MIXBENCH-TS: A MULTIVARIATE TIME SERIES FORECASTING BENCHMARK WHERE CHANNEL MIXING PAYS OFF

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariate Time Series Forecasting (MTSF) models that mix information across channels assume that the past of one channel carries information about the future of another. Yet they are evaluated on a small fixed set of standard datasets whose cross-channel structure is rarely examined. We ask two questions: *How can we reliably measure lagged, non-linear, and joint coupling in MTSF datasets?* and *Do the standard datasets actually have such coupling?* To answer the first, we test four candidate measures on synthetic datasets with planted ground-truth coupling: Granger Causality (GC), Transfer Entropy (TE), lagged Mutual Information (MI), and the CD gain, a model-based measure we introduce that compares a channel-dependent (CD) model to its channel-independent (CI) variant. Only lagged MI and the CD gain recover every planted coupling. For the second question, the answer is a definite *no*, as the standard datasets have a median of only 23% lagged-coupled channel pairs and a median CD gain of -4.9%, compared to 78% and $+4.6\%$ on chaotic ODE systems. We therefore propose **MixBench-TS**, a benchmark of 10 real-world datasets with a median of 55.5% lagged-coupled pairs and a median CD gain of $+1.7\%$. Across six state-of-the-art models tuned under one protocol, CI models win on 10/10 (MSE) and 8/10 (MAE) standard datasets, but on only 3/10 and 2/10 MixBench-TS datasets. We recommend using our benchmark for evaluating new CD models. Moreover, we propose profiling new datasets with lagged MI and the CD gain before using them to evaluate multivariate models. Code and data are available at `https://anonymous.4open.science/r/mixbench-ts-B027`.

1 INTRODUCTION

Multivariate Time Series Forecasting (MTSF) models that mix information across channels are designed on an assumption of lagged coupling: *a source channel's past carries additional information about a target channel's future*. Contemporaneous coupling alone is not enough, since the source's value at the timestep of forecasting is unknown as well. On datasets whose cross-channel coupling is weak or mostly contemporaneous, a channel-dependent (CD) model therefore gains nothing over a channel-independent (CI) model, which forecasts each channel from its own past. Such datasets are a not a good fit for evaluating CD models.

Prior work has characterized MTSF datasets along axes such as domain coverage and temporal properties (Godahewa et al., 2021; Qiu et al., 2024), but not along the lagged coupling that CD models rely on. The closest effort (Abdelmalak et al., 2026) uses ODE datasets from Gilpin (2021) to argue that current datasets lack channel coupling. However, it relies on Granger Causality (GC)(Granger, 1969), which cannot capture non-linear lagged coupling. Another limitation is that they consider only the standard and ODE datasets. This leaves two open questions:

- Q1. *How can we reliably measure lagged, non-linear, and joint coupling in MTSF datasets?*
- Q2. *Once measured, do the standard datasets used by the field (Lai et al., 2018; Zhou et al., 2021; Wu et al., 2021) actually have such coupling?*

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.

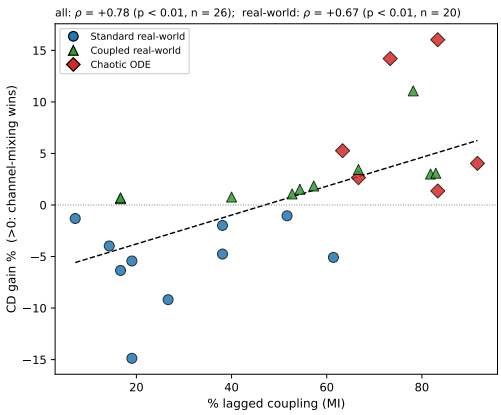


Figure 1: Lagged coupling vs. channel-mixing payoff on 26 datasets (10 standard, 10 coupled, 6 ODE). x: lagged MI ($\mathrm{MI}_{\mathrm{lag}}$), the share of channel pairs coupled more strongly at a lag than at lag 0. y: CD gain (CD_g), the MSE reduction of TSMixer over $\mathrm{TSMixer}_{\mathrm{CI}}$ (> 0: mixing helps). Dashed: linear fit. Spearman correlations above the plot.

Table 1: Retrieval of the coupling planted into the target c_0 of each synthetic dataset. ✓: exactly the planted source(s), at the planted lag where the measure resolves lags. ✗: missed, mis-located or false. CD_g uses identical hyperparameters for both variants and the independent dataset as reference (3 seeds). **Bold**: all retrieved. Details in Table 2.

Dataset	CD_g	$\mathrm{MI}_{\mathrm{lag}}$	TE	GC
independent	✓	✓	✓	✓
contemp-linear	✓	✓	✗	✓
lagged-linear	✓	✓	✓	✓
lagged-non-linear	✓	✓	✓	✗
multiplicative-joint	✓	✓	✓	✗
additive-mixed	✓	✓	✓	✗
Retrieved	**6/6**	**6/6**	5/6	3/6

To address Q1, we curate six synthetic datasets with known planted couplings. On this suite, we test four candidate measures: Granger Causality (GC; Granger, 1969), Transfer Entropy (TE; Schreiber, 2000), lagged Mutual Information ($\mathrm{MI}_{\mathrm{lag}}$; Shannon, 1948; Cover & Thomas, 2006), and the CD gain (CD_g), a model-based measure we introduce that compares a CD forecaster to its CI variant, following the landmarking concept of Pfahringer et al. (2000). GC misses the non-linear and joint couplings, and TE places the contemporaneous coupling at a wrong, positive lag. Only $\mathrm{MI}_{\mathrm{lag}}$ and CD_g recover every planted coupling (Table 1), so we adopt them as our two profiling tools.

We then address Q2 by measuring $\mathrm{MI}_{\mathrm{lag}}$ and CD_g on the 10 standard datasets, using six datasets from the chaotic ODE benchmark (Gilpin, 2021) as a known-correlated reference. The answer is a *definite no*, as the standard datasets show weak lagged coupling compared to the ODE datasets (median of 23% vs. 78% lagged-coupled pairs, Figure 1 and Table 3). Moreover, a CI variant of TSMixer beats the full channel-mixing model on all 10 standard datasets, while channel mixing helps on every ODE dataset (median CD_g of -4.9% vs. $+4.6\%$).

If the standard datasets fall short, the natural next step is to find datasets that have the lagged coupling that channel mixing needs. We therefore propose **MixBench-TS**, a benchmark of 10 real-world datasets that exist in the literature but are rarely used for MTSF evaluation (Qiu et al., 2024; Qiao et al., 2026; Stein et al., 2025; Baigutanova et al., 2025), selected by their positive CD_g. They cover environmental monitoring, energy, hydrology, marine sensing, and health, and we refer to them throughout as *the coupled datasets*. They show substantially stronger lagged coupling than the standard datasets.

To test whether this pattern carries over to actual forecasting, we evaluate six state-of-the-art models spanning both channel strategies (CI and CD), tuned under one unified protocol that also covers the lookback window (Abdelmalak et al., 2026). On the coupled datasets, where lagged coupling is substantially stronger, CI models win on only 3 of 10 (MSE), compared to all 10 standard datasets.

Our contributions are as follows:

- We introduce the CD gain (CD_g), a model-based measure of exploitable cross-channel coupling in MTSF datasets, and validate it alongside three other candidate measures on synthetic datasets with known ground-truth coupling (Sections 4 and 5).
- We show that the datasets the field treats as standard are poor tests of CD models (Section 6).

- We propose MixBench-TS, a benchmark of 10 real-world coupled datasets that are rarely used for MTSF evaluation and carry substantially stronger lagged coupling than the standard datasets (Section 7.1).
- We benchmark six models under a unified tuning protocol and show that the CI advantage collapses on the coupled datasets (Section 7.3).[1]

2 Related Work

Channel mixing strategies in MTSF models. There has been rising interest in building MTSF models with varying architectures. We categorize recent models into CI and CD models based on whether a model treats each channel as a separate time series or mixes information across channels. The CI category covers linear (Zeng et al., 2023), MLP-based (Lin et al., 2024), and even transformer-based models that still operate on each channel in isolation (Nie et al., 2023). The CD category includes transformer-based models that apply attention across channels (Zhang & Yan, 2023; Liu et al., 2024; Chen et al., 2025a) and MLP-based or linear models that mix channels (Chen et al., 2023; Yue et al., 2025). It also covers CNN-based models (Luo & Wang, 2024) and frequency-domain-based methods (Wang et al., 2024; 2025). More recent CD models mix channels selectively rather than fully, using graph-based filtering (Hu et al., 2025) or channel clustering (Qiu et al., 2025; Chen et al., 2024).

Benchmarks and evaluation. Many recent papers have extended the datasets available for MTSF in different directions, such as ODE-based benchmarks (Gilpin, 2021; Klötergens et al., 2025), bursty network-traffic benchmarks (Guthula et al., 2025), causality-driven benchmarks (Stein et al., 2025), and application-traffic benchmarks balanced across forecasting regimes (Xue et al., 2026). Other benchmarks have been proposed to extend domain coverage and characterize temporal properties of the data (Qiu et al., 2024; Godahewa et al., 2021; Aksu et al., 2024). On the *evaluation* side, Zeng et al. (2023) show that a simple linear model outperforms transformers on the standard datasets. Meanwhile, Chen et al. (2025b) analyze the existing benchmark based on transformer-based models' performance. Roque et al. (2025) investigate how papers choose datasets to evaluate on, while Brigato et al. (2026) study the effect of hyperparameter tuning and forecasting horizons on model performance on the standard datasets. Closer to our work, some studies investigate when cross-channel modeling pays off. For instance, Han et al. (2024) attribute the strong performance of CI models to their robustness to distribution shift. Shao et al. (2025) link model performance to the level of heterogeneity within datasets. Finally, Abdelmalak et al. (2026) show via Granger Causality that the standard datasets carry little linear channel coupling. We build on this literature by following rigorous evaluation practices in our protocol, and by extending the analysis to non-linear and lagged coupling.

3 Problem Formulation

Each dataset is a multivariate time series $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T) \in \mathbb{R}^{T \times C}$ with C channels (variates) observed at T equally spaced steps. Following Qiu et al. (2024), $\mathbf{Z}$ is split *chronologically* into training, validation, and test segments, and samples are built with a sliding window of stride 1. A model $f_{\boldsymbol{\theta}} : \mathbb{R}^{L \times C} \to \mathbb{R}^{H \times C}$ maps a *lookback window* of the last L steps to the next H steps (the *forecast horizon*), jointly over all C channels. Forecast targets never cross a split boundary (Appendix C). The standard datasets commonly use $H \in \{96, 192, 336, 720\}$, while for the remaining datasets we adapt the horizons to be semantically relevant (Appendix A).

4 Channel Dependency Measurement

We consider four complementary measures of cross-channel coupling, which differ in the type of coupling they target and in how they are computed. Granger Causality (GC), Transfer Entropy (TE), and Lagged Mutual Information ($\mathrm{MI}_{\mathrm{lag}}$) are statistical and are computed *pairwise and directionally*

[1] Code and data, including the profiling toolkit ($\mathrm{MI}_{\mathrm{lag}}$, CD_g) and the synthetic data generator, are available at `https://anonymous.4open.science/r/mixbench-ts-B027`.

on the data itself, whereas the CD gain (CD_g) is model-based. Each pairwise measure is evaluated on every ordered source–target pair of the channel set $\mathcal{C}$,

$$\mathcal{P} = \{(x, y) : x, y \in \mathcal{C},\, x \neq y\}, \qquad |\mathcal{P}| = C(C-1), \tag{1}$$

and then aggregated to a single dataset value. On wide datasets, $|\mathcal{C}|$ is capped at 40 by uniform random sampling of channels (sensitivity in Appendix G). Further implementation details are given in Appendix D.

Granger Causality (GC). Each pair is jointly made stationary (seasonal and ADF-selected regular differencing, Said & Dickey, 1984), so GC reflects directed predictive structure rather than shared trend or seasonality. We then fit two OLS regressions with p lags, where p is the smallest candidate lookback window of the dataset, on the N usable time steps: a univariate *restricted* model that predicts y_t from its own past $\{y_{t-1}, \ldots, y_{t-p}\}$, and a multivariate model that adds the source's past $\{x_{t-1}, \ldots, x_{t-p}\}$ (Granger, 1969). With $\mathrm{SSR_u}, \mathrm{SSR_{mv}}$ their sums of squared residuals and $\nu_{\mathrm{u}} = N - (p{+}1)$, $\nu_{\mathrm{mv}} = N - (2p{+}1)$ their residual degrees of freedom, our metric is the degrees-of-freedom corrected partial R^2 (Geweke linear feedback, Geweke, 1982), taken as the median over pairs:

$$R^2_{\mathrm{GC}}(x \to y) = 1 - \frac{\mathrm{SSR_{mv}}/\nu_{\mathrm{mv}}}{\mathrm{SSR_u}/\nu_{\mathrm{u}}}, \qquad \widetilde{R^2_{\mathrm{GC}}} = \operatorname*{median}_{(x,y)\in\mathcal{P}} R^2_{\mathrm{GC}}(x \to y). \tag{2}$$

Intuitively, R^2_{GC} is the fraction of y's residual variance that x's past removes. It is ≈ 0 under the null of no Granger causality and moves toward 1 the more x's past helps predict y. Dividing by the residual degrees of freedom cancels the mechanical decrease of SSR from the p extra regressors, which centers the null near 0 regardless of p or N.

Transfer Entropy (TE). TE uses the same joint differencing as GC, so the linear and non-linear measures see identical series. It is the directed, non-linear generalization of GC, approximated by the rank-space conditional mutual information $I(y_t; x_{t-\ell} \mid y_{t-1})$ (Schreiber, 2000; Kraskov et al., 2004). Here, y_t is the target's present, $x_{t-\ell}$ the source's value ℓ steps earlier, and y_{t-1} the target's most recent past. As in the restricted GC model, conditioning on y_{t-1} removes what the target's own past already explains, so only the *additional* information carried by the source is counted. We sweep the lags $\ell \in \{1, \ldots, p\}$, with p as for GC, and select $\ell^\star$ as the statistically significant lag with the largest effective value (Appendix D). Our metric maps the effective TE onto the R^2 scale of GC and, as for GC, takes the median over pairs:

$$R^2_{\mathrm{TE}}(x \to y) = 1 - e^{-2\,\mathrm{TE}^{\mathrm{eff}}(x \to y)}, \qquad \widetilde{R^2_{\mathrm{TE}}} = \operatorname*{median}_{(x,y)\in\mathcal{P}} R^2_{\mathrm{TE}}(x \to y), \tag{3}$$

where $\mathrm{TE}^{\mathrm{eff}}(x \to y) = \max\big(0,\, \widehat{I}(y_t; x_{t-\ell^\star} \mid y_{t-1}) - \overline{I}_{\mathrm{null}}\big)$ is the debiased TE (Marschinski & Kantz, 2002), with $\overline{I}_{\mathrm{null}}$ the mean of a permutation-based surrogate null (Runge, 2018). The mapping $R^2 = 1 - e^{-2I}$ is the Gaussian information–R^2 identity.

