Performance Uncertainty and Ranking Significance of Early-Warning Models[1]

Suman S. Basu
International Monetary Fund

Roberto A. Perrelli
International Monetary Fund

Weining Xin
University of Southern California

Abstract

Due to the small data nature of macroeconomic data in early-warning framework, it is critical to assess model performance uncertainty and test model ranking significance when conducting a horse race and selecting the best model to use. To assess model performance uncertainty, this paper explores three sources of data variation in early-warning framework and proposes three types of jackknifing methods to construct confidence intervals of model performance respectively. Additionally, this paper proposes to construct confidence intervals of conditional performance difference and performs hypothesis testing on the conditional performance difference to test model ranking significance. The approaches are illustrated in an example of predicting sudden stops in capital flows for emerging market countries. Results show that the degree of model performance uncertainty depends on the structure of model and the source of data variation. Also, our approach to construct confidence intervals of conditional performance difference presents evidence of model ranking significance which is otherwise not revealed in simply comparing confidence intervals of individual model performance.

JEL Classification Numbers: C53, E37, F47

Keywords: Early warning; performance uncertainty; statistical inference; sudden stop

[1] Contacts: Suman S. Basu (SBasu2@imf.org), Roberto Perrelli (RPerrelli@imf.org), and Weining Xin (weiningx@usc.edu). The views expressed in this paper are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

I. INTRODUCTION

Anticipating and preparing for crises lie at the heart of the mandate of central banks, which yet are intrinsically difficult tasks. Early-warning models are developed to tackle this challenge. The history of early-warning models can go far back to two decades ago, when Kaminsky et al. (1998) introduced signal extraction model and Frankel and Rose (1996) applied logit regression model to predicting currency crises. As machine learning has achieved success in many areas over the past decade, they were introduced to the literature of early-warning models and enrich the set of models by allowing more flexible relationships between crisis events and early-warning indicators.[2] Despite of the success of machine learning in many prediction areas, macroeconomic data and early-warning exercise have their unique features which may not guarantee better performance of machine learning than traditional statistical methods. Hence, in order to select one out of many to use for predicting crises and informing policy making, it is important for researchers and policymakers to conduct a horse race and rank models.

At the heart of ranking models lie model performance uncertainty arising from sampling and model ranking significance accounting for sampling errors. Model performance uncertainty arising from sampling refers to the extent to what a model performed differently if it was estimated on a different dataset, and model ranking significance accounting for sampling errors refers to whether a model performs better than another significantly, accounting for model performance uncertainty arising from sampling. Estimating model performance uncertainty and testing model ranking significance are especially important for early-warning model selection, as early-warning models are used not only for predicting crisis outcomes, but also for understanding risk factors and informing policy decisions. In case of no significant difference in performance between one traditional statistical model and one machine learning model, the former shall be recommended to policymakers for using in practice given its higher degree of interpretability and stability.[3] Hence, this paper touches on this problem by proposing approaches to estimate performance uncertainty and test rank significance and illustrating the approaches in an early-warning framework for sudden stops.

Macroeconomic panel data in the early-warning framework is small in three important aspects. First, there are not many countries in the world and even fewer crisis events in the past, which means that historical data may have only a few lessons for the future. Additionally, a high degree of heterogeneity is present among this small set of countries and their crises, which means that each country and its crises may play an important role in model estimation and testing. Hence, it follows that the set of countries and their crisis events matter for model performance and rankings. Second, infrequent but large global regime shifts limit the applicability of past lessons for future crisis prediction. The rarity of the shifts means that a few decades of available historical data cannot capture all the possible global regimes, and the large size of the shifts means that past information may become suddenly rather outdated.

[2] Relevant papers include Chamon et al. (2007), Sevim et al. (2014), Xu et al. (2018) and Basu et al. (2019) for external crises; Holopainen and Sarlin (2015), Alessi and Detken (2018), and Lang et al. (2018) for banking crises; Manasse and Roubini (2009), Savona and Vezzoli (2015) and Badia et al. (2020) for sovereign crises.

[3] Although the randomness coded in many machine learning algorithms help to prevent overfitting, it reduces the degree of interpretability and stability of such techniques.

Hence, it follows that the set of global regimes which is contained in the set of years matter for model performance and rankings. Third, in addition to global regime shifts that affect countries simultaneously, countries could have been hit by country-specific idiosyncratic shocks and therefore part of their histories might be different from what should have been. Hence, it follows that the set of country-specific histories matter for model performance and rankings.

The small data nature of macroeconomics panel data strengthens the need for assessing model performance uncertainty arising from sampling, and testing model ranking significance accounting for sampling. There is a one-to-one mapping from the above three aspects in which the macroeconomics panel data is small in the early-warning framework and three sources of data variation worth examining. First, data variation in the set of countries should be examined because countries and their crisis events are few and heterogeneous. Second, data variation in the set of years should be examined because global regime shifts are rare but significant. Third, data variation in the set of countries' histories should be examined because countries could have been hit by country-specific idiosyncratic shocks and therefore have developed in a way different from what they should have been.

To assess model performance uncertainty arising from the above three sources of data variation, we choose to make use of the jackknife resampling to obtain new samples and construct confidence intervals to represent the uncertainty. The jackknife method resamples the original dataset by dropping data, which means that new samples are subsamples of the original one. By specifying how data is selected to drop, the jackknife resampling allows us to impose priors on the dimension along which the dataset is resampled, so we are able to isolate the source of data variation and estimate the model performance uncertainty arising from specific sources. In line with the three sources of data variation discussed above, the jackknife resampling is performed along three dimensions: (1) dropping countries, that is dropping all years' data for some randomly chosen countries; (2) dropping years, that is dropping all countries' data in some randomly chosen years; and (3) dropping country-year blocks, that is dropping some randomly chosen blocks of a given number of years for some randomly chosen countries. Also, we apply the jackknife resampling to the entire dataset splitting it into training set and test set for two reasons. For one, sampling errors could rise in any part of the data, regardless of how one splits the dataset into training set and test set. Specifically, when global regime shifts happened, or countries were hit by country-specific idiosyncratic shocks did not depend on how researchers or policymakers design their training and testing scheme. For another, one should never manipulate the splitting of a given dataset into training and test set by restricting all potential data variation in only training or test set.[4] As for the percentage of data to drop, we choose to drop a fair fraction of the original data (10 percent and 5 percent), instead of dropping one single observation (i.e., a country-year pair in our data) as in standard jackknife resampling, for a few reasons behind. First, due to cross-sectional and time-series dependence of macroeconomics panel data, dropping a single observation cannot generate enough data variation, which means that we need to drop a larger fraction of data to generate enough data variation for uncertainty assessment. Second, empirically, our first and second jackknifing

[4] However, we acknowledge that this way of jackknife resampling on the entire dataset limits our ability to decompose model performance uncertainty into that arising from data variation in training set and test set, which may need more sophisticated design of resampling methods to investigate.

methods place a lower bound on the percentage of data to drop, because they are designed to drop all years' data for some countries and all countries' data in some years.[5] We also consider a fourth jackknifing method, which is to treat the data as i.i.d., and drop randomly single country-year pairs to examine how estimates of model performance uncertainty are different when accounting for the panel data structure of our data or not.

In addition to estimating model performance uncertainty arising from sampling, we also make use of the jackknife resampling to test model ranking significance. We argue that simply examining whether confidence intervals of individual model performance overlap does not provide much evidence on model ranking significance, because confidence intervals are generated by pooling performances calculated from different resamples. However, when testing the null hypothesis whether one model performs significantly better than another, the comparison should be conducted on the same subsample generated by the jackknife resampling. Hence, we propose to construct confidence intervals of the conditional model performance difference, that are confidence intervals of difference in model performance calculated from the same jackknife resampled dataset. And then a null hypothesis that the conditional model performance difference is equal to zero is tested based on the confidence interval results.

