



WP/04/52

IMF Working Paper

Assessing Early Warning Systems: How Have They Worked in Practice?

*Andrew Berg, Eduardo Borensztein, and
Catherine Pattillo*

IMF Working Paper

Research Department

Assessing Early Warning Systems: How Have They Worked in Practice?

Prepared by Andrew Berg, Eduardo Borensztein, and Catherine Pattillo¹

March 2004

Abstract

This Working Paper should not be reported as representing the views of the IMF.

The views expressed in this Working Paper are those of the author(s) and do not necessarily represent those of the IMF or IMF policy. Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate.

Since 1999, the IMF's staff has been tracking several early-warning-system (EWS) models of currency crisis. The results have been mixed. One of the long-horizon models has performed well relative to pure guesswork and to available non-model-based forecasts, such as agency ratings and private analysts' currency crisis risk scores. The data do not speak clearly on the other long-horizon EWS model. The two short-horizon private sector models generally performed poorly.

JEL Classification Numbers: F31, F47

Keywords: Currency crises, vulnerability indicators, crisis prediction, forecasting accuracy, balance of payments crisis

Authors' E-Mail Addresses: aberg@imf.org; eborensztein@imf.org; cpattillo@imf.org

¹ This paper was originally prepared for the IMF Research Department conference on early-warning systems, November 2001. We would like to thank conference participants and many IMF staff, including Paul Cashin and Robert Rennhack for useful comments, and Manzoor Gill for superb research assistance.

Contents	Page
I. Introduction.....	3
II. Early-Warning-System (EWS) Models	4
III. Value Added of EWS Models.....	8
A. EWS Models and Alternative Indicators in Asia Crisis.....	8
B. How Well Have the EWS Models Done in Practice Since Their Implementation in January 1999?	13
IV. Summary and Conclusions	29
Appendixes	
I. Description of EWS Models and Specification Issues	32
II Meaning of In-Sample and Out-of-Sample Periods in EWS Models	40
References.....	42
Boxes	
1. Crisis Definition.....	7
2. Goodness-of-Fit Measures and Trade-Offs	19
Text Tables	
1. Specification of Early-Warning Models	6
2. Risk Assessments Prior to Asia Crisis Based on KLR, DCSD, Bond Spread, Credit Rating, and Economist Intelligence Unit Forecasts	11
3. Crisis Probabilities According to Different Models, as of July 1999	14
4. Goodness of Fit: DCSD and KLR Models	18
5. Goodness of Fit: Short-Horizon Models.....	26
Appendix Tables	
6 Crisis Dates According to Different Models	37
7. Model Samples.....	41
Text Figures	
1. Korea: Forward Exchange Rate Premium, Bond Spread, and Probability of an Exchange Rate Crisis Estimated by DCSD Model, 1996–2000	12
2. DCSD and KLR In-Sample Forecasts	20
3. DCSD Model Forecasts	24
4. KLR Model Forecasts	25
5. GS Model Forecasts.....	27
6. CSFB Model Forecasts	28

I. INTRODUCTION

Research on developing early-warning-system (EWS) models of currency crisis received a strong stimulus in recent years following the Mexican and Asian crises. Both events took the international community somewhat by surprise and thus focused attention on indicators and methods that could assist in the timely identification of highly vulnerable countries. Since the beginning of 1999, the IMF's staff has been systematically tracking various models, both inhouse and from private institutions, on an ongoing basis, as part of a broader, forward-looking vulnerability assessment.

EWS models play a necessarily small part in vulnerability assessment relative to more detailed country-specific analyses. We do not review the IMF's experience with the broader vulnerability assessment process here.² The question we are interested in, and to which we return in the conclusion, is whether EWS models have any role to play at all.³

This paper looks in detail at the performance of these models in practice to date. We emphasize the distinction between in-sample and out-of-sample prediction. For an EWS model to be a useful tool in monitoring vulnerabilities, it must hold up in "real time" after the model has been formulated. We thus focus on the actual forecasts made since 1999, though we also reexamine the run-up to the Asia crisis.

A typical result from our earlier studies was that, although the model forecasts we reexamine here were statistically significant predictors of crises, whether they were economically significant is harder to say. In other words, the forecasts were informative compared with a benchmark of complete random guessing, but it was less clear whether they were useful to an already-informed observer. It is reasonable to suppose that comprehensive, country-specific, and holistic assessments by informed analysts, based on all available qualitative and quantitative information, must be better than inevitably simple EWS models. Indeed, the ability to take all the information into account is clearly a huge potential advantage. But there have been no studies on whether such comprehensive assessments are, in fact, better. We gain perspective on this issue here by comparing EWS model forecasts to non-model-based indicators such as bond spreads, agency ratings, and risk scores published by analysts.

² Some discussion of the role of EWS models in the broader vulnerability assessment can be found in IMF (2002) and in Berg and others (1999).

³ Predicting currency crises is closely related to predicting exchange rate movements, so any success of EWS models is notable in the context of the long literature, starting with Meese and Rogoff (1983), that has shown how difficult it is to predict exchange rates out of sample. Berg and others (1999) discuss this point further in the EWS context.

II. EARLY-WARNING-SYSTEM (EWS) MODELS

Policy initiatives to monitor indicators of external vulnerability can be traced to the Mexican peso crisis of December 1994. A seminal effort to use a systematic quantitative early-warning-system to predict currency crises was the “indicators” model of Kaminsky, Lizondo and Reinhart (1998). The Asia crises of 1997/1998 provided further impetus to the effort. Evidence suggested that despite the daunting challenges involved in this sort of exercise, this sort of model had some success in predicting these crises out of sample (Berg and Pattillo 1999a). It also suggested that a variety of improvements could substantially improve model performance (Berg and Pattillo 1999b).

In light of this research, IMF staff has implemented various models to predict currency and balance of payments crises since 1999, as described in Berg and others (1999). The IMF’s staff has also tracked various private sector models, including Goldman Sachs’ GS-WATCH and Credit Suisse First Boston’s (CSFB’s) Emerging Markets Risk Indicator, and a more recent model, the Deutsche Bank Alarm Clock (DBAC). Table 1 summarizes the main features of the models under consideration here. It details the crisis definition employed, the prediction horizon, the method used to generate predictions, and the predictor variables. Appendix I contains a more complete description of the models. Table 6, in Appendix I, shows all the crisis dates for these models since January 1999.⁴

As Table 1 illustrates, the specification of EWS models involves a number of decisions that, while guided in some way by economic theory, are largely empirical and judgmental in nature. Currency crises, for example, are not precisely defined events, but the models must nonetheless define crisis dates in terms of the data. (See Box 1 for a discussion of how the different models implement the currency crisis concept.) The choice of a prediction horizon depends on the objectives of the user. The in-house models adopt a relatively long horizon, which should allow time for policy changes that may prevent the crisis. The time horizon of private sector models is shorter and their criterion for evaluation of the accuracy of predictions (frequently a trading rule) is sometimes different. Nevertheless, it is still informative to consider the predictions from private sector EWS models when assessing vulnerability, not least for the fact that those predictions are disseminated widely within the investor community.

⁴ More recently, other models have been developed in the IMF, such as Mulder, Perrelli, and Rocha (2002) and Abiad (2003). Other recent developments include models designed to predict other sorts of crises, Manasse, Roubini, and Schimmelpfennig (2003). For a review of recent developments in this literature, see Abiad (2003).

The predictive variables in the models are inspired by theories of balance of payments crises but constrained by data availability, but in the end reflect what works best in fitting the data. The statistical method chosen again is an essentially empirical decision. Appendix I discusses considerations that apply to these specification choices with special reference to the models tracked in EWS exercise in the Fund.

Table 1. Specification of Early Warning Models

	DCSD ^{1/} (Berg, Borensztein, Milesi-Ferretti, and Pattillo)	KLR ^{2/} (Kaminsky, Lizondo, and Reinhart)	GS—WATCH ^{3/} (Goldman Sachs)	EMRI ^{4/} (Credit Suisse First Boston)	DB Alarm Clock ^{5/} (Deutsche Bank)
Crisis Definition	Weighted average of one-month changes in exchange rate and reserves more than 3 (country-specific) standard deviations above country average	Same as DCSD	Weighted average of three-month changes in exchange rate and reserves above country-specific threshold	Depreciation > 5% and at least double of preceding month's	Various "trigger points" Depreciation > 10% and Interest rate increase > 25% typical
Horizon	2 years	2 years	3 months	1 month	1 month
Method	Probit regression with rhs variables measured in (country specific) percentile terms	Weighted (by frequency of correct predictions) average of indicators. Variables measured as 0/1 indicators according to threshold chosen to minimize noise/signal ratio	Logit regression with (most) rhs variables measured as 0/1 indicators based on thresholds found in autoregression with dummy (SETAR)	Logit regression with rhs variables measured in logs, then deviation from mean and standardized	Logit two-equation simultaneous systems on ex rate and i rate "events" of different magnitude
Variables	Overvaluation Current account Reserve Losses Export Growth ST Debt/Reserves	Overvaluation Current Account ^{6/} Reserve Losses Export Growth Reserves/M2 (level) ^{6/} Reserves/M2 (growth) Domestic Credit Growth Money Multiplier Change Real Interest Rate Excess M1 Balances	Overvaluation Export Growth Reserves/M2 (level) Financing Requirement Stock Market Political Event Global Liquidity Contagion	Overvaluation Debt/Exports Growth of credit to private sector Reserves/Imports (level) Oil prices Stock price growth GDP growth Regional Contagion	Overvaluation Industrial Production Domestic Credit Growth Stock Market Devaluation Contagion Market Pressure Contagion Regional Dummies Interest Rate "event"

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ KLR: Kaminsky, Lizondo, and Reinhart (1998).

3/ Goldman Sachs – Ales, Masih, and Tenengauzer (1998).

4/ Credit Suisse First Boston—Roy and Tudela (2000).

5/ Deutsche Bank – Garber, Lumsdaine, and van der Leij (2000).

6/ Not included in the original KLR model.

Box 1. Crisis Definition

The EWS models considered in this paper attempt to predict currency crises, as distinct from other sorts of crises such as debt and banking crises. Although opinions differ as to what constitutes a currency crisis and when one is observed, the formulation of an EWS model requires a specific quantitative definition (Table 1 briefly describes the crisis definitions for the models discussed in detail in this paper. Table 6 lists all the crisis dates for the various models for the 1999–2001 period (1999–2002 for DCSD ^{1/} and KLR ^{2/})).

Models that attempt to predict only successful speculative attacks, such as that of Credit Suisse First Boston (CSFB), define crises solely by sufficiently large changes in the exchange rate over a short period of time. For a private sector institution, predicting sudden large changes in the exchange rate alone may be the most interesting objective. EWS models implemented in the Fund, as well as the Goldman-Sachs model, attempt to predict both successful and unsuccessful speculative attacks by calculating an “exchange market pressure” index that combines exchange rate and reserve changes.

For the CSFB model, a crisis occurs when the exchange rate moves by more than some threshold amount over a short period of time (see Table 1). For the in-house models, a crisis occurs when the exchange market pressure index is very high relative to its historical average.^{3/} The GS model also defines a crisis as an index value that is high relative to a country-specific threshold, with the threshold defined so as to separate calm periods from those of unusual volatility. The GS definition tends to produce more frequent crises, often occurring for several consecutive periods, while the DCSD/KLR crises are almost always isolated. Comparing over a common sample of 16 countries from 1996 through mid-2001, the CSFB crisis definition produces 34 crises, the DCSD/KLR definition produces 34 crises, and the GS definition 150 crises, with the latter grouped into 47 distinct episodes of one or more consecutive crisis months.

The omission of interest rates in the crisis definitions for most emerging market EWSs, due to poor availability of historical data on market-determined interest rates, is increasingly recognized as a shortcoming in identifying crises. For example, the 1995 attack on the Argentinean peso in the wake of the “tequila” effect is an event that the models should attempt to predict. However, this failed attack, which was evidenced mainly by the rapid increase in domestic interest rates, is not identified as a crisis by many EWS models. The Deutsche Bank Alarm Clock (DBAC) defines substantial exchange rate depreciations and interest rate increases as separate events but jointly estimates the probability of these two types of events. However, the model uses IFS data on money market interest rates, which are deficient for many emerging economies.

It is unlikely that any simple formula, however well thought out, will always be successful in picking out crisis periods in the data. One possible improvement is to combine the results from the quantitative definition with country-specific knowledge about exchange market developments to make some adjustments to the dating of crisis periods. Events that are “close calls” according to the crisis formula could receive particular scrutiny, and the analyst may use judgment to label some of these as crises. For example, Sri Lanka suffered a reserve loss of about 40 percent during 2000 along with a currency depreciation of nearly 15 percent, culminating with the abandonment of the crawling band exchange rate regime and further currency depreciation in January 2001. Because no single month was sufficiently traumatic, however, the formula employed in the KLR and DCSD models registered a fairly close call, but not a crisis. This episode might be called a crisis if assessed “by hand.”

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ KLR: Kaminsky, Lizondo, and Reinhart (1998).

3/ Taking into account whether the crisis index is high relative to its history in a particular country has an advantage relative to defining a crisis by the same absolute cutoff depreciation for all countries. For example, for a country that has had a pegged exchange rate regime and the rate has remained fixed for some time, a relatively small devaluation may be considered a crisis. The same size depreciation might not constitute a crisis in a country where the exchange rate is flexible and has been more volatile.

III. VALUE ADDED OF EWS MODELS

Since early 1999 the Fund has been regularly producing forecasts from two EWS models, Kaminsky, Lizondo, and Reinhart (KLR) and Developing Country Studies Division (DCSD), and monitoring two private sector models, GS and CSFB. This section examines the usefulness of these models in providing early warnings of crises.

The evaluation of EWS models requires a benchmark. Typically the question is whether a model provides a statistically and economically significant prediction of crisis. This might appear to be a low standard to meet. Given the difficulties involved in crisis predictions, however, it is ambitious. To forecast crises reliably implies systematically outperforming the market at predicting sudden changes in the exchange rate.

Assessments must focus on out-of-sample performance. Successful in-sample predictions are much easier to achieve than out-of-sample predictions but much less meaningful. First, the diligent analyst may have searched through enough truly unrelated variables until some are found that, by coincidence, happen to be correlated with crises in this particular sample. Such a spurious relationship is not likely to persist in a new sample. Second, even if a true relationship is found in the sample, the next set of crises may be fundamentally different from the last. A model that provides accurate out-of-sample forecasts has thus passed a much tougher test. Of course, only such models are useful for actual crisis prediction. For these reasons, the focus here is on out-of-sample testing.