Lagged Mutual Information ($\mathrm{MI_{lag}}$). Unlike GC and TE, MI does not condition on the target's own past. At lag 0, the MI between x_t and y_t is a *symmetric*, common-driver measure of their coupling. At a positive lag, the MI between $x_{t-\ell^\star}$ and y_t is *directional*, from source to target, as in GC and TE. We compute it on the *raw* series (rank-transformed, no differencing) with a KSG k-NN estimator (Kraskov et al., 2004), at lag 0 and at the positive lag $\ell^\star$ that maximizes the effective MI over the geometric lag grid Λ (Eq. 9, Appendix D). Our metric, the *lagged-coupling fraction* ϕ_{lag}, summarizes whether lagged or contemporaneous coupling dominates,

$$\phi_{\mathrm{lag}} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \mathbb{1}\Big[\, \widehat{I}(x_{t-\ell^\star}; y_t) > \widehat{I}(x_t; y_t) \Big], \tag{4}$$

where $\widehat{I}(a; b)$ is the effective (surrogate-debiased) KSG estimate of the MI between a and b, and $\mathbb{1}[\cdot]$ is the indicator function. ϕ_{lag} is thus the share of pairs whose lagged coupling exceeds their contemporaneous coupling (reported in percent in Tables 3 and 5). A high value means cross-channel information is predominantly *lagged* (delayed transfer) rather than *contemporaneous* (a shared instantaneous driver). This matters because contemporaneous coupling often carries little useful information for forecasting, whereas lagged coupling can supply the extra predictive signal that a multivariate model exploits. Throughout the paper, we refer to this measure as $\mathrm{MI_{lag}}$ and report ϕ_{lag} as its dataset-level value.

CD Gain (CD_g). The statistical measures above test for specific forms of coupling (linear, non-linear, lagged) between pairs of channels, and can therefore miss coupling that only emerges jointly across several channels. We therefore follow the landmarking concept of Pfahringer et al. (2000), which characterizes a dataset through the performance of simple learners on it. We train two TSMixer variants (Chen et al., 2023) that differ only in their MLP-based channel-mixing component: TSMixer, which mixes information across channels, and $\mathrm{TSMixer}_{\mathrm{CI}}$, which drops the channel-mixing layers and forecasts each channel independently. CD_g is the relative reduction in test MSE obtained through channel mixing,

$$\mathrm{CD}_g = 100 \cdot \frac{\mathrm{MSE}_{\mathrm{CI}} - \mathrm{MSE}_{\mathrm{CD}}}{\mathrm{MSE}_{\mathrm{CI}}}, \tag{5}$$

where $\mathrm{MSE}_{\mathrm{CI}}$ and $\mathrm{MSE}_{\mathrm{CD}}$ denote the test MSE of $\mathrm{TSMixer}_{\mathrm{CI}}$ and TSMixer, averaged over the forecast horizons and three random seeds. A positive CD_g indicates that a multivariate model can exploit the cross-channel information, whereas a non-positive value indicates that channel mixing brings no benefit or even hurts the forecasting performance.

5 Evaluation of Channel Dependency Measurements

5.1 Synthetic Datasets Curation

To check whether the four measures capture couplings of varying types, we build synthetic datasets with known ground-truth coupling, which allows us to verify whether each measure recovers it. The datasets span three classes of cross-channel coupling: (i) no coupling, (ii) pairwise coupling, where a single source channel drives a target channel, either contemporaneously or with a lag and either linearly or non-linearly, and (iii) joint coupling, where several source channels drive the same target channel. Each base channel c_i is an Autoregressive (AR) process of order p_i driven by independent Gaussian noise, plus a seasonal component with a channel-specific period. We then plant a coupling from one or more source channels into a target channel c_0, according to its coupling type: $\mathit{coupling} \in \{\text{contemp-linear}, \text{lagged-linear}, \text{lagged-non-linear}, \text{multiplicative-joint}, \text{additive-mixed}\}$.

Figure 2 shows the six datasets and their planted coupling structure. The independent dataset establishes the null values of the measures. The three pairwise datasets couple a single source c_1 into c_0, and the two joint datasets couple c_1 and c_2 into c_0. In multiplicative-joint, the target depends on the product of both sources, so neither source affects it on its own. Figure 3 visualizes these couplings, where each arrow links a source value to its imprint on c_0 at the planted lag. Each dataset has $C = 4$ channels and $T = 20{,}000$ timesteps (details in Appendix J).

5.2 Evaluation of Measurements on Synthetic Datasets

We summarize whether each of the four measures retrieves the planted coupling of the six datasets, with the underlying values in Table 2. For the three pairwise measures, let $D_s(\ell)$ be the lag profile of the measure from source s into the target c_0 over the lag grid $\Lambda_{\mathrm{syn}} = \{1, \dots, 48\}$ (plus lag 0 for $\mathrm{MI}_{\mathrm{lag}}$), and $D_s^{(b)}(\ell)$ its value on the b-th of B circular-shift surrogates. We compute

$$\tau_{95} = Q_{0.95}\Big(\big\{\max_{s,\, \ell \in \Lambda_{\mathrm{syn}}} D_s^{(b)}(\ell)\big\}_{b=1}^{B}\Big), \qquad r_s = \frac{\max_{\ell \in \Lambda_{\mathrm{syn}}} D_s(\ell)}{\tau_{95}}, \qquad \hat{\ell}_s = \arg\max_{\ell \in \Lambda_{\mathrm{syn}}} D_s(\ell). \tag{6}$$

Here, $Q_{0.95}(\cdot)$ is the empirical 95% quantile over the B surrogates. Taking the maximum over all sources and lags corrects for testing many source–lag combinations, so that without coupling, any source exceeds τ_{95} at any lag with only 5% probability. A source is flagged if $r_s > 1$, and $\hat{\ell}_s$ is its detected lag. GC is cumulative over VAR orders and therefore gives no lag. A dataset counts as retrieved only if exactly the planted source(s) are flagged at the planted lag. For CD_g, we compare the gain of TSMixer over $\mathrm{TSMixer}_{\mathrm{CI}}$ on each coupled dataset to that on the independent reference, and count a dataset as retrieved if this excess gain is significantly positive over 3 seeds.

All four measures correctly report no coupling on independent and retrieve the lagged-linear coupling, but differ once the coupling is contemporaneous, non-linear, or joint. GC, being linear and

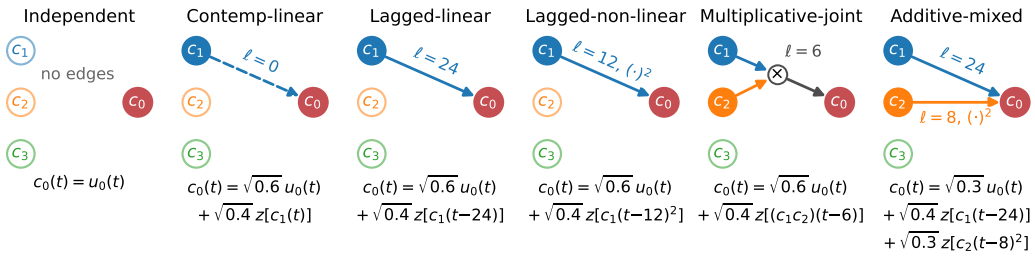


Figure 2: Planted coupling structure of the six synthetic datasets. Arrows mark the couplings into the target c_0, labeled with lag ℓ and form (dashed: contemporaneous, $(\cdot)^2$: squared, $\times$: multiplicative). Hollow nodes have no edge into c_0. Below each panel is the generating equation of c_0, with u_0 its standardized base process, $z[\cdot]$ standardization, and squared weights as variance shares (Appendix J).

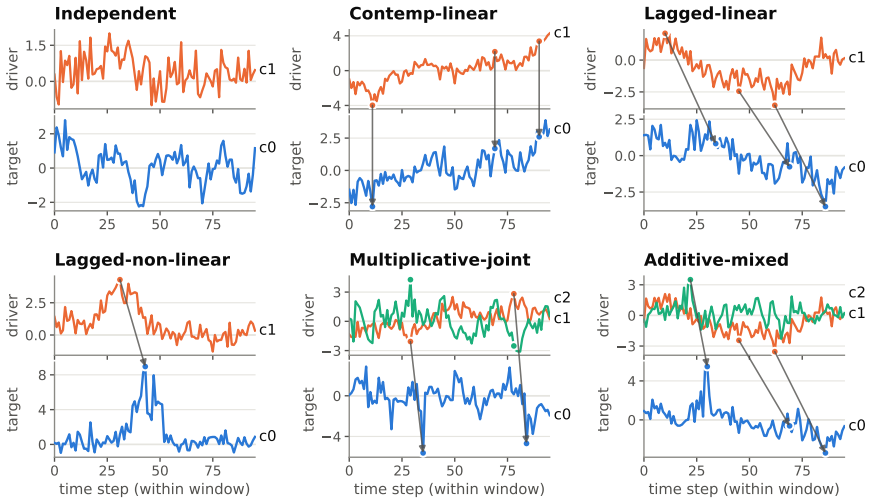


Figure 3: Planted couplings on a short window of each synthetic dataset, with the source channel(s) (labeled driver) on top and the target c_0 below. Each arrow links a source value to its imprint on c_0 after the planted lag ℓ. This is $\ell = 0$ (vertical) for contemp-linear, and 24, 12 and 6 for lagged-linear, lagged-non-linear and multiplicative-joint, respectively. In additive-mixed, c_2 arrives after $\ell = 8$ and c_1 after $\ell = 24$. independent has no coupling and no arrows.

conditioned on the source's past, flags the contemporaneous coupling only weakly ($r = 1.5$). It stays below the null on the non-linear coupling and on the squared source c_2 of additive-mixed, and barely exceeds it on the multiplicative coupling ($r = 1.1$ for c_1, below the null for c_2). TE retrieves all lagged couplings but, restricted to only positive lags, mis-locates the contemp-linear coupling at lag 24 ($r = 11$). Only $\mathrm{MI}_{\mathrm{lag}}$ and CD_g retrieve the planted coupling on all six datasets, since $\mathrm{MI}_{\mathrm{lag}}$ is non-parametric and also evaluated at lag 0, and CD_g makes no assumption on the functional form and sees all channels jointly. We therefore use them to profile the real-world datasets.

6 Channel Dependency Measurement of Standard Datasets

We apply $\mathrm{MI}_{\mathrm{lag}}$ and CD_g to the standard datasets, namely ETTh1, ETTh2, ETTm1, ETTm2 (Zhou et al., 2021), Electricity, Traffic, ExchangeRate, SolarAL (Lai et al., 2018), Weather, and ILI (Wu et al., 2021). As a reference, and similar to Abdelmalak et al. (2026), we also include six datasets from the chaotic ODE benchmark of Gilpin (2021), whose coupling is determined by the underlying ODEs (Appendix A). Table 3 reports both measures.

The two measures clearly separate the two groups. On $\mathrm{MI}_{\mathrm{lag}}$, the standard datasets have a median of only 23% lagged-coupled channel pairs, and only Weather (52%) and Electricity (61%) score above 50%. The ODE datasets reach a median of 78%, with none below 63%. CD_g confirms this

Table 2: Retrieval details behind Table 1, one row per planted source. CD_g: gain of TSMixer over $\mathrm{TSMixer}_{\mathrm{CI}}$ on c_0 minus that on the independent reference (pp), over forecast steps $h \le \ell$ (ℓ=24 if no lag is planted). Mean±std over 3 seeds, all significant (one-sided Welch t-test, $p < 0.05$). $\hat{\ell}$ and r as in Eq. equation 6, with a source flagged if $r > 1$. GC gives no lag. "< null": not flagged ($r \le 1$). Red: missed, mis-located or false.

			$\mathrm{MI}_{\mathrm{lag}}$		TE		GC
Dataset	Source	CD_g (pp)	$\hat{\ell}$	r	$\hat{\ell}$	r	r
independent	none	reference	no source flagged				
contemp-linear	c_1	$+1.5 \pm 0.5$	0	14	24	11	1.5
lagged-linear	c_1	$+25.8 \pm 0.4$	24	13	24	74	51
lagged-non-linear	c_1	$+4.8 \pm 0.6$	12	17	12	14	< null
multiplicative-joint	c_1	$+5.2 \pm 2.5$	6	3.2	6	4.3	1.1
	c_2		6	3.5	6	4.1	< null
additive-mixed	c_1	$+24.1 \pm 0.2$	24	16	24	73	41
	c_2		8	9.2	8	7.9	< null

Table 3: Channel dependency of the standard datasets and of the chaotic ODE benchmark. $\mathrm{MI}_{\mathrm{lag}}$: share of channel pairs coupled more strongly at a lag than at lag 0. CD_g: test-MSE reduction of TSMixer over $\mathrm{TSMixer}_{\mathrm{CI}}$ (Eq. 5), with MSE averaged over horizons and 3 seeds (Table 13, ODE MSEs from Abdelmalak et al. (2026)). **Bold**: positive CD_g and the better TSMixer variant.

Dataset	$\mathrm{MI}_{\mathrm{lag}}$ (%)	CD_g (%)	TSMixer	$\mathrm{TSMixer}_{\mathrm{CI}}$
Standard MTSF				
ETTh1	19	−14.9	0.498	**0.433**
ETTh2	19	−5.4	0.411	**0.390**
ETTm1	38	−4.8	0.369	**0.352**
ETTm2	38	−2.0	0.271	**0.266**
Weather	52	−1.0	0.235	**0.232**
Electricity	61	−5.1	0.166	**0.158**
Traffic	27	−9.2	0.427	**0.391**
ExchangeRate	7	−1.3	0.154	**0.152**
ILI	17	−6.3	1.992	**1.873**
SolarAL	14	−4.0	0.200	**0.193**
category median	*23*	*−4.9*		
Chaotic ODE (known-correlated reference)				
Lorenz	83	**+1.4**	**0.868**	0.880
Lorenz-Coupled	63	**+5.3**	**0.900**	0.950
Blinking-Rotlet	83	**+16.0**	**0.487**	0.580
CellCycle	67	**+2.7**	**0.771**	0.792
Double-Pendulum	92	**+4.0**	**0.737**	0.768
Hopfield	73	**+14.2**	**0.435**	0.507
category median	*78*	*+4.6*		

picture from the model-based measurement perspective. $\mathrm{TSMixer}_{\mathrm{CI}}$ outperforms TSMixer on all 10 standard datasets (median CD_g of -4.9%), while the ODE datasets reach a median of $+4.6\%$ and up to $+16.0\%$ (Blinking-Rotlet). We therefore conclude that the current standard MTSF datasets lack exploitable cross-channel coupling. Discarding the other channels does not hurt forecasting on them, and even consistently improves it. This raises the question of whether we can find real-world datasets with richer cross-channel coupling, where multivariate models can actually benefit from channel mixing.