Our approaches are illustrated in an early-warning framework of predicting sudden stops in capital flows for emerging market countries. Sudden stops have been disruptive crisis events for emerging market countries over the past three decades. The capital accounts of these countries are open enough for private capital inflows to accumulate, but not sufficiently liberalized for sudden outflows to be easily insured against. The danger of such brutal crisis events for emerging market countries have been learned the hard way so that it is important to monitor sudden stop risks, issue early warnings, and inform policy decisions. We consider two models, signal extraction model, a simple statistical method that has been extensively used and tested in the early-warning literature (Berg et al., 2005), and random forests, a machine learning method that has been proven to be successful in many prediction areas and applied to the early-warning literature in the past decade.[6] There are four main findings: (1) confidence intervals for signal extraction model are wider than those for random forests, for all types of jackknifing methods; (2) confidence intervals generated by dropping years are the widest among all for signal extraction model, while confidence intervals generated by dropping country-year blocks are the widest among all for random forests; (3) there is not much difference among model performance uncertainty arising from different sources of data variation; (4) signal extraction model performs significantly better than random forests (at 0.01 significance level) in fixed

[5] In the data in our illustrative example, there are in total 10 countries and 28 years. Hence, it implies that the smallest proportion of data to drop by dropping countries is 1/10, i.e., 10 percent, and the smallest proportion of data to drop by dropping years is 1/28, i.e., around 3.6 percent. In order to have a fair comparison among the model performance uncertainty arising from all three sources of data variation, we need to drop the same proportion of data in all three jackknifing methods. It then follows that 10 percent is the lower bound of the fraction of data to drop.

[6] Most recent papers using random forests for crisis prediction are Basu et al. (2019) for external crises, Lang et al. (2018) for banking crises, and Badia et al. (2020) for sovereign crises.

cutoff testing, while signal extraction model and random forests do not perform significantly differently in rolling cutoff testing.[7]

The rest of the paper is structured as follows: Section 2 discusses in detail the importance of assessing model performance uncertainty arising from sampling and testing model ranking performance accounting for sampling errors in an early-warning framework. Section 3 describes our data, including crisis definition and explanatory indicators, and choices of early-warning models. Section 4 explains our model estimation and testing, and approaches to assess model performance uncertainty and test model ranking significance, including resampling methodology and confidence interval construction. Section 5 presents and discusses our empirical findings on model performance uncertainty and model ranking significance. Finally, Section 6 concludes.

II. SMALL MACRO DATA

Suppose the true model of crisis prediction is

$$y_{it} = f(X_{it-1}, \varepsilon_{it}; \eta_{t-1}),$$

where y_{it} is the crisis event for country i in year t, taking the value of 0 when there is no crisis, or 1 if there is a crisis, X_{it-1} is a vector of country-specific explanatory indicators for country i in year $t-1$, f denotes a non-linear relationship between X_{it-1} and y_{it}, ε_{it} is an idiosyncratic shock, and η_{t-1} is a vector of global factors capturing the global regime which affects the relationship f between X_{it-1} and y_{it}.

Macroeconomics data in the early-warning framework is small in three importance aspects. First, there are not many countries in the world and even fewer crisis events y_{it} in the past, which means that historical data may have only a few lessons for the future. Additionally, a high degree of heterogeneity is present among this small set of countries and their crises, which means that each country and its crises may play an important role in model estimation and testing. Second, infrequent but large global regime shifts captured by variation in η_{t-1} limit the applicability of past lessons for future crisis prediction. The rarity of the shifts means that a few decades of available historical data cannot capture all the possible global regimes, and the large size of the shifts means that past information may become suddenly rather outdated. Third, in addition to global regime shifts that affect countries simultaneously, countries could have been hit by country-specific idiosyncratic shocks and therefore part of their histories reflected in X_{it-1} might be different from what should have been.

[7] In the fixed cutoff testing where year 2007 is the cutoff year, a model is estimated on the training set, consisting of data from 1990 to 2007, and then tested on the test set, consisting of data from 2008 to 2017. In the rolling cutoff testing where years 2007, 2009, 2011, 2013, and 2015 are cutoff years, a model is recursively estimated on the training set consisting of data before the cutoff year, and then tested on the test set consisting of data in the next two years. In the end, model performance on the test sets from different cutoff years are averaged.

The small data nature of macroeconomic panel data strengthens the need for assessing model performance uncertainty arising from sampling, i.e., the extent to what a model performed differently if it was estimated on a different dataset. In line with the three aspects in which macroeconomic data in the early-warning framework is small, there are three sources of data variation. First, countries and their crisis events are few and heterogeneous, which means that whether some of the countries (and thus their crisis events) are in the sample could affect model performance. Hence, data variation in the set of countries is worth examining. Second, global regime shifts are rare but significant, which means that whether some of the years (and thus global regimes) are in the sample could affect model performance. Hence, data variation in the set of years are worth examining. Third, some of the countries could have been hit by country-specific idiosyncratic shocks and therefore could have developed differently in part of their histories, which means that whether part of the histories of some countries are in the sample could affect model performance. Hence, data variation in the set of countries' histories is worth examining. Hence, we conclude that there are three sources of data variation that may affect model performance in the early-warning framework:

1. Variation in the set of countries, i.e., what if some of the countries were not in the sample?

2. Variation in the set of years, and thus the set of global regimes, i.e., what if some of the global regimes were not seen?

3. Variation in the set of countries' histories, i.e., what if some of the countries followed different trajectories in part of their histories?

III. Data and Models

A. Crisis Events

We focus on sudden stops of capital flows in emerging market countries, which have been seen as the most brutal crises for such economies. Our sample covers ten countries that have been well accepted to be emerging market countries over the past three decades, including Argentina, Brazil, Chile, Indonesia, Malaysia, Mexico, Philippines, Russia, Thailand, and Turkey. The crisis events are chosen based on the sudden stop definition in Basu et al. (2019). In the definition, a sudden stop is defined as occurring when net private capital inflows as a percentage of GDP is at least 2 percentage points lower than that in the previous year and two years before, as well as when the country gets approved to tap large IMF financial support to capture counterfactual situations in which sudden declines in private capital inflows were prevented by large IMF financial support.[8] Also, such brutal events often cause severe real economic consequences, such as large growth declines which are defined as occurring when the changes in real GDP growth relative to the previous five-year average lie in the lower 10th percentile of entire sample, as well as when the country gets approved to tap large IMF financial support to capture counterfactual situations in which large growth declines were

[8] Large IMF financial support hereafter is defined as IMF arrangements with agreed amount at least five times as large as the respective country's quota at the IMF.

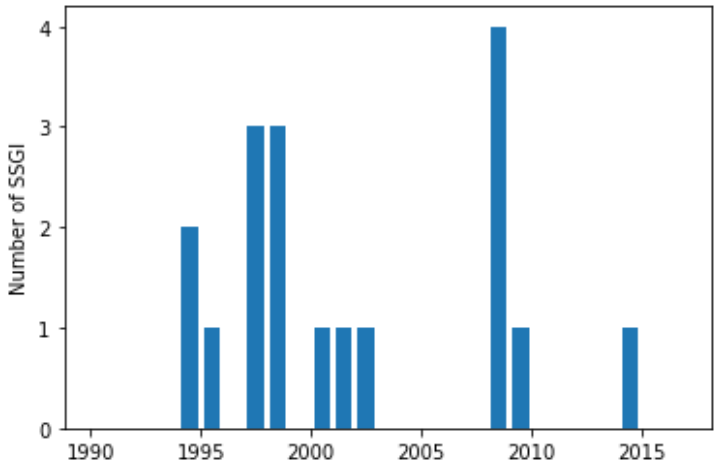
prevented by large IMF financial support.[9] Therefore, combining the two definitions, episodes of sudden stops with growth impacts (SSGI) are the main crisis events we focus on in this paper.

