both aggregation units. In general, the absolute gains in performance between the DL models are only marginal. For the outcome class **cb** the F_2-score raises from 0.590 (0.664) with the CH predictor set, to 0.616 (0.649) with the SV set to to 0.618 (0.689) with the EV predictor set for the *adm* (*bas*) districts, resulting in an absolute gain of 0.028 (0.025) points (Table A2).

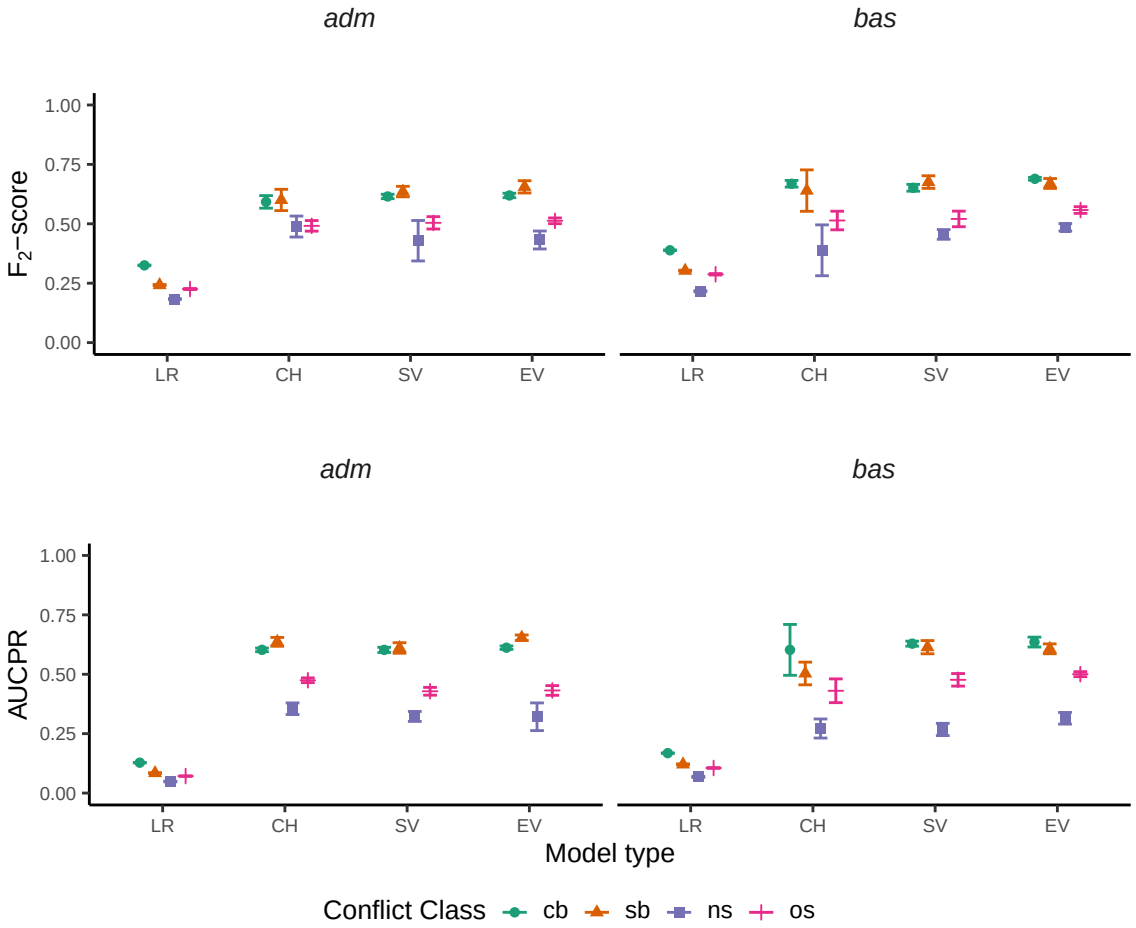


Figure 7: Global performance of the F2-score (top) and AUCPR metric (bottom). (LR: Logistic Regression, CH: Conflict History, SV: Structural Variables, EV: Environmental Variables)

For the two other conflict classes, **ns** and **os**, the picture is quite different. The first observation is that the absolute scores achieved for these classes are substantially lower. Partly, this can be explained with the lower occurrence of these classes in the data set (Table 2). As a second observation, there is evidence that for *bas* districts for both classes the performance is improving with more complex predictors (Figure 7). Considering the **ns** class, the F_2-score raises about 0.112 points from the CH to the EV theme. For the **os** conflict class, the increase is less pronounced with 0.053 points. In comparison, the performance for *adm* districts is decreasing with more complex predictor sets for outcome class **ns**. The difference between the CH and EV set is -0.029. For the **os** outcome variable, there is an increase in the F_2-score. With an absolute gain of 0.022 points from the CH to EV theme, this gain is only about half the size observed with *bas* units. These findings indicate that for the **ns** and **os** conflict classes, environmental variables based on *bas* districts play a major role in improving

the predictive performance of the DL networks. An attempt to explain this could be the nature of these conflicts. Conflicts between non-state actors might include two groups with clashing socio-economic interests contesting over the available (natural) resources (Sundberg et al., 2012) while **os** events of violence often are associated with singular terrorist acts against the state or civilians (Eck and Hultman, 2007). Both types of conflict come with high opportunity costs for the actors inducing the violence. In contrast, state-based violence in an already conflict-prone area shows lower opportunity costs for the state actor. In its own logic, a state has to defend itself against rebels or insurgencies to maintain its legitimate power position (Collier, 1998). Also, the availability of resources to engage in warfare is usually higher for the state than for other combating actors, further reducing opportunity costs. Thus, **ns** and **os** conflict classes might show a stronger relationship to natural resources. Changes in resource availability substantially reduce the opportunity costs for deprived individuals to get involved in violent actions when their livelihoods are seriously endangered.

Considering the AUPR values, the finding that DL models substantially increase performance over the LR baseline is confirmed (Figure 7). The larger variance of the CH-*bas* model becomes even more evident compared to the results of the F_2-score. Also, the tendency for *bas* districts to achieve higher performances with more complex predictor sets is more pronounced for all outcome variables. On the other hand, decreasing or stable AUPR values are observed for *adm* districts. For both, **ns** and **os** conflict classes, a substantial decline from the CH set to the EV set is observed, accounting for -0.064 and -0.041 points, respectively. In contrast, for *bas* districts and the **os** outcome class, the increase in AUPR from the CH to the EV theme amounts to 0.08 points. For the **ns** class, an increase of 0.039 points is observed. The increases for the **cb** and **sb** class are even more substantial (0.066 and 0.098, respectively). This underlines the finding that more complex predictor sets do not substantially benefit the performance of DL models when combined with *adm* districts. The selected predictors, however, do increase the performance in combination with *bas* districts, indicating that the environmental processes happening on the ground are better captured at the sub-basin watershed level.

Comparing the performance to other conflict prediction studies is complicated because (i) some studies do not directly report selected metrics, and (ii) the data sets and definitions used to identify conflicts can be different. For example, the F_2-score is seldom reported, but

it can be directly calculated for cases where the sensitivity and precision metrics are given. It should be noted that differences in the definition of conflict classes can not be accounted for directly. Hegre et al. (2019) reach an F_2-score of about 0.86 for the **sb** outcome variable on the African continent based on a country-month analysis which is substantially higher than the maximum F_2-score of 0.67 for the **sb** class achieved by the SV-*adm* model in the present analysis. They also model conflict on a regular grid with a cell size of 0.5 x 0.5 decimal degrees. For this representation, the F_2-score only reaches about 0.36, underlining the rigorous impact different aggregation units have on model performance. The F_2-scores in this study lie in between the values achieved by Hegre et al. (2019) on large and small-scale aggregation units but deliver greater spatial detail than their country-month analysis. The comparison reveals a trade-off between predictive performance and spatial detail, which needs to be considered when evaluating conflict prediction models.

Kuzma et al. (2020), on the other hand, use *adm* districts across Africa and Asia for their prediction model. They use a subtle difference in their definition of conflict by applying a moving-window of 12 months into the future and setting a threshold of 10 or more casualties to consider a given district-month as conflict. This way, they are able to obtain a F_2-score of 0.70 for African districts (0.74 overall). Given that the **cb** outcome class achieves a maximum F_2-score of 0.69 and that similar aggregation districts were used, the present study's overall performance is comparable. Halkia et al. (2020) are able to achieve an F_2-score of 0.79 based on a global country-month data set, indicating that higher performances can be achieved with simple linear regression models for a loss in spatial detail. They also use a slightly different definition of conflict by combining different types of conflicts into two classes representing sub-national conflicts and national power conflicts. While they do not report AUPR values, they are able to achieve AUC values of 0.94 for both classes. This performance is consistent with the models presented here, where AUC values between 0.944 and 0.963 are achieved for different conflict classes (Table A2). Overall AUPR scores for the complete study domain are reported by Kuzma et al. (2020) accounting for a score of 0.42. Hegre et al. (2019) report an AUPR score of 0.869 for **sb** conflicts based on country-months and of 0.277 based on the grid representation. In the current study, for the **sb** outcome class, a maximum AUPR score of 0.65 is achieved, while for the **cb**, the score is 0.639. In terms of AUPR, this study achieves better performances compared to Kuzma et al. (2020) who use similar aggregation units. In comparison to Hegre et al. (2019), the results are again in between the different aggregation units used in their study, indicating losses in performance

for increased spatial detail.

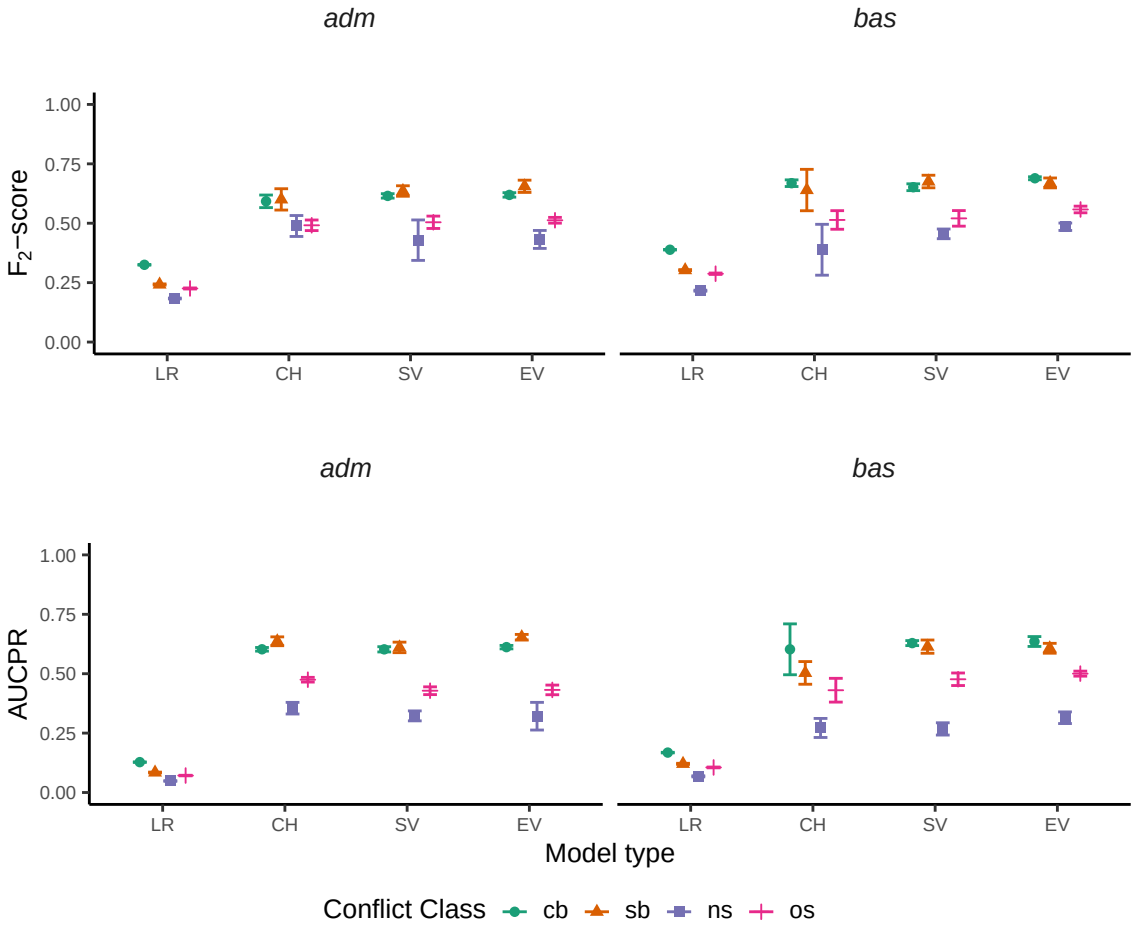


Figure 8: Global performance of precision (top) and sensitivity (bottom). (LR: Logistic Regression, CH: Conflict History, SV: Structural Variables, EV: Environmental Variables)

Both of the performance metrics previously discussed are a combination of precision and sensitivity into a single metric. The F_2-score puts more weight on sensitivity while AUPR provides a more balanced assessment. As presented in Section 3.3, precision and sensitivity behave in a fragile balance to each other. Given a specific prediction model with fitted parameters, increases in precision will decrease sensitivity and *vice versa*. In Figure 8 both of these metrics are presented. For *adm* districts, there is a substantial increase in precision from the CH to SV set. Simultaneously, this increase is associated with a decrease in sensitivity most pronounced for the **ns** outcome class. Moving to the EV set, precision is slightly reduced, and some moderate sensitivity gains are observed. This behavior indicates the EV set achieves a more balanced distribution compared to other predictor sets compromising between precision and sensitivity. In general, one would expect a simultaneous increase of both performance

metrics for a model with greater predictive ability. This is not found in the data at hand, rather more complex predictor sets achieve a more sensible balance between the metrics. The same observation holds for *bas* districts, except that the relation between sensitivity and precision is reversed. Moving from the CH to the SV set, sensitivity is increased at the cost of precision. The EV set again seems to settle on a sensible balance with slight increases in precision and decreases in precision compared to the SV set. Comparing the EV set for both aggregation units reveals that while the precision is comparable, *bas* districts are associated with higher sensitivity values compared to *adm* districts except for outcome variable **os**. This indicates that on the basis of a comparable precision rate, the models based on *bas* districts are more able to flag observed conflict district-months. Concerning the standard deviation, for *adm* districts the variance is merely decreased with more complex predictor sets. However, *bas* models seem capable to substantially reduce the variance with more complex predictor sets. The increased stability of prediction performance for *bas* districts indicates that the influence of the random initialization of the neural network's parameters does not effect the predictive performance for these models to the same degree, overall increasing the credibility of the predictions based on *bas* units.

In general, a precision greater 0.5 is only achieved once across all model configurations. This indicates that most combinations are characterized by a high false-positive rate well above 50 %. Concerning sensitivity, most models achieve performances close to a value of 0.75, indicating that about 75 % of observed conflict district-months are detected. The maximum value of 0.876 is achieved for the combination of structural variables with *bas* units for the outcome class **cb** associated with a comparatively low precision value of 0.32. Concerning the ability of a model to correctly identify peaceful district-months, the specificity, for almost all outcome variables, maximum values > 0.93 are achieved (not reported here - see Table A2 in Appendix).

Hegre et al. (2019) do not report sensitivity and precision specifically in their main article. However, for the **sb** conflict class, these metrics are given in the Supplementary Materials of their article. The final ensemble model achieves a precision (sensitivity) of 0.608 (0.963) for the country-month representation and 0.214 (0.443) for the grid-based aggregation. Considering the EV predictor set, for the *adm* representation a precision (sensitivity) of 0.39 (0.78) is achieved, and for the *bas* representation, it is 0.42 (0.79). From the comparison, it becomes evident that both metrics achieve substantially lower performance than the country-month

theme by Hegre et al. (2019). However, while the losses compared to the country-month analysis amount to approximately 1/3, the performance is nearly doubled compared to the grid representation. In the following the **cb** outcome variable's performance is considered because it most closely matches the distinct conflict definitions. In the present study precision (sensitivity) of 0.40 (0.72) is achieved for the *adm* districts, while for *bas* districts 0.42 (0.82) is achieved. Kuzma et al. (2020) report precision (sensitivity) values of 0.40 (0.85) for the African continent using similar aggregation units. While the precision is comparable, the performance of *adm* districts measured by sensitivity is substantially lower. However, for *bas* districts, the achieved performance is comparable to the results of Kuzma et al. (2020), indicating that with the selected predictor variables the *bas* representation more closely achieves comparable performance to related studies. Halkia et al. (2020) are able to achieve values of 0.577 (0.866) based on country-months. While their sensitivity is comparable to the one achieved by *bas* districts, precision is substantially higher. The comparison reveals that overall in terms of precision, the current model configurations require improvement in relation to other studies.

Precision-Recall curves (PR-cureves) are used to compare the performances of distinct models for varying values of precision and sensitivity more directly. For reasons of brevity, the discussion will focus on the PR-curves, but ROC curves are found in the Appendix (Figures A12 and A13). Figure 10 depicts the PR-curves for the *adm* representation. Note that the bold lines indicate the averaged performance across 10 repeats, while faint lines represent the curves of individual models. LR baseline is not included because of the different training process which made it mandatory to train one model for each month in the prediction horizon. For the conflict class **cb**, it becomes evident that CH and SV predictor sets can maintain a higher precision for sensitivity values up to 0.3. However, with the EV predictor set, higher precision values are obtained for sensitivity above 0.6. This shows that the EV set is superior to the other predictor sets in the range of high sensitivity values. A similar pattern emerges for the **sb** conflict class, where the EV predictor set firstly achieves lower sensitivity values at high precision rates. When sensitivity increases over a value of 0.75, the EV set is able to maintain higher precision values.

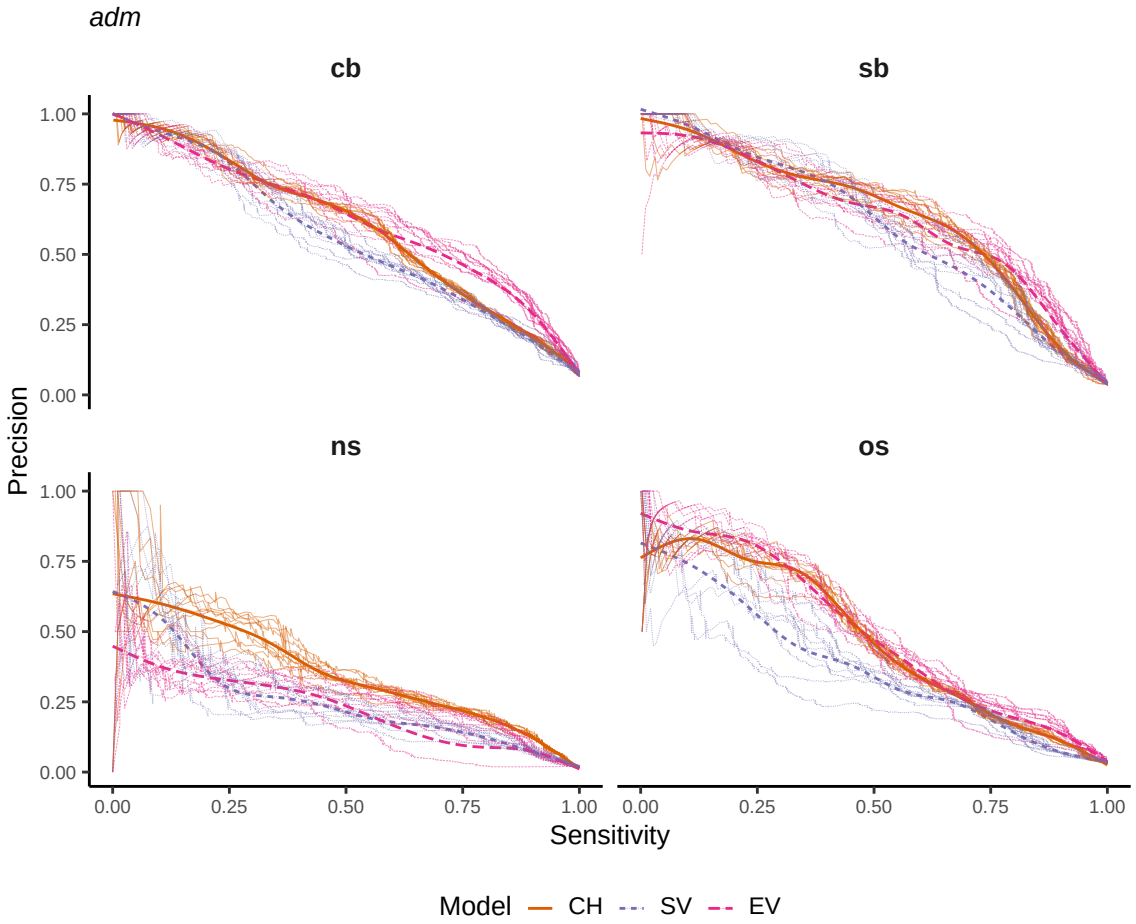


Figure 9: Precision-recall curves for *adm* district models. Bold lines indicate smoothed curves over 10 repeats. Faded lines indicate single repeats.

Concerning the **ns** class, the picture is quite different. The CH set outperforms both of the other sets almost over the complete value range. The EV set shows the lowest precision values at high rates of sensitivity, indicating that for this outcome class SV and EV predictors do not improve model performance. For the **os** class, the CH and EV sets almost perform identically. However, the EV set constantly achieves slightly higher precision values when sensitivity is above 0.5. The SV set performs substantially worse than the other predictor sets for sensitivity values between 0 and 0.7. Considering the *bas* representation for the outcome variable **cb**, CH and EV virtually show an identical performance (Figure 9). The SV predictor set obtains lower precision values for sensitivities between 0.3 to 0.9. Concerning the **sb** outcome variable, SV achieves higher precision than the other sets for low sensitivity values. Towards higher sensitivity, the CH set can maintain the highest precision while only slight differences exist between the SV and EV sets. Concerning the **ns** outcome variable, the pattern is similar to the one found for *adm* districts, indicating that the prediction of the **ns**

outcome class does not benefit from more complex predictor sets. For **os** conflicts, similar to the *adm* representation, the EV set outperforms the other two sets almost over the complete value range. However, the difference to the CH set is only marginal, while SV achieves the lowest precision for increasing sensitivity values.

