XMatch: Enhancing Covariate-Aware Time Series Forecasting through Tree-Structured Exogenous Matching

Anonymous authors
Paper under double-blind review

Abstract

Future exogenous variables provide valuable information for forecasting endogenous time series. Existing covariate-aware methods primarily learn the direct influence of exogenous variables on endogenous variables. However, these effects can be complex and change with the pattern of the exogenous variables, making them difficult to capture. Beyond this perspective, we observe that a given exogenous pattern often co-occurs with only a small set of endogenous response patterns. These associations motivate a strategy that matches future and historical exogenous patterns and uses the corresponding endogenous patterns to enhance forecasting. However, in real-world forecasting scenarios with multiple exogenous variables, each exogenous variable provides a distinct dimension for matching, creating a dilemma for this strategy between precise matching and sufficient historical support. To bridge this gap, we propose **XMatch** (**EX**ogenous **MATCH**ing), a covariate-aware forecasting model that realizes the aforementioned strategy through a tree-structured matching process that adaptively adjusts the number of exogenous variables used as matching conditions. Specifically, we first introduce the *ProtoTree Creator*, which organizes historical correspondences between exogenous and endogenous patterns into a *ProtoTree*, whose deeper levels incorporate additional exogenous variables for matching. For forecasting, we then design the *ProtoTree Matcher*, which uses future exogenous variables to query the *ProtoTree* and adaptively determines how many exogenous variables to use for matching based on exogenous pattern similarity and historical support. Finally, the matched endogenous patterns are used as explicit historical evidence to enhance forecasting. Extensive experiments on 12 real-world datasets demonstrate that XMatch outperforms state-of-the-art baselines.

1 Introduction

Time series forecasting plays a vital role in a wide range of real-world applications, including finance (Sezer et al., 2020; Huang et al., 2022), transportation (Wu et al., 2024; Shao et al., 2025), healthcare (Wu et al., 2025; Miao et al., 2021), energy (Wang et al., 2025), and AIOps (Lin et al., 2024; Pan et al., 2023). Beyond historical target observations, forecasting tasks often involve *covariates* that describe external factors related to the dynamics of the target series (Cheng et al., 2026; Wang et al., 2024). In many real-world scenarios, future covariates are also available or can be estimated in advance using dedicated forecasting systems. For example, weather forecasts provide future temperature and precipitation conditions that are informative for traffic prediction (Li et al., 2026). Since these covariates cover the same future horizon as the prediction target, they describe how external factors, such as temperature and precipitation, are expected to evolve during the prediction horizon and can substantially improve forecasting accuracy (Wang et al., 2024; Zhou et al., 2025). Effectively exploiting such future covariate information is therefore important for practical forecasting (Das et al., 2023; Yang et al., 2026).

Existing covariate-aware models primarily learn the direct influence of exogenous variables on endogenous variables (Das et al., 2023; Wang et al., 2024; Zhou et al., 2025). However, these effects can be complex and change with the pattern of the exogenous variables, making them difficult to capture with a unified interaction mechanism. Beyond this conventional perspective, we observe a

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.

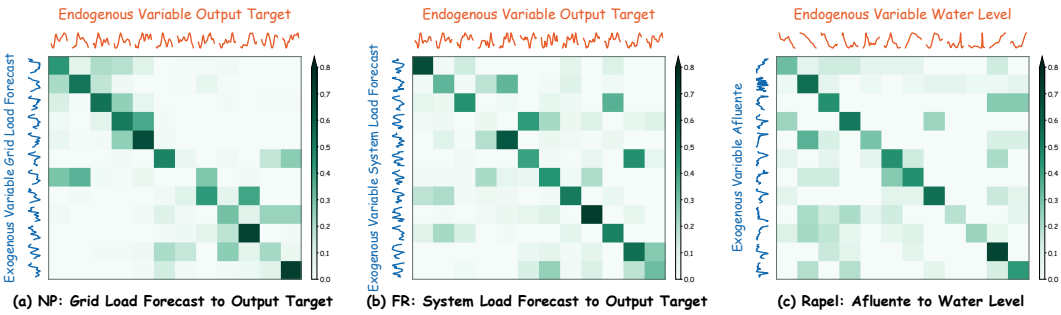


Figure 1: Illustration of the frequency of co-occurrences between exogenous and endogenous patterns in real-world datasets. Rows denote exogenous patterns and columns denote endogenous response patterns, while darker cells indicate more co-occurrences. The darker cells in each row are often concentrated in a few columns, showing that a given exogenous pattern tends to co-occur with only a small set of endogenous patterns. We provide further evidence in Appendix B.3.

recurring correspondence between exogenous and endogenous patterns in real-world time series. As illustrated in Figure 1, a given exogenous pattern often co-occurs with only a small set of endogenous response patterns. This suggests that exogenous patterns carry informative signals about which endogenous response patterns are more likely to occur. Such recurring associations can thus provide a source of information for forecasting through direct pattern matching, reducing reliance on a unified interaction mechanism to capture complex exogenous effects. Therefore, in addition to learning exogenous-to-endogenous influence, forecasting models can further leverage historical endogenous responses associated with similar exogenous patterns as explicit evidence for prediction.

Motivated by this observation, we introduce a complementary forecasting perspective that **explicitly captures recurring endogenous patterns under similar exogenous patterns.** To enhance conventional covariate-aware forecasting, we identify recurring associations between exogenous patterns and their corresponding endogenous responses in historical data, and use these associations to augment prediction. We refer to this strategy as *Exo–Endo Association Augmentation*.

However, this strategy becomes challenging in real-world forecasting scenarios with multiple exogenous variables. **This strategy faces a fundamental dilemma between precise matching and sufficient historical support.** Each additional exogenous variable requires historical references to match the future exogenous patterns along another dimension. Matching on more variables yields more specific references, but fewer historical samples satisfy these matching requirements. As a result, the resulting endogenous-pattern evidence may be supported by too few samples to be reliable (Vandermeulen et al., 2024). In contrast, matching on fewer variables provides more historical references, but may overlook differences in the remaining variables that are relevant to endogenous prediction (Gao & Hastie, 2022; Yang & van Leeuwen, 2024). The key challenge is therefore to identify historical references that closely match the future exogenous patterns while retaining enough samples to support reliable prediction.

To bridge this gap, we propose **XMatch** (**EX**ogenous **MATCH**ing), a covariate-aware forecasting model that realizes the *Exo–Endo Association Augmentation* strategy through a tree-structured matching process that adaptively adjusts the number of exogenous variables used as matching conditions to address the aforementioned dilemma. Specifically, we first introduce the *ProtoTree Creator* to organize historical Exo–Endo associations into a tree-structured memory, where each deeper level adds one exogenous variable and thus represents a more specific condition with fewer supporting samples. For forecasting, we design the *ProtoTree Matcher*, which uses future exogenous variables to query the *ProtoTree* and adaptively determines the matching depth, assigning more weight to deeper matches only when they are sufficiently similar and well supported; otherwise, the weight remains at shallower, better-supported nodes. Finally, the matched endogenous patterns are aggregated and fused with the forecasting backbone's representations to produce the forecast.

Our contributions are summarized as follows:

- We reveal recurring correspondences between exogenous and endogenous patterns in covariate-aware forecasting and introduce a new perspective that uses endogenous patterns from historical references with similar exogenous patterns as explicit forecasting evidence.

- We propose XMatch, which adaptively adjusts the number of exogenous variables used in tree-structured matching to balance precise matching and sufficient historical support for forecasting.
- We conduct extensive experiments on 12 real-world datasets with exogenous variables. The results show that XMatch outperforms state-of-the-art baselines.

2 Related Work

2.1 Time Series Forecasting with Exogenous Variables

Time series forecasting with exogenous variables has long been studied in classical statistics, where ARIMAX (Williams, 2001) and SARIMAX (Vagropoulos et al., 2016) incorporate external factors through additional regression terms. Recent deep forecasting models provide more flexible mechanisms for covariate integration. NBEATSx (Olivares et al., 2023) introduces dedicated branches for exogenous inputs, while TiDE (Das et al., 2023) concatenates static and future covariates with endogenous features. CrossLinear (Zhou et al., 2025) captures cross-variable dependencies through cross-correlation embeddings and convolutions; TFT (Lim et al., 2021), TimeXer (Wang et al., 2024), and ExoTST (Tayal et al., 2024) employ variable selection or attention mechanisms to integrate exogenous information. More recent approaches further model richer dependencies: DAG (Qiu et al., 2026) characterizes temporal and channel correlations, GCGNet (Li et al., 2026) aligns graph-structured correlations, and KITE (Cheng et al., 2026) introduces knowledge-guided covariate conditioning for probabilistic forecasting. Existing methods primarily model exogenous effects through feature-level interactions, leaving cross-variable pattern correspondences implicit. In contrast, XMatch explicitly retrieves endogenous patterns from historical references with similar exogenous patterns, complementing conventional covariate integration with historical evidence.