Table 1. Episodes of Sudden Stops with Growth Impacts

Countries	Years
Argentina	1995, 2000, 2008
Brazil	2002
Chile	1998
Indonesia	1997
Malaysia	1997, 2008
Mexico	1994, 2009
Philippines	1998
Russia	2008, 2014
Thailand	1997
Turkey	1994, 1998, 2001, 2008

Our sample spans from 1990 to 2017, during which there are eighteen sudden stops with growth impacts, accounting for 6.4 percent of the sample. Table 1 lists the eighteeen episodes of sudden stops with growth impacts (SSGI): twelve before the global financial crisis; five during the global financial crisis; and one after the global financial crisis. Figure 1 shows the crisis frequency distribution across years. It can be seen that our ten-country sample over the period 1990-2017 captures prominent historical waves of sudden stops including the Mexican peso crisis in the mid-1990s, the Asian financial crises in the late-1990s, South American crises in the early-2000s, and the global financial crisis in the late-2000s.

Figure 1. Frequency of Sudden Stops with Growth Impacts



[9] The entire sample in Basu et al. (2019) covers 53 countries and spans the period between 1990 and 2017. Hence, the ten-country sample used in this paper is a subset of the sample in Basu et al. (2019). Instead of using the lower 10th percentile of our ten-country sample, we take the value corresponding to the lower 10th percentile of the sample in Basu et al. (2019) to define large growth declines in our sample for robustness.

B. Explanatory Indicators

Our set of explanatory indicators consists of twenty-five variables that have been selected using a general-to-specific approach. Specially, we start with the set of more than seventy explanatory variables used in Basu et al. (2019).[10] Then we select a subset of twenty-five variables based on their horse race results. In particular, these twenty-five variables span multiple sectors including external, fiscal, financial, and real sectors, and can be categorized into four groups capturing different economic mechanisms: medium-term bubble building, short-term bubble bursting, buffers and mismatch, and global factors. Table 2 lists the selection of explanatory indicators.

Table 2. List of Explanatory Indicators

Variable	Source	Variable	Source
Medium-term bubble building:		**Shor-term bubble bursting:**	
5-year inflation		Change in public debt	
5-year money growth		Change in reserves	
5-year stock price growth		Change in stock price growth	
5-year housing price growth		Change in housing price growth	
5-year inter-bank liabilities growth		Change in external equity liabilities	
5-year REER growth		Change in REER appreciation	
5-year private credit growth		Change in private credit	
Buffers and mismatch:		**Global factors:**	
Current account balance		TED spread	
Amortization-to-exports ratio		Percentage of AEs in banking crises	
EMBI spread		Inter-bank liabilities to AEs in banking crises	
Foreign liabilities-to-domestic credit ratio		Export growth	
External debt			
Capital adequacy ratio			
Interest coverage ratio			

C. Model Choice

The set of early-warning models for sudden stops have been developed along the historical crisis waves in which the danger of sudden stops is learned, from simple statistical methods such as signal-extraction model (Kaminsky et al., 1998) and regression-based models (Frankel and Rose, 1996; Berg and Patillo, 1999) to recent more advanced machine learning techniques (Chamon et al., 2007; Basu et al., 2019). In this paper, we choose to focus on signal extraction model and random forests (Breiman, 2001), assessing their performance uncertainty and testing their model ranking significance. Our choices are motivated by two reasons. First, signal-extraction model has been shown to perform the best on predicting sudden stops when compared with traditional regression models and machine learning techniques including

[10] It is worth noting that their selection is supported by different generations of theoretical models on sudden stops in the literature. Explanatory variables are categorized into several groups based on different economic channels and/or mechanisms.

regularized regression models and tree-based models (Berg et al., 2005 and Basu et al., 2019), especially in terms of out-of-sample performance.[11] Second, as a well-known and widely-used machine learning method, random forests has been applied to early-warning exercises and shown to perform well on many types of crises including banking crises, currency crises and sovereign debt crises (Alessi and Detken, 2018; Basu et al., 2019; and Badia et al., 2020).

The signal-extraction model identifies one threshold for each explanatory indicator that minimizes the specified loss function. Observations (i.e., country-year pairs in our data) whose indicator variable values fall on one side of the threshold are given a 1 and flagged as risky, otherwise are given a 0 and flagged as safe. Then flags of all indicators of an observation are aggregated to generate a composite score (which sometimes is called vulnerability index in the literature and in practice) with weights given by their signal-to-noise ratio:

$$\frac{1-z}{z},$$

where z is defined as the value of the loss function achieved. Therefore, such algorithm implies that minimizing the loss function for each indicator is equivalent to maximizing the signal-to-noise ratio for each indicator and indicators with larger signaling power are given larger weights in the composite score.

We follow the literature (Berg et al., 2005) to use the loss function that is the unweighted sum of the percentages of false alarms and missed crises. The percentage of "false alarms" is defined to be the percentage of non-crisis observations (i.e., country-year pairs in our data) that the model incorrectly flags as crises, while the percentage of "missed crises" is defined to be the percentage of crisis observations (i.e., country-year pairs in our data) that the model incorrectly flags as non-crises. The threshold chosen to minimize this loss function is therefore the one for which the vertical gap between the conditional cumulative distribution function of crisis observations and the conditional cumulative distribution function of non-crisis observations is maximized.[12]

Random forests, introduced by Breiman (2001), is an ensemble method consists of a number of classification trees as building blocks. Classification tree (Breiman et al., 1984) uses a decision tree to flag an observation by going from the original complex sample to smaller and purer subsamples. Each decision tree consists of a root node, branches departing from parent nodes and entering child nodes, and multiple terminal nodes which are also called leaves. In the structure of classification tree, leaves represent the flagged classes and branches represent the conjunctions of indicators that lead to the classes.[13] Observations in the root node are sent to left or right child node according to some splitting rules that identify indicators and corresponding thresholds. Such process is repeated sequentially on each child node recursively

[11] Also, the univariate and non-parametric setting of signal-extraction model for identifying variable specific thresholds makes it more practical on macroeconomic data for which data availability varies from variable to variable.

[12] The cumulative distribution function of crisis or non-crisis observations are plotted against the value of the explanatory indicator.

[13] The class for a leaf is determined by the class with the most votes in the leaf.

until each leaf consists of only one class or some stopping criteria are met. The indicator and threshold used to split the sample at each node are chosen based on some measures of impurity, such as the Gini impurity index. Because of the recursive algorithm, such tree structure partitions the prediction space into multiple smaller spaces, which allows for complex relationship between the target and explanatory indicators. Hence, the classification tree has proven useful in many prediction areas and has been introduced to the early-warning literature (Chamon et al., 2007 and Manasse and Roubini, 2009).

The method of classification tree suffers from overfitting when a single decision tree is grown very deep and therefore includes too much noise from the sample it is estimated from. To reduce overfitting of one single classification tree, random forests grows a number of classification trees based on bootstrapped samples, i.e., random samples selected with replacement from the original sample. Additionally, instead of considering all explanatory indicators, only a random subset of indicators are considered as candidates for each split. Such algorithm, sometimes called feature bagging, effectively prevents strong correlation among trees in the forest. The final predicted class for a new observation is achieved by taking the majority vote of predictions of all trees in the forest. Bootstrap aggregating and feature bagging reduce the prediction variance on average, without increasing the prediction bias, and therefore help random forests achieve better classification performance.

IV. Model Performance Uncertainty

The estimation of model performance uncertainty consists of three steps: (1) generating new samples from the original sample, (2) estimating and testing models to collect model performance, and (3) constructing confidence intervals to represent model performance uncertainty. In this section, we first describe our methods to generate new samples (Subsection A). We then summarize how we estimate and test models (Subsection B). The last Subsection (Subsection C) discusses how we construct confidence intervals to represent model performance uncertainty. Our estimation of model performance uncertainty proceeds in the following way:

1. Perform jackknife resampling on the original entire sample S to obtain a jackknifing sample $S_j, j = 1, 2, \dots, N$.
2. Split the jackknifing sample S_j into training set and test set based on cutoff rules, either a fixed cutoff or rolling cutoffs.
3. Estimate different models (signal extraction model and random forests) on the same training set and tested on the same test set generated from the jackknifing sample S_j. Model performance on the test set are then calculated and collected.
4. Repeat 1.-2. for $N = 200$ times, and construct confidence intervals using the model performance collected.