In this section, several approaches to testing these models are pursued. First, the stage is set by a review of the performance of a model designed prior to the Asia crises, KLR, in predicting those crises. Following earlier work, the model is implemented as it might have been in early 1997 and forecasts are compared to actual outcomes. These predictions are compared to those implied by spreads on dollar-denominated sovereign bonds, sovereign credit ratings, and also the assessments of currency crisis risks produced by country experts from the Economist Intelligence Unit (EIU). These results suggest that the EWS models show promise. The model decisively outperforms all the comparators in this period.

The second part of the section, and the core of the paper, looks at how the various models that have been monitored at the Fund since early 1999 have performed in this period. First is a detailed analysis of the first set of forecasts “officially” produced within the Fund, in May 1999. These are compared to the alternative indicators, such as bond spreads, ratings and analysts’ views described above, where possible. We follow with a more systematic examination of how well the models have predicted crises over the full out-of-sample period.

A. EWS Models and Alternative Indicators in Asia Crisis

Berg and others (1999) looked at various measures of performance of a variety of EWS models, focusing in particular on their ability to predict the Asia crises of 1997–1998 out of

sample.⁵ One main conclusion was that the original KLR model, which was designed prior to 1997 and hence without the benefit of hindsight, had substantial predictive power over the Asian episodes. Column 1 of Table 2 shows the ranking of countries according to the risk of currency crisis that KLR would have produced in early 1997. These forecasts are fairly good, with many of the most vulnerable countries in fact being the hardest hit in terms of crisis severity.⁶ For example, Korea and Thailand were among the top third of countries in terms of vulnerability according to the KLR model. Although Brazil and Philippines, which were not particularly hard hit over this period, were at the top of the vulnerability table, the forecasts are informative overall. Country rank in the predicted vulnerability list is a statistically significant predictor of actual crisis incidence.

One lesson from Berg and others (1999) and other work is that there was clear scope for improvement of those earlier models. A variety of potentially important crisis indicators had not been tested, such as the current account deficit as a share of GDP and the ratio of short-term external debt to GDP. Moreover, regression-based estimation techniques that more fully exploit the information in the data seemed a promising alternative to the “indicator” based method of the KLR model. A revamped KLR-based model and the DCSD model described in section II were the result of an effort to improve upon the original KLR model. Not surprisingly, given the benefit of hindsight, these models perform substantially better in predicting the Asia crises (Column 2 of Table 2 presents results for the DCSD model).

The predictions of EWS models showed themselves to be significantly better than random guesses in predicting the Asia crises, but they were not overwhelmingly accurate. How, in comparison, did the various non-model based indicators fare over this period? Among these, **sovereign spreads** are a commonly watched indicator of country risk. While the spreads are important indicators of market access, and also market sentiment, they do not fare particularly well as currency crisis predictors over this period. The most affected countries had generally *lower* pre-crisis spreads as of Q1 1997, as shown in the third column of Table 2. The spread averaged 90 basis points in the countries that subsequently suffered a crisis, while it averaged 201 in the other countries.⁷

Second, there is some evidence that **sovereign ratings** from agencies such as Moody’s and Standard and Poor’s have also been poor predictors of recent currency crises⁸. The fourth

⁵ See also Berg and Pattillo (1999c) on the implications of EWS models for the Asia crisis.

⁶ The measure of the severity of crisis for a particular country is the maximum value reached by the exchange market pressure index within 1997, where the index itself is a weighted average of the depreciation of the exchange rate and the loss of international reserves.

⁷ Although one could rationalize the low sovereign spreads in the Asian economies on the basis of their relatively low levels of external debt, spreads did increase after October 1997, suggesting that markets may have underestimated risks. The period before the Asian crises was characterized by unusually low spreads for almost all emerging market economies.

column of Table 2 shows the sovereign ratings as of the first quarter of 1997, based on a quantitative conversion of Moody's and S&P's ratings where higher numbers correspond to a better rating. Korea, Malaysia and Thailand are the highest-rated countries while Mexico is relatively poorly ranked. Indeed, the average rating in the ten most affected countries was substantially *better* than in the ten least affected countries.

Another, non-model-based predictions of currency crises are **surveys of currency market analysts**, such as those prepared by The Economist Intelligence Unit (EIU). EIU has regularly produced estimates of currency crisis risk, defined as the risk of a 20 percent real depreciation of the currency over the two-year forecast horizon.⁹ These estimates derive from the analysis of country experts, who consider a broad set of quantitative and qualitative factors, ranging from macroeconomic and financial variables to the strength of the banking system, the quality of economic decision-making, and the stability of the political situation. The estimates are available for a large number of countries since 1996. As column 5 of Table 2 shows, these Economist Intelligence Unit forecasts gave generally positive assessments to the Asian economies that were about to suffer from severe episodes. Indeed, countries with higher risk scores in the second quarter of 1997 were systematically *less* likely to have a crisis during the 1997–1998 period.¹⁰

Along similar lines, there exist surveys of estimates of future exchange rate changes by foreign exchange traders and specialists in financial institutions and multinational corporations. Goldfajn and Valdés (1998) examined the surveys by the *Financial Times Currency Forecaster* and found that such market participants' expectations provided no useful early warning of currency crises in a large sample of emerging markets or in important cases such as Mexico in 1994 or Thailand in 1997.

To summarize, there is little evidence that “market views,” or analysts' views, as expressed in spreads, ratings and surveys are reliable crisis predictors, important as they may be in determining market access. This conclusion is illustrated for the important case of Korea in Figure 1, which shows DCSD model predictions as well as various other indicators of crisis risk.

⁸ See Sy (2003) and Reinhart (2002). As with sovereign spreads, it could be argued that these ratings are designed to predict default, not currency crisis. Against this, however, is the fact that currency crises do increase the risk of default and that, because of this, ratings have in fact been downgraded after most currency crises. This suggests that the rating agencies would have likely downgraded the countries had they seen the currency crises coming.

⁹ See Economist Intelligence Unit, *Currency Risk Handbook*, June 2001.

¹⁰ The Economist Intelligence Unit forecasts are similarly unsuccessful when compared to crises as defined by the EIU itself.

Table 2. Risk Assessments Prior to Asia Crisis

Based on KLR ^{1/}, DCSD ^{2/}, Bond Spread, Credit Rating, and Economist Intelligence Unit Forecasts

Country ^{3/}	KLR ^{4/}	DCSD ^{5/}	Spread ^{6/} 1997 Q1	Rating ^{7/} 1997 Q1	EIU 1997 Q1 Currency Risk ^{8/}
Korea	22	24	50	18	22
Thailand	20	40	51	25	42
Indonesia	16	32	109	43	38
Malaysia	14	39	37	20	36
Zimbabwe	19	na	na	na	58
Philippines	34	14	165	55	36
Taiwan Province of China	23	46	na	na	12
Colombia	15	41	129	45	35
India	10	21	na	Na	35
Brazil	31	15	233	65	51
Turkey	16	18	416	66	56
Venezuela	14	9	na	Na	53
Pakistan	20	36	na	68	49
South Africa	19	26	85	48	39
Jordan	14	15	na	Na	61
Sri Lanka	12	17	na	Na	43
Chile	11	14	na	Na	17
Bolivia	10	5	na	Na	37
Argentina	14	11	265	63	59
Mexico	14	8	231	55	55
Peru	20	26	na	70	51
Uruguay	10	14	135	50	37
Israel	14	24	44	30	46
Average					
Crisis Countries	20	34	90	34	35
Non-Crisis Countries	15	17	201	57	46
Rank Correlation ^{9/}	0.52	0.53	-0.31	-0.49	-0.33

Sources: Kaminsky, Lizondo, and Reinhart (1998); Berg and others (1999); Economist Intelligence Unit; Standard and Poor's; and Moody's.

1/ Kaminsky, Lizondo, and Reinhart (1998).

2/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

3/ Countries that suffered a crisis in 1997 are in bold. The countries are ordered by severity of crisis.

4/ Probabilities of currency crisis over 24 months horizon, from average KLR model for 1996 as implemented by the IMF's Research Department.

5/ Probabilities of currency crisis over 24 months horizon, from average of 1996 DCSD results.

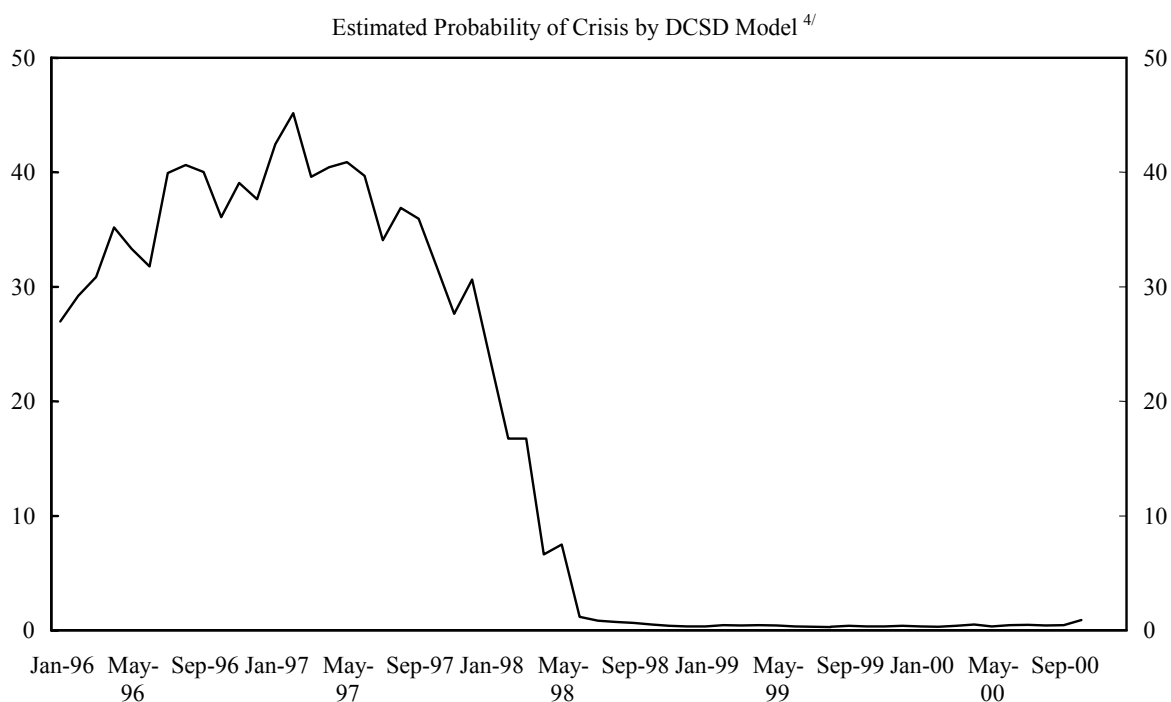
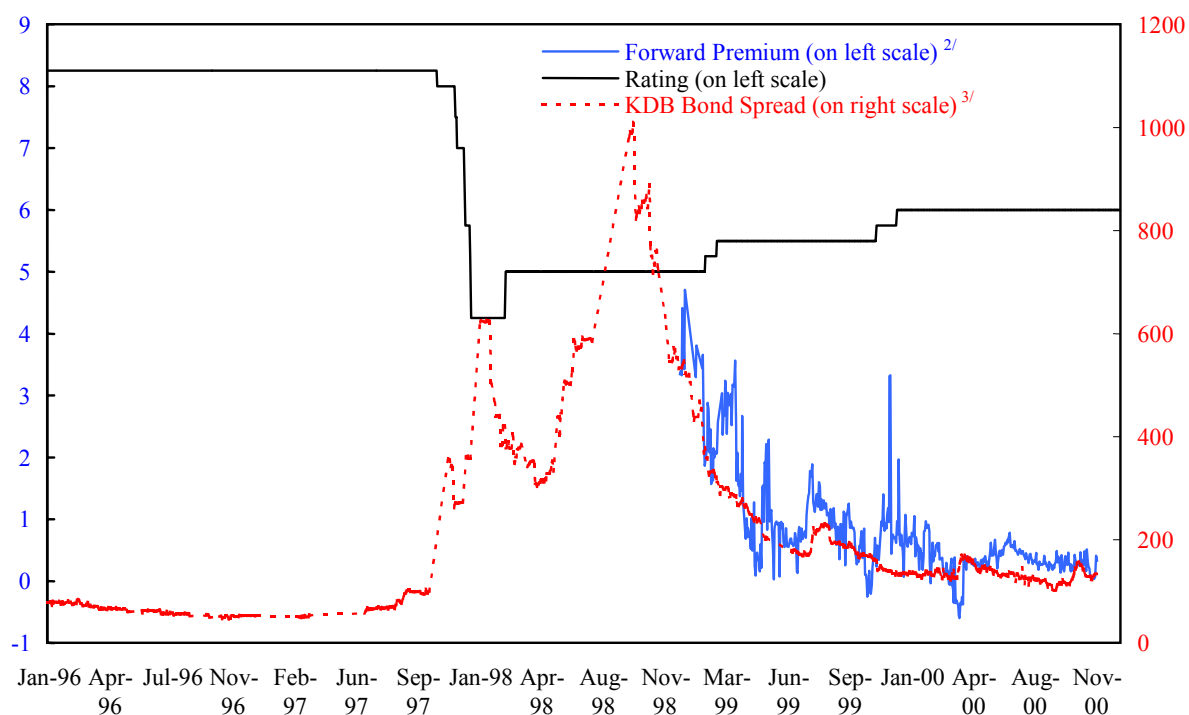
6/ The spread is expressed in basis points. It refers to the difference between the yield on US dollar denominated Foreign government eurobond and the equivalent maturity US treasury bond.

7/ Average of S&P and Moody's ratings, each converted to a numerical rating ranging from 100 (S&P SD) to 0 (S&P AAA or Moody's Aaa), following Ferri, Liu and Stiglitz (1999). A lower number means a better rating (unlike Ferri, Liu, and Stiglitz).

8/ Currency risk: "Scores and ratings assess the risk of a devaluation against the dollar of 20% or more in real terms over the two year forecast period", following the Economist Intelligence Unit (EIU).

