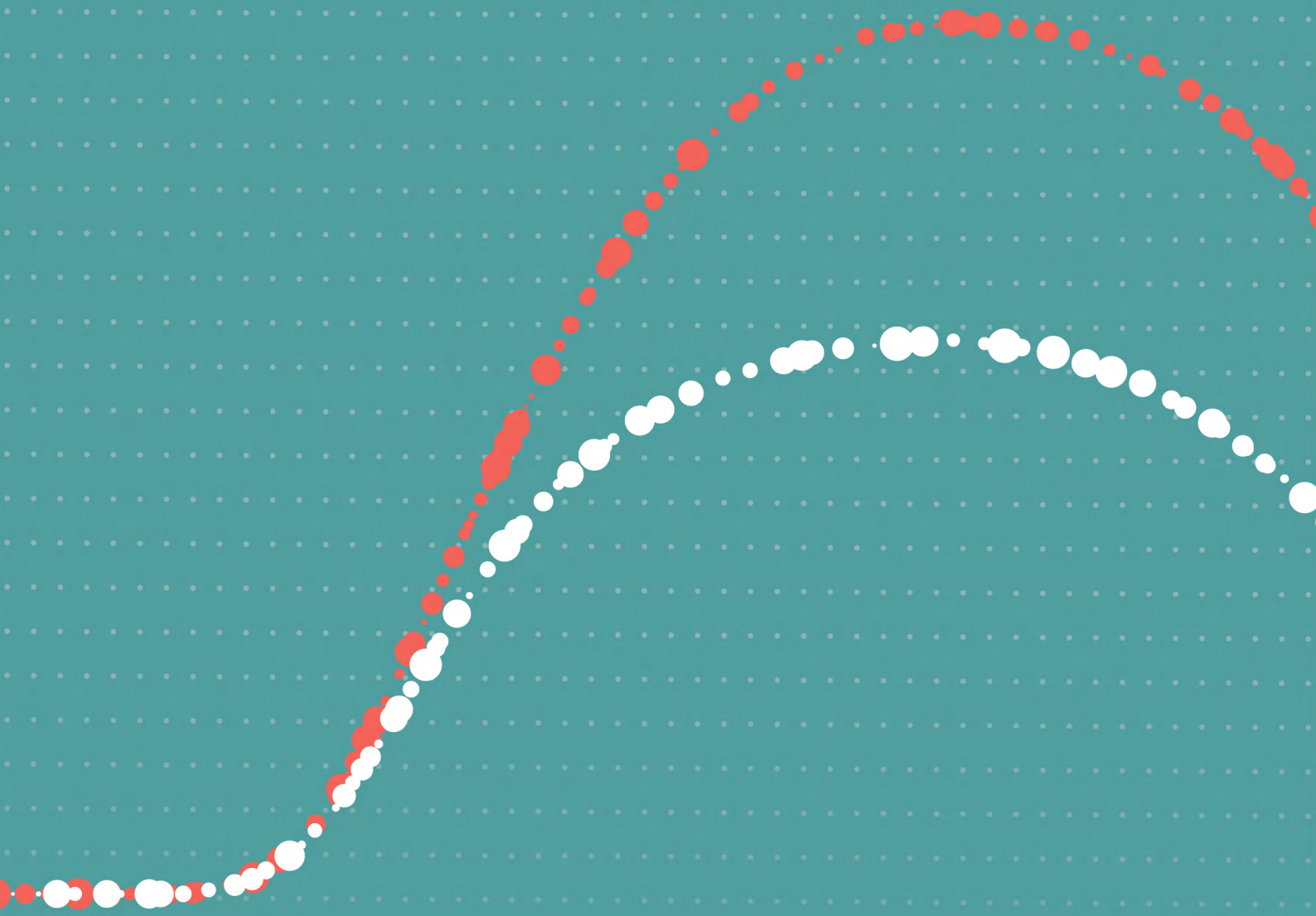
ASSESSING THE TECHNICAL FEASIBILITY OF CONFLICT PREDICTION FOR ANTICIPATORY ACTION

OCTOBER 2022



OCHA

centre for humdata

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This review was undertaken by the **United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Centre for Humanitarian Data** in The Hague. The study was written by Seth Caldwell with internal and external review by Håvard Hegre, Kim Kristensen, Erin Lentz, Leonardo Milano, Ben Parker, Josée Poirier, Manu Singh, Sarah Telford, and Marie Wagner. Graphic design is by Lena Kim. The Centre for Humanitarian Data can be contacted at **centrehumdata@un.org**.

EXECUTIVE SUMMARY

Anticipatory action enables humanitarian organizations to get ahead of a predictable shock in order to reduce its impact on vulnerable people. Since 2020, the Centre for Humanitarian Data has been supporting OCHA's anticipatory action frameworks in over a dozen countries. Our work is focused on developing the trigger mechanisms, which provide a threshold for when to release funds and take action ahead of a projected shock.

While the current anticipatory action pilots have shown that we can use data and models to predict a coming crisis, they have been limited to climate events and disease outbreaks. OCHA leadership and stakeholders have asked if it is feasible to use similar techniques to predict and act ahead of a conflict.

Predicting war, coups and riots has been a goal for a generation of social studies researchers. We reviewed a wide array of literature from several ongoing research projects across academia, the security sector and the humanitarian sector. Advances in machine learning and newly available historical datasets and predictors have given momentum to the field. Nevertheless, the problem of conflict prediction remains complex and hard to solve.

In our research, we evaluated three types of conflict prediction models – classification, risk prediction, and continuous prediction. We found insufficient justification for exclusively relying on conflict prediction models to drive anticipatory action due to several factors:

- Poor performance in predicting the onset of new conflicts.
- The lack of clear connection between predicted conflict and resulting humanitarian impact.
- The dominance of ongoing conflict as a predictor of future conflict.

To make use of conflict prediction for anticipatory action in the humanitarian sector, we recommend that future work:

- Utilize flexible models that do not pre-suppose a theoretical framework of conflict causality.
- Focus models on predicting shifts in conflicts, such as an increase in intensity or onset.
- Explore the use of human inputs through superforecasters or prediction markets and use local data to improve model performance in specific contexts.
- Improve predictions on the humanitarian impact of conflict as opposed to conflict itself.
- Ensure that model development and evaluation is done in a reproducible and transparent way that highlights the model performance in all relevant metrics.
- Learn from the state-of-the-art research underway in the academic field and ensure that applied research is relevant for humanitarian decision making.

1. INTRODUCTION

1.1 MOTIVATION

Anticipatory action enables humanitarian organizations to get ahead of a predictable shock in order to reduce its impact on vulnerable people. Since 2020, the Centre for Humanitarian Data has been supporting OCHA's anticipatory action frameworks in over a dozen countries. Our work is focused on developing the trigger mechanisms, which provide a threshold for when to release funds and take action ahead of a projected shock.

