BETA
Bureau
d'économie
théorique
et appliquée

Documents de travail

« Does oversight pay off? Independent fiscal institutions and forecast accuracy»

Auteurs

Théo Metz, Carolina Ulloa-Suárez, Oscar M. Valencia

Document de Travail n° 2026 – 14

Mai 2026

Bureau d'Économie
Théorique et Appliquée
BETA

https://www.beta-economics.fr/

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Does oversight pay off? Independent fiscal institutions and forecast accuracy

Théo Metz * Carolina Ulloa-Suárez † Oscar M. Valencia ‡

May 12, 2026

Abstract

This paper examines the causal effect of Independent Fiscal Institutions (IFIs) on the accuracy of government macroeconomic and fiscal forecasts. Using a novel dataset of real-time budget projections for 55 European and Latin American countries over 1998–2019, we exploit the staggered introduction of IFIs to identify their impact on forecast performance. We find that IFI implementation leads to statistically and economically meaningful reductions in forecast errors for GDP growth, total government revenue, and total government expenditure. Improvements in GDP growth forecasts emerge within two years of implementation, whereas gains in fiscal forecast accuracy materialise more gradually, after four to five years. This dynamic pattern is consistent with reputational and accountability mechanisms, whereby sustained independent scrutiny reshapes governments' forecasting incentives over time. The results are robust across alternative identification strategies and forecast accuracy measures. Overall, the findings underscore the role of IFIs as institutions that strengthen fiscal credibility by reshaping forecasting incentives over time, thereby complementing fiscal rules and supporting more sustainable fiscal policymaking.

Keywords: Independent Fiscal Institutions; forecast accuracy; fiscal credibility; staggered difference-in-differences; fiscal rules.
JEL classification: E62; H61; H68; C23.

1 Introduction

Government budgets are built on forecasts. Every year, when a finance ministry prepares next year's budget, the entire fiscal plan rests on a set of projections whose realism determines whether announced targets will be met or missed. These projections fall into two distinct categories that play different roles in the budget process. *Macroeconomic forecasts*, especially GDP growth projections, define the economic environment within which fiscal policy operates and constitute the baseline assumptions upon which all budgetary calculations depend. *Fiscal forecasts*, in

*Corresponding author: BETA CNRS UMR 7522, University of Strasbourg, France, theo.metz@unistra.fr
†Inter-American Development Bank, Washington, DC, USA, culloasuarez@iadb.org
‡Inter-American Development Bank, Washington, DC, USA, oscarva@iadb.org. The findings and interpretations in this paper are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank or the governments it represents.

contrast, concern the budgetary aggregates themselves, total government revenue and total government expenditure as shares of Gross Domestic Product (GDP), and reflect both the macroeconomic assumptions and the specific policy choices embedded in the budget. Those forecasts are the pillar of the budget and are determinants of its effectiveness. The accuracy of these projections is challenged by future uncertainty, economic shocks, unforeseen events, and the technical capabilities of governments.

However, a large empirical literature documents that official forecasts are not merely imprecise but systematically biased in an optimistic direction. Governments tend to overestimate future growth and revenues, particularly around elections or when fiscal rules become binding (Frankel and Schreger, 2016; Pina and Venes, 2011). This optimism bias is not random noise; it reflects political incentives to present favourable scenarios that allow higher planned expenditure or a lower apparent deficit. When growth subsequently falls short of projections, the resulting revenue shortfalls lead to fiscal slippages and unplanned debt accumulation, eroding the *informational credibility* of fiscal plans, defined as the extent to which official projections can be trusted as a reliable guide to actual budgetary outcomes.

Independent Fiscal Institutions (IFIs) have emerged as an institutional response to this forecasting bias problem. IFIs are non-partisan public bodies mandated to monitor fiscal policy, assess budgetary plans, and scrutinise the assumptions underlying official projections.[1] Crucially, IFIs operate *upstream* in the budgetary process, at the stage where forecasts are produced, and fiscal plans are justified. In a typical budget cycle, the government prepares next year's budget in the autumn based on macroeconomic projections (GDP growth, inflation, and employment). These projections then serve as inputs for the fiscal forecast, in which expected revenues and expenditures are derived. The IFI intervenes precisely at this juncture, either by producing independent macroeconomic forecasts that the government must use or justify departures from, or by publicly assessing the plausibility of the government's own projections before the budget is submitted to parliament.[2] This *ex ante* oversight creates informational incentives that raise the reputational and electoral costs of overly optimistic projections, even though IFIs have no direct decision-making power over the budget itself. By producing or certifying an independent benchmark, IFIs make it politically costly for governments to base fiscal plans on implausible assumptions, thereby strengthening the link between announcements and actual budgetary execution.

Following the sovereign debt crisis of the early 2010s, most European Union (EU) member states established IFIs under the reformed Stability and Growth Pact in 2011/2013 to strengthen fiscal frameworks and improve forecast credibility (Beetsma et al., 2019). Several Latin American and Caribbean (LAC) countries have since adopted similar institutions as part of broader fiscal responsibility frameworks (Mooney et al., 2018). Existing studies generally document positive associations between the presence of IFIs and fiscal outcomes. However, the literature provides limited causal evidence on whether IFIs effectively discipline government forecasting behaviour

[1] IFIs perform these tasks among other functions, whose breadth and depth vary across institutions in accordance with their budgetary resources and human-capital endowment.

[2] While the institutional design and mandates of IFIs vary substantially across countries, a common feature is their role in assessing compliance with fiscal rules and scrutinising the assumptions underlying official forecasts. In more institutionally mature settings, mandates may extend to the direct preparation or endorsement of macroeconomic and fiscal forecasts (Debrun, 2011; Debrun and Kinda, 2017).

during budget preparation, and offers little insight into the timing, persistence, and scope of these effects. In particular, it remains unclear whether observed improvements in forecast accuracy reflect immediate technical corrections or gradual behavioural adjustments as institutional credibility and public scrutiny take hold.

This paper fills this gap by providing a causal study of IFIs' implementation in EU and LAC countries and aims to answer the following questions: Do IFIs causally improve the accuracy of the macroeconomic and fiscal projections used in budget preparation? If so, do these effects arise immediately, reflecting technical oversight, or do they emerge gradually as credibility and accountability mechanisms become established? Moreover, does the specific institutional mandate of an IFI, namely whether it is empowered to produce or to endorse macroeconomic forecasts, fiscal forecasts, or both, condition the types of projections whose accuracy improves?

To address these questions, this paper relies on a novel cross-regional dataset of real-time, ex ante official forecasts that were actually used during budget preparation. The dataset is constructed through a systematic, country-by-country collection of official budget documents and covers projections for GDP growth, government total revenue as a share of GDP, and government total expenditure as a share of GDP for 55 countries, including 28 EU and 27 LAC economies, over the period 1998 to 2019. Forecasts for EU countries come from Stability and Convergence Programmes and Draft Budget Plans for Eurozone members, and cover horizons from t=0 to t=5, depending on data availability. Forecasts for LAC countries are drawn from official medium-term fiscal frameworks and annual budget documents published by national authorities, with the systematic compilation and harmonization of these data building on ongoing data systematization efforts conducted within the FISLAC platform and the Fiscal Management Division (FMM) of the Inter-American Development Bank, and cover horizons from t=0 to t=5, depending on data availability. By focusing on first-vintage forecasts, the data capture governments' expectations at the time fiscal decisions were made, rather than revised assessments incorporating ex post information. This feature allows to directly observe forecasting behaviour under different institutional arrangements and to assess fiscal credibility at its most relevant stage.

To identify the causal impact of IFIs, the analysis exploits the staggered adoption of these institutions across countries and over time. The empirical strategy employs modern difference-in-differences methods that address the biases of two-way fixed effects estimators under staggered treatment adoption (Callaway and Sant'Anna, 2021). This framework allows treatment effects to vary across countries and over time and uses only units that are not yet treated or have never been treated as valid controls. The approach is complemented with local projection difference-in-differences estimators (Dube et al., 2025), PanelMatch estimators (Imai et al., 2023), and propensity score matching to assess robustness and allow for flexible dynamic responses.

The results show that IFIs lead to significant reductions in absolute forecast errors for real GDP growth rate, total revenue (in % of GDP), and total expenditure (in % of GDP). Improvements in real GDP growth forecasts emerge relatively early, within 2 years of IFI establishment, while gains in fiscal forecast accuracy materialise more gradually, typically after 5 years. This staggered dynamic is consistent with reputational and accountability mechanisms, in which governments adjust their forecasting behaviour over time as institutional scrutiny becomes

credible and sustained.

Cohort-based evidence further indicates that these effects are not driven by a particular wave of IFI adoption, but instead reflect a broadly consistent improvement in forecast accuracy following activation across implementation cohorts. This finding strengthens the interpretation of the baseline estimates as capturing a genuine institutional effect, rather than compositional changes in the set of treated countries or cohort-specific dynamics. Finally, the main results are robust to alternative identification strategies, including local projection difference-in-differences, matching-based estimators, and panel matching approaches, as well as to alternative measures of forecast accuracy. Furthermore, when restricting the sample to EU countries only, the results show that when IFIs produce fiscal forecasts, forecast accuracy improves more than when IFIs just endorse them.

The remainder of the paper is structured as follows. Section 2 reviews the literature on forecast bias, political incentives, and the role of IFIs. Section 3 presents a simple theoretical framework that illustrates how IFI oversight can influence government forecasting behaviour. Section 4 presents the data and empirical strategy. Section 4.2 describes the construction of macro-fiscal forecast deviations and provides descriptive evidence. Section 6 reports the main results, while Section 7 discusses robustness checks. Section 8 concludes with implications for fiscal policy and institutional design.

2 Literature review

A growing literature has investigated the link between institutional oversight and forecast quality, though important gaps remain. Subsection 2.1 documents systematic biases in official macroeconomic and fiscal forecasts, identifying their political and institutional origins, and showing how they undermine fiscal credibility. Subsection 2.2 examines Independent Fiscal Institutions as a corrective mechanism, from their theoretical rationale and institutional design to the empirical evidence on their effectiveness, highlighting the identification gaps that motivate the present paper.

2.1 Bias in macro-fiscal forecasts and institutional responses

The literature on fiscal forecasting converges on a central insight that systematic bias in official projections is not primarily a technical forecasting problem but rather a political and institutional one (Jonung and Larch, 2006). Forecast optimism reflects strategic incentives faced by governments, particularly in the presence of electoral pressures and binding fiscal rules that can be met formally through favourable assumptions rather than through genuine adjustment. In this context, improving forecasting performance requires institutional arrangements that operate on incentives, transparency, and credibility, rather than purely technical refinements to forecasting models.

A substantial body of research documents systematic optimism bias in official fiscal forecasts. Evidence from European Union and Eurozone countries shows that government projections for growth, revenue, and deficits tend to be overly optimistic, particularly when deficit thresholds under the Stability and Growth Pact become binding (Jonung and Larch, 2006; Frankel and

Schreger, 2013, 2016). Similar patterns are observed in Latin American economies (Hadzi-Vaskov et al., 2021), indicating that optimism bias is a pervasive feature of official forecasting rather than a region-specific phenomenon.

The literature attributes these systematic biases primarily to political economy factors. Pina and Venes (2011) show that upcoming elections are associated with significantly more optimistic budget balance projections, as governments seek to signal sound economic management to voters. The same authors find that stronger budgetary institutions and stricter procedural rules are correlated with less biased forecasts. At the same time, fiscal rules in the absence of effective enforcement mechanisms may reinforce incentives to engage in optimistic forecasting, as governments attempt to comply formally with numerical targets through favourable assumptions (Frankel and Schreger, 2013).

These biases have meaningful fiscal consequences. Overly optimistic growth projections translate into revenue overestimation, and when growth outcomes fall short of expectations, fiscal balances deteriorate relative to plans (a mechanism formalised theoretically in Section 3 and documented empirically in Section 4.2). Using cross-country evidence, Ardanaz et al. (2024) show that optimism during budget preparation reduces the likelihood of ex post compliance with fiscal rules. This evidence indicates that forecast errors are not random forecasting mistakes, but instead reflect systematic political incentives and institutional weaknesses. This diagnosis naturally motivates the search for institutional arrangements that can discipline expectations and strengthen fiscal credibility, such as IFIs.

2.2 Independent fiscal institutions and macro-fiscal forecasts

IFIs have gained prominence as a response to problems of discipline in fiscal policymaking by delegating oversight tasks to non-partisan experts. The central theoretical argument is that, by scrutinising official forecasts and the assumptions underlying budget plans, IFIs enhance transparency and raise the reputational costs associated with unrealistic projections (Debrun et al., 2013; Calmfors and Wren-Lewis, 2011). Through this channel, IFIs are expected to discourage governments from engaging in systematically optimistic forecasting in the first place (Beetsma and Debrun, 2016).

Several theoretical contributions have formalised this mechanism with distinct modelling strategies. Debrun et al. (2009) develop a principal-agent framework in which the government (agent) has an incentive to manipulate fiscal projections in order to justify higher spending, while voters (principal) cannot directly observe the true state of the economy. In this setting, an IFI acts as a monitoring device that increases the probability that manipulation is detected, thereby raising the expected cost of biased forecasts. The key result is that the disciplining effect depends on the IFI's perceived independence and competence, meaning that an IFI that is captured or under-resourced produces no credible signal and thus no behavioural correction. Beetsma and Debrun (2016) extend this logic to a signalling framework in which the IFI's public assessment acts as a "second opinion" that constrains the government's ability to exploit information asymmetries. Their model predicts that IFIs are most effective when they can commit to publishing assessments regardless of political pressure, and that the effect operates through reputational costs rather than formal sanctions.

Wyplosz (2005) and Calmfors and Wren-Lewis (2011) adopt a broader institutional design perspective, comparing the "fiscal council" model, in which IFIs advise but do not decide, to full delegation of fiscal targets, analogous to independent central banks. They argue that full delegation is both politically infeasible and normatively undesirable for fiscal policy, given its inherently redistributive nature, but that advisory bodies endowed with analytical capacity and public visibility can approximate the credibility benefits of delegation without the democratic costs. Common to these theoretical contributions is the prediction that IFIs operate through informational and reputational channels rather than through direct enforcement, which implies that their effects should be gradual, building as institutional credibility and public awareness develop, rather than immediate.

A growing strand of the literature emphasises that the effectiveness of IFIs depends critically on their institutional design. Mooney et al. (2018) highlight that impactful councils typically combine independence from political interference with adequate resources, strong communication capacity, and mandates that include the evaluation or production of official forecasts. At the same time, substantial heterogeneity exists across IFIs. Some institutions focus narrowly on fiscal aggregates, while others engage more broadly with macroeconomic projections. This variation is theoretically relevant, as mandate specificity should shape which forecasts IFIs can most credibly influence (Calmfors and Wren-Lewis, 2011). However, this prediction has rarely been tested empirically.

Empirical evidence is broadly consistent with the disciplining view of IFIs, although estimates vary depending on sample coverage, identification strategy, and the measure of forecast quality employed. Debrun et al. (2009) analyse a panel of 15 EU countries over 1990–2004 using OLS and instrumental variables and find that the mere presence of an IFI is associated with reduced fiscal forecast bias, though the limited number of IFIs operational during their sample period constrains identification. Cross-country studies further suggest that countries with IFIs tend to exhibit smaller forecast errors and better fiscal outcomes, particularly when these institutions enjoy operational independence and clearly defined mandates (Debrun and Kinda, 2017). Focusing on the European context, Gilbert and de Jong (2017) examine 24 EU countries over 1999–2014 using panel fixed effects and show that the presence of IFIs attenuates the optimism bias associated with the Stability and Growth Pact's 3% deficit threshold, particularly in high-deficit Eurozone countries. Using a broader panel of 39 countries (both EU and non-EU) over 1996–2018, Beetsma et al. (2019) report that IFI presence is associated with more accurate and less optimistic GDP growth and budget balance forecasts, with stronger effects for institutions that produce or endorse macroeconomic projections. These findings are consistent with comparative evidence indicating that forecasts produced by independent institutions, such as the European Commission or the OECD, tend to be less biased than those produced by national governments (Merola and Pérez, 2013), though this comparison confounds institutional independence with differences in forecasting resources and incentives.

However, the existing evidence is subject to several important methodological limitations. Most studies rely on two-way fixed-effects (TWFE) panel regressions or on least-squares dummy-variable specifications with binary IFI indicators. These approaches are now known to produce biased estimates under staggered treatment adoption when treatment effects are heterogeneous

across countries or over time (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021). In such settings, the TWFE estimator implicitly uses already treated units as controls for newly treated units, leading to contaminated comparisons that may severely bias causal estimates and, in some cases, generate sign reversals.

In addition, endogeneity concerns remain pervasive. Countries with stronger fiscal institutions or a greater commitment to fiscal discipline may be more likely to adopt IFIs in the first place, while unobserved factors such as political culture or reform preferences may jointly influence IFI adoption and forecast accuracy. Finally, most empirical specifications model IFI presence as a time-invariant binary treatment, thereby abstracting from potentially important dynamics. As a result, the literature provides limited guidance on whether the effects of IFIs on forecasting behaviour arise immediately through technical corrections or emerge gradually as institutional credibility and public scrutiny are established over time.

The available evidence suggests that IFIs may improve forecast accuracy, although important questions remain unresolved. Existing studies provide limited guidance on the causal nature of these effects, their temporal dynamics, and the role of institutional design. In particular, it remains unclear whether observed improvements reflect immediate technical corrections or gradual credibility-building processes, and whether IFIs with different mandates influence macroeconomic and fiscal forecasts in distinct ways.

