TIME-SERIES FOUNDATION MODELS FOR SHORT-TERM WATER DEMAND FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-Term Water Demand Forecasting (STWDF) helps water-system operators schedule pumps around volatile electricity prices, but many existing forecasting methods require use-case-specific training data and maintenance effort that limit their practical deployment. We investigate whether pretrained Time-Series Foundation Models (TSFMs) can forecast short-term water demand without task-specific training. We evaluate ten openly available TSFMs in a zero-shot univariate setting on 26 real-world water demand datasets and compare against baselines. Seven of the ten TSFMs achieve lower mean Weighted Mean Absolute Percentage Error (wMAPE) than Seasonal Naïve. Chronos-2 has a mean wMAPE of 17.3%, compared with 23.8% for Seasonal Naïve, although several leading TSFMs are practically equivalent. Longer contexts and shorter horizons generally improve forecast quality. Weather and calendar covariates yield no consistent improvement in our setup, while fine-tuning can improve performance. In a comparison with 31 use-case-specific STWDF models, two TSFMs rank in the top three. Several pretrained TSFMs provide strong zero-shot baselines for STWDF in the evaluated setting. All code, data, and experimental results are publicly available on GitHub.

1 INTRODUCTION

Short-term water-demand forecasts help operators of Water Supply Systems (WSSs) schedule pumps around volatile electricity prices (Sun et al., 2026; Magagna et al., 2019; Mala-Jetmarova et al., 2017; Brás et al., 2025). Accurate forecasts can therefore support lower-cost pump schedules while satisfying future demand (Namdari et al., 2025). Existing Short-Term Water Demand Forecasting (STWDF) (Donkor et al., 2014; Ghannam & Hussain, 2024) methods range from statistical models to recent AI-based approaches (Ghannam & Hussain, 2024; Alhendi et al., 2022; Niknam et al., 2022). Recent water-demand forecasting research increasingly uses machine-learning and deep-learning methods (Liu et al., 2024). Most approaches train a time-series model on historical measurements from one water setting and use it to forecast future demand in that setting. This task-specific workflow captures local demand patterns and temporal dependencies but requires a representative dataset for each use case. Pretrained time-series models, commonly called Time-Series Foundation Models (TSFMs), offer an alternative (Liang et al., 2024; Garza et al., 2024; Rasul et al., 2023). They adapt the foundation-model approach used by GPT in natural language processing (Brown et al., 2020) and CLIP in computer vision (Radford et al., 2021). A TSFM learns from a large, heterogeneous pretraining corpus and can then produce forecasts for new tasks or domains without individual training (Dong et al., 2024; Rasul et al., 2023). Foundation models do not require a separate training procedure for each use case. This reduces the need to acquire and maintain use-case-specific data. They can also transfer patterns learned during pretraining, which may improve forecasts when local data are limited. A single pretrained TSFM can serve multiple water meters without separate training for each one, reducing training and maintenance work as deployments scale.

Thus, the goal of this paper is to analyze and evaluate TSFMs in the context of STWDF to address the following overarching research question: *How suitable are openly available TSFMs for STWDF?* We make three contributions. 1) We assess the features and characteristics of openly available TSFMs and select ten models. 2) We identify 26 public STWDF datasets for evaluation. 3) We compare the models with statistical baselines in a zero-shot univariate setting and analyze context length, forecast horizon, covariates, and fine-tuning. We also compare all ten models with 31 task-specific water-

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.

demand forecasting methods in a separate case study. The accompanying public repository[1] contains the reproducible implementation, datasets, experimental results, and additional materials. Section 2 reviews STWDF and TSFMs, Section 3 defines the experimental design, Section 4 presents the results, Section 5 discusses the implications and limitations of our findings, and Section 6 concludes and outlines future work.

2 Foundations and Related Work

STWDF combines recurring demand patterns, operational forecast horizons, and external covariates. TSFMs offer zero-shot forecasting, but their context limits, covariate support, and computational cost determine their suitability for this setting.

2.1 Short-Term Water Demand Forecasting

The literature distinguishes STWDF from long-term water-demand forecasting (Donkor et al., 2014; Ghannam & Hussain, 2024). STWDF predicts demand for a specific entity, such as a reservoir, municipality, or household, over hours to days. Long-term forecasts cover months to years and support strategic planning and infrastructure development (Liu et al., 2024). They typically use monthly or yearly data and socioeconomic or demographic variables such as population growth. STWDF instead models short-term patterns for operational decisions such as pump scheduling (Brás et al., 2025).

Three properties shape our evaluation of STWDF models. First, water-demand series exhibit recurring daily and weekly patterns that partly reflect work schedules and weather conditions (Lüdtke et al., 2021). Models therefore need enough historical context to capture these cycles. Second, operational forecasting requires hourly or sub-hourly resolution and horizons that support pump scheduling and day-ahead energy planning (Brás et al., 2025; Magagna et al., 2019). Third, weather and calendar variables can affect water demand and motivate covariate-aware forecasting (Lin et al., 2025; Wu & Kang, 2024).

2.2 Time-Series Foundation Models

TSFMs apply foundation-model pretraining to time-series tasks such as forecasting, classification, and anomaly detection (Liang et al., 2024). They learn recurring patterns from large, heterogeneous corpora that may include weather, financial, traffic, and telemetry data (Ansari et al., 2025). After pretraining, a TSFM can operate without further training (zero-shot), use a small number of task-specific examples (few-shot), or undergo additional training (fine-tuning). Our primary evaluation uses zero-shot forecasting, which requires no task-specific training. In zero-shot forecasting, the model parameters remain fixed, and the context contains the target history and any supported covariates. Fine-tuning instead updates the model or an adapter with task-specific data (Park et al., 2025).

For forecasting, TSFMs condition their predictions on a historical context window (Dong et al., 2024; Rasul et al., 2023). Depending on the model, this context can also include covariates such as weather and calendar variables. Our evaluation treats TSFMs as forecast functions over structured time-series inputs. Hybrid approaches outside our evaluation, such as ChatTime, also support natural-language inputs and outputs (Wang et al., 2025).

TSFM architectures and training procedures determine their computational requirements, pretraining-data composition, context and horizon limits, and covariate support (Kottapalli et al., 2025; Liang et al., 2024). *Computational requirements* include parameter count, training or fine-tuning time, and inference time. These costs matter when a utility forecasts thousands of water meters. The composition of the *pretraining data*, including the number, domains, and lengths of the time series, affects domain transfer and the interpretation of zero-shot results. Exposure to water-related series may aid transfer, whereas direct overlap with an evaluation collection would confound claims about unseen-data generalization. We therefore report documented domain exposure and possible direct overlap separately in Appendix B. *Context length* controls how many past time steps a model uses, while *horizon length* controls how far ahead it forecasts. For our evaluation, a model must support at

[1] https://anonymous.4open.science/r/tsfm-water-demand-forecasting-C302

least one week of context to capture daily and weekly patterns and a 48-hour horizon for operational planning (Kühnert et al., 2021). *Covariates* can improve STWDF accuracy (Papacharalampous & Langousis, 2022; Que et al., 2024), but they increase model and inference complexity. Consequently, not all TSFMs support them.

2.3 Related Work

We position this paper at the intersection of both research directions. On the one hand, there is a rich body of research on STWDF. Reviews (Ghannam & Hussain, 2024; Niknam et al., 2022; Alhendi et al., 2022; Donkor et al., 2014) characterize the field as methodologically diverse, spanning classical time-series models, regression-based approaches, machine learning, deep learning, and hybrid methods. Traditional time-series methods and artificial neural networks remain common (Niknam et al., 2022). Method performance depends on data availability, temporal granularity, and forecast horizon (Ghannam & Hussain, 2024; Niknam et al., 2022). Ghannam & Hussain (2024) organize the growing range of approaches into a taxonomy. The BWDF benchmark provides a direct comparison: 31 teams evaluated their methods across multiple datasets, and the three leading submissions used hybrid ensembles (Alvisi et al., 2025). For example, Kley-Holsteg et al. (2024) use linear and non-linear models that are specifically adapted to each dataset and aggregated using a smooth Bernstein Online Aggregation (BOA) approach (Wintenberger, 2017). Other recent works explicitly model cross-dataset correlations and meteorological inputs (Que et al., 2024), propose Transformer-based architectures (Lin et al., 2025), or employ transfer learning to reuse knowledge from related infrastructure domains such as electricity consumption (Gil-Gamboa et al., 2025). Most high-performing STWDF models still require training, tuning, or adaptation for a specific utility, water meter, or pumping station. Studies of cross-site transfer and explainable forecasting show that deployed models must handle covariate shift and changing demand patterns while meeting robustness and trust requirements (Gil-Gamboa et al., 2025; Niknam et al., 2022).

Since 2023, TSFMs have extended foundation-model pretraining to forecasting, anomaly detection, classification, and imputation (Liang et al., 2024; Garza et al., 2024; Rasul et al., 2023). Publicly available families include Chronos (Ansari et al., 2024; 2025), TimesFM (Das et al., 2024), and MOIRAI (Woo et al., 2024; Liu et al., 2025). Recent models also support related series and exogenous covariates, which are relevant to operational forecasting (Ansari et al., 2025). Evaluations increasingly cover specific domains in addition to general benchmarks. Broad benchmarks such as GIFT-Eval and *fev-bench* compare pretrained and task-specific forecasting models across multiple domains (Aksu et al., 2024; Shchur et al., 2025). Their broad coverage does not establish how TSFMs behave under the recurring patterns, operational horizons, covariates, and runtime constraints of STWDF. TSFMs have been applied to electricity-price forecasting (Marchesi et al., 2025), retail-sales forecasting (Massa & Fanucchi, 2024), and ISP-traffic forecasting (Liu et al., 2026a). Within water and environmental applications, Zheng et al. (2026) evaluate Chronos for nitrogen forecasting under data scarcity. Unlike task-specific STWDF studies, we evaluate ten pretrained models in a zero-shot setting across 26 water-demand datasets under one protocol. Complementing general TSFM benchmarks, we examine STWDF-specific operational horizons, inference cost, covariates, fine-tuning, and a bounded comparison with BWDF submissions.

3 Experimental Design

The experimental design aims to evaluate whether openly available TSFMs can serve as practical models for STWDF. For this purpose, we first define criteria for selecting suitable TSFMs (see Section 3.1). We then define a set of forecasting baselines (see Section 3.2), select publicly available water demand datasets that cover different settings (see Section 3.3), and specify the evaluation protocol (see Section 3.4). Together, these components provide the basis for the subsequent experiments.

3.1 Models

We compare TSFMs with different architectures and model characteristics. Table 1 lists the evaluated models, and the accompanying repository provides further details. We selected the models from Hugging Face on August 21, 2026. We filtered Hugging Face models by the "Time Series Forecasting" category, ranked them by download count, and examined the first 58 results. After applying the consol-

Table 1: Overview of Selected TSFMs.

