Answering the Queen: Online Machine Learning and Financial Crises *

Jérémy Fouliard
INSEE and PSE

Michael Howell
CrossBorder Capital

Hélène Rey
LBS, CEPR and NBER

Vania Stavrakeva
LBS and CEPR

November 24, 2026

Abstract

Financial crises cause economic, social and political havoc. Macroprudential policies are gaining traction but are still severely under-researched compared to monetary and fiscal policy. This paper designs an aggregation procedure for probabilistic forecasts that issues a warning up to twelve quarters before a financial crisis, identifies in real time which channel of vulnerability drives the warning, and comes with a formal performance guarantee of the type established in the online learning literature. Applied in real time to France, Germany, Sweden and the United Kingdom from 2001 to 2025, with models estimated on the crises of the early 1990s only, the procedure raises an alarm before each of the four crises of 2007–2008, with leads of one to five years, and separates pre-crisis from calm quarters with an out-of-sample AUROC between 0.83 and 0.96.

*We thank Francesco Amodeo, Anja Saskia Brüggen, Matthias Herp and Samantha Ong for excellent research assistance. We are highly indebted to Pierre Alquier, Pierre Gaillard, Francesco Orabona, Gilles Stolz and Olivier Wittenberg for their generous help. We are also very grateful to our discussants Fabrice Collard, Frank Diebold, Christian Julliard, Sebnem Kalemli Ozcan, Luc Laeven and Philip Lane as well as to Richard Portes and Nicolas Vayatis and seminar participants at the NBER IFM meeting, the ASSA Meetings, the Society for Nonlinear Dynamics and Econometrics Meetings, the Applied Machine Learning, Economics and Data Science seminar, UC Berkeley, MIT, Boston University, Cambridge Judge, Oxford, Bocconi, ENSAE, KTO-GREDEQ-OFCE, Nowcasting, the LBS AQR academic symposium, the Annual BIS Conference, Chicago conference in the honour of Alejandro Justiniano, the Bank of Spain Macroprudential Conference and the Women in Economics Chicago Conference for comments. Our thanks also go to Virginie Coudert, Julien Idier, Mikael Juselius, Anna Kovner, Marco Lo Duca, to Benoit Mojon and to Katja Taipalus for help with the database and Pierre Gaillard and Yannig Goude for the opera package. Rey thanks the ERC (Advanced Grant 695722). Jérémy Fouliard benefited greatly from the hospitality of the BIS where part of this paper was written. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the institution. Email: hrey@london.edu; jeremy.fouliard@insee.fr.

1 Introduction

In November 2008, the Queen of the United Kingdom visited the London School of Economics. After the failure of Lehman Brothers in September, the financial crisis was on everyone's mind. As she was shown graphs emphasising the scale of imbalances in the financial system, she asked a simple question: "Why didn't anybody notice?"After a rather terse reply on the spot[1] it took several months before the British Academy wrote a three-page missive to Her Majesty blaming the lack of foresight of the crisis on the "psychology of denial" that was widespread in financial and political circles who tended to believe that "financial wizards had found new and clever ways of managing risks". "So in summary, Your Majesty, the failure to foresee the timing, extent and severity of the crisis and to head it off, while it had many causes, was principally a failure of the collective imagination of many bright people, both in this country and internationally, to understand the risks to the system as a whole." This paper is an attempt to bring back some imagination in the economics of crises.

Financial crises cause economic, social and political havoc. The average cumulative output loss in a banking crisis (in deviation from its trend) is around 20% over the length of the crisis, which is on average two years, according to the database of Laeven and Valencia (2020). Systemic financial crises lead to large fiscal costs, major increases in public debt and they disrupt the fabric of our societies. In order to decrease their frequency and their severity a new set of tools has been introduced in many countries. Macroprudential policies aim at increasing the resiliency of the financial system as a whole by introducing countercyclical capital buffers for banks, liquidity coverage ratio requirements and allowing for tightening of lending standards at discretionary times chosen by the macroprudential authorities. While there is an extensive body of academic research on monetary and fiscal policies, there is still relatively little work which can guide macroprudential policies. In particular, implementing those policies requires a timely understanding of the build up of risk in the economy. As shown in the classic Reinhart and Rogoff (2009) book "This Time is Different, Eight hundred years of Financial Follies", financial crises have occurred repeatedly in emerging markets and advanced economies alike, and they exhibit some remarkable similarities. In a survey on financial crises, Sufi and Taylor (2021) emphasize that since crises do not occur randomly, it is important to understand better the pre-crisis periods. Crises are often, but not always, "credit booms gone bust" but they also display some differences in their

[1] In private correspondance Luis Garicano, who was Professor of economics at LSE indicated that he actually gave the following answer to the Queen at the time: "At every stage, someone was relying on somebody else and everyone thought they were doing the right thing." See also Financial Times, November 14 2008.

mechanics. There are many different theoretical models in macroeconomics and in finance which have been developed to understand them. Some emphasise runs as in Diamond and Dybvig (1983). Most macrofinance models focus on the bust phase of the crisis and on amplification mechanisms. A few analyze the boom phase of the financial cycle: they emphasise limited liability and asset overvaluations due to risk-shifting (Coimbra and Rey (2017)), search-for-yield in low interest rates environments (Martinez-Miera and Repullo (2017)), or deviations from rational expectations and financial constraints (Gennaioli et al. (2012)).

From an empirical point of view, a number of variables have been used to predict financial crises. Following the classic work of Kaminski and Reinhart (1999), the literature has very usefully described the behaviour of a number of key variables around crisis episodes (see e.g. Gourinchas and Obstfeld (2012)). Lowe and Borio (2002) and Schularick and Taylor (2012) underline the role of credit growth; Jordà et al. (2015) emphasize the joint importance of credit growth and asset prices; Mian and Sufi (2009) show the critical importance of household debt and Krishnamurthy and Muir (2025) shows that the interaction between credit losses and financial sector is key to understand financial crises. Most of the literature uses standard econometric methods such as panel data econometrics or event studies in order to identify early warning indicators (EWI) of financial crises[2]. Some recent attempts to introduce new forecasting methods from the machine learning literature can be found in Ward (2017) who uses classification trees. Several papers evaluate the performance of machine learning models such as decision trees, random forests, extremely randomised trees, support vector machines, or artificial neural networks(Beutel et al. (2019), Bluwstein et al. (2020), Du Plessis and Fritsche (2025)). Bi et al. (2024) recently question the role of Artificial Intelligence (AI) in financial forecasting.

Our starting point is that the ability of existing models to predict systemic crises out-of-sample early and accurately (with small type I and type II errors i.e. the ability to predict all crises which actually happened without crying wolf too often) is still limited. Turning points and non linear phenomena such as crises have been notoriously difficult to predict out-of-sample. Price-based early warning indicators tend to be more coincident indicators than good predictors. Predicting pre-crisis periods [3]in order to give macroprudential authorities the time to act proves to be ex-

[2]EWIs have been developed to predict financial crises in emerging economies (Frankel and Rose (1996); Bussiere and Fratzscher (2006)). They were also applied to large panels of advanced and emerging economies (Demirgüç-Kunt and Detragiache (1998); Eichengreen and Arteta (2002); Bordo et al. (2001)). After the 2008 crisis, a number of studies have shown the relevance of EWIs using univariate or multivariate regressions (see e.g. Borio and Drehmann (2009), Alessi and Detken (2011), Shin (2013), Frankel and Saravelos (2012), or model aggregation methodologies (Coudert and Idier (2016)). None has considered robust and agnostic aggregation methodologies like our paper.

[3]defined as the period twelve quarters before the crisis in the Early Warning Systems literature, see Detken et al. (2014)

tremely difficult. Yet financial stability policies need this type of input. The complexity and the interaction of many variables, some of them -like asset prices- very fast moving, may also render the understanding of financial crises exceptionally difficult. In such a context, the "failure of the collective imagination of many bright people" is likely to be a permanent feature of the world.

Our answer to the queen borrows from two mature bodies of statistical theory that has, to our knowledge, not been used to aggregate early-warning signals: online learning theory [4]and the sequential change-point detection framework. Ideally, we would like to forecast systemic financial crises without knowing the "true" model of the economy - if it exists, using as much information as possible (many possible models of the economy) in a way which is flexible enough to do dynamic evolving forecasting (weights put on different predictive models should vary over time). We also want to avoid the problem of overfitting which is a common issue in macroeconomic forecasting (Stock and Watson (1996)) and leads to poor out-of-sample forecasting power (see e.g. Meese and Rogoff (1983) and Rossi (2011)). Our framework first draws on online procedures in which forecasts from multiple models are aggregated in real time. The appeal of this approach is threefold. First, the aggregation scheme is provably near-optimal in a precise sense: the aggregated forecast is guaranteed to perform nearly as well as the best model in the pool—whichever model that may be—with the gap vanishing as the number of observations grows. Second, these guarantees are distribution-free: no assumption is imposed on the data-generating process, which may be non-stationary or even adversarially generated. Third, because the weights assigned to each model are updated sequentially as new observations arrive, the framework delivers a natural and interpretable measure of which model performs best at each point in time, and of how the relevant model evolves. All of these features are critical for the problem of forecasting financial crisis and the implementation of macroprudential policies.

Online learning has been a main topic of research in the theory of machine learning and has become the infrastructure of large-scale forecasting systems. It is used in many applications such as to forecast electricity consumption (Devaine et al., 2013)) and price (Mahtout and Ziel (2026)), to study sequential games such as poker games (Lee et al. (2021)), for real-time climate and weather forecasting (Flaspohler et al. (2021)), in physics and medicine (Malenica et al. (2023)), modeling network traffic(Dashevskiy and Luo (2011), predicting air quality (Mallet et al., 2009), for industrial scale recommendation (Yi et al. (2023) or the study of data stream (Althabiti et al. (2024). To our knowledge, online machine learning has not yet been applied to macroeconomics,

[4]The term online refers to the fact that the information is revealed in a sequential order

with the notable exception of Amat et al. (2018) for exchange rate forecasting[5]. However, it shares several distinctive features with many fields in economics. Since the seminal work of Bates and Granger (1969), forecast combinations are viewed as a simple and effective way to perform better than individual models (Timmermann (2006), Elliott and Timmermann (2008), Diebold and Shin (2019)). [6] However, online machine learning is different from the real-time forecast combination literature as it does not estimate the coefficients of some underlying model, but rather, chooses the best combination of weights to form the most accurate out-of-sample predictions sequentially *without making assumptions on the underlying data generating process* [7]. It is robust with respect to this out-of-sample forecast process and not with respect to an underlying stochastic process that it would have to fit (since this is not what it does)[8].

The early-warning problem is, at its core, a problem of detecting a change in real time. The policymaker observes these indicators one period at a time and must decide, at each date, whether the evidence accumulated so far is strong enough to call a regime change, knowing that a late alarm is costly and that a false alarm has a cost too. However existing early-warning frameworks test each quarter against a threshold, without accumulating evidence across quarters. This is precisely the key objective of sequential change-point detection, also known as quickest detection: to minimize the delay between the occurrence of a change and its detection, subject to a constraint on the frequency of false alarm (Page (1954), Lorden (1971)), Pollak (1985)). Indeed sequential change-point detection is one of the most widely used statistical tools in the monitoring of critical systems These procedures have several properties that we find appealing. First, it does

[5]While applications in economics remain limited, online learning has been extensively used in finance. A well-established literature studies sequential investment problems, starting with the seminal contribution of Cover (1966). This line of research re-emerged in the 1990s with Shafer and Vovk (2005) and has since generated a substantial body of work (Jun et al., 2017; Jang et al., 2023).

[6]In the presence of structural change, Diebold and Pauly (1987) and Pesaran and Timmermann (2005) argue that forecast errors can be reduced through systematic combination of forecasts. Capistrán and Timmermann (2009) shows how to combine individual survey forecasts (with entry and exit of individual forecasters) for real time forecasting. Diebold and Shin (2019) proves that a LASSO-based procedure that sets some weights to zero and shrink the surviving models towards equality outperform simple average and median forecasts (and perform almost as well as the ex post best forecaster) in an application to the European Central Bank Survey of Professional forecasters.

[7]This is why this approach is described as "meta-statistical"

[8]Online learning theory is deeply related to the foundations of the game-theoretic approach to prediction established by Hannan (1957) and Blackwell (1956) since some of the forecasters developed can be used to play zero sum repeated games. Cesa-Bianchi and Lugosi (2006) show that simple bounds for the performance of online algorithms can be seen as applications of the classical minimax theorems in game theory and reciprocally generalized minimax theorems such as Blackwell approachability theorem can be used to define good online forecasters. Even more surprisingly, a fascinating line of research shows that if all players in a repeated normal form game adopt a simple regret-minimizing prediction strategy similar to online learning algorithms, the induced dynamics leads to a certain equilibrium (see for instance Fudenberg and Levine (1995), Hart and Mas-Colell (2000), Hart and Mas-Colell (2001), Stoltz and Lugosi (2007)). It lies at the core of many novel applications in the theory of machine learning such as solving convex–concave min–max problems ((Liu and Orabona, 2022)), analyzing Generative Adversarial Networks (GANs) ((Liang and Stokes, 2019)) or multi-armed bandit algorithms ((Lee and Orabona, 2025)). It offers tools make progress on fundamental problems in statistics and game theory such as learning linear predictors with the logistic loss ((Foster et al., 2018)), stochastic convex optimization ((Cutkosky, 2019)), and the solution of large-scale games ((Daskalakis et al., 2021)).

not require feedback. A detection statistic is computed from the stream of signals alone, without waiting to observe whether past signals were correct. This matters because the feedback on early-warning signals arrives only after the pre-crisis window has elapsed, that is, after the crisis has started. Second it accumulates evidence over time. Existing early-warning systems issue an alarm when the current value of an indicator, or the current predicted probability of a model, crosses a threshold. Since Page (1954), it has been known that a procedure that accumulates evidence detects them earlier for the same rate of false alarms. Third, these algorithms are well suited to aggregating stream-specific statistics into a single global procedure when the affected stream is unknown ex ante. This feature is particularly relevant for our purpose: crises preceded by credit booms, housing booms, or a turn in the global financial cycle need not manifest in the same data streams, so a detection procedure that requires specifying in advance where the break will occur is bound to miss some of them (Tartakovsky and Veeravalli (2008), Xie and Siegmund (2013)). Sequential detection has also been extensively used in many applications. It has long been used to detect abnormal behavior in nuclear power plants (Gross and Humenik, 1991). It has also been used in many recent applications. For instance Chen et al. (2022) develop an online procedure for high-dimensional, multiscale changes and apply it to seismology data. Romano et al. (2023) propose a fast online procedure for changes of unknown size and use it to detect anomalous behavior in computer server data. [9]. To our knowledge, sequential change-point detection has not been used to aggregate early-warning signals for financial crises, nor to address the inherent delayed feedback that is intrinsic to this problem. However, several fields in economics address related questions with related methodologies such as the monitoring of econometric models (Giacomini and Rossi (2009)), the detection of business-cycle turning points (Diebold and Rudebusch (1989)) [10] and the real-time detection of asset-price bubbles (Phillips et al. (2015)). None of them, however, has been brought to bear on the prediction of financial crises.

This paper designs an aggregation procedure for probabilistic forecasts that issues a warning up to twelve quarters before a financial crisis, identifies in real time which channel of vulnerability drives the warning, and comes with a formal performance guarantee of the type established in the online learning literature. We build one forecasting model per theoretical channel of vulnerability: credit, real estate, external imbalances and the global financial cycle. Each is a penalized logit re-estimated at every date using only the information available at that date. We estimate these

[9] Another example is during the COVID-19 pandemic where quickest-detection methods were used to monitor epidemic dynamics in real time (Soldi et al. (2021), Zamba and Tsiamyrtzis (2021)).

[10] Markov-switching models which are the standard tool for real-time recession dating can be seen as a parametric counterpart of this approach.

predictive models using standard macroeconomic variables [11]. We then treat each model as a sensor, a source of observations that is monitored over time, and whose distribution may change at an unknown date. Each sensor accumulates evidence through a bounded cumulative-sum statistic, and the aggregate forecast weights each model by the posterior probability that its channel is the one currently signaling a regime change. Because this posterior responds to the models' current signals rather than to their past forecast errors, it does not need to wait for feedback that, in this problem, arrives only after the crisis has started. Two components complement the detection rule: a learning layer that updates the weights as feedback is revealed based on predictive models' past performance, and a correction that penalizes models that raise alarms in calm periods. Finally, following the standard online learning approach, a meta-aggregator guarantees that, whatever the data, the procedure performs almost as well as the detection rule and, asymptotically, as well as the best individual model. The guarantee holds under delayed feedback and its bound does not grow with the length of the sample.

We apply the procedure to France, Germany, Sweden and the United Kingdom from 1987 to 2025. All parameters are estimated or fixed using information available before 2001, and the system is then run in real time for nearly twenty-five years. Three sets of results stand out. First, the aggregate signal rises above the alarm threshold calibrated on pre-2001 data at least 9 quarters before the onset of the 2007–2008 crises in all four countries, with an average out-of-sample AUROC of 0.88. Second, the channel that carries the alarm differs across countries, even for crises with a common trigger: overvalued housing in France, Sweden and the United Kingdom, and the global equity cycle in Germany. We read this result as one of predictability, not causality. The crises of 2007–2008 originated largely in the losses on US mortgage-related assets and in the collapse of wholesale funding, and outside the United Kingdom the banks that failed did so because of their exposures abroad rather than because of domestic housing or credit. The channels the procedure identifies are therefore not the causes of the crises but the domestic markers of the phase of the financial cycle in which banks became vulnerable to such a shock. Because these markers differ across countries and cannot be known in advance, an early-warning system must weigh them by the evidence they currently carry rather than by their past performance, which is what our procedure does. Third, the gains come from detection rather than learning: without the change-point component, the aggregate still ranks quarters well, with a higher AUROC, but its level before crises falls by half, one of the four crises is missed and the policymaker's utility

[11] These variables are the ones which would have come naturally to the mind of any macroeconomist familiar for example with the important work of Kindelberger on *Manias, Panics and Crashes* (Kindleberger (1978)) or reading the *debt-deflation theory of great depressions* (Fisher (1933))

turns negative, whereas removing the learning layer leaves the results essentially unchanged. The results are robust to an eight-quarter pre-crisis horizon and to a real-time "vintage" specification that excludes series heavily revised by statistical agencies or constructed with full-sample information: the four crises remain detected, through the same channels and, within each channel, through the same indicators with the same signs.

The structure of the paper is as follows. We present our database on systemic financial crisis dates as well as the different variables which we use to build the predictive models in section 2. Section 3 describes the informationnal structure of the procedure. Section**??** the present the general procedure we use to answer the Queen and section **??** the results. Section 7 shows that our results are robust to changes in the tuning parameters, to an eight-quarter pre-crisis horizon, and to a real-time "vintage" specification.8 concludes.

2 Data and predictive models

Our framework relies on two datasets: (i) a chronology of systemic crisis episodes and (ii) a panel of macroeconomic and financial indicators used to estimate predictive models. We focus on four countries—France, the United Kingdom, Germany, and Sweden—for which at least two systemic banking crises are recorded over the period 1987–2025 and sufficiently comprehensive quarterly data are available to construct the predictors. The starting date of our dataset is chosen as the earliest date for which the maximum number of variables have non-missing observations prior to the first systemic crisis episode. [12]

For the first part of the dataset, we borrow the definition and the dates of systemic crises from the Official European database constructed by the European Central Bank and the European Systemic Risk Board (Lo Duca et al. (2017)). We also rely on their narratives of the crises. This database has been put together to establish a common ground for macroprudential oversight and policymaking in the European Union. The dating of systemic crises is in part based on quantitative indicators but it is ultimately based on the predictive model judgment of the relevant national authorities. Section 9.1.1 in Appendix describes the methodologies used to construct this database of financial crises.

For the second part, we build one elastic-net regression per theoretical channel of vulnerability: credit, real estate, external imbalances and the global financial cycle. Each of them includes 4

[12] This sample selection reflects both the historical incidence of crises and data availability. We exclude the United States and Spain because the official database used in our analysis records no systemic banking crisis episodes for these countries in the 1980s. Italy is excluded because of limitations in the availability of quarterly data during the same period.

variables with transformations (1-y, 2-y, 3-y change and gap-to-trend). The exact composition and the way they are estimated [13] is described in Section 4.2. We emphasize that the variables of our analysis are the ones which would have come naturally to the mind of the reader of Kindleberger (1978)[14] or of Minsky (1986)[15]. But really, these same variables would be considered by anyone reading in 1933 in Econometrica the *debt-deflation theory of great depressions* by Irving Fisher[16]. We do not deny that in the set of the exact measures we use some of them would not have been available historically but most of them (and actually the ones that tend to matter) would have been; and the economic concepts that all these variables measure were the ones described by this classic literature.

3 Informationnal structure

3.1 Notation

We first start by some notations. Let $i \in \mathcal{I} = \{1, \ldots, N\}$ index countries and $t \in \{1, \ldots, T\}$ quarters. Each country is observed through a vector of variables dated by the quarter t to which

[13] A more detailed description of the variables is provided in Section 9.1.2 in Appendix

[14] "By no means does every upswing in business excess lead inevitably to mania and panic. But the pattern occurs sufficiently frequently and with sufficient uniformity to merit renewed study. What happens, basically, is that some event changes the economic outlook. New opportunities for profits are seized, and overdone, in ways so closely resembling irrationality as to constitute a mania. Once the excessive character of the upswing is realized, the financial system experiences a sort of "distress," in the course of which the rush to reverse the expansion process may become so precipitous as to resemble panic. In the manic phase, people of wealth or credit switch out of money or borrow to buy real or illiquid financial assets. In panic, the reverse movement takes place, from real or financial assets to money, or repayment of debt, with a crash in the prices of commodities, houses, buildings, land, stocks, bonds -in short, in whatever has been the subject of the mania".

[15] "The economy consists of a mixture of hedge, speculative and Ponzi financing units. A hedge financing unit can fail to meet its obligations only if its gross profits after taxes fall below expectations. In the aggregate this can happen only if there is a sharp fall in aggregate demand. A speculative financing unit can fail to meet its obligations if its income is below expectations, if interest rates rise too much or if there is a breakdown in the normal functioning of some set of financial markets. A Ponzi financing unit can run into troubles for all of the reasons that a speculative unit can plus the capitalizing of interest can erode the margin of safety in equity so that lenders are unwilling to continue capitalizing interest. An economy in which the dominant financing form is hedge financing will be financially robust. The greater the proportion of firms that are speculative or Ponzi financing the more fragile the financial structure. The basic theorem of the financial instability hypothesis is that over an extended period of prosperous times the weight of speculative and Ponzi finance in the total financial picture increases, so that the economy migrates from being financially robust to being financially fragile".

[16] "While quite ready to change my opinion, I have, at present, a strong conviction that these two economic maladies, the debt disease and the price-level disease (or dollar disease), are, in the great booms and depressions, more important causes than all others put together. Some of the other and usually minor factors often derive some importance when combined with one or both of the two dominant factors. Thus over-investment and over-speculation are often important; but they would have far less serious results were they not conducted with borrowed money. That is, over-indebtedness may lend importance to over-investment or to over-speculation. The same is true as to over-confidence. I fancy that over-confidence seldom does any great harm except when, as, and if, it beguiles its victims into debt. Another example is the mal-adjustment between agricultural and industrial prices, which can be shown to be a result of a change in the general price level. Disturbances in these two factors, debt and the purchasing power of the monetary unit, will set up serious disturbances in all, or nearly all, other economic variables. On the other hand, if debt and deflation are absent, other disturbances are powerless to bring on crises comparable in severity to those of 1837,1873, or 1929-33" Fisher (1933).

they refer, and a binary systemic financial crisis indicator $C_{i,t} \in \{0,1\}$ equal to one when a systemic financial crisis is ongoing in quarter t.

Our object of interest is not the crisis itself but its beginning. We call onset the first quarter of a crisis episode, that is, a quarter in which a crisis is ongoing while none was in the previous quarter. The set of onsets for country i is :

$$\mathcal{O}_i = \{\tau : C_{i,\tau} = 1,\ C_{i,\tau-1} = 0\}. \tag{1}$$

Throughout $\Lambda(z) = (1 + e^{-z})^{-1}$ denotes the logistic function and $\text{logit}(p) = \log \frac{p}{1-p}$ its inverse. We consider the clip function which are used in Section 5.2.2. For $a \leq b$, $\text{clip}(x, a, b) = \min\{\max\{x, a\}, b\}$ denotes the projection of x onto $[a, b]$.

We assume that indicators are published with a delay: the value of $x_{i,t}$ is released in quarter $t + 1$. All series are therefore lagged by one quarter, and

$$\mathcal{F}_{i,t} = \sigma\big(x_{i,1:t-1},\ C_{i,1:t}\big) \tag{2}$$

denotes the information available about country i at date t.

We first make Assumption A1, which is a standard convention in the early-warning literature, where crisis dates are taken from an ex-post chronology (Lo Duca et al. (2017)). It is optimistic for episodes that were recognized with a lag.

Assumption A1 (Crisis dating)**.** $C_{i,t}$ is $\mathcal{F}_{i,t}$-measurable.The onset $\tau \in \mathcal{O}_i$ and the end of each crisis are observed in the quarter in which they occur.

3.2 The target

We define the pre-crisis indicator as the period occurring before the onset :

Definition 1 (Pre-crisis indicator)**.** For a horizon $h \in \mathbb{N}$,

$$y_{i,t}^{(h)} \ = \ \mathbf{1}\big\{\exists\, \tau \in \mathcal{O}_i :\ t < \tau \leq t + h\big\} \cdot \big(1 - C_{i,t}\big). \tag{3}$$

The pre-crisis indicator equals one in the h quarters preceding an onset and zero otherwise, including during crises.[17] It can be interpreted as the boom phase that precedes a systemic crisis,

[17] Coding crisis quarters as zero is standard in the early-warning literature (see, e.g., Bussiere and Fratzscher, 2006). The object of interest is the entry into a crisis, not its continuation, and counting crisis quarters among the positives would reward models that merely track contemporaneous stress. We write $y_{i,t}$ for $y_{i,t}^{(h)}$ when no confusion arises.

which is the regime a macroprudential authority wants to detect.

The choice of the horizon h reflects both a policy constraint and an economic phenomenon. On the policy side, an alarm is useful only if it leaves enough time to act. Under Basel III, an increase in the countercyclical capital buffer becomes binding twelve months after its announcement, so that a useful alarm must precede the onset by at least four quarters. Drehmann and Juselius (2014) accordingly evaluate early-warning indicators over horizons of five to twelve quarters, and Detken et al. (2014) use a twelve-quarter window in the early-warning framework of the European Systemic Risk Board. On the economic side, the build-up of vulnerabilities that precedes a crisis is a slow process. Schularick and Taylor (2012) show that credit growth over the previous five years predicts financial crises, and Greenwood et al. (2022) find that when credit and asset prices have grown rapidly over the previous three years, the probability of a crisis within the next three years approaches 40%.

The horizon also involves a statistical trade-off. A short horizon yields few positive observations, since each episode contributes only h pre-crisis quarters, which makes the predictive models hard to estimate on one or two episodes per country. A long horizon yields more positives but includes quarters that are distant from the onset and only weakly informative about it; and since the label of a calm quarter is revealed only after h quarters (Section 3.4), it also lengthens the delay with which the forecaster learns. Following the early-warning literature, we set $h = 12$ in the baseline, and Section 7 reports results for $h = 8$.

3.3 Sample definition

For every country, we split the sample into two periods with common dates : a batch sample and a online sample. The batch period runs from 1987Q1[18] to T_B = 2001Q1 and plays the role of a training sample: it is used to estimate the initial predictive models on the pooled panel, to standardize the indicators, to build the prior weights and to calibrate the alarm threshold. All the labels of the batch period are treated as known at T_B. This is exact for every quarter whose window ends before T_B [19]. From T_B onward, the procedure runs in real time: at each online date t, the predictive models are re-estimated on all the quarters of the batch period and on the online quarters whose label is resolved by t (see Section 3.4 for more a complete picture on how information is revelead during the online sample), and every forecast uses only the information

[18]Most of our macroeconomic variables start in the early 1980s at a quarterly level. Starting in 1987Q1 ensures that most of the macroeconomic variables are available with theirs transformations

[19]for the last h quarters of the batch, whose window extends beyond T_B, it amounts to treating them as provisionally negative, which coincides with their realized value in our sample, since no crisis starts between 2001Q2 and 2004Q1.

available at t. All the results reported in Section 9.11 are computed on the online period only.

The two periods differ in the information used to produce each forecast. In the batch period, which runs from 1987Q2 to T_B = 2001Q1, we stand at T_B and look back: the whole history up to T_B is available at once, and the initial models, the standardization moments, the prior weights and the alarm threshold are computed once on it. A forecast for a batch quarter s therefore uses the information of T_B, which includes what happened after s; it is an in-sample fit, not a real-time forecast, and we do not evaluate it. T_B is chosen so that the batch period contain at least one complete crisis episode per country, so that the predictive models can learn what a pre-crisis period looks like; the early-1990s banking crises in France, Sweden and the United Kingdomprovide these episodes. The online period must contain the 2007–2008 crises together with their full twelve-quarter pre-crisis windows, which start in 2004–2005; a split in 2001Q1 leaves a margin of several years of real-time calm quarters before the first pre-crisis window, so that false alarms are also tested out of sample.

In the online period, from 2001Q2 onward, we move forward one quarter at a time: the forecast of quarter t is issued with the information available at t only, that is, it is $\mathcal{F}_t$-measurable, before the outcome is known, and the models and weights are updated as labels are revealed. This distinction mirrors the one between batch learning, in which a model is fitted once on a given sample, and online learning, in which data arrive sequentially and the forecaster must commit to a forecast before each observation (Cesa-Bianchi and Lugosi, 2006). The batch period provides the starting point of the procedure; the online period reproduces what a policymaker could have done in real time, and all our results are computed on it.

3.4 Information structure and delayed feedback

The defining feature of the problem is that the label of quarter t is not observable at t. By Definition 1, $y_{i,t}$ depends on whether a crisis starts in the following h quarters, that is, on $C_{i,t+1}, \ldots, C_{i,t+h}$. Standing at the end of 2005Q1, a policymaker does not know whether 2005Q1 is a pre-crisis quarter: the answer depends on what happens until 2008Q1. The forecasting problem is therefore one of delayed feedback, and the first step is to determine when each label becomes known.

Definition 2 (Resolution date)**.** The resolution date $r_{i,t}$ is the first quarter at which $y_{i,t}$ becomes $\mathcal{F}_{i,r}$-measurable, that is, the first quarter at which its value can be computed from the information

then available. Under Assumption A1,

$$r_{i,t} = \begin{cases} \min\{\tau \in \mathcal{O}_i : \tau > t\} & \text{if } y_{i,t} = 1, \\ t & \text{if } C_{i,t} = 1, \\ t + h & \text{otherwise.} \end{cases} \tag{4}$$

The three cases correspond to three kinds of quarters. A pre-crisis quarter is recognized as such as soon as the crisis it precedes starts: its label is revealed at the next onset. A crisis quarter has a label of zero by definition, and since the crisis is observed, the label is known immediately. A calm quarter is recognized as calm only once its whole window has elapsed without a crisis: until then, a crisis could still start, and the label could still turn out to be one.

Consider, with $h = 12$, a crisis starting in 2008Q1. The labels of all the quarters from 2005Q1 to 2007Q4 are revealed at once in 2008Q1, with delays ranging from twelve quarters for 2005Q1 to one quarter for 2007Q4. During the crisis, labels are known immediately. A calm quarter such as 2003Q1 is resolved only in 2006Q1, twelve quarters later. The delay $r_{i,t} - t$ is therefore not constant: it is at most h, and it depends on the label itself.

The resolution date is not known in advance. At the end of 2007Q3, the forecaster does not know that 2007Q2 will be resolved in 2008Q1, since this depends on whether a crisis starts in the meantime; it knows only that $r_{i,s} \leq s + h$. What it can determine at any date t is whether a given label is already resolved: $r_{i,s} \leq t$ holds if the window of quarter s has fully elapsed, $s \leq t - h$, if an onset has been observed in $(s, t]$, or if s is a crisis quarter. This is all the training sets defined below require.

Lemma 1 (When the label becomes known)**.** *Under Assumption A1, for every quarter t:*

(i) at date $r_{i,t}$, the label $y_{i,t}$ is determined by the crises observed so far, $C_{i,1:r_{i,t}}$;

(ii) in the first and third cases of equation 4, it is not determined at any earlier date: for every $s < r_{i,t}$, there exist two crisis paths that coincide up to s and give different values of $y_{i,t}$.

In words, $r_{i,t}$ is the first date at which the label of quarter t is known. Part (i) states that the label is known at $r_{i,t}$; part (ii) states that it cannot be known earlier, since up to $s < r_{i,t}$ the information available is consistent both with a crisis that would make the label one and with no crisis, which would make it zero. The delay is therefore a feature of the problem, not of our procedure, and no forecaster can shorten it. The proof is in Appendix 9.3.1.

Lemma 1 has two consequences that drive the design of our procedure. First, feedback is

asymmetric. Positive labels are revealed in a block at the onset, up to h at once, whereas negative labels are revealed one at a time, each with a delay of h quarters. In calm periods, the forecaster thus learns steadily which models raise false alarms; it learns which models anticipated a crisis only when the crisis starts. Second, the information revealed at an onset concerns only the past. By the time a forecaster learns that quarters $\tau - h, \ldots, \tau - 1$ were pre-crisis quarters, the crisis has started and the warning is no longer useful for this episode. A procedure that learns only from revealed labels therefore learns which channel mattered once it no longer matters for the episode at hand. Section 5 builds on this observation.

The asymmetry of feedback has a further consequence for the learning component of the aggregation. During the build-up of a crisis, the labels revealed concern quarters that lie more than h quarters before the onset, which are calm by definition. A model that starts signalling slightly ahead of the window would thus be penalized as if it had raised a false alarm, precisely while the vulnerability builds up. The learning component therefore uses an effective label that tolerates signals issued slightly ahead of the window.

Definition 3 (Effective label and revelation date)**.** Let $H = h + 6$. For a quarter with $y_{i,t} = 0$ and $C_{i,t} = 0$, if an onset occurs in $(t, t + H]$, set $\tilde{y}_{i,t} = 1$; otherwise $\tilde{y}_{i,t} = y_{i,t}$. The revelation date is

$$\tilde{r}_{i,t} = \begin{cases} \min\{\tau \in \mathcal{O}_i : \tau > t\} & \text{if } \tilde{y}_{i,t} = 1, \\ t + H & \text{if } \tilde{y}_{i,t} = 0 \text{ and } C_{i,t} = 0, \\ t & \text{if } C_{i,t} = 1. \end{cases} \tag{5}$$

By construction, $t < \tilde{r}_{i,t} \leq t + H$ for non-crisis quarters, and $\tilde{y}_{i,t}$ is known at $\tilde{r}_{i,t}$, by the same argument as in Lemma 1. The six-quarter tolerance means that a model that raises an alarm 13 to 18 quarters before an onset is not penalized as if it had raised a false alarm. It reflects the view, common in the literature, that the length of the pre-crisis window is a convention, and that the build-up of vulnerabilities may start earlier in some episodes than in others. The effective label is used only by the aggregation: the predictive models are estimated on the true label $y_{i,t}$, and all evaluation statistics are computed on it.

Also the effective label never uses information earlier than the true label: positive labels are revealed at the onset in both cases, and negative labels are revealed later, after H rather than h quarters, since a quarter can be declared effectively calm only once no onset has occurred in the following H quarters. The tolerance is therefore paid for by a longer delay on negative labels, not by information from the future.

3.4.1 A sequential problem: predictive models, aggregation and timing

Section 3.4 established that the label of a quarter is revealed only after a delay. This section draws the consequences for the structure of the forecasting problem. The forecast is produced in two layers, a set of predictive models and an aggregation rule, and both are updated sequentially as labels are revealed. We describe what each layer knows at each date, how the weights of the aggregation evolve over the crisis cycle, and how forecasts are neutralized once a crisis has started.

Predictive models. A predictive model maps the information available at date t into a probability that quarter t is a pre-crisis quarter. We build $K = 4$ predictive models, one per channel of vulnerability (Section 4). Each predictive model is re-estimated every quarter on the labels resolved by t, so that its forecast changes over time both because its inputs change and because its parameters are updated.

Aggregation. The forecaster combines the predictive models' forecasts into a single probability,

$$\hat{p}_t = g_t \sum_{k=1}^{K} w_{k,t}\, p_{k,t}, \qquad w_{k,t} \geq 0, \quad \sum_{k=1}^{K} w_{k,t} = 1, \tag{6}$$

where $w_t = (w_{1,t}, \ldots, w_{K,t})$ is a vector of weights and $g_t \in [0, 1]$ a gate defined below (see Section 5.4). Restricting the forecast to a convex combination has two advantages. It keeps the forecast in $[0, 1]$ and on the same scale as the predictive models, so that the weights can be read as the relative importance of each channel. And it allows the forecaster to compete with the predictive models without modelling how they are generated. For the forecaster predictive models are treated as black boxes, possibly misspecified. Her problem is only to decide how much to trust each of them.[20]

Timing. Within each quarter t of the online sample, events unfold in the following order.

1. *Information arrives.* The indicators of quarter $t - 1$ are released, and the crisis indicator C_t is observed (Assumption A1).
2. *Labels are revealed.* The labels whose resolution date is t become known (Definition 2): if an onset occurs at t, the labels of the up to h preceding quarters are revealed at once as positive; otherwise, the label of quarter $t - h$ is revealed as negative.

[20] This is the framework of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006), in which no assumption is made on the process generating the forecasts or the labels.

3. *predictive models are updated.* Each predictive model is re-estimated on the quarters whose label is resolved, $\mathcal{S}_t = \{s \leq t : r_s \leq t\}$, and issues its forecast $p_{k,t}$.
4. *Weights are updated.* The forecaster updates its weights w_t using the predictive models' forecasts up to t and the effective labels revealed by t (Definition 3).
5. *The forecast is issued.* The forecaster announces $\hat{p}_t$.

Both layers are therefore subject to the same delayed feedback: the predictive models learn from resolved labels only, and the forecaster evaluates the predictive models on revealed labels only. The following definition makes precise which information the weights may use.

Definition 4 (Real-time aggregation rule)**.** An aggregation rule is a sequence of maps w_t from the information available at t to the simplex of $\mathbb{R}^K$. It is a real-time rule if w_t depends only on (i) the predictive models' forecasts $p_{k,1:t}$, (ii) the revealed labels $\{(s, \tilde{y}_s) : \tilde{r}_s \leq t\}$, and (iii) the crisis indicators $C_{1:t}$.

Lemma 2 (Real-time forecast)**.** *Under Assumption A1, if the predictive models' forecasts are $\mathcal{F}_t$-measurable and w_t is a real-time rule, then $\hat{p}_t$ is $\mathcal{F}_t$-measurable.*

In words, the aggregate forecast of quarter t can be computed by a policymaker at the end of quarter t, from the information then available: the procedure is free of look-ahead. The proof is in Appendix 9.3.2. A rule that updates the weights from revealed losses behaves very differently across the phases of the crisis cycle, because the labels revealed in each phase are of a different kind. Consider an episode with a pre-crisis window of h quarters, $[\tau - h, \tau - 1]$, followed by an onset at τ.

Calm periods. Only negative labels are revealed, each with a delay of between h and H quarters. Each revealed calm quarter penalizes the predictive models in proportion to the level of their forecast at that quarter, so that the weights drift slowly toward the predictive models that stay low in calm periods. In calm periods, the forecaster learns which predictive models raise false alarms, and only that.

Pre-crisis periods. This is the phase in which the forecast matters, and the one in which the forecaster learns least. The labels revealed during the build-up concern quarters at least h quarters earlier. Without the tolerance of Definition 3, these quarters would be revealed as negative, and a predictive model that started signalling early would be penalized for it precisely while the vulnerability builds up. The tolerance converts the labels of the $H - h = 6$ quarters before each pre-crisis window into positives, revealed at the onset, so that early signals are no longer penalized; but it does not create any positive information before the onset. During the pre-crisis

window, the weights are thus driven only by the losses of earlier quarters: they reflect which predictive models were reliable in the past, not which predictive model is currently right.

The onset. At the onset, up to H positive labels are revealed in a single quarter. The predictive models that were high during the build-up see their cumulative losses fall relative to the others, and the weights jump toward them. This jump may come too late to help detect the episode that has just started: it is useful only if the next crisis runs through the same channel.

Crises and their aftermath. During a crisis, the target is zero by convention (Definition 1), and the labels of crisis quarters are known immediately. The predictive models, however, often remain high: the indicators that signalled the build-up, such as credit or house prices, stay elevated for some time after the onset. Evaluating the predictive models on these quarters would penalize the predictive models that saw the crisis coming, and counting their forecasts as false alarms would penalize the forecaster for a crisis that is already public knowledge. Crisis quarters are therefore excluded from the cumulative losses, and the forecast is neutralized by a gate.

This description identifies the central difficulty. A rule driven by revealed losses learns in calm periods and at onsets, but not during the build-up, which is when the forecast is needed. The information that distinguishes the relevant predictive model during the build-up is not in the labels but in the predictive models' forecasts themselves, which shift as the vulnerability they monitor builds up. Section 5 uses this information to refine the standard online learning approach.

4 Predictive models

4.1 Channels

We build $K = 4$ predictive models, one per theoretical channel of vulnerability. Each predictive model uses a fixed, pre-specified set of indicators $\mathcal{V}_k$, together with their one-, two- and three-year growth rates, chosen from the literature rather than by statistical selection (Table 1). Pre-specification is what allows the aggregation to be interpreted as a diagnosis: the posterior over predictive models (Section 5.2.4) is a posterior over economic channels. The source of the data is described in Appendix section 9.1.2.

Table 1: Predictive models and channels of vulnerability

k	Channel	Indicators $\mathcal{V}_k$
1	Credit	[Basel gap, credit to households, credit to non-financial corporations, foreign-currency debt securities]
2	Real estate	[Real house prices, price-to-income, price-to-rent, rent price index]
3	External	[Current account, US exchange rate, exports, cross-border flows]
4	Global cycle	[Risk appetite, stock price index, NASDAQ , financial conditions index]

4.2 Estimation

Assumption A2 (Predictive model)**.** For each channel k, there exist $(\beta_{0,k}, \beta_k)$ such that, conditional on $z_{i,t-1,k} = (z_{i,t-1,v})_{v\in\mathcal{V}_k}$,

$$\Pr\big(y_{i,t} = 1 \mid z_{i,t-1,k}\big) = \Lambda\big(\beta_{0,k} + z'_{i,t-1,k}\beta_k\big), \tag{7}$$

with β_k sparse and possibly time-varying.

The forecast issued at t uses the predictors [21] dated $t-1$, the latest released at t, so that it relies only on information in $\mathcal{F}_{i,t}$ [22].

Assumption A2 is a working assumption, not a structural claim: each predictive model is a projection of the pre-crisis probability on its own channel, and the aggregation does not require any predictive model to be correctly specified. Sparsity reflects the fact that, within a channel, indicators are many and collinear while episodes are few. We estimate each predictive model by the elastic net (Zou and Hastie, 2005),

$$(\hat\beta_{0,k}, \hat\beta_k) = \arg\min_{\beta_0,\beta} \; -\frac{1}{n}\sum_{s\in\mathcal{S}}\Big[y_s(\beta_0 + z'_{s-1,k}\beta) - \log\big(1+e^{\beta_0+z'_{s-1,k}\beta}\big)\Big] + \lambda\Big[\alpha\|\beta\|_1 + \tfrac{1-\alpha}{2}\|\beta\|_2^2\Big], \tag{8}$$

The penalty in equation 8 combines the lasso ($\alpha = 1$) and ridge ($\alpha = 0$) penalties, and each term addresses a specific feature of our data. Within a channel, the predictors are few relative to the information in the sample, since each country contributes one or two episodes, and they are highly collinear by construction: each indicator enters in levels together with its one-, two- and three-year growth rates. The $\alpha\|\beta\|_1$ term sets most coefficients exactly to zero, which keeps each predictive model sparse and its coefficients readable. Used alone, however, the lasso handles collinearity poorly: within a group of highly correlated predictors, it tends to retain one of them arbitrarily and to discard the others (Zou and Hastie, 2005), and it is dominated by ridge regres-

[21] Variables havve been standardized (see Appendix section)

[22] In the vintage specification, which adds one quarter of lag, the predictors are dated $t-2$.

sion when predictors are strongly correlated (Tibshirani, 1996). In a real-time setting where each predictive model is re-estimated every quarter (with a delay), this would make the selected set flip between near-identical growth rates of the same indicator from one quarter to the next. The $\frac{1-\alpha}{2}\|\beta\|_2^2$ term corrects this through the grouping effect of Zou and Hastie (2005): the coefficients of highly correlated predictors are pulled toward each other, so that correlated indicators tend to enter or leave the model together. This stability is what allows us to interpret selection frequencies and sign stability across real-time estimates as evidence on the indicators themselves. The $\frac{1-\alpha}{2}\|\beta\|_2^2$ term also shrinks the coefficients, trading a small bias for a reduction in variance, which explains the good out-of-sample performance of the elastic net documented by Zou and Hastie (2005), Mol et al. (2009a), Mol et al. (2009b) and Destrero et al. (2009). We set $\alpha = 0.7$, following Addo et al. (2018): the penalty leans toward the lasso, to keep the predictive models sparse, while retaining enough of the ridge component to stabilize selection among collinear indicators.

The penalty level λ governs the number of selected indicators and is chosen by cross-validation. With one or two episodes per country, the cross-validated deviance curve is noisy, and its minimizer depends heavily on how observations are assigned to folds: two fold draws can select different values of λ, and therefore different indicators. We remove this source of instability by repeating five-fold cross-validation over $R = 50$ fold assignments, drawn with fixed seeds, averaging the R deviance curves, and taking the minimizer of the averaged curve. The selected λ, hence every estimate in the paper, is way less dependent on a particular fold draw.

Estimation proceeds in two phases.

Batch. On $\mathcal{S}_B = \{(i,s) : r_{i,s} \le T_B\}$, the set of panel observations whose label is known at T_B, we estimate one model per channel on the pooled panel. Pooling gives the initial model more than one episode to learn from. Note that $\mathcal{S}_B$ excludes the last h quarters before T_B unless their label was revealed earlier by an onset: a forecaster standing at T_B does not know them.

Real time. For each $t > T_B$ and country i, the predictive model is re-estimated on

$$\mathcal{S}_{i,t} = \{s \le t : r_{i,s} \le t\}, \tag{9}$$

that is, on country i alone and on resolved labels only, and the forecast is

$$p_{k,i,t} = \Lambda\big(\hat{\beta}_{0,k,i,t} + z'_{i,t-1,k}\hat{\beta}_{k,i,t}\big), \qquad (\hat{\beta}_{0,k,i,t}, \hat{\beta}_{k,i,t}) \text{ estimated on } \mathcal{S}_{i,t}. \tag{10}$$

An indicator with a missing value in the training window is dropped from that estimation and cannot be selected; observations are never dropped. By Lemma 1 and the construction of z, $p_{k,i,t}$ is $\mathcal{F}_{i,t}$-measurable.

