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

Julian Walterskirchen, Sonja Häffner, Christian Oswald, Marco Binetti

Center for Crisis Early Warning, University of the Bundeswehr Munich

Preprint

Abstract

Previous conflict forecasting efforts identified two areas for improvement: the importance of spatiotemporal dependencies and nonlinearities and the need to exploit the latent information contained in conflict variables further, and that complex algorithms achieve high accuracy at the expense of interpretability whereas we should aim for more interpretability. Our approach predicts future fatalities with a novel transformer-based deep learning approach which tackles both the above points. Temporal fusion transformer models have several desirable features for conflict forecasting. First, they can produce multi-horizon forecasts and probabilistic predictions through quantile regression. This offers a flexible and non-parametric approach to estimate prediction uncertainty. 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 for our prediction problem. This mechanism makes them suitable to model both long-term and short-term dependencies. Third, this approach puts a strong focus on interpretability such that we can investigate temporal dynamics more thoroughly via temporal self-attention decoders.

*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 & others (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 and feedback. 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.

1 Introduction

Conflict forecasting has attracted considerable scholarly and policy attention in recent years. Efforts such as a recent prediction competition sought to gather new ideas and approaches and identified several lessons learned and best practices to improve forecasting. We address three of these lessons: 1) a need to find useful ways to estimate and present uncertainty surrounding point estimate forecasts, 2) exploit latent information in conflict variables more by including spatio-temporal dependencies and nonlinearities more directly in the modeling process, and 3) overcome that complex algorithms achieve high accuracy at the expense of interpretability (Hegre, Vesco & Colaresi 2022). This background paper presents our contribution to the second VIEWS prediction competition (Hegre & others Forthcoming) which introduces and evaluates temporal fusion transformer models (TFTs) to predict conflict intensity. TFTs have several desirable properties for conflict forecasting such as the ability to extract more temporal and nonlinear information from the data and the possibility to model prediction distributions to fully present the uncertainty of forecasts. A major advantage of this approach is that it is possible to include known future events in the training process, in our case upcoming elections. However, all kinds of known or expected future data such as predictions of expected climate or economic changes could be included. This allows to combine available information from the past, present, and actual and expected future.

2 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 or self-attention mechanism characteristic of 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 the purpose of conflict forecasting. 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, Arık, Loeff & Pfister 2021). Lastly, TFTs, and deep learning 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 refers 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 the number of ethnic or excluded groups in a country. Time-invariant categorical features may include variables such as regime type (Lim et al. 2021).[1]

Models producing multi-horizon forecasts can generally be grouped in 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 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.

The main task of any forecasting model is to predict 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 outside academia such as for decisionmakers or international organizations (Gleditsch 2022). TFTs can use quantile regression, an extension of the stan-

[1]For all time-invariant examples mentioned the assumption would be that they do not change over time. They can change, of course, but if they do not they would fall into these categories.

dard linear regression, which estimates the conditional median of the target variable and for example the 0.25 and 0.75 quantiles, or any percentile 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. 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 decisionmakers 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.

3 Model Specification

For our implementation, we rely on the *NeuralForecast* python package provided by Nixtla (Olivares, Challú, Garza, Canseco & Dubrawski 2022). The package provides a number of convenience functions, including automatic hyperparameter tuning and various probabilistic loss functions. We mostly rely on the data provided by the VIEWS Team as part of the prediction challenge. However, we add three custom features: (1) rolling 6-month z-Scores of the lagged target variable, (2) rolling 6-month median of the lagged target variable, and (3) 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.[2] The variable produced is a binary variable

[2]The calendar can be found here: `www.ndi.org/elections-calendar`.

whether there is an election in a given country-month. We also experimented with including this feature on the priogrid-month level but found that it does not help increase predictive performance, however. Unfortunately, we are currently unable to include the election data for the final true forecast, as NDI has not released upcoming election data for 2025.

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 (HuberMQL) (Huber 1964) that 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 (Akiba, Sano, Yanase, Ohta & Koyama 2019) hyperparameter tuning. Specifically, 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 (prediction interval of 50, 60, 70, 80, and 90 percent). We use these intervals to sample 1000 draws from a normal distribution for each of the levels and investigate which level of uncertainty produces the most performant forecasts. We find that wider confidence bands (e.g., 90% intervals) produce better performance scores at the country-month level (with 80% CI producing the best results) and tighter confidence bands (e.g. 50% intervals) for the priogrid-month level.

4 Preliminary results

We provide results on both the country-month (CM) and PRIO-GRID cell-month (PGM) levels for all six test windows from 2018 to 2023 and submitted predictions for the true future forecast for July 2024 to June 2025. We use the performance metrics suggested by the VIEWS prediction challenge and compare our results to the VIEWS benchmark models for the test windows. The main evaluation metric is the Continuous Rank Probability Score (CRPS), a generalization of the mean squared error for probabilistic predictions where lower values indicate better performance. We focus on the CRPS below, additional metrics can be found in the Appendix.[3]

Figure 1 shows the CRPS for each test window per month. The TFT model outperforms the benchmark models on average across almost all years. However, the differences between the TFT and the best-performing benchmarks are not large and for individual months some benchmarks achieve lower CRPS. Furthermore, there are significant performance drops across all models in December 2020, December 2021, and October 2022.

Figure 2 shows CRPS scores on the priogrid-month level. The performance of the TFT

[3]Interested readers can further explore our results and the predictions for July 2024-June 2025 at `https://tft-prediction-explorer.streamlit.app/`.