Model	Covariate Support	Context/Horizon Steps	# Parameters	Release	Reference
MOMENT-1	No	variable / variable	346M	02/2024	Goswami et al. (2024)
Time-MoE	No	context + horizon $\leq$4096	453M	09/2024	Shi et al. (2025)
TimesFM 2.5	Yes	16384 / 1024	200M	09/2025	Das et al. (2024)
Kairos	No	2048 / 720	50M	09/2025	Feng et al. (2025)
Chronos-2	Yes	8192 / 1024	120M	10/2025	Ansari et al. (2025)
MOIRAI 2.0	Yes	variable / variable	11M	11/2025	Liu et al. (2025)
Timer-S1	No	11520 / 272	8.3B	03/2026	Liu et al. (2026b)
Toto 2.0	No	variable / variable	2.5B	05/2026	Khwaja et al. (2026)
TiRex-2	Yes	variable / variable	83M	07/2026	Podest et al. (2026)
t0-alpha	Yes	variable / 1024+	102M	08/2026	The Forecasting Company (2026)

idation and exclusion criteria below, ten models remained. We removed duplicate implementations of the same model across frameworks, such as separate PyTorch and JAX implementations, because we expected them to behave nearly identically. We consolidated model generations and sizes by selecting the latest generation and, within it, the largest publicly available checkpoint. This rule selects one checkpoint per family without using evaluation performance for selection. We inspected the complete family of each model in the ranked results, including checkpoints outside the considered ranking range. When a publication described a larger private model, we selected the largest checkpoint with publicly available weights. Finally, we excluded models that were specific to a particular application domain or for which no general zero-shot forecasting setup was provided. Eligible models had to meet six requirements. They had to (1) be pretrained on a multi-domain time-series corpus, (2) target arbitrary time-series domains, (3) support univariate zero-shot forecasting from the target history alone, (4) provide freely available weights for research, (5) have an accompanying scientific publication describing the architecture, training procedure, and evaluation, and (6) support at least one week of context and a two-day forecast horizon. One week corresponds to 672 steps for 15-minute data or 168 steps for hourly data. Two days correspond to 192 or 48 steps, respectively (Kühnert et al., 2021). Models available only through a paid API, such as TimeGPT 2, were ineligible.[2] Covariate support was desirable but not required because many TSFMs omit it and it increases model and inference complexity (Papacharalampous & Langousis, 2022).

3.2 Baseline Methods

We compare the selected TSFMs with ARIMA, Seasonal Naïve, and Exponential Smoothing (ETS), three common baselines in water-demand and time-series forecasting (Bougadis et al., 2005; Wu & Kang, 2024; Ansari et al., 2025; Das et al., 2024; Hyndman & Athanasopoulos, 2021). *ARIMA* is an established method that models a time series through autoregressive, differencing, and moving-average components. It is designed to capture serial dependence and non-stationary behavior such as changing levels or trends. In our setting, ARIMA provides a well-known statistical baseline because it can exploit temporal structure without requiring large-scale pretraining. *Seasonal Naïve* repeats the latest observation from the same seasonal position, making it suitable for stable daily or weekly demand patterns. In a water-demand forecasting competition, this baseline outperformed two-thirds of the submitted methods (Alvisi et al., 2025). ETS is a family of models that combines error, trend, and seasonal components within a unified state-space framework, where greater weight is given to more recent observations while still retaining information from older data. ETS is included as a baseline because it is widely used, interpretable, and effective for time series with evolving level, trend, and seasonality. All baseline methods are configured with weekly seasonality, if applicable, since prior studies, backed by a preliminary correlation analysis of the authors, report weekly demand patterns as the most relevant feature in water consumption data after the daily patterns (Papacharalampous & Langousis, 2022).

[2] `https://www.nixtla.io/blog/timegpt-2-announcement`; last accessed: August 21, 2026

3.3 DATASETS

We evaluate 26 datasets from five public collections (Table 2). They contain utility measurements and consumer smart-meter readings at 15-minute or hourly resolution and vary in population size, consumption behavior, and seasonality. We apply a common preprocessing pipeline to all collections.

Table 2: Overview of Datasets Used in the Experiments. Table 7 in Appendix A provides detailed statistics for every dataset.

Collection	Temp. Resolution	# Datasets	# Data Points	Location	Reference
BWDF	1 hour	10	20k	Italy	Alvisi et al. (2025)
Mauth	15 min	1	5k	Germany	Wilhelm et al. (2023)
ReCycle	1 hour	1	46k	Germany	Weyrauch et al. (2024)
SOWEKI	15 min	11	47k - 70k	Germany	Sehmer et al. (2026)
YW	15 min	3	35k	UK	Yor (2018)

For BWDF, we use the ten benchmark datasets from Alvisi et al. (2025) to compare with the reported results. For Mauth, we convert cumulative usage counters to demand rates, downsample the data from 12-second to 15-minute resolution, and aggregate 17 households over a shared period. Before evaluating any model, we select three of the first 150 YW datasets to cover different population sizes and both residential and industrial demand. ReCycle and SOWEKI require only timestamp harmonization. We standardize timestamps and retain each series' flow rate and timestamps. The accompanying repository documents the complete preprocessing procedure. We assessed potential pretraining overlap by comparing each evaluation collection with the publicly documented training corpora of all ten models. Table 8 in Appendix B reports the model- and collection-level findings.

3.4 EVALUATION PROTOCOL

We study zero-shot forecasting alongside four complementary settings.

3.4.1 EVALUATION SETTINGS

Zero-shot forecasting is the primary evaluation scenario. We evaluate every selected TSFM on all datasets using only past water demand as input and compare the resulting forecasts with the baseline methods. For each dataset, all TSFMs and baselines use the same forecast windows. Each forecast covers the subsequent 48 hours and uses up to two weeks of context. On 15-minute datasets, MOMENT-1's fixed input shortens its effective context. The two-week window spans two weekly demand cycles, whose patterns are relevant for STWDF (Papacharalampous & Langousis, 2022). This experiment tests whether pretrained TSFMs outperform statistical baselines without task-specific fitting.

The *context and forecast-horizon analysis* extends the primary zero-shot evaluation by testing how performance changes with more or less history and with shorter or longer horizons. We repeat zero-shot forecasting with one, two, or four weeks of context and horizons of 12 hours, one day, or two days. Kairos and TiRex-2 are excluded because neither supports four-week contexts.

Calendar and weather variables may capture demand patterns that past consumption alone misses. In the *covariate-aware experiment*, we compare univariate forecasts with forecasts using calendar features (day of the week and weekend indicators), weather variables (temperature and humidity), or both (Papacharalampous & Langousis, 2022; Que et al., 2024). We use models with native covariate support (Table 1) and the SOWEKI and BWDF collections, the only two with weather data. All configurations use the same forecast windows. The repository documents the covariates.

The *fine-tuning study* extends the zero-shot evaluation to examine how adaptation changes Chronos-2's univariate forecasts on the same 26 datasets. We compare global and local fine-tuning to test whether patterns shared across datasets help as much as dataset-specific adaptation. Global fine-tuning trains one adapter jointly on all datasets, whereas local fine-tuning trains a separate adapter for each dataset. Both strategies use low-rank adaptation with a learning rate of 10^{-5}, an adapter rank of 8, a scaling factor of 16, a batch size of 32, 2,500 update steps, and a random seed of 0. We evaluate the resulting models on the same held-out windows as the zero-shot configuration. With one model and one random seed, the study offers an initial view of the effect of fine-tuning.

To place the zero-shot results alongside methods developed specifically for water-demand forecasting, we add the *BWDF case study*. It compares the ten zero-shot TSFMs with 31 task-specific methods reported by the challenge (Alvisi et al., 2025). We use the ten challenge datasets and five weeks of context and evaluate forecast horizons of one day and one week. For each horizon, we calculate forecast error for every method and dataset and rank the methods by their average pairwise win rate across the ten datasets. The challenge submissions use their own training, preprocessing, and modeling procedures, while we evaluate the TSFMs zero-shot. These methodological differences restrict the analysis to a BWDF-specific case study and prevent a controlled comparison with locally trained baselines.

3.4.2 Metrics, Statistical Analysis, and Runtime Measurement

We use Weighted Mean Absolute Percentage Error (wMAPE) and Mean Absolute Error (MAE) as the main evaluation metrics. wMAPE enables scale-independent comparisons across datasets and is less sensitive to observations near zero than per-time-step percentage errors. MAE reports the average absolute error in the original unit and is therefore directly interpretable. We summarize cross-dataset wMAPE using the arithmetic mean. For the fine-tuning study, we additionally report the median and interquartile range to reduce the influence of extreme percentage errors.

We assessed differences in forecasting performance across the evaluated models using the nonparametric Friedman test (Friedman, 1937; Demšar, 2006). We treated datasets as blocks and forecasting models as treatments and ranked the models by their wMAPE on each dataset. Following a significant omnibus result, we evaluated pairwise differences using the Conover post-hoc test (Conover & Iman, 1979). We applied Holm's sequential correction (Holm, 1979) to the resulting p-values to control the family-wise error rate across multiple comparisons. We assessed practical relevance with the Bayesian signed-rank test (Benavoli et al., 2014; 2017). For each model pair, the analysis estimates the posterior probabilities that either model performs better or that their difference lies within a predefined region of practical equivalence (ROPE). We define the ROPE as ± 1 wMAPE percentage point and treat differences within this interval as practically negligible. This threshold represents a study-specific definition of practical relevance. We ran the experiments on an Apple Mac Studio M2 Ultra with a 24-core CPU, a 76-core GPU, and 128 GB of unified memory and measured inference time for each model and dataset. Figures 1 and 2 in Appendix C provide detailed total and per-window inference times. We did not measure memory usage, but the largest model, Timer-S1, requires a minimum of approximately 16.5 GB of RAM.

4 Experimental Results

The evaluation covers zero-shot forecasting, context and horizon sensitivity, covariates, Chronos-2 fine-tuning, and a comparison with BWDF submissions.

4.1 Zero-Shot Forecasting

Table 3 reports the aggregate zero-shot results. Tables 9 and 10 in Appendix C provide the dataset-level results. Forecasting error varies across datasets. Mauth has a higher aggregate wMAPE than the higher-volume SOWEKI collection because the same absolute error represents a larger share of its observed demand. Among the baselines, Seasonal Naïve has the lowest mean error, followed by ETS and ARIMA. Seasonal Naïve's performance is consistent with prior STWDF findings (Alvisi et al., 2025), where seasonal baselines are often difficult to beat. Kairos, MOMENT-1, and Time-MoE perform worse than the Seasonal Naïve baseline on average. The signed Diebold–Mariano statistics

Table 3: MAE (A) and wMAPE (B) Results for the Zero-Shot Univariate Experiment.

