Taking time seriously: Predicting conflict fatalities using temporal fusion transformers*

Julian Walterskirchen†, Sonja Häffner‡, Christian Oswald§, Marco Binetti¶

July 11, 2024

Abstract

Previous conflict forecasting efforts identified three areas for improvement: the importance of spatiotemporal dependencies and nonlinearities and the further exploitation of latent information in conflict variables, a lack of interpretability in return for high accuracy of complex algorithms, and the need to quantify prediction uncertainty. We predict conflict fatalities with temporal fusion transformers which have several desirable features for forecasting, addressing all these points. First, they can produce multi-horizon forecasts and probabilistic predictions, offering a flexible and non-parametric approach. Second, they can incorporate time-invariant covariates, known future inputs, and other exogenous time series which allows to identify globally important variables, persistent temporal patterns, and significant events. Third, this approach puts a strong focus on interpretability such that we can investigate temporal dynamics more thoroughly. Our approach outperforms benchmarks from an award-winning early warning system over several metrics and test windows and is thus a valuable addition to the forecaster's toolkit.

Word Count: ≈ 5,000

*This paper documents a contribution to the VIEWS Prediction Challenge 2023/2024. Financial support for the Prediction Challenge was provided by the German Ministry for Foreign Affairs. For more information on the Prediction Challenge please see Hegre et al. (Forthcoming) and `https://viewsforecasting.org/research/prediction-challenge-2023`. We thank the organizers and participants of the 2023 VIEWS Prediction Challenge workshop for their helpful comments. We gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the University of the Bundeswehr Munich to train the TFT models. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government.

†Center for Crisis Early Warning, University of the Bundeswehr Munich, `julian.walterskirchen@unibw.de`

‡Center for Crisis Early Warning, University of the Bundeswehr Munich, `sonja.haeffner@unibw.de`

§Center for Crisis Early Warning, University of the Bundeswehr Munich, `christian.oswald@unibw.de`

¶Institute of Intercultural and International Studies, University of Bremen, `mbinetti@uni-bremen.de`

1 Introduction

Conflict forecasting has attracted considerable attention from scholars and increasingly policymakers alike. Efforts such as a recent prediction competition sought to gather new ideas and approaches to advance the discipline, identify lessons learned, and outline avenues to improve forecasting best practices. This paper addresses three of these avenues: 1) the need to find useful ways to estimate and present uncertainty surrounding point estimate forecasts, 2) the further exploitation of latent information in conflict variables by modeling spatio-temporal dependencies and nonlinearities more directly, and 3) increased interpretability despite using complex algorithms which achieve high accuracy (Hegre, Vesco & Colaresi 2022). This paper is a contribution to the second such competition and the wider conflict forecasting literature which introduces temporal fusion transformer models (TFTs) to predict conflict fatalities.[1] TFTs have several desirable properties: they can 1) extract more temporal and nonlinear information from the data, 2) incorporate time-invariant covariates, known past and future inputs such as elections and holidays, and other exogenous time series, 3) produce sequential forecasts over the entire prediction horizon simultaneously rather than iteratively, and 4) model prediction distributions to fully present the uncertainty of forecasts (Lim, Arık, Loeff & Pfister 2021). As a result, TFTs are a valuable addition to a forecaster's toolkit well beyond conflict forecasting.

Previous conflict prediction efforts have mostly supplied point estimate forecasts, although there are examples going beyond that (Brandt, Freeman & Schrodt 2014, Brandt, D'Orazio, Khan, Li, Osorio & Sianan 2022). However, different audiences of conflict forecasts are interested in different aspects. Some may be interested in the most likely outcome, which corresponds to a point estimate or mean of a predictive distribution, whereas others might be interested in the range of possible outcomes and the probabilities of the most extreme outcomes, i.e. at the tails of the predictive distribution. The former can help answer questions such as how many fatalities are likely in country x at time y while the latter helps answer questions such as how likely is a drastic conflict escalation in country x at time y. Combining these approaches results in more nuanced and balanced forecasts with uncertainty such that users know both what the most likely outcome is, how likely it actually is, and what the range of plausible outcomes is (Gleditsch 2022). This is important information for researchers, decision-makers, or any consumer of forecasts beyond conflict.