While the current anticipatory action pilots have shown that we can use data and models to predict a coming crisis, they have been limited to climate events and disease outbreaks. Yet, conflict is a key driver of food insecurity globally, with nearly 100 million people experiencing food insecurity in 23 conflict affected countries in 2020.[1] With 36 armed conflicts reported in 2021, conflict will continue to drive and exacerbate not just food insecurity but a myriad of other humanitarian needs.[2] Given this negative impact, OCHA leadership and stakeholders have asked if it is feasible to use similar techniques to predict and act ahead of a conflict.

1.2 APPROACH

The Centre's research focused on answering two questions[3] on the feasibility of forecasting conflict for anticipatory action:

- ☐ **How accurate are conflict forecasts?**
- ☐ **How well can various sources predict different types of conflict in specific situations?**

Conflict does not appear out of nowhere. Politics, the environment or competition for resources may all contribute to a flare-up of violence. Conflict prediction generally relies on an analysis of historical conflict and contributing factors to build a model of where and when it may break out in the future.

However, the range of factors and differences across contexts can make it challenging for any data-driven model to accurately predict conflict. We reviewed existing literature and models to see how well they performed in predicting conflict and the feasibility of applying these models to anticipatory action. In this paper, we define and evaluate three types of conflict prediction models – classification, risk prediction, and continuous prediction – and conclude with a set of recommendations and next steps for the humanitarian sector.

[1] UN Office for the Coordination of Humanitarian Affairs, 2022. **Global Humanitarian Overview 2022**.

[2] Escola de Cultura de Pau, 2022. **Alert 2022!**

[3] Marie Wagner and Catalina Jaime, 2020. **An Agenda for Expanding Forecast-Based Action to Situations of Conflict**, Global Public Policy Institute working paper.

2. TYPES OF PREDICTION

The literature is dominated by models that fall under three types of conflict prediction: classification, risk prediction, and continuous prediction. The three types of models produce different types of information.

- **Classification** models categorically predict whether or not a conflict will occur in a particular area and time, through either binary (e.g., yes/no) or multiclass classification (e.g., major, minor or no conflict, or a 1-5 scale).
- **Risk prediction** is designed to generate a measure of underlying risk of conflict, usually produced as a probability of conflict at a certain geographic scale within a certain timeframe.
- **Continuous prediction** models are those that directly predict a specific measure of conflict, such as the number of fatalities or conflict events (e.g., 142), without putting the results into a category or scale.

By way of example, consider the results from the three types of models if they were designed to predict in December 2021 if there would be fatalities due to state-based violence in Mali in February 2022. For the purposes of classification and risk prediction, conflict is defined as occurring if there are 25 or more conflict-related fatalities in one month in Mali, otherwise that month is classified as peaceful. Examples of potential predictions could be:

- **Classification:** There will be conflict.
- **Risk prediction:** There is a 75 percent probability of conflict.
- **Continuous prediction:** There will be 33 fatalities due to conflict.

For interested readers, further details on each of these models are available in the Technical Annex.

2.1 MODELS IN THE HUMANITARIAN SECTOR

Each of these methods are used in the humanitarian sector to assess various hazards and shocks, including conflict. OCHA's anticipatory action frameworks in the Philippines (typhoons), Bangladesh[4] (floods) and Somalia/Ethiopia (drought) rely on **classification** models. While input models of drought,[5] floods or other events often produce probabilities of the event happening, classification models set thresholds on the probabilities. For instance, in the Philippines anticipatory action framework, trigger thresholds are set based on the probability of buildings being damaged by a typhoon: if there is over 50 percent probability of 80,000 houses being damaged in a specific geographic area, the threshold is met and anticipatory action is triggered.[6] A familiar example of a classification model in the humanitarian sector is the Integrated Food Security Phase Classification (IPC) where the five classes generated are the projected phase of acute food insecurity in a given area, five being the most severe (i.e., famine).[7]

While not calculated directly as a probability, the INFORM Risk Index, meant to measure general risk of crisis for a country based on structural factors, is an example of a **risk prediction** model.[8] INFORM itself uses a conflict risk estimate, the Global Conflict Risk Index, as one of its key inputs.[9] The global displacement forecasts produced by the Danish Refugee Council are **continuous predictions** of the number of displaced persons.[10]

[4] UN Office for the Coordination of Humanitarian Affairs, 2021. Anticipatory Action Bangladesh.

[5] Rogerio Bonifacio, Gabriela Guimaraes Nobre, and Daniela Cuellar, 2021. Drought Forecasting, Thresholds and Triggers: Implementing Forecast-Based Financing in Mozambique. *EGU Gen. Assemb. Conf.* 21.

[6] UN Office for the Coordination of Humanitarian Affairs, 2022. Anticipatory Action Philippines.

[7] Timothy R Frankenberger and René Verduijn, 2011. Integrated Food Security Phase Classification: End of Project Evaluation.

[8] The Joint Research Center of European Commission, INFORM - Global, Open-Source Risk Assessment for Humanitarian Crises and Disasters.

[9] Matina Kalkia et al., 2022. The Global Conflict Risk Index: A Quantitative Tool for Policy Support on Conflict Prevention, *Prog. Disaster Sci.* 6, 100069.

[10] Danish Refugee Council, 2021. Global Displacement Forecast 2021.

MODEL ASPECTS

For these three models types, forecasts of conflict are typically defined by a few parameters:

- **Lead time:** How far in advance the model predicts conflict. Lead time can range from one month to often between one and three years, but can go out to 50 years in some of the literature.
- **Length of forecasting period:** The temporal range of the forecast period. The length of forecast is typically one month but can go all the way to a ten year period in some models. This is different from lead time, where you could predict conflict in a single month one year into the future.
- **Geographic distribution:** The geographic scope of the prediction exercise. Is the model global or applicable only to a specific country or context? Geographic distribution is almost always at the country level or more granular, such as for states or districts within a country.
- **Type:** Type of the conflict being predicted, such as state vs. non-state actors.
- **Scale:** Scale of the conflict being predicted, often defined in terms of number of deaths. Scale can be used to define whether a conflict is happening or not, but can be more complex, such as by defining no conflict, minor levels of conflict, or major levels of conflict. Scale is only used in classification and risk prediction models; continuous predictions do not use a definition of scale, which is intrinsic to the model itself (i.e., directly predicting the number of deaths).

3. OVERALL PERFORMANCE

There are common issues affecting the usability of all three models: the scale of predicted conflict; predicting conflict onset and escalation; and the lack of linkages between predicted conflict and humanitarian impact. The last issue is critically important to the feasibility of applying these models for anticipatory action. (Details on the feasibility of each model are available in the Technical Annex).