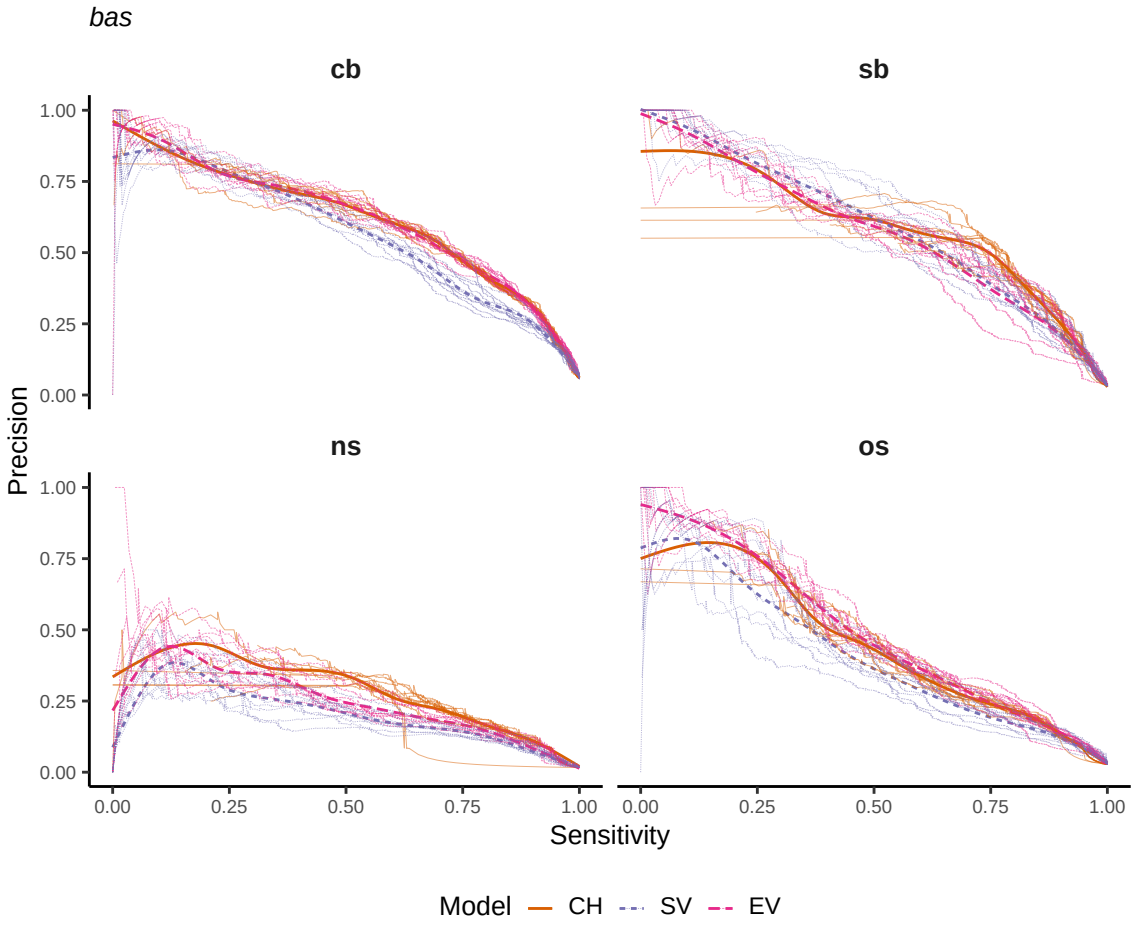


Figure 10: Precision-recall curves for *bas* district models. Bold lines indicate smoothed curves over 10 repeats. Faded lines indicate single repeats.

To summarize these findings, the EV set outperforms other model configurations for *adm* units for **cb**, **sb**, and **os** conflict classes. For the **ns** class, irrespective of the aggregation units, the CH set outperforms the other model configurations. Concerning the *bas* units, the EV set shows slightly better performance for the **os** class, but similar or worse performance to the CH set for all other classes. These findings do not support previous findings of a systematically increased performance of conflict prediction when environmental variables are included. The difference to models based only on conflict history is marginal, and the most simple model frequently outperforms complex variable configurations. However, the specific

conflict class seems to play a role in how structural and environmental variables change conflict prediction performance. Specifically, the EV seems to improve the performance on the **os** outcome class as well as on **cb** and **sb** outcome classes for *adm* districts. It should be noted that the smoothed PR-curves do not represent the performance of a specific model, but the general tendency of a predictor set's performance. The variance between the 10 repeats of training, however, can be quite high as indicated with previous results.

Despite their capability to easily compare different model configurations, PR-curves are not reported very often in related studies. Halkia et al. (2020) report distinct PR-curves for their 10-fold cross-validation strategy. Their results indicate that their model is more capable of achieving relatively high precision scores for increasing sensitivity which is in agreement with previous findings on the global performance. It underlines that the presented DL methods show weaker performance in terms of precision compared to related studies. The desirable feature of a model to maintain precision while increasing sensitivity is not observed at an equal level. Partly, this can be explained by the aggregation to country-months and differences in the definition of the conflict classes. However, comparing to existing conflict prediction tools reveals that it is possible to achieve a better balance between precision and sensitivity. Based on the selected predictor variables in this study, the models ability to reduce the False Positive Rate is limited compared to other studies. The deliberate exclusion of predictor variables delivering more context on the political and cultural compilation of a district could be a reason for the evident reduced performance in terms of precision of the proposed method. The results also illustrate the complexities in the field of conflict prediction. The focus of a machine learning procedure on one metric can adversely affect other performance metrics. It is essential to carefully select a metric for optimization, which represents the overall goal of the prediction task most closely. Because the costs associated with missing out on a conflict district-month are expected to be higher as the costs of wrongly predicting a conflict for a given district-month, the focus on the F_2-score mirrors this assumption. Optimizing towards it means putting more value on sensitivity which explains the reduced performance in terms of precision from a methodological point of view.

4.3 Temporal Performance

The occurrence of conflict is predicted 12 months into the future using a single threshold. Indications for the performance of different model configurations to distinguish between conflict and peace district-months irrespective of the time-dimension are given as density plots

in the Appendix (Figures A1 to A8). Despite using a single cut-off value, the performance of the evaluated model is different along the months in the prediction horizon. Figure 11 depicts the obtained F_2-scores for both aggregation units considering each month in the prediction horizon individually. In general, the performance is relatively stable for most model configurations. Variations over time remain mostly subtle, but changes in both directions towards lower and higher scores are observed for the DL frameworks. The performance of the LR baseline deteriorates with an increasing prediction horizon for almost all conflict classes and is substantially lower compared to the DL models. For the outcome variable **cb**, a better performance of the SV and EV sets compared to the CH theme is observed for *adm* districts, though they almost perform identically. For *bas* districts, EV slightly outperforms all other themes over the entire horizon, but SV shows lower performance than CH. Better performances of the more complex predictor sets compared to the CH theme become more evident for the **sb** conflict class.

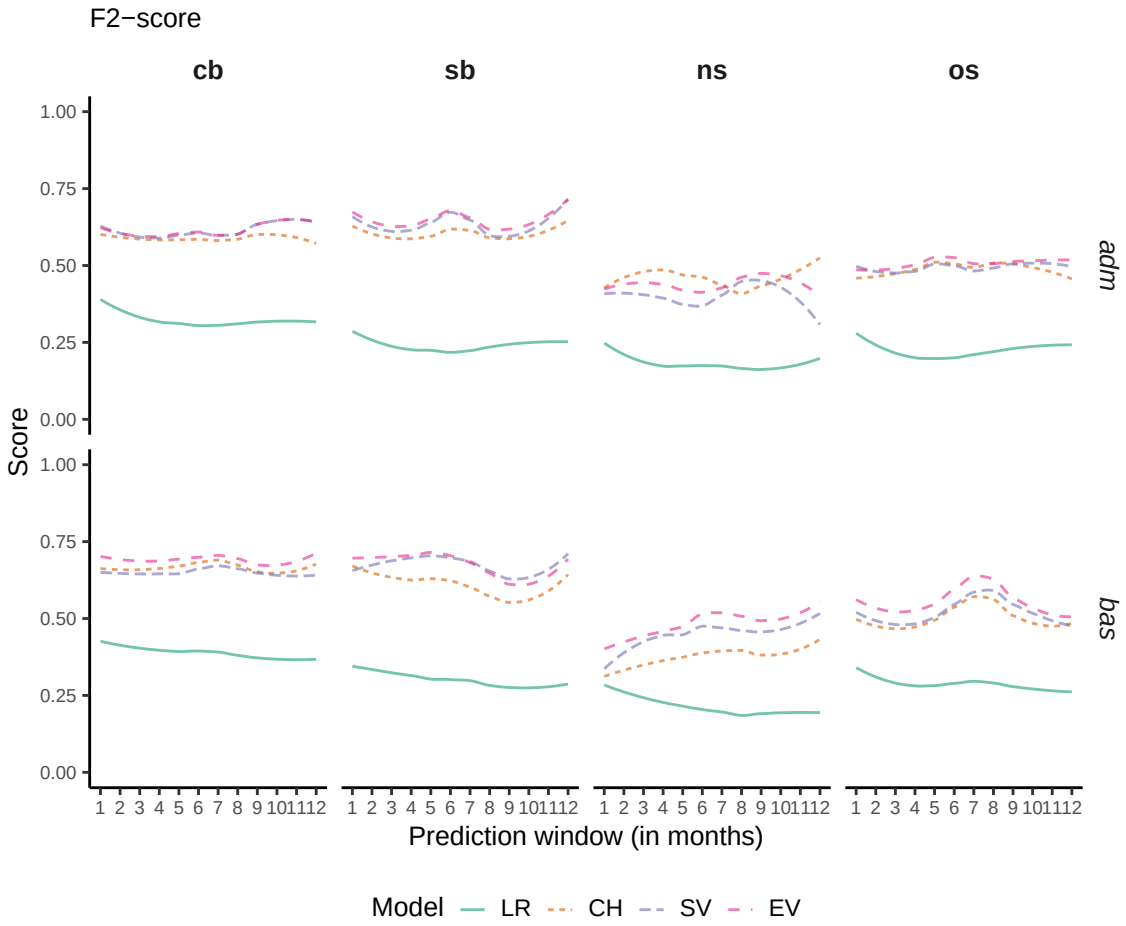


Figure 11: Time dependent performance of the F2-score for adm (top) and bas (bottom) districts. Single lines are obtained by smoothing over the individual scores obtained for 10 repeats.

For both aggregation units, the difference to the CH theme is more substantial over the complete prediction horizon. The pattern of the SV and EV theme over the prediction horizon is very similar and they perform equally well. Differences between *adm* and *bas* districts become evident for the **ns** class. Here the *bas* models almost show a monotonic increase in prediction performance over the horizon. The EV set overall achieves higher scores, especially after five months into the horizon. For *adm* districts, the CH theme outperforms the other predictor sets over the first half and towards the end of the horizon. Only for eight and nine months into the future, SV and EV achieve higher F_2-scores and then experience a substantial decrease in performance. For the **os** outcome class, the DL models perform equally well for the *adm* representation with a slightly better performance of the EV set. For *bas* districts, the increased performance of the EV is more pronounced, but towards the end of the horizon the difference compared to the SV set is reduced.

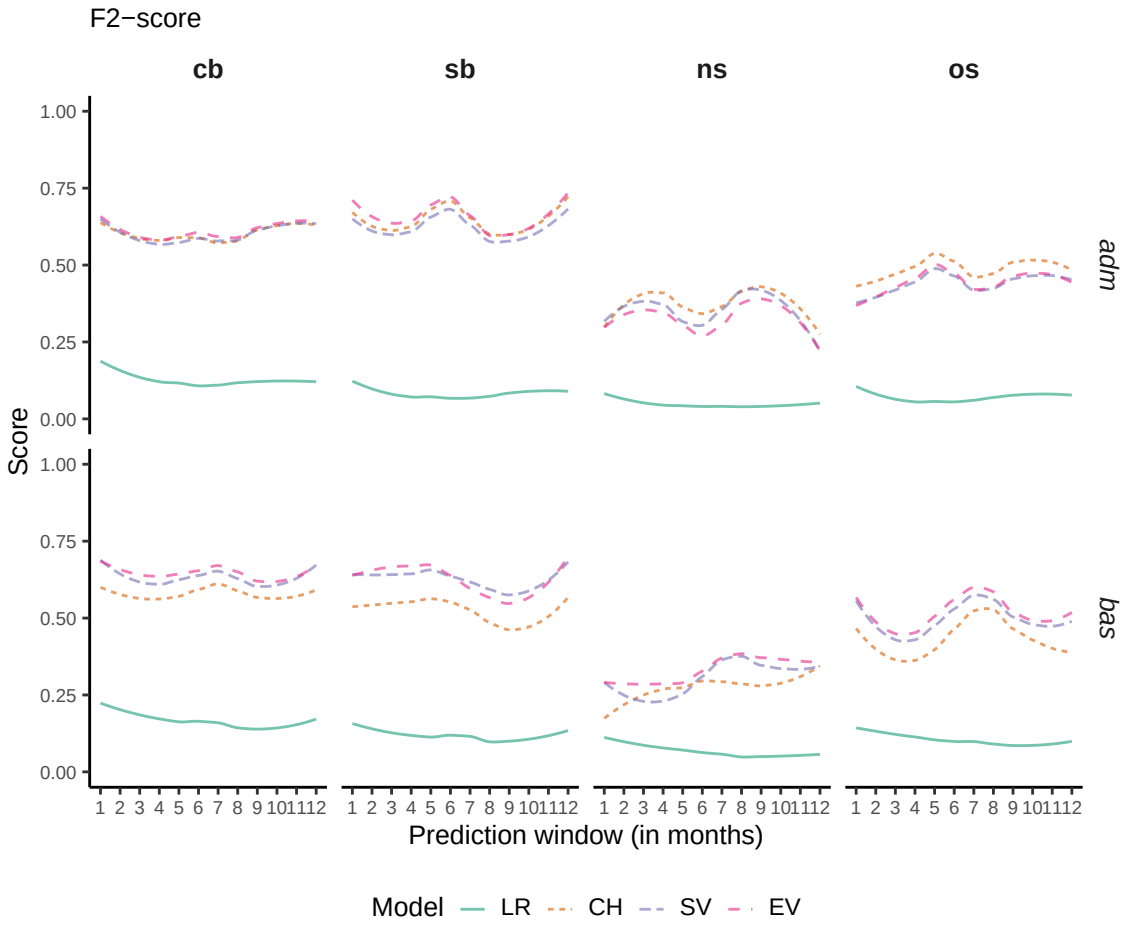


Figure 12: Time dependent performance of the AUPR metric. Single lines are obtained by smoothing over the individual scores obtained for 10 repeats.

In general, for the *adm* representation, the differences between the SV and EV sets are only marginal for all conflict class except **ns**, suggesting that environmental variables do not further increase the prediction performance. For the **ns** class, CH clearly shows some advantages over the other predictor sets for the most part of the prediction horizon, which is in agreement with previous findings. Considering the *bas* representation, EV outperforms other predictor sets except of the **sb** outcome class where the difference to the SV set is marginal. In contrast to *adm* districts, the performance compared to the CH theme is substantial, except for outcome class **cb**, suggesting that the selected predictor variables have a greater effect on prediction performance when they are aggregated on the *bas* districts.

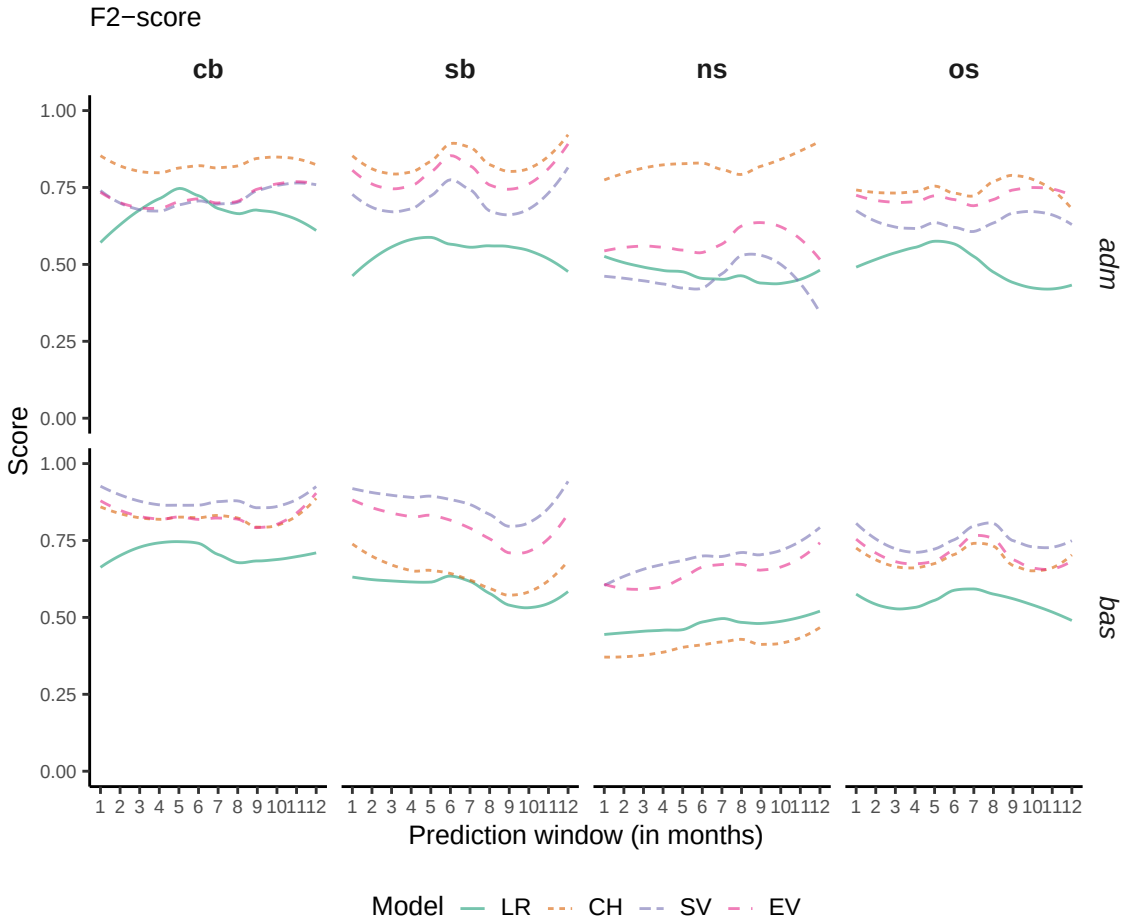


Figure 13: Time dependent performance of sensitivity. Single lines are obtained by smoothing over the individual scores obtained for 10 repeats.

Concerning the temporal performance measured by the AUPR, the general pattern is comparable to the one found for the F_2-score (Figure 11). We observe almost identical patterns for *adm* districts and outcome variables **cb** and **sb**. For outcome variables **ns** and **os** the higher

performance of the CH is even more pronounced. Concerning the *bas* districts, the pattern is reversed and the differences of the SV and EV sets to the CH set is more strongly evident. For outcome variables **cb** and **os**, the EV set shows slightly higher performances than the SV set for the complete prediction horizon. The monotonic increase in performance of the **ns** outcome variable is not as clear, reaching a plateau after five months for the CH set and after eight months for the SV and EV sets.

Directly comparing model performances in terms of precision and sensitivity delivers an imperfect picture of the comparative performances between different model configurations. Rather than the absolute precision or sensitivity values, the balance between these metrics as expressed in the F_2-score or the AUPR better captures information on their mutual dependency. An alternative perspective to compare model performance in terms of precision and sensitivity is to analyze the "cost" a gain in one metric is associated with a loss in the other. An insightful distinction between *adm* and *bas* districts is found in this regard (Figure 13 & Figure 14). While the CH set initially delivers comparatively high values of sensitivity for *adm* districts, it is associated with higher precision values for *bas* districts. Increasing the complexity of the predictor variables with the SV set leads to a substantial decrease in sensitivity for *adm* districts but simultaneously increases the precision. For *bas* districts, moving to the SV set has the opposite effect of increasing the sensitivity but decreasing precision. Finally, the EV set achieves sensitivity and precision rates in between the CH and SV sets. These findings are in agreement with the results of the global performances. The EV theme is better capable of mediating between precision and recall for both aggregation districts. A distinction between the aggregation units is the direction from which this balance is achieved. Solely based on the CH set, *adm* districts seem beneficial to reach high sensitivity values. However, the predictions are associated with a high rate of false alarms because the precision is low. Using *bas* districts, CH delivers relatively high precision values, but the rate observed conflicts are detected remains low. This contrasting behavior could be harnessed to build early-warning systems with a specialized focus on precision or sensitivity. Some applications might emphasize the importance of detecting all future conflict district-months over the precision to only select relevant cases. The results suggest that such a case would benefit from a focus on *adm* districts. For cases where high precision is more relevant than selecting all future conflicts, a model based on *bas* districts could prove a valuable choice.

Concerning related studies, only Hegre et al. (2019) report time-dependent performance of their models. The prediction window differs from the one presented here, as they predict the occurrence of conflict for up to 36 months into the future. Additionally, they restrict their time-dependent performance analysis to the AUC and AUPR metrics. For the first 12 months, similar to the results presented here, the AUPR metric shows great stability for the country-month representation but with values above 0.90. For the grid representation, there is less stability and values as low as 0.35 are observed. Consistent with previous findings, the country-month data set can achieve substantially higher scores than the presented DL models. However, because the prediction window in this study has been set to 12 months only, the long-term performance can not be compared.