A. Jackknife Resampling

As discussed in Section 2, the "small data" nature of macroeconomic panel data in the early-warning framework strengthens the need for assessing model performance uncertainty arising

from sampling, and therefore motivates three sources of data variation from which model performance uncertainty arises. We opt to make use of jackknife resampling (jackknifing) to assess model performance uncertainty, which allows us to examine and compare the model performance uncertainty arising from different source of data variation.

The jackknifing was introduced by Efron and Stein (1981) and Efron (1982) before other common resampling methods such as the bootstrap resampling (bootstrapping). The standard jackknifing simply omits one single observation of the original sample to generate a subsample. Specifically, given a sample consisting of data $x_1, x_2, \dots, x_N$, the i_{th} jackknifing subsample consists of data $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N$ for $i = 1,2, \dots, N$. However, the standard jackknifing assumes data to be i.i.d., and therefore does not work well on our macroeconomic panel data that exhibit both cross-sectional and time-series dependence. Specifically, the cross-sectional dependence is derived from the presence of global factors that affect countries simultaneously, and the time-series dependence is derived from the autocorrelation of explanatory indicators. Because of such dependence in our macroeconomic panel data, dropping one observation in the form of a country-year pair cannot generate sufficient data variation in the collection of jackknifing subsamples.

Hence, in line with the three sources of data variation discussed in Section 2, we propose three different jackknifing methods to assess the model performance uncertainty from three different sources of data variation. First, we drop countries, that is to drop all years' data of some randomly chosen countries, to assess the extent to what a model performed differently if some of the countries (and thus their crisis events) were not in the same group as others. Second, we drop years, that is to drop all countries' data in some randomly chosen years, to assess the extent to what a model performed differently if some of the global regimes were never seen or some of the global regime shifts never happened. Third, we drop country-year blocks, that is to drop data in the form of randomly chosen blocks of three years for some randomly chosen countries, to assess the extent to what a model performed differently if some of the countries were hit by idiosyncratic shocks and therefore part of their histories followed a different trajectory from what should have been.

We apply different jackknifing methods on the original sample before it is split into training and test set, instead of only resampling only the training set while keeping the test set untouched (Holopainen and Sarlin, 2015). There are two reasons: For one, sampling errors could rise in any part of the data, regardless of how one splits the dataset into training set and test set. Specifically, when global regime shifts happened, or countries were hit by country-specific idiosyncratic shocks did not depend on how researchers or policymakers design their training and testing scheme. For another, one should never manipulate the splitting of a given dataset into training and test set by restricting all potential data variation in only training or test set. However, we acknowledge that this way of jackknife resampling on the entire sample limits our ability to decompose model performance uncertainty into that arising from data

variation in training set and test set, which may need more sophisticated design of resampling methods to investigate.[14]

As for the percentage of data to drop, we choose to drop 10 percent of the dataset in our benchmark exercise, instead of dropping one single observation (i.e., a country-year pair in our data) as in standard jackknifing. There are a few reasons behind. First, as discussed before, due to the cross-sectional and time-series dependence of our macroeconomics panel data, dropping a single observation cannot generate enough data variation, so we need to drop a larger fraction of data to generate enough data variation from which we are able to draw inference of model performance uncertainty. Second, our first and second jackknifing methods place a lower bound on the percentage of data to drop, because they are designed to drop all years' data for some countries and all countries' data in some years. In our sample of ten emerging market countries spanning almost three decades from 1990 to 2017, dropping all years' data for a single country contribute to dropping 10 percent of the data below which makes it impossible to drop the entire history of one country. As a result, we choose to drop 10 percent of the data for all three jackknifing methods as our benchmark to ensure a fair comparison between the degree of model performance uncertainty arising from three sources of data variation.[15] Additionally, we explore how the degree of model performance uncertainty is affected by the percentage of data dropped. Thus, in a different exercise, we drop 5 percent of the data for all three jackknifing methods. As mentioned before, dropping 5 percent of the data makes it impossible to drop the entire history of one country, so we choose to drop half of the history of one country in this case, randomly chosen to be either the first half (i.e., from 1990 to 2007) or the second half (i.e., from 2008 to 2017).[16] To summarize, we conduct three methods of jackknifing in the following way:

1. Drop countries:
    a. Choose randomly a country.
    b. Drop all years' data for that country.
    c. Repeat (a)-(b) until the percentage of data dropped is larger than the specified value (10 percent or 5 percent).

2. Drop years:
    a. Choose randomly a year.
    b. Drop all countries' data in the year.
    c. Repeat (a)-(b) until the percentage of data dropped is larger than the specified value (10 percent or 15percent).

3. Drop country-year blocks:
    a. Choose randomly a country.

[14] And our results also show that confidence intervals are fairly wide, potentially due to the large degree of model performance uncertainty from data variation in the test set in which there are fewer observations and even fewer crisis events to calculate the percentage of false alarms and missed crises.

[15] It means dropping one country (all years), three years (all countries), or nine blocks of three years from our data.

[16] It means dropping half country (first thirteen years or last fourteen years), one year (all countries), or five blocks of three years from our data.

b. Choose randomly a block of three years for the country.
c. Drop all data in the block.
d. Repeat (a)-(c) until the percentage of data dropped is larger than the specified value (10 percent or 5 percent).

Also, we conduct a fourth type of jackknifing which mimics the standard i.i.d. jackknifing. That is, we repeatedly choose and drop single country-year pair until the percentage of data dropped is larger than the specified value (10 percent or 5 percent). By comparing model performance uncertainty estimated from our proposed three methods of jackknifing with that estimated from the i.i.d. jackknifing, we aim to (1) examine the difference in model performance uncertainty derived from different sources of data variation; and (2) contrast model performance uncertainty estimated from correct ways of resampling with that estimated from incorrect way of resampling, i.e., the i.i.d. jackknifing.[17]

B. Model Estimation and Testing

The model estimation and testing consist of three stages in which different data sets are used: (a) tuning, (b) training, and (c) testing. The data sets used are "validation set" and "its training counterpart", "training set", and "test set", respectively. Simply put, "tuning" is a process conducted before training a model to find optimal hyperparameters of the model. "Training" is the process in which a model is estimated after all hyperparameters are chosen optimally in the tuning stage. In the effort to rank model performance, "testing" is an evaluation process in which all completely estimated models are evaluated based on their out-of-sample performances on the test set(s).

Loss function

We follow the signal evaluation framework in the empirical literature on early warning systems to evaluate model performance by maximizing the signal relative to the noise. Thus, our evaluation metric is defined as the unweighted sum of "false alarms"" (i.e., the percentage of non-crisis observations that the model incorrectly flags as risky) and "missed crises" (the percentage of crisis observations that the model incorrectly flags as safe). Given that crises are rare in all our definitions and datasets, this evaluation metric attaches significantly higher costs to missed crises than to false alarms. It is also used as the loss function in both tuning and training stages to find optimal hyperparameters and estimate models in the training set, as well as calculating optimal thresholds for indicators in signal extraction model. In the testing stage, to generate binary flags on test set(s), we first find the optimal threshold over composite scores in the training set that each model produced by maximizing the signal relative to the noise, and then assign binary labels (i.e., one if a crisis is predicted to follow, zero otherwise) according to that threshold.