3 A model of Independent Fiscal Institutions, incentives, and forecast accuracy

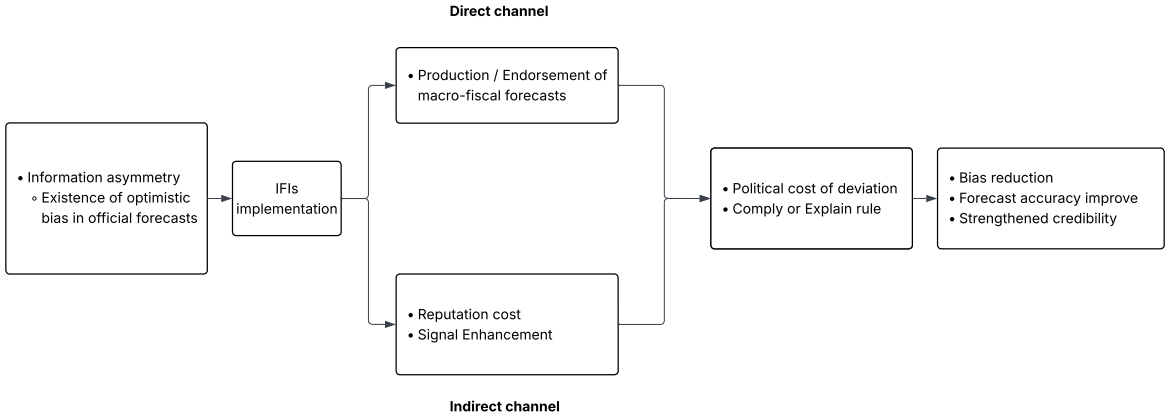
The political economy literature has developed several theoretical frameworks to explain why governments systematically deviate from accurate macroeconomic and fiscal forecasts.[3] A common insight across these models is that forecast bias reflects strategic incentives rather than purely technical forecasting errors. Governments may benefit from optimistic projections insofar as they relax budget constraints in the short run, even if such optimism increases the risk of future fiscal deviations once outcomes materialise.

Recent contributions extend these frameworks to analyse the role of IFIs in shaping fiscal behaviour through different channels. Beetsma et al. (2022) emphasise the role of IFIs in reducing informational noise in political competence signalling, Sloof et al. (2025) highlight their importance for enabling state-contingent fiscal rules, and Debrun (2011) argue that IFIs enhance democratic accountability by helping voters distinguish adverse shocks from poor policy choices. This literature suggests that IFIs operate primarily by altering incentives and accountability, rather than by mechanically improving forecasting techniques.

Figure 1 summarises the core mechanisms through which IFIs are expected to influence government forecasting behaviour. In the absence of independent oversight, governments face incentives to adopt optimistic macroeconomic assumptions that facilitate higher spending or lower apparent deficits. IFIs intervene upstream in the budgetary process by scrutinising these assumptions and, depending on their mandate, by endorsing or producing forecasts. This intervention operates through two institutional channels that jointly determine the political cost

[3] See, among others, Cukierman and Meltzer (1986); Boylan (2008); Shi and Svensson (2006).

Figure 1: Transmission channels of Independent Fiscal Institutions on economic forecasts



Source: Authors.

governments face when deviating from plausible forecasts.

A direct channel arises when IFIs have mandates to produce or formally endorse the macroeconomic and fiscal forecasts used in budget preparation, thereby imposing procedural constraints on the assumptions governments can adopt. An indirect channel operates through reputation and signalling mechanisms, whereby IFI reports and public assessments raise the visibility of forecast choices and amplify political or electoral costs when projections prove unrealistic. Both channels feed into the *political cost of deviation* represented in Figure 1, which corresponds to the marginal cost term δ introduced in the formal model (Section 3.2). This cost aggregates reputational sanctions, electoral penalties, parliamentary scrutiny, and procedural compliance requirements generated by IFI oversight. Stronger mandates, particularly those granting production or endorsement authority over official forecasts, are associated with higher values of δ, thereby increasing the marginal cost of forecast bias and reducing the government's optimal degree of optimism in equilibrium.

To formalise these mechanisms and derive testable predictions, this section develops a simple theoretical model of government forecasting behaviour under alternative institutional settings. The model captures the trade-off faced by governments between optimistic forecasting, which yields short-term political or fiscal benefits, and the expected costs of fiscal deviations when realised outcomes fall short of projections. Introducing an IFI into this environment increases the expected cost of forecast bias, thereby reducing the equilibrium degree of optimism. The framework delivers clear predictions regarding the direction, timing, and scope of IFI effects on forecast accuracy, which guide the empirical analysis that follows.

3.1 Baseline model without an IFI

Consider a government preparing its annual budget. The government chooses a forecast for next year's GDP growth, denoted by g^f, which is used to project revenues and set expenditure plans. Let μ denote the true expected growth rate based on available information. The analysis defines the forecast bias as $\beta = g^f - \mu$, where $\beta > 0$ corresponds to an optimistic forecast.

In the absence of institutional constraints, the government faces a trade-off between the political benefits of optimistic forecasting and the expected costs associated with future fiscal deviations. On the benefit side, a higher growth forecast relaxes the budget constraint by inflating projected revenues, allowing higher expenditure or a lower apparent deficit. This is captured with this incentive with a linear benefit function,

$$B(\beta) = \alpha\beta, \tag{1}$$

where $\alpha > 0$ reflects the political value of additional fiscal space.

On the cost side, optimistic forecasts increase the likelihood and magnitude of fiscal deviations when realised growth falls short of projections. These deviations may trigger penalties through fiscal rules, market reactions, or political accountability. Realised growth is given by $g^a = \mu + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$, so the ex post forecast error is $g^f - g^a = \beta - \epsilon$. Following Beetsma et al. (2022), who argue that the relevant reputational penalty is levied on the *realised* forecast error rather than on the announced bias, the cost function takes the stochastic form

$$C(\beta, \epsilon) \;=\; \tfrac{1}{2}\theta T^2(\beta - \epsilon)^2, \tag{2}$$

where $\theta > 0$ captures the government's aversion to fiscal deviations and T measures the sensitivity of the budget balance to growth surprises. Taking expectations over ϵ delivers the expected cost of forecast bias,

$$\mathbb{E}[C(\beta, \epsilon)] \;=\; \tfrac{1}{2}\theta T^2\big(\beta^2 + \sigma^2\big). \tag{3}$$

The variance σ^2 enters as a constant term that is independent of the government's forecasting strategy. Appendix A.1 provides the full derivation.

The government chooses $\beta \geq 0$ to maximise expected utility,

$$\mathbb{E}[U(\beta)] \;=\; \alpha\beta \;-\; \tfrac{1}{2}\theta T^2\big(\beta^2 + \sigma^2\big). \tag{4}$$

Because σ^2 is constant in β, the first-order condition yields the optimal forecast bias

$$\beta^* \;=\; \frac{\alpha}{\theta T^2}. \tag{5}$$

The baseline model implies that, absent external oversight, the government rationally chooses an optimistically biased forecast. The degree of bias increases with the political benefits of fiscal space and decreases with the expected cost of fiscal deviations. In this setting, forecast optimism emerges endogenously as a strategic choice rather than as a technical forecasting error. Appendix A.1 presents the formal derivation underlying the baseline model.

3.2 Introducing an IFI

The model now introduces an IFI that provides public scrutiny of the government's forecast. Rather than modelling the IFI as mechanically replacing the government's forecast, this section adopts a reduced-form approach in which IFI oversight alters the government's cost-benefit calculation by increasing the expected penalty associated with forecast bias.

IFI oversight operates through reputational and political channels. When projected growth deviates substantially from plausible benchmarks, the IFI can publicly challenge the forecast, generating scrutiny from parliament, the media, or financial markets. Consistent with the stochastic cost specification introduced above, the penalty is levied on the *realised* forecast error rather than on the announced bias[4]:

$$R(\beta, \epsilon) \;=\; \delta\,(\beta - \epsilon)^2, \qquad \mathbb{E}[R(\beta, \epsilon)] \;=\; \delta(\beta^2 + \sigma^2), \tag{6}$$

where $\delta > 0$ captures the institutional strength of the IFI, aggregating the political costs arising from both the direct channel (production/endorsement mandates) and the indirect channel (reputational and signalling effects) illustrated in Figure 1. Stronger mandates generate higher values of δ by imposing tighter procedural constraints and increasing public visibility of forecast choices.

Under IFI oversight, the government chooses $\beta \geq 0$ to maximise

$$\mathbb{E}[U(\beta)] \;=\; \alpha\beta \;-\; \tfrac{1}{2}\theta T^2(\beta^2 + \sigma^2) \;-\; \delta(\beta^2 + \sigma^2). \tag{7}$$

The σ^2 terms are constant in β, so the first-order condition delivers

$$\beta^{**} \;=\; \frac{\alpha}{\theta T^2 + 2\delta}, \tag{8}$$

which is strictly lower than the baseline bias β^*. IFI oversight therefore unambiguously reduces forecast optimism. The stronger the IFI's influence, the closer the equilibrium forecast moves toward the unbiased benchmark. Appendix A.2 provides the formal derivation.

Table 1 summarises how the key elements of the theoretical framework map into the empirical analysis. The model provides a clear interpretation of forecast errors as the outcome of strategic incentives faced by governments, and of IFI oversight as a mechanism that raises the marginal cost of forecast bias. This mapping guides both the choice of outcome variables and the identification strategy used to estimate the causal effects of IFIs on forecast accuracy.

[4]The quadratic specification of IFI costs reflects the increasing severity of reputational penalties as forecast bias becomes more pronounced. While minor deviations from IFI benchmarks may generate limited scrutiny, substantial biases trigger more visible public interventions and media attention. This convexity captures the 'spotlight' mechanism through which IFIs operate (Beetsma et al., 2019; Debrun et al., 2013), while the realised-error specification follows Beetsma et al. (2022).

Table 1: Mapping theoretical framework to empirical analysis

Parameter	**Role in the model**	**Empirical counterpart**
β	Optimism bias in government forecasts	Forecast errors in growth, revenue, and expenditure based on first-vintage official projections
α	Political benefits of optimistic forecasts	Political incentives and fiscal pressures, captured by country fixed effects
θ	Cost of fiscal deviations	Fiscal discipline and enforcement environment, captured by fixed effects
δ	Strength of IFI oversight	Presence and design of IFIs, including mandate heterogeneity
IFI adoption	Increase in the marginal cost of forecast bias	Staggered introduction of IFIs across countries and over time

This framework highlights two key implications. First, IFIs act as a commitment device by making the costs of optimistic forecasting more immediate and salient to policymakers. Second, the effectiveness of IFIs depends on institutional design. IFIs with limited authority or visibility exert only modest discipline, whereas institutions with strong mandates or forecast endorsement powers can substantially curb forecast bias.

By reducing optimism in growth assumptions, IFI oversight also improves the accuracy of revenue projections and the alignment of expenditure plans, thereby lowering the likelihood and magnitude of fiscal deviations.

The model also generates a prediction about the differential timing of improvements across forecast types, which is directly relevant to interpreting the empirical results. Macroeconomic forecasts, particularly GDP growth projections, are relatively standardised technical exercises that rely on established methodologies, publicly available data, and well-defined modelling frameworks. IFI scrutiny of these assumptions can therefore yield rapid corrections, as the technical dimension of the forecast is amenable to independent verification and benchmarking. Fiscal forecasts, revenue and expenditure projections, are, by contrast, more deeply embedded in the budgetary process and more directly shaped by political choices. Revenue projections depend not only on the macroeconomic baseline but also on assumptions about tax policy changes, compliance behaviour, and the composition of growth, all of which involve greater discretionary judgment. Expenditure projections are even more politically sensitive, reflecting distributive decisions, contingent commitments, and the scope of automatic stabilisers.

In terms of the model, this asymmetry implies that the marginal cost parameter δ operates with a lag for fiscal forecasts. While IFIs can immediately raise the cost of biased GDP assumptions (which are easy to benchmark against external sources), the cost of biased revenue and expenditure assumptions rises only as the IFI builds sufficient institutional knowledge, credibility, and public visibility to challenge the more complex and politically charged components of the budget. This prediction motivates the examination of dynamic treatment effects at different post-adoption

horizons in the empirical analysis.

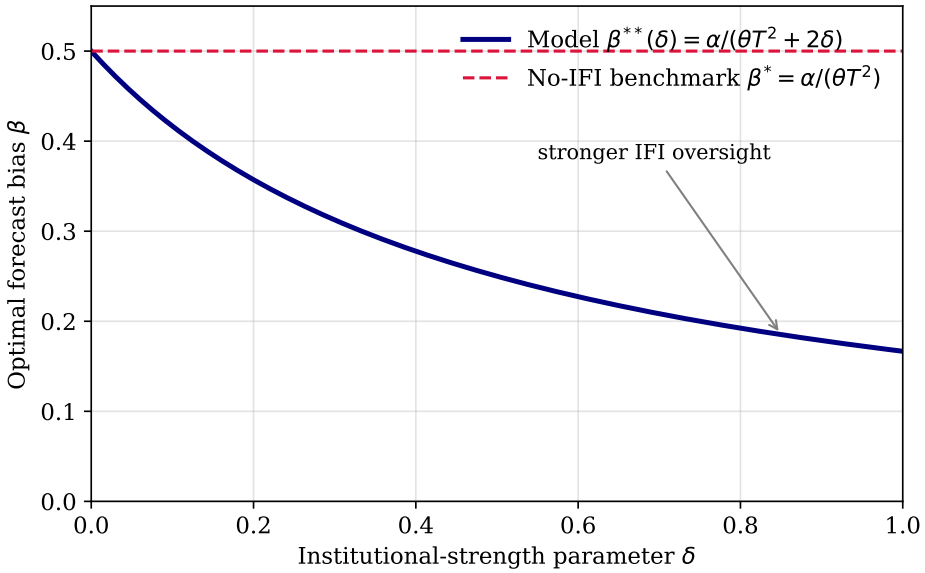
These predictions guide the empirical analysis of forecast accuracy and institutional heterogeneity that follows. Appendix A.2 provides the formal derivation of the IFI extension.

3.3 Predicted effect of IFI oversight on forecast optimism

The closed-form expression $\beta^{**} = \alpha/(\theta T^2 + 2\delta)$ yields a sharp qualitative prediction: optimal forecast optimism is strictly decreasing and convex in the institutional-strength parameter δ. At $\delta = 0$ the expression collapses to the baseline bias $\beta^* = \alpha/(\theta T^2)$; as δ rises, the marginal cost of deviating from a plausible forecast grows, and the equilibrium bias falls monotonically toward zero in the limit $\delta \to \infty$. The speed of the decline is governed by the curvature of the expected cost, θT^2: when the government already faces a steep penalty on fiscal deviations, additional IFI scrutiny delivers smaller marginal gains, whereas in environments with weak intrinsic costs the same increase in δ produces much larger reductions in optimism.

Figure 2 illustrates this prediction for an illustrative calibration with $\alpha = 0.5$ and $\theta = T = 1$. The horizontal dashed line marks the no-IFI benchmark β^*, while the solid curve traces $\beta^{**}(\delta)$ as institutional strength increases. The figure makes explicit three comparative statics that guide the empirical analysis below: (i) the *sign* of the effect—introducing an IFI always lowers equilibrium optimism; (ii) the *magnitude*, which is substantial even for moderate values of δ; and (iii) the *diminishing returns* to institutional strength, which suggest that the marginal value of additional independence is largest in countries with weak baseline oversight. These predictions motivate the staggered difference-in-differences and cohort-based analyses in Sections 6 and 7.

Figure 2: Predicted effect of IFI oversight on equilibrium forecast optimism



Notes: Theoretical schedule of the optimal forecast bias $\beta^{**}(\delta) = \alpha/(\theta T^2 + 2\delta)$ obtained from the first-order condition of the government's expected-utility problem under IFI oversight (Section 3.2). The curve is drawn for the illustrative calibration $\alpha = 0.5$ and $\theta = T = 1$. The horizontal dashed line shows the no-IFI benchmark $\beta^* = \alpha/(\theta T^2)$; the solid curve is strictly decreasing and convex, with $\beta^{**} \to 0$ as $\delta \to \infty$.

4 Data structure and forecast error measure

This section describes the construction of the dataset used to quantify the stylised facts documented in Section 4.2. The data structure is designed to capture both forecast bias and forecast accuracy, their evolution across forecast horizons, and the propagation of macroeconomic forecast errors into fiscal outcomes across countries and time.

4.1 Official macro-fiscal forecasts

A central contribution of the paper is the construction of a novel cross-country dataset combining official macroeconomic and fiscal forecasts with realised outcomes for three core variables: GDP growth, total government revenues, and total government expenditures. The dataset covers the period 1998–2019 and includes 55 countries, comprising 28 EU members and 27 LAC countries.[5] For each country-year, we compute forecast errors as the difference between realised outcomes and the projections available at the time of budget preparation, the end-of-year forecasts.

The novelty of the dataset lies in the manual collection and harmonisation of first-vintage, ex ante forecasts drawn directly from official fiscal planning documents. For EU countries, forecasts are taken from Stability and Convergence Programmes, or Draft Budgetary Plans for Euro area members, submitted to the European Commission, while realised outcomes correspond to ex post fiscal data reported by Eurostat. For LAC countries, forecasts are collected from official medium-term fiscal frameworks and annual budget documents published by national authorities, and realized outcomes are constructed from observed fiscal aggregates reported ex post and aligned with internationally comparable statistical definitions[6]. This approach ensures conceptual consistency between forecasts and outcomes across regions.

All forecasts correspond to first-vintage projections, that is, the initial forecasts produced at the time of budget preparation, prior to any intra-year revisions or ex post updates. These projections reflect the information set and expectations that effectively guided fiscal decisions when policies were designed. Focusing on first-vintage forecasts allows the dataset to capture institutional and political determinants of forecast behaviour, rather than subsequent information adjustments. We also winsorise the sample's extreme values to avoid issues with weighting arising from large forecast deviations.

This data structure is particularly well-suited to study the role of fiscal institutions. IFIs primarily operate at the budget preparation stage, aiming to improve the realism and credibility of official projections. Using revised or updated forecasts would mechanically reduce observed forecast errors, thereby blurring the channels through which institutional oversight affects forecast bias and accuracy.

[5] **EU**: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and the United Kingdom.
LAC: Argentina, Belize, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, The Bahamas, Uruguay, and Venezuela.

[6] For LAC countries, the systematic compilation and harmonization of medium-term fiscal framework data builds on ongoing data systematization efforts conducted within the FISLAC platform and the Fiscal Management Division (FMM) of the Inter-American Development Bank.

4.2 Forecast errors: measurement and stylized facts

This section defines the forecast-error measures used in the empirical analysis and documents basic stylised facts in the data. The analysis focuses on first-vintage projections, so the measures reflect the information set available at the time budgets were prepared and the expectations that effectively guided fiscal decisions, as described in Section 4.

