Supporting Information

S1 Nadaraya-Watson kernel regression is the limit of SPSb bootstrapping

The Nadaraya-Watson kernel regression (NWKR) [40] is a non-parametric method for estimating the conditional expectation of a response variable Y given a predictor X. For a set of observations $\{(X_i, Y_i)\}_{i=1}^n$, NWKR is defined as

$$\hat{m}(x) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{b})Y_i}{\sum_{i=1}^n K(\frac{x-X_i}{b})}, \tag{S1}$$

where $K_h(\cdot)$ is a kernel function with bandwidth b that controls the degree of smoothing.

The Bootstrapped Selective Predictability Scheme (SPSb) works by sampling historical analogues weighted according to a Gaussian kernel centred at the target point x. That is, an analogue at position X_i is selected with probability

$$p(Y_i|x) = \mathcal{N}(x - X_i \,|\, 0, \sigma), \tag{S2}$$

where σ is the standard deviation of the Gaussian kernel. As the number of bootstrap iterations B increases, the empirical frequency of selecting each analogue converges (by the law of large numbers) to the kernel weight $K_h(x - X_i)$. In this limit, the bootstrap averaging used in SPSb yields an expected prediction that converges to the NWKR estimator [30]:

$$E_{\text{SPSb}}(Y|x) \to \hat{m}(x). \tag{S3}$$

Additionally, it has been proven [30] that the standard deviation of the NWKR estimate—which quantifies the uncertainty in the prediction—is given by

$$\sigma^2_{\text{NWKR}}(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i^2}{\sum_{i=1}^n K_h(x - X_i)} - \left[\hat{m}(x)\right]^2. \tag{S4}$$

This formula is directly analogous to the variance computed in SPSb; indeed, in the limit of large B, the uncertainty derived from SPSb converges to $\sigma_{\text{NWKR}}(x)$ (up to a scaling factor that accounts for the size N of the sample over which each bootstrap average is computed).

The convergence of SPSb to NWKR demonstrates that the bootstrapped procedure not only produces a reliable point estimate but also a robust measure of uncertainty [30]. This equivalence justifies the use of either method in forecasting applications while highlighting the computational efficiency of the NWKR formulation [30].

S2 Further details about data preparation

This section provides additional details on the data preparation process, complementing the description in Section 2.2.

GDPpc. Following standard practice in the Economic Complexity literature, we use the `NY.GDP.PCAP.PP.KD` indicator from the World Development Indicators (WDI) dataset, which measures GDPpc per capita in constant international dollars.

Fitness. To avoid convergence issues highlighted in prior studies [41], we remove monopolies (products exported by only one country) from the M_{cp} matrices. Fitness is computed by running the Fitness-Complexity algorithm [6] for 50,000 iterations. The resulting values are transformed using $\log_{10}$.

Technological Fitness. Patents in the PATSTAT dataset are classified using an 8-digit hierarchical system, where the first three digits define broad categories and subsequent digits refine subcategories. We use the most granular 8-digit classification to define technologies. For each country, we construct country-technology matrices by counting patents with specific classification codes. Patents are assigned to countries based on inventors' addresses, with fractional counting applied for multi-country patents.

While we use the RCA-based quantization method from [3] for Fitness calculations, Technological Fitness requires a different approach due to data sparsity. We binarise matrices by setting entries to 1 if they are ≥ 1 and 0 otherwise. Technological Fitness is then calculated using the stable version of the Fitness-Complexity algorithm developed by Servedio et al. [42], which ensures convergence (see Section S4).

Technological Fitness can only be computed for countries with at least one patent in a given year. For less developed countries with intermittent patent activity, this results in alternating finite and undefined values over time. To address this noise, we apply linear interpolation to fill undefined values, creating a complete time series.

Polity. We manually reconciled country names in the Polity V dataset with ISO 3166-1 alpha-3 codes used in WDI and COMTRADE datasets. The `POLITY2` indicator is used as our institutional metric. It adjusts raw scores by treating foreign interruptions (-66) as missing, converting interregnum/anarchy (-77) to a neutral score of 0, and prorating transitions (-88) across transition periods. This standardisation ensures a continuous representation of regime changes suitable for time-series analysis.

S3 Data cleaning

This section outlines the data cleaning procedures applied to ensure consistency and reliability in the dataset.

Removal of Analogues with Low Fitness. Analogues (see Section 2.1 for the definition) with a Fitness value below -6 were excluded from the dataset.

Handling Missing Metrics. For Fitness, GDPpc, and Polity, any analogue missing a single metric was removed entirely from all models and evaluations. This approach did not significantly skew the dataset's composition toward specific types of countries. However, Technological Fitness presented unique challenges due to its tendency to fail at converging for countries with minimal patent activity, as shown in Figure S1.

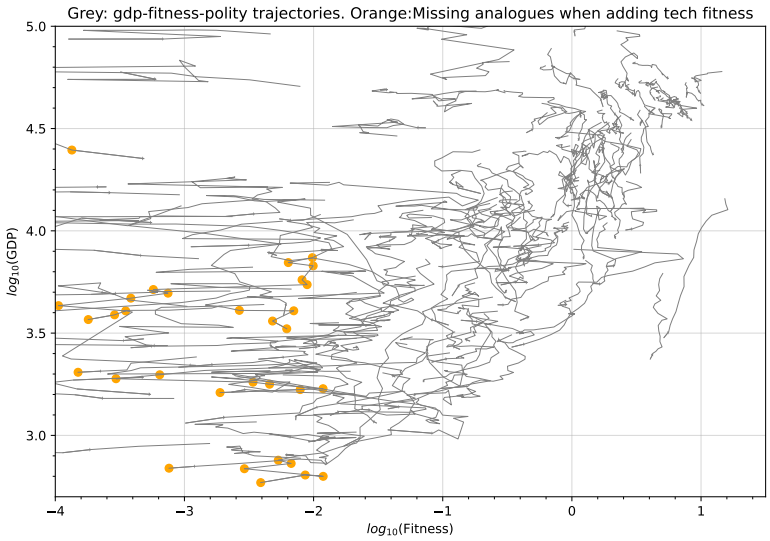


Figure S1: Analogues removed from the dataset due to missing Technological Fitness values. The plot shows that the distribution of missing values is heavily skewed towards the bottom-left part of the plane, where Technological Fitness performs worse. Not all are shown; additional analogues exist in regions with even lower Fitness values.

The strategy of removing all 43 analogues where Technological Fitness failed to converge disproportionately excluded countries with low patent activity—precisely those where Technological Fitness performs poorly. This artificially inflated the average predictive performance of models using Technological Fitness at the expense of other metrics.

To address this issue, we adopted an alternative strategy: assigning the lowest observed Technological Fitness value in the dataset to any missing Technological Fitness value at the beginning of a country's time series. This adjustment ensures that countries with no patent activity are appropriately represented with low Technological Fitness values. It also balances the dataset for backtesting, enabling richer phenomenological insights.

By applying these data cleaning procedures, we ensured that our analysis reflects a balanced and comprehensive view of country dynamics while mitigating biases introduced by missing or unreliable data.

S4 Fitness computation

Fitness quantifies a country's capabilities, particularly its industrial and manufacturing competitiveness, by analysing international trade patterns. The underlying intuition is that export competitiveness reflects a country's productive capabilities [7]. Developed countries tend to export a diverse range of products, signifying high capability levels, while developing countries export fewer products, reflecting more limited capabilities [1].

However, the exports of developed countries are less informative for assessing product complexity since they encompass almost all product types. In contrast, when a developing country exports a product, it signals that the product is less complex. Product complexity is thus inversely related to the number of countries that can export it. These observations form the basis for computing Fitness through an iterative algorithm applied to export data.

Fitness' definition can be understood directly as a consequence of observations about the world export market. Consider the binary matrix M_{cp}, where $M_{cp} = 1$ indicates that country c exports a significant quantity of product p, and $M_{cp} = 0$ otherwise. This matrix defines the bipartite graph connecting countries to their exports. The discovery that the M_{cp} matrix can consistently be rearranged to have a triangular shape [1], or equivalently that the country-product graph is nested, is itself a consequence of these observations and led to the definition of the Fitness and Complexity algorithm.

Let us first define how to compute M_{cp}. To determine the significance of product p exports for country c, we introduce the *export matrix* EXM_{cp}, representing the dollar value of product p exported by country c in a given year. Following standard practices in the literature, we compute the *Revealed Comparative Advantage* (RCA), also known as the Balassa index [43], defined as:

$$RCA_{cp} = \frac{\frac{EXM_{cp}}{\sum_j EXM_{cj}}}{\frac{\sum_i EXM_{ip}}{\sum_{kl} EXM_{kl}}}. \tag{S5}$$

This index represents the ratio of product p's exports by country c to the total exports of c, divided by the same ratio calculated for the entire world. Conven-

tionally, thresholding this matrix at a specific level (typically 1) yields M_{cp}:

$$M_{cp} = \begin{cases} 1 & \text{if} \quad RCA_{cp} \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{S6}$$

In our case, we derive M_{cp} by applying a Hidden Markov Model (HMM) to quantise the RCA_{cp} entries over time, following a regularisation procedure shown to improve results in [3].

The Fitness (F_c) and Complexity (C_p) metrics are computed using an iterative algorithm applied to M_{cp} until convergence. The equations governing this iterative process start from:

$$F_c^{(0)} = 1, \quad C_p^{(0)} = 1, \quad \forall c, p. \tag{S7}$$

The update rules for each iteration are:

$$\tilde{F}_c^{(n)} = \sum_p M_{cp} C_p^{(n-1)}, \tag{S8}$$

$$\tilde{C}_p^{(n)} = \frac{1}{\sum_c M_{cp} \frac{1}{F_c^{(n-1)}}}, \tag{S9}$$

followed by normalization:

$$F_c^{(n)} = \frac{\tilde{F}_c^{(n)}}{\langle \tilde{F}_c^{(n)} \rangle_c}, \quad C_p^{(n)} = \frac{\tilde{C}_p^{(n)}}{\langle \tilde{C}_p^{(n)} \rangle_p}. \tag{S10}$$

The Fitness of a country is proportional to the sum of Complexities of its exported products. Conversely, a product's Complexity is constrained by being lower than the smallest Fitness among its exporters. Additionally, as more countries export a given product, its Complexity decreases. These properties ensure that Fitness captures both diversification and specialization in global trade. Fitness has been shown to be highly informative in characterizing economic development trajectories across countries.

