DEFINE-UK: running and gating an unlicensed ecological stock-flow consistent model*

Vahid Ahmadi†

August 2026

Abstract

DEFINE-UK is the ecological stock-flow consistent (E-SFC) model of the United Kingdom built at SOAS by George and Dafermos (2026a): a closed monetary accounting system for seven institutional sectors, coupled to an energy and emissions block, with UK-specific housing and power-generation structures through which climate policy reaches the financial system. It is the only member of the PolicyEngine Macro suite that answers the question *who bears a climate-policy scenario*. This paper is its replication record, and it is deliberately a record of a ceiling as much as of a result. The upstream repository is public but carries **no licence**, so nothing is vendored or hosted: the adapter fetches it at pinned commit `846081a` at runtime and runs the authors' unmodified R. Four published-target gates were attempted. The manual's Table 4 baseline macro block **passes** — population and labour force exact to 0.01 m, growth matching to 0.06 pp at 2030 and 2040 — but the 2025 growth miss of −0.30 pp clears a ±0.31 pp gate *that was set after the run*, so that year's pass is a statement about the tolerance, not about agreement, and it is labelled as such at every point of claim. The baseline emissions path **diverges** and is pinned rather than tolerated: 393 vs 407 $MtCO_2e/yr$ in 2025, 343 vs 382 in 2030, 249 vs 324 by 2040 (−3.5, −10.2, −23.3 per cent, computed from the pinned ratios rather than the rounded levels). The two scenario targets are **closed at the achievable ceiling, not passed**: no machine-readable numeric scenario results are published for version 1.1, so what is gated is scenario *design* — the exact policy-switch set each cached scenario toggles — plus two coarse anchors from a pre-1.0 conference version, which are regression pins and not agreement tests; the emissions anchor admits 330–352 for a published claim of "just under 350", so a run at 351 would pass a gate named for a statement it violates. Because baseline levels sit far from outturns (2025 growth of 4.66 per cent against an ONS 1.31) and the manual itself says the baseline "should also not be seen as a prediction or forecast", the adapter serves annualised scenario-minus-baseline *deltas, never levels*. A cumulative green-public-investment multiplier of **1.78** is reported as ours and as *not comparable* to any published multiplier, for three reasons that all push the same way. In parallel, a clean-room Python reimplementation of the published equations — the only route to a hostable model given the licence — has passed milestone 1 (62 collected accounting tests,

*Acknowledgements: I thank Max Ghenis for his great suggestions and helpful discussions. DEFINE-UK is the work of Adam George and Yannis Dafermos at SOAS University of London, building on the DEFINE framework of Yannis Dafermos, Maria Nikolaidi and Giorgos Galanis; this paper documents an independent replication and reimplementation of their published model, is neither produced by nor endorsed by them, and any errors in the description of their work are mine. This paper and the accompanying codebase were prepared with the assistance of Claude (Anthropic); all remaining errors are my own.

†Research Associate at PolicyEngine. Email: vahid@policyengine.org.

44 of them Table 1/2 row and column identities) and stands at 118 contiguous equations, Eqs. (21)–(138), on milestone 2, which is **in progress and not passed**. Thirteen defects in the manual are carried as machine-readable gap records rather than absorbed into tolerances; an August 2026 audit re-characterised four of them and withdrew one claim outright.

Keywords: ecological macroeconomics, stock-flow consistent modelling, climate policy, replication, clean-room reimplementation, open source

1 Introduction

Climate policy is a macrofinancial event. A fossil-fuel ban writes down capital that somebody owns and somebody else has lent against; a retrofit subsidy is a fiscal flow that lands on households with a particular distribution of housing stock; a decarbonisation path that misses its target does so through investment decisions taken under credit rationing. Most macroeconomic models cannot say any of this, because their accounting stops short of balance sheets or their environmental block is a reduced-form elasticity bolted onto a supply-side core. Ecological stock-flow consistent (E-SFC) models are built for exactly this joint question, and DEFINE-UK is the United Kingdom's.

DEFINE-UK is documented in the *DEFINE-UK Model Manual, Version 1.1* (George and Dafermos, 2026a), released April 2026 alongside version 1.0. It is a closed monetary accounting system covering households, non-financial corporations, the power-generation sector, monetary and non-monetary financial institutions, government and the rest of the world, with a non-sectoral production module, coupled to an energy and emissions block, and it carries two UK-specific structures — a housing stock tracked by energy efficiency and a power sector with separate fossil and non-fossil capital — through which climate policy reaches the financial system rather than merely the output gap. Version 1.1 added regulatory policies that directly depreciate affected capital (stranded assets as an explicit balance-sheet event), limited forward-looking behaviour towards policy announcements, direct government ownership of non-fossil power capital, and a green sovereign bond.

Within PolicyEngine Macro (Ghenis and PolicyEngine contributors, 2025), an open suite of macroeconomic models for the UK and US, DEFINE-UK occupies a role no other member can fill: it is the only model in the suite that answers *who bears a climate-policy scenario*. It is also the suite's most constrained member, and this paper is as much a record of those constraints as of what was reproduced.

What this paper is. The suite's standard is that a model earns its place by replication: run the published code or reimplement the published equations, then compare against the published results at recorded tolerances, and report the comparison with numbers rather than adjectives — the same bar applied to the Bank of England SVAR (Ahmadi, 2026b) and the OBR macroeconometric emulator (Ahmadi, 2026a). This paper applies that standard to DEFINE-UK and records, target by target, where it passes, where it diverges, and where the publications themselves impose a ceiling on what can be checked at all.

Three constraints shape everything that follows. The first is legal. The upstream repository is public but carries **no licence**, which means its code cannot be vendored, redistributed or hosted. The adapter therefore fetches it at pinned commit `846081a` at runtime and executes it as published, in R, unmodified — and DEFINE-UK is the only member of the suite that is local-only, never hosted. The second is evidential. No machine-readable numeric scenario results are published for version 1.1: the manual's own results stop at the baseline table plus scenario *design* parameters, and both companion papers are access-restricted. There is, in short, nothing to replicate on the scenario side, and Section 6 says so rather than manufacturing a comparison. The third is calibrational. The published baseline sits far from UK outturns — 2025 real GDP growth of 4.66 per cent against an ONS 1.31 — and the manual states plainly that its baseline "should also not be seen as a prediction or forecast". Taking that literally is what makes the adapter serve annualised scenario-minus-baseline *deltas, never levels*.

Contribution. The contribution is threefold, and none of it is a forecast.

First, *a replication with its ceiling stated*. Section 6 reports four published-target gates. The manual's Table 4 macro block passes. Its emissions path diverges, and is pinned at the observed ratios so that further drift fails loudly rather than being absorbed. The two scenario targets are closed at the achievable ceiling and are explicitly not passes. Throughout, checks that cannot fail are separated from checks that can, at the point of claim: the 2025 growth pass clears a tolerance set after the run by 0.01 pp, and the two conference-vintage anchors are bands we chose to fit a verbal quantity, one of which admits values that contradict the sentence it is named for.

Second, *a calibration and a multiplier reported with their non-comparabilities attached*. Section 7 gives the baseline against ONS, DESNZ and OBR figures as a computed, regression-tested artifact rather than prose, and reports a cumulative green-public-investment multiplier of 1.78 that is ours, not the authors', and that is not like-for-like against any published multiplier on horizon, denominator or closure. A previously quoted IMF comparator is recorded as withdrawn.

Third, *a clean-room reimplementation in progress*. Because the upstream cannot be hosted, the route to a hostable DEFINE-UK is a from-scratch Python implementation of the published equations, with the upstream used strictly as a numerical output oracle and never read to write equations. Section 8 reports milestone 1 passed and milestone 2 in progress — 118 contiguous equations, Eqs. (21)–(138) — and treats the reimplementation as an instrument in its own right: reading the manual closely enough to run it has surfaced thirteen defects, which are carried as machine-readable gap records rather than papered over, and an audit of those records re-characterised four of them and withdrew one claim outright.

Road map. Section 2 places the model in the SFC, ecological-macroeconomics and climate-finance literatures. Section 3 sets out the model as the manual specifies it. Section 4 identifies the source publications, the licence position, and the replication scope. Section 5 describes the adapter and what it serves. Section 6 is the replication record. Section 7 covers external calibration and the multiplier. Section 8 covers the clean-room track. Section 9 concludes with limitations. The appendix reproduces the published baseline table, the full gap-record inventory, and the dated run record.

2 Related literature

DEFINE-UK sits at the junction of three literatures: stock-flow consistent macroeconomic modelling, ecological macroeconomics, and the climate-and-financial-stability agenda. A fourth — the practice of replication itself — is what this paper contributes to.

2.1 Stock-flow consistent modelling

The stock-flow consistent (SFC) tradition descends from Godley and Lavoie (2007), whose organising insight is that accounting comes before theory: every monetary flow leaves one sector's account and enters another's, and every flow accumulates into a stock, so that a transactions matrix sums to zero in every row and a balance-sheet matrix sums to zero in every instrument row. Nothing appears from nowhere and nothing leaks out. The discipline is unusually strong for a macroeconomic modelling framework, because it converts a large class of specification errors into arithmetic failures that a matrix will not close over.

