# High Resolution Conflict Forecasting with Spatial Convolutions and Long Short-Term Memory[*]

Benjamin J. Radford

March 17, 2022

[*] Data and replication files are available at https://dvn.iq.harvard.edu/dvn/dv/internationalinteractions. All questions regarding replication should be directed to the author.

**Abstract**

The 2020 Violence Early Warning System (ViEWS) Prediction Competition challenged participants to produce predictive models of violent political conflict at high spatial and temporal resolutions. This paper presents a convolutional long short-term memory (ConvLSTM) recurrent neural network capable of forecasting the log change in battle-related deaths resulting from state-based armed conflict at the PRIO-GRID cell-month level. The ConvLSTM outperforms the benchmark model provided by the ViEWS team and performs comparably to the best models submitted to the competition. In addition to providing a technical description of the ConvLSTM, I evaluate the model's out-of-sample performance and interrogate a selection of interesting model forecasts. I find that the model relies heavily on lagged levels of battle-related fatalities to forecast future decreases in violence. The model struggles to forecast escalations in violence and tends to underpredict the magnitude of escalation while overpredicting the spatial spread of escalation.

***Keywords***— Political conflict; forecasting; neural networks; machine learning

# 1 Introduction

The increasing availability of high-quality and high-resolution data about political violence has, in recent years, spurred interest in the use of machine learning to forecast conflict events. New datasets offer the finest-grained global-scale views into conflict events ever collected and made publicly available; these include the Armed Conflict Location and Event Dataset, the Integrated Conflict Early Warning System (ICEWS), Phoenix, and the many datasets produced by the Peace Research Institute Oslo (Raleigh et al., 2010; O'Brien, 2010; Althaus et al., 2020). The proliferation of these data has coincided with increased interest in conflict forecasting (D'Orazio and Yonamine, 2015; Hegre et al., 2017; Chiba and Gleditsch, 2017; Blair and Sambanis, 2020; Mueller and Rauh, 2020). The Violence Early Warning System, ViEWS, is one such project that endeavors to produce monthly forecasts of violence at the state and sub-state levels (Hegre et al., 2019). In 2020, the ViEWS team hosted its first ViEWS Prediction Competition, a shared task that solicited models capable of predicting changes in the severity of violent state-based conflict in monthly increments up to seven months in advance (Vesco et al., 2022*a*).

This paper describes one such model: a convolutional long short-term memory (ConvLSTM) recurrent neural network solution to the sub-state challenge of the 2020 ViEWS Prediction Competition. The sub-state challenge asks competitors to predict the change in logged levels of state-based violence at the PRIO-GRID cell-month level (Tollefsen, Strand, and Buhaug, 2012). The ConvLSTM model outperforms the ViEWS team's benchmark model in out-of-sample evaluations and performs among the best of all contestants' entries.

I proceed by first motivating and detailing the chosen modeling strategy. Next, I describe the model's performance on the various forecasting tasks. Then, I examine selected out-of-sample forecasts. Finally, the paper concludes with recommendations for future research on conflict forecasting.

## 2 Modeling Strategy

The PRIO-GRID cell-months, sometimes referred to as *pgm*, are each one-half degree longitude by one-half degree latitude. The competition requested predictions be made for every cell for 2, 3, 4, 5, 6, and 7 months in the future; these steps are indexed by $s \in \{2, \ldots, 7\}$.[1] The value of the target variable is $\ln(\texttt{ged\_best\_sb}_s + 1) - \ln(\texttt{ged\_best\_sb}_{t=0} + 1)$ where $t = 0$ denotes the time step from which forecasts are to be made. PRIO defines `ged_best_sb` as the best estimate of monthly battle deaths in state-based conflicts. For convenience, I use $\Delta_s \ln(\text{fatalities})$ as shorthand for the target values.

Previous literature on political conflict forecasting has highlighted the importance of temporal and spatial dependencies (Weidmann and Ward, 2010; Montgomery, Ward, and Hollenbach, 2011; Metternich, Minhas, and Ward, 2017). However, much of this work has

[1] I also make predictions for $s = 1$ and therefore sometimes refer to seven time steps.

relied on traditional regression modeling and, therefore, temporal and spatial lags were precomputed and entered into the model alongside other non-lagged covariates. I instead use a neural network that parameterizes the spatial and temporal lag structure; these lags are learned by the model on a per-feature basis. This strategy allows the model to forecast conflict based on nonlinear combinations of both spatially- and temporally-lagged predictors.

## *2.1 Motivation*

Inspiration for the chosen modeling solution is drawn by analogy from the conflict domain to the problem of next frame prediction in video processing (Lotter, Kreiman, and Cox, 2017). Representation of change over time and space in logged battle deaths is similar to that of video: one time axis, two spatial axes, and a features axis.[2] Framing the problem this way, the ViEWS Prediction Competition asks entrants, given a time-ordered series of images of Africa, to predict the subsequent seven "frames," or maps. In the video domain, the output features axis represents the same set of features as the input features axis (i.e. color channels). However, in the formulation of the conflict forecasting solution described here, the output features domain is change in logged fatalities, a feature not included in the input features set. This difference requires the model to translate between distinct input and output domains, distinguishing it from the typical next frame prediction task in which the input and output domains comprise the same features.[3]

## *2.2 Data Preparation*

The PRIO-GRID data include monthly observations of state-based conflict fatalities and associated covariates ("features") covering the continent of Africa. Only those cells that contain land are included and therefore do not form a full rectangular grid. I augment

[2] In video, the features axis may represent, for example, color channels. In our case, features are PRIO-GRID cell-level covariates.

[3] More specifically, the input comprises a tensor with dimensions representing time step, longitude, latitude, and features. The output comprises a tensor with dimensions representing longitude, latitude, and time leads, where the time leads correspond to $s \in \{1, ..., 7\}$.

these data by filling in all missing cell-months, those in the ocean, to produce a complete rectangular grid that includes all original PRIO-GRID cell-months. For each new cell-month, missing values are replaced with zero. An additional feature is added to every cell-month to indicate whether any missing feature value has been replaced with zero. The ViEWS team imputed missing PRIO-GRID data for all cells containing land prior to distributing the data.

I reshape the data into an array with dimensions corresponding to time steps, one-half degrees longitude, one-half degrees latitude, and features. A 12-month sliding window over the time dimension produces a new 12-month sequence for every outcome month from January 1990 through August 2020. The resulting array comprises 368 sequences, each of shape [12 months × 178 one-half degrees longitude × 169 one-half degrees latitude × 14 features]. These sequences are unique but contain overlapping observations; adjacent sequences contain 11 months' worth of identical feature values and one month each with differing feature values.

For every 12-month sequence of cell features, the corresponding target value is an array with dimensions comprising one-half degrees longitude, one-half degrees latitude, and monthly time steps. Therefore, a target value array is of shape $[178 \times 169 \times 7]$. A given value in this target array, $\Delta_{i,j,s} \ln(\text{fatalities})$, corresponds to the change in ln(fatalities) for the cell at half degree longitude $i$, half degree latitude $j$, and future time step $s$.[4] A single input sequence and output set is depicted in Figure 1. For the remainder of this paper, subscripts $i$ and $j$ are suppressed such that $\Delta_s$ will denote the changes in all cells at time step $s$.