Hyperparameter tuning

The most common tuning method in machine learning is k-fold cross-validation introduced by Stone (1974). However, it is not appropriate for our macroeconomic panel data which exhibit

[17] By correct ways of resampling, we mean the ways of resampling that account for the cross-sectional and time-series dependence in our macroeconomic panel data.

cross-sectional and time-series dependencies, because it randomly partitions the training set into non-clustered complementary subsets in which observations do not necessarily come from the same years. Hence, common global factors may be present in both the "validation set" and its training counterpart. Therefore, we make use of the random year-block tuning (Basu et al., 2019). In random year-block tuning, data in training sets are partitioned into ten year-blocks with roughly equal size as validation sets in each of which all observations are from the same years. In contrast to using only historical data for each prediction in the testing, due to the relatively small sample size, for each validation year-block we use all remainder of years as the training counterpart, to avoid depletion of data within the training set. Then the set of hyperparameters are chosen such that average performance over all validation year-blocks is maximized. To perform hyperparameters optimization, we use Bayesian optimization which works well for optimizing hyperparameters of machine learning algorithms.

Different methods have different sets of hyperparameters to be tuned. In random forests, there are three categories of hyperparameters to be tuned: hyperparameters governing tree size, maximum number of surrogate splits the tree finds at each split, and number of features to consider at each split. In standard random forests developed by Breiman (2001), bootstrapping and feature bagging help prevent overfitting, so trees can grow fully, reducing in-sample bias to help reduce out-of-sample bias. However, this approach may not be correct in the context of crisis forecasting when crisis mechanisms constantly change over time. As found in Basu et al. (2019), fully-grown random forests are prone to severe overfitting when applied to crisis forecasting problems, while controlling tree size help reduce overfitting significantly. Therefore, we include parameters which control tree size into the set of hyperparameters to be tuned for all tree-ensemble methods we are using. Among all parameters controlling tree size including minimum observations per leaf, minimum number of observations per parent node and maximum number of splits, we choose to tune minimum observations per leaf to control tree size. Instead of tuning the number of trees, we fix it to be 1000 so that there is a sufficiently large number of trees to reduce variance and thus stabilize prediction performance and variable importance.

Training and testing

We follow the testing procedure in Basu et al. (2019) to make use of two kinds of cutoff tests: a "fixed" cutoff testing and a "rolling" cutoff testing. The algorithm is quite simple: For a given cutoff year, a model is estimated using all information up to that year, and then applied to out-of-sample test sets consisting of all or some of the years after the cutoff year. By limiting the training sets to historical data for each prediction, cutoff tests replicate how the model may be used in real-time analysis, and crucially, they ensure that the testing is conducted fully out-of-sample.

Given the timing of the global financial crisis, we set 2007 as the cutoff year in the "fixed" cutoff testing. A model is estimated using all data up to 2006 to predict crises up to 2007, and is applied to an out-of-sample test set consisting of all data after 2007 to calculate its out-of-sample performance. The fixed-year cutoff testing is simple and stable because there are many crises in the testing set. However, it does not provide an assessment of how model would update and perform after the global financial crisis.

The "rolling" cutoff testing consists of multiple training and testing sets generated by multiple cutoff years: 2007, 2009, 2011, 2013, and 2015. For each cutoff year, a model is estimated using all data up to the year before the cutoff year, and then tested using a two-year test set immediately after the cutoff year. For example, when cutoff year is 2007, a model is estimated using all data up to 2006 and it is tested on a two-year set consisting of data from 2007 and 2008 to predict crises in 2008 and 2009. At the end, models are evaluated based on the average performances over all the two-year test sets.

C. Confidence Intervals

We choose to construct confidence intervals of our preferred evaluation metric, sum of errors, to assess model performance uncertainty and test model ranking significance.

Confidence intervals of individual model performance

We use $\hat{\theta}$ to denote the estimator of model performance obtained from the original sample, which in our design is the out-of-sample sum of errors, either calculated in the "fixed" cutoff exercise, or the average value calculated in the "rolling" cutoff exercise. Then we use $\hat{\theta}_j^*$ to denote the estimator of model performance obtained from the jackknifing samples $j = 1,2,\dots,J$ where $J = 200$. We make use of the estimator obtained from the original sample and the distribution of the estimator obtained from the jackknifing sample to construct the confidence interval developed by Davison and Hinkley (1997). To construct the confidence interval for individual models, we proceed as follows:

1. Order the estimators obtained from the jackknifing sample $\hat{\theta}^*$ such that $\hat{\theta}_1^* < \cdots < \hat{\theta}_J^*$, with subscript denoting the j_{th} element in the ordered list.
2. For a significance level α, select the $\lfloor J \cdot \alpha/2 \rfloor_{th}$ and $\lfloor J \cdot (1-\alpha)/2 \rfloor_{th}$ elements from the above ordered list of estimators, i.e., $\hat{\theta}^*_{\lfloor J\cdot\alpha/2 \rfloor}$ and $\hat{\theta}^*_{\lfloor J\cdot(1-\alpha)/2 \rfloor}$.
3. Construct the two-tailed confidence interval of $\hat{\theta}$ with the significance level α as $\left[2\hat{\theta} - \hat{\theta}^*_{\left\lfloor J\cdot\left(1-\frac{\alpha}{2}\right)\right\rfloor}, 2\hat{\theta} - \hat{\theta}^*_{\left\lfloor J\cdot\frac{\alpha}{2}\right\rfloor}\right]$.

By constructing the confidence intervals for individual models, we can assess to what extent models performed differently if they were estimated on different datasets. However, when we test ranking significance, simply examining whether individual confidence intervals of individual models overlap each other does not provide much evidence on model ranking significance. Because each jackknifing sample represents a possible history of the set of emerging market countries, it is not fair to compare the performance of one model estimated and tested on one history and that of another model estimated and tested on a different history. For example, signal extraction model may perform worse in a world without the crisis of Russia in year 2014 than random forests in a world with the crisis of Russia in year 2014, but this does not mean that signal extraction model cannot perform better than random forests if they are estimated on the same history. Results in Holopainen and Sarlin (2017) also show that there is almost no significant difference among models when comparing individual confidence intervals across models, that is, individual confidence intervals constructed by pooling performance obtained from jackknifing samples are so large that they overlap each other.

Hence, we emphasize that model performance should be compared on the same sample, that is to estimate models on the same training set and compare model performance obtained from the same test set, respectively, no matter whether data variation is introduced or not.

Confidence intervals of conditional performance difference

Therefore, the right way to test model ranking significance when accounting for data variation is to construct the confidence interval of the *conditional* difference in model performance estimators, that is the different in model performance obtained from the same sample. We use $\hat{\theta}_i$ to denote the estimator of model i's performance obtained from the original sample, with $i = 1$ for signal extraction model and $i = 2$ for random forests. Then we use $\hat{\theta}^*_{i,j}$ to denote the estimator of model i's performance obtained from the jackknifing sample $j = 1,2, \dots, J$ where $J = 200$. We proceed as follows:

1. Calculate the difference in performance estimators obtained from the original sample, $\Delta\,\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$.
2. Calculate the estimator differences in performance estimators obtained from the jackknifing sample, $\Delta\,\hat{\theta}^*_j = \hat{\theta}^*_{1,j} - \hat{\theta}^*_{2,j}$ for $j = 1,2, \dots, J$.
3. Order the estimator differences obtained from the jackknifing $\Delta\,\hat{\theta}^*$ such that $\Delta\,\hat{\theta}^*_1 < \cdots < \Delta\,\hat{\theta}^*_J$, with subscript denoting the j_{th} element in the ordered list.
4. For a significance level α, select the $\lfloor J \cdot \alpha/2 \rfloor_{th}$ and $\lfloor J \cdot (1-\alpha)/2 \rfloor_{th}$ elements from the above ordered list of estimators, i.e., $\Delta\,\hat{\theta}^*_{\lfloor J\cdot\alpha/2 \rfloor}$ and $\Delta\,\hat{\theta}^*_{\lfloor J\cdot(1-\alpha)/2 \rfloor}$.
5. Construct the two-tailed confidence interval of $\Delta\,\hat{\theta}$ with the significance level α as $\left[2\,\Delta\,\hat{\theta} - \Delta\,\hat{\theta}^*_{\left\lfloor J\cdot\left(1-\frac{\alpha}{2}\right)\right\rfloor}, 2\,\Delta\,\hat{\theta} - \Delta\,\hat{\theta}^*_{\left\lfloor J\cdot\frac{\alpha}{2}\right\rfloor}\right]$.