2.2 Retrieval-Augmented Time Series Modeling

Beyond encoding temporal dynamics solely in model parameters, a growing line of work introduces memory mechanisms to store and reuse features extracted from historical time series. Existing studies in this area mainly follow three directions. **1) Instance-level retrieval.** RATD (Liu et al., 2024) retrieves relevant sequences to guide diffusion-based forecasting, while RAFT (Han et al., 2025) augments predictors with the future continuations of similar historical inputs. TS-RAG (Ning et al., 2025) extends this idea to zero-shot forecasting with time series foundation models, and PFRP (Du et al., 2026) combines predictions retrieved from a global memory bank with those of a local forecaster. **2) Prototype-based memory.** PUAD (Li et al., 2023) and H-PAD (Shen, 2025) compress recurring behavior into prototypical normal patterns for anomaly detection. **3) Parametric temporal memory.** Other approaches encode historical regularities through learnable memory items (Song et al., 2023), multi-resolution attractor memory (Hu et al., 2024), or compact temporal bases (Huang et al., 2025a). Despite demonstrating the value of explicit pattern reuse, these methods retrieve or summarize patterns from historical observations without explicitly organizing the correspondences between combinations of exogenous patterns and endogenous patterns. In contrast, XMatch organizes these correspondences in a prototype tree and adaptively adjusts the number of exogenous variables used for matching, balancing precise matching with sufficient historical support.

3 Methodology

In the context of covariate-aware time series forecasting, the historical endogenous time series $X^{\text{endo}} \in \mathbb{R}^{N\times T}$ is accompanied by historical exogenous variables $X^{\text{exo}} \in \mathbb{R}^{D\times T}$ and future exogenous variables $Y^{\text{exo}} \in \mathbb{R}^{D\times F}$, where N is the number of endogenous variables, D is the number of exogenous variables, T is the number of historical time steps, and F is the forecasting horizon. XMatch further constructs a dataset-level *ProtoTree* $\mathcal{T}_\phi$ from the training set to encode historical associations between exogenous and endogenous patterns. The objective is to forecast the future endogenous series $\hat{Y}^{\text{endo}} \in \mathbb{R}^{N\times F}$ from the time-series inputs and $\mathcal{T}_\phi$:

$$\hat{Y}^{\text{endo}} = \mathcal{F}_\theta\big(X^{\text{endo}}, X^{\text{exo}}, Y^{\text{exo}}; \mathcal{T}_\phi\big)\,. \tag{1}$$

Here, $\mathcal{F}_\theta$ is the forecaster parameterized by θ. Each endogenous variable has its own *ProtoTree*; we use one variable as an example below.

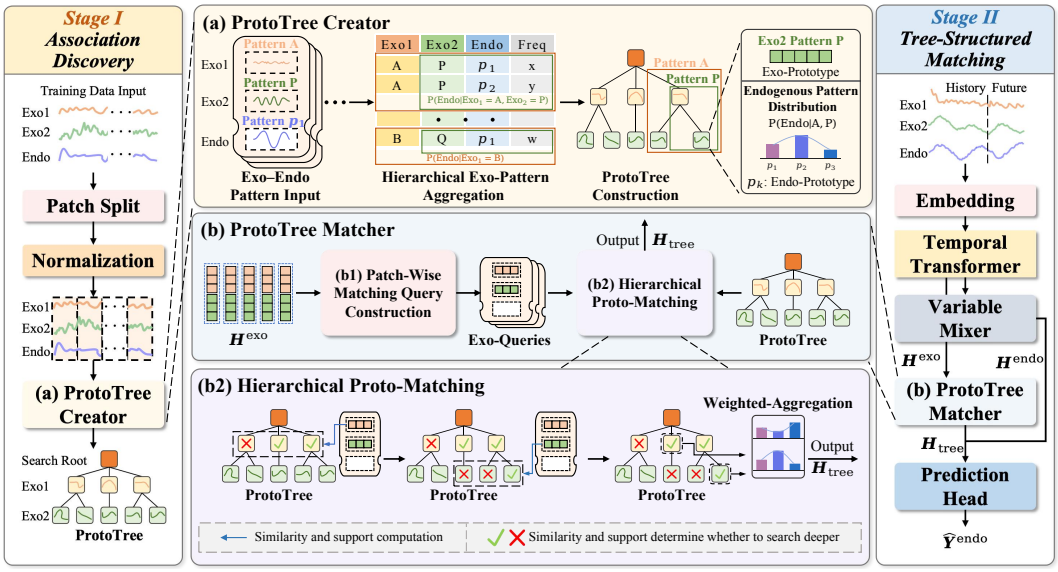


Figure 2: The architecture of XMatch. (a) The *ProtoTree Creator* organizes correspondences between exogenous and endogenous patterns according to the descending discriminative power of exogenous variables. (b) The *ProtoTree Matcher* constructs matching queries from known future exogenous variables and uses *Hierarchical Proto-Matching* to match and aggregate endogenous response prototypes, thereby improving forecasting accuracy.

3.1 Structure Overview

Figure 2 illustrates XMatch, which leverages recurring correspondences between exogenous and endogenous patterns as explicit evidence to enhance forecasting under covariate scenarios, and organizes these associations in a hierarchical *ProtoTree* to address the dilemma between precise matching and sufficient historical support in multi-exogenous matching. Our training procedure consists of two stages. In ***Stage I:*** *Association Discovery*, we mine recurring correspondences between exogenous and endogenous patterns from the training set and organize them into a *ProtoTree* whose deeper nodes match patterns across more exogenous variables, while shallower nodes retain more historical samples to support matching when deeper nodes have too few samples. In ***Stage II:*** *Tree-Structured Matching*, we encode future exogenous variables as matching queries and perform adaptive matching over the *ProtoTree*. The matched response prototypes are then fused with the temporal representation to produce $\hat{Y}^{\text{endo}}$. We detail the two stages below.

3.2 Stage I: Association Discovery

In this stage, we introduce the *ProtoTree Creator* to extract aligned local patterns from the training set, assign discrete pattern labels, and hierarchically aggregate the observed correspondences between exogenous and endogenous patterns into a *ProtoTree*. The resulting tree summarizes reusable association memory for the subsequent matching stage.

3.2.1 ProtoTree Creator

As shown in Figure 2a, the *ProtoTree Creator* extracts aligned Exo–Endo patterns, organizes their associations by matching on increasing numbers of exogenous variables, and encodes them as node prototypes for subsequent matching.

Aligned Exo–Endo Pattern Input. To assign discrete pattern labels to the correspondences between exogenous and endogenous patterns observed in the training set, we first extract M aligned patches of length P, which may overlap, from each endogenous and exogenous variable in the training set and normalize every patch independently Nie et al. (2023):

$$\widetilde{E}_j = \text{Norm}(\text{Patchify}(X_j^{\text{exo}})), \quad \widetilde{S}_i = \text{Norm}(\text{Patchify}(X_i^{\text{endo}})), \tag{2}$$

where $\widetilde{E}_j \in \mathbb{R}^{M\times P}$ and $\widetilde{S}_i \in \mathbb{R}^{M\times P}$ denote the normalized patch sequences of the j-th exogenous variable and the i-th endogenous variable, respectively, with $j \in \{1,\dots,D\}$ and $i \in \{1,\dots,N\}$.

To group patterns with similar shapes despite temporal shifts, we cluster the patterns of each variable independently using *Soft-DTW* (Cuturi & Blondel, 2017), which replaces the hard minimum over temporal alignments with a differentiable soft minimum:

$$\mathrm{sDTW}_\gamma(\boldsymbol{x},\boldsymbol{c}) = -\gamma \log \sum_{\boldsymbol{A}\in\mathcal{A}} \exp\left(-\frac{\langle \boldsymbol{A}, \boldsymbol{\Delta}(\boldsymbol{x},\boldsymbol{c})\rangle}{\gamma}\right), \tag{3}$$

where a patch is $\boldsymbol{x} \in \mathbb{R}^P$, a cluster center is $\boldsymbol{c} \in \mathbb{R}^P$, $\mathcal{A}$ is the set of valid alignments and $\boldsymbol{\Delta}$ is the pairwise cost matrix. To accommodate variable-specific pattern diversity, we adopt a *DP-means-style strategy* (Kulis & Jordan, 2012) that adaptively determines the number of cluster centers for each variable. After convergence, for each variable u, each patch m is assigned a variable-specific pattern label $z_{u,m} \in \{1,\dots,K_u\}$.

Hierarchical Exogenous Pattern Aggregation. To balance precise matching with sufficient historical support, we construct a hierarchical matching tree, the *ProtoTree*, by first ordering exogenous variables according to their informativeness about endogenous patterns and then progressively combining their pattern labels across increasing numbers of variables.