Interpretation. We report the batch coefficients $\hat{\beta}_k$ together with the sign of each indicator's raw correlation with the target. For the real-time estimates, we report the frequency with which each indicator is selected and the stability of its sign over time. An indicator is a credible predictor when it is selected often, with a stable sign consistent with theory.

5 Online aggregation by multi-sensor change-point detection

This section describes how the four predictive models forecasts are combined into a single pre-crisis probability. We both consider the standard approach of the online learning literature (Cesa-Bianchi and Lugosi, 2006) where each predictive model is aggregated according to its past performance. With a delayed feedback, this approach may learn too late: the losses that would identify the relevant predictive model for a given crisis are revealed only at its onset (Appendix **??**). This motivates a strong refinement of the standard online learning approach: weighting predictive models according to the evidence they provide, using the framework of multi-sensor change-point detection.

Our procedure sees each predictive model has a sensor : its forecasts fluctuate around a low level in normal times and shift upward when the vulnerability it monitors builds up. The four predicive form a network of four sensors, each watching one channel of vulnerability. The entry into a pre-crisis period is a change in the distribution of the sensors' observations. The problem is not to predict a single quarter, but to detect, as early as possible and without too many false alarms, that a change has occurred. Two features make the problem hard: the date of the change is unknown, and so is the channel through which it occurs, since crises differ in their origin. Each forecast carries some evidence for or against a change, measured by its log-likelihood ratio. A single high forecast is rarely conclusive. We look instead for a sequence of high forecasts . For each sensor, a CUSUM statistic accumulates this evidence over time and discards it when it turns against a change, so that the statistic stays near zero in normal times and rises once a change has occurred. Comparing the statistics of the four sensors tells us which channel is moving. We turn them into a posterior probability over sensors: the sensor whose statistic is highest receives

most of the weight. This posterior is the main output of the method. It is computed from current forecasts only and requires no label, so that it reacts before the onset; in economic terms, it is the real-time probability that each channel is the source of the vulnerability. Three slower components complement the posterior. A learning layer down-weights predictive models that made large errors on past quarters whose label has been revealed; a reliability factor penalizes predictive models that are chronically high in calm periods; and a prior built on the batch period encodes which channels mattered historically. These components use only resolved labels, so they adjust weights over years rather than quarters. In that sense, we complement the standard approach of the online learning litterature by tree other componenents. This represent our main methodological contribution. The rest of this section formalizes these notions in turn: the sensor model (Section **??**), the per-sensor CUSUM and its link with the predictive models' forecasts (Section 5.2.3), the posterior over sensors (Section 5.2.4), the learning, reliability and prior components, the gate, and the aggregate forecast.

Our contribution can be summarized by the following predicted probability $\hat{p}_t$ defined as follows :

Definition 5 (Detection-based aggregation [23])**.**

$$\hat{p}_t = g_t \sum_{k=1}^{K} w_{k,t}\, p_{k,t}, \qquad w_{k,t} \;\propto\; w_k^0 \;\cdot\; e^{-\eta_t (L_{k,t} - \min_j L_{j,t})} \;\cdot\; \pi_{k,t} \;\cdot\; (1 - F_{k,t})^{\kappa}, \tag{11}$$

normalized over the predictive models available at t.

The terms g_t, $\pi_{k,t}$ and $(1 - F_{k,t})^{\kappa}$ represents the extensions we introduce to the standard online learning approach[24].The following sections describe how each of the terms of Definition 5 is computed.

5.1 The learning layer

Following the online learning literature (Cesa-Bianchi and Lugosi, 2006), the first component of the weights depends on the past performance of the predictive model. Each predictive model k starts from an initial weight w_k^0 and, as the labels of past quarters are revealed, loses weight

[23]Every factor in equation 11 is $\mathcal{F}_t$-measurable by Lemmas 1 and 4 and Definition 3, so $\hat{p}_t$ is a real-time forecast.

[24]Combining a learning layer based on predictive models' past performance and a detection layer which measures the confidence the forecaster can have for a predictive model is the structure of aggregation with expert confidences built by (Gaillard et al., 2014). A predictive model reports a low confidence in a given round receives little weight in that round, whatever its past performance. In our setting, the confidence of a predictive model is not reported by the predictive model itself but inferred from the regime change in its forecasts.

exponentially in its cumulative loss:

$$v_{k,t} = \frac{w_k^0 \, e^{-\eta_t (L_{k,t} - \min_j L_{j,t})}}{\sum_{i=1}^{K} w_i^0 \, e^{-\eta_t (L_{i,t} - \min_j L_{j,t})}}, \qquad L_{k,t} = \sum_{s \in \mathcal{R}_t} \omega_s \left(p_{k,s} - \tilde{y}_s \right)^2, \tag{12}$$

where $\mathcal{R}_t$ is the set of quarters whose effective label has been revealed by t, ω_s the class and cost weights of Section 9.4 in Appendix, and η_t the learning rate. A predictive model that made large errors in the past receives a small weight; a predictive model that made small errors keeps its initial weight. Subtracting $\min_j L_{j,t}$ does not change the weights, since the common factor $e^{\eta_t \min_j L_{j,t}}$ cancels in the ratio; it only avoids numerical underflow.

This is the exponentially weighted average forecaster, adapted to delayed feedback in the simplest way: the cumulative loss is computed on revealed quarters only, so that the weights are updated whenever labels arrive, one at a time in calm periods and in a block at each onset. The learning rate

$$\eta_t = \frac{1}{B} \sqrt{\frac{8 \log K}{t}}, \qquad B = \frac{\theta}{0.05}, \tag{13}$$

governs the speed at which the weights move away from their initial value. It is the anytime rate of Cesa-Bianchi and Lugosi (2006, Section 2.3), which does not require the length of the sample to be known in advance, rescaled by the bound B of the weighted loss of a quarter.

The exponentially weighted average (EWA) rule, introduced by Vovk (1990) and Littlestone and Warmuth (1994), has several advantages for our purpose.[25] First, it is computable in a simple incremental way: by equation 12, the weights depend on the past only through the predictive models' cumulative losses, so that updating them when new labels are revealed requires only adding the new losses. Second, the weights depend only on the past performance of the predictive models, not on the past forecasts of the forecaster itself, so that a poor forecast in one quarter does not propagate to the following ones. Third, the weights are easy to interpret: the log-ratio of the weights of two predictive models is the log-ratio of their initial weights minus η_t times the difference between their cumulative losses, so that each weight can be traced back to the errors of the corresponding predictive model.Beyond these practical advantages (see Section in the Appendix).

Initial weights. Without information on the predictive models, the natural starting point is the uniform vector $w_k^0 = 1/K$. Since the online period offers few labels, however, the weights remain close to their starting point for a long time, and the batch period contains information on

[25] The EWA rule is also used in game theory as a variant of smooth fictitious play (Fudenberg and Levine, 1998).

the predictive models' performance. We therefore build the initial weights from it. Let $U_{k,j}$ be the maximal utility (Definition 10) of predictive model k on the batch quarters of country j whose labels are known at T_B, and $\bar{U}_k = N^{-1} \sum_j U_{k,j}$. The initial weights are

$$w_k^0 = (1 - \varphi) \frac{e^{\gamma \bar{U}_k}}{\sum_j e^{\gamma \bar{U}_j}} + \frac{\varphi}{K}, \qquad \gamma = 6, \quad \varphi = 0.10. \tag{14}$$

predictive models that performed better in the batch period start with a higher weight. Averaging across countries uses the heterogeneity of the early-1990s episodes without privileging any country. The floor φ guarantees $w_k^0 \geq \varphi/K = 0.025$ for every predictive model: a channel that was irrelevant in the 1990s but drives the next crisis is not excluded from the start. Section 7 compares these initial weights with the uniform vector.

With $K = 4$ and $B = 16$, the learning rate is small, about 0.027 at the start of the online period. In calm periods, the weights drift slowly toward the predictive models that raise fewer false alarms. At an onset, the labels of the whole pre-crisis window are revealed at once, and the weights jump toward the predictive models that signalled it. The learning layer thus records which channels mattered in past episodes, but it cannot identify the relevant channel while a build-up unfolds, since the corresponding losses are revealed only at the onset. This is the role of the detection layer of the next section.

5.2 The detection layer

5.2.1 The sensor model

In the sequential detection literature, a sensor is a sequence of observations $(X_t)_{t \geq 1}$ whose distribution may change at an unknown date ν: before ν, the observations have density f_0; from ν onward, they have density f_1. The problem is to detect the change as soon as possible after it occurs, while raising few false alarms. In our setting, each predictive model is a sensor, whose observations are its forecasts, $X_t = p_{k,t}$: in normal times, the forecasts of model k fluctuate around a low level, and when the vulnerability it monitors builds up, their distribution shifts upward.

Definition 6 (Sensor)**.** A sensor is a sequence of observations whose distribution may change at an unknown date. Predictive model k is a sensor with observations $(p_{k,i,t})_t$.

Assumption A3 (Regime change in one sensor)**.** For a given country i , there exist densities $f_{0,k}$ and $f_{1,k}$ such that, in normal times, $p_{k,t} \sim f_{0,k}$, and from the first pre-crisis quarter ν onward, $p_{k^*,t} \sim f_{1,k^*}$ for one sensor k^*, unknown in advance, with f_{1,k^*} stochastically larger than f_{0,k^*}.

Assumption A3 states that there exist densities $f_{0,k}$ and $f_{1,k}$ such that, in normal times, the forecasts of every model are drawn from $f_{0,k}$, $p_{k,t} \sim f_{0,k}$; from the first pre-crisis quarter ν onward, a single sensor k^* changes distribution, $p_{k^*,t} \sim f_{1,k^*}$, while the others keep their normal-time density. The post-change density f_{1,k^*} is stochastically larger than f_{0,k^*}: the affected model issues higher probabilities. Neither the affected sensor nor the date of the change is known in advance. Therefore this assumption formalizes the heterogeneity of crises: a credit boom moves the credit sensor, a housing boom moves the real-estate sensor, and the next crisis may run through either.[26]

This is the "one affected sensor, unknown" problem of Tartakovsky and Veeravalli (2008) and Mei (2011). To detect a change in distribution, the reference statistic is the log-likelihood ratio of each observation, $\log f_1(X_t)/f_0(X_t)$, which answers the question: is this observation more likely under the pre-crisis distribution or under the normal one? Our predictive models, however, do not produce likelihood ratios; they produce probabilities. The following lemma shows that the two are closely related: the log-odds of a calibrated forecast, that is, of a forecast whose probabilities mean what they announce, is a log-likelihood ratio shifted by a constant.

[27]

Lemma 3 (Calibrated forecasts are log-likelihood ratios)**.** *Suppose predictive model k is calibrated, in the sense that $\Pr(y_t = 1 \mid p_{k,t}) = p_{k,t}$: among the quarters in which the predictive models' forecasts p, a fraction p are pre-crisis quarters. Let $\pi = \Pr(y_t = 1)$ be the pre-crisis frequency on its estimation sample, and $f_{j,k}$ the density of $p_{k,t}$ conditional on $y_t = j$. Then*

$$\operatorname{logit}(p_{k,t}) = \log \frac{f_{1,k}(p_{k,t})}{f_{0,k}(p_{k,t})} + \log \frac{\pi}{1-\pi}. \tag{15}$$

[26] Assumption A3 rules out simultaneous changes in several sensors, which is restrictive: a housing boom, for instance, is often financed by credit or by capital inflows. Detection with several affected sensors has been studied by Mei (2010), Xie and Siegmund (2013) and Fellouris and Sokolov (2016). When several sensors change, the assumption is violated but detection is unaffected; what changes is the interpretation of the posterior over sensors (Appendix 9.10).

[27] A predictive model is calibrated in the weak sense of Dawid (1982) if $\Pr(y_t = 1 \mid p_{k,t}) = p_{k,t}$: among the quarters in which it forecasts p, a fraction p are pre-crisis quarters. This does not require the predictive model to be correctly specified, only its forecasts to mean what they say. In that sense, our predictive models are not fully calibrate. The elastic-net penalty pulls forecasts toward the base rate of the estimation sample, which makes them underconfident; estimation on few episodes can instead produce saturated probabilities; the base rate differs between the estimation sample and the forecast period; and the models are re-estimated every quarter, so that successive forecasts do not come from the same model. With four episodes, calibration cannot be tested reliably, since most bins of a reliability diagram would be empty. Lemma 3 therefore motivates the form of the detection statistic rather than establishing its exact optimality. This approximation matters little for three reasons. First, the procedure requires only that the increment drift downward in normal times and upward in pre-crisis times. Shrinkage toward the base rate rescales the log-likelihood ratio without reversing its sign, so it slows detection without creating false alarms. Second, the clip c bounds the contribution of a saturated forecast to two units per quarter. This is why this parameter is important. Third, no tuning parameter relies on calibration: the alarm threshold is calibrated on the training sample, c and $W_{\max}$ are varied over a grid in Section 7, and the regret guarantee of Proposition 1 holds for arbitrary forecast sequences.

Therefore Lemma 3 shows that we do not need to estimate densities of each sensor in each regime to analyze evidence that the current observation brings in favour of the pre-crisis regime against the normal regime. A calibrated preditvtive model computes the log-likelihood ratio for us [28], up to a known constant,

$$\log \frac{f_{1,k}(p_{k,t})}{f_{0,k}(p_{k,t})} = \operatorname{logit}(p_{k,t}) - \log \frac{\pi}{1-\pi}. \tag{16}$$

5.2.2 The increment

A single high quarter proves nothing: it may be noise. What is informative is the accumulation of evidence over several quarters. The CUSUM statistic adds up, quarter after quarter, increments that measure the evidence in favour of a pre-crisis regime, and returns to zero whenever the sum turns negative. It thus forgets calm periods and keeps only the evidence accumulated since the beginning of the most suspicious sequence of quarters. The whole procedure therefore rests on one choice: what the statistic adds up each quarter, that is, how the forecast of a predictive model is turned into an increment of evidence.

Following Page (1954), we consider the increment $\ell_{k,t} = \operatorname{clip}\big(\operatorname{logit}(p_{k,t}) - \operatorname{logit}(\pi_0), -c, c\big)$.[29] An intuitive interpretation of this increment is as a voting game. A predictive model votes in favor of a pre-crisis if it considers such an event more likely than not (i.e., $\pi_0 > 1/2$), and against it otherwise. The weight assigned to each vote reflects the model's confidence, measured in log-odds.

Concerning the reference rate, the ideal increment of Page (1954) would use the base rate [30], $\pi_0 = \pi$. We use a fixed reference for three reasons. First, the base rate differs across the samples on which the predictive models are estimated, the pooled panel in the batch period and each country's resolved history afterwards, and it changes every quarter as the training window

[28] For instance, with $\pi = 0.15$, a forecast equal to the base rate, $p_{k,t} = 0.15$, carries no evidence, since its log-likelihood ratio is zero; a forecast $p_{k,t} = 0.5$ has log-likelihood ratio $\log(0.85/0.15) \approx 1.73$, meaning that such a forecast is $e^{1.73} \approx 5.7$ times more frequent in pre-crisis than in normal times. The detection layer can therefore be built on the predictive models' forecasts alone.

[29] This increment is formulated on a logg odds scale consistenly of the literature (Page, 1954). It is also justiied by the fact that a CUSUM adds up evidence, so its increment must be measured on a scale on which evidence is additive. By Lemma 3, the log-odds of a calibrated forecast is, up to a constant, the log-likelihood ratio, which is that scale: two quarters with a log-likelihood ratio of one bring the same evidence as one quarter with a log-likelihood ratio of two. The probability scale does not have this property. An increment such as $p_{k,t} - \pi_0$ is bounded by one, so a forecast of 0.999 counts barely more than a forecast of 0.9, although the first is ten times stronger evidence in odds terms; and it has no interpretation as accumulated evidence.

[30] The reference π_0 should not be confused with the base rate π. The base rate is a feature of the data: about 15% of quarters are pre-crisis quarters. The reference is a choice of the method: it is the level above which a forecast counts as evidence for a pre-crisis. Page's statistic would set the reference at the base rate; we set it at one half, so that a forecast adds evidence only when the model considers a pre-crisis more likely than not.

grows. With $\pi_0 = \pi$, the reference would move over time and differ across sensors, and the statistics of the four sensors would no longer be on a common scale, which the comparison in the posterior equation 19 requires. A fixed reference keeps them comparable. Second, $\pi_0 = 1/2$ gives the increment a direct reading: a predictive model adds evidence exactly when it considers a pre-crisis more likely than not, and removes evidence otherwise. Third, it is conservative. Combining the increment with Lemma 3 gives its meaning in terms of evidence: $\ell_{k,t} > 0$ if and only if $f_{1,k}(p_{k,t})/f_{0,k}(p_{k,t}) > (1-\pi)/\pi \approx 5.7$, that is, if what the predictive model observes is at least 5.7 times more frequent in pre-crisis than in normal times. A quarter counts as evidence for a pre-crisis only if this evidence is strong enough to overturn the rarity of crises. Equivalently, the reference reads the predictive models' forecasts as if both regimes were equally likely a priori, so that the statistic accumulates the evidence brought by the data rather than the prior [31]

5.2.3 Per-sensor CUSUM

Following Page (1954), we compute for each predictive model a CUSUM which accumulates over time the evidence that the model has entered a pre-crisis regime. Comparing the statistics of the four models tells us which channel is moving, and therefore how much weight each model should receive.

Definition 7 (Bounded CUSUM)**.** With increments

$$\ell_{k,t} = \mathrm{clip}\big(\,\mathrm{logit}(p_{k,t}) - \mathrm{logit}(\pi_0),\ -c,\ c\big), \tag{17}$$

define recursively $W_{k,0} = 0$ and

$$W_{k,t} = \min\Big\{W_{\max},\ \max\big\{0,\ W_{k,t-1} + \ell_{k,t}\big\}\Big\}, \qquad c = 2, \quad W_{\max} = 6. \tag{18}$$

At each onset, $W_{k,t}$ is reset to zero for all k.

The CUSUM recursion of Page (1954) is nothing but a policymaker who, every quarter, asks a single question: has the economy entered a pre-crisis regime, and if so, since when? A natural way to answer it is to treat every past quarter ν as a candidate date for the change, to weigh the evidence accumulated since ν in favour of the change, $S_t(\nu) = \sum_{s=\nu}^{t} \ell_s$, and to retain the

[31] This reference is more demanding than the policymaker's costs would suggest, since with $\theta = 0.8$ an alarm is worthwhile as soon as $p > 1 - \theta = 0.2$. We argue that the two steps play distinct roles: the reference governs when a sensor starts accumulating evidence, and is deliberately cautious; the alarm threshold τ^* governs the decision and incorporates the asymmetry of costs, since it maximizes the policymaker's utility.

most favourable candidate. Each possible date ν is a candidate for the beginning of the pre-crisis regime, and the evidence accumulated since ν is the sum of the increments $S_t(\nu) = \sum_{s=\nu}^{t} \ell_s$. The CUSUM statistic is the largest of these sums over all candidate dates, $W_t = \max_\nu S_t(\nu)$ (Lemma 7 in Appendix 9.7). When the increments are log-likelihood ratios, e^{W_t} is therefore the likelihood ratio of the most likely change date: the statistic tests every possible date for the beginning of the pre-crisis regime and keeps the most favourable one, which is why it does not need to know when the change occurred. The reset to zero is what makes this possible: when the accumulated evidence turns negative, the candidate date moves forward, and calm periods, which argue against a change, are forgotten rather than allowed to mask a change that starts later [32].

We add two extensions to the the original CUSUM statistic of Page (1954). The first one directly answers the fact that penalized logits estimated on few episodes occasionally produce saturated forecasts: $p_{k,t} = 0.9999$ has log-odds 9.2. Without the clip, a single such quarter would move the statistic by more than its cap. With $c = 2$, a quarter contributes at most two units [33]. The second refinement concerns the fact that the CUSUM statistic is also bounded by a cap $W_{\max} = 6$[34]. Without a cap [35], a sensor that has signaled for a long time would build an arbitrarily large lead, and a channel that becomes active later could never take over the posterior, however strong its signal. With the cap, the lead of any sensor is at most $W_{\max}$, and a sensor starting from zero reaches the cap in $\lceil W_{\max}/c \rceil = 3$ quarters of maximal evidence. For the same reason, when several channels become active, the cap ensures that they end up sharing the weight equally rather than in proportion to the date at which each started rising (Appendix 9.10). The cap can be interpreted as a way of bouding the memory of the statistic : evidence accumulated during an episode that does not end in a crisis must be unwound before the statistic returns to zero. Without a cap, a long period of elevated forecasts would keep the statistic high for as long as it took to build up; with the cap, a sensor at $W_{\max}$ returns to zero in at least $W_{\max}/c = 3$ quarters of strong evidence against a change, and in a few years at most under the average calm-period drift. [36].

[32] Two features of this procedure follow from the question itself rather than from any modelling choice. First, the evidence for a change starting at ν is a sum, because independent pieces of evidence combine additively on the log-likelihood scale: a sustained build-up, quarter after quarter, weighs more than a single alarming observation. Second, the most favourable candidate is never earlier than the last quarter in which the accumulated evidence turned negative, since including quarters that argue against a change can only weaken the case for it. The CUSUM is a sequence of sequential probability ratio tests each started afresh as soon as the previous one has concluded that no change has occurred.

[33] c is varied over a grid in Section

[34] $W_{\max}$ is varied over a grid in Section 7

[35] The price of the cap is the loss of the exact optimality of Theorem 4, which holds for the uncapped statistic. It does not prevent detection: the alarm is based on the aggregate forecast $\hat{p}_t$ and its threshold τ^*, not on the statistic crossing a level, so that a sensor at the cap simply carries the maximal weight. The cap is varied over $W_{\max} \in \{4, 6, 8\}$ in Section 7, with little effect on the results.

[36] The cap thus plays the role of a finite memory, in the spirit of the window-limited rules of Lai (1998): in a non-stationary environment, old evidence becomes stale, and a false signal should not dominate the statistic long after it has

5.2.4 Posterior over sensors

We now have four statistics, $W_{1,t}, \ldots, W_{4,t}$, one per channel. We want to turn them into the probability that each channel is the one that is moving. The exact Bayesian answer involves, for each channel, a sum over all possible change dates; the posterior we use replaces each sum by its largest term, which is precisely what the CUSUM statistic retains.

Following Tartakovsky and Veeravalli (2008) and Mei (2011), we turn the four statistics into a posterior probability over channels:

$$\pi_{k,t} = \frac{\exp(\zeta\, W_{k,t})}{\sum_{j=1}^{K} \exp(\zeta\, W_{j,t})}, \qquad \zeta = 1. \tag{19}$$

The form of this transformation follows from Bayes' rule. If one channel has changed regime but we do not know which, the probability that it is channel k is proportional to how much more likely its recent forecasts are under the pre-crisis regime than under the normal one. By Lemma 7, $e^{W_{k,t}}$ measures exactly this likelihood ratio, evaluated at the most likely date of the change. With $\zeta = 1$, equation 19 is therefore an approximate Bayesian posterior: each channel is weighted by the evidence accumulated in its favour.

Three properties of equation 19 matter for our exercice. First, it depends only on differences between statistics: $\pi_{k,t}/\pi_{j,t} = e^{\beta(W_{k,t}-W_{j,t})}$, so that a lead of two units, one quarter of maximal evidence, multiplies the relative weight of a channel by $e^2 \approx 7.4$. When no sensor carries evidence, all statistics are at zero and the posterior is uniform: it does not distort the weights in calm periods. Second, the parameter ζ acts as a temperature. As $\zeta \to \infty$, the posterior puts all the weight on the sensor with the largest statistic, which is the max-rule of Tartakovsky and Veeravalli (2008); as $\zeta \to 0$, it becomes uniform and ignores the evidence. The value $\zeta = 1$ is not a tuning choice but the one implied by Bayes' rule. Third, the posterior is soft rather than hard. The max-rule is optimal for deciding whethera change has occurred, but a poor rule for weighting forecasts: when two channels carry similar evidence, it switches all the weight from one to the other at each small fluctuation, and it cannot share the weight when several channels are active at once (Appendix 9.10). The softmax varies continuously with the evidence and hedges between channels whose statistics are close.

The posterior has the same exponential form as the weights of the learning layer of the Exponential Weighted Average, $e^{-\eta L_{k,t}}$, with accumulated evidence in place of accumulated losses: the learning layer rewards predictive models that were right in the past, the posterior rewards

faded.

predictive models that currently carry evidence. Unlike losses, evidence requires no label: $\pi_{k,t}$ is computed from the forecasts $p_{k,1:t}$ only, so that it reacts before the onset, when no label of the pre-crisis period has yet been revealed. In economic terms, when a single channel is active, $\pi_{k,t}$ is the real-time probability that channel k is the source of the vulnerability; when several are active, it is the share of the evidence attributable to channel k.

5.3 The reliability layer

Reliability. The third factor of the weights penalizes predictive models that are chronically high in calm periods. Let $\mathcal{C}_t = \{s \in \mathcal{R}_t : \tilde{y}_s = 0,\ C_s = 0\}$ be the set of revealed calm quarters at date t, and

$$F_{k,t} = \frac{1}{|\mathcal{C}_t|} \sum_{s \in \mathcal{C}_t} p_{k,s}, \tag{20}$$

the average forecast of predictive model k on these quarters. If the forecasts of predictive model k were used as alarms, $F_{k,t}$ would be the expected share of calm quarters in which it raises an alarm: it is a continuous analogue of the false-positive rate, which does not require choosing a threshold and remains stable when few calm quarters have been revealed. The weight of a predictive model k is multiplied by $(1 - F_{k,t})^{\kappa}$, with $\kappa = 1/2$ and $F_{k,t}$ capped at 0.95.

The factor addresses a bias of the detection layer. By Lemma 3 and **??**, the drift of a sensor's statistic in calm periods is negative, but less so for a predictive model whose forecasts are high in calm periods: its statistic spends more time away from zero, and it receives a higher posterior weight $\pi_{k,t}$ without carrying more evidence. A chronically alarmist predictive model is thus systematically favoured by the posterior, and the reliability factor corrects this bias by discounting it in proportion to its observed false-alarm propensity. The learning layer also penalizes such a predictive model, since high forecasts on calm quarters raise its cumulative loss, but it does so slowly: a calm quarter carries a weight of $(1-\theta)/(1-\hat{\pi}_t) \approx 0.24$ in the loss and the learning rate is small, so that a decade of false alarms moves the weights by only a few percent. The reliability factor isolates the dimension of performance that matters most for macroprudential policy in calm periods, the propensity to raise false alarms, which is costly for a policymaker since each false alarm can lead to tightening capital requirements without cause.

The power κ acts as a temperature, playing for the false-alarm propensity the role that ζ plays for the evidence and η_t for the cumulative loss: in log-weights, the factor contributes $\kappa \log(1 - F_{k,t})$. With $\kappa = 0$, the factor disappears; with $\kappa = 1$, the weight is proportional to $1 - F_{k,t}$, a continuous analogue of the predictive model's specificity. The intermediate value $\kappa = 1/2$ makes

the correction moderate: a a predictive model whose average calm forecast is 0.05 keeps 97% of its weight, a predictive model at 0.4 keeps 77%, and the strongest possible discount, at the cap $F_{k,t} = 0.95$, is a factor of 0.22, or 1.5 units in log-weights. This is less than a single quarter of maximal evidence in the detection layer: the reliability factor arbitrates between predictive models that carry similar evidence but never overrides a sensor that carries strong evidence. Like the learning layer, it uses only revealed labels, and κ is varied over $\{0, 1/2, 1\}$ in Section 7.

5.4 The gate

Once a crisis has started, the forecasting problem changes nature. The target is zero by convention during crises (Definition 1), since the object of interest is the entry into a crisis, not its continuation. The predictive models, however, often remain high: the indicators that signalled the build-up, such as credit or house prices, stay elevated for some time after the onset, and a model trained to recognize these levels keeps forecasting a pre-crisis. Left as such, every crisis quarter in which a predictive model stays high would count as a false alarm. This would penalize the forecaster for an event that is already public knowledge, distort the calibration of the alarm threshold, and make little economic sense: once a crisis has started, a policymaker no longer needs an early warning, and would release the countercyclical buffer rather than raise it. We therefore neutralize the forecast during crises with a gate, which sets it to zero while a crisis is ongoing and brings it back gradually once the crisis has ended. The gradual return prevents the forecast from jumping back to the predictive models' level, which may still be elevated, on the first quarter after the crisis. The gate uses only the crisis indicator, which is observed in real time, and acts on the forecast only, not on the weights of the aggregation.

Definition 8 (Gate)**.** Let $e_t = \max\{s < t : C_s = 1\}$ be the last observed crisis quarter, if any. Then

$$g_t = \begin{cases} 0 & \text{if } C_t = 1, \\ \min\{1,\ (t - e_t)/\phi\} & \text{if } C_t = 0 \text{ and } e_t \text{ exists}, \\ 1 & \text{otherwise}, \end{cases} \qquad \phi = 4. \tag{21}$$

The gate sets the forecast to zero during a crisis and brings it back linearly over the $\phi = 4$ quarters that follow its end: one quarter after the end of a crisis, the forecast is multiplied by $1/4$, two quarters after by $1/2$, and from the fourth quarter onward it is no longer affected.

The value $\phi = 4$ balances two risks. A faster return would bring the forecast back to the predictive models' level, which often remains elevated, on the quarter that follows the crisis, and

would generate false alarms in the immediate aftermath of every crisis. A slower return would mask a new build-up for years, against the purpose of an early-warning system. One year is also the reference horizon of the paper: it is the time a countercyclical capital buffer takes to become binding under Basel III, and the length of the smoothing window of the signal. The value is set ex ante, and the length of the smoothing window of the signal. The value is set ex ante, and Section 7 varies it over $\phi \in \{2, 4, 8\}$.

Lemma 4. *Under Assumption A1, g_t is $\mathcal{F}_t$-measurable, and multiplying any forecast by g_t incurs zero loss on crisis quarters.*

Proof. By 21, g_t depends only on $C_{1:t}$, which is $\mathcal{F}_t$-measurable by Assumption A1. On a crisis quarter, $y_t = 0$ by Definition 1 and $g_t = 0$, so that the gated forecast $g_t\hat{p}_t = 0$ equals the target and the loss is zero. □

The gate formalizes a simple fact: once a crisis has started, a policymaker no longer needs an early warning. It uses only the crisis indicator, which is public, and does not help in pre-crisis quarters: it acts only on quarters in which the crisis has begun or has just ended. The linear fade after the end of a crisis prevents the forecast from jumping back to the predictive models' level, which may still be elevated, on the first quarter after the crisis. The gate acts on the forecast only, not on the weights: the weights keep evolving during the crisis, and the forecast resumes from them once the gate reopens.

5.5 Summary of the procedure

The following table summarizes the different layers of the procedure with its associated parameters :

5.6 A regret guarantee

Our procedure adds several layers to the standard exponentially weighted average in order to overcome the delayed feedback. One important drawback is that it is no longer optimal in the sense of the online learning literature [37], since the posterior $\pi_{k,t}$ depends on current signals rather than on past losses.

We recover a guarantee by nesting the detection rule in a second, simpler aggregation, which we call the meta-aggregator. The meta-aggregator treats the detection rule as one more candidate

[37] In the standard online literature, a strategy is optimal in the sense that it competes with the best predictive model (or the best combination of predictive model) known ex-post, see Appendix 9.5

Table 2: Parameters of the aggregation and how they are set

Layer	Parameter	Role	How it is set
Learning	$\theta = 0.8$	Relative cost of a missed crisis in the loss	Policymaker's preferences, as in the evaluation criterion (Alessi and Detken, 2011)
	$\eta_t = B^{-1}\sqrt{8 \log K / t}$	Learning rate	Anytime rate of exponential weights (Cesa-Bianchi and Lugosi, 2006); no tuning
	$B = \theta / 0.05$	Bound on the loss of a quarter	Implied by θ and the lower bound on $\hat{\pi}_t$
	$\gamma = 6$	Temperature of the prior	Set ex ante; alternative priors compared.
	$\varphi = 0.10$	Floor of the prior	Set ex ante; bounds the cost $\log(1/w_k^0)$ of tracking any channel
	$\pi_0 = 1/2$	Reference of the increment	A forecast adds evidence when a pre-crisis is more likely than not
Detection	$c = 2$	Clip of the increment	Set ex ante; grid $\{1, 2, 3\}$
	$W_{\max} = 6$	Cap of the statistic	Set ex ante; grid $\{4, 6, 8\}$
	$\zeta = 1$	Temperature of the posterior	Implied by Bayes' rule; no tuning
Reliability	$\kappa = 1/2$	Weight of the false-alarm penalty	Set ex ante; grid $\{0, 1/2, 1\}$
Gate	$\phi = 4$	Quarters of gradual return after a crisis	One year, the activation horizon of the buffer; grid $\{2, 4, 8\}$

Notes: All parameters are fixed before the online period and none is optimized on it. The grid of Section 7 varies c, $W_{\max}$, κ and ϕ jointly, over $3^4 = 81$ combinations.

alongside the K predictive models, and combines these $K+1$ forecasts with exponential weights driven by their revealed losses only. Since the detection rule is the forecast we want to use, it receives a prior weight of 0.9, the remaining 0.1 being shared among the predictive models: as long as the detection rule performs well, the meta-aggregator essentially reproduces it. If the rule were to fail in the long run, however, the losses would shift the weight toward the predictive models that perform better. The meta-aggregator thus provides insurance rather than performance: by construction, it can never do much worse than the best of its candidates, including the detection rule itself.

Two features make this guarantee useful on our sample. First, the loss is a weighted squared loss, which is exp-concave: this yields a regret bound that does not grow with the length of the sample, unlike the $\sqrt{T}$ bound of standard exponential weights. Second, the meta-aggregator is also contrainted by the delayed feedback. A standard exponentially weighted forecaster needs the loss of quarter t before forecasting quarter $t+1$, whereas here a label is revealed up to $H = h+6$ quarters later; H is not a tuning parameter but the maximal delay implied by the definition of the effective label (Definition 3). Following Joulani et al. (2013), we run several copies of the forecaster, which we call mixtures: each has its own weights over the $K+1$ candidates, and

its forecast is the corresponding weighted mixture of the candidates. Each quarter, the forecast is issued by a mixture whose previous quarters have all had their labels revealed, and a new mixture is created only if none is available. Each mixture therefore never faces a delay, and the standard bound applies to it. A mixture is busy at date t only if it awaits the label of one of the last $H-1$ quarters, so that at most $H-1$ mixtures are busy at any date and at most H are ever needed. The regret bound is the bound of a single mixture multiplied by the number of mixtures. Against the detection rule, whose prior weight is 0.9, this gives about $H\log(1/0.9)/\eta \approx 3$ in cumulative loss for $h = 12$, small relative to the total loss at stake; against each predictive model, whose prior weight is 0.025, it gives about $H\log 40/\eta \approx 106$.

Definition 9 clarifies the behavior of the meta-aggregator :

Definition 9 (Meta-aggregator)**.** Let the $K+1$ candidates be $x_{0,t} = \sum_k w_{k,t}p_{k,t}$ (the rule before the gate) and $x_{k,t} = p_{k,t}$ for $k = 1,\dots,K$, with prior $u^0 = (u_0^0, u_1^0, \dots, u_K^0)$, $u_0^0 = 0.9$, $u_k^0 = 0.1/K$. Define the loss

$$\ell_t(x) = \omega_t\,(x-\tilde{y}_t)^2, \qquad \omega_t = \theta\,\mathbf{1}\{\tilde{y}_t = 1\} + (1-\theta)\,\mathbf{1}\{\tilde{y}_t = 0\}, \tag{22}$$

and the theoretical learning rate $\xi = 1/(2\max\{\theta, 1-\theta\})$. The meta-aggregator maintains a pool of mixtures. A mixture is an exponentially weighted forecaster over the $K+1$ candidates with a theoretical learning rate ξ: it holds weights $u = (u_0,\dots,u_K)$ and forecasts the corresponding weighted average of the candidates' forecasts. The pool evolves as follows.

1. *Choosing a mixture.* At each non-crisis quarter t, the forecast is issued by a mixture that is free, that is, whose previously assigned quarters have all had their labels revealed. If no mixture is free, a new one is created, with the initial weights u^0.
2. *Forecasting.* The chosen mixture forecasts $q_t = \sum_j u_{j,t}\,x_{j,t}$, and quarter t is assigned to it.
3. *Updating.* When the label $\tilde{y}_s$ of a quarter s is revealed, the mixture to which s was assigned updates its weights, $u_j \leftarrow u_j\,e^{-\eta\,\ell_s(x_{j,s})}$, and renormalizes them. The other mixtures are unchanged.
4. *Unavailable candidates.* A candidate with no forecast at t is assigned the forecast q_t of the mixture. It then incurs the same loss as the mixture, so that its relative weight is unaffected.
5. *Final forecast.* The forecast is $g_t\,q_t$, where g_t is the gate of Definition 8; crisis quarters are assigned to no mixture.

Each mixture forecasts a new quarter only once it has learned from all its previous ones: along

the quarters assigned to it, it never faces a delay. With a delay of three quarters, for instance, three mixtures suffice, each in charge of one quarter out of three; in general, at most H mixtures are needed, since a mixture is busy only while it awaits the label of one of the last $H-1$ quarters.

We propose the following regret guarantee for the meta aggregator :

Proposition 1 (Regret of the meta-aggregator)**.** *Let $\mathcal{T} = \{t : C_t = 0\}$ and let N be the number of mixtures created. For any sequence of predictive models' forecasts and labels, and any candidate $j \in \{0, \dots, K\}$,*

$$\sum_{t \in \mathcal{T}} \ell_t(g_t q_t) \;\le\; \sum_{t \in \mathcal{T}} \ell_t(x_{j,t}) \;+\; \frac{N \,\log(1/u_j^0)}{\xi} \;+\; \max\{\theta, 1-\theta\}\, n_{\text{fade}}, \qquad N \le H, \tag{23}$$

where $\xi = 1/(2\max\{\theta, 1-\theta\})$ is the learning rate of the meta-aggregator and $n_{\text{fade}} = |\{t \in \mathcal{T} : g_t < 1,\ \tilde{y}_t = 1\}|$.

The left-hand side of 23 is the cumulative loss of the meta-aggregator over non-crisis quarters; the right-hand side is the cumulative loss of any candidate j, either the detection rule or one of the predictive models, plus two terms:

$$\underbrace{\sum_{t \in \mathcal{T}} \ell_t(g_t q_t)}_{\text{meta-aggregator}} \;\le\; \underbrace{\sum_{t \in \mathcal{T}} \ell_t(x_{j,t})}_{\text{candidate } j} \;+\; \underbrace{\frac{N \log(1/u_j^0)}{\xi}}_{\text{cost of not knowing } j} \;+\; \underbrace{\max\{\theta, 1-\theta\}\, n_{\text{fade}}}_{\text{cost of the gate}}.$$

The first term is the cost of not knowing in advance which candidate will perform best. It decreases with the prior weight of the candidate: $\log(1/u_j^0)$ is about 0.1 for the detection rule, whose prior weight is 0.9, and about 3.7 for each predictive model, whose prior weight is 0.025. The learning rate $\xi = 0.625$ is fixed, unlike the learning rate η_t of Section 5.1, which decreases over time: it is the largest rate for which the weighted squared loss is exp-concave (see Appendix 1), which is what makes the bound independent of the length of the sample. The second term is the cost of the gate. In the quarters that follow a crisis, the gate shrinks the forecast; if one of these quarters turns out to be a pre-crisis quarter, this can cost at most $\max\{\theta, 1-\theta\} = 0.8$ per quarter. The number n_{fade} of such quarters is zero in all countries but one. In short, the meta-aggregator can never lose much more than the best of its candidates, whichever it turns out to be.

The bound holds for any sequence of forecasts and labels [38], without assumptions on how the predictive models are generated, and its excess term is a constant that does not grow with

[38]The meta-aggregator provides a guarantee on the standard regret, but only by nesting the detection rule in a second aggregation; the rule itself has no useful bound. The online learning literature suggests a more direct route. Gaillard et al. (2014) show that if the base algorithm is run on modified losses, $\tilde{\ell}_{k,t} = I_{k,t}\,\ell_{k,t} + (1 - I_{k,t})\,\hat{\ell}_t$, where $I_{k,t} \in [0,1]$ is the confidence of predictive model k and $\hat{\ell}_t$ the loss of the aggregate forecast, and the weights are then rescaled by the

the length of the sample. Against the detection rule, whose prior weight is 0.9, it is about 3 in cumulative loss, less than the advantage of the rule over the predictive models in every country: the meta-aggregator cannot lose this advantage, and the realized regret, between -0.05 and 0.12, shows that it essentially reproduces the rule. Against each predictive model, the bound is about 106, uninformative on our sample; since it does not grow with time, however, it guarantees that in the long run the meta-aggregator does no worse than the best predictive model, should the detection rule fail. The guarantee is insurance, not a source of performance.

5.7 Evaluation

An early-warning system is useful only if its signals lead to good decisions. A policymaker who observes the forecast must decide, each quarter, whether to issue an alarm, and two errors are possible: missing a crisis, which leaves the banking system without the buffers it needs, and issuing a false alarm, which tightens capital requirements without cause. Statistical measures of fit, such as the mean squared error, weigh these errors symmetrically and do not say whether the signal would have triggered an alarm at the right time. Following the early-warning literature (Alessi and Detken, 2011), we therefore evaluate the forecasts by the utility of a policymaker who turns them into alarms with a threshold and weighs the two errors by their relative cost.

Definition 10 (Policymaker's utility)**.** For a signal $\hat{s}_t$ and threshold τ, let $\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$ be the counts of true positives, false positives, false negatives and true negatives over an evaluation set.

$$U(\tau) = 1 - \frac{\theta \, \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} + (1-\theta) \, \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}}{\min\{\theta, 1-\theta\}}, \qquad \theta = 0.8. \tag{24}$$

$U = 0$ is the utility of a naive rule (always or never issue an alarm) and $U = 1$ that of a perfect signal. The signal is smoothed by a backward rolling maximum over five quarters, $\hat{s}_t = \max_{0 \le j < 5} \hat{p}_{t-j}$, so that an alarm once issued remains active for a year, matching the activation

confidences, the resulting forecaster controls the confidence regret

$$R^{c}_{k,T} = \sum_{t=1}^{T} I_{k,t} \left(\hat{\ell}_t - \ell_t(p_{k,t}) \right),$$

the regret against predictive model k over the rounds in which it is confident, weighted by its confidence. Taking $I_{k,t} = \pi_{k,t}$, or $\pi_{k,t}(1 - F_{k,t})^{\kappa}$, which also lies in $[0,1]$, and handling the delay with the pool of mixtures of Joulani et al. (2013), would give the detection rule a guarantee of its own: on the quarters in which a channel is active, the aggregate forecast would do no worse than the corresponding predictive model, up to a term that grows as the square root of the confidence-weighted number of such quarters. This guarantee differs from the standard one, since it does not compare the forecaster with the best predictive model over the whole sample, but it is arguably the relevant one for early warning: what matters is that the forecaster follows the right channel when that channel is active. The confidences are computed from the predictive models' forecasts up to t, before the forecast of quarter t is issued, as the framework requires. We leave this extension for future research.

horizon of the countercyclical buffer. The threshold τ^* maximizes U on the batch evaluation set $\{t \leq T_B : r_t \leq T_B\}$ and is applied unchanged to the real-time evaluation set $\{t > T_B : r_t \leq T\}$. The last h quarters of the sample are excluded because their labels are not yet known.

We complement this assessment by providing, for each country and specification, the out-of-sample AUROC [39], the mean and minimum level of $\hat{p}_t$ on pre-crisis quarters, its mean level on calm and on crisis quarters, $U(\tau^*)$, whether the real-time episode is detected and with how many quarters of lead, the channel with the highest average weight during the pre-crisis window, and the realized regret of the meta-aggregator against its bound.

We also provide for each country an out-of-sample graphs of the time-varying probability of systemic crises. We argue that even a simple graph can provide a transparent means of assessing the performance of our methodology.A sizable, monotonic increase at the right time is a reliable and clear pre- crisis signal, which is valuable for macroprudential authorities. At which point those authorities may want to respond to that increased probability of crisis is a matter of judgement, which cannot be mechanical and will depend on external parameters, which we do not attempt to model.

6 Results

This section reports the real-time performance of the procedure over 2001Q2–2025Q4 (Section 6.1), reads the alarms it raises before the crises of 2007–2008 and after them (Sections 6.1 and 6.3), and documents, episode by episode, which channel and which indicators carry the signal (Section 6.2).

6.1 Real-time alarms

Figure 1 reports the aggregate probability of a pre-crisis regime for France, Germany, Sweden and the United Kingdom over the real-time period, together with the 12-quarter pre-crisis windows (light grey), the crisis periods (dark grey), the alarm threshold calibrated on the batch period

[39] It is worth noting that AUROCs are by no means a panacea to assess the performance of a model. The AUROC has the advantage of being an objective metric to assess the quality of a forecast by testing if the distribution of the predictions are significantly different under the two different tests, in the sens that it requires no subjective input. However, it has also several disavdantages. The interpretation of AUROC values largely depends on conventions specific to each field. Even in medical sciences, De Hond et al. (2022) note that "AUC values between 0.7 and 0.8 have been labelled as poor, moderate, fair, or good. We could not identify a justifiable reasoning on these value judgements in most papers; intuition and experience seem to have been dominant ". Moreover, AUROC can be a misleading measure of a forecast's predictive power. For instance, a predictive model who increases their pre-crisis probability from 0 to 0.00001 at the correct time would achieve an AUROC of 1, yet their forecast would not be particularly actionable or useful. More fundamentally, AUROC has been shown to be inherently incoherent with respect to misclassification costs, as it evaluates classifiers using different misclassification cost distributions. This implies that relying on AUROC is equivalent to applying different evaluation metrics to different classification rules. (Hand (2009)).

(dashed line) and the first real-time alarm (triangle).

To turn the aggregate probability into an alarm, we smooth it by a backward rolling maximum over five quarters, $\hat{s}_{i,t} = \max_{0 \le j < 5} \hat{p}_{i,t-j}$, so that an alarm, once issued, remains active for one year, the time needed for a countercyclical capital buffer to become binding. The alarm threshold τ_i^* is chosen separately for each country to maximize the relative utility $U(\tau)$ of Alessi and Detken (2011) with $\theta = 0.8$ over a grid $\tau \in \{0.02, 0.03, \ldots, 0.98\}$, using only batch-period quarters whose label is known at T_B. It is then frozen and applied unchanged to the real-time period, so that no information after T_B enters its choice. An alarm is a sequence of consecutive quarters in which $\hat{s}_{i,t}$ exceeds τ_i^*. The first real-time alarm associated with a crisis is the first quarter of the alarm still in progress in the quarter preceding the onset; if no alarm is in progress at that date, it is the first quarter of the pre-crisis window in which the threshold is crossed.[40] Because the predictive models' batch predictions are in-sample, the batch aggregate is well separated and τ_i^* can be high. If the real-time signal is less sharp, a high threshold delays or prevents the alarm: this is a conservative bias, not a source of look-ahead.