Table 4: The coupled datasets. *Timesteps*: series length, pooled over the independent series for the multi-series panels noted in *Domain*. *Granularity*: sampling interval.

Dataset	Channels	Timesteps	Granularity	Domain
AirQuality	132	35,064	1 hour	Environment
AQShunyi	11	35,064	1 hour	Environment
CzeLan	11	19,934	30 min	Environment
ZafNoo	11	19,225	30 min	Environment
Bavaria-Iller	4	59,502	15 min	Hydrology
Bavaria-Donau	11	31,250	15 min	Hydrology
WaterQuality-Darwin	6	106,606	15 min	Marine (7 sites)
Current-Velocity	6	174,425	20 min	Ocean (27 deployments)
Wearable	24	38,913	5 min	Health
HouseholdPower	7	138,352	15 min	Energy

7 THE MIXBENCH-TS BENCHMARK

7.1 DATASETS

We propose MixBench-TS, a benchmark of 10 real-world datasets collected from recent MTSF benchmarks and public data repositories. We refer to these datasets as *the coupled datasets*. AQShunyi, CzeLan, and ZafNoo are taken from TFB (Qiu et al., 2024), WaterQuality-Darwin and Current-Velocity from TIME (Qiao et al., 2026), and Bavaria-Iller and Bavaria-Donau from CausalRivers (Stein et al., 2025). Wearable is taken from Baigutanova et al. (2025), AirQuality from the Beijing multi-site air-quality data (Zhang et al., 2017), and HouseholdPower from the UCI household power consumption data (Hebrail & Berard, 2006). Although publicly available, these datasets are rarely used for MTSF evaluation. Since our aim is to complement the standard datasets with datasets that have exploitable cross-channel coupling, we screen candidate datasets with CD_g (Section 5) and retain only those on which channel mixing pays off ($\mathrm{CD}_g > 0$). $\mathrm{MI}_{\mathrm{lag}}$ is not used for this selection and only characterizes the type of coupling. They cover a wide range of domains (environmental monitoring, energy, hydrology, marine sensing, and health), numbers of channels (4 to 132), series lengths (19,225 to 174,425 timesteps), and sampling intervals (5 min to 1 hour), as summarized in Table 4. The seven datasets on which lagged coupling also dominates ($\mathrm{MI}_{\mathrm{lag}} > 50\%$) are exactly those with the highest CD_g ($\geq +1.1\%$). We call them ***MixBench-TS-Core***, a stricter subset that we consider the strongest testbed for multivariate models. We additionally report two spatio-temporal traffic datasets (METR-LA and PEMS-BAY) (Li et al., 2018) in Appendix B.

7.2 EXPERIMENTAL PROTOCOL

We compare six state-of-the-art models covering both channel strategies (Section 2): DLinear (Zeng et al., 2023), $\mathrm{TSMixer}_{\mathrm{CI}}$ (Chen et al., 2023), and CycleNet (Lin et al., 2024) as CI models, and TSMixer (Chen et al., 2023), iTransformer (Liu et al., 2024), and SimpleTM (Chen et al., 2025a) as CD models. Since insufficient tuning can distort model comparisons (Brigato et al., 2026), we tune every model on every dataset with Optuna's TPE sampler (Akiba et al., 2019; Bergstra et al., 2011) for 20 trials. We use the search space of the original paper or, if unavailable, a range around its default parameters (Appendix E). Following Abdelmalak et al. (2026), we additionally tune the lookback window. Candidate lookback windows and forecast horizons are chosen per dataset based on its sampling interval (Appendix A). All models are trained with the MSE loss, selected by validation MSE, and evaluated by test MSE and MAE, averaged over horizons and three seeds.

7.3 RESULTS

Table 5 combines the two dependency measures with the forecasting results on the coupled datasets. By construction, channel mixing helps on all of them (median CD_g of $+1.7\%$ vs. -4.9% on the standard datasets), and they have a median of 55.5% lagged-coupled pairs, compared to 23%. The winning model class follows the two measures. The four datasets with the highest CD_g (Bavaria-Donau, Bavaria-Iller, Current-Velocity, and CzeLan) all combine a high share of lagged coupling

Table 5: Channel dependency and test MSE on the coupled datasets, sorted by CD_g. $\mathrm{MI}_{\mathrm{lag}}$ and CD_g as in Table 3. MSE averaged over horizons and 3 seeds, best per row **bold**, second underlined. Win: channel strategy of the best model. **CI**: $\mathrm{TSM}_{\mathrm{CI}}$ = $\mathrm{TSMixer}_{\mathrm{CI}}$, DLin. = DLinear, CycNet = CycleNet. **CD**: TSM = TSMixer, SimTM = SimpleTM, iTr. = iTransformer. Wins: number of best rows (ties counted for each). Above the line: MixBench-TS-Core ($\mathrm{MI}_{\mathrm{lag}} > 50\%$). All datasets in Table 13.

			CI			CD			
Dataset	$\mathrm{MI}_{\mathrm{lag}}$ (%)	CD_g (%)	$\mathrm{TSM}_{\mathrm{CI}}$	DLin.	CycNet	TSM	SimTM	iTr.	Win
Bavaria-Donau	78	+11.1	0.097	0.095	0.098	**0.086**	**0.086**	0.100	CD
Bavaria-Iller	67	+3.4	0.185	0.176	0.186	0.179	**0.170**	0.186	CD
Current-Velocity	83	+3.1	0.263	0.264	0.270	0.255	**0.254**	0.265	CD
CzeLan	82	+3.0	0.212	0.218	0.216	**0.205**	0.211	0.225	CD
ZafNoo	57	+1.8	0.523	**0.486**	0.528	0.514	0.529	0.539	CI
AirQuality	54	+1.5	0.750	**0.738**	0.744	0.739	0.752	0.748	CI
AQShunyi	53	+1.1	0.684	0.672	0.682	0.677	**0.664**	0.697	CD
WaterQuality-Darwin	40	+0.8	0.270	0.281	0.271	**0.268**	0.275	0.280	CD
Wearable	17	+0.7	0.671	0.668	0.781	**0.666**	0.753	0.755	CD
HouseholdPower	17	+0.6	0.652	**0.644**	0.646	0.648	0.648	0.646	CI
median / wins	*55.5*	*+1.7*	0	3	0	**4**	**4**	0	

(67–83%) with a CD model winning. Overall, CI models win only 3 of the 10 coupled datasets by MSE and 2 of 10 by MAE, compared to 10 of 10 and 8 of 10 standard datasets (Tables 13 and 14).

There are, however, a few outliers. On ZafNoo, AirQuality, and HouseholdPower, channel mixing improves TSMixer (+0.6 to +1.8%), yet DLinear achieves the lowest MSE, suggesting that the benefit of channel mixing is smaller than the differences between architectures there. Conversely, TSMixer wins on Wearable and WaterQuality-Darwin, although both fall outside MixBench-TS-Core. Finally, DLinear remains a strong baseline, which indicates that more complex architectures do not yet fully exploit these datasets and further motivates including them in MTSF benchmarks.

8 Conclusions and Future Work

In this work, we asked two questions. First, how can we reliably measure lagged, non-linear, and joint coupling in MTSF datasets? On synthetic datasets with known ground-truth coupling, only lagged MI ($\mathrm{MI}_{\mathrm{lag}}$) and CD gain ($\mathrm{CD}_g$) retrieve all planted couplings. Second, do the standard MTSF datasets have such coupling? The answer is no. They have a median of only 23% lagged-coupled channel pairs compared to 78% in the chaotic ODE benchmark, and channel mixing hurts on every one of them. We therefore proposed MixBench-TS, a benchmark of 10 real-world coupled datasets for evaluating multivariate forecasting models. These datasets have a median of 55.5% lagged-coupled pairs and a positive CD_g, and CI models win on only 3 of them, compared to all 10 standard datasets. We recommend profiling further datasets with $\mathrm{MI}_{\mathrm{lag}}$ and CD_g. Those with $\mathrm{CD}_g > 0$ can join MixBench-TS, and those that also have $\mathrm{MI}_{\mathrm{lag}} > 50\%$ can join MixBench-TS-Core. Evaluating multivariate models requires benchmarks that actually contain multivariate structure, so that the benefits of channel mixing can be observed rather than assumed.

Limitations. First, CD_g uses a single model family (TSMixer) and thus also reflects whether this architecture can exploit the coupling, e.g., Electricity has 61% lagged pairs but a CD_g of -5.1%. Second, $\mathrm{MI}_{\mathrm{lag}}$ is a pairwise fraction that ignores coupling strength, captures joint coupling only indirectly, and uses at most 40 channels on wide datasets. Third, our synthetic suite has only four channels and fixed couplings, without the regime changes and hidden common drivers of real data.

Future Work. The landmarking concept behind CD_g can be extended beyond TSMixer to further CD models and their CI variants, making CD_g less dependent on a single architecture. Per-channel and time-windowed coupling maps could guide selective channel mixing and capture couplings that change across regimes. Finally, our criteria can broaden MixBench-TS with short-and-wide, spatio-temporal, and irregularly sampled datasets.

AI USE STATEMENT

We used generative AI tools for three purposes. First, we used them to polish the writing. Second, we used them to draft parts of individual sections, which we then revised over several rounds of proofreading. Third, we used them to help search for related work. We verified every retrieved reference and took care that no closely related work is omitted. We reviewed and validated all content of the final version and take full responsibility for it.

REPRODUCIBILITY STATEMENT

We describe all components needed to reproduce our results. The datasets, their sources, splits, horizons, and lookback windows are given in Section 7.1 and Appendices A and C. The coupling measures are defined in Section 4, with implementation details in Appendix D. The synthetic data generation, including all generation parameters, is described in Appendix J. The models, the tuning protocol, and all hyperparameter search ranges are given in Section 7.2 and Appendix E. All forecasting results are averaged over three seeds (2021, 2025, and 2026), with standard deviations in Appendix H. Our anonymized repository (`https://anonymous.4open.science/r/mixbench-ts-B027`) contains the code for all experiments, the profiling toolkit, and the synthetic data generator, together with instructions for obtaining each dataset and running the experiments.

REFERENCES

Ibram Abdelmalak, Kiran Madhusudhanan, Jungmin Choi, Christian Klötergens, Vijaya Krishna Yalavarthi, Maximilian Stubbemann, and Lars Schmidt-Thieme. Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 585–597. Springer, 2026. URL https://doi.org/10.1007/978-981-92-1462-4_46.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019. URL https://doi.org/10.1145/3292500.3330701.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024. URL https://arxiv.org/abs/2410.10393.

Aitolkyn Baigutanova, Sungkyu Park, Marios Constantinides, Sang Won Lee, Daniele Quercia, and Meeyoung Cha. A continuous real-world dataset comprising wearable-based heart rate variability alongside sleep diaries. *Scientific Data*, 12(1):1474, 2025. URL https://doi.org/10.1038/s41597-025-05801-3.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2546–2554, 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Lorenzo Brigato, Rafael Morand, Knut Strømmen, Maria Panagiotou, Markus Schmidt, and Stavroula Mougiakakou. There are no champions in supervised long-term time series forecasting. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=yO1JuBpTBB.

Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=oANkBaVci5.

Jialin Chen, Jan Eric Lenssen, Aosong Feng, Weihua Hu, Matthias Fey, Leandros Tassiulas, Jure Leskovec, and Rex Ying. From similarity to superiority: Channel clustering for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130635–130663, 2024. URL https://openreview.net/forum?id=MDgn9aazo0.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=wbpxTuXgm0.

Yu Chen, Nathalia Céspedes, and Payam Barnaghi. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=kHEVCfES4Q.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. URL https://doi.org/10.1002/047174882X.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313, 1982. URL https://doi.org/10.1080/01621459.1982.10477803.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=enYjtbjYJrf.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=I01l7rc0jcb.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. URL https://doi.org/10.2307/1912791.

Satyandra Guthula, Jaber Daneshamooz, Charles Fleming, Ashish Kundu, Walter Willinger, and Arpit Gupta. Netburst: Event-centric forecasting of bursty, intermittent time series. *arXiv preprint arXiv:2510.22397*, 2025. URL https://arxiv.org/abs/2510.22397.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7129–7142, 2024. URL https://doi.org/10.1109/TKDE.2024.3400008.

Georges Hebrail and Alice Berard. Individual household electric power consumption. UCI Machine Learning Repository, 2006. URL https://doi.org/10.24432/C58K54.

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=490VcNtjh7.

Christian Klötergens, Vijaya Krishna Yalavarthi, Randolf Scholz, Maximilian Stubbemann, Stefan Born, and Lars Schmidt-Thieme. Physiome-ode: A benchmark for irregularly sampled multivariate time-series forecasting based on biological odes. In *International Conference on Learning Representations*, volume 2025, pp. 34961–34988, 2025. URL https://openreview.net/forum?id=6ouZaBzeNO.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004. URL https://doi.org/10.1103/PhysRevE.69.066138.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018. URL https://doi.org/10.1145/3209978.3210006.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bfe7998398779dde03cad7a73b1f81b6-Abstract-Conference.html.

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. URL https://openreview.net/forum?id=JePfAI8fah.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Robert Marschinski and Holger Kantz. Analysing the information flow between financial time series: An improved estimator for transfer entropy. *The European Physical Journal B*, 30(2):275–281, 2002. URL https://doi.org/10.1140/epjb/e2002-00379-2.

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. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Bernhard Pfahringer, Hilan Bensusan, and Christophe G Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 743–750. Morgan Kaufmann, 2000. URL https://dl.acm.org/doi/10.5555/645529.658105.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's TIME: Towards the next generation of time series forecasting benchmarks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=79TgfXHbsK.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9): 2363–2377, 2024. URL https://doi.org/10.14778/3665844.3665863.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1185–1196, 2025. URL https://doi.org/10.1145/3690624.3709325.

Luis Roque, Vitor Cerqueira, Carlos Soares, and Luis Torgo. Cherry-picking in time series forecasting: How to select datasets to make your model shine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20192–20199, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34224.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 938–947. PMLR, 2018. URL https://proceedings.mlr.press/v84/runge18a.html.

