Policy Research Working Paper 9412

Predicting Food Crises

Bo Pieter Johannes Andrée
Andres Chamorro
Aart Kraay
Phoebe Spencer
Dieter Wang

WORLD BANK GROUP
Fragility, Conflict and Violence Global Theme
&
Development Economics Vice Presidency
September 2020

Abstract

Globally, more than 130 million people are estimated to be in food crisis. These humanitarian disasters are associated with severe impacts on livelihoods that can reverse years of development gains. The existing outlooks of crisis-affected populations rely on expert assessment of evidence and are limited in their temporal frequency and ability to look beyond several months. This paper presents a statistical foresting approach to predict the outbreak of food crises with sufficient lead time for preventive action. Different use cases are explored related to possible alternative targeting policies and the levels at which finance is typically unlocked. The results indicate that, particularly at longer forecasting horizons, the statistical predictions compare favorably to expert-based outlooks. The paper concludes that statistical models demonstrate good ability to detect future outbreaks of food crises and that using statistical forecasting approaches may help increase lead time for action.

This paper is a product of the Fragility, Conflict and Violence Global Theme and the Development Economics Vice Presidency. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at bandree@worldbank.org.

Produced by the Research Support Team

Predicting Food Crises

Bo Pieter Johannes Andrée[a,*], Andres Chamorro[a], Aart Kraay[a],
Phoebe Spencer[a], and Dieter Wang[a]

Keywords: Famine, Food Insecurity, Extreme Events, Unbalanced Data, Cost-sensitive learning.

JEL: C01, C14, C25, C53, O10

[a] *World Bank,* [*] *email: bandree@worldbank.org.* This work was prepared as background for the Famine Action Mechanism (FAM). Support from the State and Peace-Building (SPF) Trust Fund (Grants N. TF0A7049 and TF0A5070) is gratefully acknowledged. The authors would like to thank Nadia Piffaretti, Zacharey Carmichael, Harun Dogo, Arif Hussain, Luca Russo, José Lopez, Colin Bruce, Nick Haan, Frank Davenport, Dan Maxwell, Joanna Macrae, Soomin Park, Marco Zambotti, Sardar Azari, Therese Norman-Monroe, Jacob LaRiviere, and the IPC, WFP mVAM, and FAO teams for invaluable contributions in the initial phase of this work. In particular, we'd like to thank the participants of the FAM Workshop in Geneva on February 2018 hosted by ICRC, Artemis Working Days in Rome on April 2018 hosted by WFP, the FAM Data and Analytics meetings with global tech partners in Rome and New York on September 2018, and the participants to the Predictive Analytics workshop hosted by UN OCHA, at the Center for Humanitarian data in the Hague in April 2019. This paper reflects the views of the authors, and does not reflect the official views of the World Bank, its Executive Directors, or the countries they represent.

1 Introduction

Despite progress in reducing poverty in recent decades (World Bank Group, 2018), one in nine people in the world faces hunger (FAO et al., 2019). More than 130 million people are currently estimated to be in food crisis (Food Security Information Network, 2020), meaning that they are able to meet minimum dietary needs only through irreversible coping strategies such as liquidating livelihood assets.

Food crises reflect the complex interactions of conflict, poverty, extreme weather, climate, and food price shocks (Misselhorn, 2005; Headey, 2011; Singh, 2012; D'Souza and Jolliffe, 2013), that compound in the presence of long-standing structural factors (Maxwell and Fitzpatrick, 2012). These humanitarian disasters are characterized by high levels of acute malnutrition that lead to mortality in vulnerable populations.[1] Beyond immediate loss of life, food crises have long-lasting consequences for survivors as well, including inter-generational health and education effects (Galler and Barrett, 2001; Veenendaal et al., 2013; Galler and Rabinowitz, 2014; Asfaw, 2016; Li et al., 2017). Recognizing these costs, the international community has long responded to food crises with humanitarian aid. In recent years, this has been complemented with a growing emphasis on investment in prevention, since in many cases it is more cost effective to prevent crises than to respond to them (Meerkatt et al., 2015; Mechler, 2016).

Targeted prevention policies require a capacity to predict when, where and how food crises will emerge, ideally well in advance (Kleinberg et al., 2015). To contribute to this capacity, this paper explores statistical approaches to predicting food crisis. Statistical models can offer systematized insight into the timing and location of possible future events, using readily-observable and verifiable data. Moreover, model-based forecasts can assign meaningful probabilities to future events, allowing expected costs, expected benefits and uncertainty to be quantified.

In the context of food crisis prediction, balancing false positives (predicting food crises when they do not occur) and false negatives (failing to predict food crises that do occur) is particularly important. Outbreaks of food crises are relatively rare events, but when they occur they are persistent. Failing to act early and prevent these protracted humanitarian disasters comes at a high human cost. On the other hand, scarce humanitarian resources limit the scope for responding to false positives, and hard choices must be made between investing in competing options for prevention, including saving funds to address future crises. These difficulties are compounded by the fact that in order to be useful, preventive measures must be taken well before all relevant information is available.

In this paper, we explore how formal statistical forecasting models can help to address these challenges. First, we generate monthly predictions at the sub-national level, optimizing for alternative targeting policies that assume different costs of false negatives relative to false positives. Second, because finance is often unlocked at a country level, predictions are developed for the total number of people in crisis-affected districts in a country. The two predictions can be derived from the same model and work in tandem: first highlighting the

[1] Black et al. (2013) conclude that almost half of child deaths globally are associated with undernutrition and Devereux (2000) estimates that deaths due to famines in the 20th century equal those of World Wars I and II combined.

countries that are in need; then helping to allocate finance toward the districts that face the highest risks. The empirical application uses 10 years of sub-annual assessments on district-level food insecurity outcomes in 21 developing countries together with monthly covariates that capture known drivers of food insecurity. The predictions are generated using a Random Forest and a novel approach is deployed to adjust probabilistic predictions to optimize the forecasting objectives of the paper. The forecasts are validated, up to a full year ahead, against holdout data and the accuracy is compared to historical non-model-based outlooks. The results indicate that, particularly at longer forecasting horizons, model-based predictions compare favorably. The paper concludes that statistical models demonstrate good ability to detect future crisis outbreaks for a set level of tolerance to false alarms and that using them may help increase lead time for action.

This research contributes to understanding the ability of statistical models to anticipate future food crisis in the particular context supporting early interventions. The paper builds on several earlier efforts.[2] Mellor (1986) discussed prevention strategies with an emphasis on economic weakness, crop failure, and subsequent price signals as leading indicators of famine. Price signals have been discussed further in particular by Seaman and Holt (1980), and their statistical significance has been investigated empirically in the context of the Ethiopian famine of 1972-1974 (Cutler, 1984) and the famine of 1984-1985 in Niger (Khan, 1994). Seaman (2000) explored a dynamic modeling approach to assess risks based on rapid household survey data and simulations of coping strategies under income and food supply shocks. More recently, researchers have begun to use machine learning techniques for prediction. For example, Mwebaze et al. (2010); Okori and Obua (2011) presented attempts at predicting household famine in Uganda between 2004 and 2005 using household data.

The importance of using systematic approaches to predicting food crises and allocating humanitarian resources effectively is likely to increase in coming years as key drivers of food insecurity are expected to worsen into the 21st century. Historical achievements in eradicating poverty have largely coincided with substantial degradation of environments (Stern et al., 1996; Andrée et al., 2019). Degrading environments, climatic extremes, and growing populations will continue to put pressure on future agricultural systems (Grainger, 1990; Ingram et al., 2010; Myers et al., 2017; Diogo et al., 2017). These developments place obstacles in the path to zero hunger, specifically for the rural poor that rely on local natural assets for income and food consumption (Duraiappah, 1998; Barrett and Bevis, 2015; Barbier and Hochard, 2018). A further benefit of model-based predictions of food insecurity is that they offer an alternative to the allocating of scarce resources based on qualitative assessments, instead relying on readily-verifiable and openly available data in a systematic way to predict the deterioration in food security when action is delayed.

The paper is structured as follows. Section 2 introduces the data, section 3 develops a framework for the validation and calibration of probabilistic predictions, paying particular focus to balancing false positives and false negatives. It then details an application using a Random Forest algorithm. Section 4 presents key results, and section 5 concludes. Additional results are found in the supplementary appendices.

[2] Apart from the food insecurity literature mentioned here, the paper also relates to the work of Celiku and Kraay (2017) who explore predicting conflict outbreaks in the similar context of weighted prediction loss functions.

2 Data

2.1 Target variable: food crisis outbreaks

The aim of this paper is to forecast transitions into critical states of food insecurity with enough lead time to take action. We obtain historical data on food insecurity from periodic assessments performed across 1,162 districts in 21 developing countries over the period from 2009 to 2020, obtained from FEWS NET.[3] Food insecurity is measured using the Integrated Phase Classification (IPC) system, an analytical framework that follows evidence-based guidelines to qualitatively classify the severity of food insecurity and prescribe policies to mitigate risk (Hillbruner and Moloney (2012) review the process). The IPC scale distinguishes five phases of food insecurity: (1) minimal, (2) stressed, (3) crisis, (4) emergency and (5) famine. Our predictions are focused on a binary "crisis indicator" defined as IPC ratings of crisis or worse, i.e. phases (3), (4), and (5). This choice is motivated by the fact that the IPC scale recommends a significant shift in policies in order to mitigate, rather than manage the risk of, severe malnutrition outcomes and death once a district enters crisis. In particular, during stressed (2) conditions and below the recommended focus is on risk management, while at crisis levels (3) and above this turns into urgent action to mitigate outcomes.

The FEWS NET data are reported at a sub-national level that follows a combination of administrative boundaries (admin2) and "livelihood zones".[4] A one-on-one mapping from livelihood zones to administrative districts does not necessarily exist and the former may change over time based on expert opinions. To obtain a consistent time series, we map all the data to standard admin2 districts using a spatial overlay, assigning values to districts by majority coverage. The FEWS NET data are reported at quarterly frequency from 2009 to 2016, and then every four months afterwards. Our sample of food crisis observations used for estimation and cross-validation purposes extends through February 2019. Data for June and October 2019, and February 2020, are considered purely for temporal holdout validation. We develop predictions up to a full year ahead, through February 2020, using covariates that run through February 2019. Some of the district assessments are missing. However, our approach allows us to generate predictions for all districts where covariates are available, even if food crisis outcomes were not.

The IPC rating scale is intended to reflect the actual on-the-ground conditions inclusive of any humanitarian assistance. Since we are interested in predicting food crises themselves and not the impacts of associated humanitarian response, we need to net out the latter. Since 2012, the FEWS NET data marks locations where humanitarian assistance was significant enough to have reduced the IPC phase by one. These markers are used to reconstruct the relevant minimum IPC phases that would have been issued absent humanitarian presence. Figure A.1 in supplementary Appendix A contains the resulting data and Table 1 summarizes.

[3] FEWS NET is a leading food insecurity information provider funded by USAID that regularly publishes analysis and data on current and future food insecurity outcomes.

[4] Admin2 districts are defined by Global Administrative Unit Layers (GAUL); livelihood zones are available from FEWS NET and define sub-national geographic areas of a country, that may cut through administrative borders, in which people are assessed to share similar options for obtaining food and income and have similar access to markets.

The summary statistics reveal several important basic insights. Population-weighted average IPC ratings across districts generally are lower than simple averages, indicating that food insecurity is more common in sparsely-populated districts. The minima and maxima highlight that most countries have at some point experienced crisis or worse, but famine declarations (IPC rating of 5) are extremely rare occurrences. Zambia is the only country in our sample that has not experienced food crisis in this period. The number of observations available in each country is listed, with "Crisis+" denoting the number of food crisis observations. "Outbreaks" correspond to the subset of food crisis observations that were preceded by a phase 1 or phase 2 observation. Predicting these transitions into food crisis is the primary goal of this paper. While approximately 16% of observations correspond to food crises, only a small share of these classify as transitions into new outbreaks of food crisis (1,723/6,143 $\sim$ 28%). This highlights why the issuance of timely warnings is important: over two-third of classified transitions into crisis are followed by extended periods of crisis, at which point it is too late for prevention.

Table 1: Descriptives of FEWS NET ratings aggregated to admin2, netted of humanitarian assistance effects, used for estimation and cross-validation.