[40] The lead time is the number of quarters between the first alarm and the onset. It can exceed h when the alarm starts before the pre-crisis window. The quarters of that alarm that precede the window are nonetheless counted as false positives in the utility, so that early warnings are not rewarded in our performance statistics.

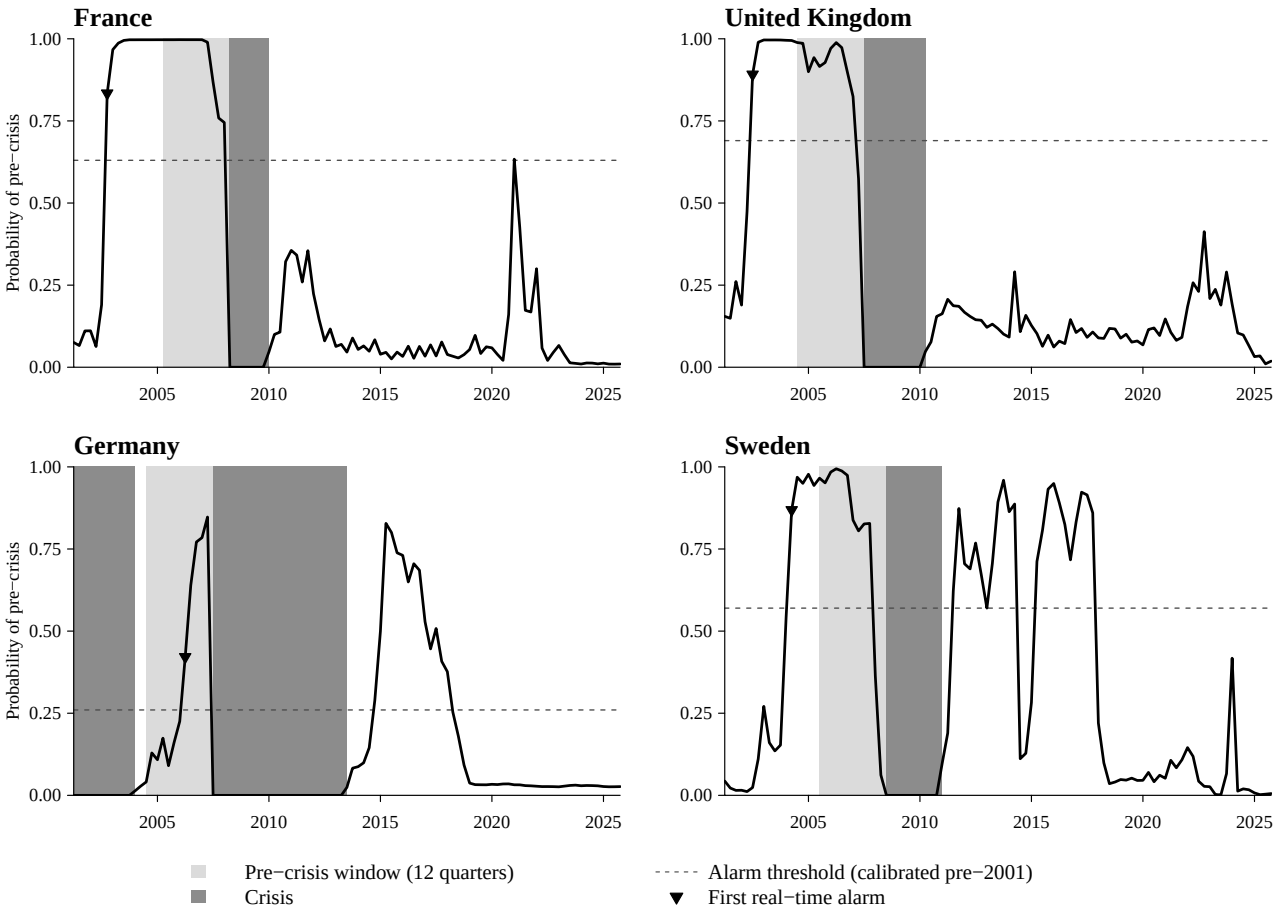


Figure 1: Probability of a pre-crisis regime, real-time period

Aggregate probability $\hat{p}_{i,t}$ from the detection-based aggregation, 2001Q2–2025Q4. Light grey: 12-quarter pre-crisis window. Dark grey: crisis. Dashed line: alarm threshold calibrated on the batch period (1987–2001) and applied unchanged. Triangle: first real-time alarm.

As seen in graph 1, we argue that the procedure we propose provides a convincing answer to the Queen. Table 3 summarizes real-time performance. All four crises of 2007–2008 are preceded by an alarm in progress at the onset, with leads of 22 quarters in France, 20 in the United Kingdom, 17 in Sweden and 5 in Germany. The procedure separates pre-crisis from calm quarters in every country: the out-of-sample AUROC ranges from 0.83 in Sweden to 0.96 in France and averages 0.88. The aggregate probability averages 0.76 in pre-crisis quarters against 0.29 in calm quarters. In France, the United Kingdom and Sweden, the signal is both early and strong: the probability exceeds 0.80 on average over the pre-crisis window, the smoothed signal stays above the threshold throughout the window, and the relative utility is positive, between 0.51 and 0.85. Germany is the exception: its signal rises only in 2006, its average probability over the window is 0.37, and its utility is negative (-1.63). More importantly, we emphasize that providing graph 1 to policymakers would convey valuable information.

Table 3: Real-time performance of the detection-based aggregation (baseline, $h = 12$)

Country	AUROC	Pre.	Min. pre.	Calm	Crisis	Thres.	Util.	TP	FP	FN	Lead	Channel
France	0.96	0.94	0.75	0.24	0	0.63	0.75	12	19	0	22	Real estate
United Kingdom	0.90	0.91	0.57	0.24	0	0.69	0.85	12	11	0	20	Real estate
Germany	0.84	0.37	0.04	0.24	0	0.26	-1.63	5	22	7	5	Real estate
Sweden	0.83	0.80	0.06	0.42	0	0.57	0.51	12	37	0	17	Real estate

Real-time period 2001Q2 onward, excluding the last h quarters whose labels are unknown. Threshold calibrated on the batch period and applied unchanged. Utility is the Alessi–Detken relative utility with $\theta = 0.8$. Lead is the number of quarters between the start of the alarm in progress at the onset (smoothed signal above the threshold without interruption) and the onset; it can exceed h. Alarm before window is the number of those quarters that precede the pre-crisis window; they are counted as false positives in FP and in the utility. Channel is the predictive model with the highest average weight in the pre-crisis window.

Two preliminary checks are particularly informative about the efficiency of the procedure. First the aggregation does as well as the best model, without knowing it in advance.Table 17 compares the cumulative loss of the detection rule with that of each predictive model. In every country, the detection rule loses far less than the credit, external and global-cycle models, and about as much as the real-estate model, which turns out ex post to be the best single model in all four countries: 2.52 against 2.47 in France, 3.65 against 6.64 in the United Kingdom, 12.10 against 13.10 in Germany and 5.58 against 4.23 in Sweden. This is the property the procedure is designed to deliver. The identity of the best channel is not known in real time: in the batch period, the credit and real-estate models have almost the same utility and receive almost the same prior weight, 0.43 and 0.41 (Table 12), and the external model is the best one in the German batch period. Second the meta-aggregator, which carries the regret guarantee of Proposition 1, leaves these conclusions unchanged (Table 11 and Figure 13 in Appendix 9.11.1). Its regret bound holds in all 20 country–candidate pairs; its realized regret against the detection rule is 0.07 on average, against a bound of 3.63; its AUROC, utility and leads are identical to those of the detection rule in every country. As expected from mixing the detection rule with the individual models, it slightly lowers the level of the signal in the pre-crisis window, from 0.94 to 0.89 in France, from 0.91 to 0.85 in the United Kingdom, from 0.80 to 0.75 in Sweden and from 0.37 to 0.36 in Germany. The guarantee is thus obtained at a small cost in sharpness and at no cost in detection.

6.2 Which channel, which variables? Reading the episodes

The aggregation produces, for each episode, three objects that can be read together: the CUSUM statistic of each predictive model (Figure 8), which measures the evidence that the model has entered a pre-crisis regime; the weights of the aggregation (Figure 3), which turn this evidence into a diagnosis of the channel at work; and the real-time re-estimations of each model (Figures 9

to 12 and Tables 13 to 16), which tell which indicators carry the signal.

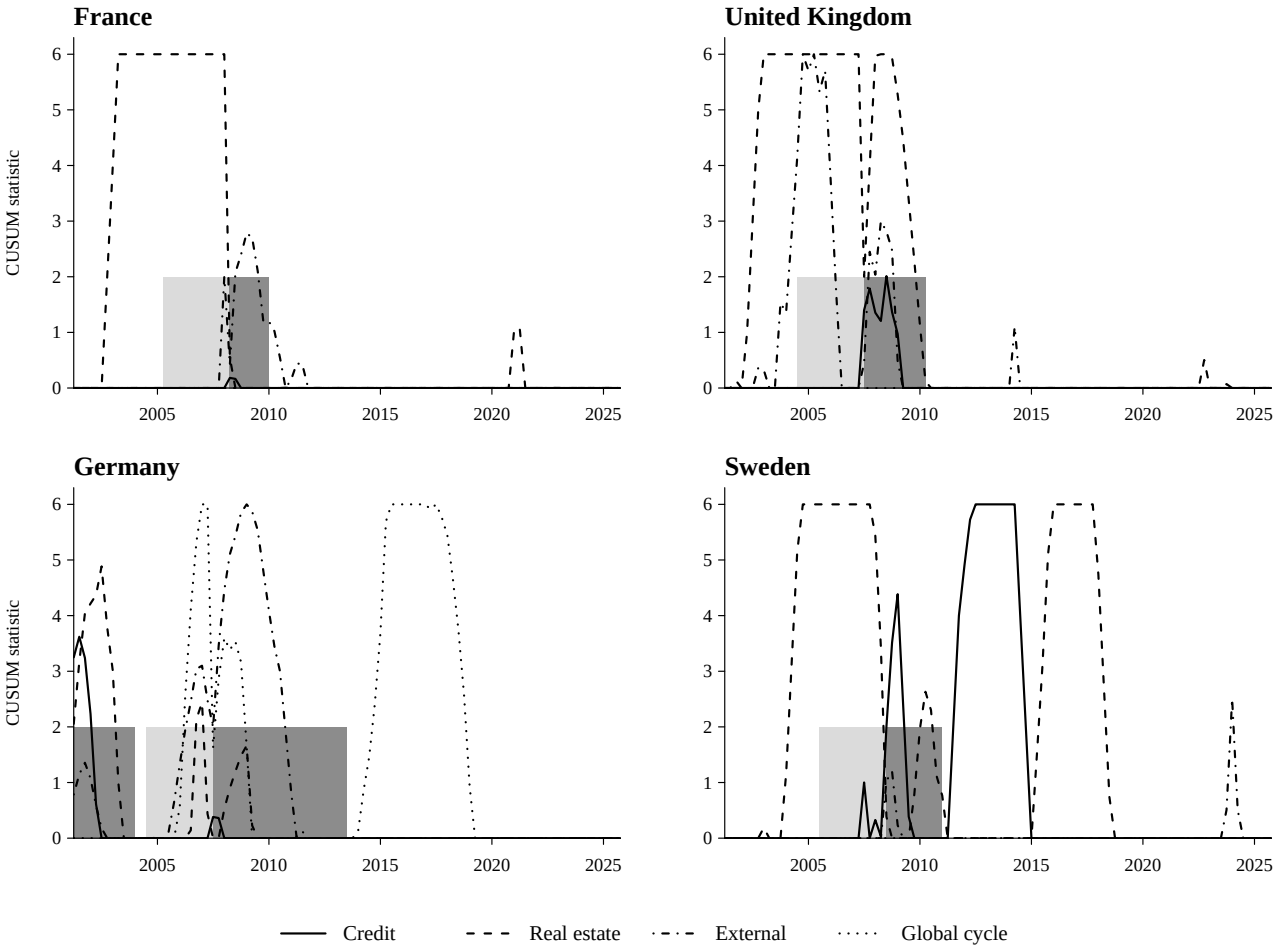


Figure 2: CUSUM statistics by predictive model
Bounded CUSUM statistic $W_{k,t}$ of each predictive model, capped at $W_{\max} = 6$ and reset at each crisis onset. Light grey: pre-crisis window. Dark grey: crisis.

In calm periods, the statistics stay at or near zero: as shown in Lemma 3, the increment of a calibrated predictive model has a negative drift when no vulnerability builds up. When a model's channel starts signalling a regime change, its statistic rises steadily until it reaches the cap $W_{\max} = 6$, and the weight of that model increases accordingly (Figure 3). In the four countries, the weight moves toward the active channel between one and three quarters after its statistic starts rising, and years before the onset reveals which channel was right.

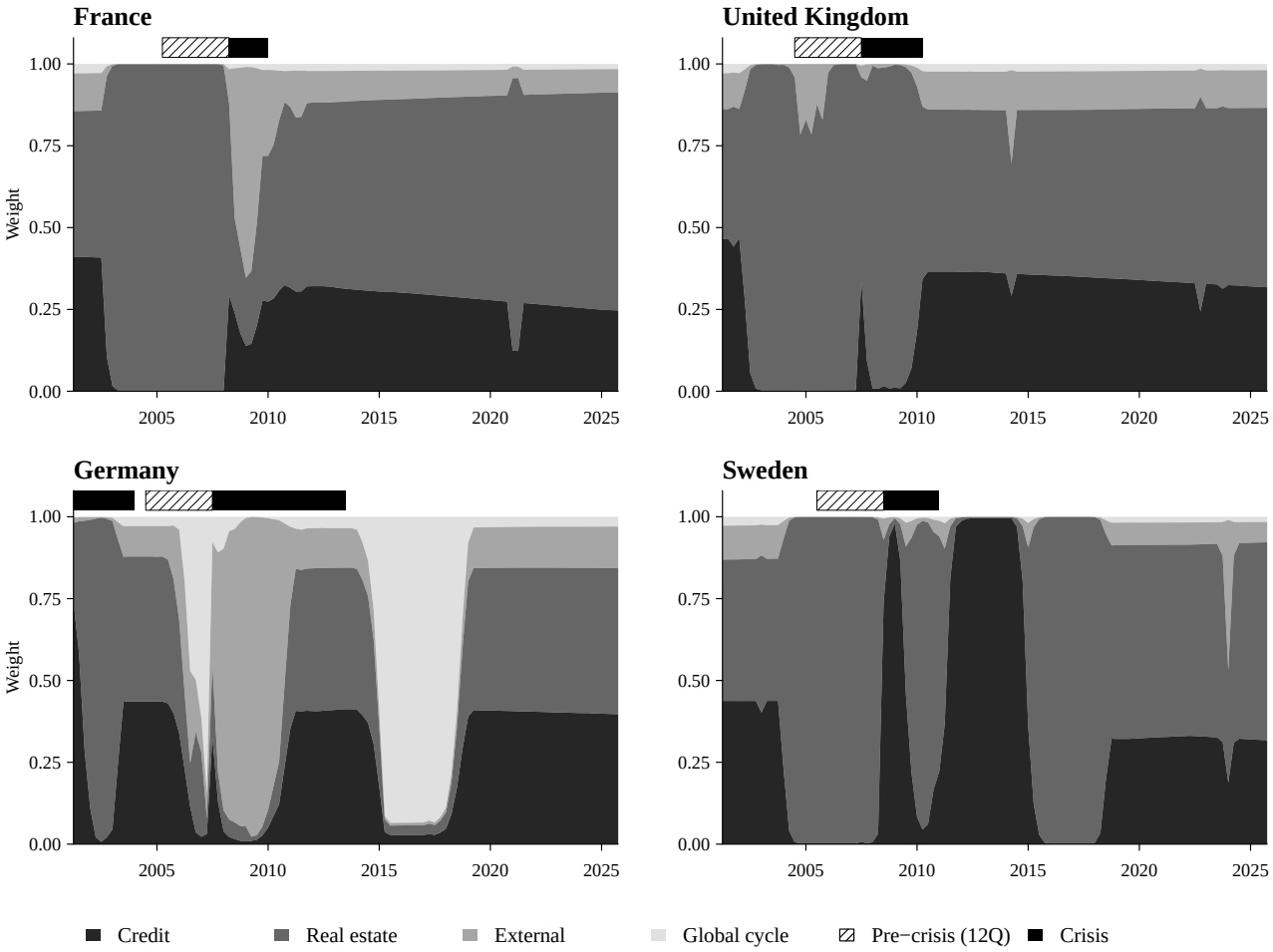


Figure 3: Weights of the aggregation, real-time period
Stacked weights $w_{k,t}$ of the four predictive models. Hatched bar: 12-quarter pre-crisis window. Black bar: crisis.

To go beyond the weights of the aggregation, we look directly at the coefficients of the predictive models. The coefficients of penalized regressions are biased toward zero by construction, their selection among correlated indicators is partly arbitrary, and they come without valid standard errors. We therefore do not interpret their magnitudes, and we interpret their signs only under some very precise conditions. Following the logic of stability selection (Meinshausen and Bühlmann, 2010), in a sparse model estimated on few events, what is informative is not the size of a coefficient but the persistence of its selection and of its sign. We exploit the fact that each model is re-estimated at every date of the real-time period, and we summarize the sequence of estimates by two statistics: the frequency with which an indicator is selected, and the stability of the sign of its coefficient when selected. Figures 9 to 12 display, for each country and model, the sign of every coefficient at every re-estimation date, with the pre-crisis window framed; Tables 13 to 16 report selection frequencies and sign stability averaged across countries.

The interpretation of a sign depends on how the indicator enters the model. Each model

contains an indicator in level together with its one-, two- and three-year growth rates, so that every coefficient is conditional on the others. For the level, this conditioning is innocuous: holding the growth rates fixed, a higher level shifts the whole past trajectory upward, and a positive coefficient reads as a higher level raising the probability of a pre-crisis. [41] The sign of a growth rate conditional on the level therefore describes the shape of the trajectory, and need not match the sign of the unconditional correlation between growth and the target; the same applies to an indicator entered together with a closely related one, such as the price-to-rent ratio given the price-to-income ratio. Because these regressors are strongly collinear, such conditional signs are also fragile. We therefore adopt the following rule. We interpret the sign of an indicator only if it is stable across re-estimations and across countries; among such indicators, we interpret the sign of a level, and the sign of a growth rate only when it agrees with economic theory, as for the growth of credit. All other signs, unstable, or stable but conditional and counterintuitive, are reported but not interpreted. An indicator is thus a credible predictor when it is selected in most re-estimations with a stable sign that admits this interpretation [42]

Three regularities emerge from the cross-country averages of Tables 13 to 16. First, within the real-estate model, the price-to-income ratio is selected in 97 percent of all re-estimations, with a positive sign in 99 percent of them, and the one-year growth of rent price index in 90 percent, with a positive sign throughout: the real-estate signal is a valuation signal, not a momentum signal. Second, within the credit model, the two- and three-year growth of credit to non-financial corporations and the one-year growth of household credit are selected in 98 percent of re-estimations with a positive and stable sign, whereas the credit-to-GDP gap in level is selected in fewer than half of them with an unstable sign: once growth rates are included, the gap adds little, which is consistent with the critique of the gap as a real-time indicator (**?**). Third, within the global-cycle model, the three-year growth of the NASDAQ is selected in 96 percent of re-estimations with a negative and stable sign, and risk appetite in 60 percent with a positive sign: conditional on the level of equity prices, a boom that has already run for three years lowers the probability of a pre-crisis, a sign that we read as the model dating the position in the global cycle rather than its level. We do not interpret the sign of the financial conditions index, which depends on the convention of the index.

[41] For a growth rate, it is not: since $\Delta_{12} x_t = x_t - x_{t-12}$, a model $a\, x_t + b\, \Delta_{12} x_t$ can be written $(a+b)\, x_t - b\, x_{t-12}$, so that a negative b does not say that growth lowers the probability of a pre-crisis, but that, for a given current level, a level reached earlier raises it.

[42] A final caution concerns timing. Since the models are estimated on resolved labels, the coefficients displayed during a pre-crisis window are those of a model that has not yet learned from that window: they are the coefficients of the model the forecaster actually used in real time.

In the following paragraphs we describe how our general procedure detects the Global Financial Crisis.

France, 2008: *la crise systémique en France* The French episode is the cleanest. The real-estate model's forecast rises above 0.9 in 2002Q4 and stays there until the onset: it averages 0.95 over the pre-crisis window, against 0.04 for the credit model, 0.17 for the external model and 0.02 for the global-cycle model (Figure 3). Its CUSUM statistic reaches half its cap in 2003Q1 and the cap in 2003Q2, where it remains for twenty consecutive quarters until the onset of 2008Q2, while the three other statistics stay at zero. Within the real-estate model, the alarm rests on two kinds of indicators. The first is a valuation ratio: the price-to-income ratio is selected with a positive sign at every re-estimation of the window. The second is the momentum of prices: the three-year growth of real house prices, in both the national and the BIS measure, and the one-year growth of the BIS real residential property price index are selected with positive and stable signs, as they are in the other countries [43]. A positive level together with positive growth rates describes prices that are both high relative to incomes and still rising: what the model reads as a pre-crisis regime is neither overvaluation alone nor price growth alone, but an overvaluation that has kept building for several years [44]. It is also the configuration of high and rising asset prices that Greenwood et al. (2022) find to predict crises, with one difference: the credit component is absent, since the forecast of the credit model stays near zero throughout the window.This reading calls for a caution. The French banking crisis of 2008 did not originate in domestic mortgage losses [45]. The housing signal should therefore be read as a marker of the phase of the domestic financial cycle, a period of rising valuations and risk-taking, rather than as the mechanism of the crisis (?). The model, estimated on the housing-driven crises of the early 1990s, recognizes in 2003–2008 the configuration that preceded them; that the crisis which ended this phase had a different proximate cause is consistent with valuations and risk-taking rising together across asset classes, and is not evidence that housing caused it.

[43] This configuration describes the French housing market of the 2000s, well described by the French Statistical Office after the Great Financial Crisis: prices of existing dwellings doubled between 2000 and 2007, whereas disposable income per household rose by 22 percent and rents by 24 percent, so that the price-to-income ratio rose by 69 percent (Rougerie and Friggit, 2010); the rise was sustained by falling interest rates and longer loan maturities, and by early 2008 the implied risk premium on housing had turned negative (Clévenot, 2011).

[44] The remaining signs of the model, negative on the two-year growth of the price-to-income ratio and on the level of the BIS price index given the national one, are conditional on closely related regressors, and we do not interpret them.

[45] House prices fell by about 7 percent in 2009 before recovering, French mortgage lending is mostly at fixed rates and underwritten on the borrower's income rather than on the value of the property (Clévenot, 2011), and the distress of French banks came from their exposure to US structured credit and their reliance on wholesale funding.

United Kingdom, 2007 ***the systemic crisis in the UK*** The British episode reflects the same underlying crisis as the French case, while adding two distinctive features: an external signal in the middle of the window, and a fading of the housing signal at its end that the accumulation of evidence absorbs. The real-estate model detects the boom early. Its forecast averages 0.80 over the calm quarters of 2001–2004, before the window opens, and 0.94 over the window; its statistic reaches the cap in 2003Q1 and stays there for eighteen consecutive quarters until the onset, and its weight exceeds 0.5 from 2002Q2 and averages 0.92 over the window. The external model adds a second signal in 2004–2005. Its forecast averages 0.37 over the window with peaks close to 1, its statistic reaches half the cap in 2004Q3 and the cap in 2004Q4, stays close to it through 2005, and returns to zero in 2006Q3. For these quarters, two channels signal at once and the posterior shares the weight between them: the external model receives up to 0.22 of the weight, against 0.08 on average over the window. The indicator behind this signal whose sign is stable across re-estimations and countries is the current account, which enters with a negative sign: a larger deficit raises the probability of a pre-crisis regime. The external signal thus marks the years in which the British boom was financed by a widening deficit, and fades before the onset.The end of the window illustrates the role of accumulation. As the Bank of England raised Bank Rate five times between August 2006 and July 2007, from 4.50 to 5.75 percent (Bank of England, 2199), and quarterly house price inflation slowed from 3.3 percent in 2006Q4 to 2.4 percent in 2007Q2 (Bank of England, 2007, Table 1.A), the real-estate forecast fell from 0.98 in 2006Q3 to 0.58 in 2007Q2. A rule based on the level of the forecast would have weakened the alarm in the last quarters before the crisis; the CUSUM statistic did not, because a forecast that remains above one half still adds evidence for the pre-crisis regime, and the alarm remained at its maximum at the onset.Unlike the French one, the British crisis did originate where the signal pointed. Northern Rock, the first British bank to fail, was a mortgage lender that had financed the housing boom by securitization and wholesale borrowing [46] The two channels that the procedure identifies, a housing boom and its external financing, are therefore the two sides of the balance sheet that failed.

Sweden, 2008: ***den systemiska krisen i Sverige*** The Swedish episode has the same core as the French and British ones, with two differences: a credit signal that appears in the final year, and a housing signal that fades before the onset. The real-estate model's forecast averages 0.44 over the calm quarters of 2001Q2–2005Q2, rising from 2004, and 0.80 over the window; its statistic

[46] According to Shin (2009), between 1998 and 2007 its balance sheet grew by a factor of 6.5, while retail deposits fell from 60 to 23 percent of its liabilities, and the run of September 2007 was a withdrawal of wholesale creditors before it was a queue of depositors.

reaches the cap in 2004Q4 and stays there for thirteen consecutive quarters, until 2007Q4, and its weight exceeds 0.5 from 2004Q1 and averages 0.99 over the window. The indicators are valuation ratios: the price-to-rent and price-to-income ratios are selected throughout with a positive sign, together with the three-year growth of the price-to-income ratio, and its one-year growth until 2007. The levels of the national and BIS real house price indices enter negatively, conditional on these ratios; we do not interpret them. The diagnosis matches that of the Riksbank at the time, which noted in December 2007 that "house prices and household debt are continuing to rise at higher rates than household income"(Sveriges Riksbank, 2007).The credit model adds a feature absent elsewhere. Its growth rates of household and corporate credit are selected with a positive sign at every horizon and at every re-estimation of the window, but its forecast remains low until 2007: the model monitors credit throughout, and credit growth reaches the range that preceded the crises of the early 1990s only in the final year. The forecast rises to 0.30 in 2007Q2 and 0.73 in 2007Q3 and averages about 0.5 over the last four quarters before the onset. Its statistic does not exceed one sixth of the cap, so that the credit model receives almost no weight (0.01 on average over the window). The sequence of a housing boom followed by an acceleration of credit is nonetheless the one that the Swedish authorities would identify again after 2010 (Section 6.3). The end of the window is less favourable to the procedure than in the other countries. The real-estate forecast falls from 0.83 in 2007Q4 to 0.36 in 2008Q1 and 0.05 in 2008Q2, and its statistic declines from the cap to 3.4. The alarm is still in progress at the onset of 2008Q3, but only because of the one-year persistence of alarms; a procedure without that persistence would have lifted the alarm one quarter before the crisis. As in France, the crisis did not originate where the signal pointed. The growth of Swedish bank lending in 2007 had occurred "in particular to borrowers abroad, mainly in the Baltic countries"(Sveriges Riksbank, 2007), and when the crisis came, the largest losses expected by the Riksbank, out of SEK 155 billion over 2009–2011, were those of the banks' Baltic operations (Sveriges Riksbank, 2009). The domestic housing signal should therefore be read, as in France, as a marker of the phase of the financial cycle in which Swedish banks expanded their balance sheets, at home and abroad, rather than as the source of their losses.

Germany, 2007: *die systemische Krise in Deutschland* Germany is the episode the procedure handles least well, and the three lenses explain why. The German crisis did not follow a domestic boom: it was imported through the exposure of German banks, the Landesbanken and banks such as IKB, to US structured credit and to the asset-backed commercial paper conduits they sponsored (Acharya and Schnabl, 2010; Puri et al., 2011). No domestic model has much to detect.

Over the window, the credit model's forecast averages 0.04 and the real-estate model's 0.21; the external model averages 0.54 and the global-cycle model 0.45, and neither is high for long. Until 2005, the weights are those of the prior corrected by the learning layer, 0.44 on credit and 0.44 on real estate, and the aggregate forecast stays below 0.2. The external model illustrates what the detection rule requires. From 2005Q3, its forecast hovers between 0.5 and 0.66: a model that leans toward the pre-crisis regime without being confident. Since the increment of the statistic is the log-odds of the forecast against one half, such a forecast adds little evidence per quarter, about 0.4 for a forecast of 0.6, and the external statistic takes five quarters to rise from 0.1 to half the cap, which it reaches in 2006Q4. The global-cycle model moves later but more sharply: its forecast rises from 0.23 in 2006Q1 to 0.92 in 2006Q2, its statistic reaches half the cap in 2006Q3 and the cap in 2007Q1, two quarters before the onset, and its weight rises from 0.04 to 0.84 within five quarters. The real-estate model adds a short signal, with a forecast of 0.93 in 2006Q4 and a weight of 0.31, which vanishes in 2007Q2. The aggregate forecast exceeds 0.6 from 2006Q3, four quarters before the onset.The indicators behind this signal are also the least readable of the four episodes. Within the global-cycle model, the signal is carried by the levels of real equity prices and of the NASDAQ and by the three-year growth of the NASDAQ, all with positive signs; within the external model, by the level of exports and by the exchange rate, with positive signs. These signs are stable across the German re-estimations but not across countries: equity prices and exports enter with different signs elsewhere, and the level of the NASDAQ enters positively in Germany only. Under the rule of Section 6.2, we report them without interpreting them. A likely reason for this German specificity lies in the dating of German crises, which is less favourable to learning than in the other countries. While the batch model is estimated on the panel of the four countries, the real-time re-estimations use the data of each country only. Germany, unlike France, Sweden and the United Kingdom, had no crisis in the early 1990s: until 2007, the only German pre-crisis quarters are the twelve quarters of 1998Q1–2000Q4 that preceded the crisis of 2001, and these coincide with the boom in technology stocks. With a single pre-crisis episode of this kind, the German models learn to associate a pre-crisis regime with high equity prices, which explains why the levels of equity prices and of the NASDAQ enter positively in Germany and not elsewhere. The timing of the next crisis then works against the procedure. The crisis of 2001 lasted until 2003Q4, so that the pre-crisis window of the 2007 crisis opens in 2004Q3, three quarters after its end, while the gate is still returning to one and while equity prices are close to the trough they reached in 2003. A model that reads high equity prices as a sign of vulnerability cannot signal at the beginning of such a window; it signals only once the recovery of equity prices is well advanced, in 2006. Nor can it

learn from the post-crisis years in time: since a calm quarter is revealed only h quarters later, the labels of 2004 become available only in 2007. The German episode therefore associates high equity prices with a pre-crisis regime and the one external indicator whose sign is stable across countries, the current account, enters negatively, so that a deficit raises the probability of a pre-crisis regime. In Germany, where the surplus widened over the window, it lowered the external forecast rather than raising it. The German signal of 2006–2007 is therefore neither domestic nor the signal of an external imbalance in the usual sense: it is a signal of the position in the global equity cycle, consistent with a crisis transmitted through the exposure of banks to foreign assets, but not one that identifies that exposure. The two domestic statistics that rise in 2001–2002, credit and real estate, do so inside the crisis of 2001–2003, when the forecasts are neutralized by the gate.

One important finding is that the channel that carries the alarm differs across countries and episodes: real estate in France, Sweden and the United Kingdom before 2008, with an external component in the United Kingdom; the global equity cycle in Germany; credit, then real estate, in Sweden after 2011; the global cycle in Germany after 2015; and real estate in France in 2021. This finding must be read for what it is: a statement about predictability, not about causality. The crises of 2007–2008 had a largely common proximate cause, the losses on US mortgage-related assets and the collapse of the wholesale funding markets that financed them, and in France, Sweden and Germany the distress of banks came from that shock and from their exposures abroad rather than from losses on domestic housing or domestic credit. The United Kingdom is the partial exception: there, the bank that failed first had financed the domestic housing boom in the wholesale market. What the procedure identifies is therefore not the cause of each crisis, but the domestic markers of the phase of the financial cycle in which banks expanded their balance sheets and became vulnerable to such a shock: overvalued housing in three countries, the global equity cycle in the fourth. That these markers differ across countries is consistent with the evidence that credit booms (Schularick and Taylor, 2012) and housing booms (Jordà et al., 2015) are distinct and partly independent signatures of financial fragility, whose relative importance varies across countries even when the trigger of the crisis is common. It is also what makes a detection-based aggregation necessary. A system calibrated on a single marker would have missed either the German episode or the three others, and a system weighting markers by their past performance would have waited for the labels, which arrive only at the onset, by which time the question of which marker was informative has lost its value for prevention.

6.3 False alarms

Outside the pre-crisis windows, the signal crosses the threshold in three episodes. In Sweden from 2011Q3 to 2018Q4, the credit model's forecast rises above 0.9 in 2011Q3, its statistic is at the cap from 2012 to early 2014, and its weight exceeds 0.5 from 2011Q3 to 2014Q4; the signal is carried by the growth rates of household and corporate credit and by the credit-to-GDP gap, with positive signs. From 2015, the real-estate model takes over: its statistic is at the cap in 2016–2017 and its weight exceeds 0.5 from 2015Q1, on the same valuation ratios as before 2008; it falls back in 2018, when house prices drop. In Germany from 2014Q4 to 2019Q1 the global-cycle model's statistic is at the cap from 2015 to 2018 and its weight exceeds 0.5 from 2015Q1 to 2018Q3, on the same indicators as in 2006–2007: this is a signal of compressed risk premia rather than of a domestic imbalance. In France from 2021Q1 to 2022Q1,the real-estate model's forecast rises to 0.74 in 2021Q1 and its statistic rises to about 2 in 2021; the weight of the real-estate model was already above 0.5 since 2010, so that the aggregate forecast follows it and crosses the threshold for four quarters.

None of these episodes was followed by a systemic crisis, and our performance statistics count them as false alarms. Each of them, however, coincides with a period in which the macroprudential authorities themselves identified a build-up of vulnerabilities and tightened policy. An early-warning system is evaluated against crises that occurred, but a warning that leads to policy action may prevent the crisis it announces; in that case the alarm is a false positive by construction, not by error. We cannot establish such a causal link, but the timing of the episodes is informative.

Sweden, 2011–2018. The Swedish signal is carried first by the credit channel, whose statistic rises in 2011 and stays at its cap from 2012 to early 2014, and from 2015 by the real-estate channel, at its cap from 2016 to the end of 2017; it raises the average probability in calm quarters to 0.42. Both episodes match the concerns of the Swedish authorities over the period. Household credit grew faster than income, and house prices rose by about 8 percent a year in 2014 (Finansinspektionen, 2015). The authorities responded with a sequence of measures: a loan-to-value cap of 85 percent on new mortgages in 2010, a risk-weight floor on mortgages of 15 percent in 2013, raised to 25 percent in 2014, the activation of the countercyclical capital buffer at 1 percent, decided in 2014, an amortisation requirement for new mortgages above 50 percent loan-to-value in June 2016, and an additional requirement for borrowers above 4.5 times their income in March 2018 (Finansinspektionen, 2018). In November 2016, the European Systemic Risk Board issued a warning to Sweden on medium-term vulnerabilities in residential real estate (European Systemic Risk

Board, 2016). House prices fell by about 8 percent in the last quarter of 2017 (Finansinspektionen, 2018), and the Swedish signal falls back in 2018.

Germany, 2014–2019. The German false alarm is carried by the global-cycle channel, whose statistic is at its cap from 2015 to 2018: the predictive model built on global risk appetite, equity prices and financial conditions signals an environment of compressed risk premia and easy financing conditions. The German authorities reached a similar assessment. The Financial Stability Committee recommended activating the countercyclical capital buffer, and BaFin raised it from zero to 0.25 percent in 2019, its first activation, on the grounds that cyclical systemic risks had built up during an extended period of favourable conditions (Bundesanstalt für Finanzdienstleistungsaufsicht, 2019). In September 2019, the European Systemic Risk Board issued a warning to Germany on vulnerabilities in residential real estate (European Systemic Risk Board, 2019).

France, 2021–2022. The French false alarm is carried by the real-estate channel, whose statistic rises in 2021. It coincides with the tightening of borrower-based measures in France. In December 2019, the High Council for Financial Stability recommended capping the debt-service-to-income ratio of new mortgages at 33 percent and their maturity at 25 years; the cap was raised to 35 percent in January 2021, and the measure was made legally binding from January 2022 (Haut Conseil de stabilité financière, 2021). In September 2019, the European Systemic Risk Board had issued a warning to France on vulnerabilities in residential real estate (European Systemic Risk Board, 2019).

These coincidences do not prove that the measures averted crises, and the cross-country evidence on the effectiveness of macroprudential policy, which finds that borrower-based measures such as loan-to-value and debt-to-income limits dampen credit growth and house prices, is not specific to these episodes (Cerutti et al., 2017). We leave this issue for future research. They suggest, however, that the false alarms of the procedure are not random: they flag the episodes that the authorities themselves considered as build-ups of vulnerabilities, through the same channels.

7 Robustness checks

This section asks whether the results depend on the choices made in building the procedure. We consider four groups of choices, in increasing order of how much they change the exercise. First, the parameters of the detection rule: the cap of the CUSUM statistic, the temperature of the posterior, the exponent of the reliability weight and the length of the return period of the gate.

Second, the prior weights of the meta aggregator, which we replace by alternatives that do not use the batch performance of the models. Third, the components of the aggregation, which we remove one at a time (the change-point detection, the learning layer and the gate) to measure what each contributes. Fourth, the definition of the exercise itself: a shorter pre-crisis horizon of eight quarters instead of twelve, and a real-time specification restricted to indicators that are neither revised nor constructed with full-sample information. For each variant, we ask not only whether the four crises are still detected and how the performance statistics change, but, for the last two, whether they are detected for the same reasons: through the same channel and, within that channel, through the same indicators with the same signs. In the section, the he baseline specification is the specification of Section 9.11: AUROC 0.884, average probability of 0.754 in pre-crisis quarters and 0.286 in calm quarters, utility 0.120, four crises detected. The utility is the most fragile of these statistics: it depends on a discrete threshold calibrated on the batch period, and its cross-country average is dominated by the negative value of Germany, so that small changes in the German threshold move it substantially. We therefore read the utility as a secondary indicator and give more weight to the AUROC, to the level of the signal in pre-crisis and calm quarters, and to the number of crises detected.We also emphasize that plots of real-time predicted probabilities provide a valuable mean of evaluating forecast performance.

7.1 Detection parameters

The parameters of the detection rule are fixed a priori rather than estimated: the clip of the log-likelihood ratio (c), the cap of the CUSUM statistic ($W_{\max}$), the weight of the reliability correction (κ) and the length of the post-crisis fade (ϕ). Table 4 reports the distribution of performance over the 81 combinations of $c \in \{1, 2, 3\}$, $W_{\max} \in \{4, 6, 8\}$, $\kappa \in \{0, 0.5, 1\}$ and $\phi \in \{2, 4, 8\}$. The results are insensitive to these choices. The AUROC lies between 0.877 and 0.893 across all combinations, the average probability between 0.721 and 0.766 in pre-crisis quarters and between 0.252 and 0.298 in calm quarters, the utility between 0.113 and 0.167, and the four crises are detected before their onset in every one of the 81 combinations. The baseline specification lies between the first quartile and the median of the distribution for the AUROC and the pre-crisis level.

Table 4: robustness grid

Statistic	AUC	pre	crise	calme	U	n_detect
Min	0.877	0.721	0	0.252	0.113	4
Q25	0.881	0.733	0	0.269	0.120	4
Median	0.888	0.752	0	0.278	0.133	4
Q75	0.890	0.757	0	0.288	0.147	4
Max	0.893	0.766	0	0.298	0.167	4

7.2 Components of the aggregation

Table 5 removes the components of the aggregation one at a time. Removing the change-point detection, so that the weights depend only on the prior, on revealed losses and on reliability, has the largest effect, and an instructive one. The AUROC rises, from 0.884 to 0.917, but the signal no longer rises before crises: its average level in pre-crisis quarters falls by half, from 0.754 to 0.377, one of the four crises is no longer detected, and the utility of the policymaker turns from 0.120 to -0.920. Without detection, the aggregate is a slowly varying average of the four models; it ranks quarters well, because the models that are high before crises are also high relative to calm periods, but it stays low in absolute terms, since in each pre-crisis window only one or two of the four models signal and the others pull the average down. Detection does not improve the ordering of quarters; it is what turns a correct ordering into a timely alarm, by moving the weight to the model that signals the change of regime as soon as its evidence accumulates. This has a cost, visible in calm quarters, where the average signal is higher with detection (0.286) than without (0.157): the same mechanism that concentrates the weight on a signalling model before a crisis also does so in the episodes after 2010, which Section 6.3 discusses.

Removing the delayed-feedback learning layer leaves the AUROC and the level of the signal unchanged (0.885 against 0.884, and 0.758 against 0.754 in pre-crisis quarters) and raises the utility from 0.120 to 0.220. This is the empirical counterpart of Theorem 3: with a feedback delay of at least twelve quarters, the losses that would move weight toward the relevant model are revealed only once the crisis has started, so that learning from them contributes nothing to the detection of the next crisis; what it contributes over the sample is a slow reweighting after the crises, which moves the signal of a few quarters across the threshold. Removing the gate lowers the AUROC to 0.815 and the utility to zero, while leaving the pre-crisis level unchanged (0.754). This reflects the definition of the target rather than an advantage given to the procedure: crisis quarters are coded zero, and without the gate the models' forecasts, which remain high during crises, count as false alarms. Removing both the learning layer and the gate gives the AUROC of the gate-free

rule (0.820), with a utility of 0.083. The gains of the procedure therefore come from detection: learning can be removed at no cost, and the gate only aligns the forecast with the coding of crisis quarters.

Table 5: robustness ablation

variant	AUROC	Pre-crisis	Calm	Utility	Detected
cusum_only	0.820	0.759	0.288	0.083	4
full	0.884	0.754	0.286	0.120	4
no_detection	0.917	0.377	0.157	-0.920	3
no_gate	0.815	0.754	0.293	-0.000	4
no_learn	0.885	0.758	0.282	0.220	4

7.3 Prior weights of the meta aggregator

Table 6 compares the pooled prior with three alternatives: a uniform prior, a prior computed on the batch period of the target country only, and a prior computed on the batch period of the three other countries. The choice matters little. All four priors detect the four crises before their onset, the AUROC varies between 0.880 and 0.886, the pre-crisis level between 0.737 and 0.767 and the calm level between 0.282 and 0.302. The uniform prior gives the highest utility (0.300) together with the highest level in calm quarters, that is, more false alarms; the own-country prior gives the lowest pre-crisis level (0.737), consistent with a prior that points to the channel of that country's previous crisis. The small size of these differences is itself a result: because the detection layer moves the weight to the active channel within a few quarters of its statistic rising, the starting point of the weights has little influence on the alarms, and the pooled prior is kept for its interpretability rather than for its performance.

Table 6: robustness priors

prior	AUROC	Pre-crisis	Calm	Utility	Detected
others_batch	0.886	0.764	0.290	0.200	4
own_batch	0.882	0.737	0.282	0.123	4
pooled	0.884	0.754	0.286	0.120	4
uniform	0.880	0.767	0.302	0.300	4

7.4 Meta-aggregator

Table 7 varies the prior weight u_0^0 that the meta-aggregator of Section 5.6 puts on the detection rule. The parameter governs a trade-off between the guarantee and the fidelity to the detection

rule. A lower u_0^0 gives a tighter bound against the individual models but mixes the detection rule with them, which lowers the pre-crisis level of the signal from 0.734 ($u_0^0 = 0.95$) to 0.541 ($u_0^0 = 0.5$) and the AUROC from 0.883 to 0.866. A higher u_0^0 keeps the signal close to the detection rule: the realized regret against it falls from 1.38 to 0.02, while the bound falls from 20.6 to 2.1. In every case the realized regret stays well below its bound and the four crises are detected. Our choice of $u_0^0 = 0.9$ keeps the pre-crisis level within 0.04 of the detection rule while guaranteeing a regret against it of at most 3.6.

Table 7: robustness meta u0

u0_cpd	AUROC	Pre-crisis	Calm	Detected	Regret vs rule	Bound
0.500	0.866	0.541	0.229	4	1.379	20.563
0.800	0.878	0.671	0.265	4	0.245	7.027
0.900	0.884	0.713	0.276	4	0.072	3.634
0.950	0.883	0.734	0.281	4	0.021	2.077

7.5 Horizon and real-time data

The last two checks change the definition of the exercise rather than a parameter of the aggregation: the length of the pre-crisis window, and the data available in real time. In both cases, the question is not only whether the crises are still detected, but whether they are detected *for the same reasons*: through the same channel, and within that channel, through the same indicators with the same signs. A procedure that detected the same crises through different variables would owe its results to chance; one that detects them through the same variables under a different target and a different information set is robust in the sense that matters. Appendix 9.11 contains all the relevant information for the two exercises; Figures 4 and 5 report the real-time re-estimations.

Eight-quarter horizon. With $h = 8$, the pre-crisis window covers the two years before the onset instead of three, and the predictive models are re-estimated on this shorter target. The four crises are again detected before their onset, with leads of 23 quarters in France, 18 in the United Kingdom, 17 in Sweden and 4 in Germany (Table 18 in Appendix 9.11.2) , and the alarms start at almost the same dates as in the baseline: 2002Q3 in France, 2003Q1 in the United Kingdom, 2004Q2 in Sweden and 2006Q3 in Germany. The channels are the same. The real-estate model carries the alarm in France, the United Kingdom and Sweden, with its statistic at the cap from 2003 to the onset in the first two countries and from 2004 in the third; in Germany, no statistic reaches the cap before 2007, and the global-cycle statistic rises in 2006 (Figures 15 and 17 in Appendix 9.11.2).