Model	ARIMA		ETS		SNaive		Chronos-2		Kairos		MOIRAI 2.0		MOMENT-1		t0-alpha		Time-MoE		Timer-S1		TimesFM 2.5		TiRex-2		Toto 2.0	
Metric	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
BWDF	**14.7**	19.2	7.5	9.9	5.9	9.0	**4.2**	**6.5**	20.8	24.3	4.4	6.7	9.2	11.7	4.5	6.8	4.5	6.8	4.3	6.5	4.5	6.7	4.5	6.6	4.3	6.5
Mauth	**0.1**	113.9	0.1	168.2	0.1	105.1	0.0	69.6	0.1	79.6	0.0	72.6	0.1	117.6	0.0	72.4	0.1	78.6	**0.0**	**66.9**	0.0	69.1	0.0	71.3	0.0	71.3
ReCycle	1180.1	50.9	283.7	12.2	269.4	11.6	**199.1**	**8.6**	1320.6	57.0	209.1	9.0	363.1	15.7	212.8	9.2	205.9	8.9	205.5	8.9	210.6	9.1	218.8	9.4	216.4	9.3
SOWEKI	8.5	58.9	13.7	63.1	3.3	29.5	**2.5**	**21.6**	8.2	59.2	2.7	23.1	5.5	52.3	2.6	22.2	5.3	38.4	2.7	22.7	3.0	23.5	2.8	22.7	2.6	22.5
YW	19.3	54.2	6.8	20.4	3.1	12.7	**2.2**	**8.9**	16.3	47.0	2.5	9.8	7.4	30.3	2.4	9.5	4.7	16.4	2.3	9.2	2.5	9.8	2.6	10.0	2.5	9.8
mean	61.3	51.4	23.6	47.8	15.2	23.8	**11.2**	**17.3**	67.9	51.9	11.9	18.5	21.8	41.7	12.0	17.9	13.7	29.6	11.7	18.1	12.1	18.8	12.4	18.3	12.1	18.1
std	225.0	26.9	56.3	49.9	51.1	23.2	**37.8**	16.3	251.8	21.0	39.7	17.2	68.9	45.9	40.4	16.6	39.0	22.2	39.0	**16.2**	40.0	16.7	41.6	16.7	41.1	16.5

indicate that differences from Seasonal Naïve are statistically significant for most datasets, while their direction varies across datasets (see Table 11 in Appendix C). Seven of the ten TSFMs achieve lower mean wMAPE than Seasonal Naïve (23.8%), led by Chronos-2 at 17.3%. Across all models, Chronos-2 achieves the best overall performance and the lowest average rank in the Friedman test. However, the subsequent Conover post-hoc test with Holm correction shows that Chronos-2 does not significantly outperform Timer-S1. The Bayesian signed-rank analysis finds small practical differences between Chronos-2 and several other high-performing models. Chronos-2, t0-alpha, Timer-S1, TimesFM 2.5, and TiRex-2 are classified as practically equivalent within the predefined ROPE. For MOIRAI 2.0 and Toto 2.0, practical equivalence is also highly probable, with posterior probabilities above 90%, although the evidence is somewhat weaker than for the models classified as practically equivalent. In contrast, Chronos-2 is clearly favored over Kairos, MOMENT-1, and Time-MoE. MOIRAI 2.0 and t0-alpha have the lowest per-window inference times, while Time-MoE and Timer-S1 have the highest. Figures 1 and 2 in Appendix C show the total and per-window inference times across the evaluated configurations. Chronos-2 has the lowest aggregate error, MOIRAI 2.0 and t0-alpha have the lowest per-window inference times, and several leading TSFMs remain practically equivalent in forecasting accuracy.

We repeat the evaluation across different context and forecast-horizon lengths. Consistent with findings from other domains (Nie et al., 2023; Hyndman & Athanasopoulos, 2021), longer contexts generally reduce error, while longer horizons increase it. The magnitude of these changes differs across models. MOMENT-1 and Time-MoE reverse the general context-length pattern in several configurations. Additional history does not consistently improve their forecasts. Figure 3 in Appendix D reports all configurations.

4.2 Covariate-Aware Forecasting

Table 4 shows no consistent accuracy improvement from adding covariates. For Chronos-2, t0-alpha, and TiRex-2, the Bayesian signed-rank analysis places most differences between configurations with and without covariates within the predefined ROPE. For MOIRAI 2.0 and TimesFM 2.5, the Friedman test followed by Conover post-hoc comparisons with Holm correction identifies statistically significant error increases after adding covariates. No evaluated model achieves a practically relevant

Table 4: wMAPE Results for Different Sets of Covariates.

Dataset	Covariates	Chronos-2	MOIRAI 2.0	t0-alpha	TimesFM 2.5	TiRex-2
BWDF	None	6.45	6.65	6.80	6.71	6.61
	Weather	6.72	12.35	6.95	9.24	6.65
	Calender	6.67	16.44	7.13	8.68	6.66
	All	6.66	16.48	7.02	10.16	6.66
SOWEKI	None	21.47	23.02	22.06	23.33	22.49
	Weather	22.46	46.66	23.07	29.40	22.39
	Calender	22.10	48.65	22.49	26.46	22.20
	All	22.10	47.65	22.63	29.24	22.23

accuracy gain from covariates in this setting. Figure 4 in Appendix E shows the corresponding per-window inference times. Covariates increase inference cost for all evaluated models, with the largest increases for MOIRAI 2.0 and TiRex-2. The models also integrate covariates differently. MOIRAI 2.0 is primarily a univariate model with limited native covariate support, whereas TimesFM 2.5 combines exogenous information with its base forecast through a separate XReg component. These architectural differences may contribute to the different computational and predictive effects of adding covariates.

4.3 Fine-Tuning Study

Table 5 compares zero-shot Chronos-2 with global and local fine-tuning across the 26 datasets. Table 13 in Appendix F provides the corresponding dataset-level results. For each configuration, the table reports the median wMAPE across datasets, the 25th–75th percentiles, and the change in median wMAPE relative to the zero-shot configuration. Negative differences indicate lower forecast error than in the zero-shot setting. Fine-tuning lowers wMAPE on 23 of 26 datasets with a global adapter and on 24 with local adapters. One global adapter achieves 91% of the median-error reduction of 26 dataset-specific adapters, so global fine-tuning preserves nearly all aggregate gain with fewer

Table 5: Fine-Tuning Results for Chronos-2. The final column reports the change in median wMAPE from the zero-shot configuration in percentage points. Negative values indicate lower error.

Configuration	Median wMAPE (%) ↓	25th–75th Percentiles (%)	Δ from Zero-Shot (pp) ↓
Zero-Shot	10.21	7.10 - 14.69	-
Global Fine-Tuning	8.59	5.38 - 12.16	-1.62
Local Fine-Tuning	8.43	5.30 - 11.97	-1.78

adapters to train and maintain. The fixed single-model, single-seed design limits this conclusion to the evaluated setup.

4.4 Comparison with the BWDF Challenge

The BWDF Challenge provides a domain-specific comparison for the zero-shot results. It reports rankings for 31 task-specific water-demand forecasting methods on the same ten datasets used here (Alvisi et al., 2025). Following the protocol in Section 3.4, we evaluate all ten TSFMs without fitting them to the challenge datasets and rank them alongside the published methods by average pairwise win rate for one-day and one-week forecasts. Table 6 shows that six of the ten highest-ranked entries

Table 6: Combined Rankings on the BWDF Collection for One-Day and One-Week Forecast Horizons. We report the ten highest-ranked entries for the one-day horizon.

Entry	Source (BWDF Rank)	One-Day Rank	One-Week Rank
Kley-Holsteg et al. (2024)	BWDF (1)	1	2
Chronos-2	Zero-shot TSFM	2	6
Time-MoE	Zero-shot TSFM	3	> 15
Zanutto et al. (2024)	BWDF (2)	4	1
Groß & Hans (2024)	BWDF (4)	5	7 / 8
Toto 2.0	Zero-shot TSFM	6	7 / 8
Timer-S1	Zero-shot TSFM	7	4
TimesFM 2.5	Zero-shot TSFM	8	12
TiRex-2	Zero-shot TSFM	9	10
Gabriele et al. (2024)	BWDF (5)	10	11

for one-day forecasts are zero-shot TSFMs, with Chronos-2 and Time-MoE placing second and third. At the one-week horizon, four of those six remain in the top ten, and Timer-S1 places fourth. These results show that TSFMs used without additional fitting can rank alongside methods developed specifically for water-demand forecasting. Their relative ranks change with the forecast horizon, but their strong placement, especially for one-day forecasts, is the central finding of this case study.

5 Discussion & Limitations

In the main univariate evaluation, zero-shot TSFMs use only historical demand, the same input available to the statistical baselines, and require no task-specific fitting. Seven of the ten models achieve lower mean wMAPE than Seasonal Naïve across 26 datasets. The BWDF case study provides a second strong result: six zero-shot TSFMs rank among the top ten for one-day forecasts alongside 31 task-specific methods, with Chronos-2 and Time-MoE placing second and third. These findings make pretrained TSFMs a strong starting point for STWDF. Inference cost and forecast horizon still matter when choosing a model.

Our audit found no documented inclusion of BWDF, Mauth, ReCycle, SOWEKI, or YW in the public pretraining records available for the ten model families (see Table 8 in Appendix B). Detailed training descriptions indicate no direct overlap for Time-MoE, Kairos, Chronos-2, or the Toto 2.0 base checkpoint. Incomplete documentation leaves direct overlap unresolved for MOMENT-1, TimesFM 2.5, MOIRAI 2.0, Timer-S1, and TiRex-2, while the undisclosed t0-alpha corpus prevents assessment. Several models include other hydrological or hydrometeorological series in their pretraining data. These corpora constitute domain exposure, but we found no documented direct overlap with the evaluation collections. The zero-shot results therefore establish performance without task-specific training, while generalization to data unseen during pretraining remains unverified.

The BWDF rankings show that several zero-shot models are competitive with leading task-specific methods on the same ten datasets. Because challenge submissions used different training and preprocessing procedures, the rankings do not isolate the effect of task-specific training. A controlled comparison would require public checkpoints or sufficiently detailed protocols, which the reviewed methods do not provide. The Chronos-2 fine-tuning study instead tests adaptation within one consistent setup.

The results show a trade-off between forecasting accuracy and computational cost. Chronos-2 achieves the strongest aggregate accuracy, whereas MOIRAI 2.0 and t0-alpha have lower per-window inference times. Covariates increase runtime without consistent accuracy gains, and fine-tuning improves Chronos-2 at the cost of training and maintaining adapters. The fine-tuning evidence applies only to Chronos-2 under the evaluated setup: one checkpoint, one random seed, and one fixed LoRA configuration without a hyperparameter-sensitivity analysis. Water-demand patterns can change over time because of shifts in consumer behavior or infrastructure, creating non-stationarity and concept drift (Zhao & Shen, 2025; Lüdtke et al., 2021). Our evaluation does not identify known distribution shifts or compare forecasting accuracy before and after them. Future work should evaluate temporal robustness and determine when adaptation or retraining becomes necessary.

