ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables

Arango, Sebastian Pineda; Mercado, Pedro; Kapoor, Shubham; Ansari, Abdul Fatir; Stella, Lorenzo; Shen, Huibin; Senetaire, Hugo; Turkmen, Caner; Shchur, Oleksandr; Maddix, Danielle C.

Total number of authors:
13

Published in:
Proceedings of 28th International Conference on Artificial Intelligence and Statistics

Publication date:
2025

Document Version
Early version, also known as pre-print

ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables

Sebastian Pineda Arango[*2†], **Pedro Mercado**[*1], **Shubham Kapoor**[1], **Abdul Fatir Ansari**[1], **Lorenzo Stella**[1], **Huibin Shen**[1], **Hugo Senetaire**[3†], **Caner Turkmen**[1], **Oleksandr Shchur**[1], **Danielle C. Maddix**[1], **Michael Bohlke-Schneider**[1], **Yuyang Wang**[1], **Syama Rangapuram**[1].
[1]Amazon Web Services [2]University of Freiburg [3]Technical University of Denmark

Abstract

Covariates provide valuable information on external factors that influence time series and are critical in many real-world time series forecasting tasks. For example, in retail, covariates may indicate promotions or peak dates such as holiday seasons that heavily influence demand forecasts. Recent advances in pretraining large language model architectures for time series forecasting have led to highly accurate forecasters. However, the majority of these models do not readily use covariates as they are often specific to a certain task or domain. This paper introduces a new method to incorporate covariates into pretrained time series forecasting models. Our proposed approach incorporates covariate information into pretrained forecasting models through modular blocks that inject past and future covariate information, without necessarily modifying the pretrained model in consideration. In order to evaluate our approach, we introduce a benchmark composed of 32 different synthetic datasets with varying dynamics to evaluate the effectivity of forecasting models with covariates. Extensive evaluations on both synthetic and real datasets show that our approach effectively incorporates covariate information into pretrained models, outperforming existing baselines.

1 Introduction

One important component for time series forecasting –regardless of the forecasting model class– is exogenous variables, also called covariates. Covariates provide external information to the forecasting model and allow the forecasting model to adjust the forecast based on this additional information. For example, the predicted energy output of a wind farm can be informed by additional weather information like the forecasted wind speed. While it is straightforward to incorporate covariate information into local statistical models and global deep learning (or other machine learning) models, it is less clear how this can be achieved for pretrained forecasting models. Pretrained forecasting models are trained across different datasets from different domains and every dataset might have a different number of variables with different properties, making it difficult to pretrain these models with covariates. Consequently, the majority of pretrained time series models do not support covariate data (Dooley et al., 2023; Goswami et al., 2024; Das et al., 2024; Ansari et al., 2024). One notable exception is Moirai, which introduces one possible way to address this problem with a mechanism called "any-variate attention" that can pretrain on datasets with a varying number of covariates (Woo et al., 2024). Nevertheless, how to include covariates into pretrained models that do not natively support covariates remains an open question.

In this paper, we present an approach called ChronosX to include covariate data into the pretrained forecasting model Chronos (Ansari et al., 2024). Chronos is a pretrained model that is trained on a time series dataset corpus without covariates. Drawing inspiration from modular deep learning (Pfeiffer et al., 2023), our proposed approach consists of attaching two modules. The first module updates the pretrained token embeddings with past covariates and the second module uses future covariates to adjust the output distribution. These light-weight adapter modules can

[*]Equal contribution. [†]Work done during internship at AWS. Correspondence: Pedro Mercado pedroml@amazon.de. Code available at https://github.com/amazon-science/chronos-forecasting/tree/chronosx.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

be quickly trained for a downstream forecasting task even when the underlying pretrained model is frozen. Moreover, we show that the same framework can be extended to other pretrained models to incorporate covariates and showcase this with state-of-the-art pretrained models such as TimesFM (Das et al., 2024) and MOMENT (Goswami et al., 2024).

To validate our approach, we introduce a novel collection of 32 datasets that emulates different kinds of covariates together with several time dynamics. We also demonstrate that ChronosX achieves low forecasting error across 18 real-world datasets with covariates. Furthermore, we show that the adapter approach can be used to finetune a pretrained time series model on a target dataset, which is much faster than full finetuning of the entire pretrained model. In summary, we make the following contributions: 1) We introduce ChronosX, an adapter approach for incorporating covariates into pretrained models for time series forecasting, 2) we show that the same framework can be extended to other pretrained models to ingest covariates and introduce TimesFMX and MOMENTX, 3) we introduce a novel benchmark of 32 synthetic datasets that allows us to evaluate how effective are forecasting models in leveraging covariate information under different dynamics, 4) we demonstrate that ChronosX achieves low forecasting error both for synthetic and real-world datasets with covariates.

2 Background and Related Work

Time series forecasting is the task of extrapolating a time series into the future based on its historical values and, optionally, a set of covariates. These covariates are external variables that potentially influence the primary time series and may enhance the forecasting model's accuracy by providing additional context (e.g., weather may influence the sales of fans). In its general form, time series forecasting can be cast as the problem of modeling the conditional distribution,

$$P(\mathbf{z}_{C+1:H}|\mathbf{z}_{1:C}, \mathbf{X}_{1:H}; \Phi), \tag{1}$$

where $\mathbf{z}_{1:C} = [z_1, \ldots, z_C]$ is the historical context of the primary time series, $\mathbf{z}_{C+1:H} = [z_{C+1}, \ldots, z_H]$ is the future target until horizon H, $\mathbf{X}_{1:H} = [\mathbf{x}_1, \ldots, \mathbf{x}_H]$, are covariates from the historical context until horizon H, and Φ denotes a set of learnable parameters. In this paper, we focus on the case where the primary time series is univariate, $z_t \in \mathbb{R}$, while the covariates may have multiple dimensions, $\mathbf{x}_t \in \mathbb{R}^c$.

A wide variety of time series forecasting models have been proposed in the literature. Based on how they leverage time series data for training, they may be categorized into: *local* and *global* models. Local models such as ARIMA (Box et al., 2015) and ETS (Hyndman et al., 2008) are fit individually for each time series. On the other hand, global models such as DeepAR (Salinas et al., 2020), TFT (Lim et al., 2021), PatchTST (Nie et al., 2023), and NHiTS (Challu et al., 2023) are trained across multiple time series from a given dataset, and are able to leverage global patterns present within a dataset.

Pretrained time series models. Recent work (Ansari et al., 2024; Woo et al., 2024; Das et al., 2024; Goswami et al., 2024) on large-scale training of time series models has given rise to a category of global models called *pretrained models* (also referred to as "foundation" models). Pretrained models are trained on a large corpus of time series data and can be used for accurate zero-shot forecasting or finetuned on downstream datasets and tasks. Moirai (Woo et al., 2024), MOMENT (Goswami et al., 2024), and TimesFM (Das et al., 2024) are pretrained models based on patching (Nie et al., 2023). Specifically, they convert time series into patches before processing these patches with transformer-based models. Further, TTM (Ekambaram et al., 2024) is a pretrained model built on top of model blocks of TSMixer (Chen et al., 2023), which is an architecture designed by stacking multiple MLPs that allow to extract information efficiently from both time and feature dimensions. On the other hand, Chronos (Ansari et al., 2024) maps time series values into tokens from a fixed vocabulary via scaling and quantization, and trains existing language models architectures on such *time series tokens*. Ansari et al. (2024) show that this simple tokenization scheme is very effective and Chronos models achieve zero-shot performance comparable to deep learning models with access to training data.

Forecasting with covariates. Covariates are essential to account for external events in forecasting models, such as peak dates in retail, planned strikes that might disrupt traffic, or sports events that generate massive online attendance. Forecasting models incorporate covariates in different ways. For instance, in DeepAR (Salinas et al., 2020) covariates are part of the input to the RNN block, whereas TFT (Lim et al., 2021) explicitly encodes past and future covariates with different encoding networks whose outputs are concatenated. ARIMA uses extra coefficients to account for covariates, and NBEATSx (Olivares et al., 2023) and N-HiTS (Challu et al., 2023) concatenate the primary time series with past and future covariates and compute a fixed input size. Moreover, Ekambaram et al. (2024) introduce TTM-CM, a fine-tuning strategy for TTM based on channel mixing that allows to integrate covariates. To the best of our knowledge, the only pretrained model that natively consumes covariates is Moirai (Woo

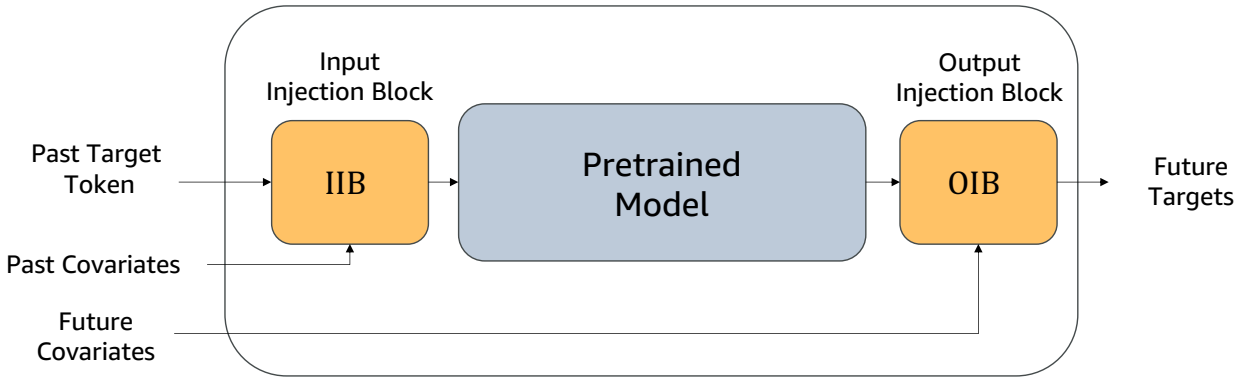


Figure 1: Architecture of ChronosX. It consists of two blocks: the first block adds covariates of the past and updates token embeddings. The second block adds future covariates updating the estimated distribution of the pretrained model.

et al., 2024), which flattens time series and covariates into a single sequence and assigns a variate ID to distinguish between target time series and covariates. In this paper, we introduce ChronosX, an adapter approach inspired by modular deep learning (Pfeiffer et al., 2023), for incorporating covariates into the univariate Chronos (Ansari et al., 2024) model which does not support covariates natively.

3 ChronosX: Covariate Integration in Pretrained Models

We now introduce the proposed model to add covariates to pretrained models. We denote by **ChronosX** the resulting model of adding covariates to the pretrained model Chronos (Ansari et al., 2024). See Figure 1 for an overview of the model architecture.

Chronos is a pretrained model for univariate probabilistic time series forecasting based on the encoder-decoder T5 transformer models. Chronos applies a preprocessing scheme where time series are first mean-scaled and then tokenized through a suitable bin-quantization approach. The resulting tokenization is then passed to generate the corresponding pretrained embedding and fed into the encoder. Chronos uses the categorical distribution over the vocabulary of tokens as output distribution, and is trained to minimize the cross entropy between the distribution of the quantized ground truth label and the predicted distribution. For more details, we refer the reader to (Ansari et al., 2024).

Our novel approach ChronosX adds the information of covariates in two different stages: 1) we add covariates from the past to update the corresponding token embeddings, and 2) we add covariates of the future to adjust the logits. Observe that, depending on the user context, one can choose to either work with covariates of the past, the future, or both. Following a modular deep learning paradigm (Pfeiffer et al., 2023), we propose modules to achieve this while minimizing the modifications to the original model.

ChronosX allows us to freeze any arbitrary portion of the original architecture and just update the proposed modules. In the following sections, we refer as Input Injection Block (IIB) and Output Injection Block (OIB) to the adapters injecting covariates from the past and the future, respectively. See Figure 2 for a depiction of these. Additionally, we will overload the notation and use $\mathbf{z}$ for tokenized time series. In what follows we will use fully connected feed-forward networks consisting of two linear transformations with a ReLU activation in between, $\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$.

Input Injection Block. To add information of covariates from the past we update the token embeddings. Specifically, for every time step $\mathbf{z}_{t-1}$ we pass the corresponding token embeddings and covariates through independent linear layers and apply an FFN to the ReLU of the concatenated mappings. Let $h_{\text{emb}}(\mathbf{z}_{t-1}) \in \mathbb{R}^{d_{\text{model}}}$ be the pretrained token embedding, the updated embeddings f_{IIB} are defined as

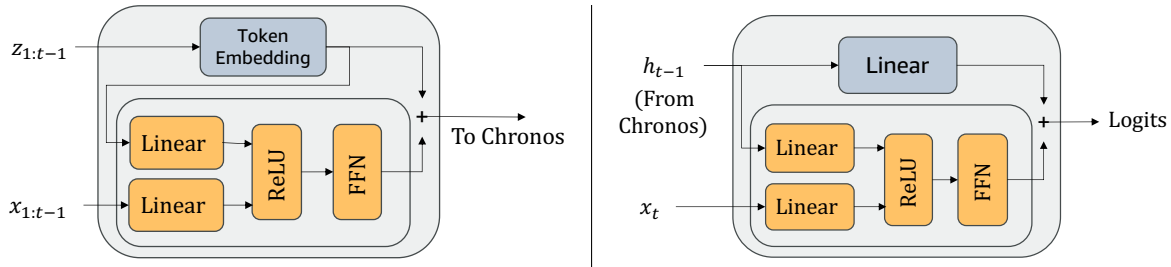
$$f_{\text{IIB}}\left(\mathbf{z}_{t-1}, \mathbf{x}_{t-1}\right) = h_{\text{emb}}(\mathbf{z}_{t-1}) + g_{\text{IIB}}(h_{\text{emb}}(\mathbf{z}_{t-1}), \mathbf{x}_{t-1}) \tag{2}$$

and g_{IIB} is defined as

$$g_{\text{IIB}}(h_{\text{emb}}(\mathbf{z}_{t-1}), \mathbf{x}_{t-1}) = \\ \mathrm{FFN}\left(\mathrm{ReLU}\left(h_{\text{emb}}(\mathbf{z}_{t-1})W_{\text{IIB}}^{(\text{emb})} \oplus \mathbf{x}_{t-1}W_{\text{IIB}}^{(\text{cov})}\right)\right) \tag{3}$$

with linear mappings $W_{\text{IIB}}^{(\text{emb})}$ and $W_{\text{IIB}}^{(\text{cov})}$ and $\oplus$ represents the concatenation operator.

The function g_{IIB} adjusts the embeddings of the tokenized time series value $\mathbf{z}_{t-1}$ by merging information of covariates from the past $\mathbf{x}_{t-1}$ and the corresponding time series embedding $h_{\text{emb}}(\mathbf{z}_{t-1})$.



(a) Input Injection Block (IIB)

(b) Output Injection Block (OIB)

Figure 2: Input and Output injection blocks used in ChronosX. Observe that IIB and OIB take past and future covariates, respectively. Figure 2a shows that the input injection block takes a pretrained tokenized embedding together with covariates of the past, whereas Figure 2b takes the pretrained logits (expressed as the final hidden state multiplied by a pretrained matrix) together with covariates of the future. The blue color indicates that the module is taken from the pretrained model.

Output Injection Block. We leverage information from future covariates $\mathbf{x}_t$ by adjusting the logits of the pretrained model. These are generated through a matrix multiplication between the last hidden state and a pretrained matrix (Vaswani et al., 2017). The adjusted logits are defined as follows: let $h_{\text{out}}(\mathbf{z}_{t-1})$ be the final hidden state corresponding to $\mathbf{z}_{t-1}$, and W_{out} the pretrained matrix producing the logits in the model, then the adjusted logits f_{OIB} are defined as

$$f_{\text{OIB}}\left(\mathbf{z}_{t-1}, \mathbf{x}_t\right) = h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{out}} + g_{\text{OIB}}(h_{\text{out}}(\mathbf{z}_{t-1}), \mathbf{x}_t) \tag{4}$$

where g_{OIB}

$$g_{\text{OIB}}(h_{\text{out}}(\mathbf{z}_{t-1}), \mathbf{x}_t) = \\ \text{FFN}\left(\text{ReLU}\left(h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{OIB}}^{(\text{out})} \oplus \mathbf{x}_t W_{\text{OIB}}^{(\text{cov})}\right)\right) \tag{5}$$

with parameter matrices $W_{\text{OIB}}^{(\text{out})}$ and $W_{\text{OIB}}^{(\text{cov})}$, and $\oplus$ represents the concatenation operator. Observe that the inputs considered for this update are covariates from the future and the last hidden state evaluated on time series values from the past.

Modular variants of ChronosX. Due to its modular nature, we can adapt ChronosX depending on the kind of covariates that we want to incorporate. In what follows we will denote by ChronosX the case where both past and future covariates are incorporated, i.e. IIB and OIB are used. Yet, we can extend our model to the case where only covariates of the past are available by only considering the input injection block. Similarly, for the case where we only want to consume future covariates, we can consider only the output injection block. When training ChronosX, we train from scratch the parameters of the Feed-Forward Networks (FFN) and new matrices $W_{\text{IIB}}^{(\text{emb})}, W_{\text{IIB}}^{(\text{cov})}, W_{\text{OIB}}^{(\text{cov})}, W_{\text{OIB}}^{(\text{out})}$. Additionally, we consider a full-finetuned variant ChronosX(FF) where all the parameters, including the pretrained ones, are updated. Whereas in the main paper we will only report results for ChronosX, in the appendix we describe and present evaluations of additional variants. For instance, we study the case where either only past or future covariates are considered.

4 Extending Injection Blocks to Other Pretrained Models

In the previous section, we introduced ChronosX, an extension for Chronos to incorporate covariates. Given the modular and minimally-invasive nature of our design, our approach can be extended to other pretrained models. We will showcase this with two pretrained models: MOMENT and TimesFM. Contrary to Chronos which is an encoder-decoder transformer, MOMENT and TimesFM are encoder-only and decoder-only models, respectively. The main differences to consider with respect to Chronos are: *i)* input patching and *ii)* point forecasts, which imply minor modifications to our framework. The resulting extensions of pretrained models with covariates will be referred to as **TimesFMX** and **MOMENTX**.

Extending on Input Patching. Patching has received a relevant amount of attention in recent years, becoming a standard operation in recent forecasting models (Nie et al., 2023; Goswami et al., 2024; Das et al., 2024). Patching is the operation of breaking a time series into potentially disjoint fixed-length subsequences.

To extend our framework to models that rely on a patched input $\tilde{\mathbf{z}}_{t-1} \in \mathbb{R}^{P_n \times P_d}$, we also create patches for the covariates $\tilde{\mathbf{x}}_{t-1} = \text{Patch}(\mathbf{x}_{t-1}) \in \mathbb{R}^{P_n \times (P_d \cdot c)}$, where P_d, P_n, c denote the patch dimension, the number of patches and the number of covariates. In this manner, we guarantee that both the primary time series input patches and the covariate patches have the same length. Subsequently, we feed these patched inputs into the modules as in Equation 3.

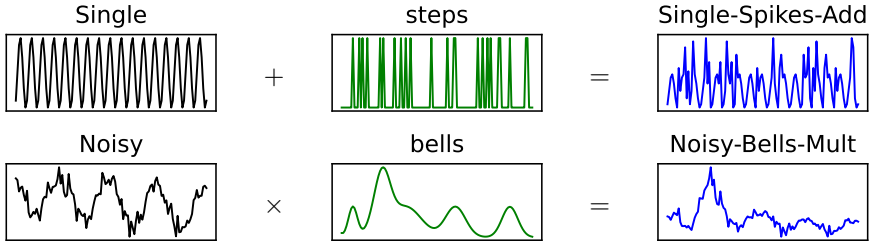


Figure 3: Illustration of the generation of synthetic time series. Different combinations of the 4 possible main signals modified by one of the 4 possible external covariates through a chosen operator ($+$ or $\times$), result in realistic synthetic time series with covariates.

Extending on Point Forecast. Our framework can be extended to the case of point forecasts as follows. Instead of considering a categorical distribution, we can take a given point forecast generated by the current model $\hat{\mathbf{z}}_t = h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{out}}$ and apply the following update rule:

$$f_{\text{OIB}}(h_{\text{out}}(\mathbf{z}_{t-1}), \mathbf{x}_t, \hat{\mathbf{z}}_t) = \hat{\mathbf{z}}_t + \text{FFN}\left(\text{ReLU}\left(h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{OIB}}^{(\text{out})} \oplus \mathbf{x}_t W_{\text{OIB}}^{(\text{cov})} \oplus \hat{\mathbf{z}}_t W_{\text{OIB}}^{(\text{p})}\right)\right) \tag{6}$$

5 Synthetic Datasets with Covariates

Although the task of forecasting time series with covariates is a highly relevant task, there is a limited amount of freely available time series datasets with covariates where forecasting models can be evaluated. To overcome this limitation in this paper we introduce a collection of 32 different synthetic datasets of time series with covariates carefully handcrafted to emulate real-life applications where external factors greatly impact the time series.

Each of the datasets contains 100 times series of daily frequency extending over 1827 days[1]. For each dataset, we consider 3 basic elements that allow for the generation of 100 times series with covariates: 1) the **main signal** ($\hat{z}_t$), 2) an **external covariate** (x_t), and 3) a **combination operator** $\odot \in \{+, \times\}$. The generation of time series with these three ingredients is defined by $z_t = \hat{z}_t \odot x_t$. In what follows we present several variations of these three basic elements with the goal of providing a comprehensive evaluation of suitable forecasting models for time series with covariates.

The time series representing the main signal $\hat{z}_t$ are divided in two types: **Simple** and **Complex**. The simple synthetic type includes single sinusoids and simple sinusoids variants, whereas the complex synthetic type includes diverse sinusoids and noisy sinusoids variants. We now describe each of the four different variants of the main signal $\hat{z}_t$ with increasing complexity with the form of Eq. (7), mainly: 1) *Single sinusoids* considers a standard sine curve with fixed zero phase, unit amplitude and period of 7 days corresponding to one week for the entire dataset; 2) *Simple sinusoids* is the sum of three sine curves each with a weekly, monthly and yearly period, together with a randomly sampled amplitude and zero phase; 3) *Diverse sinusoids* is the result of combining three sine curves with randomly sampled phases and amplitudes and with a global trend, and 4) *Noisy sinusoids* which results from adding gaussian noise (ϵ) to *diverse sinusoids* and allows us to assess model robustness. Please see the Appendix for more details on the process to generate the dataset.

$$\hat{z}_t = \sum_{i=1}^{3} a_i \sin\left(f_i t + \phi_i\right) + b_1 t + b_2 + \epsilon \tag{7}$$

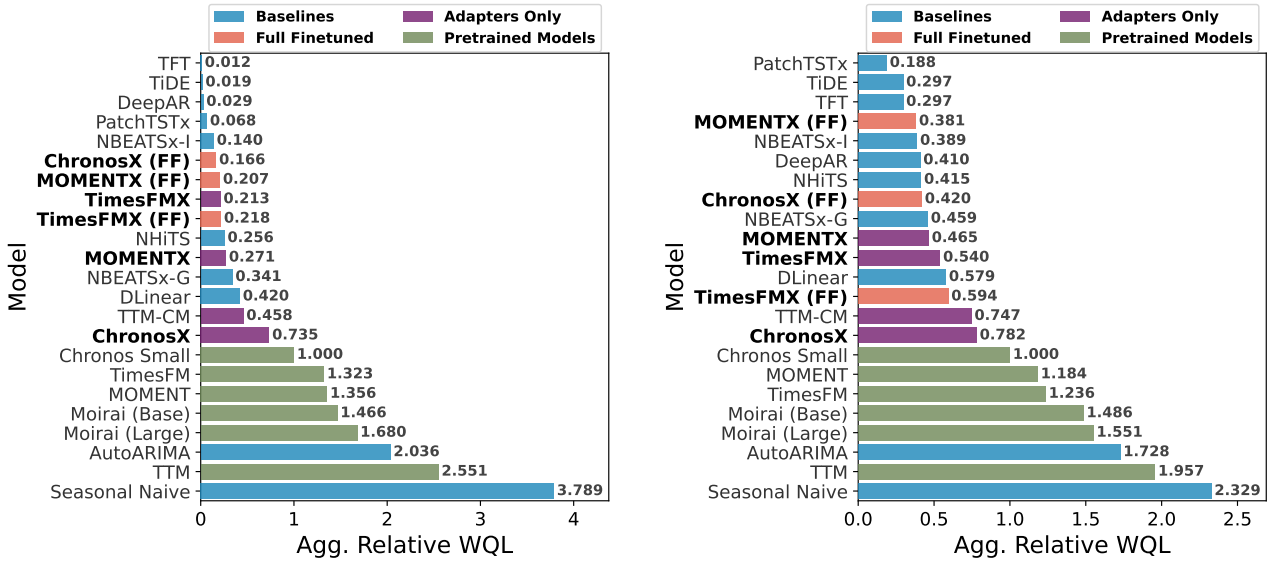
We design four different kinds of **external covariates**: 1) *spikes* models short time events such as strikes or power failure, 2) *steps* model longer events with sudden and sustained changes such as discounts in the retail industry, 3) *bells* represent smoother changes such as festive season, and 4) *autoregressive process* with randomly generated parameters that model cases where the value of covariates is determined by previous observations. Finally, the **combination operator** $\odot$ is defined either as the addition $+$ or the multiplication $\times$ between the main signal ($\hat{z}_t$) and the covariates (x_t). An illustration of the generative process for this synthetic datasets along with a few examples of the resulting time series is depicted in fig. 3.

6 Experiments and Results

Setup. In all experiments for our models we consider learning rates $\{10^{-2}, 10^{-3}, 10^{-4}\}$ and select the one with the best validation error with a validation window of length equal to the prediction length of the dataset in consideration. A detailed description of the hyperparameters is presented in the appendix.

Metrics. We consider the Weighted Quantile Loss (WQL) to evaluate probabilistic forecasts, which measure the alignment of quantile levels of the predictive distribution with respect to the ground truth (Gneiting and Raftery, 2007; Gasthaus et al., 2019; Shchur et al., 2023). In all experiments the WQL is computed on quantile levels $\{0.1, 0.2, \ldots, 0.9\}$. For methods generating sample forecasts we compute the quantiles based on 100 samples, whereas quantile forecasting methods are trained on the same quantile levels we use

[1]This corresponds to the number of days in the fictional time series between 01/01/2025 and 31/12/2029.