First, we rank exogenous variables by information gain about endogenous patterns, measured as the reduction in Shannon entropy H Shannon (1948). Let Z_i^{endo} and Z_j^{exo} denote the discrete pattern-label variables obtained by clustering endogenous variable i and exogenous variable j, respectively. We compute the information gain as follows:

$$I_{i,j} = H(Z_i^{\mathrm{endo}}) - H(Z_i^{\mathrm{endo}} \mid Z_j^{\mathrm{exo}}). \tag{4}$$

For each endogenous variable i, we order exogenous variables by descending $I_{i,j}$ from shallow to deep levels. Let l denote tree depth and π_l the exogenous variable index at that level.

Second, we combine the ordered labels into a tuple of exogenous pattern labels $\boldsymbol{\eta}_{l,m} = (z^{\mathrm{exo}}_{\pi_1,m},\dots,z^{\mathrm{exo}}_{\pi_l,m})$ at tree level l for patch m. Each distinct label tuple defines a node v that stores all endogenous patterns from matching training patches. Normalizing their pattern-label counts yields $p_v(k)$, the empirical probability of endogenous pattern k for this label tuple:

$$p_v(k) = \frac{1}{|\mathcal{M}_v|} \sum_{m\in\mathcal{M}_v} \mathbb{I}[z^{\mathrm{endo}}_{i,m} = k], \quad k \in \{1,\dots,K_i^{\mathrm{endo}}\}, \tag{5}$$

where $\mathcal{M}_v$ contains the training patches whose exogenous pattern labels match the tuple represented by node v, and $z^{\mathrm{endo}}_{i,m}$ is the pattern label of endogenous variable i at patch m. This hierarchy preserves one-to-many Exo–Endo associations, retaining more historical samples at shallow levels and matching on more exogenous variables at deeper levels.

***ProtoTree* Construction.** Each node v at level l represents a tuple of exogenous pattern labels $\boldsymbol{\eta}_{l,m}$ encoded by the tree path leading to it. To support prototype matching, we use an MLP to project the exogenous pattern center $\boldsymbol{c}^{\mathrm{exo}}_{\pi_l,z_v} \in \mathbb{R}^P$ associated with its current-level label $z_v = z^{\mathrm{exo}}_{\pi_l,m}$ into a learnable prototype key $\boldsymbol{k}_v \in \mathbb{R}^d$. We use another MLP to map the endogenous pattern centers $\boldsymbol{c}^{\mathrm{endo}}_{i,k} \in \mathbb{R}^P$, weighted by the node's empirical distribution $p_v(k)$, into a learnable response prototype $\boldsymbol{r}_v \in \mathbb{R}^d$ associated with the matched node:

$$\boldsymbol{k}_v = \mathrm{MLP}_K\left(\boldsymbol{c}^{\mathrm{exo}}_{\pi_l,z_v}\right), \qquad \boldsymbol{r}_v = \mathrm{MLP}_V\left(\sum_k p_v(k)\boldsymbol{c}^{\mathrm{endo}}_{i,k}\right). \tag{6}$$

This construction grounds prototype initialization in observed patterns and their conditional frequencies, providing an empirical basis for subsequent learning. Finally, we add a virtual *Search Root* and connect it to all first-level nodes, providing a unified entry for hierarchical matching.

3.3 Stage II: Tree-Structured Matching

In this stage, we introduce the *Transformer Backbone* and the *ProtoTree Matcher* to match and aggregate endogenous-response evidence from the *ProtoTree*. The *Transformer Backbone* encodes

the inputs into patch-wise representations, using future exogenous latents as matching queries and future endogenous latents as the base forecast. The *ProtoTree Matcher* then performs *Hierarchical Proto-Matching* to adaptively select and aggregate response prototypes, which are fused with the base representation to produce the final forecast.

3.3.1 TRANSFORMER BACKBONE

As shown in Figure 2, the Transformer backbone encodes historical observations and known future exogenous variables using the patch-wise representation introduced above Cirstea et al. (2022). We append learnable mask tokens M_i^{endo} to the historical endogenous variables and concatenate the historical exogenous variables with their known future values:

$$S_i^{\text{endo}} = \text{PatchEmbed}([X_i^{\text{endo}}; M_i^{\text{endo}}]), \qquad S_j^{\text{exo}} = \text{PatchEmbed}([X_j^{\text{exo}}; Y_j^{\text{exo}}]). \tag{7}$$

Let L denote the total number of patches. This produces $S^{\text{endo}} \in \mathbb{R}^{N\times L\times d}$ and $S^{\text{exo}} \in \mathbb{R}^{D\times L\times d}$. When future exogenous variables are unavailable, they are replaced with learnable mask tokens.

We use $\text{Transformer}_{\text{temp}}$ with causal masking to model dependencies among patches, followed by MLP_{var} with a residual connection to exchange information across variables:

$$H_{\text{time}} = \text{Transformer}_{\text{temp}}(S; M_{\text{causal}}) \in \mathbb{R}^{(N+D)\times L\times d}, \tag{8}$$

$$H_{\text{mix}} = H_{\text{time}} + \text{MLP}_{\text{var}}(H_{\text{time}}) \in \mathbb{R}^{(N+D)\times L\times d}, \tag{9}$$

where $S = [S^{\text{endo}}; S^{\text{exo}}]$ and M_{causal} blocks attention to subsequent patches. The temporal blocks capture per-variable temporal dependencies, while MLP_{var} operates along the variable dimension to exchange information across endogenous and exogenous variables.

We use the future exogenous latents from the temporal Transformer output H_{time}, before variable mixing, as matching queries H^{query} for the *ProtoTree*.

3.3.2 *ProtoTree Matcher*

As shown in Figure 2b, the *ProtoTree Matcher* constructs a matching query for each exogenous variable at each future patch, searches the *ProtoTree* level by level, and aggregates the matched response prototypes as forecasting evidence.

Patch-Wise Matching Query Construction. As shown in Figure 2b1, we construct queries by selecting the future-patch representations of the exogenous variables from the temporal Transformer output H_{time}, before the variable mixer. Let $L_f = \lceil F/P \rceil$ denote the number of future patches. With endogenous variables preceding exogenous variables in the concatenated input, we extract the query for exogenous variable j at future patch s as follows:

$$H^{\text{query}}_{j,s,:} = H_{\text{time}}[N + j, L - L_f + s, :] \in \mathbb{R}^d. \tag{10}$$

This yields $H^{\text{query}} \in \mathbb{R}^{D\times L_f\times d}$. For each future patch, we select these variable-specific queries according to the exogenous variable order $\pi_1, \ldots, \pi_D$ established during tree construction, using the query for variable π_l to match the exogenous prototype keys at the corresponding tree level l.

Hierarchical Proto-Matching. As shown in Figure 2b2, we use the variable-specific future-patch queries constructed by the preceding module to search the *ProtoTree* level by level, with $\boldsymbol{q}_s^{(l)} = H^{\text{query}}_{\pi_l,s,:}$ serving as the query at level l. For each valid exogenous prototype j at this level, we measure its cosine similarity $a_{s,l,j} = \cos(\boldsymbol{q}_s^{(l)}, \boldsymbol{k}_{l,j})$ and convert the similarity scores into relative matching weights using softmax:

$$q_{s,l,j} = \text{softmax}_{j\in\mathcal{J}_l}\left(a_{s,l,j}/\tau_{\text{route}}\right), \tag{11}$$

where τ_{route} is the routing temperature and $\mathcal{J}_l$ denotes the set of indices of all valid exogenous prototypes at level l, not only those of the children observed under the current node v. Let $c(v, j)$ denote the corresponding child, if it exists, and let $n_{c(v,j)}$ be the number of training patches matching the full sequence of exogenous pattern labels along its tree path. During inference, search continues along this branch to the next tree level only when all three conditions below are satisfied:

$$b_{s,v,j} = \mathbb{I}\big[c(v, j) \text{ exists } \wedge\ n_{c(v,j)} \geq n_{\min} \ \wedge\ a_{s,l,j} \geq \tau_{\text{sim}}\big] = 1. \tag{12}$$

Let $\mu_{s,v}$ denote the weight allocated to node v for patch s, with the root weight initialized to one for each future patch. The weight allocated to child $c(v, j)$ is then computed as the parent weight multiplied by its matching weight and branch indicator:

$$\mu_{s,c(v,j)} = \mu_{s,v} \times q_{s,l,j} \times b_{s,v,j}. \tag{13}$$

Node v retains the weight not passed to its children, yielding the following aggregation weight:

$$w_{s,v} = \mu_{s,v} \left(1 - \sum_{j \in \mathcal{J}_l} q_{s,l,j} b_{s,v,j} \right). \tag{14}$$

A leaf retains its full allocated weight, and the aggregation weights sum to one. Weights for unsupported, dissimilar, or unobserved branches remain at the parent. This allocation combines evidence from multiple matching branches while favoring matches on more exogenous variables only when they have sufficient historical support and query similarity.