The covariate analysis covers only SOWEKI and BWDF, the available weather and calendar inputs, and five models with native covariate support. The lack of consistent improvement contrasts with prior STWDF studies that found meteorological variables informative (Papacharalampous & Langousis, 2022; Que et al., 2024). One possible explanation is that TSFMs learn recurring demand patterns from the historical context, leaving these covariates with little additional information. For MOIRAI 2.0 and TimesFM 2.5, the additional inputs may then introduce noise, whereas Chronos-2, t0-alpha, and TiRex-2 appear less sensitive to the same inputs. The experiment does not isolate this mechanism, so this explanation remains tentative. Future evaluations should test this explanation by varying context length on identical forecast windows and checking whether covariate gains increase when recurring demand patterns are weaker. Such evaluations should also account for covariate quality, alignment, lags, and scaling. The findings therefore do not establish that covariates generally fail to improve STWDF.

6 Conclusion & Future Work

Several pretrained TSFMs provide strong zero-shot baselines for STWDF, although forecasting accuracy, inference cost, and sensitivity to context and horizon length differ across models. Seven of the ten models outperform the strongest statistical baseline on average. Chronos-2 achieves the best aggregate performance, although several leading models are practically equivalent, while MOIRAI 2.0 and t0-alpha have lower per-window inference times. In the single-model study, both fine-tuning strategies improve Chronos-2 across most datasets under the evaluated configuration. By contrast, the tested covariates provide no consistent accuracy improvement and increase inference time. The BWDF case study shows that several zero-shot TSFMs rank alongside strong challenge submissions, but their relative competitiveness depends on the forecast horizon. We expect future TSFMs to increase in scale and forecasting performance, much as Large Language Models (LLMs) have.

Future work should evaluate fine-tuning across multiple TSFMs, checkpoints, random seeds, adapter configurations, and optimization budgets to determine how consistently adaptation improves STWDF (Gil-Gamboa et al., 2025; Park et al., 2025). Future work could fine-tune a general-purpose TSFM on a large, diverse corpus of water-demand series and test whether the resulting domain-adapted model transfers to unseen datasets without local fine-tuning. These evaluations should compare accuracy gains with training and maintenance costs and determine whether better forecasts reduce pumping costs or improve scheduling reliability. For TSFMs that provide probabilistic forecasts, future work should evaluate calibration and interval coverage. These measures would test whether their uncertainty estimates remain reliable during high-demand periods, which our point-error metrics do not assess. The same methodology could be applied to related domains, such as short-term electric-load forecasting for operational planning and energy optimization. This would test whether the results generalize beyond water-demand forecasting and whether TSFMs also support other infrastructure forecasting tasks.

AI USE STATEMENT

In this work, we used generative AI tools for research ideation and execution, including brainstorming and refining research questions, providing feedback on experimental analyses, and supporting the interpretation of results. We have not used generative AI tools to generate experimental data, and the remaining tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools to identify potentially relevant literature and technical documentation and to draft, restructure, and polish parts of the manuscript. We have reviewed all AI-assisted work by examining original sources before citing them and checking AI-assisted claims and interpretations against the underlying experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproduction of our results, we provide an anonymous repository[3] containing all code used to conduct the evaluation. The paper documents the datasets, preprocessing steps, evaluated models, experimental configurations, evaluation protocol, metrics, and statistical analyses. The appendix provides additional information relevant to reproduction and further studies, including detailed dataset statistics, supplementary analyses, full dataset-level results, runtime measurements, and statistical test results. Together, these materials provide the resources needed to reproduce the reported evaluation and extend it in future work.

REFERENCES

Yorkshire Water leakage DMA 15 minute data, 2018. URL https://datamillnorth.org/dataset/yorkshire-water-leakage-dma-15-minute-data-2zlgn. Accessed: 2026-04-14.

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, 2024. URL https://arxiv.org/abs/2410.10393. arXiv preprint arXiv:2410.10393.

Alya A Alhendi, Ameena S Al-Sumaiti, Feruz K Elmay, James Wescaot, Abdollah Kavousi-Fard, Ehsan Heydarian-Forushani, and Hassan Haes Alhelou. Artificial intelligence for water–energy nexus demand forecasting: a review. *International Journal of Low-Carbon Technologies*, 17: 730–744, 05 2022. ISSN 1748-1317. doi: 10.1093/ijlct/ctac043. URL https://doi.org/10.1093/ijlct/ctac043.

Stefano Alvisi et al. Battle of water demand forecasting. *Journal of Water Resources Planning and Management*, 151(10):04025049, 2025. doi: 10.1061/JWRMD5.WRENG-6887. URL https://ascelibrary.org/doi/abs/10.1061/JWRMD5.WRENG-6887.

Abdul Fatir Ansari et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari et al. Chronos-2: From univariate to universal forecasting, 2025. URL https://arxiv.org/abs/2510.15821. arXiv preprint arXiv:2510.15821.

Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1026–1034, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/benavoli14.html.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017. URL http://jmlr.org/papers/v18/16-305.html.

[3] https://anonymous.4open.science/r/tsfm-water-demand-forecasting-C302/

John Bougadis, Kaz Adamowski, and Roman Diduch. Short-term municipal water demand forecasting. *Hydrological Processes*, 19(1):137–148, 2005. doi: https://doi.org/10.1002/hyp.5763. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.5763.

Tom Brown et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems (2020)*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Marlene Brás, Ana Moura, and António Andrade-Campos. Cost efficiency in water supply systems: An applied review on optimization models for the pump scheduling problem. *European Journal of Operational Research*, 323(1):1–19, 2025. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2024.07.039. URL https://www.sciencedirect.com/science/article/pii/S037722172400599X.

William J Conover and Ronald L Iman. Multiple comparisons using rank sums. *Technometrics*, 21 (4):489–495, 1979.

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 (2024)*, ICML'24. JMLR.org, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3692474.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. ISSN 07350015. URL http://www.jstor.org/stable/1392185.

Qingxiu Dong et al. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (2024)*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Emmanuel A. Donkor, Thomas A. Mazzuchi, Refik Soyer, and J. Alan Roberson. Urban water demand forecasting: Review of methods and models. *Journal of Water Resources Planning and Management*, 140(2):146–159, 2014. doi: 10.1061/(ASCE)WR.1943-5452.0000314. URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0000314.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models. *arXiv preprint arXiv:2509.25826*, September 2025. doi: 10.48550/arXiv.2509.25826. URL https://arxiv.org/abs/2509.25826.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

Annalaura Gabriele, Daniela Biondi, Rudy Gargano, and Ezio Todini. From deterministic to probabilistic forecasts of water demand. *Engineering Proceedings (2024)*, 69(1), 2024. ISSN 2673-4591. doi: 10.3390/engproc2024069122. URL https://www.mdpi.com/2673-4591/69/1/122.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1, 2024. URL https://arxiv.org/abs/2310.03589. arXiv preprint arXiv:2310.03589.

Safa Ghannam and Farookh Hussain. Short-term water demand forecasting: a review. *Australasian Journal of Water Resources*, 0(0):1–19, 2024. doi: 10.1080/13241583.2024.2350102. URL https://doi.org/10.1080/13241583.2024.2350102.

A. Gil-Gamboa, J.F. Torres, F. Martínez-Álvarez, and A. Troncoso. Energy-efficient transfer learning for water consumption forecasting. *Sustainable Computing: Informatics and Systems*, 46:101130, 2025. ISSN 2210-5379. doi: https://doi.org/10.1016/j.suscom.2025.101130. URL https://www.sciencedirect.com/science/article/pii/S2210537925000502.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, July 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Matthias Groß and Lukas Hans. Leveraging potentials of local and global models for water demand forecasting. *Engineering Proceedings (2024)*, 69(1), 2024. ISSN 2673-4591. doi: 10.3390/engproc2024069129. URL https://www.mdpi.com/2673-4591/69/1/129.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition, 2021. ISBN 9780987507136. URL https://books.google.de/books?id=gZB-zgEACAAJ.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, May 2026. doi: 10.48550/arXiv.2605.20119. URL https://arxiv.org/abs/2605.20119.

Jens Kley-Holsteg, Björn Sonnenschein, Gregor Johnen, and Florian Ziel. Water demand forecasting based on online aggregation for district meter areas-specific adaption. *Engineering Proceedings (2024)*, 69(1), 2024. ISSN 2673-4591. doi: 10.3390/engproc2024069015. URL https://www.mdpi.com/2673-4591/69/1/15.

Siva Rama Krishna Kottapalli, Karthik Hubli, Sandeep Chandrashekhara, Garima Jain, Sunayana Hubli, Gayathri Botla, and Ramesh Doddaiah. Foundation models for time series: A survey, 2025. URL https://arxiv.org/abs/2504.04011. arXiv preprint arXiv:2504.04011.

Christian Kühnert, Naga Mamatha Gonuguntla, Helene Krieg, Dimitri Nowak, and Jorge A. Thomas. Application of LSTM networks for water demand prediction in optimal pump control. *Water*, 13(5), 2021. ISSN 2073-4441. doi: 10.3390/w13050644. URL https://www.mdpi.com/2073-4441/13/5/644.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2024)*, KDD '24, pp. 6555–6565. ACM, August 2024. doi: 10.1145/3637528.3671451. URL http://dx.doi.org/10.1145/3637528.3671451.

Peijie Lin, Xiangxin Zhang, Longcong Gong, Jingwei Lin, Jie Zhang, and Shuying Cheng. Multi-timescale short-term urban water demand forecasting based on an improved patchtst model. *Journal of Hydrology*, 651:132599, 2025. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2024.132599. URL https://www.sciencedirect.com/science/article/pii/S0022169424019954.

Chanjuan Liu, Jing Xu, Xi'an Li, Zhongyao Yu, and Jinran Wu. Water resource forecasting with machine learning and deep learning: A scientometric analysis. *Artificial Intelligence in Geosciences*, 5:100084, 2024. ISSN 2666-5441. doi: https://doi.org/10.1016/j.aiig.2024.100084. URL https://www.sciencedirect.com/science/article/pii/S266654412400025X.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. MOIRAI 2.0: When less is more for time series forecasting, 2025. URL https://arxiv.org/abs/2511.11698. arXiv preprint arXiv:2511.11698.

Fan Liu, Behrooz Farkiani, and Patrick Crowley. Time-series foundation models for ISP traffic forecasting, 2026a. URL https://arxiv.org/abs/2511.17529. arXiv preprint arXiv:2511.17529.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-S1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, March 2026b. doi: 10.48550/arXiv.2603.04791. URL https://arxiv.org/abs/2603.04791.

Deike U. Lüdtke, Robert Luetkemeier, Michael Schneemann, and Stefan Liehr. Increase in daily household water demand during the first wave of the COVID-19 pandemic in Germany. *Water*, 13(3), 2021. ISSN 2073-4441. doi: 10.3390/w13030260. URL https://www.mdpi.com/2073-4441/13/3/260.