(a) Agg. Rel (WQL) on **Simple** synthetic datasets.

(b) Agg. Rel (WQL) on **Complex** synthetic datasets.

Figure 4: Evaluations on Simple and Complex datasets as introduced in Section 5. Scores are normalized by Chronos Small. We can see that our proposed models ChronosX and ChronosX(FF) effectively incorporate covariates.

for evaluation. We report the Mean Absolute Scaled Error (MASE), which measures deviations from the median forecast with the ground-truth values (Hyndman and Koehler, 2006). Finally, we follow (Fleming and Wallace, 1986; Ansari et al., 2024; Woo et al., 2024) and report the aggregated WQL and MASE computed as the geometric mean of normalized scores by the corresponding baseline.

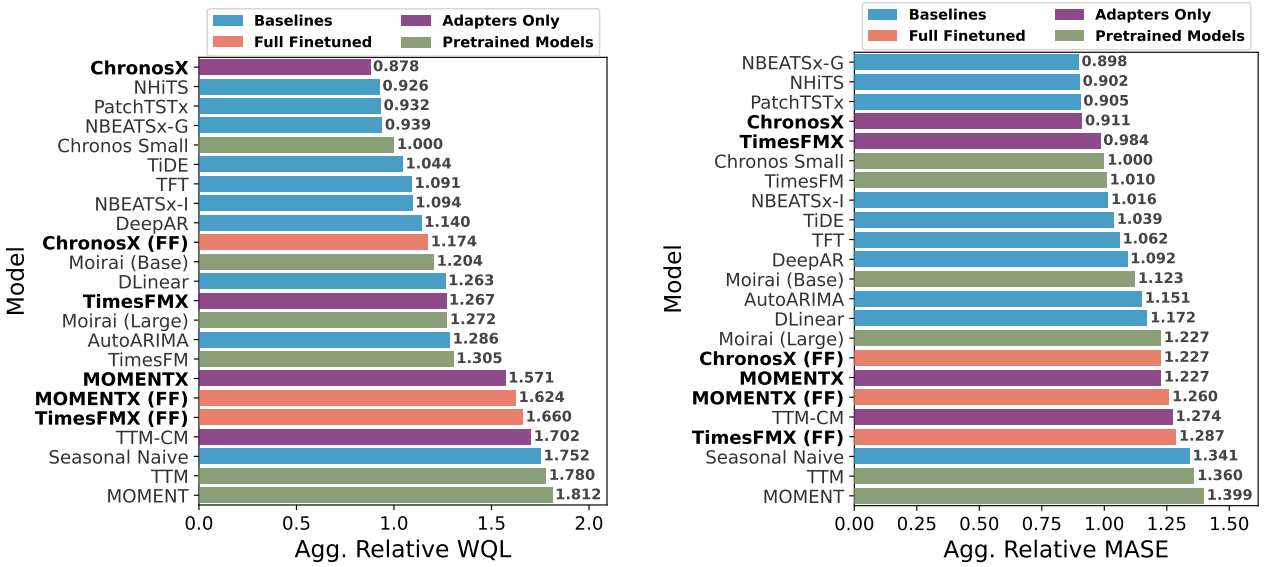
6.1 Finetuning and Adapters on Synthetic Benchmark Datasets with Covariates

In this section we evaluate how effective different forecasting models are in incorporating external information through covariates under different dynamics.

We take the 32 different datasets as introduced in Section 5 and follow the suggested splits Simple and Complex dataset collections each consisting of 16 datasets, and for each dataset we generate 100 time series of daily frequency with length 1827 and prediction length of 30. Recall that our adapter-based approach ChronosX, and its extensions MOMENTX and TimesFMX, incorporates both past and future covariates. Further, we consider a full finetuning variant per model and denote it by ChronosX(FF), MOMENTX(FF), TimesFMX(FF), which updates all parameters including those of the pretrained model. We also consider baselines that include covariates like DeepAR, PatchTSTx (an extension of PatchTST that supports covariates; see the supplementary appendix for details), NHiTS, and NBEATSx with its interpretable and general variants, together with AutoArima, which incorporates covariates through residuals. For all methods we standardize the covariates by the mean of absolute values. All scores are normalized by the scores of Chronos Small, which is a variant of Chronos with 46M parameters.

Figures 4a and 4b present the aggregated relative WQL on *Simple* and *Complex* synthetic datasets. For both types of datasets, our proposed adapter-only approaches ChronosX, TimesFMX, and MOMENTX, together with their full fine-tuning variants (ChronosX (FF), TimesFMX (FF), and MOMENTX (FF)) incorporate covariates, as they clearly outperform the zero-shot performance of their pretrained models, i.e. Chronos Small, TimesFM, and MOMENT, respectively. In particular, we can see that ChronosX outperforms Chronos by 22% in both WQL and MASE. Further, we observe that baseline models like TFT, DeepAR and PatchTSTx perform particularly well in *Simple* datasets, verifying that they effectively incorporate covariates in simple cases, whereas for *Complex* datasets the errors are larger, yet most of baselines outperform the zero-shot performance of Chronos Small. Moreover, we can see that in general full fine-tuning further improves the performance, which can be observed with ChronosX (FF) and MOMENTX (FF), whereas for TimesFMX (FF) in general it seems that there is a slight decrease in performance with complex synthetic datasets.

Overall, we can see that our proposed benchmark of synthetic datasets is useful to discern the effectivity of



(a) Agg. Rel. WQL on real datasets with covariates.

(b) Agg. Rel. MASE on real datasets with covariates.

Figure 5: Evaluations on real datasets with covariates. We report the Aggregated Relative WQL and MASE.

different models to incorporate covariates in time series. In particular, we have seen that most of the baselines here considered incorporate covariates that yield better scores than methods that do not take covariates, and that our covariate extensions for pretrained models are effectively incorporating covariate information

6.2 Finetuning and Adapters on Real Datasets with Covariates.

In this section, we evaluate our proposed models on 18 real datasets with covariates taken from diverse sources (Zhou et al., 2021; Makridakis et al., 2022; Godahewa et al., 2021; Woo et al., 2024; Wang et al., 2024). These datasets encompass data from fields such as retail, nature, transport, and mostly energy. Frequencies include hourly, daily, and 15-minute intervals, with forecast horizons up to 30 steps ahead. Please see the appendix for further details.

In Figure 5 we can see that our adapter-based approach ChronosX consistently performs best among pretrained models adapted to include covariates TimesFMX and MOMENTX. Moreover, we can see that adapter-only variants ChronosX, TimesFMX, and MOMENTX perform better than their zero-shot counterparts Chronos, TimesFM, and MOMENT, respectively, further verifying that our models incorporate covariate information. Additionally, we can see that full fine-tuning variants ChronosX (FF), TimesFMX (FF), and MOMENTX (FF) underperform with respect to several of the baselines here considered. This is likely due to the fact that several datasets consist of only one time series, and hence having a model with less trainable parameters provides an advantage in data-sparse regimes. Further, we can see that our model ChronosX performs best in WQL, and in terms of MASE it performs in the top-5 models together with TimesFMX. On the other side, despite the fact that MOMENT underperforms other models in general, we can see that our covariate extensions MOMENTX and MOMENTX (FF) significantly improve its performance, verifying that our adaption of pretrained models to incorporate covariates is effectively consuming this kind of information.

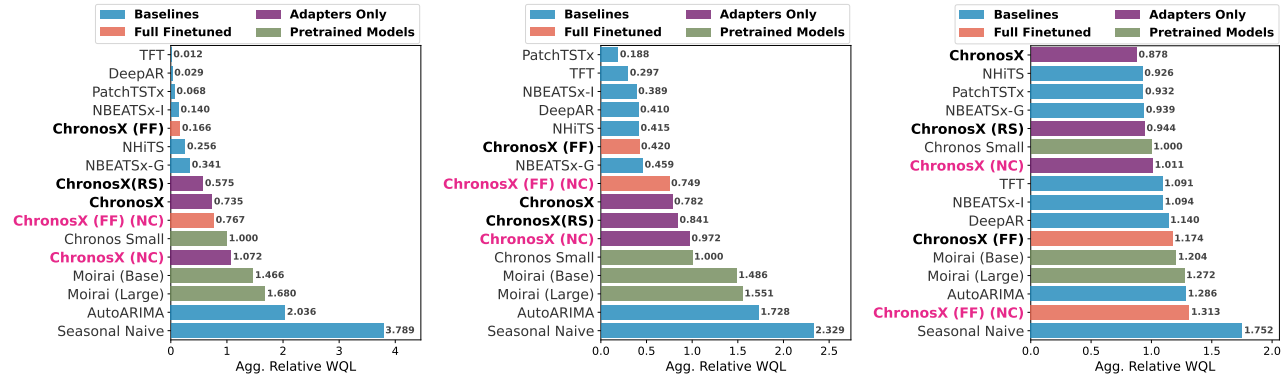
Overall we have shown that our proposed covariate extensions ChronosX, TimesFMX, and MOMENTX effectively incorporate covariate information in real datasets and ChronosX performs best in WQL and is comparable to other models in MASE.

6.3 Ablations of Covariate Adapters

In this section we explore the impact of covariates and backbone model sizes in ChronosX.

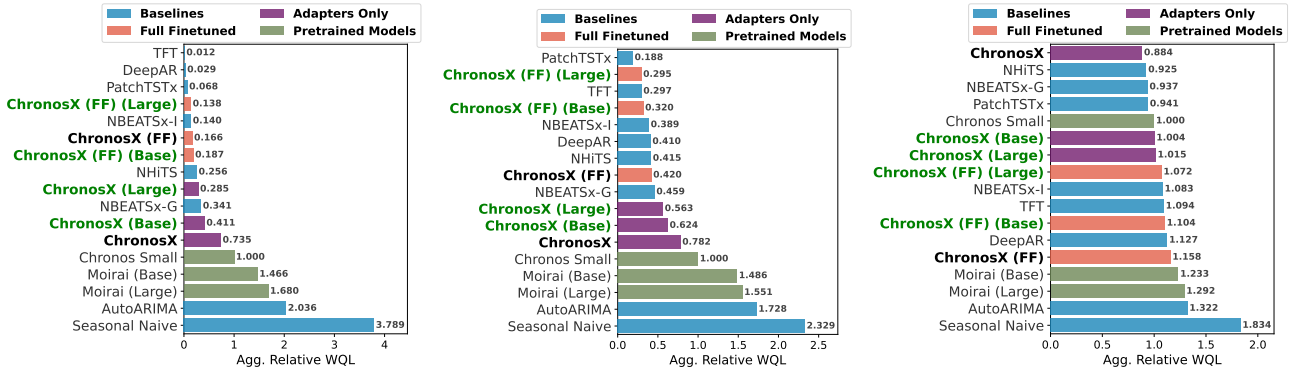
Impact of covariates in ChronosX. We analyze if the performance improvement of ChronosX is obtained from the covariate information or the extra parameters included by the covariate modules. We run experiments using the Input and Output blocks, but omitting the covariates by dropping the matrices related to covariates, inducing the following variations:

$$f_{\text{IIB-NC}}\left(\mathbf{z}_{t-1}, \mathbf{x}_{t-1}\right) = h_{\text{emb}}(\mathbf{z}_{t-1}) + \text{FFN}\left(\text{ReLU}\left(h_{\text{emb}}(\mathbf{z}_{t-1}) W_{\text{IIB}}^{(\text{emb})}\right)\right) \tag{8}$$



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets.

Figure 6: Covariate ablation study. ChronosX (NC) and ChronosX (FF) (NC) are variants with no-covariates of our models ChronosX and ChronosX (FF), respectively. Overall model ablations with no-covariates perform worse.



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets

Figure 7: Model size ablation. ChronosX (Base) and ChronosX (Large) are variants of ChronosX with backbone models with 200M and 710M parameters. Overall, larger backbone models improve performance with full fine-tuning.

$$f_{\text{OIB}-\text{NC}}\left(\mathbf{z}_{t-1}, \mathbf{x}_t\right) = h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{out}} + \text{FFN}\left(\text{ReLU}\left(h_{\text{out}}(\mathbf{z}_{t-1})W_{\text{OIB}}^{(\text{out})}\right)\right) \quad (9)$$

Evaluations of these ablations are presented in Fig. 6, where the ablations without covariates are denoted ChronosX (NC) and ChronosX (FF) (NC). We can see across synthetic and real datasets that our model ChronosX and its full-finetuned version ChronosX (FF) perform better than its corresponding ablations, ChronosX (NC) and ChronosX (FF) (NC), respectively. Thus, we verify the performance improvement observed by our proposed model ChronosX is due to effectively consuming the information encoded through covariates. Finally, we provide an adapter-only variant that consumes covariates of the future through a linear transformation, denoted by ChronosX (RS)

$$f_{\text{RS}}\left(\mathbf{z}_{1:t-1}, \mathbf{x}_t\right) = h_{\text{out}}(\mathbf{z}_{1:t-1})W_{\text{out}} + \mathbf{x}_t W_{\text{OIB}}^{(\text{cov})} \quad (10)$$

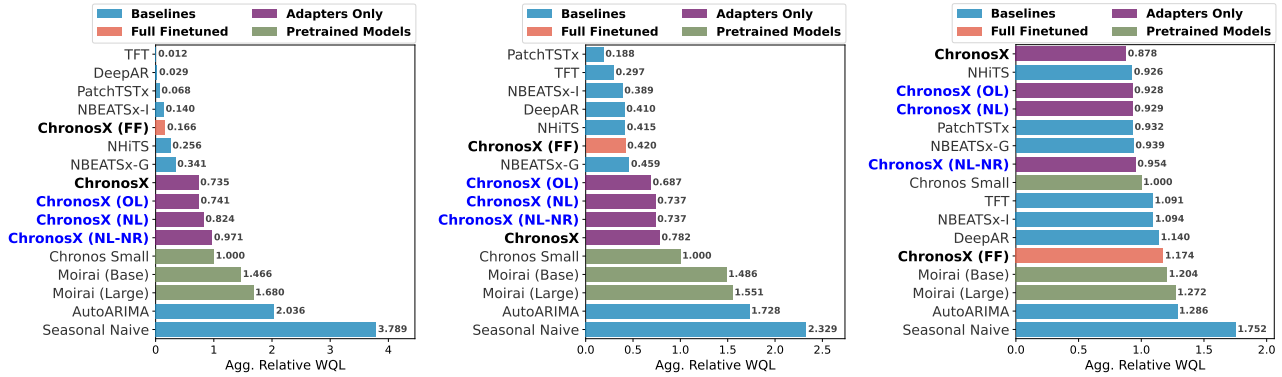
We observe that this residual-based model ChronosX (RS) systematically outperforms its adapter-only ablation ChronosX (NC), further verifying the relevance of covariate information to improve performance.

Impact of Backbone Size in ChronosX. In our experiments we have used for our model ChronosX the backbone model Chronos in its small version (46M). In this ablation, we study the effect of using the size variants *Base* and *Large* of Chronos, which have 200M and 710M parameters, respectively. The results in Fig. 7 demonstrate that increasing the model capacity is particularly helpful in complex synthetic datasets. For real datasets we can see that full fine-tuning approaches improve their performance with larger backbone models, i.e. ChronosX (FF)(Base) and ChronosX (FF)(Large), whereas adapter-only variants decrease in performance. Hence, we observe that using the backbone size *Small*, as done with our model ChronosX, provides a good balance between model size and forecast quality.

Impact of Adapters Architecture. Our proposed adapters consider two independent linear layers, followed by a Feed-Forward Network, as explained in Section 3. One linear layer operates on embeddings related to the target time series (token embedding and hidden state), while operates on the covariates. This design al-

Table 1: Definition of three alternative versions of ChronosX: with a single linear layer (OL), without linear layer but with activation (NL), and without linear layer and activation (NL-NR).

Ablation	**Input Injection Block** $g_{\text{IIB}}(h_{\text{emb}}(\mathbf{z}_{t-1}), \mathbf{x}_{t-1})$	**Output Injection Block** $g_{\text{OIB}}(h_{\text{out}}(\mathbf{z}_{t-1}), \mathbf{x}_t)$
One Linear (OL)	$\text{FFN}\left(\text{ReLU}\left(\left(h_{\text{emb}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_{t-1}\right) W_{\text{IIB}}^{(\text{OL})}\right)\right)$	$\text{FFN}\left(\text{ReLU}\left(\left(h_{\text{out}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_t\right) W_{\text{OIB}}^{(\text{OL})}\right)\right)$
No Linear (NL)	$\text{FFN}\left(\text{ReLU}\left(h_{\text{emb}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_{t-1}\right)\right)$	$\text{FFN}\left(\text{ReLU}\left(h_{\text{out}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_t\right)\right)$
No Linear - No ReLU	$\text{FFN}(h_{\text{emb}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_{t-1})$	$\text{FFN}(h_{\text{out}}(\mathbf{z}_{t-1}) \oplus \mathbf{x}_t)$



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets.

Figure 8: Results of ablations of ChronosX with different covariate adapters' architectures, i.e. ChronosX (OL), ChronosX (NL), and ChronosX (NL-NR). Overall, our proposed architecture outperforms these variants.

lows the model to learn interactions independently, and has been explored in previous work (Pineda Arango and Grabocka, 2023). However, to further justify our design, we compare it with three other alternatives. Firstly, we consider ChronosX with one linear layer *(OL)* that receives the concatenated inputs of the embeddings of the target time series and their respective covariates. We also study ChronosX without the initial linear layers *(NL)*, and just an activation function before FFN. Finally, we ablate ChronosX without initial linear layers and activations *(NL-NR)*, just using a FFN. In Table 1, we formalize these variants mathematically both for the input and output modules, following the same notation as in our original formulation. Figure 8 reports the results after evaluating the architectures in our datasets. In general, we observe that our original formulation ChronosX outperforms the other alternatives in real datasets and simple synthetic data. In complex data, ChronosX (OL) obtained the best results among the adapter-only variants, but still is outperformed by ChronosX (FF). Interestingly, we observe that ChronosX (OL) is consistently better than ChronosX (NL) and ChronosX (NL-NR), which highlights the importance of the additional linear layer. ChronosX (NL) outperforms ChronosX (NL-NR) in all datasets, further motivating the use of the activation function before the Feed-Forward Network. Please refer to the Appendix for more details.

7 Conclusion

In this paper we presented extensions of pretrained time series models to incorporate exogenous variables via modules that inject information from past and future covariates. We showcase our approach with ChronosX and show that the same approach can be applied to other pretrained models with minimal modifications, resulting in extensions like TimesFMX and MOMENTX. We study the cases where we fine-tune the entire model or we only fit the parameters of the covariate-related modules, and introduced a family of 32 synthetic datasets with covariates to evaluate the efficacy of forecasting models to integrate covariates. We test the adapters in our models ChronosX, TimesFMX, and MOMENTX and demonstrate in synthetic and real datasets that our approach enables pretrained models to capture covariate information and perform favorably when compared to other baselines.

Limitations - Future Work. A limitation of this work is that the adapter approach for integrating covariate data requires training of the adapter module, therefore losing the benefit of pretrained models to perform zero-shot inference (arguable in exchange for lower forecasting error). An interesting research direction is to integrate covariates into pretrained forecasting models in inference akin to in-context learning, using directly the covariate data to reduce the forecast error.

References

A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *The Journal of Machine Learning Research*, 21(1):4629–4634, 2020.

A. F. Ansari, K. Benidis, R. Kurle, A. C. Turkmen, H. Soh, A. J. Smola, B. Wang, and T. Januschowski. Deep explicit duration switching models for time series. *Advances in Neural Information Processing Systems*, 34, 2021.

A. F. Ansari, L. Stella, A. C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and B. Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR. Expert Certification.

G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.

S.-A. Chen, C.-L. Li, S. O. Arik, N. C. Yoder, and T. Pfister. TSMixer: An all-MLP architecture for time series forecast-ing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=wbpxTuXgm0.

A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jn2iTJas6h.

S. Dooley, G. S. Khurana, C. Mohapatra, S. Naidu, and C. White. ForecastPFN: Synthetically-trained zero-shot forecasting. In *Advances in Neural Information Processing Systems*, 2023.

V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 74147–74181. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/874a4d89f2d04b4bcf9a2c19545cf040-Paper-Conference.pdf.

P. J. Fleming and J. J. Wallace. How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, 1986.

F. Garza, M. M. Canseco, C. Challú, and K. G. Olivares. StatsForecast: Lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/statsforecast.

J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski. Probabilistic forecasting with spline quantile function rnns. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1901–1910. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/gasthaus19a.html.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

R. Godahewa, C. Bergmeir, G. I. Webb, R. J. Hyndman, and P. Montero-Manso. Monash time series forecasting archive. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. MOMENT: A family of open time-series foundation models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16115–16152. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

J. Lago, G. Marcjasz, B. De Schutter, and R. Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2021.

116983. URL https://www.sciencedirect.com/science/article/pii/S0306261921004529.

B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

S. Makridakis, E. Spiliotis, and V. Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38 (4):1346–1364, 2022.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

K. G. Olivares, C. Challú, F. Garza, M. M. Canseco, and A. Dubrawski. NeuralForecast: User friendly state-of-the-art neural forecasting models. PyCon Salt Lake City, Utah, US 2022, 2022. URL https://github.com/Nixtla/neuralforecast.

K. G. Olivares, C. Challu, G. Marcjasz, R. Weron, and A. Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

J. Pfeiffer, S. Ruder, I. Vulić, and E. Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=z9EkXfvxta. Survey Certification.

S. Pineda Arango and J. Grabocka. Deep pipeline embeddings for automl. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1907–1919, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599303. URL https://doi.org/10.1145/3580305.3599303.

D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

O. Shchur, A. C. Turkmen, N. Erickson, H. Shen, A. Shirkov, T. Hu, and B. Wang. Autogluon–timeseries: Automl for probabilistic time series forecasting. In *International Conference on Automated Machine Learning*, pages 9–1. PMLR, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Z. Wang, Q. Wen, C. Zhang, L. Sun, L. V. Krannichfeldt, S. Pan, and Y. Wang. Benchmarks and custom package for energy forecasting, 2024. URL https://arxiv.org/abs/2307.07191.

G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pages 11106–11115. AAAI Press, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] → Refer to Section 3 and Section 6.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] → Refer to Section I.1.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No] → Released upon acceptance.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes].
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] ← Refer to Section B.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] ← Refer to Section H.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] → Refer to Section C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] → Refer to Section F.

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Appendix Structure

In this appendix, we specify details on different aspects of our work. We address the following items:

- In Appendix B, we detail the hyperparameters used for the main method.
- In Appendix C, we mention the computing resource used in the experiments.
- In Appendix D, we describe the Chronos variants.
- In Appendix E, we give more information about the synthetic dataset creation.
- In Appendix F, we detail information on the datasets with covariates.
- In Appendix G, we specify the baselines configurations.
- In Appendix H, we describe the metrics that we used in our experiments.
- In Appendix I, we present additional results:
  - In Appendix I.1, we compare the execution time and the number of parameters for different chronos variants.
  - Section I.2: full set of evaluations with ChronosX, TimesFMX, and MOMENTX
  - Section I.3: evaluations on covariate injection on pretrained models with either past covariates, future covariates, or both.
  - Section I.4: ablations showing that performance improvement is brought by covariates.
  - Section I.5: ablations showing the performance with larger version of Chronos.
  - Section I.6: ablations showing the performance of different adapter architectures for covariates.
  - Section I.7: Probabilistic forecasts visualizations.

B Hyperparameter Details

For all the experiments, we use the output dimension of the linear layers and the hidden dimension of the Feed Forward Networks is 256. We train the models for 5000 steps, and checkpointing the best observed model every 100 steps. The rest of the training setup such as the optimizer and learning rate scheduler were set following previous work Ansari et al. (2024). For TimesFMX we used a hidden dimension of 256, learning rate 1.e-5, and cosine annealing scheduler as in (Das et al., 2024). For MOMENTX we used a hidden dimension of 64, maximum learning 1.e-4 with one cycle learning rate schedule as in (Goswami et al., 2024). Remaining parameters of TimesFMX a MOMENTX were set up following previous works (Das et al., 2024; Goswami et al., 2024).

C Compute Resource Information

For the experiments, we use GPU Nvidia A10G. More expensive experiments used NVIDIA Tesla V100. For baselines we used cpu instances with 16 virtual cpus and 32 GiB of memory.

D Chronos Variants

D.1 Additional Covariate Modules: Hidden State Only and Covariates Only (Residual) Injection Block.

To understand the impact of the proposed output and input injection blocks, we present the following variants. First, we propose a hidden-state base adapter that does not take covariates into account and works as an alternative to full finetuning without covariates, defined as

$$f_{\text{HS}}(h_{\text{out}}(\mathbf{z}_{1:t-1})) = \text{FFN}\left(\text{ReLU}\left(h_{\text{out}}(\mathbf{z}_{1:t-1})W_{\text{OIB}}^{(\text{out})}\right)\right) \tag{11}$$

and second, we present an adapter to incorporate covariates of the future by updating logits through a linear transformation.

$$f_{\text{RS}}\left(\mathbf{z}_{1:t-1}, \mathbf{x}_t\right) = h_{\text{out}}(\mathbf{z}_{1:t-1})W_{\text{out}} + \mathbf{x}_t W_{\text{OIB}}^{(\text{cov})} \tag{12}$$

D.2 Variants List

Here we explain the baselines, for more information please refer to the section 3. In the notation below, we refer to the IIB (Input Injection Block) as the network that receives the past covariates and updates the token embedding, while the OIB (Output Injection Block) indicates the network that receives the hidden state and the future covariates.

- Chronos is the model with the pretrained weights (not finetuned).
- Chronos(FF) is the pretrained model full-finetuned on the corresponding dataset.
- ChronosX, represented as well as ChronosX(IIB+OIB), is the frozen pretrained model with the additional Input and Output Injection Blocks.
- ChronosX(FF), represented as well as ChronosX(FF+IIB+OIB), is the full-finetuned Chronos model with the additional Input and Output Injection Block.
- ChronosX(FF+IIB) is the full-finetuned Chronos model with Input Injection Block.
- ChronosX(FF+OIB) is the full-finetuned Chronos model with Output Injection Block.
- ChronosX(OIB) is the frozen pretrained model only with Output Injection Block.
- ChronosX(IIB) is the frozen pretrained model only with Input Injection Block.
- ChronosX(HS) is the frozen pretrained model only with hidden state, as in Equation 11.
- ChronosX(RS) is the frozen pretrained model only with modules accepting covariates, as in Equation 12.