The chosen features are a subset of those included in the ViEWS team's random forest benchmark model (Jansen et al., 2020). I have removed time-lagged predictors to reduce the memory requirements of the model; however, the time-lagged information remains in the model as any given observation is represented as a sequence of variable values over the preceding 12-month period. The remaining fourteen features are described in Table 1. The thirteen PRIO-GRID features included in the model are retained from the benchmark model

[4] Here ln(fatalities) is shorthand for $\ln(\texttt{ged\_best\_sb} + 1)$.

Table 1: Features included in predictive ConvLSTM model of $\Delta \ln(\text{fatalities})$.

| Variable | Description |
|---|---|
| `ln_ged_best_sb` | The natural log of the best estimate of monthly battle deaths from PRIO. |
| `pgd_bdist3` | Distance from grid centroid to the nearest land-contiguous national border, km. |
| `pgd_capdist` | Distance from grid centroid to national capital city, km. |
| `pgd_agri_ih` | Agriculture, percentage cell area. |
| `pgd_pop_gpw_sum` | Total cell population. |
| `pgd_ttime_mean` | Average travel time to nearest major city. |
| `spdist_pgd_diamsec` | Distance to secondary diamond deposits. |
| `pgd_pasture_ih` | Pasture, percentage cell area. |
| `pgd_savanna_ih` | Savanna, percentage cell area. |
| `pgd_forest_ih` | Forest, percentage cell area. |
| `pgd_urban_ih` | Urban, percentage cell area. |
| `pgd_barren_ih` | Barren, percentage cell area. |
| `pgd_gcp_mer` | Gross cell product, USD. |
| `missing_value_indicator` | Indicates one or more missing feature values have been assigned 0.0. |

after removing all time-lagged features and features that are not originally measured at the cell-month level.[5]

The data are partitioned into *training*, *validation*, and *test* sets according to the ViEWS Competition guidelines. The partitions include years 1990–2013 (*training set*), 2014–2016 (*validation set*), and 2017–2019 ($test_{17-19}$ *set*). Forecasts made on the validation and $\text{test}_{17-19}$ sets are referred to by the ViEWS team as task 3 and task 2, respectively. Task 1 refers to forecasts made for a fourth partition that includes the six months from October 2020 through March 2021 ($test_{20-21}$ *set*). The $\text{test}_{20-21}$ set represents a hard out-of-sample test in the sense that forecasts for these months were due to the competition organizers prior to October 2020.

[5] Due to an oversight, one cell-level feature included in the benchmark model was omitted from this analysis: `spdist_pgd_petroleum`, distance to the nearest petroleum resource. Additionally, PRIO provides measures of one-sided and non-state violence similar to the state-based violence measure utilized here. While these are included in the benchmark model, I do not include them in the competition entry model. However, I reestimate the ConvLSTM to include these additional predictors and discuss the results in Section 3.1.

## *2.3 Convolutional LSTM Neural Network*

ConvLSTM recurrent neural networks build on standard fully-connected feed-forward artificial neural networks by introducing two specialized layer types. The first is the spatial convolution. A neural network that contains convolutional layers is called a convolutional neural network (CNN). Spatial convolutions can be thought of as a set of filters that, for each pixel in an image, compute a weighted combination of adjacent or nearby pixel values. A convolution's window size is typically predetermined by the researcher but the convolution's weights are learned model parameters. Stacking convolutions in a series of layers such that the output of the first set of convolutions is the input to a later convolutional layer allows a CNN to learn convolutions that act as feature extractors at differing spatial scales.

ConvLSTMs further expand on traditional CNNs by incorporating a recurrent component called a long short-term memory (LSTM) cell (Hochreiter and Schmidhuber, 1997). LSTM layers allow neural networks to model temporal dependencies in sequential data. An LSTM layer takes as input the features from the current step of a sequence as well as the layer's own output from the previous step of the sequence. Input gates and forget gates, themselves learned parameters of the LSTM, control the ratio of current and past information retained in each sequence step. ConvLSTMs combine convolutional layers with LSTMs by accepting as input a 3-dimensional tensor, a sequence of images, and convolving across the two spatial dimensions (Shi et al., 2015). Therefore, ConvLSTMs are able to model local spatial dependencies over time: the prediction for a given cell (i.e. pixel) is a function of its own past feature values and its neighbors' past feature values.

The model described below is a deep neural network comprising several layers. The inputs to subsequent layers are the outputs ("activations") of preceding layers. I refer to this model as a ConvLSTM for simplicity, but it actually comprises a series of convolutional and ConvLSTM layers.

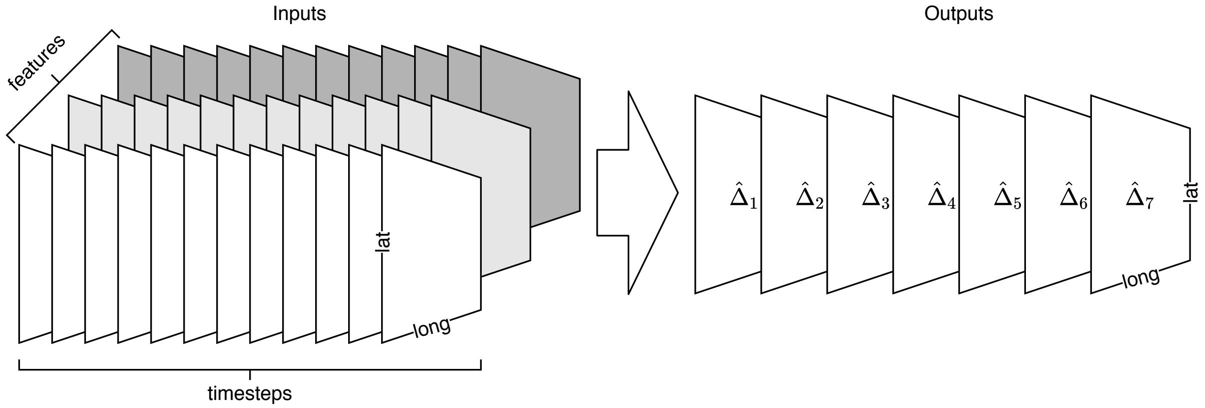


Figure 1: Input and output data shapes. A single input example, $X_i$, consists of a time series of feature maps: twelve time steps (months) of maps per feature. There are 14 unique features. The output is a set of seven grid cell maps: one per $s \in \{1, ..., 7\}$.