Two surveys map the field the model belongs to. Caverzasi and Godin (2015) trace the post-Keynesian SFC programme from its Cambridge origins through the empirical and open-economy

extensions, and note the characteristic methodological trade: SFC models buy accounting completeness at the cost of large parameter sets calibrated rather than estimated, which makes their published tables — initial values, parameter provenance — an unusually large part of what a reader must be able to check. Nikiforos and Zezza (2017) survey the same literature with emphasis on how such models are closed, and on the recurring result that a demand-led closure delivers larger fiscal multipliers than a supply-constrained one, because output is not returned to a potential path by assumption. Section 7 shows that this is not an abstract point for DEFINE-UK: it is one of the three reasons its green-investment multiplier cannot be read against short-horizon published figures.

The empirical relevance of these surveys to the present paper is direct. In an SFC model calibrated rather than estimated, the published parameter and initial-value tables *are* the specification. Section 8 is in large part a report on what happens when one takes that seriously enough to transcribe those tables in full and demand that the printed equations reproduce the printed values.

2.2 Ecological macroeconomics and the DEFINE framework

The specific model replicated here belongs to the DEFINE family. Dafermos et al. (2017) introduced the stock-flow-fund framework that DEFINE implements: a synthesis of Godley–Lavoie monetary accounting with the material flow-fund accounting of ecological economics, so that the same model tracks money and matter under separate but coupled consistency conditions. Dafermos et al. (2018) extended it to the climate–finance channel, showing how physical damages and transition dynamics feed through firm leverage and bank balance sheets to credit availability, and how monetary policy interacts with that loop — the mechanism DEFINE-UK's credit-rationing and default blocks inherit.

DEFINE-UK (George and Dafermos, 2026a) is the United Kingdom calibration and extension of that framework. Its distinguishing content relative to the generic case is the pair of UK-specific structural blocks described in Section 3: a housing block that tracks the dwelling stock by energy efficiency, so that retrofit policy operates through the stock rather than through an assumed elasticity, and a power-generation sector with separate fossil and non-fossil capital, an investment trigger tied to capacity utilisation, and — new in version 1.1 — regulatory instruments that directly depreciate fossil capital.

2.3 Climate policy, stranded assets and financial stability

The reason to want balance sheets in a climate model is the transition-risk agenda that Carney (2015) named: the losses from a rapid, credible decarbonisation fall not on the emitting capital alone but on the claims written against it, held by institutions whose solvency and lending behaviour then feed back into the real economy. Battiston et al. (2017) made that concrete with a network stress test showing how climate-policy shocks propagate from exposed sectors through equity and credit holdings to the wider financial system. DEFINE-UK's version 1.1 regulation instruments are an in-model implementation of that mechanism: a regulatory scenario writes down power-sector capital, and the resulting change in leverage, illiquidity and credit rationing is a modelled consequence rather than an add-on. Whether it does so *correctly* is not something this paper can establish for the scenarios, for reasons Section 6 sets out.

2.4 Replication as a research output

Finally, this paper belongs to the replication literature, in the sense Dewald et al. (1986) established: that published results should be reproducible from published materials, and that the systematic attempt to do so is itself a finding. Their project's central lesson — that the binding constraint is usually the completeness of what was published, not the willingness of the replicator — describes this exercise closely. Here the constraint is threefold and unusually explicit: the code is unlicensed, the scenario results are unpublished, and the manual's parameter tables and equations do not everywhere agree with each other. Each is recorded as a ceiling with a stated route to lifting it, rather than routed around.

Within the PolicyEngine Macro suite the same standard has been applied to a central bank's structural VAR (Ahmadi, 2026b) and to the OBR's macroeconometric model (Ahmadi, 2026a). The methodological point those papers share with this one is the separation, at the point of claim, between checks that can fail and checks that cannot. A tolerance chosen after seeing the deviation it admits, an identity that holds by construction, a band fitted to a verbal quantity — these are worth reporting, but they are not evidence of agreement, and a replication that presents them as such is overstating its result. Section 6 labels each one where it appears.

3 The model

This section describes DEFINE-UK as the manual specifies it (George and Dafermos, 2026a). It is a description of the authors' model, not of an implementation choice of ours; where the manual's own text and tables disagree, Section 8 reports the disagreement rather than resolving it here.

3.1 Accounting before theory

DEFINE (Dynamic Ecosystem FINance-Economy) is the stock-flow-fund framework of Dafermos et al. (2017); DEFINE-UK is its United Kingdom application, with the accounting structure derived directly from UK national accounts (ONS Blue Book and UK Economic Accounts) rather than assumed. Two matrices carry the discipline.

The *transactions flow matrix* (manual §2.2, Table 1) records every monetary flow in the model with a sign convention in which + is an inflow and − an outflow to the sector in whose column it sits. Two families of identity follow:

$$\sum_{s} T_{f,s} = 0 \qquad \text{for every flow row } f, \tag{1}$$

$$\sum_{f} T_{f,s} = \mathrm{LEND}_{s} \qquad \text{for every sector column } s, \tag{2}$$

that is, each flow leaves one sector and enters another, and each sector's column sums to its net lending position.

The *balance-sheet matrix* (§2.2, Table 2) mirrors it for stocks, with + an asset and − a liability:

$$\sum_{s} B_{i,s} = 0 \qquad \text{for every instrument row } i, \tag{3}$$

$$\sum_{i} B_{i,s} = \mathrm{FNW}_{s}, \qquad \sum_{s} \mathrm{FNW}_{s} = 0, \tag{4}$$

so one sector's asset is another's liability, financial columns sum to sector financial net worth, and financial net worth sums to zero across the economy. Adding real assets to a sector's financial column gives its total net worth.

Because the matrices are derived from national accounts rather than written down abstractly, they cannot close on model-determined flows alone. Following the treatment the manual attributes to Zezza and Zezza, two residual constructs absorb the difference: a *residual transaction* in the flow matrix, capturing the net position of flows the model does not represent, and a *residual financial instrument* $\mathrm{RES}_s = \mathrm{FNW}_s - \mathrm{FNWM}_s$ in the balance sheet, capturing the net asset position of stocks it excludes. These are part of the specification, not a plug added by this replication, and they are transcribed as such.

3.2 Sectors

The model has seven institutional sectors — non-financial corporations (NFC), the power sector (PS), households (HH), monetary financial institutions (MFI), non-monetary financial institutions (NMFI), government (GVT) and the rest of the world (RoW) — plus a *production module* that is not a sector. The production module is where GDP expenditure (consumption, gross capital formation, exports less imports) flows in and GDP income (wages, gross operating surplus, indirect taxes) flows out; it holds no assets or liabilities, so its column in the transactions matrix sums to exactly zero rather than to a net-lending position. The split between the power sector and the production module follows the NACE classification: electricity generation (D35) is the power sector and all other productive activity, public and private, sits in the production module, with input–output intermediate consumption running between them in both directions.

The rest of the world is recorded in *net* terms for property income and financial stocks (§3.4.6): net interest and net dividends in the flow matrix, and a net interest-bearing position in the balance sheet. This matters for transcription — it is the only reading of the manual's own initialisation that is consistent with its tabulated MFI financial assets and liabilities — and it is where one of the pinned manual defects lives (Section 8).

3.3 The macro block and the closure

The manual's §3.2 sets out the high-level macroeconomic variables — GDP by expenditure and income, deflators, employment, the labour force and unemployment, productivity — in 23 equations, Eqs. (21)–(43). Two of its properties do most of the work in interpreting every result in this paper.

First, the closure is **demand-led**. Output adjusts to aggregate demand; real GDP is not held to a supply ceiling. Second, productivity growth is **Kaldor–Verdoorn** (Eq. (31)): a demand impulse raises productivity growth, and does so persistently, in the tradition Kaldor (1966) set out for exactly the UK growth question. Together these mean that a sustained fiscal impulse in DEFINE-UK raises output without being returned to potential by assumption, and raises the productivity path as it goes. That is a deliberate feature of the post-Keynesian closure, and it is also the reason the model's fiscal multipliers are not comparable, without adjustment, to those of supply-constrained models such as the OBR emulator (Ahmadi, 2026a); Section 7 makes that argument with the arithmetic attached.

3.4 Production, power, and the two UK-specific blocks

Production (§3.3) divides into the domestic production module (§3.3.1, Eqs. (44)–(70)), the power generation sector (§3.3.2, Eqs. (71)–(138)) and input–output calculations (§3.3.3). The production module carries a Leontief gross-output block, mark-up pricing over lagged unit costs, a wage-share and wage-rate distribution block, direct-energy prices and costs, and the productive capital aggregates.

The power sector is where the model's climate-policy content concentrates. It carries separate fossil and non-fossil generation capital, a fossil/non-fossil cost split, marginal-cost electricity pricing, a utilisation and forward-looking-expectation block that drives investment, credit-rationed capital formation, and a full financial account running through to leverage, illiquidity and credit rationing. Government renewable investment is triggered when non-fossil capacity falls short of a utilisation threshold relative to electricity demand. Version 1.1's regulatory policies enter here: a ban or regulation directly depreciates fossil capital, which is a balance-sheet event for whoever holds the claims against it — the transmission channel of Carney (2015) and Battiston et al. (2017) implemented inside the accounting rather than bolted onto it.