E Synthetic Dataset Creation

We created a suite of 32 synthetic datasets comprising a 100 time series with length $L = 1827$ and 30 steps for the prediction length, assuming a "fictional" daily frequency. Each dataset is created by combining a **main signal** $\hat{z}_t$ with a **covariate** x_t using a **combination operator** according to Eq. 13.

$$z_t = \hat{z}_t \odot x_t, \quad t = 1, ..., L \tag{13}$$

Once a combination is chosen, we randomly sample the parameters for the main signal appendix E.1 and the covariates appendix E.2 to get 100 different time series. All possible combinations are illustrated in fig. 9 and fig. 10.

E.1 Main signal generation

The main signal to be affected by the covariates can be generated in different ways. We describe four variants in an increasing order of complexity by changing the set of possible parameters in Eq. 14.

$$\hat{z}_t = \sum_{i=1}^{3} a_i \sin\left(f_i t + \phi_i\right) + b_1 t + b_2 + \epsilon \tag{14}$$

- **Single Sinuoid**: This represents the simplest time series, keeping the same amplitude, phase and a weekly frequency among time series.

$$\hat{z}_t^{(\text{single})} = \sin\left(2\pi \frac{t}{7}\right)$$

- **Simple Sinusoids**: Every time series in the dataset is the superposition of three sinusoids with different frequencies, simulating weekly, monthly and yearly seasonal patterns. For each time series, a different amplitude is sampled from $a_i \sim U(1, 5)$.

$$\hat{z}_t^{(\text{sinusoids})} = a_1 \sin\left(2\pi \frac{t}{7}\right) + a_2 \sin\left(2\pi \frac{t}{30}\right) + a_3 \sin\left(2\pi \frac{t}{365}\right)$$

- **Diverse Sinusoids**: Every time series is similar to *simple sinusoids*, but we sample a different phase for every sinusiodal component with a random amplitude $a_i \sim U(1, 5)$ and a random phase $\phi_i \sim U(-\pi, \pi)$. A trend component is also added with random coefficient $b_1, b_2 \sim U(-1, 1)$.

$$\hat{z}_t^{(\text{diverse})} = a_1 \sin\left(2\pi \frac{t}{7} + \phi_1\right) + a_2 \sin\left(2\pi \frac{t}{30} + \phi_2\right) + a_3 \sin\left(2\pi \frac{t}{365} + \phi_3\right) + b_1 \frac{t}{365} + b_2$$

- **Noisy Sinusoids**. The Noisy Sinusoid are *Diverse sinusoids* with noise added at every step. The variance is chosen to be proportional to the scale $s = \frac{1}{L}\sum_t^L |\hat{z}_t|$ of the main diverse sinusoid $\epsilon \sim \mathcal{N}\left(0, \frac{s}{4}\right)$.

$$\hat{z}_t^{(\text{noisy})} = \hat{z}_t^{(\text{diverse})} + \epsilon$$

E.2 Covariate

Similarly to the main signal, the parameters of the covariates are randomly chosen depending on the types and the scale of the signal.

- **Spikes** are used to represent short time events by considering large value for single "active" time steps as in Eq. 15. The $N = 500$ times steps $\mathbb{T} = \{t_1, \ldots, t_N\}$ are uniformly sampled without replacement. The strength of all the spikes is randomly sampled in an interval limited to 5 times the scale $\gamma \sim U(1, 5s)$.

$$x_t = \begin{cases} \gamma & \text{if } t \in \mathbb{T}_N \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

- **Steps** are used to study the influence of sudden but sustained events. Steps covariates share the same formula as the spikes in Eq. 16 but the generation of the "active" index is different. We consider $N = 125$ non-overlapping intervals $T_i = \{t_i, t_{i+1}, \ldots, t_i + \delta_i\}$ by uniformly sampling the starting steps $t_i \sim U(0, L)$ and the associated duration $\delta_i \sim U(1, 30)$ with $T_i \cap T_j = \emptyset$. The resulting signal is then composed of a union of the different intervals : $\mathbb{T} = \cup_i^N T_i$. All steps have the save amplitude γ, randomly sampled in the same manner as the spikes.

$$x_t = \begin{cases} \gamma & \text{if } t \in \mathbb{T} \\ 1 & \text{if otherwise} \end{cases} \tag{16}$$

- **Bells** are considered to represent trends with smoother changes. In this case, the signal is a mixture of gaussian bells as in Eq. 17. We consider mixtures of $N = 125$ bells with randomly sampled mean $\hat{\mu}_i \sim U(0, L)$ and randomly sampled standard deviation $\sigma_i \sim U(1, 15)$. The scale of the bells is randomly sampled in an interval limited to 5 times the scale $\gamma \sim U(1, 5s)$.

$$x_{\mathrm{t}} = \gamma \sum_i^N \exp\left(\frac{-\|t - \hat{\mu}_i\|^2}{\sigma_i^2}\right) \tag{17}$$

- **ARP** is created following Eq. 18. Only second order auto-regressive processes are created with coefficients $a_1 \sim U(0, 1)$ and $a_2 = 1 - a_1$, and noise amplitude equal to one. The scale of the final covariates is randomly sampled in an interval limited to 5 times the scale $\gamma \sim U(1, 5s)$.

$$x_t = a_1 \cdot x_{t-1} + a_2 \cdot x_{t-2} + \epsilon \tag{18}$$

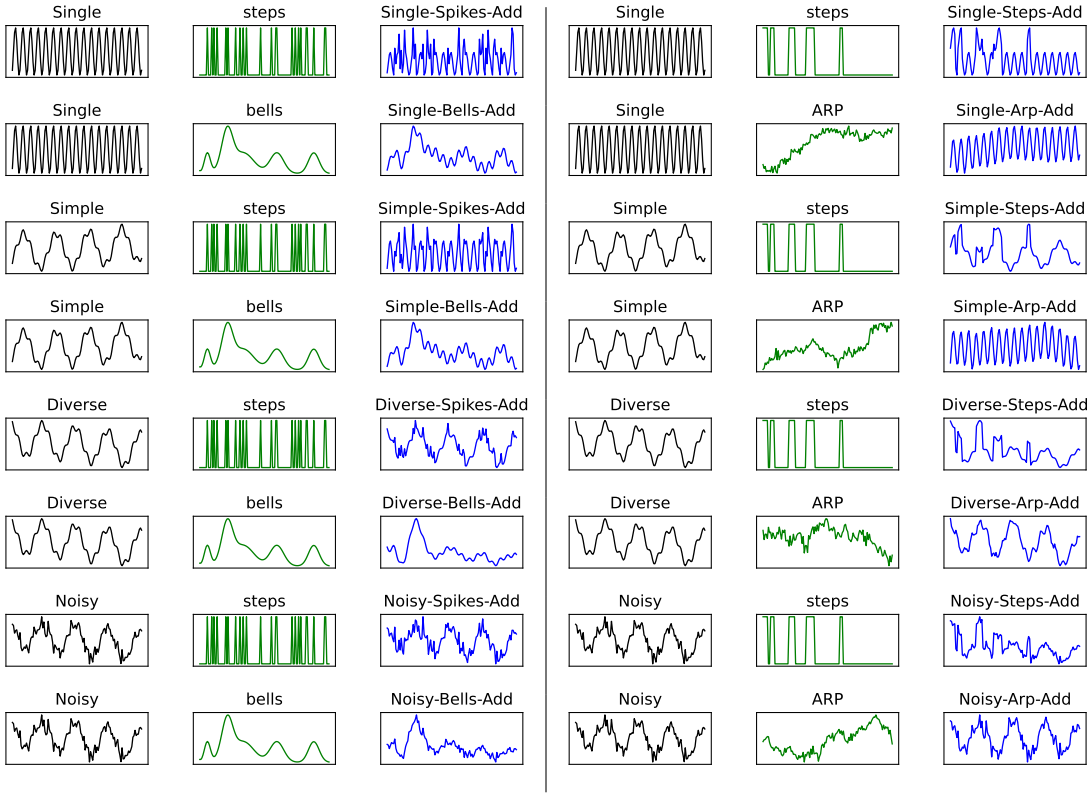


Figure 9: This figure provides a single time series example from all possible combinations of **covariates** and **main signal** with the add + **operator.**

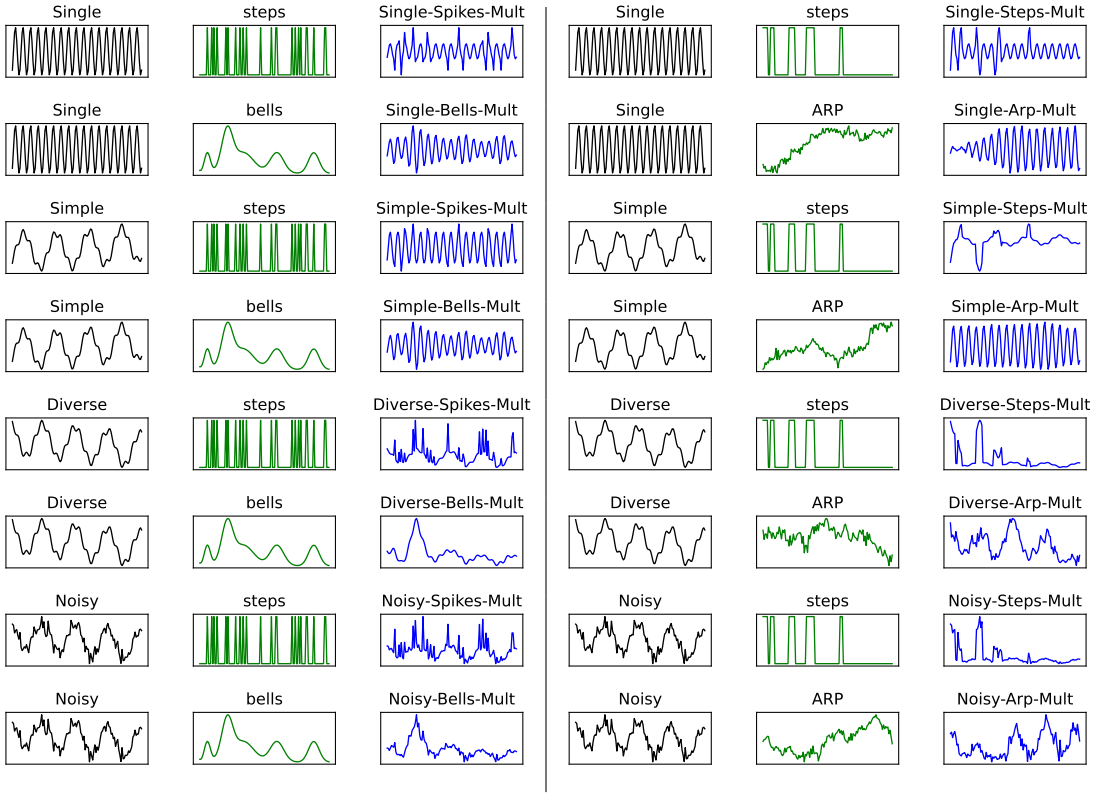


Figure 10: This figure provides a single time series example from all possible combinations of **covariates** and **main signal** with the mult $\times$ **operator.**

F Datasets With Covariates

In this section we present a summary of the datasets with covariates used for evaluations.

Table 2: Real Datasets Descriptors.

Dataset Name	Num. Series	Num. Covariates	Freq.	Prediction Length	Source
ETT (15 Min.)	2	5	15min	24	(Zhou et al., 2021)
ETT (Hourly)	2	5	1H	24	(Zhou et al., 2021)
M5	30490	12	1D	28	(Makridakis et al., 2022)
Electricity-BE	1	2	1H	24	(Lago et al., 2021; Olivares et al., 2023)
Electricity-DE	1	2	1H	24	(Lago et al., 2021; Olivares et al., 2023)
Electricity-FR	1	2	1H	24	(Lago et al., 2021; Olivares et al., 2023)
Electricity-NP	1	2	1H	24	(Lago et al., 2021; Olivares et al., 2023)
Electricity-PJM	1	2	1H	24	(Lago et al., 2021; Olivares et al., 2023)
BDG-2 Hog	24	5	1H	24	(Woo et al., 2024; Wang et al., 2024)
BDG-2 Bull	41	3	1H	24	(Woo et al., 2024; Wang et al., 2024)
BDG-2 Cockatoo	1	5	1H	24	(Woo et al., 2024; Wang et al., 2024)
Covid19Energy	1	6	1H	24	(Woo et al., 2024; Wang et al., 2024)
GEF12	20	1	1H	24	(Woo et al., 2024; Wang et al., 2024)
GEF14	1	1	1H	24	(Woo et al., 2024; Wang et al., 2024)
GEF17	8	1	1H	24	(Woo et al., 2024; Wang et al., 2024)
PDB	1	1	1H	24	(Woo et al., 2024; Wang et al., 2024)
Spanish	1	20	1H	24	(Woo et al., 2024; Wang et al., 2024)
Rideshare	156	8	1H	24	(Godahewa et al., 2021)

Table 3: Baseline models and hyperparameter choices. Hyperparameters not specified are set to defaults in their respective implementations. C stands for context length, d_h for hidden layer dimension, n_L for number of layers, n_H for number of heads, and η for learning rate.

Model	Model Type	Implementation	Probabilistic	Hyperparameters
SeasonalNaive	Local	StatsForecast	Yes	N/A
AutoARIMA	Local	StatsForecast	Yes	$C = 1000$
DeepAR	Task-specific	GluonTS	Yes	$d_h = 40, n_L = 2$
TFT	Task-specific	GluonTS	Yes	$d_h = 32, n_H = 4$
PatchTST	Task-specific	GluonTS	Yes	Patch length: 16, Stride: 8, $d_h = 32, n_L = 2, n_H = 4$
DLinear	Task-specific	GluonTS	Yes	Kernel size: 25, $d_h = 20$
TiDE	Task-specific	GluonTS	Yes	encoder ($d_h = 512$, $n_L = 2$), decoder ($d_h = 512$, $n_L = 2$, output dimension: 8), temporal hidden dimension: 128, distribution hidden dimension: 512.
N-BEATSx-I	Task-specific	NeuralForecast	No	Input size multiplier: 5, scaler: robust_scaler, stack_types: Interpretable
N-BEATSx-G	Task-specific	NeuralForecast	No	Input size multiplier: 5, scaler: robust_scaler, stack_types: Generalized
N-HiTS	Task-specific	NeuralForecast	No	Input size multiplier: 5, scaler: robust_scaler
Moirai-1.0-R	Pretrained	Reference	Yes	$C = 1024$, Patch length: auto, batch size:4
TimesFM-1.0-200m-pytorch	Pretrained	Reference	No	$C = 512$
MOMENT	Pretrained/Task-specific	Reference	No	Forecasting head training epochs: 100, $C = 512, \eta = 10^{-4}$
TTM-R2	Pretrained	Reference	No	$C = 512$
TTM-CM	Pretrained/Task-specific	Reference	No	$C = 512$

G Baselines

For task-specific deep learning architectures DeepAR (Salinas et al., 2020), PatchTST (Nie et al., 2023), TFT (Lim et al., 2021) we based evaluations on implementations in GluonTS (Alexandrov et al., 2020). However, NBEATSx-I, NBEATSx-G (Olivares et al., 2023) and NHiTS (Challu et al., 2023) experiments were based on implementations in the NeuralForecast (Olivares et al., 2022) library. The original PatchTST model does not naturally accept additional time features as input, so we extend it in the following way. Similar to target time series, covariates are converted to patches and appended to the target patches. Since PatchTST is encoder only model, we combine past and future covariates (if available) and (left) shift them by prediction length so that the covariates are aligned with the target input. We call this version PatchTSTx.

The baselines DeepAR, PatchTST, TFT, NBEATSx-I, NBEATSx-G, , and N-HiTS were trained and evaluated three times and their performance averaged in order to account for high variance inherent in their optimization.

Statistical baselines SeasonalNaive was used with their default hyperparameters in StatsForecast (Garza et al., 2022), but with season length implied by their frequencies. For example, daily frequency data had season length set to 7, hourly data 24, etc. For this heuristic, we used the helper function `get_seasonality` from GluonTS.

Default hyperparameter configurations provided in baseline implementations were kept as is, and no dataset specific or global hyperparameter tuning was performed. GluonTS-based implementations were optimized with a batch size of 128.

H Evaluation Metrics

For the evaluation metrics we follow the approach exposed in (Ansari et al., 2021). Let $\{\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,C+H}]\}_{i=1}^{N}$ be a dataset with N time series where C and H are the context length and prediction length, respectively.

The mean absolute scaled error (MASE, Hyndman and Koehler (2006)) is defined as:

$$\mathrm{MASE}(\hat{\mathbf{x}}_i, \mathbf{x}_i) = \frac{C-S}{H} \frac{\sum_{t=C+1}^{C+H} |\hat{x}_{i,t} - x_{i,t}|}{\sum_{t=1}^{C-S} |x_{i,t} - x_{i,t+S}|},$$

where S is a seasonality parameter.

The Weighted Quantile Loss is defined as

$$\mathrm{WQL} = \frac{1}{K} \sum_{j=1}^{K} \mathrm{WQL}_{\alpha_j}.$$

where K is the number of quantiles, α is the set of quantile levels to evaluate, and

$$\mathrm{WQL}_\alpha = \frac{2\sum_{i,t} \mathrm{QL}_\alpha(q_{i,t}^{(\alpha)}, x_{i,t})}{\sum_{i,t} |x_{i,t}|}.$$

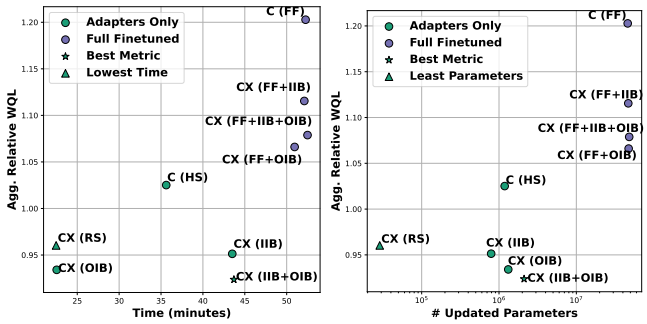
where $\mathbf{q}_i^{(\alpha)} = [q_{i,C+1}^{(\alpha)}, \ldots, q_{i,C+H}^{(\alpha)}]$ are the predicted quantiles at levels $\alpha \in (0,1)$, and

$$\mathrm{QL}_\alpha(q,x) = \begin{cases} \alpha(x-q), & \text{if } x > q, \\ (1-\alpha)(q-x), & \text{otherwise.} \end{cases} \tag{19}$$

I Additional Results

I.1 Time and Memory Efficiency of Adapters and Finetuning.

We present an empirical analysis of the trade-offs between our models using only adapters and full finetuning against time execution and number of updated parameters. For a better visualization we represent ChronosX and Chronos with CX and C, respectively. For this, we measure the average computational time consumed per run for training and testing across datasets and compared it to the aggregated WQL metric. As can be seen in Figure 11a, the adapters are among the best performing and fastest variants. This is due to the decreased number of updated parameters as we report in Figure 11b. Overall the most efficient variants in terms of time and number of updated hyperparameters is ChronosX(RS), while the best performing and still efficient is ChronosX. We can see that full finetuning models have a larger amount of updated parameters and larger execution time than our approaches using only adapters.



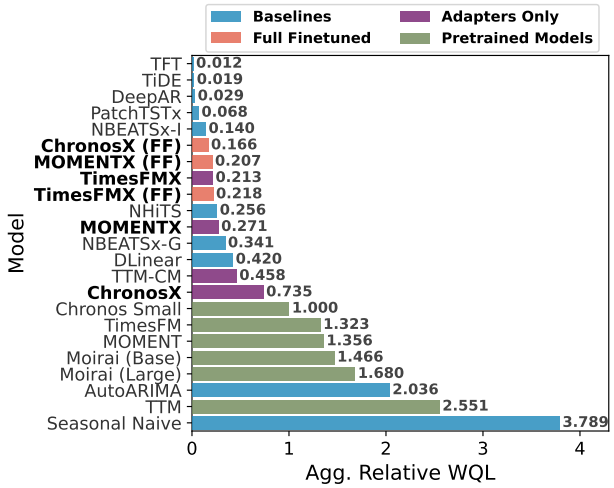
(a) WQL vs Exec. Time (b) WQL vs # Parameters

Figure 11: Average Time Execution and number of updated parameters of our proposed models against Aggregated Relative Weighted Quantile Loss on real datasets with covariates.

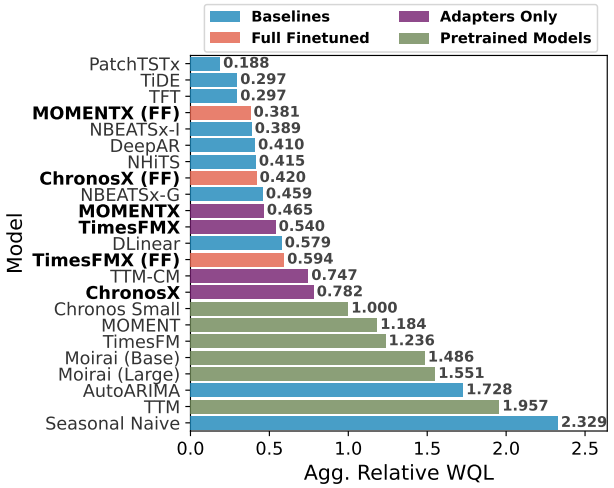
I.2 Full set of evaluations with ChronosX, TimesFMX, and MOMENTX

We present the following results:

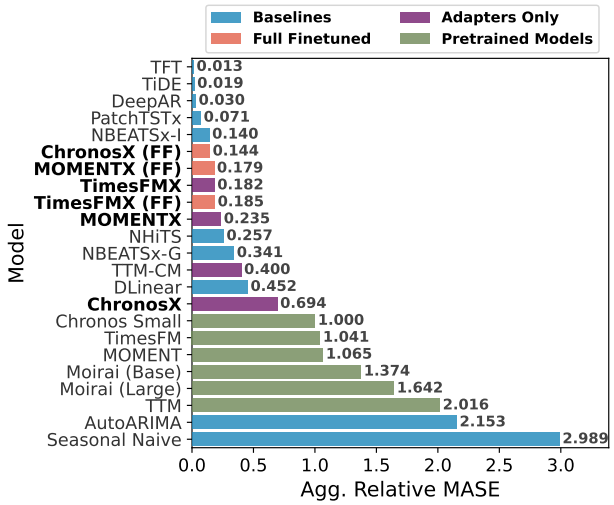
- In Fig. 12 we present the Aggregated Relative WQL and MASE on simple, complex, and real datasets,
- In Table 4 and Table 5 we present the WQL on simple and complex datasets, respectively,
- In Table 6 and Table 7 we present the MASE on simple datasets, respectively.
- In Table 8 and Table 9 we present the MASE and WQL on real datasets, respectively.



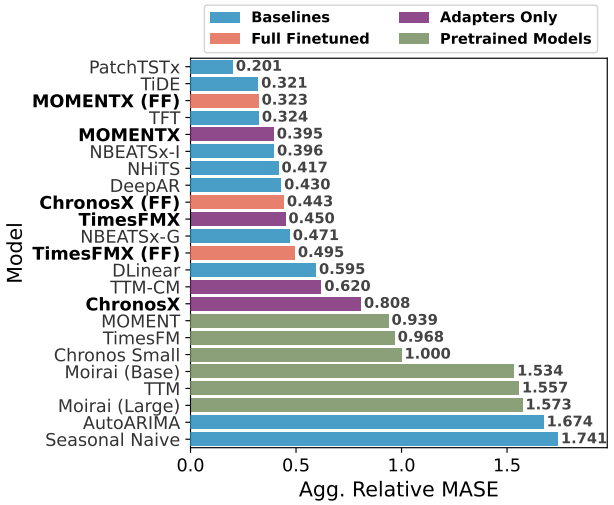
(a) Agg. Rel. WQL on **Simple** synthetic datasets.



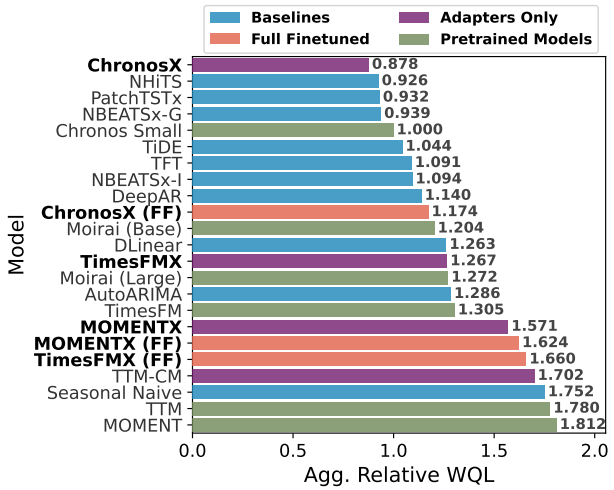
(b) Agg. Rel. WQL on **Complex** synthetic datasets.



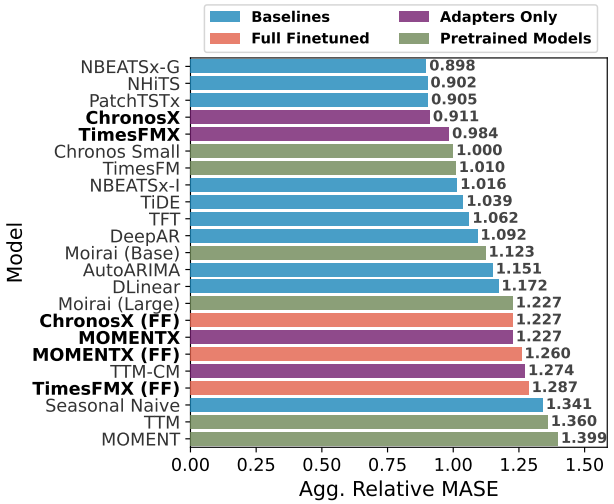
(c) Agg. Rel. MASE on **Simple** synthetic datasets



(d) Agg. Rel. MASE on **Complex** synthetic datasets



(e) Agg. Rel. WQL on **real** datasets



(f) Agg. Rel. MASE on **real** datasets

Figure 12: Aggregated Relative WQL and MASE on simple, complex, and real datasets. Emphasis on comparisons of pretrained models with extensions to incorporate covariates.

