Evaluation of Foundation Models for Water Demand Forecasting in a Zero-Shot Scenario

Luana Rathunde Sandler[1], **Pablo Andretta Jaskowiak** [1]

[1]Interdisciplinary Data and Artificial Intelligence Laboratory (IDA Lab)
Centro Tecnológico de Joinville (CTJ) – Universidade Federal de Santa Catarina (UFSC)
Joinville – Santa Catarina – Brasil

luana.sandler@grad.ufsc.br, pablo.andretta@ufsc.br

Abstract. *Deploying Time Series Foundation Models (TSFMs) in domain-specific applications without retraining remains an open challenge, particularly in systems characterized by high seasonal volatility. In this work we evaluate Chronos and Moirai for zero-shot daily water demand forecasting using data from a coastal city in Southern Brazil, where seasonal tourism drives significant population fluctuations. Results are compared against statistical and machine learning baselines from the literature. Our findings support that Transformer-based FMs achieve competitive performance, highlighting their viability for water demand forecasting in data-scarce or rapid-deployment scenarios.*

1. Introduction

Effective water demand management is essential to urban sustainability, particularly in cities subject to intense seasonal variations or population fluctuations driven by factors such as tourism and temporary migration. Such factors impose significant challenges on the operation and planning of supply systems, demanding strategies that ensure reliability and efficiency. In this context, water demand forecasting moves beyond a simple operational tool, consolidating itself as a strategic management instrument capable of anticipating consumption peaks and reducing losses in the distribution system [Fang et al. 2024].

Estimating water consumption involves statistical and computational methods designed to forecast future water demand based on historical time series and explanatory variables, which can be utilized either in isolation or in combination [Adamowski et al. 2012]. Traditionally, Autoregressive Integrated Moving Average (ARIMA) models and Recurrent Neural Networks (RNN) have been employed, including their Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants [Sit et al. 2020]. However, these methods are generally limited in their ability to generalize and transfer learned patterns across different domains, which motivated the appearance of scalable approaches for temporal analysis [Rasul et al. 2024].

In this context, Foundation Models (FMs) represent a paradigm shift in machine learning, characterized by models trained on massive, heterogeneous datasets across multiple domains. These are generally adaptable to a wide range of downstream tasks, often through zero-shot inference or minimal fine-tuning [Bommasani et al. 2022]. The development of FMs is enabled by the growing availability of data, computational advances, and, most notably, the Transformer architecture [Vaswani et al. 2017]. Examples include BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020], and DALL-E [Ramesh et al. 2021], which showcase the versatility of Transformers across different

domains such as Natural Language Processing (NLP) and Computer Vision (CV). By incorporating broad, cross-domain knowledge and vast amounts of data during pre-training, these models can perform new tasks they have never been explicitly trained on [Brown et al. 2020, Devlin et al. 2019, Ramesh et al. 2021].

Following the success of this paradigm in other domains, Time Series Foundation Models (TSFMs) have emerged as a natural extension of the FMs paradigm to the temporal domain, aiming to generalize across diverse forecasting contexts [Liang et al. 2024]. However, time series data present unique challenges that set it apart from NLP and CV. Unlike text or images, time series often exhibit domain-specific structures, such as seasonality, trends, irregular sampling, and high cross-domain variability, which can lead to statistical distribution shifts between training and target domains [Karaouli et al. 2025].

Motivated by these developments, this study investigates whether Transformer-based TSFMs [Vaswani et al. 2017] maintain their performance under zero-shot inference on water demand data from a coastal city characterized by intense tourism-driven seasonality and population variability. This scenario allowed us to assess the ability of these models to generalize the statistical properties learned during training to real-world, domain-specific contexts without any retraining or fine-tuning [Karaouli et al. 2025].

The remainder of the paper is organized as follows. In Section 2 related work considering machine learning for water demand forecasting is presented. FMs for time series forecasting are discussed on Section 3. Experimental setup is introduced in Section 4, whereas results and discussion of the experiments are carried on Section 5. Finally, Section 6 draws the main conclusions of the paper and discuss future work directions, based on our current findings.

2. Water Demand Forecasting

Water demand forecasting has been extensively studied due to its critical role in urban planning and resource management. Reviews show that statistical methods, machine learning, and deep learning approaches have been explored to improve prediction performance [de Souza Groppo et al. 2019]. Critical assessments highlight that model selection and data characteristics significantly impact forecasting quality. However, no single approach consistently outperforms others across different scenarios [Niknam et al. 2022].