[1] See Hegre, Vesco, Colaresi, Vestby, Timlick, Kazmi, Becker, Binetti, Bodentien, Bohne, Brandt, Chadefaux, Drauz, Dworschak, D'Orazio, Fritz, Frank, Gleditsch, Häffner, Hofer, Klebe, Macis, Malaga, Mehrl, Metternich, Mittermaier, Muchlinski, Mueller, Oswald, Pisano, Randahl, Rauh, Rüter, Schincariol, Seimon, Siletti, Tagliapietra, Thornhill, Vegelius & Walterskirchen (Forthcoming) for the introductory article to the prediction challenge.

Novel and sophisticated algorithms have been used recently to improve our collective ability to predict conflict intensity (Hegre, Vesco & Colaresi 2022). However, these oftentimes complex models tend to perform quite well individually at the expense of interpretability (Vesco, Hegre, Colaresi, Jansen, Lo, Reisch & Weidmann 2022). Some users or consumers of conflict forecasts might be more interested in knowing which factors influence forecasts most rather than having highly accurate forecasts without interpretability (Gleditsch 2022). After all, early warning can only work well if we know what change or action is required to prevent or mitigate escalation. We provide evidence for several forecasting windows to evaluate the performance of the TFT model and demonstrate that TFTs predict conflict intensity with high accuracy while enabling interpretability.[2]

The paper is structured as follows. First, we briefly review recent efforts to forecast conflict intensity and the particular challenges involved. We subsequently describe transformer models in general and the temporal fusion transformer model in particular. After presenting the data and providing more technical details about the modeling approach, we discuss the results on the country- and grid cell-month level as units of analysis. We conclude by outlining limitations and avenues for future research.

2 Conflict forecasting (and its limits)

Recent efforts to forecast conflict intensity have introduced both novel data sources and algorithms. Information from newspaper articles (Mueller & Rauh 2022) or internet searches (Oswald & Ohrenhofer 2022) for example have the potential to pick up early signs of tensions before escalation due to their fast-changing nature (Hegre, Vesco & Colaresi 2022). Furthermore, algorithms such as recurrent neural networks (Radford 2022, Malone 2022) or dynamic time warping (Chadefaux 2022) have shown promising predictive performance. However, the overarching challenge is that complex algorithms tend to perform well and quite accurately but oftentimes provide little interpretability (Vesco et al. 2022). Meanwhile, consumers of conflict forecasts are at least as interested in knowing which factors contribute to changes in conflict intensity as they are in having accurate forecasts (Gleditsch 2022). It may be of higher value for decision-makers to know which factors contribute to an increased conflict intensity risk since it provides them with actionable information on which prevention and mitigation strategies might be more or less effective.

Part of the modeling approach is a need to extract more latent information from conflict variables such as spatial and temporal dependencies and nonlinear relationships (Hegre,

[2]There is also a true forecast for July 2024 through June 2025 as part of the challenge. This and the test set forecasts for all models in this paper can be found at `https://tft-prediction-explorer.streamlit.app`.

Vesco & Colaresi 2022). Conflict variables are by and large the best and most reliable predictors for future intensity (Hegre, Bell, Colaresi, Croicu, Hoyles, Jansen, Leis, Lindqvist-McGowan, Randahl, Rød & Vesco 2021, Bazzi, Blair, Blattman, Dube, Gudgeon & Peck 2022). Finding algorithms which are able to exploit nonlinearities and dependencies in these conflict variables are, in the absence of new data sources and improved quality of existing data sources, the forecaster's best bet to improve predictive performance. However, this added layer of complexity should not come at the expense of interpretability.

Lastly, there are only few previous studies which supplied conflict forecasts with measures of uncertainty (Brandt, Freeman & Schrodt 2014, Brandt et al. 2022). Providing uncertainty around point estimates is useful for researchers and consumers of forecasts alike and has to be reported in clear ways (Hegre, Vesco & Colaresi 2022). Practicioners and decision-makers would not only like to know the most likely outcome, i.e. the point estimate of predicted fatalities, but also how likely this most likely outcome is, how likely more extreme outcomes are, and what the range of plausible outcomes is (Gleditsch 2022). It is also of interest to producers and consumers of forecasts to evaluate with how much confidence the models were on or off target. In sum, conflict forecasts will be more valuable for consumers if they can exploit spatial or temporal dependencies and nonlinearities from conflict variables, and provide measures of uncertainty and some level of interpretability while upholding accuracy levels. The temporal fusion transformer can provide all of these at no expense to accuracy.