Table 4: WQL scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models						Baselines								
	ChronosX (FF)	MOMENTX (FF)	TimesFMX (FF)	ChronosX	MOMENTX	TimesFMX	TTM-CM	Chronos Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Simple ARP (Add)	0.164	0.070	0.024	0.168	0.074	0.036	0.063	0.056	0.099	0.148	0.075	0.059	0.276	0.198	0.007	0.007	**0.002**	0.005	0.015	0.012	0.402	**0.003**	0.022
Simple ARP (Mult.)	0.042	0.039	0.038	0.136	0.047	0.026	0.044	0.037	0.114	0.153	0.058	0.048	0.374	0.313	**0.001**	0.004	0.003	0.012	0.018	0.014	0.572	**0.002**	0.030
Simple Bells (Add)	0.287	0.035	0.079	0.225	0.051	0.060	0.280	0.319	0.346	0.383	0.402	0.453	0.415	0.176	0.005	0.056	**0.003**	0.048	0.061	0.037	0.582	**0.005**	0.060
Simple Bells (Mult.)	0.041	0.022	0.018	0.176	0.025	0.022	0.053	0.089	0.136	0.156	0.104	0.105	0.290	0.273	**0.001**	0.005	0.004	0.040	0.037	0.033	0.417	**0.002**	0.093
Simple Spikes (Add)	0.007	0.042	0.088	0.250	0.056	0.056	0.081	0.431	0.498	0.515	0.615	0.607	0.667	0.381	**0.003**	**0.003**	0.004	0.036	0.117	0.083	0.978	0.004	0.066
Simple Spikes (Mult.)	0.007	0.018	0.017	0.022	0.024	0.023	0.046	0.142	0.196	0.212	0.201	0.194	0.278	0.446	**0.001**	**0.002**	0.099	0.014	0.116	0.101	0.440	0.003	0.113
Simple Steps (Add)	0.007	0.035	0.059	0.185	0.050	0.060	0.102	0.415	0.458	0.455	0.591	0.605	0.607	0.293	**0.004**	0.009	**0.004**	0.067	0.105	0.096	0.835	0.005	0.080
Simple Steps (Mult.)	0.007	0.019	0.022	0.030	0.025	0.025	0.062	0.169	0.269	0.275	0.254	0.255	0.347	0.486	**0.001**	**0.002**	0.143	0.019	0.046	0.032	0.577	0.003	0.169
Single ARP (Add.)	0.130	0.078	0.043	0.175	0.087	0.029	0.069	0.062	0.106	0.152	0.077	0.070	0.289	0.222	0.005	0.007	**0.003**	0.008	0.018	0.013	0.416	**0.004**	0.025
Single ARP (Mult.)	0.038	0.040	0.038	0.157	0.048	0.023	0.045	0.038	0.109	0.131	0.065	0.058	0.347	0.410	**0.001**	0.004	0.004	0.011	0.016	0.012	0.538	**0.002**	0.030
Single Bells (Add.)	0.252	0.034	0.041	0.261	0.050	0.058	0.110	0.367	0.383	0.410	0.412	0.476	0.469	0.181	**0.006**	0.036	**0.003**	0.072	0.098	0.063	0.592	0.036	0.067
Single Bells (Mult.)	0.034	0.023	0.018	0.214	0.024	0.017	0.049	0.105	0.136	0.163	0.107	0.109	0.321	0.289	**0.001**	0.005	0.004	0.040	0.055	0.038	0.461	**0.002**	0.089
Single Spikes (Add.)	0.008	0.033	0.056	0.181	0.049	0.043	0.068	0.391	0.436	0.479	0.550	0.544	0.635	0.310	**0.001**	0.002	0.242	0.025	0.139	0.073	0.954	**0.001**	0.043
Single Spikes (Mult.)	0.008	0.018	0.021	0.031	0.024	0.021	0.061	0.145	0.197	0.230	0.209	0.201	0.288	0.463	**0.001**	**0.001**	0.098	0.012	0.120	0.108	0.462	0.003	0.116
Single Steps (Add.)	0.008	0.032	0.033	0.151	0.046	0.047	0.076	0.390	0.449	0.471	0.552	0.536	0.617	0.262	**0.002**	0.003	**0.002**	0.021	0.062	0.058	0.900	0.003	0.060
Single Steps (Mult.)	0.007	0.021	0.019	0.018	0.036	0.024	0.068	0.186	0.283	0.291	0.256	0.268	0.361	0.622	**0.001**	**0.001**	0.146	0.018	0.050	0.034	0.596	0.003	0.174
Agg. Relative Score	0.166	0.207	0.218	0.735	0.271	0.213	0.458	1.000	1.466	1.680	1.356	1.323	2.551	2.036	**0.012**	0.029	0.068	0.140	0.341	0.256	3.789	**0.019**	0.420
Avg. Rank	9.250	8.000	8.438	15.125	10.250	8.250	12.688	14.812	18.000	19.188	18.188	17.938	21.562	19.312	**1.625**	3.625	5.625	6.562	11.375	9.438	22.812	**2.562**	11.375

Table 5: WQL scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models						Baselines								
	ChronosX (FF)	MOMENTX (FF)	TimesFMX (FF)	ChronosX	MOMENTX	TimesFMX	TTM-CM	Chronos Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Diverse ARP (Add)	0.092	0.052	0.050	0.117	0.052	0.052	0.052	0.063	0.138	0.167	0.061	0.063	0.227	0.265	0.023	0.032	**0.007**	**0.009**	0.019	0.012	0.250	0.019	0.028
Diverse ARP (Mult.)	0.036	0.047	0.062	0.077	0.054	0.048	0.057	0.059	0.137	0.153	0.059	0.057	0.202	0.269	0.030	0.029	**0.016**	**0.024**	0.028	0.024	0.221	0.027	0.041
Diverse Bells (Add)	0.208	0.066	0.134	0.271	0.088	0.117	0.255	0.257	0.313	0.317	0.322	0.355	0.369	0.168	**0.027**	0.161	**0.011**	1.162	0.105	0.124	0.464	0.035	0.081
Diverse Bells (Mult.)	0.051	0.040	0.057	0.116	0.059	0.061	0.096	0.089	0.155	0.162	0.104	0.113	0.194	0.233	**0.036**	0.081	**0.012**	0.050	0.067	0.055	0.213	0.040	0.084
Diverse Spikes (Add)	0.016	0.031	0.064	0.075	0.040	0.055	0.065	0.230	0.279	0.288	0.293	0.306	0.374	0.352	0.019	0.036	**0.006**	0.029	0.046	0.049	0.460	**0.015**	0.074
Diverse Spikes (Mult.)	**0.021**	0.024	0.049	0.049	0.032	0.035	0.096	0.084	0.131	0.134	0.110	0.119	0.160	0.200	**0.012**	0.027	0.055	0.029	0.068	0.061	0.202	0.021	0.079
Diverse Steps (Add)	0.046	0.029	0.106	0.108	0.046	0.095	0.074	0.273	0.293	0.296	0.337	0.350	0.392	0.326	0.028	0.076	**0.007**	0.028	0.061	0.051	0.490	**0.018**	0.107
Diverse Steps (Mult.)	0.042	**0.031**	0.085	0.087	0.052	0.061	0.114	0.126	0.188	0.185	0.156	0.174	0.236	0.264	**0.033**	0.086	0.036	0.049	0.067	0.054	0.273	0.037	0.119
Noisy ARP (Add)	0.063	0.080	0.080	0.102	0.080	0.078	0.082	0.072	0.152	0.169	0.089	0.090	0.229	0.185	0.059	0.052	**0.048**	**0.049**	0.051	0.051	0.261	0.058	0.053
Noisy ARP (Mult.)	0.064	0.066	0.073	0.077	0.070	0.068	0.072	0.066	0.134	0.144	0.077	0.075	0.196	0.253	0.053	**0.047**	0.045	0.047	0.048	0.048	0.219	0.052	0.053
Noisy Bells (Add)	0.198	0.071	0.131	0.265	0.088	0.119	0.237	0.249	0.290	0.299	0.327	0.329	0.339	0.108	0.051	0.052	**0.032**	0.062	0.078	0.066	0.439	**0.051**	0.080
Noisy Bells (Mult.)	**0.049**	0.065	0.083	0.101	0.076	0.077	0.101	0.094	0.157	0.163	0.106	0.121	0.191	0.205	0.072	0.058	**0.042**	0.088	0.091	0.081	0.213	0.068	0.090
Noisy Spikes (Add)	**0.050**	0.061	0.078	0.087	0.065	0.076	0.076	0.239	0.274	0.277	0.293	0.299	0.357	0.208	0.061	0.051	**0.044**	0.053	0.059	0.063	0.433	0.059	0.072
Noisy Spikes (Mult.)	**0.038**	0.049	0.050	0.059	0.053	0.054	0.101	0.100	0.146	0.149	0.118	0.124	0.183	0.197	0.048	**0.040**	0.050	0.050	0.084	0.076	0.216	0.048	0.088
Noisy Steps (Add)	0.065	0.058	0.097	0.127	0.066	0.095	0.097	0.277	0.267	0.266	0.301	0.315	0.349	0.197	0.066	0.053	**0.041**	**0.049**	0.071	0.060	0.425	0.065	0.096
Noisy Steps (Mult.)	**0.044**	0.053	0.074	0.072	0.063	0.069	0.118	0.120	0.154	0.158	0.146	0.146	0.190	0.223	0.057	0.050	**0.042**	0.058	0.065	0.062	0.259	0.060	0.108
Agg. Relative Score	0.420	0.381	0.594	0.782	0.465	0.540	0.747	1.000	1.486	1.551	1.184	1.236	1.957	1.728	0.297	0.410	**0.188**	0.389	0.459	0.415	2.329	**0.297**	0.579
Avg. Rank	6.625	6.188	11.750	15.188	9.312	10.500	13.875	14.938	18.500	19.250	17.688	18.375	21.375	19.875	4.750	6.062	**2.125**	5.875	8.562	7.250	22.688	**4.250**	11.000

Table 6: MASE scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models						Baselines								
	ChronosX (FF)	MOMENTX (FF)	TimesFMX (FF)	ChronosX	MOMENTX	TimesFMX	TTM-CM	Chronos Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Simple ARP (Add)	0.947	0.305	0.108	0.951	0.328	0.163	0.283	0.322	0.571	0.849	0.328	0.264	1.228	1.103	0.039	0.034	**0.013**	0.029	0.086	0.067	1.788	**0.019**	0.122
Simple ARP (Mult.)	0.161	0.130	0.134	0.500	0.160	0.093	0.148	0.157	0.454	0.642	0.204	0.175	1.260	1.616	**0.006**	0.020	0.014	0.051	0.070	0.057	1.930	**0.009**	0.128
Simple Bells (Add)	3.203	0.310	0.656	2.459	0.439	0.499	2.976	3.483	3.689	4.069	3.408	3.858	3.520	1.993	**0.055**	0.613	**0.031**	0.491	0.641	0.392	4.951	0.059	0.682
Simple Bells (Mult.)	0.220	0.098	0.081	0.858	0.111	0.095	0.240	0.473	0.691	0.847	0.466	0.477	1.305	1.698	**0.004**	0.025	0.019	0.212	0.206	0.177	1.882	**0.012**	0.583
Simple Spikes (Add)	0.010	0.112	0.189	0.372	0.160	0.126	0.205	0.653	0.756	0.823	0.740	0.732	0.862	0.644	0.013	0.009	**0.007**	0.071	0.231	0.146	1.258	**0.009**	0.162
Simple Spikes (Mult.)	0.024	0.061	0.056	0.089	0.079	0.076	0.151	0.581	0.664	0.781	0.642	0.620	0.885	1.930	**0.004**	**0.006**	0.456	0.054	0.458	0.453	1.397	0.009	0.490
Simple Steps (Add)	**0.023**	0.232	0.358	0.742	0.309	0.356	0.649	1.653	1.856	1.862	1.811	1.790	1.863	1.079	0.033	0.077	**0.018**	0.337	0.452	0.422	2.577	0.029	0.411
Simple Steps (Mult.)	0.023	0.058	0.070	0.113	0.079	0.084	0.193	0.720	1.014	1.067	0.808	0.802	1.122	2.262	**0.008**	**0.007**	0.703	0.076	0.177	0.124	1.808	0.009	0.735
Single ARP (Add.)	0.601	0.331	0.181	0.811	0.372	0.121	0.294	0.344	0.548	0.823	0.326	0.299	1.230	1.200	0.027	0.035	**0.015**	0.040	0.093	0.072	1.774	**0.020**	0.137
Single ARP (Mult.)	0.148	0.138	0.129	0.593	0.166	0.082	0.154	0.166	0.418	0.540	0.221	0.197	1.213	1.914	**0.004**	0.019	0.016	0.048	0.069	0.053	1.877	**0.009**	0.128
Single Bells (Add.)	2.675	0.273	0.331	2.599	0.402	0.469	0.886	3.728	3.835	4.079	3.324	3.834	3.779	1.945	**0.054**	0.346	**0.026**	0.687	0.965	0.616	4.779	0.067	0.673
Single Bells (Mult.)	0.165	0.099	0.080	1.122	0.103	0.073	0.212	0.539	0.684	0.851	0.462	0.470	1.384	1.611	**0.003**	0.025	0.019	0.214	0.299	0.208	1.989	**0.012**	0.534
Single Spikes (Add.)	0.010	0.043	0.073	0.309	0.064	0.056	0.089	0.653	0.690	0.826	0.716	0.709	0.826	0.540	**0.002**	0.002	0.470	0.040	0.235	0.112	1.242	**0.001**	0.072
Single Spikes (Mult.)	0.023	0.056	0.064	0.115	0.075	0.065	0.186	0.575	0.629	0.831	0.640	0.617	0.881	1.909	**0.002**	**0.004**	0.451	0.046	0.453	0.449	1.415	0.009	0.482
Single Steps (Add.)	0.019	0.080	0.083	0.514	0.118	0.119	0.193	1.273	1.501	1.566	1.400	1.358	1.564	0.886	0.005	0.009	**0.008**	0.064	0.196	0.186	2.281	0.009	0.197
Single Steps (Mult.)	0.023	0.065	0.059	0.061	0.112	0.074	0.209	0.735	0.998	1.067	0.790	0.827	1.115	2.672	**0.004**	**0.004**	0.684	0.068	0.192	0.127	1.841	0.009	0.714
Agg. Relative Score	0.144	0.179	0.185	0.694	0.235	0.182	0.400	1.000	1.374	1.642	1.065	1.041	2.016	2.153	**0.013**	0.030	0.071	0.140	0.341	0.257	2.989	**0.019**	0.452
Avg. Rank	9.250	7.562	8.125	15.125	9.875	8.062	12.312	16.000	18.750	20.312	17.125	16.812	21.000	19.625	**2.125**	3.750	5.562	6.812	11.188	9.188	22.688	**2.438**	12.312

Table 7: MASE scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models						Baselines								
	ChronosX (FF)	MOMENTX (FF)	TimesFMX (FF)	ChronosX	MOMENTX	TimesFMX	TTM-CM	Chronos Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Diverse ARP (Add)	1.004	0.547	0.509	1.282	0.551	0.511	0.520	0.768	1.848	2.056	0.608	0.613	2.130	2.801	0.317	0.419	**0.099**	**0.110**	0.242	0.148	1.947	0.270	0.318
Diverse ARP (Mult.)	0.500	0.515	0.659	0.987	0.579	0.501	0.598	0.772	1.791	1.928	0.648	0.571	1.973	3.091	0.404	0.370	**0.221**	**0.314**	0.374	0.323	1.921	0.360	0.533
Diverse Bells (Add)	3.392	0.834	1.660	4.066	1.105	1.510	3.557	3.975	4.713	4.634	3.871	4.417	4.529	2.234	**0.446**	2.513	**0.169**	12.561	1.584	1.512	5.563	0.586	1.249
Diverse Bells (Mult.)	0.925	**0.572**	0.850	1.959	0.861	0.904	1.445	1.620	2.755	2.834	1.580	1.618	2.712	3.453	0.651	1.462	**0.249**	0.920	1.171	1.000	2.698	0.721	1.619
Diverse Spikes (Add)	0.062	0.100	0.195	0.294	0.127	0.162	0.195	0.767	0.999	1.048	0.775	0.801	1.049	1.203	0.077	0.132	**0.022**	0.107	0.176	0.163	1.286	**0.059**	0.274
Diverse Spikes (Mult.)	**0.173**	0.185	0.351	0.451	0.250	0.265	0.635	0.691	1.201	1.247	0.722	0.772	1.211	1.826	**0.121**	0.240	0.491	0.264	0.537	0.516	1.427	0.199	0.492
Diverse Steps (Add)	0.393	0.198	0.691	0.896	0.303	0.598	0.475	2.099	2.388	2.387	2.019	2.107	2.396	2.182	0.252	0.584	**0.056**	0.222	0.490	0.409	2.851	**0.155**	0.837
Diverse Steps (Mult.)	0.495	**0.319**	0.842	1.054	0.542	0.607	1.093	1.464	2.195	2.110	1.421	1.574	2.165	3.165	0.404	1.021	**0.401**	0.581	0.818	0.644	2.329	0.448	1.396
Noisy ARP (Add)	0.613	0.572	0.585	0.967	0.579	0.573	0.591	0.713	1.473	1.580	0.645	0.663	1.609	1.760	0.570	0.495	**0.459**	**0.464**	0.483	0.477	1.713	0.558	0.500
Noisy ARP (Mult.)	0.750	0.578	0.633	0.884	0.615	0.603	0.646	0.778	1.619	1.647	0.666	0.657	1.703	2.510	0.632	0.552	**0.512**	0.552	0.551	0.551	1.665	0.614	0.593
Noisy Bells (Add)	2.370	0.747	1.371	3.301	0.914	1.250	2.634	3.328	3.721	3.851	3.399	3.357	3.477	1.472	0.722	0.714	**0.436**	0.805	1.027	0.867	4.558	**0.702**	1.112
Noisy Bells (Mult.)	**0.693**	0.718	0.929	1.425	0.858	0.866	1.148	1.360	2.166	2.211	1.187	1.360	2.017	2.819	1.029	0.833	**0.583**	1.187	1.219	1.105	2.155	0.987	1.263
Noisy Spikes (Add)	0.210	**0.189**	0.237	0.344	0.201	0.229	0.229	0.776	0.974	0.995	0.769	0.779	0.996	0.842	0.241	0.203	**0.176**	0.208	0.236	0.239	1.194	0.236	0.277
Noisy Spikes (Mult.)	0.362	**0.362**	**0.358**	0.547	0.389	0.391	0.664	0.777	1.265	1.281	0.746	0.786	1.287	1.717	0.451	0.364	0.445	0.453	0.660	0.629	1.401	0.450	0.742
Noisy Steps (Add)	0.524	**0.357**	0.593	0.958	0.407	0.576	0.608	1.994	2.071	2.030	1.752	1.816	2.031	1.479	0.511	0.417	**0.324**	0.381	0.546	0.457	2.424	0.504	0.746
Noisy Steps (Mult.)	0.497	**0.462**	0.638	0.806	0.551	0.601	0.976	1.190	1.574	1.597	1.204	1.163	1.555	2.123	0.630	0.558	**0.462**	0.636	0.714	0.683	2.052	0.661	1.149
Agg. Relative Score	0.443	0.323	0.495	0.808	0.395	0.450	0.620	1.000	1.534	1.573	0.939	0.968	1.557	1.674	0.324	0.430	**0.201**	0.396	0.471	0.417	1.741	**0.321**	0.595
Avg. Rank	8.188	**4.438**	10.125	15.500	7.000	8.438	12.562	16.812	20.062	20.625	15.750	16.688	20.688	20.625	6.250	6.812	**2.062**	6.875	9.312	7.875	21.688	5.500	12.125

Table 8: WQL scores per model in real datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models						Baselines									
	Chronos-X (FF)	MOMENT-X (FF)	TimesFM-X (FF)	Chronos-X	MOMENT-X	TimesFM-X	TTM-CM	Chronos-Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TPT	DeepAR	PatchTST-X	NBEATSx-G	NBEATSx-I	NBITS	Seasonal Naive	TiDE	DLinear
ETT (15 Min.)	0.022	0.036	0.036	**0.016**	0.026	0.038	0.024	0.016	0.019	0.040	0.043	0.023	0.024	0.020	0.020	0.072	0.035	0.016	0.027	**0.015**	0.065	0.034	0.028
ETT (Hourly)	**0.025**	0.050	0.045	0.025	0.050	0.049	0.041	**0.022**	0.037	0.076	0.068	0.048	0.093	0.045	0.045	0.026	0.036	0.065	0.063	0.035	0.101	0.048	0.030
M5	**0.538**	0.750	0.763	0.557	0.759	0.778	0.824	**0.549**	0.586	0.576	0.759	0.756	0.778	–	0.560	0.617	0.812	0.607	0.604	0.748	0.934	0.595	0.728
Electricity-BE	0.058	0.064	0.111	0.049	0.097	0.067	0.068	0.049	0.069	0.075	0.082	0.079	0.061	0.080	0.087	0.078	0.040	**0.036**	0.054	**0.038**	0.071	0.056	0.085
Electricity-DE	0.455	0.319	0.545	**0.219**	0.313	0.524	0.544	0.315	0.457	0.414	0.595	0.466	0.582	0.372	0.374	0.235	0.304	**0.195**	0.378	0.295	1.048	0.250	0.312
Electricity-FR	0.074	0.102	0.119	0.062	0.109	0.100	0.106	0.060	0.083	0.058	0.152	0.093	0.097	0.061	0.076	0.100	**0.048**	0.061	0.076	**0.041**	0.135	0.082	0.077
Electricity-NP	**0.020**	0.026	0.038	**0.019**	0.034	0.035	0.052	0.025	0.039	0.026	0.033	0.042	0.056	0.028	0.026	0.030	0.022	0.027	0.029	0.034	0.037	0.029	0.023
Electricity-PJM	0.059	0.067	0.050	0.038	0.069	0.054	0.103	0.039	0.062	0.051	0.052	0.038	0.108	0.075	0.027	0.030	0.030	0.027	**0.027**	**0.024**	0.037	0.028	0.040
Bull	0.077	0.268	0.260	**0.062**	0.211	0.154	0.164	0.113	0.171	0.171	0.249	0.155	0.177	0.133	0.095	0.104	0.082	**0.072**	0.073	0.077	0.234	0.112	0.145
Cockatoo	0.054	0.061	0.058	0.040	0.108	0.031	0.042	0.044	**0.029**	0.055	0.089	0.033	0.077	0.060	0.055	0.031	0.085	**0.026**	0.091	0.029	0.068	0.060	0.169
Covid19 (ProEnFo)	0.016	0.017	0.018	0.011	0.018	0.009	**0.006**	0.017	0.008	**0.006**	0.020	0.009	0.012	0.018	0.012	0.007	0.008	0.006	0.007	0.009	0.014	0.008	0.008
GFC12	0.064	0.052	0.062	0.037	0.045	0.043	0.048	**0.034**	0.047	0.040	0.052	0.045	0.045	0.047	0.040	0.056	0.037	0.039	0.042	0.042	0.064	**0.037**	0.039
GFC14	0.015	0.025	0.024	**0.013**	0.029	0.021	0.036	0.015	0.019	0.017	0.030	0.027	0.062	0.032	0.016	**0.012**	0.019	0.019	0.022	0.027	0.031	0.017	0.037
GFC17	0.025	0.040	0.043	0.019	0.035	0.045	0.037	0.035	0.034	0.031	0.042	0.035	0.036	0.021	0.018	0.046	**0.013**	0.021	0.021	0.019	0.031	**0.016**	0.031
Hog	0.031	0.055	0.051	**0.018**	0.053	0.033	0.173	0.024	0.034	0.043	0.050	0.035	0.049	0.037	**0.016**	0.019	0.020	0.020	0.021	0.020	0.022	0.019	0.026
PDB	0.027	0.021	0.027	0.017	0.026	0.018	0.020	**0.015**	0.019	0.019	0.031	0.021	0.052	0.020	0.018	0.025	**0.010**	0.017	0.016	0.020	0.030	0.019	0.020
Spain	0.059	0.039	0.052	0.038	0.029	**0.015**	0.043	0.035	0.039	0.054	0.056	0.043	0.047	0.033	0.060	0.046	**0.023**	0.071	0.038	0.030	0.057	0.031	0.028
Rideshare	**0.177**	1.035	0.241	**0.178**	0.335	0.255	1.157	0.182	0.192	0.194	0.330	0.241	0.245	0.257	0.219	0.190	0.184	0.205	0.190	0.193	0.325	0.215	0.260
Agg. Relative Score	1.174	1.624	1.660	**0.878**	1.571	1.267	1.702	1.000	1.204	1.272	1.812	1.305	1.780	1.286	1.091	1.140	0.932	0.939	1.094	**0.926**	1.752	1.044	1.263
Avg. Rank	10.111	16.167	18.056	**4.500**	16.667	13.500	15.944	**6.667**	11.389	11.500	19.389	13.833	17.500	13.059	9.111	10.778	**6.667**	6.833	9.056	6.778	18.167	8.222	11.556

Table 9: MASE scores per model in real datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