Several works incorporate exogenous variables, such as climate and socio-economic factors, to predict water demand across different sectors. Models such as Support Vector Regression (SVR) and Random Forest (RF) have been compared in this context [Shu et al. 2024]. Deep learning approaches based on recurrent architectures, such as GRU and LSTM, have also been proposed to improve short-term forecasting quality, particularly by capturing nonlinear patterns and extreme variations [Salloom et al. 2025]. Despite advances, most existing approaches rely on supervised learning pipelines, requiring models to be trained on the target dataset. Such a dependency limits their applicability in data-scarce scenarios with a small number of historical observations or when rapid deployment across multiple similar, yet distinct locations, is required.

To overcome the aforementioned limitations, recent works have adopted the Transformer architecture or incorporated it within new hybrid methods for water demand forecasting. In the context of water demand forecasting for irrigation, Transformers were

modified and integrated with well established fuzzy logic and genetic algorithms, utilizing an adapted positional encoding to capture seasonality and mitigate the memory leakage, which is common in sequential networks [González Perea et al. 2024].

Recent works have also extended zero-shot forecasting to domain-specific applications involving water-related systems. In hydrology, an evaluation of TSFMs for streamflow prediction across 531 basins of the CAMELS dataset demonstrated that models such as Chronos and Moirai can achieve competitive zero-shot performance across diverse hydroclimatic regimes, without any domain-specific training [Sun and Sun 2026]. Although no TSFM consistently surpassed the supervised LSTM baseline, the best-performing model approached comparable generalization metrics, highlighting both the potential and the remaining gaps of zero-shot inference under distribution shifts [Sun and Sun 2026]. These findings motivate the present study, which also evaluates Chronos and Moirai, but in the distinct context of urban water demand forecasting under strong seasonal changes.

Multimodal FMs have been also proposed for zero-shot water stress forecasting in precision agriculture. A framework that combines the Segment Anything Model (SAM), for visual plant phenotyping, with the Lag-Llama foundation model, for forecasting water stress levels from minute-level IoT sensor data, has been recently introduced [Nabaei et al. 2025]. In this setting, the zero-shot approach outperformed fine-tuned variants and also achieved lower error than supervised CNN and LSTM baselines, suggesting that, in high-frequency multimodal scenarios, zero-shot inference can surpass task-specific training. These findings indicate that FMs can be effectively deployed in complex environmental systems without retraining on domain-specific data, offering a computationally efficient alternative to traditional supervised pipelines.

In time series, distribution shifts refer to changes in the underlying statistical properties of data over time or across domains, including variations in seasonality and consumption dynamics. Recent works emphasize that such shifts can significantly degrade model stability and generalization, particularly in zero-shot forecasting [Ye et al. 2024, Li et al. 2025]. Although prior work highlights the limitations of zero-shot inference under these conditions [Karaouli et al. 2025], our analysis explicitly leverages the water demand context to evaluate this scenario, assessing whether FMs can maintain robust performance despite the high volatility inherent to the dataset.

3. Foundation Models for Time Series

The Transformer architecture for natural language processing tasks, represented a milestone by replacing recurrence and convolution mechanisms with the self-attention mechanism [Vaswani et al. 2017]. Self-attention captures long-range and complex contextual dependencies between elements of a sequence, regardless of their position. Furthermore, it connects all positions with a constant number of operations. This overcomes the sequential computation bottlenecks observed in recurrent models, enabling significantly greater parallelization and reduced training time [Vaswani et al. 2017]. This architecture has transcended the language domain, successfully expanding to temporal tasks such as forecasting, classification, and anomaly detection [Wen et al. 2023].

Time series exhibit distinctive features compared to other data modalities, such as text and images. Their sequential nature gives rise to temporal dependencies, in which past values influence future observations across short and long-term hori-

zons [Ye et al. 2024]. In multivariate settings, interactions among different variables reveal inter-variable dependencies that capture complementary relationships and latent patterns not observable in univariate scenarios. Furthermore, characteristics such as distribution shift, irregular sampling, and semantic diversity, where similar subsequences may carry different meanings depending on context, make learning and, consequently, generalizing particularly challenging under such conditions [Ye et al. 2024].