Said E Said and David A Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984. URL https://doi.org/10.1093/biomet/71.3.599.

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000. URL https://doi.org/10.1103/PhysRevLett.85.461.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, Gao Cong, Christian S. Jensen, and Xueqi Cheng. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):291–305, 2025. URL https://doi.org/10.1109/TKDE.2024.3484454.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. CausalRivers - scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wmV4cIbgl6.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1CLzLXSFNn.

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*, volume 34, pp. 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Siqiao Xue, Zhaoyang Zhu, Wei Zhang, Rongyao Cai, Rui Wang, Yixiang Mu, Fan Zhou, Jianguo Li, Peng Di, and Hang Yu. QuitoBench: A high-quality open time series forecasting benchmark. *arXiv preprint arXiv:2603.26017*, 2026. URL https://arxiv.org/abs/2603.26017.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. OLinear: A linear model for time series forecasting in orthogonally transformed domain. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60573–60632, 2025. URL https://openreview.net/forum?id=DAyKP1tvwI.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023. URL https://doi.org/10.1609/aaai.v37i9.26317.

S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017. URL https://doi.org/10.1098/rspa.2017.0457.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

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*, volume 35, pp. 11106–11115, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

A Datasets - More Details

All forecasting horizons H and lookback windows L are given in timesteps of each dataset's sampling interval, and the lookback is tuned jointly with the other hyperparameters (Appendix E).

Data splits. For the ETT datasets, we follow the standard protocol of 12/4/4 months for training, validation, and testing. All other single-series datasets are split chronologically into 70/10/20%. For the multi-series datasets (WaterQuality-Darwin and Current-Velocity), each series is split 70/10/20% in time, and the windows of all series are pooled. Wearable is split at the subject level (70/15/15%, i.e., 34/7/8 of 49 subjects, assigned at random with a fixed seed), and no window spans two subjects. Validation and test windows may use lookback values from the preceding split, but their forecast targets always lie within their own split. Channels are standardized with training-set statistics, and all metrics are reported on the standardized scale.

Standard real-world datasets. For comparability with prior work, we follow the established long-term forecasting protocol for these datasets: $H \in \{96, 192, 336, 720\}$ with $L \in \{48, 96, 192, 336, 720\}$, except for ExchangeRate ($H \in \{30, 90, 180, 365\}$, $L \in \{30, 60, 90, 180, 365\}$) and ILI ($H \in \{24, 36, 48\}$, $L \in \{12, 24, 36, 52, 104\}$).

- **ETTh1, ETTh2, ETTm1, ETTm2** (7 channels; hourly and 15-minute). Six power-load features (high, middle and low useful and useless load) and the oil temperature of two electricity transformers in China (2016–2018).
- **Weather** (21 channels; 10-minute). Meteorological indicators (pressure, temperatures, humidity, wind, precipitation, radiation) of a weather station in Jena, Germany.
- **Electricity** (321 channels; hourly). Electricity consumption of 321 clients (2012–2014).
- **Traffic** (862 channels; hourly). Road occupancy rates of 862 freeway sensors in the San Francisco Bay Area (2015–2016).
- **ExchangeRate** (8 channels; daily). Daily exchange rates of eight currencies (1990–2016).
- **ILI** (7 channels; weekly). Influenza-like-illness surveillance indicators of the US CDC (weighted and unweighted ILI rates, patient counts by age group, number of providers).
- **SolarAL** (137 channels; 10-minute). Solar power output of 137 photovoltaic plants in Alabama (2006).

Spatio-temporal (traffic) datasets.

- **METR-LA** (207 channels) and **PEMS-BAY** (325 channels; both 5-minute). Traffic speed from loop detectors on the highways of Los Angeles (2012) and the San Francisco Bay Area (2017). $L \in \{48, 96, 192, 288\}$ (4 h to 1 day), $H \in \{12, 48, 144, 288\}$, i.e. 1 h (the standard short-term traffic horizon), 4 h, 12 h and one full day, which spans the morning and evening peaks.

Coupled real-world datasets. Five of these datasets are part of recent forecasting benchmarks: AQShunyi, CzeLan and ZafNoo of TFB (Qiu et al., 2024), whose horizons $H \in \{96, 192, 336, 720\}$ we adopt, and WaterQuality-Darwin and Current-Velocity of TIME (Qiao et al., 2026), which targets zero-shot evaluation of foundation models. For the remaining horizons and all lookback windows, the choices cover each dataset's dominant natural cycles, from within-cycle to multi-cycle forecasts, so that each horizon corresponds to an operationally meaningful lead time.

- **AirQuality** (132 channels; hourly). Eleven air-quality and meteorological variables (PM2.5, PM10, SO_2, NO_2, CO, O_3, temperature, pressure, dew point, rain, wind speed) at each of 12 monitoring stations in Beijing (2013–2017). The stations share pollution episodes and weather. $L \in \{48, \ldots, 720\}$, $H \in \{96, 192, 336, 720\}$ (4 to 30 days), covering the multi-day build-up and dispersal of pollution episodes.
- **AQShunyi** (11 channels; hourly). The same eleven variables at the single Shunyi station, where pollutants are coupled with each other and with the local meteorology. $L \in \{24, 48, 168, 336, 720\}$ (1 day to 30 days), H as for AirQuality.
- **CzeLan** and **ZafNoo** (11 channels each; 30-minute). The sap flow of a single tree together with its environmental drivers (incoming short-wave, extraterrestrial and photosynthetically active radiation, air temperature, relative humidity, vapor-pressure deficit, precipitation, wind speed, shallow and deep soil water content), at a site in the Czech Republic and one in South Africa. Transpiration responds to radiation and vapor-pressure deficit with

a lag, and to soil moisture on longer scales. $L \in \{48, 96, 336, 672\}$ (1 day to 2 weeks), $H \in \{96, 192, 336, 720\}$ (2 to 15 days), i.e. from a few diurnal cycles to about two weeks of drying or rewetting.

- **Bavaria-Iller** (4 channels) and **Bavaria-Donau** (11 channels; both 15-minute). River discharge at gauging stations along the Iller and Danube river networks in Bavaria, where flow at a gauge propagates to the gauges further along the network with a travel-time delay. $L \in \{48, 96, 192, 336, 720\}$ (12 h to 7.5 days), $H \in \{24, 96, 336\}$, i.e. 6 h, 1 day and 3.5 days: the lead times of short-term flood warning, next-day operation and multi-day flood-wave routing.
- **WaterQuality-Darwin** (6 channels; 15-minute; 7 series). Coastal water quality at the Darwin national reference station, Australia: electrical conductivity, dissolved oxygen, salinity, temperature, turbidity and chlorophyll, which are coupled through tidal mixing and biological activity. $L \in \{48, 96, 192, 336\}$ (12 h to 3.5 days), $H \in \{24, 96\}$, i.e. 6 h (about half a semi-diurnal tidal cycle) and 1 day (two tidal and one diurnal cycle).
- **Current-Velocity** (6 channels; 20-minute; 27 series). ADCP mooring records from the Australian national mooring network: zonal, meridional and vertical current velocity, water temperature, relative pressure and acoustic backscatter intensity. Currents and pressure are coupled through the tides. $L \in \{48, 96, 192, 336\}$ (16 h to 4.7 days), $H \in \{24, 96\}$, i.e. 8 h (within one semi-diurnal tidal cycle) and 32 h (more than two tidal cycles).
- **Wearable** (24 channels; 5-minute). Wrist-worn sensor features of free-living participants: heart rate, inter-beat interval, heart-rate-variability statistics (SDNN, SDSD, RMSSD, pNN20, pNN50, LF, HF, LF/HF), accelerometer, gravity and gyroscope averages, steps, distance, calories and ambient light. Physical activity drives the cardiac channels. $L \in \{48, 96, 192, 288, 576\}$ (4 h to 2 days), $H \in \{12, 48, 144, 288\}$, i.e. 1 h, 4 h, 12 h and one day: from the next activity bout to the full circadian cycle.
- **HouseholdPower** (7 channels; 15-minute). Electricity consumption of a single household (2006–2010): global active and reactive power, voltage, current intensity and three sub-meterings (kitchen, laundry, water heater and air conditioning), where the total load is the sum of its components. $L \in \{48, 96, 192, 336, 672\}$ (12 h to 1 week), $H \in \{96, 192, 336, 672\}$, i.e. 1, 2, 3.5 and 7 days: from the daily usage cycle to the weekly one.

Chaotic ODE systems. The six systems are taken from the `dysts` forecasting benchmark of Gilpin (2021) and were used as a known-correlated reference for MTSF by Abdelmalak et al. (2026), whose setup we follow and whose TSMixer and TSMixer$_{\text{CI}}$ test MSEs we report in Table 3. Each system is simulated over $t \in [0, 3000]$ with 60,000 samples ($\Delta t \approx 0.05$), and its channels are the state variables of the system (Table 6). Since the coupling is given by the governing equations, every channel depends on the others by construction. $L \in \{48, 96, 192, 336, 720\}$ and $H \in \{96, 192, 336, 720\}$, as for the standard datasets.

- **Lorenz** (3 channels). A minimal weather model based on atmospheric convection (Lorenz-63).
- **Lorenz-Coupled** (6 channels). Two coupled Lorenz oscillators.
- **Blinking-Rotlet** (3 channels). Chaotic advection by a blinking rotlet mixer, a model of fluid mixing, in polar coordinates.
- **Double-Pendulum** (4 channels). Two coupled rigid pendula without damping.
- **CellCycle** (6 channels). A simplified model of the cell cycle.
- **Hopfield** (6 channels). A Hopfield neural network with frustrated connectivity.

B Spatio-Temporal Datasets

In addition to the standard and coupled datasets, we evaluate two spatio-temporal traffic datasets (Li et al., 2018): METR-LA (207 channels, 34,272 timesteps) and PEMS-BAY (325 channels, 52,116 timesteps), both sampled every 5 minutes. Since their channels correspond to road sensors of the

Table 6: The chaotic ODE systems, used as a known-correlated reference in Table 3. Columns as in Table 4, with *Granularity* the integration step Δt and *Domain* the modeled system.

System	Channels	Timesteps	Granularity	Domain
Lorenz	3	60,000	Δt=0.05	Chaos
Lorenz-Coupled	6	60,000	Δt=0.05	Coupled chaos
Blinking-Rotlet	3	60,000	Δt=0.05	Fluid mixing
Double-Pendulum	4	60,000	Δt=0.05	Mechanics
CellCycle	6	60,000	Δt=0.05	Biology
Hopfield	6	60,000	Δt=0.05	Neural

same traffic network, cross-channel (spatial) coupling is expected. However, they are neither part of the standard datasets nor of our proposed coupled datasets, which is why we report them separately. Table 7 shows that both datasets together have a similar median share of lagged coupling as the coupled datasets (53.5% vs. 55.5%), well above the standard datasets (23%), but a mixed CD_g: channel mixing helps on PEMS-BAY ($+6.4\%$) but not on METR-LA (-1.5%). The forecasting results in Table 13 follow the same pattern, with a CD model winning on PEMS-BAY and a CI model winning on METR-LA in terms of MSE. The corresponding MAE results are given in Table 14 and the per-horizon results in Table 18.

Table 7: Channel dependency of the spatio-temporal datasets. Columns as in Table 3.

Dataset	$\mathrm{MI}_{\mathrm{lag}}$ (%)	CD_g (%)	TSMixer	$\mathrm{TSMixer}_{\mathrm{CI}}$
METR-LA	38	-1.5	0.930	**0.916**
PEMS-BAY	69	$+\mathbf{6.4}$	**0.469**	0.501
category median	*53.5*	*+2.4*		

C Problem Formulation - Detailed

Each dataset is a multivariate time series $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T) \in \mathbb{R}^{T \times C}$ observed at T equally spaced steps, where C is the number of channels (variates) and $\mathbf{z}_t \in \mathbb{R}^C$ represents all channel values at step t. $\mathbf{Z}$ is split *chronologically* into three contiguous segments: training $\mathbf{Z}_{\mathrm{tr}} = (\mathbf{z}_1, \ldots, \mathbf{z}_{T_{\mathrm{tr}}})$, validation $\mathbf{Z}_{\mathrm{val}} = (\mathbf{z}_{T_{\mathrm{tr}}+1}, \ldots, \mathbf{z}_{T_{\mathrm{val}}})$, and test $\mathbf{Z}_{\mathrm{te}} = (\mathbf{z}_{T_{\mathrm{val}}+1}, \ldots, \mathbf{z}_T)$. The boundaries $T_{\mathrm{tr}} = \lfloor \rho_{\mathrm{tr}} T \rfloor$ and $T_{\mathrm{val}} = \lfloor (\rho_{\mathrm{tr}} + \rho_{\mathrm{val}}) T \rfloor$ are set by the split fractions $\rho_{\mathrm{tr}} + \rho_{\mathrm{val}} + \rho_{\mathrm{te}} = 1$, so that every training step precedes every validation step, which in turn precedes every test step. We write $\mathcal{T}_{\mathrm{tr}} = \{L, \ldots, T_{\mathrm{tr}} - H\}$, $\mathcal{T}_{\mathrm{val}} = \{T_{\mathrm{tr}}, \ldots, T_{\mathrm{val}} - H\}$, and $\mathcal{T}_{\mathrm{te}} = \{T_{\mathrm{val}}, \ldots, T - H\}$ for the corresponding sets of time anchors, which represent the indices of the different time series samples where L and H are the *lookback window* and *forecast horizon* lengths, respectively.

Supervised samples are constructed by sliding a fixed-length window over the series (we always use a stride of 1 throughout our experiments). A sample anchored at step t splits into a lookback window and a forecast horizon as follows,

$$\mathbf{X}^{(t)} = (\mathbf{z}_{t-L+1}, \ldots, \mathbf{z}_t) \in \mathbb{R}^{L \times C}, \qquad \mathbf{Y}^{(t)} = (\mathbf{z}_{t+1}, \ldots, \mathbf{z}_{t+H}) \in \mathbb{R}^{H \times C}. \tag{7}$$

We index each sample by its *anchor* t, the last observed step, so that the lookback spans steps $t - L + 1, \ldots, t$ and the forecast spans $t + 1, \ldots, t + H$. Each sample is assigned to the split whose anchor set $\mathcal{T}_{\mathrm{split}}$ contains it, giving

$$\mathcal{D}_{\mathrm{split}} = \{\, (\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) : t \in \mathcal{T}_{\mathrm{split}} \,\}, \qquad \mathrm{split} \in \{\mathrm{tr}, \mathrm{val}, \mathrm{te}\}, \tag{8}$$

with the anchor ranges chosen so that every lookback fits within the series and every forecast horizon stays inside its split. Consequently, forecast targets never cross a split boundary, so no future information leaks into training or evaluation, even though a lookback window may reach back into the preceding segment, as is standard in the forecasting literature. A model $f_{\boldsymbol{\theta}} : \mathbb{R}^{L \times C} \to \mathbb{R}^{H \times C}$ predicts $\widehat{\mathbf{Y}}^{(t)} = f_{\boldsymbol{\theta}}(\mathbf{X}^{(t)})$ jointly over all C channels, and its parameters are fit by minimizing the empirical training loss $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}_{\mathrm{tr}}|} \sum_{(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}) \in \mathcal{D}_{\mathrm{tr}}} \ell(\widehat{\mathbf{Y}}^{(t)}, \mathbf{Y}^{(t)})$ for a per-sample loss ℓ (e.g.

mean squared error). The MTSF benchmarks commonly use a horizon $H \in \{96, 192, 336, 720\}$. However, as we show in Appendix A, we adapt these horizons to be of semantic relevance for the remaining datasets.