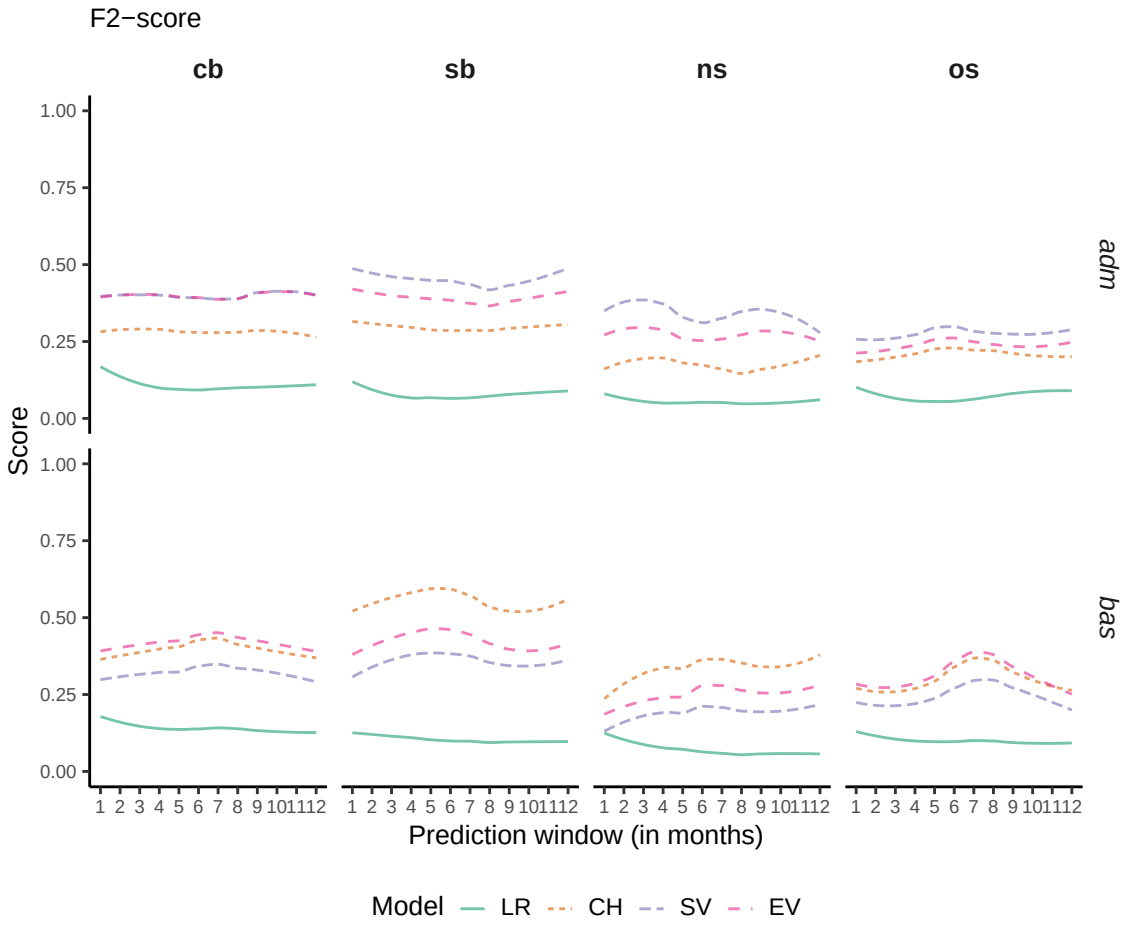


Figure 14: Time dependent performance of precision. Single lines are obtained by smoothing over the individual scores obtained for 10 repeats.

4.4 Spatial Performance

The spatial dimension of the detection task is depicted for the outcome variable **cb** in Figures 15 and 16 for *adm* and *bas* districts, respectively. Additional maps for other outcome variables are found in the Appendix (Figures A14 to A19). In the upper left, the number of observed conflict months within the validation period from January to December 2019 is indicated. The other maps represent the averaged probability of conflict risk for the CH, SV, and EV theme, respectively. The most prominent patterns of high numbers of conflict occur in Mali, the cross border region of Lake Chad, within large parts of the Democratic Republic of Kongo (D.R. Kongo), the Southern part of Somalia, and the province of Cabo Delgado in Mozambique. Lower occurrences of conflict are observed in Libya, some regions in Algeria, Tunisia, Egypt, and larger areas in Sudan, South Sudan, and Ethiopia. Considering the distribution of the observed number of conflict months reveals that for both the *adm* and *bas* districts low numbers of conflict are more frequently observed. The number of districts experiencing 10 or more months of conflicts are increasing for both aggregation units. These observations indicate a general distinction between districts on the African continent. Some of them are characterized by intense and prolonged conflicts, such as Mali, D.R. Kongo or Somalia. Districts in their spatial neighborhood are characterized by relatively low numbers of conflicts so that a pattern of spatial clusters emerges. The majority of the continent does not show conflict events occurring at all within the testing year 2019.

The averaged probability of conflict reveals an interesting distinction between *adm* and *bas* districts. Considering the distribution of predicted probabilities, for *adm* districts, we can observe an increasing separation between peace and conflict district-months. For the CH theme, the probabilities are evenly distributed over the value range from 0.3 to 0.6 with a very high number of districts with a predicted conflict risk of approximately 0.35. Here, most districts without a recent conflict history have been labeled with this value. Districts with a prolonged conflict history are found towards the end of the value range. Moving to the SV theme, this separation becomes more evident, with a clear distinction of districts with a probability of conflict below and above 0.4. Here, virtually all districts within the Sahara and areas recently experiencing high-intensity conflicts such as Mali, the Kongo basin, and the region from Sudan to Somalia are characterized by higher conflict risks than the South and West African districts. Moving to the EV theme, the distinction between districts with a low probability of conflict becomes more pronounced.

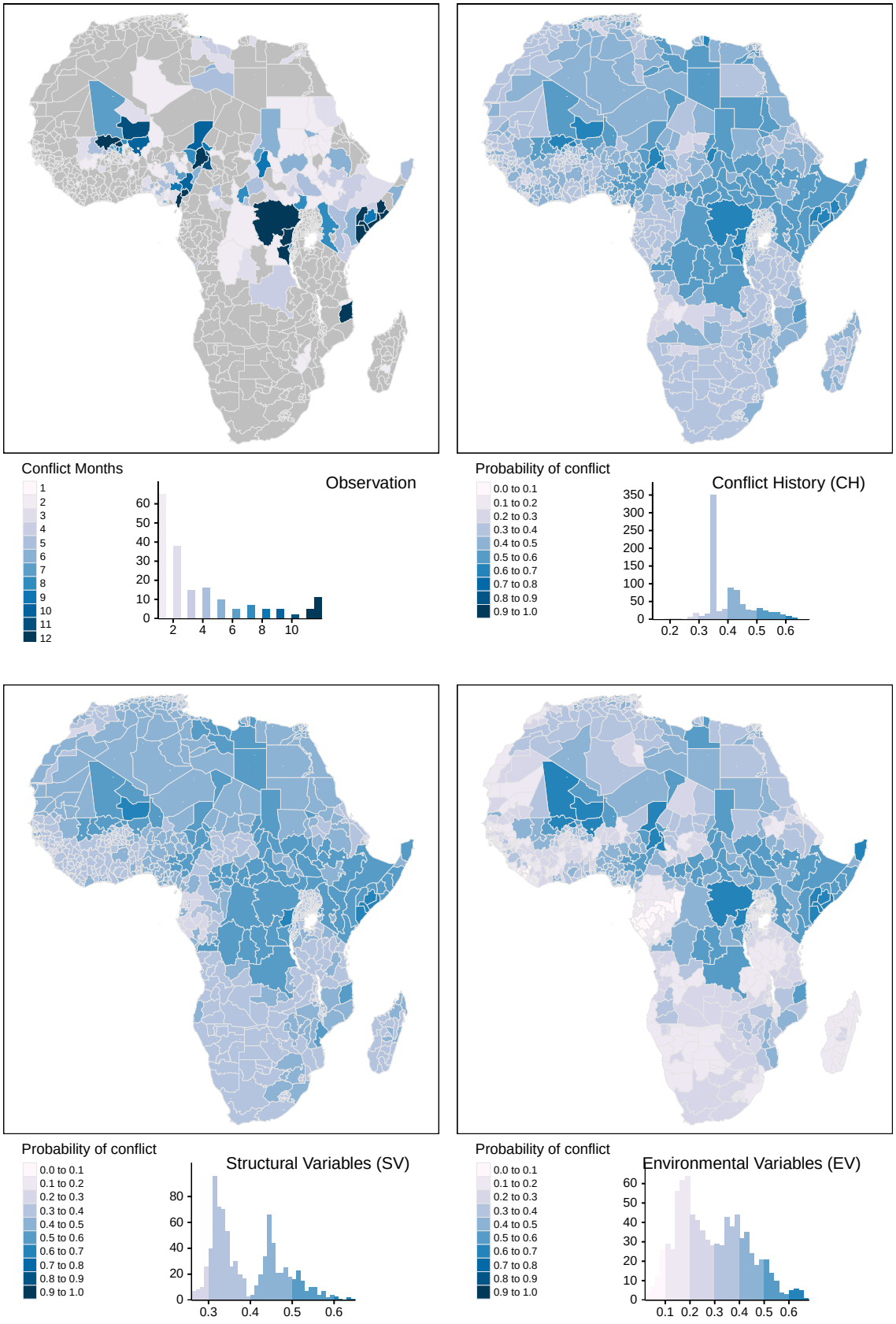


Figure 15: Spatial prediction of conflict class *cb* for *adm* districts. Observed conflict occurrences (upper left) and averaged probabilities of conflict. Grey polygons indicate zero occurrences of conflict.

For example, large parts of Namibia are attributed to a lower risk of conflict than districts in South Africa, Zambia, and Angola. This indicates that the model evaluates these regions differently, distinguishing between very low and slightly elevated conflict risks. For the other model configurations, these differences are not as apparent. On the other end of the value range, areas already experiencing ongoing conflicts are now more frequently associated with probability values above 0.6, rendering these very high conflict risk areas more distinguishable from the general elevated conflict risk in their neighborhood. For *bas* units, the pattern is strikingly different considering the SV theme. Here, the differentiation between low-risk districts and districts with an elevated risk is heavily pronounced. A similar pattern emerges when one considers the densitiy distribution of predicted conflict probability of peace versus conflict district months (Figures A1 to A8). The SV theme for *bas* districts seems to be very capable of distinguishing between districts with low and elevated conflict risk. The majority of districts are associated with predicted probabilities about 0.1, while the major conflict zones are associated with values of 0.4 and higher. This capability of the SV theme is a desirable property of a conflict prediction model. Unfortunately, this discriminatory power is not observed on the same level for the EV theme. Here, we can observe a similar pattern as with the *adm* districts distinguishing between districts with very low conflict risks and districts with slightly elevated conflict risk. However, the pattern for districts with more elevated risk is not as straightforward since most districts are associated with moderate probabilities between 0.3 and 0.4. The zones with a recent intense conflict history still show elevated conflict risks. However, their differentiation against their neighborhood is less pronounced compared to the SV theme.

Because their model is being used as an early-warning system, Hegre et al. (2019) present their conflict forecast as spatial maps based on a country basis and a grid indicating the risk of conflicts expressed as a percentage. They jointly use both prediction patterns first to identify countries at a high risk of conflict and then pinpoint areas with high conflict risk more accurately. While this approach seems reasonable to sequentially achieve higher spatial accuracy, it should be kept in mind that the grid-based performance is substantially lower and associated with high uncertainty. However, they do not spatially compare their prediction results to observed conflicts in a held-out validation set.

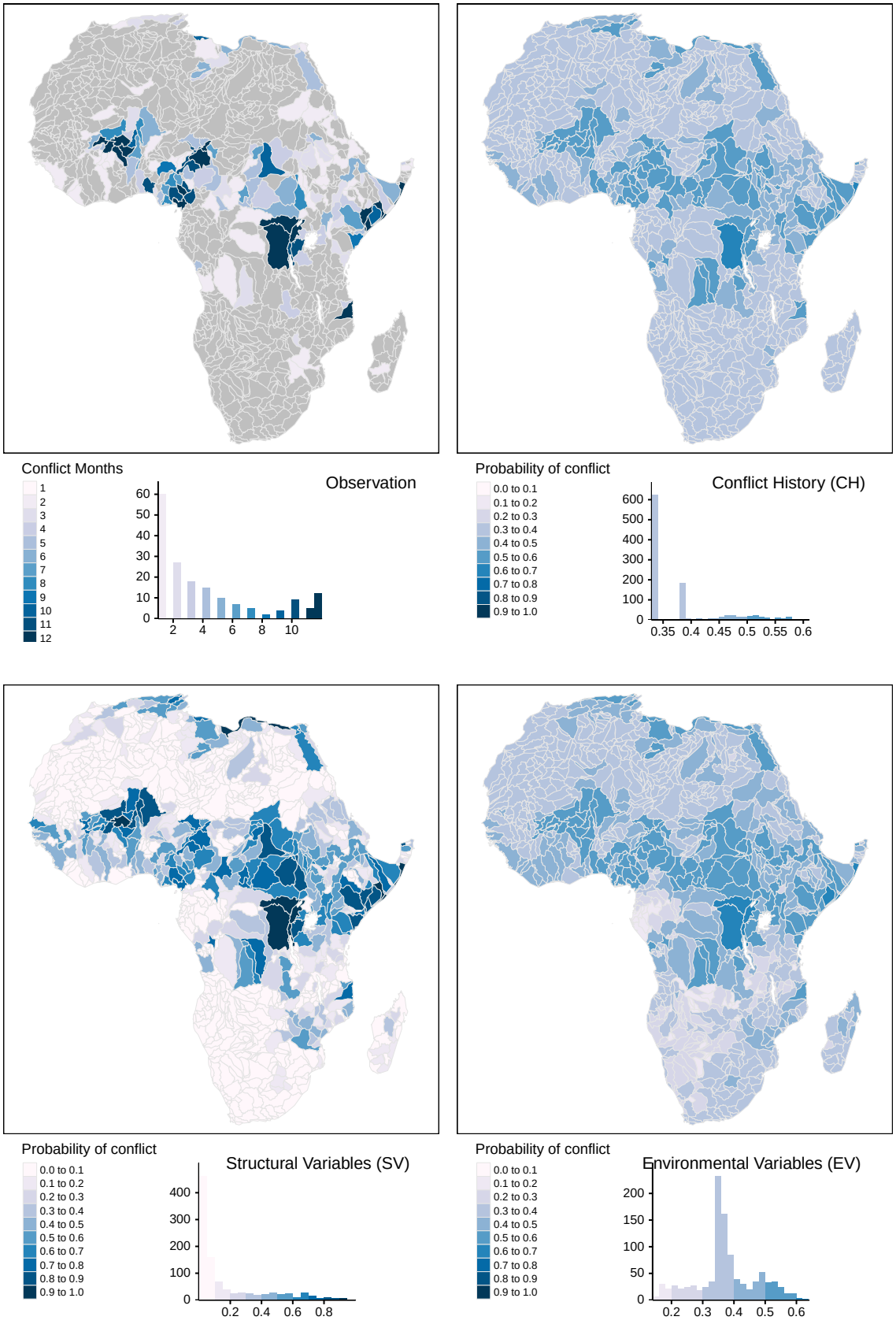


Figure 16: Spatial prediction of conflict class *cb* for *bas* districts. Observed conflict occurrences (upper left) and averaged probabilities of conflict. Grey polygons indicate zero occurrences of conflict.

A scientific publication analyzing the actual prediction for the year 2019 is still in the press as of the time of writing this thesis. Kuzma et al. (2020), although providing their prediction results as a map of conflict risk to the broader public (`https://waterpeacesecurity.org/map`), do not compare their prediction against actual observations in space. Halkia et al. (2020) deliver a map indicating the number of false positives and negatives between 1989 and 2013. Their results show a concentration of false negatives in East Africa as well as in parts of Southern Africa for subnational conflicts. For national power conflicts, these concentrations are found in East and West Africa. Overall, though most related studies serve the purpose of informing decision-makers on the potential spatiotemporal occurrence, the spatial dimension of these predictions is not thoroughly discussed in their respective scientific communication rendering more elaborated comparisons unfeasible.

4.5 Analysis of Variance

Since Leven's test for the homogeneity of variance revealed a statistically significant difference in variance between groups violating the assumption for ANOVA. For this reason, both normal distribution and variance were visually cross-checked (Figures A20 & A21). While the assumption of normal distribution holds for the visual interpretation, differences in variance were observed, especially for the LR model versus the DL models. For that reason, the non-parametric Welch-James test was applied to the interaction term between aggregation unit and predictor set (Table 23).

Table 23: Results of the Welch-James ANOVA.

	cb		**sb**		**ns**		**os**	
Term	p	Sig.	p	Sig.	p	Sig.	p	Sig.
unit	0.00e+00	***	1.77e-03	**	3.65e-01		2.41e-08	***
type	0.00e+00	***	0.00e+00	***	0.00e+00	***	0.00e+00	***
unit:type	3.77e-05	***	2.70e-03	**	1.89e-02	*	6.70e-04	***

General: The significance level is indicated according to: *p<0.05, **p<0.01, ***p<0.001

From Table 23 it becomes evident that the interaction terms between aggregation unit and predictor set are highly significant on the for outcome variables **cb**, **sb**, and **os**. For **ns** we find significant differences in the group's mean at the 95 % confidence interval. Because all

interaction terms indicate a significant difference, in the following, a discussion of the main effects will be omitted. For reasons of brevity, only the interaction terms for the contrast of the EV predictor set against all over sets will be discussed as it is also the most relevant contrast to answer H1 and H2.

Table 24: Results of the Games-Howell test for difference in mean values.

	cb		**sb**		**ns**		**os**	
Contrast	Est.	p	Est.	p	Est.	p	Est.	p
EV:LR								
adm:adm	0.29299	2.4e-10***	0.40543	4.6e-10***	0.25543	7.1e-08***	0.28137	1.4e-10***
adm:bas	0.22865	3.1e-10***	0.34574	7.9e-10***	0.22216	3.0e-07***	0.22065	0.0e+00***
bas:adm	0.36458	9.9e-12***	0.42837	3.4e-10***	0.30397	3.7e-10***	0.32709	2.2e-10***
bas:bas	0.30024	9.4e-11***	0.36869	3.9e-10***	0.27070	3.1e-10***	0.26636	5.3e-11***
EV:CH								
adm:adm	0.02770	1.2e-01	0.04354	2.1e-01	-0.02863	7.7e-01	0.02135	2.1e-01
adm:bas	-0.04619	4.9e-06***	0.04032	8.4e-01	0.06320	6.6e-01	0.00803	1.0e+00
bas:adm	0.09929	7.8e-06***	0.06649	1.6e-02*	0.01991	8.6e-01	0.06706	1.6e-05***
bas:bas	0.02540	3.0e-03**	0.06327	4.1e-01	0.11175	1.1e-01	0.05374	2.5e-02*
EV:SV								
adm:adm	0.00130	1.0e+00	0.01441	8.7e-01	0.04002	8.6e-01	0.01428	7.6e-01
adm:bas	-0.03179	5.5e-04***	-0.02382	4.8e-01	-0.01401	9.6e-01	-0.01023	9.8e-01
bas:adm	0.07289	2.0e-11***	0.03736	1.9e-02*	0.08856	1.1e-01	0.06000	3.4e-04***
bas:bas	0.03980	5.9e-05***	-0.00087	1.0e+00	0.03453	1.0e-02*	0.03549	1.1e-01
EV:EV								
bas:adm	0.07159	3.4e-11***	0.02295	4.0e-01	0.04854	4.0e-02*	0.04571	8.7e-06***

General: Contrasts are indicated by ":" reading as the left-hand side compared to the right-hand side. Est. indicates the difference in mean, p indicates the p value with significance level according to: *p<0.05, **p<0.01, ***p<0.001

First, the contrast between the EV to the LR models sets will be considered. Consistently for all outcome variables and aggregation units, the EV models achieve substantial higher F_2-scores between 0.221 and 0.428 at a highly significant level. The greatest differences are observed for the contrast between the *bas* districts compared to *adm* districts. Moving to the comparison between the EV and CH sets, there are no significant differences in the mean F_2-scores for contrasting *adm* districts across all outcome variables. This indicates that for *adm* districts, environmental predictors do not increase model performance compared to the models containing the conflict history alone. Also, for the **cb** outcome variable, the EV set for *adm* districts achieves a lower performance compared to the CH *bas* model. For the

outcome classes **cb** and **os**, *bas* districts on average achieve 0.1 and 0.067 higher F_2-scores compared to *adm* districts, respectively. For the outcome variable **sb**, the difference in mean is comparable but less significant. For **ns** no significant differences are observed. Considering the contrast between *bas* models, only for **cb** and **os** we observe higher average F_2-scores on a significant level, indicating that for **sb** and **os** including environmental predictors does not improve model performance.

This finding is supported by the contrast between EV and SV predictor sets. Again, considering only *adm* districts, no significant differences are found and in comparison to *bas* districts they achieve lower F_2-scores on average. Higher averages are found for *bas* districts compared to *adm* districts for **cb**, **sb**, and **os** outcome classes on a significant level but not for **ns**. Considering the contrasts for *bas* districts, only for the **cb** outcome class we observe a higher score moving from the SV to the EV set in the amount of 0.04 points in F_2-score. For all other outcome variables, including environmental predictors does not further improve performance further. Comparing the EV sets for *bas* with *adm* districts reveals that except for outcome class **sb** significantly higher mean values are observed for *bas* districts.

The presented evidence in terms of the analysis of variance is consistent with the visualization in Figure 7, where *bas* districts are characterized with higher F_2-scores compared to *adm* districts. However, the variance of the CH set for *bas* districts is substantially larger for **sb** and **ns** outcome classes which explains the insignificant findings for these classes.

Based on these results, H1, which states that environmental variables significantly increase the model performance, can only be confirmed considering outcome variables **cb** in combination with *bas* districts where a significant difference of 0.025 points in F_2-score is observed between the CH and EV set. For all other outcome variables and especially for *adm* districts, no significant increases in mean moving from the CH to the SV and EV predictor sets were observed. Concerning H2, stating that *bas* districts can achieve better performance than *adm* districts, the results generally indicate a confirmation of the hypothesis except for the **sb** and **ns** outcome class where no significant differences in the mean are observed. Considering the **cb** outcome class, the differences are found to be between 0.072 and 0.099 on a significant level based on the EV comparison. For the outcome class **os**, slightly lower mean differences between 0.046 and 0.067 are found.