D. Magagna, G. Bidoglio, I. Hidalgo Gonzalez, E. Peteves, M. Adamovic, B. Bisselink, M. De Felice, A. De Roo, C. Dorati, D. Ganora, H. Medarac, P. Pistocchi, W. Van De Bund, and D. Vanham. Water - energy nexus in Europe. Scientific analysis or review, Anticipation and foresight KJ-NA-29743-EN-N (online), European Commission, Luxembourg (Luxembourg), 2019.

Helena Mala-Jetmarova, Nargiz Sultanova, and Dragan Savic. Lost in optimisation of water distribution systems? a literature review of system operation. *Environmental Modelling & Software*, 93:209–254, 2017. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2017.02.009. URL https://www.sciencedirect.com/science/article/pii/S1364815216307769.

Gabriele Marchesi, Andrea Ballarino, and Alessandro Brusaferri. Assessing time series foundation models for probabilistic electricity price forecasting: Toward a unified benchmark. *Energies*, 18 (23), 2025. ISSN 1996-1073. doi: 10.3390/en18236269. URL https://www.mdpi.com/1996-1073/18/23/6269.

Kelian Massa and Dario Fanucchi. Domain-adapted lag-llama for time series forecasting in the African retail sector. In *NeurIPS Workshop on Time Series in the Age of Large Models (2024)*, 2024. URL https://openreview.net/forum?id=mTCMfeK4jR.

Hossein Namdari, Amin E. Bakhshipour, Seyed Mohammad Ashrafi, and Ali Haghighi. Pump scheduling optimization in water distribution networks, including short-term demand forecasting by deep learning. *Sustainable Water Resources Management*, 12(1):4, 2025. ISSN 2363-5045. doi: 10.1007/s40899-025-01316-2. URL https://doi.org/10.1007/s40899-025-01316-2.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations (2023)*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Azar Niknam, Hasan Khademi Zare, Hassan Hosseininasab, Ali Mostafaeipour, and Manuel Herrera. A critical review of short-term water demand forecasting tools—what method should I use? *Sustainability*, 14(9), 2022. ISSN 2071-1050. doi: 10.3390/su14095412. URL https://www.mdpi.com/2071-1050/14/9/5412.

Georgia Papacharalampous and Andreas Langousis. Probabilistic water demand forecasting using quantile regression algorithms. *Water Resources Research*, 58(6), 2022. ISSN 1944-7973. doi: 10.1029/2021wr030216. URL http://dx.doi.org/10.1029/2021WR030216.

Young-Jin Park, François Germain, Jing Liu, Ye Wang, Toshiaki Koike-Akino, Gordon Wichern, Navid Azizan, Christopher Laughman, and Ankush Chakrabarty. Probabilistic forecasting for building energy systems using time-series foundation models. *Energy and Buildings*, 348:116446, 2025. ISSN 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2025.116446. URL https://www.sciencedirect.com/science/article/pii/S0378778825011764.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, July 2026. doi: 10.48550/arXiv.2607.01204. URL https://arxiv.org/abs/2607.01204.

Qidong Que, Jinliang Gao, and Yizhou Qian. Water demand forecasting in multiple district metered areas based on a multi-scale correction module neural network architecture. *Water Research X*, 25: 100269, 2024. ISSN 2589-9147. doi: https://doi.org/10.1016/j.wroa.2024.100269. URL https://www.sciencedirect.com/science/article/pii/S2589914724000598.

Alec Radford et al. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning (2021)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Kashif Rasul et al. Lag-Llama: Towards foundation models for time series forecasting. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (2023)*, 2023. URL https://openreview.net/forum?id=jYluzCLFDM.

Felix Sehmer, Maximilian Hoffmann, Norberth Kloos, and Ralph Bergmann. SOWEKI water demand dataset, Apr. 2026. URL https://doi.org/10.5281/zenodo.19627517.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. 2025.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *The Thirteenth International Conference on Learning Representations (2025)*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Qingkai Sun, Haifeng Zheng, Zheng Zhao, Menghua Fan, Xiaojun Wang, and Yiru Shi. Negative electricity prices in electricity markets: causes, impacts, and response strategies. *Global Energy Interconnection*, 2026. ISSN 2096-5117. doi: https://doi.org/10.1016/j.gloei.2025.12.004. URL https://www.sciencedirect.com/science/article/pii/S2096511726000071.

The Forecasting Company. t0: A time-series forecasting foundation model. Hugging Face model card, 2026. URL https://huggingface.co/theforecastingcompany/t0-alpha. Model release: t0-alpha.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. ChatTime: a unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (2025)*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i12.33384. URL https://doi.org/10.1609/aaai.v39i12.33384.

Arvid Weyrauch, Oskar Taubert, Ewa Götz, Achim Streit, Markus Götz, Charlotte Debus, and Frank Wolter. Hourly water demand of a mixed district recorded by supplier, Apr. 2024. URL https://doi.org/10.5281/zenodo.11045013.

Sebastian Wilhelm, Jakob Kasbauer, Dietmar Jakob, Benedikt Elser, and Diane Ahrens. Smart meter water consumption measurements, Feb. 2023. URL https://doi.org/10.5281/zenodo.7506076.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, Jan 2017. ISSN 1573-0565. doi: 10.1007/s10994-016-5592-6. URL https://doi.org/10.1007/s10994-016-5592-6.

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 (2024)*, ICML'24. JMLR.org, 2024.

Wenhong Wu and Yunkai Kang. Ensemble empirical mode decomposition granger causality test dynamic graph attention transformer network: Integrating transformer and graph neural network models for multi-sensor cross-temporal granularity water demand forecasting. *Applied Sciences*,

14(8), 2024. ISSN 2076-3417. doi: 10.3390/app14083428. URL https://www.mdpi.com/2076-3417/14/8/3428.

Dennis Zanutto et al. A water futures approach on water demand forecasting with online ensemble learning. *Engineering Proceedings (2024)*, 69(1), 2024. ISSN 2673-4591. doi: 10.3390/engproc2024069060. URL https://www.mdpi.com/2673-4591/69/1/60.

Lifan Zhao and Yanyan Shen. Proactive model adaptation against concept drift for online time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (2025)*, KDD '25, pp. 2020–2031. ACM, July 2025. doi: 10.1145/3690624.3709210. URL http://dx.doi.org/10.1145/3690624.3709210.

Kaifeng Zheng, Guotao Cui, Yi Xie, Yimei Liu, and Xinzhong Du. Time series foundation model Chronos enhances nitrogen forecasting under data scarcity. *Water Research X*, 30:100469, 2026. ISSN 2589-9147. doi: https://doi.org/10.1016/j.wroa.2025.100469. URL https://www.sciencedirect.com/science/article/pii/S2589914725001677.

A Dataset Details

Table 7 reports descriptive statistics for all 26 preprocessed water-demand time series. *Count* denotes the number of observations, *Start* and *End* delimit the available observation period, and the remaining columns summarize the demand values in their original units.

Table 7: Characteristics of the 26 Water-Demand Time Series Used in the Evaluation.

Dataset	Count	Mean	Std	Min	Max	Start	End
BWDF 01	18901	29.61	10.33	5.04	107.09	2021-01-01T15:00	2023-03-31T21:00
BWDF 02	19071	34.49	6.05	22.57	75.66	2021-01-01T15:00	2023-03-31T21:00
BWDF 03	19574	15.03	5.33	5.36	42.03	2020-12-31T23:00	2023-03-31T21:00
BWDF 04	18719	118.41	24.68	45.58	201.44	2021-01-01T15:00	2023-03-31T21:00
BWDF 05	18921	283.31	53.09	159.66	409.09	2021-01-01T15:00	2023-03-31T21:00
BWDF 06	17775	30.35	7.57	9.44	68.78	2021-02-14T19:00	2023-03-31T21:00
BWDF 07	18163	90.70	19.74	29.68	151.57	2021-01-01T15:00	2023-03-31T21:00
BWDF 08	18566	76.46	22.52	33.70	299.28	2021-01-01T15:00	2023-03-31T21:00
BWDF 09	18168	75.85	13.73	48.52	150.28	2021-02-11T10:00	2023-03-31T21:00
BWDF 10	18761	93.88	16.14	49.47	160.07	2021-01-01T15:00	2023-03-31T21:00
Mauth	5010	0.06	0.19	0.01	5.72	2021-05-03T07:30	2021-06-24T11:45
ReCycle	46032	2314.34	1032.14	1.00	6514.00	2016-01-01T00:00	2021-04-01T23:00
SOWEKI 01	47448	5.97	2.73	-0.01	32.02	2024-09-06T22:00	2026-01-28T22:45
SOWEKI 02	70052	9.05	4.51	0.00	65.52	2024-01-01T23:00	2025-12-31T23:00
SOWEKI 03	70177	3.94	3.10	0.00	86.52	2023-12-31T23:00	2025-12-31T23:00
SOWEKI 04	47448	5.27	2.65	0.05	48.47	2024-09-06T22:00	2026-01-28T22:45
SOWEKI 05	70051	56.62	38.49	0.00	128.79	2023-12-31T23:00	2025-12-30T23:00
SOWEKI 06	70053	19.28	8.63	2.05	63.92	2023-12-31T23:00	2025-12-30T23:00
SOWEKI 07	70177	13.22	10.19	0.00	118.51	2023-12-31T23:00	2025-12-31T23:00
SOWEKI 08	67056	0.58	0.59	0.00	27.13	2023-12-31T23:00	2025-12-30T23:00
SOWEKI 09	70177	4.42	2.12	0.22	52.44	2023-12-31T23:00	2025-12-31T23:00
SOWEKI 10	68579	13.05	14.33	0.00	67.52	2023-12-31T23:00	2025-12-30T23:00
SOWEKI 11	70177	8.54	2.98	0.00	96.08	2023-12-31T23:00	2025-12-31T23:00
YW 002	35036	27.07	12.99	6.08	73.36	2016-04-01T00:00	2017-03-31T23:45
YW 017	35032	56.96	28.38	12.76	587.92	2016-04-01T00:00	2017-03-31T23:45
YW 108	35032	1.74	0.48	0.56	6.36	2016-04-01T00:00	2017-03-31T23:45

The series differ substantially in temporal coverage, number of observations, demand scale, and variability. This heterogeneity provides important context for the dataset-level forecasting errors. In particular, MAE remains dependent on the scale of each series, whereas wMAPE expresses absolute errors relative to total observed demand. Low-volume series may therefore exhibit small absolute errors but comparatively high percentage errors.

B Pretraining Contamination Audit

We distinguish direct overlap with the five evaluation collections from exposure to other hydrological or hydrometeorological time series. Table 8 summarizes our review of public model publications and checkpoint documentation. The model citations identify the publication used for each row. The notes below the table document checkpoint-specific sources and unresolved mappings.

