DaTS-Mix: Selecting Pretraining Mixtures for Time-Series Foundation Models Using Domain and Temporal Statistics

Anonymous authors
Paper under double-blind review

Abstract

Domain-aware pretraining-mixture selection has been widely studied for large language models, but its use in time-series foundation model (TSFM) pretraining is less well established. We study this problem and emphasize temporal statistics as a complement to domain information when selecting source proportions. We propose DaTS-Mix, which represents each mixture using source proportions, domain shares, and the first and second raw moments of 14 temporal statistics computed under the training sampler. We condition a TabPFN response predictor on outcomes from 128 runs of a 7.07M-parameter forecasting proxy and use it to select a mixture for larger-model pretraining. On later evaluation windows from the same 28 sources in seven domains, the selected mixture reduces geometric-mean forecasting loss by 10.85% relative to size-proportional sampling and by 5.57% relative to a matched selector without temporal statistics, under the same 20k-update budget. The mixture selected with both moment blocks achieves the lowest mean native loss and weighted quantile loss among the evaluated methods; a Mean-only ablation separates the incremental role of the second-moment block. These results support proxy-guided TSFM data-mixture design and highlight the value of temporal statistics beyond source and domain proportions.

1 Introduction

Time-series foundation models (TSFMs) learn from heterogeneous corpora containing energy, environmental, health, and web signals (Ansari et al., 2024; 2025). Once the source pool and forecasting architecture are fixed, an important remaining decision is how frequently each source is sampled. Data availability and forecasting utility need not coincide: a large dataset may dominate a size-proportional training stream without making the best use of a limited training budget. We study this allocation problem while holding the forecasting architecture and training recipe fixed.

Pretraining-data design is an active research area, particularly for language models, where domain information guides learned mixture proportions (Xie et al., 2023a; Ye et al., 2025; Chen et al., 2025). RegMix (Liu et al., 2025) trains small models on different mixtures, learns a regression model of their outcomes, and uses it to select data for larger-model training. For TSFMs, evidence connecting inexpensive forecasting experiments to source-level mixture choices and larger-model outcomes remains limited. Domain labels alone also do not describe the temporal dynamics that differ across sources, channels, and contexts. We therefore ask whether incorporating temporal statistics into mixture descriptors improves the utility of the selected training data.

We propose DaTS-Mix (Domain and Temporal Statistics for mixture selection), which augments source proportions and domain shares with 14 interpretable temporal statistics. These describe spectral structure, increments, half-window comparisons, and autocorrelation. We pool their means and second raw moments under each candidate mixture. A TabPFN response predictor (Hollmann et al., 2025; Grinsztajn et al., 2026) conditions on these descriptors and measured forecasting-proxy outcomes. Figure 1 summarizes how these statistics enter the selection procedure.

We construct 128 observed mixtures covering between-domain and within-domain allocation spread, complemented by size-informed proposals. Outcomes from 7.07M-parameter forecasting proxies support candidate scoring before committing the larger training budget. Training a Chronos-2-

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.

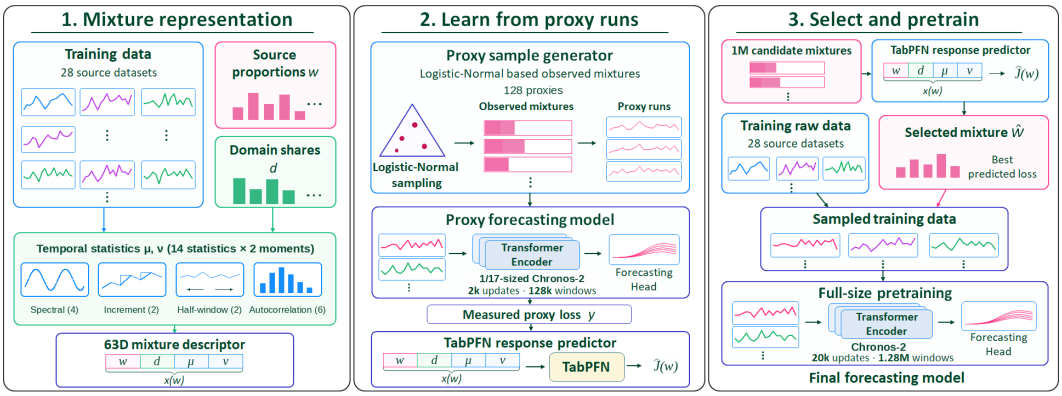

Figure 1: DATS-MIX workflow. Three anchors and 125 structured logistic-normal proposals provide 128 proxy observations (Section 4). Means and marginal second raw moments of 14 temporal statistics join source proportions and domain shares in a 63-dimensional descriptor. Descriptors and measured proxy losses condition TabPFN, which scores one million candidates from a separate Dirichlet bank. The selected source proportions define the training mixture for full-size pretraining from scratch; proxy-model parameters are not transferred. The four statistic groups follow Section 3.1; icons are schematic.

architecture model from scratch on the selected mixture reduces confirmation loss by 10.85% relative to size-proportional sampling at the same 20k-update budget. Ratio+domain removes temporal statistics, while Mean-only retains their means but removes the second-moment block. These matched ablations assess temporal statistics with and without the second-moment block under the same selection pipeline. Response-prediction diagnostics and source- and domain-level analyses describe the selected mixtures and their forecasting outcomes.

Our contributions.

- **Empirical evidence for proxy-guided TSFM pretraining.** Controlled full-size evaluations connect inexpensive mixture observations, predicted selection, and actual forecasting outcomes in a setting less explored than language-model pretraining.
- **Temporal statistics for mixture selection.** We augment domain information with the means and second raw moments of 14 sampler-consistent temporal statistics, providing summaries that jointly determine their marginal variation.
- **Matched evaluation of temporal statistics.** Shared observations, predictor settings, and candidate search support Ratio+domain and Mean-only ablations, alongside a RegMix baseline, response-prediction diagnostics, and allocation analyses.

Section 2 reviews related work. Section 3 presents the method, Section 4 describes the experiments, and Section 5 reports forecasting results and supporting analyses. Sections 6 and 7 discuss the scope and conclude.

2 RELATED WORK

2.1 LEARNING PRETRAINING MIXTURES

Learning a pretraining mixture determines how a fixed training budget is distributed across heterogeneous sources. Domain-based methods estimate how group proportions affect model performance. DoReMi learns domain weights with a small proxy using reference-relative losses and group distributionally robust optimization (Xie et al., 2023a). Data Mixing Laws fit mixture-dependent response functions together with model-size and training-step scaling laws (Ye et al., 2025). Aioli estimates a mixing law during training and updates allocations online (Chen et al., 2025).

RegMix (Liu et al., 2025) is the closest response-modeling approach to our work: it trains small language models on different mixtures, fits regressors to their outcomes, and searches the predicted response to select data for larger-model training. DATS-MIX uses this fixed-allocation approach to study TSFM mixture representation. Our Ratio+domain ablation shares TabPFN settings, observations, and the candidate bank with DATS-MIX, while omitting temporal statistics.

Quantitative data descriptions complement these strategies. DSIR matches target distributions through importance resampling in a hashed n-gram space (Xie et al., 2023b); RedPajama provides document-quality signals and metadata (Weber et al., 2024). Domain2Vec uses meta-domain distributions to guide mixture choice (Zhang et al., 2025), and CausalMix conditions mixture policies on data-pool statistics in supervised fine-tuning (Tang et al., 2026). We instead describe the temporal properties of contexts sampled under a candidate mixture to predict measured forecasting performance.

2.2 PRETRAINING TIME-SERIES FOUNDATION MODELS

TSFMs pretrain on sequences from different domains, frequencies, and channel structures. Chronos scales and quantizes observations into a discrete vocabulary and trains a Transformer with cross-entropy on real and synthetic series (Ansari et al., 2024). TimesFM uses a patched decoder-only architecture to forecast from variable-length histories (Das et al., 2024). Moirai combines a masked encoder, multiple patch sizes, and attention across variates (Woo et al., 2024). Chronos-2 combines group attention with multi-step quantile prediction and synthetic multivariate tasks (Ansari et al., 2025). These designs enable heterogeneous data to share a model; choosing source-sampling probabilities is a separate data-design decision.

Recent composition studies also examine what to mix. Nagpal et al. (2026) compare individual and mixed synthetic generators and real–synthetic compositions; their question differs from selecting proportions within a fixed real-source pool using a smaller forecasting proxy. Temporal statistics and data-quality methods characterize that data. Catch22 provides compact, interpretable descriptions of dynamics (Lubba et al., 2019). Attribute-guided pretraining selection relates training and evaluation data through temporal-feature neighborhoods (Saurav et al., 2026), while TSRating learns quality ratings from LLM judgments (Wu et al., 2026). Rather than rating individual sequences or matching source and target feature distributions, DATS-MIX predicts complete-mixture outcomes using actual forecasting experiments. Its source moments are pooled under the training sampler, with the forecasting architecture and recipe held fixed.

Compared with the language-model literature, end-to-end evidence linking reduced-scale proxy selection to larger-model TSFM pretraining remains limited. Our study examines whether a source mixture selected through inexpensive forecasting experiments is useful under a larger training recipe, and whether temporal statistics improve selection within that pipeline.

3 DATS-MIX

3.1 REPRESENTING A DATA MIXTURE

DATS-MIX represents a candidate training mixture by its source proportions, domain shares, and temporal statistics. Let K source datasets belong to G application domains, with $g(i)$ denoting the domain of source i. The source-proportion vector $w = (w_1, \ldots, w_K) \in \Delta^{K-1}$ specifies how frequently each dataset is sampled: $w_i \geq 0$ and $\sum_i w_i = 1$. Our study uses $K = 28$ sources and $G = 7$ domains.

Domain shares. Summing the probabilities of sources in each domain gives the domain-share vector $d(w) \in \mathbb{R}^G$,

$$d_g(w) = \sum_{i:g(i)=g} w_i, \qquad g = 1, \ldots, G. \tag{1}$$

Thus, w identifies the dataset allocation, while $d(w)$ expresses the same allocation at the domain level.