For each future patch s, we use the aggregation weights $w_{s,v}$ computed above to combine the node response prototypes $\boldsymbol{r}_v$:

$$H_{\text{tree}}[:, s, :] = \text{Reshape}_{N \times d} \left(\boldsymbol{W}_O \sum_v w_{s,v} \boldsymbol{r}_v \right), \tag{15}$$

where the sum runs over all tree nodes except the root, and $\boldsymbol{W}_O \in \mathbb{R}^{Nd \times d}$ projects the aggregated prototype to the endogenous representation space. We stack the outputs across future patches to obtain $H_{\text{tree}} \in \mathbb{R}^{N \times L_f \times d}$, the endogenous-response representation corresponding to the queried future exogenous patterns.

Finally, we add this output to the future endogenous representation H^{endo} from the *Transformer Backbone* and pass the fused representation through a *Prediction Head* to obtain the final endogenous forecast over the prediction horizon:

$$\hat{Y}^{\text{endo}} = \text{Head}\big(H^{\text{endo}} + H_{\text{tree}}\big) \in \mathbb{R}^{N \times F}. \tag{16}$$

4 Experiments

Experimental Setup

Datasets. We evaluate XMatch on 12 real-world datasets under the setting where future exogenous variables are provided as inputs. We conduct both short-term and long-term forecasting experiments on all datasets. For Colbun and Rapel, the short-term setting uses a lookback window of 60 with a prediction horizon of 10, while the long-term setting uses a lookback window of 180 with a prediction horizon of 30. For the other datasets, the corresponding lookback/prediction lengths are $168/24$ and $720/360$. Dataset statistics and variable descriptions are provided in Appendix A.1.

Baselines. We compare XMatch with 10 competitive time series baselines. The first group consists of methods that natively model future covariates: DAG (Qiu et al., 2026), KITE (Cheng et al., 2026), GCGNet (Li et al., 2026), TimeXer (Wang et al., 2024), TFT (Lim et al., 2021), and TiDE (Das et al., 2023). The second group contains strong general-purpose time series forecasters: DUET (Qiu et al., 2025), CrossLinear (Zhou et al., 2025), Amplifier (Fei et al., 2025), and TimeKAN (Huang et al., 2025b). For a fair comparison, models in the second group are equipped with the same MLP-based future-covariate fusion strategy used in prior covariate forecasting studies (Qiu et al., 2026) as described in Appendix A.6.

Implementation Details. To ensure consistency with previous studies, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics, where lower values indicate better performance. We use the TFB benchmark (Qiu et al., 2024) for unified evaluation, and all baseline results are obtained within this framework. Evaluation retains the final incomplete batch to ensure that all test windows are included. Further implementation details are provided in Appendix A.4.

4.1 Main Results

Table 1 presents the average forecasting results on 12 real-world datasets when future exogenous variables are available. Full results for both forecasting horizons are reported in Appendix C.1. We

Table 1: Average results on 12 real-world datasets, where the inputs are X^{endo}, X^{exo}, and Y^{exo}. **Red**: the best, Blue: the 2nd best. Full results with Y^{exo} are reported in Table 5.

Models	XMatch		DAG		KITE		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	**0.282**	**0.299**	0.362	0.344	0.325	0.323	0.370	0.348	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.420	0.418	0.405	0.419
PJM	**0.085**	**0.178**	0.093	0.180	0.096	0.179	0.095	0.187	0.108	0.198	0.114	0.207	0.142	0.246	0.102	0.197	0.112	0.223	0.137	0.246	0.139	0.262
BE	**0.420**	**0.273**	0.423	0.280	0.428	0.286	0.431	0.294	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.559	0.413	0.548	0.407
FR	0.410	**0.216**	0.414	0.219	**0.387**	0.225	0.415	0.234	0.427	0.241	0.504	0.257	0.484	0.281	0.496	0.327	0.483	0.298	0.554	0.408	0.547	0.374
DE	**0.349**	**0.369**	0.370	0.370	0.350	0.370	0.401	0.389	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.473	0.441	0.473	0.445
Energy	**0.089**	**0.231**	0.124	0.267	0.111	0.257	0.131	0.277	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.233	0.389	0.218	0.381
Sdwpfm1	**0.402**	**0.436**	0.423	0.461	0.417	0.451	0.424	0.457	0.701	0.609	0.482	0.474	0.483	0.507	0.599	0.570	0.426	0.502	0.437	0.490	0.447	0.534
Sdwpfm2	**0.469**	**0.479**	0.477	0.485	0.475	0.483	0.475	0.486	0.803	0.653	0.476	0.488	0.486	0.516	0.514	0.490	0.533	0.573	0.491	0.512	0.497	0.564
Sdwpfh1	**0.406**	**0.442**	0.448	0.486	0.441	0.491	0.450	0.500	0.746	0.643	0.479	0.491	0.453	0.508	0.539	0.516	0.557	0.593	0.537	0.598	0.577	0.638
Sdwpfh2	**0.413**	**0.451**	0.523	0.530	0.501	0.527	0.520	0.536	0.891	0.719	0.566	0.521	0.599	0.583	0.647	0.566	0.538	0.574	0.521	0.581	0.647	0.672
Colbun	0.098	**0.149**	0.098	0.154	**0.088**	0.172	0.107	0.175	0.145	0.235	0.238	0.297	0.164	0.227	0.198	0.266	0.126	0.195	0.173	0.246	0.128	0.175
Rapel	0.238	**0.284**	**0.230**	0.305	0.244	0.295	0.306	0.307	0.344	0.362	0.305	0.333	0.320	0.351	0.269	0.326	0.252	0.313	0.257	0.321	0.249	0.311
1st Count	**9**	**12**	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

have the following observations: 1) XMatch achieves the highest number of first-place rankings, with 9 in MSE and 12 in MAE. It obtains the lowest values for both metrics on nine datasets, outperforming both forecasters designed for future covariates and general-purpose forecasters augmented with future-covariate fusion. 2) The improvements are particularly pronounced on Energy and Sdwpfh2. Compared with the strongest baseline for each metric, XMatch reduces MSE and MAE by 20% and 10% on Energy, and by 18% and 13% on Sdwpfh2, respectively. These results support the effectiveness of retrieving endogenous patterns from historical references with similar exogenous patterns, as further illustrated in Appendix C.1.

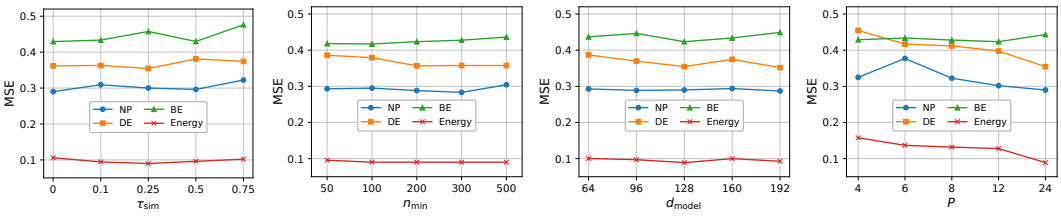
4.2 Model Analysis

Ablation Studies. To assess the contribution of the components of XMatch, we compare the full model with five variants in Table 2: (a) *w/o tree retrieval* disables the retrieval branch and sets its residual contribution to zero. Its degraded performance indicates that retrieved endogenous response patterns provide useful evidence. (b) *w/o variable ordering* retains the full model's clusters and exogenous variable set but rebuilds the prefix tree using a fixed random permutation instead of information-gain ordering. The performance drop supports prioritizing informative exogenous variables to guide hierarchical matching. (c) *w/o search stopping* disables similarity-based and reliability-based early stopping during training and inference, continuing along existing branches until a leaf or a missing branch is reached. The resulting degradation highlights the role of adaptive stopping in avoiding dissimilar or insufficiently supported matches. (d) *endo history retrieval* replaces tree retrieval with the average subsequent endogenous sequence of the five nearest training records, queried using the historical endogenous context. Its inferior results support using future exogenous patterns to retrieve relevant endogenous responses. (e) *mix exo retrieval* queries the training library using concatenated, channel-wise normalized future exogenous patches and averages the endogenous patches of the five nearest records, without clustering or a tree. Its lower accuracy supports organizing exogenous–endogenous associations into *ProtoTree* for effective matching. Ultimately, the full XMatch model, with variable ordering and adaptive matching, balances matching precision and historical support and achieves strong results.

Table 2: Average results of ablation studies for XMatch.

