

Summary for the VIEWS Prediction Competition

Simon Drauz¹ and Friederike Becker²

¹*Institute of Statistics (STAT), Karlsruhe Institute of Technology (KIT), simon.drauz@student.kit.edu*

²*Institute of Statistics (STAT), Karlsruhe Institute of Technology (KIT), friederike.becker@kit.edu*

June 27, 2024

1 Introduction - Process and Data

In the following, we present the general model selection and tuning process for our participation in the 2023/24 VIEWS prediction competition (Hegre et al., XXX), where we submit probabilistic predictions for conflict fatalities on the country-month (*cm*) level.¹ As a first step, we narrow the scope of potential modeling approaches by considering two competing model classes, which differ considerably in their general level of complexity. The first model (class) constitutes a rather simple and transparent baseline, which we expect to produce solidly performing and stable predictions. The second model class is based on Bayesian structured additive regression models, which theoretically have the capability to outperform the baseline due to the higher available degree of flexibility to model conflict fatalities.

After training and tuning both model classes via several specifications on the training set, we then choose the singular model which attains the lowest CRPS score on the years 2018-2021. Concretely, we therefore give preference to the model with superior performance in terms of the main metric by which submissions will be scored. This also means opting for a more data-driven model selection process and, in particular, giving preference to the baseline model, should the more sophisticated Bayesian model not demonstrate superior performance.

Throughout the following, we denote the outcome variable by $y_{i,t}$, where i refers to the country and t to the month. In a concrete prediction context, we additionally denote the forecast horizon by s , where $y_{i,t+s}$ is the s -step-ahead quantity to be predicted. We rely exclusively on VIEWS-provided data to train models and produce predictions. Additionally to the centrally provided data, some specifications of our Bayesian model utilize neighborhood data, which was kindly also provided by the VIEWS team. Sections 2 and 3, respectively, introduce the baseline and Bayesian model in more detail and Section 4 summarizes our results. Section 5 concludes.

¹This 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. (XXX) and <https://viewsforecasting.org/research/prediction-challenge-2023>

2 Baseline Model

The baseline model employed in this study extends beyond a standard “naive forecast” approach, which typically carries only the last observed value forward as its prediction. Instead, our model generates forecast samples by extracting empirical quantiles from multiple recent historical observations.

The concrete number of past observations, or window size, is selected individually for each combination of country and forecast horizon and is denoted by $w_{i,s}$. Given a window size, from the respective set of recent observations

$$\mathcal{T}_{i,t,s} = \{y_{i,t}, y_{i,t-1}, \dots, y_{i,t-w_{i,s}+1}\},$$

we compute empirical quantiles that directly serve as the forecasting samples for the month $t + s$. Choosing quantiles rather than random samples is due to the fact that they minimize the expected CRPS (Scheffzik et al., 2013). The quantiles are calculated as

$$Q_{i,t+s}^p = \hat{F}^{-1}(p; \mathcal{T}_{i,t,s}),$$

where $\hat{F}^{-1}$ estimates the inverse of the empirical cumulative distribution function by linearly interpolating between observed values, and p is the given probability level of interest. We calculate equidistant quantiles, concretely $p \in \{0.01, \dots, 0.99\}$.

The individual choice of $w_{i,s}$ in theory balances the trade-off between ensuring model stability and maintaining responsiveness to recent changes in the data. We generally consider values $w \in \{1, 2, \dots, 36\}$ and determine optimal values via cross-validation. We experiment with two competing cross-validation approaches, where approach (i) stratifies training and evaluation by month, only evaluating each forecast horizon on the specific month it is associated with in the final prediction task, and approach (ii) trains and evaluates each s -step-ahead forecast on all months and averages the results. We thereby obtain two optimal sets of $w_{i,s}$ and denote the resulting models by $B_{H_{MV}}$ and B_{H_V} , respectively.

Both specifications are robust baseline models and can automatically adapt to different locations and thereby capture temporal variations, as well as some other individual country’s characteristics. In particular, given its non-parametric nature, this baseline model can naturally account for excess zeroes as well as over-dispersion, as it directly reflects these characteristics from the data it computes quantiles from. Further general advantages of this model include its computational efficiency and its transparency and explainability. On the other hand, it has the notable disadvantage of *exclusively* relying on recently observed values to predict the distribution of future fatalities. In particular, this means that it is by definition unable to correctly (or falsely, for that matter) predict conflict onset in previously peaceful regions or, in a broader sense, any escalation or deescalation of conflict. Furthermore, it cannot accurately predict extreme tail behavior as it does not extrapolate beyond observed values.