Temporal statistics. For a univariate context u sampled from a channel of a source dataset, let $\phi(u) \in \mathbb{R}^{14}$ describe its temporal characteristics. The statistics comprise four spectral summaries

(spectral entropy, dominant period relative to context length, peak energy ratio, and high-frequency energy ratio), two increment statistics (difference-variance ratio and median absolute increment), two half-window comparisons (mean difference and log-variance ratio), and six autocorrelations at lags $1, 2, 4, 8, 16, 32$. Appendix B provides detailed definitions and preprocessing for each temporal statistic.

First and second moments. Let P_i be the distribution obtained by drawing an eligible channel-row uniformly in source i, then an admissible start uniformly in that row. We call the statistics sampler-consistent because their contexts follow this training sampler and the same context normalization and arcsinh transform. We summarize each source by

$$\boldsymbol{\mu}_i = \mathbb{E}_{u\sim P_i}[\phi(u)], \qquad \boldsymbol{\nu}_i = \mathbb{E}_{u\sim P_i}[\phi(u) \odot \phi(u)], \tag{2}$$

where $\odot$ is elementwise multiplication. For $P_w = \sum_i w_i P_i$, the corresponding mixture moments are

$$\boldsymbol{\mu}(w) = \sum_i w_i \boldsymbol{\mu}_i, \qquad \boldsymbol{\nu}(w) = \sum_i w_i \boldsymbol{\nu}_i. \tag{3}$$

The empirical estimates $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\nu}}_i$ average the statistics and their squares over the same sampled contexts. These are marginal second raw moments, not cross-statistic covariances. Together with the means, they determine marginal variances as $\widehat{\boldsymbol{\nu}}(w) - \widehat{\boldsymbol{\mu}}(w) \odot \widehat{\boldsymbol{\mu}}(w)$. We supply raw second moments because they pool linearly under the mixture, as in Equation (3); the marginal variances follow from both moment blocks jointly and are quadratic in w. Channel- and context-level heterogeneity contributes through the source summaries; selection changes source probabilities, not within-source sampling.

Joint descriptor and ablations. The DATS-MIX descriptor has 63 coordinates. Mean-only removes the 14 second moments, while Ratio+domain removes all temporal statistics:

$$\begin{aligned} x_{\mathrm{DaTS}}(w) &= [w, d(w), \widehat{\boldsymbol{\mu}}(w), \widehat{\boldsymbol{\nu}}(w)] \in \mathbb{R}^{63}, \\ x_{\mathrm{mean}}(w) &= [w, d(w), \widehat{\boldsymbol{\mu}}(w)] \in \mathbb{R}^{49}, \\ x_{\mathrm{ratio}}(w) &= [w, d(w)] \in \mathbb{R}^{35}. \end{aligned} \tag{4}$$

All three condition the same response predictor on the same outcomes. Here Ratio refers to source proportions; Mean-only retains both source proportions and domain shares. For a fixed source pool, the pooled coordinates are linear functions of w, so the comparisons evaluate the representation supplied to the predictor rather than access to additional outcomes.

3.2 LEARNING MIXTURE PERFORMANCE FROM PROXY EXPERIMENTS

A small forecasting model provides measured outcomes for learning the relationship between mixture descriptors and forecasting performance. For each observed mixture w_j, we train the proxy from scratch and record its selection-panel loss y_j. This produces an observation table

$$\mathcal{O} = \{(x(w_j), y_j)\}_{j=1}^{N}, \qquad y_j = J_G^{\mathrm{sel}}(w_j; a_{\mathrm{proxy}}, U_{\mathrm{proxy}}, s_{\mathrm{proxy}}), \tag{5}$$

where a_{proxy}, U_{proxy}, and s_{proxy} specify the forecasting architecture, update budget, and common training seed; the loss J_G is defined in Section 3.3. Our observation table contains $N = 128$ outcomes from 7.07M-parameter proxies trained for 2,000 updates.

TabPFN-3 (Grinsztajn et al., 2026) conditions on the descriptor–outcome pairs and uses the mean of its predictive distribution as the score $\widehat{J}(w)$ of an untested mixture. The response predictor receives mixture vectors rather than time-series windows. The two ablations and DATS-MIX share outcomes, TabPFN settings, and the candidate bank; only the descriptor differs. Response prediction and selected-mixture forecasting are evaluated separately.

3.3 FORECASTING OBJECTIVE AND MIXTURE SELECTION

An allocation determines the distribution of training windows. Source i contains R_i eligible univariate channel-rows; row r has a set $\mathcal{T}_{ir}$ of admissible window starts, each leaving the context and forecast target within the training boundary. With $T_{ir} = |\mathcal{T}_{ir}|$, the sampler draws

$$i \sim \mathrm{Categorical}(w), \qquad r \sim \mathrm{Uniform}\{1, \ldots, R_i\}, \qquad t \sim \mathrm{Uniform}(\mathcal{T}_{ir}). \tag{6}$$

Thus $\Pr(i, r, t \mid w) = w_i/(R_i T_{ir})$, and multivariate records contribute individual channel-windows.

For a model trained under w, let $L_i(w) > 0$ be the native quantile loss on dataset i, aggregated over evaluation origins and channel-rows in the model's context-normalized, arcsinh-transformed target space. The evaluation objective is

$$J_G(w) = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{D}_g} \log L_i(w), \qquad \mathcal{D}_g = \{i : g(i) = g\}. \tag{7}$$

Lower values are better. The full notation $J_G^p(w; a, U, s)$ identifies the evaluation panel $p \in \{\text{sel}, \text{conf}\}$, architecture a, number of optimizer updates U, and seed s; these are suppressed when fixed. We write J_7 for J_G with $G = 7$; full-size results use the confirmation panel unless stated otherwise. Details of the native loss are given in Appendix B.2.

After conditioning the response predictor, we select the mixture with the lowest predicted loss from a fixed candidate bank:

$$\widehat{w} = \arg\min_{w \in \mathcal{U}} \widehat{J}(w), \qquad \mathcal{U} = \{w^{(1)}, \ldots, w^{(10^6)}\}, \quad w^{(j)} \sim \text{Dirichlet}(\mathbf{1}_K). \tag{8}$$

Candidates are drawn uniformly with respect to simplex volume without using outcomes. Ties in $\widehat{J}(w)$ are broken by candidate index. The selected weights are then frozen for full-size training and confirmation-panel evaluation. Algorithm 1 in Appendix B.7 summarizes the procedure.

4 Experimental setup

Sources, panels, and forecasting models. We use 28 sources from UTSD-12G (Liu et al., 2024) spanning energy, environment, health, IoT, nature, transportation, and web. This heterogeneous pool resembles the real-data component of Chronos pretraining and shares several source datasets with it (Ansari et al., 2024). The fixed selection panel supplies proxy-response labels, and a distinct confirmation panel evaluates the larger models. The native evaluation uses later windows from the training sources, rather than unseen datasets. Both panels were inspected during development; the training boundaries and row/origin construction are given in Appendix B.3. Both models use the Chronos-2 architecture (Ansari et al., 2025), trained from random initialization with context 512 and horizon 96. The proxy has 7.07M parameters and 2,000 training updates; throughout, full-size denotes the 118.74M-parameter target model trained for 20,000 updates. Batch size is 64, with the same optimizer and schedule family (Appendix B.6).

Observed mixtures and candidate search. The observation design contains 128 mixtures: three fixed anchors, 50 global, 50 within-domain, and 25 size-informed mixtures. The generated components use logistic-normal constructions, with cap, threshold, and rounding rules for size-informed proposals (Appendix B.5). Each observation is an actual single-seed proxy selection loss. Each response predictor searches the same one-million-candidate bank drawn from a flat Dirichlet distribution over the 28 sources, with no added observed anchors. Logistic-normal observation generation and Dirichlet candidate search therefore serve distinct roles.

Comparisons and measurements. Size-proportional sampling is a fixed rule based on admissible-start counts. DATS-MIX and its Mean-only and Ratio+domain ablations share the TabPFN-3 checkpoint and constructor settings, proxy outcomes, and candidate search, with inputs defined in Equation (4). RegMix (Liu et al., 2025) with a seven-dimensional domain-share input, denoted RegMix (d_7), fits LightGBM to the same 128 labels, averages the domain shares of its 100 lowest-scored bank candidates, and distributes each domain's mass in proportion to admissible starts. It is a separate pipeline baseline; the matched ablations hold the predictor and selection rule fixed. For native 20k results, we report the mean and sample SD over three paired seeds, relative loss $100[\exp(\overline{\Delta}) - 1]$, favorable seed counts, and two-sided paired-t p-values. All reported p-values are unadjusted; Section 6 describes the scope of these comparisons.

Scale-normalized evaluation. GIFT-Eval (Aksu et al., 2024) is a public benchmark covering 23 datasets, seven domains, and ten sampling frequencies. Its open data and evaluation code provide a documented reference for complementary evaluation. Our fixed GIFT-Eval subset contains 17

datasets, 39 configurations, six domains, 11,915 contexts, and 457,671 valid target positions. The same 20k checkpoints use at most 512 context steps and predict 96 outputs; evaluation retains the configuration-specific short-horizon prefix (8–60 steps). Sampling and exclusions are specified in Appendix B.4. For configuration c, valid target positions $\mathcal{V}_c$, quantiles $\mathcal{Q} = \{0.1, \ldots, 0.9\}$, and pinball loss ρ_q, the weighted quantile loss (WQL) is

$$\mathrm{WQL}_c = \frac{2 \sum_{v \in \mathcal{V}_c} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \rho_q(y_v - \widehat{y}_{q,v})}{\sum_{v \in \mathcal{V}_c} |y_v|}. \tag{9}$$

Targets and forecasts are on their original scale, and all denominators are positive. Common contexts, masks, quantiles, and aggregation ensure a controlled comparison within this subset. We average configurations within datasets, datasets within domains, and domains equally. This WQL assessment complements the native transformed-space objective. The subset was inspected during development and is not the full benchmark or an untouched test; Appendix B.4 lists its datasets and overlap-audit scope. Tests use arithmetic paired WQL differences; Appendix C.5 also reports a separate log-ratio analysis.

Response-prediction evaluation. We also evaluate the response predictors on a fixed 48-mixture set disjoint from the 128 observed weight vectors, using two-seed mean proxy selection losses as targets. Predictions and bank selections are frozen before evaluation, but the responses for this set were observed during development, so this is a retrospective diagnostic. Its role is to assess prediction accuracy separately from the full-size outcomes of the selected mixtures. Paired-mixture bootstrap summaries use 10,000 resamples (Appendix D).