Therefore, there is a clear need for approaches capable of representing and capturing the structural variability inherent in time series data. In response, TSFMs have been proposed to analyze this diverse set of temporal patterns through large-scale pretraining on massive heterogeneous data collections, leveraging the scalability and parallel processing efficiency of the Transformer architecture [Ye et al. 2024]. Pretraining enables models to develop representations from diverse data sources that can be effectively transferred across inference scenarios [Li et al. 2025].

Three inference regimes are typically considered for time series forecasting: zero-shot, few-shot, and full-shot [Li et al. 2025, Brown et al. 2020]. In the zero-shot regime, the model is applied directly to an unseen target series without any parameter adaptation, thereby assessing its crude generalization capability. The few-shot regime entails fine-tuning the model on a small subset of the available target series (local training data), reflecting low-resource scenarios where only a limited number of observations are accessible. The full-shot regime leverages all available target series (local training and validation data) for fine-tuning, serving as an empirical upper bound on achievable performance.

The survey presented by Li et al. [Li et al. 2025] together with The Large Language Models for Time Series (LLM4TS) repository [Liao 2025], provide both a systematic overview of (TSFMs) and a mapping of their current availability. These works consolidate recent advances and organize existing methodologies into two complementary paradigms: (i) foundation models trained from scratch on large-scale time series corpora and, (ii) reprogrammed large language models that adapt pre-trained language models to time series forecasting tasks via fine-tuning [Bommasani et al. 2022, Ye et al. 2024].

Despite the rapidly growing number of models in both of the aforementioned paradigms, as we deal only with zero-shot forecasts in this work, preference was given to FMs trained from scratch on time series data. This choice ensures that the inductive bias of the models are aligned with the underlying temporal dynamics observed in numerical sequences from time series. Such alignment is critical in a zero-shot regime, as there is no parameter adaptation to the target time series, thus relying solely on the ability of the models to represent complex temporal phenomena and generalize to unseen scenarios. In this setting, Chronos [Ansari et al. 2024, Ansari et al. 2025] and Moirai [Woo et al. 2024, Liu et al. 2026] stand out, as we discuss in the following.

Chronos [Ansari et al. 2024, Ansari et al. 2025] constitutes a foundation model trained on a large-scale heterogeneous corpus of time series, including electricity, traffic, weather, finance, industrial sensors, and supply chain data. The model operates on regularly sampled time series at multiple levels of temporal granularity, ranging from sub-hourly to yearly frequencies [Ansari et al. 2024]. The Chronos architecture integrates temporal tokenization, which discretizes observations along with their associated temporal attributes, multi-scale positional encodings, and an autoregressive decoder for multi-

horizon forecasting. This design enables its application to both univariate (endogenous) and multivariate (exogenous) settings [Ansari et al. 2024]. The more recently introduced Chronos-2 [Ansari et al. 2025] extends the original Chronos framework by incorporating advancements in tokenization and model scaling, further improving multi-horizon forecasting performance and robustness across domains.

Moirai [Woo et al. 2024, Liu et al. 2026] is a general-purpose Time Series Foundation Model designed for probabilistic forecasting across diverse domains. It was trained on the Large-scale Open Time Series Archive (LOTSA) and employs a Transformer-based architecture tailored for time series, leveraging patch-based representations that encode segments of the input sequence rather than individual time steps, thereby improving computational efficiency while capturing local temporal patterns. Moirai 2.0 [Liu et al. 2026] extends this design with architectural and training refinements, including improved scaling strategies and enhanced modeling of heterogeneous temporal dynamics, resulting in stronger generalization across datasets with varying statistical properties.

Given their distinct characteristics, we adopted the most recent and improved iterations of these architectures, namely Chronos-2 [Ansari et al. 2025] and Moirai 2.0 [Liu et al. 2026]. For brevity we will refer to these models simply as Chronos and Moirai throughout the remainder of this work.

4. Experimental Setup

The evaluation scenario considered in this work consists of a real-world water demand forecasting dataset from the city of Guaratuba, Paraná, Brazil. Due to its coastal location, the city experiences a heavy influx of tourists from major nearby cities, such as Curitiba and Joinville, during summer and holiday seasons. This dynamic introduces high volatility and severe seasonal spikes into the time series, creating a challenging forecasting environment that is ideal for testing the robustness of zero-shot FMs.