5 Limitations and Recommendations

Trade-off Decisions. This thesis's results reveal that there are several trade-offs to consider in predicting violent conflict. The most obvious trade-off is balancing a model's precision with its sensitivity. For a given model, an increase in one metric will lead to a decrease in the other. Thus, it is the context the model is to be applied in governing the decision for either one of these metrics. In the present thesis, the decision has been made to give more weight to sensitivity than to precision, realized by optimizing towards the F_2-score. The argument for this decision is that not to miss actual occurrences of conflict is more critical than falsely flagging peaceful district-months as conflict. This comes at the cost that the models tend to predict the occurrence of conflict more frequently. Once a district has crossed a certain threshold, the models predict conflict for almost the entire prediction horizon, resulting in a very low precision in the temporal distribution of conflicts. However, based on the performances of existing early-warning tools, it is evident that quantitative models should not be relied on as the sole instrument in practical conflict prevention efforts (Cederman and Weidmann, 2017). Instead, predicting the risk of conflict occurrence should be considered as one link in a chain of tools for conflict prevention. Predictive models can serve the purpose of delivering information on focus areas where additional quantitative and qualitative analysis would prove as most valuable and this way helps to use limited resources more efficiently.

In this context, another trade-off becomes evident by comparing the DL models' performance with related studies. Compared to studies focusing on country-month data sets, the performance of the proposed method is substantially reduced. In relation to Kuzma et al. (2020), who used similar aggregation units on the sub-national scale, the performance is comparable. For a confirmation of this finding beyond doubt, a thorough investigation on the effect of scale is needed. However, there is some indication that a model's performance will decrease for increasing the spatial detail. Scientists, as well as policymakers, require highly detailed information on future conflicts in the spatial and temporal domain (Chadefaux, 2017). The proposed method of only using data sets that are available in a gridded format allows for almost arbitrary spatial aggregation. It also reduces the complexity of data preparation because data sets with differing spatiotemporal dimensions can easily be harmonized by free and open source tools provided by the spatial research community (Brovelli et al., 2017). Additionally, for research focusing on the interaction between environmental change and human societies, remote sensing provides spatial and temporal comprehensive data sets that

are currently used to derive a sheer magnitude of different environmental variables (Kwok, 2018). The proposed method thus seems beneficial to tailor prediction models to the specific spatiotemporal demands of real-world applications.

However, implementing DL models leads to reduced interpretability of a model's prediction. While it is relatively easy to demonstrate why and how a linear model predicts a particular outcome, DL models are sometimes referred to as black-box models (Gilpin et al., 2019). This metaphor indicates that due to the complex internal structure of DL networks, it is not always explicable how a network predicts a specific outcome. This seriously limits the effective use of DL in conflict prevention efforts because political decision-makers require recommendations on how to lower the conflict risk at a particular location. The research community has not yet fully agreed on a concise definition of interpretability. It often depends strongly on the research domain (Molnar et al., 2020). In conflict research, relatively few studies apply machine learning techniques, so robust standards of interpretability have yet to be defined.

Another trade-off is found in the comparison between *adm* and *bas* districts. Given the presented problem formulation and predictor variables, evidence has been presented that *bas* districts perform better on the conflict prediction task. However, most humans are more familiar with administrative boundaries. Familiarity is an essential factor that helps people to more quickly process visual information (Manahova et al., 2019). Changing the representation of data to something people do not expect or are less familiar with makes it harder to interpret the data. In this sense, the trade-off consists of achieving higher performances versus making data interpretation more challenging for the audience. Again, this trade-off decision needs to be based on the application context of a model. The presented results show that the difference in performance between *adm* and *bas* districts can be quite substantial, indicating that changing to a less familiar representation of the Earth's surface could prove beneficial for the conflict prediction task.

Framing the Response Variable. Within the literature of conflict prediction, various definitions of events of interest exist. Some have been focused on violent conflict (Rost et al., 2009), terrorism (Uddin et al., 2020), rebellions and insurgencies (Collier and Hoeffler, 2004), others on international wars (Beck et al., 2000) and irregular leadership changes (Ward and Beger, 2017). All of these applications require a careful semantic differentiation between different types of events. Besides focusing on a specific type of vio-

lence, various thresholds in terms of casualties have been applied to determine if an event is included in a study. In this thesis, a district-month was considered as belonging to the conflict class if at least one event in the UCDP data base was found. UCDP only includes events with at least 25 casualties (Pettersson and Öberg, 2020). The distinction between three different conflict classes found in the data base allows for some interpretation, however, a more profound semantic differentiation, e.g., in terms of involved actors, political goals, etc., was considered out of scope for the present analysis. However, backed by the finding that for **cb** and **os** conflict classes, the EV predictor set has a higher impact on increasing the prediction performance, some types of violence seem to be better predictable by environmental variables compared to others. Concentrating future investigations on these types of conflict seems promising to increase predictive performance. Additionally, the presented results do not distinguish between different modes of conflict, i.e., between a newly occurring conflict in a given district or the continuance of an ongoing conflict history. Analyzing these conflict modes distinctly can generate insights into a model's ability to differentiate between emerging and ongoing conflicts, as shown by Kuzma et al. (2020), thereby increasing the confidence one can have in a model's prediction. It requires an additional definition of emerging and ongoing conflicts to be applied to the response variable, increasing the complexity of the modeling approach, which is why it was not conducted in this thesis. Future improvements of the current approach should consider this distinction since the CH theme already, to some extent, proves capable of predicting conflict. Thus, a model's ability to capture emerging conflicts is of high relevance, comparing the performance of different model configurations.

Selection of Predictor Variables. As it has been stated above, most of the previously cited studies rely on data sets which are collected per year on a national scale and are comprehensively provided by institutions such as the World Bank. One reason to refrain from sub-national analysis might be that considering sub-national units complicates the collection of predictor variables. Spatially disaggregated variables are hard to collect on a large scale and while disaggregating national statistics to a smaller scale is possible, it adds additional complexity in data preparation and is associated with additional assumptions not necessarily matchin real-world processes (Verstraete, 2017). Disaggregating these administrative-bound variables to sub-basin watersheds is even more challenging because these units tend to cross administrative boundaries. The presented approach of variable selection was thus restricted to gridded data formats at the consequence of the deliberate exclusion of several variables which have been found valuable predictors of violent conflict. Among these are

variables associated with a population's health and education status, such as infant mortality or rate of secondary education, the economic structure on the country level, such as the rate of primary commodity exports in terms of GDP, as well as information on the political system and ethnolinguistic composition of a society, represented by indicators such as the democracy index, level of repression, or the exclusion of power for certain groups. On the one hand, evidence has been presented that despite these simplifications notable performances in conflict prediction can be achieved. On the other hand, ignoring these indicators might have reduced the overall potential of the DL models to predict conflicts more accurately.

While most of the indicators mentioned above could have been easily collected for the *adm* representation of the data, this would have hindered the direct comparison to the performance to the *bas* representation. However, there are variables originating from the research on integrated water resource management that could be collected exclusively for *bas* districts, such as indicators on the (non-)consumptive use of water, water quality, and governance as well as additional hydrological indicators characterizing water availability (Pires et al., 2017). In the future the current approach could be augmented by including *adm* and *bas* specific indicators to compare the resulting predictive performance of these approaches.

Most of the environmental predictors were derived from the MODIS twin satellites. These were chosen because their mission time started in 2001 and is still ongoing, therefore covering an extensive time window by only two instruments. Mixing measurements on the same variable from multiple instruments with differing spatiotemporal extents would have opened additional complexities during data preparation (Pasetto et al., 2018). This underlines the importance that *value-added* remote sensing products play in research questions such as conflict prediction. Different research questions can be investigated much quicker and rigorously when institutions deliver standardized products ready for analysis. Currently, such efforts are observed moving towards digital twins of processes on the Earth's surface and in the atmosphere (Bauer et al., 2021). The importance of analysis ready data sets holds for the spatial mapping of socio-economic variables such as populations counts and GDP, for which continuance and improvement over the next decades will play a decisive role in enabling innovative spatiotemporal analysis in many research fields (Head et al., 2017).

Model Architecture and Training Process. The basic CNN-LSTM architecture as presented in this thesis proved capable of learning the prediction task. The task has been

formulated as a time series problem with an increasing length starting from 48 months. Other possibilities to frame the problem exist. For example, evaluating the predictive model with a fixed size of the time window could be one option. The window size would need optimization, but the results could inform conflict theory on how much knowledge of the past is needed to make accurate conflict predictions for the future. The training strategy was based on batch gradient descent, meaning that all districts are presented to the network before weights are updated. Optimizing for different batch sizes was deemed unfeasible for this thesis because it would considerably increase training time. Additionally, because the model is not expected to generalize beyond its current spatial extent, spatial-cross validation was not considered necessary. However, regionalized models, as shown by Kuzma et al. (2020), could improve performance because the conflict pathways can not be expected to be the same in the entire study domain.

There is potential for improvement in the network design choices especially considering the latest advances in DL. Shih et al. (2019) proposed a mechanism of temporal pattern attention for multivariate time series forecasting to overcome the shortcoming of recurrent networks to memorize long-term dependencies in the data. Their attention mechanism applies convolutional filters onto the hidden state of a recurrent layer at each time step so that the network can learn which variables to pay attention to and which variables to ignore. They achieve promising results for several multivariate time series problems with this approach. Since the conflict prediction task's data structure is very similar, temporal pattern attention would lend itself to future investigations to improve performance. Another recurrent-based network architecture worth to be investigated for the conflict prediction task is Echo State Networks (ESN). These networks consist of a high number of randomly initiated recurrent cells with a trainable output layer. They are more light-weight during training than traditional LSTM and can model chaotic time-dependent systems (Jaeger, 2001). ESN have been successfully applied to highly complex time series prediction problems such as wind power forecasting (L'opez et al., 2018), rainfall estimation (Yen et al., 2019), or spatiotemporal modeling of sea surface temperature (McDermott and Wikle, 2017). Because of the high complexity associated with the occurrence of violent conflicts, ESNs could be tested as a viable alternative model architecture.

ANOVA. The influence of different model architectures on the observed differences in mean performance measured by the F_2-score can not completely be neglected with the cur-

rent study setup. To account for variances due to model architecture, training all models on exactly the same architecture would be beneficial. However, the question would arise if there exists a model architecture which performs equally well for all model configurations and how to find it. DL, like many other research activities, is a process constrained by available computation power and time. In the setup of this study, hyperparameter optimization was implemented based on the most complex predictors sets. The rationale for this decision was that a model architecture capable of learning in a complex setting would also be capable of learning in simpler contexts and not necesarily vice-versa. Additionally, hyperparameter optimization was applied for *adm* and *bas* districts under the same computational constraints simoltaniously. With fewer constraints on computational resources, a more elaborated investigation on the influence of model architechture could yield interesting results. However, in the context of this thesis, the presented evidence should not be considered as a closure on the question of the importance of environmental variables in conflict prediction, but rather as the optimized outcome of a research process associated with computational and methodological constraints.

6 Conclusion

By systematically comparing the predictive performance of different levels of predictor variables and spatial units for data aggregation, this thesis contributed to the understanding of how these components influence conflict prediction. It has been shown that vast amounts of freely available open geodata can be incorporated into complex deep learning models, delivering an edge over classical linear regression models. The occurrence of violent conflict is inherently a time-series problem and treating it as such provides high accuracies even in the absence of additional predictors. The utility of the inclusion of socio-economic and environmental variables into deep learning models shows a dependence on the definition of the outcome class. For some types of violent conflicts, the selected predictors do not decrease the prediction error. For other classes, absolute gains in performance remain low. The role the natural environment plays in the occurrence of violent conflict is still an open debate in the scientific community. The proposed methodology has shown that, due to the increased availability of dense time-series, incorporating a high number of environmental variables in prediction models is feasible. The decision on how to aggregate available predictors substantially affects the prediction outcome. While the presented results do not allow for conclusive assessments, there are indications that aggregating environmental variables based on subbasin watersheds decreases the prediction error. This comes at the cost of less familiarity with the spatial pattern of the prediction outcome. However, depending on the conflict class, the absolute gains can be quite substantial compared to more familiar sub-national administrative districts. Focusing on gridded data sets allows for almost arbitrary spatial aggregation, opening up distinct research opportunities in the field of conflict prediction. With the recently growing public focus on climate change's social consequences, evaluating its impact on conflict risk is a crucial component in ensuring sustainable development. After all, human lives are at risk and increasing our understanding of how we can prevent their losses is of uttermost importance. Prediction is a way to contribute to both supporting conflict prevention efforts and advancing the scientific understanding of the relationship between the natural environment and conflict. The usage of modern deep learning frameworks and the vast availability of open geodata allows for comprehensive spatiotemporal research designs adding value to the analysis of the complex process of violent conflict. Leveraging this potential to create impactful scientific findings and recommendations for action is a primary mandate of applied conflict research in the near future.

7 References

10 Adelaja, A., George, J., 2019. Effects of conflict on agriculture: Evidence from the Boko Haram insurgency. World Development 117, 184–195. https://doi.org/10.1016/j.worlddev.2019.01.010

Albahli, S., Alhassan, F., Albattah, W., Khan, R.U., 2020. Handwritten Digit Recognition: Hyperparameters-Based Analysis. Applied Sciences 10, 5988. https://doi.org/10.3390/app10175988

Alexeev, M., Conrad, R., 2009. The Elusive Curse of Oil. The Review of Economics and Statistics 91, 586–598. https://doi.org/10.1162/rest.91.3.586

Ali, A., Shamsuddin, S.M., Ralescu, A., 2015. Classification with class imbalance problem: A review. International Journal of Advances in Soft Computing and its Applications 5, 176–204.

Allansson, M., 2021. Methodology - Department of Peace and Conflict Research [WWW Document]. URL https://www.pcr.uu.se/research/ucdp/methodology/ (accessed 20.3.2021).

Antonakakis, N., Cunado, J., Filis, G., Perez de Gracia, F., 2015. The Resource Curse Hypothesis Revisited: Evidence from a Panel VAR. MPRA Paper, University Library of Munich. Germany.

Appel, M., Pebesma, E., 2019. On-Demand Processing of Data Cubes from Satellite Image Collections with the gdalcubes Library. Data 4, 92. https://doi.org/10.3390/data4030092

Baik, S., Choi, M., Choi, J., Kim, H., Lee, K.M., 2020. Meta-Learning with Adaptive Hyperparameters, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems. pp. 20755–20765.

Bauer, P., Dueben, P.D., Hoefler, T., Quintino, T., Schulthess, T.C., Wedi, N.P., 2021. The digital revolution of Earth-system science. Nature Computational Science 1, 104–113. https://doi.org/10.1038/s43588-021-00023-0

Bayes, Mr., Price, Mr., 1763. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in

a Letter to John Canton, A. M. F. R. S. Philosophical Transactions (1683-1775) 53, 370–418.

Beck, N., King, G., Zeng, L., 2000. Improving Quantitative Studies of International Conflict: A Conjecture. American Political Science Review 94, 21–35. https://doi.org/10.1017/S0003055400220078

Begueria, S., Vicente-Serrano, S.M., 2017. SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index [WWW Document]. URL https://sac.csic.es/spei (accessed 20.3.2021).

Bjorvatn, K., Farzanegan, M.R., Schneider, F., 2012. Resource Curse and Power Balance: Evidence from Oil-Rich Countries. World Development 40, 1308–1316. https://doi.org/10.1016/j.worlddev.2012.03.003

Boschini, A., Pettersson, J., Roine, J., 2013. The Resource Curse and its Potential Reversal. World Development 43, 19–41. https://doi.org/10.1016/j.worlddev.2012.10.007

Breiman, L., 2001. Random forests. Machine learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brovelli, M.A., Minghini, M., Moreno-Sanchez, R., Oliveira, R., 2017. Free and open source software for geospatial applications (FOSS4G) to support Future Earth. International Journal of Digital Earth 10, 386–404. https://doi.org/10.1080/17538947.2016.1196505

Brzoska, M., Fröhlich, C., 2016. Climate change, migration and violent conflict: Vulnerabilities, pathways and adaptation strategies. Migration and Development 5, 190–210. https://doi.org/10.1080/21632324.2015.1022973

Buhaug, H., Benjaminsen, T.A., Sjaastad, E., Theisen, O.M., 2015. Climate variability, food production shocks, and violent conflict in Sub-Saharan Africa. Environmental Research Letters 10, 125015. https://doi.org/10.1088/1748-9326/10/12/125015

Carmignani, F., Kler, P., 2016. The geographical spillover of armed conflict in Sub-Saharan Africa. Economic Systems 40, 109–119. https://doi.org/10.1016/j.ecosys.2015.08.002

Cederman, L.-E., Weidmann, N.B., 2017. Predicting armed conflict: Time to adjust our expectations? Science 355, 474–476. https://doi.org/10.1126/science.aal4483

Chadefaux, T., 2017. Conflict forecasting and its limits. Data Science 1, 7–17. https://doi.org/10.3233/DS-170002

Chhetri, M., Kumar, S., Pratim Roy, P., Kim, B.-G., 2020. Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan. Remote Sensing 12, 3174. https://doi.org/10.3390/rs12193174

Colaresi, M., Mahmood, Z., 2017. Do the robot: Lessons from machine learning to improve conflict forecasting. Journal of Peace Research 54, 193–214. https://doi.org/10.1177/0022343316682065

Collier, P., 1998. On economic causes of civil war. Oxford Economic Papers 50, 563–573. https://doi.org/10.1093/oep/50.4.563

Collier, P., Hoeffler, A., 2004. Greed and grievance in civil war 56, 563–595. https://doi.org/10.1093/oep/gpf064

Collier, P., Hoeffler, A., 2002. On the Incidence of Civil War in Africa. Journal of Conflict Resolution 46, 13–28. https://doi.org/10.1177/0022002702046001002

Cooney, C., Korik, A., Folli, R., Coyle, D., 2020. Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG. Sensors 20, 4629. https://doi.org/10.3390/s20164629

Cordoni, F., 2020. A comparison of modern deep neural network architectures for energy spot price forecasting. Digital Finance 2, 189–210. https://doi.org/10.1007/s42521-020-00022-2

De Jong, S., Meer, F., Clevers, J.G.P.W., 2007. Basics of Remote Sensing, in: de Jong, S.M., van der Meer, F.D. (Eds.), Remote Sensing Image Analysis: Including the Spatial Domain, Remote Sensing and Digital Image Processing. Springer Netherlands, pp. 1–15. https://doi.org/10.1007/978-1-4020-2560-0_1

Eck, K., Hultman, L., 2007. One-Sided Violence Against Civilians in War: Insights from New Fatality Data. Journal of Peace Research 44, 233–246. https://doi.org/10.1177/0022343307075124

Eklund, L., Degerald, M., Brandt, M., Prishchepov, A.V., ö, P.P., 2017. How conflict affects land use: Agricultural activity in areas seized by the Islamic State. Environmental Research Letters 12, 054004. https://doi.org/10.1088/1748-9326/aa673a

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., Dehmer, M., 2020. An Introductory Review of Deep Learning for Prediction Models With Big Data. Frontiers in Artificial Intelligence 3. https://doi.org/10.3389/frai.2020.00004

European Union, 2021. Copernicus programme [WWW Document]. URL https://www.copernicus.eu (accessed 20.3.2021).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, ROC Analysis in Pattern Recognition 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fearon, J.D., Laitin, D.D., 2003. Ethnicity, Insurgency, and Civil War. American Political Science Review 97, 75–90. https://doi.org/10.1017/S0003055403000534

Fisher, R.A., 1921. Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. The Journal of Agricultural Science 11, 107–135. https://doi.org/10.1017/S0021859600003750

Friedl, M., Sulla-Menashe, Damien, 2019. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MCD12Q1.006

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes 2, 150066. https://doi.org/10.1038/sdata.2015.66

Gates, S., Hegre, H., Nygård, H.M., Strand, H., 2012. Development Consequences of Armed Conflict. World Development 40, 1713–1722. https://doi.org/10.1016/j.worlddev.2012.04.031

GDELT, 2021. The GDELT Project [WWW Document]. URL https://www.gdeltproject.org/ (accessed 20.3.2021).

Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: Continual prediction with LSTM. Neural Computation 12, 2451–2471. https://doi.org/10.1162/089976600300015015

Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T.P., Vanwambeke, S.O., Wint, G.R.W., Robinson, T.P., 2018. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Scientific Data 5, 180227. https://doi.org/10.1038/sdata.2018.227

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv:1806.00069 [cs, stat].

Gleditsch, N.P., Wallensteen, P., Eriksson, M., Sollenberg, M., Strand, H., 2002. Armed Conflict 1946-2001: A New Dataset. Journal of Peace Research 39, 615–637. https://doi.org/10.1177/0022343302039005007

Halkia, M., Ferri, S., Schellens, M.K., Papazoglou, M., Thomakos, D., 2020. The Global Conflict Risk Index: A quantitative tool for policy support on conflict prevention. Progress in Disaster Science 6, 100069. https://doi.org/10.1016/j.pdisas.2020.100069

Hao, M., Fu, J., Jiang, D., Ding, F., Chen, S., 2020. Simulating the Linkages Between Economy and Armed Conflict in India With a Long Short-Term Memory Algorithm. Risk Analysis 40, 1139–1150. https://doi.org/10.1111/risa.13470

Hayes, M., Svoboda, M., Wall, N., Widhalm, M., 2011. The Lincoln Declaration on Drought Indices: Universal Meteorological Drought Index Recommended. Bulletin of the American Meteorological Society 92, 485–488. https://doi.org/10.1175/2010BAMS3103.1

Head, A., Manguin, M., Tran, N., Blumenstock, J.E., 2017. Can Human Development be Measured with Satellite Imagery?, in: Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, ICTD '17. Association for Computing Machinery, New York, NY, USA, pp. 1–11. https://doi.org/10.1145/3136560.3136576

Hegre, H., Allansson, M., Basedau, M., Colaresi, M., Croicu, M., Fjelde, H., Hoyles, F., Hultman, L., Högbladh, S., Jansen, R., Mouhleb, N., Muhammad, S.A., Nilsson, D., Nygård, H.M., Olafsdottir, G., Petrova, K., Randahl, D., Rød, E.G., Schneider, G., von Uexkull, N., Vestby, J., 2019. ViEWS: A political violence early-warning system. Journal of Peace Research 56, 155–174. https://doi.org/10.1177/0022343319823860

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Homer-Dixon, T.F., 1995. The Ingenuity Gap: Can Poor Countries Adapt to Resource Scarcity? Population and Development Review 21, 587–612. https://doi.org/10.2307/2137751

Homer-Dixon, T.F., 1994. Environmental Scarcities and Violent Conflict: Evidence from Cases. International Security 19, 5. https://doi.org/10.2307/2539147

Homer-Dixon, T.F., 1991. On the Threshold: Environmental Changes as Causes of Acute Conflict. International Security 16, 76–116. https://doi.org/10.2307/2539061

Hu, Y., Yan, L., Hang, T., Feng, J., 2020. Stream-Flow Forecasting of Small Rivers Based on LSTM. arXiv:2001.05681 [cs].