To determine the table entries, we searched the model publications, model cards, and available pretraining-data documentation for the names and identifiers of the five evaluation collections. Where the available documentation did not permit a direct match, we additionally compared collection release dates with the reported pretraining period and examined documented exposure to other water-related time series. Because the completeness of pretraining documentation differs substantially

Table 8: Pretraining Contamination Audit for the Five Evaluation Collections. Status entries summarize the evidence of direct overlap between each model's documented pretraining data and the evaluation collections.

Model	BWDF	Mauth	ReCycle	SOWEKI	YW	Other Water-Related Pretraining	Confidence
MOMENT-1	ND	ND	No (date)	No (date)	ND	Likely exposure to the Saugeen river-flow collection through the Monash archive (Goswami et al., 2024)	Medium
Time-MoE	No	No	No	No	No	Saugeen, precipitation, ERA5, CMIP6, and Temperature-Rain (Shi et al., 2025)	High
TimesFM 2.5	ND	ND	ND	No (date)	ND	Hydrometeorological series through GiftEval-Pretrain; synthetic and augmented data (Das et al., 2024)	Medium
Kairos	No	No	No	No	No	ERA5, CMIP6, WeatherBench, and subseasonal precipitation (Feng et al., 2025)	High
Chronos-2	No	No	No	No	No	WeatherBench, USHCN, and Temperature-Rain (Ansari et al., 2025)	Very high
MOIRAI 2.0	ND	ND	ND	No (date)	ND	Climate and weather data through public GIFT-Eval and Chronos-derived components (Liu et al., 2025)	Medium
Timer-S1	ND	ND	ND	No (date)	ND	Meteorological data; TimeBench also incorporates Chronos and LOTSA data (Liu et al., 2026b)	Medium–low
Toto 2.0	No	No	No	No	No	No documented real water data for the base checkpoint (Khwaja et al., 2026)	Very high
TiRex-2	ND	ND	ND	ND	ND	One documented training variant contains a generically named hydrology dataset (Podest et al., 2026)	Medium; checkpoint-dependent
t0-alpha	Unknown	Unknown	Unknown	Unknown	Unknown	Pretraining corpus undisclosed (The Forecasting Company, 2026)	Not assessable

Status definitions. No: the published training manifest or procedure excludes the collection. No (date): the collection postdates the documented pretraining period. This status assumes no private access before release. ND: the audit found no documented match, but incomplete disclosure prevents proof of absence. Unknown: the available documentation is insufficient for assessment. Confidence refers to the evidence against exact overlap.

across models, the assigned statuses express the strength of the available evidence rather than a uniform guarantee that overlap is absent.

The official model card identifies the audited Toto 2.0 base checkpoint as `Datadog/Toto-2.0-2.5B`.[4] The fine-tuned checkpoint uses additional public forecasting data. For TiRex-2, one published corpus variant includes a dataset labeled only as `hydrology`, whose origin remains unresolved.[5] The experimental design names only the Toto 2.0 and TiRex-2 model families. Checkpoint-specific conclusions remain provisional until we record the evaluated checkpoint identifiers.

Taken together, the audit provides no documented evidence of direct overlap, but it can rule out overlap only where the pretraining corpora or periods are sufficiently disclosed. The results should therefore be interpreted as an evidence-based contamination assessment rather than proof that every evaluation series was unseen during pretraining.

C Extended Zero-Shot Forecasting Results

This section provides the dataset-level forecasting errors and runtime measurements underlying the aggregate zero-shot results reported in the main text.

C.1 Dataset-Level Forecasting Results

Tables 9 and 10 report the results for every model, baseline, and time series. The lowest error for each series is highlighted in bold. Each time series has a fixed wMAPE denominator, so wMAPE and MAE produce the same within-series model ordering while providing complementary views of relative and absolute error.

[4] `https://huggingface.co/Datadog/Toto-2.0-2.5B`; accessed September 21, 2026.

[5] `https://huggingface.co/NX-AI/TiRex-2`; accessed September 21, 2026.

Table 9: Dataset-Level wMAPE Results for All Models and Baselines in the Zero-Shot Univariate Evaluation. Lower values indicate better forecasting performance.

Dataset	ARIMA	ETS	SNaive	Chronos-2	Kairos	MOIRAI 2.0	MOMENT-1	t0-alpha	Time-MoE	Timer-S1	TimesFM 2.5	TiRex-2	Toto 2.0
BWDF 01	22.3	16.4	16.3	12.6	28.8	12.8	17.3	13.0	12.9	12.7	12.9	12.5	**12.1**
BWDF 02	12.1	7.0	6.7	**4.5**	16.6	4.6	7.3	4.7	4.8	4.6	4.6	4.6	4.6
BWDF 03	34.9	10.8	11.7	8.2	30.4	8.2	14.6	8.4	8.4	**8.2**	8.4	8.4	8.4
BWDF 04	17.9	9.6	9.5	**7.0**	26.0	7.3	10.8	7.5	7.3	7.1	7.2	7.2	7.3
BWDF 05	15.5	7.6	3.7	**2.7**	25.3	3.0	9.4	3.1	2.9	2.9	3.1	3.2	2.9
BWDF 06	16.5	12.5	13.0	9.4	19.9	9.5	11.1	9.6	10.0	9.4	9.6	**9.4**	9.6
BWDF 07	16.3	7.1	6.5	**4.5**	25.6	4.9	10.3	5.0	5.0	4.6	4.9	4.9	4.7
BWDF 08	24.2	12.2	7.3	**4.8**	38.2	5.2	14.1	5.5	5.3	5.0	5.3	5.2	4.9
BWDF 09	15.7	8.1	8.0	5.6	13.7	5.8	10.3	5.7	6.1	5.7	5.8	5.6	**5.5**
BWDF 10	16.2	7.7	7.2	5.2	18.0	5.2	11.8	5.5	5.6	**5.1**	5.4	5.1	5.1
Mauth	113.9	168.2	105.1	69.6	79.6	72.6	117.6	72.4	78.6	**66.9**	69.1	71.3	71.3
ReCycle	50.9	12.2	11.6	**8.6**	57.0	9.0	15.7	9.2	8.9	8.9	9.1	9.4	9.3
SOWEKI 01	46.8	24.9	20.0	**14.8**	53.0	15.5	24.5	15.2	27.5	15.4	15.4	15.1	16.7
SOWEKI 02	57.0	51.1	19.6	**14.4**	56.9	15.7	26.3	15.3	28.3	15.2	15.8	15.3	16.1
SOWEKI 03	54.2	39.9	31.0	**20.8**	57.2	21.7	32.3	21.6	34.9	21.3	21.7	20.9	22.0
SOWEKI 04	45.1	32.2	23.7	**16.7**	52.6	17.3	28.5	17.2	30.0	17.6	17.4	17.0	18.5
SOWEKI 05	70.6	184.2	25.5	**20.7**	66.2	21.8	51.4	21.4	45.9	23.7	26.7	24.8	22.1
SOWEKI 06	62.5	26.5	13.2	**9.7**	51.3	11.0	21.5	10.8	23.6	10.5	10.9	10.7	10.6
SOWEKI 07	57.3	26.5	19.5	**13.5**	61.8	15.5	46.6	14.8	31.4	14.6	15.3	14.8	14.3
SOWEKI 08	75.2	116.5	81.8	**61.8**	73.6	66.7	233.5	62.7	68.6	63.0	63.1	63.0	63.8
SOWEKI 09	42.7	20.6	17.3	**12.3**	47.1	13.0	20.2	12.7	24.8	13.0	12.9	12.5	13.6
SOWEKI 10	96.5	133.7	55.0	40.4	94.2	42.9	66.9	40.0	84.3	42.8	45.9	42.2	**37.5**
SOWEKI 11	33.0	19.5	15.6	**10.7**	35.1	11.4	15.2	10.8	19.0	10.9	11.1	10.9	11.5
YW 002	76.3	13.6	10.0	**7.5**	60.0	8.5	25.0	8.3	17.1	7.8	8.6	9.1	8.8
YW 017	65.1	28.7	11.1	**7.5**	57.1	8.9	25.8	8.4	16.0	8.0	8.7	9.1	8.5
YW 108	21.1	19.0	17.2	**11.7**	23.9	12.1	40.0	12.0	16.0	11.9	12.2	11.8	12.1
mean	51.4	47.8	23.8	**17.3**	51.9	18.5	41.7	17.9	29.6	18.1	18.8	18.3	18.1
std	26.9	49.9	23.2	16.3	21.0	17.2	45.9	16.6	22.2	**16.2**	16.7	16.7	16.5

Table 10: Dataset-Level MAE Results for All Models and Baselines in the Zero-Shot Univariate Evaluation. Values are reported in the original measurement unit of each dataset. Lower values indicate better forecasting performance.

Dataset	ARIMA	ETS	SNaive	Chronos-2	Kairos	MOIRAI 2.0	MOMENT-1	t0-alpha	Time-MoE	Timer-S1	TimesFM 2.5	TiRex-2	Toto 2.0
BWDF 01	6.7	4.9	4.9	3.8	8.6	3.8	5.2	3.9	3.8	3.8	3.8	3.7	**3.6**
BWDF 02	4.2	2.4	2.3	**1.5**	5.8	1.6	2.5	1.6	1.7	1.6	1.6	1.6	1.6
BWDF 03	5.2	1.6	1.8	1.2	4.6	1.2	2.2	1.3	1.3	**1.2**	1.3	1.3	1.3
BWDF 04	21.3	11.3	11.2	**8.3**	30.8	8.6	12.8	8.9	8.6	8.4	8.6	8.5	8.7
BWDF 05	43.8	21.6	10.4	**7.7**	71.8	8.6	26.5	8.7	8.4	8.1	8.8	9.1	8.3
BWDF 06	5.0	3.8	3.9	2.8	6.0	2.9	3.3	2.9	3.0	2.8	2.9	**2.8**	2.9
BWDF 07	14.9	6.5	5.9	**4.1**	23.5	4.5	9.4	4.5	4.5	4.2	4.5	4.5	4.3
BWDF 08	18.4	9.3	5.5	**3.7**	29.1	4.0	10.7	4.2	4.0	3.8	4.0	4.0	3.7
BWDF 09	11.9	6.1	6.1	4.2	10.4	4.4	7.8	4.4	4.6	4.3	4.4	4.2	**4.2**
BWDF 10	15.3	7.2	6.8	4.9	17.0	4.9	11.1	5.2	5.2	**4.8**	5.1	4.8	4.8
Mauth	0.1	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	**0.0**	0.0	0.0	0.0
ReCycle	1180.1	283.7	269.4	**199.1**	1320.6	209.1	363.1	212.8	205.9	205.5	210.6	218.8	216.4
SOWEKI 01	2.7	1.5	1.2	**0.9**	3.1	0.9	1.4	0.9	1.6	0.9	0.9	0.9	1.0
SOWEKI 02	5.2	4.6	1.8	**1.3**	5.2	1.4	2.4	1.4	2.6	1.4	1.4	1.4	1.5
SOWEKI 03	2.1	1.6	1.2	**0.8**	2.3	0.9	1.3	0.9	1.4	0.8	0.9	0.8	0.9
SOWEKI 04	2.3	1.7	1.2	**0.9**	2.7	0.9	1.5	0.9	1.5	0.9	0.9	0.9	0.9
SOWEKI 05	40.0	104.5	14.5	**11.7**	37.5	12.4	29.2	12.1	26.0	13.4	15.1	14.1	12.5
SOWEKI 06	12.0	5.1	2.5	**1.9**	9.9	2.1	4.1	2.1	4.5	2.0	2.1	2.1	2.0
SOWEKI 07	7.6	3.5	2.6	**1.8**	8.2	2.0	6.2	2.0	4.1	1.9	2.0	2.0	1.9
SOWEKI 08	0.4	0.7	0.5	**0.4**	0.4	0.4	1.4	0.4	0.4	0.4	0.4	0.4	0.4
SOWEKI 09	1.9	0.9	0.8	**0.5**	2.1	0.6	0.9	0.6	1.1	0.6	0.6	0.6	0.6
SOWEKI 10	12.4	17.1	7.1	5.2	12.1	5.5	8.6	5.1	10.8	5.5	5.9	5.4	**4.8**
SOWEKI 11	2.8	1.7	1.3	**0.9**	3.0	1.0	1.3	0.9	1.6	0.9	1.0	0.9	1.0
YW 002	20.7	3.7	2.7	**2.0**	16.3	2.3	6.8	2.2	4.6	2.1	2.3	2.5	2.4
YW 017	36.8	16.2	6.3	**4.2**	32.4	5.1	14.6	4.8	9.0	4.5	4.9	5.2	4.8
YW 108	0.4	0.3	0.3	**0.2**	0.4	0.2	0.7	0.2	0.3	0.2	0.2	0.2	0.2
mean	61.3	23.6	15.2	**11.2**	67.9	11.9	21.8	12.0	13.7	11.7	12.1	12.4	12.1
std	225.0	56.3	51.1	**37.8**	251.8	39.7	68.9	40.4	39.0	39.0	40.0	41.6	41.1