The second UK-specific block is housing: the dwelling stock is tracked by energy efficiency, so retrofit subsidies and regulation change household energy demand and emissions through the composition of the stock rather than through a reduced-form elasticity.

3.5 The ecosystem block

The ecological side (§3.1) covers energy, emissions, and ecological efficiency and technology. Emissions decompose into electricity and non-electricity components,

$$\mathrm{EMIS} = \mathrm{EMIS}_{\mathrm{ELEC}} + \mathrm{EMIS}_{\mathrm{NELEC}}, \tag{5}$$

an identity that turns out to be diagnostically useful: in the pinned run the two components sum exactly to the total, which is one of the two checks that rule out an extraction error on our side in the emissions divergence of Section 6.

3.6 Rates of return and financial behaviour

Interest rates (§3.4.7) run off a simple Taylor rule for the Bank Rate in logs, which enforces the zero lower bound; asset rates adjust towards a long-run mark-down on the base rate subject to a floor, and liability rates towards a mark-up over the corresponding asset rate plus a risk premium that depends on the sector's liability depreciation rate and on the MFI sector's own financial-liabilities-to-assets ratio. Default rates and credit rationing then feed investment. This is the Dafermos et al. (2018) climate–finance loop in UK form: a policy shock that impairs capital raises leverage, tightens credit, and reduces investment, with the tightening depending on the state of the banking sector.

3.7 Calibration and the status of the baseline

The manual estimates its key behavioural equations econometrically on UK data and calibrates the remainder, with §5's Tables 5 and 6 giving parameters and initial values respectively. The baseline is calibrated against external projections — OBR macroeconomic forecasts to 2030 with growth rates extended thereafter, NESO current-policy emission pathways, and NGFS scenario data (§4.1).

Two statements from the manual govern how any of its baseline numbers may be used, and both are quoted here because the rest of this paper depends on them. First, on the emissions path: the manual itself notes that the baseline reduction "falls significantly short" of the UK's 2035 NDC target — by design, since it is a current-policies baseline. Second, and more consequentially, on the baseline as a whole:

> *"The baseline of the model should also not be seen as a prediction or forecast"* (George and Dafermos, 2026a, §4.1).

Section 5 takes that literally. It is the reason the adapter serves deltas and never levels, and the reason no DEFINE-UK number appears anywhere in PolicyEngine Macro's forecast surfaces.

4 The source publications, the licence, and replication scope

4.1 What is published

Four documents and one repository make up the published record of DEFINE-UK, and they differ sharply in what they let a replicator check.

The manual. *DEFINE-UK Model Manual, Version 1.1* (George and Dafermos, 2026a), by Adam George and Yannis Dafermos of SOAS University of London, April 2026. This is the specification: the full equation listing by sector (§3), the calibration approach (§4) and the parameter and initial-value tables (§5). It is also the only source of published *numeric results*, and those stop at the baseline: Table 4 (p. 50) gives fifteen baseline quantities at 2025, 2030 and 2040 with 2025–40 means and standard deviations, and Table 5 embeds the design parameters of the model's scenarios. Table 6 (pp. 63–82) gives initial values for the endogenous variables, which is a large replication surface for a single period but is not a scenario result.

The two companion papers. Both are access-restricted. *Green Fiscal Policy in the UK: A Scenario Analysis* (George and Dafermos, 2026c) is the version 1.0 paper (SSRN abstract 6541398); its scenario set, per the abstract, is carbon pricing, green power-sector subsidies and green housing subsidies — a set that predates version 1.1's regulation instruments in any case. *Evaluating climate policy mixes in the UK: an E-SFC approach* (George and Dafermos, 2026b) is the version 1.1 paper (SSRN abstract 6588778) and is likewise not obtainable. An exhaustive search, catalogued in the adapter's `validation/published_targets.json`, found no open mirror of either on RePEc, SOAS pages or author pages.

The conference version. One open document survives: *Green fiscal policy in an empirical UK E-SFC model* (George and Dafermos, 2023), the October 2023 FMM conference paper, marked by its authors "preliminary and incomplete". Its results are figure-based and it predates DEFINE-UK 1.0, so it can support qualitative or coarse checks and nothing finer. Section 6 uses exactly two numbers from it and is explicit about what they can and cannot establish.

The repository. `https://github.com/DEFINE-model/DEFINE_UK_1.1`: the authors' R implementation, public, and carrying **no licence file**.

4.2 The licence position and what follows from it

Absent a licence, the default position is that no reproduction, redistribution or hosting right has been granted. This is treated as binding rather than as an oversight to be assumed away, and three consequences follow, each of which shapes the rest of this paper.

First, *nothing upstream is vendored.* The adapter contains no upstream code and no upstream data. It fetches the repository at pinned commit `846081a` at runtime, on the machine of whoever is running it, and executes it there.

Second, *nothing upstream is hosted.* DEFINE-UK is the only member of the PolicyEngine Macro suite that is local-only. Hosted tooling returns run instructions, never results computed from the upstream code, and continuous integration never fetches or runs it — which is why the oracle comparisons of Section 6 run only where a cached pinned run exists, with a hermetic subset of the gate running everywhere.

Third, *the route to a hostable model is reimplementation.* Section 8 describes the clean-room Python implementation of the published equations that exists for this reason, and its attribution stance: the model design belongs to its authors; the implementation is ours, AGPL-3.0, and independent.

4.3 Replication scope

The exercise has two tracks, held to the same standard applied to two different references.

The **oracle track** runs the authors' unmodified R code at the pinned commit and compares its outputs against the published figures — the manual's Table 4 for the baseline, and the scenario design parameters of Table 5 plus the two FMM 2023 anchors for the scenarios. The run is calendar-anchored on the population path: 1987Q1–2040Q4, with 2025 at $t = 153$.

The **clean-room track** reimplements the published equations in Python from the manual, with the upstream repository used strictly as a numerical *output oracle* whose source is never read to write an equation. Its milestone gates compare the reimplementation against the manual's own tables first and against the oracle second.

What is out of scope. No attempt is made to re-estimate the manual's econometrically estimated behavioural equations from source data, or to reconstruct the authors' input spreadsheets. Where the reimplementation needs data it takes it from official sources rather than from the upstream repository's `input/` directory — again a consequence of the licence, and incidentally a stronger test.

Execution is not validation. On 1 August 2026 the full upstream notebook rendered end to end through the adapter's runner, on R 4.3.0 under macOS, with the upstream unmodified and an `rstudioapi` shim the only accommodation, producing 151 output files: figures and tables for all four scenario blocks plus a multiplier summary. That established that the pinned code executes. It established nothing whatever about whether its outputs mean anything, and the record says so at that date. The adapter's validation gate is explicit on the point: nothing built on the repository may present model output as meaningful except by naming the target it rests on. Section 6 is the record of what that gate has since established — including where the answer is "nothing further can be established from what is published".

5 The adapter: how the model is run, and what it serves

5.1 Code architecture

The adapter is a small Python package, `define_uk` (src layout), at https://github.com/PolicyEngine/define-uk-model. It has two halves that never touch: a harness that runs the authors' R code, and a clean-room reimplementation that must not read it.

- `upstream.py` — clones the upstream repository into a local cache at run time at the pinned commit `846081a580a6033159d5c421632ad8f0b30d0ded`, and nothing else. Bumping the pin is an explicit, reviewed change, because the validation targets are tied to that revision.
- `runner.py` — renders the upstream R Markdown notebook in place, unmodified.
- `scenarios.py` — the curated scenario surface over the *cached* outputs of that run. It never runs R; if no cached run exists it raises with instructions to produce one.
- `validation.py` — the computed comparators: the baseline-versus-external calibration table, the green-public-investment multiplier, and the caveats that travel with it.
- `model/` — the clean-room reimplementation (Section 8): an equation registry that requires a manual reference on every equation, a Gauss–Seidel per-period solver, one module per manual §3 section, the §5 parameter and initial-value tables, and the §2.2 accounting matrices.

5.2 The scenario surface

The pinned run produces 22 scenario files across four blocks: government investment, housing regulation, power-sector regulation and mixed packages. The power-sector block is run under five expectation regimes — mutual trust, poor credibility, false confidence, no forward guidance, and unexpected enforcement — of which mutual trust is treated as canonical and the others carry an explicit variant suffix. Each of the 22 is registered, and each is checked against the policy switches its published description claims (Section 6.3).

The surface returns *annualised scenario-minus-baseline delta paths only.* The convention is fixed and tested: scenario minus baseline, annual mean of quarterly levels, calendar-anchored 1987Q1–2040Q4. Eleven delta points across four scenarios and three variables are pinned to relative tolerance 10^{-6} against the cached run, and the sign and ordering structure — deltas finite, baselines identical across block folders, emissions falling in every scenario by 2035, the mixed package dominating — is pinned exactly.

5.3 Deltas, never levels

The decision to serve deltas and never levels is the single most consequential design choice in the adapter, and it is a direct reading of two facts established in Sections 3 and 7: the manual's own statement that its baseline "should also not be seen as a prediction or forecast", and the measured distance between the baseline and UK outturns — 2025 real GDP growth of 4.66 per cent against an ONS 1.31, unemployment below the outturn, emissions above the DESNZ actuals they should start from.

