WHEN DOES LEARNED NORMALIZATION ACTUALLY ADAPT? A DIAGNOSTIC ANALYSIS OF ADAPTIVE CENTRING IN TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive normalization schemes for time series forecasting introduce learnable coefficients intended to let models discover data-dependent centring behavior, but it is rarely established whether these coefficients capture genuine structure or merely reparameterize training around their initialization. We study one such mechanism, Adaptive Cross-Scale Normalisation (ACSN), across a factorial benchmark of 3,200 runs spanning four forecasting architectures, five datasets, four forecast horizons, and ten random seeds. The learned coefficient departs from its initialization in an architecture- and dataset-specific manner: on Electricity and Weather it shifts meaningfully, while on Exchange Rate and National Illness it remains close to initialization. Where the coefficient fails to move, a fixed normalization baseline reliably outperforms the adaptive variant; where it moves, adaptive normalization is more often competitive. Because the coefficient is sigmoid-parameterised and restricted to $(0, 1)$, it can approach but never equal the Fixed baseline ($\alpha = 1$). This constraint induces an inherent adaptive-versus-fixed performance gap whenever Fixed is near-optimal, complicating a direct interpretation of that comparison. Pooled performance comparisons show extreme heterogeneity (I^2 up to 99.9%), so we report conditional, dataset- and horizon-specific effects rather than a single aggregate number. On the evidence gathered here – one controlled-centring scalar, five datasets, and ten seeds per configuration – we do not claim ACSN is a general-purpose normalization improvement; we present it, and the analysis pipeline built around it, as a diagnostic method for identifying when a normalization coefficient is and is not doing meaningful work.

1 INTRODUCTION

Multi-scale forecasting architectures iteratively refine forecasts across temporal resolutions, but this refinement introduces a potential mismatch between observed inputs and model-generated representations propagated between scales. Cross-scale normalisation is commonly used to reduce this mismatch, yet is typically treated as a fixed architectural assumption rather than an empirically testable component, making it difficult to determine whether normalisation is actively required, merely tolerated, or useful only under particular combinations of dataset, architecture, horizon, and metric.

We study Adaptive Cross-Scale Normalisation (ACSN), which makes normalisation strength a learned, scale-dependent coefficient $\alpha_s \in (0, 1)$, and ask a question logically prior to "does it improve accuracy": does the coefficient adapt at all, in a way that is reproducible and task-dependent rather than optimisation noise? We combine large-scale paired benchmarking with random-effects meta-analysis, power analysis, and mechanistic analysis of the learned coefficient itself to answer this jointly with the standard performance question.

Research questions. (RQ1) Does ACSN improve accuracy relative to fixed, forced, and no-normalisation baselines? (RQ2–3) Are effects statistically and practically meaningful, and do they survive multiplicity correction? (RQ4) Are the estimated effects adequately powered? (RQ5) Does ACSN exhibit genuine adaptive behaviour rather than reproducing a fixed parameterisation? (RQ6) Is the learned coefficient itself informative about the forecasting problem, independent of accuracy? (RQ7) Do any gains justify the computational overhead?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

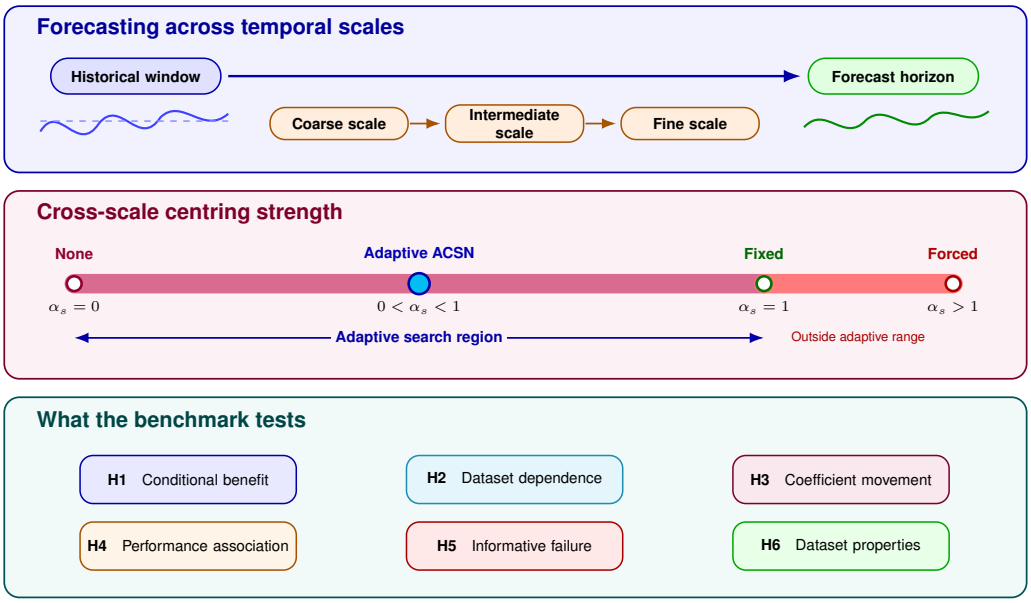

Figure 1: Problem setup and experimental logic. A historical observation window is processed through coarse-to-fine temporal refinement to produce a future forecast. The adaptation coefficient α_s controls cross-scale centring strength: $\alpha_s = 0$ applies no centring, $\alpha_s = 1$ applies full centring, and $\alpha_s > 1$ applies over-centring. Adaptive ACSN is restricted to $0 < \alpha_s < 1$. The benchmark tests whether its effects are conditional and dataset-dependent, whether the coefficient moves systematically, whether movement is associated with forecasting performance, and whether adaptation failure and dataset properties provide mechanistic information.

Contributions. (i) A large-scale, seed-matched, fully completed factorial benchmark (3,200 runs) with paired statistical tests, effect sizes, multiplicity correction, and power analysis. (ii) A mechanistic analysis showing the learned coefficient is reproducible across seeds and structured by architecture and dataset identity, but that this structure is diagnostic of *when* adaptation helps rather than a guarantee that it does. (iii) A formal proposition showing that the coefficient's sigmoid parameterisation structurally bounds the adaptive-versus-fixed comparison, which we argue must be accounted for before treating that comparison as a fair contest. (iv) An honest accounting of heterogeneity (I^2 up to 99.9%) and power (437 of 1,200 comparisons underpowered), reported as limitations rather than resolved.