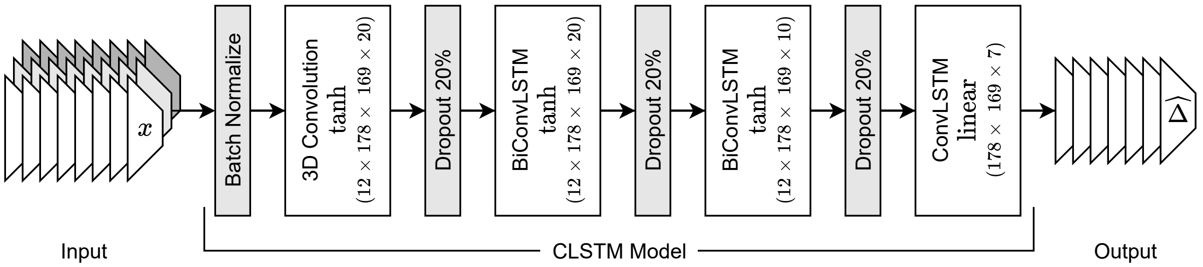


Figure 2: Model architecture diagram. Layer output sizes, per training sample, are given in parentheses. Activation functions, when appropriate, are given underneath the layer name. The input diagram does not depict the full set of features or time steps in a given training sample. Regularization layers are shaded gray.

## *2.4 Model Description*

The primary model described herein is called the *competition entry* model to indicate that predictions generated by this model were submitted to the competition. The competition entry model consists of a 3-dimensional convolutional layer followed by two bidirectional ConvLSTM (BiConvLSTM) layers and a final ConvLSTM layer. Inputs are batch normalized and 20% weight dropout is applied between all layers to help prevent overfitting the model to the training data (Ioffe and Szegedy, 2015; Srivastava et al., 2014). I do not apply any data augmentation techniques.[6]

The first convolutional layer comprises 20 filters, each of size $[1 \times 1 \times 14]$. Therefore, the same 20 convolutions are learned and applied to all cells for all time steps. Convolution weights are applied across features but not across space at this point (hence the unit values in the longitude and latitude dimensions). This can be thought of as a feature selection step: 20 features are computed for each cell-month. The computed features are linear combinations of the 14 features associated with each given cell-month. These computed values are then bound between $(-1, 1)$ with the help of the hyperbolic tangent activation function.

The two BiConvLSTM layers consist of 20 and 10 convolutional filters each, with kernel sizes of $[10 \times 10]$ and $[5 \times 5]$, applied over the spatial dimensions. Therefore, the first set of convolutions corresponds to 5° of latitude and longitude while the second corresponds to 2.5° of latitude and longitude. These layers are bidirectional, meaning that the model is trained by reading sequences in both the forward and backward time directions (Schuster and Paliwal, 1997). Each of the first three layers returns a sequence of the same spatial and temporal shape as the input.

The final layer of the model is a ConvLSTM comprising 7 filters. The model output shape is $[178 \times 169 \times 7]$, corresponding to seven "images" of the spatial grid, one per forecast time

[6] Dropout and batch normalization are often considered forms of regularization, though dropout is also equivalent to data augmentation under certain circumstances (Wager, Wang, and Liang, 2013; Zhao et al., 2019). Nevertheless, data augmentation applied to the input images (e.g., skew or rotation) may result in improved predictive performance and could be a promising avenue for future work (Shorten and Khoshgoftaar, 2019).

step. No activation function is applied to the output (i.e. linear activation) as the target values are real-valued and unbounded. The full model architecture is depicted in Figure 2.

The model itself contains 281,016 trainable parameters. I use RMSprop to fit the model and mean squared error (MSE) to compute model loss (Hinton, Srivastava, and Swersky, 2012). The model is trained with a batch size of 8 for 75 epochs on a single RTX 2080 Ti GPU.[7] The model is fit to only the first 270 12-month sequences (covering January 1990–June 2013) to avoid training on the validation and test sets. However, these 270 training sequences contain 113,709,960 unique input pixel values (not including target values). A single frame from a training example, one time slice, comprises $178 \times 169 \times 14 = 421,148$ unique values. I adjusted the model architecture, including the number and types of layers, dropout percentage, and training hyperparameters, in response to out-of-sample performance on the validation set. The test sets, those that include 2017–2019 and 2020–2021, were unobserved during model training, including while tuning, and were only utilized when final predictions were made for the contest.

### *2.5 What Doesn't the Model Know?*

This model does not have any absolute information about spatial or temporal location; no features akin to spatial or temporal fixed effects (e.g., country, season, or year) are included. Furthermore, no country-specific covariates are included. Distance to the nearest country border is included, though, and the model may learn to identify conflict-prone areas by the nearest border contours and unique geographic features.[8] However, including absolute identifiers for the grid cell, country, or time-unit would be trivial and may lead to improved predictive performance. Additional features measured at the cell-month level, or at higher-levels of aggregation, would be included by extending the features axis of the input arrays described in Section 2.2.

[7] Training time is approximately 1.5 hours.

[8] Though this is not shown here and is left for future work.

Table 2: Out-of-sample predictive performance on validation and test sets.

| | Competition Entry | | | | | | Benchmark | |
|---|---|---|---|---|---|---|---|---|
| | Validation | | $\text{Test}_{17-19}$ * | | $\text{Test}_{20-21}$ | | $\text{Test}_{20-21}$ | |
| Steps | MSE | TADDA | MSE | TADDA | MSE | TADDA | MSE | TADDA |
| $s = 2$ | 0.021 | 0.014 | 0.022 | 0.016 | 0.024 | 0.018 | 0.053 | 0.138 |
| $s = 3$ | 0.021 | 0.014 | 0.022 | 0.016 | 0.028 | 0.019 | 0.059 | 0.142 |
| $s = 4$ | 0.021 | 0.014 | 0.022 | 0.017 | 0.025 | 0.020 | 0.052 | 0.142 |
| $s = 5$ | 0.022 | 0.014 | 0.023 | 0.016 | 0.036 | 0.024 | 0.064 | 0.150 |
| $s = 6$ | 0.021 | 0.014 | 0.023 | 0.017 | 0.036 | 0.021 | 0.063 | 0.148 |
| $s = 7$ | 0.022 | 0.015 | 0.023 | 0.018 | 0.035 | 0.025 | 0.064 | 0.153 |

*The competition organizers report slightly different MSE and TADDA values (means: MSE=0.024, TADDA=0.018).