The analysis assesses forecast performance using standard measures of forecast errors, namely the Mean Error (ME), the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the Mean Absolute Percentage Error (MAPE). These measures allow us to distinguish between forecast bias and forecast accuracy. The ME captures the systematic direction of forecast deviations and therefore provides information on the fiscal stance implicit in budget projections. Negative values of ME indicate optimistic forecasts, while positive values indicate pessimism. By contrast, MAE, RMSE, and MAPE measure forecast accuracy, capturing the magnitude of forecast errors regardless of their sign. Higher values of these statistics reflect larger discrepancies between forecasts and realizations.[7]

Let $e_{i,t+h} = y_{i,t+h} - f_{i,t+h}$ denote the forecast error for country i at horizon $t+h$, where $y_{i,t+h}$ is the realized outcome and $f_{i,t+h}$ the corresponding forecast. For a sample of N_{t+h} country observations at a given forecast horizon, the analysis computes the following standard forecast-error measures:

- Mean Error:

$$\text{ME}_{t+h} = \frac{1}{N_{t+h}} \sum_{i=1}^{N_{t+h}} e_{i,t+h} \tag{9}$$

- Mean Absolute Error:

$$\text{MAE}_{t+h} = \frac{1}{N_{t+h}} \sum_{i=1}^{N_{t+h}} |e_{i,t+h}| \tag{10}$$

- Root Mean Squared Error:

$$\text{RMSE}_{t+h} = \sqrt{\frac{1}{N_{t+h}} \sum_{i=1}^{N_{t+h}} e_{i,t+h}^2} \tag{11}$$

- Mean Absolute Percentage Error:

$$\text{MAPE}_{t+h} = \frac{1}{N_{t+h}} \sum_{i=1}^{N_{t+h}} \left| \frac{e_{i,t+h}}{y_{i,t+h}} \right| \times 100 \tag{12}$$

Throughout the paper, the following sign convention is maintained: negative forecast errors (ME < 0) indicate that forecasts exceeded realisations. For GDP growth and revenues, this corresponds to optimistic projections, meaning that governments overestimated growth or fiscal capacity. For expenditure, negative errors indicate that forecast spending exceeded actual spending-that is, governments overestimated their expenditure relative to what was realised.

[7] There is no consensus in the literature on a single dominant metric for evaluating forecast performance; see Tofallis (2015).

Conversely, positive forecast errors (ME > 0) indicate that realisations exceeded forecasts. This reflects pessimistic or conservative projections. Under this definition, systematic optimism bias, the central concern of the political economy literature, manifests as negative mean errors in growth and revenue forecasts. By contrast, MAE, RMSE, and MAPE measure forecast accuracy independently of error direction. MAE and RMSE are expressed in the same units as the outcome variable (percentage points for GDP growth; percentage points of GDP for revenues and expenditures), while MAPE scales errors relative to the realised outcome, facilitating cross-country and cross-variable comparisons. RMSE assigns greater weight to large forecast misses, making it particularly sensitive to extreme errors and tail events.

As a complementary diagnostic, this section tests for forecast unbiasedness and efficiency following Holden and Peel (1990). Using pooled country–year observations of first-vintage official forecasts and their corresponding realisations, this paper estimates the following:

$$y_{i,t} = \alpha + \beta f_{i,t} + u_{i,t}, \tag{13}$$

where $y_{i,t}$ denotes the realized outcome and $f_{i,t}$ the forecast used during budget preparation. Unbiased and efficient forecasts require $\alpha = 0$ and $\beta = 1$.

Table 2 reports joint tests of these restrictions for GDP growth, government revenues, and government expenditures. The test is based on signed forecast errors (ME), which capture systematic bias, rather than accuracy measures (MAE/RMSE). The null hypothesis of unbiasedness and efficiency is rejected at the 1% level for all variables. This indicates that deviations between forecasts and realisations are systematic rather than random, consistent with the political economy literature on optimism bias in official macro-fiscal projections (Frankel, 2011; Beetsma et al., 2009).

Table 2: Joint test of macro and fiscal forecasts unbiasedness and efficiency in EU and LAC countries from 1998 to 2019.

Forecast Variable	**F-statistic**	**df (num, denom)**	**p-value**
GDP Growth	24.01	(2, 2994)	0.0000
Revenues	120.80	(2, 2824)	0.0000
Expenditures	48.26	(2, 2809)	0.0000

Notes: Up to 5-year average horizon forecasts, annual frequency. The table reports joint F-tests of forecast unbiasedness and efficiency based on regressions of realised outcomes on forecasted values. The null hypothesis is that forecasts are unbiased and efficient, implying an intercept of zero and a slope of one. Rejection of the null indicates systematic forecast bias and/or inefficiency. All tests strongly reject the null, consistent with pervasive forecast optimism and informational inefficiencies in government projections.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

4.3 Stylised facts on macro-fiscal forecast deviations

Before turning to the causal analysis, this section documents a set of stylised facts on macro-fiscal forecast deviations. These patterns show that forecast errors in growth, revenue, and expenditure are systematic, persistent, and strongly interconnected. The close co-movement between macroeconomic and fiscal deviations suggests that optimistic assumptions propagate

mechanically into budgetary outcomes, while substantial heterogeneity across countries and over time points to a role for institutional arrangements in shaping forecasting behaviour.

4.3.1 Stylised Fact 1: Systematic forecast bias

Table 3 reports summary statistics for forecast errors in GDP growth, revenues, and expenditures, disaggregated by region. Three regularities emerge. First, forecasts display systematic bias across both regions, with governments tending to overpredict GDP growth and revenues whilst underpredicting expenditures, a pattern consistent with optimistic budgetary projections documented in the political economy literature (Frankel and Schreger, 2016). Second, GDP forecast errors are considerably more dispersed in LAC than in the EU (standard deviation of 5.02 versus 1.60), with pronounced positive skewness in the Latin American sample (6.15 versus -0.03), reflecting extreme growth fluctuations tied to commodity price cycles. Revenue forecast errors follow a similar pattern, though the regional gap narrows. Third, expenditure forecast errors reverse this regional ordering, with lower dispersion in LAC (1.95) than in the EU (2.79).

Despite confronting higher macroeconomic volatility, LAC governments produce expenditure forecasts with tighter dispersion, suggesting that institutional features, namely numerical fiscal rules imposing strict expenditure ceilings, constitutional earmarking provisions, and weaker automatic stabilisers, constrain forecast variability independently of economic uncertainty. EU expenditure forecasts, by contrast, exhibit higher dispersion because fiscal frameworks accommodate countercyclical flexibility and discretionary stimulus. This divergence indicates that expenditure forecast dispersion reflects design choices about fiscal flexibility rather than forecasting capacity alone.

Finally, the pronounced kurtosis visible in Table 3, particularly for LAC GDP errors (exceeding 70), indicates that extreme forecast failures are recurring features rather than rare tail events, motivating the use of accuracy measures that do not asymmetrically penalise extreme errors in Section 7.

Table 3: Descriptive statistics for macro-fiscal deviations in EU and LAC countries from 1998 to 2019.

	Total Sample					LAC					EU				
Variable	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.
GDP	810	-0.36	2.92	7.95	158.20	214	-0.65	5.02	6.15	70.67	596	-0.25	1.60	-0.03	8.37
Revenue	723	-0.30	2.85	2.53	51.37	136	-0.31	3.10	-3.21	20.28	587	-0.30	2.79	4.34	61.70
Expenditure	731	0.45	2.68	0.05	5.55	142	1.31	1.95	-0.33	4.99	589	0.25	2.79	0.19	5.57

Notes: 5-year average horizon forecasts, annual frequency.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

4.3.2 Stylised fact 2: Forecast accuracy and horizon

Table 4 reports descriptive statistics for absolute forecast errors. Forecast accuracy is systematically lower in LAC economies across all variables, as reflected in higher mean absolute errors and greater dispersion. GDP forecasts exhibit the largest absolute errors, especially in LAC, consistent with higher macroeconomic volatility.

Table 4: Descriptive statistics for absolute macro-fiscal deviations in EU and LAC countries from 1998 to 2019.

	Total Sample					LAC					EU				
Variable	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.
Abs. GDP Error	809	2.11	2.66	10.61	193.38	213	3.24	4.49	7.52	82.24	596	1.70	1.33	1.97	8.21
Abs. revenue error	730	1.95	2.21	7.73	105.95	142	2.07	2.43	4.72	34.09	588	1.91	2.15	8.73	132.85
Abs. expenditure error	730	2.30	1.75	1.96	9.66	142	1.91	1.48	1.11	4.07	588	2.39	1.80	2.05	9.96

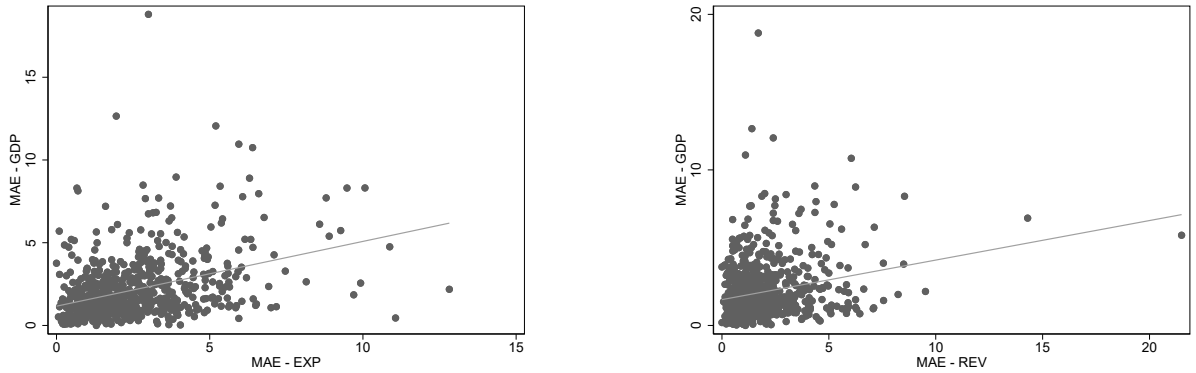
Note: 5-year average horizon forecasts, annual frequency. Statistics shown are calculated on the full sample before winsorization. Extreme kurtosis values reflect genuine crisis episodes. These observations were verified against historical crisis records and are retained in descriptive statistics but winsorized at the 1st and 99th percentiles for econometric estimation.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

Beyond accuracy levels, forecast bias varies with the projection horizon in variable-specific ways. GDP growth optimism intensifies at longer horizons, with mean errors increasing from -0.36 pp at short horizons to -1.03 at longer ones, consistent with governments extrapolating favourable growth assumptions into the medium term. Revenue forecasts display the opposite pattern, shifting from modest optimism (-0.31) to near-unbiased projections ($+0.11$) at longer horizons, suggesting that conservative revenue assumptions may serve as a budgetary cushion when uncertainty compounds over time. Expenditure forecasts exhibit increasing underestimation at longer horizons (from $+0.25$ to $+1.96$), reflecting political difficulties in credibly committing to future spending levels. These horizon-specific patterns suggest that macroeconomic projections are subject to wishful thinking, whilst fiscal projections reflect strategic conservatism or commitment problems depending on the variable. Detailed statistics by horizon are reported in Appendices B.2 and B.3.

4.3.3 Stylised fact 3: Macro–fiscal propagation

Forecast errors also exhibit strong co-movement across macroeconomic and fiscal variables. Figure 3 plots mean absolute errors in GDP growth against those in expenditures and revenues. Both panels display a clear positive association: countries with larger absolute GDP forecast errors also tend to exhibit larger fiscal forecast errors, particularly for revenues, reflecting the mechanical dependence of revenue projections on macroeconomic assumptions. This pattern confirms that errors in GDP projections propagate into fiscal forecasts through mechanical and cyclical channels.

Figure 3: Co-movement between GDP and fiscal forecast deviations, in absolute terms (pp). Full sample from 1998 to 2019



(a) GDP and Expenditure deviations

(b) GDP and Revenue deviations

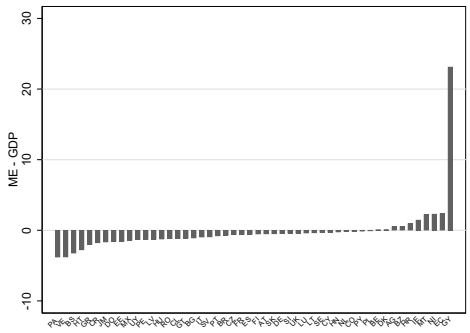
Notes: Deviations are defined as absolute forecast errors, measured as the absolute difference between first-vintage official forecasts used in budget preparation and realised outcomes, expressed in percentage points. Forecasts refer to one-year-ahead projections. The solid line represents a linear fit for illustrative purposes.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

As a complementary exercise, Appendix B.4 compares government forecasts with projections from external institutions, such as the IMF and the European Commission, and confirms the presence of systematic forecast bias.

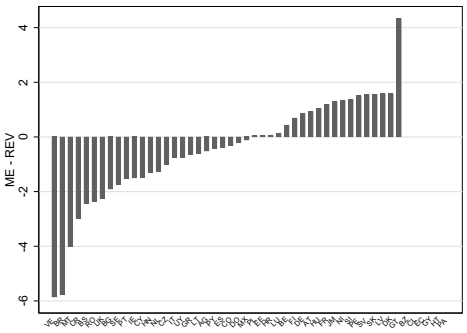
4.3.4 Stylised fact 4: Cross-country heterogeneity

Figure 4 reports country-level mean forecast errors averaged over a five-period horizon. GDP growth forecasts exhibit predominantly negative mean errors across countries, confirming that optimism bias is pervasive rather than region-specific. Revenue forecasts closely track GDP errors, reflecting the mechanical dependence documented in Stylised Fact 3. Expenditure forecasts display a distinct pattern: mean errors are more dispersed and frequently positive, indicating systematic underestimation of realised spending that worsens fiscal outcomes as actual expenditure exceeds budgeted levels. The directional biases in revenue and expenditure thus compound rather than offset each other, both contributing to ex post fiscal deterioration. Figure 5 presents the corresponding mean absolute errors. A notable finding is that expenditure forecasts, while appearing directionally conservative, do not convert this conservatism into greater accuracy, as absolute errors frequently exceed 2–4 percentage points and often surpass those observed for GDP and revenues.

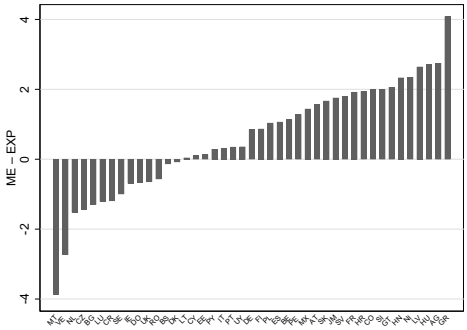
Figure 4: Average GDP growth, revenue, and expenditure forecast deviations by country, 1998-2019.



(a) Mean Error GDP



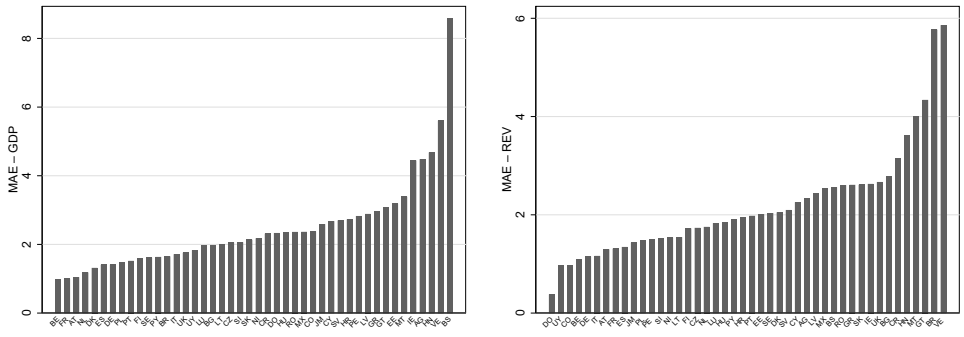
(b) Mean Error revenue



(c) Mean Error expenditure

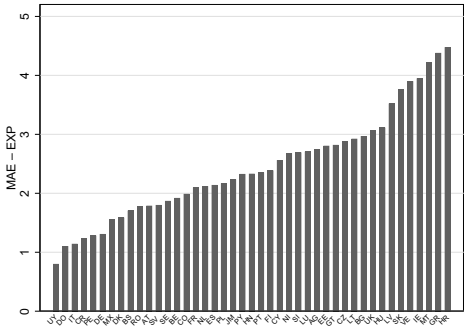
Notes: Bars report the average forecast deviation in pp by country over a five-year horizon.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

Figure 5: Mean absolute forecast deviations in GDP growth, revenue, and expenditure, 1998-2019.



(a) Mean Absolute Error GDP

(b) Mean Absolute Error Revenue



(c) Mean Absolute Error Expenditure

Notes: Bars report the mean absolute forecast error in pp by country over a five-year horizon.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

4.3.5 Stylised fact 5: Forecast accuracy before and after IFI adoption

The preceding facts document pervasive bias and heterogeneity in fiscal forecasting and therefore raise a natural question about whether the emergence of IFIs has been accompanied by any visible change in forecast performance. This subsection provides a preliminary, unconditional comparison without claiming causal interpretation. The formal identification strategy is presented in Section 6.

Table 5 reports mean errors and mean absolute errors before and after IFI activation. The average optimistic bias in GDP growth forecasts falls from −0.70 pp before adoption to −0.27 afterwards, while the MAE declines from 2.62 to 1.98, suggesting both reduced bias and improved accuracy. Revenue forecasts exhibit the most striking shift, with the mean error dropping from −0.55 to near zero (−0.04) and the MAE declining from 2.28 to 1.66. Expenditure forecasts display a more modest improvement, with the mean error declining from 0.74 to 0.42 and the MAE from 2.60 to 2.30, consistent with the role of budgetary rules and procedural constraints that shape spending projections independently of forecast oversight.

These comparisons must be read alongside Table 6, which reports the dispersion of forecast errors before and after 2009, a period during which most IFI adoptions occurred. The standard deviation of GDP growth errors rises sharply from 2.24 to 3.44, reflecting increased macroeconomic

volatility associated with the Global Financial Crisis and the European sovereign debt crisis. In contrast, the dispersion of revenue and expenditure errors *decreases* after 2009 (from 2.93 to 2.25 for revenue, from 3.00 to 2.78 for expenditure). This diverging pattern, greater volatility in GDP forecasts but tighter dispersion in fiscal forecasts, is consistent with the hypothesis that IFI oversight may help contain the transmission of macroeconomic shocks into fiscal projections, although it could also reflect learning effects or post-crisis institutional reforms.