	IPC Rating				**Crisis Indicator**		**Observations**			**Coverage**	
Country	Mean	Pop. Mean	Min	Max	Mean	Pop. Mean	Total	Crisis+	Outbreaks	Districts	Time Periods
Afghanistan	1.41	1.40	1	4	0.08	0.07	1,224	94	59	34	36
Burkina Faso	1.08	1.08	1	3	0.00	0.00	1,440	5	5	45	32^{c}
Chad	1.45	1.41	1	4	0.10	0.09	2,448	251	129	68	36
Congo, Dem. Rep.	1.91	2.00	1	4	0.21	0.26	218	46	11	50^{a}	8^{d}
Ethiopia	1.83	1.74	1	4	0.26	0.22	2,664	681	174	74	36
Guatemala	1.43	1.30	1	3	0.08	0.05	792	62	28	22	36
Haiti	1.51	1.38	1	3	0.07	0.04	1,512	103	64	42	36
Kenya	1.46	1.42	1	4	0.08	0.08	2,556	213	60	71	36
Malawi	1.48	1.48	1	4	0.16	0.16	972	152	87	27	36
Mali	1.19	1.10	1	4	0.04	0.01	1,800	68	48	50	36
Mauritania	1.43	1.44	1	4	0.07	0.09	442	32	16	13	34^{c}
Mozambique	1.18	1.13	1	3	0.05	0.04	396	21	7	11	36
Niger	1.36	1.27	1	4	0.07	0.05	2,412	173	94	67	36
Nigeria	1.78	1.20	1	4	0.20	0.02	3,564	710	231	99	36
Somalia	2.48	2.44	1	5	0.40	0.37	2,590	1,040	188	74	35^{d}
South Sudan	2.13	2.16	1	5	0.33	0.35	2,808	937	232	78	36
Sudan	1.41	1.34	1	4	0.10	0.07	2,640	265	107	80	33^{d}
Uganda	1.10	1.06	1	3	0.02	0.01	2,016	33	12	56	36
Yemen, Rep.	3.42	3.26	2	4	0.93	0.89	960	892	35	69^{b}	$14^{b,d}$
Zambia	1.02	1.01	1	2	0.00	0.00	2,304	0	0	72	32^{c}
Zimbabwe	1.54	1.44	1	4	0.17	0.13	2,160	365	136	60	36
Total	1.60	1.41	1	5	0.16	0.10	37,918	6,143	1,723	1,162	116 **months**^{e}

The descriptive statistics are calculated using FEWS NET ratings, netted of humanitarian assistance effects, taking admin2 districts as the level of observation. "IPC Rating" refers to the netted data on the original 1-5 scale, while "Crisis Indicator" refers to the binary indicators of phases (3) and above. The data dimensions under "Coverage" refer to the number of admin2 districts in the country, disregard of whether all of these districts cover livelihood zones that receive ratings from FEWS NET, and the number of assessment periods from the first assessment to the last (February 2019). The June/October 2019 and February 2020 assessments are used only for validation and not included in this table. Further guidance on data coverage is as follows: ^{a} partially covered, ^{b} missing at random, ^{c} latest assessment periods missing, ^{d} earliest assessment periods missing, ^{e} from first to last assessment: July 2009 - February 2019.

2.2 Covariates of food crisis events

We forecast transitions into food crisis using only readily-observable covariates, available monthly at the admin2 district level prior to outcomes. We deliberately do not use lagged values of the IPC ratings as predictors.[5] This allows us to generate predictions that can be updated monthly based on available data, independently of whether recent IPC ratings are available. We distinguish four key groups of covariates:

Structural factors

Structural factors capture time-invariant vulnerability to food insecurity. We proxy these at the district level with spatial trends, population counts, land area, terrain ruggedness and land use shares (cropland and pasture). In addition, we include fixed effects to capture basic country-specific seasonal variation. Nonlinear prediction models like the Random Forest we deploy can exploit interactions between any (combination) of structural factors and other time-varying covariates to capture heterogeneity.

Environmental factors

We rely on remote sensing data to track environmental factors relating to food production. Food production depends on soil moisture content, which can be monitored through the inputs (rainfall) and outputs (evapo-transpiration) of the water balance equation. We also proxy for food plant health directly using the normalized difference vegetation index (NDVI), which is a commonly-used satellite-imagery-based measure of vegetation coverage (Ross et al., 2009; Brown, 2010).

Violent conflict

Violent conflict is an important factor that can impact food insecurity, by disrupting social systems and blocking access, or may signal other factors that contribute to food insecurity. For example, riots and protests may be the sign of an ailing economy, while violent attacks on civilians by non-state actors can cause displacement of people and disrupt food production and distribution (Brück and D'Errico, 2019; Maxwell et al., 2020). We use the count of conflict events and the number of fatalities in order to capture events of differing intensity, ranging from protests and riots to lethal attacks.

Food price inflation

Food price inflation can directly cause food insecurity by raising the cost of living for vulnerable households (Baffes et al., 2008; Conceição and Mendoza, 2009). Rising prices can also signal whether fiscal space is adequate or tightened, for example due to episodes of conflict (Adam et al., 2008). Over longer time horizons, price uncertainty may lead businesses to postpone investments, see (Koomen et al., 2015; Andrée et al., 2017) for the specific case of agricultural investments. We use monthly sub-national food price inflation estimates developed by Andrée (2020a) to capture these important dynamics.[6]

[5] The impact on the forecasting performance is insignificant but the choice results in larger models. Note that, under suitable conditions, a finite-order moving average approximation of a lower-order autoregressive process remains reasonably accurate. Since innovations can be substituted by exogenous variables by defining a process for lower-level constituents, stable autoregressive processes may also be approximated using a sufficient number of lags of relevant exogenous variables.

[6] The paper assembles sporadically available food price data collected by the World Food Programme, and uses a stochastic imputation that leverages unobserved cross-correlations between intermittent market-level price time series of individual food items to impute price data for up to six staple foods for all markets and months in a given country. These are aggregated into a simple equally-weighted geometric price index and interpolated geographically using inverse distance weighting.

2.3 Data processing

Table A.1 summarizes the basic covariates and their sources. To better serve decision-making, a more transparent feature set is preferred over a highly processed one. For this reason, we perform only minimal additional data processing. Running maxima of conflict counts and intensity are calculated to capture the worst events that have occurred in the recent past. Second, food price inflation and changes therein are calculated over alternative time windows.[7] After adding these features, spatial averages using 4-nearest-neighbors are calculated for all predictors to capture regional dynamics. The entire feature set is then lagged temporally, thus including temporal lags of spatial lags, to capture dependence on the recent and more distant past. Up to the 12-th lag is included of all time-varying predictors, always ensuring that only information prior to an outcome enters the models. For example, the 4-month ahead predictions use the 4-th to 12-th lag, while the 12-month ahead predictions only use the 12-th lag.

To keep the number of predictors at a modest level, we restrict the maximum linear correlation between any two predictors in the covariate space. In particular, we calculate pair-wise correlations and select predictors that have correlations greater than 0.75, and, from this subset of highly correlated predictors, drop the one with the highest overall correlation. We repeat this process iteratively until the 0.75 threshold is no longer breached. The result is a set of processed features that is similar in dimension to the original unprocessed feature set (including temporal lags), while capturing more diverse signals, see Appendix Table A.2.

To assess the benefit of these processing steps, we also evaluate the predictive performance of simply using the original unprocessed features of Table A.1 and their temporal lags.[8] We refer to this as the *"basic"* feature set, and the more processed feature set as the *"wide"* feature set. Results show that the *"wide"* feature set provides important benefits, hence it is the main focus of the paper. Additional detail on the comparison of the *"wide"* and *"basic"* feature sets is available in supplementary Appendix B.

3 Prediction and Validation Framework

3.1 Loss functions

We calibrate our predictive model to minimize prediction loss functions that penalize false positives (incorrectly predicting food crises that do not occur) and false negatives (failing to predict food crises that do occur). We allow for a range of relative weights on these two types of errors. To specify these prediction loss functions, the following notation is helpful. Recall that our target variable is a binary variable taking the value 1 if a food crisis (defined

[7] Specifically: a 3-month running maximum of the 3-month simple moving average of both conflict counts and death intensity, a 6-month running maximum of the 6-month simple moving average of both conflict variables, a 3-month and 6-month difference of the log food price index, and a 3-month, 6-month and 12-month difference of monthly year-to-year food price inflation.

[8] Allowing for pair-wise correlation up to 0.95 to stay as close to the "unprocessed" situation but avoiding possible numerical problems with some of the other, linear, prediction algorithms that were tried. The Random Forest leads to nearly unchanged results if this step were to be dropped as it performs feature selection. Note, however, that variable importance scores may not provide a realistic estimate as two highly correlated variables can both be used to split a decision tree without concrete preference for any one variable.

as IPC categories, 3, 4, or 5) is observed, and 0 otherwise. For convenience we sometimes refer to 0 and 1 generically as "class labels" and the observations corresponding to food crises as the "positive class". Define Y^* as a column vector containing the class labels, stacking all countries/districts/months where an IPC rating is observed. Define $\hat{Y}^*$ as the corresponding column vector of binary predictions, and $\hat{P}^* \in [0, 1]$ as a column vector of predicted probabilities of the positive class.

The binary outcomes and corresponding predictions are defined for all entries in the data where an IPC rating is observed. However, since our primary focus is on predicting transitions into food crisis, we focus on a subset of observations when evaluating predictions. This subset consists of all country/district/month observations containing an IPC rating, and for which the previous IPC rating is 1 or 2, i.e. not a food crisis. That is, this subset consists of all non-crisis observations that either are followed by another non-crisis observation, or by a transition into food crisis. In this way, we eliminate from the validation set all observations that correspond to ongoing food crises and focus our validation exclusively on predictions of outbreaks of new food crises – before they occur. More specifically, define Y, $\hat{Y}$ and $\hat{P}$ as the sub-vectors of Y^*, $\hat{Y}^*$ and $\hat{P}^*$ containing the entries that were preceded by non-crises, let $\mathbb{I}$ denote a conformable vector of ones, and let $\mathrm{w} \in [0, 1]$ denote a scalar weight. We evaluate predictive performance using these two prediction loss functions:

$$L_A(Y, \hat{Y}; \mathrm{w}) = \mathrm{w}\frac{Y'(\mathbb{I} - \hat{Y})}{Y'Y} + (1 - \mathrm{w})\frac{(\mathbb{I} - Y)'\hat{Y}}{(\mathbb{I} - Y)'(\mathbb{I} - Y)}, \tag{1}$$

and

$$L_B(Y, \hat{P}; \mathrm{w}) = \mathrm{w}\frac{-Y'\ln(\hat{P})}{Y'Y} + (1 - \mathrm{w})\frac{-(\mathbb{I} - Y)'\ln(\mathbb{I} - \hat{P})}{(\mathbb{I} - Y)'(\mathbb{I} - Y)}. \tag{2}$$

Note that in L_B, the natural log operator $\ln(\cdot)$ is understood to apply element-wise. These loss functions are oriented so that lower values of L_A and L_B correspond to better predictions. Both loss functions are weighted averages of two terms corresponding to prediction errors made when the true state is 1 versus 0. L_A is a weighted average of the false negative rate and false positive rate while L_B is a similarly weighted Log Loss function.

The difference between Equation 1 and Equation 2 is that L_B simply replaces binary entries used to count the false positives and false negatives, with a continuous penalty that scales with the logarithm of the predicted probability. This means that, when a transition into food crisis occurs, models that fail to predict these transitions with a high probability are penalized exponentially-heavily for predicting values closer to 0. Vice-versa, if the model predicts a high probability of a transition into food crisis and it does not occur, the penalty again increases exponentially, only now for predictions closer to 1. Calibrating a prediction algorithm to minimize L_B will thus result in probabilities that are close to 1 (0) when transitions into food crises do (do not) occur. We calibrate our prediction models to minimize L_B, and we also report false negative and false positive rates together with L_A because of its more intuitive interpretation.[9]

[9] The motivation behind this approach is that L_B is a proper scoring rule (Gneiting and Raftery, 2007) that ranks models according to their distance to the optimal model (Andrée, 2020b). Minimizing L_B can be understood as maximizing the expected utility of a predictive distribution such that w determines utility trade-off. Propriety encourages the model to make careful predictions and to be honest about the level of

In both loss functions, the scalar parameter w governs the relative weight assigned to prediction errors made when the actual outcome is or is not a transition into food crisis. The paper considers three values with a clear policy interpretation. For identical rates,

- $\mathrm{w} = \frac{1}{3}$: failing to anticipate a crisis is half as costly as raising a false alarm,
- $\mathrm{w} = \frac{1}{2}$: failing to anticipate a crisis is as costly as raising a false alarm,
- $\mathrm{w} = \frac{2}{3}$: failing to anticipate a crisis is twice as costly as raising a false alarm.

In our context, this weighting of prediction errors is important since transitions into food crisis are rare but costly. Minimizing unweighted loss would result in a prediction model that emphasizes correct prediction of non-food crisis observations and neglects transitions into food crisis simply because they are rare. As an example, with our data, a model that predicts none of the transitions can score an overall Accuracy on all data of over 95%. In the weighted loss functions, model performance is instead quantified based on a pre-specified policy interpretable parameter rather than the data frequencies.

To calibrate predictions effectively for the different weightings, we follow the strategy proposed by Andrée and Kraay (2020). The idea here is simple. Predicted class probabilities that strongly discriminate between outcomes are more useful for policy purposes than ones that do not. For example, a prediction model that assigns probabilities of 0.49 to the 0 class and 0.51 to the 1 class is less useful to discriminate between risks than one that assigns a low (high) probability to the former (latter). Moreover, the weight in the loss function determines whether the preference for prediction bias is upward or downward. $L_B(\mathrm{w})$ can often be improved by adjusting both the level of confidence and the general direction of probabilities toward the high-cost outcome. We make use of a simple transformation of the predicted probabilities. The predicted class labels are generated according to:

$$\hat{Y} = \begin{cases} 1, \{\hat{P} = g(\bar{P}; \alpha, \beta)\} > .5 \\ 0, \{\hat{P} = g(\bar{P}; \alpha, \beta)\} \leq .5 \end{cases}. \tag{3}$$

where the predicted probability $\hat{P} = g(\bar{P}; \alpha, \beta)$ is a transformation that depends on two tuning parameters α and β and an input probability $\bar{P}$ generated by a model of interest as follows:

$$\hat{P} = g(\bar{P}; \alpha, \beta) = \begin{cases} \bar{P}^{\alpha} \beta^{(1-\alpha)} & \bar{P} \leq \beta \\ 1 - (1-\bar{P})^{\alpha}(1-\beta)^{1-\alpha} & \bar{P} > \beta \end{cases}. \tag{4}$$

The parameter $\beta \in [0, 1]$ plays the role of a cutoff probability above which the predicted probability for the positive class should be increased. The parameter $\alpha \in \mathbb{R}_{>0}$ controls the confidence of the transformed probabilities, with higher values of α bringing $\bar{P}$ closer to

certainty. In particular, L_B encourages confidence in probabilistic predictions but punishes over-confident predictions. This means that when the L_A of the optimal model is low, then minimizing L_B results in probabilities close to actual outcomes. If, on the other hand, L_A of the optimal model is high due to a high degree of randomness in the data, then minimizing L_B results in probabilities that do not strongly discriminate between outcomes. Note that when the outcome is fully random, L_B is minimized by predicting a probability equal to w. This is easily verified. This shows a straightforward connection between the two loss functions. Drawing class labels randomly with the probability that minimized $L_B(\mathrm{w})$ on a random data set results in a ratio between the false positive rate and false negative rate that is equal to their relative weight in $L_A(\mathrm{w})$.

zero (one) when the input probabilities are below (above) the cutoff value. In this paper we work with a finite number of values $\alpha \in [0.2, 2]$.[10]

The optimization and validation is thus summarized as follows. We generate predicted probabilities $\bar{P}$ from a statistical model, which may depend on classical tuning parameters. We then generate transformed probabilities $\hat{P}$ using Equation 4 and choose values for α and β that minimize L_B. The optimized probability predictions are used to generate class labels using Equation 3 which are finally used to evaluate L_A. Note that, if $\bar{P}$ is already appropriately balanced and of the preferred level of confidence, then hyper-tuning Equation 4 will lead to minimal re-scaling as the untransformed probabilities are nested by the function. Hence, the strategy can effectively be combined with other techniques for unbalanced learning tasks (including up-sampling).