5 Results

5.1 Full-size forecasting with the selected mixture

At 20k updates, DATS-MIX attains mean confirmation $J_7 = 0.7127$, compared with 0.8275 for size-proportional sampling (Table 1). This is a 10.85% reduction in geometric-mean native forecasting loss, with lower loss in all three paired seeds ($p < 0.001$). The model, training budget, source pool, and evaluation panel are shared.

Table 1: Native confirmation loss at 20k updates. DATS-MIX uses first and second raw moments; Mean-only removes its second-moment block. All learned selectors use the same 128 proxy outcomes. Values are mean ± sample SD over three paired training seeds; lower is better.

Method	Predictor	Input	Dim.	J_7
Size-proportional	–	Fixed rule	–	0.8275 ± 0.0073
Ratio+domain	TabPFN-3	$[w, d]$	35	0.7700 ± 0.0192
RegMix (d_7)	LightGBM	d	7	0.7302 ± 0.0035
Mean-only	TabPFN-3	$[w, d, \widehat{\mu}]$	49	0.7238 ± 0.0124
DATS-MIX	TabPFN-3	$[w, d, \widehat{\mu}, \widehat{\nu}]$	63	0.7127 ± 0.0063

All four learned selectors reduce mean loss by 5.59% to 10.85% relative to size-proportional sampling. DATS-MIX and Mean-only achieve 10.85% and 9.85% reductions against this common reference, each favorable in all three seeds. Relative to Ratio+domain, reductions are 4.51% for Mean-only ($p = 0.116$) and 5.57% for DATS-MIX ($p = 0.028$), also favorable in all three seeds. The direct DATS-MIX–Mean-only difference is 1.10% (two of three seeds, $p = 0.277$). DATS-MIX is 1.73% below RegMix ($p = 0.021$, three of three seeds), a separate pipeline comparison. Table 11 gives all paired contrasts.

Figure 2 shows all five methods along the same 20k schedule. DATS-MIX has lower mean loss than Size-proportional, Ratio+domain, and RegMix at each checkpoint. Its mean is also below Mean-only from 2k through 20k, by small margins at 2k and 15k. These checkpoints describe the same training runs, not independent replications. Individual checkpoint summaries are in Appendix Table 13.

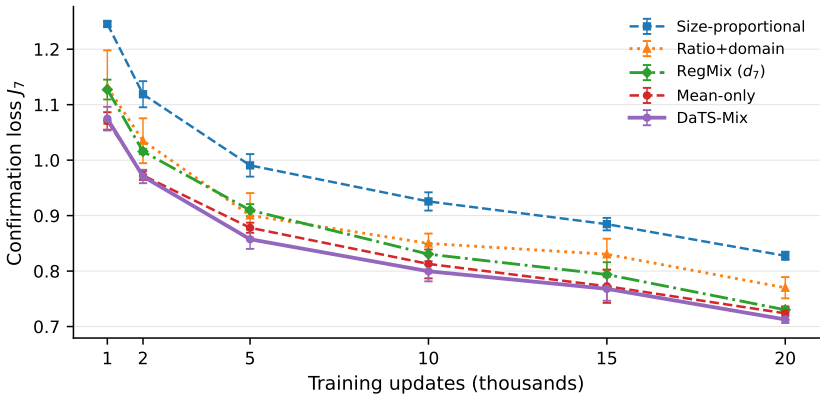

Figure 2: Confirmation trajectories for five frozen mixtures. Learned selectors use 128 proxy observations; Size-proportional is a fixed reference. Points are three-seed means and error bars are sample SDs. All checkpoints belong to the same 20k-update schedule; Mean-only removes the second moments from the DATS-MIX descriptor.

5.2 SCALE-NORMALIZED QUANTILE FORECASTING

On the GIFT-Eval subset, DATS-MIX achieves WQL 0.23593, the lowest mean among the five methods (Table 2). It improves over Size-proportional, Ratio+domain, and RegMix in all three seeds (raw-difference $p = 0.040$, 0.122, and 0.068, respectively). Relative to Mean-only, its raw WQL difference is -0.003037, favorable in two of three seeds ($p = 0.220$). Appendix C.5 reports the raw differences and a separate, consistently applied log-ratio analysis.

Table 2: Weighted quantile loss on the GIFT-Eval subset at 20k updates. The same three training seeds are evaluated for every method. Values are mean ± sample SD; lower is better.

Method	WQL
Size-proportional	0.24818 ± 0.00453
Ratio+domain	0.25174 ± 0.00973
RegMix (d_7)	0.24774 ± 0.01083
Mean-only	0.23897 ± 0.00720
DATS-MIX	0.23593 ± 0.00544

5.3 RESPONSE PREDICTION ON THE FIXED 48-MIXTURE SET

The complete descriptor lowers prediction MAE from 0.01945 to 0.01767 and RMSE from 0.02549 to 0.02183 relative to Ratio+domain, while Spearman correlation increases from 0.619 to 0.677 (Table 3). Mean-only has comparable prediction errors: its MAE is slightly lower, whereas its RMSE is slightly higher.

Table 3: Response prediction on the fixed 48-mixture set after conditioning on the same 128 observations. Targets are two-seed mean proxy selection losses. MAE, RMSE, and regret are lower-is-better; Spearman correlation is higher-is-better.

Predictor	MAE	RMSE	Spearman	Regret
Ratio+domain	0.01945	0.02549	0.619	0.02090
Mean-only	0.01753	0.02225	0.671	0.02090
DATS-MIX	0.01767	0.02183	0.677	0.02090

All three predictors have finite-set regret 0.02090, comparing the measured loss of the preferred member with the minimum among the 48 measured responses. The common regret does not

distinguish these representations at the selection level on this set; the prediction-error differences and their bootstrap summaries are in Appendix D. This diagnostic is distinct from the full-size utility of the selected training mixture. The 48 target means have sample SD 0.03100, and the mean absolute difference between their two proxy-seed outcomes is 0.02203. Appendix D also reports the set-separation check.

5.4 TEMPORAL STATISTICS UNDER THE TWO SAMPLING ALLOCATIONS

Table 4 compares the 1st and 2nd raw moments of all 14 temporal statistics under size-proportional sampling and DATS-MIX.

Table 4: First and second raw moments of context-level temporal statistics under size-proportional sampling and DATS-MIX. Size-proportional weights follow admissible training-window counts. Both use the same source statistics computed after context normalization and the arcsinh transform. Second raw moments are not variances. Values are rounded to four decimals; Appendix A.2 gives all five methods.

	1st moment		2nd moment	
Statistic	Size-proportional	DATS-MIX	Size-proportional	DATS-MIX
Spectral entropy	0.4971	0.4926	0.2880	0.2771
Dominant period / context	0.4225	0.4143	0.3220	0.3182
Peak energy ratio	0.3285	0.3417	0.1530	0.1591
High-frequency energy ratio	0.0623	0.0513	0.0152	0.0127
Difference-variance ratio	0.2680	0.2626	0.2173	0.2273
Median absolute increment	0.1262	0.1207	0.0379	0.0296
Half-window mean comparison	−0.0068	−0.0064	0.2761	0.2185
Half-window log-variance comparison	0.0268	0.1563	0.7979	2.8907
ACF lag 1	0.8628	0.8640	0.7814	0.7879
ACF lag 2	0.7690	0.7565	0.6672	0.6466
ACF lag 4	0.6880	0.6101	0.5677	0.4685
ACF lag 8	0.5244	0.3921	0.4067	0.3121
ACF lag 16	0.3184	0.2118	0.2335	0.1917
ACF lag 32	0.1307	0.0659	0.1416	0.1188

Spectral structure and local roughness. The slightly lower spectral entropy (0.4971 to 0.4926) indicates greater spectral concentration, a property related to second-order forecastability (Goerg, 2013), not a direct measure of TSFM training difficulty. The high-frequency mean falls from 0.0623 to 0.0513, and the second moment of median absolute increments from 0.0379 to 0.0296. Our high-frequency band has periods of at most four steps, shorter than a 16-step patch. These changes indicate lower average normalized roughness, not that all rapid components are noise. If some rapid fluctuations are weakly predictable over the 96-step horizon, reducing their emphasis could make the fixed training budget more useful.

Heterogeneity of temporal dependence. Lag-1 autocorrelation remains near 0.86, while lag-16 and lag-32 means decrease: the mixture retains local dependence without simply favoring persistence at every scale. Joint moments also reveal differences hidden by means. Although the lag-8 second moment decreases, its marginal variance $v_f = \widehat{\nu}_f - \widehat{\mu}_f^2$ increases from 0.1317 to 0.1584. This is greater dispersion of a dependency statistic across contexts, rather than evidence of wider support. Exposure to less uniform dependencies could encourage a shared forecaster to learn different conditional patterns.

Within-context scale contrasts. For the half-window log-variance statistic, the second moment rises from 0.7979 to 2.8907 and the marginal variance from 0.7972 to 2.8662. Whole-context normalization does not equalize the two halves. When these contrasts reflect genuine temporal changes, their greater dispersion could help learn context-dependent predictive scales, consistent with the relevance of non-stationarity to forecasting (Liu et al., 2022). Near-flat halves can also magnify this statistic through its 10^{-6} stabilizer; it is not by itself proof of regime diversity. These complementary descriptors suggest potential learning benefits, not isolated causal effects (Appendix A.2).

5.5 DOMAIN-LEVEL FORECASTING PERFORMANCE

DATS-MIX lowers mean confirmation loss in six of seven domains relative to Ratio+domain; transportation is 0.50% higher (Table 5). Appendix A reports all source probabilities and the domain heatmap; Appendix C.3 gives paired contrasts.

Table 5: Domain training shares (%) and native loss changes of DATS-MIX relative to Ratio+domain. The Lower column counts paired seeds in which DATS-MIX has lower loss. All models use the same 28-source training pool.