These patterns are suggestive but must be interpreted with caution, since countries that adopted IFIs may systematically differ from those that did not, and the timing of adoption often coincided with broader institutional reforms. The formal identification strategy in Section 6 addresses these concerns.

Table 5: Forecast errors before and after IFI adoption, full sample, 1998-2019.

	Before IFI			After IFI		
Variable	ME	MAE	N	ME	MAE	N
GDP growth (%)	-0.695	2.617	407	-0.267	1.976	403
Revenue (% GDP)	-0.547	2.282	346	-0.038	1.664	383
Expenditure (% GDP)	0.737	2.598	348	0.420	2.295	382

Notes: ME = Mean Error; MAE = Mean Absolute Error. "Before IFI" corresponds to $ifi = 0$ and "After IFI" to $ifi = 1$. 5-year horizon.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

Table 6: Dispersion of forecast errors before and after 2009, full sample, 1998-2019.

	≤ 2009		> 2009	
Variable	SD	N	SD	N
GDP growth error (ME)	2.244	263	3.437	547
GDP growth error (MAE)	1.810	263	3.119	547
Revenue error (ME)	2.928	246	2.250	483
Revenue error (MAE)	2.073	246	1.542	483
Expenditure error (ME)	3.001	241	2.778	489
Expenditure error (MAE)	1.872	241	1.824	489

Notes: SD = standard deviation; N = number of observations. 5-year horizon.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

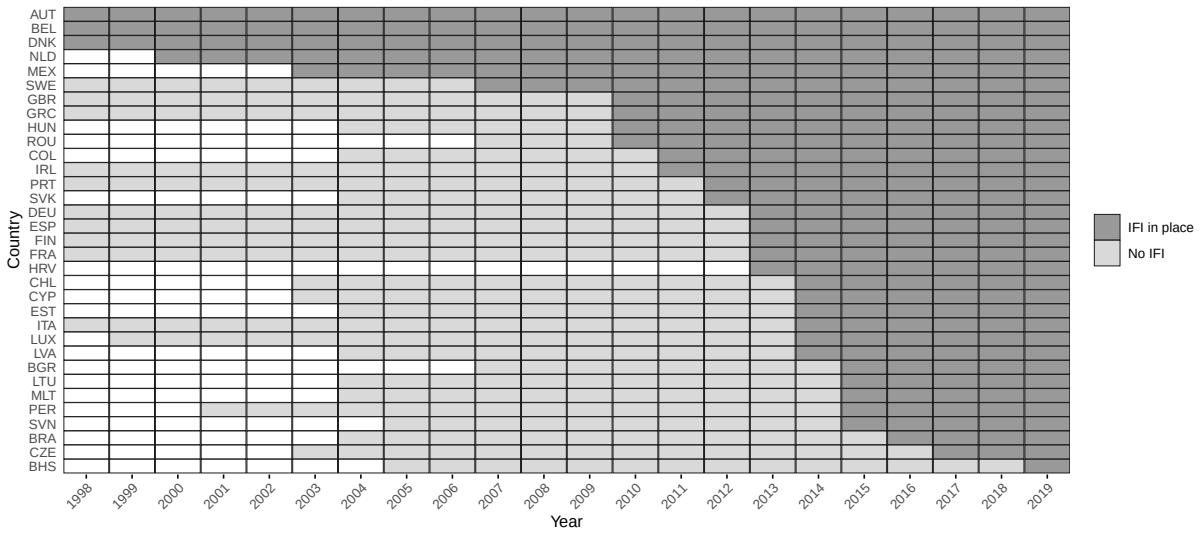
5 Empirical strategy

This section describes the dataset, variable construction, and sample restrictions used to identify the effect of IFIs' implementation on official forecast accuracy.

The main dependent variable of this paper is the mean absolute forecast error from our novel dataset presented in Section 4.1. Our empirical strategy exploits the staggered implementation of IFIs across countries and over time. We construct a binary treatment variable that equals 1 in the year an IFI becomes operational and 0 otherwise. The timing of IFI adoption is based on IMF and OECD fiscal council databases, which document the year in which councils began producing regular analyses and engaging in the budget process[8].

Figure 6 illustrates the distribution of IFI activations across countries and time. Two features are particularly relevant for identification. First, IFI adoption is highly heterogeneous, with substantial variation both across regions and over time. Second, activations are clustered in the mid-2000s and post-global financial crisis period, providing meaningful within-country pre- and post-treatment variation. This staggered adoption pattern allows us to compare forecast performance before and after IFI activation within countries, while controlling for common shocks and time trends.

Figure 6: Timing of independent fiscal institution implementation in EU and LAC countries, 1998-2019



Source: Authors' compilation based on IMF and OECD fiscal council databases. Only countries that implemented an IFI are presented.

We focus on IFI activation, rather than on cross-sectional differences in institutional design, for two reasons. First, IFIs are intended to affect fiscal outcomes primarily once they become operational and begin interacting with the budget process. Second, focusing on activation provides a transparent, policy-relevant definition of treatment that aligns naturally with an event-study and difference-in-differences framework. In additional specifications, we examine

[8]Appendix C provides the full list of IFIs included in the sample, their year of operationalization, regional coverage, and information on whether the institution produces or endorses official macroeconomic and/or fiscal forecasts.

heterogeneity across a limited set of mandate characteristics, whereas the baseline analysis treats IFI activation as the central institutional shock.

To account for confounding factors that may influence forecasting behaviour, we include a set of political, fiscal, and macroeconomic controls. Political variables capture electoral incentives and government structure through an election-year dummy and an indicator of government majority or fragmentation (Scartascini et al., 2021). Fiscal controls account for the institutional and budgetary environment, including the presence of national fiscal rules (such as debt, deficit, or expenditure ceilings) using indices from the IMF Fiscal Rules Dataset, as well as public debt and the fiscal balance (both as a share of GDP). We further control for contemporaneous economic conditions at the time of adoption, including GDP growth, inflation, and the current account balance. Finally, to isolate institutional effects from periods of heightened macroeconomic stress, we include a crisis dummy capturing financial, debt, currency, or banking crises following Nguyen et al. (2022). The descriptive statistics of the control variables are available in Appendix B, Table B.1.

5.1 Causal identification: dynamic effects of IFI activation

To identify the causal impact of IFIs on fiscal forecast performance, we exploit the staggered timing of IFI activation across countries using the difference-in-differences framework developed by Callaway and Sant'Anna (2021). As illustrated in Figure 6, the distribution of IFI adoption over time shows that countries adopt IFIs at different points in the sample period, with substantial heterogeneity across early adopters, late adopters, and countries that remain untreated. This staggered adoption pattern provides the empirical variation required to identify dynamic treatment effects. This approach is particularly well suited to settings with heterogeneous treatment effects and staggered adoption, and avoids the biases that arise in standard two-way fixed effects estimators under such conditions (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2021). It allows treatment effects to vary across cohorts and over time, which is essential in the context of institutional reforms whose effects may unfold gradually rather than instantaneously.

The treatment is defined as the activation of an operational IFI. A country is considered treated from the year in which its IFI becomes operational onward. Untreated observations include both countries that never adopted an IFI during the sample period and countries that have not yet adopted one. The estimator also removes countries that are already treated in the first period of the dataset. The outcome variable Y_{it} captures fiscal forecast performance and includes measures of forecast accuracy (mean absolute forecast errors, root mean square error, and mean absolute percentage error), as defined in Section 4.2.

The primary parameters of interest are the group-time average treatment effects on the treated, $ATT(g,t)$, which measure the effect of IFI activation for cohort g, defined as countries that activate an IFI in period g, at each subsequent period $t \geq g$. Letting $Y_{it}(1)$ and $Y_{it}(0)$ denote the potential outcomes with and without IFI activation, respectively, the group-time average treatment effect is defined as:

$$ATT(g,t) = \mathbb{E}[Y_{it}(1) - Y_{it}(0) \mid G_i = g]\,. \tag{14}$$

Identification of $ATT(g,t)$ relies on a conditional parallel trends assumption.

Specifically, conditional on a set of observable covariates X_{it} and relative to a valid control group $\mathcal{C}$ (consisting of never-treated and not-yet-treated countries), untreated potential outcomes for treated and control units are assumed to follow the same evolution from the pre-treatment period to period t. Formally:

$$\mathbb{E}[Y_{it}(0) - Y_{i,g-1}(0) \mid G_i = g, X_{it}] = \mathbb{E}[Y_{it}(0) - Y_{i,g-1}(0) \mid i \in \mathcal{C}, X_{it}], \quad \forall t \geq g-1. \tag{15}$$

The vector X_{it} includes the political, fiscal, and macroeconomic controls described in Section 4.1, allowing for differential trends driven by observable factors that may affect forecasting behaviour. This conditional formulation is particularly relevant in the present context, as the activation of IFIs may coincide with broader fiscal reforms or macroeconomic adjustments.

Under this assumption, the group-time treatment effects are estimated by comparing treated cohorts with appropriate control units at each time point. The cohort-specific $ATT(g,t)$ estimates are then aggregated, weighted by cohort size, to obtain dynamic treatment effects relative to the time of IFI activation and overall average treatment effects. These estimates provide a direct assessment of how fiscal forecast bias and accuracy evolve following IFI activation, and allow us to evaluate whether IFIs effectively discipline the forecasting process in line with the mechanisms discussed in the theoretical framework.

6 Results

The analysis now turn to the presentation of the main empirical findings on the causal effect of IFI implementation on government forecast accuracy. The analysis begins by documenting the baseline dynamic treatment effects using the Callaway and Sant'Anna (2021) estimator, focusing on how forecast accuracy evolves following IFI activation. The analysis then examines heterogeneity across adoption cohorts to assess whether the magnitude and timing of the effects differ across implementation waves.

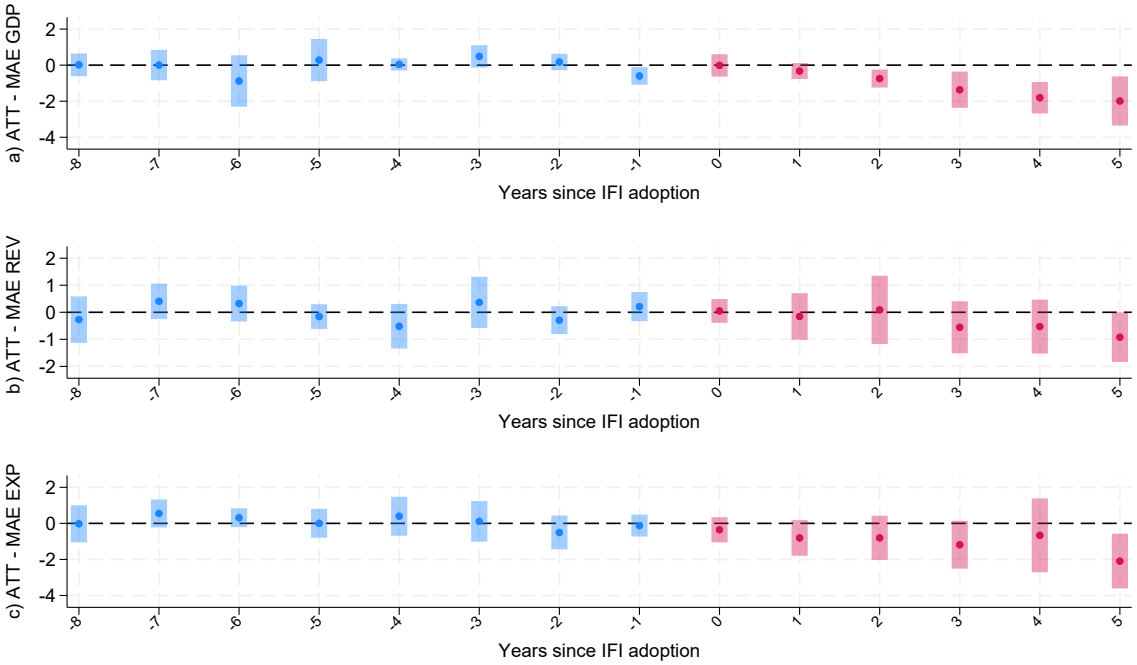
6.1 Causal effects of IFI implementation

Figure 7 reports the event-study estimates from the baseline difference-in-differences specification for GDP growth, revenue, and expenditure forecast accuracy, measured by MAE. The results provide strong support for both the identifying assumptions and the hypothesis that IFI activation causally improves government forecast performance.

Across all three panels, the pre-treatment coefficients are tightly centred around zero and display no systematic trend. This pattern provides strong visual and statistical support for the conditional parallel trends assumption underlying the identification strategy. In the absence of IFI activation, treated and control countries would have followed similar trajectories in forecast accuracy, suggesting that post-treatment differences can be credibly attributed to IFI implementation rather than pre-existing trends.

Following IFI activation (event time $t = 0$), treatment effects become increasingly negative over time, indicating sustained improvements in forecast accuracy. The timing of these effects varies across forecast variables. For GDP growth forecasts (Panel A), statistically significant improvements emerge approximately two years after IFI activation and stabilise thereafter. In contrast, improvements in revenue and expenditure forecast accuracy (Panels B and C) materialise more gradually, becoming pronounced only around five years post-adoption.

Figure 7: Dynamic Effects of IFI Adoption on GDP, Revenue, and Expenditure Forecast Accuracy



Notes: The figure reports event-time average treatment effects from the staggered difference-in-differences estimator. Each point represents the average treatment effect at a given relative year to IFI adoption, aggregated across all cohorts that are observed at that horizon. Periods before zero correspond to pre-treatment years; periods after zero correspond to post-treatment years. Grey bars denote 90% confidence intervals. All specifications include the full set of control variables.

This heterogeneity in timing is consistent with the underlying forecasting processes. GDP growth forecasts rely primarily on macroeconomic modelling, where IFI technical scrutiny can exert relatively rapid discipline. Revenue and expenditure forecasts, by contrast, are more deeply embedded in the budgetary process and more directly exposed to political incentives, requiring stronger institutional integration and sustained oversight before measurable improvements emerge.

The delayed effects, particularly for fiscal variables, align with the theoretical mechanism emphasised in Section 3. IFIs operate mainly through credibility, transparency, and reputational channels rather than immediate technical corrections. Governments anticipating public scrutiny by an independent institution face reputational costs when producing systematically biased projections, but these mechanisms take time to become effective. IFIs must first establish analytical capacity, visibility, and a track record of independent assessment before their presence meaningfully disciplines fiscal forecasting behaviour (Debrun et al., 2013; Beetsma et al., 2022).

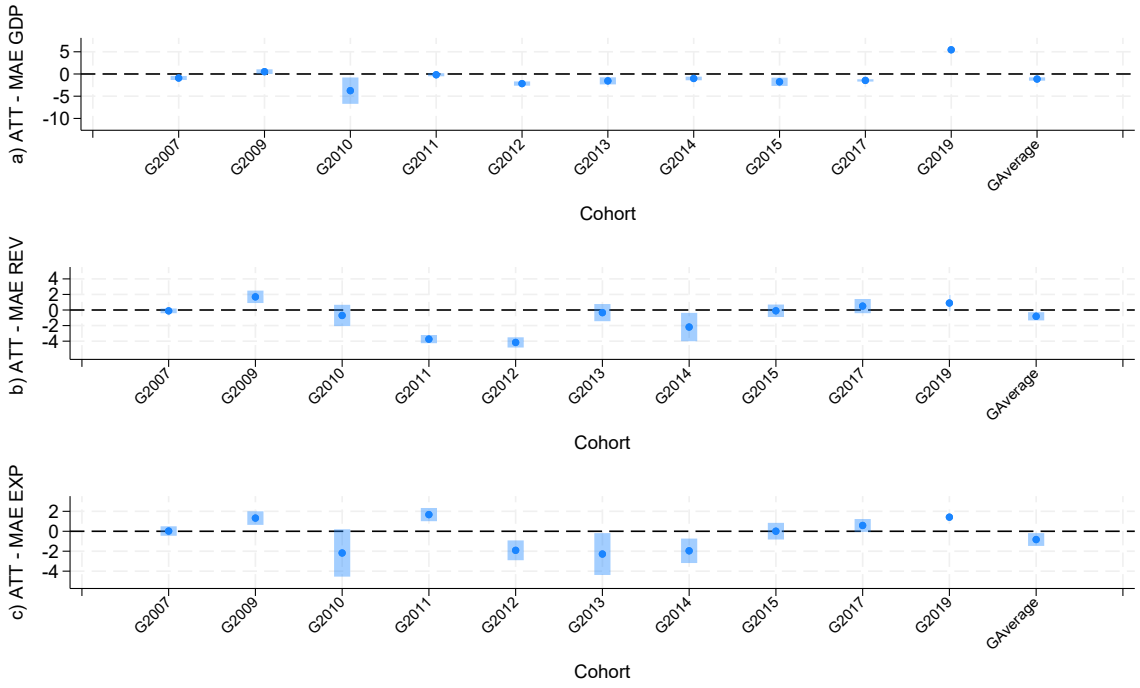
6.2 Heterogeneity across adoption cohorts

Figure 8 reports cohort-specific[9] average treatment effects, decomposing the overall impact of IFI activation by the year in which countries became treated. Each point corresponds to the group-level average treatment effect for a given adoption cohort, aggregated over the available post-treatment periods and relative to the set of countries that had not yet activated an IFI at that time. Each point represents the cohort-average effect, computed as:

$$\overline{ATT}(g) = \frac{1}{|\mathcal{T}_g|} \sum_{t \in \mathcal{T}_g} ATT(g,t) \tag{16}$$

where ATT(g,t) is the group-time specific treatment effect for cohort g at post-treatment period t, and $\mathcal{T}_g$ denotes the set of available post-treatment years for cohort g. Earlier-adopting cohorts are observed through longer post-treatment windows, contributing more time periods to the average.

Figure 8: Effects of IFI Activation on GDP, Revenue, and Expenditure Forecast Accuracy by Adoption Cohort



Notes: The figure reports cohort-specific average treatment effects ($ATT(g,t)$) from the staggered difference-in-differences estimator, where cohorts are defined by the year of IFI activation. Each point corresponds to the average post-treatment effect for a given adoption cohort. Shaded areas denote 90% confidence intervals. All specifications include the full set of control variables.