3.2 Cross-validation setup

We perform 5 repetitions of 10-fold cross-validation.[11] The full sample Y^* consists of 37,198 country/district/month observations containing an IPC rating. The sub-vector of validation cases Y contains 30,948 cases, eliminating all cases in which a food crisis observation is preceded by another food crisis observation. We train the model using the full sample, since a prediction model that uses a larger and more diverse sample of crisis cases achieves lower L_B even when it is evaluated only on the transition sample Y. The folds are mutually-exclusive and exhaustive equally-sized partitions of the full allowable validation sample Y, defined as $Y_{test,1}, Y_{test,2}, ...Y_{test,k=10}$. These partitions are generated by stratified random sampling to preserve the overall class distribution of Y^* within each fold.[12] The 10 training sets are generated by removing the validation cases from the full sample according to $Y_{train,k} = Y^* \setminus Y_{test,k}$.

Given the relatively small and unbalanced data set, we create additional synthetic cases followed by standard up-sampling. In particular, while predictors are defined for each month, $Y_{train,k}$ is only defined for the rows in the data for which IPC assessments are available. We create additional training cases by assigning each binary IPC outcome to the previous and following months. The synthetic training examples are added within folds, only to training cases and not validation samples. Standard up-sampling is then applied to the result.[13]

[10] The paper considers only values of α up to 2 although higher values are possible and would result in even stronger transformation of predictions. The reason to investigate only up to a value of 2, implicitly confining our search only to reasonably smooth transformations, is to avoid over-fitting the relatively scarce number of validation samples.

[11] Standard cross-validation can be applied in the context of spatially and temporally dependent observations if the model is flexible enough to over-fit the data, see Bergmeir et al. (2018). If the model over-fits the data strongly, it will score badly in the cross-validation. If the model under-fits the data, and residuals are still correlated, then the independence assumption is invalidated. If the level of mis-specification is mild, and residual dependence is sufficiently weak, then corrections can still be made to the LLN (Pötscher and Prucha, 1997; Andrée, 2020b). Note that Random Forest is a method that eagerly fits correlations in a data set of modest dimensions.

[12] Stratified sampling ensures folds always contain observations in both classes. Note that Equations 1 and 2 average two components, hence consistent estimation requires consistent estimation of two lower level components. Thus, the validation strategy requires a sufficient number of validation cases in both classes, but the exact class distribution in the validation sample does not matter for our loss criteria as w is pre-defined.

[13] This is important, adding synthetic training cases based on a 1 month window around the test cases would invalidate the weak dependence assumption, needed for consistent validation, that is imposed at the 3- to 4-month interval. The approximate size of each resulting training set is $\sim 105,428$ samples.

3.3 Application using Random Forest

The prediction and validation framework is applicable to any binary classification algorithm that generates predicted probabilities for the positive class. In this paper we use a Random Forest classifier.[14] We focus on this particular approach in light of the additional results provided in supplementary Appendix B which show that in this same data set, the Random Forest offers a substantial improvement in predictive performance over simple linear models, but that little further is gained by moving to more complex learners such as neural networks.

Implementing the algorithm requires specifying several tuning parameters that control the structure of the forest or its trees. We select these to optimize predictive performance using the cross-validation procedures described above. This is implemented by searching over a practical grid of values.

First, since we require predicted probabilities, we implement a probability forest in which trees are grown as regression trees following Malley et al. (2012). An important tuning parameter is minimum node size that regulates tree depth by setting the minimum number of cases at the terminal nodes of the decision trees. We use the original value of 1, as well as the default value of 10 in the implementation of Wright and Ziegler (2017).[15]

The out-of-sample performance is heavily dependent on the predictive power of individual trees and the correlation between them.[16] The relationship between the two is mainly one of trade-offs and can to a certain degree be regulated by optimizing over the number of variables that are used to grow individual decision trees.[17] This number is controlled by *mtry*, and its default value is $\sqrt(D)$ where D is the number of predictors in the data set. Performing a grid search for *mtry* is computationally challenging and in this paper is made feasible by focusing on particular combinations of tuning-parameters.[18] In particular, the grid search is performed only when the minimum node size is 10, and makes use of a

[14] The algorithm was introduced by Breiman (2001) as a generalization of the tree bagging by combining it with the random subspace method of Ho (1998). It has quickly become a popular non-parametric classification tool due to its ability to generate good results using standard tuning parameter values and its robustness to noise. It constructs prediction rules without imposing strong prior assumptions on the functional form of the relationship between predictors and the target variable. Good review articles include those of Biau (2012); Biau and Scornet (2016). Random Forest generates results that are accurate while relatively straightforward to interpret (Banerjee et al., 2012; Petkovic et al., 2018).

[15] Breiman's original algorithm grows trees to purity, that is, until each terminal node contains only observations from one class. When nodes are pure, the probability estimate at each terminal node is thus always either 0 or 1. This means that, apart from a low correlation between trees, a large number of them is needed to obtain precise estimates of probabilities by the forest. Increasing minimum node size allows the frequency of class occurrence at the terminal nodes itself to be a reasonable probability estimate.

[16] The generalization error converges *almost surely* to a limit as the number of trees in the forest grows. We use 500 trees, corresponding to the default value in the implementation by Wright and Ziegler (2017). The magnitude of expected out-of-sample error in the limit in turn depends on the predictive strength of the individual trees and the correlation between them. See also (Scornet, 2017).

[17] The predictive strength of individual trees can to a degree be improved by allowing more variables to enter individual base learners. However, while larger values for *mtry* may result in stronger base learners, it also leads to a higher correlation between them. In particular, as $mtry \to D$, the probability that two base trees contain the same variables approaches 1. If many individual trees contain a highly similar set of variables, then their predictions become more correlated. Smaller values for *mtry* may instead increase diversity of base learners, at the cost of lower individual predictive power.

[18] Larger values of *mtry* significantly increase computational cost as it becomes more complex to process the individual trees. A grid search for *mtry* is more manageable in combination with tuning values that lead to simplifications elsewhere in the calculations; higher minimum node size that leads to faster individual tree-building, and randomized splitting does not require calculating Gini coefficients.

faster splitting rule introduced by Geurts et al. (2006).[19] Table 2 lists the full combination of tuning parameters. Overall, tuning the parameters of the probability transformation in Equation 4 (α, β), had by far the biggest impact on performance with respect to our balanced evaluation criterion.

Table 2: Random Forest tuning parameter combinations.

splitrule	**mtry**	**min.node.size**	α **(probability scaling)**	β **(probability constant)**
Gini	$\sqrt{D}$	1, 10	(1:10)/5	(1:100)/100
ExtraTrees	$\sqrt{D}$	1, 10	(1:10)/5	(1:100)/100
ExtraTrees	$(0.5, 0.75, 1.25, 1.5) \times \sqrt{D}$	10	(1:10)/5	(1:100)/100

The column names correspond to the tuning-parameters considered in the tuning grid. Row entries list combinations of values that result in unique model specifications. Thus (bottom row), when **splitrule** equals ExtraTrees and **min.node.size** equals 10, a range of different values for **mtry** is considered. Moreover, for each possible combination of **splitrule**, **min.node.size** and **mtry**, 10×100 unique combinations of (α, β) are tried.

4 Results

In this section we report on the prediction performance of our main model. Section 4.1 investigates the prediction of food crisis outbreaks sub-nationally, using the prediction and validation framework of the previous section. Section 4.2 investigates forecast results aggregated to a country level, again relying on cross-validation procedures. We also validated the models in a temporal holdout exercise. Overall, we find that this additional validation exercise corroborates the main findings presented below, the additional results are contained in supplementary Appendix B, section B.2. Finally, section 4.3 describes procedures to interpret the relative importance of different factors for model predictions.

4.1 Validation of district-level predictions

To set a performance baseline, we begin by summarizing the predictive performance of a simple binary crisis indicator derived from the existing future outlooks produced by FEWS NET. These follow the identical definition as our target variable, only now applied to the future outlooks.[20] These expert-based outlooks are closely watched by policy makers, and so their near-term and medium-term binary equivalents provide a natural benchmark with

[19] The original approach splits trees based on a measure of node purity. This increases the probability that the same variables ends up at the majority of root nodes when *mtry* increases, leading to high similarity between trees. Countering the correlation through randomization can produce stronger and more diverse trees that improve forest accuracy. This paper implements randomization of splits using the *extraTrees* algorithm introduced by (Geurts et al., 2006). This selects cut-points at random, i.e., independently of the target variable. Key advantages are potentially increased forest accuracy due to lower tree correlation and faster runtime which makes tuning *mtry* more practical and more interesting.

[20] In addition to its categorical ratings of actual food insecurity conditions, FEWS NET also reports a "near-term outlook" and a "medium-term outlook". These outlooks represent preliminary assessments of the most likely food security ratings one and two reporting cycles ahead. The near-term outlook corresponds to either 3 or 4 months ahead while the medium-term outlook corresponds to either 6 or 8 months ahead, depending on the frequency of the reporting cycle which increased to 4 months in 2016. Recall that these outlooks may also follow livelihood zones, in which case we aggregated them to the district level. In an identical treatment to the outcomes, we also net the outlooks from the projected humanitarian assistance effects, and thus validate the binarized indicators purely on their ability to signal the need for action, disregard of whether that action was already anticipated. This is in line with the use case explored for model-based forecasts.

which to compare the model-based forecasts. Because FEWS NET data do not assign a probability to future outcomes, we only focus on the L_A loss function of Equation 1. Table 3 contains the results.

The most striking feature of Table 3 is the extreme imbalance between false alarms – an FPR of around 2% – and failure to predict – an FNR of over 74%. This clearly indicates that the historical performance baseline is characterized by a strong aversion to predicting transitions into food crises that subsequently do not occur. In the context of our evaluation framework, this can be interpreted as evidence of a very high implicit weight on false positives and a low weight on false negatives, i.e. a low value of w.[21] The low false alarm rate does not mean that those warnings that were raised, did also most likely predict a transition correctly. To the contrary, only 35% of the projected transitions into crisis were followed by actual transitions. Comparing the frequency of transitions into crisis in both the historical assessments and outlooks also highlights the conservative nature. Although transitions into crisis account for 5.3% of the historical assessment sample, only 3.2% of the binarized outlooks projected a transition, see Table B.1 in supplementary Appendix B.

Table 3: Historical performance baseline calculated from binarized FEWS NET outlooks.

	Error Types		$\boldsymbol{L_A}$		
	FNR	**FPR**	**w=1/3**	**w=1/2**	**w=2/3**
Near-term	74.3%	1.4%	25.7%	37.8%	50.0%
Medium-term	79.4%	2.2%	27.9%	40.8%	53.7%

FNR and **FPR** of binarized historical FEWS NET outlooks in the near-term (3-4 months, validated against outcomes one reporting cycle later) and medium-term (6-8 months, validated against outcomes two reporting cycles later) and three weighted averages ($\boldsymbol{L_A}$) with weights **w** and (1-**w**) for prediction errors in the positive (negative) class. Both the binary outcomes and outlooks are netted of humanitarian assistance effects. The validation was performed at the admin2 level, and only for observations that were preceded by non-crisis levels of food insecurity. This specifically describes prediction performance for new outbreaks of food crisis before they occur. The validation sample spans all available projection-outcome pairs from July 2009 - February 2019.

Table 4: Cross-validated performance of predictions from the wide Random Forest.

	Error Type						**Balanced Metrics**					
	FNR			**FPR**			$\boldsymbol{L_A}$			$\boldsymbol{L_B}$		
w	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**
h=4	19.7%	16.5%	16.0%	6.8%	8.2%	8.4%	11.1%	12.3%	13.5%	0.25	0.29	0.31
h=8	19.1%	16.6%	16.6%	7.3%	8.5%	8.5%	11.2%	12.5%	13.9%	0.27	0.30	0.33
h=12	20.9%	20.9%	9.9%	10.0%	10.0%	19.0%	13.6%	15.5%	13.0%	0.30	0.34	0.35

The first six columns report false negative rates and false positive rates for models optimized for the indicated value of **w**. The next six columns report weighted averages and weighted Log Loss values ($\boldsymbol{L_B}$) with weights **w** and (1-**w**) for prediction errors in the positive (negative) class. The statistics have been cross-validated for **h**=4-, 8- and 12-month ahead forecasts of crisis outbreaks. The validation was performed at the admin2 level, and only for observations that were preceded by non-crisis levels of food insecurity. This specifically describes prediction performance for new outbreaks of food crisis before they occur. The validation sample spans from July 2009 - February 2019.