3.1 PREDICTED CONFLICT AND HUMANITARIAN IMPACT

Models are often predicting conflict defined at a relatively small **scale** in terms of casualties or fatalities and across a large timeline and a wide geographic area. Academic models can perform well, but the scale can be as small as predicting a single conflict fatality in a given month and country.[11] The humanitarian impact of one fatality in a month or 10 conflict deaths in a year in a country[12] is not readily derivable from the models.

This is also an issue for defining conflict onset, where **onset** is typically defined as the first time when the scale of conflict used for the model is observed. For instance: the month with 25 battle-related deaths after 24 preceding months without 25 battle-related deaths.[13] Even if conflict prediction improved to forecast conflict escalation, it is not clear that this would directly predict the dynamics of humanitarian needs.

Predicting the humanitarian **impact** of a conflict remains a challenge. This is not just an issue for conflict prediction but also for other applications of anticipatory action, such as for climate hazards.[14] While there has been some research on this topic,[15] the dynamics of conflict and its potential impact on humanitarian needs and response require more investigation, particularly in how foreseeable these impacts are.

[11] Hannes Mueller and Christopher Rauh, 2022b. **Using Past Violence and Current News to Predict Changes in Violence**, *Int. Interact.*

[12] Samantha Kuzma et al., 2020. **Leveraging Water Data in a Machine Learning-Based Model for Forecasting Violent Conflict**, *World Resources Institute.*

[13] Ibid.

[14] Markus Enenkel et al., 2020. **Why Predict Climate Hazards If We Need to Understand Impacts? Putting Humans Back into the Drought Equation**, *Clim. Change* 162:3, 1161–76.

[15] International Committee of the Red Cross, 2021. **When Rain Turns to Dust**.

This research could build on existing work in the food insecurity space, where FEWS NET[16] and the IPC[17] work on integrating multiple data sources to project future levels of food insecurity. New models are being explored that can more accurately predict transitions between IPC phases to improve early warning.[18] And a new six year research programme into the complex impacts of armed conflict recently began in May 2022.[19]

3.2 ANTICIPATORY ACTION WITHOUT THRESHOLDS

Anticipatory action is typically defined by models with clear thresholds set in an agreed upon, transparent framework.[20] Discussed in more detail in the **Technical Annex**, it is possible that risk prediction models (e.g., there is 50 percent probability of conflict) perform well, but classification models (e.g., setting a threshold on that probability) do not. However, it is unclear how risk models would be applied in an anticipatory action framework that typically require these explicit thresholds (i.e., the use of a classification model).

Risk models are typically utilized for disaster risk reduction, peacekeeping, or security, but may not be immediately applicable to anticipatory action. Additional work is needed to explore if and how it would be possible to use risk or continuous prediction models in anticipatory action if classification models are technically infeasible due to poor performance.

4. RECOMMENDATIONS FOR FUTURE WORK

Given our findings, we do not see immediate applications of conflict prediction for triggering anticipatory action. These recommendations do not preclude the application of anticipatory action in response to other shocks in a conflict setting.

The following three areas should be considered as potential avenues towards feasibility: 1) adopt modeling best practices; 2) explore areas that are under researched; and 3) practice transparent development and evaluation of models.

4.1 ADOPT MODELING BEST PRACTICES

The modeling and prediction of conflict should learn from the findings generated by the research community in recent years, including using many predictors of conflict in models and using ensemble methods.

Include many predictors of conflict in models

A vast array of conflict predictors have been identified in the literature but they are often not consistently identified across all models.[21] The drivers of conflict vary across the political, socioeconomic and environmental landscape: from contested elections and political shifts to crop failure or unemployment. Furthermore, a literature review on the impact of environmental changes on conflict found that there was still no consensus on the relationship between the two phenomena. Findings are often not robust to changes in the model (e.g., a claim may be based on a particular model, but adding new indicators invalidates the claim) and identified relationships between predictors and conflict may not be systematic across crises.[22] Similar issues have been identified in the literature when linking climate change to conflict, where the evidence may not be as strong as assumed.[23]

[16] Richard J. Choularton and P. Krishna Krishnamurthy, 2019. **How Accurate Is Food Security Early Warning? Evaluation of FEWS NET Accuracy in Ethiopia**, *Food Sec.* 11:2, 333–44.

[17] **Timothy R Frankenberger and René Verduijn, 2011.**

[18] Joris J. L. Westerveld et al., 2021. **Forecasting Transitions in the State of Food Security with Machine Learning Using Transferable Features**, *Sci. Total Environ.* 786, 147366.

[19] **Societies at Risk research programme**, Department of Peace and Conflict Research, Uppsala University.

[20] **OCHA Anticipatory Action.**

[21] Michael D Ward, Brian D Greenhill, and Kristin M Bakke, 2010. **The Perils of Policy by P-Value: Predicting Civil Conflicts**, *J. Peace Res.* 47:4, 363–75.

[22] Thomas Bernauer, Tobias Böhmelt, and Vally Koubi, 2012. **Environmental Changes and Violent Conflict**, *Environ. Res. Lett.* 7:1, 015601.

[23] Courtland Adams et al., 2018. **Sampling Bias in Climate–Conflict Research**, *Nat. Clim. Change* 8:3, 200–03.

To avoid these kinds of errors, the use of predictive models to validate causal frameworks has been increasingly common in the literature, such as testing how well measures of democratic governance predict a country's willingness to engage in conflict.[24, 25, 26] The purpose is to justify the importance of a causal factor, such as democratic governance, climate shifts or elections, based on how well they help the model predict future conflict rather than through statistical inference. Research has found that models focused on a particular driver or theoretical framework for conflict causality often have subpar performance.[27]

New work on conflict prediction in the humanitarian sector should recognize conflict as a multi-dimensional problem. These efforts should not spend significant time or resources on developing strict theoretical frameworks for conflict causality or focus on a single driver, such as the lack of water, that may not prove fruitful if validated based on predictive performance. These efforts can be extremely valuable in many applications, but have not proven effective for the specific task of predicting conflict.

Consider ensemble methods

In a similar vein, ensemble methods such as random forests,[28] Bayesian model averaging,[29] and gradient boosting machines[30] tend to outperform single model specifications in predictive performance. State-of-the-art academic research on conflict prediction focuses almost exclusively on ensemble models. This approach is common in modeling for other complex processes, such as climate.[31] Future exploratory work in the humanitarian sector should take this into account, particularly when single model specifications might require a lot of time spent on fine tuning and may fail to learn different features of the data in ways that ensemble models can.

4.2 EXPLORE AREAS THAT ARE UNDER RESEARCHED

Based on our review of the literature, we recommend future endeavors focus on areas that have yet to be fully explored by the research community. This includes utilizing superforecasters and prediction markets; incorporating local data and analysis; predicting shifts in conflict; predicting risks, not events; and predicting the impact of conflict. As explained above, it is unlikely that a single type of model or approach will work across all conflicts relevant to humanitarian response.