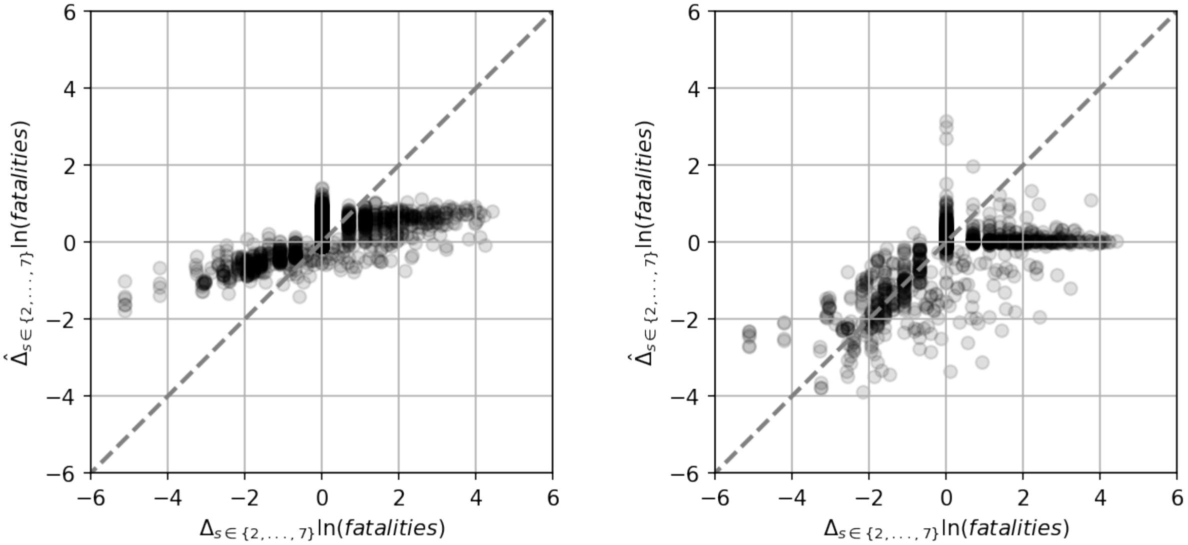
# 3 Model Evaluation

## *3.1 Test Partition Evaluation*

Model performance on the out-of-sample validation and test sets is given in Table 2. The MSEs for each time step fall between 0.020 and 0.022 for the validation set and between 0.021 and 0.023 for the $\text{test}_{17-19}$ set. This corresponds to an approximate 30% reduction in MSE when compared to the benchmark model offered by the ViEWS team in their preliminary evaluation. TADDA values for both sets fall between 0.013 and 0.018, a substantial improvement over the ViEWS team's benchmark model at 0.14. TADDA stands for "Targeted Absolute Distance with Direction Augmentation" and is calculated using the implementation found in the `OpenViEWS2` software package.[9] This metric was introduced specifically for the ViEWS Prediction Competition by Vesco et al. (2022*a*,*b*). TADDA is the mean absolute deviation between observed and predicted values with an added L1 penalty for predictions made in the wrong direction; lower values of TADDA indicate better model performance.[10] MSE and TADDA scores for the $\text{test}_{20-21}$ set are provided both for the competition entry model and the benchmark model. The competition entry model performs worse on the

[9] `https://github.com/UppsalaConflictDataProgram/OpenViEWS2`.

[10] TADDA $= (\sum_{i=1}^{N} |\Delta_i - \hat{\Delta}_i| + t_d)/N$ where $t_d$ is a penalty, $d$ represents a chosen method for handling near-zero values, and $\epsilon$ (not shown) defines "near zero." $\epsilon = 1$ is chosen following the default in `OpenViEWS2`.



(a) Benchmark model.

(b) Competition entry model.

Figure 3: Observed versus predicted change in ln(fatalities) in $\text{test}_{20-21}$ set.

$\text{test}_{20-21}$ set than it does on the $\text{test}_{17-19}$ set, especially for longer time leads. However, the competition entry model bests the benchmark model by 50% in MSE and by nearly an order of magnitude in TADDA. Of the eight total entries to the sub-state prediction task, the competition organizers ranked this ConvLSTM among the top two highest performers, with the other being the model by D'Orazio and Lin (2022) (Vesco et al., 2022*a*).

Subsetting the $\text{test}_{20-21}$ set to only those cell-months that experienced violence escalation, the competition entry model predicted an escalation of equal or greater magnitude only 0.7% of the time. The benchmark model did so for 3.2% of observations. However, when looking only at those cell-months that experienced de-escalations, the competition entry model predicts de-escalations of equal or greater magnitude 43.8% of the time, compared to the benchmark model's 4.8%. This is apparent in Figure 3. The benchmark model tends to underpredict both escalations and de-escalations while the competition entry model appears to perform better with respect to predicting the magnitude of de-escalation.

Forecast accuracy decreases as time steps increase, but not monotonically so. This may be due to the fact that all differences are taken with respect to the current time step, $t = 0$,

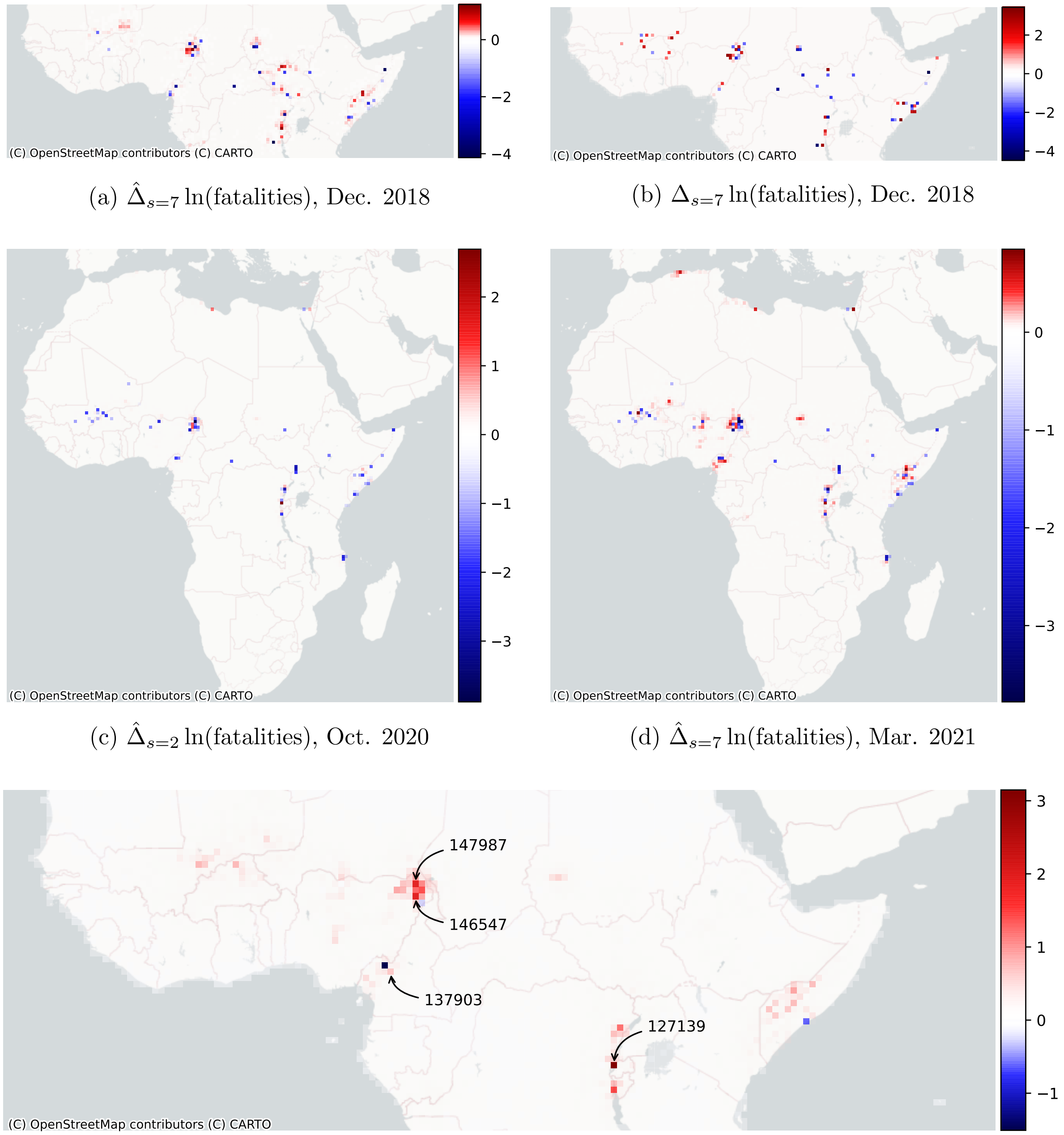
Table 3: Estimated average feature importance based on self-attention layer.