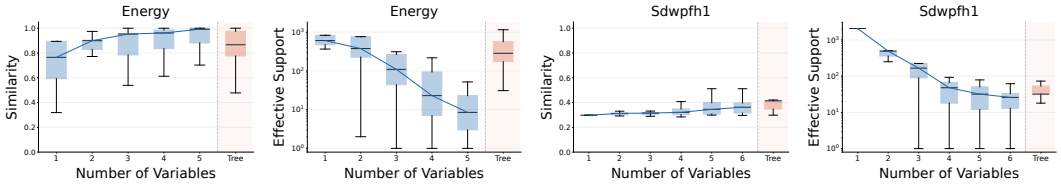
Dataset	BE		DE		Sdwpfh1	
Metrics	mse	mae	mse	mae	mse	mae
(a) w/o tree retrieval	0.440	0.298	0.371	0.373	0.442	0.477
(b) w/o variable ordering	0.438	0.289	0.410	0.396	0.457	0.481
(c) w/o search stopping	0.430	0.283	0.380	0.375	0.419	0.454
(d) endo history retrieval	0.433	0.299	0.356	0.373	0.478	0.491
(e) mix exo retrieval	0.425	0.285	0.363	0.377	0.470	0.487
XMatch (Full)	**0.420**	**0.273**	**0.349**	**0.369**	**0.406**	**0.442**

Parameter Sensitivity. We also conduct parameter sensitivity studies of XMatch, with the following observations: 1) Figures 3a and 3b illustrate the performance variations under different settings



(a) Similarity threshold (b) Minimum support (c) Model dimension (d) Patch length

Figure 3: Hyperparameter sensitivity of XMatch on NP, DE, BE, and Energy.



(a) Energy: Similarity (b) Energy: Support (c) Sdwpfh1: Similarity (d) Sdwpfh1: Support

Figure 4: Retrieval similarity and effective historical support on Energy and Sdwpfh1. Support is shown on a logarithmic scale.

of the similarity threshold τ_{sim} and minimum support count $n_{\min}$. A stricter similarity threshold does not necessarily improve forecasting accuracy. An excessively large $n_{\min}$ may reduce the number of retrieved matches, whereas an excessively small value may admit matches with insufficient historical support, reducing the reliability of the retrieved evidence. 2) Figure 3c shows that XMatch is relatively insensitive to the model dimension d, while smaller dimensions help reduce computational cost. We recommend selecting d within $[96, 160]$. 3) Figure 3d explores the effect of patch length on forecasting performance. The optimal patch length varies across datasets due to differences in temporal dependencies; however, a patch size in $[12, 24]$ tends to yield favorable results. Very small patch lengths may increase computational complexity, while excessively large patch lengths may weaken the model's ability to capture local dependencies. Additional sensitivity analyses are provided in Appendix B.1.

Matching Precision and Historical Support Analysis. We compare *ProtoTree* with direct retrieval using increasing numbers of exogenous variables on Energy and Sdwpfh1. We evaluate the similarity of retrieved endogenous patterns to the actual future endogenous patterns using weighted DTW-based similarity, together with effective historical support. These metrics jointly assess match quality and evidence sufficiency. Details are provided in Appendix A.5.

As shown in Figure 4, using more exogenous variables in direct retrieval improves similarity but reduces historical support. *ProtoTree* achieves high similarity while retaining greater support than direct retrieval using all exogenous variables, supporting its ability to balance matching precision and historical support. This balance helps avoid relying on too few historical examples when matching highly specific exogenous patterns.

5 Conclusion

In this work, we propose XMatch, a covariate-aware forecasting framework that exploits recurring correspondences between exogenous and endogenous patterns as explicit forecasting evidence. The *ProtoTree Creator* organizes historical correspondences between exogenous and endogenous patterns into a hierarchical prototype tree. The *ProtoTree Matcher* then matches future exogenous representations against exogenous prototype keys and aggregates the corresponding endogenous response prototypes, balancing precise matching with sufficient historical support. Experiments on 12 real-world datasets demonstrate the strong overall performance of XMatch against state-of-the-art forecasting baselines. The code and datasets are available at `https://anonymous.4open.science/r/XMatch-71DE/`.

AI USE STATEMENT

We used ChatGPT to assist with manuscript polishing and improve the clarity of the writing. We did not use generative AI tools to formulate the research methodology, generate experimental data or results, or draw scientific conclusions. All methodological design and experimental analyses were conducted by the authors. The authors reviewed all AI-assisted text and code for accuracy and consistency with the methodology and experiments. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on methodological research for covariate-aware time series forecasting through exogenous-conditioned tree-structured retrieval. Our experiments use existing benchmark datasets concerning electricity prices, power generation, and reservoir water levels, and do not involve human subjects or the collection of personal data. We do not release any new dataset containing sensitive or personally identifiable information. XMatch is intended for general-purpose forecasting research, and we are not aware of any direct ethical risks arising from the methodology itself.

REPRODUCIBILITY STATEMENT

The XMatch architecture is described in Section 3. The experimental settings, forecasting horizons, evaluation metrics, and baseline evaluation framework are provided in Section 4. Dataset sources, variable descriptions, and chronological data splits are detailed in Appendix A.1.

REFERENCES

Hanyin Cheng, Jingrong Zhou, Yang Shu, and Chenjuan Guo. Kite: Knowledge-guided probabilistic modeling for time series forecasting with exogenous variables. In *ICML*, 2026.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting. In *IJCAI*, pp. 1994–2001, 2022.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, volume 70, pp. 894–903, 2017.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Dazhao Du, Tao Han, and Song Guo. Predicting the future by retrieving the past. In *AAAI*, pp. 20896–20904, 2026.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *AAAI*, volume 39, pp. 11645–11653, 2025.

Zijun Gao and Trevor Hastie. Lincde: Conditional density estimation via lindsey's method. *JMLR*, 23:52:1–52:55, 2022.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan "O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *ICML*, volume 267, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Jiaxi Hu, Yuehong Hu, Wei Chen, Ming Jin, Shirui Pan, Qingsong Wen, and Yuxuan Liang. Attractor memory for long-term time series forecasting: A chaos perspective. In *NeurIPS*, 2024.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. TimeBase: The power of minimalism in efficient long-term time series forecasting. In *International Conference on Machine Learning*, volume 267, pp. 26227–26246. PMLR, 2025a.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *ICLR*, 2025b.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. In *NeurIPS*, pp. 22765–22777, 2022.

Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *ICML*, 2012.

Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, and Mingyuan Zhou. Prototype-oriented unsupervised anomaly detection for multivariate time series. In *ICML*, volume 202, pp. 19407–19424, 2023.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Chenjuan Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *ICLR*, 2026.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764, 2021.

Shengsheng Lin, Weiwei Lin, Keyi Wu, Songbo Wang, Minxian Xu, and James Z Wang. Cocv: A compression algorithm for time-series data with continuous constant values in iot-based monitoring systems. *Internet of Things*, 25:101049, 2024.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. In *NeurIPS*, 2024.

Xiaoye Miao, Yangyang Wu, Jun Wang, Yunjun Gao, Xudong Mao, and Jianwei Yin. Generative semi-supervised learning for multivariate time series imputation. In *AAAI*, volume 35, pp. 8983–8991, 2021.

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

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Y. Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *NeurIPS*, 2025.

Kin G Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

Zhicheng Pan, Yihang Wang, Yingying Zhang, Sean Bin Yang, Yunyao Cheng, Peng Chen, Chenjuan Guo, Qingsong Wen, Xiduo Tian, Yunliang Dou, et al. Magicscaler: Uncertainty-aware, predictive autoscaling. *Proc. VLDB Endow.*, 16(12):3808–3821, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pp. 1185–1196, 2025.

Xiangfei Qiu, Yuhan Zhu, Zhengyu Li, Xingjian Wu, Bin Yang, and Jilin Hu. Dag: A dual correlation network for time series forecasting with exogenous variables. In *ICML*, 2026.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Özbayoglu. Financial time series forecasting with deep learning : A systematic literature review: 2005-2019. *Appl. Soft Comput.*, 90:106181, 2020.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

Wei Shao, Ziquan Fang, Lu Chen, and Yunjun Gao. Towards trajectory anomaly detection: a fine-grained and noise-resilient framework. In *SIGKDD*, pp. 2490–2501, 2025.

Ke-Yuan Shen. Learn hybrid prototypes for multivariate time series anomaly detection. In *ICLR*, 2025.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. MEMTO: Memory-guided transformer for multivariate time series anomaly detection. In *NeurIPS 2023*, 2023.

Kshitij Tayal, Arvind Renganathan, Xiaowei Jia, Vipin Kumar, and Dan Lu. Exotst: Exogenous-aware temporal sequence transformer for time series prediction. In *ICDM*, pp. 857–862, 2024.