2 Related Work

Heterogeneous forecasting domains. Financial, meteorological, transportation, energy, and epidemiological forecasting each pose distinct temporal-dependency challenges (Epaphra, 2017; Wang, 2019; Xie, 2024; Cabrera, 2019; Pradhan, 2010; Umoru, 2023; Zhou et al., 2021; Wu et al., 2021; Jiang et al., 2023; Pham et al., 2022; Ashraf et al., 2026; Luo, 2023; Baruník, 2012), sharing a common theme: dependencies that evolve and interact across multiple temporal resolutions rather than a single characteristic horizon.

Transformer-based and multi-scale forecasting. Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and FEDformer (Zhou et al., 2022) extend attention to long-horizon forecasting through sparsity, decomposition, and frequency-domain processing; PatchTST (Nie et al., 2023) and iTransformer (Liu et al., 2024) instead improve representation quality via patch tokenisation and inverted attention. Scaleformer (Shabani et al., 2022) models hierarchical structure explicitly via iterative coarse-to-fine refinement, introducing multi-scale variants (Autoformer-MS, Informer-MS) used in our benchmark alongside PatchTST and iTransformer. Refinement across scales, however, introduces distributional mismatch between observed and generated representations that existing work

addresses with a single, architecture-specific normalisation scheme evaluated on one backbone, leaving open whether this is general or an artefact of that architecture.

Normalisation in time-series models. Reversible instance normalisation (RevIN) (Kim et al., 2022) normalises each series instance and restores its statistics on output, and is now a near-default component in long-horizon forecasting architectures. Its effectiveness, however, is known to trade off variance stabilisation against information preservation (Zhang et al., 2024) and to vary across architectures and objectives, an interaction that is rarely tested directly rather than assumed fixed. Several approaches introduce learnable normalisation behaviour rather than treating normalisation strength as fixed. Deep Adaptive Input Normalisation (DAIN) (Passalis et al., 2019) learns data-dependent transformations intended to adapt multiple aspects of the input distribution. ACSN addresses a narrower question. Rather than designing a general adaptive normalisation mechanism, ACSN isolates a single learnable centring coefficient within a RevIN-style framework, allowing adaptation itself to be studied as an observable quantity. We therefore position ACSN not as a competitor to broader adaptive schemes such as DAIN, but as a controlled-centring probe intended to identify when a learned normalisation coefficient moves, when it remains near its initialisation, and whether that behaviour corresponds to meaningful differences in forecasting performance.

Statistical validation in deep learning. Reported gains are frequently comparable in magnitude to variation from random initialisation and optimisation dynamics (Xin, 2025), motivating paired designs, confidence intervals, effect sizes, power analysis, and multiplicity correction (Zhan et al., 2026), which remain uncommon specifically within long-horizon forecasting despite normalisation effects being small and context-dependent.

Positioning. We embed four architecturally distinct backbones within a common multi-scale refinement framework, holding refinement fixed while varying backbone and normalisation strategy, and pair this with a full statistical evidence chain (significance, effect size, multiplicity correction, power, heterogeneity) applied to both forecasting performance and the learned coefficient itself.

3 Problem Setup

Given a historical window $X_{t-L+1:t} \in \mathbb{R}^{L\times d}$, the forecasting task predicts $X_{t+1:t+H}$. A *temporal scale* s is a resolution produced by a multi-scale refinement procedure; *cross-scale normalisation* is a transformation at scale s intended to reduce distributional mismatch between observed and generated representations. The *adaptation coefficient* $\alpha_s \geq 0$ controls the strength of this correction: $\alpha_s = 0$ applies no adjustment, $\alpha_s = 1$ applies full correction, and $\alpha_s > 1$ over-corrects. In the adaptive setting introduced below, α_s is further restricted to $(0, 1)$ by construction – a restriction whose consequences we formalise in Section 4 and apply throughout Section 6.

We test six falsifiable hypotheses: adaptive normalisation is conditionally rather than universally beneficial (H1); effect magnitude and direction are dataset-dependent (H2); genuine adaptivity is reflected in systematic, non-trivial coefficient movement (H3); learned adaptation is associated with forecasting performance (H4); adaptation failure carries mechanistic information (H5); and dataset-level properties partially explain adaptation behaviour (H6).

4 Method

Adaptive Cross-Scale Normalisation. ACSN is a controlled-centring extension of RevIN (Kim et al., 2022). At scale s, given input x_s and centring statistic μ_s,

$$\hat{x}_s = \frac{x_s - \alpha_s \mu_s}{\sigma_{\alpha,s}}, \qquad \sigma_{\alpha,s} = \sqrt{\mathrm{Var}(x_s - \alpha_s \mu_s) + \varepsilon}, \tag{1}$$

with the original scale restored by $x_s = \hat{x}_s\, \sigma_{\alpha,s} + \alpha_s \mu_s$. When μ_s is computed over the same dimension that $\mathrm{Var}(\cdot)$ reduces over, variance is shift-invariant, so α_s acts only on centring, not on variance normalisation; we therefore describe ACSN as a controlled-centring probe rather than a general adaptive-normalisation mechanism.

Parameterisation. ACSN maintains an unconstrained raw parameter $a_s \in \mathbb{R}$ and sets $\alpha_s = \sigma(a_s) \in (0, 1)$, initialised at $a_s = 0$ so $\alpha_s = 0.5$. We distinguish these explicitly because it is easy to conflate