Domain	Ratio+domain	DATS-MIX	Relative loss	Lower
Energy	21.26	12.24	-13.06%	3/3
Environment	14.21	18.18	-3.47%	3/3
Health	13.69	12.56	-2.09%	3/3
IoT	8.71	14.33	-5.14%	3/3
Nature	24.86	20.15	-6.79%	3/3
Transportation	16.49	8.55	$+0.50\%$	1/3
Web	0.78	13.99	-8.27%	3/3
Overall	–		-5.57%	3/3

Sensitivity to evaluation weights. Without reselecting mixtures, we also average the 28 dataset log losses equally, $J_{\text{data}} = 28^{-1} \sum_i \log L_i$. Table 6 reports improvements over Ratio+domain, RegMix, and size-proportional sampling in all three seeds. Mean loss is lower than Ratio+domain on 19 of 28 datasets; Table 16 gives detailed contrasts.

Table 6: Evaluation-weight sensitivity: relative change in the equally weighted mean of 28 dataset log losses. Models and evaluation windows are unchanged.

Contrast	Relative loss	p	Lower
DATS-MIX − Ratio+domain	-5.56%	0.020	3/3
DATS-MIX − Mean-only	-0.76%	0.296	2/3
DATS-MIX − RegMix (d_7)	-4.38%	< 0.001	3/3
DATS-MIX − Size-proportional	-12.13%	< 0.001	3/3

6 DISCUSSION AND LIMITATIONS

Analysis scope. The results concern selected-mixture utility, not cross-scale rank preservation. Paired comparisons are exploratory; unadjusted p-values do not account for model-development or representation-selection decisions. Table 11 retains all contrasts. The three seeds vary fixed-mixture training, not observation collection, candidate selection, or TabPFN random states. Evaluation sets were inspected during development.

Scope and cost. We evaluate one source pool, architecture family, recipe, and candidate distribution, without alternative coordinate controls or broader fixed-rule baselines. Proxy acquisition is about 0.76 full-size runs by parameter-count times window-count, excluding candidate scoring; this is not a runtime or compute-matched comparison (Appendix B.6). The tenfold larger full-size window budget also increases repeated-window exposure for small sources.

7 CONCLUSION

The DATS-MIX mixture reduces mean confirmation loss by 10.85% versus size-proportional sampling ($p < 0.001$), compared with 9.85% for Mean-only; both improve in all three seeds. Reductions are 5.57% versus Ratio+domain ($p = 0.028$) and 1.73% versus RegMix ($p = 0.021$), the latter comparing distinct selection pipelines. DATS-MIX also attains the lowest mean WQL. These results support temporal statistics as descriptors for domain-aware, proxy-guided TSFM mixture selection.

AI USE STATEMENT

Generative AI assisted with research ideation, experimental-design feedback, literature review, method-implementation and analysis-code drafting and review, interpretation of experiment reports, manuscript drafting and translation, figure preparation, and numerical checks. The authors operated the training and evaluation experiments and made the experimental decisions; AI was a coding and analytical aid, not an independent experimental operator. Reported outcomes come from recorded runs, not AI-generated measurements. Verification included comparing tables with per-seed outputs, recomputing paired contrasts, checking source weights and temporal-statistics identities, and inspecting rendered figures and manuscripts. The authors are responsible for the scientific content, implementation, and citations.

REPRODUCIBILITY STATEMENT

The supplementary materials include a numerical-reproduction package and a fixed-mixture training implementation. The supplementary material provides the exact observed and selected mixtures, proxy labels, source statistics, native and WQL results, per-seed training trajectories, predictions for the fixed 48-mixture set, and scripts that regenerate the tables and plots. The RegMix top-100 candidate mapping is retained with the selection metadata. The accompanying code provides Chronos-2 scratch training and native evaluation for the frozen mixture selected by DATS-MIX, including its FP64 weights, model configuration, pinned dependencies, deterministic sampler, checkpoint/resume support, and execution instructions. Reference hashes and CPU validation records support consistency checks. Running the trainer requires the original UTSD-12G Arrow snapshot and the specified CUDA environment; corpus and checkpoint tensors are not redistributed. Appendix B.6 specifies the validation coverage and the upstream selection and evaluation stages not implemented by this training supplement.

REFERENCES

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.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Mayee Chen, Michael Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. In *International Conference on Learning Representations*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10148–10167, 2024.

Georg Goerg. Forecastable component analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 64–72. PMLR, 2013. URL https://proceedings.mlr.press/v28/goerg13.html.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger,

Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026. URL https://arxiv.org/abs/2605.13986.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9881–9893, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32369–32399, 2024.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: CAnonical time-series CHaracteristics. *Data Mining and Knowledge Discovery*, 33: 1821–1852, 2019. doi: 10.1007/s10618-019-00647-x.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15630–15649. PMLR, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

Aaryan Nagpal, Debdeep Sanyal, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. Mix, Don't Pick: Why synthetic corpus composition matters for time series foundation model pretraining. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://icml-structured-fm-workshop.github.io/accepted-papers/.

Kumar Saurav, Vinamra Baghel, Ayush Jain, and Ranjini B. Guruprasad. Time series attributes guided pretraining data selection for time series foundation models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1461–1465, 2026. doi: 10.1109/ICASSP55912.2026.11462456.

Zinan Tang, Yukun Zhang, Shaomian Zheng, Zhuoshi Pan, Qizhi Pei, Dingnan Jin, Jun Zhou, Yujun Wang, and Biqing Huang. CausalMix: Data mixture as causal inference for language model training. *arXiv preprint arXiv:2607.01104*, 2026.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. RedPajama: An open dataset for training large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. Datasets and Benchmarks Track.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53140–53164, 2024.

Shunyu Wu, Dan Li, Wenjie Feng, Haozheng Ye, Jian Lou, and See-Kiong Ng. Rating quality of diverse time series data by meta-learning from LLM judgment. In *International Conference on Learning Representations*, 2026. URL

https://proceedings.iclr.cc/paper_files/paper/2026/hash/3e8abf66abca02c8aed4a5f6e742c8d3-Abstract-Conference.html.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data Mixing Laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, 2025.

Mozhi Zhang, Howe Tissue, Lu Wang, and Xipeng Qiu. Domain2Vec: Vectorizing datasets to find the optimal data mixture without training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 76139–76158, 2025.

A SOURCES AND SELECTED-MIXTURE STATISTICS

A.1 SOURCE DICTIONARY AND FROZEN PROPORTIONS

The 28-source pool contains 289,495 eligible channel-rows and 640,511,129 admissible context/target starts. It comprises three energy, three environment, eight health, one IoT, eleven nature, one transportation, and one web source. Overlapping starts are sampling opportunities rather than independent time series. The pool and domain labels follow UTSD-12G (Liu et al., 2024), with source datasets also used in Chronos pretraining (Ansari et al., 2024).

Evaluation follows Equation (7): each domain has weight $1/7$, divided equally among its datasets. Size-proportional training uses the current admissible-start counts, $w_i = T_i / \sum_j T_j$, where $T_i = \sum_{r=1}^{R_i} T_{ir}$ counts admissible starts across all eligible rows of source i. The row/start sampler is unchanged. Table 7 reports all 28 source probabilities. Figure 3 compares their sums within the seven domains for all five methods.

Table 7: All 28 source probabilities (%) for DATS-MIX, the Mean-only ablation, and the baselines. RegMix probabilities are its averaged domain shares distributed in proportion to admissible starts within each domain.

ID	Source	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
1	Australian electricity	0.1770	5.6670	0.1278	5.3989	5.8445
2	London smart meters	22.2823	15.0715	16.0867	9.4371	1.8495
3	Wind (4 seconds)	1.1542	0.5227	0.8333	0.0003	4.5414
4	Australia rainfall	1.7996	2.5565	4.9376	2.4112	0.2642
5	Beijing PM2.5	0.5661	1.9085	1.5533	0.8108	0.7281
6	Benzene concentration	2.5462	9.7411	6.9863	12.9798	17.1913
7	BIDMC32 HR	4.1004	4.5965	5.6286	0.4163	0.5997
8	IEEE PPG	0.3601	0.2773	0.4943	11.1352	0.9970
9	Motor imagery	4.3398	1.3882	5.9572	0.2133	5.6472
10	Pig arterial pressure	0.0316	0.4029	0.0434	2.7787	1.6308
11	Pig CVP	0.0316	0.3853	0.0434	0.0164	0.0740
12	Self-regulation SCP1	0.0510	3.3034	0.0700	0.9666	1.4051
13	Self-regulation SCP2	0.0934	0.2825	0.1283	3.8338	1.6735
14	TDBrain	4.1680	3.0566	5.7214	10.1489	0.5332
15	Baian	25.2027	8.7088	5.6488	3.8215	14.3294
16	EigenWorms	3.3231	3.3023	2.5191	0.3692	0.3913
17	ERA5 pressure	17.5786	2.5922	13.3255	2.2638	0.2729
18	ERA5 surface	8.7893	1.1351	6.6628	5.9988	2.8168
19	KDD Cup 2018	0.2787	3.7597	0.2112	0.5556	3.2922
20	Phoneme	0.0530	1.0904	0.0402	0.7345	4.6801
21	Saugeen daily flow	0.0030	3.4480	0.0023	1.6331	1.8663
22	StarLightCurves	0.2322	2.6947	0.1760	1.2367	0.1823
23	Sunspot	0.0109	1.9215	0.0082	2.6723	2.5900
24	Temperature / rain	0.1102	0.4524	0.0835	0.5406	1.0577
25	US births	0.0005	1.3187	0.0004	0.1948	2.4977
26	Worms	0.0040	3.1479	0.0030	0.6847	0.5068
27	Pedestrian counts	0.4478	16.4879	14.2852	12.6108	8.5500
28	Kaggle web traffic	2.2648	0.7806	8.4223	6.1362	13.9871

Relative to size-proportional sampling, DATS-MIX lowers the IoT share from 25.20% to 14.33%, energy from 23.61% to 12.24%, and nature from 30.38% to 20.15%. Environment rises from 4.91% to 18.18%, transportation from 0.45% to 8.55%, and web from 2.26% to 13.99%. Within energy, the wind source receives 4.54% of the total training probability, compared with 0.52% under Ratio+domain, while London smart meters falls from 15.07% to 1.85%. The source and domain probabilities describe the selected training distribution, not isolated source-level performance contributions.