3 The temporal fusion transformer model

A transformer model is a neural network, a deep learning approach, which can learn context and meaning by observing connections in sequential data such as time series. The attention mechanism characterizing transformer models is capable of discovering influences and dependencies between variables and data points. A transformer model consists of encoder/decoder blocks which process data. These positional encoders mark data points going through the network and attention units follow these markers to create a map of relationships between elements (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin 2017).

Temporal fusion transformers (TFTs) are attention-based deep neural networks designed to provide both good performance and interpretability. The main characteristic differentiating them from standard transformer models is that they specifically utilize the self-attention mechanism to identify complex temporal patterns in multiple time series. TFTs have additional desirable features for conflict forecasting and beyond. First, a model can be trained on multiple multivariate time series. Second, TFTs provide multi-horizon forecasts with prediction intervals. Third, they support different types of features such as

time-variant and -invariant exogenous variables. Fourth, they provide interpretability via variable importance, seasonality, and extreme event detection (Lim et al. 2021). Lastly, TFTs, and transformer models more generally, have been shown to outperform statistical, machine learning and other deep learning models, including, for example, gradient-boosted trees and Deep Space-State Models (Elsayed, Thyssens, Rashed, Jomaa & Schmidt-Thieme 2021, Lim et al. 2021, Makridakis, Spiliotis, Assimakopoulos, Semenoglou, Mulder & Nikolopoulos 2023).

Incorporating different types of features distinguishes TFTs from other well-known deep learning time-series models such as Deep AR (Salinas, Flunkert, Gasthaus & Januschowski 2020). More specifically, it can include known and unknown time-varying and real and categorical time-invariant features. Known features refer to for example holidays or election dates which are known for both the past and the future. Unknown features are e.g. the number of violent events or deaths in the past which we do not know for the prediction period. Examples of time-invariant real, meaning numerical, features can be country or grid-cell identifiers. Time-invariant categorical features may include variables such as geographic region or continent (Lim et al. 2021).

Models producing multi-horizon forecasts can generally be grouped into two categories: iterated and direct approaches. Iterated approaches use autoregressive models and forecast one step into the future to subsequently feed this prediction into the model to generate the forecast two steps into the future and so forth. Deep AR as a Long Short-term Memory (LSTM) network and Deep State-Space Models fall into this category. In contrast, direct approaches are rooted in sequence-to-sequence models and produce all forecasts for the defined prediction horizon simultaneously. TFTs fall in this category of direct approaches. The main advantage of direct approaches, and the TFT in particular, is that they can incorporate past and future time-varying inputs easily whereas iterated approaches rest on the assumption that all feature values are known for the future such that only the outcome needs to be fed into the model again repeatedly (Lim et al. 2021). The ability to incorporate different types of features and to produce direct forecasts go hand in hand for TFTs.

A common criticism of so-called black-box models is that they are overly complex and contain nonlinear interactions between numerous parameters, which is one reason for limited interpretability (Lim et al. 2021). At its core, a TFT consists of four individual components: a Gated Residual Network (GRN), a Variable Selection Network (VSN), a LSTM Encoder Decoder Layer, and an Interpretable Multi-Head Attention mechanism. In addition, a fifth component in the form of quantile regression can be added to estimate prediction intervals alongside point predictions. The GRN controls if and to what extent we allow nonlinearities to enter the modeling process. Put differently, TFTs have the ability and flexibility to

decide whether it is necessary with the data at hand to introduce nonlinear relationships or whether it is better to keep the model simple and suppress nonlinear contributions. It thus enables complexity to enter the modeling process while keeping all desirable features mentioned above, in contrast to introducing complexity by default (Lim et al. 2021).

The main task of any forecasting model is to predict a point estimate of the target variable. However, we are oftentimes as much or even more interested in the uncertainty of predictions in the form of prediction intervals. Forecasts have to convey the uncertainty surrounding point predictions if we want research to be useful for decision-makers or international organizations (Gleditsch 2022). TFTs can use quantile regression, an extension of the standard linear regression, which estimates the conditional median of the target variable and for example the 0.25 and 0.75 quantiles, or any percentiles we want, such that the model can deliver a prediction interval around the actual point prediction. We can choose whether to minimize the quantile loss function or other functions such as the mean squared error or the mean absolute percentage error (Lim et al. 2021). TFTs thus provide several options to convey uncertainty and compute prediction intervals.