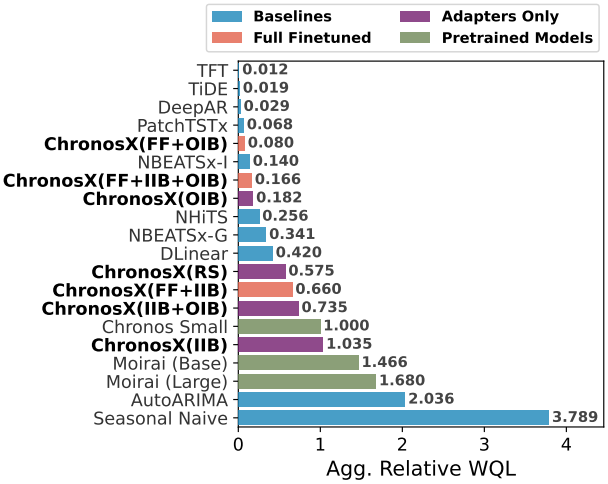
	Full Finetuned			Adapters Only				Pretrained Models						Baselines									
	Chronos-X (FF)	MOMENT-X (FF)	TimesFM-X (FF)	Chronos-X	MOMENT-X	TimesFM-X	TTM-CM	Chronos-Small	Moirai (Base)	Moirai (Large)	MOMENT	TimesFM	TTM	AutoARIMA	TPT	DeepAR	PatchTST-X	NBEATSx-G	NBEATSx-I	NBITS	Seasonal Naive	TiDE	DLinear
ETT (15 Min.)	0.338	0.361	0.368	**0.213**	0.263	0.394	0.258	0.224	0.259	0.513	0.440	0.261	0.254	0.225	0.269	0.790	0.451	**0.208**	0.404	0.239	0.677	0.472	0.377
ETT (Hourly)	0.354	0.475	0.433	0.352	0.490	0.456	0.401	**0.297**	0.489	0.900	0.635	0.437	0.834	0.469	0.599	0.351	0.463	0.752	0.677	0.443	0.950	0.618	**0.311**
M5	**0.907**	1.071	1.104	0.925	1.091	1.151	1.182	0.910	0.959	0.925	1.088	1.064	1.094	–	**0.911**	0.938	1.038	0.946	0.953	1.326	1.268	0.919	1.056
Electricity-BE	0.432	0.363	0.626	0.374	0.547	0.380	0.388	0.380	0.458	0.510	0.465	0.448	0.343	0.306	0.621	0.568	0.292	**0.276**	0.406	**0.276**	0.405	0.414	0.663
Electricity-DE	1.001	0.481	0.823	**0.378**	0.472	0.791	0.821	0.636	0.970	0.820	0.898	0.704	0.878	0.699	0.671	0.458	0.625	**0.381**	0.646	0.526	1.582	0.506	0.539
Electricity-FR	0.676	0.729	0.851	0.570	0.782	0.717	0.762	0.548	0.801	0.520	1.085	0.662	0.695	0.532	0.691	0.881	**0.451**	0.562	0.628	**0.382**	0.966	0.734	0.736
Electricity-NP	**0.548**	0.552	0.817	**0.518**	0.719	0.754	1.115	0.654	1.070	0.674	0.713	0.895	1.200	0.717	0.683	0.769	0.560	0.685	0.682	0.869	0.781	0.816	0.561
Electricity-PJM	0.389	0.312	0.230	0.261	0.318	0.250	0.476	0.267	0.372	0.308	0.240	0.174	0.500	0.194	0.155	0.166	0.176	**0.154**	0.162	**0.147**	0.169	0.169	0.190
Bull	0.830	1.869	1.882	**0.601**	1.486	1.142	1.224	0.990	1.577	1.497	1.755	1.157	1.279	1.248	0.851	0.940	0.783	**0.723**	0.727	0.769	1.681	0.972	1.261
Cockatoo	0.911	0.666	0.634	0.652	1.170	0.342	0.460	0.706	**0.324**	0.797	0.969	0.354	0.836	0.881	0.761	0.442	1.107	**0.308**	1.184	0.471	0.737	0.869	2.200
Covid19 (ProEnFo)	0.407	0.351	0.370	0.269	0.369	0.185	**0.123**	0.411	0.183	**0.141**	0.402	0.179	0.250	0.423	0.330	0.204	0.224	0.157	0.174	0.214	0.276	0.237	0.194
GFC12	1.193	0.801	0.943	0.738	0.711	0.671	0.705	0.657	0.769	0.799	0.809	0.613	**0.581**	0.718	0.738	1.138	**0.599**	0.709	0.667	0.688	0.925	0.640	0.636
GFC14	0.499	0.586	0.553	**0.414**	0.674	0.495	0.835	0.433	0.517	0.501	0.694	0.631	1.439	1.010	0.468	**0.339**	0.524	0.521	0.587	0.693	0.722	0.492	1.053
GFC17	0.525	0.672	0.700	0.474	0.587	0.743	0.622	0.732	0.782	0.644	0.668	0.585	0.614	0.535	0.421	0.904	**0.303**	0.486	0.430	0.413	0.533	**0.395**	0.605
Hog	0.468	0.729	0.630	0.284	0.714	0.389	1.041	0.302	0.378	0.574	0.629	0.450	0.564	0.720	**0.239**	0.261	0.339	0.324	0.317	0.308	**0.227**	0.311	0.397
PDB	0.658	0.376	0.482	0.404	0.470	0.328	0.355	0.355	0.347	0.432	0.563	0.376	0.937	0.497	0.398	0.580	**0.208**	0.352	**0.297**	0.388	0.545	0.395	0.420
Spain	0.834	0.406	0.553	0.589	0.304	**0.155**	0.456	0.454	0.505	0.780	0.595	0.449	0.492	0.390	0.817	0.585	**0.304**	0.850	0.482	0.357	0.606	0.417	0.403
Rideshare	**0.461**	2.007	0.468	0.467	0.650	0.495	2.242	0.472	0.473	0.476	0.640	0.467	0.474	0.630	0.528	0.475	**0.460**	0.497	0.470	0.469	0.630	0.529	0.666
Agg. Relative Score	1.227	1.260	1.287	0.911	1.227	0.984	1.274	1.000	1.123	1.227	1.399	1.010	1.360	1.151	1.062	1.092	0.905	**0.898**	1.016	**0.902**	1.341	1.039	1.172
Avg. Rank	12.833	13.500	15.944	**7.056**	14.944	10.833	13.889	8.500	13.389	13.778	18.000	10.056	15.000	12.912	10.778	12.000	**7.333**	7.500	9.556	7.889	16.472	10.500	12.778

I.3 Evaluations on covariate injection on pretrained models with either IIB, OIB or both.

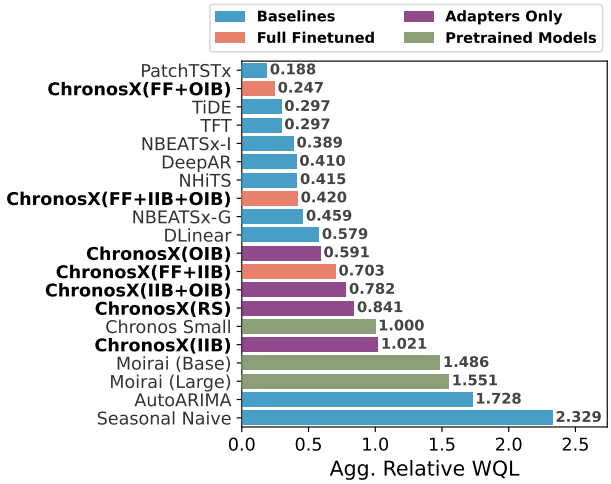
In this section we present results where where ChronosX has either covariates of the past (IIB), covariates of the future (OIB), or both (IIB and OIB).

We present the following results:

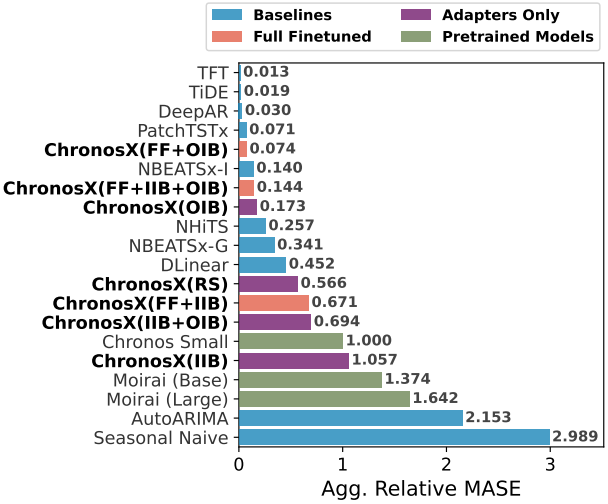
- In Fig. 13 we present the Aggregated Relative WQL and MASE on simple, complex, and real datasets,
- In Table 10 and Table 11 we present the WQL on simple and complex datasets, respectively,
- In Table 12 and Table 13 we present the MASE on simple datasets, respectively.
- In Table 14 and Table 15 we present WQL and MASE on real datasets, respectively.



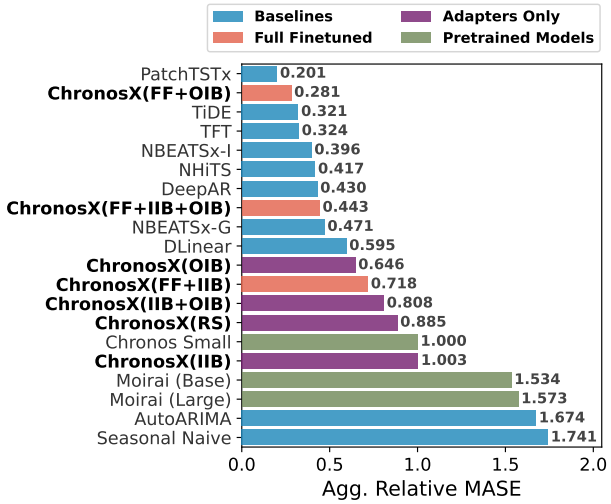
(a) Agg. Rel. WQL on **Simple** synthetic datasets.



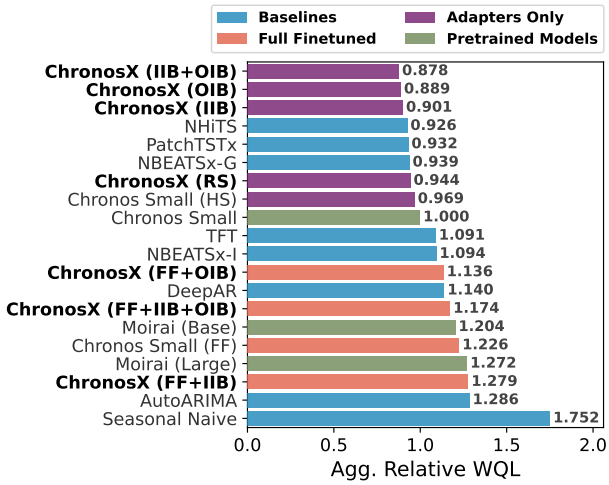
(b) Agg. Rel. WQL on **Complex** synthetic datasets.



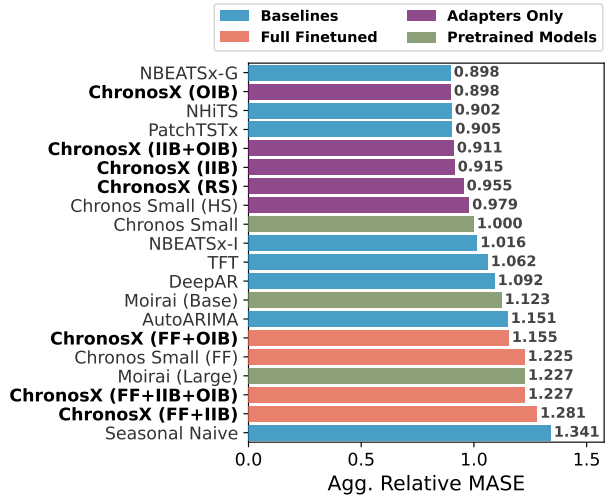
(c) Agg. Rel. MASE on **Simple** synthetic datasets



(d) Agg. Rel. MASE on **Complex** synthetic datasets



(e) Agg. Rel. WQL on real datasets with covariates.



(f) Agg. Rel. MASE on real datasets with covariates.

Figure 13: Aggregated Relative WQL and MASE on simple, complex, and real datasets. Emphasis on ablations of ChronosX with either past covariates, future covariates, or both.

Table 10: WQL scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models			Baselines								
	ChronosX(FF-IB-OB)	ChronosX(FF-OB)	ChronosX(FF-IB)	ChronosX(IB-OB)	ChronosX(OB)	ChronosX(IB)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Simple ARP (Add)	0.164	0.018	0.045	0.168	0.043	0.060	0.055	0.056	0.148	0.099	0.198	0.007	0.007	**0.002**	0.005	0.015	0.012	0.402	**0.003**	0.022
Simple ARP (Mult.)	0.042	0.007	0.028	0.136	0.024	0.041	0.037	0.037	0.153	0.114	0.313	**0.001**	0.004	0.003	0.012	0.018	0.014	0.572	**0.002**	0.030
Simple Bells (Add)	0.287	0.271	0.288	0.225	0.185	0.252	0.318	0.319	0.383	0.346	0.176	0.005	0.056	**0.003**	0.048	0.061	0.037	0.582	**0.005**	0.060
Simple Bells (Mult.)	0.041	0.008	0.040	0.176	0.022	0.083	0.089	0.089	0.156	0.136	0.273	**0.001**	0.005	0.004	0.040	0.037	0.033	0.417	**0.002**	0.093
Simple Spikes (Add)	0.007	0.007	0.338	0.250	0.065	0.458	0.157	0.431	0.515	0.498	0.381	**0.003**	**0.003**	0.004	0.036	0.117	0.083	0.978	0.004	0.066
Simple Spikes (Mult.)	0.007	0.007	0.117	0.022	0.008	0.143	0.026	0.142	0.212	0.196	0.446	**0.001**	**0.002**	0.099	0.014	0.116	0.101	0.440	0.003	0.113
Simple Steps (Add.)	0.007	0.007	0.273	0.185	0.039	0.412	0.165	0.415	0.455	0.458	0.293	**0.004**	0.069	**0.004**	0.067	0.105	0.096	0.835	0.006	0.080
Simple Steps (Mult.)	0.007	0.007	0.078	0.030	0.007	0.116	0.064	0.169	0.275	0.269	0.486	**0.001**	**0.002**	0.143	0.019	0.046	0.032	0.577	0.003	0.169
Single ARP (Add.)	0.130	0.011	0.048	0.175	0.041	0.066	0.061	0.062	0.152	0.106	0.222	0.005	0.007	**0.003**	0.008	0.018	0.013	0.416	**0.004**	0.025
Single ARP (Mult.)	0.038	0.007	0.025	0.157	0.023	0.231	0.037	0.038	0.131	0.109	0.410	**0.001**	0.004	0.004	0.011	0.016	0.012	0.538	**0.002**	0.030
Single Bells (Add.)	0.252	0.159	0.283	0.261	0.188	0.268	0.368	0.367	0.410	0.383	0.181	**0.006**	0.036	**0.003**	0.072	0.098	0.063	0.592	0.006	0.067
Single Bells (Mult.)	0.034	0.008	0.034	0.214	0.022	0.078	0.104	0.105	0.163	0.136	0.289	**0.001**	0.005	0.004	0.040	0.055	0.038	0.461	**0.002**	0.089
Single Spikes (Add.)	0.008	0.008	0.303	0.181	0.051	0.409	0.161	0.391	0.479	0.436	0.310	**0.001**	0.002	0.242	0.025	0.139	0.073	0.954	**0.001**	0.043
Single Spikes (Mult.)	0.008	0.008	0.120	0.031	0.008	0.150	0.031	0.145	0.230	0.197	0.463	**0.001**	**0.001**	0.098	0.012	0.120	0.108	0.462	0.003	0.116
Single Steps (Add.)	0.008	0.008	0.268	0.151	0.020	0.407	0.214	0.390	0.471	0.449	0.262	**0.002**	0.003	**0.002**	0.021	0.062	0.058	0.900	0.003	0.060
Single Steps (Mult.)	0.007	0.007	0.087	0.018	0.007	0.125	0.059	0.186	0.291	0.283	0.622	**0.001**	**0.001**	0.146	0.018	0.050	0.034	0.596	0.003	0.174
Agg. Relative Score	0.166	0.080	0.660	0.735	0.182	1.035	0.575	1.000	1.680	1.466	2.036	**0.012**	0.029	0.068	0.140	0.341	0.256	3.789	**0.019**	0.420
Avg. Rank	9.125	6.000	12.438	13.312	7.562	14.688	12.312	14.875	17.875	16.875	16.938	**1.625**	3.500	5.250	6.750	9.688	8.125	19.812	**2.562**	10.688

Table 11: WQL scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models			Baselines								
	ChronosX(FF-IB-OB)	ChronosX(FF-OB)	ChronosX(FF-IB)	ChronosX(IB-OB)	ChronosX(OB)	ChronosX(IB)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Diverse ARP (Add)	0.092	0.034	0.037	0.117	0.054	0.062	0.063	0.063	0.167	0.138	0.265	0.023	0.032	**0.007**	**0.009**	0.019	0.012	0.250	0.019	0.028
Diverse ARP (Mult.)	0.036	0.031	0.036	0.077	0.048	0.060	0.059	0.059	0.153	0.137	0.269	0.030	0.029	**0.016**	**0.024**	0.028	0.024	0.221	0.027	0.041
Diverse Bells (Add)	0.208	0.192	0.211	0.271	0.210	0.247	0.259	0.257	0.317	0.313	0.168	**0.027**	0.161	**0.011**	1.162	0.105	0.124	0.464	0.035	0.081
Diverse Bells (Mult.)	0.051	**0.020**	0.052	0.116	0.074	0.097	0.088	0.089	0.162	0.155	0.233	0.036	0.081	**0.012**	0.050	0.067	0.055	0.213	0.040	0.084
Diverse Spikes (Add)	0.016	**0.008**	0.195	0.075	0.069	0.246	0.108	0.230	0.288	0.279	0.352	0.019	0.036	**0.006**	0.029	0.046	0.049	0.460	0.015	0.074
Diverse Spikes (Mult.)	0.021	**0.008**	0.071	0.049	0.043	0.093	0.065	0.084	0.134	0.131	0.200	**0.012**	0.027	0.055	0.029	0.068	0.061	0.202	0.021	0.079
Diverse Steps (Add)	0.046	**0.014**	0.194	0.108	0.101	0.276	0.194	0.273	0.296	0.293	0.326	0.028	0.076	**0.007**	0.028	0.061	0.051	0.490	0.018	0.107
Diverse Steps (Mult.)	0.042	**0.013**	0.093	0.087	0.066	0.145	0.123	0.126	0.185	0.188	0.264	**0.033**	0.086	0.036	0.049	0.067	0.054	0.273	0.037	0.119
Noisy ARP (Add)	0.063	0.059	0.063	0.102	0.065	0.077	0.071	0.072	0.169	0.152	0.185	0.059	0.052	**0.048**	**0.049**	0.051	0.051	0.261	0.058	0.053
Noisy ARP (Mult.)	0.064	0.049	0.067	0.077	0.059	0.076	0.066	0.066	0.144	0.134	0.253	0.053	**0.047**	**0.045**	0.047	0.048	0.048	0.219	0.052	0.053
Noisy Bells (Add)	0.198	0.070	0.170	0.265	0.155	0.205	0.243	0.249	0.299	0.290	0.108	0.051	0.052	**0.032**	0.062	0.078	0.066	0.439	**0.051**	0.080
Noisy Bells (Mult.)	0.049	**0.043**	0.050	0.101	0.072	0.093	0.094	0.094	0.163	0.157	0.205	0.072	0.058	**0.042**	0.088	0.091	0.081	0.213	0.068	0.090
Noisy Spikes (Add)	0.050	**0.043**	0.211	0.087	0.082	0.259	0.117	0.239	0.277	0.274	0.208	0.061	0.051	**0.044**	0.053	0.059	0.063	0.433	0.059	0.072
Noisy Spikes (Mult.)	**0.038**	**0.034**	0.083	0.059	0.054	0.104	0.075	0.100	0.149	0.146	0.197	0.048	0.040	0.050	0.050	0.084	0.076	0.216	0.048	0.088
Noisy Steps (Add)	0.065	**0.042**	0.128	0.127	0.110	0.243	0.202	0.277	0.266	0.267	0.197	0.066	0.053	**0.041**	0.049	0.071	0.060	0.425	0.065	0.096
Noisy Steps (Mult.)	0.044	**0.038**	0.062	0.072	0.068	0.117	0.117	0.120	0.158	0.154	0.223	0.057	0.050	**0.042**	0.058	0.065	0.062	0.259	0.060	0.108
Agg. Relative Score	0.420	**0.247**	0.703	0.782	0.591	1.021	0.841	1.000	1.551	1.486	1.728	0.297	0.410	**0.188**	0.389	0.459	0.415	2.329	0.297	0.579
Avg. Rank	6.938	**3.875**	11.188	13.562	10.062	15.000	13.375	14.875	18.000	17.250	17.375	5.250	5.875	**2.062**	5.812	8.000	6.812	19.688	4.750	10.250

Table 12: MASE scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models			Baselines								
	ChronosX(FF-IB-OB)	ChronosX(FF-OB)	ChronosX(FF-IB)	ChronosX(IB-OB)	ChronosX(OB)	ChronosX(IB)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Simple ARP (Add)	0.947	0.139	0.262	0.951	0.261	0.344	0.317	0.322	0.849	0.571	1.103	0.039	0.034	**0.013**	0.029	0.086	0.067	1.788	**0.019**	0.122
Simple ARP (Mult.)	0.161	0.033	0.127	0.500	0.110	0.176	0.156	0.157	0.642	0.454	1.616	**0.006**	0.020	0.014	0.051	0.070	0.057	1.930	**0.009**	0.128
Simple Bells (Add)	3.203	3.044	3.227	2.459	2.119	2.805	3.474	3.483	4.069	3.689	1.993	**0.055**	0.613	**0.035**	0.491	0.641	0.392	4.951	0.059	0.682
Simple Bells (Mult.)	0.220	0.046	0.214	0.858	0.130	0.462	0.472	0.473	0.847	0.691	1.698	**0.004**	0.025	0.019	0.212	0.206	0.177	1.882	**0.012**	0.583
Simple Spikes (Add)	0.010	0.010	0.515	0.372	0.101	0.733	0.292	0.653	0.823	0.756	0.644	0.013	0.009	**0.007**	0.071	0.231	0.146	1.258	**0.009**	0.162
Simple Spikes (Mult.)	0.024	0.024	0.469	0.089	0.024	0.600	0.093	0.581	0.781	0.664	1.930	**0.004**	**0.006**	0.456	0.054	0.458	0.453	1.397	0.009	0.490
Simple Steps (Add)	0.023	**0.022**	1.163	0.742	0.141	1.691	0.780	1.653	1.862	1.856	1.079	0.033	0.077	**0.018**	0.337	0.452	0.422	2.577	0.029	0.411
Simple Steps (Mult.)	0.023	0.023	0.306	0.113	0.023	0.490	0.254	0.720	1.067	1.014	2.262	**0.008**	**0.007**	0.703	0.076	0.177	0.124	1.808	0.009	0.735
Single ARP (Add.)	0.601	0.058	0.275	0.811	0.232	0.358	0.340	0.344	0.823	0.548	1.200	0.027	0.035	**0.015**	0.040	0.093	0.072	1.774	**0.020**	0.137
Single ARP (Mult.)	0.148	0.034	0.111	0.593	0.106	1.084	0.162	0.166	0.540	0.418	1.914	**0.004**	0.019	0.016	0.048	0.069	0.053	1.877	**0.009**	0.128
Single Bells (Add.)	2.675	1.759	3.126	2.599	1.946	2.762	3.735	3.728	4.079	3.835	1.945	**0.054**	0.346	**0.026**	0.687	0.965	0.616	4.779	0.067	0.673
Single Bells (Mult.)	0.165	0.046	0.168	1.122	0.127	0.426	0.539	0.539	0.851	0.684	1.611	**0.003**	0.025	0.019	0.214	0.299	0.208	1.989	**0.012**	0.534
Single Spikes (Add.)	0.010	0.010	0.496	0.309	0.078	0.722	0.264	0.653	0.826	0.690	0.540	**0.002**	0.002	0.470	0.040	0.235	0.112	1.242	**0.001**	0.072
Single Spikes (Mult.)	0.023	0.023	0.466	0.115	0.023	0.586	0.089	0.575	0.831	0.629	1.909	**0.002**	**0.004**	0.451	0.046	0.453	0.449	1.415	0.009	0.482
Single Steps (Add.)	0.019	0.019	0.969	0.514	0.059	1.367	0.702	1.273	1.566	1.501	0.886	**0.005**	0.009	**0.008**	0.064	0.196	0.186	2.281	0.009	0.197
Single Steps (Mult.)	0.023	0.023	0.361	0.061	0.023	0.486	0.215	0.735	1.067	0.998	2.672	**0.004**	**0.004**	0.684	0.068	0.192	0.127	1.841	0.009	0.714
Agg. Relative Score	0.144	0.074	0.671	0.694	0.173	1.057	0.566	1.000	1.642	1.374	2.153	**0.013**	0.030	0.071	0.140	0.341	0.257	2.989	**0.019**	0.452
Avg. Rank	8.844	5.844	12.375	13.062	7.562	14.875	12.312	15.000	18.000	16.750	17.000	**3.250**	3.500	5.188	6.812	9.438	7.938	19.688	**2.500**	11.062