[21] Recall that unbiased Log Loss sets w to the frequency of the positive class in the validation sample, 5.3% on average in random draws, hence an unbiased model that assumes both the false positive rate and false negative rate carry the same cost, would assume a value of around w$\sim$ 0.05. The binarized outlooks produce even fewer alarms, at a rate of 3.2%, meaning that implicitly w$<$ 0.05. At, w= 0.03 the rate of false alarms is weighted 30 times as much as the rate with which anticipating crises fails in a weighted cost function like L_A.

The statistical model-based predictions we develop have the potential to improve on this baseline in several ways. We explore this using Table 4, which provides cross-validation results for 4-, 8- and 12-month ahead forecasts. Note that the 4- and 8-month forecast horizons are roughly comparable to the near- and medium-term outlooks considered by FEWS NET.

First, when we set a conservative value of w= $\frac{1}{3}$ to impose a heavier penalty on false alarms, the FPR increases slightly relative to the performance baseline produced with binarized FEWS NET outlooks, to around 7% for the 4- and 8-month forecast horizon. However, this comes with a very large reduction in the FNR to below 20% which is substantially down from the baseline rates of 74% and 79%. Taken together, the slightly higher FPR and much lower FNR results in an improvement in L_A from the baseline value of around 26% using binarized FEWS NET outlooks, to around 11% for the model predictions.

Second, the model-based approach allow us to fine-tune predictions for different use cases corresponding to different weights in the prediction loss function. As we have discussed, the severe costs of food crises suggest that for some policy purposes, a greater penalty should be imposed on failure to predict crises, i.e. that a higher value of w would be warranted. The FEWS NET outlooks offer just one fixed set of forward-looking ratings, and increasing the weight on FNR in the prediction loss function in Table 3 sharply increases L_A, indicating a sharp decline in predictive performance with regard to the less conservative forecasting objectives. For example, for the medium-term outlook, L_A increases from 28% to 54% when w increases from $\frac{1}{3}$ to $\frac{2}{3}$. In contrast, the model predictions maintain comparable predictive performance across the range of weights, as evidenced by minimal changes in the weighted performance metrics.

Third, models can learn different predictive associations depending on the forecast horizon and tailor predictions to the lead-time that is required. The complexity associated with unlocking funds and investing in prevention implies that near-term forecasts have lower utility than medium-term forecasts when equally accurate and so it would be helpful to predict further into the future with similar accuracy. The FEWS NET outlooks offer projections that run roughly 8 months into the future which, when binarized, score an L_A of 28% for the conservative weight of w=$\frac{1}{3}$. We assessed the ability to predict 12 months ahead, and find that the model maintains an L_A below 14%, still roughly half of the medium-term performance baseline.

Fourth, models have the ability to predict the probability of an event which means they can rank districts according to risk. L_B measures how well the predicted probabilities align with the outcomes and rewards predictions that trend in the right direction. The minimum value of L_B depends on the degree of noise in the target variable but it is well-known that an uninformative model, that predicts only the correct average probability, attains a Log Loss of .69.[22] Hence, any value below .69 for any value of w implies that the model not only

[22] As a reference, it can be shown that in a system, where you would know all the possible information about its future state other than some unknown internal level of randomness, so that, given that all model parameters are at their correct values, an outcome can still be different than predicted. Say for example that in this system the best one can do is predict 0.1 when the outcome is 0, and .9 when the outcome is 1, and that for one out of ten predictions the outcome falls into the other category. Then, Log Loss attains a minimum of approximately .33.

learned to predict the preferred average risk level but that, on average, the probabilities also trend in the right direction. The L_B values thus indicate that the models predict good probabilities.

Note that standard tuning parameters do not explicitly re-distribute predictions toward a high-cost outcome. The ability to tune predictions for the differential in costs instead relies on the transformation of the predicted probabilities controlled by tuning parameters (α, β).[23] For example, the untransformed "Vanilla" model at $h = 4$, tuning only over default parameters – and not (α, β) – benefits from up-sampling but has no explicit flexibility to deal with the unbalanced prediction problem and reaches an $L_B\left(\mathrm{w} = (\frac{1}{3}, \frac{1}{2}, \frac{2}{3})\right)$ of $(.31, .41, .51)$. The associated values of L_A are $(15.3\%, 21.9\%, 28.5\%)$. Finally, it is also possible to tune (α, β) to obtain an extremely conservative model. For example, when minimizing FNR under the constraint that FPR may not increase over the 1.4% set by the performance baseline, the model produces an FNR of 49% and an L_A of 17.3% at w=$\frac{1}{3}$, down from the baseline value of 26%.

Table 5: Country-specific validation results of binarized FEWS NET outlooks and predictions from the wide Random Forest.

	Binarized Outlooks				**Random Forest**			
	Near-term		**Med.-term**		**h=4**		**h=8**	
w	**1/3**	**2/3**	**1/3**	**2/3**	**1/3**	**2/3**	**1/3**	**2/3**
Afghanistan	27%	53%	24%	46%	17%	28%	17%	27%
Chad	30%	59%	33%	63%	7%	9%	7%	8%
Ethiopia	30%	57%	29%	55%	20%	19%	19%	18%
Haiti	25%	50%	31%	59%	27%	49%	28%	49%
Kenya	25%	47%	26%	50%	16%	26%	20%	35%
Malawi	27%	52%	30%	58%	11%	14%	11%	13%
Mali	24%	47%	28%	55%	11%	16%	8%	14%
Niger	22%	41%	32%	60%	9%	14%	10%	16%
Nigeria	31%	62%	31%	60%	5%	4%	5%	4%
Somalia	25%	47%	33%	59%	30%	21%	33%	21%
South Sudan	24%	41%	28%	50%	17%	13%	17%	13%
Sudan	25%	50%	30%	60%	16%	22%	15%	21%
Zimbabwe	27%	53%	23%	44%	11%	11%	10%	10%
Balanced average	26%	51%	29%	55%	15%	19%	15%	19%
Burkina Faso	0%	0%	33%	67%	28%	45%	33%	67%
Congo, Dem. Rep.	34%	67%	33%	67%	19%	28%	17%	21%
Guatemala	30%	59%	30%	59%	23%	41%	23%	41%
Mauritania	35%	67%	35%	67%	22%	40%	23%	40%
Mozambique	29%	57%	24%	48%	21%	31%	26%	44%
Uganda	33%	67%	34%	67%	21%	37%	21%	36%
Yemen, Rep.	6%	9%	24%	16%	36%	20%	28%	15%

The first four columns report weighted averages of false negative rates and false positive rates (L_A) for binarized FEWS NET outlooks in the near-term (3-4 months, validated against outcomes one reporting cycle later) and medium-term (6-8 months, validated against outcomes two reporting cycles later), for the indicated value of **w**. Both the binary outcomes and outlooks are netted of humanitarian assistance effects. The next four columns report cross-validated statistics for the wide Random Forest optimized for the indicated value of **w**, for **h**=4- and 8-month ahead forecasts of crisis outbreaks. The validation was performed at the admin2 level, and only for observations that were preceded by non-crisis levels of food insecurity. This specifically describes prediction performance for new outbreaks of food crisis before they occur. Statistics for countries with approximately fewer than 50 historical transitions into crisis are in the bottom part. Zambia is not included in the validation sample due to an absence of positive class values. The validation sample spans from July 2009 - February 2019.

[23] For w=$\frac{1}{2}$ and w=$\frac{2}{3}$ we find $\alpha = 2$, our maximum considered value, and β close to 0, indicating that nearly the entire range of predicted probabilities are skewed upward. Recall that the natural frequency of the data favors a w value around 0.05 and so our predictions need a strong transformation to do well on a loss function that penalizes FNR heavily. Future work with more validation samples could consider higher values for α. For now, we simply note that the results are already extremely competitive for high values of w.

It is also interesting to analyze how the error rates vary by country. Table 5 provides a country-level comparison between the previous baseline results and cross-validated model predictions. Recall from Table 1 that the number of historical transitions is low in some countries. This means that L_A and L_B are easily influenced by a few errors. Countries with approximately 50 or more outbreaks are indicated in the top half of the table. The L_A values in these countries vary from 23% in Zimbabwe to 33% in Chad and Somalia at w$= \frac{1}{3}$. The conservative bias in the performance baseline is a feature that persists at the individual country level. When w$= \frac{2}{3}$, the balanced error rates in multiple countries reach over 50%, in some countries reaching around 60%. These validation results could in expectation be beaten by randomly guessing at the desired frequency.[24] Overall, model-based predictions improve a simple cross-country average of balanced error rates (w$=\frac{1}{3}$) from 26% in the near-term and 29% in the medium-term, to respectively 15% and 19%.

4.2 Validation of national-level predictions

Our modeling approach thus far has focused on generating predictions of food crisis at the sub-national level, focusing particularly on calibrating the predictions to minimize the prediction loss function with varying weights of prediction errors in the positive class. Since finance is often unlocked at a country level, we also explore the properties of the predictions when they are aggregated to form country-level predictions. Specifically, for each country in our sample, we construct the share of the country's population living in districts that are in food crisis, as measured by an IPC rating of 3, 4 or 5. We then investigate how well the country-level aggregate of our district-level predictions forecasts this aggregate country-level food crisis indicator. This exercise is of interest because scarce global humanitarian resources typically are prioritized between regions of the world and countries first, while the allocations to communities and beneficiaries at sub-national level occurs with a separate targeting process. For example, the World Bank's base allocation is predetermined at country-level. Having a good forecast of country-level food crisis affected populations can therefore help move resources to the appropriate countries, potentially early enough to finance preventive measures.[25]

We construct the country-level aggregate food crisis-affected population prediction as a weighted average of the district-level probabilities in the country, with weights equal to populations living in each district, i.e.

$$\frac{\sum_{i=1}^{N}\left(\hat{P}_{ijt}\cdot \text{pop}_{ijt}\right)}{\sum_{i=1}^{N}\text{pop}_{ijt}}, \tag{5}$$

where i, j, and t index districts, countries and time periods and pop_{ijt} represents population. Since the scale of the aggregated country-level predictions depends on the bias introduced

[24] Recall that the uninformative value of L_A(w) equals 50% when labels are generated randomly with frequency w. L_A(w) will only go above 50% if w (preference for error types) moves in the opposite direction of the model's bias. For example, it will reach 100% if w=1 while the frequency of positive predictions is 0 (and all the class labels are thus in the negative class, e.g. FPR equals 0 and FNR equals 100%, the latter being fully weighted in L_A).

[25] For an alternative approach to modeling country-level food insecurity over long horizons, up to three years, see Wang et al. (2020) who utilize panel vector auto-regression techniques.

by the chosen value of w, the scale of the prediction result may differ from the scale of the actual outcomes. To adjust for this simple bias, we regress the country-level result on the available country-level historical outcomes and use the fitted values as final forecasts.[26] We validate the forecasts by calculating Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), reported in Table 6.[27]

Table 6: Country-specific validation results of crisis-affected population forecasts from binarized FEWS NET outlooks and predictions from the wide Random Forest.

	Binarized Outlooks				**Random Forest**			
	Near-term		**Medium-term**		**h=4**		**h=8**	
Country	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Afghanistan	7.9	3.9	15.5	9.8	5.4	3.2	5.9	3.9
Burkina Faso	1.9	0.4	2.7	0.8	0.8	0.3	1.1	0.5
Chad	5.2	2.2	13.2	9.0	6.0	4.5	6.4	4.8
Congo, Dem. Rep.	5.6	2.5	3.2	1.5	16.9	12.5	18.9	14.6
Ethiopia	3.8	2.4	16.5	12.8	8.8	7.4	9.5	7.9
Guatemala	1.1	0.4	6.3	3.5	3.9	3.0	4.0	3.3
Haiti	3.4	2.1	6.6	4.7	3.8	3.3	4.8	4.4
Kenya	5.0	2.7	4.2	2.5	6.2	5.2	7.2	6.6
Malawi	11.6	4.9	31.5	20.7	10.0	7.4	11.1	9.4
Mali	3.4	1.6	5.4	2.8	1.9	1.1	2.1	1.2
Mauritania	9.0	3.7	15.5	7.2	10.5	7.7	11.9	9.7
Mozambique	2.1	0.4	8.0	2.8	4.3	3.5	5.3	4.3
Niger	10.2	4.1	15.0	6.5	3.8	2.7	4.3	3.4
Nigeria	0.2	0.0	1.9	1.1	1.1	1.0	1.0	0.9
Somalia	18.7	8.6	26.5	16.2	21.5	18.2	22.2	18.8
South Sudan	14.4	8.6	17.4	14.5	11.8	10.7	11.9	10.8
Sudan	4.8	2.1	9.3	6.5	4.1	3.5	4.3	3.7
Uganda	0.0	0.0	1.0	0.5	0.9	0.7	1.2	1.0
Yemen, Rep.	12.4	4.8	16.1	7.1	14.5	8.5	14.3	8.4
Zambia	0.5	0.1	0.5	0.1	0.0	0.0	0.0	0.0
Zimbabwe	6.4	3.1	20.6	11.3	7.9	6.2	9.3	7.9
Balanced average	6.4	2.9	11.8	7.1	7.2	5.5	7.8	6.3
Balanced average*	5.6	2.7	11.9	7.3	5.6	4.4	6.2	5.2

The first four columns report Root Mean Squared Error and Mean Absolute Error for forecasts of country-level populations in crisis-affected districts, constructed from near-term and medium-term binarized FEWS NET outlooks. Both the binary outcomes and outlooks are netted of humanitarian assistance effects. The second set of four columns report cross-validated statistics of for the wide Random Forest, for 4- and 8-month-ahead forecast horizons. *Zambia, Uganda, the Republic of Yemen, Somalia and the Democratic Republic of Congo are excluded for reasons described in the text. The validation sample spans from July 2009 - February 2019.