Lastly, interpretability and explainability have received increased attention lately. Many complex algorithms tend to achieve high accuracy at the expense of explainability (Hegre, Vesco & Colaresi 2022). However, especially in early warning settings, it might be more useful for decision-makers to know which variables or factors are more important than others to get a better idea of what can be done to prevent or mitigate conflicts from intensifying rather than having highly accurate forecasts but not knowing which factors were most influential in producing these forecasts (Gleditsch 2022). TFTs provide interpretability for features, seasonality, and extreme events. The integrated Variable Selection Network can compute feature importance scores by analyzing weights attached to them in the test set. Seasonality is taken into consideration by identifying persistent temporal patterns via attention weight patterns to identify more important past time steps which influence the forecasts. Lastly, time series are vulnerable to sudden shocks caused by rare events, and oftentimes we do not know whether there are hidden persistent patterns in the data which a model cannot identify or whether it is simply random noise. TFTs provide the option of analyzing each feature across its entire distribution of values to check the robustness of the model and reveal whether hidden patterns might be present (Lim et al. 2021). There are thus several layers of interpretability and explainability directly tied to the model. In sum, a TFT is well positioned to contribute to our collective conflict forecasting efforts.

4 Data and methods

The outcome of interest is the count of fatalities resulting from state-based armed conflict, i.e. from violence between a government and an organized rebel group which leads to at least 25 battle-related deaths per year. Data from 1990 until April 2024 come from the Uppsala Conflict Data Program (UCDP)'s Georeferenced Event and Candidate Events Datasets (Davies, Pettersson & Öberg 2023, Hegre, Croicu, Eck & Högbladh 2020). We forecast the outcome for country- and PRIO-GRID cell-months as units of analysis. Countries are defined per the updated and revised Gleditsch-Ward list of independent states (Gleditsch & Ward 1999). PRIO-GRID is a standardized and static spatial grid structure of the world which consists of quadratic cells at a resolution of 0.5 x 0.5 decimal degrees (Tollefsen, Strand & Buhaug 2012). Analyses on the country-month level have global coverage whereas analyses on the PRIO-GRID-month level are restricted to Africa and the Middle East.

We forecast the number of fatalities twelve months into the future based on data up until and including October of the preceding year. We repeat this for six individual years from 2018 through 2023 and evaluate the TFT against benchmark models from the award-winning Violence and Impacts Early-Warning System (Hegre et al. 2021).[3] We mostly rely on data provided by VIEWS as part of the prediction challenge to train models. However, we add three custom features: 1) rolling 6-month z-Scores of the lagged target variable, 2) the rolling 6-month median of the lagged target variable, and 3) the rolling 6-month mean of the lagged target variable. We furthermore include information about past and future elections from the National Democratic Institute (NDI) Global Election Calendar in our country-month models.[4] The variable produced is a binary variable, indicating whether there is an election in a given country-month (CM). We also experimented with including this feature on the PRIO-GRID-month (PGM) level but found that it does not help increase predictive performance, however. We were, unfortunately, unable to include the election data for the true forecast, as NDI had not released upcoming election data for 2025.

We rely on the *NeuralForecast* python package provided by Nixtla (Olivares, Challú, Garza, Canseco & Dubrawski 2022) to implement TFTs.[5] The package provides a number of convenience functions, including automatic hyperparameter tuning and various probabilistic loss functions. We train a TFT model for each test window (2018-2023) and level of analysis (CM and PGM). We use a Huberized Multi-Quantile loss (Huber 1964) which

[3]There is also a true forecast for July 2024 through June 2025, using data up until and including April 2024, which will be evaluated separately and in comparison to other prediction challenge contributions. Visualizations can be found at `https://predcomp.viewsforecasting.org/`.

[4]The calendar can be found at `www.ndi.org/elections-calendar`.

[5]We opted for this implementation over other alternatives for reproducibility and computational reasons.

is used frequently in regression tasks with outliers or heavy tails for training our models. We use the same loss function to conduct our Optuna-based hyperparameter tuning (Akiba, Sano, Yanase, Ohta & Koyama 2019). We optimize ten model parameters: the hidden layer size, number of attention heads, learning rate, scaler type, maximum number of steps, batch size, windows batch size, random seed, input size, and step size. We obtain predictions for differing levels of uncertainty in the form of prediction intervals of 50, 60, 70, 80, and 90%. We use these intervals to sample 1000 draws from a normal distribution for each level and investigate which one produces the most performant forecasts. We find that wider confidence bands produce better performance scores at the CM level, the 80% CI producing the best results, and tighter confidence bands, e.g. 50% intervals, at the PGM level.