3 Bayesian Model

The Bayesian Penalized Structured Regression (BPSR) model outlined in the following is rooted in the comprehensive review of structured additive regression (STAR) models by Fahrmeir et al. (2021). For longitudinal data that incorporates geographical information, as is the case in the VIEWS prediction task, Fahrmeir et al. (2004) introduce a variant of STAR models that can be described as geoadditive models with temporal components. The linear predictor in these models is formulated as

$$\eta_{it} = f_1(x_{it1}) + \dots + f_k(x_{itk}) + f_{time}(t) + f_{spat}(s_i) + u_{it}\gamma.$$

$f_1, \dots, f_k$ are smooth functions of the respective covariates. f_{time} captures a time trend and f_{spat} represents a spatial effect of the location information s_i (Fahrmeir et al., 2004). Both may be modeled via structured or unstructured effects (Klein et al., 2015b). Finally, $u_{it}\gamma$ represents linear effects.

To account for the over-dispersion and zero-inflation in conflict fatality data, we consider the negative binomial (NB) or its additionally zero-inflated counterpart (ZINB) as response distributions (Abbas Moghimbeigi and Mcardle, 2008). Both are parameterized via mean μ_{it} and scale parameter $\delta > 0$, with an additional zero-probability parameter π_i for the zero-inflated version:

$$y_{it}|\eta_{it}, \delta \sim \text{NB}(\mu_{it}, \delta) \quad \text{or} \quad y_{it}|\eta_{it}, \delta, \pi_i \sim \text{ZINB}(\mu_{it}, \delta, \pi_i)$$

We use the standard log link function to ensure positivity of the mean parameter, that is, $\eta_{it} = \log(\mu_{it})$. We model the smooth functions f_j via a penalized cubic spline (P-Spline) basis as defined in Lang and Brezger (2004). While splines are highly flexible functions that can fit the data to an almost arbitrary degree, the penalization component prevents overfitting while still retaining an appropriate degree of model flexibility in responsiveness to the data (Eilers and Marx, 1996).

For selection of prior distributions, we opt for non-informative choices that are standard in the literature (cf. Klein et al. (2015a), Fahrmeir et al. (2021)) – that is, Gaussian, (inverse-)gamma, or beta-priors, depending on the range of the respective model parameter.

Given these models’ computational requirements, we employ a pragmatic two-tier approach to find an optimal specification: in a first step, five competing model specifications are fit. All specifications include a single smooth function on conflict history and differ in the additional level of complexity, that is, the choice of response distribution, and the inclusion and composition of spatial and temporal effects. To the best performing model, we additionally incorporate a small and a large set of linear predictors related to additional country characteristics. All seven model specifications can be found in Table S1. The models are implemented in Stan (Stan Development Team, 2024).

4 Results

Model	2018	2019	2020	2021	Mean
VIEWS Poisson	20.03	9.62	13.68	37.11	20.11
VIEWS Bootstrap	23.44	22.02	21.22	35.99	25.67
$B_{H MV}$	13.68	6.86	13.42	26.79	15.19
$B_{H V}$ (quantile.forecast)	13.16	7.31	13.10	27.16	15.18
M_7	34.10	22.68	19.99	35.39	28.04

Table 1: Comparison of CRPS values by model for years 2018-2021. Note that scores are computed on data available as of September 2023, to accurately reflect the model selection process. Scores are displayed separately for each year and as an average over all years, with minimum values indicated in bold. VIEWS Poisson and VIEWS Bootstrap refer to the benchmark models supplied by the VIEWS team. $B_{H MV}$ and $B_{H V}$ are our baseline specifications (that differ in their choice of window length) and M_7 is the best performing specification of our Bayesian model class. As our best performing model, $B_{H V}$ is entered into the competition under the label `quantile.forecast`.

Table 1 displays mean CRPS values achieved by both baseline model specifications, the best performing Bayesian specification and the VIEWS benchmarks (Hegre et al., 2023). M_7 represents the best