S4.1 Servedio's method

Used in cases where data is noisy or sparse (e.g., Technological Fitness derived from patent data), an alternative version of the algorithm ensures convergence by modifying Equation (S9)[42]:

$$F_c^{(n)} = \phi_c + \sum_p M_{cp} C_p^{(n-1)}, \qquad C_p^{(n)} = \pi_p + \frac{1}{\sum_c M_{cp} \frac{1}{F_c^{(n-1)}}} \tag{S11}$$

$$\tag{S12}$$

and Eq.S10 is skipped. In this work, we set $\phi_c = 10^{-6} \quad \forall c$ and $\pi_p = 1 \quad \forall p$.

S5 Polity definition details

The `POLITY2` index is calculated as the difference between `DEMOCRACY` and `AUTOCRACY` scores of the Polity V dataset [21], each rated on 11-point scales (0-10). We give an overview of the dataset's content and metric's calculation since neither have previously been used in EC forecasting models. The `DEMOCRACY` score considers four weighted indicators: "Competitiveness of political participation", "Openness of executive recruitment", "Competitiveness of executive recruitment", and "Constraints on the executive". `AUTOCRACY` adds "Regulation of political participation" to these four. As an example of indicator coding, "Competitiveness of political participation" uses a 5-category scale:

- Not Applicable
- Repressed (no political participation allowed)
- Suppressed (restricted participation outside government)
- Factional (ethnic-based factions compete)
- Transitional
- Competitive (stable political groups compete nationally)

The coding evaluates how openly citizens can express political preferences, considering factors like policy/leadership alternatives, restrictions on political groups, group stability, and power transfer patterns. See the extensive Polity V Dataset Users' Manual [21] for the complete methodology.

S6 Supporting information for Section 2.6: Timescale analysed

Please note that in the comparisons of this section and elsewhere we include a baseline model, the *autocorrelation baseline*, which naively predicts future GDPpc growth as equal to the last observed growth rate over the same interval. For instance, if U.S. GDPpc grew by 1% from 2010–2012, this model forecasts 1% growth for 2012–2014; similarly, a 1.5% growth from 2010–2015 predicts 1.5% for 2015–2030.

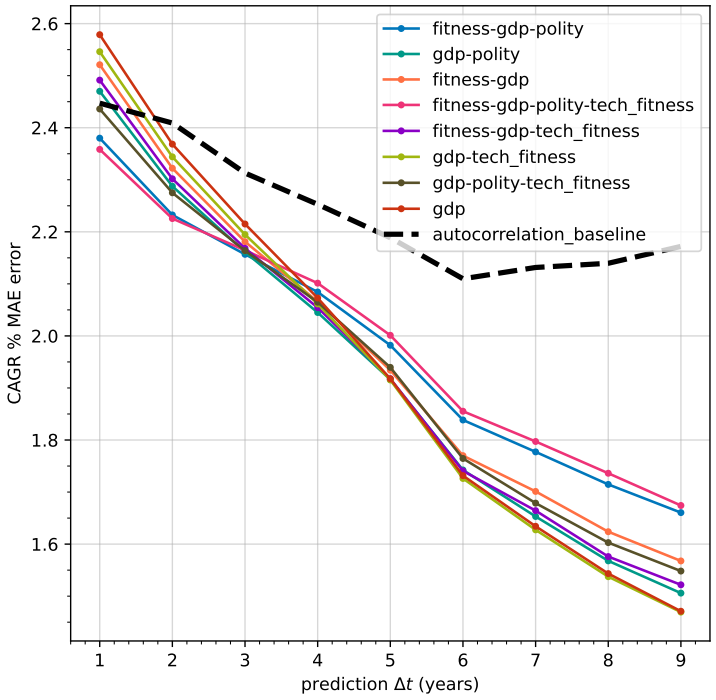


Figure S2: Average CAGR MAE % error for all of the possible models at all possible Δt's. As explained in Section 2.6, at short timescales ($\Delta t = 1-2$ years) the behaviour of GDPpc growth becomes quite unpredictable, due to high-frequency noise. At longer timescales, instead, the growth behaviour becomes increasingly predictable, with average prediction error decreasing for any SPSb model, and that many differences among countries tend to be discounted into the GDPpc metric. This results in other metrics such as Fitness losing their comparative predictive power compared to GDPpc — in other words, the best predictive performance of all models converges to that of the single-dimensional GDPpc-only SPSb model, with additional metrics potentially worsening results. Other effects appear in this plot that could open new lines of research. For $\delta t = 1$, the only models to perform better than the baseline are higher-dimensional and including Polity. However, at longer timescales these are the same models that perform worse - suggesting that the information they leverage is discounted in the GDPpc and Technological Fitness metrics for the purpose of predicting on long horizons.

S7 Supporting information for Section 2.7: Overall comparison of models

In this section we show that varying the metrics that we add to SPSb models, we do not observe significant overall changes in the models' predictive performance. This is the reason why we focus on differences in predictive performance on the basis of groups of countries.

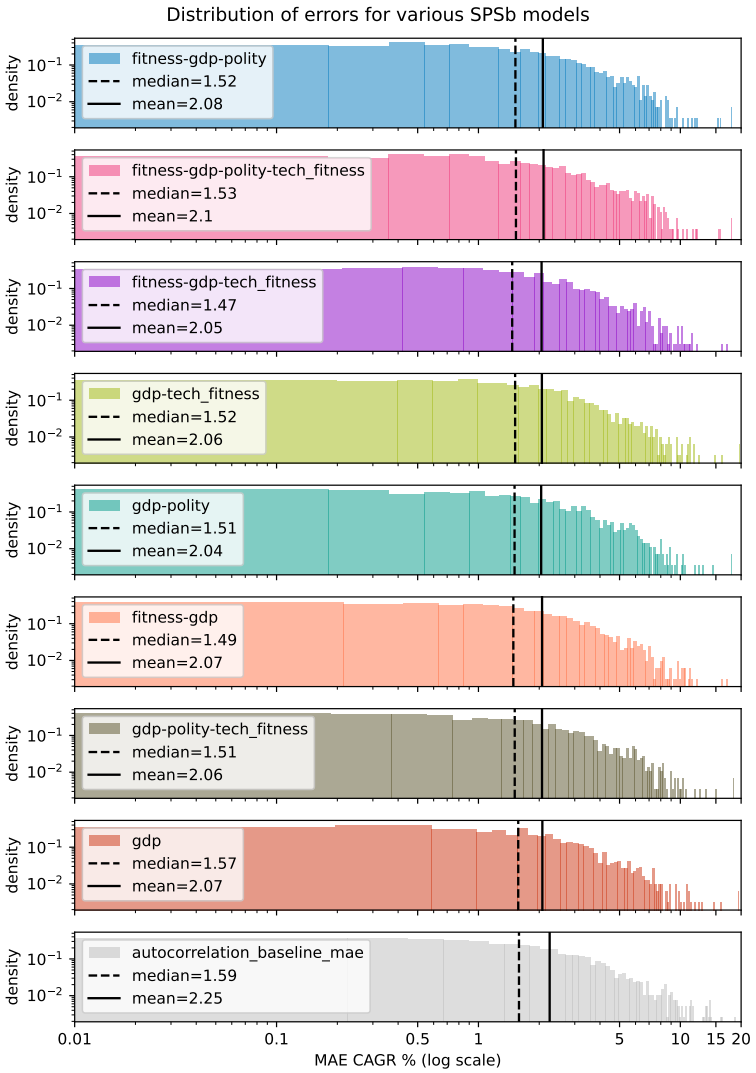


Figure S3: Global distribution of CAGR MAE % errors for all of the considered models at $\Delta t = 4$. We only consider the last 5 available time periods. We perform two-sided Kolmogorov-Smirnov tests for each available pair of models to test that the overall empirical distribution is the same. We find that the lowest p-value obtained is 0.5, clearly showing that there are no significant differences in the distribution of errors among the different models. This result is mentioned in Section 2.7. Please note that we define the autocorrelation baseline in Section S6.

	fitness gdp polity	fitness gdp polity techfit	fitness gdp tech fitness	gdp techfit	gdp polity	fitness gdp	gdp polity techfit	gdp
fitness gdp polity		0.89	0.75	0.85	0.89	0.99	0.89	0.73
fitness gdp polity techfit	0.89		0.80	0.85	0.87	0.67	0.89	0.73
fitness gdp techfit	0.75	0.80		0.93	0.85	0.99	0.78	0.51
gdp techfit	0.85	0.85	0.93		0.93	1.00	0.64	0.70
gdp polity	0.89	0.87	0.85	0.93		0.99	0.94	0.85
fitness gdp	0.99	0.67	0.99	1.00	0.99		0.89	0.67
gdp polity techfit	0.89	0.89	0.78	0.64	0.94	0.89		0.93
gdp	0.73	0.73	0.51	0.70	0.85	0.67	0.93	

Table S1: P-values of Kolmogorov-Smirnov test for all possible pairs of models. No p-value is below 0.5, indicating that the differences among distributions are not statistically significant. This result is mentioned in Section 2.7.

S8 GDPpc-Polity plane

To better understand the distribution of countries for which Polity improves predictability, we analyse the GDPpc-Polity plane. This representation compares the prediction error of the GDPpc-Polity model with that of the Fitness-GDPpc model. The methodology follows that of Figure 2, with one key difference: instead of plotting trajectories, we represent available analogues as points for improved readability. Additionally, since Polity is a quantised metric, we apply Gaussian jitter with $\mu = 0$ and $\sigma = 0.14$ to the Polity values to better visualise density. The results are shown in Figure S4.