9/ Countries are ranked according to each indicator as well as according to crisis severity (in both cases a lower number implies a worse actual or predicted crisis). The rank correlation relates these two rankings.

Figure 1. Korea: Forward Exchange Rate Premium, Bond Spread, and Probability of an Exchange Rate Crisis Estimated by DCSD Model, 1996–2000 ^{1/} (Through November 20, 2000)



Sources: Bloomberg, J.P. Morgan; and IMF staff estimates.

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ The forward premium is defined as the log of the ratio of the 12-month forward to the spot exchange rate.

3/ Spread of KDB Eurobond over comparable US Treasuries

4/ The probability of a crisis over the next 24 months estimated using the DCSD model.

The results of this round of EWS testing were sufficiently promising to suggest the continued implementation of these models on an on-going basis, along with further research and development. The in-sample results were fairly good. More remarkably, the out-of-sample results were also promising. Here, however, the results of the DCSD model results must be discounted, as this model benefited from hindsight in its formulation, though only pre-Asia-crisis data was used to produce the forecasts.¹¹ Moreover, it may be that this good Asia-crisis performance was just lucky, or alternatively that subsequent crisis episodes may have been sufficiently different that these models stopped working. This suggests an examination of the recent performance of the EWS models.

B. How Well Have the EWS Models Done in Practice Since Their Implementation in January 1999?

We examine this question in two ways. First, we look at the results of the models the first time the forecasts were produced “officially” in a forward-looking exercise, in July 1999.¹² Second, more systemic measures of the “goodness of fit” of the models are examined, emphasizing the comparison of in-sample and out-of-sample model performance, and the tradeoff between missing crisis and generating false alarms.

July 1999 forecasts

Table 3 shows the predicted probabilities of crisis for the DCSD and KLR models.¹³ Countries that suffered crises are in bold, with the dates of the crisis noted. The KLR and particularly the DCSD model did fairly well. Both the countries with DCSD probabilities of crisis above 50 percent subsequently had crises, and no crisis country had a probability below 26 percent. Using the models own definition, there have been only three crises in the roughly two years since July 1999.

As before, it is useful to compare these forecasts to other indicators and estimates. Columns 3, 4 and 5 of Table 3 show spreads on dollar-denominated bonds, sovereign ratings, and the

¹¹ The main benefit from this hindsight was the inclusion of short-term-debt/reserves as a predictive variable. The original KLR model had focused on M2/reserves instead. This latter variable also works, though not as well.

¹² Appendix II explains how in-sample and out-of-sample periods are determined for each of the models considered.

¹³ The two private sector models monitored at that time, GS and CSFB, forecast only over a one to three month horizon. The “snapshot” of the first “official” July 1999 results is thus not informative. Their performance is examined in the discussion of overall goodness of fit, below.

Table 3. Crisis Probabilities According to Different Models, as of July 1999

Country ^{1/}	KLR ^{2/}	DCSD ^{3/}	Spread ^{4/}	Ratings ^{5/}	Economist Intelligence Unit ^{6/}
Colombia (Aug. 99)	42	61	544	45	42
Turkey (Feb. 01)	45	50	554	68	58
Zimbabwe (Aug. 00)	24	26	na	na	77
Bolivia	na	36	na	na	42
Chile	11	36	na	na	29
Venezuela	42	34	na	na	58
Argentina	20	31	471	58	62
Peru	32	26	210	58	35
Uruguay	32	23	216	45	36
Brazil	24	21	451	68	47
Mexico	14	19	296	55	42
Pakistan	na	14	2270	90	69
Jordan	14	14	n.a.	n.a.	34
South Africa	32	9	141	48	40
India	11	9	na	na	38
Sri Lanka	na	7	na	na	51
Israel	11	6	78	30	30
Thailand	32	4	192	48	41
Philippines	14	3	401	50	28
Malaysia	11	3	174	45	36
Indonesia	32	1	872	78	50
Korea	24	1	238	45	30
Average					
Crisis Countries	37	46	549	56	59
Non-Crisis Countries	22	16	462	55	42

Sources: Kaminsky, Lizondo, and Reinhart (1998); Berg and others (1999); Economist Intelligence Unit; Standard and Poor's; and Moody's.

1/ Countries with crises between June 1999 and June 2001 in bold.

2/ Probabilities of currency over 24 months horizon, from KLR model. Estimated using data through March 1999 except for Brazil, Jordan, Korea, Mexico, Venezuela, and Zimbabwe (December 1998), Chile and Israel (May 1999), India and Indonesia (January 1999), Malaysia (April 1999), South Africa (February 1999), and Turkey (November 1998).

3/ Probabilities of currency crisis over 1 month horizon (annualized), from DCSD model.

4/ The spread is expressed in basis points. It refers to the difference between the yield on US dollar denominated foreign government eurobond and the equivalent maturity US treasury bond.

5/ Average of S&P and Moody's ratings, each converted to a numerical rating ranging from 100 (S&P SD) to 0 (S&P AAA or Moody's AAA), following Ferri, Liu, and Stiglitz (1999). A lower number means a better rating (unlike Ferri, Liu and Stiglitz).

6/ Currency risk: "Scores and ratings assess the risk of a devaluation against the dollar of 20 percent or more in real term over the two year forecast period", following Economist Intelligence Unit (2000).

Economist Intelligence Unit's currency crisis risk scores, all as of the second quarter of 1999. The alternative predictors fared better than prior to the Asia crises but still not well. Spreads are moderately higher at 549 for the three crisis countries compared to 462 for the others. The average sovereign rating is 56 for the three crisis countries while it is slightly worse at 55 for the rest of the countries. The Economist Intelligence Unit estimates, in contrast, are substantially better here than they were before the Asia crises. For example, the average risk score of the three crisis countries was 59, compared to 42 for the others.¹⁴

The improved performance of the non-model-based indicators compared to before the Asia crisis, combined with the low incidence of crises over the out-of-sample period, suggests that the challenge for the models over this period was more to avoid a large number of false alarms than to call otherwise unforeseen crises.

The distinction between sovereign, or default, risk and currency crisis risk surely plays an important role in explaining the performance of ratings and spreads in some important recent cases—a role that it did not play in the Asia crises. Colombia's crisis in August 1999 involved a drop in the exchange rate as the country abandoned a crawling band exchange rate regime, but there was little subsequent concern about sovereign default. Thus, it is perhaps not surprising that the ratings and spreads do not predict this incident. Conversely, Pakistan suffered a debt crisis but had no currency crisis over the period, and its exceedingly high spreads, that started to widen since the economic sanctions following the nuclear tests in 1998, greatly increase the non-crisis-country average.¹⁵

Overall goodness-of-fit since January 1999

A look at the goodness-of-fit of the models over the entire out-of-sample period provides a more systematic assessment of the models (see Box 2 for details on these measures). The computation of goodness-of-fit measures requires selecting a cut-off probability value, above which the prediction is classified as an "alarm," implying that the model expects a crisis to ensue at some point along the prediction horizon. The threshold probability for an alarm can be chosen to minimize a weighted sum of the share of false alarms and the share of missed crises, where the user must decide how much weight to place on missed crises vs. false alarms. Throughout this chapter, equal weight is placed on the share of alarms that are false and the share of crises that are missed.¹⁶

¹⁴ Using the Economist Intelligence Unit's own crisis definition, the forecasts perform somewhat worse, with the average risk for the crisis countries below the average for the non-crisis countries.

¹⁵ Excluding Pakistan, the average spread for non-crisis countries declines to 312 from 462.

¹⁶ The relative weights depend implicitly on the cost imputed to each type of error. Demirguc-Kunt and Detragiache (1999), who employ a similar loss-function approach in
(continued...)

A first result is that the in-sample goodness-of-fit of the models has also been reasonably stable as new data have come in, as Table 4 shows. For example, consider the results for the DCSD model when estimated for Berg and others (1999) in 1998 (column 1) with the same model when used to produce the first set of “official” internal forecasts, produced in July 1999 (column 5). The model’s accuracy is actually slightly improved over the longer period. Moreover, the models themselves have remained fairly stable as new data have come in and as the country coverage has changed slightly. For the DCSD model, for example, the values and statistical significance of the coefficients have not changed much.

What about out-of-sample performance? As Appendix II explains, the out-of-sample period for the KLR and DCSD models extends from January 1999 through December 2000, since it is too early to fully judge more recent forecasts. For the private sector GS and CSFB models, which have three-month and one-month horizons respectively, it is possible to look at goodness-of-fit through April and August 2001 respectively.

The out-of-sample results vary substantially by model. The KLR model performs better out-of-sample than in-sample, calling 58 percent of pre-crisis months correctly. The forecasts were highly informative: when the crisis probability was below the cut-off, a crisis ensued in fact only 9 percent of the time, compared to 35 percent of the time when the crisis probability was above the cut-off. The DCSD model’s performance deteriorated substantially over this sample, with only 31 percent of pre-crisis months correctly called. The model remained somewhat informative, with crises following above-cut-off signals 22 percent of the time and below-cut-off-signals only 14 percent of the time.

Figures 3 and 4 give a more complete picture of the models’ performance. The top panels of Figure 3 show, for example, that the DCSD model has more false alarms out of sample than in sample for any cut-off level. At the in-sample optimal cut-off, the model has about the same fraction of false alarms but calls many fewer crises correctly. The bottom panel shows that a much higher cutoff (around 50 percent) would have been desirable. It would have avoided many false alarms without increasing the number of missed crises very much. This is reflected in the top panel for the DCSD model where the out-of-sample curve lies on the in-sample curve in the neighborhood of point C, which corresponds to a cutoff probability of 47 percent.

The out-of-sample period studied in this paper was comparatively calm. According to the specific definition applied in the models, there were two crises per year during the recent period, while the average number of crisis per year in the estimation period is almost three. It is perhaps not surprising then that the models tended to predict the crises well in the recent

looking at banking crisis prediction, discuss the relative weights in terms of the costs of policies and regulations to increase the resilience of the banking system versus the costs of rescue of failed institutions.

period, but that they also registered a comparatively high number of false alarms, that is, crisis predictions that did not materialize.

Of course, it is impossible to know until it is too late what is the best cut-off probability to call crises. However, the successful goodness-of-fit performance of the models for *some* cut-off does imply that the models were able to *rank* the observations reasonably well according to crisis probability, with the higher probabilities assigned to observations that correspond to pre-crisis months.

Table 4. Goodness of Fit: DCSD^{1/} and KLR^{2/} Models

	Asia Crisis				Recent Experience			
	In Sample 1985:12 to 1995:4		Out of Sample 1995:5 to 1996:12		In Sample 1985:12 to 1997:5		Out of Sample 1999:1 to 2000:12	
	DCSD	KLR	DCSD	KLR	DCSD	KLR	DCSD	KLR
Cut-off ^{3/}	18	15	18	15	23	15	23	15
Value of loss function ^{4/}	59	73	63	76	58	68	90	63
Percent of obs correctly called	73	57	62	57	76	70	72	76
Percent of crisis in 24 mo. correctly called ^{5/}	65	73	84	75	63	60	31	58
Percent of tranquil in 24 mo. correctly called ^{6/}	75	54	53	49	79	72	80	79
False alarms as percent of total alarms ^{7/}	69	78	60	62	64	71	78	65
Probability of crisis given signal ^{8/}	31	22	40	38	37	29	22	35
Probability of crisis given no signal ^{9/}	8	8	10	18	8	10	14	9
Statistical Tests of Forecasts ^{10/}								
Coefficient in regression of actual on predicted	1.19	1.02	2.13	1.42	1.14	0.87	0.61	1.47
Standard error	0.30	0.25	0.38	0.74	0.18	0.22	0.39	0.52
p-value(coefficient=0)	0.00	0.00	0.00	0.06	0.00	0.00	0.12	0.01

Actual	Predicted		Predicted		Predicted		Predicted		Predicted	
	T	C	T	C	T	C	T	C	T	C
	T 1571	533	1196	1009	170	151	159	166	1838	711
	C 128	242	108	286	19	101	34	101	196	291
					1970	531			353	90
					180	305			59	26
									351	92
									36	49

Source: Author's calculations.

Notes: See Box 2 for further explanation of items in this table.

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ Kaminsky, Lizondo, and Reinhart (1998).

3/ This is the cutoff probability above which a forecast is deemed to signal a crisis.

4/ The loss function is equal to the sum of false alarms as a share of total tranquil periods and missed crises as a share of total pre-crisis periods.

5/ This is the number of pre-crisis periods correctly called (observations for which the estimated probability of crisis is above the cutoff probability and a crisis ensues within 24 months) as a share of total pre-crisis periods.

6/ This is the number of tranquil periods correctly called (observations for which the estimated probability of crisis is below the cutoff probability and no crisis ensues within 24 months) as a share of total tranquil periods.

7/ A false alarm is an observation with an estimated probability of crisis above the cutoff (an alarm) not followed by a crisis within 24 months.

8/ This is the number of pre-crisis periods correctly called as a share of total predicted pre-crisis periods (observations for which the estimated probability of crisis is above the cutoff probability.

9/ This is the number periods where tranquility is predicted and a crisis actually ensues as a share of total predicted tranquil periods (observations for which the estimated probability of crisis is below the cutoff probability.)

10/ Crisis dummy (1 if pre-crisis month, 0 otherwise) is regressed on forecast probabilities, with HAC standard errors. See text for explanation.

Box 2. Goodness-of-Fit Measures and Trade-Offs

Early warning systems typically produce a predicted probability of crisis. In evaluating their performance, it would be simple and informative to compare these predicted probabilities with actual crisis probabilities. As the latter are not directly observable, goodness-of-fit calculations measure how well the predicted probabilities compare to the subsequent incidence of crisis. The first step is to convert predicted probabilities of crisis into alarms, which signals that a crisis will ensue within 24 months (assuming that is the model's horizon). An alarm is defined as a predicted probability of crisis above some threshold level (the cutoff threshold). Then, each observation (a particular country in a particular month) is categorized as to whether it is an alarm (i.e. whether the predicted probability is above the cutoff threshold) and also according to whether it is an actual pre-crisis month.