| Feature | Importance |
|---:|---:|
| ln_ged_best_sb | 0.284 |
| pgd_pop_gpw_sum | 0.271 |
| pgd_urban_ih | 0.207 |
| pgd_ttime_mean | 0.051 |
| pgd_agri_ih | 0.040 |
| pgd_gcp_mer | 0.035 |
| pgd_forest_ih | 0.029 |
| spdist_pgd_diamsec | 0.017 |
| pgd_barren_ih | 0.016 |
| pgd_bdist3 | 0.014 |
| pgd_savanna_ih | 0.012 |
| pgd_pasture_ih | 0.011 |
| missing_indicator | 0.010 |
| pgd_capdist | 0.010 |

not with respect to the preceding time step, $s-1$. Therefore, a sequence of ln(fatalities) that looks like $(1,0,0,0,0,0,0,0)$ for time steps $t=0$ through $s=7$ would produce $\Delta_{s\in\{1,\ldots,7\}}$ of $(-1,-1,-1,-1,-1,-1,-1)$. This may minimize the propagation of errors over time as bad predictions at $s=2$ would compound with future prediction errors were the predicted differences always taken with respect to the preceding month (rather than month $t=0$).

Figures 4 (a) and (b) depict predicted versus observed values for a portion of Africa in December 2018 (from test$_{17-19}$). The gradient legends reveal that the model underpredicts increases in violence; predicted increases fall far below observed increases in magnitude. However, the model predicts low level increases in violence for many more cells than actually experience increases in violence, highlighting the difficulty of predicting conflict onset and escalation. Forecasts of decreases in violence appear to be more spatially constrained, presumably because the model relies on the existence of violence at $t=0$ to predict decreases in later time steps. Despite the spatial overprediction of increases in violence, those predictions do appear to cluster around actual violence hot spots.

In a second model, I add a softmax attention layer (Bahdanau, Cho, and Bengio, 2015)



(a) $\hat{\Delta}_{s=7}\ln(\text{fatalities})$, Dec. 2018

(b) $\Delta_{s=7}\ln(\text{fatalities})$, Dec. 2018

(c) $\hat{\Delta}_{s=2}\ln(\text{fatalities})$, Oct. 2020

(d) $\hat{\Delta}_{s=7}\ln(\text{fatalities})$, Mar. 2021

(e) $\max(\hat{\Delta}_{s\in[2,\ldots,7]}\ln(\text{fatalities}))$, Oct. 2020–Mar. 2021

Figure 4: **Top row:** Predicted (a) and observed (b) values for $\Delta_{s=7}\ln(\text{fatalities})$ for December 2018. **Middle row:** Predicted values for $\Delta\ln(\text{fatalities})$ for the PRIO-GRID months of October 2020 (c) and March 2021 (d). **Bottom row:** Maximum predicted values for every cell in the period from October 2020 until March 2021 are depicted in (e). Color values are scaled per map. The zero value is white for all maps.

Table 4: Out of sample predictive performance for the single feature and expanded predictors models. All performance metrics given for the $\text{test}_{17-19}$ set.

| | Expanded Features | | Single Feature | |
|---|---|---|---|---|
| Steps | MSE | TADDA | MSE | TADDA |
| $s = 2$ | 0.022 | 0.016 | 0.022 | 0.013 |
| $s = 3$ | 0.022 | 0.016 | 0.022 | 0.013 |
| $s = 4$ | 0.022 | 0.016 | 0.022 | 0.014 |
| $s = 5$ | 0.023 | 0.016 | 0.022 | 0.013 |
| $s = 6$ | 0.023 | 0.016 | 0.022 | 0.013 |
| $s = 7$ | 0.023 | 0.016 | 0.022 | 0.014 |

to the normalized input and prior to the first convolutional layer.[11] I then compute feature importance scores by averaging each feature's activation values on the attention layer for all cell-months. Estimated average feature importance scores are given in Table 3. Logged fatalities, population, and urban status appear to be the most influential (highly-weighted) predictors.

In fact, when the competition entry model is reestimated with only a single feature, ln(fatalities), the scores are similar to or better than those achieved by the full model. The single feature model achieves mean MSE and TADDA scores of 0.021 and 0.011 on the validation set and 0.022 and 0.013 on the $\text{test}_{17-19}$ set. Out-of-sample performance metrics for the single feature model are given in Table 4. Similarly, adding in three additional features that were included in the benchmark model but were not included in the competition entry model, non-state violence, one-sided violence, and the distance to the nearest petroleum resource, results in nearly identical performance to the competition entry model. Test set scores for this model, called the expanded features model, are also given in Table 4. These results reinforce the importance of ln(fatalities) as a predictor of $\Delta_s \ln(\text{fatalities})$ relative to all other evaluated predictors.

[11] MSE for the model that includes the attention layer is nearly identical to the model without the attention layer.

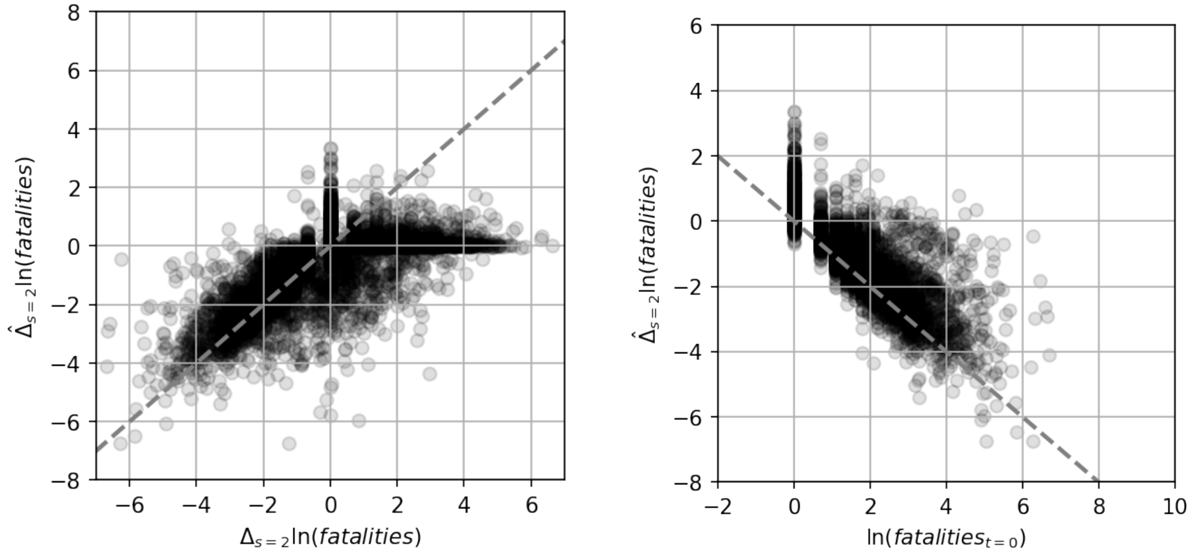
## *3.2 What Does This Model Tell Us About Conflict?*

Even with the relative black box nature of neural networks, much less ConvLSTMs, investigation of the model's predictions can reveal information about conflict dynamics.