Table 13: MASE scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only				Pretrained Models			Baselines								
	ChronosX(FF-IB-OB)	ChronosX(FF-OB)	ChronosX(FF-IB)	ChronosX(IB-OB)	ChronosX(OB)	ChronosX(IB)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-I	NBEATSx-G	NHiTS	Seasonal Naive	TiDE	DLinear
Diverse ARP (Add)	1.004	0.461	0.488	1.282	0.687	0.758	0.769	0.768	2.056	1.848	2.801	0.317	0.419	**0.099**	**0.110**	0.242	0.148	1.947	0.270	0.318
Diverse ARP (Mult.)	0.500	0.446	0.484	0.987	0.655	0.754	0.767	0.772	1.928	1.791	3.091	0.404	0.370	**0.221**	**0.314**	0.374	0.323	1.921	0.360	0.533
Diverse Bells (Add)	3.392	3.178	3.508	4.066	3.378	3.838	4.000	3.975	4.634	4.713	2.234	**0.446**	2.513	**0.169**	12.561	1.584	1.512	5.563	0.586	1.249
Diverse Bells (Mult.)	0.925	**0.456**	0.929	1.959	1.414	1.731	1.591	1.620	2.834	2.755	3.453	0.651	1.462	**0.249**	0.920	1.171	1.000	2.698	0.721	1.619
Diverse Spikes (Add)	0.962	**0.036**	0.681	0.294	0.283	0.808	0.439	0.767	1.048	0.999	1.203	0.077	0.132	**0.022**	0.107	0.176	0.163	1.286	0.059	0.277
Diverse Spikes (Mult.)	0.173	**0.072**	0.587	0.451	0.410	0.786	0.604	0.691	1.247	1.201	1.826	**0.121**	0.240	0.491	0.264	0.537	0.516	1.427	0.199	0.682
Diverse Steps (Add)	0.393	**0.124**	1.624	0.896	0.870	2.196	1.583	2.099	2.387	2.388	2.182	0.252	0.584	**0.056**	0.222	0.490	0.409	2.851	0.155	0.837
Diverse Steps (Mult.)	0.495	**0.185**	1.091	1.054	0.847	1.508	1.446	1.464	2.110	2.195	3.165	0.404	1.021	**0.401**	0.581	0.818	0.644	2.329	0.448	1.396
Noisy ARP (Add)	0.613	0.577	0.614	0.967	0.642	0.747	0.703	0.713	1.580	1.473	1.760	0.570	0.495	**0.459**	**0.464**	0.483	0.477	1.713	0.558	0.500
Noisy ARP (Mult.)	0.750	0.583	0.750	0.884	0.716	0.779	0.783	0.778	1.647	1.619	2.510	0.632	0.552	**0.512**	0.552	0.551	**0.551**	1.665	0.614	0.593
Noisy Bells (Add)	2.370	0.962	2.089	3.301	2.125	2.689	3.274	3.328	3.851	3.721	1.472	0.722	0.714	**0.436**	0.805	1.027	0.867	4.558	**0.702**	1.112
Noisy Bells (Mult.)	0.693	**0.626**	0.704	1.425	1.079	1.333	1.369	1.360	2.211	2.166	2.819	1.029	0.833	**0.583**	1.187	1.219	1.105	2.155	0.987	1.263
Noisy Spikes (Add)	0.210	**0.184**	0.675	0.344	0.329	0.822	0.456	0.776	0.995	0.974	0.842	0.241	0.203	**0.176**	0.208	0.236	0.239	1.194	0.236	0.277
Noisy Spikes (Mult.)	**0.362**	**0.335**	0.655	0.547	0.501	0.815	0.679	0.777	1.281	1.265	1.717	0.451	0.364	0.445	0.453	0.660	0.629	1.401	0.450	0.742
Noisy Steps (Add)	0.524	**0.347**	0.971	0.958	0.865	1.752	1.540	1.994	2.030	2.071	1.479	0.511	0.417	**0.324**	0.381	0.546	0.457	2.424	0.504	0.746
Noisy Steps (Mult.)	0.497	**0.447**	0.694	0.806	0.741	1.192	1.172	1.190	1.597	1.574	2.123	0.630	0.558	**0.462**	0.636	0.714	0.683	2.052	0.661	1.149
Agg. Relative Score	0.443	**0.281**	0.718	0.808	0.646	1.003	0.885	1.000	1.573	1.534	1.674	0.324	0.430	**0.201**	0.396	0.471	0.417	1.741	0.321	0.595
Avg. Rank	7.188	**4.000**	10.938	13.500	10.250	14.875	13.812	14.750	18.250	17.625	17.688	5.500	5.938	**1.812**	5.750	7.812	6.375	18.938	4.812	10.188

Table 14: WQL scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned				Adapters Only					Pretrained Models			Baselines							
	ChronosX (FF-IB-OB)	ChronosX (FF-OB)	ChronosX (FF-IB)	Chronos Small (FF)	ChronosX (IB-OB)	ChronosX (OB)	ChronosX (IB)	ChronosX (BS)	Chronos Small (BS)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTST	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.022	0.020	0.026	0.023	0.016	0.017	0.016	**0.015**	0.017	0.016	0.019	0.040	0.020	0.020	0.072	0.035	0.016	0.027	**0.015**	0.065
ETT (Hourly)	0.025	0.024	0.025	0.024	0.025	**0.023**	0.024	0.024	0.023	**0.022**	0.037	0.076	0.045	0.045	0.026	0.036	0.065	0.063	0.035	0.101
M5	0.538	0.538	**0.537**	**0.537**	0.557	0.551	0.552	0.551	0.549	0.549	0.586	0.576	-	0.560	0.617	0.812	0.607	0.604	0.748	0.934
Electricity-BE	0.058	0.056	0.058	0.054	0.049	0.047	0.051	0.047	0.047	0.049	0.069	0.075	0.080	0.087	0.078	0.040	**0.036**	0.054	**0.038**	0.071
Electricity-DE	0.455	0.620	0.470	0.729	**0.219**	0.262	0.254	0.245	0.304	0.315	0.457	0.414	0.372	0.374	0.235	0.304	**0.195**	0.378	0.295	1.048
Electricity-FR	0.074	0.072	0.072	0.065	0.062	0.062	0.050	0.059	0.061	0.060	0.083	0.058	0.061	0.076	0.100	**0.048**	0.061	0.076	**0.041**	0.135
Electricity-NP	0.020	**0.019**	0.022	0.025	**0.019**	0.020	0.021	0.022	0.023	0.025	0.039	0.026	0.028	0.026	0.030	0.022	0.027	0.029	0.034	0.037
Electricity-PJM	0.059	0.053	0.067	0.069	0.038	0.039	0.033	0.042	0.043	0.039	0.062	0.051	0.075	0.077	0.030	0.030	0.027	**0.027**	**0.024**	0.037
Bull	0.077	0.064	0.080	0.155	**0.062**	**0.061**	0.071	0.092	0.115	0.113	0.171	0.171	0.133	0.095	0.104	0.082	0.072	0.073	0.077	0.234
Cockatoo	0.054	0.063	0.170	0.039	0.040	**0.026**	0.032	0.033	0.034	0.044	0.029	0.055	0.060	0.055	0.031	0.085	**0.026**	0.091	0.029	0.068
Covid19 (ProEnFo)	0.016	0.015	0.017	0.017	0.011	0.012	0.015	0.016	0.013	0.017	0.008	**0.006**	0.018	0.012	0.007	0.008	**0.006**	0.007	0.009	0.014
GFC12	0.064	0.048	0.051	0.058	0.037	**0.033**	0.035	0.033	**0.033**	0.034	0.047	0.040	0.047	0.040	0.056	0.037	0.039	0.042	0.042	0.064
GFC14	0.015	0.017	0.015	0.016	**0.013**	0.014	0.014	0.015	0.014	0.015	0.019	0.017	0.032	0.016	**0.012**	0.019	0.019	0.022	0.027	0.031
GFC17	0.025	0.033	0.031	0.038	0.019	0.033	0.031	0.034	0.029	0.035	0.034	0.031	0.021	**0.018**	0.046	**0.013**	0.021	0.021	0.019	0.031
Hog	0.031	0.034	0.042	0.031	**0.018**	0.022	0.019	0.020	0.021	0.024	0.034	0.043	0.037	**0.016**	0.019	0.020	0.020	0.021	0.020	0.022
PDB	0.027	0.028	0.029	0.026	0.017	**0.013**	0.015	0.015	0.016	0.015	0.019	0.019	0.020	0.018	0.025	**0.010**	0.017	0.016	0.020	0.030
Spain	0.059	**0.030**	0.039	0.034	0.038	0.034	0.033	0.034	0.035	0.035	0.039	0.054	0.033	0.060	0.046	**0.023**	0.071	0.038	0.030	0.057
Rideshare	**0.177**	**0.177**	0.177	0.178	0.178	0.179	0.191	0.179	0.177	0.182	0.192	0.194	0.257	0.219	0.190	0.184	0.205	0.190	0.193	0.325
Agg. Relative Score	1.174	1.136	1.279	1.226	**0.878**	**0.889**	0.901	0.944	0.969	1.000	1.204	1.272	1.286	1.091	1.140	0.932	0.939	1.094	0.926	1.752
Avg. Rank	11.778	10.778	12.722	12.500	**6.278**	**6.333**	6.500	7.389	7.833	9.056	13.944	13.389	14.353	11.556	11.944	8.000	8.556	11.389	8.222	17.167

Table 15: MASE scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

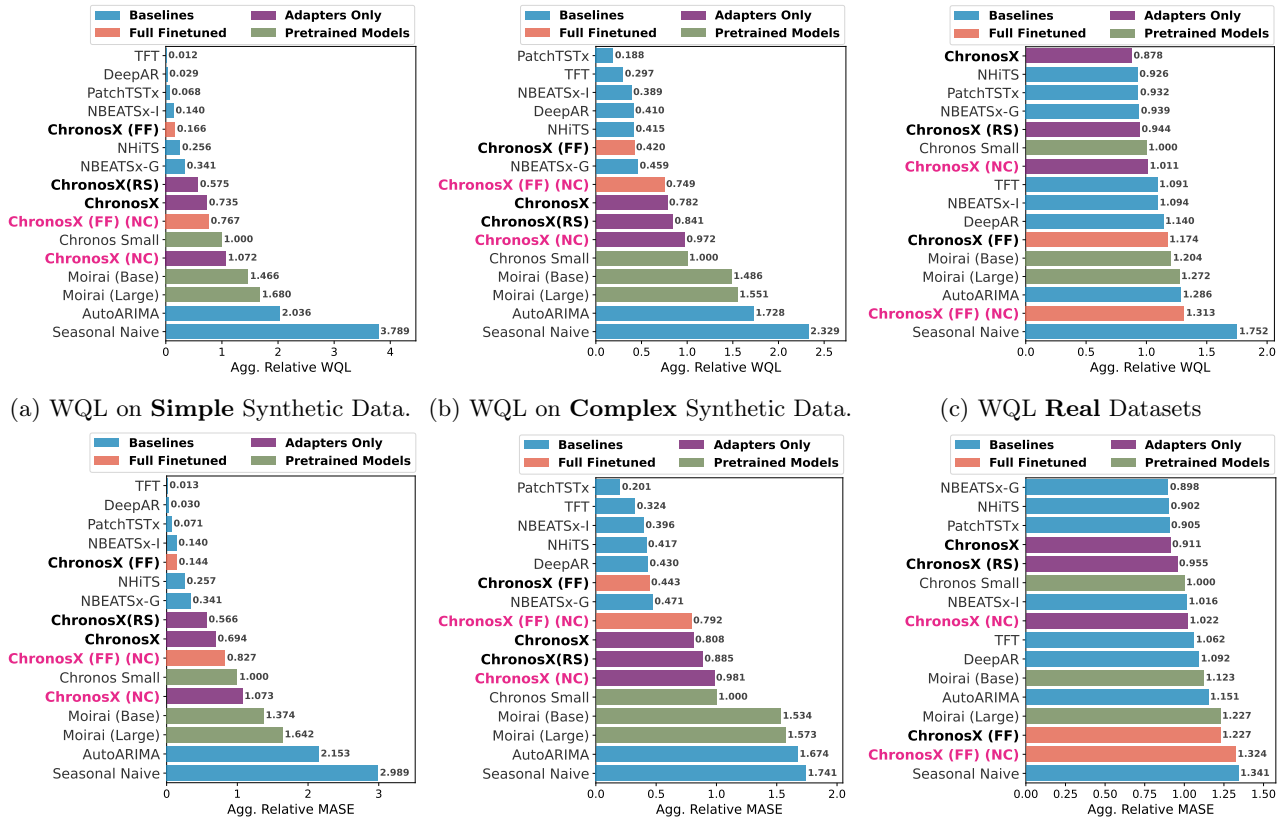
	Full Finetuned				Adapters Only					Pretrained Models			Baselines							
	ChronosX (FF-IB-OB)	ChronosX (FF-OB)	ChronosX (FF-IB)	Chronos Small (FF)	ChronosX (IB-OB)	ChronosX (OB)	ChronosX (IB)	ChronosX (BS)	Chronos Small (BS)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTST	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.338	0.297	0.352	0.337	0.213	**0.207**	0.212	**0.206**	0.211	0.224	0.259	0.513	0.225	0.269	0.790	0.451	0.208	0.404	0.239	0.677
ETT (Hourly)	0.354	0.341	0.346	0.333	0.352	0.331	0.331	0.338	**0.313**	**0.297**	0.489	0.900	0.469	0.599	0.351	0.463	0.752	0.677	0.443	0.950
M5	0.907	0.907	**0.904**	**0.903**	0.925	0.918	0.925	0.918	0.915	0.919	0.959	0.925	-	0.911	0.938	1.038	0.946	0.953	1.226	1.268
Electricity-BE	0.432	0.413	0.435	0.400	0.374	0.373	0.378	0.372	0.374	0.380	0.458	0.510	0.306	0.621	0.568	0.292	**0.276**	0.406	**0.276**	0.405
Electricity-DE	1.001	1.391	1.011	1.489	**0.378**	0.518	0.550	0.466	0.782	0.636	0.970	0.820	0.699	0.671	0.458	0.625	**0.381**	0.646	0.526	1.582
Electricity-FR	0.676	0.646	0.648	0.577	0.570	0.561	0.463	0.537	0.553	0.548	0.801	0.520	0.532	0.691	0.881	**0.451**	0.562	0.628	**0.382**	0.966
Electricity-NP	0.548	**0.546**	0.608	0.661	**0.518**	0.553	0.536	0.595	0.631	0.654	1.070	0.674	0.717	0.683	0.769	0.560	0.685	0.682	0.869	0.781
Electricity-PJM	0.389	0.365	0.438	0.445	0.261	0.258	0.216	0.281	0.285	0.267	0.372	0.308	0.194	0.155	0.166	0.176	**0.154**	0.162	**0.147**	0.169
Bull	0.830	0.623	0.850	1.463	**0.601**	**0.546**	0.675	0.840	1.027	0.990	1.577	1.497	1.248	0.851	0.940	0.783	0.723	0.727	0.769	1.681
Cockatoo	0.911	0.881	2.064	0.500	0.652	0.382	0.500	0.558	0.484	0.706	**0.324**	0.797	0.881	0.761	0.442	1.107	**0.308**	1.184	0.471	0.737
Covid19 (ProEnFo)	0.407	0.389	0.421	0.425	0.269	0.320	0.371	0.407	0.348	0.411	0.183	**0.141**	0.423	0.330	0.204	0.224	**0.157**	0.174	0.214	0.276
GFC12	1.193	0.923	1.009	1.027	0.738	**0.627**	0.733	0.644	0.638	0.657	0.769	0.799	0.718	0.738	1.138	**0.599**	0.709	0.667	0.688	0.925
GFC14	0.499	0.546	0.471	0.506	**0.414**	0.435	0.481	0.466	0.456	0.433	0.517	0.501	1.010	0.468	**0.339**	0.524	0.521	0.587	0.693	0.722
GFC17	0.525	0.650	0.630	0.792	0.474	0.700	0.617	0.717	0.620	0.732	0.782	0.644	0.535	0.421	0.904	**0.303**	0.486	0.430	**0.413**	0.533
Hog	0.468	0.402	0.575	0.377	0.284	0.289	0.287	0.277	0.286	0.302	0.378	0.574	0.720	**0.239**	0.261	0.339	0.324	0.317	0.308	**0.227**
PDB	0.658	0.742	0.753	0.656	0.404	0.344	0.377	0.366	0.396	0.355	0.347	0.432	0.497	0.398	0.580	**0.208**	0.352	**0.297**	0.388	0.545
Spain	0.834	0.402	0.452	0.498	0.589	0.487	0.365	0.438	0.462	0.454	0.505	0.780	0.390	0.817	0.585	**0.394**	0.850	0.482	**0.357**	0.606
Rideshare	0.461	**0.460**	0.462	0.463	0.467	0.466	0.498	0.467	0.463	0.472	0.473	0.476	0.630	0.528	0.475	**0.460**	0.497	0.470	0.469	0.630
Agg. Relative Score	1.227	1.155	1.281	1.225	0.911	**0.898**	0.915	0.955	0.979	1.000	1.123	1.227	1.151	1.062	1.092	0.905	**0.898**	1.016	0.902	1.341
Avg. Rank	12.944	11.444	13.111	13.056	7.667	**6.167**	7.667	**7.500**	8.111	9.278	13.444	13.611	12.559	11.333	11.889	7.722	8.556	10.556	8.056	14.917

I.4 Ablations showing that performance improvement is brought by covariates

In this section we present results where ChronosX (NC) and ChronosX (FF) (NC) are variants with no-covariates of our models ChronosX and ChronosX (FF), respectively.

We present the following results:

- In Fig. 14 we present the Aggregated Relative WQL and MASE on simple, complex, and real datasets.
- In Table 16 and Table 17 we present the WQL on simple and complex datasets, respectively,
- In Table 18 and Table 19 we present the MASE on simple datasets, respectively.
- In Table 20 and Table 21 we present WQL and MASE on real datasets, respectively.



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets

(d) MASE on **Simple** Synthetic Data. (e) MASE on **Complex** Synthetic Data. (f) MASE on **Real** Datasets

Figure 14: Covariate ablation study. ChronosX (NC) and ChronosX (FF) (NC) are variants with no-covariates of our models ChronosX and ChronosX (FF), respectively. Overall model ablations with no-covariates perform worse.

Table 16: WQL scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	NHiTS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.164	0.045	0.168	0.071	0.055	0.056	0.148	0.099	0.012	0.198	0.007	**0.005**	0.015	0.007	**0.002**	0.402
Simple ARP (Mult.)	0.042	0.027	0.136	0.051	0.037	0.037	0.153	0.114	0.014	0.313	**0.001**	0.012	0.018	0.004	**0.003**	0.572
Simple Bells (Add)	0.287	0.292	0.225	0.347	0.318	0.319	0.383	0.346	0.037	0.176	**0.005**	0.048	0.061	0.056	**0.003**	0.582
Simple Bells (Mult.)	0.041	0.062	0.176	0.085	0.089	0.089	0.156	0.136	0.033	0.273	**0.001**	0.040	0.037	0.005	**0.004**	0.417
Simple Spikes (Add)	0.007	0.338	0.250	0.449	0.157	0.431	0.515	0.498	0.083	0.381	**0.003**	0.036	0.117	**0.003**	0.004	0.978
Simple Spikes (Mult.)	0.007	0.117	0.022	0.130	0.026	0.142	0.212	0.196	0.101	0.446	**0.001**	0.014	0.116	**0.002**	0.099	0.440
Simple Steps (Add)	0.007	0.314	0.185	0.464	0.165	0.415	0.455	0.458	0.096	0.293	**0.004**	0.067	0.105	0.009	**0.004**	0.835
Simple Steps (Mult.)	0.007	0.126	0.030	0.140	0.064	0.169	0.275	0.269	0.032	0.486	**0.001**	0.019	0.046	**0.002**	0.143	0.577
Single ARP (Add.)	0.130	0.051	0.175	0.067	0.061	0.062	0.152	0.106	0.013	0.222	**0.005**	0.008	0.018	0.007	**0.003**	0.416
Single ARP (Mult.)	0.038	0.028	0.157	0.044	0.037	0.038	0.131	0.109	0.012	0.410	**0.001**	0.011	0.016	0.004	**0.004**	0.538
Single Bells (Add.)	0.252	0.282	0.261	0.385	0.368	0.367	0.410	0.383	0.063	0.181	**0.006**	0.072	0.098	0.036	**0.003**	0.592
Single Bells (Mult.)	0.034	0.068	0.214	0.118	0.104	0.105	0.163	0.136	0.038	0.289	**0.001**	0.040	0.055	0.005	**0.004**	0.461
Single Spikes (Add.)	0.008	0.302	0.181	0.485	0.161	0.391	0.479	0.436	0.073	0.310	**0.001**	0.025	0.139	**0.002**	0.242	0.954
Single Spikes (Mult.)	0.008	0.120	0.031	0.151	0.031	0.145	0.230	0.197	0.108	0.463	**0.001**	0.012	0.120	**0.001**	0.098	0.462
Single Steps (Add.)	0.008	0.297	0.151	0.462	0.214	0.390	0.471	0.449	0.058	0.262	**0.002**	0.021	0.062	0.003	**0.002**	0.900
Single Steps (Mult.)	0.007	0.131	0.018	0.156	0.059	0.186	0.291	0.283	0.034	0.622	**0.001**	0.018	0.050	**0.001**	0.146	0.596
Agg. Relative Score	0.166	0.767	0.735	1.072	0.575	1.000	1.680	1.466	0.256	2.036	**0.012**	0.140	0.341	**0.029**	0.068	3.789
Avg. Rank	6.188	9.000	9.688	11.938	8.438	10.812	13.812	12.625	5.438	13.062	**1.438**	4.375	6.688	**2.750**	3.938	15.812

Table 17: WQL scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.asdf

	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	NHiTS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	0.092	0.039	0.117	0.062	0.063	0.063	0.167	0.138	0.012	0.265	0.023	**0.009**	0.019	0.032	**0.007**	0.250
Diverse ARP (Mult.)	0.036	0.036	0.077	0.058	0.059	0.059	0.153	0.137	0.024	0.269	0.030	**0.024**	0.028	0.029	**0.016**	0.221
Diverse Bells (Add)	0.208	0.220	0.271	0.271	0.259	0.257	0.317	0.313	0.124	0.168	**0.027**	1.162	0.105	0.161	**0.011**	0.464
Diverse Bells (Mult.)	0.051	0.060	0.116	0.093	0.088	0.089	0.162	0.155	0.055	0.233	**0.036**	0.050	0.067	0.081	**0.012**	0.213
Diverse Spikes (Add)	**0.016**	0.180	0.075	0.219	0.108	0.230	0.288	0.279	0.049	0.352	0.019	0.029	0.046	0.036	**0.006**	0.460
Diverse Spikes (Mult.)	**0.021**	0.064	0.049	0.080	0.065	0.084	0.134	0.131	0.061	0.200	**0.012**	0.029	0.068	0.027	0.055	0.202
Diverse Steps (Add)	0.046	0.191	0.108	0.242	0.194	0.273	0.296	0.293	0.051	0.326	**0.028**	0.028	0.061	0.076	**0.007**	0.490
Diverse Steps (Mult.)	0.042	0.083	0.087	0.119	0.123	0.126	0.185	0.188	0.054	0.264	**0.033**	0.049	0.067	0.086	**0.036**	0.273
Noisy ARP (Add)	0.063	0.061	0.102	0.073	0.071	0.072	0.169	0.152	0.051	0.185	0.059	**0.049**	0.051	0.052	**0.048**	0.261
Noisy ARP (Mult.)	0.064	0.050	0.077	0.065	0.066	0.066	0.144	0.134	0.048	0.253	0.053	0.047	0.048	**0.047**	**0.045**	0.219
Noisy Bells (Add)	0.198	0.230	0.265	0.259	0.243	0.249	0.299	0.290	0.066	0.108	**0.051**	0.062	0.078	0.052	**0.032**	0.439
Noisy Bells (Mult.)	**0.049**	0.077	0.101	0.093	0.094	0.094	0.163	0.157	0.081	0.205	0.072	0.088	0.091	0.058	**0.042**	0.213
Noisy Spikes (Add)	**0.050**	0.194	0.087	0.234	0.117	0.239	0.277	0.274	0.063	0.208	0.061	0.053	0.059	0.051	**0.044**	0.433
Noisy Spikes (Mult.)	**0.038**	0.081	0.059	0.095	0.075	0.100	0.149	0.146	0.076	0.197	0.048	0.050	0.084	**0.040**	0.050	0.216
Noisy Steps (Add)	0.065	0.191	0.127	0.253	0.202	0.277	0.266	0.267	0.060	0.197	0.066	**0.049**	0.071	0.053	**0.041**	0.425
Noisy Steps (Mult.)	**0.044**	0.089	0.072	0.109	0.117	0.120	0.158	0.154	0.062	0.223	0.057	0.058	0.065	0.050	**0.042**	0.259
Agg. Relative Score	0.420	0.749	0.782	0.972	0.841	1.000	1.551	1.486	0.415	1.728	**0.297**	0.389	0.459	0.410	**0.188**	2.329
Avg. Rank	4.750	7.938	9.812	10.500	9.750	11.312	14.000	13.250	5.125	13.625	**3.750**	4.375	6.125	4.438	**1.562**	15.688

Table 18: MASE scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	NHiTS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.947	0.268	0.951	0.403	0.317	0.322	0.849	0.571	0.067	1.103	0.039	**0.029**	0.086	0.034	**0.013**	1.788
Simple ARP (Mult.)	0.161	0.126	0.500	0.200	0.156	0.157	0.642	0.454	0.057	1.616	**0.006**	0.051	0.070	0.020	**0.014**	1.930
Simple Bells (Add)	3.203	3.255	2.459	3.717	3.474	3.483	4.069	3.689	0.392	1.993	**0.055**	0.491	0.641	0.613	**0.035**	4.951
Simple Bells (Mult.)	0.220	0.377	0.858	0.485	0.472	0.473	0.847	0.691	0.177	1.698	**0.004**	0.212	0.206	0.025	**0.019**	1.882
Simple Spikes (Add)	0.010	0.515	0.372	0.676	0.292	0.653	0.823	0.756	0.146	0.644	0.013	0.071	0.231	**0.009**	**0.007**	1.258
Simple Spikes (Mult.)	0.024	0.469	0.089	0.486	0.093	0.581	0.781	0.664	0.453	1.930	**0.004**	0.054	0.458	**0.006**	0.456	1.397
Simple Steps (Add)	**0.023**	1.424	0.742	1.685	0.780	1.653	1.862	1.856	0.422	1.079	0.033	0.337	0.452	0.077	**0.018**	2.577
Simple Steps (Mult.)	0.023	0.641	0.113	0.686	0.254	0.720	1.067	1.014	0.124	2.262	**0.008**	0.076	0.177	**0.007**	0.703	1.808
Single ARP (Add.)	0.601	0.294	0.811	0.371	0.340	0.344	0.823	0.548	0.072	1.200	**0.027**	0.040	0.093	0.035	**0.015**	1.774
Single ARP (Mult.)	0.148	0.128	0.593	0.185	0.162	0.166	0.540	0.418	0.053	1.914	**0.004**	0.048	0.069	0.019	**0.016**	1.877
Single Bells (Add.)	2.675	3.107	2.599	3.868	3.735	3.728	4.079	3.835	0.616	1.945	**0.054**	0.687	0.965	0.346	**0.026**	4.779
Single Bells (Mult.)	0.165	0.402	1.122	0.613	0.539	0.539	0.851	0.684	0.208	1.611	**0.003**	0.214	0.299	0.025	**0.019**	1.989
Single Spikes (Add.)	0.010	0.496	0.309	0.810	0.264	0.653	0.826	0.690	0.112	0.540	**0.002**	0.040	0.235	**0.002**	0.470	1.242
Single Spikes (Mult.)	0.023	0.466	0.115	0.585	0.089	0.575	0.831	0.629	0.449	1.909	**0.002**	0.046	0.453	**0.004**	0.451	1.415
Single Steps (Add.)	0.019	1.172	0.514	1.467	0.702	1.273	1.566	1.501	0.186	0.886	**0.005**	0.064	0.196	0.009	**0.008**	2.281
Single Steps (Mult.)	0.023	0.626	0.061	0.742	0.215	0.735	1.067	0.998	0.127	2.672	**0.004**	0.068	0.192	**0.004**	0.684	1.841
Agg. Relative Score	0.144	0.827	0.694	1.073	0.566	1.000	1.642	1.374	0.257	2.153	**0.013**	0.140	0.341	**0.030**	0.071	2.989
Avg. Rank	6.000	9.000	9.438	11.875	8.375	10.875	14.062	12.688	5.312	13.188	**1.812**	4.438	6.688	**2.688**	3.875	15.688