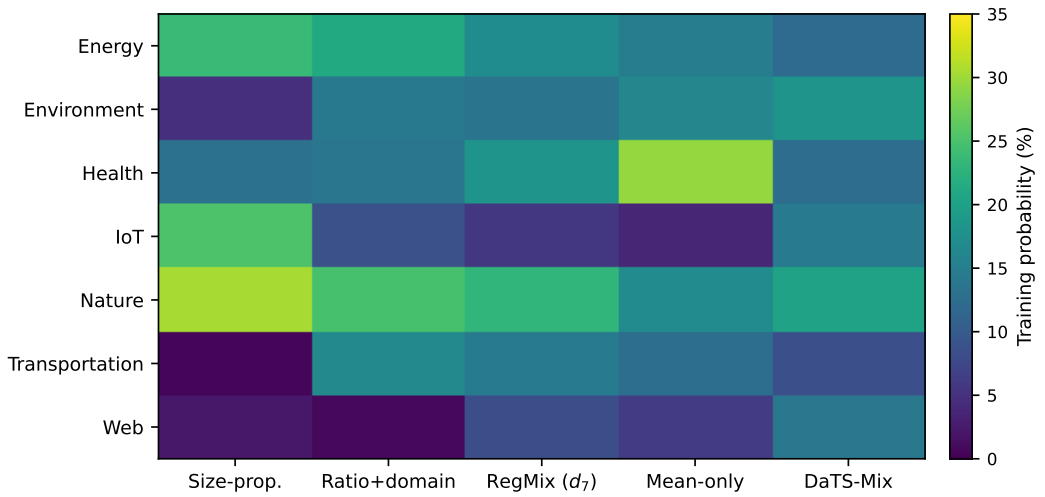

Figure 3: Domain shares of the frozen mixtures across the evaluated methods. Each cell is the percentage of source draws assigned to that domain, obtained by summing the source probabilities in Table 7. Columns sum to 100%, and a common color scale is used across methods. The values are training probabilities, not forecasting scores.

A.2 Complete first and second temporal moments

Table 8 reports all mixture means, Table 9 their second raw moments, and Table 10 the derived marginal variances. DaTS-Mix inputs the first two moments, whereas Mean-only inputs only the means. These summaries are deterministic for fixed weights and a source-statistics cache; they are not replicated across training seeds.

Table 8: Pooled means of all 14 temporal statistics for each frozen mixture. Values are computed from the same source-statistics cache.

Statistic	Size-prop.	Ratio+domain	RegMix	Mean-only	DaTS-Mix
Spectral entropy	0.497055	0.494206	0.532111	0.515402	0.492564
Dominant period / context	0.422454	0.344559	0.364244	0.343657	0.414269
Peak energy ratio	0.328458	0.346681	0.303602	0.321340	0.341729
High-frequency energy ratio	0.062327	0.046713	0.063783	0.046601	0.051258
Difference-variance ratio	0.267960	0.250522	0.304082	0.252701	0.262554
Median absolute increment	0.126223	0.123245	0.144397	0.141475	0.120719
Half-window mean comparison	−0.006782	−0.043239	−0.002152	−0.017432	−0.006386
Half-window log-variance comparison	0.026797	−0.005341	0.093033	0.065554	0.156288
ACF lag 1	0.862802	0.870418	0.844581	0.870699	0.864025
ACF lag 2	0.768989	0.758102	0.722297	0.755037	0.756515
ACF lag 4	0.688004	0.598548	0.587579	0.596608	0.610087
ACF lag 8	0.524434	0.315388	0.326613	0.302674	0.392055
ACF lag 16	0.318397	0.141322	0.161313	0.092515	0.211843
ACF lag 32	0.130742	0.021501	0.031824	0.003891	0.065943

Table 9: Pooled second raw moments of all 14 temporal statistics for each frozen mixture. Values are computed from the same source-statistics cache.

Statistic	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
Spectral entropy	0.288005	0.277743	0.315153	0.296249	0.277059
Dominant period / context	0.321993	0.258249	0.274856	0.251503	0.318249
Peak energy ratio	0.153013	0.161830	0.126861	0.141296	0.159077
High-frequency energy ratio	0.015208	0.009218	0.014574	0.009239	0.012718
Difference-variance ratio	0.217263	0.175024	0.230813	0.175342	0.227259
Median absolute increment	0.037923	0.028219	0.039304	0.033525	0.029580
Half-window mean comparison	0.276121	0.207134	0.180841	0.159477	0.218475
Half-window log-variance comparison	0.797932	0.989065	1.730018	1.381752	2.890650
ACF lag 1	0.781447	0.787332	0.748768	0.786644	0.787900
ACF lag 2	0.667201	0.634430	0.588478	0.625438	0.646601
ACF lag 4	0.567725	0.444139	0.429531	0.428789	0.468453
ACF lag 8	0.406690	0.282660	0.274148	0.235017	0.312147
ACF lag 16	0.233499	0.179175	0.155839	0.137060	0.191688
ACF lag 32	0.141605	0.119147	0.096067	0.094136	0.118806

Table 10: Derived marginal variances of all 14 temporal statistics for each frozen mixture. Values are computed from the same source-statistics cache. Variances equal $\widehat{\nu}_f(w) - \widehat{\mu}_f(w)^2$ and are not training-seed variability.

Statistic	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
Spectral entropy	0.040941	0.033504	0.032010	0.030610	0.034440
Dominant period / context	0.143525	0.139528	0.142183	0.133403	0.146631
Peak energy ratio	0.045128	0.041642	0.034687	0.038037	0.042298
High-frequency energy ratio	0.011324	0.007036	0.010506	0.007067	0.010091
Difference-variance ratio	0.145460	0.112263	0.138347	0.111484	0.158324
Median absolute increment	0.021991	0.013029	0.018454	0.013509	0.015007
Half-window mean comparison	0.276075	0.205264	0.180837	0.159173	0.218434
Half-window log-variance comparison	0.797214	0.989036	1.721363	1.377454	2.866224
ACF lag 1	0.037019	0.029705	0.035450	0.028527	0.041361
ACF lag 2	0.075857	0.059711	0.066765	0.055357	0.074287
ACF lag 4	0.094376	0.085878	0.084282	0.072848	0.096247
ACF lag 8	0.131658	0.183190	0.167472	0.143405	0.158440
ACF lag 16	0.132122	0.159203	0.129817	0.128501	0.146811
ACF lag 32	0.124511	0.118684	0.095055	0.094121	0.114457

Reading the moments as training-distribution summaries. For a fixed allocation, a first moment describes the average context statistic, whereas $v_f(w) = \widehat{\nu}_f(w) - \widehat{\mu}_f(w)^2$ describes its marginal dispersion. Neither is a calibrated learning-difficulty score. A higher variance need not imply broader support, more distinct patterns, or greater multivariate diversity: rare extremes can also increase it. The lag-8 example in Section 5.4 uses the unrounded moments, not the four-decimal entries in Table 4. The mixture also has a lower high-frequency-ratio variance (0.0113 to 0.0101) and half-window mean-difference second moment (0.2761 to 0.2185), so its changes are selective rather than a uniform increase in variability.

The spectral and dependency summaries characterize processed contexts, not forecast innovations. High-frequency periodic structure can remain predictable, and neither a lower autocorrelation nor a higher spectral entropy identifies redundant gradients or irreducible pinball loss. The RHO-LOSS framework (Mindermann et al., 2022) motivates considering learnability and usefulness rather than difficulty alone; it is not evidence that DATS-MIX implements that criterion. Likewise, heterogeneous within-context scale contrasts motivate the local-adaptation hypothesis but do not establish that the trained model changes its attention or quantile calibration in that way.

The half-window statistic is $\log[(v_{\text{second}} + 10^{-6})/(v_{\text{first}} + 10^{-6})]$. A nearly flat half and a non-flat half can produce a large magnitude. Its first two moments alone cannot determine the frequency of such contexts or distinguish temporal changes from near-flat or missing-value-affected segments. No

context-level tail or stabilizer-sensitivity audit is used to attribute the reported forecasting gains. The descriptor cache and preprocessing remain unchanged.

B TEMPORAL STATISTICS AND REPRODUCIBLE SETTINGS

B.1 TEMPORAL-STATISTIC DEFINITIONS

Each source moment estimate uses 4,096 calibration draws from the row/start sampler with seed 260907601. Draws are sampled contexts, not necessarily distinct windows. Contexts are viewed as the model sees them: FP32 context-only `InstanceNorm` with `eps=1e-5` and an arcsinh transform, after which statistics are computed in FP64. An unnormalized-context variance below 10^{-12}, a nonfinite input, or a nonfinite normalized view triggers the protected-window rule: invalid statistic values are replaced by zero before averaging, and every sampled context remains in the denominator.

Exact coordinate definitions. Let z_t, $t = 0, \dots, 511$, be the processed context after subtracting its mean, and let $\epsilon = 10^{-6}$. For the untapered real FFT, let $P_j = |\,\mathrm{rFFT}(z)_j|^2$, $j = 1, \dots, 256$, excluding DC, $p_j = P_j / \max\{\sum_k P_k, \epsilon\}$, and $j^* = \arg\max_j P_j$ (first maximizer). The four spectral coordinates are

$$-\frac{\sum_j p_j \log \max\{p_j, \epsilon\}}{\log 256}, \qquad \frac{1}{j^*}, \qquad \max_j p_j, \qquad \sum_{j:\, j/512 \geq 0.25} p_j, \tag{10}$$

where $1/j^*$ is the dominant period divided by the context length. With population variances and first differences $\Delta z_t = z_{t+1} - z_t$, the increment and half-window coordinates are

$$\frac{\mathrm{Var}(\Delta z)}{\max\{\mathrm{Var}(z), \epsilon\}}, \qquad \mathrm{median}\,|\Delta z|, \qquad \overline{z}_{256:511} - \overline{z}_{0:255}, \qquad \log \frac{\mathrm{Var}(z_{256:511}) + \epsilon}{\mathrm{Var}(z_{0:255}) + \epsilon}. \tag{11}$$

The two half-window coordinates are signed. Finally, autocorrelation at lag k is

$$\phi_{\mathrm{acf},k} = \frac{\sum_{t=0}^{511-k} z_t z_{t+k}}{\max\{\sum_{t=0}^{511} z_t^2, \epsilon\}}, \qquad k \in \{1, 2, 4, 8, 16, 32\}, \tag{12}$$

with no lag-specific rescaling of the denominator. Statistic values are unstandardized and unclipped across the source pool. We write $\widetilde{\phi}$ for the protected, finite statistic after the invalid-value rule; this is ϕ in Section 3. For processed statistic $\widetilde{\phi}_f$, the empirical source estimate and pooled descriptor are

$$\widehat{\mu}_{if} = \frac{1}{M_i} \sum_{k=1}^{M_i} \widetilde{\phi}_f(u_{ik}), \qquad \widehat{\nu}_{if} = \frac{1}{M_i} \sum_{k=1}^{M_i} \widetilde{\phi}_f(u_{ik})^2. \tag{13}$$

Here $M_i = 4{,}096$ is the number of calibration draws for source i; mixture estimates pool both quantities with the same weights, $\widehat{\mu}_f(w) = \sum_i w_i \widehat{\mu}_{if}$ and $\widehat{\nu}_f(w) = \sum_i w_i \widehat{\nu}_{if}$. The marginal variance decomposes into $\sum_i w_i(\widehat{\nu}_{if} - \widehat{\mu}_{if}^2)$ within sources and $\sum_i w_i(\widehat{\mu}_{if} - \widehat{\mu}_f(w))^2$ between sources. Both refer to variation of the context statistic, not raw signal variance.