Figure 5 (a) depicts predicted values of $\Delta_{s=2}\ln(\text{fatalities})$ versus observed values. When observed values are negative, predicted values are strongly correlated with the observed values. In other words, if there exist non-zero fatalities at $t = 0$, the model generally predicts a decrease in fatalities at $s = 2$. Given the presence of violence, de-escalation is simply more likely than further escalation, so much so that the model generally favors negative predicted values. However, when the observed value is zero or greater, the model often (but not always) predicts no change. This can be seen in the bunching of predicted values around $y = 0$ when $x > 0$. The values at $x = 0$ and $y > 0$ represent predicted escalations that did not occur. Overall, the model appears to better predict de-escalation than it does escalation. This probably reflects the fact that violence onset and escalation are relatively rare and that the available features contain little to no leading signal of escalation.

Figure 5 (b) further illustrates the nonlinear relationship between ln(fatalities) at $t = 0$ and $\hat{\Delta}_{s=2}\ln(\text{fatalities})$. When observed ln(fatalities) values are greater than 0, the model often predicts a corresponding negative change in fatalities. In other words, the model predicts that the number of fatalities will revert to zero over time. However, the amount by which fatalities change is sometimes less than the observed number of fatalities, perhaps indicating that reversion to zero fatalities does not always happen within $s = 2$ steps. For lower magnitude positive fatality values, the model sometimes even predicts further increases; this can be seen in the portion of the graph bound by (0,0) and (6,4). The predictions for that lie above $\hat{\Delta}_{s=2}\ln(\text{fatalities}) = 0$ represent predicted conflict escalations. When there are no logged fatalities at $t = 0$, the model has learned to predict 0 or a positive value; fatalities can only stay the same or increase.[12] Predicted increases from $\ln(\text{fatalities}) = 0$ represent

[12] There are some negative predicted values when logged fatalities are 0, but these are very small in magnitude.



(a) Actual vs. predicted $\Delta_{s=2}\ln$(fatalities).

(b) ln(fatalities) at $t = 0$ vs. $\hat{\Delta}_{s=2}\ln$(fatalities).

Figure 5: Predicted versus observed values at $s = 2$ (left) and predicted values at $s = 2$ versus observed ln(fatality) values at $t = 0$ (right). Dashed lines indicate $\hat{\Delta}_{s=2}\ln(\text{fatalities}) = \Delta_{s=2}\ln(\text{fatalities})$ (left) and $\hat{\Delta}_{s=2}\ln(\text{fatalities}) = \ln(\text{fatalities}_{t=0})$ (right), the expected change in fatalities if the model simply predicted the negative of the logged casualty count at time $t = 0$. Plots include all cell-months in the validation set and test$_{17-19}$ set .

conflict onsets.

# 4 Forecasts

## *4.1 Notable Predicted Escalations*

The single largest predicted increase in violence for the period from October 2020 through March 2021 is $\hat{\Delta}_{s=4}\ln(\text{fatalities}) = 3.148$ obtained by PRIO-GRID cell 127139. This is expected in December 2020, though similarly large increases are also predicted for the same cell in October 2020 and November 2020. This particular cell is located in Rwanda and is near the border with Uganda and the Democratic Republic of the Congo.[13] Figure 4 (e) depicts the maximum predicted change in log fatalities for every cell in the equatorial region of Africa from October 2020 through March 2021.

Two cells in Nigeria and near the border with Cameroon, 146547 and 147987, obtain large predicted increases in violence, both with expected changes in ln(fatalities) of between 1 and 2. These are predicted to occur in November 2020 and March 2021, respectively. Both cells experienced escalations in violence, relative to October 2020, in the expected months. These are included in *SetA*, a collection of cell-months described below and of particular interest to the ViEWS Competition organizers.

There is also evidence that the model doesn't rely solely on past violence within a given cell when predicting future violence for that cell. One cell, 137903, obtains a predicted increase in logged battle deaths of $\hat{\Delta}_{s=7}\ln(\text{fatalities}) = 0.695$ where no previous increase in logged battle deaths of similar magnitude had been recorded prior to the October 2020–March 2021 period. This grid cell is located in Cameroon and the increase is predicted to occur in November 2020. Because this prediction is not the result of past violence within this particular cell, it is likely due instead to nearby cells that experienced relatively high levels of violence in August 2020. However, this escalation did not occur.

[13] *pgm* 127139 did not experience any escalation in violence during the period of $\text{test}_{20-21}$.

## *4.2 Selected Cases*

Organizers of the ViEWS Prediction Competition requested that a selection of cell-months on the borders of Nigeria, Cameroon, and Chad as well the border between Mozambique and Tanzania be assessed specifically for the months of October 2020 through March 2021. These will be referred to as *SetA* and *SetB*, respectively. I explored many techniques for model inspection, but none were obviously appropriate for this particular request. Example-based explanation methods often estimate some form of feature importance based on a single given observation (i.e. data point) (Lundberg and Lee, 2017). In the case of the ConvLSTM, because a single data point corresponds to a $[178 \times 169 \times 7]$ array, many of these techniques would report average feature scores over the entire Africa grid for seven time steps. Local Interpretable Model-agnostic Explanations (LIME) allows for estimating feature importance in an image at the pixel-level, but it is not obvious how this would be leveraged to provide cell-specific feature weights in a ConvLSTM model with both temporal and spatial dependencies (Ribeiro, Singh, and Guestrin, 2016).

Instead, I simulate a set of interesting counterfactual examples and compare the model predictions for the chosen cell-months to predictions based on the associated counterfactual examples. I first create two counterfactual sets: *within set* counterfactuals and *outside set* counterfactuals. For *within set* counterfactuals, given features are replaced with their overall median values for the cells in *SetA* and *SetB*. For the *outside set* counterfactuals, the selected features are replaced with their overall median values for all cells not included in either *SetA* or *SetB*. Due to computational constraints, all cells of interest are treated at once rather than cell-by-cell. The *within set* counterfactuals represent changes to the time series of features within the selected cells of interest. The *outside set* counterfactuals correspond to the spatial effects of each feature from the cells not in the set of cells of interest. For each of the 13 model features, not including the missing value indicator, a set of predictions is generated on the counterfactual example created by replacing the values of that feature with the feature's

Table 5: Pearson's correlation coefficient between counterfactual examples and predictions with the full feature set (i.e., no features having been replaced). *Within set* indicates that variables in the counterfactual examples are replaced with their overall median values for the grid cells within *SetA* and *SetB*; *outside set* indicates that the variables in the counterfactual examples are replaced with their overall median values for all grid cells except those in *SetA* or *SetB*.