Chronos-2 achieves the lowest or jointly lowest error on most time series, but it does not dominate every dataset. Other models perform best on individual BWDF series, Mauth, and SOWEKI 10, demonstrating that the aggregate ranking does not translate into a uniform advantage. The com-

paratively high wMAPE values for low-volume series such as Mauth and SOWEKI 08 show how normalized errors can differ from original-unit errors.

C.2 Statistical Test Results

Table 11 reports the signed Diebold–Mariano test statistic for each model relative to the Seasonal Naïve baseline on every dataset. Positive values favor the respective model, whereas negative values favor Seasonal Naïve. Larger absolute values provide stronger evidence against the null hypothesis of equal predictive accuracy. They should not be interpreted as the absolute size of the error difference, which is reported in Tables 9 and 10.

Table 11: Signed Diebold–Mariano Test Statistics (Diebold & Mariano, 1995) for Each Model (Columns) Relative to the Seasonal Naïve Benchmark Across Datasets (Rows). Positive values favor the model and negative values favor Seasonal Naïve. Larger absolute values indicate stronger evidence against equal predictive accuracy.

Dataset	Chronos-2	Kairos	MOIRAI 2.0	MOMENT-1	t0-alpha	Time-MoE	Timer-S1	TimesFM 2.5	TiRex-2	Toto 2.0
BWDF 01	27.4	-69.9	26.2	-6.0	24.8	27.3	27.8	26.1	27.7	31.6
BWDF 02	44.9	-101.7	45.2	-8.7	41.3	36.4	44.0	44.7	44.0	40.5
BWDF 03	40.6	-109.9	41.5	-25.5	39.0	40.0	43.1	40.4	38.3	37.6
BWDF 04	42.4	-122.1	38.1	-15.7	32.7	35.4	41.9	37.8	38.8	34.0
BWDF 05	24.1	-167.4	17.9	-75.3	18.8	18.4	20.5	15.1	10.8	17.7
BWDF 06	46.2	-54.5	45.5	21.6	43.1	37.2	46.7	44.2	45.4	41.9
BWDF 07	36.5	-135.4	31.0	-43.5	30.0	27.5	35.4	29.5	28.6	31.4
BWDF 08	36.4	-154.2	34.2	-57.6	28.4	28.9	33.5	31.4	30.0	35.8
BWDF 09	50.5	-67.0	47.6	-27.1	47.0	37.6	51.1	47.2	50.6	49.7
BWDF 10	35.8	-106.8	39.4	-49.7	30.8	31.1	40.4	35.4	39.6	37.6
Mauth	6.3	4.6	5.8	~~-2.1~~	5.8	4.7	6.8	6.4	6.0	6.0
ReCycle	69.1	-256.7	59.1	-51.2	55.1	59.9	65.2	58.8	44.7	48.3
SOWEKI 01	67.4	-182.5	56.8	-34.9	61.4	-58.2	60.3	54.6	59.6	34.6
SOWEKI 02	64.9	-235.5	46.5	-50.5	52.8	-73.7	56.9	44.0	49.2	38.9
SOWEKI 03	44.1	-97.0	40.3	-5.0	40.8	-15.9	41.2	39.6	42.8	37.9
SOWEKI 04	69.9	-148.4	62.5	-31.4	64.9	-42.0	62.7	60.1	63.7	46.6
SOWEKI 05	31.4	-163.1	24.3	-104.4	27.6	-105.9	10.6	-7.2	3.5	18.6
SOWEKI 06	74.9	-281.1	45.6	-89.4	51.9	-112.2	63.8	46.1	48.4	48.4
SOWEKI 07	66.2	-158.8	46.7	-109.3	58.2	-65.7	59.4	49.8	51.1	53.7
SOWEKI 08	66.3	24.6	47.5	-164.6	63.6	41.6	62.3	61.6	61.4	57.5
SOWEKI 09	80.9	-221.3	67.5	-28.8	73.8	-77.4	73.1	68.7	74.9	54.3
SOWEKI 10	46.8	-89.8	38.1	-31.1	50.3	-83.1	41.7	30.2	36.3	51.3
SOWEKI 11	57.6	-171.3	51.1	4.3	58.3	-35.9	57.1	53.4	55.1	46.2
YW 002	54.2	-254.1	28.4	-108.4	37.1	-73.7	53.5	25.7	14.9	21.0
YW 017	24.7	-199.2	15.4	-69.3	19.6	-30.1	22.4	17.2	13.9	18.0
YW 108	70.7	-59.5	63.5	-72.7	66.7	12.5	69.2	64.7	66.8	60.8

C.3 Runtime Measurements

Figures 1 and 2 report inference times for the main zero-shot configuration and the evaluated context–horizon combinations. Total inference time reflects the cost of completing the respective evaluation, whereas inference time per forecast window supports a more direct comparison of computational cost across configurations and models. The horizontal lines in Figure 2 show the corresponding per-window runtimes of the statistical baselines.

Differences between models are generally larger than the runtime variation caused by changing the context or forecast horizon. MOIRAI 2.0 and t0-alpha have the lowest per-window inference times among the evaluated TSFMs, whereas Time-MoE and Timer-S1 require substantially more computation. These results show that similar forecasting accuracy among the leading models can be associated with markedly different inference costs.

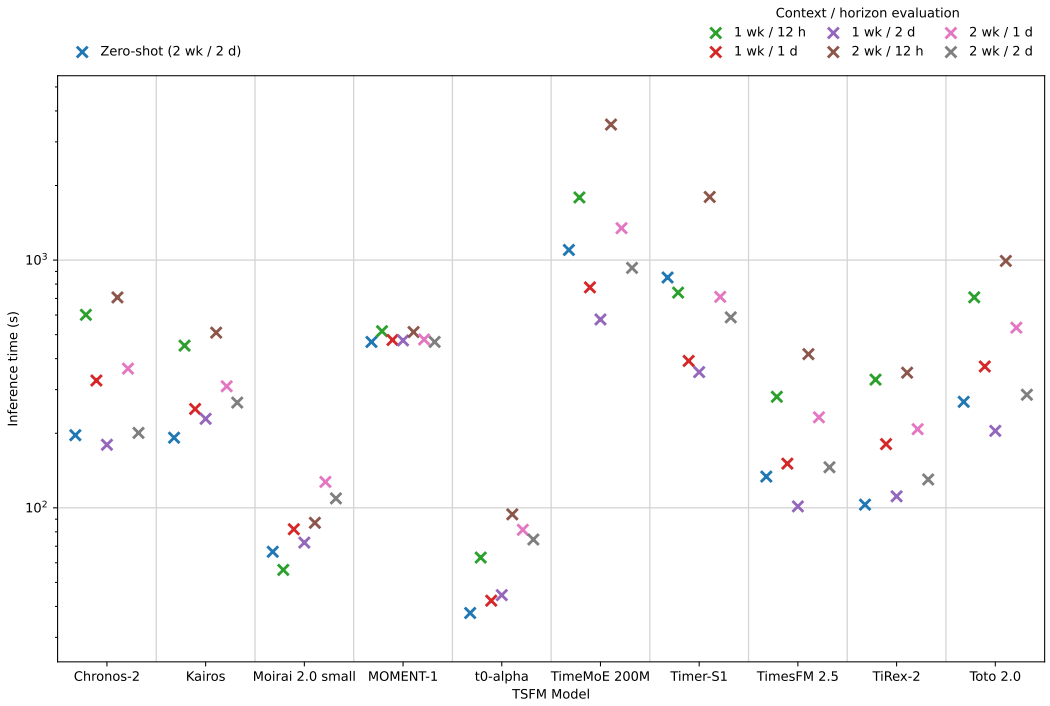

Figure 1: Total Inference Time, in Seconds, for Each Model Under the Main Zero-Shot Configuration and the Evaluated Context–Horizon Combinations. Lower values indicate faster inference.



Figure 2: Inference Time per Forecast Window, in Milliseconds, for Each Model Under the Main Zero-Shot Configuration and the Evaluated Context–Horizon Combinations. Horizontal lines show the corresponding runtimes of the statistical baselines. Lower values indicate faster inference.

D Context and Forecast-Horizon Sensitivity

Context length determines how much historical information is available to a model, whereas the forecast horizon determines how far into the future it must predict. To examine sensitivity to these parameters, we evaluate combinations of one, two, and four weeks of context with forecast horizons of 12 hours, one day, and two days. Kairos and TiRex-2 are excluded because their supported context lengths do not cover the four-week window required for the complete comparison. Figure 3 reports the mean wMAPE across datasets for every configuration and model included in this analysis.