Utilize superforecasters and prediction markets

The use of human inputs for prediction, such as superforecasters, prediction tournaments, or markets have been shown to outperform purely quantitative approaches in certain predictive tasks, such as election outcome predictions across a long time horizon.[32] This approach can generate knowledge for policy decisions even when quantitative modeling might not be possible or has limited performance.[33] Humans can identify difficult to discern patterns or unmeasurable quantities that machine learning algorithms can miss. Similarly, shifting dynamics of state borders, geopolitics and other predictors are difficult to capture quantitatively. The utility of purely quantitative predictions of conflict remains an open question.[34]

[24] Michael Ward et al., 2013. **Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction**, *Int. Stud. Rev.* 15, 473–90.

[25] Michael D. Ward, Randolph M. Siverson, and Xun Cao, 2007. **Disputes, Democracies, and Dependencies: A Reexamination of the Kantian Peace**, *Am. J. Political Sci.* 51:3, 583–601.

[26] Vito D'Orazio, 2020. **Conflict Forecasting and Prediction**, *Oxford Research Encyclopedia of International Studies*.

[27] Nathaniel Beck, Gary King, and Langche Zeng, 2000. **Improving Quantitative Studies of International Conflict: A Conjecture**, *Am. Political Sci. Rev.* 94:1, 21–35; Andreas Beger, Richard K. Morgan, and Michael D. Ward, 2021. **Reassessing the Role of Theory and Machine Learning in Forecasting Civil Conflict**, *J. Conflict Resolut.* 65: 7–8, 1405–26.

[28] **Samantha Kuzma et al., 2020**.

[29] Jacob M. Montgomery, Florian M. Hollenbach, and Michael D. Ward, 2012. **Improving Predictions Using Ensemble Bayesian Model Averaging**, *Polit. Anal.* 20:3, 271–91.

[30] Jonas Vestby et al., 2022. **Predicting (de-)Escalation of Sub-National Violence Using Gradient Boosting: Does It Work?**, *Int. Interact.*

[31] Climate Information, **Why use a model ensemble?**

[32] Joyce E. Berg, Forrest D. Nelson, and Thomas A. Rietz, 2008. **Prediction Market Accuracy in the Long Run**, *Int. J. Forecast.* 24:2, 285–300

[33] Philip E. Tetlock, Barbara A. Mellers, and J. Peter Scoblic, 2017. **Bringing Probability Judgments into Policy Debates via Forecasting Tournaments**, *Science* 355:6324, 481–83.

[34] Thomas Chadefaux, 2017. **Conflict Forecasting and Its Limits**, *Data Sci.* 1:1–2, 7–17.

Extensive time and resources are required to manage a human forecasting network over a sufficient period of time for model development, testing and use.[35] These methods are not frequently found in the research community, and most examples have been generated from communities less likely to publish rigorous validations. This includes the intelligence and defense communities, community or crowd-sourced early warning systems, and conflict monitoring mechanisms. Some results have been promising,[36] although the findings have been questioned.[37] When trying this approach, models could use only human inputs without any additional quantitative data, or human inputs could be utilized as additional predictors alongside other quantitative data.

Incorporate local data and analysis

Socioeconomic, political and other input data for conflict prediction are often missing or insufficient, particularly in the most fragile and conflict-affected states. In order to predict conflict, it is required to fill in these missing inputs. This work is sometimes done directly by conflict modelers[38] or by implicitly relying on modeled data, such as World Bank poverty headcounts or the INFORM Risk Index.[39] The use of these proxies presents issues for detecting rapid shifts in conflict dynamics. In addition, the poorer quality of the data, compared with directly measured variables, likely limits the overall forecasting potential of quantitative models.[40]

Prediction of conflict at the subnational level is typically done by disaggregating data across subnational boundaries[41] or using a grid approach (e.g., predicting conflict within a 55km area).[42] While data available globally can be subnational, using locally captured data may be more sensitive to dynamic shifts that best predict conflict.

There is also a potential to develop models that predict where conflict will occur,[43] or hotspots of higher risk,[44] rather than when.[45] This may be more feasible and still useful for humanitarian operations. However, it should be recognized that even where local data is available, model performance is often still too poor for operationalization and extensive testing remains a baseline requirement.[46] While these models may improve performance to specific contexts, they will also be much more difficult to generalize to a regional or global level.

Predict conflict shifts

Recent academic research has begun to focus on predicting the probability of conflict cessation or onset.[47] This requires a more nuanced set of error metrics and validation for models that might underperform more broadly in predicting conflict but better capture conflict escalations. A recent conflict escalation prediction competition hosted by ViEWS has produced a set of metrics useful for assessing continuous predictions.[48] New developments should focus on linking this work to classification models and assessing their performance in predicting large scale conflict onset or escalation that is more directly useful in anticipatory action for humanitarian response.

This will require the identification and use of time-variant predictors that can go beyond estimating the underlying risk of conflict to better measure when risks are changing due to dynamic factors, and may require the use of more localized or human inputs to prove successful.

[35] Michael C. Horowitz, Julia Ciocca, and Lauren Kahn, 2021. **Keeping Score: A New Approach to Geopolitical Forecasting**, *Perry World House*.

[36] Bradley J. Stastny and Paul E. Lehner, 2018. **Comparative Evaluation of the Forecast Accuracy of Analysis Reports and a Prediction Market**, *Judgm. Decis. Mak.* 13:2, 202-11.

[37] David R Mandel, 2018. **Too Soon to Tell If the US Intelligence Community Prediction Market Is More Accurate than Intelligence Reports: Commentary on Stastny and Lehner**, *Judgm. Decis. Mak.* 14:3, 288.

[38] Håvard Hegre et al., 2019. **ViEWS: A Political Violence Early-Warning System**, *J. Peace Res.* 56:2, 155–74.

[39] **Samantha Kuzma et al., 2020**.

[40] **Lars-Erik Cederman and Nils B. Weidmann, 2017**.

[41] Siri Camilla Aas Rustad et al., 2011. **All Conflict Is Local: Modeling Sub-National Variation in Civil Conflict Risk**, *Confl. Manag. Peace Sci.* 28:1, 15–40.

[42] **Håvard Hegre et al., 2022**.

[43] Sebastian Schutte, 2017. **Regions at Risk: Predicting Conflict Zones in African Insurgencies**, *Political Sci. Res. Methods* 5:3, 447–65.

[44] May Lim, Richard Metzler, and Yaneer Bar-Yam, 2007. **Global Pattern Formation and Ethnic/Cultural Violence**, *Science* 317:5844, 1540–44.

[45] **Siri Camilla Aas Rustad et al., 2011**.