B.2 TRAINING LOSS AND EVALUATION REDUCTION

For each context/target window, the native loss applies the context-derived normalization and arcsinh transform to the future target using the context location and scale. Writing the transformed target as $\widetilde{y}_h$ and the prediction at quantile q as $\widehat{y}_{q,h}$, with $\mathcal{Q} = \{0.1, 0.2, \dots, 0.9\}$,

$$\ell_{iro} = \sum_{q \in \mathcal{Q}} \frac{1}{96} \sum_{h=1}^{96} m_h \, 2\rho_q(\widetilde{y}_h - \widehat{y}_{q,h}), \qquad \rho_q(e) = (q - \mathbf{1}[e < 0])e, \tag{14}$$

where m_h combines the target-validity and future-covariate masks and is identical across quantiles; the horizon denominator remains 96. Training averages ℓ over the batch. The evaluator records each context's loss and uses normalized origin and row weights,

$$L_i = \sum_r b_{ir} \sum_o a_{iro} \ell_{iro}, \qquad J_G = \frac{1}{G} \sum_g \frac{1}{|\mathcal{D}_g|} \sum_{i \in \mathcal{D}_g} \log L_i, \tag{15}$$

giving each selected row equal dataset weight and each of its origins equal row weight. Dataset aggregation precedes the logarithm, so $\exp(J_G(A) - J_G(B))$ is a weighted geometric mean of dataset native-loss ratios; it is not a ratio of raw-scale errors, an average training loss, or an arithmetic mean of per-dataset percentage changes. For a fixed pair of mixtures (A, B) with $S = 3$ paired training seeds, the reported contrast and relative loss are

$$\overline{\Delta} = \frac{1}{S}\sum_{s=1}^{S}[J(A,s) - J(B,s)], \qquad \text{RelLoss} = 100[\exp(\overline{\Delta}) - 1]. \tag{16}$$

B.3 Training boundaries and forecasting panels

For an eligible channel-row of length n, indices run from zero to $n-1$. Let $e = n - 96$ be the latest origin admitting a complete 96-step target, and define slack $b = e - 608 \geq 0$, the number of origins $n_o = \min\{8, 1 + \lfloor b/192 \rfloor\}$, and stride $\delta = \min\{512, \lfloor b/[2(n_o - 1)] \rfloor\}$ when $n_o > 1$ (zero otherwise). The origins and effective training boundary are

$$o_j = e - (n_o - 1)\delta + j\delta, \quad j = 0, \ldots, n_o - 1, \qquad e_{\text{train}} = o_0. \tag{17}$$

The last origin is e and its target ends at the exclusive index $e + 96 = n$. Training contexts and targets end at or before e_{train}. Each evaluation target occupies $[o_j, o_j + 96)$ and uses $[o_j - 512, o_j)$ as its context.

When a source has at least 128 eligible rows, a fixed permutation selects disjoint groups of 64 rows for the two panels; smaller sources use all eligible rows in both panels. For $n_o > 1$, selection uses the first $\lceil n_o/2 \rceil$ origins and confirmation the rest. For $n_o = 1$, both use that origin only when their row groups are disjoint; the constructor rejects the case of a source with fewer than 128 rows and a single origin rather than duplicating the same row/origin pair. The panel seed is 260907005. The panel construction verifies distinct row/origin pairs and post-boundary targets. Historical contexts may overlap, and both panels were inspected during development.

The fixed 48-mixture diagnostic uses the selection panel with forecasting seeds 2609157211 and 2609157212, and a Dirichlet mixture generator with seed 2609157311. Its outcome table was available during method development; predictions and candidate-bank selections were fixed before this retrospective diagnostic. It is not an untouched forecasting test.

B.4 GIFT-Eval subset construction

The metric and results appear in Sections 4 and 5.2. The 17 dataset identifiers are `LOOP_SEATTLE`, `SZ_TAXI`, `bitbrains_fast_storage`, `bitbrains_rnd`, `bizitobs_application`, `bizitobs_l2c`, `bizitobs_service`, `car_parts_with_missing`, `covid_deaths`, `electricity`, `ett1`, `ett2`, `hierarchical_sales`, `hospital`, `jena_weather`, `restaurant`, and `solar`. The preparation code excludes the GIFT-Eval collections corresponding to US births, Saugeen daily flow, Temperature / rain, and KDD Cup 2018 because their collection families overlap training sources. It also excludes `M_DENSE` for unresolved provenance and the M4 collections because their underlying collection provenance could not establish family disjointness. These rules precede model-score inspection during subset construction; the resulting subset was subsequently inspected during development.

For each configuration, the preparation code uses NumPy `default_rng` with the pair consisting of seed 2609091401 and the integer represented by the first eight hexadecimal SHA256 digits of its configuration name. A random permutation selects at most 256 eligible channel-rows. Eligibility requires length at least $16 + h_c$, where h_c is the configuration's short forecast horizon. At most three trailing forecast origins are used per channel-row. Specifically, with minimum series length $n_{\min,c}$ in the configuration, the number is $\min\{3, \max\{1, \lceil 0.1 n_{\min,c}/h_c \rceil\}\}$. Origins are $n - jh_c$ for the corresponding trailing values of j. A window is retained if its context has at least 16 finite values and its target at least one. Contexts use the last at most 512 values, with NaN left-padding. Evaluation uses only the first h_c outputs of the 96-output forecast, masking nonfinite targets. The included horizons are 8, 12, 30, 48, and 60; medium- and long-horizon benchmark settings are not evaluated.

The supplementary configuration manifest specifies frequencies, horizons, selected-row counts, valid-window counts, and target counts for all 39 configurations. It matches the fixed subset's 11,915

contexts and 457,671 valid targets. The original family audit also records no exact full-length finite-series hash matches with training series. This recorded check does not cover all partial or transformed duplicates. Accordingly, the reported scores are controlled comparisons on this documented GIFT-Eval subset, not official full-benchmark scores or proof of complete cross-corpus independence. Construction settings and original file hashes are provided in `data/verified_metadata/`.

B.5 Observation design and response predictors

The observation set contains 128 measured proxy outcomes: three fixed anchors, 50 global mixtures, 50 within-domain mixtures, and 25 size-informed mixtures. The 125 generated rows follow a 2:2:1 allocation across the three strata. The budget gives comparable emphasis to changes between domains and within domains, complemented by proposals that reflect data availability. All proxy labels use forecasting seed 260907610 and the same 2k-update recipe.

Hierarchical logistic-normal proposals. For independent standard-normal vectors z and z_g, global mixtures use

$$\pi = \mathrm{softmax}(\sigma z), \quad v_g = \mathrm{softmax}(\tau_g z_g), \quad w_i = \pi_{g(i)} v_{g(i),i}, \tag{18}$$

where $\sigma \sim U(0.3, 2)$ and each $\tau_g \sim U(0.15, 2)$. Within-domain mixtures use the same internal construction with $\pi_g = 1/7$. This separates the spread of domain shares from that of source proportions inside each domain. The fixed anchors are uniform-source, the historical source-size prior, and equal-domain with uniform proportions inside each domain. Their exact weights are included with the observed data. The historical size prior is used for design; the fixed baseline instead uses the current admissible-start inventory. Their total-variation distance is 0.09411, and the largest source-share difference is 3.58 percentage points (Baian, the only IoT source). Both vectors are provided in `data/prior_and_canonical.csv`; they are not interchangeable.

Size-informed logistic-normal proposals. Let p be the historical source-size generation prior and $q_i = \sqrt{p_i} / \sum_j \sqrt{p_j}$. A proposal is

$$w = \mathrm{softmax}(\log q + z), \qquad z \sim N(0, I_{28}). \tag{19}$$

The logit scale is one; q specifies the logit location, not the arithmetic expectation of the final weights. Proposals are rejected if a component exceeds $\min(15p_i, 1)$. Entries below 0.0002 are zeroed, weights are normalized, rounded to a 0.0002 grid, lexicographically ordered, and deduplicated by adjacent-row L_1 distance 10^{-5}. Exact weight hashes exclude duplicates of existing observations, the fixed 48-mixture set, and previously generated candidate weights. Uniform sampling without replacement chooses the required number of eligible rows, which are then normalized again. The resulting observation table is therefore a stratified, postprocessed design, not an IID sample from one logistic-normal law.

Observation records and random-number settings. The 128 rows are specified by generator family and index independently of their response values. Previously measured labels are reused only with matching weight, seed, recipe, and evaluation records. The global/within generator uses PCG64 via NumPy `default_rng(20260917)`. The size-informed proposals were generated in two blocks: the first 18 use NumPy MT19937 and Python `random` seeded at 2609089504, and the remaining seven use seed 2609259711. The second block completes the 25-row size-informed stratum; these seeds document generation order, not a different design size. Each iteration generates 15 Gaussian-logit draws and chooses one uniformly before applying the component-cap filter; the other 14 are not screened or ranked. This is distributionally equivalent to one draw before filtering, but the 15-draw construction preserves the specified RNG stream. Each block uses 1,600 iterations and extends only until enough eligible rows are available. The supplementary files provide the exact weight/label arrays and row order for reproduction.