Enforcing that is not left to documentation. Every scenario result carries a result type of "scenario deltas" and a mandatory caveat list, held in code and gated by tests: that the

exercise is experimental and gated at the ceiling described in Section 6.3; that baseline levels are not validated against outturns and must never be presented as forecasts; that the closure is demand-led, so multipliers sit above mainstream estimates; and that the pinned commit's baseline emissions path diverges from the manual's published table. Near-term DEFINE-UK levels are not competitive with the suite's OBR emulator (Ahmadi, 2026a) or its Bank of England SVAR (Ahmadi, 2026b), and are never presented as forecasts.

5.4 Integration

Within PolicyEngine Macro the model is exposed through the `pe-macro` command-line interface — `define-scenarios` to list the registry, `define-scenario` to produce delta paths — marked local-only and experimental, for the licence reason of Section 4.2. A further command bridges a scenario's household income deltas into PolicyEngine's UK microsimulation as a pre-tax earnings overlay, the same economic-assumptions pattern the suite uses for its other macro members, so that a climate-policy scenario can be read household by household through the tax and benefit system. Only numbers travel across that bridge; the unlicensed upstream code does not. The suite's reform-scoring entry point continues to refuse this model.

5.5 What continuous integration can and cannot check

Because the upstream is unlicensed, continuous integration never fetches or runs it, so the oracle comparisons skip there and run only where a cached pinned run exists. This would be a serious weakness if the gate were left at that, so the gate is split.

Every pinned reference number the oracle tests compare against lives in one committed artifact, `validation/reference_outputs.json`, rather than in test code — so that a local run and a CI run gate identical numbers, and no tolerance or reference can drift silently inside a test. The *hermetic* half of the suite runs everywhere, on every pull request, without R or any upstream fetch: it checks that the reference artifact is well-formed and internally consistent (for instance, that the committed multiplier reproduces from its committed cumulants), that it is pinned to the same commit the code is, that the committed calibration table keeps its exact schema, pinned external observations, gap arithmetic and headline divergences, that the scenario registry matches, and that the repository's own documentation still names the pinned commit. Changing any reference number requires editing the artifact deliberately, with a dated run-record entry — never editing a test.

6 The replication record

Four published-target gates were attempted against the pinned upstream run. One passes, one diverges and is pinned, and two are closed at the ceiling the publications allow rather than passed. This section reports each with its tolerance, its arithmetic, and — where the check could not have come out any other way — a statement to that effect at the point of claim rather than in a footnote.

Table 1 is the summary. Every gate in it is enforced by a test in the adapter repository against the cached pinned run, so drift fails loudly rather than silently.

6.1 Target 1a: the baseline macro block — PASS, with one year that proves nothing

The manual's Table 4 publishes fifteen baseline quantities at 2025, 2030 and 2040. Four of them — real GDP growth, unemployment, population aged 16 and over, and the labour force — form the macro block, and it is the one published target this replication passes. Tolerances were population and labour force ±0.01 m, growth ±0.31 pp, and unemployment ±0.2 pp.

Population and labour force reproduce exactly to 0.01 m at all three dates. Unemployment is within 0.2 pp at all three. Real GDP growth matches to **0.06 pp** at 2030 and 2040. The 2025–40 means and standard deviations of growth and unemployment reproduce the table at its printing precision. The run is calendar-anchored on the population path, 1987Q1–2040Q4, with 2025 at $t = 153$.

The 2025 growth gate is not independent of the deviation it admits. At 2025 the run gives 4.66 per cent against Table 4's 4.96 — a miss of −0.30 pp against a ±0.31 pp gate. The gate was set *after* the run, and the only year that uses any of it clears by 0.01 pp. As a test of agreement in 2025 it is doing no work: any tolerance chosen once the deviation is known will admit that deviation. What it is doing is stopping that year drifting further, which is worth having and is not the same claim. **The informative part of this gate is the pair of 0.06 pp matches at 2030 and 2040**, which were not fitted to anything.

The recorded reason for the 2025 gap is a convention we could not resolve. Table 4's note says only that "all quarterly values are annualised"; it does not say which annualisation, and the readings differ by more than the gap. The adapter's own comparator table uses the annual-mean convention, which is what gives 4.66. This is recorded as an unresolved convention mismatch, not as a demonstrated agreement — and it is the reason the tolerance exists at that width rather than a tighter one.

6.2 Target 1b: the emissions path — DIVERGENCE, pinned

The pinned code runs *below* the published Table 4 emissions, and the gap widens monotonically with the horizon:

	2025	2030	2040
Pinned run ($MtCO_2e$/yr)	393	343	249
Manual Table 4 ($MtCO_2e$/yr)	407	382	324
Pinned ratio, run/table	0.965	0.898	0.767
Deviation	−3.5%	−10.2%	−23.3%

The percentages are computed from the pinned ratios themselves — 0.965, 0.898 and 0.767, held in the committed reference artifact — and not from the rounded levels printed in the first two rows, which would give slightly different figures.

Two checks rule out an extraction error on this side. First, the components satisfy the model's own identity (5): $\text{EMIS}_{\text{NELEC}} + \text{EMIS}_{\text{ELEC}}$ sums exactly to EMIS in the run outputs. Second, the baseline is identical across every scenario folder, so no scenario contamination is involved. What remains is a vintage or calibration gap between the published table and commit `846081a`: the table was produced by some code vintage, and it is not this one.

The divergence has been raised with the upstream authors on their own issue tracker, so that the record here and the record there stay in step. Meanwhile the tests gate at the *observed* ratios, within ±0.02, rather than at agreement. That is a deliberate and limited claim, and it belongs in the first block of Table 1: a gate set at the divergence cannot fail on account of the divergence. It can and will fail if the gap moves, which is the only thing it is for.

6.3 Targets 2 and 3: closed at the achievable ceiling, not passed

The two scenario targets are the ones a reader of a climate-policy model most wants replicated, and they are the ones that cannot be. An exhaustive search, catalogued with locators in the adapter's `validation/published_targets.json`, established that **no machine-readable numeric scenario results are published for version 1.1**. The manual's results stop at Table 4's baseline plus the scenario *design* parameters embedded in Table 5; both companion papers (George and Dafermos, 2026c,b) are access-restricted with no open mirror; and the version 1.0 paper's scenario set — carbon pricing and subsidies — predates version 1.1's regulation instruments in any case.

There is therefore no scenario replication in this paper, and none is manufactured. What is verifiable has been verified, and it is design rather than results.

Scenario design, pinned exactly. Each cached scenario is checked to toggle *exactly* the policy switches its published description claims — no more and no fewer. The fossil-fuel-ban scenarios toggle the ban flag with its capacity and investment ban timings; the power subsidy toggles its subsidy switch alone; green public investment toggles government investment, green bonds and green power exactly; and the housing-subsidy variant pins the published 40 per cent subsidy rate. This is a real check against a published description, and it can fail — but it constrains the scenario's *inputs*, not its outputs.

Two anchors, and what they are worth. Two coarse numeric statements survive in the open FMM 2023 conference version (George and Dafermos, 2023), and both are used, at wide tolerances, because that document is a pre-1.0 vintage marked preliminary by its authors. The current-policies baseline is described as emitting "just under 350 $MtCO_2e$" in 2030; the cached run gives **342.8**, inside a band of 330–352. And the green-public-investment scenario is described in terms of "1% of GDP"; the cached run's real GDP delta peaks at +0.92 per cent, inside a band of 0.7–1.3.

Both should be read as regression pins, not as agreement tests, and there are three separate reasons, each of which would be sufficient on its own.

First, *the bands are ours.* Neither was published as a tolerance. Both were chosen by us to fit a verbal quantity after the run was in hand, which puts them in the same category as the 2025 growth gate of Section 6.1.

Second, *they cannot discriminate.* The GDP band is ±30 per cent around "around 1 per cent". The emissions band runs to 352, which is *above* the 350 that the phrase "just under" is supposed to bound — so a run at 351 would pass a gate named for a statement it violates. A test that a value contradicting the claim would pass is not testing the claim.

Third, and specific to the GPI anchor, *the two sides may not be the same quantity.* The transcription of the FMM text held in `published_targets.json` records the green-public-investment scenario as one in which "government invests an additional 1% of GDP in green projects by 2030" — that is, 1 per cent of GDP is the size of the policy *input*. The gate compares it against the

peak of the model's real GDP *response*. Those are different objects that happen to sit near each other numerically, and this replication has not been able to obtain the conference paper's figures to establish whether a comparable published GDP-response reading exists. On the reading in the record, the anchor is not like-for-like at all; on the more generous reading it is like-for-like inside a ±30 per cent band. Neither reading supports quoting +0.92 against ≈ 1 as a replication result, so it is not quoted as one anywhere.

What the two anchors do establish is worth stating plainly, because it is not nothing: the cached run has not moved since they were recorded.

The upgrade path. This ceiling is not permanent and its lifting conditions are specific. If the authors publish scenario tables, or either SSRN paper becomes accessible, the published-figure gate reopens and targets 2 and 3 can be attempted as replications rather than closed as design checks. Until then, the honest verdict is the one in the adapter's own record: *closed at the achievable ceiling*, which is not a pass.