[46] Samuel Bazzi et al., 2022. **The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia**, *Rev. Econ. Stat.* 104:4, 764–79.

[47] **Samantha Kuzma et al., 2020**; **Jonas Vestby et al., 2022**.

[48] Paola Vesco et al., 2022. **United They Stand: Findings from an Escalation Prediction Competition**, *Int. Interact.*

Predict risk, not events

Although we do not preclude the possibility that classification models become feasible for application in anticipatory action in the future, we recommend focusing on risk prediction models. The use of human forecasters, agent-based modeling and simulation,[50] or even scenario building,[51] could be used to improve upon and develop conflict risk models. Answering the open questions around probabilistic risk prediction is likely simpler than improving the performance of classification. Risk models may have good performance and if well-calibrated would present valuable information for use in humanitarian response, although how to apply it for anticipatory action remains unclear.

Predict the impact of conflict

In response to the war in Ukraine, predictions have proliferated on the impact of the crisis on global food insecurity due to a number of factors, including disruption of agricultural production in the region, disruption to supply chains due to sanctions, and reduced funding for other crises.[51] The current intensity and interstate nature of the Ukraine conflict make it a relatively unique context for generating predictions.

However, this is not always the case and it is therefore a critical area for further research. The humanitarian impacts of conflict, depending on its type, scale and many other factors, are extremely diffuse. Meaningful anticipatory action on conflict would require not just the prediction of conflict events or risk, but also an ability to extend that prediction to the humanitarian impact of the conflict. Only by knowing the impact could anticipatory action be effective in reducing vulnerability. Understanding the relationship between conflict and humanitarian impact is therefore critical for understanding how to set the parameters for any model: defining conflict onset or escalation or the type of conflict to predict relies on how linked these elements are to humanitarian impact.

To better link predictions of conflict to humanitarian impact, we should explore, among other things, the use of more locally generated data, the incorporation of human judgment, and alternative modeling techniques to make models more sensitive to dynamics on the ground.

4.3 PRACTICE TRANSPARENT DEVELOPMENT AND EVALUATION OF MODELS

While some of the methods described above have been implemented, there is often a lack of clear and transparent testing and publication of model performance, including through the use of peer review processes. All conflict prediction research, development and implementation should use transparent and reproducible methods that clearly communicate their performance. Open communication efforts are critical to building trust in complex models and are a prerequisite for achieving feasibility.

Transparency should not be restricted to the development of quantitative modeling. All conflict prediction exercises, even human forecast-based methods, should strive to test and validate approaches. Validation will allow end users to assign credibility to high performing models and recognize where further improvements are needed prior to implementation.[53] Although beyond the scope of this paper, ethical review of models and their applications is also critical given the sensitivities of conflict-related programming.

[50] Marina Andrijevic et al., 2020. **Governance in Socioeconomic Pathways and Its Role for Future Adaptive Capacity**, *Nat. Sustain.* 3:1, 35–41.

[51] World Food Programme, 2022. **Food Security Implications of the Ukraine Conflict**; Food and Agriculture Organization, 2022. **Note on the Impact of the War on Food Security in Ukraine**.

[52] Dan Maxwell et al., 2021. **Seeing in the Dark: Real-Time Monitoring in Humanitarian Crises**, *Tufts Feinstein International Center*.

[53] **Philip E. Tetlock, Barbara A. Mellers, and J. Peter Scoblic, 2017**.

5. CONCLUSION

Based on the review of the current literature, we return to the original questions[54] framing our research.

How accurate are conflict forecasts?

For classification models, the current evidence shows that model performance, particularly when predicting conflict onset or benchmarked against a simple model using conflict history, is not sufficient for application in humanitarian response. This is in addition to the missing link between predicted conflict and humanitarian impact.

Risk and continuous prediction modeling requires additional evidence on how accurate risk estimates correspond to observed risk for responsible implementation and adoption. However, their applications in anticipatory action are less clear due to the lack of thresholds, which are required to automate the release of funds.

How well can various sources predict different types of conflict in specific situations?

For the purpose of this paper, we mainly focused on the overall performance of conflict prediction models rather than the predictive power of various sources. However, it is clear that conflict history is the best predictor of future conflict but its utility is limited to ongoing conflicts. While some sources can improve model performance, the best performing models are often those that use the widest range of available data. But critically, these models often predict conflict at a scale too small for meaningful application in humanitarian response, and do not significantly outperform simple models solely built on conflict history.[55]

We therefore conclude that the current set of classification models in use or under development do not have sufficient predictive performance for anticipatory action. The feasibility of application for risk and continuous models remains an open question that requires further research.

5.1 RECOMMENDATIONS AND NEXT STEPS

To make use of conflict prediction for anticipatory action in the humanitarian sector, we recommend to focus on six areas.

Utilize flexible models that do not pre-suppose a theoretical framework of conflict causality. Future work on developing conflict prediction models should build on the current literature. The most fruitful areas of exploration will likely be found using ensemble models or non-linear techniques that require little theoretical knowledge and the use of as many input predictors as available.

Focus models on predicting shifts in conflicts, such as an increase in intensity or onset. This could help transition the usefulness of conflict prediction from use in preparedness and disaster risk reduction to anticipatory action, where acting prior to a new or intensifying shock is the point.

Explore the use of human inputs through superforecasters or prediction markets and use local data to improve model performance in specific contexts. Work should be done to include and validate the usefulness of human inputs into model frameworks, such as through superforecasters or prediction markets. Models can be validated and tested against their ability to predict the timing of specific events or conflict shifts.

[54] Marie Wagner and Catalina Jaime, 2020.

[55] Hannes Mueller and Christopher Rauh, 2022b.

Improve predictions on the humanitarian impact of conflict as opposed to conflict itself. Work on predicting the humanitarian impact of conflict should be prioritized. Understanding the relationship between humanitarian needs and conflict is necessary for any predictions to be acted on. Even without predicting the onset or escalation of a conflict, improvements in predicting the humanitarian impacts of conflict could potentially enable anticipatory action on the impacts of observed conflict. This work could be undertaken through better integration of conflict data and modeling with existing forecasting systems, which is already being explored for predicting food insecurity.[56]

Ensure that model development and evaluation is done in a reproducible and transparent way that highlights the model performance in all relevant metrics. Humanitarian work on measuring risk,[57] predicting displacement,[58] and other applications of predictive analytics[59] often do not communicate validation statistics and evidence of good performance. Recent efforts by governments[60] and academics to predict conflict for use in humanitarian response[61] also fail to show performance sufficient to what is required for use in decision making. Regardless of the approach or methodology,[62] it is critical that all models are transparent and that their performance is validated.