| | *SetA* | | *SetB* | |
|---|---|---|---|---|
| Feature | Within | Outside | Within | Outside |
| `ln_ged_best_sb` | -0.157 | 0.862 | 0.036 | 0.978 |
| `pgd_bdist3` | 0.810 | 0.809 | 0.975 | 0.972 |
| `pgd_capdist` | 0.805 | 0.792 | 0.967 | 0.979 |
| `pgd_agri_ih` | 0.811 | 0.801 | 0.974 | 0.968 |
| `pgd_pop_gpw_sum` | 0.802 | 0.807 | 0.973 | 0.978 |
| `pgd_ttime_mean` | 0.804 | 0.842 | 0.974 | 0.974 |
| `spdist_pgd_diamsec` | 0.807 | 0.798 | 0.969 | 0.974 |
| `pgd_pasture_ih` | 0.803 | 0.813 | 0.973 | 0.976 |
| `pgd_savanna_ih` | 0.808 | 0.824 | 0.969 | 0.977 |
| `pgd_forest_ih` | 0.805 | 0.805 | 0.973 | 0.972 |
| `pgd_urban_ih` | 0.803 | 0.819 | 0.973 | 0.977 |
| `pgd_barren_ih` | 0.797 | 0.823 | 0.973 | 0.971 |
| `pgd_gcp_mer` | 0.804 | 0.803 | 0.974 | 0.976 |

median value in the entire dataset.[14]

Exploring these counterfactual predictions may help to reveal those features the model relies on most heavily given the cases in question, but these should not be interpreted as causal effects. Furthermore, the change in logged fatalities may be overdetermined by the given data and dropping variables from the model does not necessarily tell us whether the remaining set of variables would be unable to match the full model's predictive power were it to be re-trained using only those remaining variables. What these counterfactual examples do reveal is the full model's dependence on a given feature when making predictions for the given set of cell-months.

The results of this counterfactual analysis are given in Table 5. The table reports the Pearson's correlation coefficient ($\rho$) values between the predictions made with the full set of features and those made with the set in which a given feature has been replaced by its median

[14] This method was developed specifically for this paper but bears some resemblance to the occlusion sensitivity method described by Zeiler and Fergus (2014).

value. The table makes clear that the model relies heavily on the past logged battle deaths within *SetA* and *SetB*. When `ln_ged_best_sb` is replaced with its median value for those cells within *SetA* and *SetB*, the resulting predictions are essentially uncorrelated with the original predictions. No other variables appear to have such a substantial impact on the model's predictions when replaced with their median values. For *SetA*, all other counterfactual examples, both *within set* and *outside set*, result in predictions that are highly correlated with the original predictions ($\rho \approx 0.80$). This is even more true for *SetB* in which any single median substitution results in predictions that are nearly perfectly correlated with the original predictions ($\rho \approx 0.97$). However, while these counterfactuals should highlight the relative importance of past battle deaths within the cells of interest to forecasts of future changes in violence, they only tell us that this feature is a necessary predictor, not that it is sufficient on its own. Interestingly, replacing `ln_ged_best_sb` with its median value for all cells not in *SetA* or *SetB*, the *outside set* counterfactual, appears to make near the least amount of difference, as measured by $\rho$, among all features. This may change if the analysis were to be redone cell-by-cell rather than simultaneously for all cells in both *SetA* and *SetB*.

Line graphs depicting the time series of observed and predicted changes in ln(fatalities) for all grid cells between July 2017 and March 2021 are given in Figure 6. The line graphs depict values for $s = 7$, seven-month differences in logged fatalities. Values before September 2020 are observed while values after that point are predicted based on features from at least seven months prior. Each line represents a single cell over time. The graphs provide further evidence that the model underpredicts conflict escalation—very few the forecasts after September 2020 extend far above $y = 0$. This is despite the fact that escalations are not particularly rare in *SetA* or *SetB* during the preceding months. On the other hand, the model predicts large decreases in violence on the same order of magnitude as those observed prior to September 2020.

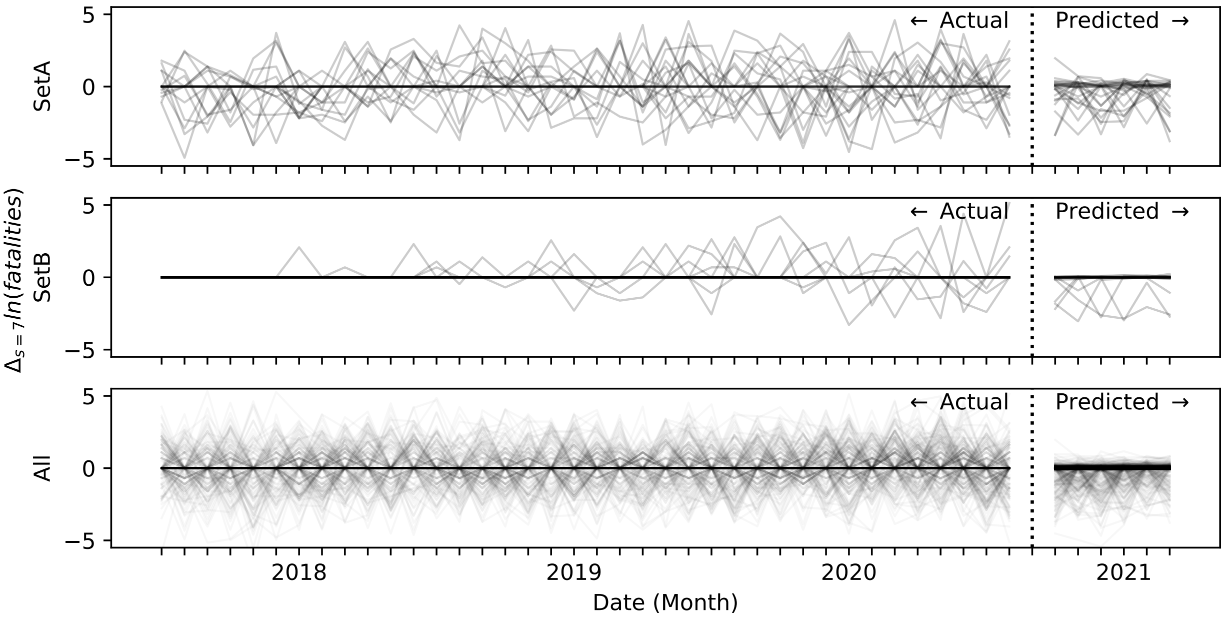


Figure 6: $\Delta_{s=7}\ln(fatalities)$ over time for *SetA* (top), *SetB* (middle), and all grid cells (bottom). The apparent diamond pattern in the bottom graph is an artifact due to the semi-transparent color and the time trends having been removed from the time series by differencing. All values are seven-month lagged differences. September 2020 is omitted because that month's `ln_ged_best_sb` was not available at the time that predictions were made and participants were not asked to forecast that month.