The performance statistics move in the direction that the shorter window implies. The quarters of 2002–2005 that were pre-crisis quarters under $h = 12$ become calm quarters under $h = 8$ while the signal is already high, so that the false positives rise from 19 to 28 in France and the AUROC falls from 0.96 to 0.85 in France and from 0.90 to 0.80 in the United Kingdom; in Sweden, where the signal rises later, the AUROC rises from 0.83 to 0.95. The level of the signal in the window is unchanged in France (0.95) and Sweden (0.89), lower in the United Kingdom (0.64, against 0.91), and the utility stays positive in the three countries (0.66, 0.86 and 0.65) and negative in Germany (-1.23). The meta-aggregator again reproduces the detection rule, with identical utilities and leads and a regret against it between -0.02 and 0.23 for a bound of 2.36 ($H = 14$), and the detection rule again loses about as much as the best single model, which is the real-estate model in every country (Table 19 in Appendix 9.11.2)

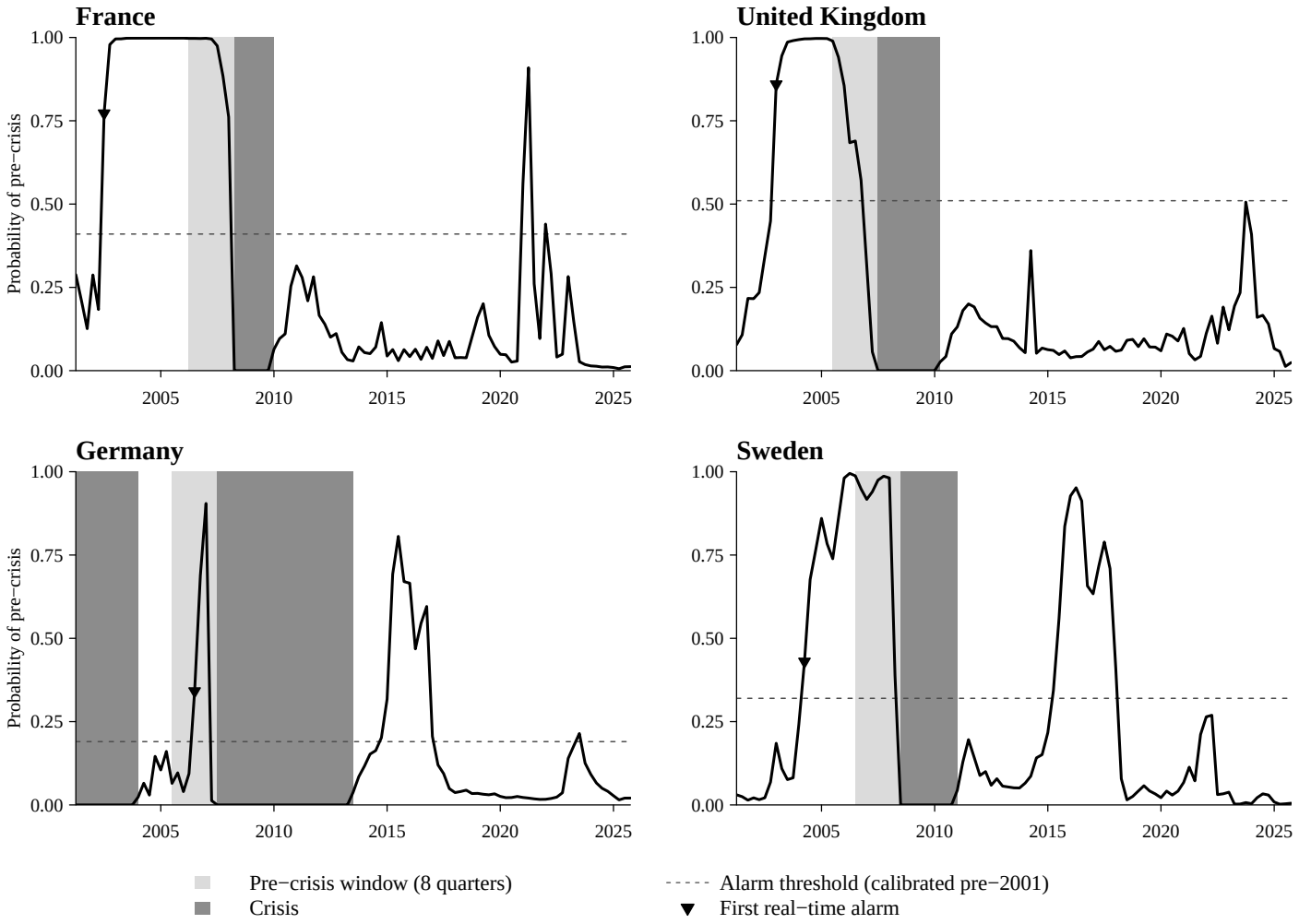


Figure 4: Probability of pre-crisis period - $h = 8$

The indicators are the same as well (Figures 18 to 21 in Appendix 9.11.2). In the real-estate model, the price-to-income ratio is selected with a positive sign at every re-estimation of the window in France, the United Kingdom and Sweden, as in the baseline, with the price-to-rent

ratio entering negatively conditional on it in France and the United Kingdom and positively in Sweden, as in the baseline. In the external model of the United Kingdom, the three-year growth of exports enters positively and the current account, the real exchange rate and cross-border flows negatively, as in the baseline. In the credit model of Sweden, the growth rates of credit to non-financial corporations and the credit-to-GDP gap enter positively. In Germany, the global-cycle model again selects the level of real equity prices and of the NASDAQ with a positive sign and the financial conditions index with a negative one, and the external model again selects exports and the growth of the real exchange rate positively and the current account negatively. The only sign that changes is that of the three-year growth of the NASDAQ in the German global-cycle model, negative in the baseline and positive with $h = 8$: a growth rate conditional on a level, whose sign we did not interpret in the baseline either. Changing the target from three years to two years therefore changes the dating of the windows and the performance statistics, but neither the channels nor the variables through which the crises are announced.

Real-time vintage data. The vintage specification replaces the indicators that are revised or constructed with full-sample information by their real-time counterparts: no ratio to GDP or to income, no trend-based gap, quantities and prices in growth rates, and series published with a lag shifted by one additional quarter (see Appendix 9.1.3 for the list of variables[47]). The exercise is more demanding, and the results are less sharp, but the conclusions hold. The four crises are detected (see Table 25 in Appendix 9.11.3), with leads of 23 quarters in France, 24 in the United Kingdom, 14 in Sweden and 3 in Germany. The AUROC is 0.82 in France, 0.94 in the United Kingdom, 0.82 in Sweden and 0.63 in Germany, and the utility is positive in the first three countries (0.84, 0.80 and 0.64) and negative in Germany (-2.39). The channel is again real estate in France, the United Kingdom and Sweden, and the batch model of the real-estate channel is built on the same configuration as in the baseline (see Figure 23 in Appendix 9.11.3): the price-to-rent ratio and the level of real house prices enter with a positive sign (Table 29 in Appendix 9.11.3), the price-to-income ratio being excluded from the vintage set as not available in real time. Two differences deserve mention. First, the credit model is almost absent from the vintage exercise: most of its real-time indicators, bank credit aggregates and money, start after 2001 and are not available for the batch estimation, so that the credit model receives a prior weight of 0.03 and the real-estate model of 0.82. This limits the comparison for the credit channel but does not affect the three real-estate episodes. Second, in France the real-estate statistic, at its cap from

[47] The number of variables in each predictive model is limited to four, including their transformations. To account for the exclusion of certain important variables, we therefore supplement the credit model with six additional variables.

2003, falls back to zero in the last quarters before the onset, so that the minimum of the signal in the window is zero; the alarm is nonetheless in progress at the onset because of the one-year persistence of alarms, and the lead is unchanged. Germany is, as in the baseline, the episode the procedure handles least well: the global-cycle channel, which carried the German signal of 2006 in the baseline, is weaker in real-time data, and the signal reaches only 0.13 on average over the window.

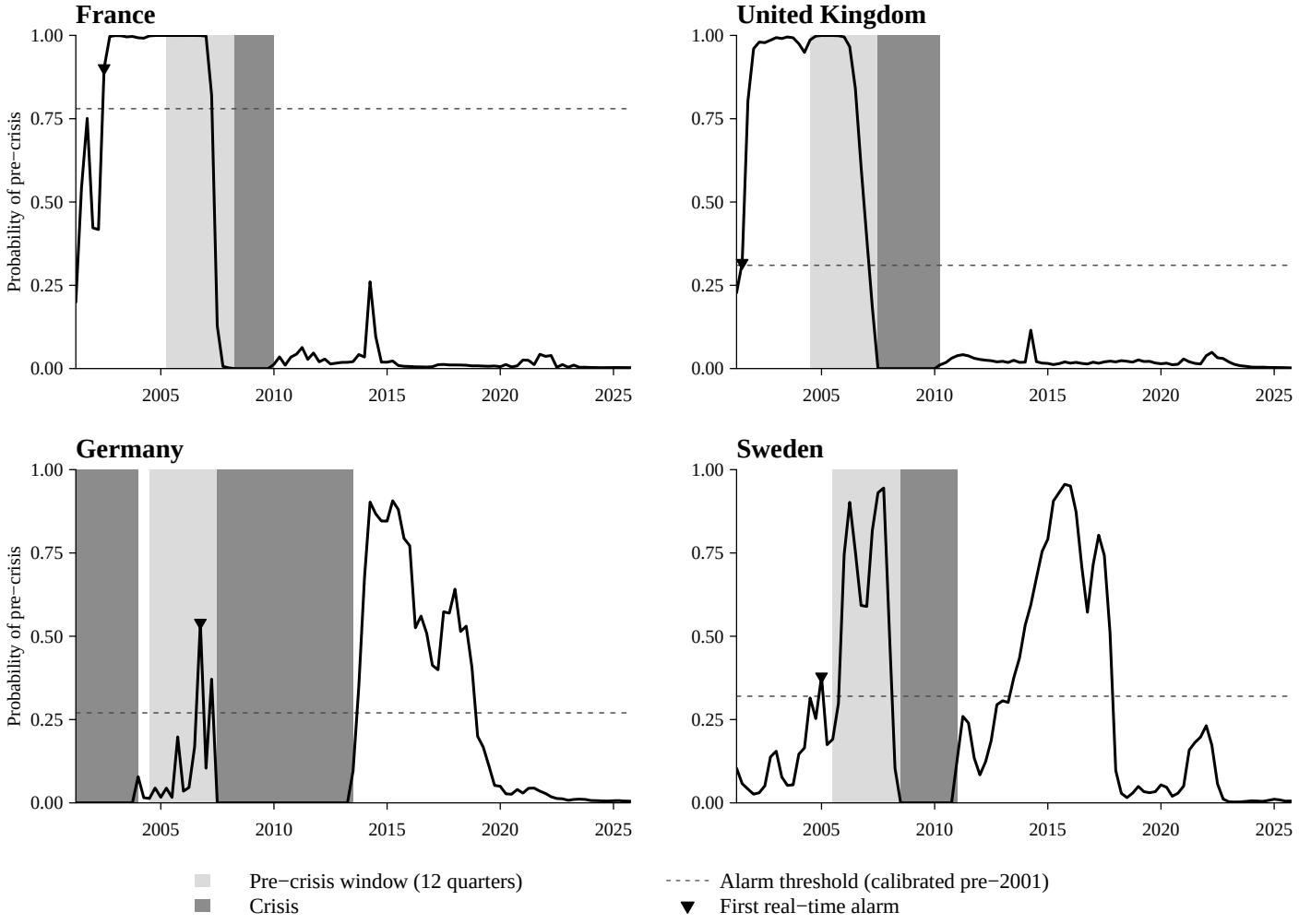


Figure 5: Probability of pre-crisis period - vintage specification

Taken together, the two exercises support the reading of Section 6.2. The crises of 2007–2008 are announced in France, the United Kingdom and Sweden by valuation ratios of the housing market, selected with the same sign whether the target is two or three years ahead and whether the indicators are taken from the latest release or from real-time data; the British episode has an external component with the same signs in both specifications; and the German episode, which did not originate in a domestic boom, is the one that every specification finds hardest to anticipate.

8 Conclusion

In response to the Queen, we develop a general real-time aggregation procedure that integrates key features of standard online learning theory and sequential detection methods. Applied in real time to France, Germany, Sweden and the United Kingdom from 2001 to 2025, with models estimated on the crises of the early 1990s only, the procedure raises an alarm before each of the four crises of 2007–2008, with leads of one to five years, and separates pre-crisis from calm quarters with an out-of-sample AUROC between 0.83 and 0.96. It does as well as the single best model in every country without knowing in advance which model that is, and its results survive changes in the tuning parameters, in the prior, in the horizon of the target and in the vintage of the data, not only in the sense that the crises remain detected, but in the sense that they remain detected through the same channels and the same indicators with the same signs. The gains come from detection: removing the learning layer leaves the results unchanged, whereas without the change-point component the signal no longer rises in time.

Three findings go beyond the performance of the procedure. First, the channel that announces a crisis differs across countries, even for crises with a common trigger: overvalued and still-rising housing prices in France, Sweden and the United Kingdom, with an external component in the United Kingdom, and the global equity cycle in Germany, whose crisis was imported through the exposure of its banks to foreign assets rather than built up at home. These channels are markers of vulnerability rather than causes: the crises of 2007–2008 originated largely in US mortgage-related losses and in the collapse of wholesale funding, and outside the United Kingdom the banks that failed did so because of their exposures abroad. But markers are what an early-warning system needs, and they cannot be chosen in advance: a system built on housing alone would have missed the German episode, and one built on the global cycle alone would have missed the three others. This is the case for weighting channels by the evidence they currently carry rather than by their past performance. Second, the indicators that carry the signal are not the ones the regulatory framework privileges. The credit-to-GDP gap, the reference indicator of the countercyclical buffer, is selected in fewer than half of the real-time re-estimations with an unstable sign, whereas the price-to-income ratio and the growth rates of credit are selected almost always with a stable sign.

The approach has limits that point to further work. The German episode shows that a procedure built on domestic vulnerabilities can only partly anticipate a crisis transmitted through the balance sheets of banks exposed abroad; a sensor built on cross-border bank exposures, which the BIS data now allow, would be a natural addition. The posterior over channels assumes that one

channel moves at a time; when several do, as in the United Kingdom in 2005, it shares the weight between them but no longer identifies a single source, and a posterior allowing for several affected sensors would be more informative. The regret guarantee is obtained for the meta-aggregator, not for the detection rule itself; the online learning framework suggests a route to a direct guarantee, in terms of regret weighted by the posterior, which we leave for future work. Finally, the procedure is evaluated on four countries and four crises, the number that the data allow when models must be estimated on earlier episodes; extending it to a wider panel, with the heterogeneity of crises that this entails, is the natural next test.

Another limitation is that these findings concern predictability rather than causality, a distinction that matters for how they should be interpreted and applied. The procedure identifies the indicators whose behaviour, in real time, distinguished the years before a crisis from calm years; it does not identify what caused the crisis. In 2007–2008, the proximate cause was largely common, the losses on US mortgage-related assets and the collapse of the wholesale funding markets that financed them, and in France, Sweden and Germany the banks that came under stress did so because of their exposures abroad rather than because of losses on domestic housing or credit. The domestic channels that carried the alarm are therefore best read as markers of the phase of the financial cycle in which banks expanded their balance sheets and became vulnerable to such a shock. This reading limits what the alarms can justify. They indicate when vulnerabilities are building up and which domestic indicators move with them, which is what a macroprudential authority needs to decide when to act and where to look; they do not show that acting on the marker, for instance tightening mortgage lending when house prices are overvalued, would have prevented a crisis that came from elsewhere. Establishing that requires identifying the effect of the instruments, which an early-warning exercise cannot do [48]. We leave this direction for future research.

A second direction concerns the panel dimension, which the procedure exploits only in part. The batch models are estimated on the pooled panel, but the real-time re-estimations use the data of each country alone, so that a country learns only from its own crises. The German episode shows the cost: with a single pre-crisis episode before 2007, coinciding with the technology boom, the German models learned a country-specific association between equity prices and crises that we cannot interpret. A multi-task estimation would let each country borrow strength from the others while keeping its own specificities, for instance by decomposing the coefficients of each

[48] The same logic applies to the alarms after 2010: because the authorities acted on the vulnerabilities the procedure signals, the absence of a subsequent crisis can neither confirm nor refute the alarms, and the prediction problem becomes entangled with the policy it informs.

model into a component common to all countries and a country-specific deviation, both penalized (Evgeniou and Pontil, 2004; Ollier and Viallon, 2017), or by selecting the same indicators jointly across countries while letting their coefficients differ (Obozinski et al., 2010). The common component would carry the regularities that the four episodes share, such as the positive and stable sign of the price-to-income ratio, and the deviation would be penalized toward zero unless a country's own data support it. The same idea extends to detection: since the global cycle is common to all countries, the evidence that a global-cycle model has entered a pre-crisis regime could be accumulated jointly across countries rather than separately, so that a signal building up in one country raises the prior on the same channel in the others. Both extensions would make the procedure more informative on a wider panel, where crises are more numerous but each country still has few of its own.

A limit of a different kind concerns the nature of the crises the procedure can anticipate. Its predictive models are built on the build-up of vulnerabilities in observed aggregates: credit, property prices, external balances and the global price of risk. It can therefore only detect crises that are preceded by such a build-up, crises that are endogenous in the sense that the financial system accumulates, over several years, the fragility that a shock later reveals. It cannot anticipate a crisis triggered by a purely exogenous shock that hits a system without visible imbalances, such as a pandemic, a war or a large operational failure; in such a case, there is no pre-crisis regime to detect, and the absence of an alarm is the correct output of the procedure rather than a failure. A subtler limit concerns endogenous vulnerabilities that build up outside the data the predictive models monitor. A boom in the valuation of a narrow set of firms, such as the current concerns about valuations linked to artificial intelligence, may be financed through private credit, non-bank intermediaries or off-balance-sheet commitments that do not appear in bank credit aggregates, and may be concentrated in a few markets that national indicators average out. The procedure would see such a boom only insofar as it moves the indicators it monitors, here mainly equity prices and the global price of risk, and would miss the channel through which it reaches the banking system. This is not specific to our approach: an early-warning system is only as good as the data on which its predictive models are built, and extending it to new forms of intermediation requires new sensors rather than a new method of aggregation, which the framework accommodates by construction.This limit, however, bears on the experts rather than on the aggregation, and we emphasize that our contribution lies in the aggregation procedure we propose. Any model that a policymaker trusts, whether a structural model, a machine-learning classifier, a market-based indicator or a supervisory assessment of a specific exposure, can be added as an additional predic-

tive model without changing anything else. If it is informative, its statistic rises when it signals and the posterior moves the weight to it, as it did for the real-estate model before 2008; if it is not, it receives little weight, and the meta-aggregator bounds the cost of having included it, since the regret against each candidate grows only with the logarithm of its inverse prior weight. Extending the system to new risks thus requires new sensors, not a new method of aggregation: the framework is designed to combine whatever predictive model are available, and to tell, in real time, which of them is currently carrying the signal.

Finally, for a macroprudential authority, the message is twofold. The information that distinguishes a build-up of vulnerabilities from calm is available in real time, in the persistence of the signals of models monitoring distinct channels, well before the labels that would confirm it; and the diagnosis of which channel is at work is as useful as the alarm itself, since the instruments that address a housing boom, a credit boom or an exposure to the global cycle are not the same.

To conclude, we could add to the letter of the British Academy addressed to the Queen that we should use coming from online learning and sequential change-point detection in our policy institutions. Neither has been applied to the problem of a macroprudential authority, although that problem is a detection problem by nature. They can give precious hints to guide humans in charge of financial stability regarding when and where they should up their game, summon their imaginative capacity and exercise their best judgment.

References

Acharya, Viral V. and Philipp Schnabl (2010) "Do Global Banks Spread Global Imbalances? Asset-Backed Commercial Paper during the Financial Crisis of 2007–09," *IMF Economic Review*, 58 (1), 37–73.

Addo, Peter Martey, Dominique Guégan, and Bertrand Hassani (2018) "Credit Risk Analysis using Machine and Deep learning models," *Documents de travail du Centre d'Economie de la Sorbonne* (18003).

Alessi, Lucia and Carsten Detken (2011) "Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity," *European Journal of Political Economy*, 27 (3), 520–533.

Althabiti, Mashail, Manal Abdullah, and Omaima Almatrafi (2024) "Multi-label Classification of IoT Data Stream: A Survey," *Computación y Sistemas*, 28 (3), 1523–1536.

Amat, Christophe, Tomasz Michalski, and Gilles Stoltz (2018) "Fundamentals and exchange rate forecastability with simple machine learning methods," *Journal of International Money and Finance*, 88, 1–24.

Bank of England (2007) "Inflation Report,"Technical report, Bank of England.

——— , "Official Bank Rate History," `https://www.bankofengland.co.uk/boeapps/database/Bank-Rate`, Accessed October 2026.

Bates, John M and Clive WJ Granger (1969) "The combination of forecasts," *Journal of the Operational Research Society*, 20 (4), 451–468.

Beutel, Johannes, Sophia List, and Gregor von Schweinitz (2019) "Does machine learning help us predict banking crises?" *Journal of Financial Stability*, 45, 100693, https://doi.org/10.1016/j.jfs.2019.100693.

Bi, Shuochen, Jue Xiao, and Tingting Deng (2024) "The role of AI in financial forecasting: ChatGPT's potential and challenges," in *Proceedings of the 4th Asia-Pacific Artificial Intelligence and Big Data Forum*, 1064–1070.

Blackwell, David (1956) "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, 6 (1), 1–8.

Bluwstein, K, M Buckmann, A Joseph, M Kang, S Kapadia, and O Simsek (2020) "Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach (Working Paper, No. 848)," *London: Bank of England. Retrieved August*, 14, 2020.

Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria (2001) "Is the crisis problem growing more severe?" *Economic policy*, 16 (32), 52–82.

Borio, Claudio EV and Mathias Drehmann (2009) "Assessing the risk of banking crises–revisited," *BIS Quarterly Review, March*.

Bundesanstalt für Finanzdienstleistungsaufsicht (2019) "General Administrative Act on the Rate for the Domestic Countercyclical Capital Buffer," Rate raised to 0.25 percent, binding on institutions from 1 July 2020.

Bussiere, Matthieu and Marcel Fratzscher (2006) "Towards a new early warning system of financial crises," *Journal of International Money and Finance*, 25 (6), 953–973.

Capistrán, Carlos and Allan Timmermann (2009) "Forecast combination with entry and exit of experts," *Journal of Business & Economic Statistics*, 27 (4), 428–440.

Cerutti, Eugenio, Stijn Claessens, and Luc Laeven (2017) "The Use and Effectiveness of Macroprudential Policies: New Evidence," *Journal of Financial Stability*, 28, 203–224.

Cesa-Bianchi, Nicolo and Gabor Lugosi (2006) *Prediction, Learning and Games*: Cambridge University Press.

Chen, Yudong, Tengyao Wang, and Richard J Samworth (2022) "High-dimensional, multiscale online changepoint detection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84 (1), 234–266.

Clévenot, Mickaël (2011) "Les prix immobiliers en France : une évolution singulière," in *Note de conjoncture*, 23–35: Insee.

Coimbra, Nuno and Hélène Rey (2017) "Financial Cycles with Heterogeneous Intermediaries," *NBER Working Papers* (23245), `https://ideas.repec.org/p/nbr/nberwo/23245.html`.

Coudert, Virginie and Julien Idier (2016) "An Early Warning System for Macro-prudential Policy in France," *Banque de France Working Paper 609*.

Cover, Thomas M (1966) *Behavior of sequential predictors of binary sequences* (7002): Stanford University, Stanford Electronics Laboratories, Systems Theory ....

Cutkosky, Ashok (2019) "Anytime online-to-batch, optimism and acceleration," in *International conference on machine learning*, 1446–1454, PMLR.

Dashevskiy, Mikhail and Zhiyuan Luo (2011) "Time series prediction with performance guarantee," *IET communications*, 5 (8), 1044–1051.

Daskalakis, Constantinos, Maxwell Fishelson, and Noah Golowich (2021) "Near-optimal no-regret learning in general games," *Advances in Neural Information Processing Systems*, 34, 27604–27616.

Dawid, A. Philip (1982) "The well-calibrated Bayesian," *Journal of the American Statistical Association*, 77 (379), 605–610.

De Hond, Anne AH, Ewout W Steyerberg, and Ben Van Calster (2022) "Interpreting area under the receiver operating characteristic curve," *The Lancet Digital Health*, 4 (12), e853–e855.

Demirgüç-Kunt, Asli and Enrica Detragiache (1998) "The determinants of banking crises in developing and developed countries," *Staff Papers*, 45 (1), 81–109.

Destrero, Augusto, Christine Mol, Francesca Odone, and Alessandro Verri (2009) "A Regularized Framework for Feature Selection in Face Detection and Authentication," *International Journal of Computer Vision*, 83, 164–177, 10.1007/s11263-008-0180-2.

Detken, Carsten, Olaf Weeken, Lucia Alessi et al. (2014) "Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options," *ESRB: Occasional Paper Series* (2014/5).

Devaine, Marie, Pierre Gaillard, Yannig Goude, and Gilles Stoltz (2013) "Forecasting electricity consumption by aggregating specialized experts," *Machine Learning*, 90, 231–260.

Diamond, Douglas W and Philip H Dybvig (1983) "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 91 (3), 401–19.

Diebold, Francis X and Peter Pauly (1987) "Structural change and the combination of forecasts," *Journal of Forecasting*, 6 (1), 21–40.

Diebold, Francis X and Glenn D Rudebusch (1989) "Long memory and persistence in aggregate output," *Journal of monetary economics*, 24 (2), 189–209.

Diebold, Francis X. and Minchul Shin (2019) "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives," *International Journal of Forecasting*, 35 (4), 1679–1691, https://doi.org/10.1016/j.ijforecast.2018.09.006.

Drehmann, Mathias and Mikael Juselius (2014) "Evaluating early warning indicators of banking crises: Satisfying policy requirements," *International Journal of Forecasting*, 30 (3), 759–780.

Du Plessis, Emile and Ulrich Fritsche (2025) "New forecasting methods for an old problem: Predicting 147 years of systemic financial crises," *Journal of Forecasting*, 44 (1), 3–40.

Duprey, Thibaut, Benjamin Klaus, and Tuomas Peltonen (2017) "Dating Systemic Financial Stress Episodes in the EU Countries," *Journal of Financial Stability*, 32, 10.1016/j.jfs.2017.07.004.

Eichengreen, Barry and Carlos Arteta (2002) "Banking crises in emerging markets: presumptions and evidence," *Financial policies in emerging markets*, 47–94.

Elliott, Graham and Allan Timmermann (2008) "Economic forecasting," *Journal of Economic Literature*, 46 (1), 3–56.

European Systemic Risk Board (2016) "Vulnerabilities in the EU Residential Real Estate Sector,"Technical report, European Systemic Risk Board, Warnings to Austria, Belgium, Denmark, Finland, Luxembourg, the Netherlands, Sweden and the United Kingdom, 28 November 2016.

——— (2019) "Vulnerabilities in the Residential Real Estate Sectors of the EEA Countries,"Technical report, European Systemic Risk Board, Warnings to the Czech Republic, Germany, France, Iceland and Norway, 23 September 2019.

Evgeniou, Theodoros and Massimiliano Pontil (2004) "Regularized Multi-Task Learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109–117.

Fellouris, Georgios and Grigory Sokolov (2016) "Second-order asymptotic optimality in multisensor sequential change detection," *IEEE Transactions on Information Theory*, 62 (1), 533–555.

Finansinspektionen (2015) "Decision on the Countercyclical Buffer Rate,"Technical report, Finansinspektionen.

——— (2018) "The Swedish Mortgage Market 2018,"Technical report, Finansinspektionen.

Fisher, Irving (1933) "The Debt-Deflation Theory of Great Depressions," *Econometrica*, 1 (4), 337–357, http://www.jstor.org/stable/1907327.

Flaspohler, Genevieve E, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey (2021) "Online learning with optimism and delay," in *International Conference on Machine Learning*, 3363–3373, PMLR.

Foster, Dylan J, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan (2018) "Logistic regression: The importance of being improper," in *Conference on learning theory*, 167–208, PMLR.

Frankel, Jeffrey A and Andrew K Rose (1996) "Currency crashes in emerging markets: An empirical treatment," *Journal of International Economics*, 41 (3-4), 351–366.

Frankel, Jeffrey and George Saravelos (2012) "Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis," *Journal of International Economics*, 87 (2), 216–231.

Freund, Yoav, Robert E Schapire, Yoram Singer, and Manfred K Warmuth (1997) "Using and combining predictors that specialize," in *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, 334–343.

Fudenberg, Drew and David K Levine (1995) "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, 19 (5-7), 1065–1089.

——— (1998) *The theory of learning in games*, 2: MIT press.

Gaillard, Pierre, Gilles Stoltz, and Tim Van Erven (2014) "A second-order bound with excess losses," in *Conference on Learning Theory*, 176–196, PMLR.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2012) "Neglected risks, financial innovation, and financial fragility," *Journal of Financial Economics*, 104 (3), 452 – 468.

Giacomini, Raffaella and Barbara Rossi (2009) "Detecting and predicting forecast breakdowns," *The Review of Economic Studies*, 76 (2), 669–705.

Glenn, W Brier et al. (1950) "Verification of forecasts expressed in terms of probability," *Monthly weather review*, 78 (1), 1–3.

Gneiting, Tilmann and Adrian E Raftery (2007) "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, 102 (477), 359–378.

Gourinchas, Pierre-Olivier and Maurice Obstfeld (2012) "Stories of the Twentieth Century for the Twenty-First," *American Economic Journal: Macroeconomics*, 4 (1), 226–65, 10.1257/mac.4.1.226.

Greenwood, Robin, Samuel G Hanson, Andrei Shleifer, and Jakob Ahm Sørensen (2022) "Predictable financial crises," *The Journal of Finance*, 77 (2), 863–921.

Gross, Kenny C and Keith E Humenik (1991) "Sequential probability ratio test for nuclear plant component surveillance," *Nuclear Technology*, 93 (2), 131–137.

Hand, David J (2009) "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine learning*, 77 (1), 103–123.

Hannan, J (1957) "Approximation to Bayes risk in repeated plays. Dresher M, Tucker A, Wolfe P, eds. Contributions to the Theory of Games, Vol. III."

Hart, Sergiu and Andreu Mas-Colell (2000) "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, 68 (5), 1127–1150.

——— (2001) "A general class of adaptive strategies," *Journal of Economic Theory*, 98 (1), 26–54.

Haut Conseil de stabilité financière (2021) "Décision D-HCSF-2021-7 relative aux conditions d'octroi de crédits immobiliers," Adopted 29 September 2021, applicable from 1 January 2022.

Jang, Kyoungseok, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona (2023) "Tighter PAC-Bayes bounds through coin-betting," in *The Thirty Sixth Annual Conference on Learning Theory*, 2240–2264, PMLR.

Jordà, Oscar, Moritz Schularick, and Alan M. Taylor (2015) "Leveraged bubbles," *Journal of Monetary Economics*, 76, S1–S20, https://doi.org/10.1016/j.jmoneco.2015.08.005.

Joulani, Pooria, Andras Gyorgy, and Csaba Szepesvári (2013) "Online learning under delayed feedback," in *International Conference on Machine Learning*, 1453–1461.

Jun, Kwang-Sung, Francesco Orabona, Stephen Wright, and Rebecca Willett (2017) "Online learning for changing environments using coin betting."

Kaminski, Graciela and Carmen Reinhart (1999) "The Twin Crises: The Causes of Banking and Balance of Payments Problems," *American Economic Review*, 89 (3), 473–500.

Kindleberger, Charles (1978) "Manias, Panic and Crashes," *New York Basic Books*.

Krishnamurthy, Arvind and Tyler Muir (2025) "How credit cycles across a financial crisis," *The Journal of Finance*, 80 (3), 1339–1378.

Laeven, Luc and Fabian Valencia (2020) "Systemic Banking Crises Database: A Timely Update in COVID-19 Times," *CEPR Discussion Paper No. 14569*.

Lai, Tze Leung (1998) "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information theory*, 44 (7), 2917–2929.

Lee, Chung-Wei, Christian Kroer, and Haipeng Luo (2021) "Last-iterate convergence in extensive-form games," *Advances in Neural Information Processing Systems*, 34, 14293–14305.

Lee, Wei-Cheng and Francesco Orabona (2025) "A Best-of-Both-Worlds Proof for Tsallis-INF without Fenchel Conjugates," *arXiv preprint arXiv:2511.11211*.

Liang, Tengyuan and James Stokes (2019) "Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 907–915, PMLR.

Littlestone, Nick and Manfred K Warmuth (1994) "The weighted majority algorithm," *Information and computation*, 108 (2), 212–261.

Liu, Mingrui and Francesco Orabona (2022) "On the initialization for convex-concave min-max problems," in *International Conference on Algorithmic Learning Theory*, 743–767, PMLR.

Lo Duca, Marco, Anne Koba, Marisa Basten et al. (2017) "A New Database for Financial Crises in European Countries," *ECB Occasional Paper* (194).

Lorden, Gary (1971) "Procedures for reacting to a change in distribution," *The annals of mathematical statistics*, 1897–1908.

Lowe, Philip and Claudio Borio (2002) "Asset prices, financial and monetary stability: exploring the nexus," BIS Working Papers 114, Bank for International Settlements, `https://ideas.repec.org/p/bis/biswps/114.html`.

Mahtout, Btissame El and Florian Ziel (2026) "Electricity Price Forecasting: Bridging Linear Models, Neural Networks and Online Learning," *arXiv preprint arXiv:2601.02856*.

Malenica, Ivana, Rachael V Phillips, Antoine Chambaz, Alan E Hubbard, Romain Pirracchio, and Mark J van der Laan (2023) "Personalized online ensemble machine learning with applications for dynamic data streams," *Statistics in Medicine*, 42 (7), 1013–1044.

Mallet, Vivien, Gilles Stoltz, and Boris Mauricette (2009) "Ozone ensemble forecast with machine learning algorithms," *Journal of Geophysical Research: Atmospheres*, 114 (D5).

Martinez-Miera, David and Rafael Repullo (2017) "Search for yield," *Econometrica*, 85 (2), 351–378.

Meese, Richard A and Kenneth Rogoff (1983) "Empirical exchange rate models of the seventies: Do they fit out of sample?" *Journal of International economics*, 14 (1-2), 3–24.

Mei, Yajun (2010) "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, 97 (2), 419–433.

——— (2011) "Quickest detection in censoring sensor networks," in *2011 IEEE International Symposium on Information Theory Proceedings*, 2148–2152, IEEE.

Meinshausen, Nicolai and Peter Bühlmann (2010) "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72 (4), 417–473.

Mian, Atif and Amir Sufi (2009) "The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis," *The Quarterly Journal of Economics*, 124 (4), 1449–1496, 10.1162/qjec.2009.124.4.1449.

Minsky, Hyman (1986) "Global Consequences of Financial Deregulation," *Washington University Working Paper* (96).

Mol, Christine, Ernesto De Vito, and Lorenzo Rosasco (2009a) "Elastic-Net Regularization in Learning Theory," *Journal of Complexity*, 25, 201–230, 10.1016/j.jco.2009.01.002.

Mol, Christine, Sofia Mosci, Magali Traskine, and Alessandro Verri (2009b) "A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data," *Journal of computational biology*, 16, 677–90, 10.1089/cmb.2008.0171.

Moustakides, George V (1986) "Optimal stopping times for detecting changes in distributions," *the Annals of Statistics*, 14 (4), 1379–1387.

Obozinski, Guillaume, Ben Taskar, and Michael I. Jordan (2010) "Joint Covariate Selection and Joint Subspace Selection for Multiple Classification Problems," *Statistics and Computing*, 20 (2), 231–252.

Ollier, Edouard and Vivian Viallon (2017) "Regression Modelling on Stratified Data with the Lasso," *Biometrika*, 104 (1), 83–96.

Page, Ewan S (1954) "Continuous inspection schemes," *Biometrika*, 41 (1/2), 100–115.

Pesaran, Hashem and Allan Timmermann (2005) "Real-Time Econometrics," *Econometric Theory*, 21 (1), 212–231, 10.1017/S0266466605050139.

Phillips, Peter CB, Shuping Shi, and Jun Yu (2015) "Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500," *International economic review*, 56 (4), 1043–1078.

Pollak, Moshe (1985) "Optimal detection of a change in distribution," *The Annals of Statistics*, 206–227.

Puri, Manju, Jörg Rocholl, and Sascha Steffen (2011) "Global Retail Lending in the Aftermath of the US Financial Crisis: Distinguishing between Supply and Demand Effects," *Journal of Financial Economics*, 100 (3), 556–578.

Reinhart, Carmen M. and Kenneth S. Rogoff (2009) *This Time Is Different: Eight Centuries of Financial Folly*, Princeton, New Jersey: Princeton University Press, `http://www.reinhartandrogoff.com/`.

Romano, Gaetano, Idris A Eckley, Paul Fearnhead, and Guillem Rigaill (2023) "Fast online changepoint detection via functional pruning CUSUM statistics," *Journal of Machine Learning Research*, 24 (81), 1–36.

Rossi, Barbara (2011) "Advances in Forecasting Under Instability," *Handbook of Economic Forecasting*, 2, 10.1016/B978-0-444-62731-5.00021-X.

Rougerie, Catherine and Jacques Friggit (2010) "Prix des logements anciens : dans les années 2000, ils ont augmenté beaucoup plus vite que les loyers et les revenus," *Insee Première* (1297).

Schularick, Moritz and Alan M. Taylor (2012) "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008," *American Economic Review*, 102 (2), 1029–61, 10.1257/aer.102.2.1029.

Shafer, Glenn and Vladimir Vovk (2005) *Probability and finance: it's only a game!*, 491: John Wiley & Sons.

Shin, Hyun Song (2009) "Reflections on Northern Rock: The Bank Run That Heralded the Global Financial Crisis," *Journal of Economic Perspectives*, 23 (1), 101–119.

——— (2013) "Procyclicality and the search for early warning indicators," *International Monetary Fund Working Paper 258*.

Soldi, Giovanni, Nicola Forti, Domenico Gaglione, Paolo Braca, Leonardo M Millefiori, Stefano Marano, Peter K Willett, and Krishna R Pattipati (2021) "Quickest detection and forecast of pandemic outbreaks: Analysis of COVID-19 waves," *IEEE Communications Magazine*, 59 (9), 16–22.

Stock, James H and Mark W Watson (1996) "Evidence on structural instability in macroeconomic time series relations," *Journal of Business & Economic Statistics*, 14 (1), 11–30.

Stoltz, Gilles and Gábor Lugosi (2007) "Learning correlated equilibria in games with compact sets of strategies," *Games and Economic Behavior*, 59 (1), 187–208.

Sufi, Amir and Alan M Taylor (2021) "Financial crises: A survey," *National Bureau of Economic Research Working Paper 29155*, 10.3386/w29155.

Sveriges Riksbank (2007) "Financial Stability Report 2007:2,"Technical report, Sveriges Riksbank.

——— (2009) "Financial Stability Report 2009:2,"Technical report, Sveriges Riksbank.

Tartakovsky, Alexander G and Venugopal V Veeravalli (2008) "Asymptotically optimal quickest change detection in distributed sensor systems," *Sequential Analysis*, 27 (4), 441–475.

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

Timmermann, Allan (2006) "Chapter 4: Forecast Combinations," 1 of Handbook of Economic Forecasting, 135–196: Elsevier, https://doi.org/10.1016/S1574-0706(05)01004-9.

Vovk, Volodimir G (1990) "Aggregating strategies," in *Proceedings of the third annual workshop on Computational learning theory*, 371–386.

Ward, Felix (2017) "Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors," *Journal of Applied Econometrics*, 32, 359–378.

Weinberger, Marcelo J and Erik Ordentlich (2002) "On delayed prediction of individual sequences," *IEEE Transactions on Information Theory*, 48 (7), 1959–1976.

Xie, Yao and David Siegmund (2013) "Sequential multi-sensor change-point detection," in *2013 Information theory and applications workshop (ITA)*, 1–20, IEEE.

Yi, Xinyang, Shao-Chuan Wang, Ruining He et al. (2023) "Online matching: A real-time bandit system for large-scale recommendations," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 403–414.

Zamba, KD and Panagiotis Tsiamyrtzis (2021) "Sequential detection framework for real-time biosurveillance based on Shiryaev-Roberts procedure with illustrations using COVID-19 incidence data," *Sequential Analysis*, 40 (2), 149–169.

Zou, Hui and T. Hastie (2005) "Regularization and Variable Selection via the Elastic Nets," *Journal of the Royal Statistical Society: Series B*, 67, 301–320.

9 Appendix

9.1 Data and predictive models

9.1.1 Data on financial crisis

The methodology used to date systemic financial crises episode is a two-step approach. Following Duprey et al. (2017), it aims at first identifying historical episodes of elevated financial stress which were also associated with real economic slowdowns using a quantitative analysis. The financial stress indicator captures three financial market segments: i) equity market: stock price index, ii) bond market: 10-year government yields and iii) foreign exchange market: real effective exchange rate. Industrial production growth is used as measure of real economic activity. At the end of this first step, a list of potential systemic crisis events, characterized by six consecutive months of real economic slowdown occurring within one year of the financial stress period is drawn.

The second step uses a qualitative approach. Each national authority distinguishes between systemic crises and residual episodes of financial stress following common criteria. An event is classified as a systemic crisis event if it fulfills one or more of the following three criteria: i) A contraction in the supply of financial intermediation or funding to the economy took place during the financial stress event, ii) The financial system was distressed (market infrastructures were dysfunctional and/or there were bankruptcies among large financial institutions) and iii) Policies were adopted to preserve financial stability (external support, extraordinary provision of central bank liquidity, direct interventions of the state). Residual events are episodes of financial stress which are not as wide and serious as systemic crises. National authorities are asked whether they want to complement the list of events or change the timing of events already flagged. Their judgments prevail. This official database of crisis episodes is available for European countries. We replicated the exact same methodology for the United States and find that, under this methodology, there is no systemic crisis before the 2008 financial crisis [49].

9.1.2 Data for predictive models - baseline and $h = 8$ specification

The following table describes the variable used for each predictive model in the baseline and horizon $h = 8$ specification.

[49] We are very grateful to the New York Fed and to Anna Kovner in particular for the US data.

Table 8: Indicators by predictive model, baseline specification

Model	Code	Description	Source
Credit	BASEL_GAP	Credit-to-GDP gap (Basel III, one-sided HP filter)	BIS
	TC_H	Total credit to households, % of GDP	BIS
	TC_NFC	Total credit to non-financial corporations, % of GDP	BIS
	DEBT_SECU_FC	Debt securities issued by financial corporations	BIS
Real estate	RHP	Real house prices	OECD
	PTI	House price-to-income ratio	OECD
	PTR	House price-to-rent ratio	OECD
	SPP_R	Residential property prices, real	BIS
External	CA	Current account balance, % of GDP	OECD
	XRU	Exchange rate against the US dollar	BIS
	cross_border_flows_adj_xfma	Cross-border capital flows	CrossBorder Capital
	X	Exports	OECD
Global cycle	risk_appetite	Risk appetite index	CrossBorder Capital
	SHP	Real share prices	OECD
	financial_conditions_index_fci	Financial conditions index	CrossBorder Capital
	NASDAQCOM	NASDAQ Composite index	FRED

Each indicator enters its model in level and in one-, two- and three-year growth rates. Indicators are standardized by country over the batch period.

9.1.3 Data for predictive models - vintage specification

The following table describes the vintage specification :

Table 9: Indicators by predictive model, real-time (vintage) specification

Model	Code	Description	Source
Credit	TCLA_B	Cross-border claims on banks	BIS LBS
	TLIA_B	Cross-border liabilities to banks	BIS LBS
	TCLA_NB	Cross-border claims on non-banks	BIS LBS
	DEBT_SECU_FC	Debt securities issued by financial corporations	BIS
	M3	Broad money	OECD
	excess_credit_growth_including_shadow_banks	Excess credit growth, including shadow banks	CrossBorder Capital
Real estate	RHP	Real house prices	OECD
	NHP	Nominal house prices	OECD
	PTR	House price-to-rent ratio	OECD
	SPP_R	Residential property prices, real	BIS
External	XRU	Exchange rate against the US dollar	BIS
	foreign_exchange_reserves	Foreign exchange reserves	CrossBorder Capital
	cross_border_flows_xfm	Cross-border capital flows	CrossBorder Capital
	forex_risk_index_fri	Foreign exchange risk index	CrossBorder Capital
Global cycle	risk_appetite	Risk appetite index	CrossBorder Capital
	SHP	Real share prices	OECD
	equity_exposure_index_eei	Equity exposure index	CrossBorder Capital
	NASDAQCOM	NASDAQ Composite index	FRED
	VIX	CBOE Volatility Index	FRED

Each indicator enters its model in level and in one-, two- and three-year growth rates. Indicators are standardized by country over the batch period. The vintage specification excludes ratios to GDP or income and trend-based gaps; the BIS series and M3 are lagged by one additional quarter to account for their publication delay.

9.2 Standardization of variable

The variables used in each predictive models are standardized. Let T_B denote the end of the batch period (2001Q1). For each country i and indicator v, we standardize the indicator with its moments over the batch period:

$$z_{i,t,v} = \frac{x_{i,t,v} - \mu_{i,v}}{\sigma_{i,v}}, \qquad \mu_{i,v} = \frac{1}{|\mathcal{B}i|} \sum s \in \mathcal{B}ixi, s, v, \qquad \sigma^2_{i,v} = \frac{1}{|\mathcal{B}i| - 1} \sum s \in \mathcal{B}i(xi, s, v - \mu_{i,v})^2, \tag{25}$$

where $\mathcal{B}_i = \{s \leq T_B - 1\}$ collects the observations released by T_B. The moments are computed once, at T_B, so $z_{i,t-1,v}$ is $\mathcal{F}_{i,t}$-measurable for $t > T_B$. Standardization serves two purposes. First, it makes indicators comparable across countries, which is necessary for the batch model estimated on the pooled panel: the same indicator can differ across countries by an order of magnitude in level and volatility, and without standardization a pooled model would mostly capture these cross-country differences rather than each country's deviation from its own history. Second, it expresses each coefficient as the effect on the log-odds of a one standard deviation deviation of the indicator from its historical mean in the country, which makes coefficients comparable across indicators and across countries. For the country-specific online models, standardization is an affine transformation of the regressors and leaves the estimated probabilities unchanged, since the elastic net penalty is applied to internally standardized regressors.

9.3 Proofs

Conventions. Throughout this appendix we fix a country and drop the index i. The information available at the end of quarter t is the σ-field $\mathcal{F}_t = \sigma(x_{1:t-1}, C_{1:t})$ generated by the indicators released up to t and by the crisis indicator up to t. We use only two properties of this object. First, a random variable Z is $\mathcal{F}_t$-measurable if and only if it is a (measurable) function of the observed data $(x_{1:t-1}, C_{1:t})$; in words, Z is *determined* by what has been observed up to t. Second, information accumulates: $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leq t$, so that a variable determined at s is also determined at t. To show that a variable is *not* $\mathcal{F}_t$-measurable, we exhibit two possible histories that coincide up to t, so that they are indistinguishable at t, but give the variable different values. This requires that the observed history never rules out the future: we assume, as is implicit in the model, that for any history up to t with $C_t = 0$, both an onset at $t+1$ and the absence of an onset at $t+1$ are possible continuations.

Recall that an *onset* is a date τ with $C_{\tau-1} = 0$ and $C_\tau = 1$, that $\mathcal{O}$ denotes the set of onsets, and that by Assumption A1 the indicator C_t is observed at t. Hence whether an onset occurs at a date $\tau \leq t$ is determined at t: it is the event $\{C_{\tau-1} = 0,\ C_\tau = 1\}$, which depends only on $C_{1:t}$. We use this fact repeatedly.

9.3.1 Proof of Lemma 1

Lemma 1 (Measurability). *Under Assumption A1, y_t is $\mathcal{F}_{r_t}$-measurable. In the first and third cases of equation 4, it is not $\mathcal{F}_{r-1}$-measurable for any $r \leq r_t$.*

Proof. Recall from Definition 1 that $y_t = 1$ if and only if $C_t = 0$ and an onset occurs in the window $(t, t+h]$, and that $y_t = 0$ otherwise, in particular whenever $C_t = 1$. The resolution date of equation 4 distinguishes three cases, which we treat in turn. For each case we show two things: that y_t is determined by the data observed up to r_t (measurability), and, in the first and third cases, that it is not determined by the data observed up to $r_t - 1$ (minimality). Since $\mathcal{F}_{r-1} \subseteq \mathcal{F}_{r_t-1}$ for $r \leq r_t$, it is enough to prove the second point for $r = r_t$: a variable that is not determined at $r_t - 1$ is not determined at any earlier date either.