The dataset is described in detail by Vieira Filho et al. [Vieira Filho et al. 2024]. In brief, it contains the daily volume of potable water produced at the local treatment station, which is highly correlated with actual water consumption. In addition to the production volumes, climate and calendar data were incorporated by the authors. The time series spans a continuous four-year period, from 2016 to 2019. This particular dataset has been previously explored in the context of time series forecasting by other authors, who considered both statistical [Stefaniak et al. 2024] and machine learning models in their works [Vieira Filho et al. 2024]. To ensure direct comparability, we adopt the same evaluation scenario previously established by Vieira Filho et al. [Vieira Filho et al. 2024], which is made available by the author in a public repository[1].

The evaluation setup established by these two benchmark studies is the same, which consists of generating rolling forecasts throughout the entire year of 2019. Since these prior works utilized supervised statistical and machine learning methods, the time series was temporally partitioned: data from 2016 through 2018 were allocated for model training and validation (with hyperparameter tuning), while the entirety of 2019 was reserved as the hold-out test set. Two windowing strategies were considered by the authors in both studies: sliding window (SW) and expanding window (EW).

[1] https://github.com/jesuinovieira/bachelor-thesis

In the sliding window approach, a fixed-length context is used, and at each iteration, the input window is shifted forward by seven days, incorporating recent observations while discarding older ones. This setup is suitable for non-stationary scenarios, as it reduces the influence of outdated patterns [Brownlee 2020]. In contrast, the expanding window strategy progressively increases the historical context by incorporating all past observations prior to each forecast, allowing the evaluation of performance gains from larger training contexts [Bergmeir and Benítez 2012].

Following the evaluation protocol established by previous works, we generate rolling forecasts for the year 2019 to guarantee direct comparability of our results [Vieira Filho et al. 2024, Stefaniak et al. 2024]. However, a fundamental distinction in our methodology is the complete bypass of local model training. By discarding the local training phase entirely, we evaluate the models in a rigorous zero-shot, cold-start scenario. Historical data from 2016–2018 is used exclusively as input context for the FMs, without any parameter adaptation. For the 2019 inference phase, we apply the same sliding and expanding window strategies solely to construct the input context for the FMs.

In both evaluation strategies (SW and EW) short term forecasts are obtained, specifically, seven day ahead forecasts are generated at each step. These are concatenated to form continuous forecast series that, ultimately, allows for comparison to the original series. Performance is assessed using standard forecasting metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), providing complementary perspectives on prediction performance and robustness of the methods under evaluation [Hyndman and Athanasopoulos 2021].

The implementation is made publicly available to ensure reproducibility at: `https://github.com/lrsandler/inferencia-em-series-temporais`. The evaluated models, namely, Chronos-2 [Ansari et al. 2025] and Moirai 2.0 [Liu et al. 2026], are applied using their official pipelines, as suggested by their authors. The repository also provides detailed information on the software environment, including the versions of the libraries and dependencies used in the experiments.

5. Results and Discussion

To evaluate the efficacy of zero-shot inference for urban water demand, the FMs deployed in this work were assessed over the year of 2019 from the data previously described. Their predictive performance is evaluated against that of ARIMA and MLP, which previous studies have reported as the best-performing supervised models for this dataset [Stefaniak et al. 2024, Vieira Filho et al. 2024]. Because we adhered to the established experimental setup and evaluation metrics, the results of our zero-shot models can be directly and fairly compared to the locally trained baselines from prior works. The comparative results are summarized in Table 1.

It is important to note that the standard Chronos architecture operates primarily as a univariate foundation model, relying exclusively on the tokenized historical target series to generate predictions. Urban water demand is, however, intrinsically linked to external factors. To account for this, we explicitly evaluate a multivariate configuration, denoted as Chronos with covariates (Chronos+Cov). The covariates, obtained from the same dataset used by Vieira Filho et al. [Vieira Filho et al. 2024], include meteorological variables, namely mean temperature, mean radiation, mean relative humidity and mean

precipitation. By integrating these exogenous climate and calendar variables into the inference process, this variant allows us to investigate whether supplementing a pre-trained zero-shot model with localized, domain-specific signals can further mitigate the impact of the severe seasonal volatility observed in this dataset.

Table 1 is organized into three sections according to the time series windowing strategies: expanding window (EW), sliding window with one-year context (SW-1Y), and sliding window with two-year context (SW-2Y). Performance metrics were computed over the rolling forecasting iterations and averaged to estimate overall model performance. The standard deviation is reported for each metric, allowing the assessment of prediction stability. The best result for each window and metric is highlighted in bold.