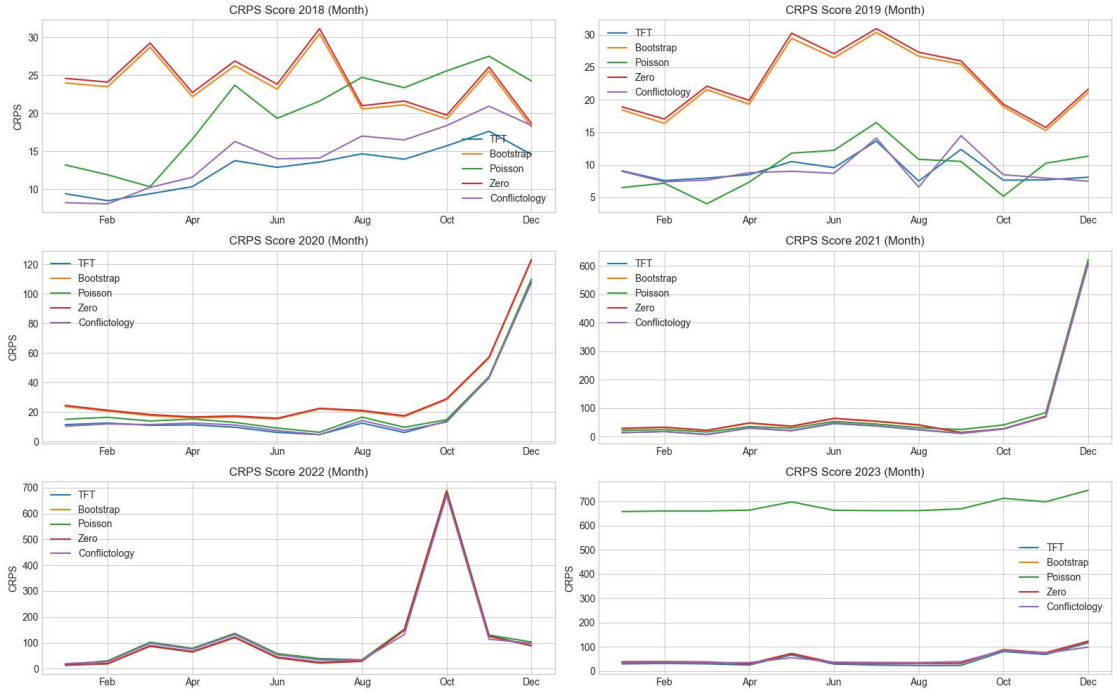

Figure 1: Continuous Rank Probability Score results 2018-2023 (CM)

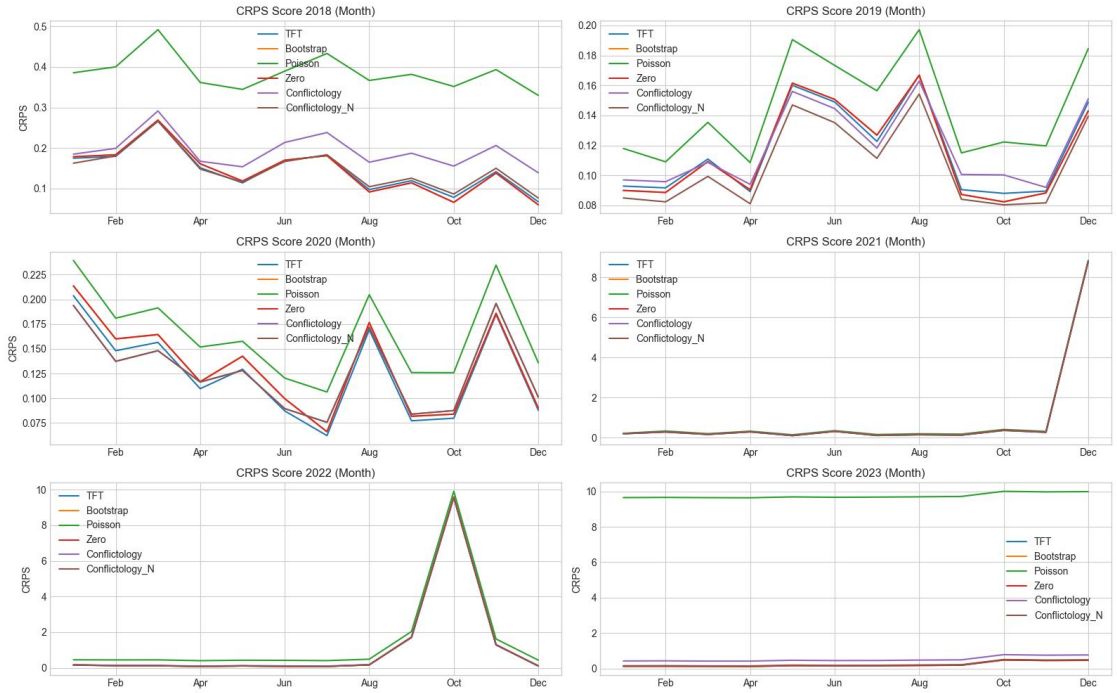

Figure 2: Continuous Rank Probability Score results 2018-2021 (PGM)

model is very similar to the best-performing benchmark models (Conflictology N and Bootstrap). The drops in performance in individual months are similar to the country-month

level, with peaks in December 2021 and October 2022. More detailed results, including additional evaluation metrics, can be found in the Appendix.

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.

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

Hegre, Håvard & others. Forthcoming. "The 2023/24 VIEWS Prediction Competition." *Journal of Peace Research* XXX.

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

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.

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.

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

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.

Appendix

Below are the full results for our TFT and the benchmark models provided by VIEWS. In addition to the CRPS, we also report Ignorance scores (IGN) and Mean Interval Scores (MIS) for each year and averaged over all six test window years.

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

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