Two features merit attention. First, estimated effects are predominantly negative across

[9]2007 cohort: Sweden; 2009 cohort: Hungary; 2010 cohort: UK, Greece, Romania; 2011 cohort: Colombia, Ireland; 2012 cohort: Portugal, Slovakia; 2013 cohort: Germany, Spain, Finland, France, Croatia; 2014 cohort: Chili, Cyprus, Estonia, Italy, Luxemburg; 2015 cohort: Bulgaria, Lithuania, Malta, Peru, Slovenia; 2016 cohort: Brazil; 2017 cohort: Czechia; 2019 cohort: Bahamas.

cohorts, indicating that forecast accuracy improvements are not driven by a single adoption wave but emerge consistently across implementation cohorts spanning nearly two decades. This cross-cohort consistency strengthens confidence in the causal interpretation of the baseline results. Second, earlier adopters display considerable heterogeneity in magnitude, with some variation appearing to reflect earlier versus later adoption timing, though the relationship is not monotonic. On average, cohorts adopting IFIs earlier display somewhat larger negative point estimates than cohorts, particularly for revenue and expenditure forecasts. This gradient, however, reflects primarily the interaction between the dynamic response documented in Figure 7 and the finite sample window, rather than genuine cohort-level differences. Because IFI effects strengthen over time (Figure 7), and because early adopters are observed in more mature post-treatment years, whilst later cohorts are observed only in initial years, cohort-specific averages mechanically increase with earlier adoption.

The apparent heterogeneity thus recapitulates the within-cohort dynamics already established, rather than revealing distinct institutional trajectories. Substantial heterogeneity also exists within the early and late cohort groups, reflecting country-specific institutional contexts, ex ante forecasting capacity, and political environments that shape the rate at which IFI credibility is established. The cohort analysis shows that improvements following IFI activation are consistent across diverse adoption contexts, supporting the external validity of the baseline findings.

7 Robustness checks

This section assesses the robustness of the baseline results along two dimensions. First, Subsection 7.1 examines whether the estimated effects of IFI activation on forecast accuracy are sensitive to the choice of identification strategy. While the baseline analysis relies on the Callaway and Sant'Anna (2021) estimator to account for staggered adoption and treatment heterogeneity, alternative designs impose different identifying assumptions and weighting schemes. Comparing results across estimators allows us to verify that the main findings are not driven by a specific empirical implementation.

Second, Subsection 7.2 evaluates the robustness of the results using alternative measures of forecast accuracy. The baseline analysis focuses on mean absolute errors, which capture the magnitude of forecast deviations independently of their direction. The analysis complements this analysis using the RMSE and the MAPE, which place different weights on large deviations and scale errors relative to outcomes. Consistent results across these alternative outcome measures support the interpretation that IFI activation systematically improves fiscal forecast performance.

7.1 Alternative identification strategies

The analysis assesses the robustness of the baseline results using a sequence of alternative identification strategies that rest on distinct and complementary assumptions. Each approach exploits varying sources of identifying variation and imposes different requirements for comparability between treated and control units. This analytical shift is pivotal for validating the core results, as it moves the focus from a single estimator to a consensus across diverse methodologies.

The analysis begins with propensity score matching, which relies on selection on observables

and provides a conservative, static benchmark. While this approach does not exploit the panel structure of the data or dynamic treatment timing, it allows us to verify that the results are not driven by simple differences in observable characteristics between IFI adopters and non-adopters. The analysis then considers alternative difference-in-differences implementations that recover dynamic treatment effects without relying on the baseline aggregation scheme. These approaches provide an intermediate robustness check, relaxing some assumptions while preserving the core panel structure of the analysis. Finally, the analysis turns to PanelMatch, which imposes the most stringent identification requirements by explicitly matching treated and control units on both treatment and covariate histories. Because these requirements are difficult to satisfy in settings with early and widespread treatment adoption, this estimator is implemented on a restricted subsample where overlap conditions are more plausible. Consistency of results under this demanding design provides strong support for the causal interpretation of the baseline findings.

7.1.1 Propensity Score Matching

As a first robustness exercise, the analysis evaluates whether the baseline results persist under an identification strategy that relies on selection on observables rather than on difference-in-differences assumptions. Specifically, Propensity Score Matching is implemented (PSM) to compare countries that activate IFI with observationally similar countries that do not, conditioning on pre-treatment political, fiscal, and macroeconomic characteristics measured in the year prior to activation. This approach provides a conservative benchmark that abstracts from dynamic treatment effects and relies exclusively on cross-sectional comparability, thereby offering a useful lower-bound check on the baseline causal estimates.

The motivation for this exercise is that IFI adoption is unlikely to be random. Countries may be more likely to activate IFIs following fiscal crises, political transitions, or external pressures related to European integration or fiscal surveillance reforms (Beetsma et al., 2019; Debrun and Kinda, 2017). If such factors are correlated with forecast accuracy, baseline estimates could partly reflect selection effects. PSM provides a useful benchmark by explicitly conditioning on observable determinants of IFI adoption.

PSM is implemented in two steps, with particular attention to the timing of covariates to avoid conditioning on post-treatment outcomes. First, propensity scores are estimated using a logit model that predicts the probability of IFI adoption as a function of pre-treatment covariates measured in year t-1, where t denotes the year of potential IFI activation. The covariate set includes lagged variables presented in Section 4. This temporal restriction is particularly important for fiscal outcome variables such as public debt and fiscal balance, which may themselves respond to IFI oversight, as well as for fiscal rules, which are sometimes adopted as part of the same reform package as IFIs. By conditioning only on lagged values, the propensity score reflects the observable factors that predict IFI adoption without mechanically absorbing treatment effects through the conditioning strategy, thereby preserving the integrity of the selection-on-observables assumption underlying PSM identification. Second, treated observations are matched to untreated observations with similar propensity scores, and the average treatment effect on the treated (ATT) is computed as the difference in forecast accuracy between matched

groups (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 2002; Smith and Todd, 2005).

A key challenge in our setting is the presence of time-varying common shocks that may simultaneously affect IFI adoption propensity and forecast performance across countries. To address this issue whilst maintaining the pre-treatment timing of covariates, two complementary matching strategies are implemented. First, nearest-neighbour matching is performed with one neighbour, estimating the propensity score on the full sample using t-1 covariates whilst including year fixed effects in the logit specification. This approach controls for common time shocks affecting all countries without compromising the pre-treatment restriction on conditioning variables. Second, a within-year matching strategy is implemented to better absorb common factors. For each calendar year t, this paper estimates propensity scores using covariates measured in year t-1 and perform matching between treated and untreated countries observed in that year. This approach ensures that treated countries in, for instance, 2013 are matched only to control countries also observed in 2013, using their respective 2012 covariate values to predict treatment probability. The year-specific ATT estimates are then pooled using inverse-variance weighting, allowing common factors such as global financial conditions or coordinated EU fiscal reforms to be absorbed through temporal stratification whilst maintaining the pre-treatment timing of all conditioning variables.

Table 7 reports the resulting ATT estimates for mean absolute forecast errors. Two features warrant attention. First, the within-year pooled matching approach (column 2) yields negative, statistically significant effects across all three variables, with IFI adoption associated with reductions of approximately 0.30 percentage points in GDP forecast errors, 0.54 percentage points in revenue forecast errors, and 0.53 percentage points in expenditure forecast errors. These estimates are precisely estimated and corroborate the baseline findings documented in Section 6. Second, the nearest-neighbour matching with year fixed effects (column 1) produces estimates of the same sign for GDP and expenditure but with substantially larger standard errors and loss of statistical significance. This specification suffers from limited overlap, as evidenced by the small ratio of matched control observations (approximately 80) to treated observations (approximately 280). This imbalance indicates that the control pool is reused extensively, with each untreated observation matched to multiple treated units, raising concerns about the quality of the matches and potential violations of the common support assumption. Balance diagnostics (reported in Appendix B, Table B.7) confirm that the within-year matching specification achieves adequate covariate balance, with standardised mean differences below 0.10 for all covariates except lagged public debt (0.13), which remains within the conventional tolerance threshold of 0.25.

By contrast, full-sample nearest-neighbour matching yields substantially larger imbalances, particularly for the fiscal rules index and inflation, reflecting the structural differences between IFI-adopting and non-adopting countries that cross-sectional matching cannot fully absorb. These diagnostics support the use of the within-year specification as the preferred PSM robustness check. Given these diagnostics, primary emphasis is placed on the within-year pooled estimates in column 2, which represent the preferred PSM specification. The consistency of these estimates with the baseline results strengthens confidence that IFI effects on forecast accuracy are robust to alternative identification strategies, even when relying solely on pre-treatment selection on observables rather than difference-in-differences assumptions. The column 1 results, whilst

directionally consistent, primarily serve to illustrate the importance of achieving adequate overlap in PSM designs, particularly in settings where treatment adoption is clustered, and the pool of suitable control units is constrained.

Table 7: Average treatment effects of IFI implementation on mean absolute forecast errors in EU and LAC countries, from 1998 to 2019: PSM estimates.

	(1) **NN(1) + Year FE**	**(2)** **Within-Year Pooled**
`Variable: GDP`		
[1] ATT	-0.690	-0.300***
	(0.794)	(0.052)
$N_{Treated}$	279	251
$N_{Controlled}$	81	245
`Variable: Revenue`		
[2] ATT	0.334	-0.544***
	(0.459)	(0.130)
$N_{Treated}$	277	272
$N_{Controlled}$	80	275
`Variable: Expenditure`		
[3] ATT	-0.325	-0.525***
	(0.325)	(0.130)
$N_{Treated}$	277	272
$N_{Controlled}$	80	275

Notes: ATT denotes the average treatment effect on the treated, with MAE as the outcome variable. Column (1) reports nearest-neighbour propensity score matching with one match and year fixed effects included in the propensity score estimation. Column (2) reports within-year propensity score matching, where treated and control units are matched within the same calendar year, and year-specific ATTs are pooled using inverse-variance weighting. Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. NN(1) refers to the first nearest-neighbour matching.

7.1.2 Local Projection Difference-in-Differences

As a second robustness exercise, The analysis assesses the sensitivity of the baseline results using the Local Projection Difference-in-Differences (LP-DiD) estimator proposed by Dube et al. (2025). This approach provides a flexible alternative to standard event-study designs by estimating dynamic treatment effects through a sequence of local projections, without imposing parametric restrictions on the shape or persistence of the treatment response.

LP-DiD is particularly well-suited to this setting for three reasons. First, it allows treatment effects to evolve gradually over time without imposing parametric restrictions on the shape

of the response path, which is appropriate for institutional reforms whose impact materialises through learning, credibility, and reputation building. Second, the estimator accommodates rich sets of time-varying controls and their dynamics, making identification robust to differential pre-treatment trends driven by observable factors. The parallel trends assumption is conditional on these controls, requiring that treated and control units follow similar trajectories in the absence of treatment after conditioning on covariate histories. Third, by estimating effects separately at each horizon through local projections, LP-DiD avoids imposing homogeneity across event times and permits flexible nonlinear dynamics that may be obscured in pooled specifications.

Specifically, for each post-treatment horizon $h \in \{0, 1, \ldots, 5\}$, this paper estimates a separate regression of the form:

$$\Delta Y_{i,t+h} = \alpha_i^h + \gamma_t^h + \beta_h D_{it} + X_{it}'\delta_h + \varepsilon_{i,t+h}^h \tag{17}$$

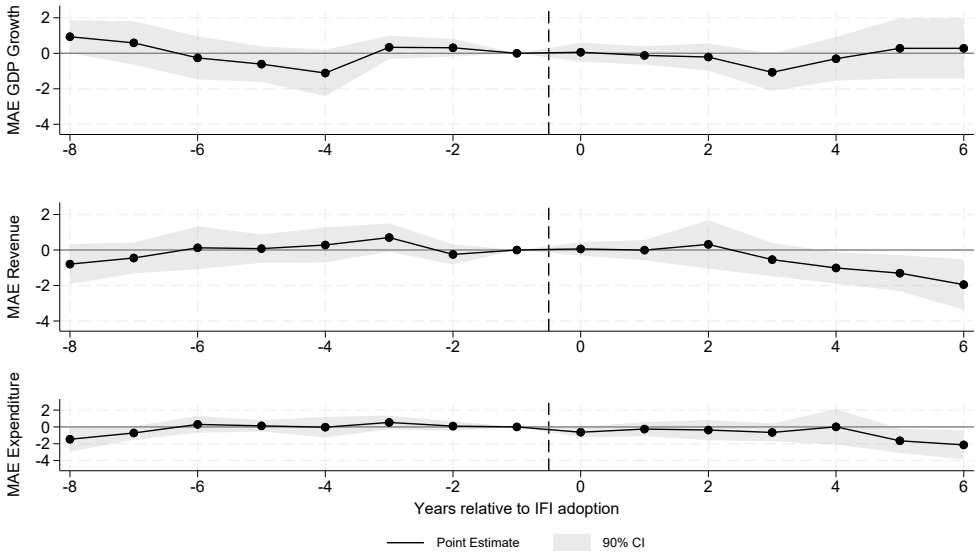
where $\Delta Y_{i,t+h} = Y_{i,t+h} - Y_{i,t-1}$ denotes the change in mean absolute forecast error from the pre-treatment period to h years after potential IFI activation, D_{it} is an indicator equal to one if country i has an active IFI in year t, X_{it} includes the full set of political, fiscal, and macroeconomic controls described in Section 4, and α_i^h and γ_t^h represent country and year fixed effects that vary by horizon. The coefficient β_h captures the average effect of IFI activation on forecast accuracy h periods after adoption, conditional on covariates. Unlike standard event-study specifications, which jointly estimate all leads and lags in a single regression, LP-DiD estimates a separate regression for each horizon, allowing all parameters to vary flexibly across event times. Standard errors are computed using a bootstrap procedure with 500 replications to account for estimation uncertainty. Figure 9 reports the dynamic LP-DiD estimates for GDP growth, revenue, and expenditure forecast accuracy. The pre-treatment coefficients are stable and display no systematic trends, providing supportive evidence for the identifying assumptions. Following IFI activation, treatment effects emerge gradually. For GDP growth forecasts, improvements in accuracy become visible after a short implementation lag. For revenue and expenditure forecasts, effects strengthen over time, with economically meaningful reductions in forecast errors appearing in the medium to long run.

The dynamic patterns closely mirror those obtained from the baseline estimator. In particular, the absence of sharp discontinuities at the time of adoption and the gradual post-treatment improvements are consistent with an institutional mechanism operating through credibility and external scrutiny rather than through immediate technical corrections. Overall, the LP-DiD results reinforce the interpretation that IFI activation leads to sustained improvements in forecast accuracy and that the baseline findings are not driven by the specific structure of the main difference-in-differences estimator.

7.1.3 PanelMatch

As a final and most stringent robustness check, The analysis assesses the sensitivity of the baseline results using the PanelMatch estimator developed by Imai et al. (2023). PanelMatch is widely regarded as a benchmark design for causal inference in panel settings with staggered treatment adoption, as it combines difference-in-differences with explicit matching on treatment and covariate histories. Compared with previous robustness exercises, this approach imposes

Figure 9: Dynamic effects of IFI implementation on forecast accuracy: Local projection DiD estimates



Notes: Dynamic treatment effects estimated using Local Projection Difference-in-Differences. For each horizon hh h, the figure plots the coefficient β_h from a separate regression of the change in mean absolute forecast error on the IFI treatment indicator and controls. Pre-treatment estimates test for differential trends; post-treatment estimates capture dynamic effects. Standard errors bootstrapped with 500 replications. Vertical bars represent 90% confidence intervals.

substantially stronger comparability and overlap requirements, thereby providing a demanding test of whether the estimated effects of IFI activation persist under highly restrictive identifying assumptions. The applicability of PanelMatch, however, critically depends on the availability of suitable control units with sufficiently long, comparable, untreated histories. In our context, the empirical structure of IFI adoption poses several challenges for implementing this estimator on the full sample.

First, IFI activation is highly front-loaded. A substantial share of countries adopt IFIs early in the sample period, particularly between the late 1990s and the early 2010s, and remain treated thereafter. This early treatment saturation substantially reduces the pool of potential control units available for later periods. Second, treatment is largely absorbing, since once an IFI becomes operational, it is rarely deactivated. As treatment accumulates over time, the number of countries that remain untreated for multiple consecutive periods declines sharply. Third, PanelMatch requires close alignment of treatment histories over a fixed number of pre-treatment lags. When combined with early and widespread adoption, this requirement makes it increasingly difficult to construct matched sets, especially for early and mid-period treated units. In the full EU–LAC sample, these features lead to limited overlap and, in many cases, the absence of valid matched control groups.