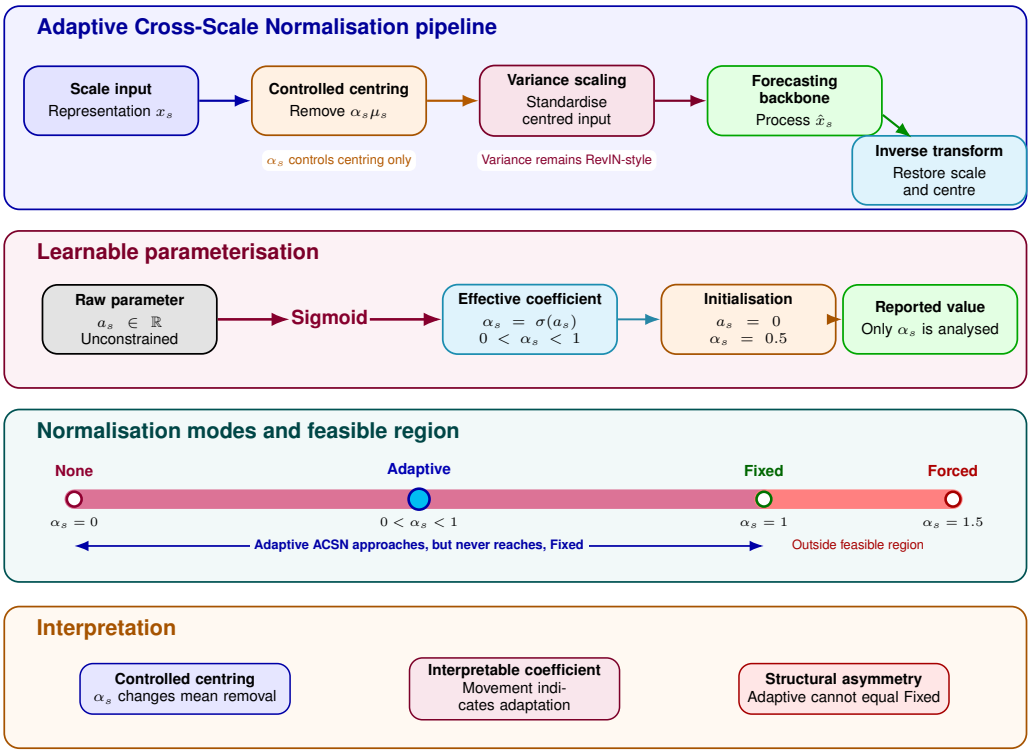

Figure 2: Adaptive Cross-Scale Normalisation. At each scale, ACSN removes a learned proportion α_s of the centring statistic before applying standard RevIN-style variance scaling. The unconstrained parameter a_s is transformed through a sigmoid to obtain the effective coefficient $\alpha_s \in (0, 1)$, initialised at 0.5. Adaptive ACSN therefore interpolates between no centring and full centring, but can approach rather than exactly attain the Fixed condition and cannot represent the Forced condition. The learned coefficient is consequently interpreted as a diagnostic measure of preferred centring strength.

them: a_s is never bounded and is not the quantity we report; every value of "the learned coefficient" in Section 6 is α_s, after the sigmoid.

This bound has a direct consequence for the baselines below: because $\alpha_s \in (0, 1)$ strictly, Adaptive ACSN can approach but never exactly reach $\alpha_s = 1$ (Fixed) and cannot represent $\alpha_s > 1$ (Forced) at all. We formalise this as follows.

Proposition 1 *Let $\alpha_s = \sigma(a_s)$ for $a_s \in \mathbb{R}$, and let $\mathcal{L}(\alpha)$ denote population forecasting loss for a fixed dataset, architecture, and horizon, as a function of a coefficient held constant at α across scales. If $\mathcal{L}$ is minimised on $[0, 1]$ at $\alpha^\star = 1$, then for every $a_s \in \mathbb{R}$, $\mathcal{L}(\sigma(a_s)) \geq \mathcal{L}(1)$ strictly, and $\inf_{a_s} \mathcal{L}(\sigma(a_s)) = \mathcal{L}(1)$ is not attained by any finite a_s.*

The proof is immediate from $\sigma(a_s) < 1$ for all finite a_s. The proposition converts an empirical observation (Section 6: Adaptive loses to Fixed on primary metrics) into a partly structural one: whenever Fixed is near-optimal, some of that gap is not attributable to adaptation *failing*, but to the parameterisation being unable to reach the optimum at all. It does not tell us how much of the empirical gap is explained this way versus by the optimiser not approaching the boundary as closely as training allows; we return to this in Section 7.

Normalisation modes. *Adaptive*: $\alpha_s = \sigma(a_s)$, learned. *Fixed*: $\alpha_s = 1$ (standard RevIN), the primary baseline. *Forced*: $\alpha_s = 1.5$, a deliberate over-correction stress test outside the feasible region. *None*: $\alpha_s = 0$, no mean removal.

5 BENCHMARK DESIGN

The benchmark is a full factorial design: 5 datasets $\times$ 4 architectures (Autoformer-MS, Informer-MS, PatchTST, iTransformer) $\times$ 4 forecast horizons $\times$ 4 normalisation modes $\times$ 10 seeds $=$ 3,200 runs, all of which completed successfully with no missing or duplicate configurations (Table 1). Datasets are Exchange Rate (Lai et al., 2018), Weather, Traffic, Electricity, and National Illness (Wu et al., 2021), following the configuration of Shabani et al. (2022): horizons $\{96, 192, 336, 720\}$ for the first four, $\{24, 32, 48, 64\}$ for Illness (its shorter weekly sampling means dataset identity and absolute horizon are confounded for Illness specifically; we use each dataset's own relative horizon position, H1–H4, wherever this matters). All comparisons are paired by dataset, architecture, horizon, and seed. Primary metrics are MSE, MAE, RMSE; MAPE/MSPE are supplementary and become unstable near zero, so we do not use them as primary evidence when they disagree with the primary metrics. Full dataset statistics, training protocol, and window sizes are in Appendix A.

Table 1: Benchmark completion by architecture.

Architecture	Completed	Planned	% complete
PatchTST	800	800	100.0
iTransformer	800	800	100.0
Autoformer-MS	800	800	100.0
Informer-MS	800	800	100.0
Total	3200	3200	100.0

6 RESULTS

Does ACSN learn meaningful adaptation? Across all 800 adaptive configurations, the learned coefficient deviates only modestly from its initialisation: mean $\alpha = 0.5115$ versus 0.5, corresponding to a mean absolute shift of 0.0185 (Table 2, Figure 3). Despite this small average movement, the effect is highly reproducible, with 73.8% of configurations ending above the initial value and a one-sample test against $\alpha = 0.5$ yielding $t = 12.06$ ($p = 6.97 \times 10^{-31}$). Importantly, the benchmark-wide average masks substantially larger dataset- and scale-specific behaviour. While the overall mean remains close to 0.5, individual scale coefficients range from approximately 0.43 to 0.58, indicating adaptation magnitudes roughly four times larger than the benchmark-wide average shift. We therefore interpret the statistical significance not as evidence of a large universal effect, but as evidence that coefficient movement is systematic and reproducible rather than random optimisation noise. Scale-specific coefficients reach $\alpha \approx 0.58$ and $\alpha \approx 0.43$, indicating that checkpoint averaging masks substantially more structured adaptation (Appendix B, Figure 7).