Ipsen, N., Mattei, P.-A., Frellsen, J., 2020. How to deal with missing data in supervised deep learning?, in: Art of Learning with MissingValues (Artemiss).

Jaeger, H., 2001. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note'. German National Research Center for Information Technology GMD. Technical Report. Bonn, Germany. 148.

James, G.S., 1951. The Comparison of Several Groups of Observations When the Ratios of the Population Variances are Unknown. Biometrika 38, 324–329. https://doi.org/10.2307/2332578

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-filled seamless SRTM data V4. Tech. Rep. International Centre for Tropical Agriculture (CIAT).

Jones, B.T., Mattiacci, E., Braumoeller, B.F., 2017. Food scarcity and state vulnerability: Unpacking the link between climate variability and violent unrest. Journal

of Peace Research 54, 335–350. https://doi.org/10.1177/0022343316684662

Koren, O., 2018. Food Abundance and Violent Conflict in Africa. American Journal of Agricultural Economics 100, 981–1006. https://doi.org/10.1093/ajae/aax106

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60, 84–90. https://doi.org/10.1145/3065386

Kummu, M., Taka, M., Guillaume, J.H.A., 2018. Gridded global datasets for Gross Domestic Product and Human Development Index over 1990. Scientific Data 5, 180004. https://doi.org/10.1038/sdata.2018.4

Kuzma, S., Kerins, P., Saccoccia, E., Whiteside, C., Roos, H., Iceland, C., 2020. Leveraging Water Data in a Machine Learning-Based Model for Forecasting Violent Conflict. Technical note. [WWW Document]. URL https://www.wri.org/publication/leveraging-water-data (accessed 20.3.2020).

Kwok, R., 2018. Ecology's remote-sensing revolution. Nature 556, 137–138. https://doi.org/10.1038/d41586-018-03924-9

L'opez, E., Valle, C., Allende, H., Gil, E., Madsen, H., 2018. Wind Power Forecasting Based on Echo State Networks and Long Short-Term Memory. Energies 11, 526. https://doi.org/10.3390/en11030526

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Lee, H., Largman, Y., Pham, P., Ng, A.Y., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks, in: Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09. Curran Associates Inc., Red Hook, NY, USA, pp. 1096–1104.

Lee, S., Lee, D.K., 2018. What is the proper way to apply the multiple comparison test? Korean Journal of Anesthesiology 71, 353–360. https://doi.org/10.4097/kja.d.18.00242

Lehner, B., Verdin, K.L., Jarvis, A., 2008. New global hydrography derived from spaceborne elevation data 89, 2. https://doi.org/10.1029/2008EO100001

Li, T., Hua, M., Wu, X., 2020. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). IEEE Access 8, 26933–26940. https://doi.org/10.1109/ACCESS.2020.2971348

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Doll'ar, P., 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs].

Manahova, M.E., Spaak, E., de Lange, F.P., 2019. Familiarity Increases Processing Speed in the Visual System. Journal of Cognitive Neuroscience 32, 722–733. https://doi.org/10.1162/jocn_a_01507

McDermott, P.L., Wikle, C.K., 2017. An Ensemble Quadratic Echo State Network for Nonlinear Spatio-Temporal Forecasting. arXiv:1708.05094 [stat].

McKee, T.B., Doesken, N.J., Kleist, J., 1993. The Relationship of Drought Frequency and Duration to Time Scales, in: Proceedings of the 8th Conference on Applied Climatology. Anaheim, California.

Mehtab, S., Sen, J., Dasgupta, S., 2020. Robust Analysis of Stock Price Time Series Using CNN and LSTM-Based Deep Learning Models, in: Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, Coimbatore, pp. 1481–1486. https://doi.org/10.1109/ICECA49313.2020.9297652

Mockus, J., Tiesis, V., Zilinskas, A., 2014. The application of Bayesian methods for seeking the extremum, in: Hey, A.M. (Ed.), Towards Global Optimization 2. pp. 117–129.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. arXiv:2010.09337 [cs, stat].

Muchlinski, D., Siroky, D., He, J., Kocher, M., 2016. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis 24, 87–103. https://doi.org/10.1093/pan/mpv024

Murdoch, J.C., Sandler, T., 2004. Civil Wars and Economic Growth: Spatial Dispersion. American Journal of Political Science 48, 138–151. https://doi.org/10.1111/j.0092-5853.2004.00061.x

Murdoch, J.C., Sandler, T., 2002. Economic Growth, Civil Wars, and Spatial Spillovers. Journal of Conflict Resolution 46, 91–110. https://doi.org/10.1177/0022002702046001006

Nelson, A., 2008. Estimated Travel Time to the Nearest City of 50000 or More People in Year 2000 [WWW Document]. URL https://forobs.jrc.ec.europa.eu/products/gam/index.php (accessed 20.3.2021).

Nomura, M., 2020. Simple and Scalable Parallelized Bayesian Optimization. arXiv:2006.13600 [cs, stat].

Owain, E.L., Maslin, M.A., 2018. Assessing the relative contribution of economic, political and environmental factors on past conflict and the displacement of people in East Africa. Palgrave Communications 4, 47. https://doi.org/10.1057/s41599-018-0096-6

Parr, T., Howard, J., 2018. The Matrix Calculus You Need For Deep Learning. arXiv:1802.01528 [cs, stat].

Pasetto, D., Arenas-Castro, S., Bustamante, J., Casagrandi, R., Chrysoulakis, N., Cord, A.F., Dittrich, A., Domingo-Marimon, C., Serafy, G.E., Karnieli, A., Kordelas, G.A., Manakos, I., Mari, L., Monteiro, A., Palazzi, E., Poursanidis, D., Rinaldo, A., Terzago, S., Ziemba, A., Ziv, G., 2018. Integration of satellite remote sensing data in ecosystem modelling at local scales: Practices and trends. Methods in Ecology and Evolution 9, 1810–1821. https://doi.org/10.1111/2041-210X.13018

Perry, C., 2013. Machine Learning and Conflict Prediction: A Use Case. Stability: International Journal of Security and Development 2, Art. 56. https://doi.org/10.5334/sta.cr

Pettersson, T., Öberg, M., 2020. Organized violence, 1989–2019. Journal of Peace Research 57, 597–613. https://doi.org/10.1177/0022343320934986

Pezzulo, C., Hornby, G.M., Sorichetta, A., Gaughan, A.E., Linard, C., Bird, T.J., Kerr, D., Lloyd, C.T., Tatem, A.J., 2017. Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. Scientific Data 4, 170089. https://doi.org/10.1038/sdata.2017.89

Phillips, B.J., 2015. Civil war, spillover and neighbors' military spending. Conflict Management and Peace Science 32, 425–442. https://doi.org/10.1177/0738894214530853

Pires, A., Morato, J., Peixoto, H., Botero, V., Zuluaga, L., Figueroa, A., 2017. Sustainability Assessment of indicators for integrated water resources management. Science of The Total Environment 578, 139–147. https://doi.org/10.1016/j.scitotenv.2016.10.217

Probst, P., Boulesteix, A.-L., 2018. To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research 18.

Radočaj, D., š, J.O., Juriši'c, M., Gašparovi'c, M., 2020. Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review. Land 9, 402. https://doi.org/10.3390/land9110402

Rajagukguk, R.A., Ramadhan, R.A.A., Lee, H.-J., 2020. A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power. Energies 13, 6623. https://doi.org/10.3390/en13246623

Raleigh, C., Linke, A., Hegre, H., Karlsen, J., 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset. Journal of Peace Research 47, 651–660.

Raschka, S., 2020. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808 [cs, stat].

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning, Adaptive computation and machine learning. MIT Press, Cambridge, Massachusetts.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Computation 29, 2352–2449. https://doi.org/10.1162/NECO_a_00990

Reuveny, R., 2007. Climate change-induced migration and violent conflict. Political Geography 26, 656–673. https://doi.org/10.1016/j.polgeo.2007.05.001

Riley, S., Degloria, S., Elliot, S.D., 1999. A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity. Intermountain Journal of Sciences 5, 23–27.

Rost, N., Schneider, G., Kleibl, J., 2009. A global risk assessment model for civil wars. Konstanzer Online-Publikations-System (KOPS). University of Konstanz, Germany. 13.

Running, Steve, Mu, Q., Zhao, M., 2017a. MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MOD16A2.006

Running, Steve, Mu, Q., Zhao, M., 2017b. MYD16A2 MODIS/Aqua Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MYD16A2.006

Running, S., Mu, Q., Zhao, M., 2015a. MOD17A2H MODIS/Terra Gross Primary Productivity 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MOD17A2H.006

Running, S., Mu, Q., Zhao, M., 2015b. MYD17A2H MODIS/Aqua Gross Primary Productivity 8-Day L4 Global 500m SIN Grid V006. https://doi.org/10.5067/MODIS/MYD17A2H.006

Running, Steven, Mu, Q., Zhao, M., Moreno, A., 2017. MODIS Global Terrestrial Evapotranspiration (ET) Product (NASA MOD16A2/A3) NASA Earth Observing System MODIS Land Algorithm 34.

Sachs, J.D., Warner, A.M., 1995. Natural Resource Abundance and Economic Growth. National Bureau of Economic Research. https://doi.org/10.3386/w5398

Schellens, M.K., Belyazid, S., 2020. Revisiting the Contested Role of Natural Resources in Violent Conflict Risk through Machine Learning. Sustainability 12, 6574. https://doi.org/10.3390/su12166574

Schutte, S., 2017. Regions at Risk: Predicting Conflict Zones in African Insurgencies. Political Science Research and Methods 5, 447–465. https://doi.org/10.1017/psrm.2015.84

Senay, G.B., Kagone, S., Velpuri, N.M., 2020. Operational Global Actual Evapotranspiration: Development, Evaluation, and Dissemination. Sensors 20, 1915. https://doi.org/10.3390/s20071915

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., de Freitas, N., 2016. Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 104, 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Shih, S.-Y., Sun, F.-K., Lee, H., 2019. Temporal Pattern Attention for Multivariate Time Series Forecasting. arXiv:1809.04206 [cs, stat].

Song, J., Gao, S., Zhu, Y., Ma, C., 2019. A survey of remote sensing image classification based on CNNs. Big Earth Data 3, 232–254. https://doi.org/10.1080/20964471.2019.1657720

South, A., 2017. Rnaturalearth: World Map Data from Natural Earth [WWW Document]. URL https://CRAN.R-project.org/package=rnaturalearth (accessed 20.3.2020).

Srinivas, N., Krause, A., Kakade, S.M., Seeger, M., 2012. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. IEEE Transactions on Information Theory 58, 3250–3265. https://doi.org/10.1109/TIT.2011.2182033

Sulla-Menashe, D., Friedl, M.A., 2018. User Guide to Collection 6 MODIS Land Cover (MCD12Q1 and MCD12C1) Product [WWW Document]. URL https://lpdaac.usgs.gov/documents/101/MCD12_User_Guide_V6.pdf (accessed 20.3.2020).

Sun, S., Cao, Z., Zhu, H., Zhao, J., 2019. A Survey of Optimization Methods from a Machine Learning Perspective. arXiv:1906.06821 [cs, math, stat].

Sundberg, R., Eck, K., Kreutz, J., 2012. Introducing the UCDP Non-State Conflict Dataset. Journal of Peace Research 49, 351–362. https://doi.org/10.1177/0022343311431598

Tharwat, A., 2020. Classification assessment methods. Applied Computing and Informatics 17. https://doi.org/10.1016/j.aci.2018.08.003

Tollefsen, A.F., Strand, H., Buhaug, H., 2012. PRIO-GRID: A unified spatial data structure. Journal of Peace Research 49, 363–374. https://doi.org/10.1177/0022343311431287

Uddin, M.I., Zada, N., Aziz, F., Saeed, Y., Zeb, A., Ali Shah, S.A., Al-Khasawneh, M.A., Mahmoud, M., 2020. Prediction of Future Terrorist Activities Using Deep Neural Networks. Complexity. https://doi.org/10.1155/2020/1373087

Verstraete, J., 2017. The Spatial Disaggregation Problem: Simulating Reasoning Using a Fuzzy Inference System. IEEE Transactions on Fuzzy Systems 25, 627–641. https://doi.org/10.1109/TFUZZ.2016.2567452

Vicente-Serrano, S.M., Beguer'ia, S., L'opez-Moreno, J.I., 2010. A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index. Journal of Climate 23, 1696–1718. https://doi.org/10.1175/2009JCLI2909.1

Wan, Z., Hook, S., Hulley, G., 2015a. MOD11C3 MODIS/Terra Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MOD11C3.006

Wan, Z., Hook, S., Hulley, G., 2015b. MYD11C3 MODIS/Aqua Land Surface Temperature/Emissivity Monthly L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MYD11C3.006

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., De Feitas, N., 2016. Bayesian Optimization in a Billion Dimensions via Random Embeddings. Journal of Artificial Intelligence Research 55, 361–387. https://doi.org/10.1613/jair.4806

Ward, M.D., Bakke, K., 2005. Predicting Civil Conflicts: On the Utility of Empirical Research, in: Proccedings of the Conference on Disaggregating the Study of Civil War and Transnational Violence. University of California. San Diego, USA. p. 22.

Ward, M.D., Beger, A., 2017. Lessons from near real-time forecasting of irregular leadership changes. Journal of Peace Research 54, 141–156. https://doi.org/10.1177/0022343316680858

Ward, M.D., Greenhill, B.D., Bakke, K.M., 2010. The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research 47, 363–375. https://doi.org/10.1177/0022343309356491

Ward, M.D., Metternich, N.W., Dorff, C.L., Gallop, M., Hollenbach, F.M., Schultz, A., Weschle, S., 2013. Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction. International Studies Review 15, 473–490. https://doi.org/10.1111/misr.12072

Welch, B.L., 1951. On the Comparison of Several Mean Values: An Alternative Approach. Biometrika 38, 330–336. https://doi.org/10.2307/2332579

Witmer, F.D., Linke, A.M., O'Loughlin, J., Gettelman, A., Laing, A., 2017. Subnational violent conflict forecasts for sub-Saharan Africa, 2015, using climate-sensitive models. Journal of Peace Research 54, 175–192. https://doi.org/10.1177/0022343316682064

WorldPop, 2018. Global High Resolution Population Denominators Project [WWW Document]. URL https://doi.org/10.5258/SOTON/WP00654 (accessed 20.3.2021).

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019. Hyper-parameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology 17, 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Łukasz, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs].

Yang, S., Wang, Y., Chu, X., 2020. A Survey of Deep Learning Techniques for Neural Machine Translation. arXiv:2002.07526 [cs].

Yao, Q., Wang, M., Chen, Y., Dai, W., Li, Y.-F., Tu, W.-W., Yang, Q., Yu, Y., 2019. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv:1810.13306 [cs, stat].

Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E., Chen, C.-C., 2019. Application of the deep learning for the prediction of rainfall in Southern Taiwan. Scientific Reports 9. https://doi.org/10.1038/s41598-019-49242-6

Yu, T., Zhu, H., 2020. Hyper-Parameter Optimization: A Review of Algorithms and Applications. arXiv:2003.05689 [cs, stat].

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Computation 31, 1235–1270. https://doi.org/10.1162/neco_a_01199

Zhang, Y., Wallace, B., 2016. A Sensitivity Analysis of (and Practitioners Guide to) Convolutional Neural Networks for Sentence Classification. arXiv:1510.03820 [cs].

Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L., 2014. Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks, in: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Kobsa, A., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Terzopoulos, D., Tygar, D., Weikum, G., Li, F., Li, G., Hwang, S., Yao, B., Zhang, Z. (Eds.), Web-Age Information Management. Springer International Publishing, Cham, pp. 298–310. https://doi.org/10.1007/978-3-319-08010-9_33

Zou, Q., Xie, S., Lin, Z., Wu, M., Ju, Y., 2016. Finding the Best Classification Threshold in Imbalanced Classification. Big Data Research 5, 2–8. https://doi.org/10.1016/j.bdr.2015.12.001

A Appendix

Density Plots of Predicted Conflict Probability

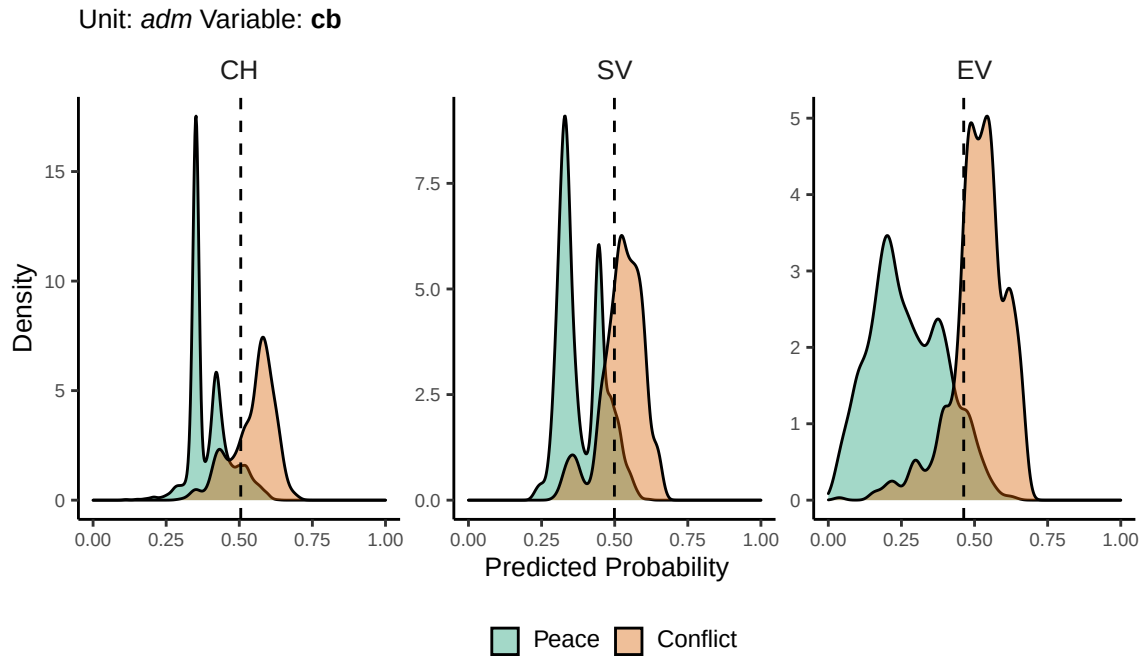


Figure A1: Predicted probability of **cb** conflicts for *adm* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

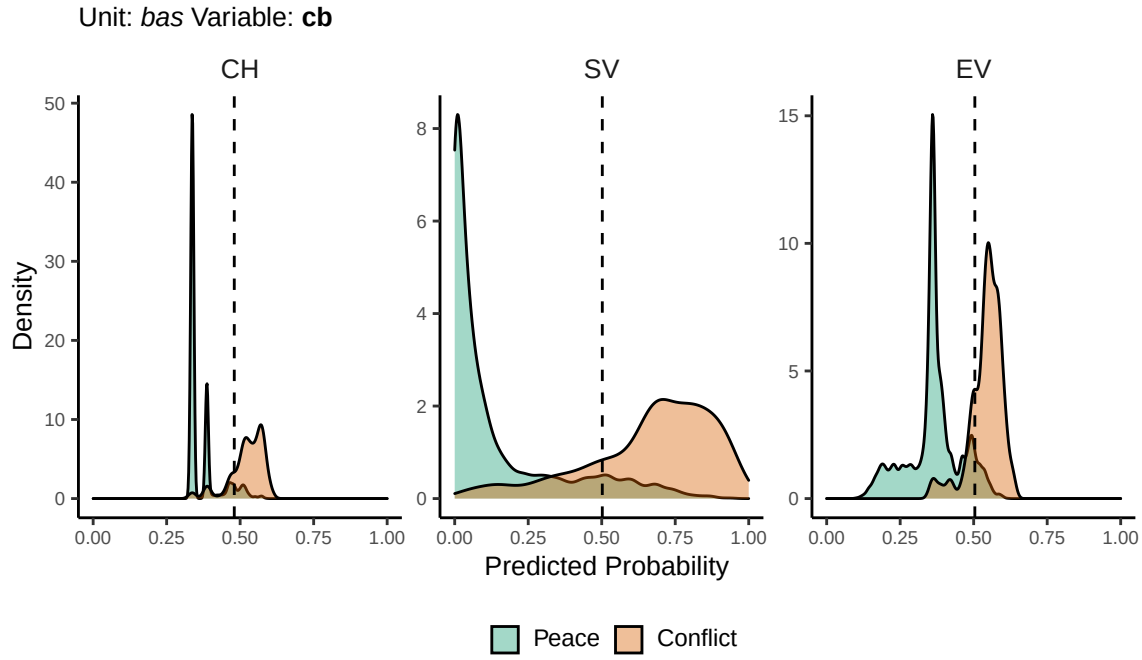


Figure A2: Predicted probability of **cb** conflicts for *bas* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