Table 19: MASE scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX(RS)	Chronos Small	Moirai (Large)	Moirai (Base)	NHITS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	1.004	0.513	1.282	0.769	0.769	0.768	2.056	1.848	0.148	2.801	0.317	**0.110**	0.242	0.419	**0.099**	1.947
Diverse ARP (Mult.)	0.500	0.491	0.987	0.740	0.767	0.772	1.928	1.791	0.323	3.091	0.404	**0.314**	0.374	0.370	**0.221**	1.921
Diverse Bells (Add)	3.392	3.608	4.066	4.079	4.000	3.975	4.634	4.713	1.512	2.234	**0.446**	12.561	1.584	2.513	**0.169**	5.563
Diverse Bells (Mult.)	0.925	1.191	1.959	1.661	1.591	1.620	2.834	2.755	1.000	3.453	**0.651**	0.920	1.171	1.462	**0.249**	2.698
Diverse Spikes (Add)	**0.062**	0.599	0.294	0.733	0.439	0.767	1.048	0.999	0.163	1.203	0.077	0.107	0.176	0.132	**0.022**	1.286
Diverse Spikes (Mult.)	**0.173**	0.508	0.451	0.676	0.604	0.691	1.247	1.201	0.516	1.826	**0.121**	0.264	0.537	0.240	0.491	1.427
Diverse Steps (Add)	0.393	1.761	0.896	2.036	1.583	2.099	2.387	2.388	0.409	2.182	0.252	**0.222**	0.490	0.584	**0.056**	2.851
Diverse Steps (Mult.)	0.495	1.125	1.054	1.386	1.446	1.464	2.110	2.195	0.644	3.165	**0.404**	0.581	0.818	1.021	**0.401**	2.329
Noisy ARP (Add)	0.613	0.596	0.967	0.715	0.703	0.713	1.580	1.473	0.477	1.760	0.570	**0.464**	0.483	0.495	**0.459**	1.713
Noisy ARP (Mult.)	0.750	0.592	0.884	0.767	0.783	0.778	1.647	1.619	**0.551**	2.510	0.632	0.552	0.551	0.552	**0.512**	1.665
Noisy Bells (Add)	2.370	3.193	3.301	3.441	3.274	3.328	3.851	3.721	0.867	1.472	0.722	0.805	1.027	**0.714**	**0.436**	4.558
Noisy Bells (Mult.)	**0.693**	1.114	1.425	1.334	1.369	1.360	2.211	2.166	1.105	2.819	1.029	1.187	1.219	0.833	**0.583**	2.155
Noisy Spikes (Add)	0.210	0.643	0.344	0.748	0.456	0.776	0.995	0.974	0.239	0.842	0.241	0.208	0.236	**0.203**	**0.176**	1.194
Noisy Spikes (Mult.)	**0.362**	0.640	0.547	0.756	0.679	0.777	1.281	1.265	0.629	1.717	0.451	0.453	0.660	**0.364**	0.445	1.401
Noisy Steps (Add)	0.524	1.552	0.958	1.898	1.540	1.994	2.030	2.071	0.457	1.479	0.511	**0.381**	0.546	0.417	**0.324**	2.424
Noisy Steps (Mult.)	**0.497**	1.017	0.806	1.128	1.172	1.190	1.597	1.574	0.683	2.123	0.630	0.636	0.714	0.558	**0.462**	2.052
Agg. Relative Score	0.443	0.792	0.808	0.981	0.885	1.000	1.573	1.534	0.417	1.674	**0.324**	0.396	0.471	0.430	**0.201**	1.741
Avg. Rank	5.000	8.188	9.562	10.625	9.938	11.062	14.250	13.625	4.750	13.875	**4.000**	4.312	6.000	4.438	**1.438**	14.938

Table 20: WQL scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX (RS)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.022	0.022	0.016	0.018	**0.015**	0.016	0.019	0.040	0.020	0.020	0.072	0.035	0.016	0.027	**0.015**	0.065
ETT (Hourly)	0.025	0.024	0.025	**0.022**	0.024	**0.022**	0.037	0.076	0.045	0.045	0.026	0.036	0.065	0.063	0.035	0.101
M5	**0.538**	**0.537**	0.557	0.552	0.551	0.549	0.586	0.576	-	0.560	0.617	0.812	0.607	0.604	0.748	0.934
Electricity-BE	0.058	0.059	0.049	0.051	0.047	0.049	0.069	0.075	0.080	0.087	0.078	0.040	**0.036**	0.054	**0.038**	0.071
Electricity-DE	0.455	0.553	**0.219**	0.243	0.245	0.315	0.457	0.414	0.372	0.374	0.235	0.304	**0.195**	0.378	0.295	1.048
Electricity-FR	0.074	0.071	0.062	0.062	0.059	0.060	0.083	0.058	0.061	0.076	0.100	**0.048**	0.061	0.076	**0.041**	0.135
Electricity-NP	0.020	**0.019**	**0.019**	0.027	0.022	0.025	0.039	0.026	0.028	0.026	0.030	0.022	0.027	0.029	0.034	0.037
Electricity-PJM	0.059	0.070	0.038	0.033	0.042	0.039	0.062	0.051	0.075	0.027	0.030	0.030	0.027	**0.027**	**0.024**	0.037
Bull	0.077	0.154	**0.062**	0.105	0.092	0.113	0.171	0.171	0.133	0.095	0.104	0.082	**0.072**	0.073	0.077	0.234
Cockatoo	0.054	0.141	0.040	0.068	0.033	0.044	**0.029**	0.055	0.060	0.055	0.031	0.085	**0.026**	0.091	0.029	0.068
Covid19 (ProEnFo)	0.016	0.017	0.011	0.013	0.016	0.017	0.008	**0.006**	0.018	0.012	0.007	0.008	**0.006**	0.007	0.009	0.014
GFC12	0.064	0.055	0.037	0.035	**0.033**	**0.034**	0.047	0.040	0.047	0.040	0.056	0.037	0.039	0.042	0.042	0.064
GFC14	0.015	0.020	**0.013**	0.014	0.015	0.015	0.019	0.017	0.032	0.016	**0.012**	0.019	0.019	0.022	0.027	0.031
GFC17	0.025	0.038	0.019	0.035	0.034	0.035	0.034	0.031	0.021	**0.018**	0.046	**0.013**	0.021	0.021	0.019	0.031
Hog	0.031	0.032	**0.018**	0.021	0.020	0.024	0.034	0.043	0.037	**0.016**	0.019	0.020	0.020	0.021	0.020	0.022
PDB	0.027	0.021	0.017	0.015	**0.015**	0.015	0.019	0.019	0.020	0.018	0.025	**0.010**	0.017	0.016	0.020	0.030
Spain	0.059	0.043	0.038	0.039	0.034	0.035	0.039	0.054	0.033	0.060	0.046	**0.023**	0.071	0.038	**0.030**	0.057
Rideshare	**0.177**	**0.178**	0.178	0.229	0.179	0.182	0.192	0.194	0.257	0.219	0.190	0.184	0.205	0.190	0.193	0.325
Agg. Relative Score	1.174	1.313	**0.878**	1.011	0.944	1.000	1.204	1.272	1.286	1.091	1.140	0.932	0.939	1.094	**0.926**	1.752
Avg. Rank	9.056	10.556	**4.667**	7.500	**5.389**	6.722	10.667	10.111	11.353	8.722	9.444	6.056	6.611	8.722	6.444	13.722

Table 21: MASE scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

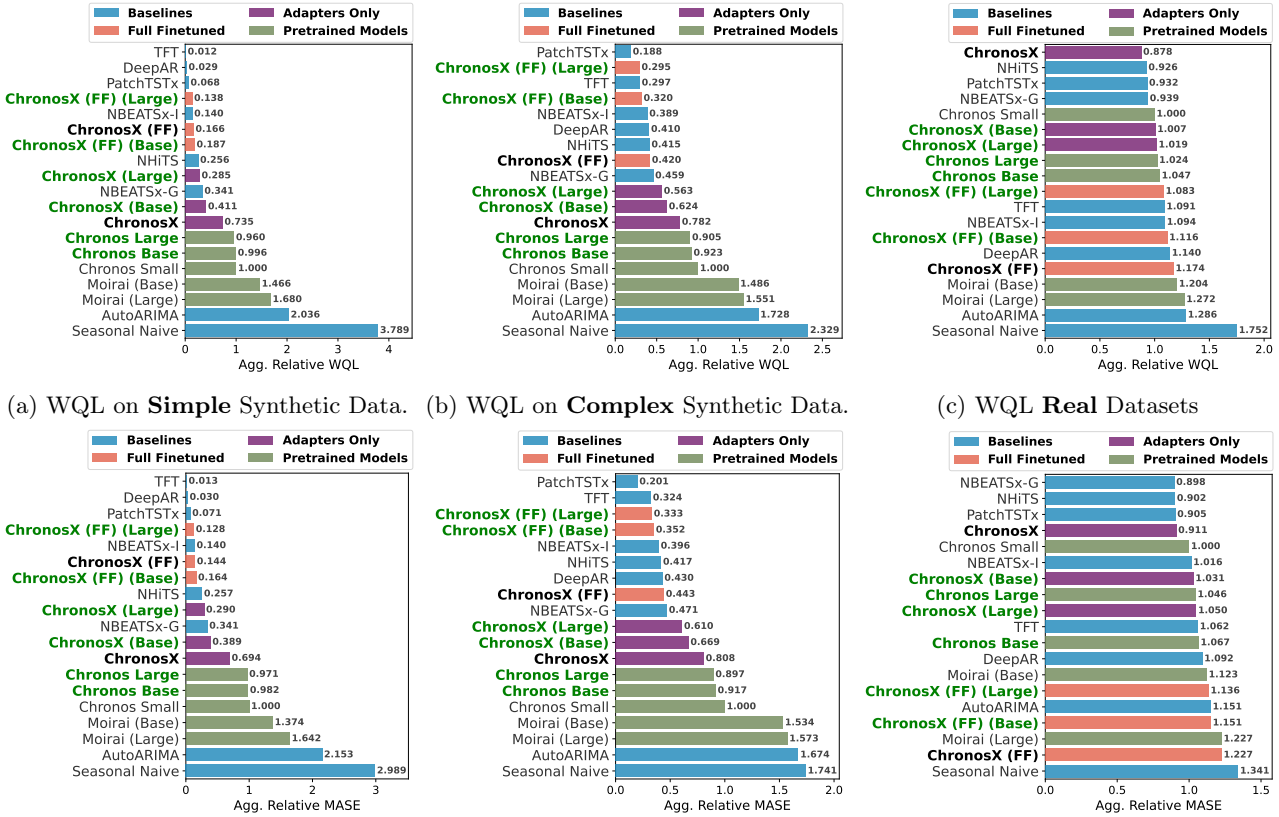
	Full Finetuned		Adapters Only			Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX (FF) (NC)	ChronosX	ChronosX (NC)	ChronosX (RS)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.338	0.327	0.213	0.210	**0.206**	0.224	0.259	0.513	0.225	0.269	0.790	0.451	**0.208**	0.404	0.239	0.677
ETT (Hourly)	0.354	0.343	0.352	**0.294**	0.338	**0.297**	0.489	0.900	0.469	0.599	0.351	0.463	0.752	0.677	0.443	0.950
M5	**0.907**	**0.905**	0.925	0.930	0.918	0.919	0.959	0.925	-	0.911	0.938	1.038	0.946	0.953	1.326	1.268
Electricity-BE	0.432	0.454	0.374	0.382	0.372	0.380	0.458	0.510	0.306	0.621	0.568	0.292	**0.276**	0.406	**0.276**	0.405
Electricity-DE	1.001	1.208	**0.378**	0.473	0.466	0.636	0.970	0.820	0.699	0.671	0.458	0.625	**0.381**	0.646	0.526	1.582
Electricity-FR	0.676	0.640	0.570	0.558	0.537	0.548	0.801	0.520	0.532	0.691	0.881	**0.451**	0.562	0.628	**0.382**	0.966
Electricity-NP	0.548	**0.526**	**0.518**	0.742	0.595	0.654	1.070	0.674	0.717	0.683	0.769	0.560	0.685	0.682	0.869	0.781
Electricity-PJM	0.389	0.450	0.261	0.227	0.281	0.267	0.372	0.308	0.194	0.155	0.166	0.176	**0.154**	0.162	**0.147**	0.169
Bull	0.830	1.447	**0.601**	0.954	0.840	0.990	1.577	1.497	1.248	0.851	0.940	0.783	**0.723**	0.727	0.769	1.681
Cockatoo	0.911	2.097	0.652	1.037	0.558	0.706	**0.324**	0.797	0.881	0.761	0.442	1.107	**0.308**	1.184	0.471	0.737
Covid19 (ProEnFo)	0.407	0.432	0.269	0.338	0.407	0.411	0.183	**0.141**	0.423	0.330	0.204	0.224	**0.157**	0.174	0.214	0.276
GFC12	1.193	1.044	0.738	0.692	**0.644**	0.657	0.769	0.799	0.718	0.738	1.138	**0.599**	0.709	0.667	0.688	0.925
GFC14	0.499	0.569	0.414	**0.403**	0.466	0.433	0.517	0.501	1.010	0.468	**0.339**	0.524	0.521	0.587	0.693	0.722
GFC17	0.525	0.820	0.474	0.743	0.717	0.732	0.782	0.644	0.535	0.421	0.904	**0.303**	0.486	0.430	**0.413**	0.533
Hog	0.468	0.364	0.284	0.293	0.277	0.302	0.378	0.574	0.720	**0.239**	0.261	0.339	0.324	0.317	0.308	**0.227**
PDB	0.658	0.530	0.404	0.369	0.366	0.355	0.347	0.432	0.497	0.398	0.580	**0.208**	0.352	**0.297**	0.388	0.545
Spain	0.834	0.628	0.589	0.638	0.438	0.454	0.505	0.780	0.390	0.817	0.585	**0.304**	0.850	0.482	**0.357**	0.606
Rideshare	**0.461**	0.463	0.467	0.614	0.467	0.472	0.473	0.476	0.630	0.528	0.475	**0.460**	0.497	0.470	0.469	0.630
Agg. Relative Score	1.227	1.324	0.911	1.022	0.955	1.000	1.123	1.227	1.151	1.062	1.092	0.905	**0.898**	1.016	**0.902**	1.341
Avg. Rank	10.167	10.889	**5.833**	8.167	**5.500**	7.000	10.444	10.500	10.088	8.833	9.222	6.167	6.611	7.944	6.333	11.972

I.5 Ablation: performance with larger version of Chronos

In this section we present results where ChronosX is based in Chronos-Small, Chronos-Base and Chronos-Large pretrained models.

We present the following results:

- In Fig. 15 we present the Aggregated Relative WQL and MASE on simple, complex, and real datasets, with ChronosX on Chronos-Small, Chronos-Base and Chronos-Large model versions.
- In Table 22 and Table 23 we present the WQL on simple and complex datasets, respectively,
- In Table 24 and Table 25 we present the MASE on simple datasets, respectively.
- In Table 26 and Table 27 we present WQL and MASE on real datasets, respectively.



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets

(d) MASE on **Simple** Synthetic Data. (e) MASE on **Complex** Synthetic Data. (f) MASE on **Real** Datasets

Figure 15: Ablations of **ChronosX** and **ChronosX (FF)** with larger versions of Chronos.

Table 22: WQL scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Large)	Moirai (Base)	NBEATS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.164	0.545	0.043	0.168	0.079	0.052	0.056	0.057	0.054	0.148	0.099	0.012	0.198	0.007	**0.005**	0.015	0.007	**0.002**	0.402
Simple ARP (Mult.)	0.042	0.020	0.021	0.136	0.097	0.042	0.037	0.039	0.038	0.153	0.114	0.014	0.313	**0.001**	0.012	0.018	0.004	**0.003**	0.572
Simple Bells (Add)	0.287	0.281	0.283	0.225	0.297	0.274	0.319	0.307	0.296	0.383	0.346	0.037	0.176	**0.005**	0.048	0.061	0.056	**0.003**	0.582
Simple Bells (Mult.)	0.041	0.081	0.052	0.176	0.183	0.086	0.089	0.089	0.090	0.156	0.136	0.033	0.273	**0.001**	0.040	0.037	0.005	**0.004**	0.417
Simple Spikes (Add)	0.007	0.007	0.007	0.250	0.093	0.033	0.431	0.376	0.358	0.515	0.498	0.083	0.381	**0.003**	0.036	0.117	**0.003**	0.004	0.978
Simple Spikes (Mult.)	0.007	0.007	0.007	0.022	0.014	0.036	0.142	0.144	0.140	0.212	0.196	0.101	0.446	**0.001**	0.014	0.116	**0.002**	0.099	0.440
Simple Steps (Add)	0.007	0.007	0.007	0.185	0.043	0.024	0.415	0.403	0.404	0.455	0.458	0.096	0.293	**0.004**	0.067	0.105	0.009	**0.004**	0.835
Simple Steps (Mult.)	0.007	0.007	0.007	0.030	0.010	0.010	0.169	0.196	0.189	0.275	0.269	0.032	0.486	**0.001**	0.019	0.046	**0.002**	0.143	0.577
Single ARP (Add.)	0.130	0.139	0.042	0.175	0.122	0.060	0.062	0.059	0.059	0.152	0.106	0.013	0.222	**0.005**	0.008	0.018	0.007	**0.003**	0.416
Single ARP (Mult.)	0.038	0.021	0.019	0.157	0.087	0.039	0.038	0.039	0.035	0.131	0.109	0.012	0.410	**0.001**	0.011	0.016	0.004	**0.004**	0.538
Single Bells (Add.)	0.252	0.293	0.292	0.261	0.314	0.293	0.367	0.343	0.316	0.410	0.383	0.063	0.181	**0.006**	0.072	0.098	0.036	**0.003**	0.592
Single Bells (Mult.)	0.034	0.108	0.058	0.214	0.182	0.112	0.105	0.100	0.100	0.163	0.136	0.038	0.289	**0.001**	0.040	0.055	0.005	**0.004**	0.461
Single Spikes (Add.)	0.008	0.008	0.008	0.181	0.102	0.030	0.391	0.366	0.343	0.479	0.436	0.073	0.310	**0.001**	0.025	0.139	**0.002**	0.242	0.954
Single Spikes (Mult.)	0.008	0.008	0.008	0.031	0.014	0.024	0.145	0.151	0.140	0.230	0.197	0.108	0.463	**0.001**	0.012	0.120	**0.001**	0.098	0.462
Single Steps (Add.)	0.008	0.008	0.008	0.151	0.033	0.020	0.390	0.385	0.400	0.471	0.449	0.058	0.262	**0.002**	0.021	0.062	0.003	**0.002**	0.900
Single Steps (Mult.)	0.007	0.007	0.007	0.018	0.009	0.011	0.186	0.195	0.182	0.291	0.283	0.034	0.622	**0.001**	0.018	0.050	**0.001**	0.146	0.596
Agg. Relative Score	0.166	0.187	0.138	0.735	0.411	0.285	1.000	0.996	0.960	1.680	1.466	0.256	2.036	**0.012**	0.140	0.341	**0.029**	0.068	3.789
Avg. Rank	7.250	7.812	6.375	12.375	11.000	9.000	13.250	13.188	12.375	16.750	15.625	7.188	15.625	**1.438**	5.875	8.562	**2.875**	4.688	18.750

Table 23: WQL scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.asdf

	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Large)	Moirai (Base)	NBEATS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	0.092	0.037	0.036	0.117	0.052	0.047	0.063	0.052	0.047	0.167	0.138	0.012	0.265	0.023	**0.009**	0.019	0.032	**0.007**	0.250
Diverse ARP (Mult.)	0.036	0.030	0.033	0.077	0.062	0.041	0.059	0.044	0.042	0.153	0.137	0.024	0.269	0.030	**0.024**	0.028	0.029	**0.016**	0.221
Diverse Bells (Add)	0.208	0.215	0.221	0.271	0.243	0.235	0.257	0.259	0.250	0.317	0.313	0.124	0.168	**0.027**	1.162	0.105	0.161	**0.011**	0.464
Diverse Bells (Mult.)	0.051	0.046	0.056	0.116	0.108	0.081	0.089	0.081	0.084	0.162	0.155	0.055	0.233	**0.036**	0.050	0.067	0.081	**0.012**	0.213
Diverse Spikes (Add)	0.016	0.010	**0.008**	0.075	0.043	0.037	0.230	0.210	0.207	0.288	0.279	0.049	0.352	0.019	0.029	0.046	0.036	**0.006**	0.460
Diverse Spikes (Mult.)	0.021	**0.010**	**0.010**	0.049	0.033	0.039	0.084	0.076	0.077	0.134	0.131	0.061	0.200	0.012	0.029	0.068	0.027	0.055	0.202
Diverse Steps (Add)	0.046	0.021	**0.016**	0.108	0.097	0.090	0.273	0.258	0.259	0.296	0.293	0.051	0.326	0.028	0.028	0.061	0.076	**0.007**	0.490
Diverse Steps (Mult.)	0.042	**0.022**	**0.022**	0.087	0.081	0.073	0.126	0.118	0.123	0.185	0.188	0.054	0.264	0.033	0.049	0.067	0.086	0.036	0.273
Noisy ARP (Add)	0.063	0.057	0.057	0.102	0.076	0.063	0.072	0.069	0.065	0.169	0.152	0.051	0.185	0.059	**0.049**	0.051	0.052	**0.048**	0.261
Noisy ARP (Mult.)	0.064	0.068	0.050	0.077	0.071	0.055	0.066	0.060	0.056	0.144	0.134	0.048	0.253	0.053	0.047	0.048	**0.047**	**0.045**	0.219
Noisy Bells (Add)	0.198	0.282	0.127	0.265	0.180	0.198	0.249	0.243	0.232	0.299	0.290	0.066	0.108	**0.051**	0.062	0.078	0.052	**0.032**	0.439
Noisy Bells (Mult.)	0.049	0.045	**0.043**	0.101	0.099	0.108	0.094	0.093	0.094	0.163	0.157	0.081	0.205	0.072	0.088	0.091	0.058	**0.042**	0.213
Noisy Spikes (Add)	0.050	**0.042**	**0.042**	0.087	0.061	0.053	0.239	0.217	0.212	0.277	0.274	0.063	0.208	0.061	0.053	0.059	0.051	0.044	0.433
Noisy Spikes (Mult.)	0.038	**0.034**	**0.033**	0.059	0.053	0.048	0.100	0.091	0.089	0.149	0.146	0.076	0.197	0.048	0.050	0.084	0.040	0.050	0.216
Noisy Steps (Add)	0.065	**0.047**	0.047	0.127	0.126	0.126	0.277	0.277	0.289	0.266	0.267	0.060	0.197	0.066	0.049	0.071	0.053	**0.041**	0.425
Noisy Steps (Mult.)	0.044	**0.039**	**0.037**	0.072	0.070	0.061	0.120	0.117	0.110	0.158	0.154	0.062	0.223	0.057	0.058	0.065	0.050	0.042	0.259
Agg. Relative Score	0.420	0.320	**0.295**	0.782	0.624	0.563	1.000	0.923	0.905	1.551	1.486	0.415	1.728	0.297	0.389	0.459	0.410	**0.188**	2.329
Avg. Rank	6.688	5.000	**4.188**	13.125	11.062	9.188	14.062	12.812	12.625	16.875	16.125	6.750	16.188	5.188	5.750	7.500	5.812	**2.375**	18.688

Table 24: MASE scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Large)	Moirai (Base)	NBEATS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.947	2.741	0.254	0.951	0.532	0.308	0.322	0.326	0.319	0.849	0.571	0.067	1.103	0.039	**0.029**	0.086	0.034	**0.013**	1.788
Simple ARP (Mult.)	0.161	0.091	0.095	0.500	0.362	0.175	0.157	0.172	0.172	0.642	0.454	0.057	1.616	**0.006**	0.051	0.070	0.020	**0.014**	1.930
Simple Bells (Add)	3.203	3.143	3.171	2.459	3.275	3.119	3.483	3.353	3.328	4.069	3.689	0.392	1.993	**0.055**	0.491	0.641	0.613	0.035	4.951
Simple Bells (Mult.)	0.220	0.409	0.312	0.858	0.908	0.457	0.473	0.494	0.497	0.847	0.691	0.177	1.698	**0.004**	0.212	0.206	0.025	**0.019**	1.882
Simple Spikes (Add)	0.010	0.010	0.010	0.372	0.157	0.071	0.653	0.585	0.557	0.823	0.756	0.146	0.644	0.013	0.071	0.231	**0.009**	**0.007**	1.258
Simple Spikes (Mult.)	0.024	0.024	0.024	0.089	0.056	0.143	0.581	0.540	0.562	0.781	0.664	0.453	1.930	**0.004**	0.054	0.458	**0.006**	0.456	1.397
Simple Steps (Add)	0.023	0.023	**0.023**	0.742	0.225	0.216	1.653	1.603	1.663	1.862	1.856	0.422	1.079	0.033	0.337	0.452	0.077	**0.018**	2.577
Simple Steps (Mult.)	0.023	0.023	0.023	0.113	0.033	0.029	0.720	0.769	0.757	1.067	1.014	0.124	2.262	**0.008**	0.076	0.177	**0.007**	0.703	1.808
Single ARP (Add.)	0.601	0.636	0.241	0.811	0.608	0.341	0.344	0.336	0.340	0.823	0.548	0.072	1.200	**0.027**	0.040	0.093	0.035	**0.015**	1.774
Single ARP (Mult.)	0.148	0.092	0.091	0.593	0.318	0.166	0.166	0.171	0.154	0.540	0.418	0.053	1.914	**0.004**	0.048	0.069	0.019	**0.016**	1.877
Single Bells (Add.)	2.675	3.162	3.174	2.599	3.254	3.088	3.728	3.514	3.327	4.079	3.835	0.616	1.945	**0.054**	0.687	0.965	0.346	**0.026**	4.779
Single Bells (Mult.)	0.165	0.526	0.339	1.122	0.864	0.559	0.539	0.507	0.523	0.851	0.684	0.208	1.611	**0.003**	0.214	0.299	0.025	**0.019**	1.989
Single Spikes (Add.)	0.010	0.010	0.010	0.309	0.167	0.046	0.653	0.593	0.569	0.826	0.690	0.112	0.540	**0.002**	0.040	0.235	**0.002**	0.470	1.242
Single Spikes (Mult.)	0.023	0.023	0.023	0.115	0.046	0.086	0.575	0.555	0.536	0.831	0.629	0.449	1.909	**0.002**	0.046	0.453	**0.004**	0.451	1.415
Single Steps (Add.)	0.019	0.019	0.019	0.514	0.094	0.053	1.273	1.263	1.338	1.566	1.501	0.186	0.886	**0.005**	0.064	0.196	0.009	**0.008**	2.281
Single Steps (Mult.)	0.023	0.023	0.023	0.061	0.028	0.037	0.735	0.736	0.710	1.067	0.998	0.127	2.672	**0.004**	0.068	0.192	**0.004**	0.684	1.841
Agg. Relative Score	0.144	0.164	0.128	0.694	0.389	0.290	1.000	0.982	0.971	1.642	1.374	0.257	2.153	**0.013**	0.140	0.341	**0.030**	0.071	2.989
Avg. Rank	6.875	7.688	6.250	12.188	11.000	8.938	13.500	13.062	12.688	16.875	15.562	7.062	15.750	**2.062**	5.875	8.562	**2.812**	4.625	18.625