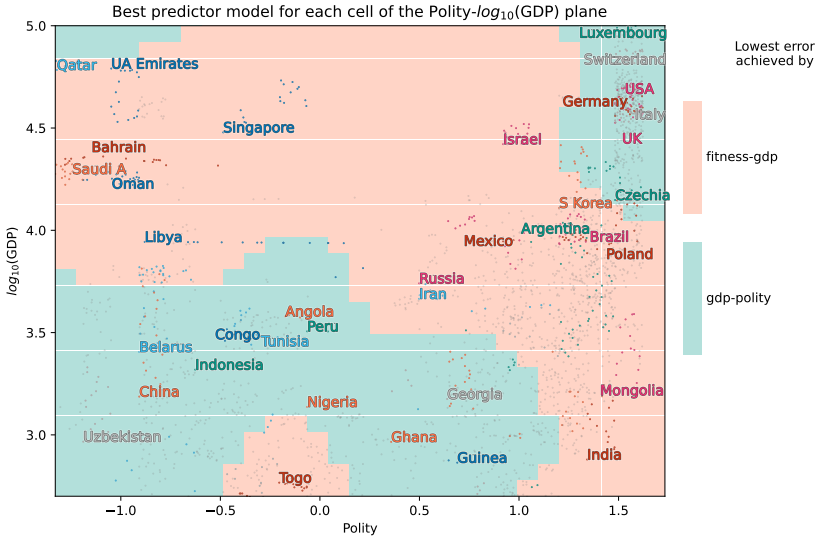


Figure S4: Comparison of the GDPpc-Polity and Fitness-GDPpc models on the GDPpc-Polity plane at $\Delta t = 4$. Analogues are plotted as points with jitter added to Polity values to highlight density. The Polity model performs best for low-GDPpc and less-than-highest Polity countries but does not resolve well for mature democracies clustered on the far right of the plane. Results are mentioned in Section 3.1.

Key observations are:

- Countries appear to follow a "U"-shaped distribution, with most high-GDPpc countries having either very high or very low Polity values. This pattern highlights the clustering of mature democracies at maximum Polity scores and fossil-fuel-dependent economies at low scores (with the notable exception of Singapore).

- The GDPpc-Polity model performs best for countries with low GDPpc

(below 10^4, or $10,000$/year per capita) and intermediate Polity values (below 1.0). These are typically developing countries that have not fully democratised.

- The model struggles with high-GDPpc, low-Polity countries, primarily fossil-fuel-dependent economies in the Middle East (e.g., Qatar, Kuwait, UAE), which are situated in the "chaotic" region of the Fitness-GDPpc plane.

- High-GDPpc, high-Polity countries (e.g., mature democracies) cluster in the top-right corner of the plane. While these countries appear better predicted by the Polity model in this representation, this is not a strong signal; results from other analyses (Figure 2) indicate that Fitness-GDPpcperforms comparably for these countries. Both models face limitations due to saturation effects in their respective metrics.

The results suggest that Polity's primary contribution lies in improving predictions for low-income, less-democratic countries. For high-income countries, particularly those with saturated Polity scores, neither model achieves significant differentiation or predictive advantage. This limitation may reflect sparsity in analogues or a transition where institutional quality (as measured by Polity) becomes less relevant above certain income thresholds.
The sparsity of high-income, low-Polity countries in the dataset complicates robust conclusions about this group. The upper-left quadrant of Figure S4 contains relatively few countries (e.g., UAE, Qatar, Kuwait), making it difficult to distinguish between genuine trends and random artifacts.

S9 Supporting information for Section 3.1: The role of institutions in predicting economic growth

We present here alternative ways to analyse Polity's effect on predictions. Each introduces a confounding factor, and is therefore a weaker result than the comparison presented in the main body of the paper. All nonetheless contribute to confirming the solidity of the result that Polity is beneficial for "poverty trap" countries, and not conclusively beneficial for highly developed countries.

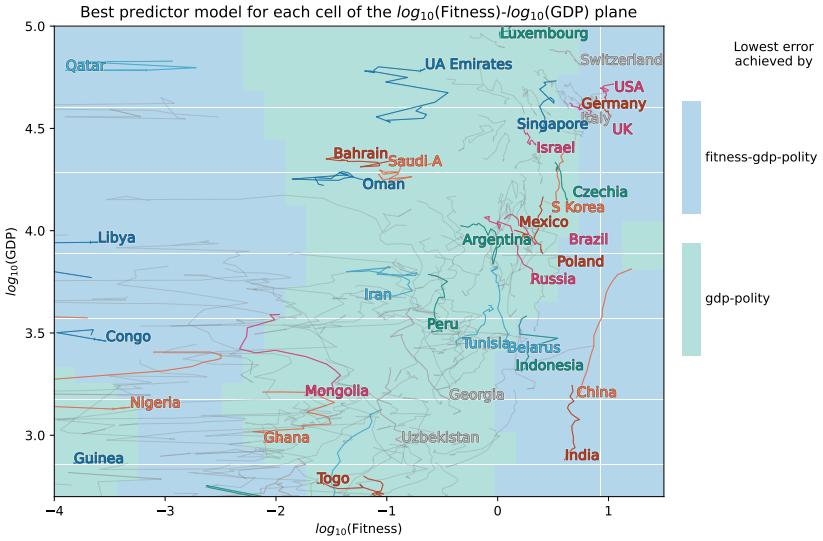


Figure S5: Comparing the 2-dimensional Fitness-GDPpc model with the 3-dimensional Fitness-GDPpc-Polity. We find that, similarly to Figure 2, the Polity model shows a demarcation line around Fitness > 0.5 that separates regions where the Polity model performs better versus worse. The reason for this is discussed in Section S8, where we show that most of the countries where Polity does worse share the identical level of Polity, which limits the SPSb model's capacity of finding good analogues. We also see that the Polity-only model does better for poverty trap countries.

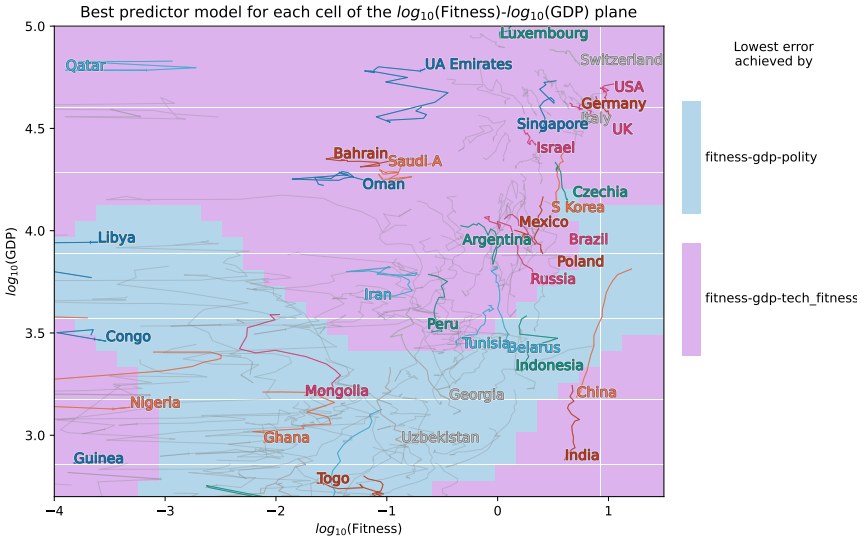


Figure S6: Comparing the 3-dimensional Fitness-GDPpc-Technological Fitness model with Fitness-GDPpc-Polity. We find that, similarly to Figure 2, the Polity-containing model performs best on countries in the "poverty trap" area. The model containing Technological Fitness instead is the best predictor on the developed countries. We posit that this reflects the predictive contribution of Technological Fitness, the best-performing metric for this group of countries as remarked in Section 3.2. The findings are extremely similar to those of Figure S7.

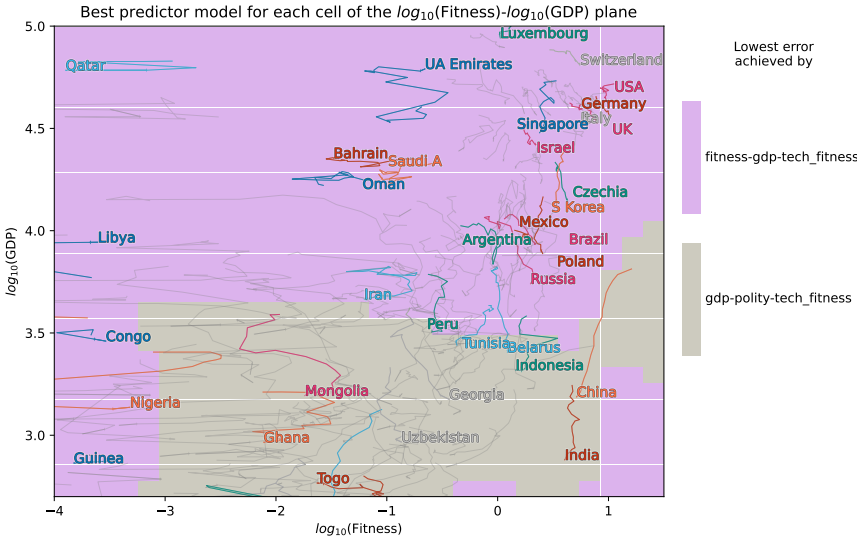


Figure S7: Comparing the 3-dimensional Fitness-GDPpc-Technological Fitness model with GDPpc-Polity-Fitness. The findings are extremely similar to those of Figure S6 regarding both "poverty trap" and highly developed countries. However, we note that both China and India are best predicted by the model containing Polity. This is because the combination of these two metrics is one of the best for outlier countries under $\$3,000$ GDPpc, as discussed in Section 3.3, Section S14, and Section S12.