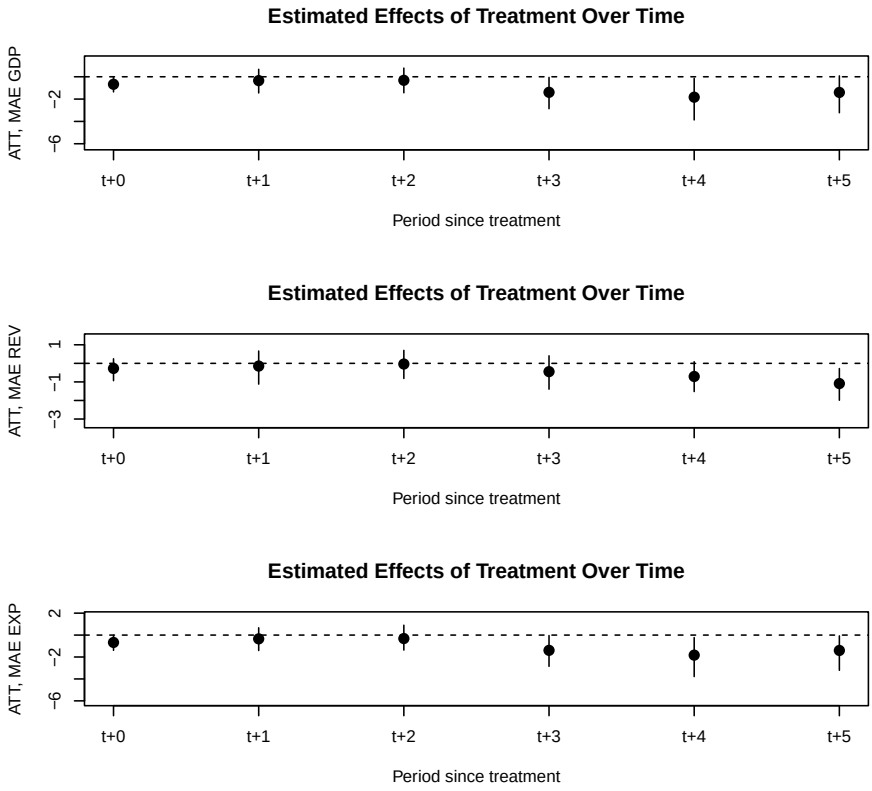
For these reasons, PanelMatch cannot be reliably implemented on the full sample without violating its core overlap and feasibility conditions. Importantly, this limitation reflects the empirical pattern of IFI adoption rather than shortcomings of the estimator itself. The features

that challenge PanelMatch—early, absorbing, and widespread treatment—are well-suited to the Callaway and Sant'Anna estimator, which exploits staggered adoption timing to construct valid comparison groups without requiring a persistent pool of untreated units. This design advantage makes Callaway-Sant'Anna particularly appropriate for institutional reforms that diffuse widely over time. Nevertheless, PanelMatch can be meaningfully applied to a more homogeneous subsample. this robustness exercise is therefore restricted to EU countries, where IFI adoption is more evenly staggered over time, and institutional environments are more comparable. This restriction mitigates concerns related to early treatment saturation and improves the availability of suitable control units with comparable treatment histories. While overlap remains imperfect even within the EU, matched sets can be constructed for a non-trivial subset of treatment onsets, allowing us to assess whether the direction and timing of the baseline effects are preserved under a more restrictive identification strategy.

The PanelMatch results, reported in Figure 10, show dynamic treatment effects that are qualitatively consistent with the baseline estimates in their directional patterns and broad temporal evolution, though precise timing should not be directly compared given the substantial differences in sample composition and identification strategy. Point estimates are predominantly negative across post-treatment horizons, particularly for GDP and expenditure forecast accuracy, indicating improvements following IFI activation.

The estimates reflect the EU subsample only and show substantially smaller effective sample sizes after matching than in the baseline specification.

Figure 10: Estimated average effects of IFI activation on GDP, Revenue, and Expenditure forecast errors (MAE)



Notes: Estimates are based on the PanelMatch estimator using Mahalanobis matching on treatment and covariate histories over the two years prior to IFI activation. The figure reports average treatment effects from the year of implementation ($F = 0$) up to five years after implementation ($F = 5$). Vertical bars indicate 95% confidence intervals obtained via block bootstrap with 1,000 replications.

Several qualitative patterns align with the baseline findings. First, effects for GDP growth forecasts become negative and economically meaningful within the first two to three years following activation, suggesting a relatively rapid response once IFI oversight becomes operational, though the exact horizon at which effects become statistically distinguishable from zero varies with the matching procedure and available sample. Second, revenue and expenditure forecasts display more gradual improvements, with effect sizes strengthening over time and reaching larger magnitudes at horizons of 4 to 5 years post-adoption. This temporal ordering, whereby macroeconomic forecast improvements precede or coincide with fiscal forecast improvements, mirrors the baseline pattern documented in Figure 7. Third, expenditure forecasts show larger and more persistent negative effects than revenue forecasts, which are estimated less precisely and display confidence intervals that cross zero at longer horizons, consistent with the greater difficulty of projecting discretionary spending relative to revenue tied mechanically to GDP.

These patterns confirm the core baseline findings under a more demanding identification design that conditions on detailed pre-treatment histories and restricts inference to matched pairs of observationally similar countries. Overall, despite reduced precision due to smaller effective sample sizes and the stringent nature of the matching procedure, these results confirm that the

baseline findings are not driven by functional form assumptions or by the specific estimator employed in the main analysis.

Beyond serving as a robustness check, this specification also allows us to exploit variation in a key institutional characteristic of IFIs, namely, their formal forecasting remit. In particular, the analysis uses the PanelMatch framework to examine whether the effects of IFI activation differ depending on whether the institution produces or formally endorses official macroeconomic or fiscal forecasts. This distinction is central to the transmission mechanisms emphasised in the theoretical discussion, as mandates over forecasts directly affect the stage at which optimism bias may be constrained. The corresponding analysis, together with a detailed description of the PanelMatch implementation for mandate-specific treatments, is presented in Appendix D.

7.2 Alternative measures of forecast accuracy

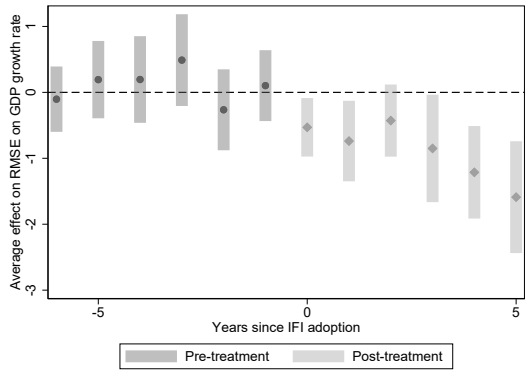
This subsection examines whether the baseline results are sensitive to the choice of forecast accuracy metric. The main analysis focuses on the mean absolute error, which provides a transparent measure of the average magnitude of forecast deviations, independently of their direction. While MAE is widely used in the fiscal forecasting literature, alternative metrics emphasise different dimensions of forecast performance. Re-estimating the baseline specifications using these alternatives allows an assessment of whether the estimated effects of IFI activation are driven by a specific loss function or reflect more general improvements in forecasting performance.

The analysis considers two additional measures commonly used in forecast evaluation (Armstrong and Collopy, 1992; Hyndman and Koehler, 2006). First, the RMSE, which assigns greater weight to large forecast errors and is therefore more sensitive to extreme forecast misses. This metric is particularly relevant in a fiscal policy context, where large forecasting errors can have disproportionate consequences for budget execution and compliance with fiscal rules. Second, the MAPE scales forecast errors by realised outcomes and facilitates comparisons across countries with different economic sizes and across variables measured in different units.

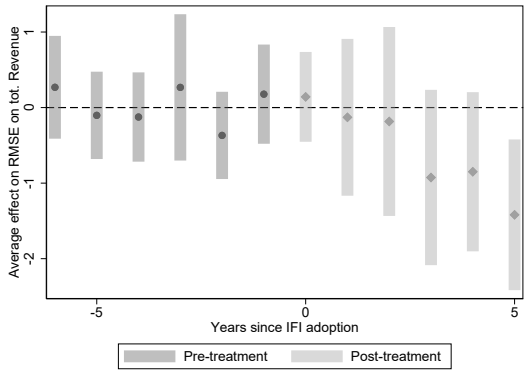
Importantly, changing the error metric also changes the implicit notion of improvement. Under RMSE, improvements reflect a reduction in large, infrequent forecast failures, while under MAPE, improvements indicate greater proportional accuracy relative to economic size. If IFIs primarily operate by disciplining excessive optimism and reducing extreme forecast deviations, their effects would be expected to be particularly visible when using RMSE. If, instead, IFIs improve forecasting practices more broadly and systematically, improvements should also be detectable under MAPE.

Figure 11 reports event-study estimates using RMSE as the dependent variable. Across GDP growth, revenue, and expenditure forecasts, post-treatment coefficients are predominantly negative, indicating reductions in squared forecast errors following IFI activation. The decline is especially pronounced for revenue and expenditure forecasts, where RMSE decreases steadily in the medium to long run after adoption. This pattern suggests that IFIs are effective not only at reducing typical forecast errors but also at constraining large and economically costly forecast misses.

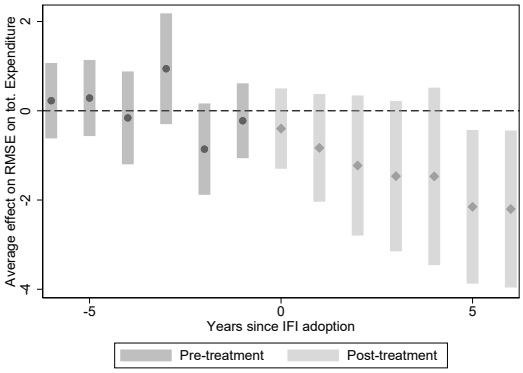
Figure 11: Dynamic effects of IFI adoption on RMSE forecast errors



(a) GDP Growth RMSE



(b) Revenue RMSE



(c) Expenditure RMSE

Notes: Event-study estimates from the staggered difference-in-differences estimator using the RMSE as the dependent variable. Time $t = 0$ denotes IFI activation, with $t = -1$ as the reference period. Point estimates are shown with 90% confidence intervals. All specifications include the full set of control variables.

Pre-treatment coefficients fluctuate around zero and show no systematic trend. Confidence intervals are wider than in the MAE specifications, reflecting the higher intrinsic volatility of RMSE. By construction, RMSE assigns disproportionate weight to extreme forecast realisations and is therefore more sensitive to macroeconomic shocks and tail events. As a result, greater dispersion in both point estimates and confidence intervals is expected and, by itself, does not signal violations of the identifying assumptions. Importantly, there is no evidence of persistent pre-treatment divergence or anticipatory dynamics, supporting the validity of the conditional parallel trends assumption underlying the identification strategy.

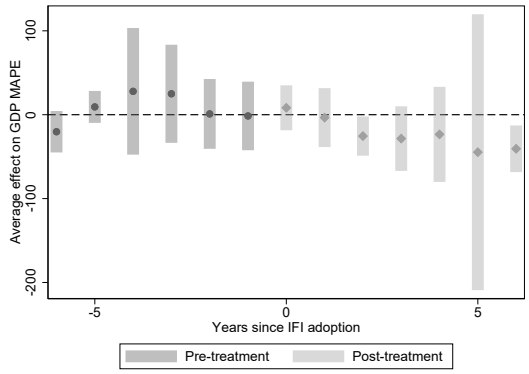
Figure 12 presents the corresponding event-study estimates using MAPE, which scales forecast errors relative to the size of the realised outcome. This metric facilitates cross-country comparisons by normalising errors, but it introduces a well-known measurement challenge. MAPE becomes unstable when realised outcomes are close to zero (Hyndman and Koehler, 2006; Tofallis, 2015). This limitation is particularly salient for GDP growth forecasts (Panel a). GDP growth rates can fluctuate near zero or turn negative during recessions, making the denominator of the MAPE formula arbitrarily small and causing percentage errors to explode. The resulting estimates display extremely wide confidence intervals and an explosive y-axis scale (ranging from -200 to

+100), indicating that a small number of observations with near-zero realised growth dominate the calculation. Under these conditions, the GDP growth MAPE specification provides little additional information beyond the main MAE and RMSE results. These estimates are therefore not interpreted as the GDP MAPE estimates as supportive evidence, but are presented for completeness.

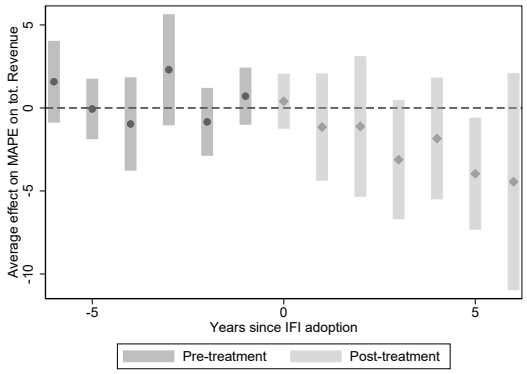
By contrast, the MAPE estimates for revenue and expenditure (Panels b and c) are considerably more informative. Both fiscal variables are measured as shares of GDP and rarely approach zero, making the percentage-error metric well-defined and reasonably scaled. For revenues, post-treatment coefficients shift systematically negative, with effect sizes of approximately 4 percentage points at longer horizons. Expenditure MAPE displays a similar pattern, with improvements emerging 5 periods. These fiscal MAPE results corroborate the baseline findings using alternative metrics that are not sensitive to absolute scales or GDP volatility. Taken together, the robustness checks using RMSE and MAPE confirm that the baseline MAE results are not artefacts of a particular error metric. IFI activation is associated with improvements across measures that emphasise large forecast misses (RMSE) and proportional accuracy for fiscal variables (MAPE), while maintaining the central finding that fiscal forecast improvements materialise more gradually than macroeconomic forecast improvements

As with RMSE, confidence intervals widen at longer horizons, reflecting both the accumulation of uncertainty over time and the sensitivity of percentage-based errors to small denominators in adverse macroeconomic conditions. Despite this increased variability, the post-treatment pattern remains consistent with the MAE and RMSE results. Taken together, these alternative specifications confirm that the baseline findings are not specific to a particular error metric. IFI activation is associated with improvements in forecast accuracy across measures that emphasise average deviations, extreme errors, and proportional accuracy, strengthening the interpretation that IFIs contribute to a broad-based and economically meaningful improvement in fiscal forecasting practices.

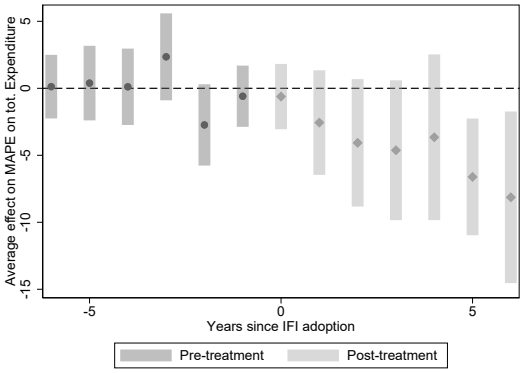
Figure 12: Dynamic effects of IFI adoption on MAPE forecast errors



(a) GDP Growth MAPE



(b) Revenue MAPE



(c) Expenditure MAPE

Notes: Notes. Event-study estimates from the staggered difference-in-differences estimator using MAPE as the dependent variable. Panel (a) for GDP growth displays very wide confidence intervals due to MAPE's sensitivity to near-zero denominators; these estimates should be interpreted with caution. Panels (b) and (c) for revenue and expenditure are better-suited to MAPE measurement as fiscal shares of GDP rarely approach zero. Time t = 0 denotes the adoption of IFI. Point estimates are reported with 90% confidence intervals. All specifications include the full set of control variables.

8 Conclusion

This paper examines whether IFIs implementation causally improve the accuracy of official macroeconomic and fiscal forecasts. The analysis combines a simple theoretical framework with a novel cross-regional dataset of first-vintage forecasts and modern causal inference methods designed for staggered treatment adoption. The evidence suggests that IFI activation is associated with statistically and economically meaningful reductions in forecast errors for GDP growth, government revenues, and government expenditures. These findings emerge from an analysis of 28 EU member states and 27 LAC economies over the period 1998–2019, exploiting variation in the timing of IFI adoption across countries.

The improvements in forecast accuracy do not appear uniformly across variables or time horizons. GDP forecast accuracy improves relatively early, within approximately 2 years of IFI

establishment, while gains in fiscal forecast accuracy emerge more gradually, typically requiring 4 to 5 years. This differential timing is consistent with reputational and accountability mechanisms operating through distinct channels. Macroeconomic forecasts, being more standardised and less directly tied to distributive choices, may respond more quickly to independent scrutiny, while fiscal forecasts, involving politically sensitive assumptions about revenues and expenditures, may require more sustained oversight before governments adjust their behaviour.

Cohort-based analysis indicates that improvements following IFI activation emerge across diverse adoption contexts, lending support to the external validity of the baseline estimates. However, early-adopting cohorts display somewhat larger effects, which may reflect either genuine institutional maturation, as IFIs accumulate credibility, expertise, and public visibility over calendar time, or compositional differences between countries that adopted early versus late. The available evidence cannot cleanly distinguish between these interpretations, cautioning against precise quantitative conclusions about effect magnitudes.

Several limitations warrant acknowledgement. First, while the staggered difference-in-differences design addresses many identification concerns, the possibility remains that IFI adoption coincides with unobserved reforms or shifts in political commitment to fiscal discipline. The combination of EU and LAC samples partially addresses this concern, since LAC countries adopted IFIs without any supranational mandate, but selection on unobservables cannot be fully ruled out. Second, the analysis focuses on official forecast accuracy. An extension would be to study the effect of IFIs on private forecast expectations. A further extension would be to link independent fiscal oversight with sovereign spread dynamics by analysing the market impact of IFI opinions on daily sovereign yields.

From a policy perspective, the findings highlight the central role of reputation and accountability mechanisms in fiscal governance. IFIs appear to improve forecast credibility not through immediate enforcement or technical correction, but through the gradual construction of institutional credibility and the progressive increase in reputational costs associated with overly optimistic projections. This suggests that premature evaluations of IFI effectiveness may underestimate their impact if they fail to account for institutional maturation. The results also carry implications for the design of fiscal frameworks. Optimistic forecasts represent a key channel through which governments can comply with fiscal rules *ex ante* while deviating *ex post*. By constraining forecast bias at the preparation stage, IFIs may reduce the likelihood of fiscal slippages driven by unrealistic assumptions, thereby strengthening the operational credibility of rules-based frameworks. In this sense, IFIs complement fiscal rules not by enforcing them directly, but by improving the informational foundations on which fiscal commitments rest.

Finally, the gradual emergence of IFI effects suggests that institutional design matters beyond formal mandate. Effective oversight requires not only legal authority but also technical capacity, public visibility, and sustained engagement with the budget process. Countries establishing IFIs should anticipate that meaningful effects on forecasting behaviour may require several years to materialise as institutions build recognition and credibility.

The forecast credibility examined in this paper represents one dimension of the broader informational environment that shapes fiscal outcomes. More credible forecasts improve the foundations on which budgets are prepared, but they constitute only part of the information

that investors, voters, and other stakeholders use to assess fiscal sustainability.

References

Ardanaz, M., Ulloa-Suárez, C., and Valencia, O. (2024). Why don't we follow the rules? drivers of compliance with fiscal policy rules in emerging markets. *Journal of International Money and Finance*, 142:103046.

Armstrong, J. S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80.

Barro, R. J. and Gordon, D. B. (1983). A positive theory of monetary policy in a natural rate model. *Journal of Political Economy*, 91(4):589–610.