Learn from the state-of-the-art research underway in the academic field and ensure that applied research is relevant for humanitarian decision making. Significant time and resources will be required to collect data, develop models and test performance of new observations, particularly if models are extended to predict humanitarian impact. It is only through engagement between conflict researchers, academia, the private sector and the humanitarian community that this work will be sustainably financed and developed.

We hope that the above recommendations provide an initial framework and understanding for these conversations. We welcome questions and feedback at centrehumdata@un.org.

[56] Joris J. L. Westerveld et al., 2021.

[57] The Joint Research Center of European Commission, INFORM - Global, Open-Source Risk Assessment for Humanitarian Crises and Disasters.

[58] Danish Refugee Council, 2021; Christopher Earney and Rebeca Moreno Jimenez, 2019. Pioneering Predictive Analytics for Decision-Making in Forced Displacement Contexts, *Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study*, 101-19.

[59] Kevin Hernandez and Tony Roberts, 2020. Predictive Analytics in Humanitarian Action: A Preliminary Mapping and Analysis, *Institute of Development Studies*.

[60] Sarah Bressan, 2021. Crisis Early Warning: Berlin's Path From Foresight to Prevention, *Peace Lab*; Predicting Conflict – a Year in Advance, *The Alan Turing Institute*.

[61] Håvard Hegre et al., 2022.

[62] Laura Hielkema and Jasmijn Suidman, 2021. Multi-Hazard Risk Analysis Methodologies, *Anticipation Hub*.

TECHNICAL ANNEX

1. TYPES OF PREDICTION

As described in the body of this paper, the three model types are:

- Classification: There will be conflict.
- Risk prediction: There is a 75 percent probability of conflict.
- Continuous prediction: There will be 33 fatalities due to conflict.

We describe each model and assess their usefulness and performance in more detail below.

1.1 CLASSIFICATION

Classification models categorically state whether or not a conflict will occur in a particular area and time, through either binary (yes/no) or multiclass classification (e.g., major, minor or no conflict, or a 1-5 scale). For binary models, the classifications are whether or not conflict will or will not occur in that single time period, which may or may not be different from the previous period (e.g., predicting peace in one period but conflict in the next).

Classification models are typically built on top of a probabilistic model. Thresholds are then determined that convert the probability of conflict into discrete classes, such as by predicting that conflict will only occur if the probability is greater than a specific value. For instance, recall the Mali example where conflict is defined as occurring if there are 25 or more conflict-related fatalities in one month. Using a probability threshold of 50 percent, a risk estimate of 75 percent would result in a classification forecast that conflict will occur in February 2022 resulting in 25 or more fatalities.

1.2 RISK PREDICTION

Risk prediction is designed to generate a measure of underlying risk of conflict, usually produced as a probability of conflict at a certain geographic scale within a certain timeframe. These are often the probabilistic models that underlie classification models.

Predictions are typically generated from a set of indicators, e.g., gross domestic product or number of conflict events, that are fed into the model. The probabilities typically range from near zero percent in contexts that have not seen conflict in years to close to 100 percent in active conflict zones. Examples of this type of model are the ViEWS Risk Monitor[63] and related conflict forecasting systems,[64] which usually publish forecasts as risk predictions. For instance, in the December 2021 ViEWS Risk Monitor, the risk of state-based violence resulting in 25 or more fatalities in Mali by February 2022 was estimated to be greater than 75 percent.

1.3 CONTINUOUS PREDICTION

Continuous prediction models are those that predict a specific measure of conflict, such as the number of fatalities or conflict events.[65] Since these models do not define conflict thresholds, they are potentially better able to capture gradations in severity and magnitude that may be missed in risk prediction or classification. In the case of Mali, the continuous prediction for February 2022 could be an exact figure of 33 fatalities.

[63] The ViEWS team, 2021. **The Risk Monitor: December 2021**.

[64] **Conflict Forecast system**, for example.

[65] Frank DW Witmer et al., 2017. **Subnational Violent Conflict Forecasts for Sub-Saharan Africa, 2015–65, Using Climate-Sensitive Models**, *J. Peace Res.* 54:2, 175–92; Benjamin E. Bagozzi, 2015. **Forecasting Civil Conflict with Zero-Inflated Count Models**, *Civil Wars* 17:1, 1–24; N. Johnson et al., 2018. **Self-Exciting Point Process Models for Political Conflict Forecasting**, *Eur. J. Appl. Math.* 29:4, 685–707.

1.4 MODEL USEFULNESS

Classification outputs are often more useful to policy makers due to their ease of interpretation and actionability, but may not always be preferable to other models based on measures of predictive performance. For example, imagine a model that generates probabilities of conflict with the prediction based on a specific threshold. Predictions greater than or equal to the threshold will be predicted as being in conflict, and those below as not in conflict. This requires decent separability of the data, which means that the threshold clearly delineates when conflict will occur versus when it will not. If that is the case, there is high confidence that conflict will occur if the threshold is met.

Conversely, risk prediction models can be used even if we do not have the confidence in predicting a specific event. Intuitively, when a risk model predicts a 50 percent probability of conflict, the model performs well if conflict occurs approximately 50 percent of the time. However, a threshold of 50 percent for classification would generate false positives 50 percent of the time.

2. MODEL PERFORMANCE

There are a variety of metrics used to assess model performance. The main difference is between measuring calibration or discrimination.

Calibration is how well the probabilities of a model match observed frequencies. Perfect calibration means if we predict a 50 percent probability of conflict, it is likely to occur 50 percent of the time.[66] Poor calibration means that the predicted probability is not representative of the actual likelihood of conflict.

Discrimination is how well the model is able to separate two or more classifications (e.g., conflict and no conflict) in the data using a specific threshold. If there was a threshold of 50 percent probability of conflict, perfect discrimination would be that there is always conflict if the probability is higher than 50 percent, and never conflict if the probability is lower. Poor discrimination could be where conflict is just as likely to occur whether or not the prediction is above the threshold.

Models can be well-calibrated even when they have poor discrimination (and vice-versa).[67]

MODEL PERFORMANCE METRICS

Measures of calibration for classification and risk prediction models include:

- Brier score: Measures how close predicted probabilities of conflict are to the observed truth.
- Continuous risk probability score (CRPS): Measures the accuracy of probabilistic forecasts against the observed distribution of outcomes.[68]
- Hosmer-Lemeshow test: Measures how closely the predicted probabilities match the observed occurrence of events.[69]

[66] Colin G. Walsh, Kavya Sharman, and George Hripcsak, 2017. **Beyond Discrimination: A Comparison of Calibration Methods and Clinical Usefulness of Predictive Models of Readmission Risk**, *J. Biomed. Inform.* 76, 9–18.

[67] R. B. D'Agostino and Byung-Ho Nam, 2003. **Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures**, *Handb. Stats.* 23, 1–25.

[68] Tilmann Gneiting and Adrian E Raftery, 2007. **Strictly Proper Scoring Rules, Prediction, and Estimation**, *J. Am. Stat. Assoc.* 102:477, 359–78.