Table 1. Evaluation metrics for the 2019 test year, considering all models.

Model	RMSE $\pm \sigma$	MAE $\pm \sigma$	MAPE $\pm \sigma$
Expanding Window			
MLP†	1709.70 ± 593.10	1408.97 ± 515.16	11.81 ± 4.70
ARIMA‡	1771.39 ± 1044.06	1523.65 ± 928.79	11.99 ± 6.38
Moirai	1703.25 ± 650.03	1408.27 ± 552.22	11.34 ± 4.14
Chronos	1558.69 ± 659.70	1301.25 ± 560.91	10.58 ± 4.34
Chronos+cov	**1500.36 ± 568.72**	**1248.59 ± 474.34**	**10.23 ± 4.06**
Sliding Window – 1 Year Training/Context			
MLP†	2142.58 ± 927.99	1796.94 ± 855.90	14.11 ± 5.84
ARIMA‡	**1745.08 ± 1165.27**	**1499.44 ± 999.01**	**11.52 ± 6.21**
Moirai	1886.14 ± 942.89	1577.69 ± 803.46	12.44 ± 4.71
Chronos	1922.94 ± 966.72	1626.37 ± 826.08	12.68 ± 4.88
Chronos+cov	1860.74 ± 888.14	1576.91 ± 768.90	12.26 ± 4.45
Sliding Window – 2 Years Training/Context			
MLP†	1927.41 ± 732.00	1620.15 ± 630.53	13.08 ± 4.91
ARIMA‡	1652.67 ± 1078.71	1419.40 ± 975.04	11.15 ± 6.65
Moirai	1717.60 ± 678.51	1422.91 ± 573.28	11.43 ± 4.16
Chronos	1558.62 ± 665.27	1293.36 ± 562.47	10.49 ± 4.33
Chronos+cov	**1509.01 ± 568.91**	**1251.87 ± 466.09**	**10.27 ± 4.03**

† Results from [Vieira Filho et al. 2024].
‡ Results from [Stefaniak et al. 2024].

Overall, the results show that Transformer-based models achieve competitive performance compared to traditional approaches. To place the results in perspective, the test set contains daily water production values ranging from approximately 5,000 to 26,000m³, with peak consumption occurring during the summer, which sees a high influx of tourists. In the expanding window (EW) configuration, Chronos+cov achieved the best performance across all metrics, with an RMSE of 1500.36m³ and MAE of 1248.59m³, outperforming both the MLP baseline (RMSE of 1709.70m³ and MAE of 1408.97m³) and ARIMA (RMSE of 1771.39m³ and MAE of 1523.65m³).

In the sliding window with one-year context (SW-1Y), ARIMA achieved the lowest RMSE of 1745.08m³, suggesting that statistical methods remain strong baselines under limited historical context. It is important to note in this case, however, that ARIMA

has the largest standard deviation across all models. Additionally, Chronos and Moirai not only show a lower standard deviation but also outperform MLP, with results close to ARIMA, the best performer. For the two-year sliding window (SW-2Y), Chronos+cov again obtained the best RMSE of 1509.01m³ benefiting from the larger input context. Although the inclusion of covariates leads to consistent improvements in some scenarios, the overall gains are relatively modest across the evaluated metrics. This suggests that, in the zero-shot setting, the models may not fully exploit domain-specific auxiliary information, possibly due to the mismatch between pretraining distributions and the target data.