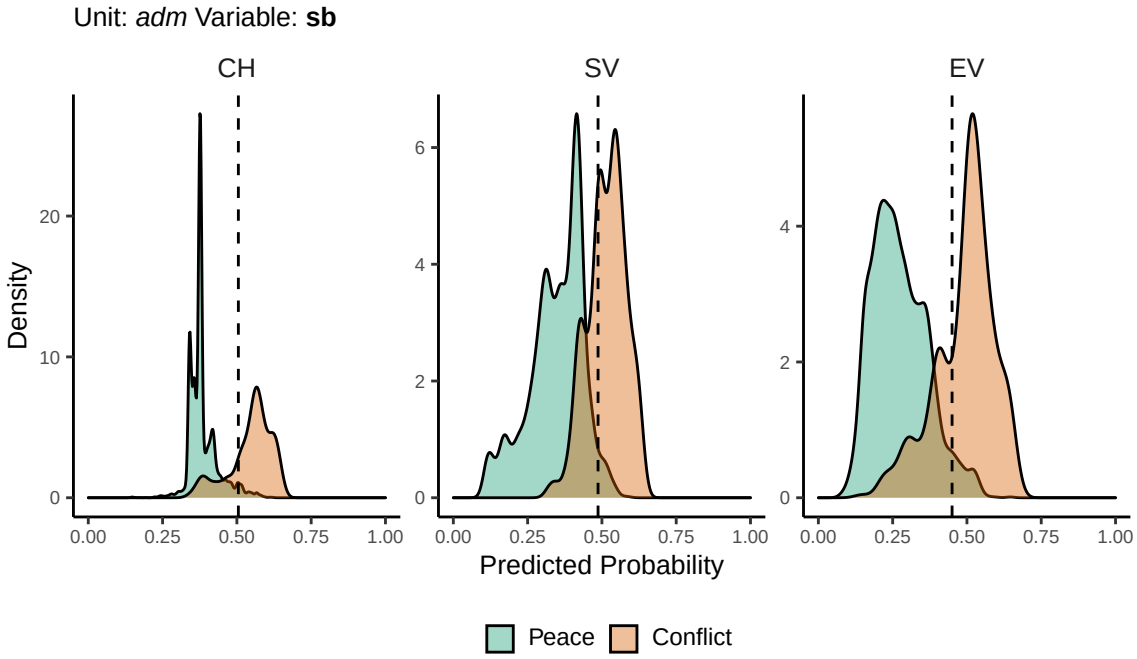


Figure A3: Predicted probability of **sb** conflicts for *adm* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

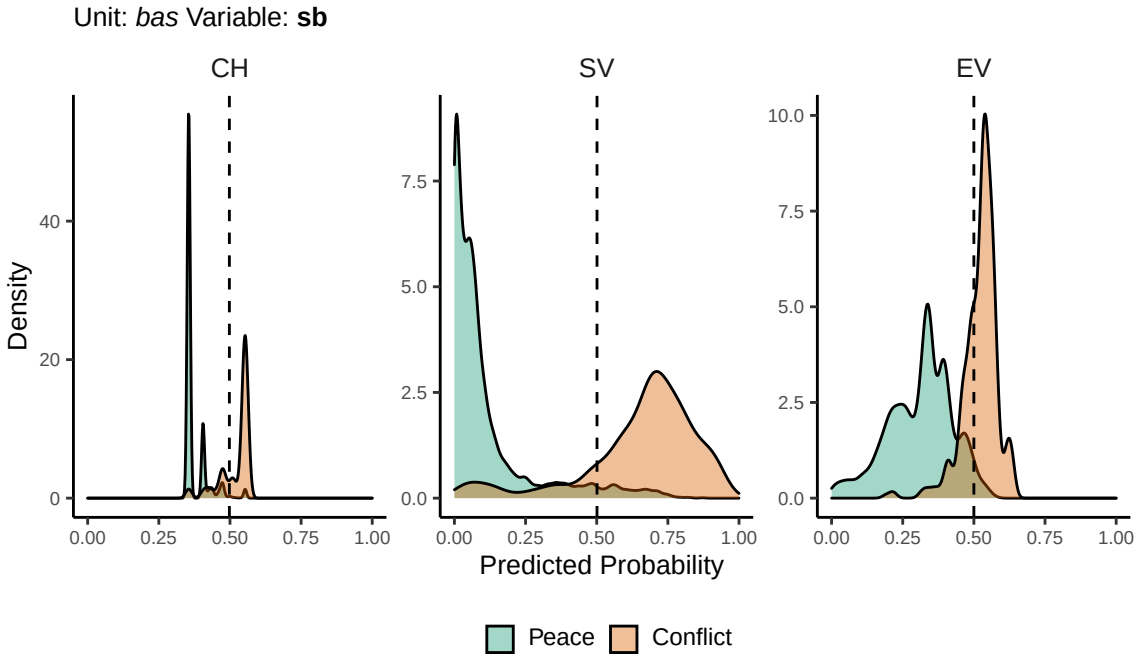


Figure A4: Predicted probability of **sb** conflicts for *bas* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

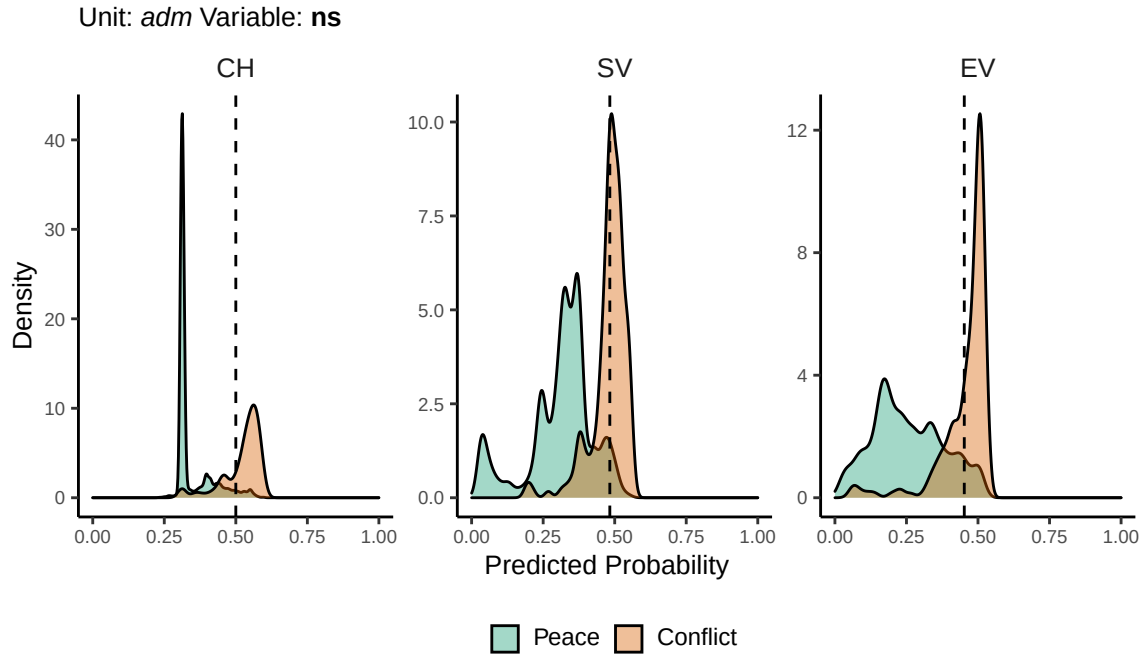


Figure A5: Predicted probability of **ns** conflicts for *adm* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

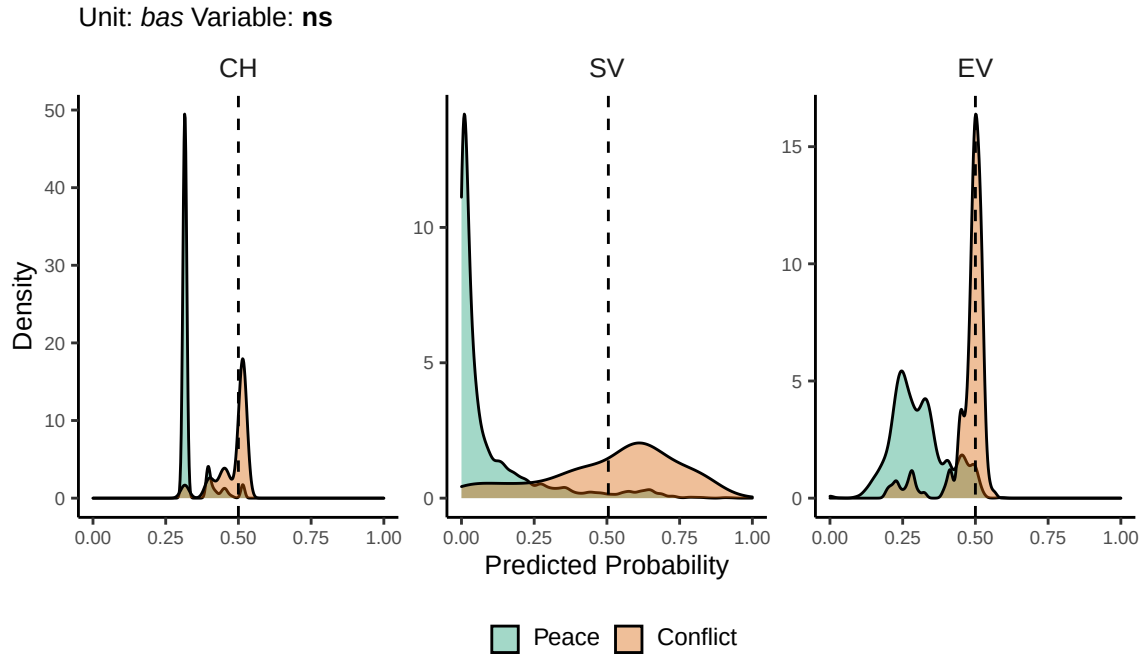


Figure A6: Predicted probability of **ns** conflicts for *bas* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

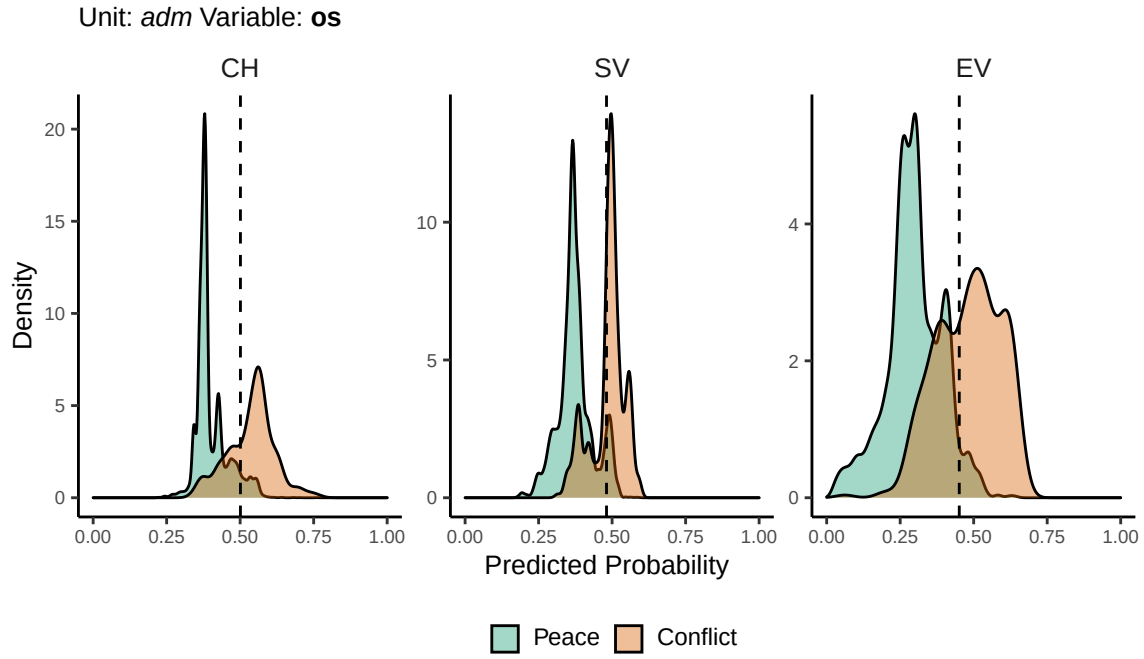


Figure A7: Predicted probability of **os** conflicts for *adm* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

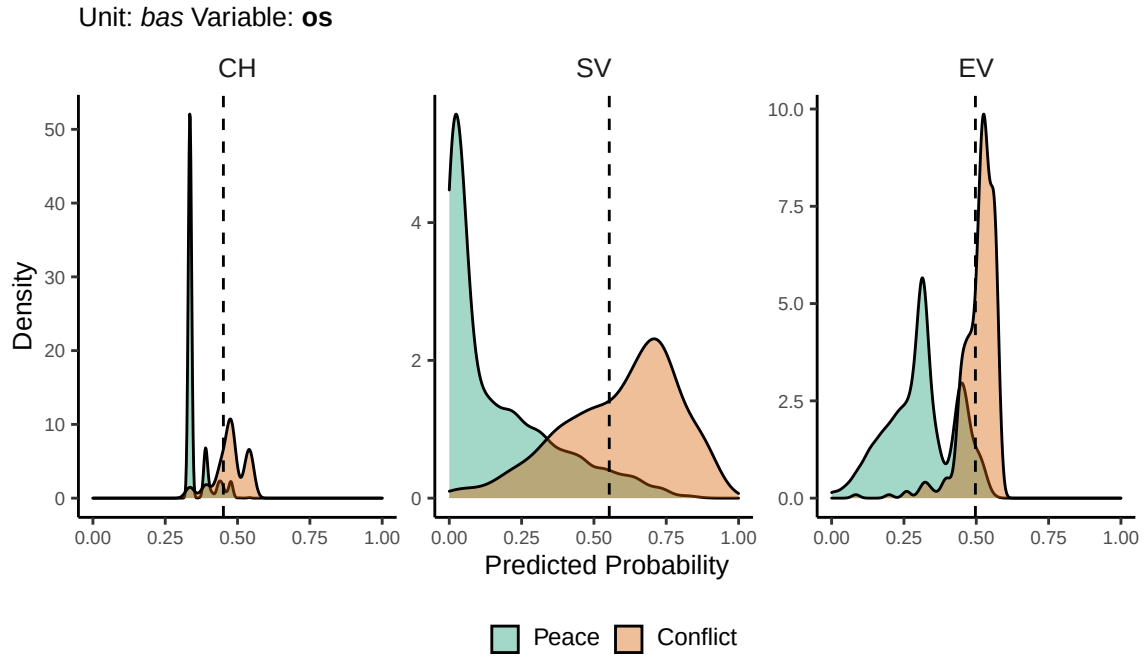


Figure A8: Predicted probability of **os** conflicts for *bas* districts. Note that in order to increase visibility the scale on the y-axis differs from one facet to another.

Additional Global Performance Metrics

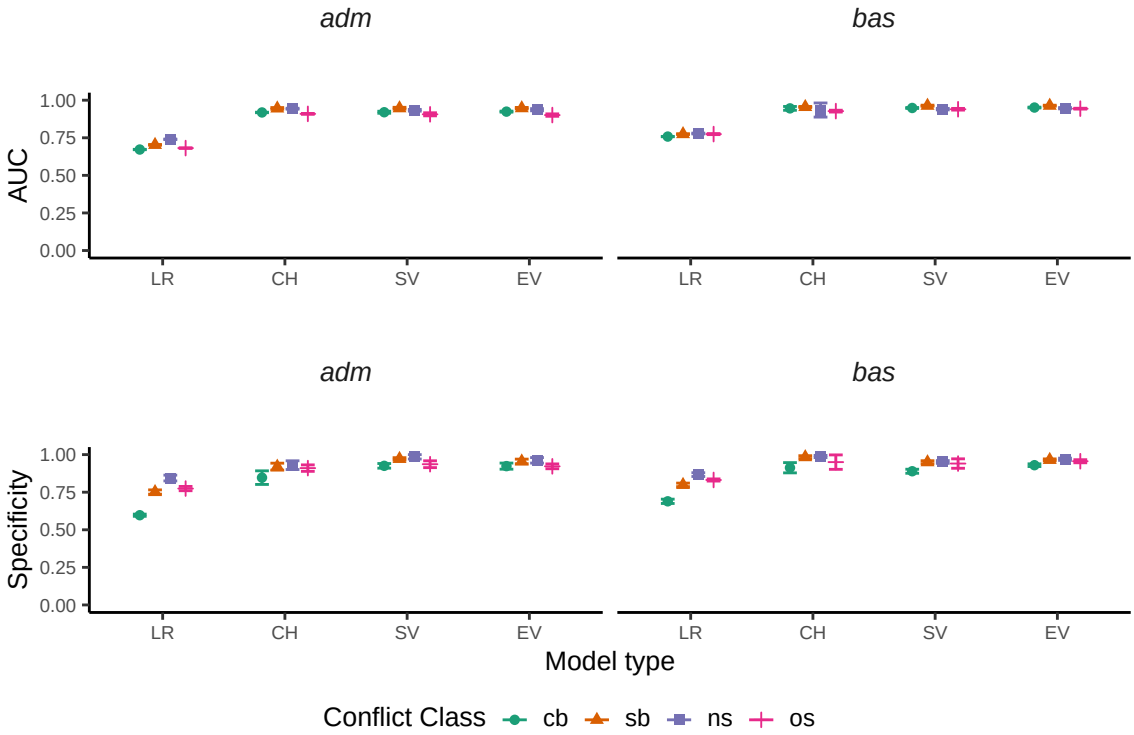


Figure A9: Global performance of the AUC metric (top) and pecificity (bottom).

Additional Temporal Performance Metrics

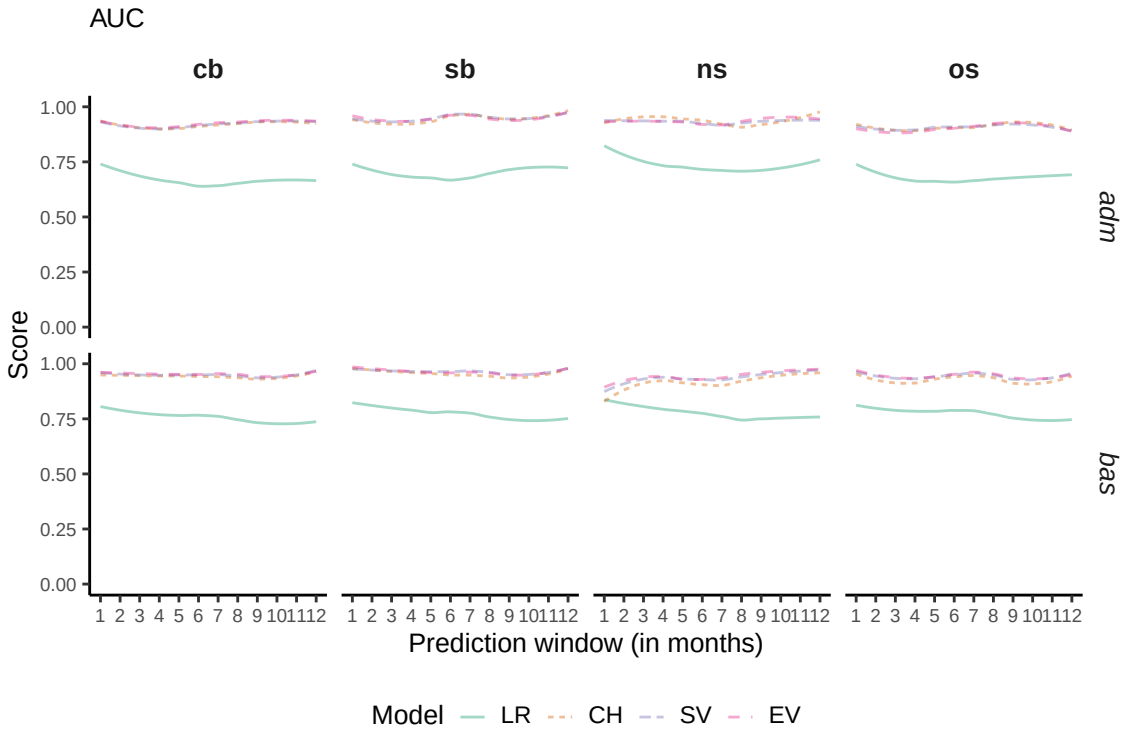


Figure A10: Time dependent performance of the AUC metric.

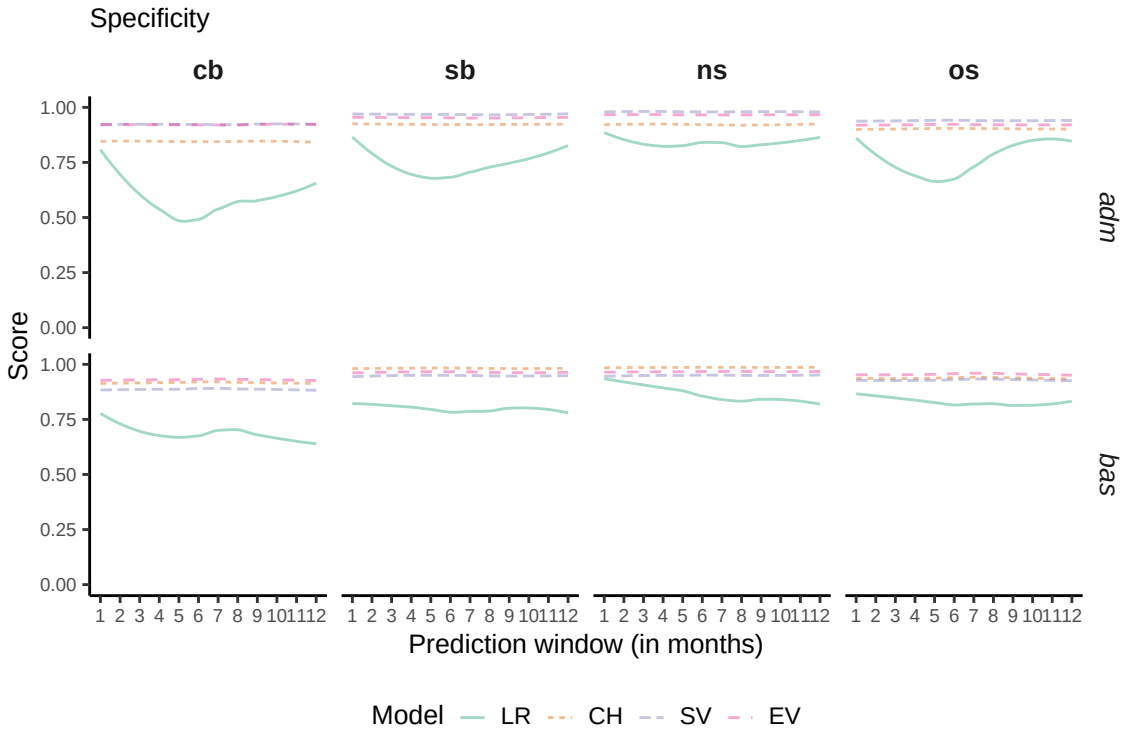


Figure A11: Time dependent performance of specificity.