Case 2: $C_t = 1$, so that $r_t = t$. By definition, $y_t = 0$ as soon as $C_t = 1$, whatever happens afterwards. Since C_t is observed at t, the value $y_t = 0$ is determined at $t = r_t$. No minimality claim is made in this case.

Case 1: $y_t = 1$, so that $r_t = \tau^* := \min\{\tau \in \mathcal{O} : \tau > t\}$. Since $y_t = 1$, there is at least one onset in $(t, t+h]$, so τ^* is well defined and $t < \tau^* \leq t + h$.

- *Measurability.* At date τ^*, the observer has seen $C_{1:\tau^*}$. In particular, she has seen $C_t = 0$, $C_{\tau^*-1} = 0$ and $C_{\tau^*} = 1$. The last two equalities say that an onset occurs at τ^*, and since $t < \tau^* \leq t + h$, this onset lies in $(t, t+h]$. Together with $C_t = 0$, this is exactly the definition of $y_t = 1$. Hence y_t is determined by $C_{1:\tau^*}$, that is, y_t is $\mathcal{F}_{\tau^*}$-measurable.
- *Minimality.* At date $\tau^* - 1$, the observer has seen $C_{1:\tau^*-1}$, and by definition of τ^* no onset has occurred in $(t, \tau^* - 1]$. Moreover $C_s = 0$ for all $s \in [t, \tau^* - 1]$: a crisis quarter in this interval would either be the continuation of a crisis in progress at t, which is excluded by $C_t = 0$, or follow an onset in $(t, \tau^* - 1]$, which is excluded by the choice of τ^*. Consider now two continuations of this history. In the first, an onset occurs at τ^* (this is the actual history), and $y_t = 1$. In the second, no onset occurs in $[\tau^*, t + h]$; this continuation is possible because $C_{\tau^*-1} = 0$, and it gives $y_t = 0$, since then no onset occurs anywhere in $(t, t+h]$. The two histories coincide up to $\tau^* - 1$ and give different values to y_t, so y_t is not $\mathcal{F}_{\tau^*-1}$-measurable.

Case 3: $y_t = 0$ and $C_t = 0$, so that $r_t = t + h$. In this case no onset occurs in $(t, t+h]$.

- *Measurability.* At date $t + h$, the observer has seen $C_{1:t+h}$, hence knows, for every $\tau \in (t, t+h]$, whether an onset occurs at τ. She therefore knows that no onset occurs in $(t, t+h]$, and, together with $C_t = 0$, that $y_t = 0$. Hence y_t is $\mathcal{F}_{t+h}$-measurable.

- *Minimality.* At date $t+h-1$, the observer has seen $C_{1:t+h-1}$ and knows that no onset occurs in $(t, t+h-1]$. As in Case 1, $C_s = 0$ for all $s \in [t, t+h-1]$. Two continuations are possible: no onset at $t+h$, which is the actual history and gives $y_t = 0$; or an onset at $t+h$, which is possible since $C_{t+h-1} = 0$, and which gives $y_t = 1$ because $t+h \in (t, t+h]$. The two histories coincide up to $t+h-1$ and give different values to y_t, so y_t is not $\mathcal{F}_{t+h-1}$-measurable.

This proves both claims in each case. □

Remark 1 (The decision rule behind the lemma). The proof can be summarized by the rule that an observer applies at each date $s \geq t$ to decide the label of quarter t: if $C_t = 1$, declare $y_t = 0$; otherwise, if an onset has been observed at some date in $(t, \min\{s, t+h\}]$, declare $y_t = 1$; otherwise, if $s \geq t+h$, declare $y_t = 0$; otherwise, wait. The rule uses only $C_{1:s}$, and the first date at which it stops waiting is r_t. In the language of stochastic processes, r_t is a stopping time with respect to $(\mathcal{F}_s)$, since the event $\{r_t \leq s\}$ is determined by $C_{1:s}$, and the lemma states that y_t is measurable with respect to the σ-field of this stopping time.

9.3.2 Proof of Lemma 2

Lemma 2. *Under Assumption A1, if the predictive models' forecasts are $\mathcal{F}_t$-measurable and w_t is a real-time rule, then $\hat{p}_t$ is $\mathcal{F}_t$-measurable.*

Proof. By equation 6, $\hat{p}_t = g_t \sum_{k=1}^{K} w_{k,t}\, p_{k,t}$. By Definition 4, a real-time rule computes the weights w_t from three inputs: the predictive models' forecasts $p_{k,1:t}$, the labels revealed by t, and the crisis indicators $C_{1:t}$. The forecast $\hat{p}_t$ is therefore a function of four ingredients: these three inputs and the gate g_t. We show that each ingredient is $\mathcal{F}_t$-measurable. Since sums, products and compositions of measurable functions are measurable, this proves that $\hat{p}_t$ is $\mathcal{F}_t$-measurable. We use throughout that $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leq t$.

Step 1: forecasts. For every k and every $s \leq t$, $p_{k,s}$ is $\mathcal{F}_s$-measurable by assumption, hence $\mathcal{F}_t$-measurable.

Step 2: crisis indicators. $C_{1:t}$ is $\mathcal{F}_t$-measurable by Assumption A1. As noted in the conventions, it follows that for every $\tau \leq t$, the event "an onset occurs at τ" is determined at t.

Step 3: revealed labels. The rule uses, for each $s \leq t$, whether the label of s has been revealed by t and, if so, its value. Formally, it uses the pair

$$\big(\mathbf{1}\{\tilde{r}_s \leq t\},\ \tilde{y}_s\, \mathbf{1}\{\tilde{r}_s \leq t\}\big),$$

whose second component is the revealed label when there is one and zero otherwise. We show that both components are functions of $C_{1:t}$. Fix $s \leq t$ and distinguish two cases according to Definition 3.

If $C_s = 1$, then $\tilde{y}_s = y_s = 0$ and $\tilde{r}_s = s \leq t$: the label is revealed, with value zero, and both facts are determined by C_s, hence at t.

If $C_s = 0$, Definition 3 gives $\tilde{y}_s = 1$ if an onset occurs in $(s, s+H]$ and $\tilde{y}_s = 0$ otherwise, with revelation date $\tilde{r}_s = \min\{\tau \in \mathcal{O} : \tau > s\}$ in the first case and $\tilde{r}_s = s+H$ in the second. (In the first case the minimum is at most $s+H$, so $\tilde{r}_s \leq s+H$ in both cases.) Hence the label of s is revealed by t, that is $\tilde{r}_s \leq t$, if and only if one of two events occurs:

(a) an onset occurs at some date in $(s, \min\{t, s+H\}]$; or

(b) $s+H \leq t$.

Indeed, if $\tilde{y}_s = 1$, the first onset after s lies in $(s, s+H]$, and it is at most t exactly when (a) holds; if $\tilde{y}_s = 0$, then $\tilde{r}_s = s+H \leq t$ exactly when (b) holds, and (a) cannot hold since there is no onset in $(s, s+H]$. Event (a) concerns onsets at dates at most t, hence is determined at t by Step 2; event (b) is a deterministic comparison of dates. Therefore $\mathbf{1}\{\tilde{r}_s \leq t\}$ is $\mathcal{F}_t$-measurable. On the event $\{\tilde{r}_s \leq t\}$, the value of the label is also determined at t: if (a) holds, $\tilde{y}_s = 1$; if (b) holds and (a) does not, the whole window $(s, s+H]$ lies in the past, no onset has been observed in it, and $\tilde{y}_s = 0$. Hence $\tilde{y}_s\, \mathbf{1}\{\tilde{r}_s \leq t\}$ is $\mathcal{F}_t$-measurable.

Step 4: gate. By Definition 8, g_t is a function of $C_{1:t}$ only, hence $\mathcal{F}_t$-measurable by Step 2; this is the first part of Lemma 4, proved below.

Conclusion. By Steps 1 to 3 and Definition 4, the weight vector w_t is a function of $\mathcal{F}_t$-measurable quantities, hence is $\mathcal{F}_t$-measurable. The forecast $\hat{p}_t = g_t \sum_k w_{k,t}\, p_{k,t}$ is a sum of products of $\mathcal{F}_t$-measurable quantities, hence is $\mathcal{F}_t$-measurable. □

9.3.3 Proof of Lemma 3

Lemma 3 (Calibrated forecasts are log-likelihood ratios)**.** *Suppose predictive model k is calibrated, in the sense that $\Pr(y_t = 1 \mid p_{k,t}) = p_{k,t}$: among the quarters in which the predictive model forecasts p, a fraction p are pre-crisis quarters. Let $\pi = \Pr(y_t = 1)$ be the pre-crisis frequency on its estimation sample, and $f_{j,k}$ the density of $p_{k,t}$ conditional on $y_t = j$. Then*

$$\operatorname{logit}(p_{k,t}) = \log \frac{f_{1,k}(p_{k,t})}{f_{0,k}(p_{k,t})} + \log \frac{\pi}{1-\pi}.$$

Proof. Fix a predictive model k and a quarter t, and drop both indices: p denotes the forecast $p_{k,t}$, a random variable with values in $(0,1)$; y the label y_t, a random variable with values in $\{0,1\}$; and f_j the density $f_{j,k}$ of p conditional on $y = j$. We assume $0 < \pi < 1$ and $f_0(x) > 0$, $f_1(x) > 0$ at every point x considered, so that all ratios below are well defined. The proof is Bayes' rule, written for a pair made of a discrete variable, the label, and a continuous one, the forecast.

Step 1: the joint distribution of (y, p). The label is a Bernoulli variable with $\Pr(y = 1) = \pi$ and $\Pr(y = 0) = 1 - \pi$. By definition of a conditional density, for any interval $A \subseteq (0,1)$,

$$\Pr(p \in A \mid y = j) = \int_A f_j(x)\, dx, \qquad j \in \{0,1\}.$$

Multiplying by $\Pr(y = j)$ gives the joint distribution of the pair:

$$\Pr(y = 1,\ p \in A) = \pi \int_A f_1(x)\, dx, \qquad \Pr(y = 0,\ p \in A) = (1-\pi) \int_A f_0(x)\, dx.$$

In other words, the joint distribution has density $\pi f_1(x)$ on the line $y = 1$ and density $(1-\pi) f_0(x)$ on the line $y = 0$.

Step 2: the marginal density of p. The events $\{y = 0\}$ and $\{y = 1\}$ partition the sample space, so by the law of total probability

$$\Pr(p \in A) = \Pr(y = 1,\ p \in A) + \Pr(y = 0,\ p \in A) = \int_A \big[\pi f_1(x) + (1-\pi) f_0(x)\big]\, dx.$$

Hence p has a density, namely

$$f(x) = \pi\, f_1(x) + (1-\pi)\, f_0(x),$$

which is positive wherever f_0 or f_1 is.

Step 3: Bayes' rule. The conditional probability of the label given the forecast is the ratio of the joint density to the marginal density:

$$\Pr(y = 1 \mid p = x) = \frac{\pi\, f_1(x)}{f(x)}, \qquad \Pr(y = 0 \mid p = x) = \frac{(1-\pi)\, f_0(x)}{f(x)}.$$

This is the continuous version of the elementary formula $\Pr(B \mid A) = \Pr(A \cap B)/\Pr(A)$: conditioning on $p = x$ is conditioning on p falling in a small interval around x, and the ratio of the two probabilities of Step 1 to the probability of Step 2 converges, as the interval shrinks, to the ratio of the densities. By Step 2, the two conditional probabilities sum to one, as they should.

Step 4: the posterior odds. Dividing the first expression of Step 3 by the second, the marginal

density $f(x)$ cancels:

$$\frac{\Pr(y=1 \mid p=x)}{\Pr(y=0 \mid p=x)} = \frac{\pi\, f_1(x)}{(1-\pi)\, f_0(x)} = \frac{f_1(x)}{f_0(x)} \cdot \frac{\pi}{1-\pi}. \tag{26}$$

In words: the posterior odds of a pre-crisis quarter, given the forecast, equal the likelihood ratio of the forecast times the prior odds. This identity holds for any forecast, calibrated or not.

Step 5: calibration. The calibration assumption says that $\Pr(y = 1 \mid p = x) = x$, hence $\Pr(y = 0 \mid p = x) = 1 - x$. Substituting into the left-hand side of equation 26 gives

$$\frac{x}{1-x} = \frac{f_1(x)}{f_0(x)} \cdot \frac{\pi}{1-\pi}.$$

This is the content of calibration: for a calibrated predictive model, the forecast itself *is* the posterior probability, so its odds are the posterior odds.

Step 6: logarithms. Both sides of the last display are positive, since $x \in (0,1)$, $\pi \in (0,1)$ and the densities are positive. Taking logarithms and using $\log(ab) = \log a + \log b$,

$$\log \frac{x}{1-x} = \log \frac{f_1(x)}{f_0(x)} + \log \frac{\pi}{1-\pi}.$$

The left-hand side is $\operatorname{logit}(x)$ by definition.

Step 7: conclusion. The identity holds for every $x \in (0,1)$ at which the densities are positive. Evaluating it at the realized forecast $x = p_{k,t}$ gives the statement of the lemma. □

9.3.4 Proof of Lemma 4

Lemma 4. *Under Assumption A1, g_t is $\mathcal{F}_t$-measurable, and multiplying any forecast by g_t incurs zero loss on crisis quarters.*

Proof. Measurability. By Definition 8, g_t is computed from two quantities: the current indicator C_t, and the date $e_t = \max\{s < t : C_s = 1\}$ of the last crisis quarter before t, when such a quarter exists. The first is observed at t by Assumption A1. The second is a function of $C_{1:t-1}$: whether a crisis quarter exists before t, and if so which is the last one, is read off the sequence $C_1, \ldots, C_{t-1}$. Hence g_t is a function of $C_{1:t}$, which is $\mathcal{F}_t$-measurable by Assumption A1. Explicitly, $g_t = 0$ if $C_t = 1$; $g_t = \min\{1, (t - e_t)/\phi\}$ if $C_t = 0$ and a crisis quarter exists before t; and $g_t = 1$ otherwise. In every case the value is determined by $C_{1:t}$.

Zero loss on crisis quarters. Let t be a crisis quarter, $C_t = 1$. Then $g_t = 0$ by Definition 8, so that for any forecast $\hat{p}_t \in [0,1]$ the gated forecast is $g_t \hat{p}_t = 0$. On the other hand, the target is $y_t = 0$ by Definition 1, since the factor $(1 - C_t)$ vanishes, and the effective label is $\tilde{y}_t = y_t = 0$ by Definition 3. The gated forecast therefore coincides with the label, and any loss that vanishes when forecast and label coincide, in particular the weighted square loss $\ell_t(x) = \omega_t(x - \tilde{y}_t)^2$ of Definition 9, equals $\ell_t(0) = \omega_t(0 - 0)^2 = 0$. □

9.3.5 Proof of Lemma ??

Lemma 5 (Exp-concavity)**.** *Lemmalem:expconcave For $y \in \{0,1\}$, $\omega \in (0,1]$ and $x \in [0,1]$, the map $x \mapsto \exp\{-\eta\, \omega\, (x-y)^2\}$ is concave whenever $\eta \le 1/(2\omega)$.*

Proof. Let $\psi(x) = \exp\{-\eta\omega(x-y)^2\}$ for $x \in [0,1]$, with y, ω and $\eta > 0$ fixed. The function ψ is twice differentiable, and a twice differentiable function on an interval is concave if and only if its second derivative is non-positive on that interval. We compute the derivatives with the chain rule.

First derivative. Writing $\psi(x) = e^{u(x)}$ with $u(x) = -\eta\omega(x-y)^2$ and $u'(x) = -2\eta\omega(x-y)$,

$$\psi'(x) = u'(x)\, e^{u(x)} = -2\eta\omega\, (x-y)\, \psi(x).$$

Second derivative. Differentiating the product $-2\eta\omega(x-y)\cdot\psi(x)$,

$$\psi''(x) = -2\eta\omega\,\psi(x) \;+\; \big(-2\eta\omega(x-y)\big)\,\psi'(x) = -2\eta\omega\,\psi(x) \;+\; 4\eta^2\omega^2(x-y)^2\,\psi(x).$$

Factoring $2\eta\omega\,\psi(x)$, which is positive,

$$\psi''(x) = 2\eta\omega\,\psi(x)\,\big[\,2\eta\omega(x-y)^2 - 1\,\big].$$

Sign. Since $2\eta\omega\,\psi(x) > 0$, the sign of $\psi''(x)$ is that of the bracket, and $\psi''(x) \le 0$ if and only if $2\eta\omega(x-y)^2 \le 1$. Now $x \in [0,1]$ and $y \in \{0,1\}$, so $|x-y| \le 1$ and $(x-y)^2 \le 1$. Hence, if $\eta \le 1/(2\omega)$, then $2\eta\omega(x-y)^2 \le 2\eta\omega \le 1$ for every $x \in [0,1]$, so that $\psi'' \le 0$ on $[0,1]$ and ψ is concave there.

The condition is also necessary: at $x = 1-y$, which lies in $[0,1]$, $(x-y)^2 = 1$, and $\psi''(1-y) \le 0$ requires $2\eta\omega \le 1$. □

Remark 2. Concavity of $x \mapsto e^{-\eta\ell(x)}$ is the property, called exp-concavity of the loss, that makes the exponentially weighted average forecaster achieve a regret that does not grow with the number of rounds (Cesa-Bianchi and Lugosi, 2006, Section 3.3). The lemma shows that the weighted square loss is exp-concave on $[0,1]$ as soon as the learning rate is at most $1/(2\omega)$; the choice $\eta = 1/(2\max\{\theta, 1-\theta\})$ in Definition 9 is the largest learning rate that satisfies this condition for both values $\omega \in \{\theta, 1-\theta\}$ of the weight.

9.3.6 Proof of Proposition 1

Proposition 1 (Regret of the meta-aggregator)**.** *Let $\mathcal{T} = \{t : C_t = 0\}$ and let N be the number of mixtures created. For any sequence of predictive models' forecasts and labels, and any candidate $j \in \{0, \dots, K\}$,*

$$\sum_{t\in\mathcal{T}} \ell_t(g_t q_t) \;\le\; \sum_{t\in\mathcal{T}} \ell_t(x_{j,t}) \;+\; \frac{N\,\log(1/u_j^0)}{\eta} \;+\; \max\{\theta, 1-\theta\}\; n_{\text{fade}}, \qquad N \le H,$$

where $n_{\text{fade}} = |\{t \in \mathcal{T} : g_t < 1,\ \tilde{y}_t = 1\}|$.

Proof. The proof proceeds in four steps. Step (i) proves the regret bound for a single exponentially weighted mixture that observes the label of each round before forecasting the next; this is the classical argument for exp-concave losses. Step (ii) handles the delay by splitting the rounds among the N mixtures of the pool, each of which faces no delay on its own rounds, and bounds N. Step (iii) checks that the convention for unavailable candidates does not affect the argument. Step (iv) accounts for the gate. Throughout, all forecasts lie in $[0,1]$: the predictive models' forecasts are probabilities, $x_{0,t}$ is a convex combination of them, and so is q_t.

(i) A single mixture without delay. Consider one mixture and let $t_1 < t_2 < \cdots < t_n$ be the rounds assigned to it, in the order in which it forecasts them. By construction of the pool, when the mixture forecasts round t_m, the labels of all its earlier rounds $t_1, \dots, t_{m-1}$ have been revealed, so that its weights have been updated with the losses of those rounds and of no others. Writing

$$L_{j,m} = \sum_{l=1}^{m} \ell_{t_l}(x_{j,t_l})$$

for the cumulative loss of candidate j on the first m rounds of the mixture, with $L_{j,0} = 0$, the weights used at round t_m are therefore

$$u_{j,t_m} = \frac{u_j^0\, e^{-\eta L_{j,m-1}}}{\Phi_{m-1}}, \qquad \Phi_m := \sum_{j=0}^{K} u_j^0\, e^{-\eta L_{j,m}},$$

since applying the update $u_j \leftarrow u_j e^{-\eta\ell_s(x_{j,s})}$ once for each revealed round and renormalizing gives exactly $u_j^0 e^{-\eta L_{j,m-1}}$ divided by the sum over j. Note that $\Phi_0 = \sum_j u_j^0 = 1$. The quantity

Φ_m is the *potential* of the mixture; the proof consists in showing that the loss of the mixture at each round is controlled by the decrease of the potential.

Jensen's inequality. At round t_m, the mixture forecasts $q_{t_m} = \sum_j u_{j,t_m} x_{j,t_m}$, a convex combination of the candidates' forecasts. Let $\psi(x) = \exp\{-\eta\, \ell_{t_m}(x)\} = \exp\{-\eta\, \omega_{t_m}(x - \tilde{y}_{t_m})^2\}$. Since $\omega_{t_m} \in \{\theta, 1-\theta\}$, we have $\omega_{t_m} \le \max\{\theta, 1-\theta\}$ and hence $\eta = 1/(2\max\{\theta, 1-\theta\}) \le 1/(2\omega_{t_m})$: by Lemma **??**, ψ is concave on $[0,1]$. Jensen's inequality for a concave function states that the value at a convex combination is at least the convex combination of the values, so that

$$\exp\{-\eta\, \ell_{t_m}(q_{t_m})\} = \psi\Big(\sum_j u_{j,t_m} x_{j,t_m}\Big) \;\ge\; \sum_j u_{j,t_m}\, \psi(x_{j,t_m}) = \sum_j u_{j,t_m}\, e^{-\eta\, \ell_{t_m}(x_{j,t_m})}. \tag{27}$$

The right-hand side is a ratio of potentials. Substituting the expression of the weights,

$$\sum_j u_{j,t_m}\, e^{-\eta\, \ell_{t_m}(x_{j,t_m})} = \sum_j \frac{u_j^0\, e^{-\eta L_{j,m-1}}}{\Phi_{m-1}}\, e^{-\eta\, \ell_{t_m}(x_{j,t_m})} = \frac{1}{\Phi_{m-1}} \sum_j u_j^0\, e^{-\eta(L_{j,m-1} + \ell_{t_m}(x_{j,t_m}))} = \frac{\Phi_m}{\Phi_{m-1}},$$

where the last equality uses $L_{j,m-1} + \ell_{t_m}(x_{j,t_m}) = L_{j,m}$. Hence equation 27 reads

$$\exp\{-\eta\, \ell_{t_m}(q_{t_m})\} \;\ge\; \frac{\Phi_m}{\Phi_{m-1}}, \qquad m = 1, \dots, n.$$

Telescoping. Multiplying these n inequalities, whose sides are all positive, the ratios of potentials cancel in chain:

$$\exp\Big\{-\eta \sum_{m=1}^{n} \ell_{t_m}(q_{t_m})\Big\} = \prod_{m=1}^{n} \exp\{-\eta\, \ell_{t_m}(q_{t_m})\} \;\ge\; \prod_{m=1}^{n} \frac{\Phi_m}{\Phi_{m-1}} = \frac{\Phi_n}{\Phi_0} = \Phi_n.$$

Comparison with one candidate. All terms of the sum defining Φ_n are non-negative, so Φ_n is at least any one of them: for any candidate j,

$$\Phi_n \;\ge\; u_j^0\, e^{-\eta L_{j,n}}.$$

Combining the last two displays and taking logarithms, which preserves the inequality since the logarithm is increasing,

$$-\eta \sum_{m=1}^{n} \ell_{t_m}(q_{t_m}) \;\ge\; \log u_j^0 - \eta L_{j,n}.$$

Dividing by $-\eta < 0$, which reverses the inequality, and recalling the definition of $L_{j,n}$,

$$\sum_{m=1}^{n} \ell_{t_m}(q_{t_m}) \;\le\; \sum_{m=1}^{n} \ell_{t_m}(x_{j,t_m}) + \frac{\log(1/u_j^0)}{\eta}. \tag{28}$$

This is the regret bound of a single mixture on its own rounds (Cesa-Bianchi and Lugosi, 2006, Proposition 3.1): it holds for any sequence of forecasts and labels, and its right-hand side does not depend on the number of rounds n.

(ii) Delay and the number of mixtures. By Definition 9, every non-crisis round $t \in \mathcal{T}$ is assigned to exactly one mixture, and the meta-aggregator's forecast q_t at that round is the forecast of that mixture. The sets of rounds of the N mixtures therefore form a partition of $\mathcal{T}$. Applying equation 28 to each mixture and summing over the N mixtures, the sums of losses over the parts recombine into sums over $\mathcal{T}$:

$$\sum_{t \in \mathcal{T}} \ell_t(q_t) \;\le\; \sum_{t \in \mathcal{T}} \ell_t(x_{j,t}) + \frac{N \log(1/u_j^0)}{\eta}. \tag{29}$$

This is the argument of Joulani et al. (2013): a delayed problem is solved by a pool of undelayed forecasters, at the price of multiplying the regret bound by the size of the pool.

It remains to bound N. Say that a mixture is *busy* at round t if it has been assigned a round

$s < t$ whose label has not been revealed by t, that is, $\tilde{r}_s > t$; a mixture is used at t only if it is not busy. By Definition 3, $\tilde{r}_s \leq s + H$ for every non-crisis quarter s, so that $\tilde{r}_s > t$ implies $s > t - H$, that is, $s \in \{t - H + 1, \ldots, t - 1\}$. This set contains $H - 1$ dates, and each date is assigned to at most one mixture, so at most $H - 1$ mixtures are busy at any round. A new mixture is created only when all existing mixtures are busy; if the pool already contained H mixtures, at least one of them would be free, and no mixture would be created. Hence the pool never exceeds H mixtures: $N \leq H$.

(iii) Unavailable candidates. At some rounds, a predictive model k has no forecast because all its indicators are missing; Definition 9 then sets $x_{k,t} = q_t$. This convention is consistent: writing $\mathcal{A}_t$ for the set of available candidates at t, the defining equation $q_t = \sum_j u_{j,t} x_{j,t}$ becomes

$$q_t = \sum_{j \in \mathcal{A}_t} u_{j,t}\, x_{j,t} + q_t \sum_{j \notin \mathcal{A}_t} u_{j,t}, \qquad \text{i.e.} \qquad q_t = \frac{\sum_{j \in \mathcal{A}_t} u_{j,t}\, x_{j,t}}{\sum_{j \in \mathcal{A}_t} u_{j,t}},$$

so that q_t is the weighted average of the available candidates, with weights renormalized to sum to one; it is well defined as long as one candidate is available, and it lies in $[0, 1]$. Since the unavailable candidates are assigned the value $q_t \in [0, 1]$, every step of (i) applies verbatim: Jensen's inequality is applied to a convex combination of points of $[0, 1]$, one of which is q_t itself, and the potential argument is unchanged. The bound equation 29 therefore holds for every candidate j. For a round at which j is unavailable, $\ell_t(x_{j,t}) = \ell_t(q_t)$, so that this round contributes the same amount to both sides of equation 29 and cancels. The bound thus compares the meta-aggregator with candidate j on the rounds where j is available, which is the meaningful comparison (Freund et al., 1997). The detection rule, $j = 0$, is available whenever at least one predictive model is.

(iv) Gate. The final forecast is $g_t q_t$, and the proposition bounds its loss on the non-crisis rounds $t \in \mathcal{T}$. (On crisis rounds, $t \notin \mathcal{T}$, its loss is zero by Lemma 4; these rounds are excluded from both sides of the bound.) We compare $\ell_t(g_t q_t)$ with $\ell_t(q_t)$ for $t \in \mathcal{T}$, using $g_t \in [0, 1]$ and $q_t \in [0, 1]$.

If $g_t = 1$, the two losses are equal.

If $g_t < 1$ and $\tilde{y}_t = 0$, then $\ell_t(g_t q_t) = \omega_t (g_t q_t)^2 \leq \omega_t q_t^2 = \ell_t(q_t)$, since $0 \leq g_t q_t \leq q_t$: shrinking a forecast toward zero cannot increase the loss when the label is zero.

If $g_t < 1$ and $\tilde{y}_t = 1$, the gate may increase the loss, but the loss remains bounded: since $g_t q_t \in [0, 1]$, $(g_t q_t - 1)^2 \leq 1$, and $\ell_t(g_t q_t) \leq \omega_t \leq \max\{\theta, 1 - \theta\}$. By definition, there are n_{fade} such rounds in $\mathcal{T}$.

Summing over $t \in \mathcal{T}$, and bounding the loss by $\ell_t(q_t)$ on the rounds of the first two kinds and by $\max\{\theta, 1 - \theta\}$ on the rounds of the third kind,

$$\sum_{t \in \mathcal{T}} \ell_t(g_t q_t) \;\leq\; \sum_{t \in \mathcal{T}} \ell_t(q_t) + \max\{\theta, 1 - \theta\}\, n_{\text{fade}},$$

where we have used $\ell_t(q_t) \geq 0$ to add the losses of the rounds of the third kind to the first sum. Combining with equation 29 gives the proposition. □

Remark 3. Against the detection rule ($j = 0$, $u_0^0 = 0.9$), the bound is $18 \times \log(1/0.9)/0.625 \approx 3.0$ for $h = 12$, hence $H = 18$, small relative to the total loss at stake: the meta-aggregator essentially reproduces the rule. Against each predictive model ($u_k^0 = 0.025$) the bound is approximately 106, uninformative on our sample but, unlike Theorem 3, independent of T, so that the average regret vanishes at rate $H \log(1/u_j^0)/(\eta T)$. The guarantee is insurance against the detection rule failing in the long run, not a source of performance. Regret is measured against $\tilde{y}$, which differs from y only on the six tolerance quarters before each window.

9.4 The loss function

The weights of the learning layer depend on the predictive models' past losses, so the loss defines what performing well means. We require four properties: (i) the loss rewards forecasts that are informative about the probability of a pre-crisis quarter, which is the object the predictive models estimate; (ii) it reflects the policymaker's asymmetric costs; (iii) it is not dominated by the calm

quarters, which make up about 85 percent of the sample; and (iv) it is bounded, since every regret bound scales with the range of the loss.

The base loss. We use the squared loss, or Brier score (Glenn et al., 1950), $(p-y)^2$ for a forecast $p \in [0,1]$ and a label $y \in \{0,1\}$. It satisfies (i) because it is a *strictly proper* scoring rule (Gneiting and Raftery, 2007): if the label is a Bernoulli variable with probability q, the expected loss of forecasting p is

$$\mathbb{E}\big[(p-y)^2\big] = q\,(1-p)^2 + (1-q)\,p^2,$$

a quadratic function of p with derivative $-2q(1-p) + 2(1-q)p = 2(p-q)$, which vanishes only at $p = q$ and is negative before and positive after: the expected loss is uniquely minimized by reporting the true probability. The Brier score is also bounded in $[0,1]$, which gives (iv). The two natural alternatives each fail one requirement. The absolute loss $|p-y|$ is not proper: its expected value $q(1-p) + (1-q)p = q + (1-2q)\,p$ is linear in p, hence minimized at $p = 1$ if $q > 1/2$ and at $p = 0$ if $q < 1/2$; it rewards extreme forecasts rather than accurate probabilities. The logarithmic loss $-y \log p - (1-y)\log(1-p)$ is proper but unbounded: a single quarter in which a penalized logit forecasts $p = 10^{-4}$ before a crisis costs $\log(10^4) \approx 9.2$, more than nine calm quarters forecast at $p = 1/2$, and would dominate the cumulative loss. The policymaker's utility of Definition 10, finally, is computed through a threshold, is not a sum of per-quarter terms and is not convex in the forecast, so it cannot drive the weights, although it is the criterion by which we evaluate them.

Class and cost weights. Requirements (ii) and (iii) are met by weighting the squared loss of each revealed quarter $s \in \mathcal{R}_t$ by

$$\begin{aligned} \omega_s &= \frac{\theta}{\hat{\pi}_t}\,\mathbf{1}\{\tilde{y}_s = 1\} + \frac{1-\theta}{1-\hat{\pi}_t}\,\mathbf{1}\{\tilde{y}_s = 0\}, \\ \hat{\pi}_t &= \min\Big\{\max\big\{\bar{\pi}_t,\ 0.05\big\},\ 0.5\Big\}, \qquad \bar{\pi}_t = \frac{|\{s \in \mathcal{R}_t : \tilde{y}_s = 1\}|}{|\mathcal{R}_t|}, \end{aligned} \tag{30}$$

where $\hat{\pi}_t$ is the share of positive labels among the quarters revealed by t, clipped to $[0.05, 0.5]$, and $\theta = 0.8$ is the policymaker's relative cost of a missed crisis, as in the evaluation criterion. The two factors of each weight have distinct roles.

Balancing the classes. Suppose that $n = |\mathcal{R}_t|$ quarters have been revealed and that the share of positives, π, lies within the clipping bounds, so that $\hat{\pi}_t = \pi$. The $n\pi$ positive quarters each receive the weight θ/π, for a total of $n\pi \cdot \theta/\pi = n\theta$; the $n(1-\pi)$ negative quarters each receive $(1-\theta)/(1-\pi)$, for a total of $n(1-\theta)$. The total weight of each class therefore does not depend on π. Without this correction, with $\pi \approx 0.15$, about 85 percent of the cumulative loss would come from calm quarters, and the predictive model that always forecasts zero, whose loss is zero on every calm quarter, would be the best predictive model by a wide margin.

Pricing the classes. The factors θ and $1-\theta$ then set the ratio of the total weights of the two classes to $\theta/(1-\theta) = 4$: a unit of squared error on a pre-crisis quarter costs four times a unit of squared error on a calm quarter, which is the policymaker's trade-off in Definition 10.

Bounding the loss. The clipping of $\hat{\pi}_t$ serves two purposes. Early in the online period, few labels have been revealed and their share of positives can be zero or one; the bounds prevent the weights from exploding or being undefined. They also bound the weighted loss of any quarter: since $(p-y)^2 \le 1$,

$$\omega_s\,(p-\tilde{y}_s)^2 \ \le\ \max\Big\{\frac{\theta}{0.05},\ \frac{1-\theta}{1-0.5}\Big\} = \max\{16,\ 0.4\} = 16 =: B.$$

This is the constant B that enters the learning rate $\eta_t = \sqrt{8\log K/t}/B$ and the regret bounds below.

Remark 4 (Time-varying weights). Because $\hat{\pi}_t$ is re-estimated as labels arrive, the weight ω_s of a given past quarter changes with t: the cumulative loss $L_{k,t} = \sum_{s\in\mathcal{R}_t} \omega_s\,(p_{k,s} - \tilde{y}_s)^2$ is a reweight-

ing of past errors rather than a sum of fixed per-round losses. This departs from the framework of prediction with expert advice, in which the loss of a round is fixed once revealed, and it is one reason why the learning layer of the detection rule is not given a regret guarantee of its own. The meta-aggregator of Definition 9, for which Proposition 1 is proved, uses the fixed weights $\omega_t \in \{\theta, 1-\theta\}$, without the class-balancing factor; this is what makes it a standard instance of the protocol. The regret bounds of the next subsection, which are stated for fixed losses, therefore describe the learning layer in the special case where $\hat{\pi}_t$ is held fixed; they are the guarantees that a pure exponentially weighted average would have, and the point of Section 9.6 is that these guarantees are vacuous under the delay of our problem.

9.5 Regret of exponentially weighted averages with a prior

This subsection states and proves the regret bound of the exponentially weighted average forecaster (Vovk, 1990; Littlestone and Warmuth, 1994) in the form used by the learning layer: a non-uniform initial weight vector, the batch prior, and a learning rate that decreases with time, so that the forecaster does not need to know the horizon. The proofs follow Cesa-Bianchi and Lugosi (2006, Chapter 2); we give them in full because the extension to a prior is where the batch information enters the guarantee.

Setting. There are K predictive models. At each round $t = 1, \dots, T$, predictive model k incurs a loss $\ell_{k,t} \in [0, B]$, and the forecaster, who forecasts $\hat{p}_t = \sum_k w_{k,t}\, p_{k,t}$ with weights w_t in the simplex, incurs a loss $\hat{\ell}_t$. We assume that the loss is *convex in the forecast*, so that

$$\hat{\ell}_t \;\le\; \sum_{k=1}^{K} w_{k,t}\, \ell_{k,t}; \tag{31}$$

this holds for the weighted squared loss, since $x \mapsto \omega\,(x-y)^2$ is convex and Jensen's inequality applies. We write $L_{k,t} = \sum_{s \le t} \ell_{k,s}$ and $\hat{L}_t = \sum_{s \le t} \hat{\ell}_s$ for the cumulative losses, and the *regret* with respect to the predictive model k is $\mathrm{Reg}_T(k) = \hat{L}_T - L_{k,T}$. The losses are arbitrary: no assumption is made on how they are generated, and the bounds hold for every sequence.

The forecaster. Given a prior w^0 in the simplex with $w^0_k > 0$ for all k, and a sequence of learning rates $\eta_1 \ge \eta_2 \ge \cdots > 0$, the exponentially weighted average forecaster uses at round t the weights

$$w_{k,t} = \frac{w^0_k\, e^{-\eta_t L_{k,t-1}}}{\sum_{j=1}^{K} w^0_j\, e^{-\eta_t L_{j,t-1}}}, \qquad L_{k,0} = 0. \tag{32}$$

Subtracting $\min_j L_{j,t-1}$ from every cumulative loss, as the learning layer does for numerical stability, multiplies numerator and denominator by the same factor and leaves the weights unchanged.

The proofs rest on one inequality from probability theory.

Lemma 6 (Hoeffding's lemma). *Let X be a random variable with values in an interval $[a, b]$. Then for every real s,*

$$\log \mathbb{E}\big[e^{sX}\big] \;\le\; s\,\mathbb{E}[X] + \frac{s^2 (b-a)^2}{8}.$$

Proof. Let $\psi(s) = \log \mathbb{E}[e^{sX}]$, the cumulant generating function of X; it is finite for every s because X is bounded, and it is twice differentiable, with

$$\psi'(s) = \frac{\mathbb{E}[X e^{sX}]}{\mathbb{E}[e^{sX}]}, \qquad \psi''(s) = \frac{\mathbb{E}[X^2 e^{sX}]}{\mathbb{E}[e^{sX}]} - \Big(\frac{\mathbb{E}[X e^{sX}]}{\mathbb{E}[e^{sX}]}\Big)^2.$$

These two expressions have a probabilistic meaning. Define the *tilted* probability measure $\Pr_s$ by $d\Pr_s = e^{sX}\, d\Pr\, / \mathbb{E}[e^{sX}]$; it is a probability measure because its density is positive and integrates

to one. Then $\psi'(s) = \mathbb{E}_s[X]$ and $\psi''(s) = \mathbb{E}_s[X^2] - \mathbb{E}_s[X]^2 = \mathrm{Var}_s(X)$, the variance of X under the tilted measure. Under any probability measure, a variable with values in $[a, b]$ has variance at most $(b-a)^2/4$: writing $m = (a+b)/2$ for the midpoint, $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2] \le \mathbb{E}[(X-m)^2] \le \big((b-a)/2\big)^2$, the first inequality because the mean minimizes $c \mapsto \mathbb{E}[(X-c)^2]$ and the second because $|X - m| \le (b-a)/2$. Hence $\psi''(s) \le (b-a)^2/4$ for every s. Finally, $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbb{E}[X]$, so Taylor's formula with Lagrange remainder gives, for some ξ between 0 and s,

$$\psi(s) = \psi(0) + s\,\psi'(0) + \frac{s^2}{2}\,\psi''(\xi) \;\le\; s\,\mathbb{E}[X] + \frac{s^2(b-a)^2}{8}. \qquad \square$$

Theorem 1 (Fixed learning rate, Cesa-Bianchi and Lugosi (2006))**.** *For a constant learning rate $\eta_t = \eta > 0$, the forecaster equation 32 satisfies, for every sequence of losses in $[0, B]$ and every predictive model k,*

$$\mathrm{Reg}_T(k) \;\le\; \frac{\log(1/w_k^0)}{\eta} + \frac{\eta\, B^2\, T}{8}. \tag{33}$$

For $\eta = \sqrt{8\log K/T}/B$ and a uniform prior, the bound equals $B\sqrt{(T/2)\log K}$.

Proof. Define the potential $\Phi_t = \sum_k w_k^0\, e^{-\eta L_{k,t}}$, so that $\Phi_0 = \sum_k w_k^0 = 1$ and, by equation 32, $w_{k,t} = w_k^0 e^{-\eta L_{k,t-1}}/\Phi_{t-1}$.

Step 1: one round. Using $L_{k,t} = L_{k,t-1} + \ell_{k,t}$,

$$\frac{\Phi_t}{\Phi_{t-1}} = \sum_k \frac{w_k^0\, e^{-\eta L_{k,t-1}}}{\Phi_{t-1}}\, e^{-\eta \ell_{k,t}} = \sum_k w_{k,t}\, e^{-\eta \ell_{k,t}}.$$

The right-hand side is $\mathbb{E}[e^{-\eta X}]$ for a random variable X that takes the value $\ell_{k,t}$ with probability $w_{k,t}$; it has values in $[0, B]$. By Lemma 6 with $s = -\eta$,

$$\log \frac{\Phi_t}{\Phi_{t-1}} \;\le\; -\eta \sum_k w_{k,t}\, \ell_{k,t} + \frac{\eta^2 B^2}{8} \;\le\; -\eta\, \hat{\ell}_t + \frac{\eta^2 B^2}{8},$$

where the second inequality is equation 31.

Step 2: summing over rounds. Adding these inequalities for $t = 1, \dots, T$, the logarithms of the ratios telescope:

$$\log \Phi_T = \log \Phi_T - \log \Phi_0 = \sum_{t=1}^{T} \log \frac{\Phi_t}{\Phi_{t-1}} \;\le\; -\eta\, \hat{L}_T + \frac{\eta^2 B^2 T}{8}.$$

Step 3: comparison with one predictive model. Every term of Φ_T is positive, so $\Phi_T \ge w_k^0\, e^{-\eta L_{k,T}}$, and $\log \Phi_T \ge \log w_k^0 - \eta L_{k,T}$. Combining with Step 2,

$$\log w_k^0 - \eta L_{k,T} \;\le\; -\eta\, \hat{L}_T + \frac{\eta^2 B^2 T}{8}.$$

Rearranging and dividing by $\eta > 0$ gives equation 33. For the last claim, with $w_k^0 = 1/K$ the bound is $\log K/\eta + \eta B^2 T/8$, a function of η of the form $a/\eta + b\eta$ with $a = \log K$ and $b = B^2 T/8$, minimized at $\eta = \sqrt{a/b} = \sqrt{8\log K/T}/B$ with minimal value $2\sqrt{ab} = B\sqrt{(T/2)\log K}$. $\square$

The bound 33 shows the role of the prior: the price of competing with predictive model k is $\log(1/w_k^0)/\eta$, which is small for a predictive model that the prior already favours and large for one it neglects, but finite for every predictive model since $w_k^0 > 0$. This is why the learning layer keeps a floor on the prior weights: a predictive model with prior weight zero could never be recovered. The bound also shows the cost of a fixed learning rate: the optimal η depends on the horizon T. The next result removes this dependence.

Theorem 2 (Anytime learning rate, Cesa-Bianchi and Lugosi (2006))**.** *For the decreasing learning rates $\eta_t = \sqrt{8\log K/t}/B$, the forecaster equation 32 satisfies, for every T, every sequence of losses in $[0, B]$ and every predictive model k,*

$$\mathrm{Reg}_T(k) \;\le\; \frac{\log(1/w_k^0)}{\eta_T} + \frac{B^2}{8}\sum_{t=1}^{T}\eta_t \;\le\; B\sqrt{T}\Big(\frac{\log(1/w_k^0)}{\sqrt{8\log K}} + \sqrt{\frac{\log K}{2}}\Big). \tag{34}$$

With a uniform prior, the right-hand side is $\frac{3}{2}B\sqrt{(T/2)\log K}$.

Proof. The learning rate now changes between rounds, so the potential must be indexed by both the round and the rate. For $\eta > 0$ and $t \ge 0$, define

$$\Phi_t(\eta) = \frac{1}{\eta}\log\sum_k w_k^0\, e^{-\eta L_{k,t}}, \qquad \Phi_0(\eta) = \frac{1}{\eta}\log 1 = 0.$$

Note that the weights of round t in equation 32 are computed with the rate η_t, so that $w_{k,t} = w_k^0 e^{-\eta_t L_{k,t-1}}/\sum_j w_j^0 e^{-\eta_t L_{j,t-1}}$.

Step 1: one round, at the current rate. Exactly as in Theorem 1, with $\eta = \eta_t$,

$$\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t) = \frac{1}{\eta_t}\log\sum_k w_{k,t}\, e^{-\eta_t \ell_{k,t}} \;\le\; -\hat{\ell}_t + \frac{\eta_t B^2}{8}.$$

Hence $\hat{\ell}_t \le \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t) + \eta_t B^2/8$.

Step 2: summing. Adding over $t = 1, \dots, T$,

$$\hat{L}_T \;\le\; \sum_{t=1}^{T}\big[\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t)\big] + \frac{B^2}{8}\sum_{t=1}^{T}\eta_t.$$

The sum no longer telescopes directly, because consecutive terms are evaluated at different rates. Writing it out,

$$\sum_{t=1}^{T}\big[\Phi_{t-1}(\eta_t) - \Phi_t(\eta_t)\big] = \Phi_0(\eta_1) - \Phi_T(\eta_T) + \sum_{t=2}^{T}\big[\Phi_{t-1}(\eta_t) - \Phi_{t-1}(\eta_{t-1})\big],$$

as one checks by expanding both sides: each Φ_{t-1} for $2 \le t \le T$ appears once with the rate η_t and once with the rate η_{t-1}.

Step 3: the potential is nondecreasing in the rate. We claim that for fixed t, $\eta \mapsto \Phi_t(\eta)$ is nondecreasing. Differentiating, and writing $q_k(\eta) = w_k^0 e^{-\eta L_{k,t}}/\sum_j w_j^0 e^{-\eta L_{j,t}}$ for the normalized weights at rate η,

$$\begin{aligned}\Phi_t'(\eta) &= -\frac{1}{\eta^2}\log\sum_k w_k^0 e^{-\eta L_{k,t}} - \frac{1}{\eta}\sum_k q_k(\eta)\, L_{k,t} \\ &= \frac{1}{\eta^2}\sum_k q_k(\eta)\Big[-\log\sum_j w_j^0 e^{-\eta L_{j,t}} - \eta L_{k,t}\Big] = \frac{1}{\eta^2}\sum_k q_k(\eta)\log\frac{q_k(\eta)}{w_k^0},\end{aligned}$$

where the last equality uses $\log q_k(\eta) = \log w_k^0 - \eta L_{k,t} - \log\sum_j w_j^0 e^{-\eta L_{j,t}}$. The sum is the Kullback–Leibler divergence of $q(\eta)$ from w^0, which is non-negative. Hence $\Phi_t' \ge 0$. (As $\eta \to 0$, $\Phi_t(\eta)$ tends to $-\sum_k w_k^0 L_{k,t}$, the prior-weighted average loss, and as $\eta \to \infty$ to $-\min_k L_{k,t}$; the potential interpolates between minus the average and minus the minimum, and a larger rate brings it closer to the minimum.)