We evaluate model performance with three metrics: the Continuous Rank Probability Score (CRPS), the log or ignorance score (IGN), and Mean Interval Scores (MIS). CRPS measures accuracy and can be thought of as the mean absolute error equivalent for predictive distributions. Values get closer to 0 if the prediction distribution has low variance and is centered around the actual values. IGN is the log of the predictive density evaluated at the actual observation and complements CRPS. It is less concerned with the uncertainty around a prediction or the distance between prediction and observation but with the probability attributed to the actual event. Lastly, MIS strikes a balance between having fairly narrow prediction intervals and a good coverage rate. It focuses on the most likely values, penalizes increasing prediction interval size and rewards coverage (Gneiting & Raftery 2007).

Metric	Calibration	Sharpness	Focus	Nearness	Propriety
CRPS	X	X	-	x	X
IGN	-	X	X	-	x
MIS	X	X	-	-	X

Table 1: Performance metrics overview

These metrics capture, to varying degrees, five qualities of probabilistic forecasting systems: calibration, sharpness, focus, nearness, and propriety. Calibration captures how well the predicted frequency of y-values matches the observed frequency of y in unseen data. Sharpness refers to the concentration of the predictive distribution around the true value. Focus captures the aim of having distinctive peak(s) or high-density regions in the predictive distributions for unseen data, i.e. high probability density for events which materialize. Nearness refers to how "near", both in time and space, predictive distributions are to actual values. Lastly, propriety means that predictive distributions represent the honest beliefs of a model which is ensured by using proper scoring rules, in contrast to improper scoring rules

which might reward increased certainty. Table 1, reproduced from Hegre et al. (Forthcoming), gives an overview to what extent the three metrics capture these five qualities, where X means to a large extent and x to a small extent.[6]

5 Results

We provide results for both the country-month (CM) and PRIO-GRID cell-month (PGM) levels for six test windows from 2018 through 2023.[7] We compare the TFT to several VIEWS-provided benchmark models. These include on the country-month level 1) a model that bootstraps predictions from the last 20 years (*Bootstrap*), 2) a model that predicts from a Poisson distribution centered around the last observed values for each unit of analysis (*Poisson*), 3) a model that predicts zero fatalities (*Zero*), and 4) a model that uses the previous 12 lags of the target variable for each unit of analysis as the basis of its predictive distribution (*Conflictology*). On the PGM level, these benchmarks are complemented by another model which extends the *Conflictology* model to also include observed values from neighbouring grid cells (*Conflictology N*). The main evaluation metric is the Continuous Rank Probability Score (CRPS), while the Ignorance Score (IGN) and Mean Interval Score (MIS) are used as supplementary metrics. We present results for all three metrics.[8]

Table 2 shows that the TFT outperforms the benchmarks on average, as well as for almost all years when looking at the CRPS for the CM predictions. Only in 2019, where *Conflictology* is marginally better, and in 2022, where the *Bootstrap* and *Zero* models reach a slightly lower CRPS, does the TFT not perform best. This pattern is similar or even better when looking at the IGN and MIS evaluation. The TFT is outperforming the benchmark models across all years and on average quite considerably for the Ignorance Score and has the lowest MIS for all years except 2022. However, Table 2 also shows that the differences between the TFT and the best-performing benchmarks are at times not large. There are, furthermore, significant differences in performance across all models for individual months and between countries as further exemplified in the supplemental information. Still, there is substantial evidence that the TFT does perform well on the country-month level and helps improve our collective ability to forecast conflict fatalities with uncertainty.

Table 3 shows the annual aggregation evaluation scores for the TFT and benchmark

[6]Definitions and a more detailed discussion of these qualities are in the supplemental information and Hegre et al. (Forthcoming).

[7]Furthermore, VIEWS will continuously evaluate the true future forecast (July 2024-June 2025), which can be found at `https://predcomp.viewsforecasting.org/`.

[8]A country and grid-cell overview of Ignorance Scores for selected test windows is in the supplemental information. All our results, all metrics, and all predictions for the six test windows and the July 2024-June 2025 true forecast can be found at `https://tft-prediction-explorer.streamlit.app/`.