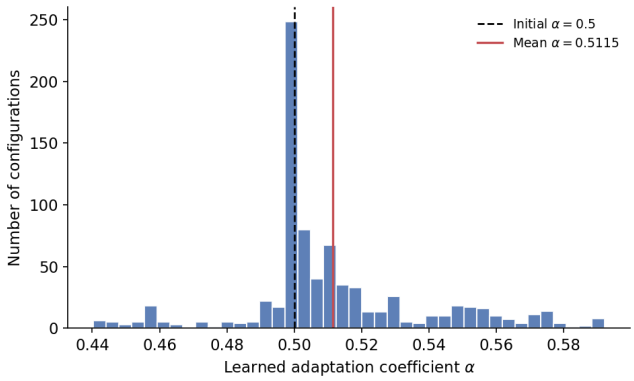

Figure 3: Distribution of learned α across 800 adaptive configurations. Dashed line: initialisation (0.5); solid line: benchmark mean (0.5115).

Table 2: Descriptive statistics for learned α; test is against $H_0 : \bar{\alpha} = 0.5$.

Mean α / median / SD	0.5115 / 0.5030 / 0.0270
Min / Max α	0.4404 / 0.5922
Mean abs. shift from 0.5	0.0185
t-statistic / p-value	12.06 / 6.97×10^{-31}

What determines the learned adaptation? A discriminability-ratio analysis (between-group variance relative to total variance; Figure 4) shows architecture (ratio 0.669, ANOVA $p = 3.6 \times 10^{-70}$) and dataset (0.476, $p = 1.2 \times 10^{-54}$) dominate horizon (0.158, $p = 7.4 \times 10^{-9}$), while random seed shows no detectable effect (Kruskal–Wallis $p = 0.999997$; range of seed means $< 10^{-3}$) – the coefficient is reproducible and structured by task, not optimisation noise (H3, confirmed). Electricity and Weather show the largest departures (mean $\alpha \approx 0.53$); Exchange Rate and National Illness remain within 0.005 and 0.001 of initialisation respectively (H6, confirmed; Appendix B, Table 6).

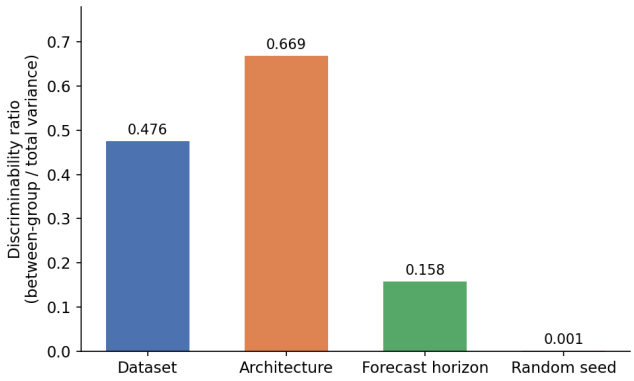

Figure 4: Discriminability ratios by experimental factor.

When does adaptive normalisation help? Table 3 reports random-effects (DerSimonian–Laird) pooled effects against each baseline for MSE, MAE, and RMSE. Adaptive beats None and Forced on all three metrics, but Fixed beats Adaptive by 16.9% (MSE), 12.4% (MAE), and 8.1% (RMSE) – a comparison that Proposition 1 shows is not evaluated on equal footing. Every pooled estimate carries extreme heterogeneity ($I^2 \in [96, 100]\%$), indicating that the effects vary substantially across datasets, architectures, and horizons rather than representing a single uniform phenomenon. To make this heterogeneity interpretable, we summarise the Adaptive-vs-Fixed comparison at two complementary levels. Table 4 reports simple arithmetic mean MSE differences within each dataset, providing a descriptive overview of the observed effects. Figure 5 further decomposes these effects by dataset and architecture, revealing substantial interactions. Later, Table 7 reports dataset-specific random-effects pooled estimates that explicitly account for between-configuration heterogeneity. Together, these analyses support H1 and H2 while avoiding over-interpretation of the pooled benchmark-wide estimates. Traffic is the most consistent failure case, with Fixed outperforming Adaptive by a relatively small but stable margin. Exchange Rate and National Illness are the largest failure cases and are also the datasets where α remains closest to its initial value, supporting H4 and H5 (Appendix B, Table 9). Weather illustrates why pooled summaries alone can be misleading: although its dataset-level average favours Fixed, PatchTST and iTransformer frequently favour Adaptive, producing a strong dataset-by-architecture interaction visible in Figure 5. Pooled across all datasets, the Adaptive-vs-Fixed gap narrows monotonically from $+21.5\%$ at the shortest relative horizon to $+13.0\%$ at the longest (Appendix B, Table 8), consistent with, though not proof of, the boundary in Proposition 1 becoming less influential as forecasting difficulty increases.

Coefficient behaviour as a diagnostic signal. Figure 6 relates the mean learned coefficient to win/loss outcomes across all 80 dataset–architecture–horizon cells: configurations where α stagnates near 0.5 are not exclusively wins or losses, but the two clearest failure regimes (Exchange Rate; National Illness for three of four architectures) are exactly the two lowest-adaptation regimes in the benchmark. We read this as evidence that coefficient stagnation reliably flags *when* adapta-

Table 3: Pooled random-effects meta-analysis, all completed configurations.

Comparison	Metric	Effect (%)	95% CI	I^2 (%)
Adaptive vs. None	MSE	−12.49	[−13.66,−11.30]	99.0
Adaptive vs. None	MAE	−8.20	[−9.14,−7.25]	98.9
Adaptive vs. None	RMSE	−6.45	[−7.08,−5.82]	99.0
Adaptive vs. Fixed	MSE	+16.85	[14.86,18.88]	99.8
Adaptive vs. Fixed	MAE	+12.44	[10.94,13.96]	99.8
Adaptive vs. Fixed	RMSE	+8.10	[7.17,9.03]	99.8
Adaptive vs. Forced	MSE	−8.84	[−11.52,−6.08]	99.9
Adaptive vs. Forced	MAE	−6.57	[−8.83,−4.26]	99.9
Adaptive vs. Forced	RMSE	−4.52	[−5.94,−3.09]	99.9



Figure 5: Mean Adaptive-vs-Fixed MSE difference by dataset and architecture (negative favours Adaptive). The strong interaction explains the heterogeneity in Table 3.

tion did not occur, but not, by itself, *why* – stagnation is consistent both with the midpoint being genuinely near-optimal and with the optimiser simply not moving a_s far from initialisation, and our data cannot distinguish these (Section 7).