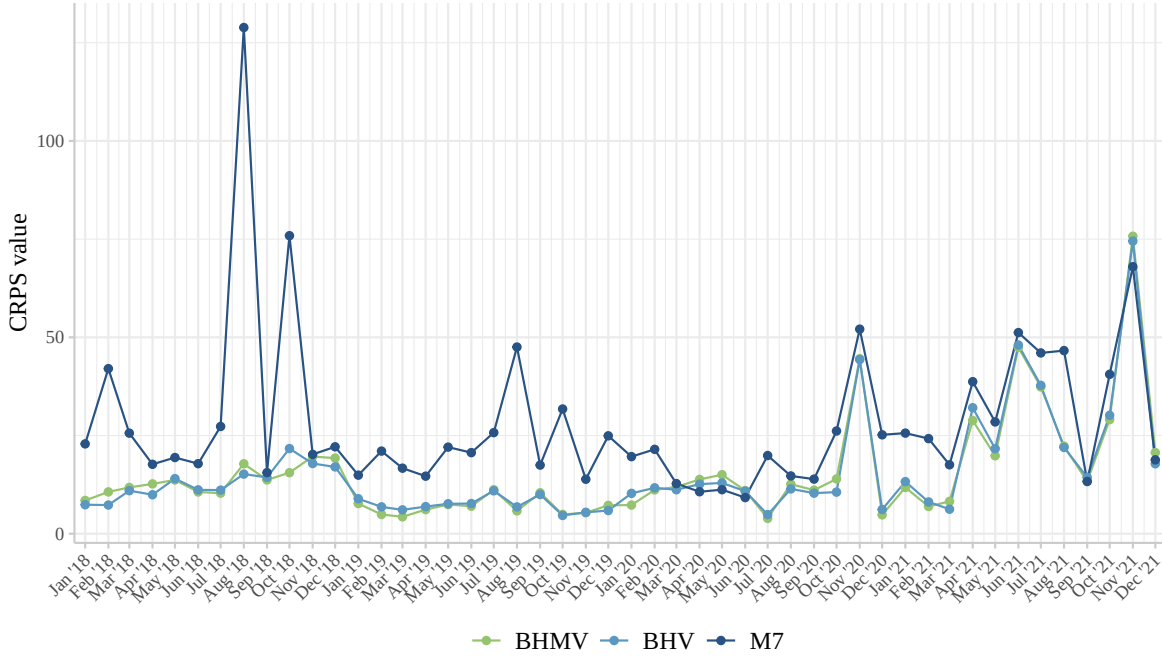


Figure 1: CRPS values by month, for the two baseline variants B_{HV} and B_{HMV} and the best performing Bayesian model M_7 .

performing Bayesian model in our study, with a zero-inflated negative binomial as response, a structured spatial component, an unstructured temporal component, and a large set of linear predictors, additional to the smooth function of conflict history. Supplemental Table S2 additionally displays IGN, MIS and CRPS values for all models, including for all other considered specifications of the Bayesian model. Note that all scores are computed on data available as of September 2023, to accurately reflect the model selection process.

Both baseline models perform substantially better than any Bayesian model specification or either of the VIEWS benchmarks, in terms of all individual years' scores as well as in the aggregate. Conversely, even the best performing Bayesian model M_7 only manages to achieve comparable scores to the VIEWS benchmarks in some years and eventually performs worse in the aggregate.

Considering these scores and our general model selection process outlined in Section 1, we choose to submit forecasts generated by the baseline model B_{HV} (denoted `quantile.forecast` in submissions) for the VIEWS competition and put a primary focus on communicating this model's performance characteristics in the following.

Figure 1 shows monthly CRPS values of the two baseline variants and the best Bayesian model, for countries with at least one observed fatality in 1990-2021, as the predictions of the baseline model by definition don't differ for zero-fatality countries. It is evident that the highly similar aggregate performance of the baseline specifications also translates to the monthly level. Additionally notable are the models' CRPS spikes in months in 2021, where the models failed to predict some larger spikes in conflict fatalities.

Considering the calibration of model B_{HV} , the PIT histogram in Figure 2 exhibits a characteristic (partial) U-shape, with PIT values at the margins substantially deviating from a uniform distribution. This tail behavior is somewhat unsurprising given the fact that the model constructs predictive samples from the range of previous observations and, in particular, does not extrapolate beyond them. Interestingly, extreme observations occur with the same relative frequency at both tail ends of the

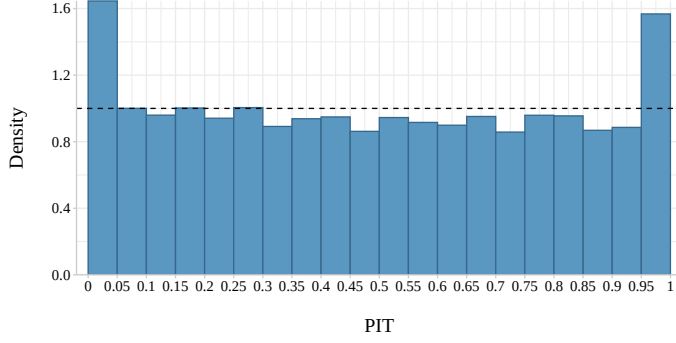


Figure 2: PIT histogram for baseline model B_{HV} . Uniformly distributed PIT values indicate a well-calibrated model.