D Statistical Profiling - Implementation Details

In this appendix, we provide more detailed implementation details which should be helpful in both reproducing the results and understanding our design choices for these metrics. For datasets with more than 40 channels, all three pairwise measures are computed on a subset of 40 channels drawn by uniform random sampling without replacement. A sensitivity analysis over this subsampling is reported in Appendix G.

D.1 Granger Causality

Granger Causality was originally proposed by Granger (1969) and has been widely used in econometrics and time series analysis. The basic idea is that if a time series X Granger-causes another time series Y, then past values of X should contain information that helps predict Y beyond the information contained in past values of Y alone. What makes this method more challenging to apply in practice are the following factors:

- **Stationarity**: Granger Causality assumes that the time series are stationary. In practice, many real-world time series exhibit trends and seasonality, which can violate this assumption. To address this, we apply seasonal differencing and regular differencing to make the series stationary before applying Granger Causality tests.
- **Lag Selection**: The choice of lag length is crucial in Granger Causality analysis. For comparability with deep learning models, we set the VAR order p equal to the smallest candidate lookback window of the dataset (the order p of the main text). This allows us to capture the relevant temporal dependencies without introducing unnecessary complexity.

The seasonal period P is detected once per dataset, before any channel subsampling, as the strongest local peak of the autocorrelation function averaged over all detrended channels. It therefore does not depend on the channel subset, and both series of a pair receive the same seasonal difference at period P. For the multi-series panels, we use a fixed period instead ($P = 96$ for WaterQuality-Darwin and $P = 72$ for Current-Velocity), and the synthetic lag profiles use $P = 24$. The seasonal difference is skipped if the series is shorter than $3P$. We then use the Augmented Dickey-Fuller (ADF) (Said & Dickey, 1984) test to determine if additional regular differencing is needed. The ADF test checks for the presence of a unit root in the time series, which indicates non-stationarity. For each pair, the ADF test selects a differencing order $d \leq 2$ separately for the source and the target, and both series are then differenced $\max(d_{\mathrm{src}}, d_{\mathrm{tgt}})$ times. This common order keeps the two series aligned and equally transformed.

With these settings fixed, both models are fitted with Ordinary Least Squares (OLS) on the preprocessed series. The sums of squared residuals (SSR) of the restricted and multivariate models are then used to compute the degrees-of-freedom corrected partial R^2 (Geweke linear feedback (Geweke, 1982)), which quantifies the linear directed predictability of one time series from another. As with TE, this per-pair R^2 is aggregated to a single dataset value by taking the median over all directed channel pairs, as reported in the main text.

D.2 Transfer Entropy (TE)

Now regarding the Transfer Entropy, we use the same joint differencing as Granger Causality, so the linear and non-linear axes see identical series. The main idea of TE is to measure the amount of directed information between two time series, capturing non-linear coupling that GC might miss. TE is computed based on KSG estimator (Kraskov et al., 2004) quantifying the conditional mutual information between the present state of one time series and the past state of another, conditioned on the target's own most recent value. Conditioning on a single past value keeps the conditioning set low-dimensional, which the k-nearest-neighbor estimator requires to remain reliable. Computing the CMI on the ranks of the series (rather than their raw values) makes the estimate invariant to

monotone transformations and robust to outliers, while the non-parametric estimator captures non-linear coupling of any form. Unlike GC, TE does not directly produce a p-value, so we use a surrogate data approach to generate a null distribution of TE values. Following Runge (2018), we cast TE as the conditional independence test $X \perp Y \mid Z$ with source $X = x_{t-\ell}$, target $Y = y_t$, and conditioning set $Z = y_{t-1}$ (the target's own past). Rather than a full random shuffle, the surrogates are then drawn with a *local-permutation* scheme: each source sample is permuted only among its k_{perm} nearest neighbors in the conditioning space Z, which breaks the $X \to Y$ information flow while preserving the marginal distributions and the dependence on Z. By comparing the observed TE value to this null distribution, we can assess the statistical significance of the detected information flow.

The test has a few essential parameters, which we set as follows:

- **TE Max Lag ($\ell^{\mathrm{TE}}_{\max}$)**: We set the maximum lag $\ell^{\mathrm{TE}}_{\max}$ for TE computation to be equal to the smallest candidate lookback window of the dataset (the order p of GC). This is done to make sure that every coupling identified by TE is relevant to the forecasting task and not just a coupling in a lag that the model is not able to use.
- **Number of Surrogates**: We generate 100 surrogate time series for each pair of time series to create a robust null distribution. This number is taken directly from the literature (Runge, 2018) and is a good trade-off between computational cost and statistical power (as a higher number of surrogates does not yield higher stability in results anymore).
- k_{CMI}: The number of nearest neighbors used in the KSG estimator of the CMI. This value also followed the original paper's rule of thumb of picking the k value based on the number of usable samples n in the dataset, $k_{\mathrm{CMI}} = \min(0.2\,n, 200)$. Scaling k with n lowers the estimator's variance on long series, and the cap bounds the computational cost.
- k_{perm}: The number of nearest neighbors (in the conditioning space Z) used by the local-permutation scheme for the surrogate null distribution. This is set universally for all datasets as 5, which was shown to work well for a variety of the datasets in the original paper (Runge, 2018).
- **Lag Sweep & Peak Selection ($\ell^\star$)**: For each ordered pair we evaluate the effective TE at every candidate lag $\ell \in \{1, \ldots, \ell^{\mathrm{TE}}_{\max}\}$. The reported value is the *effective* (debiased) transfer entropy $\mathrm{TE}^{\mathrm{eff}} = \max\big(0,\, \widehat{I}(y_t; x_{t-\ell} \mid y_{t-1}) - \overline{I}_{\text{null}}\big)$, which subtracts the surrogate-null mean $\overline{I}_{\text{null}}$ from the raw KSG estimate (the same debiasing applied to MI). We test each swept lag against its surrogate null and apply a Benjamini–Hochberg false-discovery-rate correction across the swept lags (Benjamini & Hochberg, 1995). This controls the expected proportion of lags falsely flagged as significant. The reported lag $\ell^\star$ is the FDR-significant lag with the largest effective TE, and the effective TE is set to 0 if no lag survives the correction. This is the sweep procedure referenced from the main text (Section 4).

Finally, based on the above preprocessing and parameter settings, we compute the TE values for all pairs of channels in each dataset and aggregate them into the median TE R^2 used in the main text (results in Appendix F).

D.3 Mutual Information (MI)

Now regarding the Mutual Information, unlike GC and TE it does not condition on the target's own past. At lag 0, it is a symmetric measure of the overall coupling shared between a pair of channels. At positive lags, it is directional, from source to target, as in GC and TE. It also differs in preprocessing: MI is estimated on the *raw* series (rank-transformed for scale invariance, *without* the seasonal and regular differencing applied for GC and TE), so that shared trend and seasonality are retained as legitimate common-driver information rather than removed. The estimator itself is the same KSG estimator used for TE (Kraskov et al., 2004), and, as with TE, every MI estimate is debiased against a surrogate null (see the **Number of Surrogates** parameter below). We mainly use MI to estimate whether cross-channel coupling is lagged or contemporaneous. We set its parameters as follows:

- k_{MI}: The number of nearest neighbors used in the KSG estimator for MI computation. We set this value to 5, a common choice for the KSG estimator (Kraskov et al., 2004).

- **Number of Surrogates**: Similar to TE, we generate 100 surrogate time series for each pair of channels to create a null distribution for the MI values, chosen at this value for consistency with TE. The surrogates are drawn by randomly permuting the source series, which destroys any coupling with the target while preserving its marginal distribution. Every reported MI value — both contemporaneous and lagged — is *effective*, i.e. debiased by subtracting the mean of this surrogate null from the raw KSG estimate. The peak-lag and lag-0 quantities that enter ϕ_{lag} are thus compared on the same debiased scale.

- **Lag Grid**: We evaluate MI in two setups, contemporaneous and lagged. The contemporaneous MI is computed between the present states of the two channels ($\ell = 0$), while the lagged MI is evaluated over the geometrically spaced candidate lags in Λ (Eq. 9) and retains the positive lag $\ell^\star \in \Lambda$ that maximizes the effective MI. This is the lag reported in the main text. The selection is thus driven by the MI values on this fixed lag grid, not by any cross-correlation criterion. Here is how the lag grid is defined for each dataset:

$$\Lambda \;=\; \{0\} \;\cup\; \Big\{ \big\lfloor \ell_{\mathrm{top}}^{\,i/(m-1)} \big\rceil \;:\; i = 0, 1, \ldots, m-1 \Big\}, \qquad \ell_{\mathrm{top}} \;=\; \min\!\big(\ell_{\max},\, \ell_{\mathrm{cap}}\big), \tag{9}$$

  where $m = 6$ is the number of positive-lag grid points, $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, $\ell_{\mathrm{cap}} = 300$ caps the largest lag, and $\ell_{\max}$ is the dataset-specific lag window. The resulting grid of each dataset is listed in Table 8. Duplicate integers introduced by rounding are collapsed, so Λ is a set. The geometric spacing concentrates grid points at short lags—where cross-channel coupling typically resides—while still reaching ℓ_{top}. For example, $\ell_{\mathrm{top}} = 300$ yields $\Lambda = \{0, 1, 3, 10, 31, 96, 300\}$ and $\ell_{\mathrm{top}} = 48$ yields $\Lambda = \{0, 1, 2, 5, 10, 22, 48\}$.

Table 8: Lag grid of the lagged-MI measure. For each dataset, MI is evaluated at lag 0 and at six log-spaced lags in $[1, \ell_{\mathrm{top}}]$, where $\ell_{\mathrm{top}} = \min(\ell_{\max}, \ell_{\mathrm{cap}})$ is the dataset's lag window capped at $\ell_{\mathrm{cap}} = 300$ steps († = capped). The synthetic grid is used for profiling only, while the retrieval in Section 5.2 uses the dense grid Λ_{syn}.

Dataset(s)	Sampling	ℓ_{top}	Lag grid (steps)	Span of ℓ_{top}
Standard real-world				
ETTh1, ETTh2, Electricity, Traffic	1 h	168	1, 3, 8, 22, 60, 168	1 week
ETTm1, ETTm2	15 min	300†	1, 3, 10, 31, 96, 300	3.1 days
Weather, SolarAL	10 min	168	1, 3, 8, 22, 60, 168	28 h
ExchangeRate	1 day	60	1, 2, 5, 12, 26, 60	2 months
ILI	1 week	52	1, 2, 5, 11, 24, 52	1 year
Spatio-temporal (traffic)				
METR-LA, PEMS-BAY	5 min	288	1, 3, 10, 30, 93, 288	1 day
Coupled real-world				
AirQuality, AQShunyi	1 h	168	1, 3, 8, 22, 60, 168	1 week
CzeLan, ZafNoo	30 min	300†	1, 3, 10, 31, 96, 300	6.25 days
Bavaria-Iller, Bavaria-Donau, HouseholdPower	15 min	300†	1, 3, 10, 31, 96, 300	3.1 days
WaterQuality-Darwin	15 min	192	1, 3, 8, 23, 67, 192	2 days
Current-Velocity	20 min	144	1, 3, 7, 20, 53, 144	2 days
Wearable	5 min	288	1, 3, 10, 30, 93, 288	1 day
Chaotic ODE systems	$\Delta t = 0.05$	50	1, 2, 5, 10, 23, 50	2.5 time units
Synthetic datasets	1 step	168	1, 3, 8, 22, 60, 168	168 steps

Finally, based on the above preprocessing and parameter settings, we compute the contemporaneous and lagged MI for all pairs of channels in each dataset. These are aggregated into the *lagged-coupling fraction* ϕ_{lag} defined in the main text, the fraction of pairs whose coupling is stronger at the peak lag $\ell^\star$ than at lag 0, since it directly separates lagged transfer from contemporaneous (potentially redundant) coupling.

E HYPERPARAMETER RANGES

Every (model, dataset, horizon) combination is tuned separately with Optuna's TPE sampler (20 trials), selecting the trial with the lowest validation MSE. All trials share the same training protocol: Adam, at most 30 epochs with early stopping on the validation MSE (patience 5), and a learning rate halved after every epoch. Table 9 lists the tuned hyperparameters of each model. Ranges that depend on the number of channels C of a dataset, which bound the memory footprint on wide datasets, are given in Table 10. The lookback window is tuned jointly with these hyperparameters, and its candidates are listed per dataset in Appendix A.

Table 9: Tuned hyperparameters and their search ranges. $\mathcal{B}$, $\mathcal{D}$, $\mathcal{H}$ and $N_{\max}$ are the channel-dependent ranges of Table 10. $\log\mathcal{U}$ denotes log-uniform and $\mathcal{U}$ uniform sampling.

Model	Hyperparameter	Range
All except CycleNet	learning rate	$\log\mathcal{U}[10^{-4}, 10^{-2}]$
All models	batch size	$\mathcal{B}$
TSMixer, TSMixer$_{\text{CI}}$	number of mixer blocks	$\{1, \dots, N_{\max}\}$
	hidden size of the mixing MLPs	$\mathcal{H}$
	dropout	$\mathcal{U}[0.01, 0.3]$
	per-channel output head	{yes, no}
DLinear	per-channel linear layers	{yes, no}
CycleNet	learning rate	$\{2\cdot10^{-3}, 5\cdot10^{-3}, 10^{-2}\}$
	backbone	{linear, MLP}
iTransformer	model dimension d_{model}	$\mathcal{D}$
	attention heads	$\{4, 8, 16\}$
	encoder layers	$\{1, 2, 3\}$
	dropout	$\mathcal{U}[0.01, 0.3]$
SimpleTM	model dimension d_{model}	$\mathcal{D}$
	encoder layers	$\{1, 2, 3\}$
	dropout	$\mathcal{U}[0.01, 0.3]$
	wavelet	{`db1`, `bior3.1`, `db4`}
	wavelet decomposition levels m	$\{1, 2, 3\}$
	geometric-attention weight α	$\mathcal{U}[0, 1]$
	geometric-attention dropout	$\mathcal{U}[0.1, 0.5]$
	L1 attention regularization weight	$\{0, 5\cdot10^{-5}, 5\cdot10^{-4}, 5\cdot10^{-3}\}$

Table 10: Channel-dependent search ranges (C = number of channels).