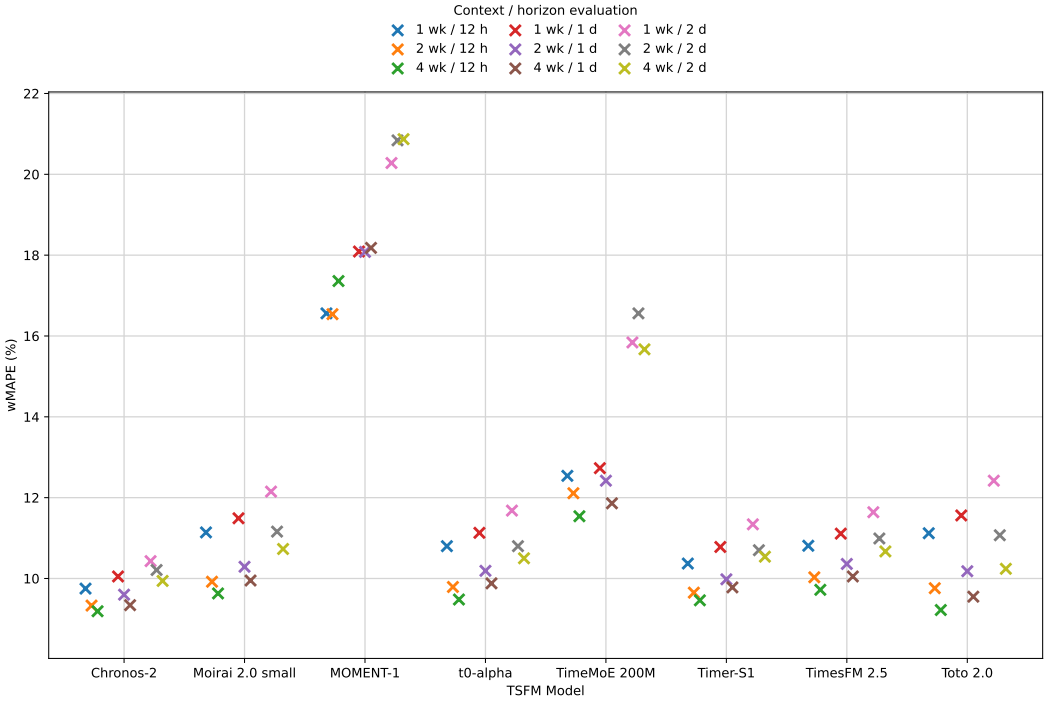

Figure 3: Mean wMAPE Across Datasets for Each Model, Context Length, and Forecast Horizon. Lower values indicate better forecasting performance.

For most models, shorter forecast horizons and longer context windows reduce forecasting error. The improvements from extending the context tend to diminish as more history becomes available. MOMENT-1 and Time-MoE deviate from this general pattern: their performance does not improve with longer context and, in several configurations, deteriorates.

To examine whether these changes in absolute error also affect the relative model ordering, we rank the models by their wMAPE on each dataset and average these ranks across datasets. Table 12 reports the resulting average rank for each context–horizon combination.

Table 12: Average Model Ranks Across Datasets Based on wMAPE for Each Context Length and Forecast Horizon. Lower ranks indicate better forecasting performance.

Context	**1 week**			**2 week**			**4 week**		
Horizon	**12h**	**1d**	**2d**	**12h**	**1d**	**2d**	**12h**	**1d**	**2d**
Chronos-2	1.7	1.4	1.5	1.5	1.2	1.4	1.8	1.5	1.6
MORIAI 2.0	5.2	5.2	5.3	4.7	4.5	4.4	4.9	4.8	4.6
MOMENT-1	8.0	8.0	7.7	8.0	8.0	7.7	8.0	8.0	7.7
t0-alpha	5.0	4.9	4.4	4.5	4.4	4.1	4.4	4.2	4.3
Time-MoE	5.2	5.5	5.9	5.9	6.0	6.8	6.0	6.3	6.8
Timer-S1	2.5	2.7	2.5	2.6	2.8	2.5	3.3	3.4	3.3
TimesFM 2.5	4.4	4.1	3.8	5.5	5.4	4.8	5.7	5.5	5.1
Toto 2.0	4.1	4.3	4.8	3.3	3.7	4.3	2.0	2.3	2.7

Chronos-2 maintains the best average rank across all nine configurations, while MOMENT-1 remains last. The intermediate ordering is more sensitive to the configuration: Toto 2.0 improves markedly with longer contexts, whereas the average ranks of Time-MoE and TimesFM 2.5 generally deteriorate. Thus, context and horizon choices affect both absolute forecasting accuracy and the relative performance of several models, although the strongest and weakest positions remain stable.

E Covariate-Aware Forecasting

Figure 4 complements the forecasting results in the main text by comparing the average inference time per forecast window for the four covariate configurations. The figure reports the configurations without covariates, with calendar covariates, with weather covariates, and with both covariate sets on a logarithmic scale.

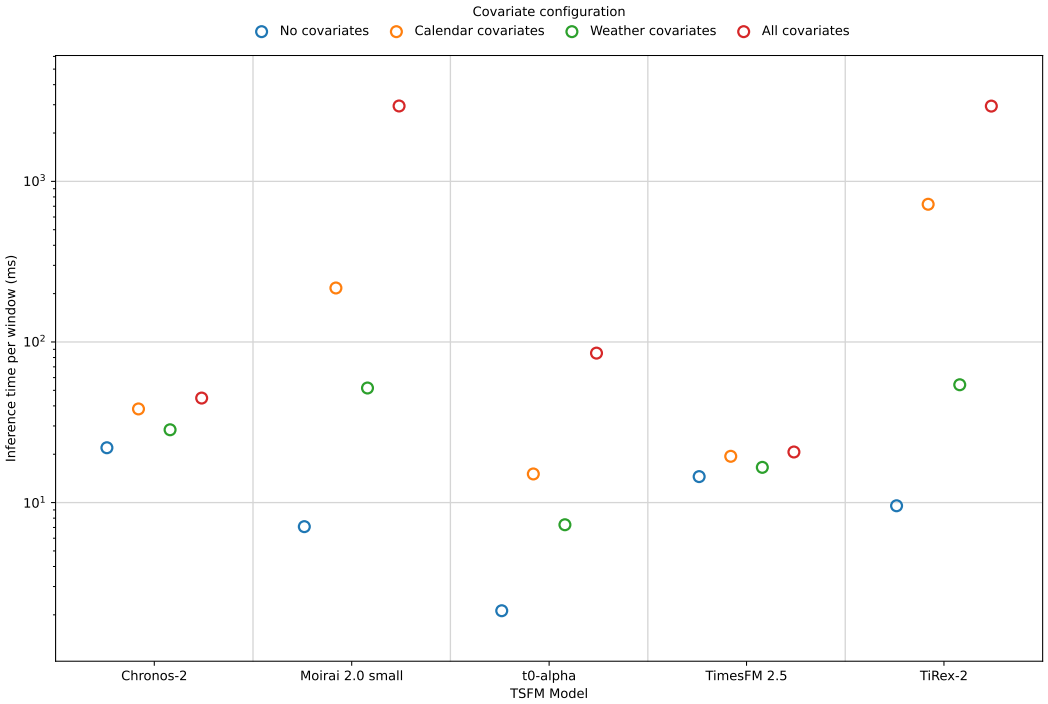

Figure 4: Inference Time per Forecast Window, in Milliseconds, for Each Model and Covariate Configuration on a Logarithmic Scale. Lower values indicate faster inference.

Adding covariates increases inference time for all five models, but the magnitude of this increase differs substantially. The additional cost is comparatively moderate for Chronos-2 and TimesFM 2.5, whereas the calendar and combined configurations increase the runtime of MOIRAI 2.0 and TiRex-2 by orders of magnitude. Weather covariates alone generally incur a smaller increase than calendar covariates. Thus, in the evaluated setting, the absence of consistent accuracy improvements from covariates is accompanied by additional and, for some models, substantial computational cost.

F Fine-Tuning Results

Table 13 reports the dataset-level MAE and wMAPE values for zero-shot Chronos-2 and its globally and locally fine-tuned configurations. The global configuration uses one adapter trained jointly across all datasets, whereas the local configuration uses a separate adapter for each dataset. Because MAE is expressed in the original unit and scale of each time series, its values should be compared only within a dataset.

Both fine-tuning strategies reduce wMAPE for most datasets, with particularly large improvements for several BWDF and SOWEKI series. Local fine-tuning achieves a slightly lower median error than

Table 13: Dataset-Level MAE and wMAPE Results for Zero-Shot Chronos-2 and Its Global and Local Fine-Tuned Configurations. Lower values indicate better forecasting performance.

	Global Fine-Tuning		Local Fine-Tuning		Zero-Shot	
Dataset	**MAE**	**wMAPE**	**MAE**	**wMAPE**	**MAE**	**wMAPE**
BWDF 01	0.56	9.03	0.54	8.69	3.77	12.64
BWDF 02	0.20	2.31	0.20	2.33	1.54	4.46
BWDF 03	0.14	5.24	0.14	5.18	1.23	8.19
BWDF 04	1.76	6.11	1.74	6.03	8.26	6.97
BWDF 05	1.21	1.64	1.18	1.60	7.67	2.71
BWDF 06	0.62	6.62	0.61	6.54	2.85	9.42
BWDF 07	0.71	2.99	0.70	2.96	4.12	4.50
BWDF 08	0.82	3.78	0.82	3.74	3.69	4.85
BWDF 09	1.02	4.66	1.00	4.57	4.22	5.57
BWDF 10	0.95	4.41	0.95	4.44	4.92	5.21
Mauth	0.04	57.21	0.04	57.95	0.05	69.57
ReCycle	49.61	7.82	48.56	7.65	199.10	8.59
SOWEKI 01	0.73	11.90	0.72	11.73	0.87	14.80
SOWEKI 02	0.94	12.24	0.93	12.05	1.30	14.35
SOWEKI 03	0.81	17.81	0.83	18.13	0.82	20.80
SOWEKI 04	0.66	14.20	0.66	14.23	0.86	16.68
SOWEKI 05	6.03	11.04	5.95	10.89	11.72	20.67
SOWEKI 06	1.39	7.44	1.38	7.40	1.87	9.73
SOWEKI 07	1.52	11.36	1.47	10.98	1.78	13.49
SOWEKI 08	0.39	65.97	0.38	65.08	0.36	61.81
SOWEKI 09	0.54	12.38	0.54	12.27	0.55	12.31
SOWEKI 10	2.53	21.19	2.31	19.35	5.19	40.43
SOWEKI 11	0.76	9.54	0.76	9.50	0.91	10.69
YW 002	1.58	5.82	1.54	5.69	2.03	7.50
YW 017	6.10	10.70	6.10	10.70	4.22	7.46
YW 108	0.16	8.15	0.16	8.17	0.20	11.67

global fine-tuning, but it does not outperform the global configuration on every dataset. The exceptions to the overall improvement are also visible: both strategies increase wMAPE for SOWEKI 08 and YW 017, while global fine-tuning produces a small increase for SOWEKI 09. These results show that the aggregate gains from fine-tuning are broadly distributed but not universal.