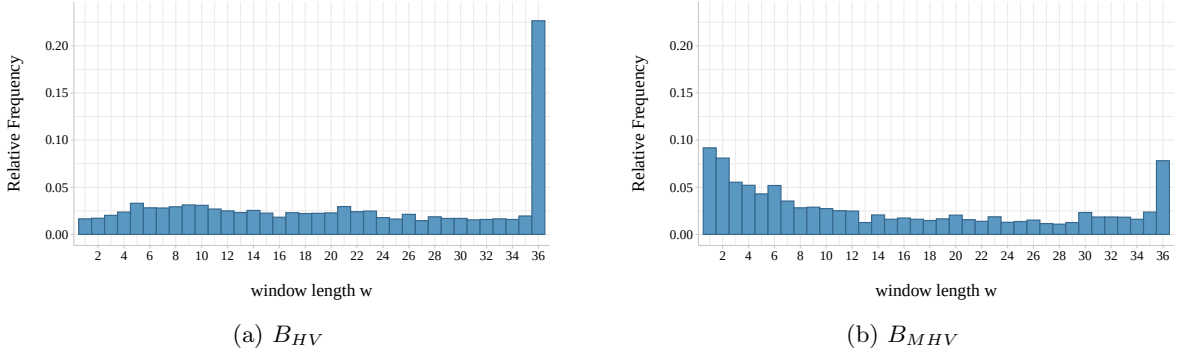


Figure 3: Optimal window lengths w for baseline models B_{HV} and B_{HMV} with respect to the CRPS.

model’s predictive distributions, indicating that the model is equally “surprised” by unusually small as well as unusually large observations. Furthermore, PIT values are mostly in line with uniformity and thus appear well calibrated in the mid-range.

Lastly, we examine the potential impact of the window length on the baseline model’s forecasting performance. The respective distributions of $w_{i,s}$ under the two baseline specifications B_{HV} and B_{HMV} differ considerably, see Figure 3. Baseline B_{HMV} generally favors shorter window lengths, indicating a preference for immediate trends in conflict dynamics. Conversely, baseline model B_{HV} , where each forecast horizon is trained on all available months, predominantly defaults to the maximum window length of 36 months (default value when performance was constant for all choices of w), with a more uniform distribution across alternative window lengths. Notwithstanding these substantial differences in optimal window lengths, the final average as well as monthly CRPS scores (see Table 1 and Figure 1) show only marginal differences, suggesting that within the scope of this study, the parameter w in fact did not have a substantial influence on the predictive performance.

5 Conclusion

To conclude, the baseline model B_{HV} (denoted `quantile_forecast` in submissions) demonstrates commendable predictive power in terms of the CRPS and outperforms the more sophisticated Bayesian model class. While the baseline model has significant advantages (cost, stability, interpretability, ease of implementation), we also acknowledge its shortcomings, such as its inability to predict conflict onset or any other change in conflict dynamics. Nevertheless, given its good performance, we argue that it might serve as an alternative benchmark to compare more sophisticated models against.

References

- ABBAS MOGHIMBEIGI, MOHAMMED REZA ESHRAGHIAN, K. M. AND B. MCARDLE (2008): “Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros,” *Journal of Applied Statistics*, 35, 1193–1202.
- EILERS, P. H. AND B. D. MARX (1996): “Flexible smoothing with B-splines and penalties,” *Statistical science*, 11, 89–121.
- FAHRMEIR, L., T. KNEIB, AND S. LANG (2004): “Penalized Structured Additive Regression for space-time data: a Bayesian perspective,” *Statistica Sinica*, 14, 731–761.
- FAHRMEIR, L., T. KNEIB, S. LANG, AND B. D. MARX (2021): *Regression Models*, Berlin, Heidelberg: Springer Berlin Heidelberg, 23–84.
- HEGRE, H., P. VESCO, M. COLARESI, AND J. VESTBY (2023): “Invitation to the 2023/24 VIEWS Prediction competition. Predicting the number of fatalities in armed conflict, with uncertainty,” .
- HEGRE, H., P. VESCO, M. COLARESI, J. VESTBY, ET AL. (XXX): “The 2023/24 VIEWS prediction competition,” *Journal of Peace Research*.
- KLEIN, N., T. KNEIB, AND S. LANG (2015a): “Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data,” *Journal of the American Statistical Association*, 110, 405–419.
- KLEIN, N., T. KNEIB, S. LANG, AND A. SOHN (2015b): “Bayesian structured additive distributional regression with an application to regional income inequality in Germany,” *The Annals of Applied Statistics*, 9, 1024 – 1052.
- LANG, S. AND A. BREZGER (2004): “Bayesian P-splines,” *Journal of computational and graphical statistics*, 13, 183–212.
- SCHEFZIK, R., T. L. THORARINSDOTTIR, AND T. GNEITING (2013): “Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling,” *Statistical Science*, 28, 616 – 640.
- STAN DEVELOPMENT TEAM (2024): “Stan Modeling Language User’s Guide and Reference Manual,” Version 2.34.0.