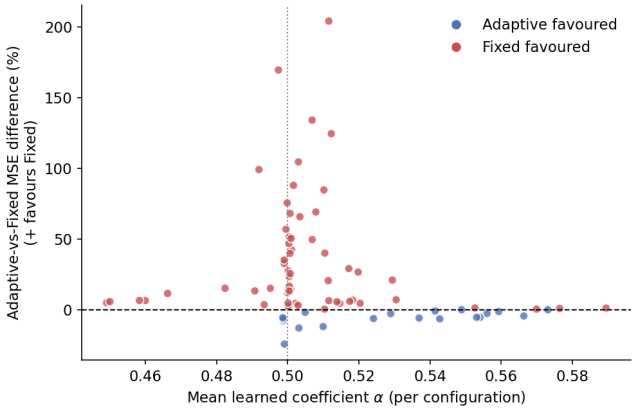

Figure 6: Learned coefficient vs. Adaptive-vs-Fixed MSE difference, all 80 cells. Blue: Adaptive favoured; red: Fixed favoured.

Robustness and efficiency. 763 of 1,200 paired comparisons (63.6%) reach power ≥ 0.80 (median required $N_{95} = 9$); the 437 underpowered comparisons concentrate in the supplementary met-

Table 4: Arithmetic mean Adaptive-vs-Fixed MSE difference across all configurations within each dataset. These descriptive means are provided for interpretability and differ from the random-effects pooled estimates reported later in Table 7.

Dataset	Mean effect (%)	Favours
Traffic	$+5.11$	Fixed
Electricity	$+10.15$	Fixed
Weather	$+15.65$	Mixed
National Illness	$+22.59$	Fixed
Exchange Rate	$+71.32$	Fixed

rics and the Adaptive-vs-None/Forced comparisons, not the primary Adaptive-vs-Fixed claims (Appendix B, Table 10). We disclose this as an unresolved limitation rather than a resolved one: closing it for the worst cells would require sample sizes in the tens or hundreds of thousands by our own calculations, outside this study's compute budget. Parameter overhead is negligible (one scalar per scale) and mean training runtime differs by at most 1.9% across modes (Appendix A, Table 5) – any trade-off concerns forecasting behaviour, not compute (RQ7).

7 Discussion

RQ1–2. Adaptive beats None and Forced (both outside or at the edge of its feasible range) but loses to Fixed, pooled, on all three primary metrics. Fixed therefore remains the strongest-performing centring strategy in the benchmark, and we cannot recommend ACSN as a drop-in replacement when forecasting accuracy alone is the objective. Proposition 1 does not change that conclusion; rather, it affects how the Adaptive-versus-Fixed gap should be interpreted, because Adaptive can asymptotically approach but never exactly attain the $\alpha = 1$ parameterisation represented by Fixed. We therefore view the comparison as informative about both forecasting performance and the consequences of the parameterisation itself, rather than attributing the entire gap solely to adaptation failure. Nearly all pooled effects are significant, but $I^2 \in [96, 100]\%$ means the pooled numbers alone are not practically meaningful; the dataset-conditional analyses provide the more informative level of interpretation. **RQ3–4.** Failure-case statistics remain significant after FDR correction for three of four architectures on National Illness (Appendix B); power is adequate for the primary comparisons but not uniformly so. **RQ5–6.** The coefficient moves systematically, reproducibly across seeds, and is structured by architecture and dataset identity – genuine in the sense of not being optimisation noise, though not unconstrained, since it remains a bounded interpolation between two fixed endpoints. Importantly, the value of ACSN is not limited to whether it improves forecasting accuracy. The learned coefficient consistently distinguishes settings in which adaptation occurs from settings in which it does not, revealing strong dataset- and architecture-specific behaviour that would be hidden by a purely performance-focused analysis. We therefore interpret the coefficient as a diagnostic signal of adaptation rather than a guarantee of improved forecasting performance. At the same time, the evidence collected here does not allow us to determine why adaptation stagnates in some settings. In particular, without training-time a_s trajectories or a dense fixed-α sweep on the failure datasets, we cannot distinguish between two competing explanations: (i) the midpoint is genuinely near-optimal for the task, or (ii) optimisation did not move the coefficient sufficiently far from its initial value. Resolving that distinction remains an important direction for future work.

Limitations. (1) The Adaptive-vs-Fixed comparison is structurally asymmetric (Proposition 1); we cannot precisely quantify how much of the empirical gap this explains. (2) National Illness confounds dataset and horizon. (3) No external normalisation baselines beyond internal ablations of ACSN itself; we do not compare to other adaptive normalisation schemes (Passalis et al., 2019). (4) Five datasets, none at high-frequency/sub-daily sampling, is a narrow basis for general claims – though the datasets do span a wide range of coefficient of variation (0.145–0.813), and the two lowest-CoV datasets are exactly where adaptation failed. (5) A single centring scalar cannot differentiate between fundamentally different normalisation strategies; we treat ACSN as a controlled-centring probe, not a general adaptive-normalisation method, deliberately. (6) $n = 10$ seeds is adequate for the primary comparisons but leaves 437/1,200 comparisons underpowered; we did not

retain training-time coefficient trajectories, per-sample residuals, or GPU memory measurements. Full detail on each point is in Appendix C.

8 Conclusion

We asked whether ACSN's normalisation coefficient adapts at all before asking whether adaptation improves forecasting. The answer is yes: the coefficient adapts reproducibly, is structured by architecture and dataset identity, and exhibits behaviour that cannot be explained by optimisation noise alone. Whether that adaptation improves forecasting is a separate question. Fixed centring outperforms Adaptive on the primary metrics when results are pooled across the benchmark, although the magnitude of this comparison is influenced by the bounded ACSN parameterisation (Proposition 1). The primary contribution of ACSN is therefore not a new state-of-the-art forecasting method, but a controlled-centring probe that makes adaptation itself observable. Across 3,200 experiments, the learned coefficient revealed systematic and interpretable differences in how architectures and datasets respond to centring, even when these differences did not translate into improved forecasting performance. We view the coefficient as a diagnostic signal of when adaptation occurs and when standard centring is already sufficient, and hope this framework motivates future work on richer but equally interpretable adaptive normalisation mechanisms.