Since $\eta_t \le \eta_{t-1}$, Step 3 gives $\Phi_{t-1}(\eta_t) \le \Phi_{t-1}(\eta_{t-1})$: every bracket in the last sum of Step 2 is

non-positive, and dropping these terms can only increase the right-hand side. With $\Phi_0(\eta_1) = 0$,

$$\hat{L}_T \;\le\; -\Phi_T(\eta_T) + \frac{B^2}{8}\sum_{t=1}^{T}\eta_t.$$

Step 4: comparison with one predictive model. As in Theorem 1, $\sum_k w_k^0 e^{-\eta_T L_{k,T}} \ge w_k^0 e^{-\eta_T L_{k,T}}$, so that $-\Phi_T(\eta_T) \le L_{k,T} + \log(1/w_k^0)/\eta_T$. This gives the first inequality of equation 34.

Step 5: the explicit rates. With $\eta_t = \sqrt{8\log K/t}/B$, we have $1/\eta_T = B\sqrt{T/(8\log K)}$, and, comparing the sum with an integral, $\sum_{t=1}^{T} t^{-1/2} \le 1 + \int_1^T x^{-1/2}dx = 2\sqrt{T} - 1 \le 2\sqrt{T}$, so that

$$\frac{B^2}{8}\sum_{t=1}^{T}\eta_t \;\le\; \frac{B^2}{8}\cdot\frac{\sqrt{8\log K}}{B}\cdot 2\sqrt{T} = B\sqrt{\frac{T\log K}{2}}.$$

Substituting gives the second inequality. With $w_k^0 = 1/K$, the first term is $B\sqrt{T\log K/8} = \frac{1}{2}B\sqrt{(T/2)\log K}$, and the total is $\frac{3}{2}B\sqrt{(T/2)\log K}$. □

Remark 5 (What the bounds say for the learning layer). Three features of equation 34 matter for the paper. First, the regret grows as $\sqrt{T}$, so the average regret per round, $\mathrm{Reg}_T(k)/T$, vanishes: in the long run, the forecaster does at least as well as every predictive model, for any sequence of losses. Second, the dependence on the number of predictive models is logarithmic, and the dependence on the prior is through $\log(1/w_k^0)$: a prior that puts weight 0.1 on a predictive model costs $\log 10 \approx 2.3$ in place of $\log 4 \approx 1.4$ for the uniform prior, a modest price for using the batch information. Third, the bounds are stated for losses revealed at the end of each round. Under the delay of our problem, the losses of round t are revealed up to H rounds later, and the guarantee deteriorates as described next.

9.6 Why loss-based aggregation fails under delayed feedback

We provide this part to justify the extension we introduce to the standard online theory approach. Consider the protocol of Section 9.5, with one change: the loss of round s is revealed at a date $r_s \le s + D$ rather than at the end of round s, so that at round t the forecaster knows only the losses of the rounds s with $r_s \le t$. The exponentially weighted average forecaster then cannot be run as in equation 32, since $L_{k,t-1}$ is not available. The standard remedy, due to Weinberger and Ordentlich (2002), is to run several copies of the forecaster in parallel, each on a subsequence of rounds spaced far enough apart that the delay is invisible to it.

Theorem 3. *Regret under delayed feedback, Weinberger and Ordentlich (2002) Suppose that the label of quarter t is revealed at the latest at $t + D$. Running $D + 1$ independent copies of an exponentially weighted average forecaster, each in charge of one quarter out of $D + 1$, yields, for losses in $[0, B]$ and uniform initial weights,*

$$\mathrm{Reg}_T(k) \;\le\; B\sqrt{\tfrac{1}{2}(D+1)\,T\log K}.$$

Proof. The pool. Index the copies by $c \in \{0, 1, \dots, D\}$ and assign round t to copy $c(t) = t \bmod (D + 1)$. Copy c is in charge of the rounds $t \equiv c \pmod{D+1}$, which are spaced exactly $D + 1$ apart; let n_c be their number, so that $\sum_c n_c = T$ and $n_c \le \lceil T/(D+1)\rceil$.

Each copy faces no delay. When copy c forecasts a round t, its previous round was $t - (D+1)$, whose loss is revealed at the latest at $t - (D+1) + D = t - 1 < t$. By induction, the losses of all its earlier rounds are known when it forecasts, so that on its own subsequence copy c runs the undelayed forecaster equation 32. Giving it the constant learning rate $\eta_c = \sqrt{8\log K/n_c}/B$ and a uniform prior, Theorem 1 bounds its regret on its n_c rounds by $B\sqrt{(n_c/2)\log K}$.

Summing over the copies. The forecaster's loss at round t is the loss of copy $c(t)$, and the subsequences partition $\{1, \dots, T\}$, so the regret over all rounds is the sum of the regrets of the

copies:

$$\mathrm{Reg}_T(k) \;\le\; \sum_{c=0}^{D} B\sqrt{\tfrac{1}{2} n_c \log K} = B\sqrt{\tfrac{1}{2}\log K}\,\sum_{c=0}^{D}\sqrt{n_c}.$$

By the Cauchy–Schwarz inequality applied to the vectors $(1,\dots,1)$ and $(\sqrt{n_0},\dots,\sqrt{n_D})$,

$$\sum_{c=0}^{D}\sqrt{n_c} \;\le\; \sqrt{(D+1)\sum_{c=0}^{D} n_c} = \sqrt{(D+1)\,T},$$

which gives the bound. □

Weinberger and Ordentlich (2002) also show that this rate cannot be improved in general: for a delay D, the minimax regret over all forecasters and all loss sequences is of order $B\sqrt{(D+1)T\log K}$. The factor $\sqrt{D+1}$ is therefore the price of delay for any loss-based procedure, not a weakness of the pool construction. Joulani et al. (2013) extend the construction to variable delays, which is the case of our problem, where $r_s - s$ ranges from 1 to H; the bound then holds with D equal to the maximal delay.

Corollary 1. *In our problem, $D = H = 18$, $K = 4$, and about $T = 90$ online quarters are evaluated per country, with losses bounded by $B = 16$. The bound of Theorem 3 is then $B\sqrt{\frac{1}{2}\cdot 19\cdot 90\cdot\log 4} \approx 34\,B$, or about $0.38\,B$ per quarter. An uninformative forecast of $1/2$ in every quarter incurs a weighted loss of at most $\omega_s/4 \le 0.25\,B$ per quarter.*

The guarantee is therefore vacuous on our sample: it allows the forecaster to lose more, on every quarter, than a forecaster that ignores the predictive models. This is a quantitative statement, but the obstacle it reflects is structural. By Lemma 1, the labels of the pre-crisis quarters of an episode, and with them the losses that would identify the predictive model that was right during the build-up, are revealed at the onset. A loss-based forecaster therefore moves its weight toward the relevant predictive model only once the crisis has started, whatever its learning rate: the information arrives, but too late to be useful for the episode that generates it. This is what the ablation without detection in Section 7 shows empirically, where the weights are those of the prior corrected by revealed losses: the aggregate forecast ranks quarters as well as the full procedure but no longer rises before the crises. A procedure that succeeds in this problem must use information that does not depend on feedback, namely the predictive models' current forecasts. This is what detection does.

9.7 The CUSUM as a maximum over change dates

This subsection proves the interpretation of Page's recursion given in Section 5.2.3. Recall that the statistic is defined by $W_0 = 0$ and $W_t = \max\{0,\; W_{t-1} + \ell_t\}$ for $t \ge 1$, where ℓ_t is the increment of date t. For $\nu \le t$, let

$$S_t(\nu) = \sum_{s=\nu}^{t} \ell_s$$

be the sum of the increments since date ν, and set $S_t(t+1) = 0$, an empty sum, which corresponds to the hypothesis that the change has not occurred by t.

Lemma 7 (The CUSUM is a maximum over possible change dates)**.** *For every $t \ge 0$,*

$$W_t = \max_{1\le\nu\le t+1} S_t(\nu).$$

Proof. By induction on t.

Initialization. For $t = 0$, the only candidate date is $\nu = 1$, and $S_0(1) = 0 = W_0$ by the empty-sum convention.

Induction step. Let $t \ge 1$ and suppose that $W_{t-1} = \max_{1\le\nu\le t} S_{t-1}(\nu)$. Split the candidate dates at time t into two groups.

- For $\nu \leq t$, the sum since ν gains the new increment: $S_t(\nu) = S_{t-1}(\nu) + \ell_t$. Adding the same constant ℓ_t to every element of a finite set adds it to the maximum, so the maximum over this group is $\max_{1 \leq \nu \leq t} S_{t-1}(\nu) + \ell_t = W_{t-1} + \ell_t$ by the induction hypothesis.
- For $\nu = t + 1$, $S_t(t + 1) = 0$ by convention.

The maximum over all candidate dates is the larger of the two group maxima, $\max\{W_{t-1} + \ell_t,\ 0\}$, which is W_t by definition of the recursion. □

Interpretation. Suppose that the increments are log-likelihood ratios, $\ell_s = \log f_1(X_s)/f_0(X_s)$, for independent observations X_s with density f_0 before the change and f_1 after. Then

$$S_t(\nu) = \log \prod_{s=\nu}^{t} \frac{f_1(X_s)}{f_0(X_s)} = \log \frac{\prod_{s<\nu} f_0(X_s) \prod_{s \geq \nu} f_1(X_s)}{\prod_{s \leq t} f_0(X_s)}$$

is the log-likelihood ratio of the hypothesis "the change occurred at date ν" against the hypothesis "no change has occurred by t": under the first, the observations from ν onward have density f_1; under the second, all observations have density f_0; and the observations before ν, which have density f_0 under both, cancel. By Lemma 7,

$$e^{W_t} = \max_{1 \leq \nu \leq t+1} \prod_{s=\nu}^{t} \frac{f_1(X_s)}{f_0(X_s)},$$

is the *generalized* likelihood ratio: the likelihood ratio evaluated at the most likely change date, with the value 1 for the hypothesis of no change. This is why the CUSUM does not require the change date to be known, and why it starts accumulating evidence only from the date at which the evidence starts.

Example. Suppose that the increments over six quarters are -1, -1, $+2$, -1, $+2$ and $+2$. The sums since each candidate date, evaluated at $t = 6$, are

ν	1	2	3	4	5	6	7
$S_6(\nu)$	3	4	5	3	4	2	0

Their maximum, 5, is reached at $\nu = 3$, the quarter in which the positive increments start. The recursion gives the same value: $W_1 = 0$, $W_2 = 0$, $W_3 = 2$, $W_4 = 1$, $W_5 = 3$, $W_6 = 5$. By contrast, the plain cumulative sum from the first quarter is $S_6(1) = 3$: the two calm quarters at the start reduce it, whereas the CUSUM discards them.

The bounded statistic. Lemma 7 holds for Page's recursion. The statistic of Definition 7 adds a clip c to each increment and a cap $W_{\max}$ to the statistic. The clip only changes the increments, so the lemma applies verbatim to the clipped increments: the statistic is the maximum over change dates of the sums of clipped increments. The cap breaks the exact equality: once the statistic reaches $W_{\max}$, the evidence beyond it is discarded, and later declines are measured from the cap, so that the capped statistic is at most the maximum of the lemma and can be strictly smaller after a period at the cap. The interpretation remains exact as long as the statistic has not reached the cap, and at the cap the evidence in favour of a change is already conclusive, a likelihood ratio of $e^{W_{\max}} \approx 400$, at which point the posterior puts almost all the weight on the corresponding sensor (Lemma 9 and Section 9.10).

9.8 Optimality of the CUSUM

The CUSUM is not only a convenient recursion: it is the optimal sequential detector for a single sensor, in a sense made precise by the following theorem. A *stopping time* τ is a rule that decides, at each date and on the basis of the observations so far, whether to raise an alarm. Two quantities

characterize its performance. The *average run length to false alarm*, $\mathbb{E}_\infty[\tau]$, is the expected time before an alarm when no change ever occurs, under the measure $\Pr_\infty$ under which all observations have density f_0; it measures the cost of false alarms, and a constraint $\mathbb{E}_\infty[\tau] \geq \gamma$ requires alarms to be rare in calm periods. The *detection delay* is the time between the change and the alarm. Since the change date ν is unknown and the observations before it can be arbitrarily unfavourable, Lorden (1971) measures the delay in the worst case over both: $\Pr_\nu$ denotes the measure under which the change occurs at ν, and the essential supremum is over the observations before ν.

Theorem 4 (Lorden, 1971; Moustakides, 1986)**.** *Let observations be independent with density f_0 before an unknown change date ν and f_1 from ν onward, and let $\ell_t = \log f_1(p_t)/f_0(p_t)$. Among all stopping times τ with average run length to false alarm $\mathbb{E}_\infty[\tau] \geq \gamma$, the CUSUM stopping time $\tau_b = \min\{t : W_t \geq b\}$, with the threshold b chosen so that $\mathbb{E}_\infty[\tau_b] = \gamma$, minimizes the worst-case conditional expected delay*

$$\bar{\mathbb{E}}_1[\tau] = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu\big[(\tau - \nu + 1)^+ \mid \mathcal{F}_{\nu-1}\big]. \tag{35}$$

Lorden proved that the CUSUM is asymptotically optimal as $\gamma \to \infty$; Moustakides proved exact optimality for every γ. Lai (1998) extends the asymptotic optimality to dependent observations, under conditions on the rate at which the log-likelihood ratios accumulate, which is the relevant case for our quarterly forecasts. The theorem is the reason for building the posterior on CUSUM statistics rather than on cumulative sums or moving averages: for the question "has this sensor changed regime?", no other rule detects faster at the same rate of false alarms.

The two modifications in Definition 7 trade this exact optimality for robustness. The clip c bounds the contribution of a single quarter: a penalized logit estimated on a small sample can produce a forecast of 0.999 or 0.001, whose log-odds exceed 6 in absolute value and would by themselves move the statistic across its whole range; clipping each increment to $[-c, c]$ with $c = 2$ requires evidence to accumulate over several quarters. The cap $W_{\max}$ bounds the lead that a sensor can accumulate, so that a sensor whose evidence has been conclusive for years does not retain its weight for years after its evidence has vanished, and so that any other sensor can take over the posterior within $W_{\max}/c = 3$ quarters of maximal evidence (Section 9.10). Both devices are standard in the control-chart literature, and both are varied over a grid in Section 7. The reset of the statistics at an onset follows the logic of repeated detection: once a crisis has started, the question is whether the next build-up has begun, and the evidence accumulated before the onset no longer bears on it.

9.9 The posterior over sensors

This subsection derives the softmax posterior equation 19 from a Bayesian model with a single affected sensor, and shows that it is the approximation of the exact posterior by the most likely change date of each sensor. Throughout, $\Lambda_{k,t}(\nu) = \prod_{s=\nu}^{t} f_{1,k}(p_{k,s})/f_{0,k}(p_{k,s})$ denotes the likelihood ratio of sensor k for a change at date $\nu \leq t$, with $\Lambda_{k,t}(t+1) = 1$ for no change.

Proposition 2 (Exact posterior over the affected sensor)**.** *Under Assumption A3, suppose that the observations of the sensors are independent across sensors and over time, that the affected sensor $k^* \in \{1, \ldots, K\}$ has a uniform prior, that the change date $\nu \in \{1, \ldots, t\}$ has a uniform prior, and condition on the event that a change has occurred by t. Then*

$$\Pr(k^* = k \mid \mathcal{F}_t) = \frac{R_{k,t}}{\sum_{j=1}^{K} R_{j,t}}, \qquad R_{k,t} = \sum_{\nu=1}^{t} \Lambda_{k,t}(\nu).$$

The quantity $R_{k,t}$ is the Shiryaev–Roberts statistic of sensor k.

Proof. Step 1: the likelihood of a scenario. A scenario is a pair (k, ν): sensor k is affected and its change occurs at ν. Under this scenario, the observations of every sensor $j \neq k$ have density

$f_{0,j}$ at every date, and those of sensor k have density $f_{0,k}$ before ν and $f_{1,k}$ from ν onward. By independence, the likelihood of all observations up to t is the product of the densities:

$$\mathcal{L}(k,\nu) = \prod_{j\neq k}\prod_{s=1}^{t} f_{0,j}(p_{j,s}) \times \prod_{s<\nu} f_{0,k}(p_{k,s}) \prod_{s=\nu}^{t} f_{1,k}(p_{k,s}).$$

Multiplying and dividing by $\prod_{s=\nu}^{t} f_{0,k}(p_{k,s})$,

$$\mathcal{L}(k,\nu) = \underbrace{\prod_{j=1}^{K}\prod_{s=1}^{t} f_{0,j}(p_{j,s})}_{=:\mathcal{L}_0} \times \prod_{s=\nu}^{t} \frac{f_{1,k}(p_{k,s})}{f_{0,k}(p_{k,s})} = \mathcal{L}_0\, \Lambda_{k,t}(\nu),$$

where $\mathcal{L}_0$ is the likelihood of the scenario in which no sensor has changed. In words: one starts from "everything normal" and, for the affected sensor and the dates from ν on only, replaces $f_{0,k}$ by $f_{1,k}$, which multiplies the likelihood by the corresponding ratio.

Step 2: Bayes' rule. The prior of the scenario (k,ν) is uniform, $1/(Kt)$, so the posterior is proportional to the likelihood: $\Pr(k^* = k, \nu \mid \mathcal{F}_t) \propto \mathcal{L}_0\, \Lambda_{k,t}(\nu)$.

Step 3: marginalizing over the change date. The event $\{k^* = k\}$ is the union over ν of the disjoint events $\{k^* = k, \nu\}$, so

$$\Pr(k^* = k \mid \mathcal{F}_t) \propto \sum_{\nu=1}^{t} \mathcal{L}_0\, \Lambda_{k,t}(\nu) = \mathcal{L}_0\, R_{k,t}.$$

Step 4: normalizing. The posterior probabilities of the K sensors sum to one. The factor $\mathcal{L}_0$ is common to all sensors and cancels in the normalization, which gives $\Pr(k^* = k \mid \mathcal{F}_t) = R_{k,t} / \sum_j R_{j,t}$. □

From the exact posterior to the softmax of the CUSUMs. The exact posterior requires the Shiryaev–Roberts statistics, sums over all possible change dates. The CUSUM retains only the largest term of each sum. The following lemma shows that the two differ by at most a factor that is common to all sensors.

Lemma 8 (CUSUM and Shiryaev–Roberts)**.** *Let $W_{k,t}$ be the CUSUM of sensor k built on the increments $\ell_{k,s} = \log f_{1,k}(p_{k,s})/f_{0,k}(p_{k,s})$, without clip or cap. Then*

$$W_{k,t} \;\leq\; \log\big(1 + R_{k,t}\big) \;\leq\; W_{k,t} + \log(t+1).$$

Proof. With the convention $\Lambda_{k,t}(t+1) = 1$, the sum $1 + R_{k,t} = \sum_{\nu=1}^{t+1} \Lambda_{k,t}(\nu)$ has $t+1$ positive terms, and $\log \Lambda_{k,t}(\nu) = S_t(\nu)$ in the notation of Lemma 7. By that lemma, the largest term is $\max_\nu \Lambda_{k,t}(\nu) = e^{W_{k,t}}$. A sum of $t+1$ positive terms is at least its largest term and at most $t+1$ times it:

$$e^{W_{k,t}} \;\leq\; 1 + R_{k,t} \;\leq\; (t+1)\, e^{W_{k,t}}.$$

Taking logarithms gives the claim. □

The term $1 + R_{k,t}$, rather than $R_{k,t}$, is the Shiryaev–Roberts sum extended by the hypothesis that sensor k has not changed; it is the natural object when one does not condition on a change having occurred. By the lemma, $\log(1 + R_{k,t})$ and $W_{k,t}$ differ by at most $\log(t+1)$, a quantity that does not depend on k. Replacing $1 + R_{k,t}$ by $e^{W_{k,t}}$ in the posterior of Proposition 2 gives

$$\pi_{k,t} = \frac{e^{\zeta W_{k,t}}}{\sum_{j=1}^{K} e^{\zeta W_{j,t}}}, \qquad \zeta = 1,$$

which is equation 19: the posterior computed by keeping, for each sensor, only its most likely change date. This is a Laplace approximation of each sum by its largest term, and the error it

introduces is controlled by the lemma: since the omitted factor lies in $[1, t+1]$ for every sensor, the ratio of two posterior weights, $\pi_{k,t}/\pi_{j,t} = e^{W_{k,t}-W_{j,t}}$, is within a factor $t+1$ of the exact ratio $(1+R_{k,t})/(1+R_{j,t})$ in either direction. In practice the gap is much smaller than this worst case: when a sensor has changed, the terms of its sum for change dates far from the true one are exponentially smaller than the largest term, and the sum is dominated by a few dates around it.

The role of the temperature. The parameter ζ interpolates between two rules. As $\zeta \to \infty$, the softmax puts all the weight on the sensor with the largest CUSUM: this is the max rule of multi-sensor detection (Mei, 2010), which retains a single channel and ignores the others. As $\zeta \to 0$, the weights become uniform and the statistics play no role. The value $\zeta = 1$ is the one for which the softmax is the approximate Bayesian posterior, with the CUSUM read as a log-likelihood ratio; values of ζ different from one correspond to tempering the likelihood, and Section 7 reports the sensitivity of the results to this choice.

Lemma 9 (The posterior never vanishes). *For every sensor k and every date t, $\pi_{k,t} \geq e^{-\zeta W_{\max}}/K$.*

Proof. The numerator of $\pi_{k,t}$ is $e^{\zeta W_{k,t}} \geq e^0 = 1$, since $W_{k,t} \geq 0$. In the denominator, sensor k contributes at least 1 and each of the other $K-1$ sensors at most $e^{\zeta W_{\max}}$, since $W_{j,t} \leq W_{\max}$; but for an upper bound on the denominator we may bound every term by $e^{\zeta W_{\max}}$, which gives $\sum_j e^{\zeta W_{j,t}} \leq K e^{\zeta W_{\max}}$. Hence

$$\pi_{k,t} \;\geq\; \frac{1}{K\, e^{\zeta W_{\max}}} = \frac{e^{-\zeta W_{\max}}}{K}. \qquad \square$$

Numerically, with $\zeta = 1$, $W_{\max} = 6$ and $K = 4$, the bound is $e^{-6}/4 \approx 6 \times 10^{-4}$: no channel is ever eliminated by the posterior, and a channel whose statistic starts rising recovers weight immediately, but the gap between a channel at the cap and a channel at zero is a factor $e^6 \approx 400$. This lower bound is what prevents a regret bound for the detection rule itself: the posterior acts as a multiplicative perturbation of the exponential weights, bounded below only by $e^{-\zeta W_{\max}}/K$, and carrying this constant through the argument of Theorem 1 yields a bound larger than the losses at stake, which is why the guarantee is obtained for the meta-aggregator instead (Section 5.6).

9.10 Simultaneous changes in several sensors

Assumption A3 states that a single sensor changes distribution at the beginning of a pre-crisis period. This is restrictive: a housing boom is often financed by credit or by capital inflows, so that several channels may become active at the same time. This subsection describes how the posterior equation 19 behaves when several sensors change, explains the role of the cap $W_{\max}$, and discusses the interpretation of $\pi_{k,t}$. Throughout, $K = 4$, $\zeta = 1$, $c = 2$ and $W_{\max} = 6$, as in the baseline.

9.10.1 Several sensors at the cap

Recall that

$$\pi_{k,t} = \frac{e^{\zeta W_{k,t}}}{\sum_{j=1}^{K} e^{\zeta W_{j,t}}}. \tag{36}$$

Two sensors at the cap. Suppose that sensors 1 and 2 have changed distribution and that their statistics have reached the cap, while sensors 3 and 4 are at zero:

$$(W_{1,t}, W_{2,t}, W_{3,t}, W_{4,t}) = (6, 6, 0, 0).$$

The numerator of equation 36 for sensor 1 is e^6. The denominator sums the four terms, $e^6 + e^6 + e^0 + e^0 = 2e^6 + 2$, since $e^0 = 1$. Hence

$$\pi_{1,t} = \pi_{2,t} = \frac{e^6}{2e^6 + 2} = \frac{1}{2 + 2e^{-6}}, \qquad \pi_{3,t} = \pi_{4,t} = \frac{1}{2e^6 + 2}, \tag{37}$$

where the second expression for $\pi_{1,t}$ divides the numerator and the denominator by e^6. Numerically, $e^6 \approx 403.4$, so that the denominator is $2 \times 403.4 + 2 = 808.9$ and

$$\pi_{1,t} = \pi_{2,t} \approx \frac{403.4}{808.9} \approx 0.499, \qquad \pi_{3,t} = \pi_{4,t} \approx \frac{1}{808.9} \approx 0.001.$$

The four weights sum to one. The second form in equation 37 shows why the active sensors receive almost exactly one half: $2e^{-6} \approx 0.005$ is negligible relative to 2. The inactive sensors are weighted $e^6 \approx 400$ times less than the active ones.

General case. With m sensors at the cap and $K - m$ at zero, the same computation gives, for each active sensor,

$$\pi_{k,t} = \frac{e^{\zeta W_{\max}}}{m\, e^{\zeta W_{\max}} + (K - m)} = \frac{1}{m + (K - m)\, e^{-\zeta W_{\max}}}. \tag{38}$$

Writing $\varepsilon = (K - m)\, e^{-\zeta W_{\max}}$, the distance to $1/m$ satisfies

$$0 \;\le\; \frac{1}{m} - \pi_{k,t} \;=\; \frac{1}{m} - \frac{1}{m + \varepsilon} \;=\; \frac{\varepsilon}{m(m + \varepsilon)} \;\le\; \frac{\varepsilon}{m^2} = \frac{(K - m)\, e^{-\zeta W_{\max}}}{m^2}. \tag{39}$$

For $m = 2$, the bound is $2e^{-6}/4 \approx 0.0012$, which matches the gap between 0.499 and 0.5. For $m = 3$, each active sensor receives $\pi_{k,t} \approx 0.333$. The softmax therefore splits the weight equally among the sensors at the cap.

9.10.2 Before saturation: the posterior depends on differences

When the active sensors are not all at the cap, the posterior depends only on the differences between statistics. For any two sensors k and j, 36 gives, since the denominators are equal,

$$\frac{\pi_{k,t}}{\pi_{j,t}} = e^{\zeta(W_{k,t} - W_{j,t})}. \tag{40}$$

With $(W_{1,t}, W_{2,t}, W_{3,t}, W_{4,t}) = (6, 4, 0, 0)$, the denominator is $e^6 + e^4 + 2 \approx 403.4 + 54.6 + 2 = 460.0$, so that $\pi_{1,t} \approx 0.877$, $\pi_{2,t} \approx 0.119$ and $\pi_{3,t} = \pi_{4,t} \approx 0.002$. The ratio $\pi_{1,t}/\pi_{2,t} = e^2 \approx 7.4$ reflects a lead of two units, that is, one quarter of maximal evidence.

9.10.3 The role of the cap

Equation 40 implies that, without a cap, a sensor that starts rising first keeps its lead indefinitely. Consider the following sequence. Sensor 1 changes distribution at $t = 1$ and sensor 2 at $t = 4$; both receive the maximal increment $c = 2$ in every quarter after their change, while sensors 3 and 4 receive negative increments and stay at zero. Table 10 reports the statistics and the posterior with and without the cap.

Without the cap, the difference $W_{1,t} - W_{2,t}$ remains equal to 6 from $t = 4$ onward, since both statistics grow by 2 per quarter. By equation 40, sensor 2 then receives about $e^{-6} \approx 0.0025$ of the weight of sensor 1, forever, although both sensors carry the same evidence in every quarter. With the cap, sensor 1 stops at 6 while sensor 2 catches up: a sensor at zero that receives the maximal increment reaches the cap in $\lceil W_{\max}/c \rceil = 3$ quarters, and the two sensors share the weight equally from $t = 6$ onward. The cap is therefore what makes the split symmetric: it prevents the timing of the first signal from determining the attribution.

9.10.4 Consequence for the aggregate forecast

Ignoring the prior, learning and reliability factors, the aggregate forecast before the gate is $\sum_k \pi_{k,t}\, p_{k,t}$. Suppose that the two active predictive models forecast $p_{1,t} = 0.7$ and $p_{2,t} = 0.6$, and the two

Table 10: Posterior over sensors with and without the cap

	With cap ($W_{\max} = 6$)				Without cap			
t	$W_{1,t}$	$W_{2,t}$	$\pi_{1,t}$	$\pi_{2,t}$	$W_{1,t}$	$W_{2,t}$	$\pi_{1,t}$	$\pi_{2,t}$
1	2	0	0.711	0.096	2	0	0.711	0.096
2	4	0	0.948	0.017	4	0	0.948	0.017
3	6	0	0.993	0.002	6	0	0.993	0.002
4	6	2	0.977	0.018	8	2	0.997	0.002
5	6	4	0.877	0.119	10	4	0.997	0.002
6	6	6	0.499	0.499	12	6	0.998	0.002
7	6	6	0.499	0.499	14	8	0.998	0.002

Notes: $K = 4$, $\zeta = 1$, $c = 2$. Sensor 1 changes at $t = 1$, sensor 2 at $t = 4$; both receive the increment c in every subsequent quarter. Sensors 3 and 4 stay at zero and receive the remaining weight.

inactive predictive models $p_{3,t} = p_{4,t} = 0.1$. With the weights of equation 37,

$$\sum_k \pi_{k,t}\, p_{k,t} = 0.499 \times 0.7 + 0.499 \times 0.6 + 0.001 \times 0.1 + 0.001 \times 0.1 \approx 0.649.$$

If only sensor 1 were at the cap, $(W_{k,t}) = (6, 0, 0, 0)$, the weights would be $(0.993, 0.002, 0.002, 0.002)$ and the forecast 0.697. In both cases, the aggregate forecast is a weighted average of the forecasts of the active predictive models: detection is not weakened when several channels are active. The aggregation lies between the max rule, which retains only the largest statistic, and the sum rule, which is more powerful when several sensors are affected (Mei, 2010; Xie and Siegmund, 2013; Fellouris and Sokolov, 2016).

9.10.5 Interpretation of the posterior

The softmax imposes $\sum_k \pi_{k,t} = 1$. It answers the question "which channel?" and cannot answer "these two channels". Under Assumption A3, $\pi_{k,t}$ is the probability that channel k is the source of the vulnerability. When several sensors change, it should instead be read as the share of the evidence attributable to channel k: two channels at 0.5 each are two channels carrying the same evidence, not two channels that are each the source with probability one half.

A model that allows for several affected sensors gives a posterior for each sensor separately. Suppose that each sensor is affected independently with probability p_0, and that sensors are independent. The likelihood of a subset A of affected sensors is then proportional to $\prod_{k \in A} p_0 \, \mathrm{LR}_{k,t} \prod_{k \notin A} (1 - p_0)$, where $\mathrm{LR}_{k,t}$ is the likelihood ratio of sensor k. This product factorizes across sensors, so that the posterior odds that sensor k is affected are $\frac{p_0}{1-p_0}\,\mathrm{LR}_{k,t}$, independently of the other sensors. Replacing $\mathrm{LR}_{k,t}$ by $e^{W_{k,t}}$, as in the main text, gives

$$\Pr\left(k \text{ affected} \mid \mathcal{F}_t\right) \approx \Lambda\Big(W_{k,t} + \log \frac{p_0}{1 - p_0}\Big), \tag{41}$$

where Λ is the logistic function. These probabilities do not sum to one. With $p_0 = 1/K = 0.25$, a sensor at the cap has probability $\Lambda(6 - \log 3) \approx 0.993$ of being affected, and a sensor at zero $\Lambda(-\log 3) = 0.25$, its prior. In the example of equation 37, sensors 1 and 2 are each affected with probability 0.993, which expresses directly that both channels are active.

The United Kingdom before 2007 is an instance of this situation: the real-estate statistic is at the cap from 2003 and the external statistic joins it in 2004Q4, after which the weight is shared between the two channels, consistent with a housing boom financed by capital inflows (Section 6.2).

9.11 Results

9.11.1 Baseline specification

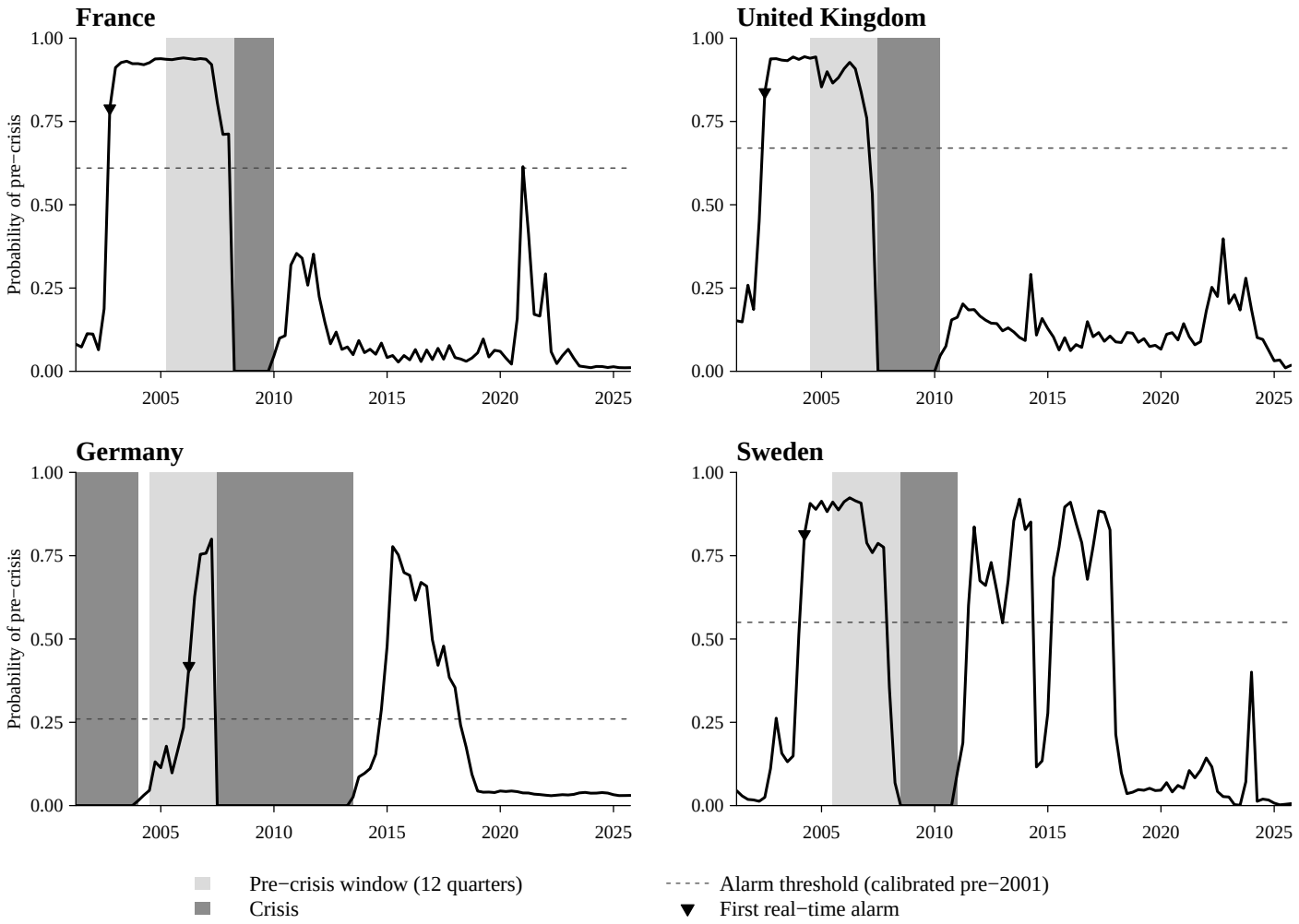


Figure 6: Probability of pre-crisis period - meta aggregator

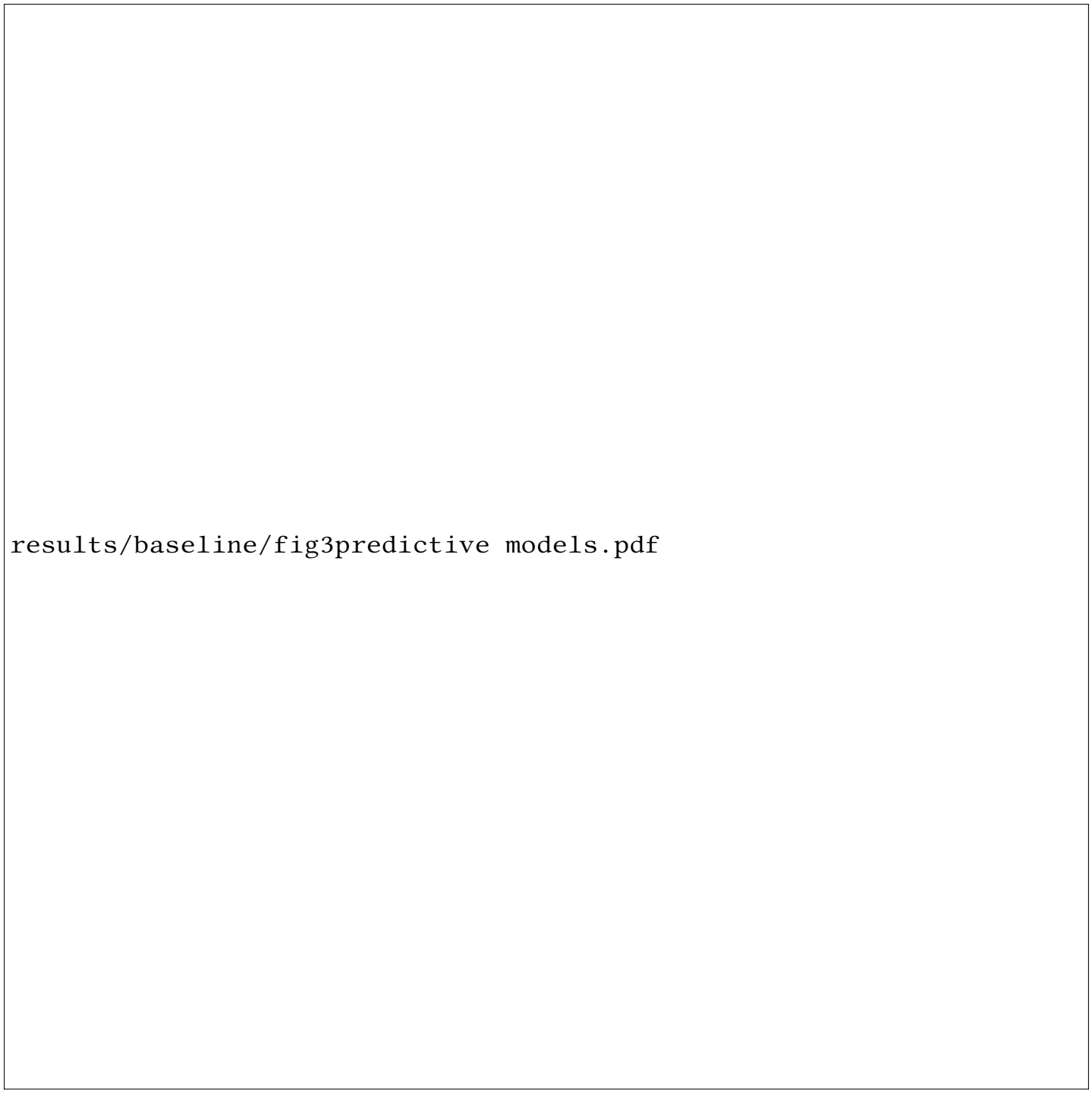


Figure 7: Individual forecasting models

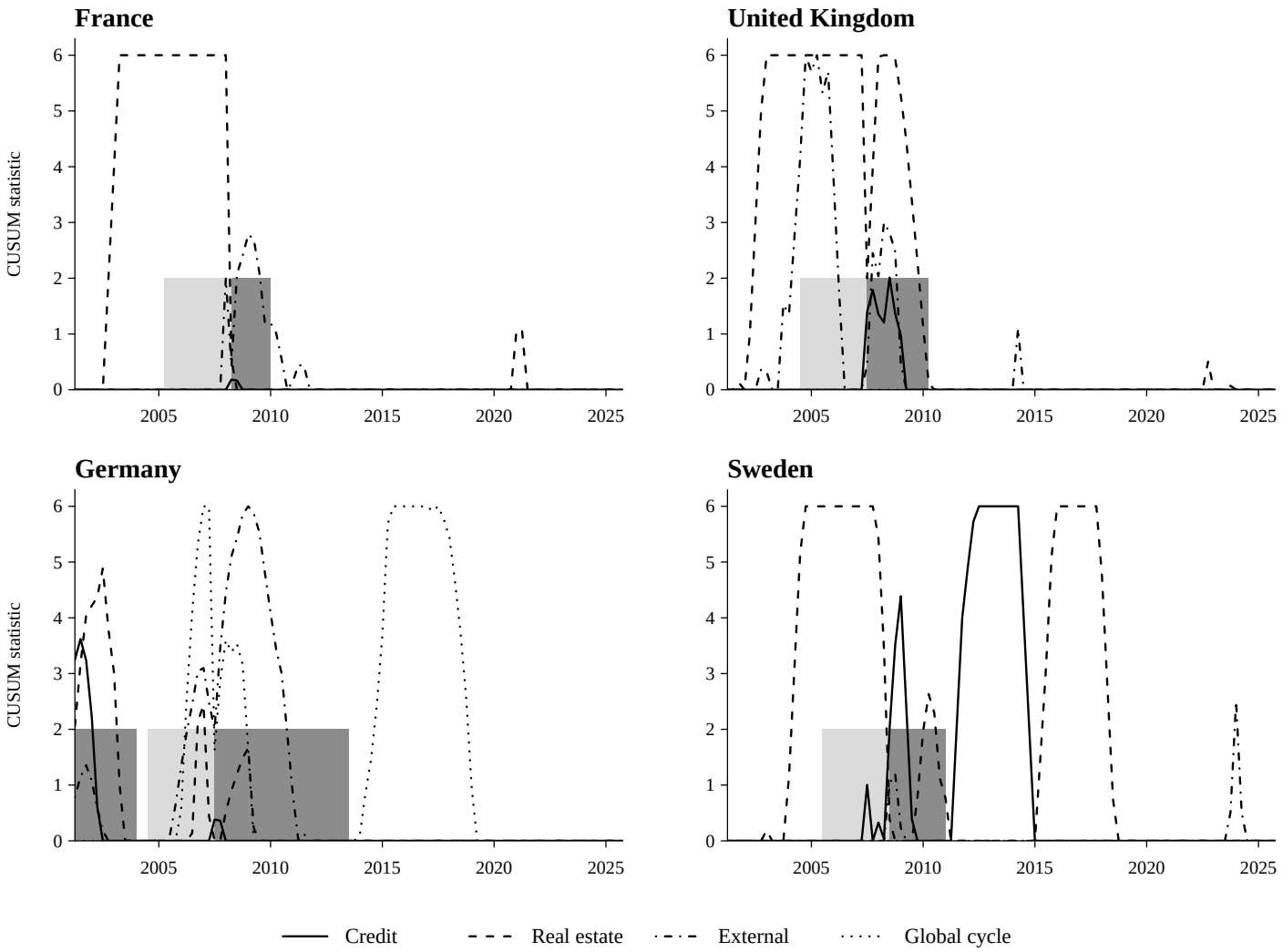
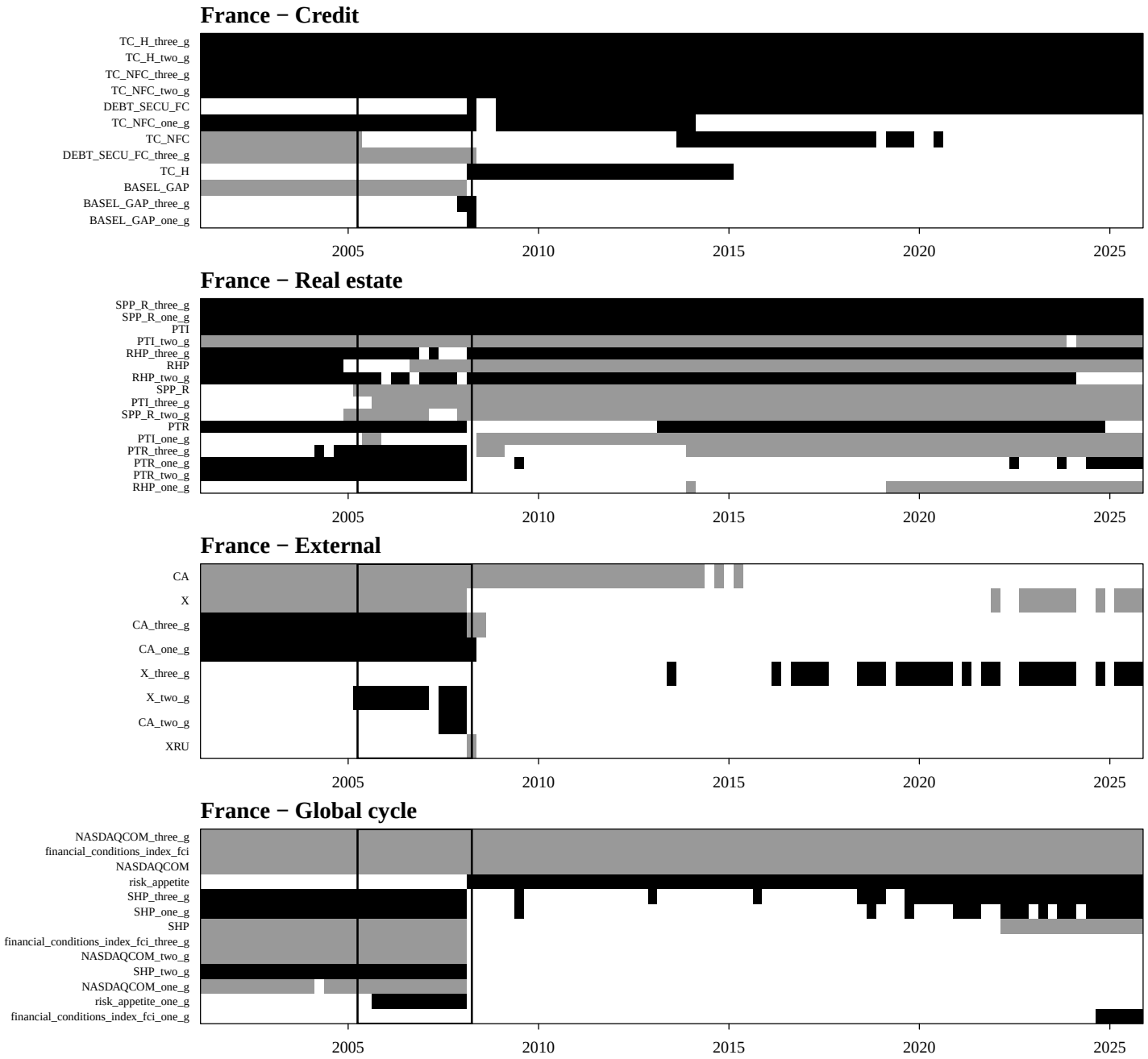
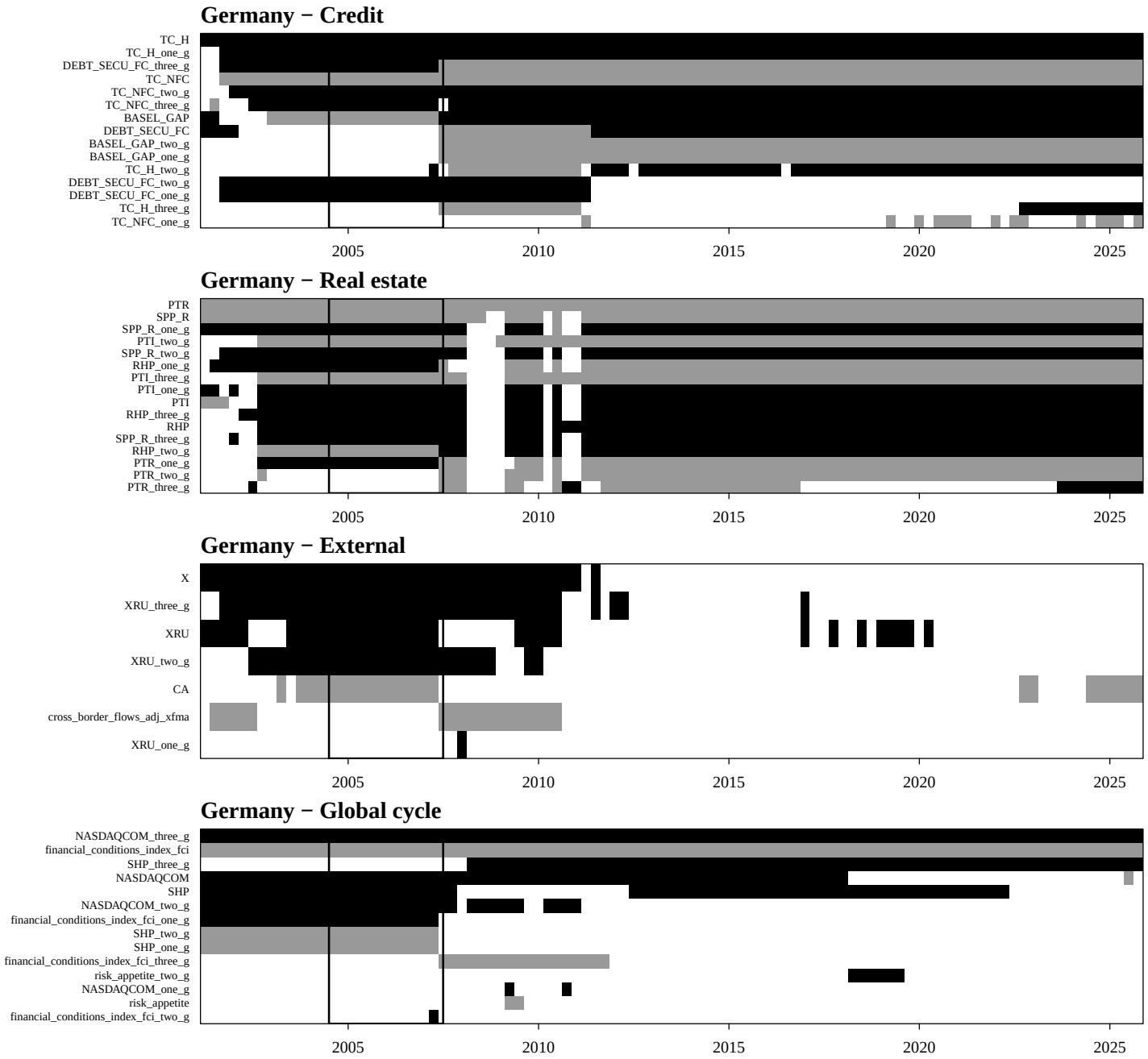