A Appendix

Model	Complexity	Distribution	Non-linear Feat.	Spatial C.	Temporal C.	Linear Feat.	Zero-Probability
M_1	Low	NB	Conflict History	Unstruct.	-	Intercept	-
M_2	High	ZINB	Conflict History	Struct.	Unstruct.	Intercept	Spatial Struct.
M_3	High	ZINB	Conflict History	Unstruct.	Unstruct.	Intercept	Spatial Unstruct.
M_4	Medium	ZINB	Conflict History	M_{2-3}^*	Unstruct.	Intercept	Beta Prior
M_5	Medium	ZINB	Conflict History	M_{2-4}^*	-	Intercept	M_{2-4}^*
M_6	-	M_{1-5}^*	Conflict History	M_{1-5}^*	M_{1-5}^*	Intercept, Small Linear Predictor Set	M_{1-5}^*
M_7	-	M_{1-5}^*	Conflict History	M_{1-5}^*	M_{1-5}^*	Intercept, Large Linear Predictor Set	M_{1-5}^*

Table S1: Overview of Evaluated BPSR Model Variants: This table presents seven distinct BPSR models, detailing their complexity, assumed distributions, and the types of components utilized. M_{2-3}^* indicates the specification of the best-performing model out of M2 and M3 based on the CRPS score. Similarly, M_{2-4}^* and M_{1-5}^* represent the best specifications chosen out of the ranges M2 to M4, and M1 to M5, respectively.

Model	Error Metric	2018	2019	2020	2021	Year Average
VIEWS Poisson	CRPS	20.03	9.62	13.68	37.11	20.11
	IGN	1.19	1.04	1.09	1.23	1.14
	MIS	738.02	338.35	497.83	1429.31	750.88
VIEWS Bootstrap	CRPS	23.44	22.02	21.22	35.99	25.67
	IGN	1.11	1.09	1.08	1.12	1.10
	MIS	828.71	765.92	715.05	1292.25	900.48
$B_{H MV}$	CRPS	13.679	6.859	13.424	26.792	15.188
	IGN	0.739	0.668	0.678	0.722	0.701
	MIS	167.684	72.727	192.260	447.968	220.160
$B_{H V}$	CRPS	13.159	7.311	13.102	27.156	15.182
	IGN	0.683	0.656	0.626	0.685	0.662
	MIS	158.073	72.259	182.363	438.930	212.906
M_1	CRPS	274.008	68.556	73.213	151.296	141.768
	IGN	0.855	0.848	0.867	0.916	0.871
	MIS	17152.181	2001.062	2420.564	4302.727	6469.133
M_2	CRPS	34.351	27.087	27.594	40.236	32.317
	IGN	0.831	0.862	0.869	0.923	0.871
	MIS	781.194	505.939	631.042	738.676	664.213
M_3	CRPS	36.932	27.937	27.045	41.055	33.242
	IGN	0.828	0.858	0.870	0.918	0.868
	MIS	870.064	505.762	574.276	741.032	672.784
M_4	CRPS	301.203	71.563	65.255	169.846	151.967
	IGN	0.856	0.845	0.869	0.916	0.871
	MIS	16542.350	2105.421	2369.798	4848.388	6466.489
M_5	CRPS	31.273	25.864	28.217	38.283	30.909
	IGN	0.827	0.859	0.872	0.919	0.869
	MIS	725.301	483.960	586.963	662.067	614.573
M_6	CRPS	35.739	25.989	25.760	40.113	31.900
	IGN	0.831	0.857	0.864	0.917	0.867
	MIS	845.898	486.335	480.449	743.366	639.012
M_7	CRPS	34.099	22.684	19.986	35.387	28.039
	IGN	0.799	0.828	0.840	0.888	0.839
	MIS	874.070	422.428	301.856	540.347	534.675

Table S2: Comparison of Error Metrics by model for 2018-2021: This table summarizes the performance of various models over the years 2018 to 2021 on the three scoring rules: CRPS, IGN, MIS. The baseline models, particularly $B_{H V}$, consistently exhibit superior performance across all metrics. Among the BPSR models, M_7 shows the best performance, albeit with marginal differences compared to the other variants. The table also highlights the year-to-year variability in model performance for CRPS and MIS.