	$C \le 50$	$50 < C \le 200$	$200 < C \le 500$	$C > 500$
batch size $\mathcal{B}$	$\{16, 32, 64, 128\}$	$\{16, 32, 64\}$	$\{8, 16, 32\}$	$\{4, 8, 16\}$
d_{model} $\mathcal{D}$	$\{64, 128, 256, 512\}$	$\{64, 128, 256\}$	$\{64, 128, 256\}$	$\{64, 128\}$
TSMixer hidden size $\mathcal{H}$	$\{32, 64, 128, 256, 512\}$	$\{32, 64, 128, 256\}$	$\{32, 64, 128\}$	$\{32, 64, 128\}$
TSMixer blocks $N_{\max}$	6	4	3	3

Some settings are fixed rather than tuned. The feed-forward width of the transformers is $4\,d_{\text{model}}$, and the DLinear moving-average kernel is 25. The CycleNet MLP backbone has hidden size 512 and uses RevIN without affine parameters, and its cycle length is set per dataset to the dominant seasonal period. TSMixer$_{\text{CI}}$ shares the search space of TSMixer, with the channel-mixing MLP disabled.

F STATISTICAL PROFILING - MORE RESULTS

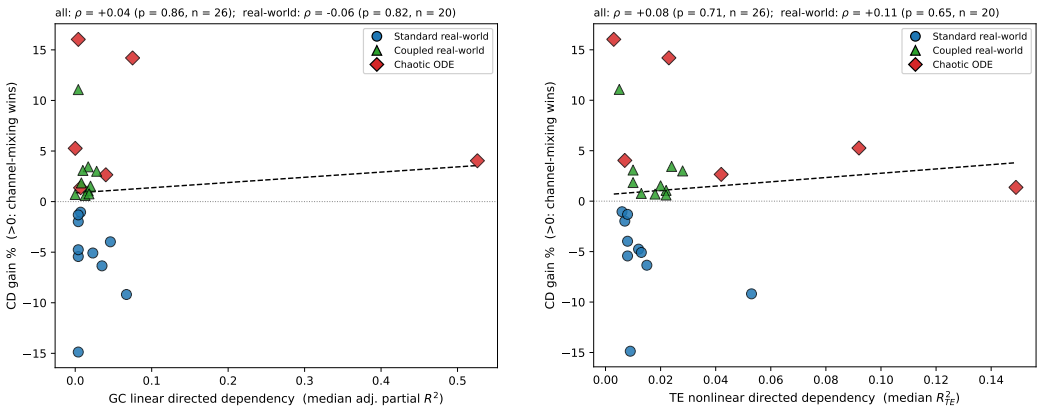
Table 11 reports GC and TE, the two statistical measures that do not retrieve all planted couplings in Section 5, next to MI_{lag} and CD_g for all datasets. At the suite level, both measures are higher on the coupled datasets than on the standard datasets. The median $\widetilde{R^2_{\text{GC}}}$ is 0.015 vs. 0.005 (about $3\times$),

and the median $\widetilde{R^2_{\mathrm{TE}}}$ is 0.019 vs. 0.009 (about $2.1\times$). TE exceeds the standard median on 9 of the 10 coupled datasets, and GC on 8 of 10.

At the dataset level, however, neither measure reliably separates the two suites or tracks the channel-mixing payoff. Traffic has the highest GC (0.067) and TE (0.053) of all real-world datasets, yet one of the most negative CD_g values (-9.2%). Conversely, Bavaria-Donau, the coupled dataset with the largest CD_g ($+11.1\%$), has among the lowest GC and TE values (0.004 and 0.005). The same holds for the ODE suite, where Blinking-Rotlet combines the largest CD_g ($+16.0\%$) with GC and TE values close to zero. Across all 26 datasets, the Spearman correlation with CD_g is only 0.04 for GC and 0.08 for TE (Figure 4), compared to 0.78 for $\mathrm{MI}_{\mathrm{lag}}$ (Figure 1). This matches the synthetic results, where GC misses the non-linear and joint couplings and TE mis-locates the contemporaneous coupling (Table 2). The magnitude of GC and TE therefore reflects the strength of a specific functional form rather than whether channel mixing pays off, which is why we use $\mathrm{MI}_{\mathrm{lag}}$ and CD_g in the main analysis.

Table 11: Median GC and TE over all ordered channel pairs ($\widetilde{R^2_{\mathrm{GC}}}$ and $\widetilde{R^2_{\mathrm{TE}}}$, Section 4), next to $\mathrm{MI}_{\mathrm{lag}}$ and CD_g as in Tables 3 and 5.

Dataset	$\widetilde{R^2_{\mathrm{GC}}}$	$\widetilde{R^2_{\mathrm{TE}}}$	$\mathrm{MI}_{\mathrm{lag}}$ (%)	CD_g (%)
Standard MTSF				
ETTh1	0.004	0.009	19	−14.9
ETTh2	0.004	0.008	19	−5.4
ETTm1	0.004	0.012	38	−4.8
ETTm2	0.004	0.007	38	−2.0
Weather	0.007	0.006	52	−1.0
Electricity	0.023	0.013	61	−5.1
Traffic	0.067	0.053	27	−9.2
ExchangeRate	0.004	0.008	7	−1.3
ILI	0.035	0.015	17	−6.3
SolarAL	0.046	0.008	14	−4.0
category median	*0.005*	*0.009*	*23*	*−4.9*
Coupled				
Bavaria-Donau	0.004	0.005	78	+11.1
Bavaria-Iller	0.017	0.024	67	+3.4
Current-Velocity	0.010	0.010	83	+3.1
CzeLan	0.028	0.028	82	+3.0
ZafNoo	0.008	0.010	57	+1.8
AirQuality	0.020	0.020	54	+1.5
AQShunyi	0.017	0.022	53	+1.1
WaterQuality-Darwin	0.018	0.013	40	+0.8
Wearable	0.000	0.018	17	+0.7
HouseholdPower	0.013	0.022	17	+0.6
category median	*0.015*	*0.019*	*55.5*	*+1.7*
Spatio-temporal (traffic)				
METR-LA	0.011	0.025	38	−1.5
PEMS-BAY	0.021	0.015	69	+6.4
category median	*0.016*	*0.020*	*53.5*	*+2.4*
Chaotic ODE (known-correlated reference)				
Lorenz	0.007	0.149	83	+1.4
Lorenz-Coupled	0.000	0.092	63	+5.3
Blinking-Rotlet	0.004	0.003	83	+16.0
CellCycle	0.040	0.042	67	+2.7
Double-Pendulum	0.526	0.007	92	+4.0
Hopfield	0.075	0.023	73	+14.2
category median	*0.024*	*0.033*	*78*	*+4.6*



(a) GC (b) TE

Figure 4: Median GC (a) and TE (b) vs. CD_g (%) of TSMixer over $\mathrm{TSMixer}_{\mathrm{CI}}$ on 26 datasets (10 standard, 10 coupled, 6 ODE). The dashed line is a linear fit, and Spearman correlations are given above each plot. Compare with $\mathrm{MI}_{\mathrm{lag}}$ in Figure 1.

G Parameter Sensitivity

Pairwise profiling scales quadratically with the number of channels, so datasets with more than 40 channels are profiled on a random subset of $|\mathcal{C}| = 40$ channels, drawn uniformly without replacement with a fixed seed (2021). All $|\mathcal{P}| = |\mathcal{C}|(|\mathcal{C}| - 1)$ ordered pairs of the subset are analyzed. We test whether this choice affects the dataset-level dependency measures on the two widest datasets with a completed sweep, Electricity (321 channels) and PEMS-BAY (325 channels). Following a one-factor-at-a-time design, we vary the subset size $|\mathcal{C}| \in \{20, 40, 80\}$ at the default seed, and the sampling seed over $\{2021, 7, 13\}$ at the default size $|\mathcal{C}| = 40$ (for PEMS-BAY also seed 7 at $|\mathcal{C}| = 20$). All other settings (estimators, lag grids and deseasonalization) are those of the main analysis. Table 12 reports the median degrees-of-freedom corrected partial R^2 of GC, $\widetilde{R^2_{\mathrm{GC}}}$, the median TE R^2, $\widetilde{R^2_{\mathrm{TE}}}$, and the lagged-coupling fraction ϕ_{lag} of $\mathrm{MI}_{\mathrm{lag}}$, all computed over the analyzed pairs (Section 4).

Table 12: Channel-subsampling sensitivity of $\widetilde{R^2_{\mathrm{GC}}}$ (GC), $\widetilde{R^2_{\mathrm{TE}}}$ (TE), and $\mathrm{MI}_{\mathrm{lag}}$ over all ordered pairs of the subset, for subset size $|\mathcal{C}|$ and sampling seed s. Δ: deviation from the default setting ($|\mathcal{C}| = 40$, $s = 2021$, marked *), relative for GC and TE and in percentage points (pp) for $\mathrm{MI}_{\mathrm{lag}}$.

Dataset	Setting	Pairs	GC	ΔGC	TE	ΔTE	$\mathrm{MI}_{\mathrm{lag}}$ (%)	$\Delta\mathrm{MI}_{\mathrm{lag}}$ (pp)
Electricity	$\lvert\mathcal{C}\rvert = 20$, $s = 2021$	380	0.0215	−4.8%	0.0129	−3.7%	66.6	+5.2
	$\lvert\mathcal{C}\rvert = 40$, $s = 2021^*$	1560	0.0225	—	0.0134	—	61.4	—
	$\lvert\mathcal{C}\rvert = 80$, $s = 2021$	6320	0.0210	−6.7%	0.0137	+2.2%	62.9	+1.5
	$\lvert\mathcal{C}\rvert = 40$, $s = 7$	1560	0.0247	+9.5%	0.0147	+10.1%	67.2	+5.8
	$\lvert\mathcal{C}\rvert = 40$, $s = 13$	1560	0.0204	−9.7%	0.0125	−6.7%	68.4	+7.0
PEMS-BAY	$\lvert\mathcal{C}\rvert = 20$, $s = 2021$	380	0.0248	+20.1%	0.0152	+1.3%	64.2	−4.5
	$\lvert\mathcal{C}\rvert = 20$, $s = 7$	380	0.0170	−17.3%	0.0124	−17.0%	74.2	+5.5
	$\lvert\mathcal{C}\rvert = 40$, $s = 2021^*$	1560	0.0206	—	0.0150	—	68.7	—
	$\lvert\mathcal{C}\rvert = 80$, $s = 2021$	6320	0.0188	−8.8%	0.0143	−4.7%	67.5	−1.2
	$\lvert\mathcal{C}\rvert = 40$, $s = 7$	1560	0.0193	−6.2%	0.0142	−5.3%	67.2	−1.5
	$\lvert\mathcal{C}\rvert = 40$, $s = 13$	1560	0.0151	−26.9%	0.0142	−5.3%	69.7	+1.0

Number of sampled channels. Doubling the subset from $|\mathcal{C}| = 40$ to $|\mathcal{C}| = 80$ (four times as many pairs) changes the medians by at most 9% (GC −6.7% and −8.8%, TE +2.2% and −4.7% on Electricity and PEMS-BAY). Halving it to $|\mathcal{C}| = 20$ is less stable: with only 380 pairs, the PEMS-BAY GC median deviates by +20% at seed 2021 and by −17% at seed 7. $\mathrm{MI}_{\mathrm{lag}}$ is similarly stable.

It moves by $+1.5$ and -1.2 pp at $|\mathcal{C}| = 80$ and by at most 5.5 pp at $|\mathcal{C}| = 20$. A cap of $|\mathcal{C}| = 40$ is therefore large enough for the medians to settle, while keeping the pairwise TE tests tractable.

Sampling seed. At $|\mathcal{C}| = 40$, the choice of subset moves the Electricity medians by up to $\pm 10\%$ (coefficient of variation across seeds: 9.6% for GC, 8.4% for TE). On PEMS-BAY, TE is stable (CV 3.2%), whereas GC varies more (CV 15.8%, with seed 13 giving -27%), reflecting the heterogeneity of its sensor network. $\mathrm{MI}_{\mathrm{lag}}$ shifts by up to $+7.0$ pp on Electricity and by at most 1.5 pp on PEMS-BAY.

Effect on the cross-dataset comparison. These deviations are small relative to the differences between datasets, whose median GC effect sizes span more than an order of magnitude (from about 0.004 on the ETT datasets to 0.067 on Traffic) and whose $\mathrm{MI}_{\mathrm{lag}}$ values range from 7% (ExchangeRate) to 83% (Current-Velocity). In particular, Electricity keeps the same rank among the profiled real-world datasets on GC and TE under all five settings. On $\mathrm{MI}_{\mathrm{lag}}$, both datasets stay above 50% lagged-coupled pairs under all settings (61–68% on Electricity and 64–74% on PEMS-BAY), and Electricity remains the standard dataset with the highest $\mathrm{MI}_{\mathrm{lag}}$. The subsampling therefore does not change the dataset ordering on which the coupling analysis relies.

H Seed-Averaged Results and Standard Deviations

I Per-horizon Results

In this section, we detail based on the horizons discussed in A the results for all models on all datasets. For better organization of results, we split them into three seperate tables, Table 17, Table 18, and Table 19 showing the results for the standard, spatio temporal and the coupled datasets respectively.

J Synthetic Data Generation Description

Each synthetic dataset has four channels $c_0, \dots, c_3$ observed over $T = 20{,}000$ hourly timesteps. Channels c_1, c_2, c_3 are independent sources. The target c_0 receives the planted coupling. All six datasets share the same four channel bases, drawn from the same random streams, and differ *only* in the edges planted into c_0. Generation proceeds in the five steps below. Throughout, $z[\cdot]$ denotes standardization over the full sample,

$$z[v](t) = \frac{v(t) - \bar{v}}{\hat{\sigma}_v}, \qquad \bar{v} = \tfrac{1}{T}\textstyle\sum_t v(t), \quad \hat{\sigma}_v^2 = \tfrac{1}{T}\textstyle\sum_t \big(v(t) - \bar{v}\big)^2, \tag{10}$$

so every weight below is a variance share.