Candidate bank and TabPFN. All response predictors use the same candidate bank: one million FP64 $\mathrm{Dirichlet}(\mathbf{1}_{28})$ rows generated with PCG64 seed 2609089602. No observed anchors are added, and ties use the first original index. The DATS-MIX, Mean-only, and Ratio+domain selections are `U_SEL_0443203`, `U_SEL_0279861`, and `U_SEL_0249154`, respectively. All three are distinct from the 128 observed weight vectors.

We use TabPFN-3 (Grinsztajn et al., 2026) through package version 8.5.0. The checkpoint is `tabpfn-v3-regressor-v3_default.ckpt`, with 58,274,944 parameters. The original checkpoint-file hash and the canonical FP64 parameter digest are recorded separately in `data/verified_metadata/tabpfn_checkpoint_provenance.json`; the 128-observation preflight matches the latter to the loaded predictor. The software version alone is not used to identify model weights.

All three descriptors use the same checkpoint and constructor settings: FP64 inference, TF32 off, random state zero, `fit_mode=fit_preprocessors`, `n_estimators=auto`, `auto_scale_n_estimators=True`, and `softmax_temperature=0.9`. Version 8.5.0 can increase an automatic ensemble size for feature coverage. The resolved ensemble counts for the 35-, 49-, and 63-coordinate inputs were not exported, so the matched-control claim concerns the recorded settings, not an independently verified equality of those runtime counts. Candidate queries follow the same order in fixed 1,000-row batches. Scores are predictive means returned by `predict(X)`.

RegMix domain-only baseline. RegMix (d_7) uses LightGBM 4.7.0 with the same 128 actual proxy labels and original one-million-candidate bank. Three-fold validation keeps repeated domain-share inputs together: all 51 equal-domain observations lie in one validation fold. The other two validation folds each contain 26 observations, and the first contains 76. The saved fold assignments produce best iterations 603, 110, and 338; their median gives 338 rounds for fitting all labels. Exact assignments are included with the fit metadata. For the 100 lowest predicted-loss candidates, let $\overline{d}_g$ be their mean domain share. The training mixture is

$$w_i^{\mathrm{RegMix}} = \overline{d}_{g(i)} \frac{T_i}{\sum_{j:g(j)=g(i)} T_j}. \tag{20}$$

Thus it averages top-ranked domain mixtures and uses a fixed within-domain allocation, rather than choosing the single source-level argmin used by the three TabPFN selectors. The selected identifier is `REGMIX128_D7_TOP100`. This is the domain-only baseline, not a claim of reproducing every setting in the original RegMix paper. The supplementary selection metadata includes the top-100 identifiers, scores, and exact resulting weights.

B.6 Forecasting models and computation

Both models instantiate the Chronos-2 architecture and train from random initialization. The proxy has hidden size 256, feed-forward size 1024, six layers, and four heads; the larger model uses 768, 3072, twelve layers, and twelve heads. Both use key/value dimension 64 and dropout zero. The exact parameter counts are 7,069,728 and 118,740,000, respectively; the initialization check verifies the latter as $\sum_{\theta\in\Theta} \mathrm{numel}(\theta)$. Context length is 512, horizon 96, patch size and stride 16, and the quantiles are 0.1 through 0.9. AdamW uses peak learning rate 3×10^{-4}, weight decay 0.01, and gradient clipping at one. Warmup/stable/decay fractions are 2%/73%/25%, with final learning rate $0.1\times$ peak. Forward computation is BF16; parameters and optimizer states are FP32. Training examples are independent univariate channel-windows from the real source pool; the experiment does not reproduce public Chronos-2 pretraining or joint multivariate training. In contrast, the public model configuration specifies context length 8,192, up to 64 output patches, 21 quantiles, and dropout 0.1.[1] Our controlled configuration uses context 512, six output patches, nine quantiles, and dropout zero.

Batch size is 64 with accumulation one. Each proxy outcome corresponds to 2,000 successful updates and 128,000 sampled-window presentations; the larger model uses 20,000 successful updates and 1,280,000 presentations. The complete 128-observation set represents 16,384,000 proxy-window presentations. Source-statistic estimation uses $28 \times 4096 = 114{,}688$ calibration contexts. As a parameter-count-times-window-count proxy for workload, observation acquisition is $128(7{,}069{,}728)(128{,}000)/[(118{,}740{,}000)(1{,}280{,}000)] = 0.7621$ times one full-size run. This is not a measured FLOP or wall-clock ratio and excludes statistic estimation and candidate scoring. It does not establish a compute-matched advantage.

[1] Amazon, `chronos-2/config.json`, accessed 2026-09-26: `https://huggingface.co/amazon/chronos-2/blob/main/config.json`.

The six saved full-size checkpoints lie within one 20k-update schedule. The 10k checkpoint is not the endpoint of a separately trained 10k schedule. Full-size seeds are 2609118111, 2609118112, and 2609118113, paired across methods. The analysis includes previously completed runs when weight, seed, initialization, model, optimizer, and evaluation records match. The five reported mixtures are distinct and contribute three training runs each; checkpoint reuse is recorded as reuse rather than an additional replicate.

The supplementary material contains numerical records and scripts for the tables and plots, including all five methods' per-seed native checkpoints, per-dataset native losses, and configuration-level WQL scores. These numerical records support trajectory SDs and domain and dataset reaggregation; raw forecast arrays and checkpoint tensors are not redistributed.

The accompanying fixed-mixture training implementation provides scratch training and native evaluation of the mixture selected by DATS-MIX; the supplementary README identifies its entry point and configuration files. Its 28 FP64 weights exactly match `U_SEL_0443203`. Model settings, pinned dependencies, deterministic source/row/start sampling, checkpoint/resume instructions, and reference hashes are included; the original UTSD-12G Arrow snapshot must be supplied separately. The code starts after selection and does not implement proxy-observation collection, predictor fitting/search, or GIFT-Eval inference.

The accompanying `VALIDATION.json` records CPU self-tests and exact checks of three full-size initializations, optimizer/scheduler states, 18 sampling prefixes, and both native panels. It reports no new 20k GPU run or end-to-end resume validation. The numerical package reproduces the reported summaries without training; running the fixed-mixture trainer additionally requires the specified CUDA environment. Neither package redistributes corpus, trained checkpoint tensors, or fitted predictors.

B.7 SELECTION PROCEDURE

Algorithm 1 DATS-MIX mixture selection

Require: Source pool, sampler, context statistics ϕ, observed mixtures, candidate bank $\mathcal{U}$.
1: Compute source moments $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\nu}}_i$ from sampler-drawn contexts.
2: **for** each observed mixture w_j **do**
3: Obtain its measured 7.07M/2k forecasting selection loss y_j.
4: **end for**
5: Construct $x(w_j) = [w_j, d(w_j), \sum_i w_{j,i}\widehat{\boldsymbol{\mu}}_i, \sum_i w_{j,i}\widehat{\boldsymbol{\nu}}_i]$.
6: Condition TabPFN on $(x(w_j), y_j)$ and score the candidate bank.
7: Select $\widehat{w} = \arg\min_{w\in\mathcal{U}} \widehat{J}(w)$, breaking ties by index.
8: Freeze $\widehat{w}$, train the larger forecasting model, and evaluate confirmation loss.

C DETAILED FULL-SIZE RESULTS

C.1 MAIN ENDPOINT CONTRASTS

For each comparison, we form the loss difference within the same training seed, then report its mean $\overline{\Delta}$, standard error $s_\Delta/\sqrt{3}$, and two-sided paired-t p-value with two degrees of freedom. The numerical interval is $\overline{\Delta} \pm t_{0.975,2} s_\Delta/\sqrt{3}$. All tests are unadjusted. Section 6 describes the scope of these comparisons. Table 11 gives the main endpoint contrasts, and Table 12 reports every training seed.

Table 11: Native 20k endpoint contrasts from the same three paired seeds. Relative loss is $100[\exp(\overline{\Delta}) - 1]$. Two-sided paired-$t$ tests use two degrees of freedom; all p-values are unadjusted. Section 6 describes the scope of these comparisons.

Contrast	$\overline{\Delta} J_7$	Relative	p	95% interval	Lower
DATS-MIX − Ratio+domain	−0.057266	−5.57%	0.028	[−0.099605, −0.014928]	3/3
DATS-MIX − Mean-only	−0.011076	−1.10%	0.277	[−0.043276, +0.021124]	2/3
DATS-MIX − RegMix (d_7)	−0.017450	−1.73%	0.021	[−0.028482, −0.006417]	3/3
DATS-MIX − Size-proportional	−0.114804	−10.85%	< 0.001	[−0.123434, −0.106173]	3/3
Mean-only − Ratio+domain	−0.046190	−4.51%	0.116	[−0.120509, +0.028128]	3/3
Mean-only − Size-proportional	−0.103728	−9.85%	0.003	[−0.128124, −0.079332]	3/3
Ratio+domain − Size-proportional	−0.057537	−5.59%	0.040	[−0.108325, −0.006750]	3/3
RegMix (d_7) − Size-proportional	−0.097354	−9.28%	< 0.001	[−0.107365, −0.087343]	3/3
RegMix (d_7) − Ratio+domain	−0.039817	−3.90%	0.077	[−0.090346, +0.010713]	3/3
Mean-only − RegMix (d_7)	−0.006374	−0.64%	0.379	[−0.030835, +0.018088]	2/3

Table 12: Individual 20k confirmation J_7 values. Each column uses one fixed mixture across all three seeds.

Seed	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
2609118111	0.835461	0.762505	0.734233	0.736014	0.717518
2609118112	0.821020	0.755710	0.727836	0.724246	0.705623
2609118113	0.826124	0.791778	0.728475	0.711162	0.715053

Table 13: Confirmation trajectories along the same 20k-update schedules. Every entry is a three-seed mean ± sample SD calculated from individual checkpoint values.