Beetsma, M. R. M. and Debrun, M. X. (2016). *Fiscal councils: rationale and effectiveness*. International Monetary Fund.

Beetsma, R., Debrun, X., Fang, X., Kim, Y., Lledó, V., Mbaye, S., and Zhang, X. (2019). Independent fiscal councils: Recent trends and performance. *European Journal of Political Economy*, 57:53–69.

Beetsma, R., Debrun, X., and Sloof, R. (2022). The political economy of fiscal transparency and independent fiscal councils. *European Economic Review*, 145:104118.

Beetsma, R., Giuliodori, M., and Wierts, P. (2009). Planning to cheat: Eu fiscal policy in real time. *Economic policy*, 24(60):753–804.

Boylan, R. T. (2008). Political distortions in state forecasts. *Public Choice*, 136(3):411–427.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230.

Calmfors, L. and Wren-Lewis, S. (2011). What should fiscal councils do? *Economic Policy*, 26(68):649–695.

Cukierman, A. and Meltzer, A. H. (1986). A positive theory of discretionary policy, the cost of democratic government and the benefits of a constitution. *Economic Inquiry*, 24(3):367–388.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, 110(9):2964–2996.

Debrun, X. (2011). Democratic accountability, deficit bias, and independent fiscal agencies.

Debrun, X., Hauner, D., and Kumar, M. S. (2009). Independent fiscal agencies. IMF Working Paper 09/135, International Monetary Fund.

Debrun, X. and Kinda, T. (2017). Strengthening post-crisis fiscal credibility: fiscal councils on the rise–a new dataset. *Fiscal Studies*, 38(4):667–700.

Debrun, X., Kinda, T., Curristine, T., Eyraud, L., Harris, J., and Seiwald, J. (2013). The functions and impact of fiscal councils. *IMF Policy Paper*, 16.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Dube, A., Girardi, D., Jorda, O., and Taylor, A. M. (2025). A local projections approach to difference-in-differences. *Journal of Applied Econometrics*.

Frankel, J. (2011). Over-optimism in forecasts by official budget agencies and its implications. *Oxford Review of Economic Policy*, 27(4):536–562.

Frankel, J. and Schreger, J. (2013). Over-optimistic official forecasts and fiscal rules in the eurozone. *Review of World Economics*, 149:247–272.

Frankel, J. A. and Schreger, J. (2016). Bias in official fiscal forecasts: can private forecasts help? Technical report, National Bureau of Economic Research.

Gilbert, N. D. and de Jong, J. F. (2017). Do european fiscal rules induce a bias in fiscal forecasts? evidence from the stability and growth pact. *Public Choice*, 170:1–32.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277.

Hadzi-Vaskov, M., Werner, A. M., and Zamarripa, R. (2021). *Authorities' fiscal forecasts in Latin America: are they optimistic?* International Monetary Fund.

Holden, K. and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *The Manchester School of Economic & Social Studies*, 58(2):120–127.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Imai, K., Kim, I. S., and Wang, E. H. (2023). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605.

Jonung, L. and Larch, M. (2006). Improving fiscal policy in the eu: the case for independent forecasts. *Economic Policy*, 21(47):492–534.

Larch, M. and Santacroce, S. (2025). Trailing the fiscal frontier: the track record of independent fiscal institutions. Efb working paper, European Fiscal Board.

Merola, R. and Pérez, J. J. (2013). Fiscal forecast errors: governments versus independent agencies? *European Journal of Political Economy*, 32:285–299.

Mooney, H., Wright, A., and Grenade, K. (2018). Fiscal councils: Evidence, common features and lessons for the caribbean.

Nguyen, T. C., Castro, V., and Wood, J. (2022). A new comprehensive database of financial crises: Identification, frequency, and duration. *Economic Modelling*, 108:105770.

Pina, Á. M. and Venes, N. M. (2011). The political economy of edp fiscal forecasts: an empirical assessment. *European Journal of Political Economy*, 27(3):534–546.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Scartascini, C., Cruz, C., and Keefer, P. (2021). The database of political institutions 2020 (dpi2020).

Shi, M. and Svensson, J. (2006). Political budget cycles: Do they differ across countries and why? *Journal of public economics*, 90(8-9):1367–1389.

Sloof, R., Beetsma, R., and Steinweg, A. (2025). Debt ceilings with fiscal intransparency and imperfect electoral accountability. *International Economic Review*.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, 225(2):175–199.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362.

Wyplosz, C. (2005). Fiscal policy: institutions versus rules. *National Institute Economic Review*, 191:64–78.

Appendix

Appendix A Model Derivations

A.1 Baseline model: derivations

This appendix provides the derivations underlying the baseline model presented in Section 3.

Let actual GDP growth be given by:

$$g^a = \mu + \epsilon, \tag{A.1}$$

where ϵ is a random shock with mean zero and variance σ^2. The government chooses a forecast $g^f = \mu + \beta$, where β denotes the optimism bias.

Tax revenues are assumed to be proportional to GDP growth, with a sensitivity parameter T. When realised growth falls short of the forecast, the resulting unanticipated fiscal deviation is given by:

$$\tilde{D} = -T(g^a - g^f) = T(\beta - \epsilon). \tag{A.2}$$

The government incurs a penalty for fiscal deviations, which may reflect the enforcement of fiscal rules, market discipline, or political accountability. The model assumes this penalty is quadratic in the deviation. The expected cost is therefore:

$$\mathbb{E}[\tilde{D}^2] = T^2(\beta^2 + \sigma^2), \tag{A.3}$$

implying an expected cost function:

$$C(\beta) = \frac{1}{2}\theta\mathbb{E}[\tilde{D}^2] = \frac{1}{2}\theta T^2(\beta^2 + \sigma^2). \tag{A.4}$$

Note that $E[\bar{D}] = T\beta$ and $E[\bar{D}^2] = T^2(\beta^2 + \sigma^2)$ since $E[\epsilon] = 0$.
The government's expected utility[10] is given by

$$U(\beta) = \alpha\beta - \frac{1}{2}\theta T^2(\beta^2 + \sigma^2), \tag{A.5}$$

where $\alpha > 0$ captures the political benefit of forecast optimism. Since the constant term $\frac{1}{2}\theta T^2\sigma^2$ does not affect the choice of β, the optimization problem reduces to

$$\max_{\beta \geq 0} \ \alpha\beta - \frac{1}{2}\theta T^2\beta^2. \tag{A.6}$$

The first-order condition yields

$$\alpha - \theta T^2\beta = 0, \tag{A.7}$$

[10]The linear benefit function captures the direct mechanical relationship between forecast bias and projected fiscal space through the revenue channel. The quadratic cost function reflects the canonical specification in time-inconsistency models (Barro and Gordon, 1983), representing policymakers' aversion to large deviations and the accelerating credibility costs of systematic forecast biases. This specification provides analytical tractability while preserving the key trade-off between short-term political incentives and longer-term reputational concerns

implying the optimal forecast bias

$$\beta^* = \frac{\alpha}{\theta T^2}. \tag{A.8}$$

This result shows that, in the absence of institutional constraints, governments optimally choose an optimistic forecast whenever political benefits are positive and the expected cost of deviations is finite.

A.2 IFI extension: formal derivation

This appendix presents the formal derivation of the IFI extension discussed in Section 3.

Building on the baseline model, the IFI introduces an additional reputational penalty on the *realised* forecast error, as suggested by Beetsma et al. (2022). This cost captures the political penalties arising from public scrutiny of forecast bias once the outcome is observed and takes the quadratic form

$$R(\beta, \epsilon) = \delta\,(\beta - \epsilon)^2, \qquad \mathbb{E}[R(\beta, \epsilon)] = \delta(\beta^2 + \sigma^2), \tag{A.9}$$

where $\delta > 0$ measures the strength and credibility of the IFI.

The government's expected utility becomes

$$\mathbb{E}[U(\beta)] = \alpha\beta - \tfrac{1}{2}\theta T^2(\beta^2 + \sigma^2) - \delta(\beta^2 + \sigma^2). \tag{A.10}$$

Ignoring constant terms that do not affect the choice of β, the optimisation problem reduces to

$$\max_{\beta \geq 0}\ \alpha\beta - \left(\frac{1}{2}\theta T^2 + \delta\right)\beta^2. \tag{A.11}$$

The first-order condition yields

$$\alpha - (\theta T^2 + 2\delta)\beta = 0, \tag{A.12}$$

implying the optimal forecast bias under IFI oversight:

$$\beta^{**} = \frac{\alpha}{\theta T^2 + 2\delta}. \tag{A.13}$$

Since $\delta > 0$, it follows that $\beta^{**} < \beta^*$, where β^* denotes the equilibrium bias in the absence of an IFI. In the limiting case where the IFI fully constrains the forecast used for budgeting, $\delta \to \infty$ and $\beta^{**} \to 0$.

Appendix B Additional descriptive statistics

B.1 Descriptive statistics of control variables.

Table B.1: Descriptive statistics of control variables by region

	All					EU					LAC				
Variable	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Crises	0.18	0.39	0.00	1.00	948	0.16	0.37	0.00	1.00	633	0.23	0.42	0.00	1.00	315
GDP growth rate (in %)	2.49	5.04	-30.00	62.29	972	2.32	3.74	-14.69	24.62	631	2.79	6.82	-30.00	62.29	341
Inflation rate (in %)	96.61	2195.17	-1.68	65374.08	970	2.65	2.85	-1.68	19.45	632	272.30	3715.94	-1.55	65374.08	338
CAB (in % of GDP)	-1.25	5.98	-68.78	26.19	971	-0.03	5.43	-23.89	12.65	631	-3.52	6.28	-68.78	26.19	340
Election	0.30	0.46	0.00	1.00	974	0.31	0.46	0.00	1.00	633	0.28	0.45	0.00	1.00	341
Majority of government	0.54	0.12	0.03	0.97	843	0.55	0.08	0.25	0.80	549	0.52	0.17	0.03	0.97	294
Public debt (in % of GDP)	59.61	35.48	3.77	329.10	973	63.13	35.44	3.77	213.15	632	53.08	34.66	3.90	329.10	341
Fiscal balance (in % of GDP)	-2.60	3.40	-32.11	7.96	972	-2.44	3.46	-32.11	6.73	631	-2.89	3.27	-30.98	7.96	341
FRI	2.15	1.13	0.00	4.00	974	2.77	0.61	0.00	4.00	633	1.01	0.98	0.00	3.00	341

Note: Summary statistics are reported by region and for the pooled sample. Differences in N across variables reflect missing values.

B.2 Forecast deviations by horizon

Table B.2: Short- and long-term macro-fiscal forecast deviations descriptive statistics in EU and LAC countries from 1998 to 2019.

	Total Sample					LAC					EU				
Variable	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.
GDP	810	-0.36	2.92	7.95	158.20	214	-0.65	5.02	6.15	70.67	596	-0.25	1.60	-0.03	8.37
GDP (ST)	810	-0.36	3.15	6.36	118.69	214	-0.58	5.04	6.03	69.23	596	-0.28	2.09	-0.07	9.27
GDP (LT)	468	-1.03	3.77	-0.50	10.31	67	-1.36	4.19	-0.34	5.10	401	-0.98	3.70	-0.53	11.62
Revenue	723	-0.30	2.85	2.53	51.37	136	-0.31	3.10	-3.21	20.28	587	-0.30	2.79	4.34	61.70
Revenue (ST)	726	-0.31	2.87	2.25	49.09	139	-0.10	3.20	-2.90	18.41	587	-0.36	2.78	4.06	61.54
Revenue (LT)	426	0.11	3.64	2.16	27.67	37	0.03	2.71	-0.33	3.54	389	0.12	3.71	2.22	27.72
Expenditure	731	0.45	2.68	0.05	5.55	142	1.31	1.95	-0.33	4.99	589	0.25	2.79	0.19	5.57
Expenditure (ST)	730	0.25	2.82	0.13	5.31	142	1.24	2.09	0.80	9.87	588	0.01	2.92	0.18	4.94
Expenditure (LT)	426	1.96	4.06	0.44	7.61	37	2.23	2.23	-0.21	2.83	389	1.93	4.20	0.45	7.32

Notes: This table compares forecast error distributions across time horizons. For each variable, the first row (without ST/LT designation) reproduces the baseline descriptive statistics from Table 3, pooling all forecast horizons. The ST (short-term) and LT (long-term) rows decompose the sample by forecast horizon: ST includes forecasts 0-2 years ahead, and LT includes forecasts 3-5 years ahead. The baseline row and horizon-specific rows report the same total sample size in the column headers, but the ST and LT statistics are computed using different subsets of observations corresponding to their respective horizon definitions. All forecast errors are computed as e = y - f (realisation minus forecast). Negative values indicate optimistic forecasts for GDP and revenue; positive values indicate an underestimation of expenditure.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

B.3 Absolute forecast deviations by horizon

Table B.3: Short- and long-term macro-fiscal absolute forecast deviations descriptive statistics in EU and LAC countries from 1998 to 2019.

	Total Sample					LAC					EU				
Variable	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.	N	Mean	SD	Skew.	Kurt.
Abs. GDP Error	809	2.11	2.66	10.61	193.38	213	3.24	4.49	7.52	82.24	596	1.70	1.33	1.97	8.21
Abs. GDP (ST)	810	1.69	2.68	10.51	191.14	214	2.52	4.40	8.21	93.75	596	1.39	1.59	2.75	13.53
Abs. GDP (LT)	468	2.46	3.03	2.81	13.51	67	3.03	3.18	1.73	5.63	401	2.37	3.00	3.03	15.37
Abs. Revenue	730	1.95	2.21	7.73	105.95	142	2.07	2.43	4.72	34.09	588	1.91	2.15	8.73	132.85
Abs. Revenue (ST)	726	1.83	2.23	7.39	103.15	139	2.02	2.48	4.59	32.56	587	1.78	2.17	8.35	130.47
Abs. Revenue (LT)	426	2.47	2.67	5.82	68.95	37	1.90	1.91	1.05	3.01	389	2.52	2.72	5.92	68.88
Abs. Expenditure	730	2.30	1.75	1.96	9.66	142	1.91	1.48	1.11	4.07	588	2.39	1.80	2.05	9.96
Abs. Expenditure (ST)	730	2.07	1.93	1.90	8.20	142	1.75	1.68	2.55	15.05	588	2.15	1.98	1.79	7.27
Abs. Expenditure (LT)	426	3.42	2.93	2.33	15.59	37	2.63	1.72	0.54	2.42	389	3.50	3.01	2.30	15.06

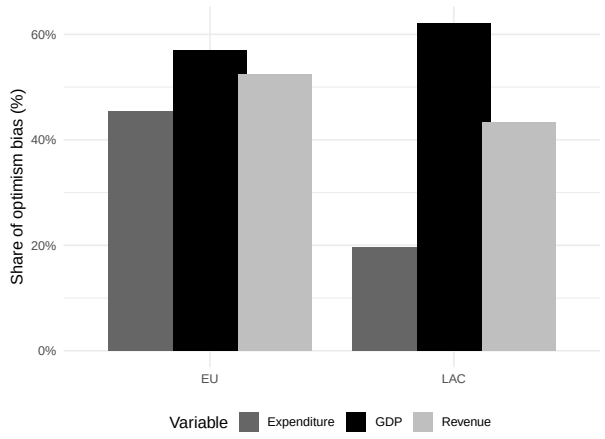
Notes: Absolute deviations measure forecast accuracy independently of the direction of errors. Larger values indicate lower forecast precision. ST and LT denote short- and long-term forecast horizons. Short horizon is up to two periods (t=0 to t+2), while long horizon is from t=3 to t=5.
Sources: SGP and DPB, and FISLAC for EU and LAC countries, respectively.

B.4 Forecast optimism

This appendix presents complementary descriptive evidence on forecast optimism. The aim is to document the prevalence and direction of optimism bias using alternative summary indicators and external benchmarks. These results complement the stylised facts discussed in the main text and provide additional context on the nature of forecast errors, without implying causal interpretation.

Figure B.1 reports the share of optimistic forecasts by region and macro-fiscal variable. A forecast is classified as optimistic when projected values exceed realised values for GDP growth and revenues, or when projected values underestimate realised values for expenditures. This indicator captures the frequency of optimism bias, abstracting from its magnitude.

Figure B.1: Share of optimism in macro-fiscal forecasts by regions from 1998 to 2019



Source: Authors' computation with SGP and DPB, and FISLAC for EU and LAC countries, respectively.

Two regularities emerge. First, optimistic forecasts are widespread across regions and variables, particularly for GDP growth. Second, optimism appears more prevalent in LAC countries than in the EU, especially for GDP and revenues. While informative, this measure remains descriptive and does not account for country characteristics, forecast horizons, or institutional features.

To assess whether optimism bias is specific to government forecasts, I also compare official projections with forecasts produced by external institutions. Table B.4 contrasts government forecasts with IMF World Economic Outlook projections, while Table B.5 compares government forecasts with those produced by the European Commission.

Table B.4: Differences in forecast errors between IMF and governments

Variable	Mean Err. Gov.	Mean Err. IMF	Diff Bias	Obs	Comment
Expenditure	0.388	0.207	-0.181	1758	IMF less optimistic
GDP	-0.262	-0.202	0.060	1994	IMF less optimistic
Revenue	-0.246	-0.275	-0.029	1724	IMF more optimistic

Note: Data from 2012 to 2024. Average mean forecast error over horizons $t = 0$ to $t = 5$.
Source: IMF forecasts are taken from WEO Fall vintages. Authors' computation.

Table B.5: Comparison between European Commission (EC) and Government Forecasts

Statistic	European Commission	Government	Diff Bias	Comment
GDP Mean Error	2.16	-0.673	-2.83	Government more optimistic
GDP SD	1.93	3.60		
GDP Observations	678	760		

Notes: Data from 2000 to 2024. One-year-ahead forecast errors. Forecast errors are computed as the realisation minus the forecast. Negative mean errors indicate optimistic forecasts (forecasts exceed realisations). Positive mean errors indicate pessimistic forecasts (realisation exceeds forecast).
Sources: The sample includes EU countries with both the EC *Forecast Tracker* from Larch and Santacroce (2025) and the new government forecast database.