6.4 What the whole record does and does not license

The adapter's validation file carries a status header, and for a period after target 1a passed that header still read "NOT VALIDATED. No published DEFINE-UK result has been reproduced" — which put it in contradiction with every page that cited it as the record. The correction belongs in this paper because this paper cites that file as the record. It now reads **PARTIALLY VALIDATED** and breaks the claim down target by target: 1a pass, 1b divergence pinned, targets 2 and 3 closed at the achievable ceiling rather than passed, reimplementation milestone 1 pass and milestone 2 outstanding.

Nothing in that licenses presenting DEFINE-UK scenario output as a validated result. The operative rule, held in the repository and applied throughout this paper, is that a claim about this model must name the target it rests on.

7 External calibration and the multiplier

Replication against a publication and calibration against the world are different questions, and a model can pass the first while failing the second badly. DEFINE-UK does. This section reports the second, and then reports the one number from the model that has circulated most widely — a cumulative green-investment multiplier of 1.78 — with the reasons it cannot be read against anybody else's multiplier.

7.1 The baseline against outturns

Since 5 August 2026 this comparison is an artifact rather than a paragraph. A function in the adapter recomputes every row from the cached pinned run against vintage-exact external observations; the result is committed as `validation/baseline_vs_external.csv`; and a test fails on any drift between the committed table and a recomputation. Table 2 reproduces it.

The 2025 growth gap is the one to keep in view: 4.66 per cent against an ONS outturn of 1.31, a gap of well over three percentage points on the single most familiar macroeconomic quantity there is. Unemployment sits about half a point below the outturn, the labour force about a million and a half high, and 2024 emissions some 30 $MtCO_2e$ above the DESNZ actuals the model should be starting from. None of this is presented as a defect of the model as its

authors intend it to be used — the manual is explicit that the baseline is not a forecast — but it is decisive for how the adapter may serve it. It is the measured basis for Section 5.3's deltas-only rule.

7.2 The 1.78 multiplier, and why it is not comparable to any published one

The model's cumulative green-public-investment multiplier is **1.78** by *our own* computation: cumulative ΔGDP of 209.9 over cumulative ΔSPEND_GVT of 118.2, nominal, over the full simulation, on the green-public-investment-plus-green-bonds scenario against its baseline. The arithmetic reproduces ($209.9/118.2 = 1.7758$) and the number is pinned in the test suite to ± 0.01 with its cumulants pinned to ± 0.5.

Two labels must travel with it. The first is that **the number is ours, not the authors'**: no published DEFINE-UK multiplier corresponds to it, and it is a quantity this replication defined and computed, not one it reproduced. The second is that **it is not like-for-like against any published multiplier**, on three separate grounds, all of which push in the same direction.

Horizon. Ours is a cumulative ratio over the whole simulation to 2040. Published green-spending multipliers, and the OBR's capital-spending impact multiplier of roughly 1.0, are short-horizon figures — impact or first-few-year. A long-horizon cumulative ratio is *expected* to sit above them, so setting 1.78 beside a short-horizon number and reading a disagreement is a units error dressed as a finding.

Denominator. SPEND_GVT is total government spending excluding wages. The manual's Table 6 components imply it is other government consumption plus social benefits plus government gross capital formation ($58.03 + 82.8 + 20.76 = 161.59$ against a tabulated 161.6), of which government investment — the actual policy lever — is only **12.8 per cent**. The denominator is therefore an endogenous aggregate rather than the policy impulse, and it moves the wrong way for the ratio: the scenario cuts unemployment, social benefits fall, the denominator shrinks, and the measured multiplier is **biased up**. The clean comparison — cumulative ΔGDP over cumulative ΔGCF_GVTG, the actual impulse — **has never been computed**. Until it is, 1.78 should be read as upper-leaning, and that is stated wherever the number is served.

Closure. As Section 3 sets out, the model is demand-led with Kaldor–Verdoorn productivity (Eq. (31)) and real GDP not held to a supply ceiling (Eq. (38)). A multiplier above one over fifteen years is what this closure *implies*. It is not a symptom of an implementation error, and it is not a finding about the world; it is a property of the model one has chosen to run. That is why the number is labelled rather than corrected.

These three caveats are not editorial: they are carried in code as a constant and gated hermetically by a test that fails if the number is served without them.

A withdrawn comparator. Earlier versions of this project's surfaces set 1.78 beside "IMF green-spending multipliers of 1.1–1.5", attributed to a 2021 IMF publication. That publication could not be located in either repository or anywhere else, and the attribution has been **withdrawn rather than repaired**, on the principle that an unidentifiable citation is worse than none. The only IMF reference this project actually holds is Batini et al. (2014), which is a different paper reporting a different quantity: first-year multipliers of 0–1 for advanced economies.

No published multiplier is quoted alongside 1.78 in this paper, because none of them measures the same object.

The upstream multiplier table is unusable. The upstream run also produces its own `tables/Multiplier_Summary.csv`, and it is documented unusable rather than quietly ignored. Its impact, four-quarter and eight-quarter multiplier columns are *identical across all eight scenarios* (−32.22, −0.88, 2.6), which no correct scenario-specific multiplier can be; and its cumulative government-spending figure of 104.22 matches the cumulative delta of *no* variable in the scenario file. One column does reconcile — its cumulative GDP delta of 250.76 is the cumulative quarterly real-GDP delta of the green-public-investment scenario, 250.6 by our own sum — which is what makes the failure of the others diagnostic rather than ambiguous. Neither the table nor the multiplier of roughly 2.4 once derived from it is quoted anywhere.

8 The clean-room reimplementation

The upstream code cannot be hosted, so the route to a hostable DEFINE-UK is to implement the published equations from scratch — the same play the suite ran for the OBR macroeconometric model (Ahmadi, 2026a). This section reports where that stands, and treats it as more than an engineering task: reading a manual closely enough to run it is itself a form of validation, and it has produced findings the R source could not.

8.1 Protocol

The specification sources are the manual (George and Dafermos, 2026a), the companion papers, and Dafermos et al. (2017) for framework-level derivations the manual references. Three rules make the exercise clean-room rather than a translation.

1. **The upstream repository is an output oracle only.** Its numerical outputs are compared against ours in the oracle tests. Its source code is never read to write an equation here.

2. **Every equation carries its provenance.** Each registered equation declares a `manual_ref` of the form "§3.3.2 eq. (84)", validated by a regular expression at construction time, and an equation without one does not merge. Two registrations citing the same reference are rejected, because a copy-pasted provenance would make an equation *look* transcribed while implementing something else — a failure the per-section contiguity checks cannot see on their own, since a duplicate shows up there only as a hole somewhere else.

3. **Data come from official sources**, not from the upstream repository's input spreadsheets — again a licence consequence, and incidentally a stronger test.

The architecture is an equation registry, a Gauss–Seidel per-period solver that iterates all registered equations to convergence, one module per manual §3 section, the §5 parameter and initial-value tables, and the §2.2 accounting matrices. Stock-flow norms are asserted each period, not assumed.

8.2 Milestone 1: PASS

Milestone 1 (4 August 2026) transcribes the §2.2 transactions and balance-sheet matrices and the full §5 parameter and initial-value tables, and requires that every Table 1 and Table 2 row and column identity hold on the §5 initial values within the manual's own four-significant-figure printing precision.

It passes. The accounting test module collects **62** tests, of which **44** are the Table 1/2 row and column identities themselves: 20 transactions rows, 6 transactions columns, 7 balance-sheet rows and 11 balance-sheet columns. The remaining 18 are the residual-instrument and model-financial-net-worth definitions, the economy-wide closure $\sum_s \mathrm{FNW}_s = 0$, and the two tests that pin the one manual inconsistency this milestone found.

That inconsistency is worth stating because it is the template for everything in Section 8.4. Table 6's rest-of-world lending entry omits the net-dividends term of Eq. (383), so the MFI and RoW transaction columns miss lending by ∓ 5.44. This is not an inference: the manual itself tabulates overall lending at 5.44 where, by its own account, it "should equal 0". Rather than adjust a value or widen a tolerance until the matrix closed, the discrepancy is pinned exactly, so the implementation reproduces the manual including its defect and a future correction upstream fails the test rather than passing silently.

8.3 Milestone 2: IN PROGRESS, not passed

Milestone 2 covers the high-level macro and production blocks (§3.2–3.3), gated on the baseline GDP path against the oracle. Three of its four sections have landed:

Manual section	Content	Equations	Range
§3.2	High-level macroeconomic variables	23	(21)–(43)
§3.3.1	Domestic production module	27	(44)–(70)
§3.3.2	Power generation sector	68	(71)–(138)
Total		**118**	**(21)–(138)**, contiguous

Each section solves as a system under its own test module, and contiguity is checked structurally rather than asserted: the equation numbers parsed out of the provenance strings must form an unbroken range, so a skipped equation fails a test rather than going unnoticed.

Milestone 2 is not passed. §3.3.3, the input–output calculations, is not implemented, and the oracle baseline comparison — the milestone's actual gate — has not been run. Reporting 118 contiguous equations as a milestone pass would be reporting effort as evidence. What can be said is narrower and true: three of four sections are transcribed and internally consistent against the manual's own initial values, at the precision those values are printed to, with every disagreement enumerated below rather than absorbed.