Broader Impacts

This work studies adaptive normalisation in time-series forecasting and introduces a diagnostic framework for analysing learned centring behaviour. The contribution is methodological and uses only publicly available benchmark datasets. Potential benefits include improved understanding of forecasting model behaviour in applications such as energy, transportation, epidemiology, and finance. Potential risks are limited to those associated with forecasting systems generally and are not materially increased by the proposed method.

AI Use Statement

Generative AI tools were used in a limited capacity for language editing, manuscript revision, code drafting, and figure preparation. All experimental results, analyses, calculations, and figures were generated from benchmark outputs and verified by the authors. No reported results were generated, selected, or modified by AI systems.

Ethics Statement

This work uses only publicly available, previously published benchmark datasets (Exchange Rate, Weather, Traffic, Electricity, National Illness) that contain no personal or sensitive information. We are not aware of ethical concerns specific to this study.

Reproducibility Statement

The full factorial benchmark design, training protocol, and analysis pipeline are described in Section 5 and Appendix A; every table and figure in the paper is computed deterministically from the released per-seed results tables. The proof of Proposition 1 is given in full in Section 4. An anonymous repository containing the source code, benchmark configuration, statistical analysis scripts, and figure-generation code used in this study is available at: https://github.com/anon23researcher/acsn-diagnostic-study

References

Imran Ashraf, Inzamam Mashood Nasir, Muhammad Awais, Wided Bouchelligua, Sabrina Mansour, Enas Alyounis, and Marriam Nawaz. Deep sequence learning with multi-task supervision for scalable population health monitoring. In *Frontiers in Public Health*, volume 14, 2026.

Jozef Baruník. Modeling and forecasting exchange rate volatility in time-frequency domain. *Journal of Financial Econometrics*, 10(3):404–447, 2012.

José Cabrera. Modeling and projection of the mexican exchange rate. *Economia Mexicana*, 28(1): 45–78, 2019.

Manamba Epaphra. Modeling exchange rate volatility: Application of the GARCH and EGARCH models. *Journal of Mathematical Finance*, 7(1):121–143, 2017.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Mengying Luo. COVID-19 and FX market analysis. *Finance Research Letters*, 51:103450, 2023.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Deep adaptive input normalization for time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3760–3765, 2019.

Chi Hieu Pham, Quang Minh Nguyen, Duc Thang Nguyen, and Tran Quoc Anh Tao. Short-term electricity load forecasting based on temporal fusion transformer model. *IEEE Access*, 10: 106296–106304, 2022.

Rudra P. Pradhan. Forecasting exchange rate in india using artificial neural network. *Journal of Forecasting*, 29(5):456–471, 2010.

Mohammad Amin Shabani, Amir Hosein Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. In *International Conference on Learning Representations*, 2022.

David Umoru. Forecasting exchange rate dynamics in developing countries. *Journal of African Finance*, 12(2):88–104, 2023.

Lu Wang. Chinese currency exchange rates forecasting with EMD-based neural network. *Expert Systems with Applications*, 125:55–65, 2019.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, 2021.

Rui Xie. Study on exchange rate forecasting with stacked optimisation. *Journal of Forecasting*, 43 (1):110–125, 2024.

Dong Xin. Multi-source heterogeneous data fusion and intelligent prediction modeling for chemical engineering construction projects based on improved transformer architecture. *Scientific Reports*, 15, 2025.

Qing Zhan, Min Hu, Liang He, Guang Wang, and Jun Liu. A tale of two variances: When single-seed benchmarks fail in bayesian deep learning. *arXiv preprint arXiv:2604.23114*, 2026.

Jiawen Zhang, Xiao Wen, Zhiqiang Zhang, Shuicheng Zheng, Jian Li, and Jiang Bian. ProbTS: Benchmarking point and distributional forecasting across diverse prediction horizons. In *Advances in Neural Information Processing Systems*, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 2022.

A Extended Benchmark and Dataset Detail

A.1 Datasets

Exchange Rate contains daily exchange rates for 8 countries (7,588 timestamps, 1990–2010) (Lai et al., 2018). Weather contains 21 meteorological variables sampled at 10-minute intervals through 2020 (52,696 timestamps). Traffic aggregates occupancy measurements from San Francisco Bay Area freeway sensors (17,544 timestamps, 862 variables, 2016–2018). Electricity (ECL) records electricity consumption for 321 customers (26,304 timestamps, 2016–2019). National Illness (ILI) contains weekly influenza-like-illness indicators (966 timestamps, 7 variables, 2002–2020). All dataset splits follow the standard chronological 70%/10%/20% train/validation/test protocol.

A.2 Dataset Availability and Licensing

All datasets used in this study are publicly available benchmarks commonly used in long-horizon forecasting research. Exchange Rate, Weather, Traffic, Electricity, and National Illness were obtained from publicly distributed benchmark collections used by Informer (Zhou et al., 2021), Autoformer (Wu et al., 2021), and subsequent forecasting studies. Users should consult the original dataset providers for the applicable licence terms and conditions. No proprietary, confidential, restricted-access, or personally identifiable data were used in this work.

A.3 Training Protocol

AdamW, learning rate 10^{-4}, up to 10 epochs with early stopping (patience 3). Batch size 16 for Exchange Rate/Weather/Illness, 8 for Traffic/Electricity given their larger variable counts. Encoder input $L_{\text{input}} = 96$, label window $L_{\text{label}} = 48$ for the four higher-frequency datasets; $L_{\text{input}} = 32$, $L_{\text{label}} = 16$ for Illness. All four architectures, dataset splits, and seeds are held fixed within a matched comparison; normalisation mode is the only intended difference.

Table 5: Mean training runtime by normalisation mode (relative units), pooled across all 3,200 runs.

Mode	Mean runtime
Adaptive	9.624
Fixed	9.669
Forced	9.553
None	9.808

B Extended Results

National Illness is the only dataset where Adaptive-vs-Forced favours Forced ($+6.87\%$), consistent with near-total coefficient stagnation making Adaptive behave like Forced rather than Fixed there.