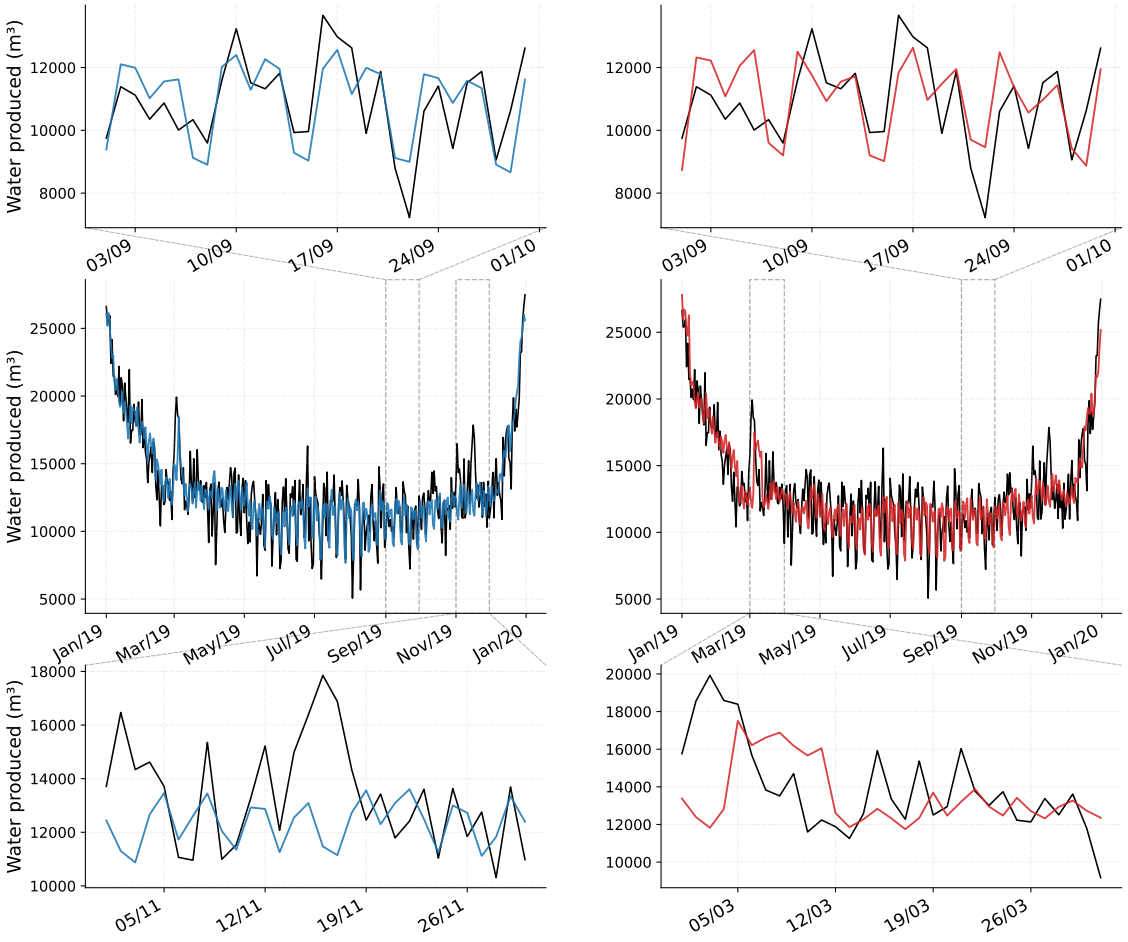
Moirai shows stable performance across all settings, although it does not achieve the best results, indicating a more conservative behavior in the zero-shot scenario. Despite not achieving best performance, in all cases it outperformed MLP, which was fine-tuned for this application, as discussed by the authors [Vieira Filho et al. 2024]. These results are particularly promising given that Moirai operates in a zero-shot regime, thereby avoiding the training and hyperparameter tuning costs associated with MLP adoption. The FMs exhibited standard deviations generally comparable to those of MLP, and consistently lower than ARIMA across all configurations, suggesting that zero-shot inference does not introduce greater prediction instability than supervised approaches [Karaouli et al. 2025].

In addition to the aggregated metrics presented in Table 1, a more detailed analysis of the temporal behavior of the models is provided in Figures 1 and 2. These figures allow for a detailed evaluation of model performance across different periods of the year, highlighting variations that are not captured by global error metrics alone.

Figure 1 presents a comparison between observed and predicted values for selected periods, including months with best and worst predictive performance. It can be observed that both models are generally able to capture the overall trend of water demand. The higher errors observed in March and November can be partly explained by seasonal events and tourism-related demand fluctuations. In particular, Carnival, a major Brazilian festive period associated with increased tourism and population mobility, occurred in February in 2017 and 2018 but shifted to March in 2019, introducing a non-fixed seasonal pattern that was difficult for the zero-shot models to anticipate. This shift may have contributed to the high error in March, as the historical context did not contain an equivalent demand peak in that month. In November, the increased error may instead reflect the gradual rise in tourism and water consumption preceding the summer.

Meanwhile, September and October exhibit the lowest prediction errors, coinciding with periods of lower seasonal variability and stable consumption patterns. This behavior is consistent with known challenges in time series forecasting, where non-stationarity and seasonal effects introduce complexities [Ye et al. 2024].