These comparisons show that optimism bias is not exclusive to government forecasts. External forecasters, including the IMF and the European Commission, also exhibit systematic forecast errors, sometimes in the same direction as official projections. This suggests that forecast optimism reflects broader informational constraints and institutional frictions, rather than purely strategic behaviour by governments.

Table B.6: Covariate balance diagnostics before and after matching in full sample nearest neighbour matching specification

Variable	Type	Diff. Unadj.	V.Ratio Unadj.	Diff. Adj.	V.Ratio Adj.	Threshold status
distance	Distance	1.4931	1.0145	1.0327	1.2387	
l.GDP	Contin.	-0.2537	0.5723	-0.2027	0.9129	Not balanced, > 0.1
lInflation	Contin.	-116.6363	0.0000	-0.6527	0.3926	Not balanced, > 0.1
l.CAB	Contin.	0.9299	0.4196	0.5765	0.6924	Not balanced, > 0.1
l.Public Debt	Contin.	0.4237	1.2311	0.3307	1.5575	Not balanced, > 0.1
l.Fiscal Balance	Contin.	0.1424	0.6318	0.1342	0.6245	Not balanced, > 0.1
l.Crises	Binary	-0.0021	.	0.0091	.	Balanced, < 0.1
l.FRI	Contin.	1.5684	0.4463	0.7905	1.1144	Not balanced, > 0.1
l.Election	Binary	0.0153	.	0.0122	.	Balanced, < 0.1
l.Majority	Contin.	0.0657	0.7640	0.0074	1.5902	Balanced, < 0.1
Balance tally (mean differences)						
Balanced (< 0.1): 3		Not balanced (> 0.1): 6				
Largest adjusted mean difference: lag(SUMFR), Diff. Adj. = 0.7905						

Notes: Diff. Unadj. and Diff. Adj. are standardized mean differences before and after matching. V.Ratio denotes the treated-to-control variance ratio. For binary covariates, variance ratios are not reported (“.”). A common benchmark is |Diff. Adj.| < 0.1 for adequate balance.
N treated = 80, N controlled = 279

Table B.7: Covariate balance diagnostics before and after matching in within-year matching specification

Variable	Type	Diff. Unadj.	Diff. Adj.	Threshold status
distance	Distance	1.4931	0.0705	Balanced, < 0.1
l.GDP	Contin.	-0.2537	-0.0320	Balanced, < 0.1
l.Inflation	Contin.	-116.6363	0.0266	Balanced, < 0.1
l.CAB	Contin.	0.9299	0.0364	Balanced, < 0.1
l.Public Debt	Contin.	0.4237	0.1336	Not balanced, > 0.1
l.Fiscal Balance	Contin.	0.1424	0.0713	Balanced, < 0.1
l.Crises	Binary	-0.0021	-0.0085	Balanced, < 0.1
l.FRI	Contin.	1.5684	0.0116	Balanced, < 0.1
l.Election	Binary	0.0153	-0.0508	Balanced, < 0.1
l.Majority	Contin.	0.0657	0.0708	Balanced, < 0.1
year	Contin.	0.9077	0.0000	Balanced, < 0.1
Balance tally (mean differences)				
Balanced (< 0.1): 10		Not balanced (> 0.1): 1		
Largest adjusted mean difference: lag(Actual_Debt), Diff. Adj. = 0.1336				

Notes: Diff. Unadj. and Diff. Adj. denote standardized mean differences before and after matching, respectively. A common rule-of-thumb considers absolute standardized differences below 0.1 as acceptable balance.
N treated = 251, N controlled = 245.

Appendix C Implementation timing and mandates of independent fiscal institutions

Table C.1: List of IFI implementation in the EU and LAC countries, operational year

Country	Region	IFI	Produce/endorse Macro forecast	Produce/endorse Fiscal forecast	Country	Region	IFI	Produce/endorse Macro forecast	Produce/endorse Fiscal forecast
Austria	EU	1970	✓		Argentina	LAC			
Belgium	EU	1959	✓		Bahamas	LAC	2019	✓	
Bulgaria	EU	2015	✓		Barbados	LAC			
Croatia	EU	2013	✓		Belize	LAC			
Cyprus	EU	2014	✓		Bolivia	LAC			
Czech Republic	EU	2017			Brazil	LAC	2016	✓	
Denmark	EU	1962	✓		Chile	LAC	2014	✓	
Estonia	EU	2014	✓		Colombia	LAC	2011	✓	
Finland	EU	2013	✓		Costa Rica	LAC	2021	✓	
France	EU	2013	✓		Cuba	LAC			
Germany	EU	2013	✓	✓	Dominica	LAC			
Greece	EU	2010	✓	✓	Dominican Rep.	LAC			
Hungary	EU	2009	✓	✓	Ecuador	LAC			
Ireland	EU	2011	✓		El Salvador	LAC			
Italy	EU	2014	✓		Grenada	LAC	2017		
Latvia	EU	2014	✓	✓	Guatemala	LAC			
Lithuania	EU	2015	✓		Guyana	LAC			
Luxembourg	EU	2015			Haiti	LAC			
Malta	EU	2015	✓		Honduras	LAC			
Netherlands	EU	1945	✓		Jamaica*	LAC	2023		
Poland	EU				Mexico	LAC	1998	✓	
Portugal	EU	2012	✓	✓	Nicaragua	LAC			
Romania	EU	2010	✓		Panama*	LAC	2018	✓	
Slovakia	EU	2012	✓	✓	Paraguay*	LAC	2016		
Slovenia	EU	2015	✓		Peru	LAC	2015	✓	
Spain	EU	2014	✓		Uruguay	LAC	2021	✓	
Sweden	EU	2007	✓		Venezuela	LAC			
United Kingdom		2010	✓	✓					

Source: IMF Fiscal Council Dataset, OECD dataset on PBOs and IFIs.
Note: Blank cells in the "IFI" column indicate countries that had not established an IFI by the end of the sample period (2019); blank cells in the "Produce/endorse" columns indicate that the IFI's mandate does not include the production or endorsement of macroeconomic or fiscal forecasts, respectively. Countries marked with an asterisk (*) had legally established IFIs that were not yet operational; these observations are coded as untreated in all empirical specifications.

Appendix D PanelMatch design for differentiating IFI mandate effects

This appendix describes the PanelMatch design used to assess whether the effects of IFIs on forecast accuracy depend on their formal forecasting mandate. The analysis is restricted to European Union countries, where institutional environments are more homogeneous and where detailed information on IFI mandates is consistently available. This restriction helps mitigate concerns related to selection on broader institutional quality, although mandate assignment remains non-random even within this subsample.

D.1 PanelMatch design

Let i index countries and t years. The outcome variable Y_{it} is the mean absolute forecast error, with expenditure forecasts as the baseline specification. Results for GDP growth and revenue forecasts are analogous. The treatment variable $X_{it} \in \{0, 1\}$ indicates whether an IFI with a given mandate is operational in country i at time t. The analysis distinguishes between IFIs that produce or endorse official macroeconomic forecasts and those that produce or endorse fiscal forecasts.

Following Imai et al. (2023), The analysis defines dynamic average treatment effects for event time $F \in \{0, \ldots, 5\}$ using $L = 2$ pre-treatment lags. The estimand is

$$\delta(F, L) = \mathbb{E}\Big[Y_{i,t+F}\Big(1, 0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\Big) - Y_{i,t+F}\Big(0, 0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\Big) \;\Big|\; X_{it} = 1,\ X_{i,t-1} = 0\Big], \quad \text{(D.1)}$$

which allows for dynamic effects following mandate activation.

Identification relies on three assumptions. First, there is no interference across units and no carryover effects beyond the specified lag length. Second, untreated potential outcomes satisfy a conditional parallel trends assumption given past treatment, outcomes, and covariates. Third, sufficient overlap exists so that each treated onset can be matched to comparable control units.

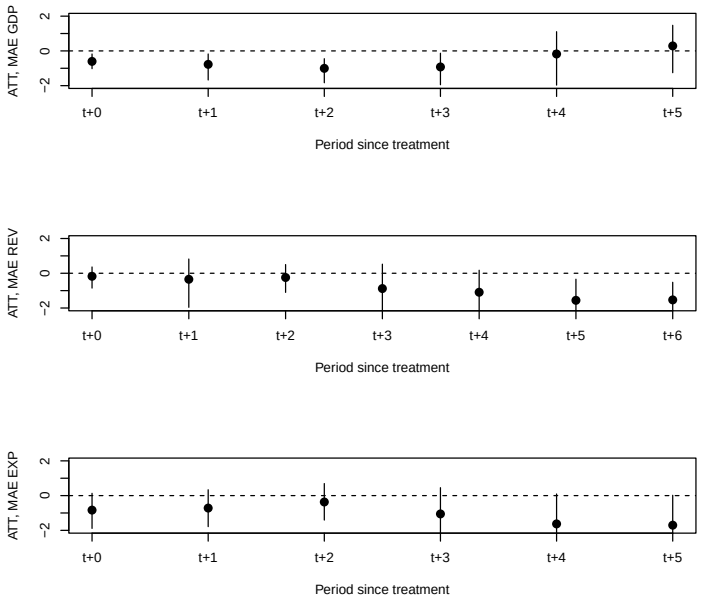
For each treatment onset (i, t), the analysis constructs a time-exact risk set of control units with identical treatment histories over the previous two periods. These candidate controls are refined using Mahalanobis distance matching based on pre-treatment covariates, including realized GDP, public debt, fiscal balance, inflation, crisis indicators, electoral timing, government majority, the current account balance, and the presence of fiscal rules. Two nearest neighbors are selected with replacement, and treatment onsets without valid matches are discarded to preserve internal validity.

For each event time F, treatment effects are computed as matched difference-in-differences estimators relative to the pre-treatment period. Inference is based on design-based uncertainty, using both analytic variance formulas and unit-level block bootstrap procedures that account for serial correlation within countries. This design yields dynamically interpretable treatment effects under a restrictive but transparent set of identifying assumptions.

D.2 PanelMatch results by IFI remit

Beyond serving as a robustness check, this specification also allows us to exploit variation in a key institutional characteristic of IFIs, namely their formal forecasting remit. In particular, the analysis examines whether the effects of IFI adoption differ depending on whether the institution produces, formally endorses, or both official macroeconomic and fiscal forecasts. This distinction is central to the transmission mechanisms emphasised in the theoretical discussion, as mandates over forecasts directly affect the stage at which optimism bias may be constrained. The heterogeneity analysis follows a comparative logic. Figure D.1 establishes baseline effects for IFIs with macroeconomic remits, showing how these institutions affect GDP, revenue, and expenditure forecast accuracy. Figure D.2 then tests whether IFIs that also have fiscal remits achieve differential improvements in fiscal forecast accuracy beyond what macroeconomic oversight alone achieves. The analysis focuses this comparison on revenue and expenditure outcomes, as these are the forecast domains where institutional differences in fiscal mandate should matter most. GDP forecast accuracy is excluded from Figure D.2 because both macro-remit and fiscal-remit IFIs should affect GDP similarly (many fiscal-remit IFIs also hold macroeconomic mandates, as shown in Table C.1), making GDP uninformative for distinguishing between mandate types. The central question is whether direct institutional authority over fiscal projections disciplines fiscal forecasting behaviour more effectively than indirect oversight operating through macroeconomic scrutiny alone.

Figure D.1: Estimated average effects of IFIs' **macroeconomic** remit on GDP, Revenue, and Expenditure mean absolute forecast error

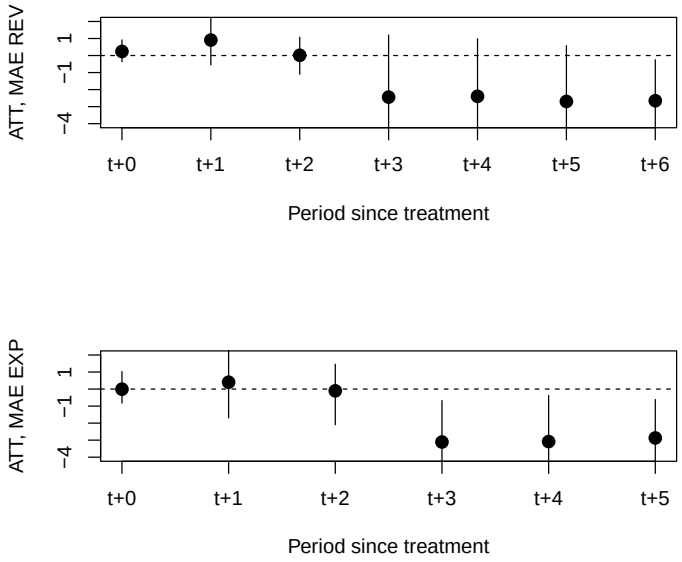


Notes: Estimates are based on the Mahalanobis matching method that adjusts for the treatment and covariates' histories during the 2-year period prior to the treatment. The estimates are the average effects of IFI are shown for the period of 5 years after the implementation, with 95% asymptotic confidence intervals as vertical bars. Standard errors are bootstrapped using 1000 replications.

IFIs with mandates over macroeconomic forecasts exhibit relatively rapid improvements in GDP growth forecast accuracy, with statistically significant effects emerging within two years of implementation. Improvements in fiscal forecast accuracy materialise more gradually, consistent with an indirect transmission mechanism whereby more realistic macroeconomic assumptions constrain subsequent revenue and expenditure projections.

Figure D.2 examines whether IFIs with explicit mandates over fiscal forecasts generate differential improvements in fiscal forecast accuracy relative to the macro-remit baseline shown in Figure D.1. The comparison focuses exclusively on revenue and expenditure outcomes, as these are the forecast domains where fiscal mandate should create institutional differentiation. The results reveal three patterns. First, for revenue forecasts (Panel a), fiscal-remit IFIs produce improvements of larger magnitude after 6 periods than those observed under macro-remit institutions. This differential suggests that direct authority to scrutinise or endorse revenue projections disciplines forecasting behaviour more effectively than macroeconomic oversight alone. Second, for expenditure forecasts (Panel b), significant improvements emerge approximately three years after adoption, with effect sizes of around 3pp. Comparing these dynamics to Figure D.1, expenditure accuracy improves faster under fiscal remits. Third, the differential timing between revenue and expenditure improvements suggests that fiscal-remit IFIs may prioritise revenue scrutiny (which is directly linked to macroeconomic assumptions and thus easier to challenge) over expenditure scrutiny (which involves discretionary policy choices and is more politically sensitive). These patterns indicate that mandate specificity matters for forecast improvements. IFIs with direct institutional authority over fiscal projections generate larger and more immediate improvements in fiscal forecast accuracy than institutions that influence fiscal variables only indirectly through macroeconomic oversight. This finding supports the theoretical mechanism in Section 3, where IFIs operate through reputational costs that are strongest in domains where they possess explicit institutional authority.

Figure D.2: Estimated average effects of IFIs' **fiscal** remit on Revenue, and Expenditure mean absolute forecast error

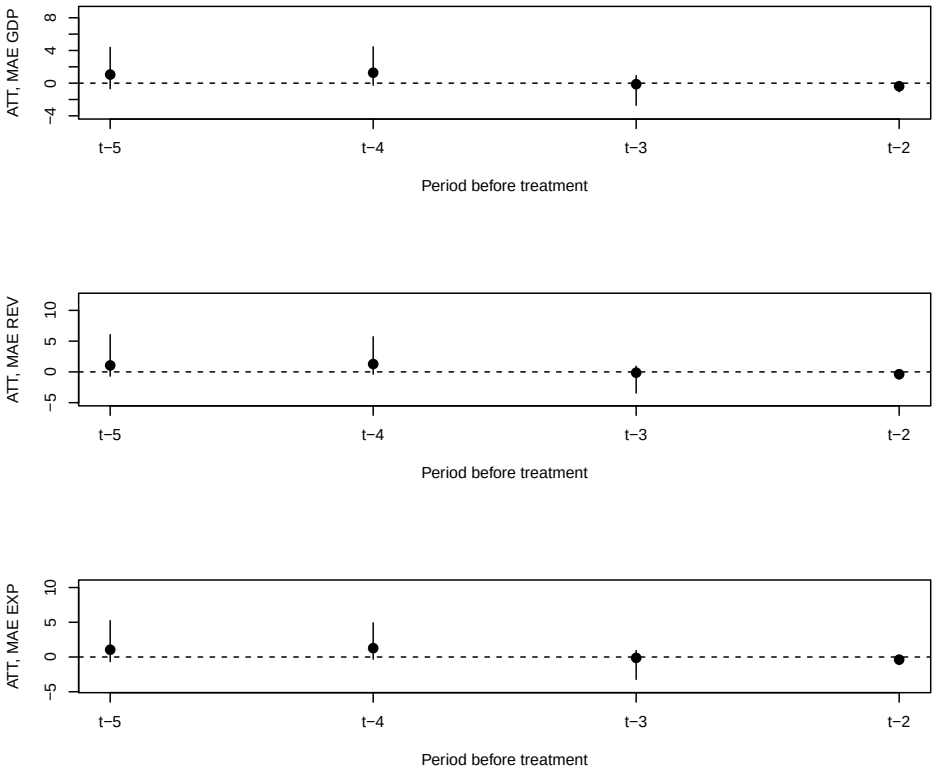


Notes: Estimates are based on the Mahalanobis matching method that adjusts for the treatment and covariates' histories during the 2-year period prior to the treatment. The estimates are the average effects of IFI are shown for the period of 5 years after the implementation, with 95% asymptotic confidence intervals as vertical bars. Standard errors are bootstrapped using 1000 replications.

The results confirm the importance of mandate specificity. While both types of IFIs contribute to improved forecast accuracy, institutions with direct fiscal forecasting responsibilities exert stronger and faster effects on fiscal variables, where political incentives for optimism bias are typically most pronounced.

Figure D.3 shows that placebo ATTs in the pre-treatment window are centered around zero and do not display a systematic trend across leads. This indicates no evidence of differential pre-treatment dynamics between treated and matched control units, supporting the validity of the identification strategy. Overall, the placebo diagnostics are satisfactory, suggesting that the research design does not require further refinement (Imai et al., 2023).

Figure D.3: Placebo (pre-treatment) test for PanelMatch estimates



Notes: The figure reports placebo average treatment effects on the treated (ATT) in pre-treatment periods ($t-5$ to $t-2$) for GDP, revenue, and expenditure forecast-error outcomes. The absence of systematic pre-trends supports the identifying assumptions of the PanelMatch design.