S10 Supporting information for Section 3.2: Comparative analysis of predictive metrics

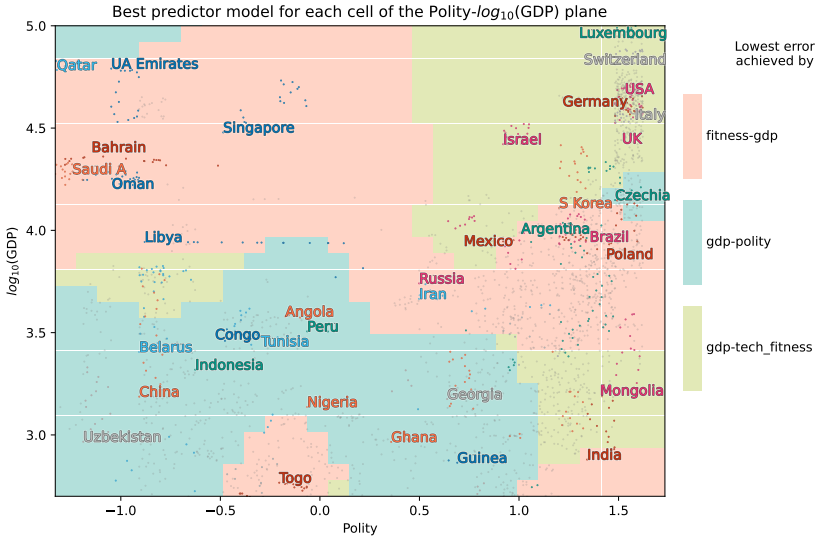


Figure S8: Comparison of the GDPpc-Polity, GDPpc-Technological Fitness, and Fitness-GDPpc models, on the GDPpc-Polity plane, at $\Delta t = 4$. The treatments are identical to those of Figure S4. Results are commented in Section 3.2.

S11 Contributions of metrics to higher-dimensional models

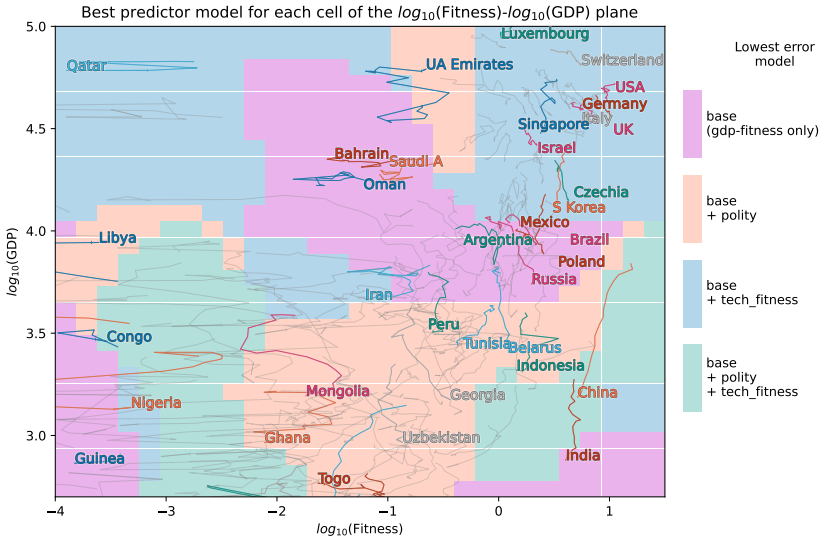


Figure S9: Comparison of all possible 2-, 3-, and 4-dimensional models constructed using the Fitness, GDPpc, Polity, and Technological Fitness metrics (listed in Table 1) on the Fitness-GDPpc plane at $\Delta t = 4$. Cells are coloured based on the presence or absence of Polity and Technological Fitness in the model with the lowest prediction error for that cell. The treatments are otherwise identical to those in Figure 2. In the bottom half of the figure (off the development diagonal), 4-dimensional models are the best forecasters. For $\log_{10}(GDPpc) \geq 4$, using Polity appears to have no positive effect on forecasts. Results are mentioned in Section 3.3

Having shown in Section 3.3 that higher-dimensional models are useful in certain cases, we investigate whether regularities exist in how different combinations of metrics benefit specific classes of countries.

Similar to Figure 4, we compare all possible 2-, 3-, and 4-dimensional models (listed in Table 1). In Figure S9, cells are coloured based on which combination of metrics is included in the best predictive model for that cell. Specifically, we distinguish between models containing:

- only Polity but not Technological Fitness,
- only Technological Fitness but not Polity,
- both metrics, and

- neither metric (the Fitness-GDPpc model).

From previous results (Figure 4), most countries along the "development diagonal" are best predicted by a 2-dimensional model. Moving left to right along this diagonal in Figure S9, countries are sequentially best predicted by:

- a 2-dimensional GDPpc-Polity model,
- then a Fitness-GDPpc model,
- and finally a Technological Fitness-GDPpc model.

These cases can be excluded from our analysis of higher-dimensional models since they do not benefit from additional dimensions. Instead, we focus on off-diagonal countries.

Excluding a few "noise" patches (e.g., Guinea, Libya, UAE, Iran), a key regularity emerges for countries with $\log_{10}(GDPpc) > 4$ ($> 10,000$/year per capita): only Technological Fitness is useful when included in higher-dimensional models. For these countries in the upper half of the plane, the best high-dimensional model is consistently Fitness-GDPpc-Technological Fitness. Conversely, for countries with $\log_{10}(GDPpc) < 4$ (bottom half), the best-performing higher-dimensional model is exclusively the 4-dimensional Fitness-GDPpc-Polity-Technological Fitness.

These findings suggest that Polity is irrelevant for predictions in high-income countries ($\log_{10}(GDPpc) > 4$). This could indicate a transition where the kind of institutional quality measured by Polity becomes less relevant (suggesting that fewer predictive dimensions suffice in this regime) above a certain income threshold. However, it might also result from data sparsity. The upper quadrant of Figure S9 contains far fewer countries—primarily developed or developing clusters along the development diagonal—where simpler 2-dimensional models already suffice.

The full list of upper-left quadrant countries best predicted by higher-dimensional models includes UAE, Qatar, Kuwait, Equatorial Guinea, Libya, Gabon, and Suriname. This low density raises concerns about random chance influencing results. A similar analysis on the GDPpc-Polity plane (Figure S10) confirms that Polity remains almost irrelevant for $\log_{10}(GDPpc) > 4$ countries but provides no additional insights due to limited data availability.

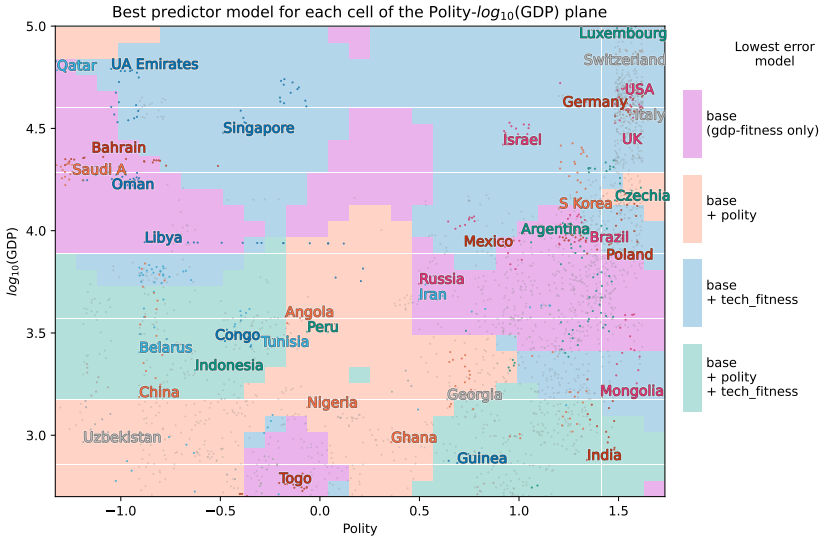


Figure S10: Comparison of all possible 2, 3, and 4-dimensional models that can be constructed using the Fitness, GDPpc, Polity, Technological Fitness metrics on the GDPpc-Polity plane, at $\Delta t = 4$. We colour each cell according to the presence or absence of the Polity and Technological Fitness metrics in the model having the lowest prediction error in that cell. The treatments are otherwise identical to those of Figure 2. Results are mentioned in Section 3.3.

S12 Interpretable outliers

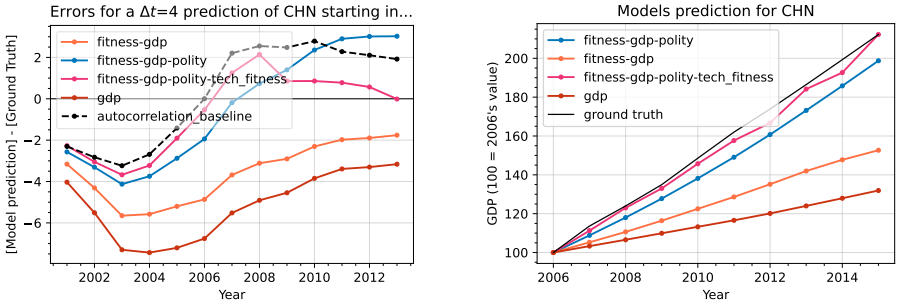
We analyse China and India in greater depth, as they are best predicted by higher-dimensional models. Their outlier status makes it easier to disentangle and interpret the contributions of additional dimensions. While these two countries are close to each other in the Fitness-GDPpc space and distant from other countries, they differ significantly along the Polity dimension (approximately -0.9 for China and 1.3 for India, as shown in Figure S4). Under these conditions, both countries primarily use past versions of themselves as analogues.

China is best predicted by the 4-dimensional model. As shown in Figures S11a and S11b, starting with a one-dimensional GDPpc-only model and incrementally adding metrics where China is an outlier progressively separates it from other countries, making predictions increasingly resemble a weighted sum of its past performance. To highlight this effect, we include the autocorrelation baseline model (defined in Section S6) in Figure S11a.

During China's period of high growth (2002–2007), the model's error decreases as more high-growth analogues from its past are included. After 2007, as China's growth slows, the model's error increases because it relies on ana-

logues from its earlier high-growth phase. Notably, the 4-dimensional model is the only one that corrects the autocorrelation baseline's error during the 2008–2012 period.

India, on the other hand, is best predicted by the 3-dimensional GDPpc-Polity-Technological Fitness model, which narrowly outperforms the 4-dimensional model. As shown in Figures S12a and S12b, Fitness and Technological Fitness tend to pull India closer to high-growth China, overestimating its growth. However, India's relatively high Polity value compared to China draws it closer to slower-growing developed countries. The combined effect of Polity's underestimation tendency and Fitness and Technological Fitness's overestimation tendency achieves a balanced prediction with the lowest error.



(a) Forecast errors for a $\Delta t = 4$ prediction on China's GDPpc growth as a function of the start year.