Using these confidence intervals of the *conditional* performance difference, we test model ranking significance using a two-sided hypothesis test of the null $H_0: \Delta\,\theta = 0$ and examine whether zero is inside the two-tailed confidence interval with the significance level α. Given that the estimator difference is calculated as the difference between performance estimator of signal extraction model and random forests, if the confidence interval lies on the left side of zero (not including zero), then we reject the null hypothesis and conclude that signal extraction model performs significantly better than random forests at the α significance level. If the confidence interval lies on the right side of zero (not including zero), then we reject the null hypothesis and conclude that random forests perform significantly better than signal extraction model at the α significance level. If zero is inside the confidence interval, then we fail to reject the null hypothesis and conclude that signal extraction model and random forests do not perform significantly different from each other.

V. EMPIRICAL RESULTS

In this section, we present our results of model performance uncertainty and model ranking significance in terms of confidence intervals. (Subsection A) first presents the confidence interval results in fixed cutoff testing exercise and (Subsection B) shows the confidence interval results in rolling cutoff exercise.

A. Fixed Cutoff Testing

In fixed cutoff testing exercise, we choose to drop 10 or 5 percent of the data and perform four jackknifing methods for each. Table 3 and Figure 2 show and plot results for confidence intervals at the significant level of 0.01.

To begin with, we discuss the results of dropping 10 percent of the data. First, we note that most of confidence intervals are wide, especially for signal extraction model. When dropping 10 percent of the data, confidence intervals of signal extraction model have a width greater than 0.35, implying that sum of errors at the 95th percentile is larger than that at the 5th percentile by at least one third of the possible range from 0 to 1. As for random forests, confidence intervals are narrower, but still wider than 0.2 in three jackknifing methods (dropping years, dropping country-year blocks, and dropping country-year pairs (as i.i.d. observations)). Second, for signal extraction model, among all four jackknifing methods, dropping years produce the widest confidence interval which ranges from a sum of error 0.332 to 0.888, covering more than half of the possible range. For random forests, dropping country-year blocks yield the widest confidence interval which has a width almost 0.3. In contrast, dropping countries generate the narrowest confidence interval for random forests, which is about half of all other three. Third, confidence intervals of signal extraction model and random forests overlap in all four jackknifing methods, although the sum of errors of signal extraction model calculated from the original dataset (0.758) is much lower than that of random forests (0.970). However, as discussed in previous section, overlapping confidence intervals of individual models does not necessarily mean there are no significant difference in performance among models. The correct way to proceed is to examine the confidence interval of *conditional* performance difference, that is the difference in sum of errors of different models calculated from the same jackknifing sample. We will discuss the results on this later. Fourth, there seems no much difference between confidence intervals generated by different jackknifing methods including dropping country-year pairs (as i.i.d. observations), except for the wider one generated by dropping years in signal extraction model and the narrower one generated by dropping countries in random forests.

Table 3: Confidence interval results on unconditional performance in fixed cutoff testing

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.695	0.515	0.874
Drop years	0.610	0.332	0.888
Drop country-year blocks	0.701	0.515	0.888
Drop i.i.d.	0.703	0.515	0.890

(a) Signal extraction model, dropping 10 percent of the data

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.894	0.811	0.978
Drop years	0.905	0.762	1.048

Drop country-year blocks	0.904	0.738	1.069
Drop i.i.d.	0.939	0.809	1.069

(b) Random forests, dropping 10 percent of the data

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.690	0.515	0.866
Drop years	0.628	0.430	0.826
Drop country-year blocks	0.706	0.532	0.880
Drop i.i.d.	0.692	0.529	0.855

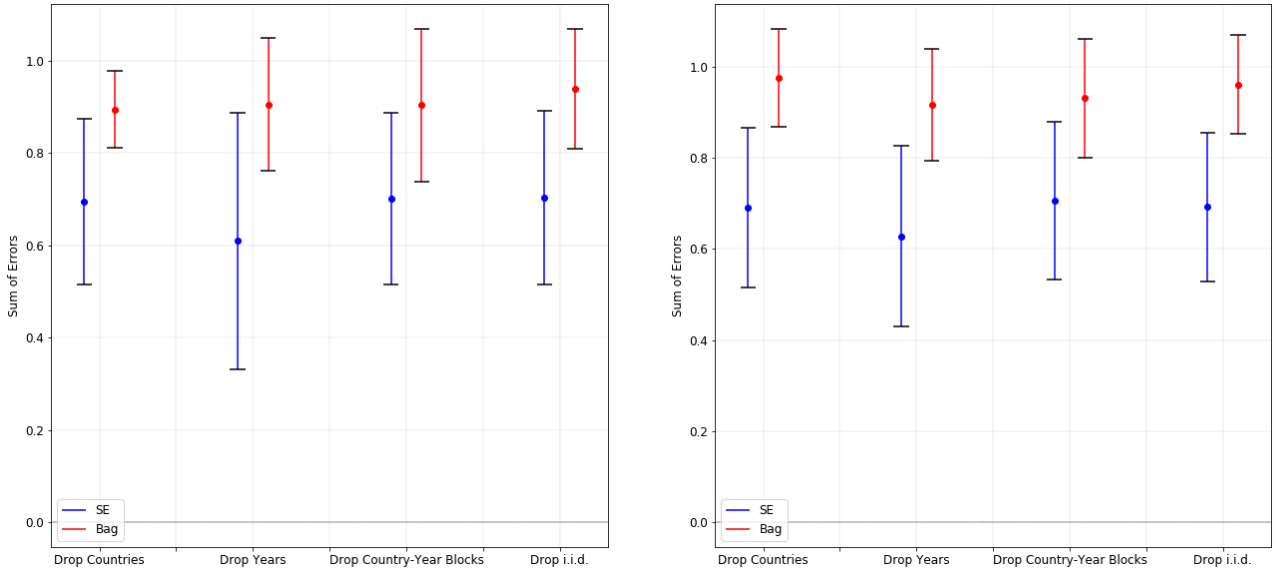
(c) Signal extraction model, dropping 5 percent of the data

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.975	0.867	1.083
Drop years	0.916	0.793	1.039
Drop country-year blocks	0.931	0.801	1.061
Drop i.i.d.	0.961	0.853	1.070

(d) Random forests, dropping 5 percent of the data

When dropping 5 percent of the data, the results do not change much, except that the overlaps between confidence intervals seem to shrink. In the case of dropping countries and dropping country-year pairs (as i.i.d. observations), there are almost no overlaps between confidence intervals of signal extraction model and random forests. It is also worth noting that for signal extraction model, reducing the percentage of data dropped does not shrink confidence intervals much, except for dropping years. Confidence interval generated by dropping years is shrunk by 0.1 when switching from dropping 10 percent to 5 percent of the data, while in all other three jackknifing methods, the reductions in width are small. Combining with the finding that dropping years produce the widest confidence interval for signal extraction model, it may imply that the presence of global regimes and their shifts plays an important role in determining model performance uncertainty for signal extraction model. Because it extracts information by distinguishing between the set of crises and non-crises, whether certain years capturing global regimes that contain certain generations of crises are present in the data may be a determinant factor for its performance variation. In contrast, for random forests, confidence interval generated by dropping country-year blocks is shrunk the most when switching from dropping 10 percent to 5 percent of the data. Combining with the finding to hat dropping country-year blocks yield the widest confidence interval for random forests, it may indicate that the presence of certain individual crisis events plays an important role in determining model performance uncertainty for random forests. Because the recursive partitioning and bootstrap aggregating algorithms make random forests better at accounting for heterogeneity and learning from individual observations, whether certain slices of countries' histories that contain certain crisis observations are present in the data may affect its model performance the most.