The threshold probability for an alarm can be chosen to minimize a "loss function" equal to the weighted sum of false alarms (as a share of total tranquil periods) and missed crises (as a share of total crisis periods). In this paper, equal weight is placed on the share of alarms that are false and the share of crises that are missed. (The former might be thought of as Type 1 errors and the latter as Type 2 errors, if the null hypothesis is of no crisis.) A higher weight on missed crises would imply a lower cut-off threshold for calling a crisis, and the model would generate both fewer missed crises and more false alarms. Note that only in-sample information can be used to calculate a threshold for actual forecasting purposes. When using the model out-of-sample to make predictions, however, there is no guarantee that another threshold would not provide better goodness-of-fit.

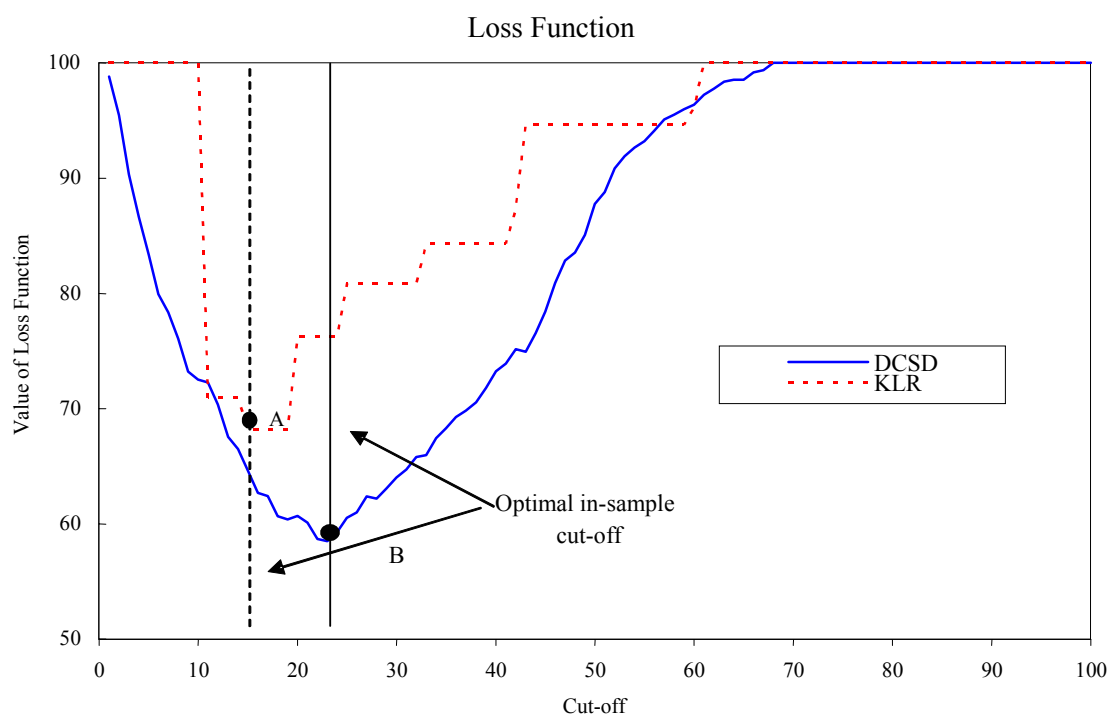
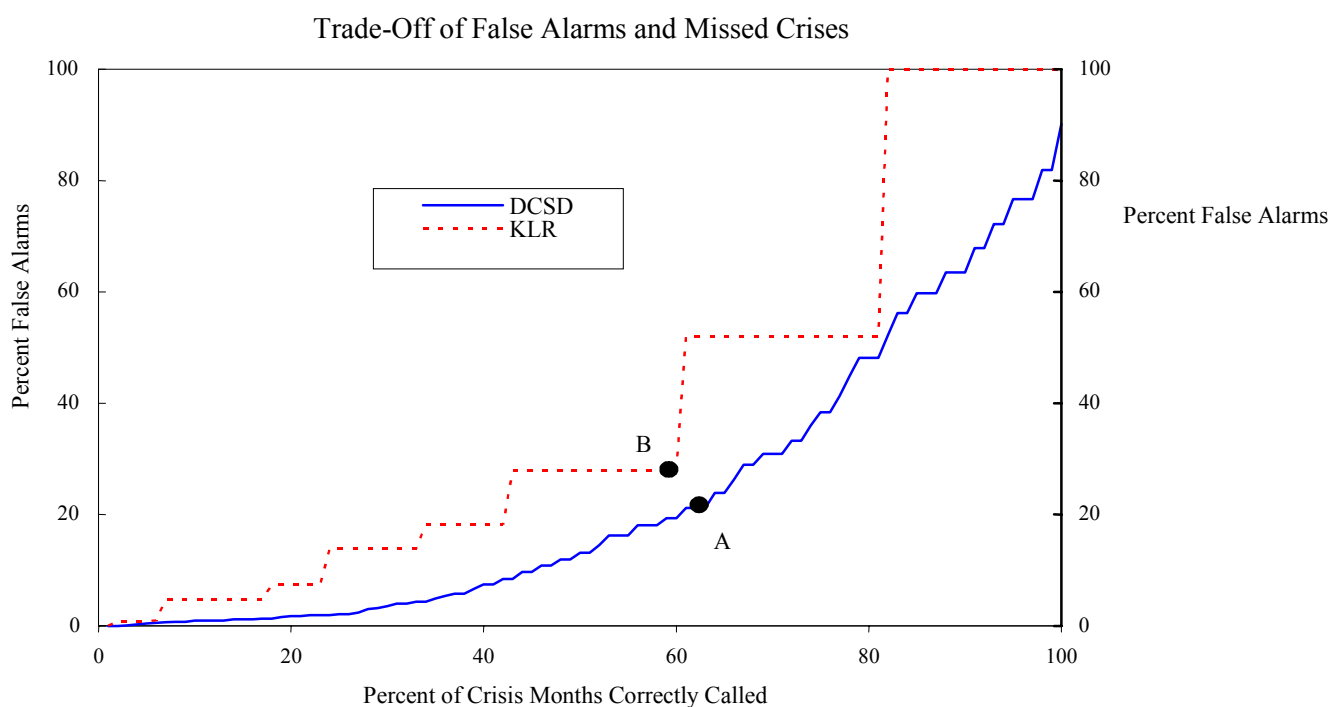
The columns of text Table 4 and Table 5 show how the signals from the various models compare to actual outcomes, over various periods. Each number in the goodness-of-fit table represents the number of observations that satisfy the criteria listed in the rows and columns. For example, for the DCSD model (Table 4, column 7) over the 1999:1 to 2000:12 period, there were a total of 443 tranquil months, and for 90 of them the probability was above the cutoff threshold.

From this table various measures of accuracy can be calculated. For example, the percent of crises correctly called is equal to the number of observations for which the alarm was sounded and a crisis in fact ensued divided by the total number of actual crises. The footnotes to Table 4 define all these various measures.

It is possible that one model might do better when great weight is placed on avoiding false alarms, i.e. when cut-off thresholds are relatively high, while another might excel when the optimal cut-off threshold is low. This turns out not generally to be the case for the models examined here. To see this, figures can be produced that show how the models perform for any cut-off and independent of the loss function chosen. For a given model over a given sample, each candidate cutoff threshold produces a certain percentage of crises correctly called and percentage of false alarms. For example, a cutoff of 0 produces 100 percent crises correctly called but also 100 percent false alarms (as a share of tranquil periods) because the model would predict a crisis every time, while a cutoff of 100 produces 0 percent crises correctly called but 0 percent false alarms because the model will never predict a crisis. The upper panel of Figure 2 traces out all these points for each cut-off between 1 and 100, for the DCSD and the KLR models over the in-sample period. Points that are further towards the lower right are unambiguously preferred for any loss function, in that the percent of crises correctly called is higher while the percent of false alarms is lower. As shown in the figure, the DCSD model dominates for all cut-off frequencies, in that the DCSD curve lies to the right and below the KLR curve. For any given percent of crises correctly called, the DCSD model calls fewer false alarms. Text Figures 3 through 6 show similar results, though they each show one model's in-sample and out-of-sample results.

Another way to see how models compare that does not depend on the cut-off is the loss function graph shown in the lower panel of Figure 2. This shows how the models perform for various cut-off thresholds, for a given loss function. To read this figure, note that the loss function is the number of false alarms (in percent of total tranquil periods) plus the number of missed crises (in percent of total pre-crisis periods). Thus, a loss function value of 50, for example, implies 20 percentage points fewer false alarms and/or missed crises than a loss function of 70.

Figure 2. DCSD ^{1/} and KLR ^{2/} In-Sample Forecasts



Source: Author's calculations.

Notes: See Box 2 for an explanation of these Figures. Point A corresponds to the cut-off that minimizes the loss function in-sample for the KLR model. Point B indicates the same point for the DCSD model.

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ Kaminsky, Lizondo, and Reinhart (1998).

How statistically good (in the case of KLR) or bad (in the case of DCSD) are these results? Table 4 also shows the results of regressing the actual value of the crisis variable on the model's predicted probability of crisis for various models and sample periods. Thus, we run a regression of the form:

$$c24_{it} = \alpha + \beta * PredProb_{it} + \varepsilon_{it}$$

where $c24_{it} = 1$ if there is a crisis in the 24 months after period t (for country i) and 0 otherwise. $PredProb_{it}$ is the predicted crisis probability for period t and country i . For informative forecasts, β should be significant; a coefficient of 1 implies that they are unbiased.¹⁷

The regression results confirm the strong out-of-sample KLR forecasts and suggest that the data are not revealing for the DCSD model. First, the strong KLR and DCSD in-sample results are clear. The estimated β is always statistically different from 0 and a value of 1 cannot be rejected. Turning to the recent out-of-sample period, the KLR model's forecasts are highly significant, while the hypothesis that the true β is 1 cannot be rejected. The DCSD forecasts are not significant at the traditional confidence levels, with a p-value for $\beta = 0$ of 12 percent. However, the data are more consistent with the hypothesis that the forecasts are accurate than that they are useless: the p-value for $\beta = 1$ is 31 percent.

As suggested by the fact that neither the hypothesis that $\beta = 0$ nor that $\beta = 1$ can be decisively rejected, the tests lack power. This in turn reflects the small amount of information in the out-of-sample period. We illustrate this lack of power by carrying out the following simulation exercise. We suppose that in fact, and that the data generating process for the out-of-sample period is exactly the DCSD model as estimated in-sample. This implies that $\beta = 1$ and any

¹⁷ We estimate this regression using OLS with HAC standard errors. This solves two sorts of problems. First, the $C24$ and $PredProb$ variables are highly serially correlated, which causes the OLS standard errors to be incorrect. Monte carlo exercises suggest that in our setup, the OLS standard errors are substantial underestimates but that a HAC correction largely solves this problem. Second, the $C24$ variable is qualitative, resulting in a heteroskedastic ε , as is well known from the "linear probability" literature. The usual solution is to run a probit or logit regression. Here, though, the relationship between $PredProb$ and $C24$ will be linear under either the null (with $\beta = 0$) or the alternative (with $\beta = 1$.) The heteroskedasticity is of known form, suggesting FGLS. However, some observations will produce negative variances. The usual solution is to apply some ad hoc adjustment to these observations, such as dropping them. Our own experience and some monte carlo exercises confirm much earlier conclusions that these procedures are unsatisfactory and suggest that OLS with HAC standard errors produces reasonable results with only a small loss of efficiency compared to GLS. (See Judge, Griffiths, Hill, and Lee (1980) on the linear probability model.) Berg and Coke (2004) discuss similar problems in the estimation of EWS models themselves. Harding and Pagan (2003) address related issues in a different context.

remaining errors are due to the noise inherent in the data generating process – the forecasts are as good as they could be. We then simulate the data-generating process implied by the estimated DCSD model 500 times, creating 500 sets of out-of-sample observations and associated model forecasts. We ask, how often would these ideal forecasts look as bad as those actually produced by the DCSD model using the true out-of-sample data? The answer is that for these ideal forecasts, the hypothesis that $\beta = 0$ would *not* be rejected 28 percent of the time.

Neither of the short-horizon private-sector models performs well (Table 5 and Figures 5 and 6).¹⁸ While in-sample goodness-of-fit was adequate, the models' out-of-sample forecasts deteriorate sharply.¹⁹ The statistical tests reflect this, in that the forecasts are not significant predictors of actual crisis incidence out of sample. Again, the data are not completely definitive, however, particularly for the GS model. Here, the p-value for the hypothesis that the forecasts should be given no weight is 0.17, while the p-value that the forecasts are unbiased predictors of crisis risk is 0.28.

All of these results should be interpreted cautiously. The number of crises actually observed has been limited. This translates into a small effective sample size.²⁰ In this context, small changes in sample can make a large difference in the goodness of fit indicators yielded by the models. A previous version of this paper found that the DCSD model forecasts performed as well in the 1999:1 to 1999:6 out-of-sample period as in-sample.

The private sector models set out to accomplish a distinctly different task than DCSD or KLR, namely to predict the timing of a crisis with precision. The adoption of a shorter horizon may make prediction easier, since signs of crisis may emerge more clearly right before. On the other hand, the exact timing of a currency crisis may be more difficult to

¹⁸ Each model's own definition of crisis used to evaluate its performance. See Box 1 on crisis definitions.

¹⁹ This assessment is based on this paper's metric for evaluating performance. CSFB's own method uses a different loss function to choose a cut-off, putting more weight on missed crises. In effect, their objective is to minimize false alarms, subject to achieving a certain share of correctly called crises. Also note that CSFB uses the probabilities in a more complex way to generate various levels of risk warnings for clients, based on changes in the probabilities over the most recent one to six months. We have not evaluated how well this system does in predicting crises.

²⁰ To put this problem another way, the pre-crisis observations and the predicted probabilities are highly serially correlated; adjusting for this factor greatly increases the standard errors in the model. This also implies that adding observations through extending the time dimension of the out-of-sample period is not as helpful as the increase in the number of total observations would suggest.