Stylianos I Vagropoulos, GI Chouliaras, Evaggelos G Kardakos, Christos K Simoglou, and Anastasios G Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *ENERGYCON*, pp. 1–6, 2016.

Robert A. Vandermeulen, Wai Ming Tai, and Bryon Aragam. Breaking the curse of dimensionality in structured density estimation. In *NeurIPS*, 2024.

Hao Wang, Pan Li, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *NeurIPS*, 2025.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, volume 37, pp. 469–498, 2024.

Billy M Williams. Multivariate vehicular traffic flow prediction: Evaluation of arimax modeling. *Transportation Research Record*, 1776(1):194–200, 2001.

Binqing Wu, Zongjiang Shang, Jianlong Huang, and Ling Chen. Millgnn: Learning multi-scale lead-lag dependencies for multi-variate time series forecasting. In *CIKM*, pp. 3344–3354, 2025.

Xinle Wu, Xingjian Wu, Bin Yang, Lekui Zhou, Chenjuan Guo, Xiangfei Qiu, Jilin Hu, Zhenli Sheng, and Christian S Jensen. AutoCTS++: zero-shot joint neural architecture and hyperparameter search for correlated time series forecasting. *The VLDB Journal*, 33(5):1743–1770, 2024.

Lincen Yang and Matthijs van Leeuwen. Conditional density estimation with histogram trees. In *NeurIPS*, 2024.

Linxiao Yang, Xue Jiang, Gezheng Xu, Tian Zhou, Min Yang, Zhaoyang Zhu, Linyuan Geng, Zhipeng Zeng, Qiming Chen, Xinyue Gu, Rong Jin, and Liang Sun. Baguan-ts: A sequence-native in-context learning model for time series forecasting with covariates. *CoRR*, abs/2603.17439, 2026.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *SIGKDD*, 2025.

A Experimental Details

A.1 Datasets

Table 3: Statistics of datasets. #Num denotes the number of exogenous variables. Ex. and En. are abbreviations for the exogenous variable and endogenous variable, respectively.

Dataset	#Num	Ex. Descriptions	En. Descriptions	Sampling Frequency	Lengths	Split
NP	2	Grid load, wind power	Nord Pool electricity price	1 Hour	52,416	7:1:2
PJM	2	System load, COMED zonal load	COMED zonal electricity price	1 Hour	52,416	7:1:2
BE	2	Belgian load, French generation	Belgian electricity price	1 Hour	52,416	7:1:2
FR	2	Generation, load	French electricity price	1 Hour	52,416	7:1:2
DE	2	Wind power, Amprion zonal load	German electricity price	1 Hour	52,416	7:1:2
Energy	5	Battery, geothermal, hydroelectric, solar, wind	Thermoelectric generation	1 Hour	13,064	7:1:2
Colbun	2	Precipitation, tributary inflow	Water level	1 Day	2,958	7:1:2
Rapel	2	Precipitation, tributary inflow	Water level	1 Day	3,366	7:1:2
Sdwpfh	6	Climate Feature	Active power	1 Hour	14,641	7:1:2
Sdwpfm	6	Climate Feature	Active power	30 Minutes	29,281	7:1:2

We conduct experiments on 12 real-world datasets with exogenous variables, including five electricity price datasets from the EPF benchmark (Lago et al., 2021; Wang et al., 2024) and seven energy and hydropower datasets collected by DAG (Qiu et al., 2026), following KITE (Cheng et al., 2026) and GCGNet (Li et al., 2026). We evaluate forecasting under the setting where future exogenous variables are provided as inputs. **NP** contains hourly electricity prices from the Nord Pool market, with grid-load and wind-power forecasts as exogenous variables. **PJM** records zonal electricity prices in the Commonwealth Edison (COMED) area of the Pennsylvania–New Jersey–Maryland Interconnection, with system-load and COMED zonal-load forecasts as exogenous variables. **BE** and **FR** contain Belgian and French electricity prices, respectively; BE uses Belgian load and French generation forecasts as covariates, while FR uses French generation and load forecasts. **DE** records hourly German electricity prices, with Amprion zonal-load forecasts and wind-power forecasts as exogenous variables. **Energy** provides hourly power-generation data from Chile, where thermoelectric generation is the endogenous target and battery, geothermal, hydroelectric, solar, and wind generation are exogenous variables. **Colbun** and **Rapel** contain daily measurements from two Chilean reservoirs, with water level as the target and precipitation and tributary inflow as covariates. **Sdwpfh1**, **Sdwpfh2**, **Sdwpfm1**, and **Sdwpfm2** record wind-power generation from two turbines at the Longyuan wind farm at hourly and half-hourly resolutions, respectively. Their target is active power output (Patv), and their exogenous variables are six meteorological measurements from ERA5 (Hersbach et al., 2020): temperature, surface pressure, relative humidity, wind speed, wind direction, and total precipitation. All datasets are split chronologically into training, validation, and test sets in a 7:1:2 ratio, with detailed statistics provided in Table 3.

A.2 Baselines

We compare XMatch with 10 competitive baselines for forecasting. Following the treatment of future exogenous variables, we organize these methods into two groups.

Methods with native future-covariate support. This group includes DAG (Qiu et al., 2026), KITE (Cheng et al., 2026), GCGNet (Li et al., 2026), TimeXer (Wang et al., 2024), TFT (Lim et al., 2021), and TiDE (Das et al., 2023). These methods incorporate future exogenous information within their forecasting architectures. They cover several approaches to covariate modeling, including temporal and cross-variable correlation modeling, graph-based interactions, attention mechanisms, and MLP-based fusion. KITE additionally uses exogenous information to guide generative forecasting; here, we evaluate its point forecasts using the same deterministic metrics as the other methods.

Methods augmented with future-covariate fusion. This group consists of DUET (Qiu et al., 2025), CrossLinear (Zhou et al., 2025), Amplifier (Fei et al., 2025), and TimeKAN (Huang et al., 2025b). To incorporate known future exogenous variables, we equip these baselines with the MLP fusion module described in Appendix A.6, following prior covariate forecasting studies (Qiu et al., 2026). The module combines each backbone's output with future exogenous inputs to produce the final endogenous forecast. All baselines are evaluated within the TFB framework (Qiu et al., 2024), using the same dataset splits and forecasting horizons as XMatch.

The code repositories for the baseline models are listed in Table 4.

Table 4: Code repositories for the baseline models.

Model	Code repository	Model	Code repository
DAG	decisionintelligence/DAG	TiDE	thuml/Time-Series-Library
KITE	decisionintelligence/KITE	DUET	decisionintelligence/DUET
GCGNet	decisionintelligence/GCGNet	CrossLinear	mumiao2000/CrossLinear
TimeXer	thuml/TimeXer	Amplifier	aikunyi/amplifier
TFT	google-research/.../tft	TimeKAN	huangst21/TimeKAN

A.3 Evaluation Metrics

To evaluate model performance, we employ Mean Absolute Error (MAE) and Mean Squared Error (MSE) for deterministic forecasting, following the TFB benchmark Qiu et al. (2024).

Mean Squared Error (MSE). The Mean Squared Error (MSE) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual observation. Unlike MAE, MSE penalizes larger errors more severely due to the squaring operation, making it more sensitive to outliers. The mathematical representation of MSE is defined as:

$$\text{MSE} = \frac{1}{K \times T} \sum_{k=1}^{K} \sum_{t=1}^{T} (x_t^k - \hat{x}_t^k)^2, \tag{17}$$

where x_t^k denotes the ground truth and $\hat{x}_t^k$ denotes the predicted value.

Mean Absolute Error (MAE). The Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the verification sample of the absolute differences between prediction and actual observation. The mathematical representation of MAE is given by:

$$\text{MAE} = \frac{1}{K \times T} \sum_{k=1}^{K} \sum_{t=1}^{T} |x_t^k - \hat{x}_t^k|, \tag{18}$$

where K represents the number of time series and T represents the number of time steps.

A.4 Implementation Details

All experiments are conducted using PyTorch (Paszke et al., 2019) in Python 3.10 and executed on an NVIDIA GeForce RTX 3090 GPU. We do not apply the "Drop Last" trick to ensure a fair comparison, retaining the final incomplete batch during evaluation. Testing uses the rolling batch interface of TFB (Qiu et al., 2024) to process all test windows, including the final batch when it contains fewer windows than the batch size. The evaluation code additionally verifies that the number of predictions matches the number of target windows.

A.5 Matching Precision and Historical Support Evaluation

We compare direct matching with *ProtoTree* on Energy and Sdwpfh1 using a history length of 168, a forecast length of 24, and a patch length of $P = 12$. Direct matching uses increasing numbers of exogenous variables, from one to five on Energy and one to six on Sdwpfh1, while *Tree* denotes the full soft-matching result. Both methods use training patches as historical references and are evaluated by weighted DTW-based similarity to the ground-truth future endogenous pattern and effective historical support. The boxplots summarize these metrics over successfully matched queries, with effective support shown on a logarithmic scale.