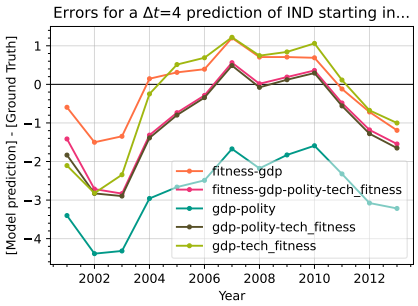
(b) Full series of predictions of GDPpc growth at different Δt for China, comparing different models, normalised to 2006's GDPpc.

Figure S11: Forecast errors as a function of time for China. Colours code the performance of models including different sets of metrics. Adding one more metric to the model can improve predictions for specific country groups. The results are discussed in more detail in Section S12.

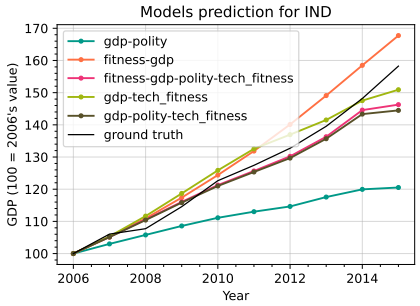
In summary, China benefits most from a 4-dimensional model that incorporates all available metrics, while India is best predicted by a 3-dimensional model that balances competing tendencies among metrics (Polity, Fitness, and Technological Fitness). These cases highlight how higher-dimensional models can improve predictions by leveraging country-specific variation across metrics.

S13 Statistical validation

To assess whether observed differences in model performance across the fitness-GDP plane are statistically significant, we employ a bootstrap-based hypothesis testing procedure [44] combined with False Discovery Rate (FDR) control for multiple comparisons [45].



(a) Forecast errors for a $\Delta t = 4$ prediction on India's GDPpc growth as a function of the start year.



(b) Full series of predictions of GDPpc growth at different Δt for India, comparing different models, normalised to 2006's GDPpc.

Figure S12: Forecast errors as a function of time for India. Colours code the performance of models including different sets of metrics. The combined effects of Fitness and Technological Fitness's overestimation and Polity's underestimation tendencies result in optimal predictions when all three metrics are included. The results are discussed in more detail in Section S12.

S13.1 Spatial Model Selection via Nadaraya-Watson Kernel Regression

For each candidate model $m \in \mathcal{M}$ and each location $\mathbf{x} = (\text{fitness}, \text{GDP})$ in the plane, we leverage the Nadaraya-Watson kernel regression technique to compute the expected prediction error. Let $\{(\mathbf{x}_i, e_{i,m})\}_{i=1}^{n}$ denote the observed data, where $e_{i,m}$ is the prediction error of model m for country i at location $\mathbf{x}_i$. The smoothed error surface is estimated as:

$$\hat{\mu}_m(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{x}_i) e_{i,m}}{\sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{x}_i)}$$

where $K_h(\cdot)$ is a Gaussian kernel with bandwidth h chosen via cross-validation. For each pixel in a discretized grid of the fitness-GDP plane, we identify the model with minimum estimated error: $m^*(\mathbf{x}) = \arg\min_{m \in \mathcal{M}} \hat{\mu}_m(\mathbf{x})$.

By means of this procedure we identify the model with the lowest error in each of the cells of the various figures in this work. We describe below a procedure to test that the model with the lowest error in a given cell has an error difference with the second-best model in that cell that is statistically significant.

S13.2 Bootstrap Hypothesis Testing

To test whether the observed best model $m^*(\mathbf{x})$ at each location $\mathbf{x}$ significantly outperforms the second-best model $m^{(2)}(\mathbf{x})$, we employ a pairs bootstrap procedure [44]:

1. **Bootstrap sampling:** Generate $B = 500$ bootstrap samples by resampling n pairs $\{(\mathbf{x}_i, e_{i,m})\}$ with replacement from the original data.

2. **Model refitting:** For each bootstrap sample $b = 1, \ldots, B$, refit all models using Nadaraya-Watson regression to obtain bootstrap error surfaces $\hat{\mu}_m^{(b)}(\mathbf{x})$ for each model $m \in \mathcal{M}$.

3. **Pairwise comparison:** At each pixel location $\mathbf{x}$, extract the bootstrap prediction errors for the original best and second-best models:

$$\Delta^{(b)}(\mathbf{x}) = \hat{\mu}^{(b)}_{m^*(\mathbf{x})}(\mathbf{x}) - \hat{\mu}^{(b)}_{m^{(2)}(\mathbf{x})}(\mathbf{x})$$

4. **P-value computation:** Compute a two-sided bootstrap p-value as:

$$p(\mathbf{x}) = 2 \cdot \min\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\Delta^{(b)}(\mathbf{x}) < 0\}, \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\Delta^{(b)}(\mathbf{x}) > 0\}\right)$$

   This measures the proportion of bootstrap replicates in which the ordering of the two models reverses.

S13.3 Multiple Testing Correction

Since we perform hypothesis tests at N spatial locations simultaneously, we control the False Discovery Rate using the Benjamini-Hochberg (BH) procedure [45]. This method is preferred over Bonferroni correction for two reasons: (1) it provides greater statistical power while controlling the expected proportion of false discoveries, and (2) it remains valid under spatial dependence between tests, which is inherent in our kernel-smoothed estimates.

Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ denote the ordered p-values across all pixels. The BH procedure finds the largest index k such that:

$$p_{(k)} \leq \frac{k}{N} \cdot \alpha$$

where $\alpha = 0.05$ is the target FDR level. All hypotheses with $p_{(i)} \leq p_{(k)}$ are rejected, indicating significant differences between the best and second-best models at those locations.

S13.4 Testing results

We report the results for Figures 2,3,4. We omit reporting results for the figures in the Supporting Information, which give qualitatively similar results. The discussion of each figure is given in the captions.

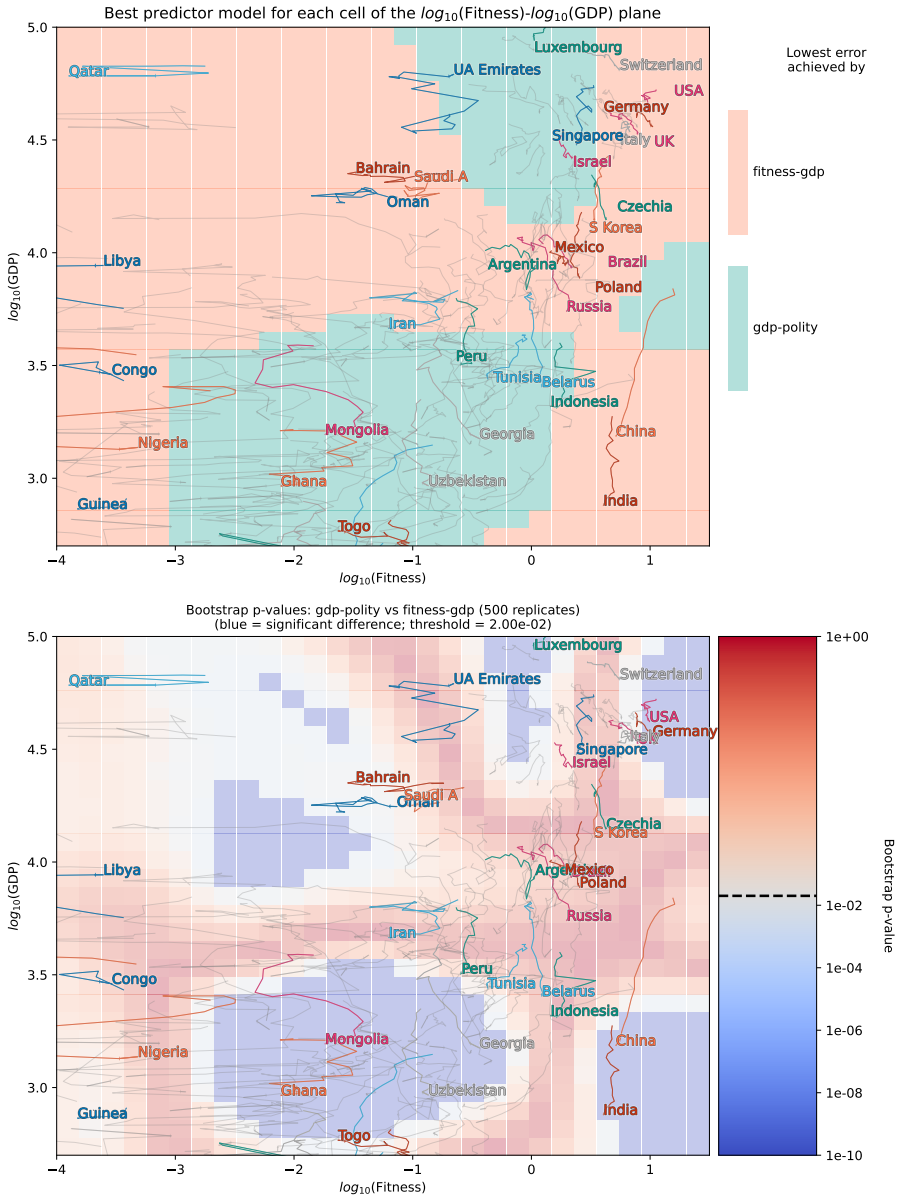


Figure S13: Statistical significance results for Figure 2. The procedure for computing these results is described in Section S13. Blue colour indicates the cells where the results are significant. The main result is the high significance of the countries in the poverty trap cluster, centered on Mongolia. As already highlighted and discussed in Section 3.1, the results in developed cluster are ambiguous because this cluster is split in two. Statistical significance analysis reveals that the results are not valid for all the countries in this cluster, and reconfirms our finding that Polity has inconclusive effects in this area of the plane.