Metric	2018	2019	2020	2021	2022	2023	Avg.
CRPS							
TFT	**12.874**	9.186	**20.890**	**75.938**	122.191	**46.342**	**47.903**
Bootstrap	23.577	22.458	31.417	86.626	**120.249**	52.722	56.175
Poisson	20.173	9.480	23.698	85.605	131.017	678.960	158.156
Zero	24.130	23.019	32.041	87.339	120.968	53.543	56.840
Conflictology	14.483	**9.146**	21.339	76.849	123.995	50.357	49.362
IGN							
TFT	**0.657**	**0.687**	**0.746**	**0.714**	**0.740**	**0.878**	**0.737**
Bootstrap	1.123	1.111	1.115	1.152	1.155	1.154	1.135
Poisson	1.198	1.046	1.110	1.228	1.124	1.125	1.139
Zero	1.558	1.558	1.549	1.615	1.632	1.615	1.588
Conflictology	1.237	1.212	1.193	1.224	1.238	1.241	1.224
MIS							
TFT	**102.325**	**80.716**	**318.659**	**1381.931**	2296.786	**835.996**	**836.069**
Bootstrap	454.090	426.006	606.003	1708.304	2380.744	1030.987	1101.022
Poisson	380.623	172.686	455.806	1690.711	2599.278	13523.463	3137.095
Zero	482.609	460.375	640.812	1746.780	2419.363	1070.864	1136.800
Conflictology	186.554	89.058	344.964	1435.555	**2142.128**	1042.916	873.529

Table 2: Full results for TFT and VIEWS benchmark models - country-month level

models at the PGM level. It outperforms all benchmark models on average regarding the CRPS and IGN scores and gets only marginally beaten by *Conflictology N* with regards to the MIS score. The TFT also dominates all benchmark models for the individual years on the IGN score again. It is slightly different regarding the CRPS score for individual years, where the TFT performs best just once but is mostly quite close to the best-performing model to achieve the best averaged score across all test windows. Results are similar for the MIS score, where the TFT performs best in one year and is close to the best-performing models for other years.

It is important to reiterate what the different metrics capture and what it tells us about the TFT's performance. The CRPS is the mean absolute error equivalent for probabilistic predictions and tells us how accurate forecasts are. It tells us that the TFT is among the best calibrated, sharpest, nearest, and most honest models on the PGM and even more so on the CM level. This means that the frequency of predicted and observed outcome values matches well, and that the predictive distribution is both around the true value in time and space (nearness) and concentrated around the true value (sharpness).[9] The Ignorance Score likewise measures sharpness and, as the only evaluation metric, focus, which tells us whether a predictive distribution has distinct high probability densities for materialized events, or in other words whether the prediction is on target with some observable level of certainty. The

[9]The MIS is similar to CRPS but does not capture nearness. The overall interpretation is thus quite similar.

Metric	2018	2019	2020	2021	2022	2023	Avg.
CRPS							
TFT	0.145	0.117	**0.125**	0.932	1.135	0.227	**0.447**
Bootstrap	**0.144**	0.115	0.132	0.940	1.137	**0.223**	0.449
Poisson	0.386	0.144	0.165	0.970	1.457	9.750	2.145
Zero	0.144	0.115	0.132	0.940	1.137	0.224	0.449
Conflictology	0.192	0.118	0.127	0.930	1.142	0.524	0.506
Conflictology N	0.147	**0.107**	0.127	**0.928**	**1.131**	0.250	0.448
IGN							
TFT	**0.083**	**0.085**	**0.088**	**0.102**	**0.102**	**0.110**	**0.095**
Bootstrap	0.093	0.095	0.107	0.118	0.118	0.120	0.108
Poisson	0.118	0.105	0.116	0.129	0.145	0.151	0.127
Zero	0.092	0.094	0.108	0.119	0.120	0.121	0.109
Conflictology	0.859	0.856	0.860	0.865	0.867	0.869	0.863
Conflictology N	0.177	0.175	0.182	0.189	0.190	0.192	0.184
MIS							
TFT	**2.535**	1.977	2.209	18.356	22.409	4.189	8.612
Bootstrap	2.888	2.309	2.637	18.796	22.749	4.472	8.975
Poisson	7.149	2.617	2.993	19.080	28.527	193.974	42.390
Zero	2.888	2.309	2.637	18.796	22.749	4.472	8.975
Conflictology	2.834	1.889	**2.073**	**17.870**	**22.277**	13.218	10.027
Conflictology N	3.062	**1.879**	2.115	18.106	22.475	**4.033**	**8.612**

Table 3: Full results for TFT and VIEWS benchmark models - priogrid-month level

TFT was thus most on target on both the CM and PGM level across all test windows.