Figure 8: CUMSUM statistics



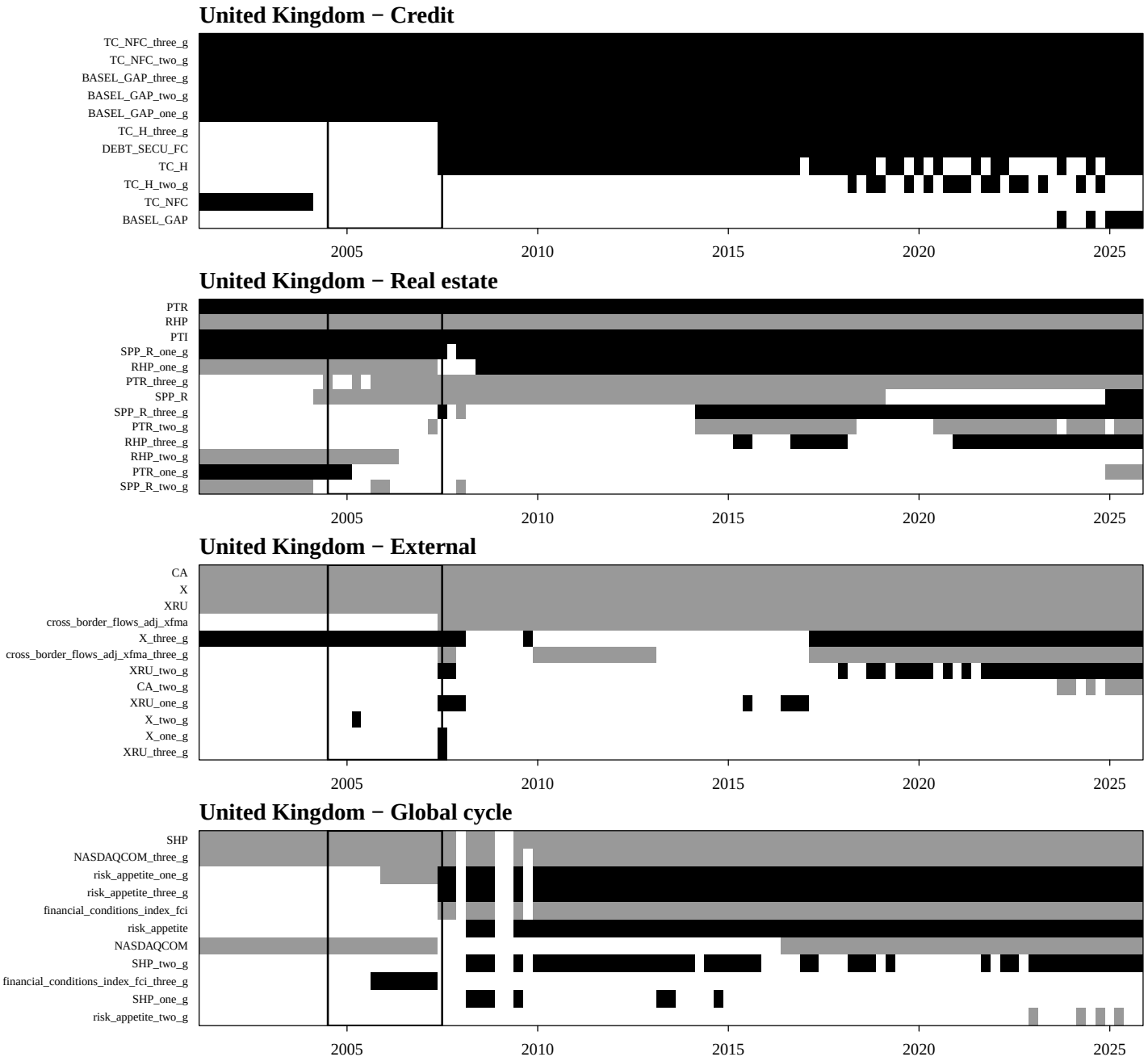
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 9: Coefficients- France



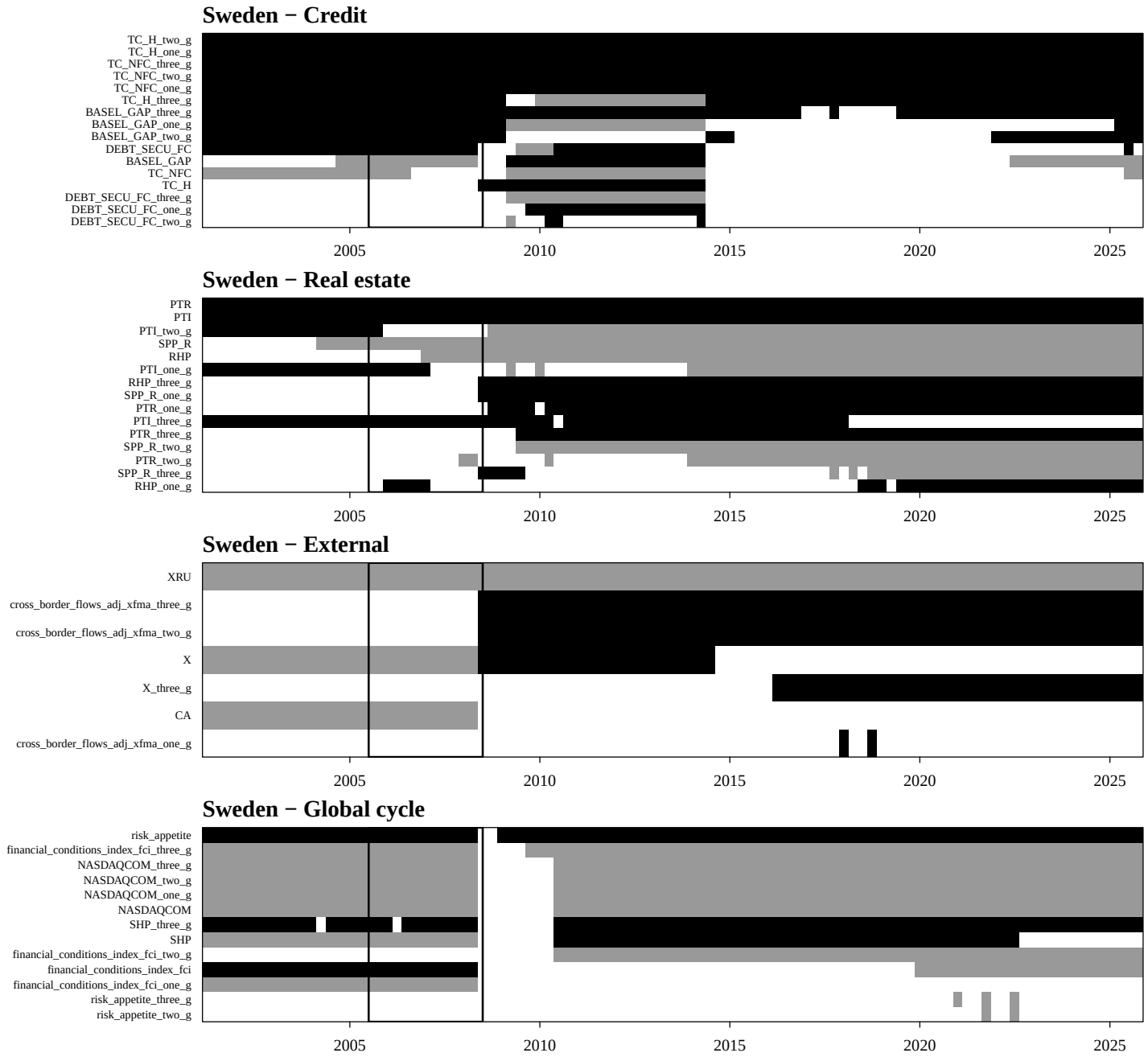
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 10: Coefficients- Germany



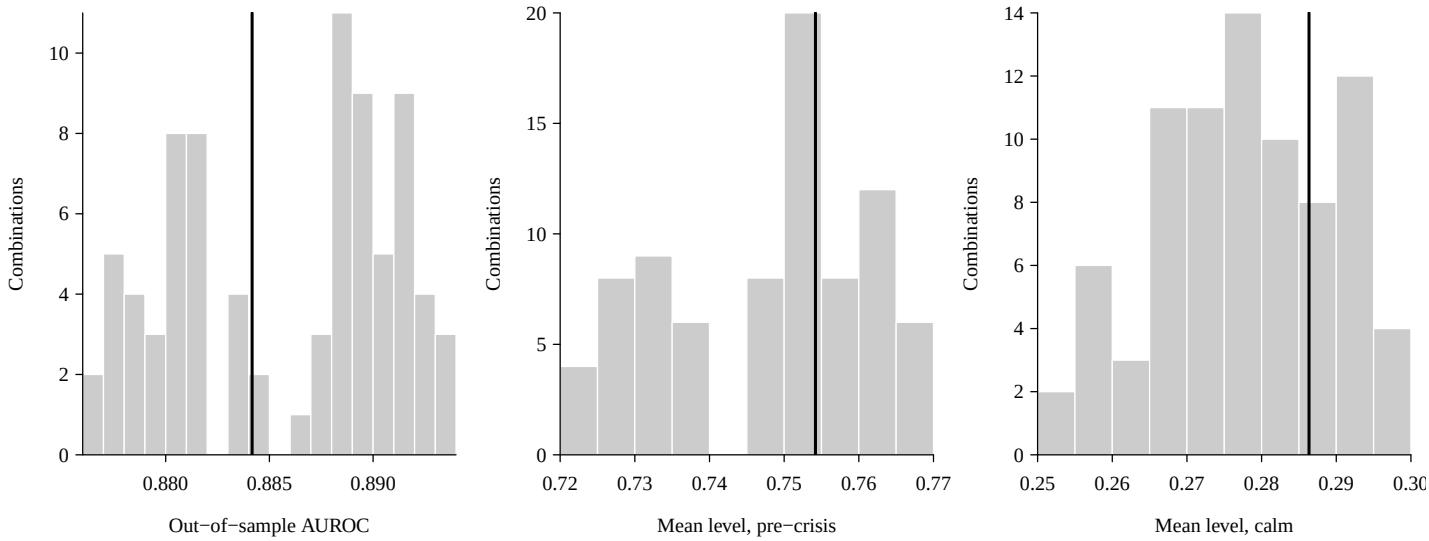
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 11: Coefficients- United Kingdom



Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 12: Coefficients- Sweden



Histograms over the 81 combinations of (c, Wmax, kappa, fade). Vertical line: baseline specification.

Figure 13:

Table 11: Detection rule versus meta-aggregator

country	AUC_online_CPD	AUC_online_META	pre_CPD	pre_META	calme_CPD	calme_META	U_o
France	0.96	0.95	0.94	0.89	0.24	0.23	
United Kingdom	0.90	0.92	0.91	0.85	0.24	0.24	
Germany	0.84	0.85	0.37	0.36	0.24	0.24	
Sweden	0.83	0.82	0.80	0.75	0.42	0.40	

Table 12: Batch utility by expert and country, and pooled prior

Expert	Prior weight	Lcr	Lho	Lext	Lgfc
Credit	0.426	0.886	0.886	0.795	0.455
Real estate	0.413	0.977	0.977	0.909	0.455
External	0.124	0.977	0.977	0.364	0.000
Global cycle	0.037	0.864	0.841	0.705	0.455

Utility is the maximal Alessi–Detken utility of each expert on the batch period of each country. The prior is a softmax of the cross-country mean with temperature 6 and a floor of 0.10.

Table 13: Real-time re-estimations, Credit expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
TC_NFC_two_g	98.25	0.34	1.00	1.00
TC_H_one_g	98.00	0.42	1.00	1.00
TC_NFC_three_g	96.75	0.50	1.00	0.99
BASEL_GAP_two_g	78.67	0.20	0.67	1.00
TC_H_three_g	77.00	0.51	0.83	0.76
TC_H_two_g	71.00	0.27	0.95	0.90
DEBT_SECU_FC	67.75	0.12	0.92	0.83
BASEL_GAP_one_g	57.00	-0.00	0.69	0.87
TC_NFC_one_g	55.67	0.10	0.79	0.75
TC_H	51.00	2.17	1.00	1.00
BASEL_GAP_three_g	49.00	0.25	0.75	1.00
TC_NFC	44.25	-1.36	0.43	0.85
BASEL_GAP	41.25	0.20	0.55	0.80
DEBT_SECU_FC_three_g	34.00	0.09	0.29	0.91
DEBT_SECU_FC_two_g	22.00	0.28	0.90	0.80
DEBT_SECU_FC_one_g	17.67	0.10	0.67	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 14: Real-time re-estimations, Real estate expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
PTI	95.75	3.21	0.99	0.98
PTR	95.50	-1.41	0.75	1.00
PTI_two_g	90.67	-0.57	0.07	0.86
RHP	89.50	0.58	0.34	0.82
SPP_R_one_g	87.25	0.37	1.00	0.99
SPP_R	83.75	-1.30	0.05	0.90
PTI_one_g	81.67	-0.07	0.43	0.81
PTI_three_g	78.67	-0.42	0.33	1.00
RHP_three_g	69.25	1.12	1.00	1.00
SPP_R_three_g	68.75	0.70	0.78	0.92
PTR_three_g	65.50	-0.04	0.42	0.66
SPP_R_two_g	61.00	-0.22	0.25	1.00
RHP_one_g	60.50	-0.03	0.49	0.76
RHP_two_g	59.00	0.21	0.62	0.91
PTR_two_g	51.50	-0.06	0.26	0.99
PTR_one_g	49.25	0.37	0.68	0.73

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 15: Real-time re-estimations, External expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
cross_border_flows_adj_xfma_two_g	70.00	0.22	1.00	1.00
XRU	58.00	-0.30	0.25	1.00
X	56.00	-0.09	0.34	0.82
cross_border_flows_adj_xfma_three_g	54.50	0.12	0.50	1.00
CA	48.75	-0.56	0.00	1.00
X_three_g	44.67	0.08	1.00	1.00
CA_three_g	30.00	0.15	0.93	0.87
CA_one_g	29.00	0.13	1.00	1.00
XRU_two_g	28.00	0.06	1.00	1.00
cross_border_flows_adj_xfma	24.50	-0.06	0.51	0.99
XRU_three_g	21.00	0.22	1.00	1.00
X_two_g	6.50	0.01	1.00	1.00
CA_two_g	6.00	0.01	0.50	1.00
XRU_one_g	4.50	0.02	1.00	1.00
cross_border_flows_adj_xfma_one_g	2.00	0.00	1.00	1.00
X_one_g	1.00	0.01	1.00	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 16: Real-time re-estimations, Global cycle expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
NASDAQCOM_three_g	96.00	-0.12	0.25	1.00
NASDAQCOM	79.50	-0.53	0.25	0.99
financial_conditions_index_fci	78.00	-0.40	0.11	0.77
SHP_three_g	72.33	0.20	1.00	1.00
SHP	68.00	0.15	0.42	0.85
risk_appetite	59.75	0.39	0.75	1.00
NASDAQCOM_two_g	49.67	-0.08	0.33	1.00
risk_appetite_one_g	43.00	0.05	0.96	0.92
NASDAQCOM_one_g	39.67	-0.08	0.33	1.00
risk_appetite_three_g	36.50	0.09	0.50	1.00
financial_conditions_index_fci_three_g	34.75	-0.15	0.25	1.00
financial_conditions_index_fci_two_g	32.00	-0.04	0.50	1.00
SHP_two_g	30.33	-0.04	0.67	1.00
SHP_one_g	24.00	0.02	0.67	1.00
financial_conditions_index_fci_one_g	16.67	-0.07	0.67	1.00
risk_appetite_two_g	7.00	0.09	0.52	0.70

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 17: Realized regret of the meta-aggregator against each candidate and theoretical bound

Country	Candidate	Prior	Loss cand.	Loss meta	Regret	Bound	Holds
France	Detection rule	0.90	2.52	2.63	0.11	3.03	yes
France	Credit	0.03	17.57	2.63	-14.94	106.24	yes
France	Real estate	0.03	2.47	2.63	0.15	106.24	yes
France	External	0.03	16.81	2.63	-14.18	106.24	yes
France	Global cycle	0.03	21.71	2.63	-19.08	106.24	yes
United Kingdom	Detection rule	0.90	3.65	3.85	0.20	3.03	yes
United Kingdom	Credit	0.03	16.18	3.85	-12.33	106.24	yes
United Kingdom	Real estate	0.03	6.64	3.85	-2.79	106.24	yes
United Kingdom	External	0.03	12.36	3.85	-8.51	106.24	yes
United Kingdom	Global cycle	0.03	21.35	3.85	-17.50	106.24	yes
Germany	Detection rule	0.90	12.10	12.18	0.09	5.43	yes
Germany	Credit	0.03	16.06	12.18	-3.87	108.64	yes
Germany	Real estate	0.03	13.10	12.18	-0.92	108.64	yes
Germany	External	0.03	13.20	12.18	-1.02	108.64	yes
Germany	Global cycle	0.03	17.22	12.18	-5.03	108.64	yes
Sweden	Detection rule	0.90	5.58	5.47	-0.10	3.03	yes
Sweden	Credit	0.03	13.97	5.47	-8.49	106.24	yes
Sweden	Real estate	0.03	4.23	5.47	1.24	106.24	yes
Sweden	External	0.03	18.94	5.47	-13.47	106.24	yes
Sweden	Global cycle	0.03	21.52	5.47	-16.05	106.24	yes

Losses are cumulated weighted square losses on real-time non-crisis quarters with revealed labels. Bound from Proposition 2.

9.11.2 Horizon $h = 8$

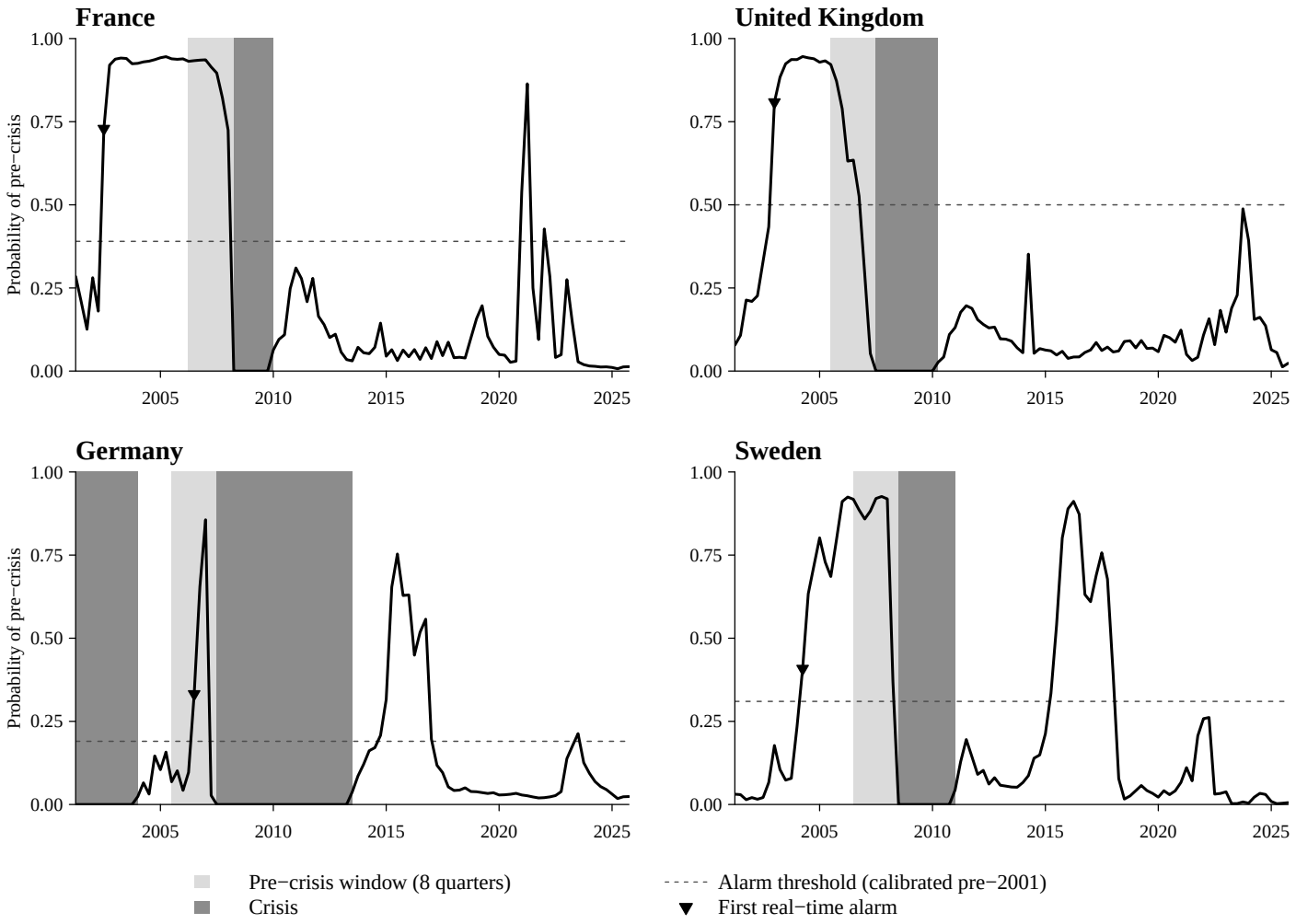


Figure 14: Probability of pre-crisis period - meta aggregator

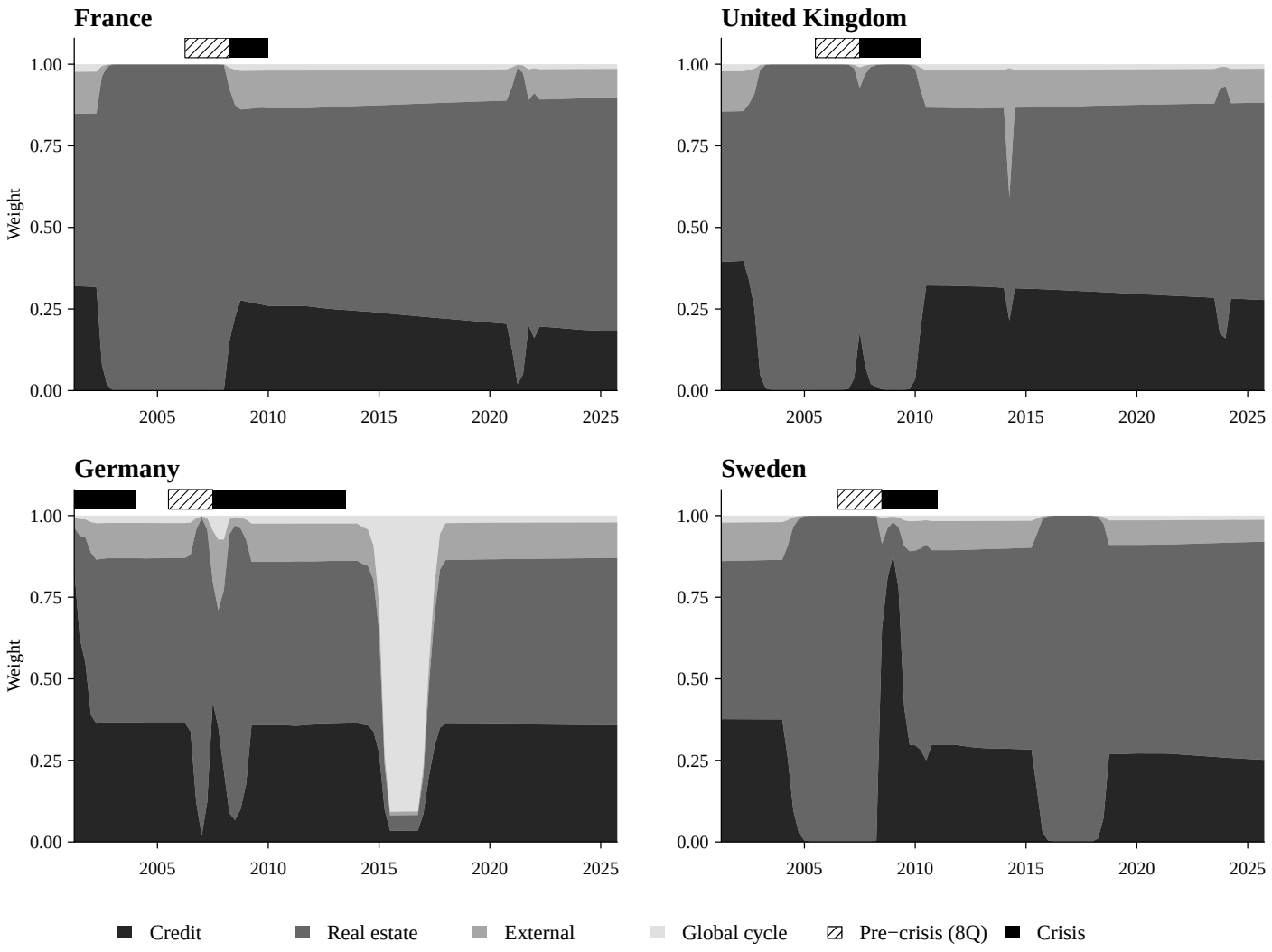


Figure 15: Weights

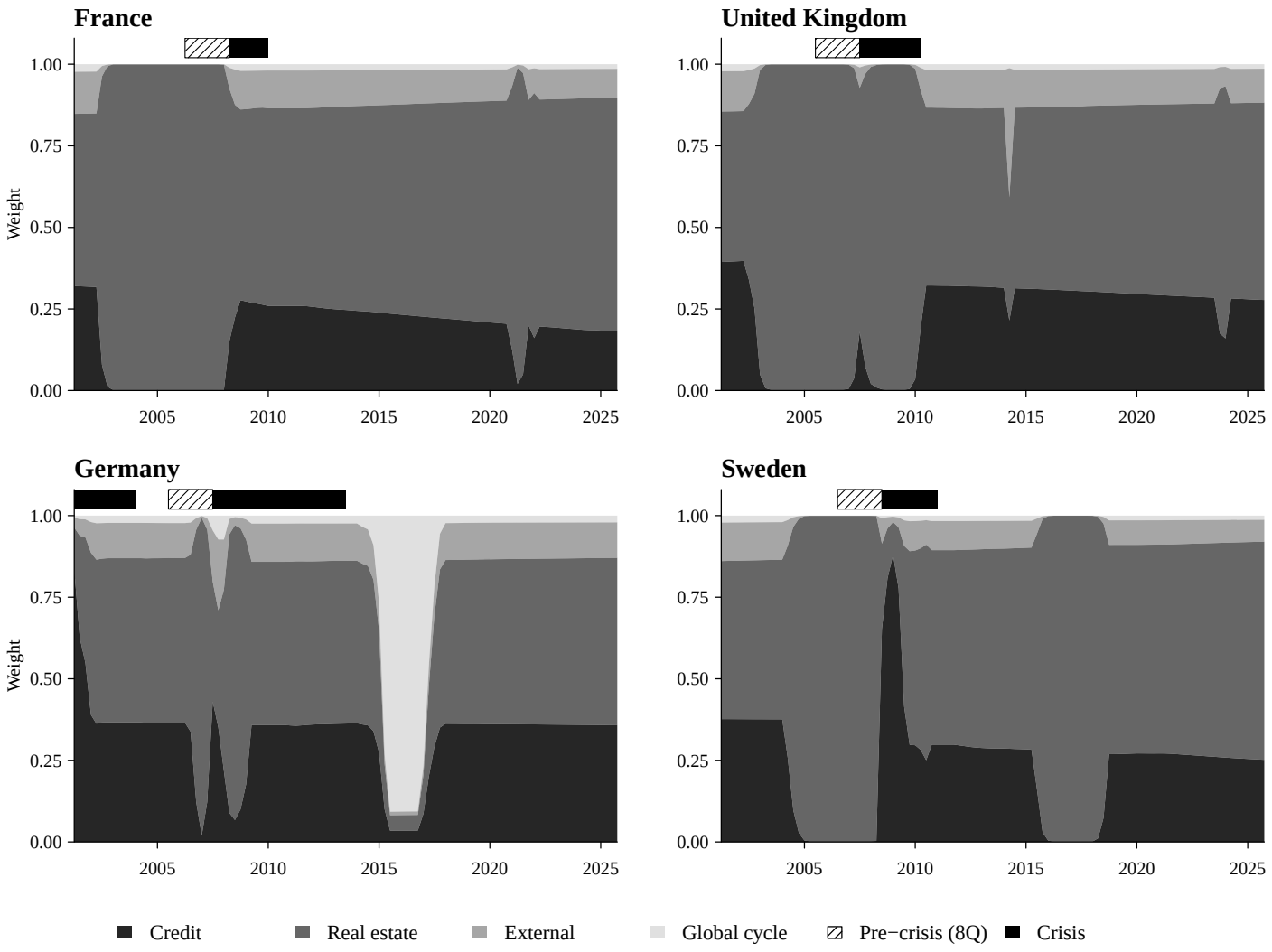


Figure 16: Individual forecasting models

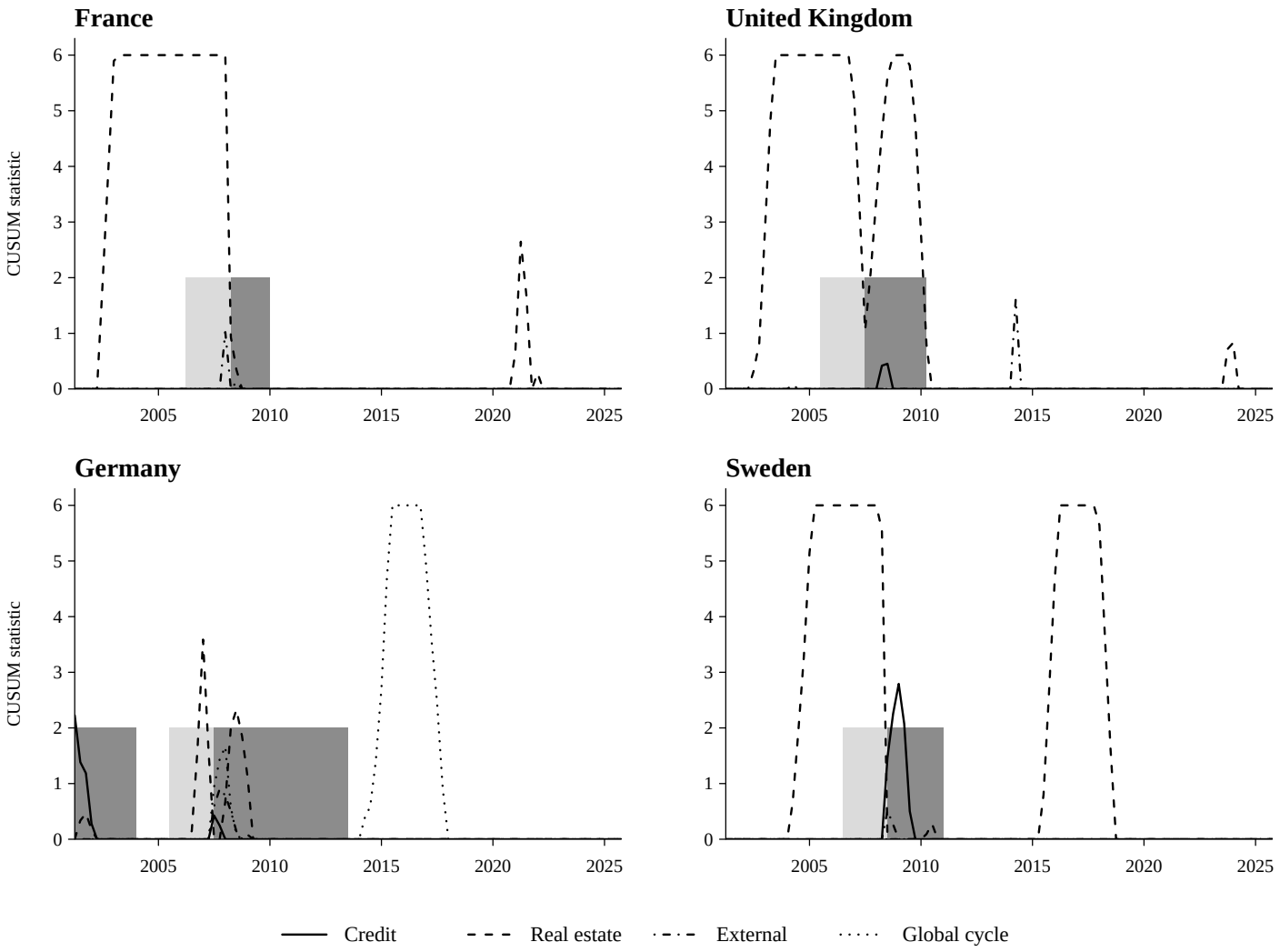
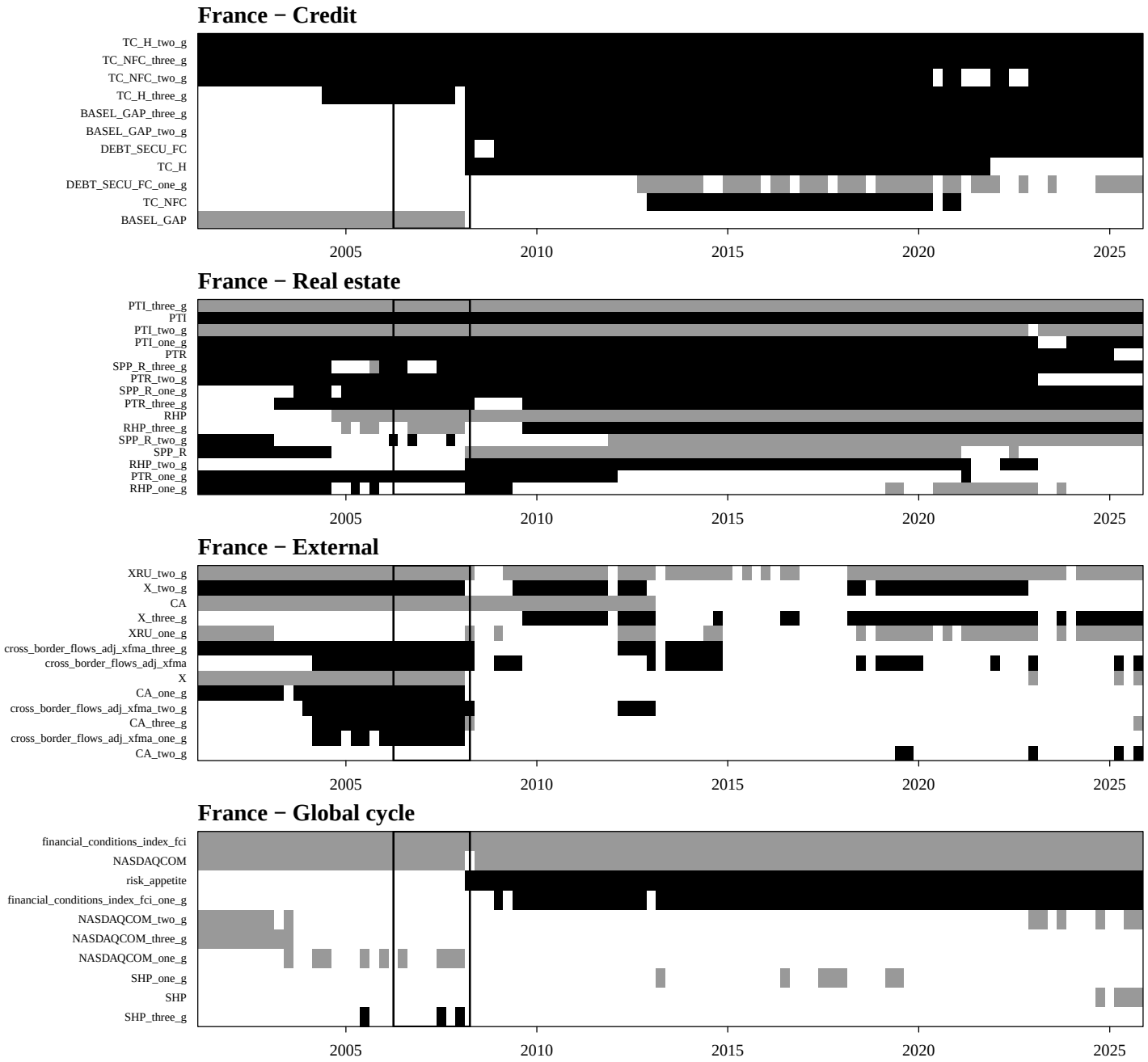
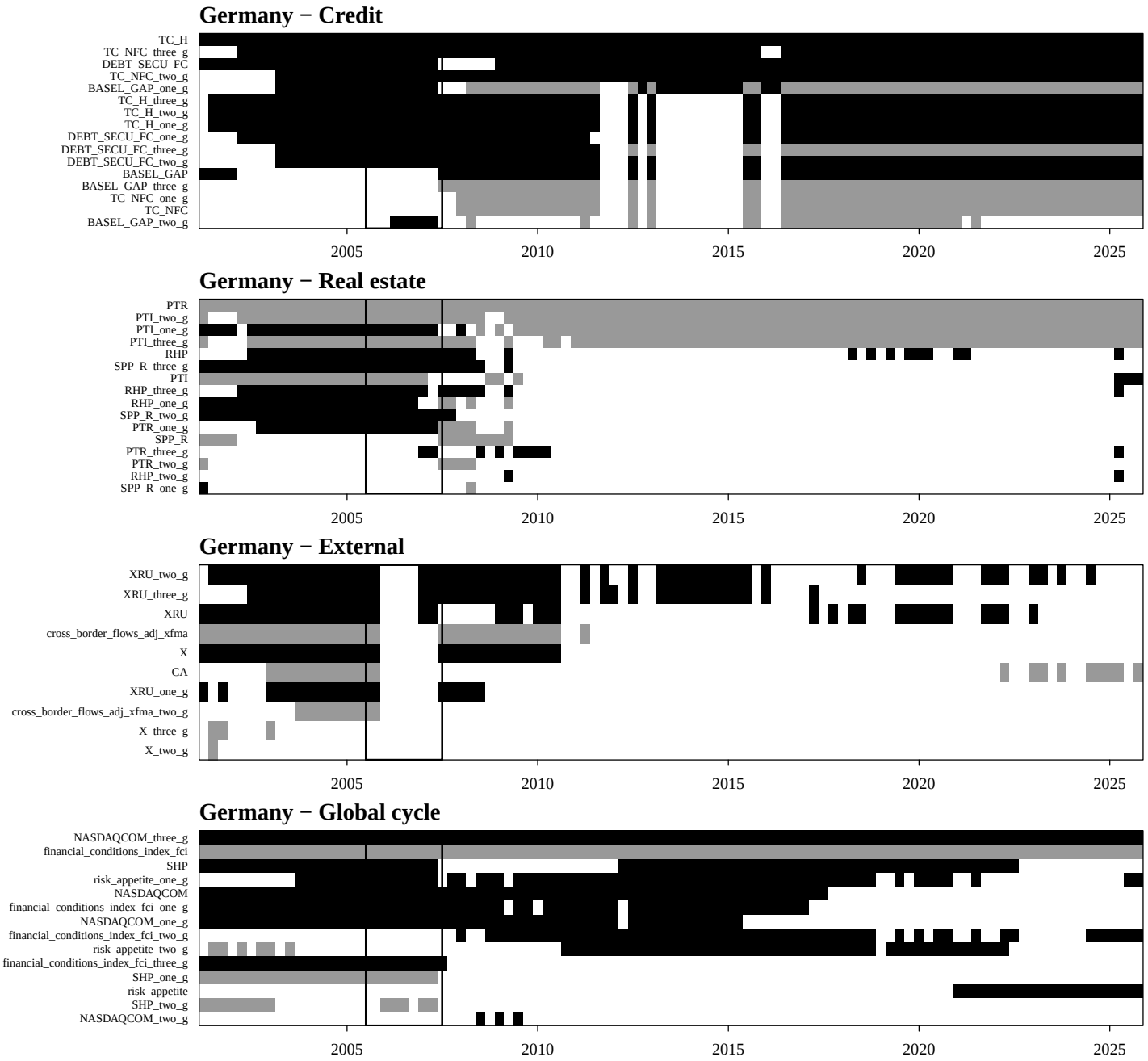


Figure 17: CUMSUM statistics



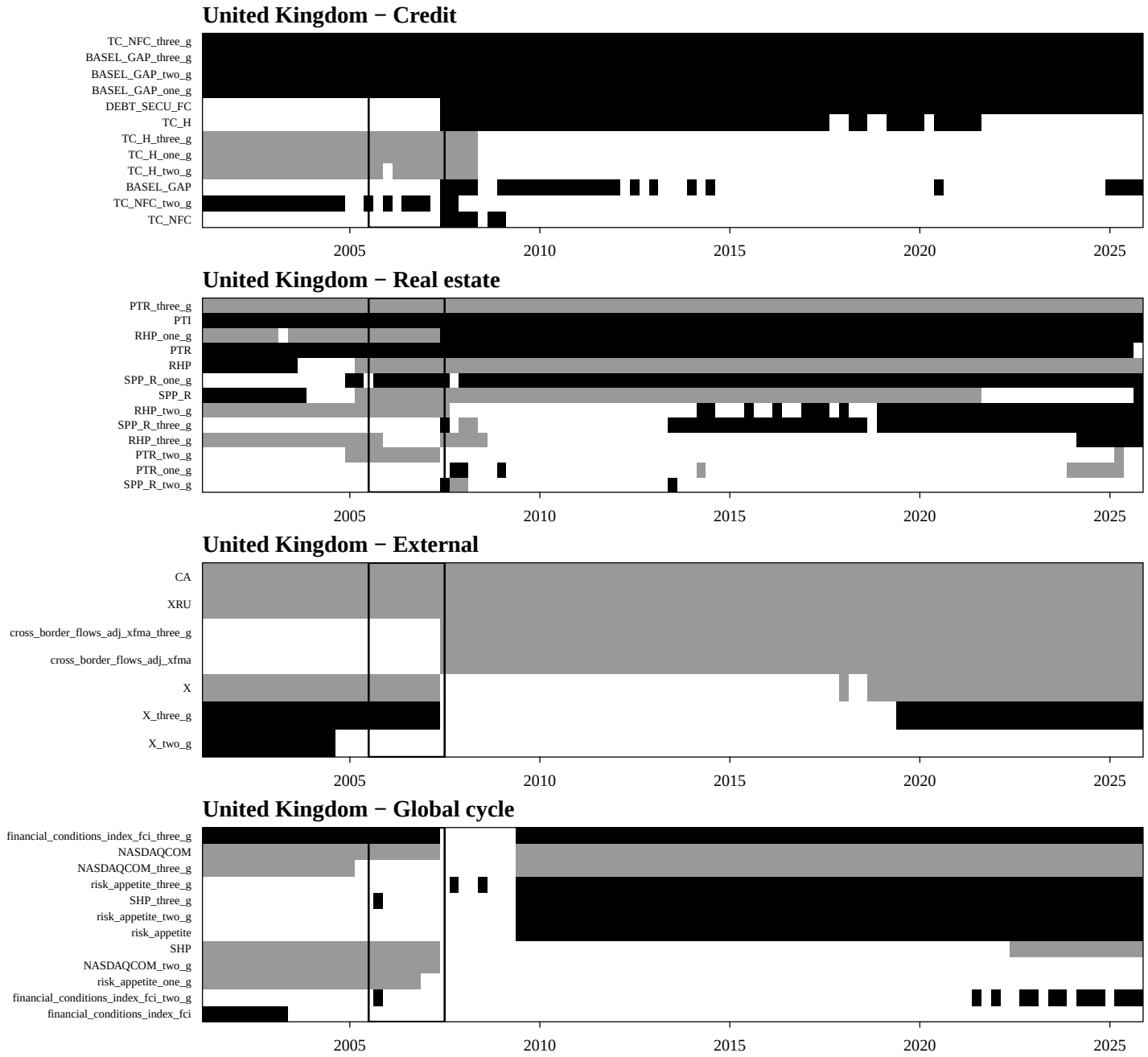
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 18: Coefficients- France



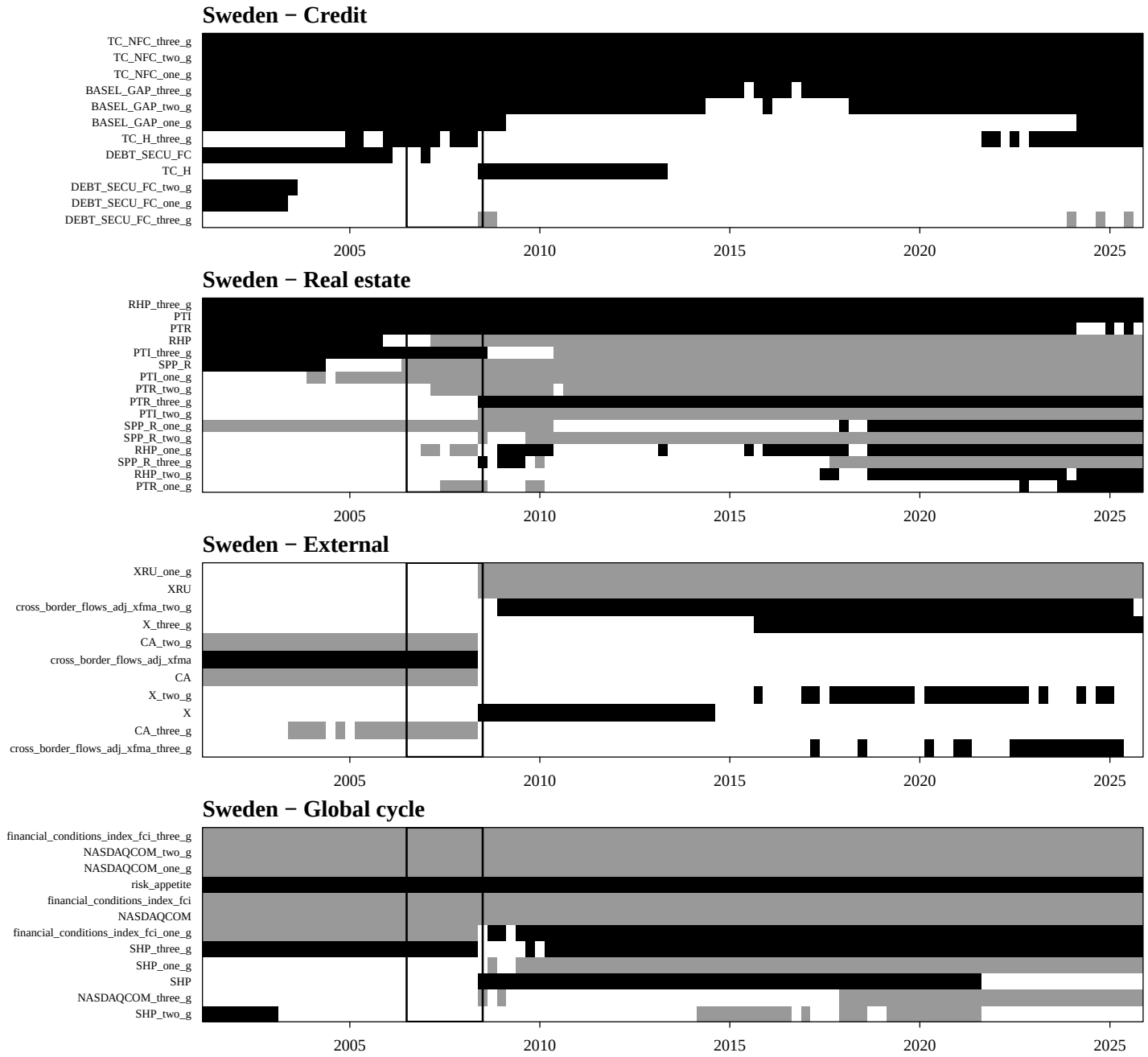
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 19: Coefficients- Germany



Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 20: Coefficients- United Kingdom



Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 21: Coefficients- Sweden

Table 18: Real-time performance of the detection-based aggregation (horizon8, $h = 8$)

Country	AUROC	Pre-crisis	Min. pre-crisis	Calm	Crisis	Threshold	Utility	TP	FP	FN	Lead (q
France	0.85	0.95	0.76	0.30	0	0.41	0.66	8	28	0	2
United Kingdom	0.80	0.64	0.06	0.24	0	0.51	0.86	8	12	0	1
Germany	0.77	0.28	0.01	0.16	0	0.19	-1.23	4	19	4	
Sweden	0.95	0.89	0.39	0.27	0	0.32	0.65	8	29	0	1

Real-time period 2001Q2 onward, excluding the last h quarters whose labels are unknown. Threshold calibrated on the batch period an relative utility with $\theta = 0.8$. Lead is the number of quarters between the start of the alarm in progress at the onset (smoothed signal a onset; it can exceed h. Alarm before window is the number of those quarters that precede the pre-crisis window; they are counted as the expert with the highest average weight in the pre-crisis window.