ROC Curves

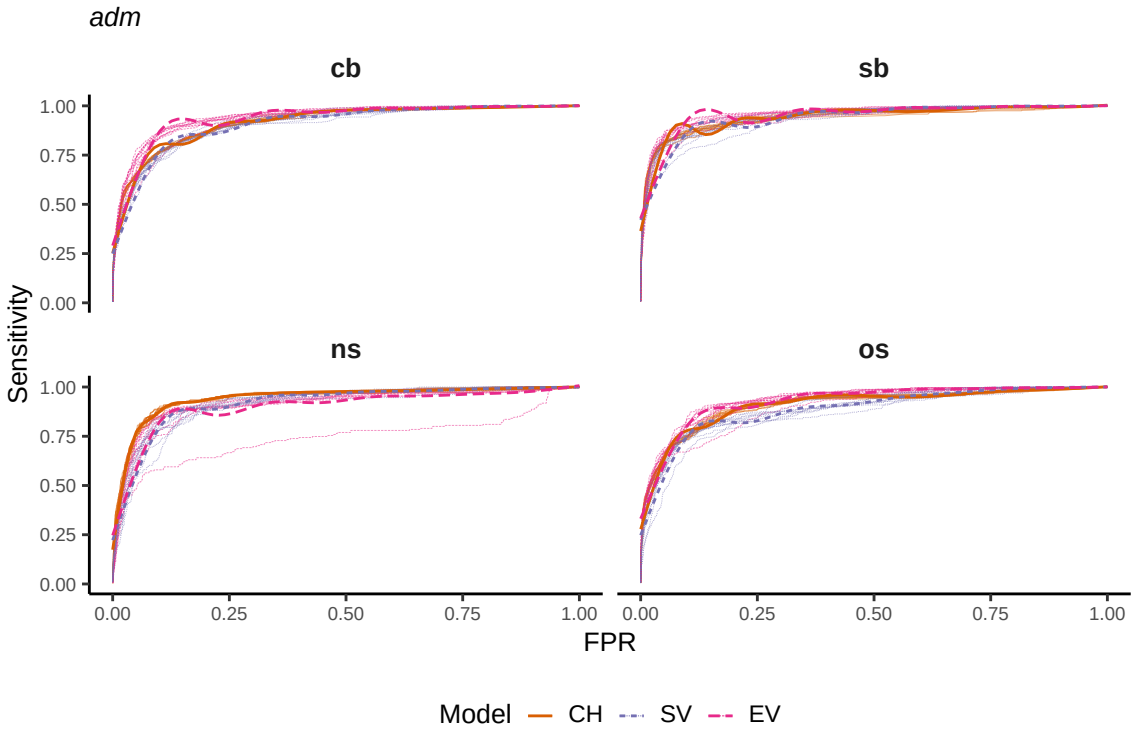


Figure A12: ROC curves for *adm* district models.

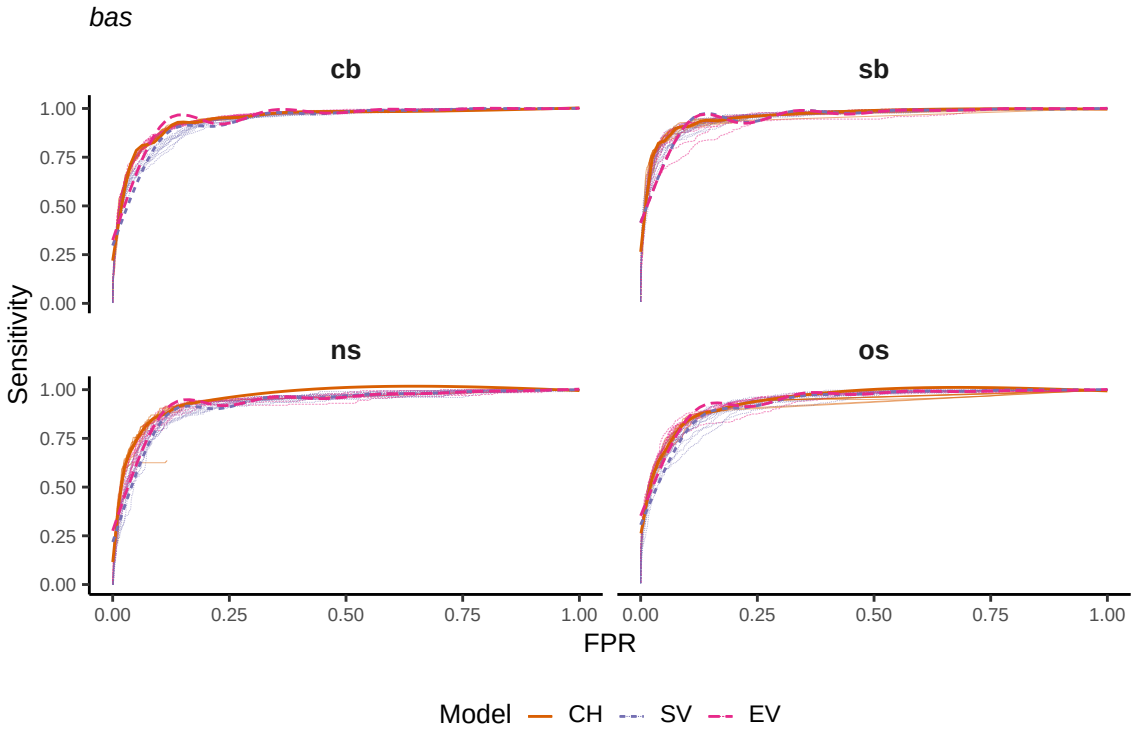


Figure A13: ROC curves for *bas* district models.

Additional Spatial Predictions

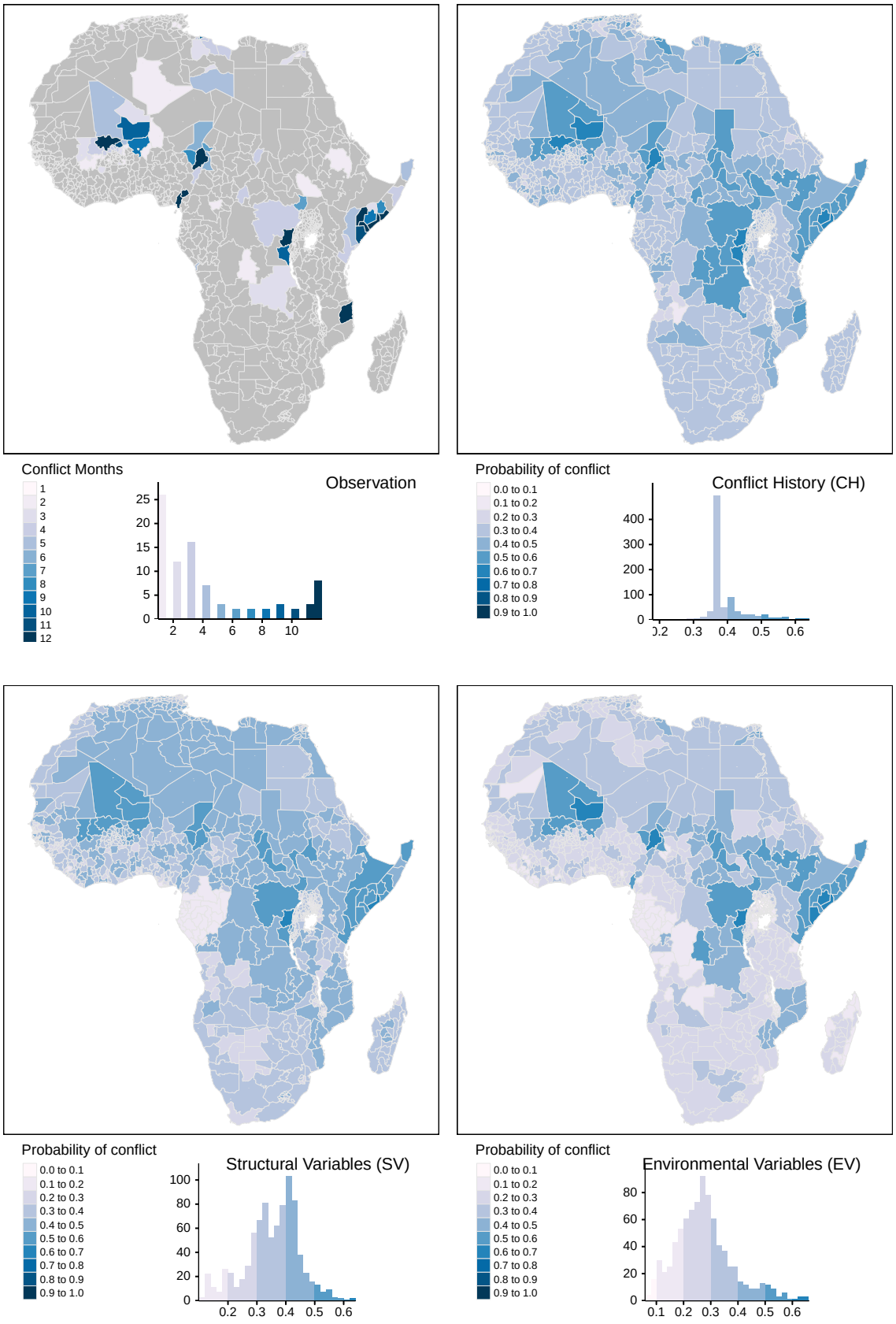


Figure A14: Spatial prediction of conflict class **sb** for *adm* districts.

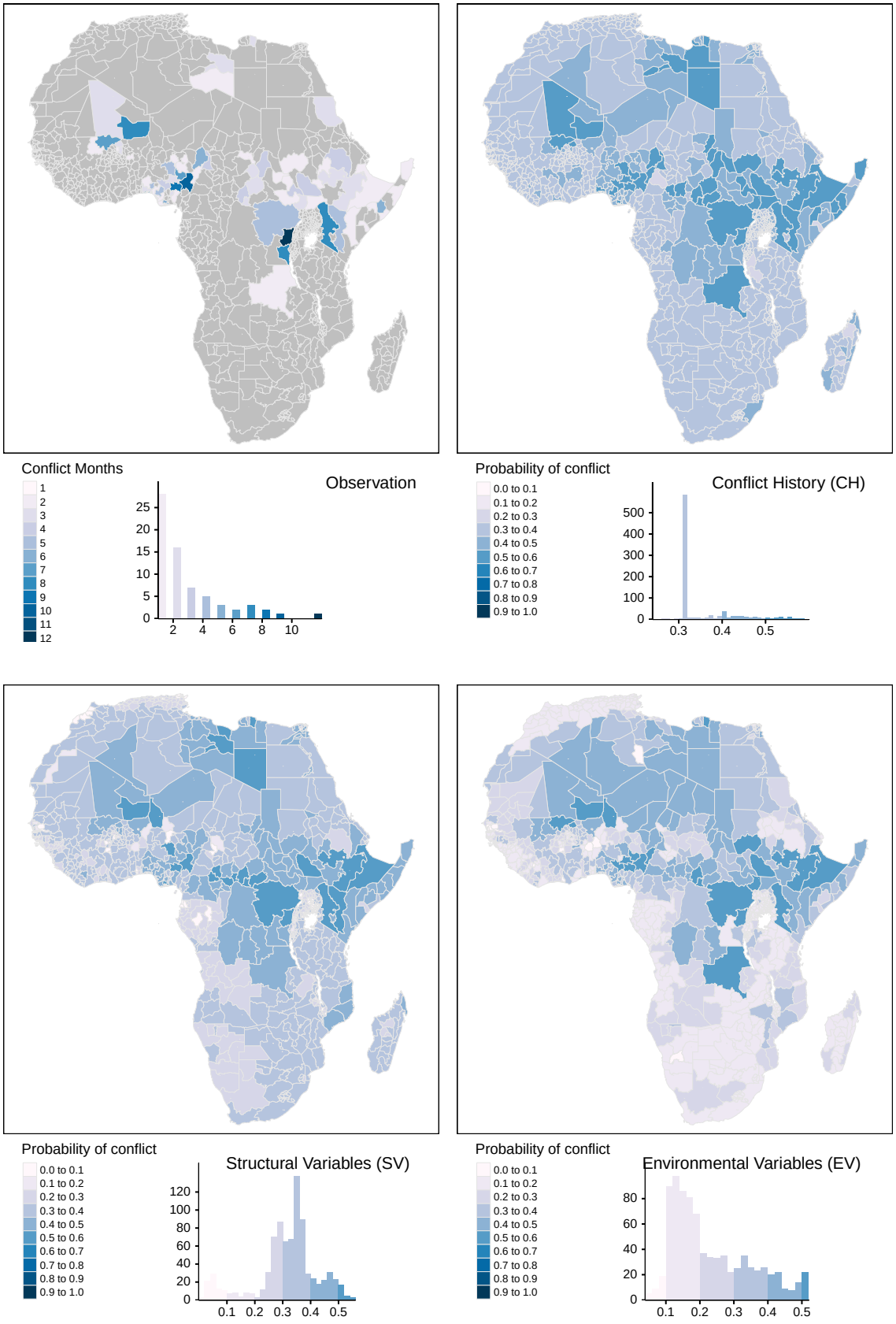


Figure A15: Spatial prediction of conflict class **ns** for *adm* districts.

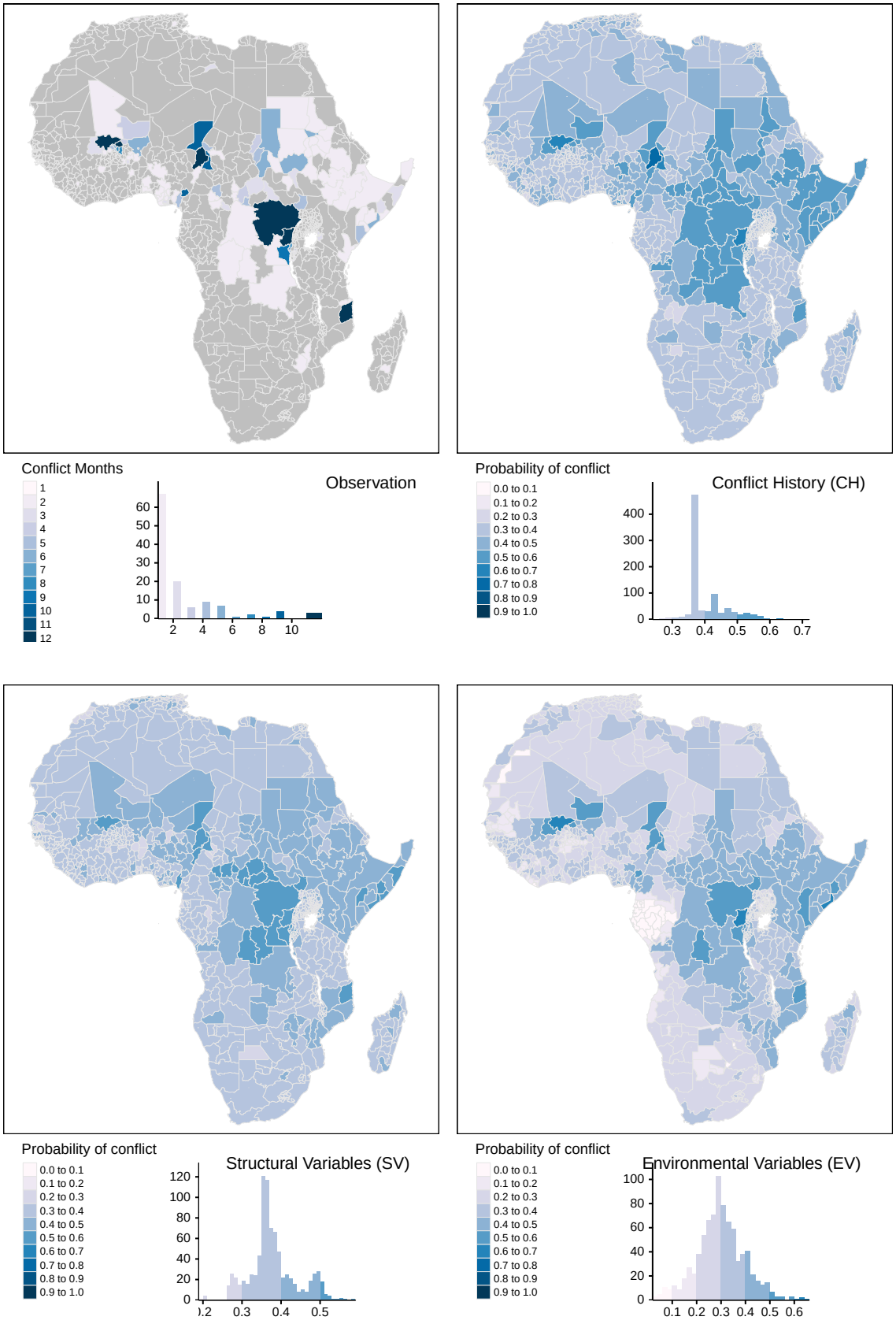


Figure A16: Spatial prediction of conflict class **os** for *adm* districts.

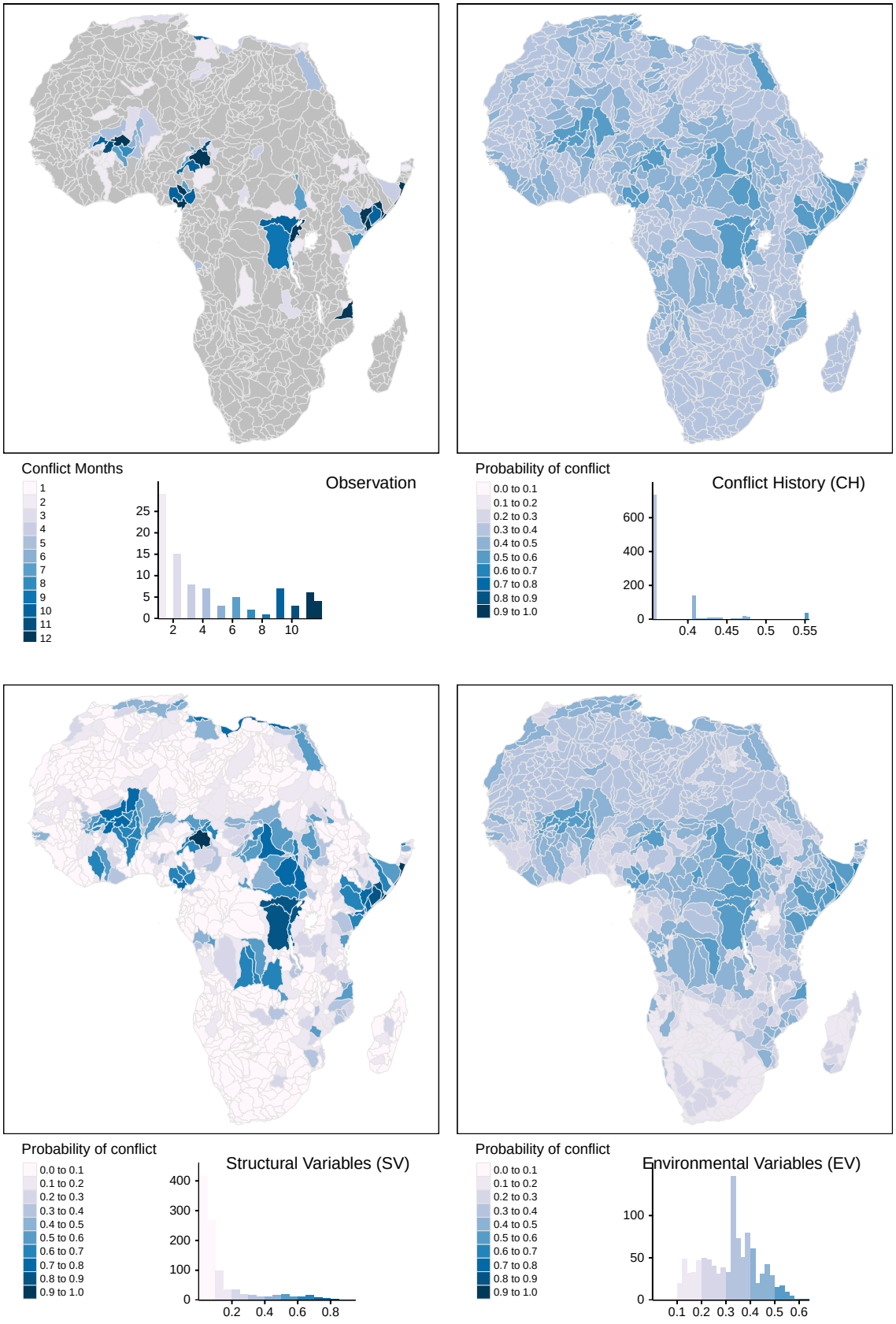


Figure A17: Spatial prediction of conflict class **sb** for *bas* districts.