[69] A. Donati et al., 2004. **A New and Feasible Model for Predicting Operative Risk**, *Br. J. Anaesth.* 93:3, 393–99.

Measures of discrimination for classification and risk prediction models include:

- Area Under the Receiver Operating Characteristic Curve (ROC AUC): The ability of the model to separate conflict events from non-conflict events at each potential probability threshold.
- Area Under the Precision-Recall Curve (AUPRC): Measure of the balance between precision and recall that does not use true negatives in its calculation. This makes it potentially important for conflict prediction to ensure that model evaluation is not solely driven by true predictions of 'no conflict'.
- **Precision:** The ability of the model to only predict true conflicts. Only applicable to classification models with set thresholds.
- **Recall:** The ability of the model to predict all conflicts. Only applicable to classification models with set thresholds.
- **F1:** Balanced measure of the model based on precision and recall. Only applicable to classification models with set thresholds.

Metrics for continuous prediction models include:

- Mean Squared Error (MSE): Common and simple measure of how close the predicted value is to the observed value.
- Pseudo-Earth Mover Divergence (pEMDiv): A more complex measure that ensures that predictions that are only slightly off geographically or temporally are still slightly rewarded, rather than static accuracy measured for each prediction only against observations in that exact location and point in time.
- Targeted Absolute Difference with Direction Augmentation (TADDA): Relevant for predictions of change but not point predictions. It is a measure of general accuracy that provides an additional penalty if the predicted direction of change is incorrect.

2.1 PERFORMANCE OF CLASSIFICATION MODELS

The performance of classification models is typically assessed in the literature by a few key metrics and are calculated on the actual outcomes (e.g., conflict did or did not occur compared to the prediction). Taken together, they give a sense of a model's usefulness and how it should or should not be applied.

Advancements in model performance

Over the past two decades, the academic research around conflict prediction has vastly expanded, and overall, model performance has improved.[70] These improvements have come on the back of developments in improving global databases of conflict, such as ACLED,[71] the UCDP events database,[72] or CAMEO event data,[73] which provide easier access to data for model development. New techniques for non-linear learning of complex data, such as tree-based models, random forests, and neural networks, have allowed a more complex use of predictor variables without presupposing a theoretical framework for conflict causality, producing more powerful and robust predictions.[74, 75]

[70] Michael Ward et al., 2013.

[71] Clionadh Raleigh et al., 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature, *J.Peace Res.* 47:5, 651–60.

[72] Ralph Sundberg and Erik Melander, 2013. Introducing the UCDP Georeferenced Event Dataset, *J. Peace Res.* 50:4, 523–32.

[73] Deborah J. Gerner, Philip A. Schrodt, and Ömür Yilmaz, 2008. Conflict and Mediation Event Observations (CAMEO): An Event Data Framework for a Post-Cold War World, *International Conflict Mediation*, 287–304.

[74] Patrick T. Brandt et al., 2022. Conflict Forecasting with Event Data and Spatio-Temporal Graph Convolutional Networks, *Int. Interact.*

[75] Nathaniel Beck, Gary King, and Langche Zeng, 2000.

Results in the past few years have seen increased predictive performance, with new iterations of the ViEWS models released,[76] specific contextual models developed and benchmarked against ViEWS,[77] and automated machine learning models outperforming these baselines.[78]

Predictive power of conflict history

While these advances have been driven by the development of complex methodologies, much of the predictive power of these models is based on a history of conflict. Where reported in the literature, models with additional predictors often fail to significantly outperform simple models that just use ongoing conflict to predict future conflict.

	WPS model	Simple model
Recall	0.86	0.71
Precision	0.47	0.73
F2	0.74	0.71
ROC AUC	0.89	0.84
AUPRC	0.42	0.55
Brier score	0.084	0.057

Water, Peace and Security model underperformed a simple model on key metrics such as AUPR and Brier score[79]

Above, we can see that a simple model, built solely using only ongoing conflict, outperforms the complex model in certain key metrics, in particular AUPR, a robust statistic for rare-event classification, and the Brier score, a key forecast loss statistics.[80] New research on deep learning models trained solely on historical conflict data show the capacity of models only relying on conflict history to outperform those with more complex inputs.[81]

These findings highlight two points on the state of conflict prediction:

1. Much of the performance seen in model evaluations is based on predicting no change in the state of conflict, simply a continuation of peace or conflict seen in the previous time period.

2. Until complex models significantly outperform ongoing conflict as a predictor, allocating resources based on where conflict is already occurring would more efficiently reach current and future affected populations than allocating based on complex models.

Difficulty in predicting conflict onset

There is a distinction between predicting whether conflict is going to continue in its current state (e.g., classifying the forecasted period to be the same as the current period) and predicting the onset of a conflict where there has not been one previously (e.g., predicting a change in classification between two periods).

[76] Håvard Hegre et al., 2021.

[77] Samantha Kuzma et al., 2020.

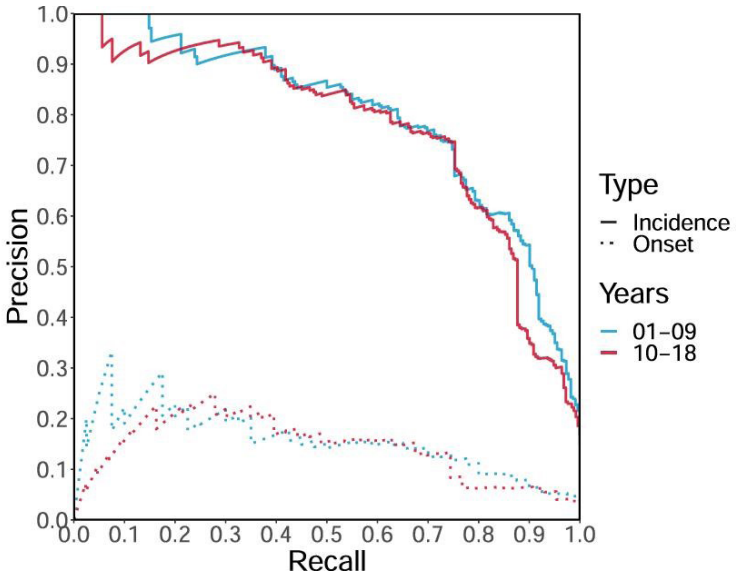
[78] Vito D'Orazio and Yu Lin, 2022. Forecasting Conflict in Africa with Automated Machine Learning Systems, *Int. Interact.*; Vito D'Orazio et al., 2019. Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning, *2019 Int. Conf. Big Data*, 4714–23.

[79] Samantha Kuzma et al., 2020.

[80] Ibid.