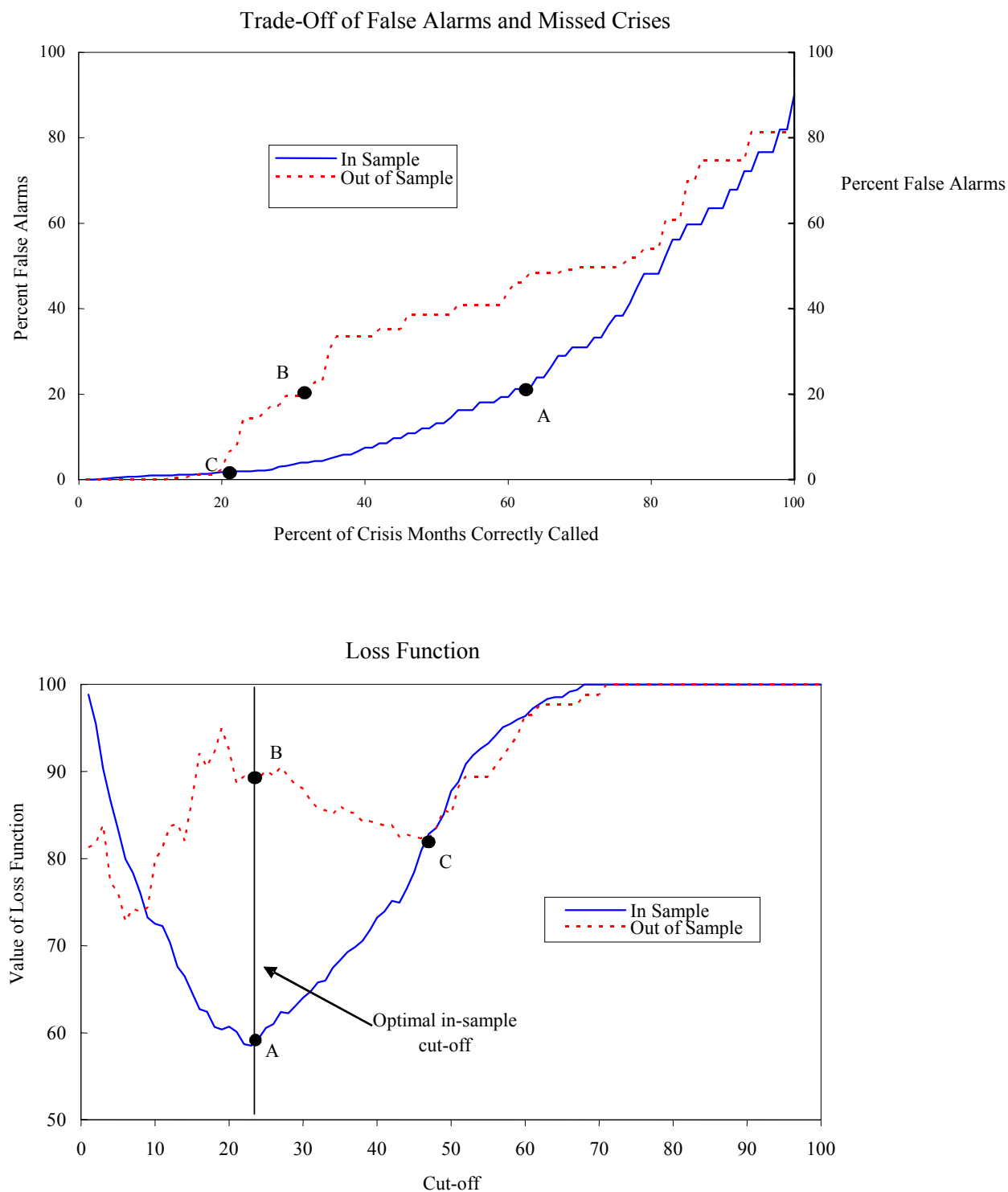
predict than vulnerability over an interval of time as wide as two years, in part because of the possible existence of multiple equilibria and resulting difficulty in predicting the timing of speculative attacks. In any case, the comparison of the short-horizon models and the long-horizon models is not direct and must be treated with caution.²¹

A final and important qualification is that the terms “false alarm” and “missed crisis” should not be taken too literally. An alarm is considered false if no crisis in fact ensues. However, this signal may have been appropriate. First, the crisis definitions employed may miss classifying some events as crises that we might well want a model to warn about. Second, a warning may be followed by policy adjustment or luck that causes the crisis to be avoided; the warning might nonetheless have been useful. Indeed, an examination of the 90 observations that generated false alarms in the 1999:1 to 2000:12 period in the DCSD model suggests that half of them (46) can be readily classified in one or the other of these cases.²²

²¹ In addition, there are some differences in the way the out-of-sample forecasts were generated. Those for the GS model come directly from contemporary monthly publications, so they necessarily reflect incomplete data that had to be filled with estimates for various predictive variables. For example, the January 1999 crisis probability (for April 1999) uses GS estimates of data as of January 1999. The CSFB estimates, in contrast, may use revised predictive variables, although it is not clear how substantial are revisions. This issue is somewhat less serious for the DCSD and KLR models as they forecast over much longer horizons. The July 1999 forecasts, for example, made use of data available only through April for many series, but since the forecast horizon is so long, the use of such data did not make the forecasts obsolete.

²² The countries involved in these 46 observations of technically false, but still useful alarms are Argentina, Chile (before July 1999), Pakistan, Venezuela, Turkey, and Uruguay.

Figure 3. DCSD^{1/} Model Forecasts

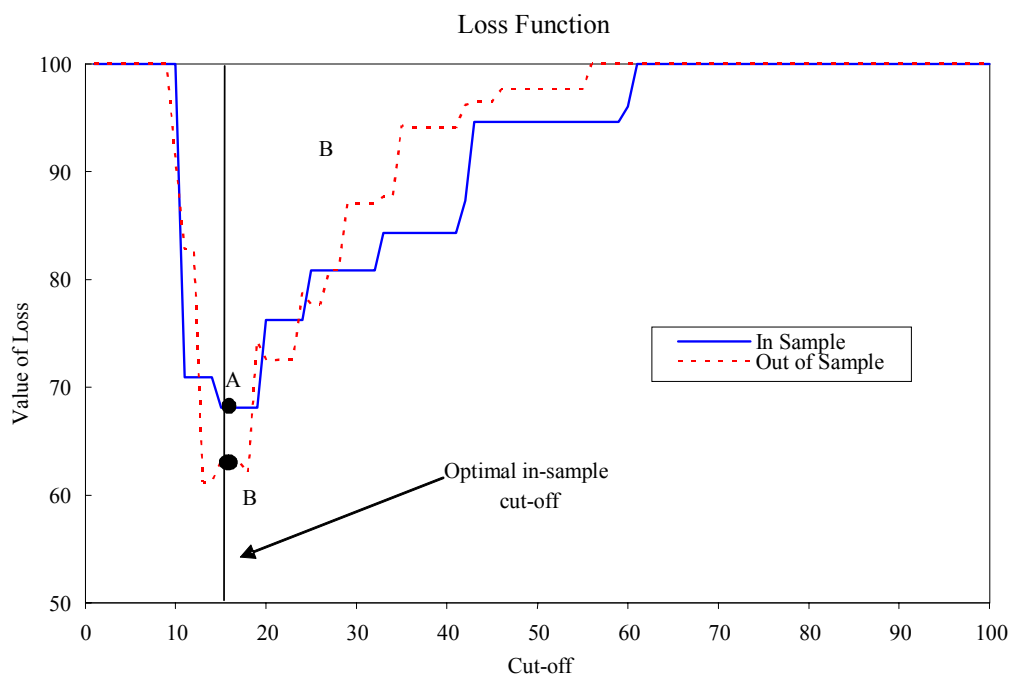
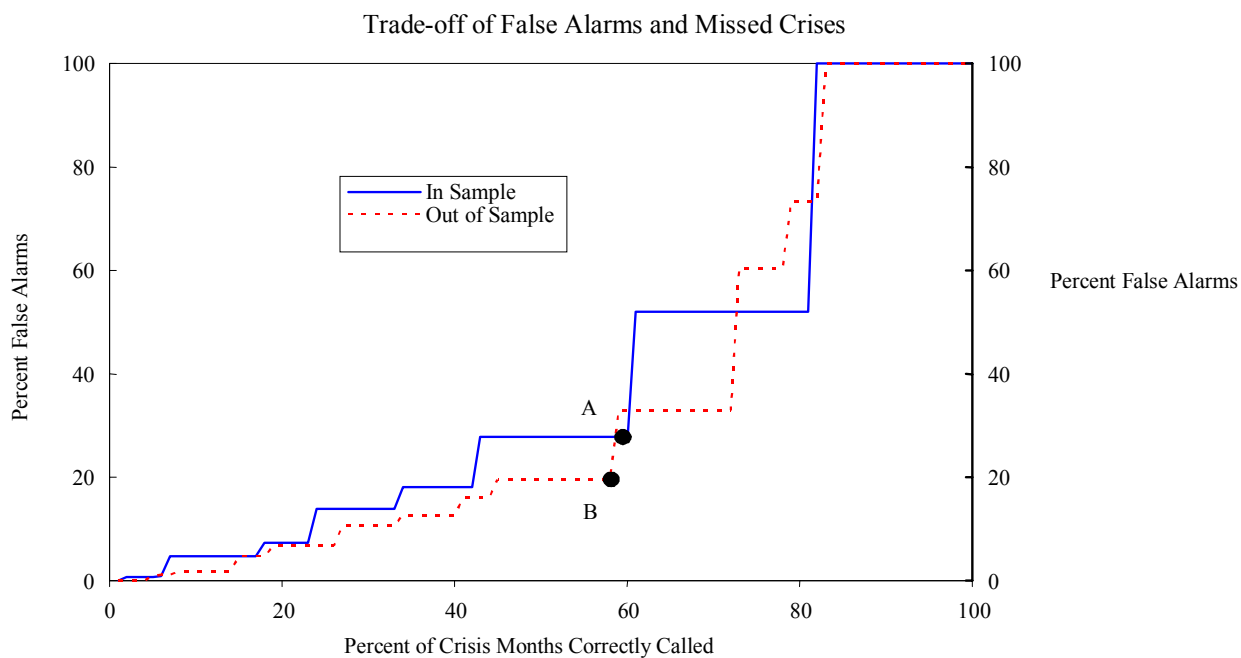


Source: Author's calculations.

Notes: See Box 2 for an explanation. Point A corresponds to the cut-off that minimizes the loss function in-sample. Point B indicates the out-of-sample results corresponding to the same cut-off. Point C corresponds to a cut-off of 47 percent.

^{1/} Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

Figure 4. KLR ^{1/} Model Forecasts



Source: Author's calculations.

Note: See Box 2 for an explanation. Point A corresponds to the cut-off that minimizes the loss function in-sample. Point B indicates the out-of-sample results corresponding to the same cut-off.

1/ Kaminsky, Lizondo, and Reinhart (1998).

Table 5. Goodness of Fit: Short-Horizon Models

	In Sample		Out of Sample	
	GS	CSFB	GS	CSFB
	1996:1 to 1998:12	1994:1 to 2000:7	1999:1 to 2001:4	2000:8 to 2001:8
Cut-off ^{1/}	10	35	10	35
Value of loss function ^{2/}	66	58	97	88
Percent of obs correctly called	66	76	50	83
Percent of crisis correctly called ^{3/}	66	65	54	27
Percent of tranquil correctly called ^{4/}	66	76	50	85
False alarms as percent of total alarms ^{5/}	74	92	87	96
Probability of crisis given signal ^{6/}	26	8	14	4
Probability of crisis given no signal ^{7/}	8	2	12	2
Statistical Tests of Forecasts ^{8/}				
Coefficient in regression of actual on predicted	1.41	0.17	0.56	0.06
Standard error	0.42	0.03	0.41	0.06
p-value(coefficient<=0)	0.00	0.00	0.17	0.29
Actual	Predicted		Predicted	
	Tranquil	Crisis	Tranquil	Crisis
	540	279	325	332
	50	98	45	52
Crisis	Predicted		Predicted	
	Tranquil	Crisis	Tranquil	Crisis
	1951	618	361	65
	29	54	8	3

Source: Author's calculations.

Notes: See Box 2 for further explanation of items in this table.

1/ This is the cutoff probability above which a forecast is deemed to signal a crisis.

2/ The loss function is equal to the sum of false alarms as a share of total tranquil periods and missed crises as a share of total pre-crisis periods.

3/ This is the number of precrisis periods correctly called (observations for which the estimated probability of crisis is above the cutoff probability and a crisis ensues in 3 months (GS), or 1 month (CSFB)) as a share of total pre-crisis periods.

4/ This is the number of tranquil periods correctly called (observations for which the estimated probability of crisis is below the cutoff probability and no crisis ensues in 3 months (GS), or 1 month (CSFB)) as a share of total tranquil periods.

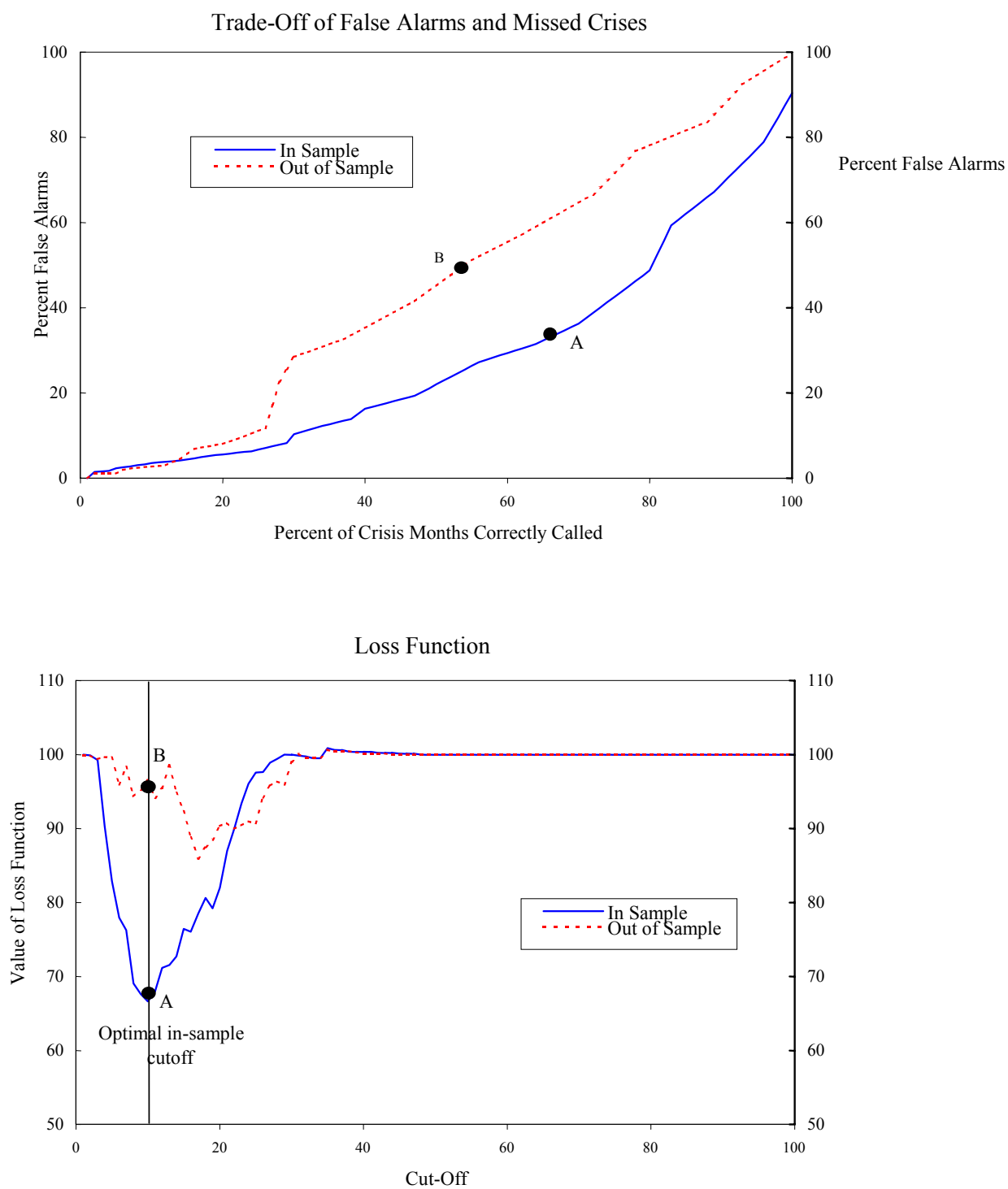
5/ A false alarm is an observation with an estimated probability of crisis above the cutoff (an alarm) not followed by a crisis in 3 months (GS), or 1 month (CSFB).