The results suggest that aggregated district-level predictions are reasonable predictors of country-level outcomes. At both 4- and 8-month forecast horizons, the model produces error rates that are broadly comparable to those obtained with binarized near-term outlooks.[28]

[26] A simple equation that involves only one scalar, e.g. of the form $Y = \beta X + \varepsilon$. The district-level predictions used here, $\hat{P}_{ijt}$, are generated setting w=$\frac{1}{3}$ in the prediction loss function and are optimized as before to minimize loss at the district level.

[27] Specifically, we generate holdout predictions for all observations in the data set under optimized hyper-parameters, and using 5 repetitions of 10-fold cross validation, as described in the previous section. It is important to note that the model hyper-parameters have been optimized to forecast crisis outbreaks, not total crisis-populations. Future efforts could test possible forecast gains when designing a prediction model specifically for the latter purpose.

[28] The balanced average of the model is heavily impacted by Somalia and the Democratic Republic of Congo. Inspection of the predictions in Somalia revealed that a considerable share of the errors was due to over-predicting in the period between 2014-2017 when outcomes briefly hit historical lows of 0%, and under-predicting when the outcomes hit 100%. Generally, we found that the modeled results were actually good leading indicators. For example, in a holdout (June and October 2019), the binarized near- and medium-term outlooks from FEWS NET had an absolute error of 13.4 and 47.0 points respectively, while the model errors were 1.2 and 2.0 points on the 4 and 8 month horizons. In Democratic Republic of Congo,

Importantly, the error of model increases only slightly as we move to the 8-month forecast horizon, while errors made with binarized outlooks nearly double when moving from the near-term to the medium-term horizon. This means that the model-based predictions are better suited for forecasting further into the future, which may help increase lead time for preventive action.

4.3 Model interpretability

The previous sections showed that the Random Forest model provides good predictive capabilities. While predictive performance is important, the modeling structure of the transformed classifier may be difficult to interpret. Out-of-the-box diagnostics may be used to help determine which variables are important in the forest, but do not inform about the signs and magnitude of (local) relationships. Parametric models are more transparent in that regard, but while it can be argued that these models are more interpretable, they might not be comprehensible given the number of predictors involved. Moreover, standard methods do not account for effects induced by the transformation of probabilities. To overcome these issues, this section develops a prediction decomposition strategy that relies on assessing the change in prediction values when the model is confronted with observed values compared to reference input data. This provides comprehensible interpretation at an observational level and takes the following algorithmic strategy.

1. Partition the spatio-temporally covariates into a fully exhaustive set of non-overlapping groups. We use the groups defined in section 2.2 but split rainfall variables from the other environmental factors.[29]
2. Construct a country data set by leaving all time-invariant district-level data at local values and setting all spatio-temporally varying covariates at a reference value. We use country means.[30] Make reference predictions with this data set.
3. For each group of covariates defined in step 1, set the levels of the data to the observed values while keeping all the other covariates at reference values as defined in the previous step, then generate a new set of predictions.
4. For each set of predictions generated in the previous step, subtract the reference predictions generated in step 2. Then, from this prediction difference, calculate the lowest value in each district across all time periods and subtract these.
5. Standardize the prediction differences generated in the previous step to [0,1] by dividing them by their sum, then multiply by probabilities predicted when all data are set to observed levels.

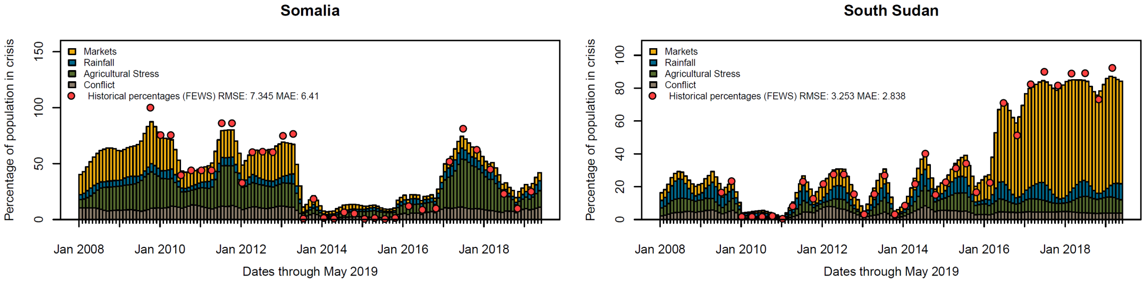
only 8 assessment cycles are available for cross-validation and errors are heavily driven by the first two periods. Again, good performance was established in a holdout, binarized outlooks scoring respectively 9.3 and 16.3 points while the model-based predictions missed the outcome by only 1.6 and 3.2 points. Finally, few assessments are available for the Republic of Yemen which has stayed near 100% crisis in all but the first period, no difference was ever recorded between near-term preliminary assessments and final outcomes in Uganda, and Zambia never experienced crisis. When focusing on the remaining 16 countries for validation, RMSE in the near-term was identical for model-based predictions and binarized outlooks, while on the 8-month target the model reduced RMSE by 48% increasing only 0.6 points compared to the binarized near-term outlook.

[29] "Markets" includes all features calculated from food prices and "Conflict" all features derived from the conflict data. The Environmental Factors are divided in two groups: "Rainfall" and "Agricultural Stress", the latter combining all the features derived from NDVI and evapo-transpiration.

[30] It would also be possible to construct specific alternatives with which to compare observed values such as conflict counts of 0 rather than a non-zero average.

The approach seeks to calculate the increase in crisis probability from 0, attributed to observing current values of variables within a related group when other variables are kept at means, then standardizes the relative prediction contributions and rescales the result by the final prediction results. Naturally, the results can be scaled to align with the percentages of population in crisis districts. The exercise is performed by country because the country indicators are defined uniquely at this level of observation. Figure 1 shows two example countries, while results for the other countries are covered by Figure B.1 in the supplementary Appendix.

Figure 1: Prediction decompositions of population share in crisis districts in two key food crisis countries.



Prediction decompositions, aggregated to the country-level by taking a population weighted average of the district-level results. The normalized prediction decompositions are scaled by the predicted share of crisis-effected populations. A 3-month centered moving average has been applied to smoothen the results. Historical outcomes and measures of fit are added for reference. The height of each bar indicates the predicted share of the country population living in crisis-affected districts, the colors indicate the share of the modeled value that is attributed to variables that cover four key groups of food insecurity drivers.

The results in Figure 1 highlight that this strategy is able to clearly illustrate the groups of variables that are used to predict certain outcomes at the specified forecasting interval. For example, all features derived from NDVI and evapo-transpiration together drive the largest share of predictions in Somalia. Over time, a change in this pattern is visible as well. Whereas predictions in the first few years that cover the complexities of the famine period are driven by a combination of environmental variables and food price inflation, the 2017-2018 drought-induced crises is modeled almost exclusively by drought indicators. Stark differences between the results in Somalia and South Sudan are also highlighted. While both countries have experienced extremely severe situations, the crisis predictions in South Sudan are mainly traced back to price inflation. Note that while these narratives fit the countries, the results are approximations, and not exact prediction contributions. They should also not be interpreted in a causal sense.

5 Discussion and Conclusion

Food crises impose heavy human costs that are likely to increase in light of worsening climatic drivers and growing populations. Outbreaks of food crises are relatively rare events, accounting for just 4.5% of our historical data. But when they occur, they are prolonged events, leading to widespread suffering, loss of human capital, household asset depletion, and death. In this context, the capacity to predict new outbreaks of food crisis events well in advance opens important opportunities to mitigate and avoid the worst. In this paper, we have explored the benefits of using formal statistical models to forecast food crises, using sub-national data provided by FEWS NET for 21 countries since 2009.

As a baseline, it is instructive to compare model-based forecasts with existing early warning systems. In our validation framework, these historical outlooks turn out as quite conservative, with extremely low false positive rates and relatively few warnings when compared with the actual incidence of food crisis outbreaks. As a consequence, existing outlooks have high false negative rates, i.e. many food crisis outbreaks were not anticipated. In addition, the overall forecast performance of existing outlooks deteriorates with longer forecast horizons.

Using standard cross- and temporal holdout validation methods, we have shown that relatively simple statistical models provide good predictive performance up to 12 months in advance. We used verifiable monthly data capturing environmental factors, conflict, and food price shocks as key predictors for this. We find that the model-based forecasts can deliver much lower false negative rates at the cost of only modest increases in false positive rates. Moreover, our metrics of forecast performance are fairly stable as the forecast horizon increases from 4 to 12 months, suggesting good predictive performance over longer forecast horizons.

A key benefit of statistical forecasts is that they can be optimized relative to an explicit trade-off between false positives and false negatives. This trade-off may be different for different policy purposes. When the costs of full-blown food crises are large and preventive measures have high returns, a forecasting framework with a high tolerance for false positives might be appropriate. Conversely, when resources for prevention are scarce, a greater tolerance for false negatives is more suitable. Our validation framework made this trade-off explicit, by tuning the parameters of our models to minimize prediction loss functions that balance false positives and false negatives. Depending on the use case, calibrated predictions can be generated that strike the appropriate balance between false positives and false negatives dictated by the application at hand. We acknowledge, however, that putting this principle into practice is a difficult task. While specifying a loss function that trades off false positives and false negatives is mathematically straightforward, setting the appropriate weights requires careful deliberation of the human and economic implications of competing errors, both of which are subject to considerable uncertainty.

Combined, the results show that statistical models can provide new and attractive forecasting capabilities in various ways, including an improved ability to detect new crisis outbreaks for a specified tolerance for raising (false) alarms; calibrating error types to take the costs and benefits of specific interventions into account; increasing lead time for preventive action by forecasting further into the future, and updating predictions continuously using readily-available, verifiable and open high-frequency data. Realizing these benefits will require close coordination between statistical modeling work and policy dialog, to ensure that the design of warning systems is tailored to the policies they are intended to inform.

References

Adam, C., Collier, P., and Davies, V. A. B. (2008). Postconflict Monetary Reconstruction. *The World Bank Economic Review*, 22:87–112.

Andrée, B. P. J. (2020a). Estimating Food Price Inflation from Partial Surveys. *In Progress*.

Andrée, B. P. J. (2020b). *Theory and Application of Dynamic Spatial Time Series Models*. Rozenberg Publishers and the Tinbergen Institute, Amsterdam.

Andrée, B. P. J., Chamorro, A., Spencer, P., Koomen, E., and Dogo, H. (2019). Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission. *Renewable and Sustainable Energy Reviews*, 114:109221.

Andrée, B. P. J., Diogo, V., and Koomen, E. (2017). Efficiency of second-generation biofuel crop subsidy schemes: Spatial heterogeneity and policy design. *Renewable and Sustainable Energy Reviews*, 67:848–862.

Andrée, B. P. J. and Kraay, A. (2020). Balancing predicted probabilities to minimize weighted prediction loss functions. *In Progress*.

Asfaw, A. (2016). The Inter-Generational Health Effect of Early Malnutrition: Evidence from the 1983-85 Ethiopian Famine. *SSRN Electronic Journal*.

Baffes, J., Mitchell, D., Riordan, E. M., Streifel, S., Timmer, H., and Shaw, W. (2008). *Global Economic Prospects: Commodities at the Crossroads 2009*. World Bank Publications, Washington, DC.

Banerjee, M., Ding, Y., and Noone, A. M. (2012). Identifying representative trees from ensembles. *Statistics in Medicine*, 31(15):1601–1616.

Barbier, E. B. and Hochard, J. P. (2018). Land degradation and poverty. *Nature Sustainability*, 1(11):623–631.

Barrett, C. B. and Bevis, L. E. M. (2015). The self-reinforcing feedback between low soil fertility and chronic poverty. *Nature Geoscience*, 8(12):907–912.

Bergmeir, C. and Benítez, J. M. (2012). Neural networks in R using the Stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13 (2012):1063–1095.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., de Onis, M., Ezzati, M., Grantham-McGregor, S., Katz, J., Martorell, R., Uauy, R., and Maternal and Child Nutrition Study Group (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The Lancet*, 382(9890):427–451.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, M. E. (2010). *Famine early warning systems and remote sensing data*. Springer-Verlag Berlin Heidelberg.

Brück, T. and D'Errico, M. (2019). Food security and violent conflict: Introduction to the special issue. *World Development*, 117:167–171.

Celiku, B. and Kraay, A. (2017). Predicting Conflict. *World Bank Policy Research Working Papers*.

Conceição, P. and Mendoza, R. U. (2009). Anatomy of the global food crisis. *Third World Quarterly*.

Cutler, P. (1984). Famine forecasting; Prices and peasant behaviour in Northern Ethiopia. *Disasters*, 8(1):48–56.

Devereux, S. (2000). Famine in the twentieth century. *IDS Working Paper 105.*

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Diogo, V., Reidsma, P., Schaap, B., Andrée, B. P. J., and Koomen, E. (2017). Assessing local and regional economic impacts of climatic extremes and feasibility of adaptation measures in Dutch arable farming systems. *Agricultural Systems*, 157:216–229.

D'Souza, A. and Jolliffe, D. (2013). Conflict, food price shocks, and food insecurity: The experience of Afghan households. *Food Policy*, 42:32–47.

Duraiappah, A. K. (1998). Poverty and environmental degradation: A review and analysis of the nexus. *World Development*, 26(12):2169–2179.

FAO, IFAD, UNICEF, WFP, and WHO (2019). *The State of Food Security and Nutrition in the World 2019. Safeguarding against economic slowdowns and downturns.* FAO, Rome.

Food Security Information Network (2020). Global Report on Food Crises. Technical report, FAO, Rome.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Galler, J. and Rabinowitz, D. G. (2014). The intergenerational effects of early adversity. *Progress in molecular biology and translational science*, 128:177–98.

Galler, J. R. and Barrett, L. R. (2001). Children and famine: long-term impact on development. *Ambulatory Child Health*, 7(2):85–95.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

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

Grainger, A. (1990). *The threatening desert: Controlling desertification.* John Wiley & Sons, Ltd, London.