Table 6: Dataset characteristics and learned ACSN adaptation behaviour.

Dataset	Rows	Vars.	CoV	Mean α (mean abs. shift)
Electricity	26304	321	0.449	0.529 (0.0379)
Weather	52696	21	0.476	0.528 (0.0293)
Traffic	17544	862	0.813	0.496 (0.0202)
Exchange Rate	7588	8	0.145	0.504 (0.0045)
National Illness	966	7	0.554	0.500 (0.0006)

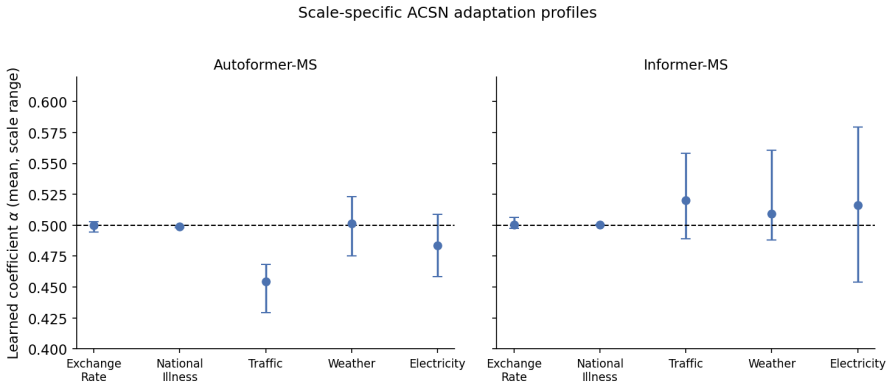

Figure 7: Scale-specific ACSN adaptation profiles for Autoformer-MS and Informer-MS. Points: mean learned coefficient per dataset; whiskers: mean per-run min–max across scales. Traffic shows the largest downward adaptation, Electricity the largest upward.

Table 9: Failure-case statistics (Adaptive vs. Fixed, MSE): National Illness and Exchange Rate. d_z is paired Cohen's d; p_{FDR} is Benjamini–Hochberg corrected.

Dataset	Model	Horizon	d_z	p_{FDR}
Exchange Rate	Autoformer-MS	720	2.42	8.2×10^{-5}
Exchange Rate	Informer-MS	720	-1.42	2.9×10^{-3}
Exchange Rate	PatchTST	720	-1.29	5.0×10^{-3}
Exchange Rate	iTransformer	336	3.81	3.1×10^{-6}
National Illness	Autoformer-MS	24–64	$\lvert d_z \rvert < 1$	≈ 0.02
National Illness	Informer-MS	64	7.26	2.9×10^{-8}
National Illness	PatchTST	64	31.51	9.3×10^{-13}
National Illness	iTransformer	64	17.80	5.3×10^{-11}

Autoformer-MS is the sole architecture on National Illness where sign consistently favours Adaptive, with effect sizes an order of magnitude smaller than the other three architectures – the basis for treating it as a partial exception rather than reading National Illness as uniformly a failure case. On Exchange Rate, horizon 720 is where Informer-MS and PatchTST flip sign relative to their own shorter-horizon results, the basis for the success cases discussed in Section 6.

C Extended Limitations

1. **Parameterisation asymmetry.** Proposition 1 shows Adaptive can only approach, never equal, Fixed; we cannot precisely partition the empirical gap between this structural bound and genuine adaptation failure, since we lack training-time gradient logs for a_s or a dense fixed-α sweep on $(0.5, 1]$.
2. **Dataset–horizon confound.** National Illness is the only dataset evaluated at horizons $\{24, 32, 48, 64\}$; relative-horizon-position analysis restores comparability but cannot fully separate a horizon effect from a dataset effect for Illness specifically.
3. **No external normalisation baselines.** Fixed/Forced/None are internal ablations of ACSN, all built on RevIN-style centring-and-scaling; we do not compare to other adaptive schemes (e.g. DAIN, Passalis et al., 2019).
4. **Dataset diversity.** Five datasets, none at high-frequency/sub-daily sampling. They span coefficient of variation 0.145–0.813, and the two lowest-CoV datasets are exactly where adaptation

Table 7: Per-dataset DerSimonian–Laird random-effects meta-analysis of MSE effects. Unlike the descriptive arithmetic means reported in Table 4, these pooled estimates explicitly account for between-configuration heterogeneity and therefore need not coincide with the corresponding dataset averages.

Dataset	Comparison	Effect (%)	95% CI	I^2 (%)
Traffic	vs. None / Fixed / Forced	-5.74 / $+4.94$ / -21.04	–	98.0/91.1/100.0
Electricity	vs. None / Fixed / Forced	-8.97 / $+8.84$ / -15.50	–	98.9/99.6/99.9
Weather	vs. None / Fixed / Forced	-8.44 / $+3.89$ / -1.28	–	97.3/98.6/97.4
National Illness	vs. None / Fixed / Forced	-7.56 / $+20.59$ / $+6.87$	–	99.2/99.6/96.3
Exchange Rate	vs. None / Fixed / Forced	-37.90 / $+59.65$ / -13.48	–	99.1/98.7/96.1

Table 8: Per-relative-horizon random-effects meta-analysis of MSE effects. H1 is each dataset's shortest evaluated horizon, H4 its longest.

Horizon	vs. None (%)	vs. Fixed (%)	vs. Forced (%)
H1	-13.78	$+21.49$	-4.28
H2	-13.81	$+15.72$	-6.72
H3	-12.31	$+16.35$	-12.44
H4	-11.04	$+13.03$	-11.67

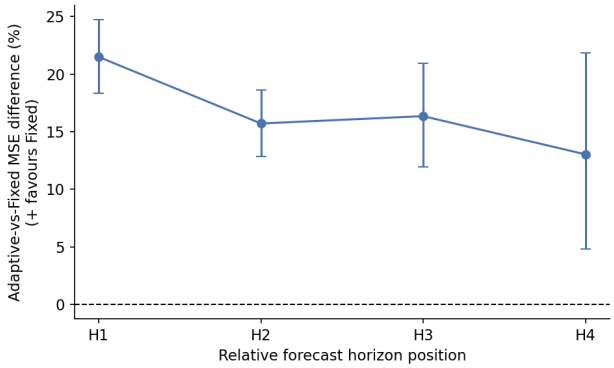

Figure 8: Adaptive-vs-Fixed MSE effect by relative horizon (pooled), corresponding to Table 8.