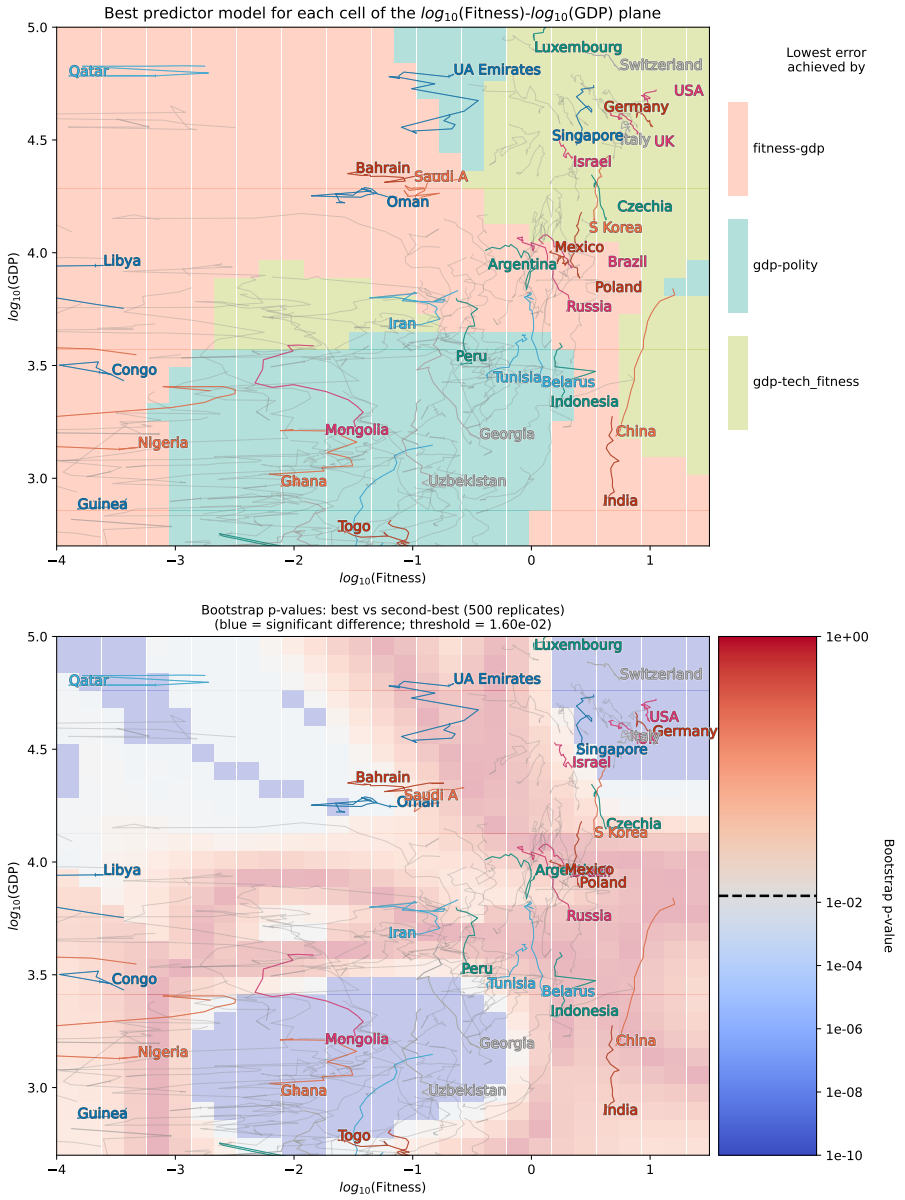


Figure S14: Statistical significance results for Figure 3. The procedure for computing these results is described in Section S13. Blue colour indicates the cells where the results are significant. Like in Figure S13, we reconfirm the high significance of the countries in the poverty trap cluster, centered on Mongolia. Compared to the same figure, the situation of the developed cluster changes, showing now strong signal for the high significance of the Tech Fitness metric in this area. The cells in the developing cluster are not significant. This could be due to the fact that it is a transition area. Still, the fact that in the developing cluster we find Fitness consistently positioned as the best model across a high number of contiguous cells, each containing a high density of countries provides collective evidence that complements cell-level testing. China and India are not significant, for reasons discussed in the caption of Figure S15.

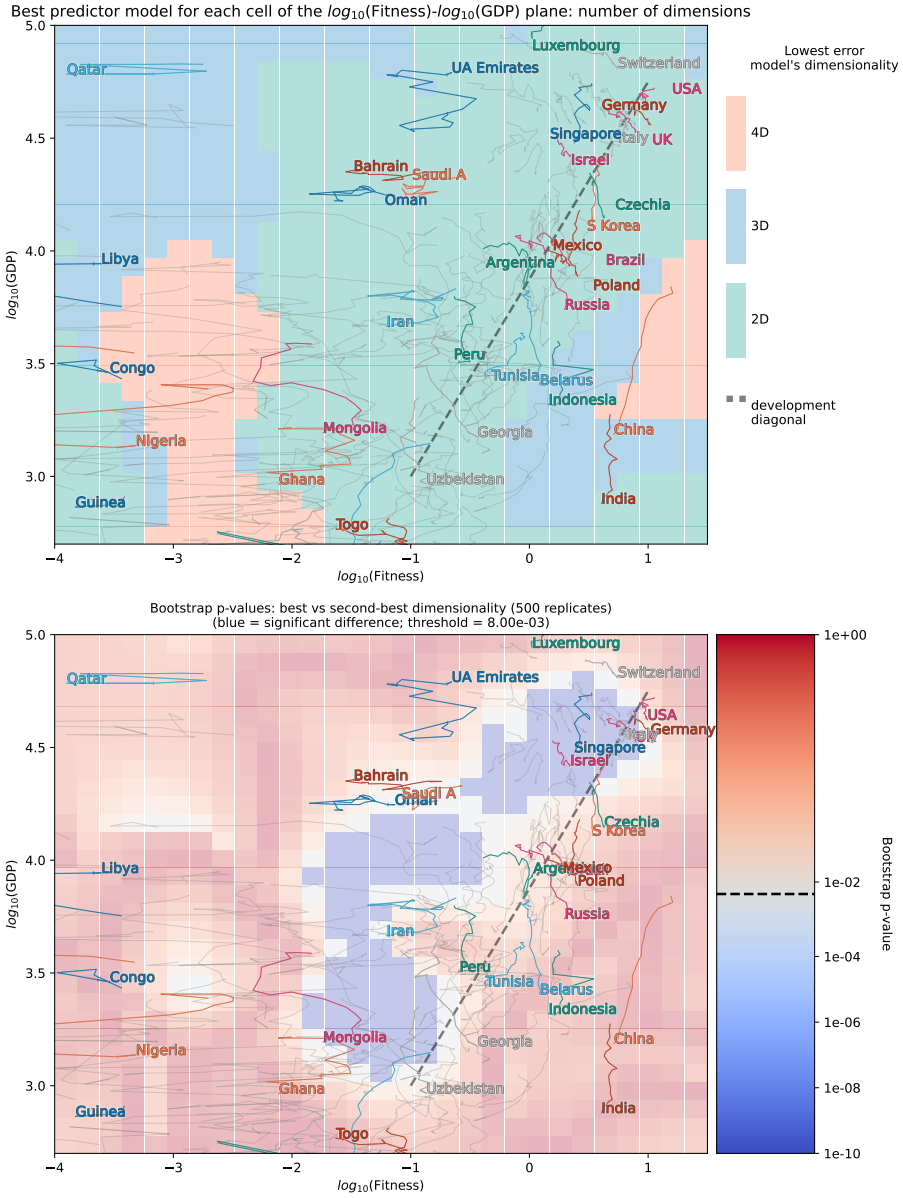


Figure S15: Statistical significance results for Figure 4. The procedure for computing these results is described in Section S13. Blue colour indicates the cells where the results are significant. Here, we find that for most of the off-diagonal countries the results are not significant on a cell-by-cell basis, with the exception of one area around the end of Nigeria's trajectory and Venezuela (with name not printed and gray trajectory: it is the country above Libya). For the countries to the left of the development diagonal, the lack of statistical significance cell by cell is weighed by the fact that large patches of several contiguous cells each containing relatively high density of countries all consistently show higher-dimensional models to have lower error. Results in showing significance for this area of the plane are in Figure S16. China and India, on the other hand, are not significant. This is explained by clarifying our methodology and Figures S12a and S11a. This figure shows the statistical significance of the best model in a cell with the second best (by dimensionality). In Figs.S12a,S11a we see that the 4-dimensional and 3-dimensional models are best and second-best and they are very close to each other. Therefore this metric is not useful for determining the significance of our findings that higher-dimensional models are best for these two countries.

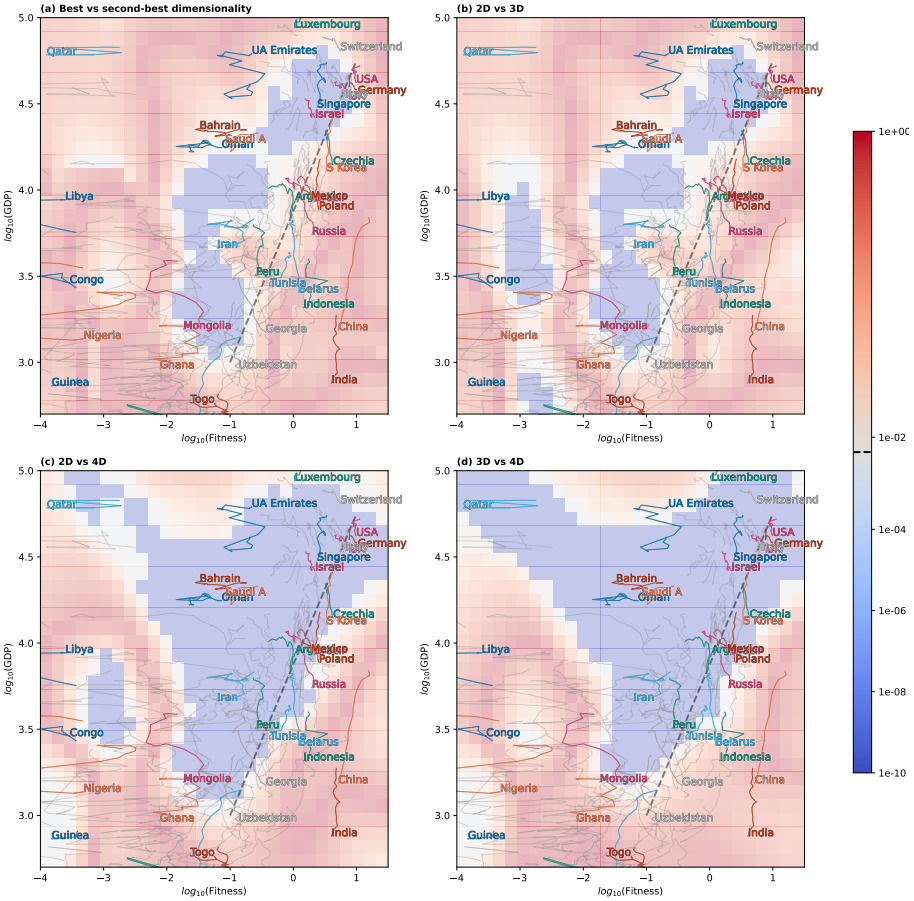


Figure S16: Further statistical significance results for Figure 4. The procedure for computing these results is described in Section S13. Blue colour indicates the cells where the results are significant. Panel (a) in this figure is the same as Figure S15. In the other panels, we compare only two dimensionalities at a time to gather more insight on whether the results of Figure 4 are significant. For the countries to the left of the development diagonal, the lack of statistical significance in panel (a) is explained by panels (b) and (c). Here, we can see clear indication that for many of these countries there is a significant difference between either the best 2d model and the best 3d one, or the best 2d model and the best 4d one. This implies that in these areas of the plane the best model and the second best model are respectively 3- and 4- dimensional, but they do not have significant differences between them. However, they both have significantly lower error than the best 2-dimensional model. China and India remain not significant. The argument that higher-dimensional models do better than 2-dimensional ones for most of the histories of both countries (as is evident from Figures S12a and S11a) still offers a signal that there is a significant difference for these countries as well. The reader can understand the situation more in-depth in Section S12.