Headey, D. (2011). Rethinking the global food crisis: The role of trade shocks. *Food Policy*, 36(2):136–146.

Hillbruner, C. and Moloney, G. (2012). When early warning is not enough—Lessons learned from the 2011 Somalia Famine. *Global Food Security*, 1(1):20–28.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844.

Ingram, J., Ericksen, P. J., and Liverman, D. (2010). *Food security and global environmental change.* Earthscan, London.

Khan, M. (1994). Market-based early warning indicators of famine for the pastoral households of the Sahel. *World Development*, 22(2):189–199.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–495.

Koomen, E., Diogo, V., Dekkers, J., and Rietveld, P. (2015). A utility-based suitability framework for integrated local-scale land-use modelling. *Computers, Environment and Urban Systems*, 50:1–14.

Li, J., Liu, S., Li, S., Feng, R., Na, L., Chu, X., Wu, X., Niu, Y., Sun, Z., Han, T., Deng, H., Meng, X., Xu, H., Zhang, Z., Qu, Q., Zhang, Q., Li, Y., and Sun, C. (2017). Prenatal exposure to famine and the development of hyperglycemia and type 2 diabetes in adulthood across consecutive generations: a population-based cohort study of families in Suihua, China. *The American Journal of Clinical Nutrition*, 105(1):221–227.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A. (2012). Probability Machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51(1):74–81.

Maxwell, D. and Fitzpatrick, M. (2012). The 2011 Somalia famine: Context, causes, and complications. *Global Food Security*, 1(1):5–12.

Maxwell, D., Khalif, A., Hailey, P., and Checchi, F. (2020). Determining famine: Multidimensional analysis for the twenty-first century. *Food Policy*, 92:101832.

Mechler, R. (2016). Reviewing estimates of the economic efficiency of disaster risk management: opportunities and limitations of using risk-based cost–benefit analysis. *Natural Hazards*, 81(3):2121–2147.

Meerkatt, H., Kolo, P., and Renson, Q. (2015). UNICEF/WFP Return on Investment for Emergency Preparedness study. Technical report, The Boston Consulting Group.

Mellor, J. W. (1986). Prediction and prevention of famine. *Federation Proceedings*, 45(10):2427–2431.

Misselhorn, A. A. (2005). What drives food insecurity in southern Africa? a meta-analysis of household economy studies. *Global Environmental Change*, 15(1):33–43.

Mwebaze, E., Okori, W., and Quinn, J. A. (2010). Causal structure learning for famine prediction. In *AAAI Spring Symposium - Technical Report*.

Myers, S. S., Smith, M. R., Guth, S., Golden, C. D., Vaitla, B., Mueller, N. D., Dangour, A. D., and Huybers, P. (2017). Climate Change and Global Food Systems: Potential Impacts on Food Security and Undernutrition. *Annual Review of Public Health*, 38(1):259–277.

Nunn, N. and Puga, D. (2012). Ruggedness: The Blessing of Bad Geography in Africa. *Review of Economics and Statistics*, 94(1).

Okori, W. and Obua, J. (2011). Machine learning classification technique for famine prediction. In *Proceedings of the World Congress on Engineering 2011, WCE 2011*.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Petkovic, D., Altman, R., Wong, M., and Vigil, A. (2018). Improving the explainability of Random Forest classifier – User centered approach. In *Pacific Symposium on Biocomputing*, number 23, pages 204–215. World Scientific Publishing Co. Pte Ltd.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ripley, B. D. (1993). Statistical aspects of neural networks. In *Networks and Chaos — Statistical and Probabilistic Aspects*, pages 40–123. Springer US.

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:409–456.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Ross, K. W., Brown, M. E., Verdin, J. P., and Underwood, L. W. (2009). Review of FEWS NET biophysical monitoring requirements. *Environmental Research Letters*, 4(2):024009.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162.

Seaman, J. (2000). Making Exchange Entitlements Operational: The Food Economy Approach to Famine Prediction and the RiskMap Computer Program. *Disasters*, 24(2):133–152.

Seaman, J. and Holt, J. (1980). Markets and Famines in the Third World. *Disasters*, 4(3):283–297.

SEDAC and CIESIN (2015). Gridded Population of the World Version 4. *Center for International Earth Science Information Network*.

Shawn J., R., Stephen D., D., and Elliot, R. (1999). Index that quantifies topographic heterogeneity. *International Journal of sciences*, 5(1-4):23–27.

Singh, R. B. (2012). Climate Change and Food Security. In Tuteja, N., Gill, S. S., and Tuteja, R., editors, *Improving Crop Productivity in Sustainable Agriculture*, chapter 1, pages 1–22. Wiley-VCH Verlag GmbH & Co. KGaA.

Stern, D. I., Common, M. S., and Barbier, E. B. (1996). Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Development*, 24(7):1151–1160.

Teluguntla, P. G., Thenkabail, P. S., Xiong, J. N., Gumma, M. K., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T. T., Massey, R., Phalke, A., and Yadav, K. (2015). Global Cropland Area Database (GCAD) derived from Remote Sensing in Support of Food Security in the Twenty-first Century: Current Achievements and Future Possibilities.

Tibshirani, R. (1996). Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.

van Velthuizen, H., Huddleston, B., Fischer, G., Salvatore, M., Ataman, E., Nachtergaele, F. O., Zanetti, M., Bloise, M., Antonicelli, A., Bel, J., De Liddo, A., De Salvo, P., and Franceschini, G. (2007). *Mapping biophysical factors that influence agricultural production and rural vulnerability.* Food and Agriculture Organization of the United Nations, Rome.

Veenendaal, M., Painter, R., de Rooij, S., Bossuyt, P., van der Post, J., Gluckman, P., Hanson, M., and Roseboom, T. (2013). Transgenerational effects of prenatal exposure to the 1944-45 Dutch famine. *BJOG: An International Journal of Obstetrics & Gynaecology*, 120(5):548–554.

Wang, D., Andrée, B. P. J., Chamorro, A. F., and Spencer, P. G. (2020). Stochastic modeling of food insecurity. *World Bank Policy Research Working Papers.*

World Bank Group (2018). Poverty and Shared Prosperity 2018. Technical report, World Bank, Washington, DC.

Wright, M. N. and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Supplementary Appendix

A Data

Table A.1: Covariates of food crisis.

Covariates	***d***	**Time-varying**	**Description**	**Source**
Dummies	23	No	Quarterly and by country, one left out.	Generated.
Spatial trends	2	No	Coordinates of center points.	Admin2 boundaries (FEWS NET).[a]
District size	1	No	Area in square kilometers calculated using Mollweide projection.	Admin2 boundaries.
Population	1	No	Areal sum of grids within districts.	GPWv4, adjusted (SEDAC and CIESIN, 2015).
Terrain Ruggedness	1	No	Areal means of 100m grids.	Index introduced by Shawn J. et al. (1999), data from Nunn and Puga (2012).
Cropland and pastures area shares.	2	No	Areal means.	Cropland estimates (1km×1km, 2010) taken from the GCAD[b], produced with methods described by Teluguntla et al. (2015). Pasture estimates (5 arc-minute ∼ 10km×10km, 2000) taken from FAO[c], produced with methods described by van Velthuizen et al. (2007).
NDVI	2	Yes	Monthly means and anomalies (deviations from average seasonal values).	Calculated from 16-day 250m×250m MODIS / Aqua Vegetation Indices, v.6.[d]
Rainfall	2	Yes	Monthly means and anomalies (deviations from average seasonal values).	Rainfall estimates (0.05°) from rain gauge and satellite observations (CHIRPS).[e]
Evapo-Transpiration	1	Yes	Monthly means.	From MOD16A2, v.6 (8-day composite, 500m resolution).[f]
Conflict events	2	Yes	Monthly counts of violent conflict events within districts and the average number of fatalities per event. The district-level data are spatially interpolated using inverse distance weighting to capture cross-border exposure.	Daily conflict event data taken from ACLED, aggregated to the monthly interval by sum.[g]
Food prices	2	Yes	Monthly log nominal food price index and monthly year-on-year differences.	Methods following Andrée (2020a), raw market prices from WFP.[h]

Covariates used to predict food crisis: ***d*** marks the number of covariates excluding lags. Sources as follows:

[a] https://fews.net/fews-data/334

[b] https://lpdaac.usgs.gov/products/gfsad1kcmv001/

[c] "Occurrence of pasture and browse (FGGD)" obtained at http://www.fao.org/geonetwork

[d] https://doi.org/10.5067/MODIS/MYD13Q1.006

[e] https://chc.ucsb.edu/data/chirps

[f] https://lpdaac.usgs.gov/products/mod16a2v006/

[g] https://www.acleddata.com/

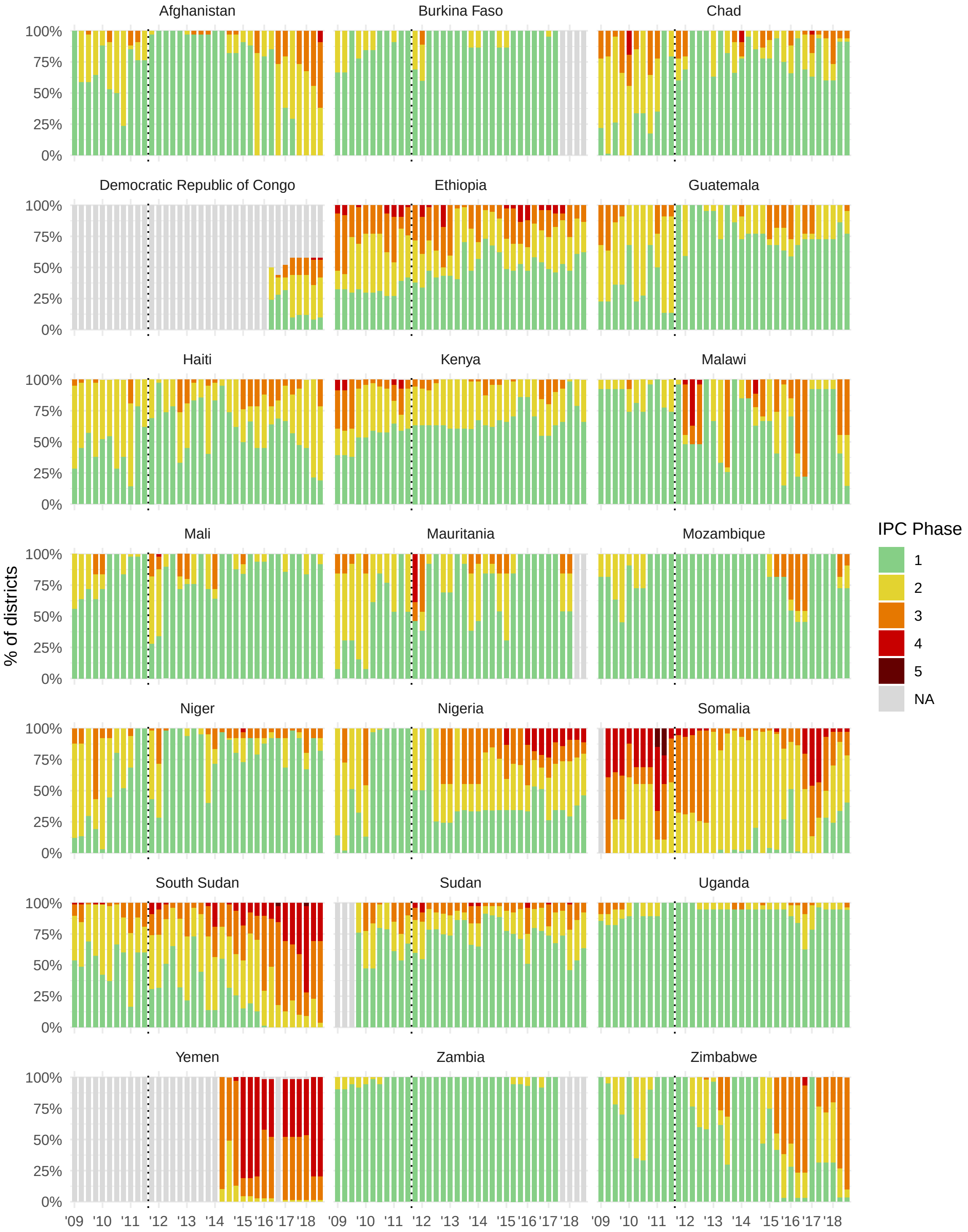
[h] https://dataviz.vam.wfp.org/economic_explorer/prices

Table A.2: Number of predictors of the un-processed versus processed data set.

	Basic	**Wide**
h=4	104	112
h=8	72	80
h=12	39	47

Number of predictors after calculating lags and dropping features until pair-wise correlations are below .95 (Basic); number of predictors after calculating additional features, spatial averages and temporal lags, and dropping features until pair-wise correlations are below .75 (Wide).

Figure A.1: Share of districts in different IPC phases netted of humanitarian assistance.



Share of districts in different IPC phases by reporting cycle based on historical FEWS NET assessments, netted of humanitarian assistance effects. Vertical lines mark the first period in which humanitarian assistance factors are available. The crisis indicator used for modeling corresponds to districts in phases (3), (4) and (5). The data spans from July 2009 - February 2019.

B Additional Results

B.1 Additional validation metrics

Several metrics other than FPR and FNR are commonly used to validate classification accuracy. These are commonly derived from the confusion matrix, provided in Table B.1, which in turn can be calculated from the FPR and FNR that the paper focuses on.

Table B.1: Confusion matrix for binarized FEWS NET outlooks and predictions from the wide Random Forest.