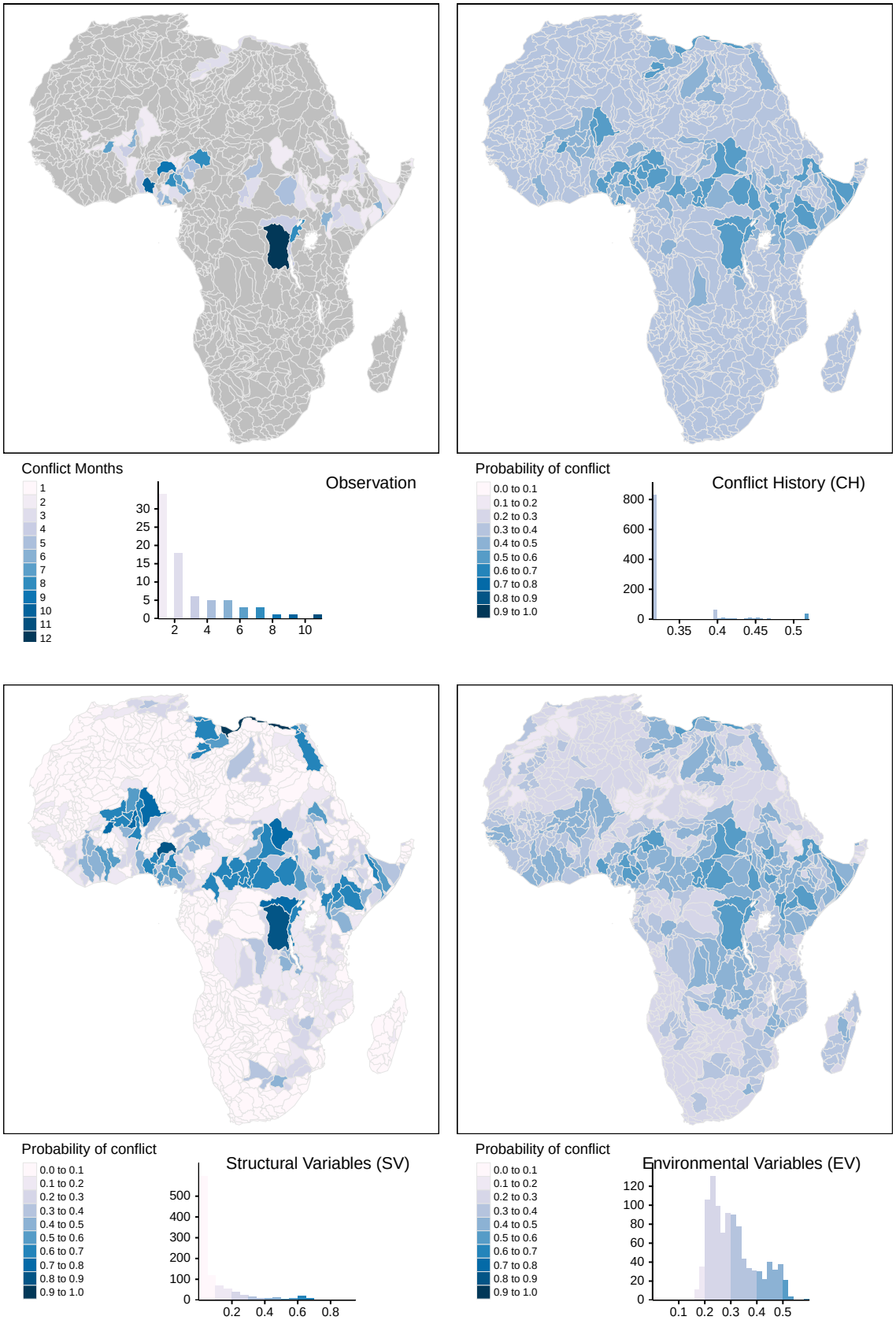


Figure A18: Spatial prediction of conflict class **ns** for *bas* districts.

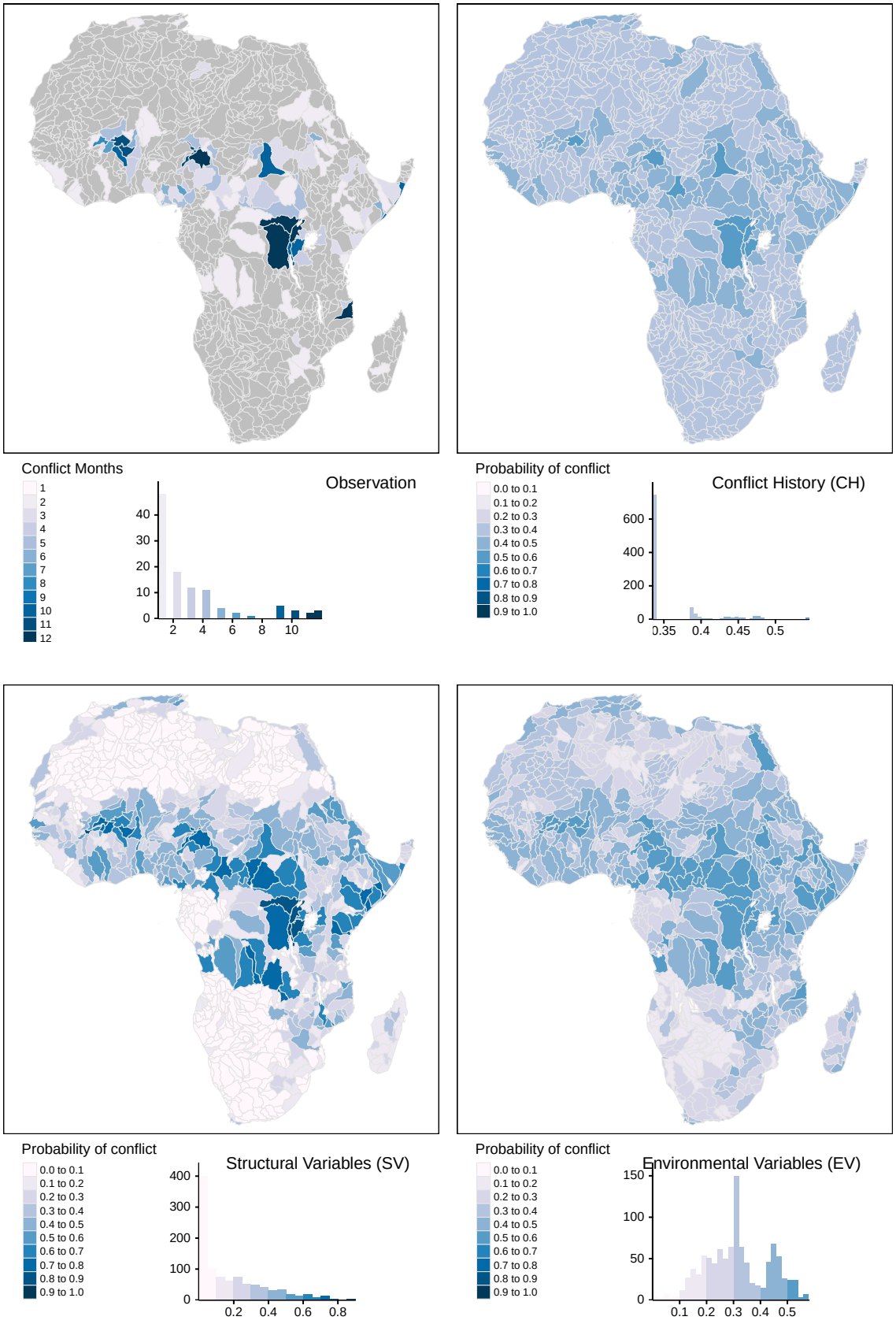


Figure A19: Spatial prediction of conflict class **os** for *bas* districts.

QQ-Plots and Residual Plots of Interaction Model

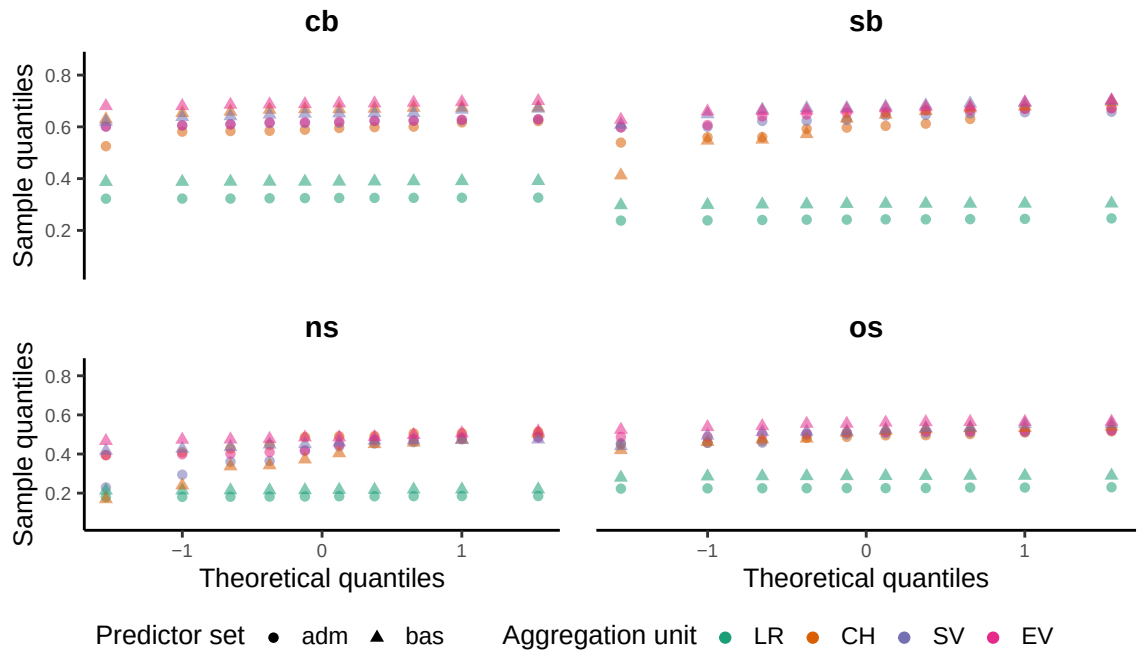


Figure A20: QQ-plots of the linear interaction model for the F2-score. Point shapes indicate the aggregation districts, colors the predictor set.

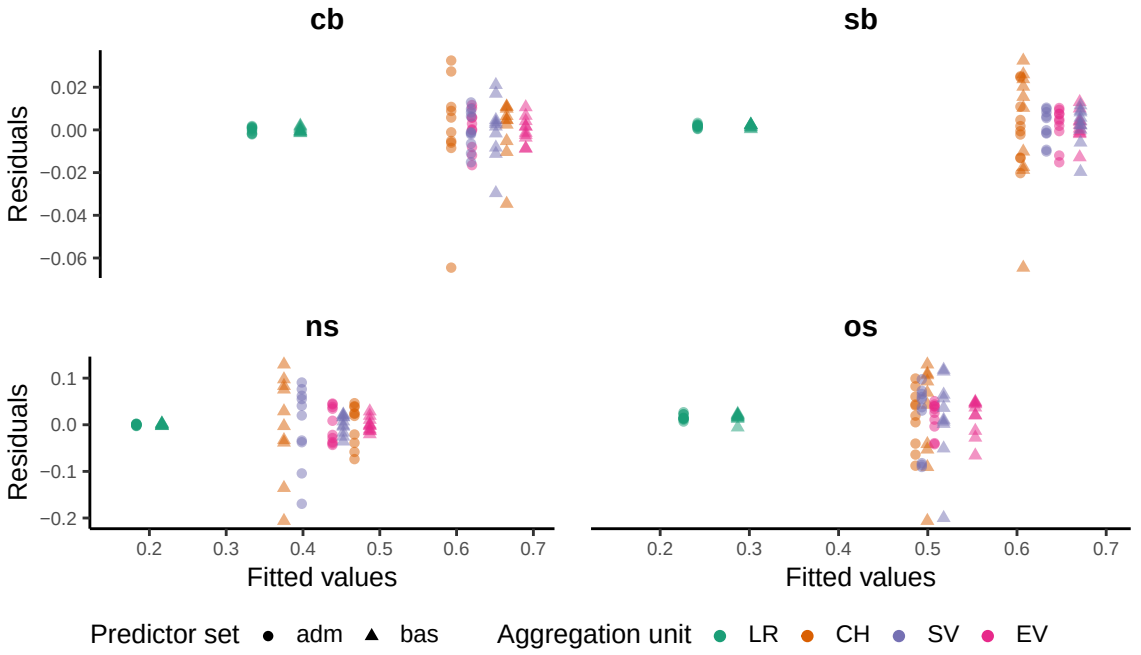


Figure A21: Residual plot of the linear interaction model for the F2-score. Point shapes indicate the aggregation districts, colors the predictor set.

Descriptive Statistics of Interaction Variables with Agricultural Mask

Table A1: Descriptive statistics of agricultural interaction variables. (Unit of measurment: AGRET - kg/m²; AGRGPP - kg C/m²; AGRLST - K; AGRPREC - mm AGRANOM - mm; others - dimensionless)

Spatial Unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
AGRET							
adm	0.000	0.253	9.996	118.725	138.828	1273.573	643
bas	0.000	0.000	0.647	28.715	12.693	973.495	39358
AGRGPP							
adm	0.000	0.491	20.158	232.349	272.386	6251.350	478
bas	0.000	0.000	1.337	55.510	27.987	2465.299	38957
AGRLST							
adm	0.000	0.137	8.428	60.278	104.179	312.276	0
bas	0.000	0.000	0.066	18.272	7.486	279.603	0
AGRPREC							
adm	0.000	0.003	0.526	17.237	13.174	429.346	0
bas	0.000	0.000	0.001	4.473	0.500	304.854	0
AGRANOM							
adm	-154.115	-0.148	0.000	0.535	0.050	265.607	0
bas	-138.119	-0.001	0.000	0.082	0.000	125.800	0
AGRSPI1							
adm	-3.554	-0.006	0.000	0.011	0.004	5.854	1100
bas	-3.471	0.000	0.000	0.003	0.000	4.494	2571
AGRSPI3							
adm	-4.829	-0.004	0.000	0.012	0.007	5.773	2654
bas	-2.411	0.000	0.000	0.003	0.000	3.517	3331
AGRSPI6							
adm	-3.935	-0.003	0.000	0.012	0.008	4.157	5180
bas	-2.596	0.000	0.000	0.003	0.000	3.035	6005
AGRSPI12							
adm	-4.750	-0.003	0.000	0.012	0.008	3.555	10233
bas	-2.481	0.000	0.000	0.003	0.000	2.726	12059
AGRSPEI1							
adm	-4.126	-0.005	0.000	0.007	0.006	3.306	973

Table A1: Descriptive statistics of agricultural interaction variables. (Unit of measurment: AGRET - kg/m²; AGRGPP - kg C/m²; AGRLST - K; AGRPREC - mm AGRANOM - mm; others - dimensionless) *(continued)*

Spatial Unit	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
bas	-2.446	0.000	0.000	0.002	0.000	2.427	2230
AGRSPEI3							
adm	-2.740	-0.005	0.000	0.008	0.006	5.006	2651
bas	-1.819	0.000	0.000	0.002	0.000	2.636	3310
AGRSPEI6							
adm	-2.423	-0.004	0.000	0.008	0.007	4.259	5183
bas	-1.856	0.000	0.000	0.002	0.000	2.730	6005
AGRSPEI12							
adm	-3.134	-0.004	0.000	0.007	0.007	3.558	10246
bas	-2.689	0.000	0.000	0.002	0.000	2.668	12059

Table of Global Performance Metrics

Table A2: Global performance metrics for all models configurations. Bold number indicate the best performance of the respective performance metric per conflict class. Standard deviation is given in brackets.

Theme	Unit	Variable	F2	AUC	AUPR	Precision	Sensitivity	Specificity
LR	*adm*	**cb**	0.325 (±0.001)	0.671 (±0.001)	0.128 (±0.001)	0.11 (±0.002)	0.668 (±0.007)	0.598 (±0.007)
LR	*adm*	**sb**	0.242 (±0.003)	0.703 (±0.003)	0.084 (±0.002)	0.08 (±0.002)	0.541 (±0.02)	0.751 (±0.014)
LR	*adm*	**ns**	0.183 (±0.002)	0.739 (±0.002)	0.049 (±0.001)	0.055 (±0.002)	0.468 (±0.032)	0.842 (±0.019)
LR	*adm*	**os**	0.226 (±0.002)	0.681 (±0.002)	0.072 (±0.001)	0.075 (±0.001)	0.494 (±0.02)	0.777 (±0.014)
LR	*bas*	**cb**	0.389 (±0.001)	0.758 (±0.001)	0.168 (±0.001)	0.141 (±0.003)	0.708 (±0.016)	0.687 (±0.014)
LR	*bas*	**sb**	0.302 (±0.002)	0.775 (±0.003)	0.121 (±0.002)	0.105 (±0.003)	0.592 (±0.019)	0.8 (±0.012)
LR	*bas*	**ns**	0.216 (±0.002)	0.779 (±0.003)	0.069 (±0.002)	0.073 (±0.003)	0.477 (±0.022)	0.867 (±0.013)
LR	*bas*	**os**	0.287 (±0.003)	0.774 (±0.003)	0.105 (±0.002)	0.1 (±0.002)	0.549 (±0.012)	0.83 (±0.007)
CH	*adm*	**cb**	0.59 (±0.027)	0.918 (±0.002)	0.6 (±0.008)	0.283 (±0.046)	0.824 (±0.041)	0.845 (±0.045)
CH	*adm*	**sb**	0.604 (±0.045)	0.943 (±0.006)	0.634 (±0.019)	0.297 (±0.069)	0.832 (±0.022)	0.923 (±0.023)
CH	*adm*	**ns**	0.467 (±0.044)	0.943 (±0.003)	**0.359 (±0.024)**	0.178 (±0.043)	**0.828 (±0.073)**	0.922 (±0.029)
CH	*adm*	**os**	0.486 (±0.022)	0.909 (±0.003)	0.472 (±0.01)	0.207 (±0.027)	0.743 (±0.031)	0.902 (±0.021)
CH	*bas*	**cb**	0.664 (±0.014)	0.942 (±0.013)	0.573 (±0.107)	0.395 (±0.084)	0.822 (±0.073)	0.916 (±0.034)
CH	*bas*	**sb**	0.607 (±0.087)	0.954 (±0.003)	0.518 (±0.048)	**0.552 (±0.08)**	0.635 (±0.125)	**0.982 (±0.01)**
CH	*bas*	**ns**	0.375 (±0.107)	0.92 (±0.047)	0.266 (±0.04)	0.335 (±0.054)	0.404 (±0.151)	**0.986 (±0.007)**
CH	*bas*	**os**	0.5 (±0.039)	0.925 (±0.006)	0.42 (±0.05)	0.295 (±0.114)	0.683 (±0.151)	0.937 (±0.048)
SV	*adm*	**cb**	0.616 (±0.009)	0.92 (±0.007)	0.599 (±0.011)	0.402 (±0.036)	0.715 (±0.037)	0.923 (±0.015)
SV	*adm*	**sb**	0.633 (±0.022)	0.947 (±0.006)	0.613 (±0.022)	0.455 (±0.044)	0.706 (±0.044)	0.968 (±0.008)
SV	*adm*	**ns**	0.398 (±0.085)	0.933 (±0.006)	0.322 (±0.021)	**0.345 (±0.09)**	0.45 (±0.155)	0.98 (±0.014)
SV	*adm*	**os**	0.493 (±0.026)	0.906 (±0.01)	0.426 (±0.016)	0.275 (±0.048)	0.638 (±0.1)	0.94 (±0.023)
SV	*bas*	**cb**	0.649 (±0.014)	0.948 (±0.003)	0.625 (±0.011)	0.32 (±0.021)	**0.876 (±0.016)**	0.886 (±0.013)
SV	*bas*	**sb**	**0.671 (±0.026)**	**0.963 (±0.004)**	0.617 (±0.028)	0.356 (±0.036)	**0.865 (±0.025)**	0.948 (±0.01)
SV	*bas*	**ns**	0.452 (±0.02)	0.94 (±0.005)	0.275 (±0.026)	0.191 (±0.022)	0.696 (±0.046)	0.95 (±0.009)
SV	*bas*	**os**	0.518 (±0.033)	0.941 (±0.006)	0.48 (±0.026)	0.241 (±0.041)	**0.747 (±0.074)**	0.928 (±0.031)

Table A2: Global performance metrics for all models configurations. Bold number indicate the best performance of the respective performance metric per conflict class. Standard deviation is given in brackets. *(continued)*

Theme	Unit	Variable	F2	AUC	AUPR	Precision	Sensitivity	Specificity
EV	*adm*	**cb**	0.618 (±0.009)	0.924 (±0.003)	0.609 (±0.007)	0.401 (±0.044)	0.72 (±0.048)	0.922 (±0.02)
EV	*adm*	**sb**	0.647 (±0.026)	0.947 (±0.005)	**0.65 (±0.012)**	0.394 (±0.072)	0.783 (±0.039)	0.953 (±0.015)
EV	*adm*	**ns**	0.438 (±0.038)	0.936 (±0.007)	0.298 (±0.058)	0.275 (±0.091)	0.565 (±0.144)	0.967 (±0.018)
EV	*adm*	**os**	0.508 (±0.012)	0.903 (±0.008)	0.431 (±0.02)	0.237 (±0.027)	0.719 (±0.042)	0.921 (±0.016)
EV	*bas*	**cb**	**0.689 (±0.006)**	**0.951 (±0.003)**	**0.639 (±0.021)**	**0.418 (±0.026)**	0.824 (±0.017)	**0.93 (±0.009)**
EV	*bas*	**sb**	0.67 (±0.021)	0.963 (±0.003)	0.616 (±0.021)	0.42 (±0.036)	0.791 (±0.038)	0.964 (±0.007)
EV	*bas*	**ns**	**0.487 (±0.016)**	**0.947 (±0.004)**	0.305 (±0.025)	0.249 (±0.028)	0.647 (±0.051)	0.967 (±0.007)
EV	*bas*	**os**	**0.553 (±0.014)**	**0.944 (±0.003)**	**0.5 (±0.011)**	**0.308 (±0.035)**	0.697 (±0.04)	**0.954 (±0.01)**

Declaration of Authorship

I hereby confirm that I have authored this Master's Thesis independently and without use of others than the indicated sources. The Master's Thesis has not yet been submitted to another university in its current or similar form and has not yet served any other examination purposes

Marburg, 24 March, 2021

. . . . . .

Darius A. Görgen

Declaration of Consent for the Inspection of the Thesis

◯ I agree that my thesis may be viewed by third parties in the department/university archives for scientific purposes.

◯ I do not agree that my thesis may be viewed by third parties in the department/university archives for scientific purposes.

Marburg, 24 March, 2021

. . . . . .

Darius A. Görgen