S14 Supporting information for Section 3.3: Higher dimensions

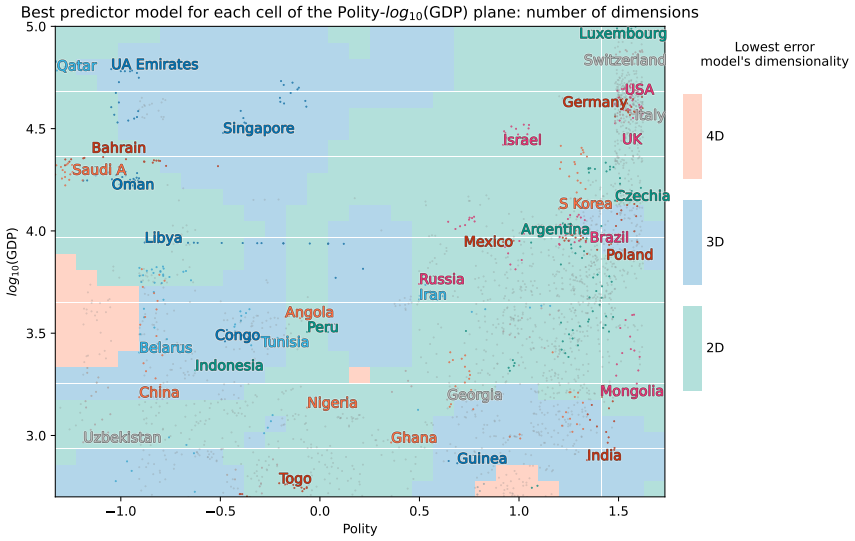


Figure S17: Comparison of all possible 2, 3, and 4-dimensional models that can be constructed using the Fitness, GDPpc, Polity, Technological Fitness metrics on the GDPpc-Polity plane, at $\Delta t = 4$. We colour each cell according to the number of dimensions of the model having the lowest prediction error in that cell. The treatments are otherwise identical to those of Figure S4. Results are commented in Section 3.3.

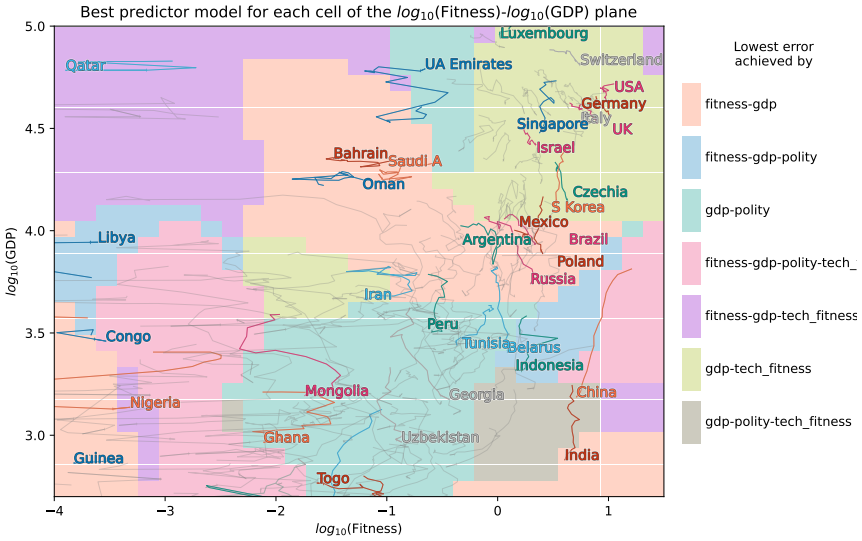


Figure S18: Comparison of all possible 2, 3, and 4-dimensional models that can be constructed using the Fitness, GDPpc, Polity, Technological Fitness metrics on the Fitness-GDPpc plane, at $\Delta t = 4$. We colour each cell according to the model having the lowest prediction error in that cell. The treatments are otherwise identical to those of Figure 2. This is a figure containing a large amount of information, which makes it hard to interpret. We advise the reader to familiarise themselves with the other plots and the meaning of different zones of the plane before attempting to read this figure.

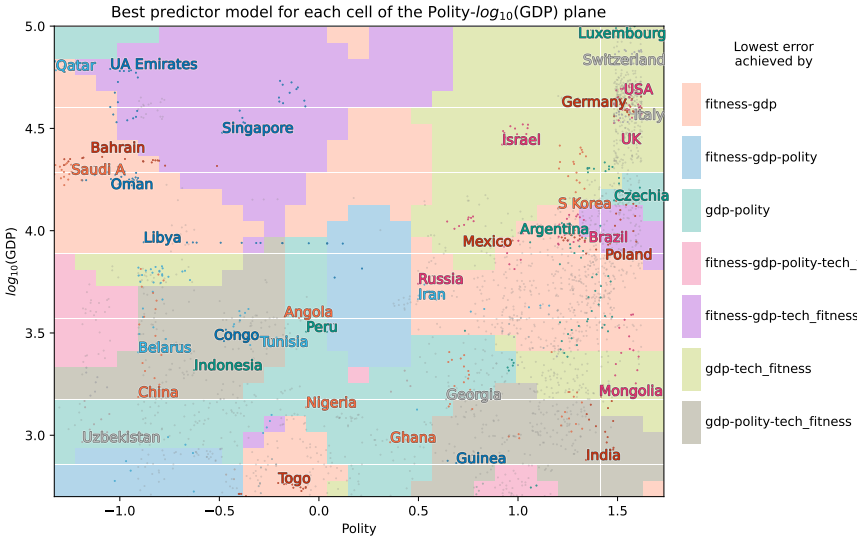


Figure S19: Comparison of all possible 2, 3, and 4-dimensional models that can be constructed using the Fitness, GDPpc, Polity, Technological Fitness metrics on the GDPpc-Polity plane, at $\Delta t = 4$. We colour each cell according to the model having the lowest prediction error in that cell. The treatments are otherwise identical to those of Figure 2. This is a figure containing a large amount of information, which makes it hard to interpret. We advise the reader to familiarise themselves with the other plots and the meaning of different zones of the plane before attempting to read this figure.

Table S4: Best GDPpc forecast model by country.

Country	Best model
Albania	`fitness gdp techfit`
Algeria	`fitness gdp polity`
Argentina	`gdp techfit`
Armenia	`gdp polity techfit`
Australia	`gdp polity techfit`
Austria	`gdp polity techfit`
Azerbaijan	`fitness gdp polity techfit`
Bahrain	`fitness gdp`
Bangladesh	`fitness gdp polity techfit`
Belarus	`fitness gdp`
Belgium	`fitness gdp`

Continued on next page

Table S4: Best GDPpc forecast model by country.

Country	Best model
Benin	`fitness gdp polity techfit`
Bolivia, Plurinational State of	`gdp techfit`
Brazil	`gdp techfit`
Bulgaria	`fitness gdp`
Burkina Faso	`fitness gdp techfit`
Cambodia	`gdp polity`
Cameroon	`gdp polity`
Canada	`gdp polity techfit`
Chile	`gdp polity`
China	`gdp polity techfit`
Colombia	`fitness gdp`
Costa Rica	`gdp polity`
Croatia	`gdp techfit`
Cuba	`gdp techfit`
Cyprus	`gdp techfit`
Czechia	`gdp techfit`
Côte d'Ivoire	`fitness gdp techfit`
Denmark	`gdp polity`
Dominican Republic	`fitness gdp techfit`
Ecuador	`fitness gdp`
Egypt	`gdp polity`
El Salvador	`gdp polity techfit`
Estonia	`fitness gdp`
Fiji	`gdp techfit`
Finland	`gdp polity`
France	`fitness gdp polity techfit`
Gambia	`fitness gdp techfit`
Georgia	`fitness gdp techfit`
Germany	`fitness gdp techfit`
Ghana	`gdp techfit`
Greece	`gdp polity`
Guatemala	`gdp polity techfit`
Guyana	`gdp techfit`
Haiti	`fitness gdp polity techfit`
Honduras	`gdp polity techfit`
Hungary	`gdp polity`
India	`fitness gdp`
Indonesia	`fitness gdp`
Iran, Islamic Republic of	`fitness gdp`
Ireland	`gdp polity`
Israel	`fitness gdp polity`

Continued on next page

Table S4: Best GDPpc forecast model by country.