6/ This is the number of precrisis periods correctly called as a share of total predicted precrisis periods (observations for which the estimated probability of crisis is above the cutoff probability).

7/ This is the number periods where tranquility is predicted and a crisis actually ensues as a share of total predicted tranquil periods (observations for which the estimated probability of crisis is below the cutoff probability.)

8/ Crisis dummy (1 if pre-crisis month, 0 otherwise) is regressed on forecast probabilities, with HAC standard errors. See text for explanation.

Figure 5. GS ^{1/} Model Forecasts

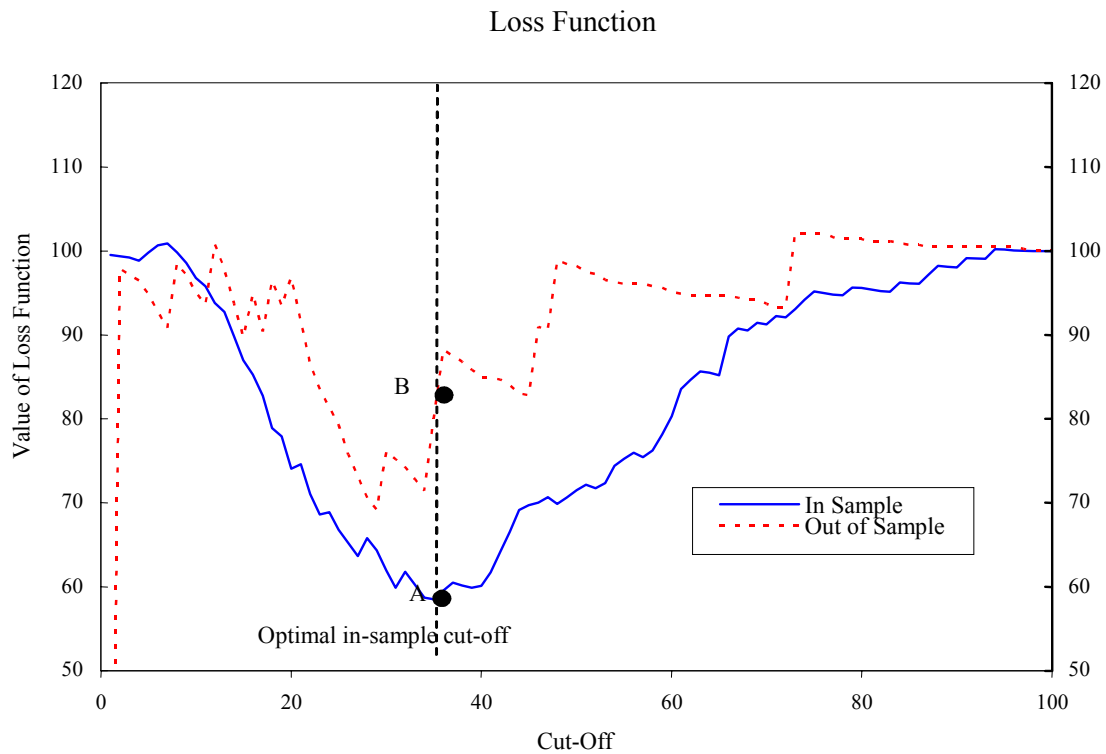
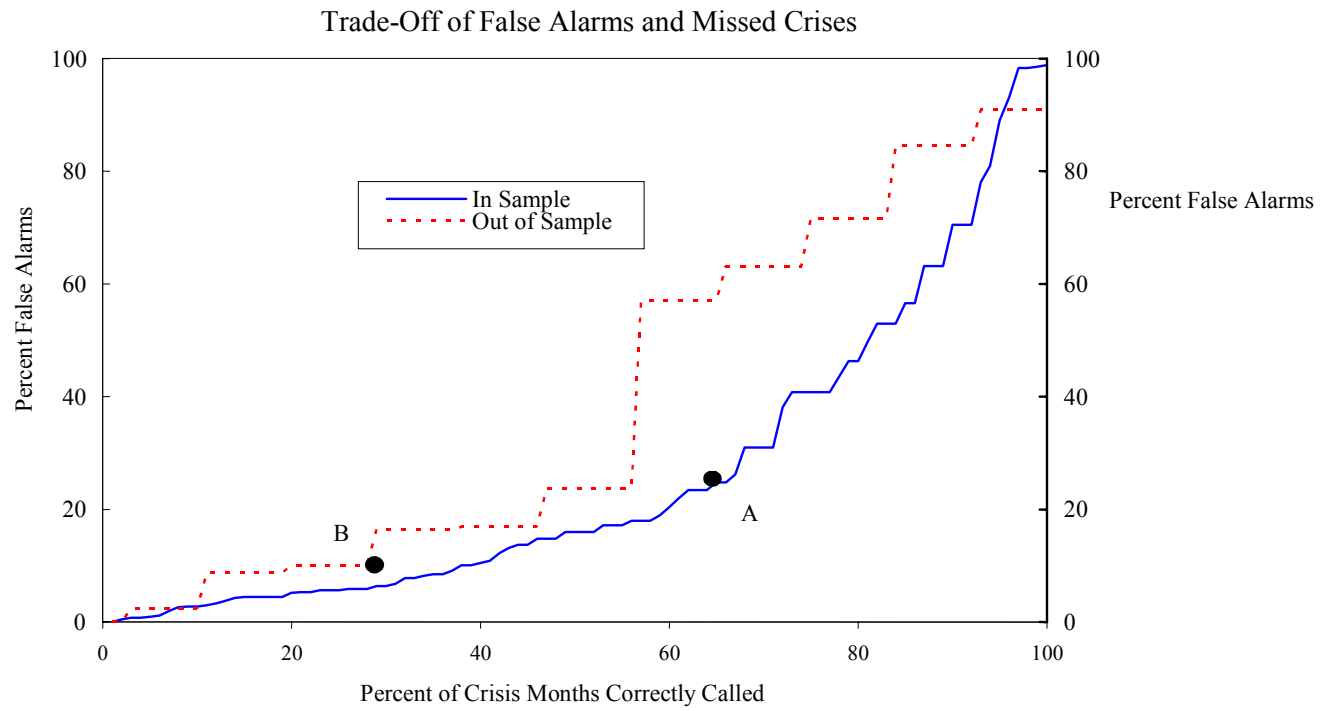


Source: Goldman-Sachs and author's calculations.

Notes: See Box 2 for an explanation of these Figures. Point A corresponds to the cut-off that minimizes the loss function in-sample. Point B indicates the out-of-sample results corresponding to this same cut-off.

^{1/} Goldman-Sachs.

Figure 6. CSFB ^{1/} Model Forecasts



Source: Credit Suisse First Boston and author's calculations.

Notes: See Box 2 for an explanation of these Figures. Point A corresponds to the cut-off that minimizes the loss function in-sample. Point B indicates the out-of-sample results corresponding to this same cut-off.

^{1/} Credit Suisse First Boston.

IV. SUMMARY AND CONCLUSIONS

Since the beginning of 1999, the IMF's staff has been systematically tracking several models, both inhouse and from private institutions, on an ongoing basis, as part of its broader, forward-looking vulnerability assessments. This paper looks in detail at the performance of these models in practice.

We have monitored two long-horizon, in-house models (DCSD and KLR) and two short-horizon, private sector models (GS and CSFB) since 1999. This paper has analyzed the forecasts made between January 1999 and December 2000 by the 24-month-horizon DCSD and KLR models, between January 1999 and April 2001 by the GS model, and between April 2000 and June 2001 by the CSFB model. These forecasts were "pure" out-of-sample forecasts, in that no information about actual outcomes was used in the forecasts or, more generally, in the development or estimation of the models themselves.

The results are mixed. The forecasts of the KLR model are statistically and economically significant predictors of actual crises. The forecast accuracy in the out-of-sample period is only slightly inferior to the accuracy in the estimation period. The DCSD model performs substantially worse out of sample than in sample. The forecasts are still somewhat informative, however; and the hypothesis that the forecasts are unbiased and informative is more likely (p value = 0.31) than the hypothesis that the model's forecasts were useless (p value = 0.12). The ambiguous statistical results reflect the fact that the out-of-sample period contains 528 observations but contains only eight crises; the latter number is important in determining the amount of information in the data.²³

On the whole, the short-horizon, private sector models we examined performed poorly out of sample, despite stellar in-sample performance. Both sets of crisis predictions were largely uninformative—the probability of a crisis was about the same whether the forecast probability was above a cutoff threshold or not.

At least for the KLR model and the DCSD model, the forecasts were statistically significant predictors of crisis (at only the 12 percent level for the latter). This means that they are likely better than could have been produced with a dartboard and a suitable target. How do they compare, though, with more challenging benchmark of alternative forecasts? Here we have compared them with bond spreads, agency ratings, and, perhaps most relevant for IMF work, overall currency crisis risk scores published by analysts. We find that during the Asian crisis, these alternative indicators fared very poorly, much worse than the DCSD and KLR EWS

²³ The serial correlation in the data also reduces the effective amount of information, as discussed in footnote 17. An earlier version of this paper analyzed data through end-1999 and found that the DCSD model performed as well out of sample as in sample. This dependence of the results on the sample is captured by the low power of the tests.

models. During the recent period, the non-model forecasts performed somewhat better, though still generally not as well as the models.

Overall, these results reinforce the view that EWS models are not accurate enough to be used as the sole method to anticipate crises. However, they can contribute to the analysis of vulnerability in conjunction with more traditional surveillance methods and other indicators. It is worth underlining the relatively high standard to which these models are being held. It is plausible to suppose that comprehensive assessments by informed analysts, based on all available qualitative and quantitative information, must be better than the inevitably simple EWS models. But the evidence we have examined with respect to this question is not encouraging concerning these more comprehensive assessments.

The advantage of EWS models lies in their objective, systematic nature. The models process data in a mechanical way and are not clouded by conventional misperceptions or biases based on past experiences. For example, as shown in Section III, Korea, a country that had one of the most successful economic records in recent years, was showing some serious signs of vulnerability to an external crisis in 1996–97 according to EWS models. However, possibly on account of that successful record, analysts and markets did not signal any increase in risk prior to the country's December 1997 currency crisis.

Over the Asia crisis periods, the best EWS models did dramatically better than non-model-based predictors, such as spreads, ratings, and assessments of informed analysts. Over the more recent period, the performance of some of these alternative predictors improved somewhat, so that the relative superiority of the models declined. This suggests that recent crises have simply not been the surprises that the Asian crises were, either because they were easier to predict or because analysts' sensitivity was heightened. In general, most analysts foresaw important risks in the crisis countries in question. The expected strength of the EWS models is in identifying important crisis risks that other forms of analysis do not expose, while avoiding an excessive number of false alarms that would dilute the credibility of the crisis signals. The crises of 1999–2002 have not, fortunately, afforded such an opportunity.

Looking at events of the past few years, it is clear that several developments are underway that are changing the landscape for currency-crisis-prediction models. First, we have observed a resurgence of crises in which sovereign and domestic debt dynamics play a central role. Debt and currency crises are related but distinct: most debt crises are associated with currency crises, but the reverse is not true. Recent work has, appropriately, turned to predict these sorts of crises.²⁴

²⁴ Detragiache and Spilimbergo (2001) and Manasse, Roubini, and Schimmelpfennig (2003) look at determinants of debt crises. Hemming, Kell, and Schimmelpfennig (2003) look at fiscal vulnerabilities in emerging market economies. Sy (2003) emphasizes that debt and currency crises are distinct events.

A second trend is the increased importance of floating exchange rates in emerging markets. Over the past decade, in large part in the aftermath of currency crises, there has been a sharp increase in the number of emerging market countries in which the capital account is broadly open, the de jure exchange rate regime is floating, and there is substantial de facto flexibility. Ten years ago perhaps only South Africa, among large developing countries, fit these characteristics. Now, a large number, including Brazil, Chile, Colombia, Korea, Mexico, Poland, and Thailand have joined the ranks. This sort of arrangement, often augmented by an inflation-targeting monetary policy, is indeed now more the rule than the exception for such countries. Going forward, what does this imply for currency-crisis models?

In principle, a sharp depreciation of the exchange rate can happen under a floating exchange rate regime as easily as under any other regime. Although floating rate regimes would help to avoid situations of extreme overvaluation, particularly those driven by policy inconsistencies, the economies could still be vulnerable to sudden changes in market sentiment, unsustainable levels of debt or financial sector weaknesses, among other factors. Moreover, regimes that are broadly floating may, under speculative attack, evolve towards de facto pegs as policymakers resist the downward pressure on the currency. In fact, according to the IMF's de jure classification, there is no evidence that floating rates have been more resistant to currency crises.²⁵ We draw no firm conclusions here. We suspect, though, that painful currency crises will remain a feature of emerging markets for the foreseeable future.