Figure 2. Confidence Intervals of Performance in Fixed Cutoff Testing



(a) Drop 10 percent of the data

(b) Drop 5 percent of the data

In order to test model ranking significance for signal extraction model and random forests in fixed cutoff testing accounting for model performance uncertainty, we construct confidence intervals of *conditional* performance different between them, i.e., the difference between sum of error of signal extraction model and random forests calculated from the same sample. Table 4 and Figure 3 show confidence interval results on conditional performance difference generated by four jackknifing methods, dropping 10 percent and 5 percent of the data, respectively. First, confidence intervals of conditional performance difference are also wide. All confidence intervals are wider than 0.4 when dropping 10 percent of the data. Although they are shrunk when switching to dropping 5 percent of the data, the reductions in width are not large. Second and more importantly, these confidence intervals present evidence of significant difference in model performance between signal extraction model and random forests. When dropping 10 percent of the data, all confidence intervals of conditional performance difference lie on the left side of zero, except that the upper bounds of those generated by dropping countries and country-year blocks are slightly larger than zero.

Table 4: Confidence interval results on conditional performance difference in fixed cutoff testing

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	-0.206	-0.450	0.037
Drop years	-0.281	-0.486	-0.075
Drop country-year blocks	-0.213	-0.437	0.010
Drop i.i.d.	-0.268	-0.476	-0.060

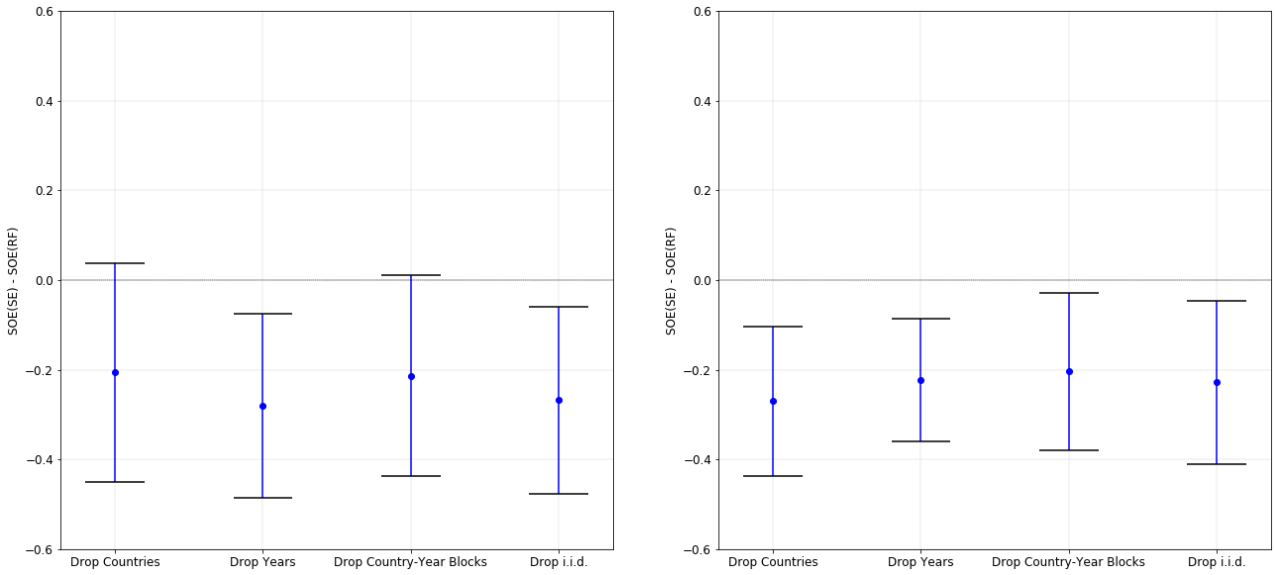
(a) Conditional difference in sum of errors, dropping 10 percent of the data

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	-0.270	-0.437	-0.104
Drop years	-0.224	-0.360	-0.087
Drop country-year blocks	-0.204	-0.379	-0.029
Drop i.i.d.	-0.228	-0.411	-0.046

(b) Conditional difference in sum of errors, dropping 5 percent of the data

These results indicate that in fixed cutoff testing, signal extraction model performs significantly better than random forests at the significance level of 0.01 accounting for three sources of data variation in: (1) the set of countries, (2) the set of years, and (3) the set of countries' histories, in contrast to previous insignificant difference based on comparison between individual confidence intervals. Such different results in testing model ranking significance illustrate the fact we mentioned before: one model may perform better in one version of emerging market history than another model in another version of emerging market history, which makes their confidence intervals overlap. Nevertheless, simply comparing their individual confidence intervals is not the correct way to test their performance difference and may lead to biased insignificant difference. We should always look at the performance difference between models on the same history, and then draw inferences of the difference.

Figure 3. Confidence Intervals of Performance Difference in Fixed Cutoff Testing



(a) Drop 10 percent of the data

(b) Drop 5 percent of the data

When dropping 5 percent of the data, signal extraction model still performs robustly better than random forests when accounting for the three sources of data variation, as all confidence intervals of conditional performance difference lie on the left size of zero. It is also noted that all conditional confidence intervals are narrower when dropping 5 percent of the data. Additionally, different results in testing model ranking significance between comparing

confidence intervals of individual model performance and conditional confidence intervals are observed: Comparing confidence intervals of individual model performance do not reject the null hypothesis when accounting for data variation in the set of years and countries' histories, while examining confidence intervals of conditional performance difference provide evidence of their significant difference.

It is interesting that there is not much difference among confidence intervals generated by different types of jackknifing methods, which essentially account for different types of data variation. There is one potential reason behind such wide and similar confidence intervals. Our approach to jackknifing on the entire dataset encompasses two dimensions of data variation, that in the training set and that in the test set. It could be the case that one dimension of data variation leads to different degrees of model performance uncertainty across the methods, but there is another dimension of data variation that generate larger degree of performance variation which is similar across the methods and dominate the figures, i.e., performance variations derived from the training set variation are different but those derived by the test set variation are huge and similar across different sources. Since there are only six crisis events in the test set of the "fixed" cutoff testing exercise, performance variations derived from the test set variation are very likely to be huge because whether some of the crisis events are included in the test set may alter the percentage of missed crises in a non-smooth way.

B. Rolling Cutoff Testing

We now assess model performance uncertainty and test model ranking significance in "rolling" cutoff testing exercise. Table 5 and Figure 4 show confidence interval results generated by four types of jackknifing methods and dropping 10 percent of the data.

Table 3: Confidence interval results in rolling cutoff testing

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.744	0.506	0.981
Drop years	0.750	0.474	1.027
Drop country-year blocks	0.742	0.520	0.965
Drop i.i.d.			

(a) Signal extraction model, dropping 10 percent of the data

Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI
Drop countries	0.934	0.793	1.074
Drop years	0.913	0.791	1.035
Drop country-year blocks	0.894	0.745	1.043
Drop i.i.d.			