8.4 Thirteen gap records

Where the manual is incomplete, silent, or inconsistent with itself, the implementation does not patch, tune or tolerate. It implements what is printed and records the gap as a machine-readable entry, pinned by a test that fails if the situation changes — including if the DEFINE team later publishes a value. There are **thirteen** such records: one in §3.2, two in §3.3.1 and ten in §3.3.2. Appendix B lists them all. Three general observations matter more than the individual entries.

They are not all the same kind of thing, and one is not a manual defect at all. The thirteen cover missing parameters, symbols the body uses and the tables never define, variables the tables carry and the body never uses, a printed inequality that contradicts the prose above it, and quantities Table 6 simply omits. One of them — the expected capital path over the planning horizon of Eq. (89) — records *our* modelling choice rather than a defect: the manual never says what capital stock the sector expects to hold at $t+\tau$ and the model carries no forecast of it, so real capital and electricity demand are held at current values across the horizon. It is recorded in the same place as the others precisely so that a choice of ours is not mistaken for a reading of theirs. It is inert whenever the baseline credibility switch is zero.

Some of them have consequences, and the consequences are stated. §3.3.2 as published *cannot be simulated forward*: six parameters it uses are absent from Table 5, and two variables — the fossil and non-fossil capital profit rates of Eq. (99) — are never defined anywhere in the manual. At the defaults this is not cosmetic. Credit rationing degenerates to a constant, the green/fossil investment split degenerates to 50:50 against the 69:31 that Table 6 implies, and desired power-sector investment turns negative: held at its own initial values the section returns power-sector capital formation of -0.021 against a tabulated $+2.15$. The 50:50 default is visibly wrong and is pinned as such rather than tuned to fit, because tuning it would replace a documented hole with an undocumented guess. This is the single largest gap in the section, since that split is what steers investment green.

Where the manual could have been guessed, it was not. Table 5 reuses non-financial-corporation parameter values for the power sector in at least nine places, so reusing them once more for Eq. (138)'s credit-rationing slopes would have been a defensible inference — but it would have been *our* inference rather than the manual's instruction, and it does not reproduce Table 6 in any case. It was not done. Similarly, Eq. (96)'s missing intercept can be backed out as roughly 0.0210 if Table 6 is read as a steady state; that value is recorded and *not used*, because the equation's two lagged terms make the steady-state reading an assumption rather than a measurement.

8.5 The largest disagreements

Beyond the gaps, a number of printed identities disagree with Table 6 by one to three orders of magnitude more than its printing noise. All are implemented as printed and pinned individually. Four deserve statement here.

Eq. (84): a factor of 3.04. The electricity price rule, a fixed mark-up over marginal cost, gives 0.9725 against a tabulated 0.3198, implying a mark-up of -0.267 — electricity sold below marginal cost. Both sides are independently corroborated: marginal cost by Eqs. (82) and (83) to 4.5×10^{-5}, and the price by Eq. (74) to 3.7×10^{-4}. So the manual's two chains meet at a contradiction. The most economical account is not a wrong number but a *missing equation*: §5 calibrates an electricity-price long-run formation rule — a switch time and a long-run price — that §3 never prints, and Eq. (84) is the only price equation printed, with no long-run term and no switch. Nothing is implemented for the unprinted rule, because there is nothing printed to implement.

One deflator, counted once. Eq. (27) gives real gross capital formation of 111.88 against a tabulated 112.30 — 0.37 per cent, roughly ten times the rounding noise — and Eq. (25) reproduces the tabulated real GDP only using 112.30. This is emphatically *one* finding and not five: gross capital formation, its NFC, household and government components, and the two capital stocks all show nominal-to-real ratios of 1.0309–1.0312 where the equations prescribe a production deflator of 1.035, and the same 0.4 per cent reappears in §3.3.2's Eqs. (104), (105), (132) and (133). Table 6's entire capital block is deflated at 1.031 — most plausibly a national-accounts investment deflator where the equations want the production deflator. Counting it once matters: reported as five independent discrepancies it would look like a section riddled with errors rather than one systematic property of one table's capital rows.

Eq. (107): a model-wide sign problem, not a power-sector one. Eq. (107) has a determinate-sign right-hand side and Table 6 tabulates the opposite. This qualification belongs with the finding wherever it is quoted: **it is not a power-sector defect**. The manual prints the same rule for three other sectors — Eqs. (177), (231) and (284) — every coefficient is positive, and Table 6 tabulates all four transfers negative, with magnitudes out by factors of 1.7 to 4.6 as well. Whatever is wrong is wrong in the interest-bearing asset transfer rule, or in Table 6's sign convention for it, *model-wide*; §3.3.2 is simply where it is met first. Confirming that requires §3.4.1, §3.4.3 and §3.4.4, which are not implemented yet, so the claim is scoped accordingly.

A normalisation that propagates. Eq. (61) gives a fuel price of 0.6788 against a tabulated 1, where Table 5 says of the relevant parameter that "at the initial condition fuel price is normalised to equal 1". A third witness supports the normalisation: Table 6 tabulates the power sector's nominal and real fuel intermediate consumption at the same value, which is only possible at a fuel price of 1. The gap does not contaminate any §3.3.2 identity, because the affected quantity is tabulated. In a *simulation* it cuts that input 32 per cent and carries essentially undiluted into marginal cost and the electricity price — and it does not cancel Eq. (84): 3.04×0.68 still leaves the price at 2.07 times Table 6's.

8.6 The audit, and what it withdrew

On 12 August 2026 the pinned findings were audited against the manual PDF and recomputed independently of the sector modules: all 661 transcribed §5 values were re-checked against the page each source comment names, and every pinned discrepancy was reproduced. The arithmetic held. The interpretations did not entirely, and the corrections are recorded here because a replication that only ever adds findings is not being audited.

Four findings re-characterised. Eqs. (51), (59), (94) and (111) had been reading as the manual contradicting itself. They are not. Table 5 describes each of the parameters concerned as a *historical mean* — "calibrated so the long-run mark-up at initial utilisation equals the mean of" two sample windows, "calculated using the mean over past data", "set as the mean of past implied values", "taken as the mean of past data". A parameter described that way never claimed to reproduce the manual's own initial period. These are therefore **first-period jumps** that a run started from Table 6 will take in its first quarter, not defects. They still matter — one of them moves a price that steers green investment, by 60 per cent — and they are still pinned. They are simply not evidence that the manual disagrees with itself, and the distinction turns

on the manual's own stated provenance for each parameter, which is why Table 5's remarks are treated as part of the specification.

One claim withdrawn outright. An earlier version of the reimplementation notes claimed of Eq. (51) that "the equation is the corroborated one", because Eqs. (49) and (50) imply a value close to it. That inference does not hold and is withdrawn: Eq. (49) reads *lagged* unit costs, and the corroborating value is what it gives only if unit costs were flat into the initial period. Honouring the lag, the tabulated mark-up implies +0.91 per cent quarterly unit-cost growth and the equation's implies −0.12 per cent — deflation — and the manual's own initial period is growing and inflating, which favours the tabulated value if anything. On the evidence a single-period snapshot provides, neither side is the outlier, and the finding now says so.

One finding unified. Eqs. (27), (104), (105), (132) and (133) were counted as five and are now counted as one: the single deflator finding above.

One new gap found. The unprinted electricity-price rule described above was identified by the audit, and is the likeliest account of Eq. (84).

Two of the findings have been filed publicly on the authors' own tracker — the emissions vintage divergence of Section 6.2, and §3.3.2's missing parameters and undefined variables together with Eqs. (84) and (61) — so that the record here and the record there stay in step. Both were re-verified in the audit.

8.7 Attribution

The suite's standard adapted-model stance applies. The **model design belongs to its authors**; the **implementation is ours**. Every surface that presents results names DEFINE-UK and its authors, links the manual and papers, and states that this is an independent Python implementation of the published equations — not the authors' code, and not endorsed by them. The implementation is AGPL-3.0, which is what makes hosting possible without any upstream licence.

9 Limitations and conclusion

Six limitations should frame any use of this model, and each is disclosed in the codebase as well as here.

No scenario result has been replicated. This is the limitation that matters most, because scenarios are what the model is for. No machine-readable numeric scenario results are published for version 1.1, so targets 2 and 3 are closed at the achievable ceiling rather than passed (Section 6.3). What is gated is scenario *design* — the exact policy-switch set each cached scenario toggles — plus two coarse anchors from a pre-1.0 conference version that are regression pins, not agreement tests, one of which admits values contradicting the sentence it is named for. Nothing in this paper establishes that DEFINE-UK's scenario *outputs* reproduce anything published. The condition for lifting this is specific: the authors publish scenario tables, or either SSRN paper becomes accessible.

The baseline is not a forecast, and its levels are far from outturns. 2025 real GDP growth of 4.66 per cent against an ONS 1.31; unemployment below the outturn; 2024 emissions some 30 $MtCO_2e$ above the DESNZ actuals. The manual itself says the baseline "should also not be seen as a prediction or forecast". The adapter therefore serves deltas and never levels, and no DEFINE-UK number appears in any of the suite's forecast surfaces. Users who want UK levels should use the suite's OBR emulator or its Bank of England SVAR.