Figure 2 further analyzes this behavior by presenting the distribution of absolute errors for each month using boxplots. The results reveal that prediction errors are not uniformly distributed throughout the year. Instead, certain months exhibit higher median errors and larger dispersion, indicating periods of reduced model reliability. In particular, months with higher variability show wider interquartile ranges and more extreme outliers, suggesting that both models are more sensitive to changes in data distribution during these intervals. This observation aligns with the concept of distribution shift, in which the statistical properties of the data vary over time, negatively impacting model stability,



(a) Chronos (b) Moirai

Figure 1. Forecast performance for whole year, best and worst months (zoomed).

especially in zero-shot settings [Karaouli et al. 2025, Ye et al. 2024].

Moreover, the variability observed in the error distributions is consistent with the standard deviations reported in Table 1, where TSFMs present prediction variability slightly lower than (or comparable) to MLP, and consistently below ARIMA. Overall, our results highlight that evaluating forecasting models solely based on aggregated metrics may obscure important temporal dynamics. A detailed analysis across different periods reveals that model performance is influenced by seasonal patterns and domain-specific characteristics, which remain challenging for TSFMs in zero-shot.

6. Conclusion

This study evaluated TSFMs under a zero-shot, cold-start protocol for urban water demand forecasting in a highly seasonal coastal city. The results demonstrate that TSFMs achieve competitive performance, particularly under the expanding window and two-year sliding window strategies, where Chronos with covariates consistently outperformed the baselines across the evaluated metrics in the highest-context configurations (EW and SW-2Y). These results have to be analyzed from two intrinsically linked perspectives.

On one hand, the margin of improvement over supervised baselines such as

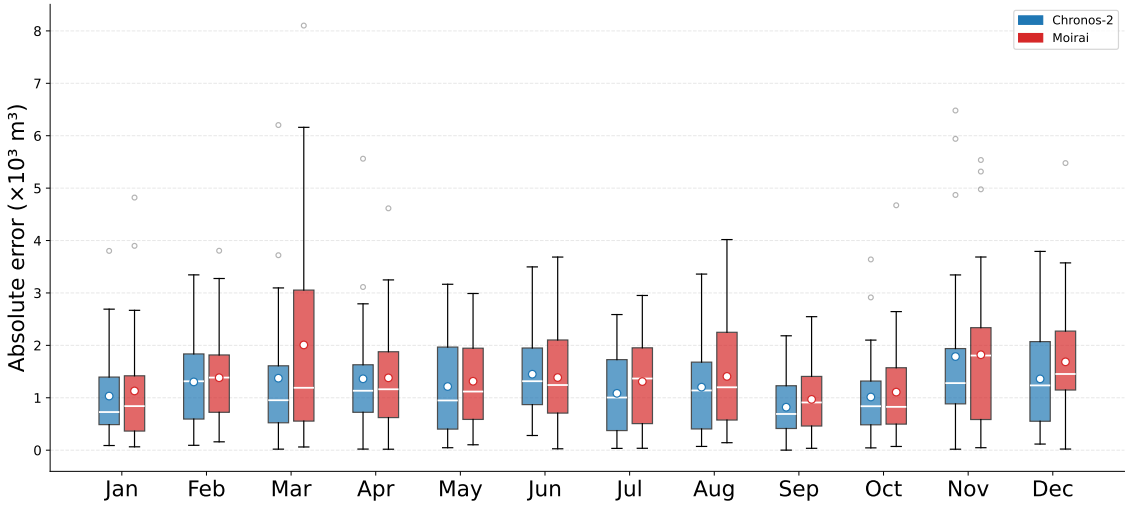


Figure 2. Monthly distribution of absolute forecast errors.

ARIMA and MLP is modest, and in several configurations, the differences fall within one standard deviation, suggesting that the zero-shot paradigm has not yet reached a level of superiority that would justify replacing well-tuned supervised methods in operational settings.

On the other hand, TSFMs achieved these results without fine-tuning or training, unlike ARIMA and MLP, which required an extensive hyperparameter search. Regarding computational efficiency, the cost comparison between TSFMs and supervised models is uneven, since FMs concentrate their computational cost in a single pretraining phase conducted by the model developer, delivering to the final user an already-trained model whose operational cost reduces to inference. In contrast, ARIMA and MLP require their parameters to be estimated or re-tuned at each forecasting window, so that fitting and inference occur jointly at every iteration. Thus, although it would be misleading to claim that TSFMs demand only a fraction of the computational cost, given the high pretraining cost, it can be argued that, in the deployment context, the marginal cost per forecast is low, as it does not require local tuning/fitting at each iteration.