Step 1: noise. Each channel c_i has its own pseudo-random generator, seeded from the dataset seed (21) and the channel name, so a channel's draws do not depend on the order in which channels are generated. From it we draw Gaussian innovations and one seasonal phase:

$$\varepsilon_i(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad t = -B, \dots, T-1,\ B = 200; \qquad \phi_i \sim \mathcal{U}[0, 2\pi). \tag{11}$$

Innovations of different channels are independent, so the bases built from them are mutually independent processes.

Step 2: stochastic AR(p) core. The core of c_i is an autoregression of order p_i, started at zero and run from $t = -B$, with the first B samples discarded as burn-in:

$$\eta_i(t) = \sum_{k=1}^{p_i} a_{i,k}\, \eta_i(t-k) + \varepsilon_i(t), \qquad a_{i,k} = \rho_i \frac{e^{-(k-1)/\tau_i}}{\sum_{j=1}^{p_i} e^{-(j-1)/\tau_i}}, \quad \tau_i = \tfrac{p_i}{2}. \tag{12}$$

The coefficients decay exponentially and sum to the persistence $\rho_i < 1$. Since $a_{i,k} > 0$, for $|x| \le 1$ we have $\left|\sum_k a_{i,k} x^k\right| \le \rho_i < 1$, so the characteristic polynomial $1 - \sum_k a_{i,k} x^k$ has no root in the closed unit disc and every core is stationary. The orders and persistences are $(p_i, \rho_i) = (4, 0.7)$, $(5, 0.8)$, $(3, 0.6)$, $(4, 0.7)$ for $i = 0, \dots, 3$. For example, $(a_{0,1}, \dots, a_{0,4}) = (0.319,\ 0.193,\ 0.117,\ 0.071)$.

Table 13: Seed-averaged test MSE, i.e., the mean over seeds 2021/2025/2026 of the horizon-averaged MSE. Best per row in **bold** and second underlined, with ties all bold and Win listing the strategies. Models grouped by channel strategy and shaded: CI channel-independent (TSM_{CI}=$\text{TSMixer}_{\text{CI}}$, DLin.=DLinear, CycNet=CycleNet), CD channel-dependent (TSM=TSMixer, SimTM=SimpleTM, iTr.=iTransformer).

	CI			CD			
Dataset	TSM_{CI}	DLin.	CycNet	TSM	SimTM	iTr.	Win
Standard MTSF							
ETTh1	0.433	0.456	**0.413**	0.498	0.427	0.451	CI
ETTh2	0.390	0.441	**0.370**	0.411	0.379	0.382	CI
ETTm1	**0.352**	0.359	0.354	0.369	0.357	0.367	CI
ETTm2	0.266	0.263	**0.249**	0.271	0.265	0.273	CI
Weather	0.232	**0.222**	0.232	0.235	0.235	0.238	CI
Electricity	**0.158**	0.161	0.159	0.166	0.162	0.162	CI
Traffic	**0.391**	0.410	0.408	0.427	0.399	0.418	CI
ExchangeRate	0.152	**0.128**	0.173	0.154	0.182	0.165	CI
ILI	**1.873**	2.388	1.952	1.992	1.900	2.382	CI
SolarAL	**0.193**	0.224	0.202	0.200	0.200	0.198	CI
Spatio-temporal (traffic)							
METR-LA	0.916	**0.898**	0.919	0.930	0.951	1.020	CI
PEMS-BAY	0.501	0.613	0.528	0.469	**0.423**	0.446	CD
Coupled							
AirQuality	0.750	**0.738**	0.744	0.739	0.752	0.748	CI
AQShunyi	0.684	0.672	0.682	0.677	**0.664**	0.697	CD
CzeLan	0.212	0.218	0.216	**0.205**	0.211	0.225	CD
ZafNoo	0.523	**0.486**	0.528	0.514	0.529	0.539	CI
Bavaria-Iller	0.185	0.176	0.186	0.179	**0.170**	0.186	CD
Bavaria-Donau	0.097	0.095	0.098	**0.086**	**0.086**	0.100	CD
WaterQuality-Darwin	0.270	0.281	0.271	**0.268**	0.275	0.280	CD
Current-Velocity	0.263	0.264	0.270	0.255	**0.254**	0.265	CD
Wearable	0.671	0.668	0.781	**0.666**	0.753	0.755	CD
HouseholdPower	0.652	**0.644**	0.646	0.648	0.648	0.646	CI

Step 3: seasonal and trend components. Each channel carries one random-phase sinusoid and, optionally, a linear trend:

$$s_i(t) = \sin\Big(\frac{2\pi t}{P_i} + \phi_i\Big), \qquad g_i(t) = \beta_i\, t. \tag{13}$$

The periods are distinct across channels, $(P_0, \ldots, P_3) = (24, 168, 8, 12)$, so no two channels share a frequency. A shared frequency would create an artificial periodic cross-correlation between channels that are not coupled. For the same reason, the trend is disabled in the suite ($\beta_i = 0$). Two independent trending series are correlated in levels (spurious regression) and are non-stationary, which would contaminate the ground truth.

Step 4: channel base. The components are combined *additively* at an explicit variance split $\pi_i = (\pi_{i,\eta}, \pi_{i,s}, \pi_{i,g})$, with $\sum_m \pi_{i,m} = 1$:

$$u_i = z\Big[\sqrt{\pi_{i,\eta}}\; z[\eta_i] + \sqrt{\pi_{i,s}}\; z[s_i] + \sqrt{\pi_{i,g}}\; z[g_i]\Big]. \tag{14}$$

The seasonal and trend terms are added to the output of the AR recursion, not fed through it, so the AR polynomial does not filter the season. Because the components are standardized before weighting, $\pi_{i,m}$ is (up to finite-sample correlation) the fraction of the base's variance explained by component m, whatever the AR persistence or the sinusoid amplitude. We use $\pi_{i,g} = 0$ and $(\pi_{i,\eta}, \pi_{i,s}) = (0.75, 0.25)$, $(0.80, 0.20)$, $(0.85, 0.15)$, $(0.80, 0.20)$ for $i = 0, \ldots, 3$. With $z[s_i] = \sqrt{2}\, s_i$, the target's base is thus

$$u_0(t) \approx 0.866\; z[\eta_0](t) + 0.707\; \sin\Big(\frac{2\pi t}{24} + \phi_0\Big). \tag{15}$$

The sources have no incoming edges and are final at this point: $c_i = z[u_i]$ for $i = 1, 2, 3$.

Table 14: Seed-averaged test MAE, i.e., the mean over seeds 2021/2025/2026 of the horizon-averaged MAE. Best per row in **bold** and second underlined, with ties all bold and Win listing the strategies. Models grouped by channel strategy and shaded: CI channel-independent (TSM_{CI}=$\text{TSMixer}_{\text{CI}}$, DLin.=DLinear, CycNet=CycleNet), CD channel-dependent (TSM=TSMixer, SimTM=SimpleTM, iTr.=iTransformer).

	CI			CD			
Dataset	TSM_{CI}	DLin.	CycNet	TSM	SimTM	iTr.	Win
Standard MTSF							
ETTh1	0.437	0.457	**0.425**	0.473	0.440	0.459	CI
ETTh2	0.408	0.449	**0.401**	0.422	0.406	0.406	CI
ETTm1	**0.380**	0.382	0.381	0.392	0.383	0.394	CI
ETTm2	0.320	0.330	**0.313**	0.325	0.321	0.330	CI
Weather	0.271	0.279	0.271	0.272	**0.270**	0.275	CD
Electricity	**0.251**	0.258	0.252	0.266	0.256	0.258	CI
Traffic	**0.268**	0.282	0.278	0.306	0.282	0.298	CI
ExchangeRate	0.253	**0.246**	0.267	0.255	0.273	0.266	CI
ILI	**0.902**	1.095	0.932	0.909	**0.902**	1.027	CD/CI
SolarAL	**0.240**	0.287	0.253	0.261	0.257	0.261	CI
Spatio-temporal (traffic)							
METR-LA	0.555	0.587	0.551	0.550	**0.546**	0.580	CD
PEMS-BAY	0.338	0.401	0.352	0.323	**0.298**	0.300	CD
Coupled							
AirQuality	0.502	0.504	0.495	**0.494**	0.496	0.499	CD
AQShunyi	0.507	0.508	0.505	0.504	**0.497**	0.510	CD
CzeLan	**0.256**	0.277	0.264	0.259	0.262	0.279	CI
ZafNoo	0.462	**0.440**	0.461	0.459	0.470	0.472	CI
Bavaria-Iller	0.187	0.207	0.188	0.184	**0.182**	0.187	CD
Bavaria-Donau	0.169	0.168	0.173	**0.155**	0.156	0.172	CD
WaterQuality-Darwin	0.188	0.190	0.190	**0.185**	0.187	0.190	CD
Current-Velocity	0.269	0.271	0.274	0.263	**0.262**	0.271	CD
Wearable	0.592	**0.590**	0.599	**0.590**	0.591	0.594	CD/CI
HouseholdPower	0.509	0.514	0.505	0.508	**0.504**	**0.504**	CD

Step 5: planting the coupling. The target mixes its own base with the standardized contributions of its incoming edges $e \in \mathcal{E}_0$:

$$c_0 = z\Big[\sqrt{\omega}\, u_0 + \sum_{e\in\mathcal{E}_0} \sqrt{w_e}\, z\big[f_e(c_{\text{src}_e})\big]\Big], \qquad \omega + \sum_e w_e = 1. \tag{16}$$

ω is the own share and w_e the strength of edge e. u_0 is independent of every source, and the edges of a dataset use independent sources, so the terms are uncorrelated and each explains its nominal share of $\text{Var}(c_0)$. The realized shares match to within 0.01. Standardizing each contribution before weighting ties the coupling strength to w_e alone, independent of the functional form of f_e. Each f_e reads its source through the causal lag operator $(\text{L}^\ell x)(t) = x(t-\ell)$ for $t \geq \ell$, and 0 otherwise, and is one of

$$f(x)(t) \in \Big\{\, x(t),\;\; x(t-\ell),\;\; x(t-\ell)^2,\;\; (c_a c_b)(t-\ell) \,\Big\}, \tag{17}$$

i.e. contemporaneous linear, lagged linear, lagged non-linear, and multiplicative (joint) coupling. The multiplicative coupling is a single edge with two sources c_a, c_b. With $\omega = 0.6$ and $w_e = 0.4$ for every single-edge dataset, the six targets are given below, as in Figure 2. We omit the outer $z[\cdot]$ of Eq. equation 16, which is close to the identity here, since the terms are uncorrelated with unit

Table 15: Across-seed std of the horizon-averaged test MSE (seeds 2021/2025/2026), reported to gauge significance. Columns: TSM_{CI}=$\text{TSMixer}_{\text{CI}}$, DLin.=DLinear, CycNet=CycleNet (CI), and TSM=TSMixer, SimTM=SimpleTM, iTr.=iTransformer (CD).

	CI			**CD**		
Dataset	TSM_{CI}	DLin.	CycNet	TSM	SimTM	iTr.
Standard MTSF						
ETTh1	0.001	0.012	<0.001	0.009	<0.001	0.003
ETTh2	0.001	0.012	0.003	0.003	0.001	0.001
ETTm1	<0.001	0.006	<0.001	0.002	0.003	0.004
ETTm2	0.001	0.003	0.001	0.002	0.001	0.001
Weather	0.001	0.002	0.001	0.002	0.001	0.002
Electricity	<0.001	<0.001	<0.001	0.001	0.009	<0.001
Traffic	<0.001	<0.001	<0.001	0.001	0.007	0.018
ExchangeRate	<0.001	0.003	0.003	0.001	0.001	0.004
ILI	0.062	0.018	0.042	0.087	0.006	0.223
SolarAL	<0.001	<0.001	0.001	0.001	0.004	0.001
Spatio-temporal (traffic)						
METR-LA	0.002	0.002	0.004	0.011	0.003	0.009
PEMS-BAY	<0.001	<0.001	0.005	0.002	0.001	0.003
Coupled						
AirQuality	0.001	<0.001	<0.001	0.003	0.002	0.003
AQShunyi	<0.001	<0.001	<0.001	0.005	0.004	0.001
CzeLan	<0.001	<0.001	0.001	0.001	0.003	0.002
ZafNoo	0.004	0.002	0.001	0.004	0.006	0.002
Bavaria-Iller	<0.001	0.001	<0.001	0.001	<0.001	<0.001
Bavaria-Donau	<0.001	0.001	0.003	0.001	0.001	0.002
WaterQuality-Darwin	0.002	0.001	0.004	0.004	0.001	0.002
Current-Velocity	<0.001	<0.001	0.001	0.001	<0.001	0.002
Wearable	<0.001	<0.001	0.001	0.001	<0.001	<0.001
HouseholdPower	<0.001	<0.001	0.001	0.001	0.001	0.002

variance and their shares sum to one:

$$
\begin{aligned}
\textsf{independent}: &\quad c_0(t) = u_0(t),\\
\textsf{contemp-linear}: &\quad c_0(t) = \sqrt{0.6}\,u_0(t) + \sqrt{0.4}\,z[c_1(t)],\\
\textsf{lagged-linear}: &\quad c_0(t) = \sqrt{0.6}\,u_0(t) + \sqrt{0.4}\,z[c_1(t-24)],\\
\textsf{lagged-non-linear}: &\quad c_0(t) = \sqrt{0.6}\,u_0(t) + \sqrt{0.4}\,z[c_1(t-12)^2],\\
\textsf{multiplicative-joint}: &\quad c_0(t) = \sqrt{0.6}\,u_0(t) + \sqrt{0.4}\,z[(c_1c_2)(t-6)],\\
\textsf{additive-mixed}: &\quad c_0(t) = \sqrt{0.3}\,u_0(t) + \sqrt{0.4}\,z[c_1(t-24)] + \sqrt{0.3}\,z[c_2(t-8)^2].
\end{aligned}
$$

For the linear contributions, the inner $z[\cdot]$ is the identity, because the sources are already standardized. In each dataset, the sources without an edge into c_0 are independent controls. The four channels $c_0, \ldots, c_3$ are written as one CSV with an hourly timestamp and split chronologically 70/10/20 into train/validation/test, as for the single-series real-world datasets (Appendix A).

Table 16: Across-seed std of the horizon-averaged test MAE (seeds 2021/2025/2026), reported to gauge significance. Columns: TSM_{CI}=$TSMixer_{CI}$, DLin.=DLinear, CycNet=CycleNet (CI), and TSM=TSMixer, SimTM=SimpleTM, iTr.=iTransformer (CD).