One published match is a tolerance, not an agreement. Target 1a passes, and the 0.06 pp matches at 2030 and 2040 are real. The 2025 growth pass is not: it clears a ±0.31 pp gate set after the run by 0.01 pp, on an annualisation convention Table 4 does not state precisely enough to resolve. It is reported as a pass of the gate and not as a match.

The emissions path diverges, and the cause is not established. The pinned commit runs 3.5 per cent below the published table in 2025 and 23.3 per cent below by 2040. Two checks rule out extraction error on this side, which leaves a vintage or calibration gap between the published table and this commit — but which vintage produced Table 4 is not known here, and the question is with the authors. Anyone using the adapter's emissions deltas should know that the level path they are differenced from is not the published one.

The multiplier is ours and is not comparable. 1.78 is a quantity this replication defined, on a horizon, a denominator and a closure that all bias it upward relative to any published short-horizon, impulse-based multiplier (Section 7.2). The clean number — cumulative ΔGDP over cumulative ΔGCF_GVTG — has never been computed, and computing it is the obvious next piece of work. A previously quoted IMF comparator has been withdrawn as unlocatable.

The reimplementation is early, and the model is local-only until it is not. Milestone 1 is passed and milestone 2 is in progress with 118 of the manual's equations transcribed and its oracle gate unrun; §3.3.3, the seven sectoral sections, the ecosystem block and the scenarios remain. Until it is complete, DEFINE-UK cannot be hosted at all, because the upstream is unlicensed. Two operational caveats attach: the full-notebook runtime and the upstream's R dependency set are untested in continuous integration, and the oracle half of the validation suite runs only where a cached pinned run exists.

Not an official product. This is an independent open-source replication and reimplementation, built from the published manual and papers. It is not produced by, endorsed by, or speaking for the DEFINE team; where this work and the publications disagree, the publications are ground truth and the disagreement is recorded as such. The model design belongs to its authors; the implementation is ours.

Conclusion

The useful thing about DEFINE-UK for a replication programme is that it fails in an instructive way. The code runs, unmodified, at a pinned commit. One published table's macro block reproduces, two of its rows to 0.06 pp. Another row of the same table diverges by 23 per cent at the horizon. And the results that would matter most — what happens to output, emissions, employment and balance sheets under a fossil-fuel ban or a green investment programme —

cannot be checked at all, not because the model is wrong but because the numbers were never published.

That combination is common and it is usually reported as a pass. The alternative this paper argues for is a record with three separations held throughout: between checks that can fail and checks that cannot, at the point of claim rather than in a footnote; between a number we computed and a number somebody published; and between a ceiling imposed by the publications and a limitation of the replication, with the lifting condition named in each case. Applied here, that record supports a narrow set of claims — a baseline macro block reproduced, an emissions divergence pinned, scenario design verified, deltas served with mandatory caveats, and a clean-room reimplementation 118 equations into its second milestone — and refuses a much larger set that the same underlying work could have been made to look like.

Thirteen defects in a published manual, four of them re-characterised and one claim withdrawn on audit, is the other half of the return. Reimplementing published equations is expensive, and its usual justification is licensing. The justification on display here is different: it is the only reading of a model close enough to find out that a section as published cannot be simulated forward, that a price identity is off by a factor of three because an equation is missing from the document, and that one table's capital block is deflated at the wrong price. None of that was visible from running the code.

References

Ahmadi, V. (2026a). An open OBR macro model for UK fiscal scoring. Policyengine macro working paper, PolicyEngine.

Ahmadi, V. (2026b). An open replication of the Bank of England structural VAR. Policyengine macro working paper, PolicyEngine.

Batini, N., Eyraud, L., Forni, L., and Weber, A. (2014). Fiscal multipliers: Size, determinants, and use in macroeconomic projections. Technical Notes and Manuals 14/04, International Monetary Fund.

Battiston, S., Mandel, A., Monasterolo, I., Schütze, F., and Visentin, G. (2017). A climate stress-test of the financial system. *Nature Climate Change*, 7:283–288.

Carney, M. (2015). Breaking the tragedy of the horizon — climate change and financial stability. Speech at Lloyd's of London, 29 September.

Caverzasi, E. and Godin, A. (2015). Post-Keynesian stock-flow-consistent modelling: A survey. *Cambridge Journal of Economics*, 39(1):157–187.

Dafermos, Y., Nikolaidi, M., and Galanis, G. (2017). A stock-flow-fund ecological macroeconomic model. *Ecological Economics*, 131:191–207.

Dafermos, Y., Nikolaidi, M., and Galanis, G. (2018). Climate change, financial stability and monetary policy. *Ecological Economics*, 152:219–234.

Department for Energy Security and Net Zero (2025). UK territorial greenhouse gas emissions national statistics. 2024 provisional figures; 2023 final figures.

Dewald, W. G., Thursby, J. G., and Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking project. *American Economic Review*, 76(4):587–603.

George, A. and Dafermos, Y. (2023). Green fiscal policy in an empirical UK E-SFC model. Technical report, Forum for Macroeconomics and Macroeconomic Policies (FMM) Conference. Conference version, marked "preliminary and incomplete".

George, A. and Dafermos, Y. (2026a). DEFINE-UK model manual, version 1.1. SOAS University of London.

George, A. and Dafermos, Y. (2026b). Evaluating climate policy mixes in the UK: An E-SFC approach. Working paper, SSRN. SSRN abstract 6588778; DEFINE-UK 1.1. Access-restricted.

George, A. and Dafermos, Y. (2026c). Green fiscal policy in the UK: A scenario analysis. Working paper, SSRN. SSRN abstract 6541398; DEFINE-UK 1.0. Access-restricted.

Ghenis, M. and PolicyEngine contributors (2025). PolicyEngine: Open-source tax–benefit microsimulation and macroeconomic models. https://policyengine.org.

Godley, W. and Lavoie, M. (2007). *Monetary Economics: An Integrated Approach to Credit, Money, Income, Production and Wealth*. Palgrave Macmillan, Basingstoke.

Kaldor, N. (1966). *Causes of the Slow Rate of Economic Growth of the United Kingdom: An Inaugural Lecture*. Cambridge University Press, Cambridge.

Nikiforos, M. and Zezza, G. (2017). Stock-flow consistent macroeconomic models: A survey. *Journal of Economic Surveys*, 31(5):1204–1239.

Office for Budget Responsibility (2026). Economic and fiscal outlook, march 2026. Technical report, Office for Budget Responsibility.

Office for National Statistics (2026a). Gross domestic product: Chained volume measures. Quarterly national accounts, series ABMI; vintage 26 July 2026.

Office for National Statistics (2026b). Unemployment rate (aged 16 and over, seasonally adjusted), series MGSX. Labour market statistics; vintage 26 July 2026.

Appendices

A The published baseline table

Table 3 reproduces the manual's Table 4 (George and Dafermos, 2026a, p. 50) in full, since it is the only published numeric result this replication has to work with and only its first four rows plus emissions are gated. Its note reads: "All quarterly values are annualised and the mean and standard deviation are calculated from 2025-2040." It does not state *which* annualisation, which is the unresolved convention behind the 2025 growth gap of Section 6.1.

B The thirteen gap records

Table 4 lists every machine-readable gap record carried by the clean-room reimplementation as of 12 August 2026. Each is pinned by a test that fails if the situation changes — including if the DEFINE team publishes a value for a defaulted parameter.

C Run record

Every entry below is dated and held in the adapter's `VALIDATION.md`. Changing any pinned reference number requires editing the committed reference artifact deliberately and adding an entry here — never editing a test.

Table 5: Dated run record

Date	Basis	Result
2026-08-01	`846081a`, R 4.3.0	Full upstream notebook renders end to end, unmodified, with an `rstudioapi` shim the only accommodation. 151 output files. **Execution is not validation**: no comparison against published figures yet, so every target remains pending.
2026-08-04	`846081a`, cached run	Table 4 replication run. Macro block passes (population and labour force exact to 0.01 m; growth exact at 2030/2040, −0.30 pp at 2025; unemployment within 0.2 pp; 2025–40 means and standard deviations match). Emissions diverge as target 1b records.
2026-08-04	clean-room, manual only	**Milestone 1 PASS.** §2.2 matrices and §5 Tables 5–6 transcribed in full; every Table 1/2 row and column identity holds on the §5 initial values at the manual's four-significant-figure printing precision. The rest-of-world lending inconsistency pinned exactly.
2026-08-04	`846081a`, cached run	Scenario surface gated against the pinned oracle: all 22 scenario files across four blocks registered, annualised delta paths only with mandatory caveats, 11 pinned delta points, sign and ordering structure pinned.
2026-08-05	`846081a`, cached run	**Targets 2 and 3 closed at the achievable ceiling**; scenario-design gate added with the two FMM 2023 anchors; external comparator table computed and committed; the 1.78 multiplier computed, replacing the unusable upstream multiplier table; published sources catalogued.
2026-08-11	clean-room, manual only	§3.2, §3.3.1 and §3.3.2 landed: 118 contiguous equations, Eqs. (21)–(138). §3.3.3 and the oracle gate outstanding, so **milestone 2 is not passed**.
2026-08-12	audit, manual + committed artifacts	All 661 transcribed §5 values re-checked against the manual page each source comment names; every pinned discrepancy recomputed independently of the sector modules. All arithmetic reproduced. Four findings re-characterised as first-period jumps, one claim withdrawn outright, five discrepancies unified into one deflator finding, one new gap found. Suite: 236 passed, 59 skipped.