Overall, the TFT produces consistently good and solid results on the CM level, with a particular advantage when evaluated using the Ignorance Score. The TFT likewise almost always performs best regarding the CRPS and MIS scores and has the lowest average score across all test windows. The results are less dominant on the PGM level, although the TFT scores best again when evaluated against the Ignorance Score across all test windows. Unlike on the CM level, it only has the best average score for the CRPS and is very narrowly beaten by one benchmark model for the MIS. This also indicates that the TFT performs close to, or at times better, than the benchmark models on the PGM level as well. While we expected the TFT to benefit from the higher disaggregation of input data on the PGM level and reach higher performance levels, the possible performance gain from more granular data may have been truncated by the problem of exceedingly high zero-inflation on the grid-cell level.

6 Conclusion

Conflict forecasting has important scholarly and practical value. Initiatives like the VIEWS prediction challenge offer opportunities to pool the knowledge of experts and experiment with new models and data sources to advance our collective ability to produce good and

useful forecasts for researchers and practitioners alike (Hegre et al. Forthcoming). Building on some lessons learned from the first prediction competition (Hegre, Vesco & Colaresi 2022, Vesco et al. 2022), we contribute to these efforts by introducing the novel Temporal Fusion Transformer (TFT) model to predict armed conflict fatalities with uncertainty. The TFT exhibits several desirable characteristics suitable to the prediction task at hand. TFTs can be trained on multiple multivariate time series and provide probabilistic multi-horizon forecasts. They also support different types of features such as time-variant and -invariant exogenous variables as well as known future inputs, such as future election dates or holidays. Finally, they also provide interpretability via variable importance, seasonality, and extreme event detection, an important characteristic for prevention efforts. These characteristics, coupled with the TFTs ability to model linear and complex non-linear relationships, make it a promising tool for forecasting conflict and any other social phenomena.

Results for the six test windows underscore that the TFT is indeed a suitable algorithm for conflict prediction. The TFT outperforms all benchmark models for almost all years at the country-month level across all evaluation metrics. The TFT likewise shows a strong performance on the grid cell-month level, although it is less dominant compared to the country-month level. Still, there are only marginal differences to the best-performing benchmark models and it performs best on average for two out of three metrics, the third being a very close tie. It is worth noting again that the TFT dominates regarding the Ignorance Score across all levels of analysis and years, which means it was the most on-target model with some level of certainty. Overall, the TFT has theoretically desirable qualities which produce promising results in a true prediction task with real-world data. These results further underline the need to continuously evaluate a diverse set of models for conflict forecasting. We also hope to have demonstrated that the Temporal Fusion Transformer model can be an important tool for forecasting beyond armed conflict and that researchers will find our paper convincing to implement the TFT for their substantive interests in political or social science more generally.

References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta & Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19 New York, NY, USA: Association for Computing Machinery pp. 2623–2631.

Bazzi, Samuel, Robert A. Blair, Christopher Blattman, Oeindrila Dube, Matthew Gudgeon & Richard Peck. 2022. "The Promise and Pitfalls of Conflict Prediction: Evidence from

Colombia and Indonesia." *The Review of Economics and Statistics* 104(4):764–779.

Brandt, Patrick T., John R. Freeman & Philip A. Schrodt. 2014. "Evaluating Forecasts of Political Conflict Dynamics." *International Journal of Forecasting* 30(4):944–962.

Brandt, Patrick T., Vito D'Orazio, Latifur Khan, Yi-Fan Li, Javier Osorio & Marcus Sianan. 2022. "Conflict Forecasting with Event Data and Spatio-Temporal Graph Convolutional Networks." *International Interactions* 48(4):800–822.

Chadefaux, Thomas. 2022. "A Shape-Based Approach to Conflict Forecasting." *International Interactions* 48(4):633–648.

Davies, Shawn, Therése Pettersson & Magnus Öberg. 2023. "Organized Violence 1989–2022, and the Return of Conflict between States." *Journal of Peace Research* 60(4):691–708.

Elsayed, Shereen, Daniela Thyssens, Ahmed Rashed, Hadi Samer Jomaa & Lars Schmidt-Thieme. 2021. "Do We Really Need Deep Learning Models for Time Series Forecasting?".