	Near-term / h=4			Medium-term / h=8		
	Binarized Outlooks	Random Forest		Binarized Outlooks	Random Forest	
		w=1/3	w=2/3		w=1/3	w=2/3
TP	443	1384	1447	354	1394	1437
FP	435	2104	2600	675	2259	2631
FN	1280	339	276	1369	329	286
TN	30513	28844	28348	30273	28689	28317
Accuracy	0.948	0.925	0.912	0.937	0.921	0.911
Kappa	0.316	0.496	0.462	0.227	0.481	0.456
Specificity	0.986	0.932	0.916	0.978	0.927	0.915
Pos Pred Value	0.505	0.397	0.358	0.344	0.382	0.353
Neg Pred Value	0.960	0.988	0.990	0.957	0.989	0.990
Precision	0.505	0.397	0.358	0.344	0.382	0.353
Recall	0.257	0.803	0.840	0.206	0.809	0.834
F1	0.341	0.531	0.502	0.257	0.519	0.496
Prevalence	0.053	0.053	0.053	0.053	0.053	0.053
Detection rate	0.014	0.042	0.044	0.011	0.043	0.044
Detection Prevalence	0.027	0.107	0.124	0.032	0.112	0.125
Balanced Accuracy	0.622	0.868	0.878	0.592	0.868	0.875

Historical validation results for binarized near- and medium-term FEWS NET outlooks and cross-validated predictions of crisis outbreaks from the wide Random Forest. Both the binary outcomes and outlooks are netted of humanitarian assistance effects, the validation only considers observations preceded by a phase rating below crisis. The validation sample spans from July 2009 - February 2019. Model validation is based on 5 repetitions of 10-fold cross-validation for models calibrated respectively for balanced Log Loss with weights **w**, indicated in the columns, and **w**-1 for prediction errors in the positive and negative classes.

B.2 Temporal holdout validation

The model validation results have so far relied on "pseudo-out-of-sample" methods that repeatedly partition the data randomly into training and test data sets. Here we also investigate the predictive performance of the model fully out-of-sample using data for June 2019, October 2019, and February 2020, that were reserved exclusively for validation purposes. We perform a final temporal holdout validation by estimating and optimizing our prediction model using data through February 2019, then generate 4-, 8-, and 12-month ahead forecasts to compare against the assessments published respectively 4, 8 and 12 months later.

Temporal holdout validation is a method that is heavily dependent on the composition of the holdout sample and the local behavior of the data generating process (Diebold, 2015; Andrée, 2020b). This means that single holdout results may deviate from average accuracy. For this reason it is important to first consider the frequency of food crisis events in the validation periods. Historically, only 16% of the data consisted of crises while the share of crises in the June validation sample is more than double, around 34%. Of all the cases in which a transition into food crisis could occur, it occurred in more than 11% from February to June, double the historical frequency. This indicates that the validation sample constitutes a period of severe crisis, and the predictive performance of the model in this holdout may be quite different from that in average periods.

Table B.2: Validation of February 2019 near/medium-term binarized FEWS NET outlooks and 4- and 8-month ahead model-based forecasts, compared to June/October 2019 outcomes.

	Error Type		L_A	
	FNR	**FPR**	**w=1/3**	**w=2/3**
Binarized Outlooks				
Near-term	63%	1%	30%	55%
Medium-term	80%	4%	30%	55%
Random Forest				
w=1/3, h=4	49%	8%	22%	-
w=2/3, h=4	32%	14%	-	26%
w=1/3, h=8	68%	3%	25%	-
w=2/3, h=8	54%	7%	-	39%

The first two columns report on **FNR** and **FPR** of binarized FEWS NET outlooks and out-of-sample forecasts from the wide Random Forest calibrated for the specified value of **w**. The near-term (**h**=4) validation is based on 620 non-crisis samples and 79 crisis samples, the medium-term (**h**=8) validation is based on 614 non-crisis samples and 85 crisis samples. The final two columns contain weighted averages, $\boldsymbol{L_A}$. Entries are intentionally left blank. Both the binary outcomes and outlooks are netted of humanitarian assistance effects, the validation only considers observations with a February phase rating below crisis.

As in previous sections, Table B.2 sets a baseline by reporting the performance of the binarized FEWS NET near- and medium-term outlooks. The predictive performance in this holdout sample is broadly similar to the historical results reported earlier, with very high false negative rates and extremely low false positive rates. With a weight of w=$\frac{1}{3}$, the L_A is around 30% for both near- and medium-term binarized outlooks, which is slightly worse than in the historical validation. The wide Random Forest (w=$\frac{1}{3}$) also performs somewhat worse but, nevertheless, provides a modest improvement in predictive power over these baseline results. For example, in the near-term, L_A is down to 22%, while in the medium-term, the model-based forecast performance improved the weighted error rate by 5% points.

Since FEWS NET does not provide 12-month ahead predictions, a comparison against a baseline result is not possible as it is not possible to compare single holdout validation results across different validation samples. Nevertheless, for observations that were non-crisis in respectively February, June or October 2019, but were in crisis in February 2020, the 12-month ahead forecast made in February 2019 scored an $L_A(\text{w} = \frac{1}{3})$ of 32%, 26% and 19%, which are in range of the model's 4- and 8-month ahead holdout results. The accuracy of crisis-affected populations also show that the 12-month forecast improved over medium-term forecasts constructed from binarized FEWS NET outlooks, see Table B.3. Country-specific validation results are available in Tables B.4 and B.5.

Table B.3: Holdout validation of forecasts for crisis-affected population percentages.

	MAE	**MAE***
Random Forest		
h=4	7.9	4.2
h=8	9.0	5.8
h=12	12.4	10.0
Binarized Outlooks		
Near-term	8.7	7.9
Medium-term	16.0	18.3

Mean Absolute Error for forecasts of crisis-affected population percentages (0-100), made with binarized near- and medium-term outlooks from FEWS NET and model-based predictions from the wide Random Forest produced in February, validated against June, October and February 2020 outcomes. Both the binary outcomes and outlooks are netted of humanitarian assistance effects. *Afghanistan, Uganda, the Republic of Yemen and Zimbabwe excluded in this calculation as possible outliers.

Table B.4: Country-specific holdout validation of forecasts for crisis-affected population percentages constructed from binarized FEWS NET outlooks.

	Outcomes				Forecasts			Errors		
Country	**Feb'19**	**Jun'19**	**Oct'19**	**Feb'20**	**Jun'19 (h4)**	**Oct'19 (h8)**	**Feb'20 (h12)**	**h4 error**	**h8 error**	**h12 error**
Afghanistan	51.3	3.9	16.1	47.8	41.4	4	-	37.5	12.1	-
Burkina Faso	-	-	-	-	-	-	-	-	-	-
Chad	4	7.6	3.8	3.8	4	4	-	3.6	0.2	-
Congo, Dem. Rep.	37.3	39.7	47.1	32.3	49	30.8	-	9.3	16.3	-
Ethiopia	11.1	20	16.1	16.1	15.5	13.3	-	4.5	2.8	-
Guatemala	1.9	7.8	1.9	1.9	5.3	13.1	-	2.5	11.2	-
Haiti	11	12.1	30.1	38.7	12.7	9.4	-	0.6	20.7	-
Kenya	0	9	22.7	0	0	0	-	9	22.7	-
Malawi	40.2	0	27	26.9	40.2	0	-	40.2	27	-
Mali	0	4.8	0	0	0	4.8	-	4.8	4.8	-
Mauritania	-	-	-	-	-	-	-	-	-	-
Mozambique	5	10	17.6	17.6	5	0	-	5	17.6	-
Niger	3.3	6.3	9.6	3.9	8.2	8.2	-	1.9	1.4	-
Nigeria	2	3.2	2	2.12	2	3	-	1.2	1	-
Somalia	24.9	46.2	49.5	10.4	32.8	2.5	-	13.4	47	-
South Sudan	92.3	92.8	76.8	96.4	92.3	94	-	0.5	17.2	-
Sudan	4.4	22.9	8	5.6	18	21.7	-	4.9	13.7	-
Uganda	2	2	1.2	0	2	0	-	0	1.2	-
Yemen, Rep.	99.5	99.5	99.5	99.5	99.5	99.5	-	0	0	-
Zambia	-	-	-	-	-	-	-	-	-	-
Zimbabwe	73.6	56	79.9	100	73.6	9.4	-	17.6	70.5	-

The first four columns report outcomes of crisis-affected population percentages in each country. The next two columns report the forecasts made with binarized near- and medium-term outlooks from FEWS NET. The 12-month ahead forecasts for February 2020 are intentionally left blank as FEWS NET does not provide full-year ahead outlooks. Both the binary outcomes and outlooks are netted of humanitarian assistance effects. The final three columns report on absolute errors of the near- and medium-term forecasts. The outcomes and errors for Burkina Faso, Mauritania and Zambia are left blank as new outcomes were not available for validation in these countries.

Table B.5: Country-specific holdout validation of forecasts for crisis-affected population percentages constructed from predictions from the wide Random Forest.

	Outcomes				Forecasts			Errors		
Country	**Feb'19**	**Jun'19**	**Oct'19**	**Feb'20**	**Jun'19 (h4)**	**Oct'19 (h8)**	**Feb'20 (h12)**	**h4 error**	**h8 error**	**h12 error**
Afghanistan	51.3	3.9	16.1	47.8	35.9	39.7	43.6	32.0	23.6	4.2
Burkina Faso	-	-	-	-	1.0	0.5	1	-	-	-
Chad	4	7.6	3.8	3.8	22.2	10.2	16	14.6	6.4	12.2
Congo, Dem. Rep.	37.3	39.7	47.1	32.3	41.3	43.9	40	1.6	3.2	7.7
Ethiopia	11.1	20	16.1	16.1	24.6	15.8	14.7	4.6	0.3	1.4
Guatemala	1.9	7.8	1.9	1.9	8.0	1.0	3.9	0.2	0.9	2
Haiti	11	12.1	30.1	38.7	9.6	4.7	6.3	2.5	25.4	32.4
Kenya	0	9	22.7	0	6.9	9.5	7.9	2.1	13.2	7.9
Malawi	40.2	0	27	26.9	2.4	28.4	33.3	2.4	1.4	6.4
Mali	0	4.8	0	0	4.9	0.2	1.3	0.1	0.2	1.3
Mauritania	-	-	-	-	25.3	6.9	12.9	-	-	-
Mozambique	5	10	17.6	17.6	1.2	7.5	13.9	8.8	10.1	3.7
Niger	3.3	6.3	9.6	3.9	10.2	4.2	3.2	3.9	5.4	0.7
Nigeria	2	3.2	2	2.12	4.1	1.9	2.6	0.9	0.1	0.48
Somalia	24.9	46.2	49.5	10.4	47.4	47.4	49.5	1.2	2.1	39.1
South Sudan	92.3	92.8	76.8	96.4	82.6	78.6	84.4	10.2	1.8	12
Sudan	4.4	22.9	8	5.6	29.0	18.4	18.8	6.1	10.4	13.2
Uganda	2	2	1.2	0	1.5	0.4	1.7	0.5	0.8	1.7
Yemen, Rep.	99.5	99.5	99.5	99.5	95.9	95.8	96	3.6	3.7	3.5
Zambia	-	-	-	-	0.1	0.1	0.1	-	-	-
Zimbabwe	73.6	56	79.9	100	8.9	26.1	27.1	47.1	53.8	72.9

The first four columns report outcomes of crisis-affected population percentages in each country. The next three columns report the forecasts made by aggregating the 4-, 8- and 12-month ahead predictions from the wide Random Forest using the methods described in the paper. The outcomes and predictions are netted of humanitarian assistance effects. The final three columns report on absolute errors of the three forecasts. Burkina Faso, Mauritania and Zambia are left blank as outlooks and outcomes were not available for validation in these countries.

B.3 Alternative models

The sections below detail two alternative algorithms and their application to our data. Table B.6 contains results of a simple linear classifier and a more complex nonlinear approach. The results highlight the attractive properties of the Random Forest application.

Table B.6: Cross-validation results for all models.

	Error Type						**Balanced Metrics**					
	FNR			**FPR**			L_A			L_B		
w	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**	**1/3**	**1/2**	**2/3**
Basic Random Forest												
h=4	19.7%	16.1%	16.1%	7.7%	9.3%	9.3%	11.7%	12.7%	13.8%	0.27	0.30	0.32
h=8	19.2%	16.7%	16.0%	8.5%	9.7%	10.6%	12.1%	13.2%	14.2%	0.28	0.31	0.33
h=12	24.7%	23.1%	10.6%	11.7%	12.6%	23.6%	16.1%	17.9%	14.9%	0.34	0.37	0.38
Wide Random Forest												
h=4	19.7%	16.5%	16.0%	6.8%	8.2%	8.4%	11.1%	12.3%	13.5%	0.25	0.29	0.31
h=8	19.1%	16.6%	16.6%	7.3%	8.5%	8.5%	11.2%	12.5%	13.9%	0.27	0.30	0.33
h=12	20.9%	20.9%	9.9%	10.0%	10.0%	19.0%	13.6%	15.5%	13.0%	0.30	0.34	0.35
Basic Logit												
h=4	49.3%	30.7%	5.1%	13.7%	24.4%	52.5%	25.6%	27.6%	20.9%	0.48	0.51	0.47
h=8	54.9%	27.7%	4.4%	14.7%	32.7%	58.3%	28.1%	30.2%	22.3%	0.52	0.55	0.51
h=12	53.8%	31.2%	4.3%	15.2%	28.0%	60.6%	28.0%	29.6%	23.1%	0.51	0.55	0.51
Wide Logit												
h=4	49.4%	21.6%	4.6%	13.5%	29.8%	52.2%	25.5%	25.7%	20.5%	0.48	0.51	0.47
h=8	57.2%	28.4%	4.8%	14.2%	32.6%	59.1%	28.5%	30.5%	22.9%	0.52	0.56	0.51
h=12	54.8%	31.5%	3.9%	15.1%	27.9%	60.0%	28.3%	29.7%	22.6%	0.51	0.55	0.50
Basic Multi-layer Perceptron												
h=4	26.9%	16.4%	9.2%	11.9%	18.2%	26.1%	16.9%	17.3%	14.8%	0.37	0.39	0.36
h=8	28.4%	16.4%	9.3%	12.7%	20.9%	29.9%	17.9%	18.6%	16.2%	0.39	0.41	0.38
h=12	28.6%	15.5%	8.6%	17.5%	26.4%	36.1%	21.2%	21.0%	17.8%	0.43	0.46	0.42
Wide Multi-layer Perceptron												
h=4	26.0%	12.6%	10.5%	11.6%	21.3%	23.9%	16.4%	17.0%	15.0%	0.37	0.39	0.35
h=8	27.1%	17.4%	11.1%	12.2%	18.9%	26.2%	17.2%	18.2%	16.1%	0.39	0.41	0.38
h=12	30.0%	17.6%	8.9%	15.2%	23.2%	35.8%	20.1%	20.4%	17.8%	0.42	0.45	0.41

The first six columns report false negative rates and false positive rates when the model is optimized for the indicated value of **w**. The next six columns report three weighted averages ($\boldsymbol{L_A}$) of **FNR** and **FPR** and weighted Log Loss ($\boldsymbol{L_B}$) with weights **w** and (1-**w**) for predictions in the positive (negative) class. The results have been produced using identical methods as described in the main text.