Table 19: Detection rule versus meta-aggregator

country	AUC_online_CPD	AUC_online_META	pre_CPD	pre_META	calme_CPD	calme_META	U_o
France	0.85	0.86	0.95	0.89	0.30	0.29	
United Kingdom	0.80	0.80	0.64	0.59	0.24	0.23	
Germany	0.77	0.79	0.28	0.27	0.16	0.16	
Sweden	0.95	0.95	0.89	0.83	0.27	0.26	

Table 20: Batch utility by expert and country, and pooled prior

Expert	Prior weight	Lcr	Lho	Lext	Lgfc
Credit	0.371	0.646	0.854	0.729	0
Real estate	0.469	0.917	1.000	0.917	0
External	0.134	0.979	0.979	0.354	0
Global cycle	0.027	0.875	0.750	0.646	0

Utility is the maximal Alessi–Detken utility of each expert on the batch period of each country. The prior is a softmax of the cross-country mean with temperature 6 and a floor of 0.10.

Table 21: Real-time re-estimations, Credit expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
TC_NFC_three_g	97.50	0.81	1.00	1.00
BASEL_GAP_three_g	81.50	0.22	0.75	1.00
TC_NFC_one_g	78.00	-0.27	0.50	1.00
TC_NFC_two_g	76.25	0.30	1.00	1.00
BASEL_GAP_one_g	74.33	0.30	0.78	0.77
BASEL_GAP_two_g	71.50	0.28	0.79	0.92
TC_H_two_g	70.00	0.27	0.67	1.00
DEBT_SECU_FC	64.25	0.13	1.00	1.00
TC_H	56.50	1.10	1.00	1.00
TC_H_one_g	56.00	0.07	0.50	1.00
TC_H_three_g	55.75	0.37	0.75	1.00
DEBT_SECU_FC_two_g	43.00	0.24	1.00	1.00
DEBT_SECU_FC_one_g	41.67	0.05	0.67	1.00
DEBT_SECU_FC_three_g	40.50	0.04	0.22	0.55
BASEL_GAP	39.00	0.35	0.67	1.00
TC_NFC	31.67	-1.01	0.67	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 22: Real-time re-estimations, Real estate expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
PTR	96.75	0.64	0.75	1.00
PTI_three_g	93.00	-1.37	0.11	0.78
PTI_one_g	92.00	-0.08	0.42	0.82
PTI_two_g	87.33	-0.74	0.00	1.00
PTI	81.75	2.01	0.78	0.95
RHP	76.50	-0.14	0.33	0.85
PTR_three_g	66.00	0.17	0.75	1.00
SPP_R	62.00	-0.32	0.13	0.75
SPP_R_one_g	59.75	0.14	0.74	0.53
RHP_three_g	57.75	1.13	0.78	0.83
SPP_R_three_g	53.25	0.60	0.76	0.92
RHP_one_g	52.75	0.24	0.78	0.55
PTR_two_g	44.50	-0.17	0.25	1.00
SPP_R_two_g	41.00	-0.04	0.42	0.67
RHP_two_g	37.75	0.10	0.90	0.79
PTR_one_g	24.00	0.45	0.67	0.54

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 23: Real-time re-estimations, External expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
XRU_two_g	73.00	0.00	0.50	1.00
XRU	70.00	-0.32	0.33	1.00
CA	49.25	-0.56	0.00	1.00
XRU_three_g	45.00	0.13	1.00	1.00
cross_border_flows_adj_xfma_three_g	43.33	0.04	0.67	1.00
cross_border_flows_adj_xfma	43.25	-0.13	0.50	1.00
XRU_one_g	43.00	-0.02	0.33	1.00
X	35.75	0.04	0.50	1.00
X_three_g	34.75	0.04	0.75	1.00
cross_border_flows_adj_xfma_two_g	32.67	0.06	0.67	1.00
CA_one_g	27.00	0.09	1.00	1.00
X_two_g	25.25	0.01	0.75	1.00
CA_three_g	18.00	0.04	0.44	0.89
CA_two_g	17.00	-0.02	0.50	1.00
cross_border_flows_adj_xfma_one_g	14.00	0.03	1.00	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 24: Real-time re-estimations, Global cycle expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
NASDAQCOM	88.50	-0.54	0.25	1.00
financial_conditions_index_fci	76.50	-0.54	0.25	1.00
financial_conditions_index_fci_one_g	74.67	0.06	0.90	0.80
financial_conditions_index_fci_three_g	72.00	0.03	0.67	1.00
risk_appetite_three_g	68.00	0.29	1.00	1.00
risk_appetite	64.00	0.69	1.00	1.00
risk_appetite_two_g	59.00	0.11	0.94	0.88
NASDAQCOM_three_g	56.25	-0.03	0.25	1.00
NASDAQCOM_one_g	54.67	-0.06	0.33	1.00
SHP_three_g	54.33	0.13	1.00	1.00
risk_appetite_one_g	44.50	-0.05	0.50	1.00
SHP	40.75	0.20	0.50	1.00
NASDAQCOM_two_g	35.50	-0.10	0.25	1.00
financial_conditions_index_fci_two_g	34.00	0.04	1.00	1.00
SHP_one_g	33.00	-0.19	0.00	1.00
SHP_two_g	22.50	-0.07	0.12	0.75

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

9.11.3 Vintage

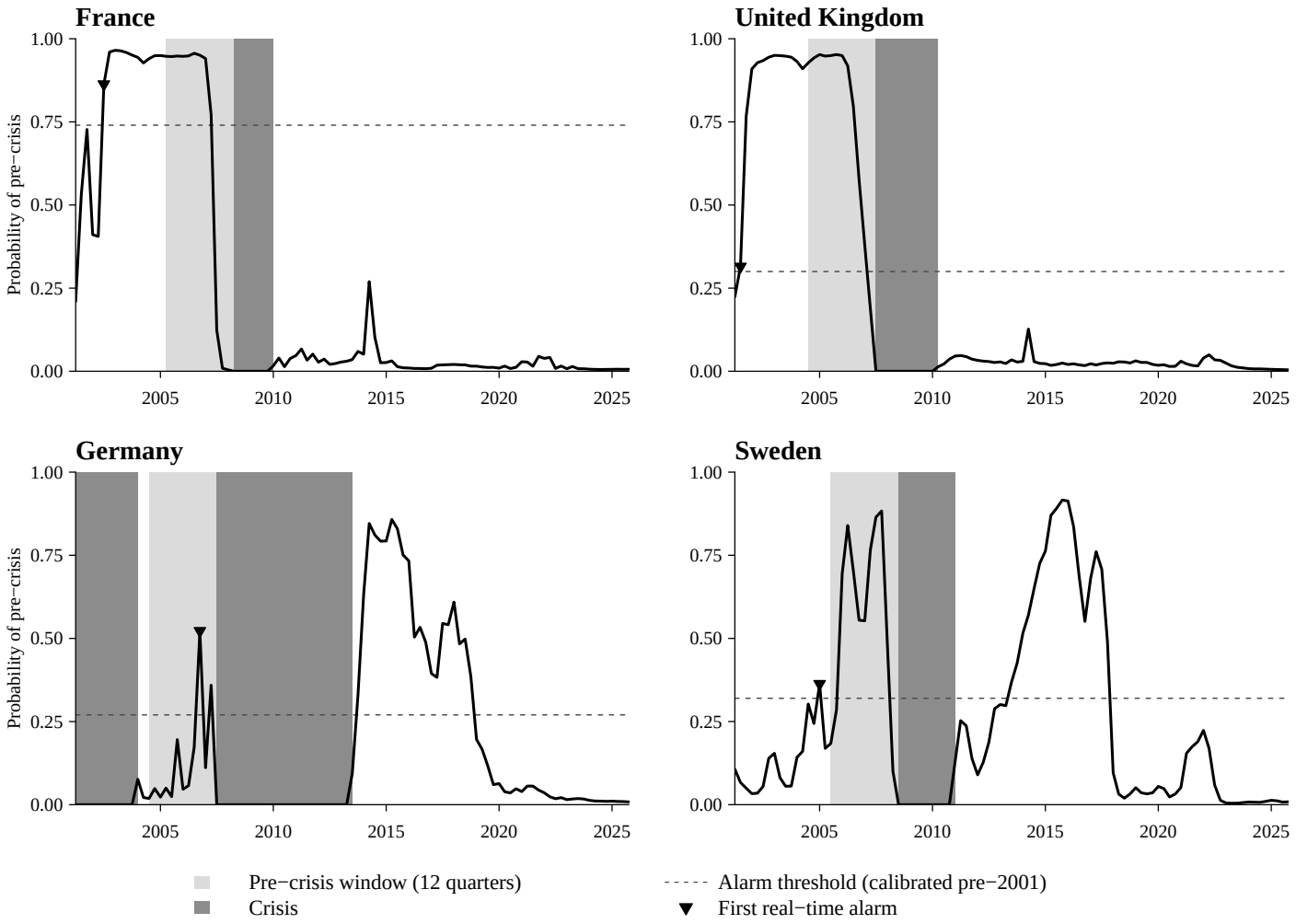


Figure 22: Probability of pre-crisis period - meta aggregator

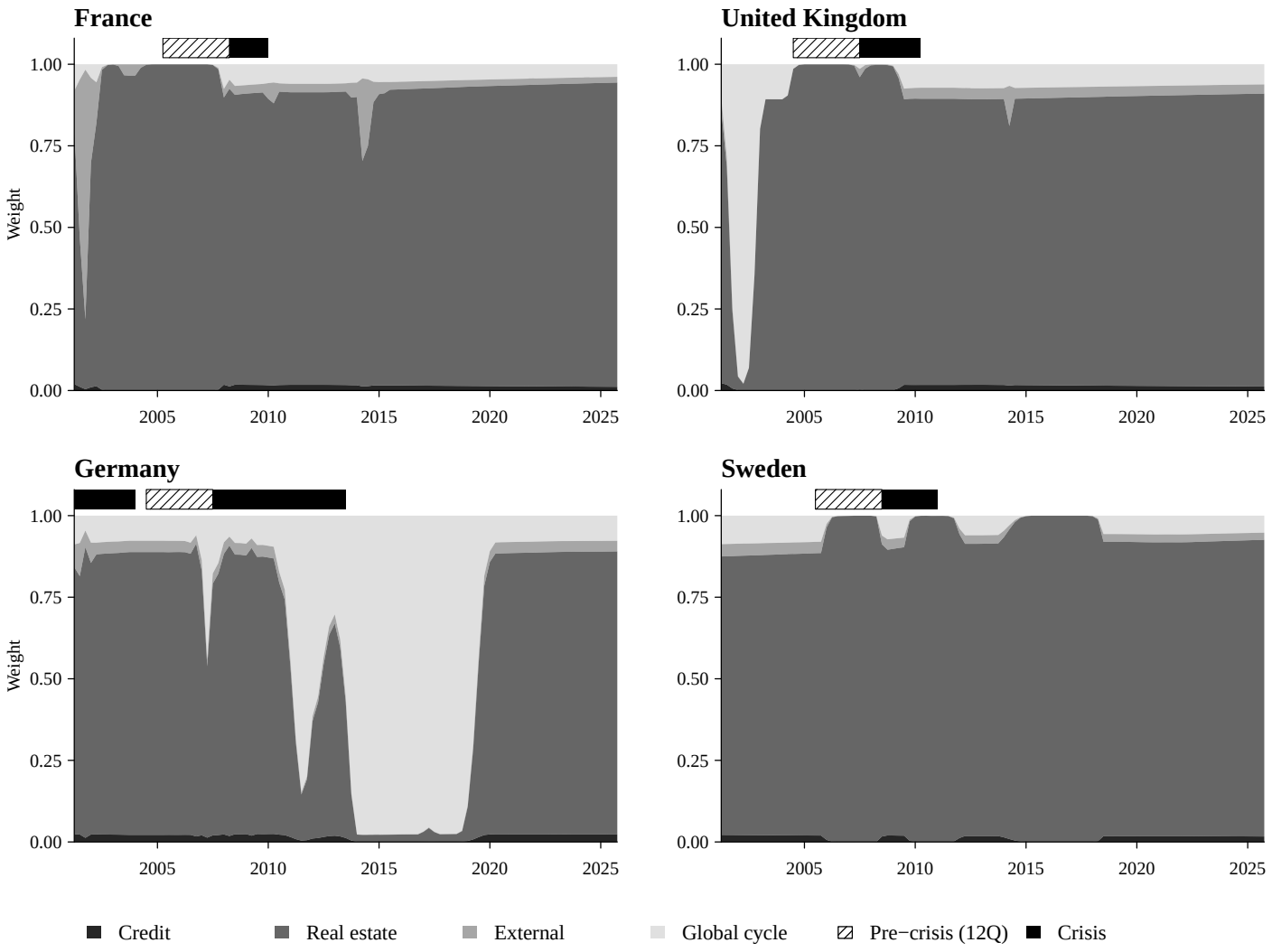


Figure 23: Weights

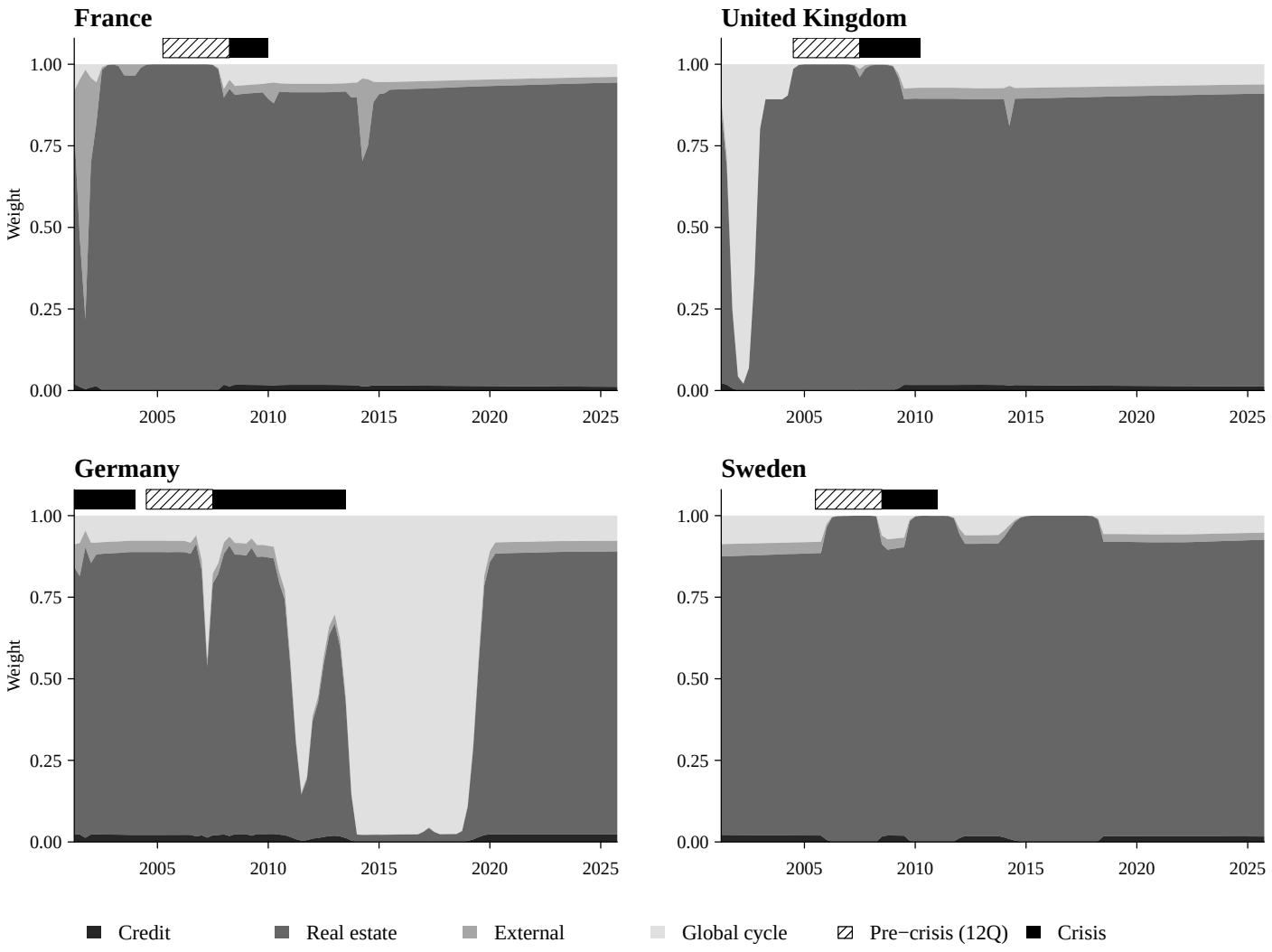


Figure 24: Individual forecasting models

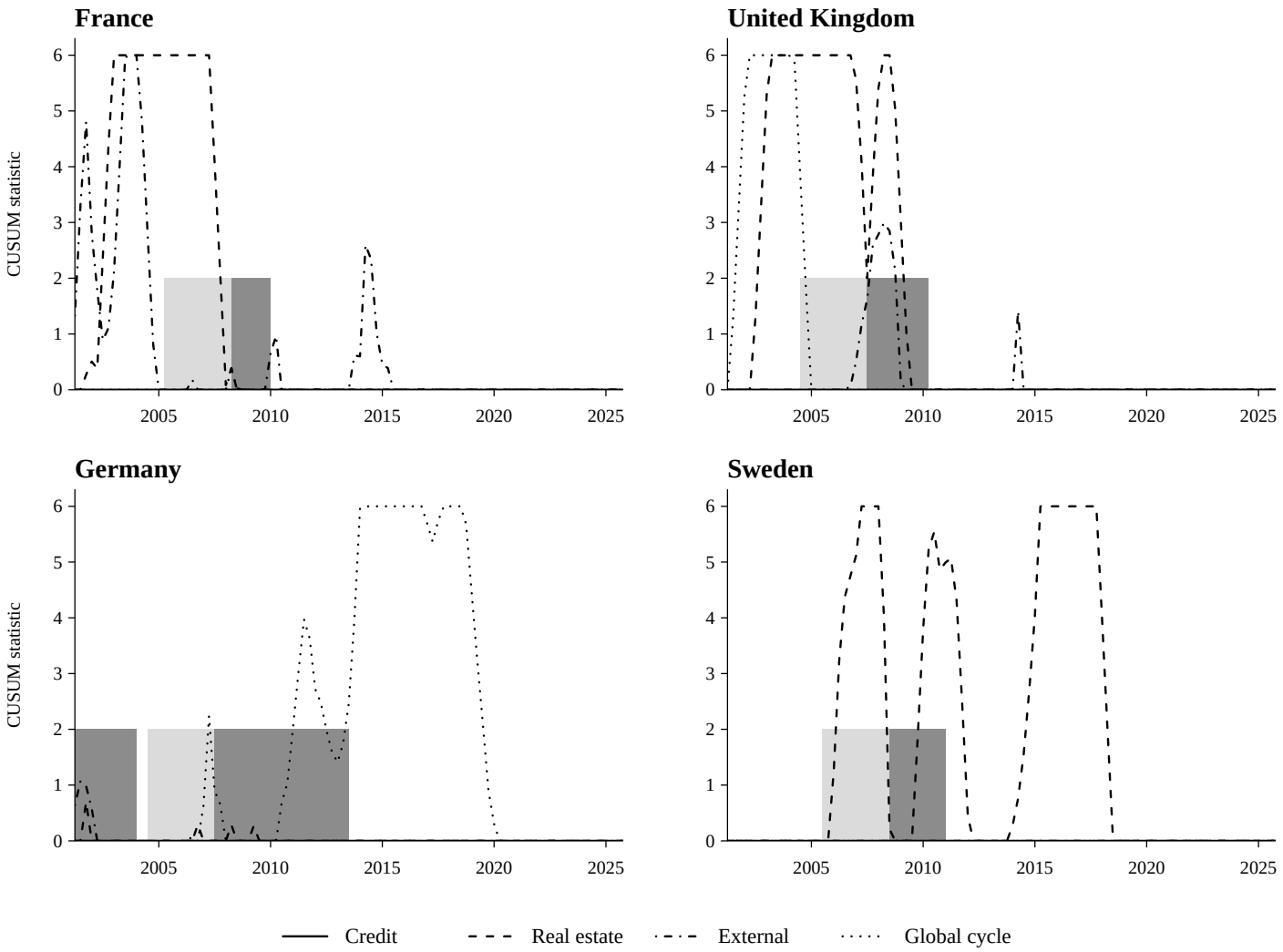
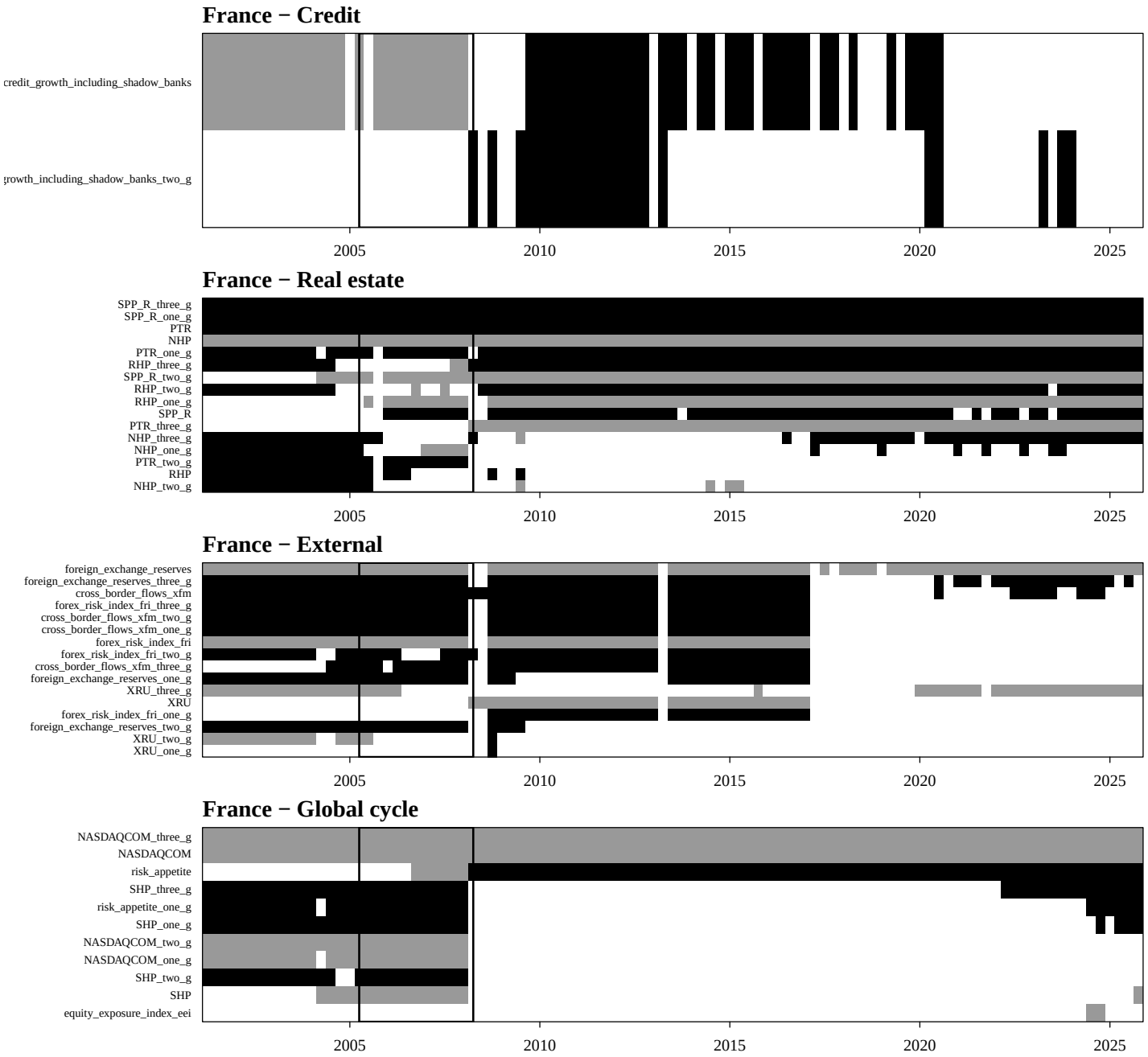
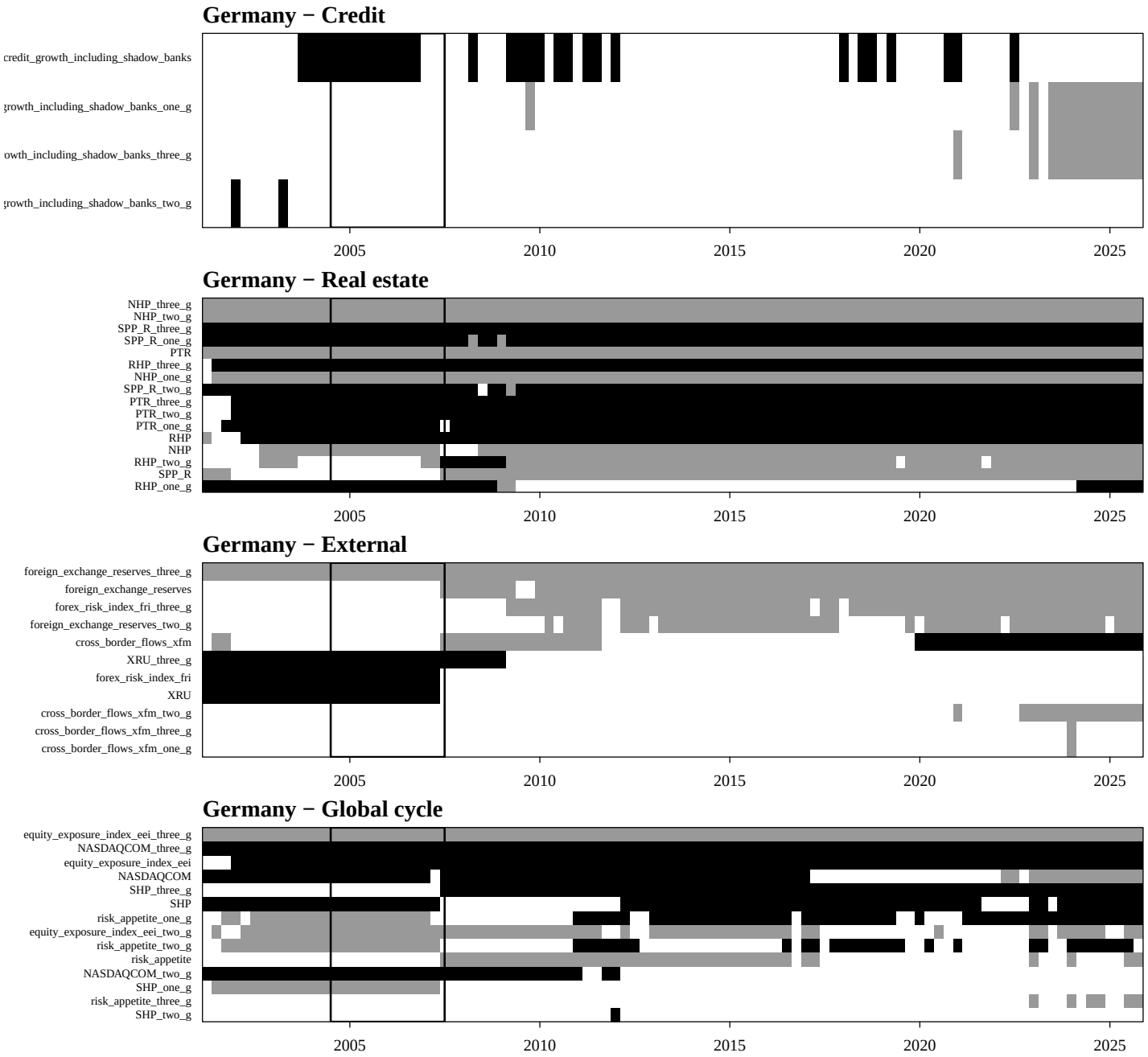


Figure 25: CUMSUM statistics



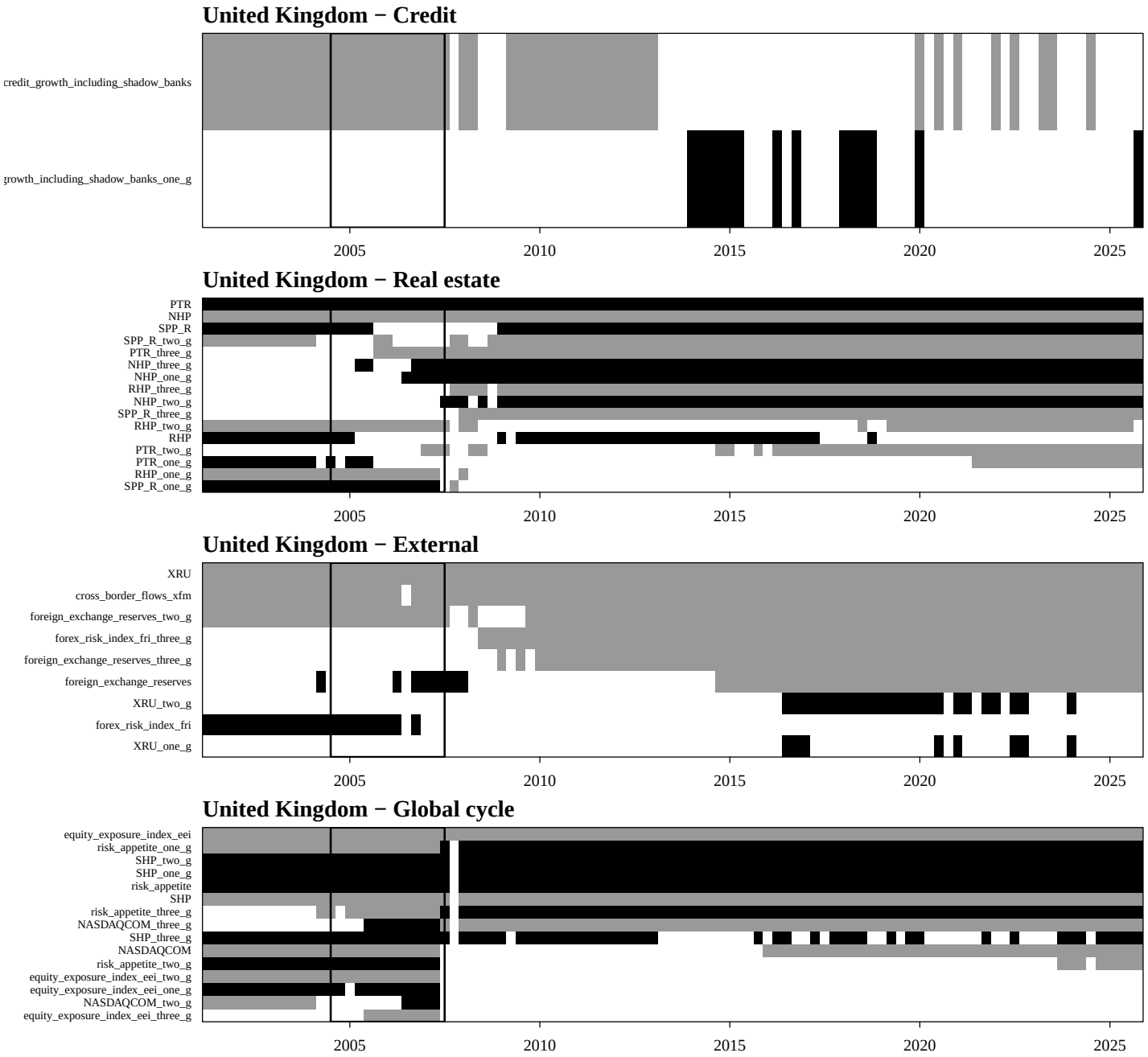
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 26: Coefficients- France



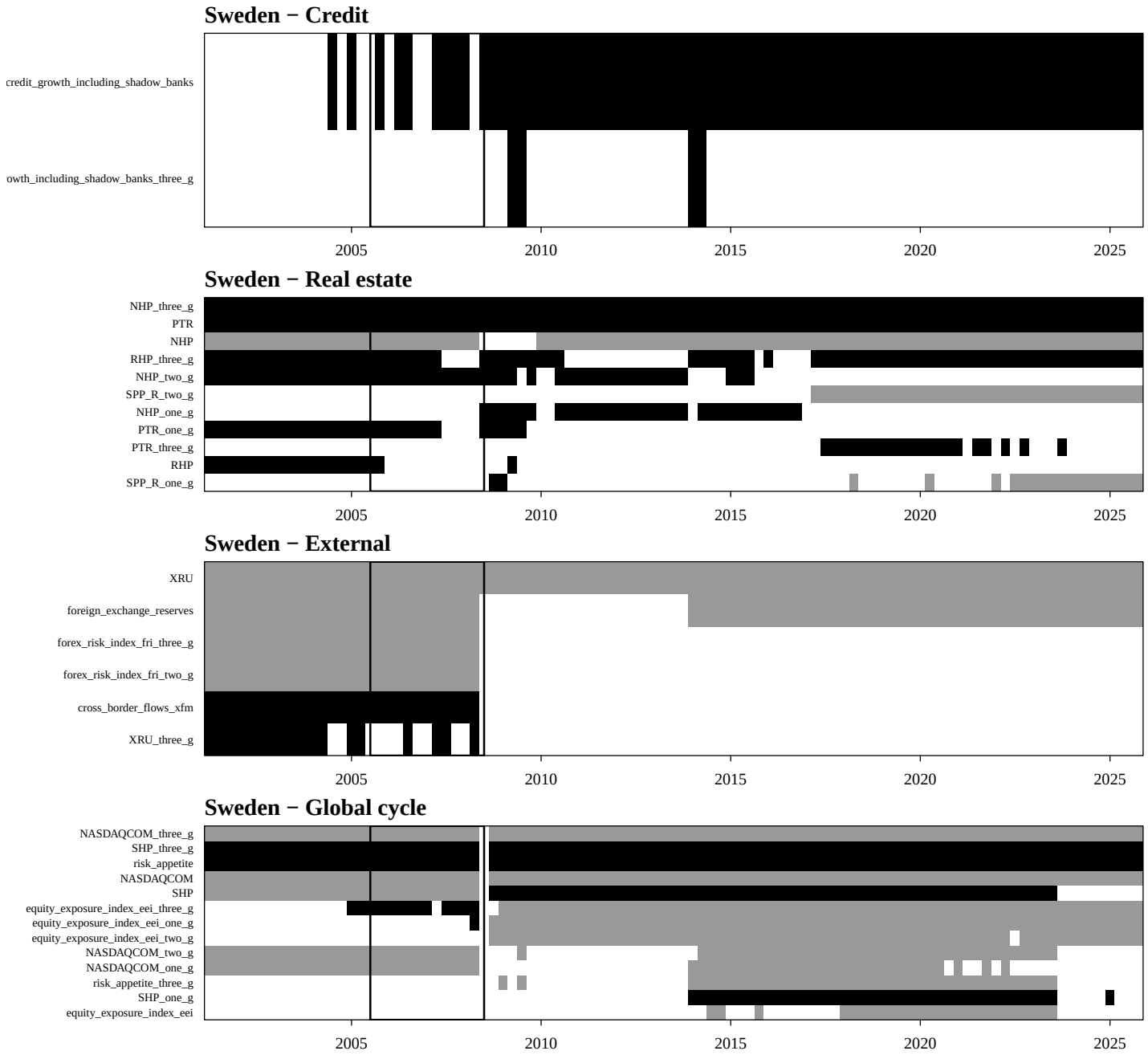
Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 27: Coefficients- Germany



Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 28: Coefficients- United Kingdom



Black: selected with positive coefficient. Grey: selected with negative coefficient. White: not selected. Box: pre−crisis window.

Figure 29: Coefficients- Sweden

Table 25: Real-time performance of the detection-based aggregation (vintage, $h = 12$)

Country	AUROC	Pre-crisis	Min. pre-crisis	Calm	Crisis	Threshold	Utility	TP	FP	FN	Lead (q
France	0.82	0.75	0.00	0.21	0	0.78	0.84	12	12	0	2
United Kingdom	0.94	0.83	0.18	0.19	0	0.31	0.80	12	15	0	2
Germany	0.63	0.13	0.01	0.36	0	0.27	-2.39	3	29	9	
Sweden	0.82	0.62	0.10	0.29	0	0.32	0.64	12	27	0	1

Real-time period 2001Q2 onward, excluding the last h quarters whose labels are unknown. Threshold calibrated on the batch period an relative utility with $\theta = 0.8$. Lead is the number of quarters between the start of the alarm in progress at the onset (smoothed signal a onset; it can exceed h. Alarm before window is the number of those quarters that precede the pre-crisis window; they are counted as the expert with the highest average weight in the pre-crisis window.

Table 26: Detection rule versus meta-aggregator

country	AUC_online_CPD	AUC_online_META	pre_CPD	pre_META	calme_CPD	calme_META	U_o
France	0.82	0.77	0.75	0.71	0.21	0.21	
United Kingdom	0.94	0.92	0.83	0.79	0.19	0.19	
Germany	0.63	0.63	0.13	0.14	0.36	0.35	
Sweden	0.82	0.81	0.62	0.58	0.29	0.28	

Table 27: Batch utility by expert and country, and pooled prior

Expert	Prior weight	Lcr	Lho	Lext	Lgfc
Credit	0.028	0.000	0.886	0.159	0.636
Real estate	0.818	0.000	0.955	0.091	0.818
External	0.045	0.000	0.977	0.045	0.000
Global cycle	0.109	0.000	0.841	0.909	0.705

Utility is the maximal Alessi–Detken utility of each expert on the batch period of each country. The prior is a softmax of the cross-country mean with temperature 6 and a floor of 0.10.

Table 28: Real-time re-estimations, Credit expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
excess_credit_growth_including_shadow_banks	50.75	0.01	0.64	0.78
excess_credit_growth_including_shadow_banks_one_g	10.00	0.02	0.50	1.00
excess_credit_growth_including_shadow_banks_three_g	9.00	-0.00	0.50	1.00
excess_credit_growth_including_shadow_banks_two_g	7.00	0.00	1.00	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 29: Real-time re-estimations, Real estate expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
PTR	99.00	1.36	0.75	1.00
NHP	92.75	-1.87	0.00	0.99
SPP_R_three_g	90.67	1.41	0.68	0.98
NHP_three_g	84.50	-0.19	0.75	0.99
RHP_three_g	84.25	0.71	0.75	0.98
SPP_R_two_g	74.75	-0.61	0.25	0.99
PTR_three_g	67.75	-0.95	0.50	0.99
PTR_one_g	61.00	1.19	0.83	0.83
SPP_R	60.75	-0.86	0.75	1.00
NHP_two_g	59.75	-0.13	0.68	0.86
SPP_R_one_g	58.75	0.54	0.76	0.91
NHP_one_g	58.50	-0.26	0.70	0.91
PTR_two_g	57.00	0.55	0.67	0.99
RHP_two_g	55.25	0.05	0.53	0.91
RHP_one_g	48.00	-0.76	0.32	0.97
RHP	47.00	1.90	1.00	0.99

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 30: Real-time re-estimations, External expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
foreign_exchange_reserves_three_g	79.67	-0.13	0.33	1.00
foreign_exchange_reserves	77.50	-0.39	0.03	0.93
XRU	62.75	-0.43	0.26	0.99
cross_border_flows_xfm	58.25	0.19	0.64	0.76
foreign_exchange_reserves_two_g	55.00	0.00	0.33	1.00
forex_risk_index_fri_three_g	55.00	-0.06	0.25	1.00
forex_risk_index_fri_two_g	41.00	0.04	0.50	1.00
cross_border_flows_xfm_two_g	34.50	0.05	0.50	1.00
forex_risk_index_fri	33.67	-0.11	0.67	1.00
forex_risk_index_fri_one_g	32.00	0.10	1.00	1.00
cross_border_flows_xfm_one_g	30.50	0.04	0.50	1.00
XRU_three_g	27.67	0.02	0.67	1.00
foreign_exchange_reserves_one_g	27.00	-0.04	0.50	1.00
cross_border_flows_xfm_three_g	25.50	0.10	0.50	1.00
XRU_two_g	13.67	-0.00	0.69	0.96
XRU_one_g	4.50	0.01	1.00	1.00

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.

Table 31: Real-time re-estimations, Global cycle expert: selection and sign stability

variable	Selection freq.	Mean coef.	Share positive	Sign stability
NASDAQCOM_three_g	95.75	-0.09	0.27	0.95
NASDAQCOM	81.50	-0.73	0.20	0.90
risk_appetite	78.25	0.28	0.73	0.96
SHP	68.50	0.19	0.43	0.87
SHP_three_g	66.00	0.25	1.00	1.00
risk_appetite_one_g	64.33	0.02	0.84	0.68
equity_exposure_index_eei_three_g	64.00	-0.33	0.07	0.87
equity_exposure_index_eei	57.00	-0.24	0.25	1.00
SHP_one_g	50.00	0.14	0.75	1.00
equity_exposure_index_eei_two_g	48.33	-0.18	0.00	1.00
NASDAQCOM_one_g	48.00	-0.13	0.00	1.00
risk_appetite_three_g	47.67	0.00	0.29	0.91
equity_exposure_index_eei_one_g	44.50	-0.07	0.51	0.99
risk_appetite_two_g	44.00	0.47	0.72	0.45
SHP_two_g	41.00	0.20	1.00	1.00
NASDAQCOM_two_g	39.50	-0.13	0.32	0.87

Averages across the four countries of per-country statistics over all real-time re-estimations. Sign stability is |mean(sign)|: 1 means the sign never changes.