Gleditsch, Kristian S. & Michael D. Ward. 1999. "A Revised List of Independent States since the Congress of Vienna." *International Interactions* 25(4):393–413.

Gleditsch, Kristian Skrede. 2022. "One without the Other? Prediction and Policy in International Studies." *International Studies Quarterly* 66(3):sqac036.

Gneiting, Tilmann & Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102(477):359–378.

Hegre, Håvard, Curtis Bell, Michael Colaresi, Mihai Croicu, Frederick Hoyles, Remco Jansen, Maxine Ria Leis, Angelica Lindqvist-McGowan, David Randahl, Espen Geelmuyden Rød & Paola Vesco. 2021. "ViEWS2020: Revising and Evaluating the ViEWS Political Violence Early-Warning System." *Journal of Peace Research* 58(3):599–611.

Hegre, Håvard, Mihai Croicu, Kristine Eck & Stina Högbladh. 2020. "Introducing the UCDP Candidate Events Dataset." *Research & Politics* 7(3):2053168020935257.

Hegre, Håvard, Paola Vesco & Michael Colaresi. 2022. "Lessons from an Escalation Prediction Competition." *International Interactions* 48(4):521–554.

Hegre, Håvard, Paola Vesco, Michael Colaresi, Jonas Vestby, Alexa Timlick, Noorain Syed Kazmi, Friederike Becker, Marco Binetti, Tobias Bodentien, Tobias Bohne, Patrick T. Brandt, Thomas Chadefaux, Simon Drauz, Christoph Dworschak, Vito D'Orazio, Cornelius Fritz, Hannah Frank, Kristian Skrede Gleditsch, Sonja Häffner, Martin Hofer, Finn L. Klebe, Lucas Macis, Alexandra Malaga, Marius Mehrl, Nils W. Metternich, Daniel Mittermaier, David Muchlinski, Hannes Mueller, Christian Oswald, Paola Pisano, David Randahl, Christopher Rauh, Lotta Rüter, Thomas Schincariol, Benjamin Seimon, Elena Siletti, Marco Tagliapietra, Chandler Thornhill, Johan Vegelius & Julian Walterskirchen. Forthcoming. "The 2023/24 VIEWS Prediction Competition." *Journal of Peace Research* XXX.

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics* 35(1):73–101.

Lim, Bryan, Sercan Ö. Arık, Nicolas Loeff & Tomas Pfister. 2021. "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting." *International Journal of Forecasting* 37(4):1748–1764.

Makridakis, Spyros, Evangelos Spiliotis, Vassilios Assimakopoulos, Artemios-Anargyros Semenoglou, Gary Mulder & Konstantinos Nikolopoulos. 2023. "Statistical, Machine Learning and Deep Learning Forecasting Methods: Comparisons and Ways Forward." *Journal of the Operational Research Society* 74(3):840–859.

Malone, Iris. 2022. "Recurrent Neural Networks for Conflict Forecasting." *International Interactions* 48(4):614–632.

Mueller, Hannes & Christopher Rauh. 2022. "Using Past Violence and Current News to Predict Changes in Violence." *International Interactions* 48(4):579–596.

Olivares, Kin G., Cristian Challú, Federico Garza, Max Mergenthaler Canseco & Artur Dubrawski. 2022. "NeuralForecast: User Friendly State-of-the-Art Neural Forecasting Models." PyCon Salt Lake City, Utah, US 2022.

Oswald, Christian & Daniel Ohrenhofer. 2022. "Click, Click Boom: Using Wikipedia Data to Predict Changes in Battle-Related Deaths." *International Interactions* 48(4):678–696.

Radford, Benjamin J. 2022. "High Resolution Conflict Forecasting with Spatial Convolutions and Long Short-Term Memory." *International Interactions* 48(4):739–758.

Salinas, David, Valentin Flunkert, Jan Gasthaus & Tim Januschowski. 2020. "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks." *International Journal of Forecasting* 36(3):1181–1191.

Tollefsen, Andreas Forø, Håvard Strand & Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49(2):363–374.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*. Vol. 30 Curran Associates, Inc.

Vesco, Paola, Håvard Hegre, Michael Colaresi, Remco Bastiaan Jansen, Adeline Lo, Gregor Reisch & Nils B. Weidmann. 2022. "United They Stand: Findings from an Escalation Prediction Competition." *International Interactions* 48(4):860–896.