Updates	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
1k	1.2455 ± 0.0043	1.1316 ± 0.0664	1.1272 ± 0.0179	1.0706 ± 0.0158	1.0747 ± 0.0213
2k	1.1188 ± 0.0235	1.0350 ± 0.0404	1.0156 ± 0.0048	0.9724 ± 0.0081	0.9705 ± 0.0121
5k	0.9905 ± 0.0203	0.9004 ± 0.0402	0.9098 ± 0.0109	0.8781 ± 0.0091	0.8574 ± 0.0174
10k	0.9255 ± 0.0164	0.8498 ± 0.0178	0.8305 ± 0.0188	0.8129 ± 0.0261	0.7997 ± 0.0182
15k	0.8846 ± 0.0112	0.8303 ± 0.0281	0.7938 ± 0.0221	0.7726 ± 0.0299	0.7680 ± 0.0215
20k	0.8275 ± 0.0073	0.7700 ± 0.0192	0.7302 ± 0.0035	0.7238 ± 0.0124	0.7127 ± 0.0063

C.2 MEAN-ONLY ABLATION

Mean-only uses the 49-dimensional descriptor $[w, d(w), \widehat{\mu}(w)]$ and removes the 14 second raw moments from the 63-dimensional DATS-MIX representation. Observations, source cache, TabPFN settings, and candidate bank are shared. It selects `U_SEL_0279861`; DATS-MIX selects `U_SEL_0443203`. Table 14 reports their direct comparison.

Table 14: Second-moment ablation: DATS-MIX minus Mean-only. Native contrasts are differences of J_7; WQL contrasts are raw-scale differences.

Metric	Mean difference	p	Lower
Native J_7	-0.011076	0.277	2/3
WQL	-0.003037	0.220	2/3

Against the same size-proportional reference, the mixtures selected by DATS-MIX and Mean-only reduce mean native loss by 10.85% and 9.85%, respectively. Their direct comparison gives 1.10% lower mean native loss for DATS-MIX ($p = 0.277$) and WQL lower by 0.003037 ($p = 0.220$), with both comparisons favoring DATS-MIX in two of three seeds. These measurements compare fixed selected mixtures rather than independent repetitions of the complete selection procedure.

C.3 DOMAIN-LEVEL CONTRASTS

Table 15 complements the main domain summary with all seven domain contrasts and the overall J_7 across the paired training runs. Every model is trained on the same 28-source pool. The domain and source probabilities change jointly under each selected mixture, so their association with loss does not isolate an individual-source intervention.

Table 15: Domain and aggregate contrasts, DATS-MIX minus Ratio+domain at 20k updates. Domain-wise tests are unadjusted secondary summaries. All four comparator-specific domain contrasts are supplied in the numerical data.

Metric	Mean ΔJ	p	95% interval
Energy	-0.139928	0.060	$[-0.294499, +0.014643]$
Environment	-0.035309	0.087	$[-0.083165, +0.012547]$
Health	-0.021135	0.222	$[-0.073056, +0.030785]$
IoT	-0.052753	0.021	$[-0.086198, -0.019307]$
Nature	-0.070335	0.029	$[-0.123091, -0.017579]$
Transportation	$+0.004965$	0.518	$[-0.022477, +0.032408]$
Web	-0.086369	0.016	$[-0.133422, -0.039316]$
J_7	-0.057266	0.028	$[-0.099605, -0.014928]$

C.4 DATASET-EQUAL PAIRED CONTRASTS

Table 16 gives the detailed contrasts for the evaluation-weight sensitivity in Section 5.5. The supplementary material includes source-level paired differences for all 28 datasets.

Table 16: Detailed paired contrasts under dataset-equal aggregation, with no mixture reselection. The aggregation weights differ from the main domain-equal objective.

Contrast	Mean Δ	p	95% interval
DATS-MIX − Ratio+domain	-0.057237	0.020	$[-0.092844, -0.021630]$
DATS-MIX − Mean-only	-0.007645	0.296	$[-0.031123, +0.015832]$
DATS-MIX − RegMix (d_7)	-0.044755	< 0.001	$[-0.050140, -0.039371]$
DATS-MIX − Size-proportional	-0.129322	< 0.001	$[-0.140784, -0.117859]$

C.5 WQL ENDPOINT CONTRASTS

For Section 5.2, Table 17 retains the raw seed-wise difference $\delta_s = \mathrm{WQL}_{A,s} - \mathrm{WQL}_{B,s}$ as the main WQL contrast. We additionally report a log-ratio analysis with $\ell_s = \log(\mathrm{WQL}_{A,s}/\mathrm{WQL}_{B,s})$ and relative change $100[\exp(\bar{\ell}) - 1]$. Table 18 reports that companion analysis for all comparisons, rather than mixing its p-values with raw-difference tests. Both use three paired seeds, two degrees of freedom, and unadjusted two-sided t tests. For DATS-MIX versus RegMix, the raw difference is -0.011808 ($p_{\mathrm{raw}} = 0.068$), and the geometric-mean relative change is -4.72% ($p_{\log} = 0.059$). Table 19 gives the seed-level WQL values. These are secondary endpoint analyses, not additional model runs.

Table 17: WQL contrasts on raw paired seed differences. This is the scale used for tests in the main text. Two-sided paired-t tests use two degrees of freedom; p-values are unadjusted.

Contrast	ΔWQL	p_{raw}	95% interval	Lower
DATS-MIX − Ratio+domain	-0.015803	0.122	$[-0.042009, +0.010404]$	3/3
DATS-MIX − Mean-only	-0.003037	0.220	$[-0.010454, +0.004380]$	2/3
DATS-MIX − RegMix (d_7)	-0.011808	0.068	$[-0.025766, +0.002149]$	3/3
DATS-MIX − Size-proportional	-0.012248	0.040	$[-0.023135, -0.001361]$	3/3
Mean-only − Ratio+domain	-0.012766	0.244	$[-0.046389, +0.020858]$	3/3
Mean-only − Size-proportional	-0.009212	0.161	$[-0.027419, +0.008996]$	3/3
Ratio+domain − Size-proportional	$+0.003554$	0.434	$[-0.012212, +0.019320]$	1/3
RegMix (d_7) − Size-proportional	-0.000440	0.944	$[-0.024430, +0.023550]$	2/3
RegMix (d_7) − Ratio+domain	-0.003994	0.693	$[-0.041627, +0.033639]$	2/3
Mean-only − RegMix (d_7)	-0.008772	0.062	$[-0.018618, +0.001075]$	3/3

Table 18: Companion WQL log-ratio analysis. For $\ell_s = \log(\mathrm{WQL}_{A,s}/\mathrm{WQL}_{B,s})$, relative change is $100[\exp(\bar{\ell}) - 1]$. This supplementary scale analysis does not replace the raw-difference tests.

Comparison	Relative WQL	$p_{\log}$	95% relative interval	Lower
DATS-MIX / Ratio+domain	-6.25%	0.117	$[-15.56, +4.09]\%$	3/3
DATS-MIX / Mean-only	-1.26%	0.220	$[-4.27, +1.84]\%$	2/3
DATS-MIX / RegMix (d_7)	-4.72%	0.059	$[-9.64, +0.46]\%$	3/3
DATS-MIX / Size-proportional	-4.94%	0.041	$[-9.15, -0.54]\%$	3/3
Mean-only / Ratio+domain	-5.05%	0.241	$[-17.08, +8.72]\%$	3/3
Mean-only / Size-proportional	-3.73%	0.163	$[-10.73, +3.82]\%$	3/3
Ratio+domain / Size-proportional	$+1.39\%$	0.436	$[-4.66, +7.83]\%$	1/3
RegMix (d_7) / Size-proportional	-0.23%	0.927	$[-9.33, +9.78]\%$	2/3
RegMix (d_7) / Ratio+domain	-1.60%	0.687	$[-15.22, +14.21]\%$	2/3
Mean-only / RegMix (d_7)	-3.51%	0.053	$[-7.00, +0.11]\%$	3/3

Table 19: Individual WQL values for the same 20k checkpoints and GIFT-Eval subset.

Seed	Size-prop.	Ratio+domain	RegMix	Mean-only	DATS-MIX
2609118111	0.250621	0.250957	0.260249	0.247133	0.242099
2609118112	0.242959	0.242421	0.241638	0.236285	0.231813
2609118113	0.250968	0.261833	0.241341	0.233495	0.233891

D RESPONSE-PREDICTION DIAGNOSTICS

For the fixed 48-mixture set $\mathcal{H}$, we compute

$$r_{\mathcal{H}} = y_{\arg\min_{j \in \mathcal{H}} \widehat{y}_j} - \min_{j \in \mathcal{H}} y_j. \tag{21}$$

Targets are two-seed mean proxy selection losses. The minimum is restricted to these 48 observed responses, not the full candidate bank. The response predictors conditioned on 128 observations share the same 48-mixture set and targets and use 10,000 paired-mixture resamples. Table 20 reports prediction-error differences; their unit is the absolute metric difference, not a percentage change.

Table 20: Paired-mixture bootstrap differences on the fixed 48-mixture set (10,000 resamples). These describe prediction-error uncertainty, not repeated observation collection or full-size training.

Comparison	Metric	Difference	95% interval
DATS-MIX − Ratio+domain	MAE	-0.001775	$[-0.006407, +0.002579]$
DATS-MIX − Ratio+domain	RMSE	-0.003661	$[-0.008673, +0.001394]$
DATS-MIX − Mean-only	MAE	$+0.000142$	$[-0.001181, +0.001397]$
DATS-MIX − Mean-only	RMSE	-0.000418	$[-0.001941, +0.000901]$
Mean-only − Ratio+domain	MAE	-0.001917	$[-0.006521, +0.002410]$
Mean-only − Ratio+domain	RMSE	-0.003243	$[-0.008149, +0.001700]$

DATS-MIX has MAE/RMSE 0.01767/0.02183, compared with 0.01753/0.02225 for Mean-only and 0.01945/0.02549 for Ratio+domain; all three have finite-set regret 0.02090. Prediction accuracy and selected-mixture forecasting utility remain separate evaluations. The targets were already observed during development, and predictor outputs were fixed before this retrospective evaluation. None of the 48 weight vectors exactly matches any of the 128 observations; the smallest cross-set total-variation distance is 0.28652. Across the 48 two-seed target means, sample variance is 0.000961 (SD 0.03100). The paired proxy seeds differ by 0.02203 in mean absolute value and 0.03293 in root-mean-square value. These observed differences describe seed variation, not a proven lower bound on achievable prediction error.