[81] Iris Malone, 2022. Recurrent Neural Networks for Conflict Forecasting, *Int. Interact.*

When reported, the performance of models evaluated against onset prediction is relatively poor,[82] including older models re-evaluated against new historical data,[83] new models built on text data (e.g., newspaper reports)[84] and newer research that calculates these benchmarks by default.[85] These difficulties are seen across related disciplines, from forecasting genocide[86] to political instability and coups.[87]



Difference in performance, predicting all conflict versus predicting onset[88]

The figure above plots the precision and recall of a model at all potential probability thresholds, where values to the top of the plot have better precision and to the right of the plot better recall. We can see the significant drop in performance when predicting the onset of conflict (dotted lines) compared to predicting all incidents of conflict (solid lines). This problem also applies when predicting changes in the level of conflict.

Efficacy of classification models in predicting conflict

Overall, the evidence from the literature shows that classification models lack the performance and granularity needed for anticipatory action. Given the high levels of humanitarian need driven by ongoing conflict and the dominance of ongoing conflict in predicting future conflict, responding to current needs would reach populations most likely to experience the impacts of future conflict. This is different from other shocks, such as flooding, where you cannot anticipate future flooding simply by looking at current flooding.

Given the poor performance in predicting the onset of new conflicts or intensification of ongoing conflict, there is little justification for using classification models to drive anticipatory action on conflict. Improvements in conflict prediction are needed for these methods to become operational.

2.2 PERFORMANCE OF RISK PREDICTION

Assessing the feasibility of risk prediction is a separate question from the feasibility of classification. Risk prediction requires the predicted probability of conflict to closely match the observed frequency, (i.e., calibration described in the above methods section).

[82] Samantha Kuzma et al., 2020.

[83] Harvard Hegre, Harvard Mokleiv Nygard, and Peder Landsverk, 2021. Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality, *Int. Stud. Q.* 65:3, 660–68.

[84] Thomas Chadefaux, 2014. Early Warning Signals for War in the News, *J. Peace Res.* 51:1, 5–18; Hannes Mueller and Christopher Rauh, 2022a. The Hard Problem of Prediction for Conflict Prevention, *J. Eur. Econ. Assoc.* jvac025.

[85] Michael Ward et al., 2013.

[86] Benjamin Goldsmith and Charles Butcher, 2017.

[87] Drew Bowlsby et al., 2020.

[88] Harvard Hegre, Harvard Mokleiv Nygard, and Peder Landsverk, 2021.

Lack of validation on calibration

Many classification models actually present their results as estimates of conflict risk.[89] Yet, the model validation and testing often use metrics designed to test how well the model discriminates between when conflict will or will not occur.[90] These measurements of error do not inform the direct usefulness of the probability itself, which requires testing the calibration of the model as described on page fourteen.

Proving the performance of risk models requires model validation on its calibration to gauge how well predicted probabilities perform relative to observed distributions.[91] These critical prerequisites for the responsible adoption of risk modeling for conflict are not always presented in the conflict forecasting literature, which is often focused on the larger goal of predicting exactly when conflict will occur (and thus relies on measures of discrimination). Although measures of calibration such as the Brier score are frequently used, they are not explored in the same detail as measures of discrimination, with limited plotting of metrics or additional measures presented. Having these readily available would provide more evidence on how the risk probability measures match empirically observed risk and allow end users to interpret and apply the probabilities in planning humanitarian operations.

As with classification models, we propose there should also be evidence that risk estimates perform well when predicting the risk of conflict onset, not just in existing conflict scenarios. Applications of risk models should also take into account the model limitations when extending predictions of conflict risk to humanitarian risk, given the scale of conflict and the lack of clear linkages to the risk of humanitarian need that anticipatory action responds to. This may indicate a need for measuring the calibration of models when predicting large-scale escalation or onset, where the scale and potential humanitarian impact are higher, a method being explored for forecasts of food security.[92]

Efficacy of risk models in predicting conflict

Unlike the evidence around classification models, there is often a lack of sufficient exploration of model calibration and the performance of conflict risk estimates relative to observed risk. To determine feasibility of application to anticipatory action, measures of risk estimate performance from other fields, such as the Hosmer-Lemeshow test and calibration plots, should be used. Validating how well-calibrated risk prediction models are is a prerequisite for determining feasibility alongside connecting conflict risk to risk of humanitarian impact.

2.3 PERFORMANCE OF CONTINUOUS PREDICTION

Continuous prediction models have historically been less common in the conflict prediction space. However, focus on these models has increased in recent years.

The ViEWS research team organized an escalation prediction competition in 2020 and 2021, producing a spate of new research on the topic specifically aimed at solving issues mentioned previously, particularly the ability of models to predict significant changes in conflict.[93] While the competition models predicted change in fatalities due to conflict, rather than point estimates of conflict, they are still continuous prediction models. A suite of appropriate metrics were defined for the competition that aim to reward different aspects of a model's performance, and which would be useful for any future research to build on.[94]

[89] Håvard Hegre et al., 2022.

[90] The Joint Research Center of European Commission, 2017. The Global Conflict Risk Index (GCRI) Regression Model: Data Ingestion, Processing, and Output Methods.

[91] Colin G. Walsh, Kavya Sharman, and George Hripcsak, 2017.

[92] Yujun Zhou et al., 2022. Machine Learning for Food Security: Principles for Transparency and Usability, *Appl. Econ. Perspect. Policy* 44:2, 893–910.

[93] Håvard Hegre, Paola Vesco, and Michael Colaresi, 2022. Lessons From an Escalation Prediction Competition, *Int. Interact.*

[94] Paola Vesco et al., 2022.

Burgeoning research but similar issues

The research spurred on from the prediction competition has attempted to address many of the problems discussed above. New models that allow for representation of complex conflict dynamics show increased performance in predicting when existing conflict may escalate.[95,96] Other models that are built upon dynamic data that captures shifts in conflict risk, such as text data, perform better on predicting the outbreak of conflict in peaceful settings.[97] A multitude of other advances and approaches are detailed in the summary paper on the ViEWS competition.[98]

Efficacy of continuous prediction models in predicting conflict

Unlike risk prediction models, substantive work has been put into developing metrics for assessing performance of continuous prediction models. The scale of conflict predicted remains too small for likely application in anticipatory action. Performance at the extremes of the distributions should also be tested to see if models are able to predict large-scale escalations in conflict. Most importantly, work on fatality prediction still has no direct linkage to humanitarian impact prediction.

[95] Iris Malone, 2022.

[96] Ibid; Patrick T. Brandt et al., 2022.

[97] Christian Oswald and Daniel Ohrenhofer, 2022. Click, Click Boom: Using Wikipedia Data to Predict Changes in Battle-Related Deaths, *Int. Interact.*

[98] Håvard Hegre, Paola Vesco, and Michael Colaresi, 2022.