Table 10: Underpowered comparisons (power < 0.80) by metric and comparison; 437 of 1,200 total.

Metric	Count	Comparison	Count
MSPE	123	Adaptive vs. Forced	169
MAPE	111	Adaptive vs. None	159
MAE	77	Adaptive vs. Fixed	109
MSE	64		
RMSE	62		

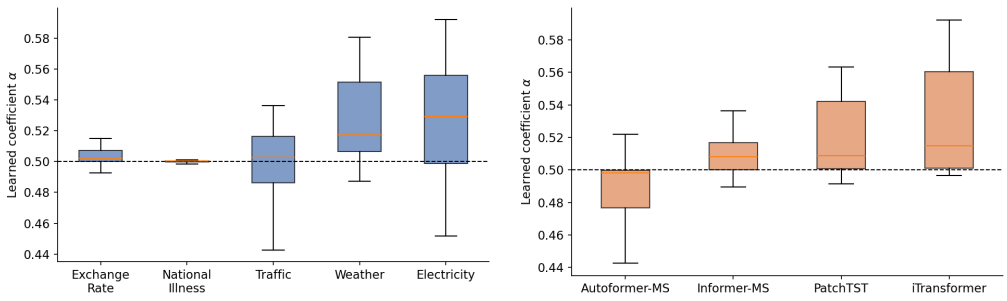

Figure 9: Learned α by dataset (left) and architecture (right).

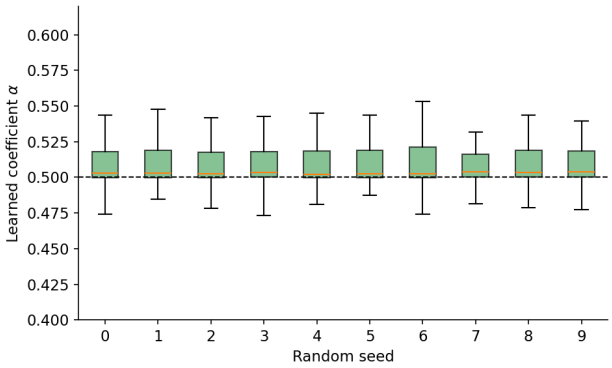

Figure 10: Learned α by random seed, illustrating the near-zero seed discriminability ratio.

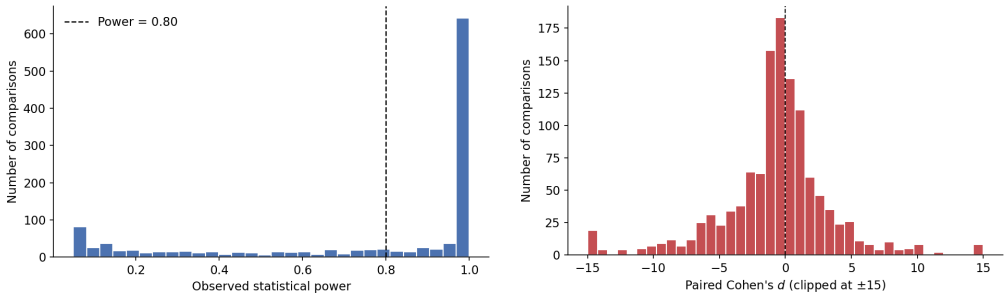

Figure 11: Distribution of achieved power (left) and paired Cohen's d, clipped at ± 15 (right), across all 1,200 comparisons.

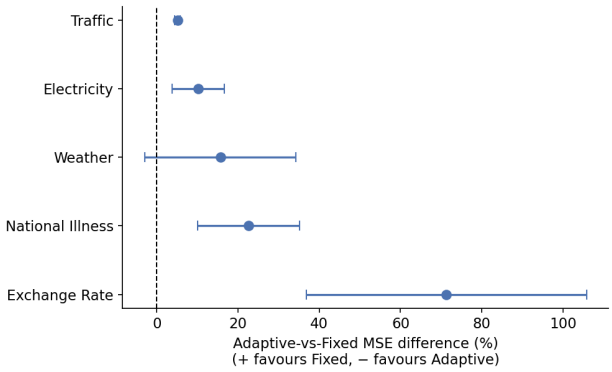

Figure 12: Dataset-level Adaptive-vs-Fixed effects (MSE) with 95% CIs, corresponding to Table 4.

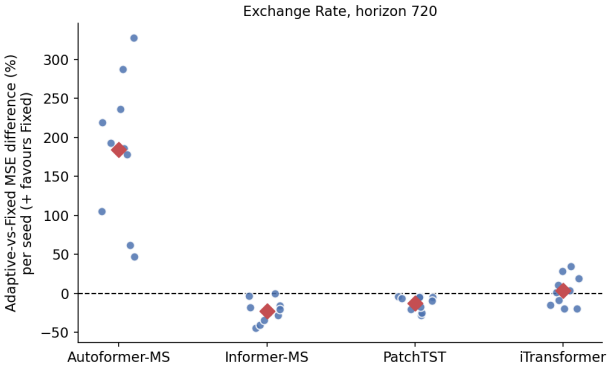

Figure 13: Seed-level Adaptive-vs-Fixed MSE differences, Exchange Rate, horizon 720, by architecture.

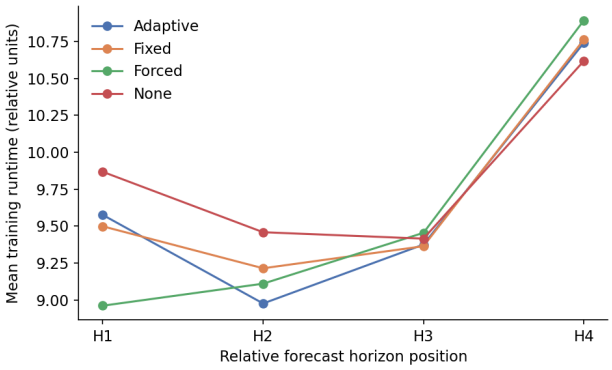

Figure 14: Mean training runtime by relative horizon and normalisation mode.