Country	Best model
Italy	`fitness gdp polity techfit`
Jamaica	`fitness gdp polity techfit`
Japan	`fitness gdp polity techfit`
Jordan	`gdp polity`
Kazakhstan	`gdp polity`
Kenya	`gdp techfit`
Korea, Republic of	`fitness gdp`
Kyrgyzstan	`fitness gdp`
Lao People's Democratic Republic	`gdp polity`
Latvia	`fitness gdp`
Lithuania	`gdp polity techfit`
Madagascar	`fitness gdp techfit`
Malawi	`gdp techfit`
Malaysia	`fitness gdp polity`
Mali	`fitness gdp polity techfit`
Mauritius	`gdp polity`
Mexico	`fitness gdp polity techfit`
Moldova, Republic of	`fitness gdp`
Mongolia	`gdp polity`
Morocco	`fitness gdp polity techfit`
Myanmar	`fitness gdp polity techfit`
Nepal	`fitness gdp polity`
Netherlands	`gdp polity techfit`
New Zealand	`gdp polity`
Nicaragua	`gdp polity`
Niger	`fitness gdp polity techfit`
Nigeria	`gdp polity`
North Macedonia	`fitness gdp techfit`
Norway	`fitness gdp techfit`
Oman	`gdp polity`
Pakistan	`gdp techfit`
Panama	`fitness gdp polity techfit`
Paraguay	`gdp techfit`
Peru	`fitness gdp`
Philippines	`fitness gdp`
Poland	`fitness gdp techfit`
Portugal	`gdp polity`
Romania	`fitness gdp`
Russian Federation	`gdp techfit`
Saudi Arabia	`fitness gdp`
Senegal	`gdp polity techfit`

Continued on next page

Table S4: Best GDPpc forecast model by country.

Country	Best model
Sierra Leone	`fitness gdp techfit`
Singapore	`fitness gdp techfit`
Slovakia	`fitness gdp polity`
Slovenia	`gdp polity`
Spain	`gdp polity techfit`
Sri Lanka	`fitness gdp polity`
Sudan	`fitness gdp`
Sweden	`gdp polity techfit`
Switzerland	`fitness gdp polity techfit`
Tajikistan	`gdp techfit`
Tanzania, United Republic of	`gdp techfit`
Thailand	`gdp polity`
Togo	`fitness gdp`
Tunisia	`gdp polity`
Turkmenistan	`fitness gdp polity techfit`
Türkiye	`fitness gdp`
Uganda	`fitness gdp polity`
Ukraine	`gdp polity`
United Arab Emirates	`fitness gdp`
United Kingdom	`gdp polity techfit`
United States	`fitness gdp polity`
Uruguay	`gdp polity`
Uzbekistan	`gdp polity`
Venezuela, Bolivarian Republic of	`fitness gdp`
Viet Nam	`fitness gdp polity`
Zambia	`gdp polity`
Zimbabwe	`gdp polity`

S15 Sensitivity to Technological Fitness imputation strategy

The baseline imputation strategy for missing Technological Fitness values assigns the lowest observed Technological Fitness value in the dataset to any missing entry at the beginning of a country's time series (Section S3). Because this missingness is deterministic (countries lack patent data before a development threshold, rather than through random censoring) multiple imputation methods are inappropriate. Instead, we bracket the plausible range by comparing three strategies:

- **Baseline** (used throughout the paper): fill all missing Technological Fitness values with the global minimum observed value.

Region	Income	N_o	N_c	Best model
Latin America & Caribbean	All	299	23	`gdp techfit`
South Asia	All	78	6	`fitness gdp polity techfit`
Sub-Saharan Africa	All	429	41	`gdp polity techfit`
Europe & Central Asia	All	588	46	`gdp techfit`
Middle East & North Africa	All	204	17	`fitness gdp techfit`
East Asia & Pacific	All	221	17	`fitness gdp polity techfit`
North America	All	26	2	`gdp polity techfit`
All	High	546	42	`fitness gdp techfit`
All	Low	253	24	`fitness gdp polity techfit`
All	Lower middle	537	43	`gdp polity`
All	Upper middle	509	43	`gdp techfit`
Latin America & Caribbean	High	26	2	`gdp polity`
Europe & Central Asia	High	338	26	`gdp techfit`
Middle East & North Africa	High	91	7	`fitness gdp techfit`
East Asia & Pacific	High	65	5	`fitness gdp techfit`
North America	High	26	2	`gdp polity techfit`
Latin America & Caribbean	Low	13	1	`fitness gdp polity techfit`
South Asia	Low	13	1	`fitness gdp polity`
Sub-Saharan Africa	Low	227	22	`gdp polity techfit`
Latin America & Caribbean	Lower middle	65	5	`gdp polity techfit`
South Asia	Lower middle	65	5	`fitness gdp polity techfit`
Sub-Saharan Africa	Lower middle	148	13	`gdp polity`
Europe & Central Asia	Lower middle	91	7	`gdp polity`
Middle East & North Africa	Lower middle	64	5	`gdp polity`
East Asia & Pacific	Lower middle	104	8	`fitness gdp polity techfit`
Latin America & Caribbean	Upper middle	195	15	`fitness gdp techfit`
Sub-Saharan Africa	Upper middle	54	6	`gdp polity`
Europe & Central Asia	Upper middle	159	13	`fitness gdp polity`
Middle East & North Africa	Upper middle	49	5	`fitness gdp`
East Asia & Pacific	Upper middle	52	4	`gdp polity techfit`

Table S3: Best GDPpc forecast model by region and income level, based on the World Development Indicators regions categorisation. N_c is the number of countries that fall in the category, while N_o is the number of predictions (country-year pairs) available. The income thresholds are defined by the following GDPpc brackets: Low-income: $1,145 or less; Lower-middle-income: between $1,146 and $4,515; Upper-middle-income: between $4,516 and $14,005; High-income: above $14,005.

- **Drop**: remove all country-year pairs where Technological Fitness is undefined.

- **Country minimum**: for each country, backfill leading missing values with that country's earliest observed Technological Fitness value.

Only six countries are affected: Bhutan, Mozambique, Cape Verde, Togo, Laos, and Ethiopia, totalling 107 observations (0.76% of the dataset). These countries occupy the low-Fitness, low-GDPpc region of the Fitness-GDPpc plane, distant from the high-Fitness region where Technological Fitness-containing models show their distinctive advantage.

The country minimum strategy produces predictions indistinguishable from baseline: model rankings are identical (Spearman $\rho = 1.0$) and the maximum MAE difference across all 15 models is < 0.00001. This is because the per-country earliest observed Technological Fitness values are close to the global minimum.

The drop strategy, which removes the 107 observations, produces the largest perturbation. It disproportionately affects Polity-containing models (mean MAE increase +2.5%) compared to models without Polity (+0.8%), because the dropped countries are low-income states where institutional variation is informative. The best overall model shifts from GDPpc-Polity (baseline) to Fitness-GDPpc-Technological Fitness (drop), but the MAE gap between them is only 0.002 — less than 10% of the typical prediction error. All model MAE differences remain below 3.1%.

On the Fitness-GDPpc plane (Figure S20), model ranking changes between baseline and drop occur in 16.3% of cells, concentrated exclusively along colour boundaries where competing models have near-identical smoothed MAE. The interior of each model's dominance region is completely unaffected (Figure S21).

We conclude that the paper's findings are robust to the treatment of missing Technological Fitness values.

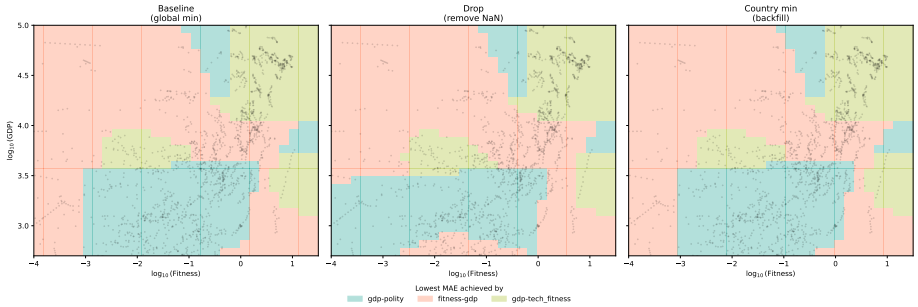


Figure S20: Best-model maps on the Fitness-GDPpc plane at $\Delta t = 4$ under three Technological Fitness imputation strategies: baseline (global minimum fill), drop (remove missing), and country minimum (per-country backfill). The three-model comparison (GDPpc-Polity vs Fitness-GDPpc vs GDPpc-Technological Fitness) is shown. Baseline and country minimum maps are visually indistinguishable. The drop strategy produces changes only at colour boundaries where models are effectively tied.

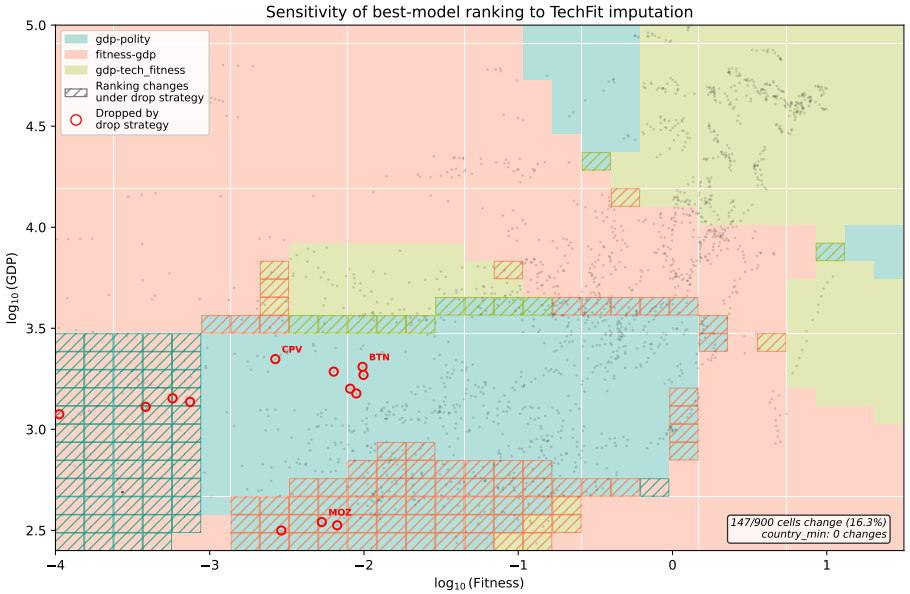


Figure S21: Spatial stability of model rankings under the drop imputation strategy. The baseline best-model map is shown with hatched cells indicating locations where the best model changes under the drop strategy. Red circles mark the positions of dropped observations (country-year pairs with missing Technological Fitness). Changes are concentrated at colour boundaries; the interior of each dominance region is stable.