# 5 Conclusion

The analysis here highlights the difficulty of predicting conflict onset. Onsets, escalations in violence from zero, are challenging due to their rarity and the lack of time-varying predictors that reliably anticipate the initiation of violence. Most predictors used in this paper, for example, typically see substantial change only over many years, not months. The primary exception to this is the lagged battle deaths estimate which is particularly helpful for anticipating conflict de-escalation. When lagged fatalities are zero, the only direction for change in fatalities to go is up. However, due to the extreme rarity of onset, models seem to predict very small escalations at most; zero lagged fatalities is not strongly predictive of substantial escalation in fatalities at any particular time. Inclusion of more rapidly time-varying and geographically localizable covariates may improve the performance of conflict forecasting models with respect to escalation. One promising source of such predictors is political event data, now widely available from a variety of sources including ICEWS and the Phoenix project; though, Chiba and Gleditsch (2017) report only modest gains from incorporating event data into predictive models of conflict.

While the convolutional component of this model parameterizes spatial weights allowing for flexibility with respect to spatial dependencies, it also assumes an equal-area grid. However, the PRIO-GRID data are degrees-based: one-half degrees latitude and longitude per cell. Therefore, cell areas vary from 3098.0 km$^2$ to 4521.5 km$^2$. Because of this, the learned convolutions result in differing spatial weights for different regions of the map. More specifically: the spatial convolutions scale with the centroid distances between adjacent cells and therefore represent differing spatial areas for different parts of the globe. The easiest solution to this problem would involve interpolating values to produce an equal-area grid and then transforming predictions back into the original grid. Alternatively, it may be possible to correct for unequal cell sizes by reformulating the problem as prediction on a graph where edge weights are inversely proportional to the distances between centroids of adjacent cells. This may result in better predictive performance and would be especially important were future work to focus on interrogating the learned convolutions. Simply controlling for cell land area by including it as an additional feature will likely not allow the model to correct for the fact that the convolutional filters represent differing spatial areas in different regions.[15]

Indeed, future plans for this research project involve interrogation of the learned spatial convolutions. These convolutions represent spatial weights associated with predictive features of conflict and may provide insight about the appropriate methods for evaluating the decay of an effect with distance. For example, we could determine whether the predictive power of various features decays linearly or geometrically with distance. This could have implications for how spatially-lagged variables are computed for spatial regression models.

Similarly, an attention layer on the LSTM portion of this model would reveal learned temporal dynamics of conflict. These attention activation values are conditional on each sample's feature values and, therefore, this modified ConvLSTM model may help researchers to identify distinguishing characteristics of recurring political violence and one-time violent events.

[15] Given the noisiness of social processes and data collection, as well as the natural geographic and social heterogeneity of PRIO-GRID cells, I am not convinced that correcting for land area differences of up to 50% will result in substantially improved predictive performance.

# References

Althaus, Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. 2020. "Cline Center Historical Event Data." *Cline Center for Advanced Social Research. v.1.3.0. University of Illinois Urbana-Champaign.*

Bahdanau, Dzmitry, KyungHyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate." Paper presented at the *Proceedings of ICLR 2015*,.

Blair, Robert A., and Nicholas Sambanis. 2020. "Forecasting Civil Wars: Theory and Structure in an Age of "Big Data" and Machine Learning." *Journal of Conflict Resolution* 64(10): 1885–1915.

Chiba, Daina, and Kristian Skrede Gleditsch. 2017. "The shape of things to come? Expanding the inequality and grievance model for civil war forecasts with event data." *Journal of Peace Research* 54(2): 275–297.

D'Orazio, Vito, and James E. Yonamine. 2015. "Kickoff to Conflict: A Sequence Analysis of Intra-State Conflict-Preceding Event Structures." *PloS one* 10.

D'Orazio, Vito, and Yu Lin. 2022. "Forecasting Political Instability in Africa with Automated Machine Learning Systems." *International Interactions* 48.

Hegre, Håvard, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. 2019. "ViEWS: A political violence early-warning system." *Journal of Peace Research* 56(2): 155–174.

Hegre, Håvard, Nils W. Metternich, Håvard Mokleiv Nygård, and Julian Wucherpfennig. 2017. “Introduction: Forecasting in peace research.” *Journal of Peace Research* 54: 113–124.

Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. 2012. “Neural Networks for Machine Learning. (Presentation).” .

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Comput.* 9(8): 1735–1780.

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” Paper presented at the *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*,. ICML’15 , 448–456.

Jansen, Remco, Håvard Hegre, Michael Colaresi, and Frederick Hoyles. 2020. “Benchmark models for the ViEWS prediction competition.” *Online.*

Lotter, William, Gabriel Kreiman, and David Cox. 2017. “Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning.” *Proceedings of ICLR 2017.*

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *Presented at Advances in Neural Information Processing Systems* 30: 4765–4774.

Metternich, Nils W., Shahryar Minhas, and Michael D. Ward. 2017. “Firewall? or Wall on Fire? A Unified Framework of Conflict Contagion and the Role of Ethnic Exclusion.” *Journal of Conflict Resolution* 61(6): 1151–1173.

Montgomery, Jacob, Michael Ward, and Florian Hollenbach. 2011. “Dynamic Conflict Forecasts: Improving Conflict Predictions Using Ensemble Bayesian Model Averaging.” Paper presented at the *Annual meeting of the International Studies Association Annual Conference*,.

Mueller, Hannes, and Christopher Rauh. 2020. "The Hard Problem of Prediction for Conflict Prevention." *Working Paper.*

O'Brien, Sean P. 2010. "Crisis early warning and decision support: Contemporary approaches and thoughts on future research." *International Studies Review* 12: 87–104.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED-Armed Conflict Location and Event Data." *Journal of Peace Research* 47: 651–660.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." Paper presented at the *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, , 1135–1144.

Schuster, M., and K.K. Paliwal. 1997. "Bidirectional Recurrent Neural Networks." *Trans. Sig. Proc.* 45(11): 2673–2681.

Shi, Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting." Paper presented at the *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA. Cambridge, MA, USA NIPS'15 , 802–810.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. "A survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15: 1929–1958.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49: 363–374.

Vesco, Paola, Michael Colaresi, Håvard Hegre, Remco Jansen, Adeline Lo, Gregor Reisch, and Nils Weidmann. 2022*a*. "United They Stand: Findings from an Escalation Prediction Competition." *International Interactions* 48.

Vesco, Paola, Michael Colaresi, Håvard Hegre, Remco Jansen, Adeline Lo, Gregor Reisch, and Nils Weidmann. 2022*b*. "United They Stand: Findings from an Escalation Prediction Competition Online Appendix." *International Interactions* 48.

Wager, Stefan, Sida Wang, and Percy Liang. 2013. "Dropout Training as Adaptive Regularization." Paper presented at the *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, Red Hook, NY, USA. Red Hook, NY, USA NIPS'13 , 351–359.

Weidmann, Nils B., and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54(6): 883–901.

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." Paper presented at the *Computer Vision – ECCV 2014*, Cham. Cham , 818–833.

Zhao, Dazhi, Guozhu Yu, Peng Xu, and Maokang Luo. 2019. "Equivalence between dropout and data augmentation: A mathematical check." *Neural Networks* 115: 82–89.