B.3.1 Logistic Regression with Lasso Penalty

The logistic regression framework shows what can be achieved with simple and interpretable tools. We apply the LASSO, or L_1 norm, following Tibshirani (1996); Zou (2006) to penalize the Maximum Likelihood Estimator by the absolute sum of the coefficients, thereby discouraging high parameter values. The training data was balanced with up-sampling, and for computational efficiency duplicate cases are converted into case weights. This means that the estimation problem is of the following form.

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i^*, \beta_0 + \beta' x_i^*) + \lambda ||\beta||_1, \tag{6}$$

in which w_i are case weights generated under the up-sampling scheme, and $l(\varepsilon)$ is the negative log-likelihood contribution for observation i. The term $\lambda||\beta||_1$ applies absolute shrinkage and can set parameters of noise predictors to zero. The strength of the penalty is controlled by λ, which is a tuning parameter. To find the right level of penalization, λ is set through cross-validation. The ridge, or L_2 penalty following Park and Hastie (2008) and the elasticnet of Friedman et al. (2010), that combines both by using a penalty of the form $\lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$, were explored and led to near identical results as obtained using equation 6. Because this model is a linear method, we do not use the Food Price Index levels; instead we only use Food Price Inflation.

B.3.2 Multi-Layer Perceptron Neural Network

Neural networks came into active development as a classification tool after the mid-1980s when their relation to statistical pattern recognition was studied (Ripley, 1993, 1994). The basic architecture upon which neural networks are built is the perceptron of (Rosenblatt, 1958) which consists of artificial neurons that take inputs, process by taking a weighted sum, and pass the result through a transformation to return an output. In a directed graph, the perceptron consists of three layers that: (1) represent the inputs to the model, (2) define the computational operations on the inputs, and (3) define model outputs. In its simplest form, one computational neuron, this model could be represented as an equation encountered frequently in econometrics:

$$\hat{y} = f(\mathbf{x}\hat{\omega}) \tag{7}$$

where $\hat{y}$ is a predicted value using a vector of observations $\mathbf{x}$ multiplied by some estimated linear weights $\hat{\omega}$, possibly adding also a scalar bias component. The product $\mathbf{x}\hat{\omega}$ is transformed through an *activation function* f to obtain the final prediction. Indeed, this is essentially a Generalized Linear Model in which f would be referred to as the *link function*: if f is chosen to be a logistic function, equation 7 is simply a logistic regression model. Because the non-linear transformation always occurs after the linear operation, it is trivial to see that regardless of the shape of f, a single computational neuron can perform classification only on linearly separable patterns. It is possible to increase the number of units to create more dividing lines, but those lines must somehow be combined to form more complex classifications. This is the novelty of multi-layer perceptron. The neural network for $k \in [0, 1]$ class output, with output function f_o, multiple computational units and an activation function for hidden units f_h, can be written as

$$\hat{y}_k = f_o\left(\hat{\alpha}_k + \sum_j \hat{\omega}_{jk} f_h\left(\hat{\alpha}_j + \sum_i \hat{\omega}_{ij} x_{ij}\right)\right). \tag{8}$$

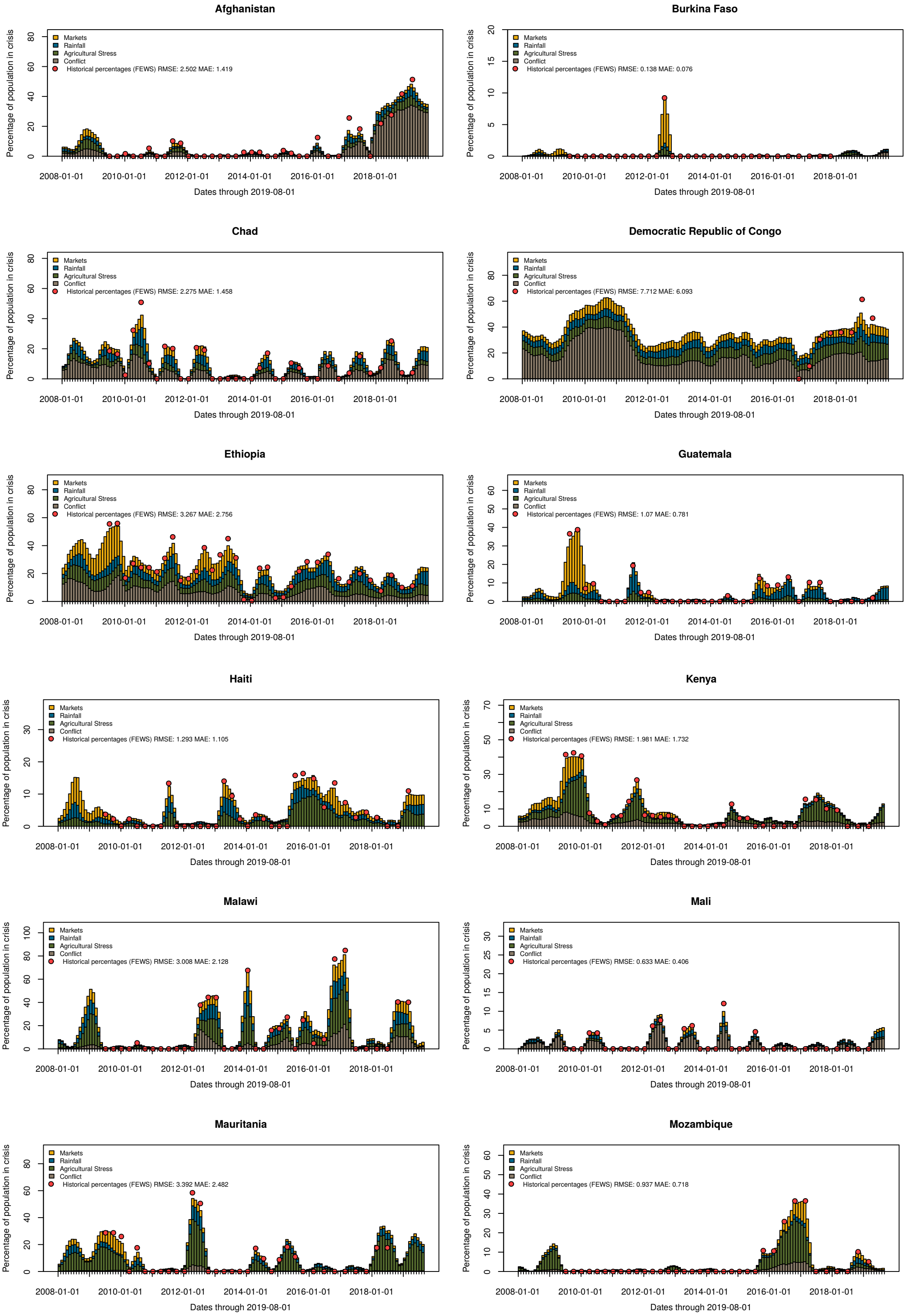
This is often referred to as the multi-layer perceptron with one hidden layer. The weights in this model can be estimated using a standard back-propagation algorithm (Rumelhart et al., 1986). The input units distribute the inputs to the 'hidden' units in the second layer. The hidden units then sum their inputs, add a constant and take a function f_h of the result. The output units apply the same procedure using an output function f_o. In this paper, both f_o and f_h are parameterized with logistic functions hence the model can be simply understood as a nonlinear logistic regression model.

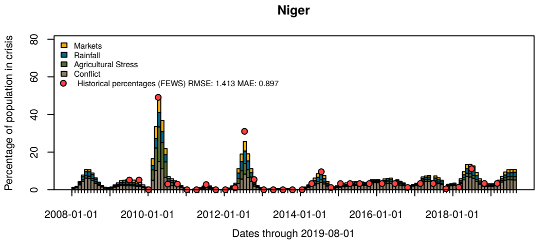
By increasing the network complexity, deep neural networks can learn increasingly complex types of non-linearity. A random grid search was performed across 20 randomly generated architectures keeping other learning parameters at their recommended values of Bergmeir and Benítez (2012), whose implementation has been used here. The number of hidden units in the first hidden layer is randomly assigned taking a number from 2 to one-third of the number of predictors, and the number of hidden units in the second and third layers is any number from zero to one-third of the number of predictors. A hidden layer of 0 units is removed. The best configuration was picked by cross-validation and compared against a systematic search across single hidden layer networks to confirm that the preferred model was deep.

B.4 Forecast decompositions

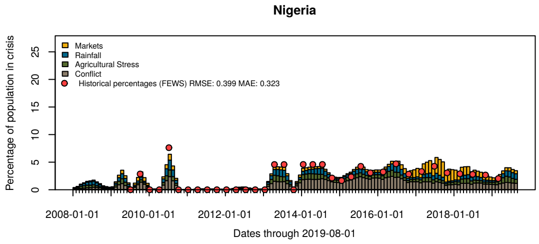
The paper suggested a prediction decomposition strategy. Figure B.1 contains the results for all countries in the data set. The variety in results highlights that predictions for different major crisis events are traced back to different input variables, highlighting the heterogeneity in fitted relationships.

Figure B.1: National-level 1-month ahead prediction decompositions from the final wide Random Forest.

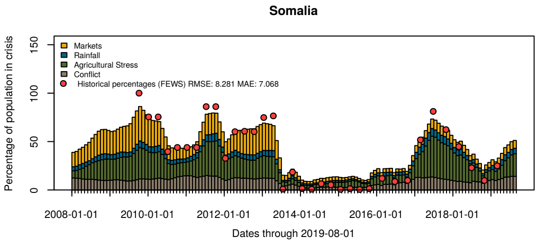




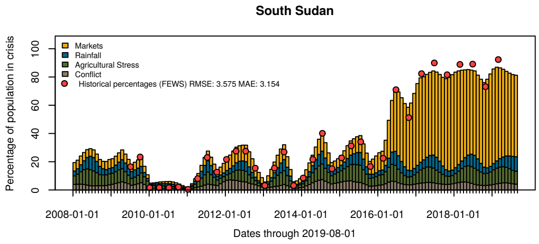
Niger



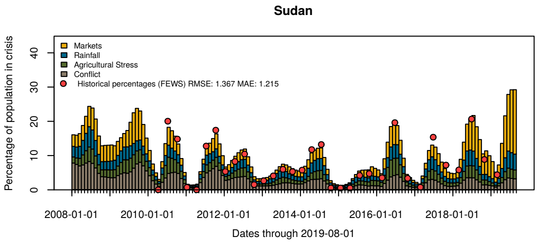
Nigeria



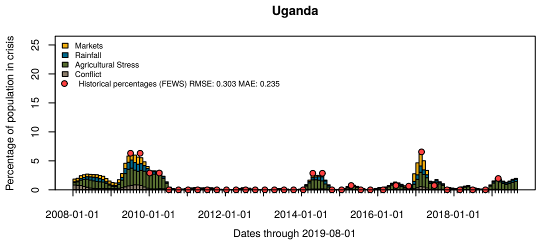
Somalia



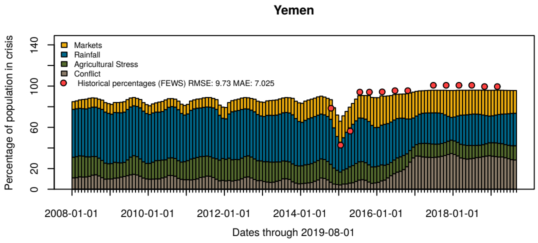
South Sudan



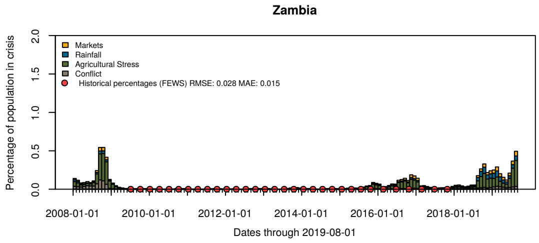
Sudan



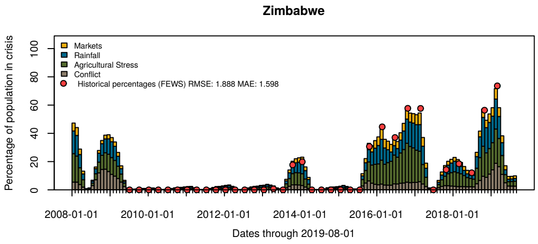
Uganda



Yemen



Zambia



Zimbabwe