	CI			**CD**		
Dataset	TSM_{CI}	DLin.	CycNet	TSM	SimTM	iTr.
Standard MTSF						
ETTh1	0.001	0.007	<0.001	0.004	<0.001	0.002
ETTh2	<0.001	0.006	0.002	0.001	0.001	0.001
ETTm1	<0.001	0.006	<0.001	0.002	0.001	0.003
ETTm2	<0.001	0.003	<0.001	0.001	0.001	<0.001
Weather	0.001	0.002	0.001	<0.001	0.001	0.002
Electricity	<0.001	<0.001	<0.001	0.001	0.010	<0.001
Traffic	<0.001	<0.001	<0.001	0.002	0.006	0.013
ExchangeRate	<0.001	0.003	0.002	<0.001	0.001	0.002
ILI	0.021	0.004	0.019	0.016	0.006	0.059
SolarAL	<0.001	0.001	<0.001	0.001	0.005	<0.001
Spatio-temporal (traffic)						
METR-LA	<0.001	0.001	0.004	0.002	0.002	0.004
PEMS-BAY	<0.001	<0.001	<0.001	0.002	0.001	0.001
Coupled						
AirQuality	0.001	<0.001	<0.001	0.001	0.001	0.002
AQShunyi	<0.001	<0.001	<0.001	0.002	0.001	0.001
CzeLan	<0.001	0.001	0.002	0.001	0.001	0.001
ZafNoo	0.002	0.001	<0.001	0.002	0.003	<0.001
Bavaria-Iller	0.001	0.002	<0.001	0.001	<0.001	0.001
Bavaria-Donau	<0.001	0.002	0.003	0.001	0.001	0.002
WaterQuality-Darwin	0.002	<0.001	0.001	<0.001	0.003	0.001
Current-Velocity	<0.001	<0.001	0.001	0.001	<0.001	0.002
Wearable	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
HouseholdPower	<0.001	<0.001	<0.001	0.001	<0.001	0.002

Table 17: Per-horizon test MSE/MAE on the standard MTSF datasets (mean over seeds 2021/2025/2026). Horizon averages are in Tables 13/14. Best per row and metric **bold**, second underlined (ties at three decimals all bold), and 1st count = wins. Shading and abbreviations as in Table 13.

		CI						CD					
Dataset	H	TSM_{CI}		DLin.		CycNet		TSM		SimTM		iTr.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.378	**0.395**	0.374	0.396	**0.372**	**0.395**	0.416	0.417	0.379	0.397	0.397	0.413
	192	0.419	0.426	0.453	0.448	**0.401**	**0.413**	0.456	0.447	0.426	0.435	0.454	0.454
	336	0.445	0.443	0.469	0.457	**0.431**	**0.430**	0.460	0.455	0.444	0.449	0.460	0.461
	720	0.490	0.486	0.528	0.527	**0.448**	**0.463**	0.658	0.574	0.457	0.478	0.496	0.507
ETTh2	96	0.293	0.343	0.283	0.347	**0.281**	**0.338**	0.311	0.355	0.296	0.349	0.293	0.349
	192	0.376	0.396	0.380	0.414	**0.361**	**0.393**	0.415	0.416	0.378	0.396	0.390	0.399
	336	0.439	0.441	0.421	0.446	**0.403**	**0.426**	0.484	0.463	0.414	0.431	0.420	0.433
	720	0.453	0.454	0.680	0.588	0.436	0.449	0.436	0.455	0.427	0.447	**0.426**	**0.445**
ETTm1	96	0.289	0.339	**0.286**	**0.335**	0.306	0.354	0.288	0.342	0.289	0.343	0.304	0.353
	192	0.330	0.366	0.335	**0.364**	0.337	0.375	0.331	0.368	**0.329**	0.368	0.355	0.389
	336	**0.359**	0.389	0.369	**0.384**	0.365	0.386	0.401	0.408	0.369	0.391	0.374	0.396
	720	0.430	0.427	0.447	0.443	**0.409**	**0.410**	0.456	0.448	0.442	0.432	0.436	0.436
ETTm2	96	0.170	0.254	0.163	0.257	**0.160**	**0.250**	0.171	0.256	0.170	0.256	0.180	0.269
	192	0.236	0.301	0.220	0.300	**0.215**	**0.289**	0.234	0.305	0.234	0.299	0.244	0.312
	336	0.291	0.336	0.283	0.347	**0.269**	**0.328**	0.289	0.340	0.284	0.338	0.288	0.341
	720	0.366	0.390	0.385	0.415	**0.353**	**0.383**	0.390	0.400	0.371	0.391	0.380	0.398
Weather	96	0.150	0.200	**0.145**	0.211	0.148	0.199	0.149	0.205	0.150	**0.198**	0.159	0.206
	192	0.200	0.252	**0.185**	0.248	0.206	0.257	0.191	**0.243**	0.197	**0.243**	0.204	0.254
	336	0.246	0.286	**0.237**	0.294	0.261	0.298	0.251	0.289	0.245	**0.284**	0.264	0.299
	720	0.334	0.346	0.322	0.365	**0.313**	**0.333**	0.348	0.350	0.347	0.356	0.323	0.341
Electricity	96	**0.129**	**0.223**	0.133	0.230	0.131	0.225	0.142	0.240	0.131	0.225	0.133	0.230
	192	**0.147**	**0.239**	**0.147**	0.244	0.148	0.241	0.152	0.251	0.154	0.246	0.156	0.252
	336	0.162	**0.256**	0.162	0.261	**0.159**	**0.256**	0.169	0.273	0.172	0.269	0.169	0.265
	720	0.195	0.287	0.201	0.298	0.197	0.287	0.202	0.298	**0.189**	**0.285**	0.190	**0.285**
Traffic	96	0.359	**0.251**	0.385	0.269	0.381	0.266	0.380	0.267	**0.355**	0.258	0.436	0.319
	192	**0.379**	**0.262**	0.396	0.274	0.397	0.269	0.426	0.319	0.391	0.274	0.395	0.292
	336	**0.398**	**0.272**	0.410	0.282	0.405	0.278	0.431	0.300	0.412	0.294	0.406	0.282
	720	**0.430**	**0.288**	0.450	0.305	0.447	0.300	0.472	0.337	0.440	0.302	0.433	0.298
ExchangeRate	30	**0.028**	**0.113**	**0.028**	**0.113**	0.029	0.114	0.029	0.115	0.029	0.115	0.029	0.114
	90	0.077	0.192	**0.074**	**0.191**	0.077	0.193	0.077	0.193	0.076	0.192	0.094	0.216
	180	0.160	0.284	**0.144**	**0.282**	0.171	0.292	0.162	0.285	0.162	0.283	0.164	0.286
	365	0.344	0.424	**0.266**	**0.399**	0.414	0.468	0.350	0.427	0.462	0.503	0.375	0.445
ILI	24	1.857	0.850	2.326	1.073	1.918	0.884	2.059	0.920	**1.731**	**0.819**	1.839	0.879
	36	**1.756**	**0.878**	2.294	1.082	1.824	0.905	1.997	0.891	1.813	0.898	2.091	0.919
	48	2.007	0.979	2.545	1.131	2.113	1.006	**1.920**	**0.915**	2.156	0.989	3.216	1.281
SolarAL	96	**0.173**	**0.225**	0.222	0.294	0.196	0.236	0.196	0.245	0.183	0.232	0.177	0.242
	192	0.193	**0.233**	0.211	0.272	0.199	0.253	**0.187**	0.258	0.197	0.259	0.192	0.259
	336	**0.198**	**0.245**	0.228	0.286	0.203	0.257	0.202	0.269	0.210	0.266	0.208	0.274
	720	**0.207**	**0.257**	0.236	0.294	0.211	0.265	0.216	0.273	0.210	0.270	0.212	0.269
1st count		11	14	9	7	14	14	2	2	4	5	1	2

Table 18: Per-horizon test MSE/MAE on the spatio-temporal (traffic) datasets (mean over seeds 2021/2025/2026). Horizon averages are in Tables 13/14. Best per row and metric **bold**, second underlined (ties at three decimals all bold), and 1st count = wins. Shading and abbreviations as in Table 13.

		CI						**CD**					
Dataset	H	TSM_{CI}		DLin.		CycNet		TSM		SimTM		iTr.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
METR-LA	12	0.431	0.337	0.430	0.345	0.433	0.344	**0.422**	**0.321**	0.424	0.322	0.453	0.355
	48	0.817	0.522	0.838	0.569	**0.806**	0.521	0.810	0.543	0.892	**0.506**	0.881	0.546
	144	1.162	0.659	**1.106**	0.699	1.156	**0.645**	1.170	0.656	1.178	0.662	1.346	0.695
	288	1.255	0.701	**1.217**	0.736	1.281	0.693	1.319	**0.678**	1.310	0.694	1.400	0.725
PEMS-BAY	12	0.294	0.260	0.317	0.267	0.273	0.245	**0.247**	0.225	0.264	0.233	0.252	**0.224**
	48	0.465	0.324	0.799	0.497	0.508	0.343	0.483	0.326	0.441	0.310	**0.435**	**0.293**
	144	0.600	0.368	0.656	0.414	0.641	0.394	0.547	0.353	**0.468**	**0.318**	0.534	0.335
	288	0.646	0.401	0.678	0.425	0.689	0.427	0.601	0.387	**0.520**	**0.329**	0.564	0.349
1st count		0	0	2	0	1	1	2	2	2	3	1	2

Table 19: Per-horizon test MSE/MAE on the coupled datasets (mean over seeds 2021/2025/2026). Horizon averages are in Tables 13/14. Best per row and metric **bold**, second underlined (ties at three decimals all bold), and 1st count = wins. Shading and abbreviations as in Table 13.

		CI						**CD**					
Dataset	H	TSM_{CI}		DLin.		CycNet		TSM		SimTM		iTr.	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AirQuality	96	0.693	0.476	0.678	0.474	0.683	**0.467**	**0.674**	0.469	0.687	0.473	0.691	0.474
	192	0.737	0.494	**0.720**	0.496	0.726	**0.487**	0.769	0.501	0.730	0.490	0.727	0.490
	336	0.756	0.506	0.748	0.515	0.750	**0.499**	**0.740**	0.502	0.752	0.502	0.751	0.501
	720	0.815	0.530	0.805	0.532	0.817	0.527	**0.773**	**0.504**	0.839	0.521	0.822	0.532
AQShunyi	96	0.639	0.484	**0.621**	0.480	0.628	**0.476**	0.652	0.491	0.630	**0.476**	0.678	0.491
	192	0.676	0.502	0.659	0.502	0.665	**0.494**	0.667	0.499	**0.652**	**0.494**	0.670	0.500
	336	0.687	0.509	0.681	0.516	0.694	0.516	0.681	0.507	**0.670**	**0.501**	0.699	0.514
	720	0.733	0.533	0.725	0.535	0.743	0.533	0.707	0.520	**0.705**	**0.516**	0.742	0.534
CzeLan	96	0.172	**0.219**	0.167	0.227	0.178	0.236	**0.162**	0.221	0.169	0.228	0.171	0.235
	192	0.200	**0.246**	0.195	0.253	0.202	**0.246**	**0.193**	0.247	0.204	0.252	0.234	0.289
	336	**0.222**	**0.265**	0.225	0.284	0.231	0.276	0.228	0.281	0.236	0.284	0.242	0.284
	720	0.253	0.296	0.286	0.343	0.255	0.299	0.239	0.287	**0.234**	**0.284**	0.254	0.308
ZafNoo	96	0.441	0.405	**0.419**	**0.393**	0.447	0.408	0.431	0.405	0.445	0.410	0.457	0.419
	192	0.506	0.453	**0.477**	**0.433**	0.511	0.446	0.496	0.452	0.514	0.463	0.507	0.449
	336	0.544	0.479	**0.508**	**0.447**	0.560	0.476	0.538	0.477	0.553	0.494	0.552	0.474
	720	0.603	0.513	**0.541**	**0.485**	0.594	0.512	0.589	0.503	0.604	0.512	0.641	0.545
Bavaria-Iller	24	0.040	0.079	0.041	0.082	0.047	0.083	**0.035**	0.077	0.036	**0.072**	0.043	0.080
	96	0.172	0.179	0.171	0.201	0.174	0.182	**0.154**	**0.171**	0.156	0.174	0.172	0.179
	336	0.344	0.304	**0.315**	0.339	0.336	**0.300**	0.348	0.304	0.318	0.301	0.342	0.302
Bavaria-Donau	24	0.012	0.060	0.011	0.058	0.011	0.059	**0.009**	**0.051**	0.010	0.055	0.011	0.057
	96	0.058	0.142	0.066	0.155	0.073	0.166	**0.050**	**0.130**	0.054	0.131	0.064	0.153
	336	0.221	0.306	0.210	0.291	0.210	0.295	0.200	0.284	**0.194**	**0.281**	0.224	0.305
WaterQuality-Darwin	24	0.222	0.169	0.230	0.171	0.229	0.170	**0.218**	**0.165**	0.230	0.171	0.229	0.173
	96	0.317	0.207	0.332	0.209	**0.314**	0.209	0.317	0.205	0.321	**0.204**	0.332	0.208
Current-Velocity	24	0.230	0.243	0.237	0.248	0.237	0.248	**0.223**	**0.236**	0.224	**0.236**	0.234	0.248
	96	0.295	0.295	0.291	0.293	0.302	0.299	0.287	0.289	**0.285**	**0.288**	0.296	0.295
Wearable	12	0.647	0.576	**0.638**	0.571	0.734	0.576	0.639	0.571	0.723	**0.570**	0.727	0.574
	48	0.673	0.591	**0.669**	0.591	0.778	0.599	0.672	0.592	0.758	**0.590**	0.761	0.593
	144	0.678	0.596	0.676	0.595	0.802	0.605	**0.671**	**0.593**	0.778	0.598	0.779	0.599
	288	0.685	**0.604**	0.687	0.605	0.809	0.615	**0.683**	0.605	0.754	0.607	0.754	0.608
HouseholdPower	96	0.629	0.493	0.623	0.501	0.623	0.493	0.625	0.494	0.622	0.490	**0.621**	**0.488**
	192	0.643	0.505	0.637	0.510	**0.635**	0.502	0.644	0.508	0.642	**0.498**	0.641	0.501
	336	0.663	0.515	**0.650**	0.518	0.652	0.509	0.653	0.512	0.654	**0.507**	0.651	0.508
	672	0.674	0.523	**0.666**	0.528	0.674	**0.516**	0.671	0.520	0.672	0.521	0.669	0.518
1st count		1	4	11	4	2	8	13	7	6	14	1	1