Weighted DTW-based endogenous pattern similarity. Let $\widetilde{\boldsymbol{y}}_q \in \mathbb{R}^P$ denote the standardized ground-truth future endogenous patch corresponding to query q, and let $\widetilde{\boldsymbol{c}}_r$ denote the standardized retrieved endogenous pattern center r. The ground-truth future endogenous patch is used only as an

evaluation reference and is not used to construct the retrieval query or determine retrieval weights. We compute its DTW-based similarity to each retrieved endogenous pattern and take the weighted average under the retrieved pattern distribution π_q:

$$K_{q,r} = 1 - \frac{\mathrm{DTW}_{\mathrm{band}=2}(\widetilde{\boldsymbol{y}}_q, \widetilde{\boldsymbol{c}}_r)}{2P}, \qquad S_q = \sum_r \pi_{q,r} K_{q,r}. \tag{19}$$

The DTW band permits temporal offsets of at most two time steps. Higher scores indicate that the retrieved endogenous patterns more closely resemble the ground-truth future endogenous pattern for the query. Direct matching assigns equal weight to every matched historical patch, so $\pi_{q,r}$ is the empirical frequency of endogenous pattern r among the matches for query q. For brevity, we omit the query subscript in the aggregation weights and pattern distribution below.

For Tree, we retain the complete aggregation weights w_v from the *ProtoTree Matcher*, excluding the virtual root, and normalize them for this analysis:

$$\alpha_v = \frac{w_v}{\sum_{u \neq v_{\mathrm{root}}} w_u}, \qquad \pi_r = \sum_{v \neq v_{\mathrm{root}}} \alpha_v p_v(r). \tag{20}$$

A Tree query is included when its total non-root weight is positive. We do not select a maximum-weight node or truncate to top-k nodes. Here, $p_v(r)$ is the empirical endogenous pattern distribution at node v. Thus, the similarity score can equivalently be written as $S_q = \sum_{v \neq v_{\mathrm{root}}} \alpha_v \sum_r p_v(r) K_{q,r}$, incorporating both node aggregation weights and within-node pattern frequencies. Applying the same S_q to both methods evaluates the agreement of their retrieved shape evidence with the ground-truth future endogenous pattern before MLP_V, rather than the smoothness of an averaged response curve.

Effective support. Let $\mathcal{M}_v$ be the set of historical patches supporting node v and $n_v = |\mathcal{M}_v|$. The normalized contribution of historical patch i and the effective support are

$$\beta_i = \sum_{v:\, i \in \mathcal{M}_v} \frac{\alpha_v}{n_v}, \qquad N_{\text{eff}} = \frac{1}{\sum_i \beta_i^2}. \tag{21}$$

Contributions of the same patch are combined across nodes before squaring, accounting for overlapping support between ancestors and descendants. For direct matching with n equally weighted historical patches, this definition reduces exactly to $N_{\text{eff}} = n$. Effective support therefore measures the amount of historical evidence under the aggregation weights, rather than the sum of node counts.

A.6 MLP Fusion Module

For a fair comparison, we equip conventional forecasting methods with an efficient MLP fusion module (Qiu et al., 2026), allowing them to incorporate future exogenous information. Given the historical endogenous variables $X^{\text{endo}} \in \mathbb{R}^{N \times T}$ and historical exogenous variables $X^{\text{exo}} \in \mathbb{R}^{D \times T}$, the forecasting backbone first produces a latent representation:

$$z = \theta_{\text{Model}}(X^{\text{endo}}, X^{\text{exo}}), \tag{22}$$

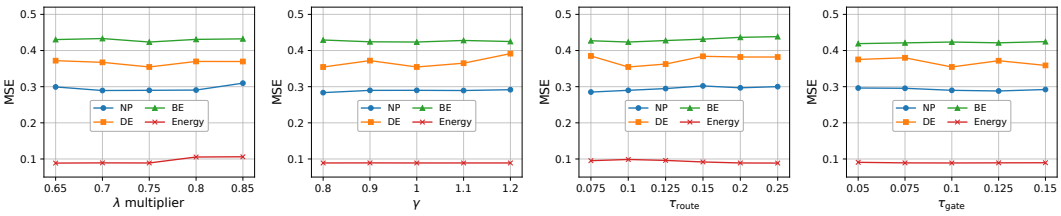
where $z \in \mathbb{R}^{N \times F}$ and θ_{Model} denotes the parameters of the backbone. The resulting representation z is then concatenated with the future exogenous variables $Y^{\text{exo}} \in \mathbb{R}^{D \times F}$ and fed into an MLP to obtain the final forecast:

$$\hat{Y}^{\text{endo}} = \theta_{\text{MLP}}\big(\text{Concat}(z, Y^{\text{exo}})\big), \tag{23}$$

where θ_{MLP} denotes the parameters of the MLP fusion module.

B More Analysis

B.1 Additional Parameter Sensitivity Analysis



(a) Clustering penalty (b) Soft-DTW smoothing (c) Routing temp. (d) Stopping gate temp.

Figure 5: Additional hyperparameter sensitivity of XMatch.

We further study the parameter sensitivity of XMatch, including the clustering penalty multiplier and Soft-DTW smoothing parameter γ in the *ProtoTree Creator*, and the routing and stopping gate temperatures in the *ProtoTree Matcher*. 1) Figure 5a shows the impact of the clustering penalty. The multiplier scales the 90th percentile of Soft-DTW divergences (normalized by P^2) between each variable's normalized training patches and their mean patch to initialize the clustering penalty. Although performance is relatively stable overall, forecasting errors increase slightly toward the lower or upper ends of the tested range in some cases. A small penalty may produce overly fragmented clusters, whereas an excessively large penalty favors fewer, coarser pattern centers that may obscure useful pattern distinctions. A moderate penalty is therefore preferable for preserving distinct patterns while retaining sufficient samples in each cluster. We recommend selecting the penalty multiplier within $[0.7, 0.75]$. 2) Figure 5b illustrates the impact of γ. Performance remains relatively stable across much of the tested range, but larger values increase forecasting errors on some datasets. This suggests that excessive smoothing across temporal alignments may weaken the distinction between different patch shapes. We recommend selecting γ within $[0.8, 1.0]$ to allow flexible temporal alignment while preserving discriminative shape information. 3) As illustrated in Figure 5c, the routing temperature modulates the distribution of aggregation weights across branches, with moderate performance variations as the weights become more diffuse. We recommend $[0.075, 0.125]$ as an initial tuning range, with higher temperatures considered when more diffuse aggregation improves validation performance. 4) Figure 5d indicates low sensitivity to the stopping gate temperature, which controls the smoothness of the continue-or-stop transition during training; inference still uses a hard threshold. We recommend 0.1 as a starting value, with tuning within $[0.05, 0.15]$ if needed. These recommendations serve as practical starting points, with final values selected on the validation set.

B.2 Limitations and Future Work

Despite the promising performance of XMatch, this work still has several limitations.

1) Fixed Variable Ordering. The current framework determines a fixed exogenous variable order based on the information gain between each individual exogenous variable and endogenous patterns. Although this ordering prioritizes informative variables for hierarchical matching, individual information gain may not fully reflect the joint contributions of multiple variables, and the most informative variable combinations may vary across forecasting contexts. Exploring tree construction based on conditional information gain, or dynamically adapting variable selection and matching order to each query, is a direction for future work.

2) Fixed Patch Length. The current framework uses a fixed patch length to construct pattern prototypes. While this design enables aligned pattern extraction and matching, a single temporal scale may not fully capture exogenous–endogenous associations underlying both short-term fluctuations and long-term changes. Developing a multi-scale *ProtoTree* or adaptively selecting patch lengths based on the data is another direction for future work.

B.3 Temporal Recurrence of Exogenous and Endogenous Patterns



Figure 6: **Temporal recurrence of exogenous and endogenous patterns** on (a) NP, (b) FR, and (c) Rapel. The horizontal axis shows calendar time, and the vertical axis indexes patterns. Green and orange marks indicate the occurrence intervals of exogenous and endogenous patterns.

Figure 6 complements the pattern correspondences in Figure 1 by showing when exogenous and endogenous patterns occur. Both types of patterns recur at multiple, separated times. Together, the two figures motivate XMatch from complementary perspectives: Figure 1 shows that an exogenous pattern is often associated with a small set of endogenous patterns, while Figure 6 shows that both types of patterns recur over time. This suggests that their correspondences from the training set may remain useful in forecasting when similar exogenous patterns reappear. Given future exogenous patterns, XMatch can therefore match them against historical exogenous patterns and use the corresponding endogenous patterns to narrow down the plausible future responses and guide prediction.