Table 1: DEFINE-UK replication gates

Gate	Verdict	Enforced by
Cannot fail as evidence of agreement — pins and post-hoc tolerances, reported as such		
2025 real GDP growth within ±0.31 pp	clears by 0.01 pp	tolerance set after the run
FMM 2023 anchor: GPI GDP peak in $[0.7, 1.3]$ %	in band (+0.92)	band chosen by us, ±30 %
FMM 2023 anchor: 2030 baseline emissions in $[330, 352]$	in band (342.8)	band admits values it is named against
Emissions ratios within ±0.02 of 0.965/0.898/0.767	holds	pin at the *observed* divergence
11 scenario delta points, rel. 10^{-6}	holds	pin of the cached run against itself
Reference artifact well-formed and self-consistent	holds	hermetic artifact check
Can fail — comparisons against a publication		
Manual Table 4 macro block, S1 baseline	**PASS**	`test_replication_baseline.py`
population 16+ and labour force, ±0.01 m	exact	
real GDP growth at 2030 and 2040	match to 0.06 pp	
unemployment, ±0.2 pp	within tolerance	
Manual Table 4 total emissions, S1 baseline	**DIVERGENCE**	`test_replication_baseline.py`
Scenario policy-switch sets vs published descriptions	exact	`test_scenario_design.py`
Published numeric v1.1 scenario results	**CEILING** — none exist	`published_targets.json`
Baseline vs ONS/DESNZ/OBR outturns	**COMPUTED** — gaps material	`test_validation.py`
Clean-room milestone 1 (§2.2 matrices on §5 values)	**PASS**	`test_accounting.py`
Clean-room milestone 2 (§3.2–3.3)	**IN PROGRESS**	`test_macro.py`

Notes: The first block is separated because none of those six checks can produce evidence that this replication agrees with a publication. The growth tolerance was set after the deviation it admits was known; the two anchor bands were chosen by us to fit a verbal quantity and were never published as tolerances; the emissions gate is set *at* the observed divergence, so it detects further drift and asserts no agreement; the delta-point and artifact checks compare the cached run and the committed artifacts against themselves. All verdicts are against upstream commit `846081a`, dated 4–5 August 2026 for the oracle gates and 11 August 2026 for milestone 2.

Table 2: DEFINE-UK baseline against external observations

Quantity	Model	External	Units	Verdict
Real GDP growth, 2025	4.66	1.31	% y/y	far above outturn; never a forecast
Real GDP growth, mean 2025–40	2.4	1.65	% y/y	high vs official long-run
Unemployment, 2025	4.39	4.88	%	below outturn
Population 16+, 2025	55.66	56.0	millions	consistent
Labour force, 2025	35.97	34.5	millions	slightly high
Emissions, 2024	401.5	371.0	$MtCO_2e/yr$	above the actuals it should start from
Emissions, 2030 vs NDC	342.8	260.0	$MtCO_2e/yr$	short of the NDC by design

Notes: Model column from the cached pinned run at commit `846081a`, annual-mean convention — which is why 2025 growth reads 4.66 where the manual's Table 4 prints 4.96, the same anchoring gap recorded under target 1a (Section 6.1). Sources: ONS quarterly national accounts, vintage 26 July 2026 (Office for National Statistics, 2026a); ONS series MGSX, same vintage, 2025 quarterly path 4.6/4.7/5.0/5.2 giving an annual mean of 4.88 (Office for National Statistics, 2026b); ONS 16+ population and labour force; DESNZ UK territorial greenhouse gas emissions, 2024 provisional, with 2023 final at 384 (Department for Energy Security and Net Zero, 2025); OBR March 2026 long-run growth assumption (Office for Budget Responsibility, 2026); and the UK NDC of 68 per cent below 1990 by 2030. The final row is not a failure: the manual states that its current-policies baseline falls short of the NDC by design.

Table 3: DEFINE-UK Model Manual v1.1, Table 4: key characteristics of the baseline scenario

Variable	2025	2030	2040	Mean	St. dev.
Real GDP growth (%)	4.96	2.35	2.01	2.35	0.71
Unemployment (%)	4.31	4.73	4.72	4.71	0.14
Population (millions)	55.66	57.74	61.61	58.68	1.87
Labour force (millions)	35.97	37.12	39.13	37.62	1.00
Proportion of non-fossil electricity generation (%)	57.93	61.77	77.14	65.92	6.35
Total emissions ($MtCO_2e$/year)	407.32	381.76	324.37	367.10	25.46
Emission price (£/tCO_2e)	8.01	10.58	15.71	11.86	2.44
Green investment (% of GDP)	1.00	1.07	1.30	1.13	0.10
NFC default rate (%)	0.28	0.33	0.45	0.36	0.05
NFC credit rationing (%)	3.81	2.56	4.79	3.49	0.74
Power sector default rate (%)	0.16	0.11	0.13	0.12	0.01
Power sector credit rationing (%)	47.22	35.44	21.09	30.64	9.04
Imports (% of GDP)	39.99	39.97	41.18	40.31	0.51
Exports (% of GDP)	38.58	40.19	44.15	41.21	1.80
Real effective exchange rate (index, 2022 = 1)	1.00	1.00	0.98	0.99	0.01

Notes: "Population" is the population aged 16 and over, not total population. Rows 1–4 are the macro block of target 1a (Section 6.1); row 6 is target 1b (Section 6.2). The remaining nine rows are published but not gated: no tolerance has been set for them and no verdict is claimed on them here.

Table 4: Machine-readable gap records in the clean-room reimplementation

Section	Record	What it records
§3.2	intercept of Eq. (31)	Table 5 never tabulates the Kaldor–Verdoorn intercept; the α_0 block ends one entry earlier. Defaulted to 0.0.
§3.3.1	direct-energy symbols	Eqs. (59)–(63) use a "direct energy" subscript family that appears nowhere in Tables 5–6, while the tables carry a non-electric-energy family that appears nowhere in the body. They are the same variables: Eqs. (60) and (62) reproduce the tabulated values to 1.1×10^{-5} and 5.8×10^{-5} under that identification. The tabulated names are used.
§3.3.1	split green capital stocks	Eq. (68) sums the NFC and government real green capital stocks; Table 6 tabulates only their sum. Nothing defaulted — both are exogenous to the section — and the identity is checked against the tabulated sum instead.
§3.3.2	power-sector depreciation rate	Seven equations use one rate; Table 6 tabulates only the two split rates, which are equal, so the identification is unambiguous and the rate is seeded from them. If a revision ever splits them, two equations need rewriting rather than reseeding.
§3.3.2	intercept of Eq. (96)	Not in Table 5. The table carries the two intercepts of a *split* investment rule that v1.1's body replaced, and neither can be substituted without guessing which. Defaulted to 0.0; a steady-state reading of Table 6 implies about 0.0210, recorded and deliberately not used.
§3.3.2	Eq. (99) split parameters and profit rates	Two logistic parameters untabulated and two *variables* — the fossil and non-fossil capital profit rates — never defined anywhere in the manual. All four defaulted to 0.0, which forces a 50:50 green/fossil investment split against the 69:31 Table 6 implies. The largest hole in the section, since this split is what steers investment green.
§3.3.2	Eq. (138) credit-rationing slopes	Three of four slopes untabulated. The NFC equation has identical structure and its slopes *are* tabulated, and Table 5 reuses NFC values for the power sector in at least nine other places — but reusing them here would be our inference, not the manual's instruction, and it does not reproduce Table 6 anyway. Defaulted to 0.0.
§3.3.2	government non-fossil capital	Eq. (86) puts it in a denominator; Table 6 does not tabulate it. Measured to be zero in the baseline twice over, via Eq. (86) and via §3.1's Eq. (5). It is a scenario instrument.
§3.3.2	forward-looking switches	Eqs. (88)–(89) introduce four quantities §5 never tabulates (ban period, credibility, planning horizon, hyperbolic discount); the manual gives only bounds. Treated as scenario switches with an inert baseline.
§3.3.2	expected capital path	*Our choice, not a manual defect.* Eq. (89) averages expected utilisation over the planning horizon, but the manual never says what capital stock the sector expects to hold, and the model carries no forecast of it. Real capital and electricity demand are held at current values; inert whenever credibility is zero.
§3.3.2	Eq. (88) inequality direction	The printed condition and the prose above it point opposite ways. The printed inequality is implemented — transcription beats interpretation — with the baseline switch set so the no-ban case returns current utilisation, which is what the manual's own annotation asserts.
§3.3.2	unprinted electricity-price rule	§5 calibrates a long-run electricity-price formation rule — a switch quarter and a long-run price — that §3 never prints. Nothing is implemented, because nothing is printed to implement. The most economical account of Eq. (84)'s factor of 3.04.
§3.3.2	five untabulated endogenous variables	Table 6 omits five variables the section determines. Three are recoverable from what Table 6 does print and are checked that way; one recovery is confirmed by Eq. (106) reproducing a tabulated lending figure to 2.7×10^{-5}.