These results indicate a promising direction for the application of FMs in scenarios where the observed error range is operationally acceptable and where local training data is scarce or unavailable. For contexts demanding higher accuracy, particularly during peak demand periods driven by seasonal tourism, few-shot fine-tuning represents a promising direction for closing the remaining performance gap, as it would allow the models to adapt their pretrained representations to the specific distributional characteristics of the target domain without requiring full retraining.

References

Adamowski, J. et al. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada. *Water Resources Research*, 48(1):W01528.

Ansari, A. F. et al. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*.

Ansari, A. F. et al. (2025). Chronos-2: Scaling time series foundation models. *arXiv preprint arXiv:2510.15821*.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bommasani, R. et al. (2022). On the opportunities and risks of foundation models.

Brown, T. et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Brownlee, J. (2020). *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery, 1.9 edition.

de Souza Groppo, G., Costa, M. A., and Libânio, M. (2019). Predicting water demand: a review of the methods employed and future possibilities. *Water Supply*, 19(8):2179–2198.

Devlin, J. et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, Minneapolis, Minnesota.

Fang, X., Liu, J., Zhou, M., Zhang, H., and Zhao, J. (2024). Review of the mechanism and methodology of water demand forecasting in the socio-economic system. *Water*, 16(11).

González Perea, R., Camacho Poyato, E., and Rodríguez Díaz, J. (2024). Attention is all water need: Multistep time series irrigation water demand forecasting in irrigation disctrics. *Computers and Electronics in Agriculture*, 218:108723.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Melbourne, 3 edition.

Karaouli, N., Coquenet, D., Fromont, E., Mermillod, M., and Reyboz, M. (2025). How foundational are foundation models for time series forecasting? In *NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT2S)*, San Diego, CA, USA.

Li, Z. et al. (2025). TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 2, pages 5595–5606, Toronto, Canada.

Liang, Y. et al. (2024). Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565, Barcelona, Spain.

Liao, Y. (2025). LLM4TS: Large language & foundation models for time series. GitHub repository.

Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Savarese, S., Sahoo, D., Xiong, C., and Li, J. (2026). Moirai 2.0: When less is more for time series forecasting.

Nabaei, S. H., Zheng, Z., Chen, D., and Heydarian, A. (2025). Multimodal foundation models for zero-shot water stress forecasting in precision agriculture. In *2025 International Conference on Machine Learning and Applications (ICMLA)*, pages 1330–1337.

Niknam, A., Zare, H. K., Hosseininasab, H., Mostafaeipour, A., and Herrera, M. (2022). A critical review of short-term water demand forecasting tools—what method should i use? *Sustainability*, 14(9).

Ramesh, A. et al. (2021). Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8821–8831.

Rasul, K. et al. (2024). Lag-Llama: Towards foundation models for probabilistic time series forecasting. In *R0-FoMo: Robustness of Few-Shot and Zero-Shot Learning in Foundation Models, NeurIPS Workshop*.

Salloom, T., Kaynak, O., and He, W. (2025). A novel deep neural network architecture for real-time water demand forecasting. *arXiv preprint arXiv:2512.06714*.

Shu, J., Xia, X., Han, S., He, Z., Pan, K., and Liu, B. (2024). Long-term water demand forecasting using artificial intelligence models in the tuojiang river basin, china. *PLOS ONE*, 19(5):e0302558.

Sit, M. et al. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, 82(12):2635–2670.

Stefaniak, A. K., Jaskowiak, P. A., and Weihmann, L. (2024). A case study on water demand forecasting in a coastal tourist city. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 3–17, Belém, Brazil.

Sun, A. Y. and Sun, A. A. (2026). Zero-shot forecasting of streamflow using time series foundation models: are we there yet? *Machine Learning: Earth*, 2(1):010501.

Vaswani, A. et al. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, volume 30, Long Beach, CA, USA.

Vieira Filho, J., Scortegagna, A., Vieira, A. P. S. D., and Jaskowiak, P. A. (2024). Machine learning for water demand forecasting: Case study in a brazilian coastal city. *Water Practice and Technology*, 19(5):1586–1602.

Wen, Q. et al. (2023). Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, number 759, pages 6778–6786.

Woo, J., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. (2024). Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*.

Ye, J. et al. (2024). Empowering time series analysis with foundation models: A comprehensive survey.