(b) Random forests, dropping 10 percent of the data

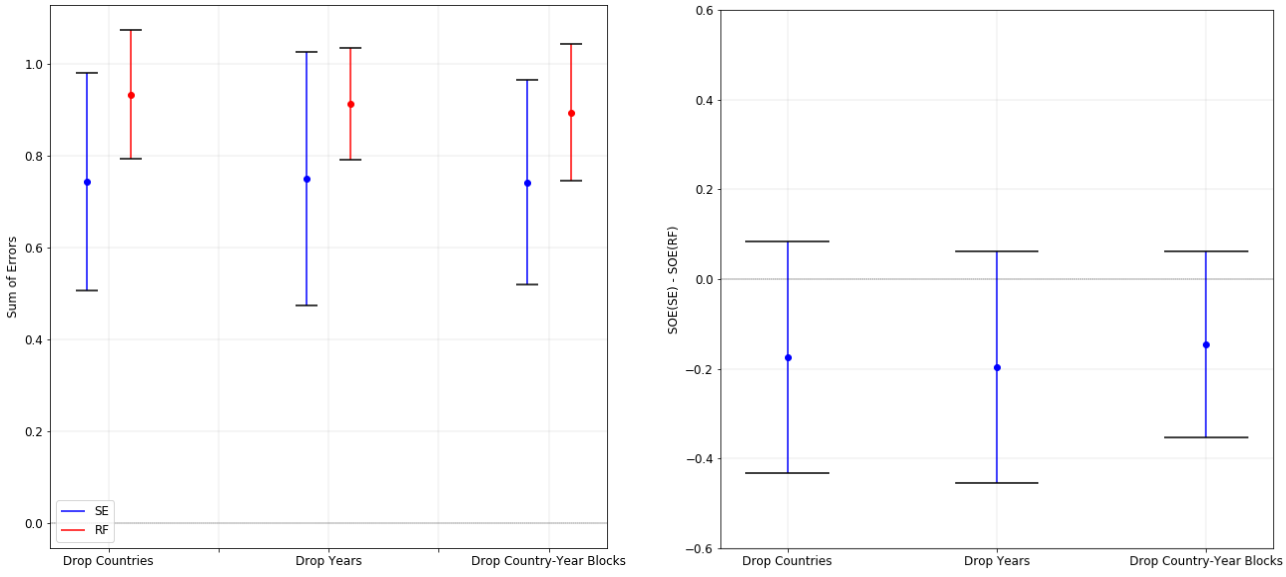
Jackknifing method	Median	Lower bound of 90th CI	Upper bound of 90th CI

Drop countries	-0.174	-0.432	0.084
Drop years	-0.197	-0.456	0.061
Drop country-year blocks	-0.145	-0.352	0.062
Drop i.i.d.			

(c) Conditional performance difference

Similar results as in the "fixed" cutoff testing exercise are seen in the "rolling" cutoff testing exercise. First, confidence intervals of signal extraction model are very wide for all types of jackknifing methods, while those of random forests are much narrower. Second, confidence interval generated by dropping years is the widest for signal extraction model, while confidence interval generated by dropping country-year blocks is the widest for random forests. Consistent with previous results, these results provide evidence of the importance of global regimes for signal extraction model and the importance of individual countries' histories for random forests, potentially due to their model algorithm and structure. Third and most importantly, both comparing confidence intervals of individual model performance and examining confidence intervals of conditional performance difference do not reject the null hypothesis and therefore imply insignificant performance difference between signal extraction model and random forests. Confidence intervals of signal extraction model and random forests overlap for all types of jackknifing methods, and zero is inside in all confidence intervals of conditional performance difference. It implies that when evaluating model performance in such a recursive way, signal extraction model no longer performs significantly better than random forest accounting for any types of data variation. However, although signal extraction model and random forests do not exhibit significant difference in their performance, we will prefer signal extraction model in practical use given its stability and interpretability.

Figure 4. Confidence Intervals in Rolling Cutoff Testing



(a) Confidence intervals of individual models

(b) CIs of conditional performance difference

VI. CONCLUSION

Recent technological advances introduce more complex models (machine learning and deep learning) to the literature on early-warning models and expand the set of early-warning models available for policymakers. Hence, it becomes increasingly important for policymakers to conduct a horse race and rank models in order to select the best one to use in practice. Given the small data nature of macroeconomic data in the early-warning framework, one model that performs best based on past histories may perform worse when new data is available and models are re-estimated. Hence, it is critical to assess the uncertainty in model performance and test the significance in performance difference accounting for data variation. We emphasize the importance of three sources of data variation in macroeconomic data used for early-warning models and propose three types of jackknifing methods to account for these data variations respectively. Then we construct confidence intervals based on model performance calculated from different jackknifing samples to assess model performance uncertainty and perform hypothesis testing to examine whether there is significant difference in performance between models.

Our results show that model performance uncertainty, i.e., the extent to what a model performed different if it was estimated on a different dataset depends on the model structure and the source of data variation. For signal extraction model which looks at the difference between the set of crises and non-crises and does not learn aggressively from individual crisis events, its performance varies the most if there is change in the history of global regimes. For random forest which is designed to tackle heterogeneity and learn aggressively from individual crisis events, its performance varies the most if there is change in individual countries' histories. Also, we show that simply comparing confidence intervals of individual model performance does not provide much evidence on the significance of model performance difference accounting for data variation, and sometimes may lead to incorrect inferences. The correct way to proceed is to construct confidence intervals of the *conditional* performance difference, that is to focus on difference in model performance on the same sample and assess variation in this difference arising from sampling.

As we observed, most of the confidence intervals are wide. Out conjecture is that resampling on the entire dataset does not distinguish between two dimensions of data variation, training set variation and test set variation, and test set variation leads to so large degree of model performance uncertainty that it dominates that derived from training set variation. Hence, for the future, it is worth decomposing overall model performance uncertainty into that derived from training set variation and test set variation and examining them separately.

REFERENCES

Alessi, Lucia, and Carsten Detken. "Identifying excessive credit growth and leverage." *Journal of Financial Stability* 35 (2018): 215-225.

Basu, Suman S., Roberto A. Perrelli, and Weining Xin. "External Crisis Prediction Using Machine Learning: Evidence from Three Decades of Crises Around the World." Paper presented at *Computing in Economics and Finance*, Ottawa, Canada, June 2019. 2019.

Berg, Andrew, Eduardo Borensztein, and Catherine Pattillo. "Assessing early warning systems: how have they worked in practice?." *IMF staff papers* 52, no. 3 (2005): 462-502.

Berg, Andrew, and Catherine Pattillo. "Are currency crises predictable? A test." *IMF Staff papers* 46, no. 2 (1999): 107-138.

Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and regression trees*. CRC press, 1984.

Chamon, Marcos, Paolo Manasse, and Alessandro Prati. "Can we predict the next capital account crisis?." *IMF Staff Papers* 54, no. 2 (2007): 270-305.

Davison, Anthony Christopher, and David Victor Hinkley. *Bootstrap methods and their application*. Vol. 1. Cambridge university press, 1997.

Efron, Bradley, and Charles Stein. "The jackknife estimate of variance." *The Annals of Statistics* (1981): 586-596.

Efron, Bradley. *The jackknife, the bootstrap, and other resampling plans*. Vol. 38. Siam, 1982.

Frankel, Jeffrey A., and Andrew K. Rose. "Currency crashes in emerging markets: An empirical treatment." (1996).

Holopainen, Markus, and Peter Sarlin. "Toward robust early-warning models: A horse race, ensembles and model uncertainty." *Quantitative Finance* 17, no. 12 (2017): 1933–1963.

Kaminsky, Graciela, Saul Lizondo, and Carmen M Reinhart. "Leading indicators of currency crises." *Staff Papers* 45, no. 1 (1998): 1–48.

Lang, Jan Hannes, Tuomas A Peltonen, and Peter Sarlin. "A framework for early-warning modeling with an application to banks," 2018.

Manasse, Paolo, and Nouriel Roubini. ""Rules of thumb" for sovereign debt crises." *Journal of International Economics* 78, no. 2 (2009): 192-205.

Moreno Badia, Marialuz, Franziska Ohnsorge, Pranav Gupta, and Yuan Xiang. "Debt is not Free." (2020).

Savona, Roberto, and Marika Vezzoli. "Fitting and forecasting sovereign defaults using multiple risk signals." *Oxford Bulletin of Economics and Statistics* 77, no. 1 (2015): 66-92.

Sevim, Cuneyt, Asil Oztekin, Ozkan Bali, Serkan Gumus, and Erkam Guresen. "Developing an early warning system to predict currency crises." *European Journal of Operational Research* 237, no. 3 (2014): 1095–1104.

Stone, Mervyn. "Cross-validation and multinomial prediction." *Biometrika* 61, no. 3 (1974): 509-515.

Xu, Lei, Takuji Kinkyo, and Shigeyuki Hamori. "Predicting Currency Crises: A Novel Approach Combining Random Forests and Wavelet Transform." *Journal of Risk and Financial Management* 11, no. 4 (2018): 86.