²⁵ A more complete analysis should correct for the influence of other factors that contribute to currency crises and would consider de facto classifications such as those presented in Levi-Yeyati and Sturzenegger (2001) and Reinhart and Rogoff (2003).

DESCRIPTION OF EWS MODELS AND SPECIFICATION ISSUES

1. Models Implemented at the IMF

Kaminsky-Lizondo-Reinhart (KLR) model

Perhaps the most prominent model for predicting currency crises proposed before the Asia crisis is the indicators approach of Kaminsky, Lizondo, and Reinhart (1998) (KLR) who monitor a large set of monthly indicators that signal a crisis whenever they cross a certain threshold. The model attempts to predict the probability of a crisis within the next 24 months, where a crisis occurs when there are extreme changes in a weighted average of the monthly exchange rate depreciation and reserve loss. A variable-by variable approach is chosen so that a surveillance system based on the method would provide assessments of which variables are “out-of-line”. In addition to overvaluation, the current account, reserve losses and export growth, the model also includes reserves to broad money as a measure of reserve adequacy and several monetary variables such as domestic credit growth, real interest rates, and excess M1 balances.²⁶ The information from the separate variables is combined, using each variables forecasting track record, to produce a composite measure of the probability of crisis (Kaminsky, 1998). Staff has implemented a version of the KLR model, supplemented with several additional variables.

Developing Country Studies Division (DCSD) model

The current structure of the Developing Country Studies Division (DCSD) model has been influenced by the path of its development. Its origins were a project testing the out-of-sample performance of the KLR and other models in predicting the Asian crisis. Later work tested the usefulness of interpreting predictive variables in terms of discrete thresholds, the crossing of which signals a crisis (Berg and Pattillo, 1999a,b). Using the same crisis definition and prediction horizon as KLR, but embedding the KLR approach into a multivariate probit regression, they found that a better simple assumption is that the probability of crisis goes up linearly with changes in the predictive variables. The variables are measured in percentiled form, that is, relative to their own history.

The resulting “linear” probit model in that paper was composed of five variables: real exchange rate deviations from trend, the current account to GDP ratio, export growth reserve growth, and the level of M2 to reserves. This set of predictor variables was the result of

²⁶ Goldstein, Kaminsky, and Reinhart (2000) add some new indicators and update the KLR model. They find that the best monthly indicators for predicting currency crisis were: real exchange rate appreciation, a banking crisis, a decline in equity prices, a fall in exports, a high ratio of broad money to reserves and a recession; while the best annual indicators were a large current account deficit relative to both GDP and investment.

starting with an extensive set of KLR variables plus our additions, and selecting the most important variables through a specification search process.

Since the role of short-term debt into weak financial systems was brought to the forefront by the Asian crises, a measure of short-term debt to reserves was added to the model (OP 186). It was found to be highly significant, while the ratio of M2 to reserves lost its significance and was dropped from the model, resulting in the current five variable DCSD model.²⁷

Policy Development and Review (PDR) model

A third EWS model recently developed in the Fund is the PDR model. This EWS adds balance sheet variables and proxies for standards to the DCSD model (Mulder, Perrelli and Rocha, 2001). Variables found to be important in predicting the probability of a crisis are: at the corporate level, leveraged financing and a high ratio of short-term debt to working capital; balance sheet indicators of bank and corporate debt to foreign banks as a share of exports; and a legal regime variable proxying shareholders rights.

The corporate sector data are available only on an annual basis and with a significant lag. These variables are often slow moving, however, so they can still contribute to forecasting accuracy. More up-to-date data, as well as a larger and more stable underlying sample of corporations, would increase the analytical and forecasting usefulness of the model.

2. Private Sector Models

The interest of investment banks in developing EWSs as tools for advising their clients has fluctuated following both what is “in fashion” and whether crises are in the daily headlines. Following the Asian crisis, most major banks developed in-house models attempting to predict currency crashes. These models were designed either for explicit use in advising on foreign currency trading strategies, or more generally to assess values and risks in emerging market currencies and supplement economic forecasts provided to investors. Since that time, a number of these systems have ceased operation: Lehman Brothers has abandoned its Currency Jump Probability model; Citicorp no longer implements its Early Warning System for anticipating balance-of-payments crises in Latin America; and JP Morgan has substituted a simple weighted vulnerability index for its Event Risk Indicator model. However, as volatility in emerging markets ratcheted up again in late 2000 and 2001, a number of new private models were brought out. For example, Deutsche Bank introduced its Deutsche Bank Alarm Clock (DBAC) and Morgan Stanley Dean Witter has recently set up an Early Warning System for Currency Crises.

²⁷ The model uses mainly monthly data, but also some quarterly or, for some countries, annual data. These latter series are interpolated or extrapolated to generate monthly crisis predictions.

Goldman Sachs's GS-WATCH

Staff regularly tracks Goldman Sachs' GS-WATCH model and Credit Suisse First Boston's (CSFB) Emerging Markets Risk Indicator (Roy and Tudela (2000)), both of which have been in operation since 1998. GS-WATCH (Ades, Masih, and (1998)) predicts the likelihood of a crisis in three-month time, defined as a weighted average of three-month exchange rate and reserve changes. The predictions are generated through a logit regression where most explanatory variables are converted into zero/one signals. The predictor variables include macro fundamentals such as measures of credit booms, real exchange rate misalignment, export growth, reserve growth, and external financing requirements, as well as changes in stock prices, political risk, contagion and global liquidity. The latter two variables are measured continuously, making the overall crisis probabilities follow a smoother path. While inclusion of political risk makes sense, the simple zero/one variable (one around the time of elections, or when a revolution, coup, major riot or strike takes place) only partially captures this type of risk. The model is estimated using monthly data, but predictions are updated weekly for inclusion in analysts' reports. On a week-to-week basis, changes in the contagion variable drive much of the movements in the crisis predictions. Contagion is measured for each country as a weighted average of the changes in the exchange rate and reserve change index for the other countries in the sample, where the weights are the historical relationships in those indices across countries.

Credit Suisse First Boston's Emerging Markets Risk Indicator (CSFB)

CSFB re-specified its model in September 2000, changing some of the predictor variables and reducing the number of variables (CSFB, 2000). A logit model predicts the one-month ahead probability of depreciation greater than 5 percent and at least double the preceding month's depreciation. The variables are standardized, that is, measured relative to the country-specific mean and variability for that variable. Many variables similar to other models are included: real exchange rate deviations from trend, the ratio of debt to exports, growth in credit to the private sector, output changes, reserves to imports, changes in stock prices, oil prices and a regional contagion dummy, measured simply as the number of countries in the region recently experiencing a crisis.

Deutsche Bank Alarm Clock (DBAC)

The DBAC model defines separate exchange rate and interest rate "events" as depreciations greater than a certain size (estimated separately for levels ranging from 5 to 25 percent) and increases in the money market interest rates of more than 25 percent in a month (Deutsche Bank, 2000). It uses a methodology to jointly estimate the probability of these two types of events, allowing the probability of a simultaneous interest rate event to influence the likelihood of an exchange rate crisis and the probability of a depreciation event to affect the predictions of an interest rate crisis. Relatively few predictors are included in the exchange rate event model: changes in stock prices, domestic credit, industrial production, and real exchange rate deviations, as well as a contagion variable. All of the investment bank models claim to demonstrate that an investor using trading strategies based on their model could earn

substantial profits over a particular period. DBAC adds a twist to these calculations by proposing an “action trigger” to identify cutoff probability levels at which an alarm should be sounded and investors should change their positions. The trigger is calculated to maximize profits, assuming a strategy where the investor will be long the local currency when the probability of a depreciation crisis is below the trigger and short whenever the probability crosses above the trigger.

3. Specification Issues

EWS models are econometric methods for generating predictions of currency crises, precisely defined. Although there have long been empirical studies of currency crises, it was not until after the 1994–95 Mexican tequila crisis that the literature focused on finding methods for predicting crises, rather than on explaining a particular set of historical crises or testing specific theories. The largely unexpected Asian crisis, however, provided the real impetus for a new wave of papers, and the development of systems for a continuous monitoring of crisis vulnerabilities at various institutions.

What is being predicted?

Most would agree that a sudden, large depreciation of the exchange rate constitutes a currency crisis. Further, a situation of intense pressure on the foreign exchange market, resulting in large losses of international reserves and/or a hike in domestic interest rates can also be considered a crisis, even if step devaluation is avoided. In any event, one may be interested in forecasting both successful (those resulting in an exchange rate depreciation) and unsuccessful attacks on the currency, so that both types of event would be considered a crisis for the purpose of a forecasting model. Text Box 1 discusses the difficulties involved in operationalizing the concept of currency crisis and how they are addressed in the models considered in this paper. Table 6 lists crisis dates for the various models for the 1999–2001 period.

What variables to include?

After identifying a set of crises, the next issue is the choice of a set of variables that may be useful in predicting crises. Berg and others (1999) surveys the literature on currency crises and looks for common symptoms of crises in past historical episodes. The starting place for drawing up a list of potential predictive variables is the theoretical models of currency crisis. “First-generation” models focus on macroeconomic imbalances that lead to a depletion of foreign exchange reserves and make devaluation inevitable. In second-generation models, the government weighs the cost and benefits of defending the currency. Since expectations affect the trade-off faced by the policymaker, crises may be self-fulfilling, and thus much more difficult to predict. More recent models stress elements such as market failure in international capital markets, and distortions in domestic financial markets. For example, information failures can lead to investor herding and contagion, and implicit government guarantees of private sector liabilities can generate moral hazard and unsustainable implicit deficits.

The theoretical literature suggests classifying variables into three groups. First, macroeconomic fundamentals such as measures of real exchange rate overvaluation, the fiscal deficit, excess money growth, terms of trade, domestic credit, the current account deficit and output growth. Second, variables that provide a gauge of a country's vulnerability to attacks, if, given relatively weak fundamentals, an attack was to take place. These include measures of the adequacy of international reserves relative to possible short-run liabilities of external and domestic origin, external financing needs, and soundness of the financial sector.

Third, indicators of market expectations or sentiment, such as interest rate differentials, bond spreads, the forward exchange rate, number of crises elsewhere or other contagion channels, and variables representing investors' "risk appetite".

The task of specifying a model with variables that are useful predictors of crisis does not involve simply assembling all the a-priori plausible variables. There is an important danger of "overfitting" a model, by adding more and more variables through "data-mining". Typically, such a model will do very well in explaining a particular historical episode of crisis, but will have little power in forecasting the next set of crises. The objective of finding the best method to forecast crisis probabilities argues for a parsimonious model: a robust set of variables useful for predicting both past and future crises. There is a deeper problem associated with the statistical one. If the nature of crises changes from one episode to the next, how can a model be robust to those changes? The answer is to focus on the "symptoms" that may be common to all external crisis episodes even when the ultimate causes of those crises are different.

It should also be kept in mind that the different indicators are related among themselves, so that the inclusion of all of them is not necessary. The indicators may be covered indirectly, in that the variables employed in the model may capture many of the important manifestations of these other problems. For example, a large fiscal deficit and high inflation may contribute to the risk of crisis, but may be already accounted for in a model that includes real exchange rate overvaluation and the current account deficit.

Finally, there are the issues of availability of consistent data over time and across countries, and at a desirably high frequency. Data on the health of the financial sector, such as rates of non-performing loans, is an important example of factors that do not meet those standards. Political risk is another example of a factor that is intrinsically difficult to consistently measure. In addition, some variables may not fit well into the structure of a given model. A good example is the phenomenon of contagion. The transmission of crises from country to country, particularly if the mechanism operates through financial channels, seems to occur quite rapidly. Thus, it is difficult to incorporate contagion in models attempting to predict the likelihood of crisis over a longer horizon, such as the next two years. Also, there are other idiosyncratic variables (for example, oil prices) that while particularly important for some countries, may have insignificant or contrary effects over other emerging markets.

Table 6. Crisis Dates According to Different Models

DCSD ^{1/} KLR ^{2/} Crisis Dates (1999:1–2003:3)

Brazil	Jan 99
Colombia	Aug 99; Jul 02
South Africa	Dec 01
Turkey	Feb 01
Uruguay	Jul 02
Venezuela	Feb 02
Zimbabwe	Aug 00

GS ^{3/} Crisis Dates (1998:7–2001:4)

Brazil	Jul 98–Jan 99; Jun–Jul 99; Mar–May 00; Sep 00; Jan–Apr 01
Bulgaria	Jan–Apr 99; Jan–Feb 00; Feb 01
Chile	Jun–Aug 99; Mar–Apr 01
China	Jul–Aug 98
Colombia	Jul–Aug 98; Mar–Jul 99; Mar–May 00
Czech Rep.	Nov 98–Jan 99
Ecuador	Jul–Aug 98; Nov 98–Feb 99; Apr–May 99; Jul–Nov 99; Jan 01
Egypt	Jul 99; Sep–Oct 00
Hong Kong	Jul–Sep 98; May–Jul 99
Hungary	Jul 98
India	May 00
Indonesia	Dec 98; Jun–Jul 99; Mar–May 00; Jan–Mar 01
Israel	Jul–Sep 98; Jul 99
Korea	Sep–Nov 00
Malaysia	Aug 00
Mexico	Jul 98
Peru	Jul–Nov 98; Jul 99; Nov 00
Philippines	Jul 98; Jun–Jul 99; May–Oct 00
Poland	Nov 98; Jul 99; Mar 00
Russia	Jul–Nov 98; Jul–Aug 99
South Africa	Sep–Oct 00
Singapore	Oct 98; Mar 01
Taiwan	Sep–Nov 00; Apr 01
Thailand	Jun–Jul 99; Feb–Aug 00
Turkey	Sep–Nov 98; Nov 00–Mar 01

CSFB ^{4/} Crisis Dates (1998:7–2001:8)

Brazil	Mar 99
Colombia	Nov 98; Sep 99
Croatia	Apr 99
Czech Rep.	Apr 99
Ecuador	Mar–Apr 99; Aug 99; Nov–Dec 99; Feb 00
Indonesia	Jul 98; Mar 99; Oct 99; Jan 00; Nov 00; Jun 01
Israel	Dec 98
Korea	Feb 01
Mexico	Oct 98
Nigeria	May 99; May 01
Pakistan	Nov 00
Philippines	Dec 00
Poland	Apr 99
Russia	Oct–Nov 98; Feb 99; Jun 99; Mar 00
South Africa	Aug 98
Slovakia	Oct 98; Jul 00
Sri Lanka	Aug 00; Mar 01
Thailand	Aug 98; Nov 99
Turkey	Jan 99; Apr–May 01; Aug 01
Zimbabwe	Jul 98; Oct 98; Mar 99; Oct 00

Source: Author's calculations.