C FULL RESULTS

C.1 FORECASTING WITH FUTURE EXOGENOUS VARIABLES

Table 5: Full results on forecasting with historical and future exogenous variables across 12 real-world datasets, where the inputs are (X^{endo}, X^{exo}, and Y^{exo}). **Red**: the best, Blue: the 2nd best. Avg means the average results from two forecasting horizons.

Models		XMatch		DAG		KITE		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics		mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	24	0.185	**0.221**	0.202	0.237	**0.179**	0.223	0.208	0.237	0.236	0.266	0.219	0.249	0.284	0.301	0.246	0.287	0.210	0.266	0.252	0.303	0.273	0.310
	360	**0.380**	**0.377**	0.521	0.451	0.471	0.422	0.531	0.459	0.600	0.475	0.539	0.501	0.601	0.498	0.576	0.528	0.531	0.508	0.587	0.534	0.538	0.529
	Avg	**0.282**	**0.299**	0.362	0.344	0.325	0.323	0.370	0.348	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.420	0.418	0.405	0.419
PJM	24	0.060	0.147	0.057	**0.143**	**0.056**	**0.143**	0.060	0.150	0.075	0.166	0.095	0.195	0.106	0.214	0.072	0.166	0.088	0.191	0.096	0.208	0.115	0.244
	360	**0.110**	**0.209**	0.130	0.218	0.135	0.214	0.129	0.223	0.140	0.231	0.133	0.219	0.177	0.279	0.131	0.228	0.135	0.254	0.177	0.285	0.162	0.281
	Avg	**0.085**	**0.178**	0.093	0.180	0.096	0.179	0.095	0.187	0.108	0.198	0.114	0.207	0.142	0.246	0.102	0.197	0.112	0.223	0.137	0.246	0.139	0.262
BE	24	0.361	0.241	0.361	**0.229**	**0.348**	0.240	0.350	0.248	0.392	0.253	0.426	0.272	0.426	0.285	0.432	0.272	0.391	0.259	0.471	0.339	0.451	0.319
	360	**0.479**	**0.304**	0.485	0.330	0.507	0.332	0.511	0.340	0.512	0.327	0.482	0.310	0.571	0.364	0.597	0.436	0.568	0.416	0.646	0.487	0.645	0.495
	Avg	**0.420**	**0.273**	0.423	0.280	0.428	0.286	0.431	0.294	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.559	0.413	0.548	0.407
FR	24	0.357	0.174	0.355	**0.171**	**0.311**	0.178	0.347	0.188	0.366	0.208	0.543	0.253	0.418	0.255	0.384	0.251	0.390	0.226	0.459	0.348	0.454	0.296
	360	0.463	**0.258**	0.473	0.268	**0.462**	0.271	0.482	0.279	0.489	0.273	0.465	0.261	0.551	0.308	0.607	0.403	0.575	0.370	0.648	0.468	0.641	0.452
	Avg	0.410	**0.216**	0.414	0.219	**0.387**	0.225	0.415	0.234	0.427	0.241	0.504	0.257	0.484	0.281	0.496	0.327	0.483	0.298	0.554	0.408	0.547	0.374
DE	24	0.271	**0.320**	0.277	0.322	**0.262**	0.327	0.280	0.331	0.339	0.362	0.380	0.383	0.367	0.383	0.376	0.378	0.387	0.396	0.394	0.407	0.399	0.412
	360	**0.427**	0.418	0.462	0.418	0.437	**0.413**	0.523	0.447	0.610	0.474	0.599	0.509	0.630	0.511	0.589	0.482	0.583	0.507	0.551	0.474	0.547	0.479
	Avg	**0.349**	**0.369**	0.370	0.370	0.350	0.370	0.401	0.389	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.473	0.441	0.473	0.445
Energy	24	**0.061**	**0.191**	0.079	0.215	0.071	0.208	0.081	0.221	0.122	0.273	0.093	0.235	0.103	0.248	0.117	0.283	0.241	0.418	0.138	0.306	0.135	0.298
	360	**0.117**	**0.271**	0.169	0.320	0.151	0.305	0.182	0.334	0.204	0.357	0.167	0.331	0.202	0.355	0.288	0.452	0.237	0.385	0.328	0.472	0.302	0.464
	Avg	**0.089**	**0.231**	0.124	0.267	0.111	0.257	0.131	0.277	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.233	0.389	0.218	0.381
Sdwpfm1	24	0.380	0.409	**0.351**	**0.400**	0.384	0.410	0.376	0.415	0.558	0.533	0.366	0.421	0.474	0.488	0.551	0.564	0.355	0.473	0.364	0.445	0.418	0.503
	360	**0.423**	**0.463**	0.495	0.522	0.450	0.491	0.472	0.499	0.845	0.684	0.597	0.528	0.492	0.526	0.646	0.577	0.497	0.532	0.510	0.535	0.476	0.566
	Avg	**0.402**	**0.436**	0.423	0.461	0.417	0.451	0.424	0.457	0.701	0.609	0.482	0.474	0.483	0.507	0.599	0.570	0.426	0.502	0.437	0.490	0.447	0.534
Sdwpfm2	24	0.446	0.453	**0.372**	**0.414**	0.443	0.450	0.421	0.441	0.627	0.570	0.411	0.458	0.461	0.492	0.445	0.452	0.477	0.536	0.394	0.462	0.474	0.538
	360	**0.492**	**0.505**	0.583	0.556	0.506	0.515	0.529	0.531	0.978	0.736	0.541	0.519	0.511	0.540	0.584	0.528	0.589	0.611	0.587	0.563	0.520	0.589
	Avg	**0.469**	**0.479**	0.477	0.485	0.475	0.483	0.475	0.486	0.803	0.653	0.476	0.488	0.486	0.516	0.514	0.490	0.533	0.573	0.491	0.512	0.497	0.564
Sdwpfh1	24	**0.391**	**0.424**	0.408	0.438	0.424	0.464	0.435	0.486	0.651	0.587	0.401	0.460	0.434	0.489	0.527	0.513	0.548	0.585	0.576	0.627	0.511	0.582
	360	**0.422**	**0.461**	0.489	0.534	0.457	0.517	0.465	0.514	0.841	0.700	0.557	0.523	0.472	0.527	0.551	0.519	0.566	0.601	0.497	0.569	0.643	0.694
	Avg	**0.406**	**0.442**	0.448	0.486	0.441	0.491	0.450	0.500	0.746	0.643	0.479	0.491	0.453	0.508	0.539	0.516	0.557	0.593	0.537	0.598	0.577	0.638
Sdwpfh2	24	**0.417**	**0.446**	0.438	0.465	0.477	0.498	0.473	0.506	0.820	0.677	0.474	0.493	0.579	0.553	0.629	0.563	0.468	0.540	0.473	0.533	0.580	0.614
	360	**0.409**	**0.457**	0.608	0.595	0.524	0.555	0.566	0.565	0.962	0.761	0.657	0.549	0.619	0.614	0.665	0.569	0.608	0.609	0.569	0.628	0.713	0.729
	Avg	**0.413**	**0.451**	0.523	0.530	0.501	0.527	0.520	0.536	0.891	0.719	0.566	0.521	0.599	0.583	0.647	0.566	0.538	0.574	0.521	0.581	0.647	0.672
Colbun	10	0.074	**0.093**	0.061	0.094	**0.055**	0.101	0.065	0.108	0.113	0.172	0.092	0.135	0.089	0.131	0.089	0.134	0.071	0.102	0.071	0.121	0.061	0.101
	30	0.122	**0.206**	0.135	0.215	**0.121**	0.243	0.149	0.243	0.176	0.299	0.383	0.460	0.240	0.322	0.307	0.397	0.182	0.288	0.275	0.370	0.195	0.249
	Avg	0.098	**0.149**	0.098	0.154	**0.088**	0.172	0.107	0.175	0.145	0.235	0.238	0.297	0.164	0.227	0.198	0.266	0.126	0.195	0.173	0.246	0.128	0.175
Rapel	10	0.187	0.213	**0.151**	**0.203**	0.192	0.231	0.211	0.230	0.301	0.308	0.201	0.253	0.228	0.271	0.174	0.219	0.163	0.209	0.181	0.227	0.174	0.231
	30	**0.289**	**0.355**	0.309	0.408	0.296	0.358	0.401	0.384	0.387	0.416	0.409	0.414	0.411	0.432	0.365	0.432	0.340	0.417	0.333	0.416	0.325	0.390
	Avg	0.238	**0.284**	**0.230**	0.305	0.244	0.295	0.306	0.307	0.344	0.362	0.305	0.333	0.320	0.351	0.269	0.326	0.252	0.313	0.257	0.321	0.249	0.311
1st Count		**22**	**29**	4	6	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0