1/ Berg and others (1999). DCSD stands for Developing Country Studies Division, the Division of the Fund in which the model was originally formulated.

2/ Kaminsky, Lizondo, and Reinhart (1998).

3/ Goldman-Sachs.

4/ Credit Suisse First Boston.

How to generate predictions?

Two conceptual questions underlie the choice of a methodology that uses the variables to predict the crises. First, how should the information content of each explanatory variable be assessed? One option is the “signaling” approach, where each indicator is said to issue a signal of impending crisis when its value exceeds a particular threshold. For example, if the country specific threshold for the current account deficit to GDP were 3 percent, a ratio below 3 percent would not contribute to the risk of crisis; while, ratios at any level above 3 percent would contribute equally to the probability of a crisis. A second option is to introduce the variables continuously so that, for example, any small increase in the current account/GDP ratio could marginally increase the crisis prediction.

It is also necessary to decide how the variables should be measured. Some models include the variables in raw form, often in growth rates or ratios. Alternatively, the variables could be measured relative to their history for each country. For example, what matters in the DCSD model is not the level of the current account deficit per se but whether the deficit corresponds to a high percentile relative to the history of the current account deficit in each country considered individually.

The second question is how to aggregate the information from the different variables into a single prediction. A method associated with the signaling approach is the calculation of a composite probability as the weighted sum of the number of indicators that are signaling, where each indicator is weighted by its reliability in predicting crises.²⁸ An alternative is to use a probit (or logit) regression, that is a regression where the dependent variable takes the value of one when there is a crisis and zero otherwise.²⁹

What are the relative advantages and disadvantages of each approach? The indicator approach is a popular one because the framework of monitoring key variables for signs of “unusual” behavior accords well with the intuition of early warning. But, by evaluating each variable separately, the method does not consider how an inter-related set of conditions could make an economy more vulnerable to crisis. A practical difficulty with the indicator approach is that the crisis probabilities tend to be “jumpy”, as variables move in and out of

²⁸ The BIS adopts a less common approach, in that after each variable is converted to a score from a set scale, the scores are aggregated by summing using judgmental weights.

²⁹ There are also a number of new approaches that are being explored in the literature. For example, Burkart and Coudert (2000) use linear discriminant analysis; Vlaar (1999) and Fratzcher (1999) develop switching regime models; and Osband and Van Rijckeghem (2000) use non-parametric methods to identify safe zones.

the signaling territory, making interpretation difficult.³⁰ A probit regression addresses many of the problems with the indicator approach: it generates predictions considering the correlations between all the predictive variables, and allows testing of the statistical significance of individual variables. However, since the probit is a nonlinear model, the contribution of a particular variable depends on the magnitude of all the other variables. This makes the relationship between changes in the variables themselves and changes in their contribution to the crisis prediction not always transparent. In the final analysis, the relative merits of the two approaches are decided by one key factor: how successful is each method in predicting crises?

Forecasting horizon

Another important design issue for models that attempt to predict both the cross-country incidence and timing of crises is how far in advance the prediction is to be made. Both the KLR and the DCSD model do not attempt to predict the exact timing of the crisis (which may be much harder or impossible), but rather the likelihood that a crisis will occur sometime in the following 24 months. The relatively long prediction window could be useful for the Fund because it would permit sufficient lead time for the authorities to make some policy adjustments. In fact, research on the DCSD model has indicated relatively little difference in the estimated model using any horizon between nine months and two years.

Private sector models tend to attempt predicting the probability of a crisis over a shorter horizon, from one to three months. Some investment banks provide weekly updates of crisis predictions to their clients, although only a small subset of the variables change at this frequency. This prediction horizon clearly relates to these firms' objectives of providing advice to clients participating mainly in foreign exchange markets, who may use shifting short-term forecasts to continually adjust their portfolios or hedge their positions. Different sets of variables may be important predictors at short horizons. For example, the three private sector models tracked by staff all include a measure of contagion in the model, reflecting the fact that contagion can occur relatively rapidly in emerging markets. Both changes in stock prices and domestic credit to the private sector are also found to be important predictive variables in all three private sector models.

³⁰ The Goldman-Sachs GS-Watch model also uses predictive indicators in zero/one form, but these are used as regressors in a logit model. Therefore, the probabilities are less "jumpy" than in the KLR indicators model.

MEANING OF IN-SAMPLE AND OUT-OF-SAMPLE PERIODS IN EWS MODELS

The text emphasizes the distinction between in-sample and out-of-sample performance. This appendix defines these terms and explains their implementation in this paper. The designer of an EWS chooses the variables and estimates the parameters of the model so as to best fit the observations in a particular sample (the estimation sample). In-sample testing measures how well the models fit the crises in this sample, i.e. *in-sample*. Good in-sample testing is a sign of a useful model but must be interpreted cautiously. The good in-sample performance may be a coincidence, perhaps resulting from a search through a large number of specifications until a good fit occurs by chance. Moreover, the determinants of crises may vary through time.

In out-of-sample testing, the predictions of an existing model are compared to a new set of observations, not belonging to the estimation sample. An unavoidable difficulty with out-of-sample testing is that a forecast can only be properly judged after the entire forecast window has closed. This paper examines the forecasts through June 1999 for the KLR and DCSD models, since it is too early to fully judge more recent forecasts. A prediction of risks as of August 1999, for example, cannot be fully judged until September 2001. Before then, it is not known whether August 1999 was in fact a pre-crisis or a tranquil month, since it is not yet known whether a crisis followed within 24 months. Given the two-year model horizon, these forecasts apply to realizations during a two-and-a-half year period, through July 2001. For the private sector GS and CSFB models, which have three-month and one-month horizons respectively, it is possible to look at goodness-of-fit through April and August 2001.³¹

Out-of-sample testing should mimic the process by which a forecasting model would be used in practice. In the strictest and most interesting form of out-of-sample testing, the modeler has no knowledge of the out-of-sample observations when generating the forecasts to be tested. Sometimes, in contrast, the modeler may withhold the most recent observations from the estimation sample, using them for subsequent out-of-sample testing. The modeler may none-the-less use information from these observations to create the model. For example, the DCSD model was estimated over the pre-Asia-crisis period and used to predict the Asia crises out-of-sample in OP 186. However, the authors created the model in 1998, after the Asia crises, and they added the short-term debt/reserves variable because they knew it was likely important in explaining the Asia crises.

³¹ Similarly, in-sample estimation periods for KLR and DCSD must end some 24 months before the time the model is being estimated, as it is not yet known whether later observations are pre-crisis or tranquil periods. For example, the in-sample period for the DCSD model in OP 186 ended in May 1995 so that the estimation did not reflect knowledge of the Asia crises that began in July 1997.

Table 7. Model Samples

	Sample
Asia crisis	
<i>In-sample</i>	
KLR/DCSD	1985:12 to 1995:4
<i>Out-of-sample</i>	
KLR/DCSD	1995:5 to 1996:12
Recent experience	
<i>In-sample</i>	
KLR/DCSD	1985:12 to 1997:5
GS	1996:1 to 1998:12
CSFB	1994:1 to 2000:7
<i>Out-of-sample</i>	
KLR/DCSD	1999:1 to 2000:12
GS	1999:1 to 2001:4
CSFB	2000:8 to 2001:8

Source: Author's calculations.

The start dates for the out-of-sample periods examined in this paper are chosen to be after the date at which they could have informed the estimation of the models. The KLR and DCSD forecasts examined here, for the 1999:1 to 2000:12 period, correspond to the versions used for the “official” internal July 1999 forecasts and subsequent internal forecasts. The model specifications were finalized in late-1998. The GS out-of-sample forecasts come directly from contemporary monthly publications over the 1999:1 to 2001:4 period, so they could not have reflected out-of-sample information. The CSFB out-of-sample estimates for the April 2000 to June 2001 period were produced in August 2001 using the model as it had been estimated a year earlier, so in principal they should not have been influenced by out-of-sample events.

REFERENCES

- Abiad, Abdul, 2003, "Early Warning Systems: A Survey and a Regime-Switching Approach," Working Paper No. 03/32 (Washington: International Monetary Fund).
- Ades, Alberto, Rumi Masih, and Daniel Tenengauzer, 1998, "Gs-Watch: A New Framework for Predicting Financial Crisis in Emerging Markets," (New York: Goldman Sachs).
- Berg, Andrew, Eduardo Borensztein, Gian Maria Milesi-Ferretti, and Catherine Pattillo, 1999, "Anticipating Balance of Payments Crises: The Role of Early Warning Systems," IMF Occasional Paper 186 (Washington, D.C.: International Monetary Fund).
- Berg, Andrew, and Catherine Pattillo, 1999a, "Are Currency Crises Predictable? A Test," *IMF Staff Papers*, Vol. 46, Issue 2 (June), pp. 107–38 (also issued as IMF WP 98/154 and published in popularized form in "The Challenge of Predicting Economic Crises," *Economic Issues* No. 22, Washington, D.C.: International Monetary Fund.).
- , 1999b, "Predicting Currency Crises: The Indicators Approach and an Alternative," *Journal of International Money and Finance*, Vol. 18, Issue 4 (August), pp. 561–86.
- , 1999c, "What Caused the Asian Crises: An Early Warning System Approach," *Economic Notes*, Vol. 28, Issue 3 (November).
- , and Rebecca Coke, 2004, "Autocorrelation-Corrected Standard Errors in Panel Probits: An Application to Currency Crisis Prediction," Working Paper No. 04/39 (Washington: International Monetary Fund).
- Burkart, Olivier, and Virginie Coudert, 2000, "Leading Indicators of Currency Crises in Emerging Economies," Notes d'Etudes et de Recherche # 74 (Paris: Banque de France).
- Detragiache, Enrica, and Antonio Spilimbergo, 2001, "Crises and Liquidity: Evidence and Interpretation," Working Paper WP/01/02 (Washington, D.C.: International Monetary Fund).
- Ferri, G., L. G. Liu, and J. E. Stiglitz, 1999, "The Procyclical Role of Rating Agencies: Evidence from the East Asian Crisis," *Economic Notes*, Vol. 28, Issue 3 (November), pp. 335–55.
- Garber, Peter M., Robin L. Lumsdaine, and Marco van der Leij, 2000, "Deutsche Bank Alarm Clock: Forecasting Exchange Rate and Interest Rate Events in Emerging Markets," (New York: Deutsche Bank).
- Goldfajn, Ilan, and Rodrigo O. Valdes, 1998, "Are Currency Crises Predictable?," *European Economic Review*, Vol. 42, Issue 3–5 (May), pp. 873–85.

- Goldstein, Morris, Graciela L. Kaminsky, and Carmen M. Reinhart, 2000, *Assessing Financial Vulnerability: An Early Warning System for Emerging Markets* (Washington, D.C.: Institute for International Economics).
- Harding, Don, and Adrian Pagan, 2003, "Synchronization of Cycles," manuscript (Victoria: University of Melbourne).
- Hemming, Richard, Michael S. Kell, and Axel Schimmelpfennig, 2003, "Fiscal Vulnerability and Financial Crises in Emerging Market Economies," IMF Occasional Paper No. 218 (Washington, D.C.: International Monetary Fund).
- IMF, 2002, "Early Warning System Models: The Next Steps Forward." in *Global Financial Stability Report (March)* (Washington: International Monetary Fund).
- Judge, George G., William E. Griffiths, R.Carter Hill, and Tsoung-Chao Lee, 1980, *The Theory and Practice of Econometrics* (New York: John Wiley and Sons).
- Kamin, Steven B., John W. Schindler, and Shawna L. Samuel, 2001, "The Contribution of Domestic and External Factors to Emerging Market Devaluation Crises: An Early Warning Systems Approach," International Finance Discussion Paper No. 711 (Washington, D.C.: Board of Governors of the Federal Reserve System).
- Kaminsky, Graciela L., 1999, "Currency and Banking Crises: The Early Warnings of Distress," Working Paper WP/99/178 (Washington: International Monetary Fund).
- , Saúl Lizondo, and Carmen M. Reinhart, 1998, "Leading Indicators of Currency Crises," *International Monetary Fund Staff Papers*, Vol. 45, Issue 1 (March), pp. 1–48.
- Levy-Yeyati, Eduardo, and Federico Sturzenegger, 2001, "Exchange Rate Regimes and Economic Performance," *IMF Staff Papers*, Vol. 47, Issue 0 (Special), pp. 62–98.
- Manasse, Paolo, Nouriel Roubini, and Axel Schimmelpfennig, 2003, "Predicting Sovereign Debt Crises," forthcoming working paper (Washington: International Monetary Fund).
- Meese, Richard A., and Kenneth Rogoff, 1983, "Empirical Exchange Rate Models of the Seventies: Do They Fit out of Sample?," *Journal of International Economics*, Vol. 14, Issue 1–2 (February), pp. 3–24.
- Mulder, Christian, Roberto Perrelli, and Manuel Rocha, 2002, "The Role of Corporate, Legal and Macroeconomic Balance Sheet Indicators in Crisis Detection and Prevention," Working Paper WP/02/59 (Washington: International Monetary Fund).
- Reinhart, Carmen M., 2002, "Default, Currency Crises, and Sovereign Credit Ratings," *World Bank Economic Review*, Vol. 16, Issue 2 pp. 151–70.

- Roy, Amlan, and Maria M. Tudela, 2000, "Emerging Market Risk Indicator (Emri): Re-Estimated Sept 00," Amlan Roy and Maria M. Tudela ed. (New York: Credit Suisse/First Boston).
- Sy, Amadou, 2003, "Rating the Ratings Agencies: Anticipating Currency Crises or Debt Crises," Working Paper No. 03/122 (Washington: International Monetary Fund).
- Vlaar, Peter J.G., 1999, "Currency Crisis Models for Emerging Markets," (Amsterdam: De Nederlandsche Bank).