Table 25: MASE scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Large)	Moirai (Base)	NBEATS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	1.004	0.491	0.505	1.282	0.657	0.566	0.768	0.645	0.566	2.056	1.848	0.148	2.801	0.317	**0.110**	0.242	0.419	**0.099**	1.947
Diverse ARP (Mult.)	0.500	0.427	0.460	0.987	0.796	0.544	0.772	0.569	0.545	1.928	1.791	0.323	3.091	0.404	**0.314**	0.374	0.370	**0.221**	1.921
Diverse Bells (Add)	3.392	3.596	3.664	4.066	3.880	3.810	3.975	4.111	3.960	4.634	4.713	1.512	2.234	**0.446**	12.561	1.584	2.513	**0.169**	5.563
Diverse Bells (Mult.)	0.925	0.911	1.136	1.959	1.786	1.443	1.620	1.419	1.478	2.834	2.755	1.000	3.453	**0.651**	0.920	1.171	1.462	**0.249**	2.698
Diverse Spikes (Add)	0.062	0.039	**0.036**	0.294	0.189	0.169	0.767	0.693	0.669	1.048	0.999	0.163	1.203	0.077	0.107	0.176	0.132	**0.022**	1.286
Diverse Spikes (Mult.)	0.173	**0.086**	**0.090**	0.451	0.327	0.348	0.691	0.607	0.605	1.247	1.201	0.516	1.826	0.121	0.264	0.537	0.240	0.491	1.427
Diverse Steps (Add)	0.393	0.191	**0.153**	0.896	0.824	0.796	2.090	2.013	2.099	2.387	2.388	0.409	2.182	0.252	0.222	0.490	0.584	**0.056**	2.851
Diverse Steps (Mult.)	0.495	**0.298**	**0.314**	1.054	1.019	0.917	1.464	1.437	1.435	2.110	2.195	0.644	3.165	0.404	0.581	0.818	1.021	0.401	2.329
Noisy ARP (Add)	0.613	0.556	0.556	0.967	0.724	0.619	0.713	0.669	0.631	1.580	1.473	0.477	1.760	0.570	**0.464**	0.483	0.495	**0.459**	1.713
Noisy ARP (Mult.)	0.750	0.763	0.587	0.884	0.811	0.671	0.778	0.687	0.672	1.647	1.619	**0.551**	2.510	0.632	0.552	0.551	0.552	**0.512**	1.665
Noisy Bells (Add)	2.370	3.280	1.742	3.301	2.353	2.718	3.328	3.274	3.162	3.851	3.721	0.867	1.472	0.722	0.805	1.027	**0.714**	**0.436**	4.558
Noisy Bells (Mult.)	0.693	0.641	**0.623**	1.425	1.402	1.478	1.360	1.332	1.340	2.211	2.166	1.105	2.819	1.029	1.187	1.219	0.833	**0.583**	2.155
Noisy Spikes (Add)	0.210	0.178	**0.175**	0.344	0.257	0.227	0.776	0.680	0.664	0.995	0.974	0.239	0.842	0.241	0.208	0.236	0.203	**0.176**	1.194
Noisy Spikes (Mult.)	0.362	**0.336**	**0.326**	0.547	0.506	0.456	0.777	0.696	0.683	1.281	1.265	0.629	1.717	0.451	0.453	0.660	0.364	0.445	1.401
Noisy Steps (Add)	0.524	0.387	0.385	0.958	0.968	0.949	1.994	1.934	1.977	2.030	2.071	0.457	1.479	0.511	**0.381**	0.546	0.417	**0.324**	2.424
Noisy Steps (Mult.)	0.497	**0.457**	**0.437**	0.806	0.773	0.681	1.190	1.173	1.101	1.597	1.574	0.683	2.123	0.630	0.636	0.714	0.558	0.462	2.052
Agg. Relative Score	0.443	0.352	0.333	0.808	0.669	0.610	1.000	0.917	0.897	1.573	1.534	0.417	1.674	**0.324**	0.396	0.471	0.430	**0.201**	1.741
Avg. Rank	6.812	4.938	**4.312**	13.000	11.438	9.500	14.188	12.688	12.188	17.250	16.625	6.250	16.562	5.312	5.500	7.500	5.875	**2.125**	17.938

Table 26: WQL scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTST*	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.022	0.026	0.025	**0.016**	0.020	0.019	0.016	0.017	0.018	0.019	0.040	0.020	0.020	0.072	0.035	0.016	0.027	**0.015**	0.065
ETT (Hourly)	0.025	**0.024**	0.027	0.025	0.026	0.026	**0.022**	0.025	0.026	0.037	0.076	0.045	0.045	0.026	0.036	0.065	0.063	0.035	0.101
M5	0.538	**0.533**	**0.533**	0.557	0.551	0.550	0.549	0.544	0.544	0.586	0.576	-	0.560	0.617	0.812	0.607	0.604	0.748	0.934
Electricity-BE	0.058	0.062	0.065	0.049	0.052	0.045	0.049	0.051	0.044	0.069	0.075	0.080	0.087	0.078	0.040	**0.036**	0.054	**0.038**	0.071
Electricity-DE	0.455	0.218	0.419	0.219	**0.210**	0.324	0.315	0.267	0.393	0.457	0.414	0.372	0.374	0.235	0.304	**0.195**	0.378	0.295	1.048
Electricity-FR	0.074	0.067	0.060	0.062	0.059	0.050	0.060	0.054	0.049	0.083	0.058	0.061	0.076	0.100	**0.048**	0.061	0.076	**0.041**	0.135
Electricity-NP	0.020	0.020	**0.016**	0.019	**0.016**	0.025	0.025	0.020	0.023	0.039	0.026	0.028	0.026	0.030	0.022	0.027	0.029	0.034	0.037
Electricity-PJM	0.059	0.038	0.058	0.038	0.031	0.034	0.039	0.030	0.033	0.062	0.051	0.075	0.027	0.030	0.030	0.027	**0.027**	**0.024**	0.037
Bull	0.077	**0.061**	0.144	**0.062**	0.070	0.110	0.113	0.117	0.125	0.171	0.171	0.133	0.095	0.104	0.082	0.072	0.073	0.077	0.234
Cockatoo	0.054	0.169	0.035	0.040	0.072	0.080	0.044	0.070	0.058	**0.029**	0.055	0.060	0.055	0.031	0.085	**0.026**	0.091	0.029	0.068
Covid19 (ProEnFo)	0.016	0.012	0.014	0.011	0.013	0.011	0.017	0.013	0.011	0.008	**0.006**	0.018	0.012	0.007	0.008	**0.006**	0.007	0.009	0.014
GFC12	0.064	0.051	0.043	0.037	**0.032**	0.036	0.034	**0.030**	0.036	0.047	0.040	0.047	0.040	0.056	0.037	0.039	0.042	0.042	0.064
GFC14	0.015	0.014	0.016	0.013	0.015	**0.013**	0.015	0.014	0.013	0.019	0.017	0.032	0.016	**0.012**	0.019	0.019	0.022	0.027	0.031
GFC17	0.025	0.029	0.026	0.019	0.027	**0.018**	0.035	0.043	0.019	0.034	0.031	0.021	0.018	0.046	**0.013**	0.021	0.021	0.019	0.031
Hog	0.031	**0.017**	0.020	0.018	0.018	0.021	0.024	0.021	0.023	0.034	0.043	0.037	**0.016**	0.019	0.020	0.020	0.021	0.020	0.022
PDB	0.027	0.033	0.019	0.017	0.032	0.017	**0.015**	0.018	0.018	0.019	0.019	0.020	0.018	0.025	**0.010**	0.017	0.016	0.020	0.030
Spain	0.059	0.067	0.053	0.038	0.083	0.094	0.035	0.092	0.097	0.039	0.054	0.033	0.060	0.046	**0.023**	0.071	0.038	**0.030**	0.057
Rideshare	**0.177**	**0.177**	0.177	0.178	0.178	0.178	0.182	0.177	0.178	0.192	0.194	0.257	0.219	0.190	0.184	0.205	0.190	0.193	0.325
Agg. Relative Score	1.174	1.116	1.083	**0.878**	1.007	1.019	1.000	1.047	1.024	1.204	1.272	1.286	1.091	1.140	0.932	0.939	1.094	**0.926**	1.752
Avg. Rank	10.944	9.167	10.111	**5.833**	8.278	8.000	8.500	8.333	8.444	13.167	12.667	13.882	10.500	11.056	**7.722**	8.056	10.722	7.944	16.389

Table 27: MASE scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

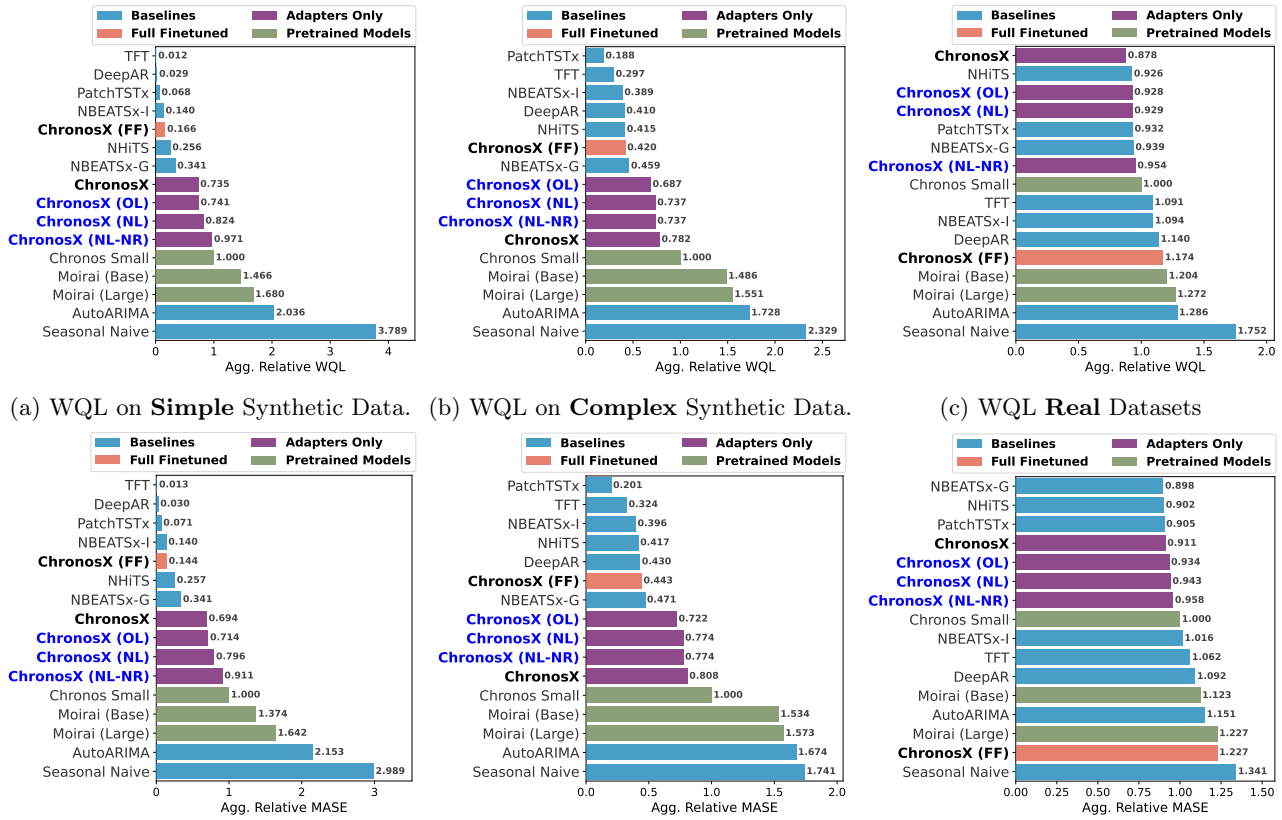
	Full Finetuned			Adapters Only			Pretrained Models					Baselines							
	ChronosX (FF)	ChronosX (FF) (Base)	ChronosX (FF) (Large)	ChronosX	ChronosX (Base)	ChronosX (Large)	Chronos Small	Chronos Base	Chronos Large	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTST*	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.338	0.363	0.344	**0.213**	0.287	0.272	0.224	0.242	0.243	0.259	0.513	0.225	0.269	0.790	0.451	**0.208**	0.404	0.239	0.677
ETT (Hourly)	0.354	**0.341**	0.373	0.352	0.365	0.366	**0.297**	0.347	0.373	0.489	0.900	0.469	0.599	0.351	0.463	0.752	0.677	0.443	0.950
M5	0.907	**0.902**	**0.902**	0.925	0.914	0.916	0.919	0.916	0.917	0.959	0.925	-	0.911	0.938	1.038	0.946	0.953	1.326	1.268
Electricity-BE	0.432	0.482	0.499	0.374	0.380	0.326	0.380	0.383	0.329	0.458	0.510	0.306	0.621	0.568	0.292	**0.276**	0.406	**0.276**	0.405
Electricity-DE	1.001	0.443	0.906	**0.378**	0.396	0.691	0.636	0.559	0.856	0.970	0.820	0.699	0.671	0.458	0.625	**0.381**	0.646	0.526	1.582
Electricity-FR	0.676	0.598	0.531	0.570	0.541	0.467	0.548	0.496	0.459	0.801	0.520	0.532	0.691	0.881	**0.451**	0.562	0.628	**0.382**	0.966
Electricity-NP	0.548	0.606	**0.464**	0.518	**0.467**	0.662	0.654	0.559	0.589	1.070	0.674	0.717	0.683	0.769	0.560	0.685	0.682	0.869	0.781
Electricity-PJM	0.389	0.260	0.378	0.261	0.202	0.227	0.267	0.186	0.218	0.372	0.308	0.194	0.155	0.166	0.176	**0.154**	0.162	**0.147**	0.169
Bull	0.830	**0.596**	1.322	**0.601**	0.615	1.023	0.990	1.032	1.113	1.577	1.497	1.248	0.851	0.940	0.783	0.723	0.727	0.769	1.681
Cockatoo	0.911	1.888	0.548	0.652	1.096	1.180	0.706	1.040	0.917	**0.324**	0.797	0.881	0.761	0.442	1.107	**0.308**	1.184	0.471	0.737
Covid19 (ProEnFo)	0.407	0.320	0.359	0.269	0.334	0.302	0.411	0.334	0.302	0.183	**0.141**	0.423	0.330	0.204	0.224	**0.157**	0.174	0.214	0.276
GFC12	1.193	1.079	0.879	0.738	0.678	0.754	0.657	**0.651**	0.742	0.769	0.799	0.718	0.738	1.138	**0.599**	0.709	0.667	0.688	0.925
GFC14	0.499	0.459	0.521	0.414	0.473	**0.392**	0.433	0.436	0.394	0.517	0.501	1.010	0.468	**0.339**	0.524	0.521	0.587	0.693	0.722
GFC17	0.525	0.590	0.556	0.474	0.560	0.424	0.732	0.895	0.427	0.782	0.644	0.535	0.421	0.904	**0.303**	0.486	0.430	**0.413**	0.533
Hog	0.468	0.313	0.322	0.284	0.287	0.304	0.302	0.321	0.329	0.378	0.574	0.720	**0.239**	0.261	0.339	0.324	0.317	0.308	**0.227**
PDB	0.658	0.790	0.472	0.404	0.778	0.431	0.355	0.466	0.449	0.347	0.432	0.497	0.398	0.580	**0.208**	0.352	**0.297**	0.388	0.545
Spain	0.834	0.909	0.852	0.589	1.091	1.162	0.454	1.160	1.193	0.505	0.780	0.390	0.817	0.585	**0.304**	0.850	0.482	**0.357**	0.606
Rideshare	**0.461**	0.461	0.462	0.467	0.463	0.463	0.472	0.462	0.464	0.473	0.476	0.630	0.528	0.475	**0.460**	0.497	0.470	0.469	0.630
Agg. Relative Score	1.227	1.151	1.136	0.911	1.031	1.050	1.000	1.067	1.046	1.123	1.227	1.151	1.062	1.092	0.905	**0.898**	1.016	**0.902**	1.341
Avg. Rank	11.833	10.333	11.389	**6.667**	8.944	9.222	8.389	8.944	9.722	12.500	12.778	11.971	9.944	10.722	7.500	7.889	9.833	**7.167**	13.861

I.6 Ablation: performance with architecture of adapter architecture

In this section we present results where ChronosX has different adapter architectures for covariates.

We present the following results:

- In Fig. 16 we present the Aggregated Relative WQL and MASE on different versions of ChronosX with different adapter blocks.
- In Table 28 and Table 29 we present the WQL on simple and complex datasets, respectively,
- In Table 30 and Table 31 we present the MASE on simple and complex datasets, respectively.
- In Table 32 and Table 33 we present WQL and MASE on real datasets, respectively.



(a) WQL on **Simple** Synthetic Data. (b) WQL on **Complex** Synthetic Data. (c) WQL **Real** Datasets

(d) MASE on **Simple** Synthetic Data. (e) MASE on **Complex** Synthetic Data. (f) MASE on **Real** Datasets

Figure 16: Ablations of **ChronosX** with simpler adapter blocks.

Table 28: WQL scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Large)	Moirai (Base)	NHITS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.164	0.168	0.089	0.088	0.066	0.056	0.148	0.099	0.012	0.198	0.007	**0.005**	0.015	0.007	**0.002**	0.402
Simple ARP (Mult.)	0.042	0.136	0.127	0.108	0.057	0.037	0.153	0.114	0.014	0.313	**0.001**	0.012	0.018	0.004	**0.003**	0.572
Simple Bells (Add)	0.287	0.225	0.302	0.342	0.347	0.319	0.383	0.346	0.037	0.176	**0.005**	0.048	0.061	0.056	**0.003**	0.582
Simple Bells (Mult.)	0.041	0.176	0.165	0.151	0.133	0.089	0.156	0.136	0.033	0.273	**0.001**	0.040	0.037	0.005	**0.004**	0.417
Simple Spikes (Add)	0.007	0.250	0.206	0.233	0.223	0.431	0.515	0.498	0.083	0.381	**0.003**	0.036	0.117	**0.003**	0.004	0.978
Simple Spikes (Mult.)	0.007	0.022	0.020	0.025	0.136	0.142	0.212	0.196	0.101	0.446	**0.001**	0.014	0.116	**0.002**	0.099	0.440
Simple Steps (Add)	0.007	0.185	0.133	0.122	0.118	0.415	0.455	0.458	0.096	0.293	**0.004**	0.067	0.105	0.009	**0.004**	0.835
Simple Steps (Mult.)	0.007	0.030	0.059	0.105	0.204	0.169	0.275	0.269	0.032	0.486	**0.001**	0.019	0.046	**0.002**	0.143	0.577
Single ARP (Add.)	0.130	0.175	0.105	0.121	0.076	0.062	0.152	0.106	0.013	0.222	**0.005**	0.008	0.018	0.007	**0.003**	0.416
Single ARP (Mult.)	0.038	0.157	0.114	0.089	0.062	0.038	0.131	0.109	0.012	0.410	**0.001**	0.011	0.016	0.004	**0.004**	0.538
Single Bells (Add.)	0.252	0.261	0.318	0.391	0.398	0.367	0.410	0.383	0.063	0.181	**0.006**	0.072	0.098	0.036	**0.003**	0.592
Single Bells (Mult.)	0.034	0.214	0.157	0.110	0.124	0.105	0.163	0.136	0.038	0.289	**0.001**	0.040	0.055	0.005	**0.004**	0.461
Single Spikes (Add.)	0.008	0.181	0.210	0.241	0.361	0.391	0.479	0.436	0.073	0.310	**0.001**	0.025	0.139	**0.002**	0.242	0.954
Single Spikes (Mult.)	0.008	0.031	0.040	0.070	0.140	0.145	0.230	0.197	0.108	0.463	**0.001**	0.012	0.120	**0.001**	0.098	0.462
Single Steps (Add.)	0.008	0.151	0.156	0.195	0.185	0.390	0.471	0.449	0.058	0.262	**0.002**	0.021	0.062	0.003	**0.002**	0.900
Single Steps (Mult.)	0.007	0.018	0.052	0.073	0.213	0.186	0.291	0.283	0.034	0.622	**0.001**	0.018	0.050	**0.001**	0.146	0.596
Agg. Relative Score	0.166	0.735	0.741	0.824	0.971	1.000	1.680	1.466	0.256	2.036	**0.012**	0.140	0.341	**0.029**	0.068	3.789
Avg. Rank	5.938	9.875	9.375	9.750	10.375	10.250	13.875	12.375	5.562	13.375	**1.438**	4.375	6.812	**2.750**	4.062	15.812

Table 29: WQL scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.asdf

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Large)	Moirai (Base)	NHITS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	0.092	0.117	0.061	0.063	0.065	0.063	0.167	0.138	0.012	0.265	0.023	**0.009**	0.019	0.032	**0.007**	0.250
Diverse ARP (Mult.)	0.036	0.077	0.061	0.061	0.060	0.059	0.153	0.137	0.024	0.269	0.030	**0.024**	0.028	0.029	**0.016**	0.221
Diverse Bells (Add)	0.208	0.271	0.252	0.265	0.266	0.257	0.317	0.313	0.124	0.168	**0.027**	1.162	0.105	0.161	**0.011**	0.464
Diverse Bells (Mult.)	0.051	0.116	0.096	0.098	0.100	0.089	0.162	0.155	0.055	0.233	**0.036**	0.050	0.067	0.081	**0.012**	0.213
Diverse Spikes (Add)	**0.016**	0.075	0.077	0.075	0.075	0.230	0.288	0.279	0.049	0.352	0.019	0.029	0.046	0.036	**0.006**	0.460
Diverse Spikes (Mult.)	**0.021**	0.049	0.051	0.047	0.048	0.084	0.134	0.131	0.061	0.200	**0.012**	0.029	0.068	0.027	0.055	0.202
Diverse Steps (Add)	0.046	0.108	0.109	0.108	0.107	0.273	0.296	0.293	0.051	0.326	**0.028**	0.028	0.061	0.076	**0.007**	0.490
Diverse Steps (Mult.)	0.042	0.087	0.079	0.091	0.090	0.126	0.185	0.188	0.054	0.264	**0.033**	0.049	0.067	0.086	**0.036**	0.273
Noisy ARP (Add)	0.063	0.102	0.083	0.077	0.081	0.072	0.169	0.152	0.051	0.185	0.059	**0.049**	0.051	0.052	**0.048**	0.261
Noisy ARP (Mult.)	0.064	0.077	0.071	0.069	0.066	0.066	0.144	0.134	0.048	0.253	0.053	0.047	0.048	**0.047**	**0.045**	0.219
Noisy Bells (Add)	0.198	0.265	0.163	0.284	0.302	0.249	0.299	0.290	0.066	0.108	**0.051**	0.062	0.078	0.052	**0.032**	0.439
Noisy Bells (Mult.)	**0.049**	0.101	0.098	0.098	0.090	0.094	0.163	0.157	0.081	0.205	0.072	0.088	0.091	0.058	**0.042**	0.213
Noisy Spikes (Add)	**0.050**	0.087	0.087	0.087	0.088	0.239	0.277	0.274	0.063	0.208	0.061	0.053	0.059	0.051	**0.044**	0.433
Noisy Spikes (Mult.)	**0.038**	0.059	0.057	0.061	0.062	0.100	0.149	0.146	0.076	0.197	0.048	0.050	0.084	**0.040**	0.050	0.216
Noisy Steps (Add)	0.065	0.127	0.126	0.140	0.137	0.277	0.266	0.267	0.060	0.197	0.066	**0.049**	0.071	0.053	**0.041**	0.425
Noisy Steps (Mult.)	**0.044**	0.072	0.069	0.097	0.096	0.120	0.158	0.154	0.062	0.223	0.057	0.058	0.065	0.050	**0.042**	0.259
Agg. Relative Score	0.420	0.782	0.687	0.737	0.737	1.000	1.551	1.486	0.415	1.728	**0.297**	0.389	0.459	0.410	**0.188**	2.329
Avg. Rank	4.688	10.188	8.938	9.750	9.625	10.500	13.938	13.188	5.375	13.812	**3.688**	4.312	6.188	4.375	**1.750**	15.688

Table 30: MASE scores per model in 16 **simple** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Large)	Moirai (Base)	NHITS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Simple ARP (Add)	0.947	0.951	0.567	0.559	0.429	0.322	0.849	0.571	0.067	1.103	0.039	**0.029**	0.086	0.034	**0.013**	1.788
Simple ARP (Mult.)	0.161	0.500	0.479	0.410	0.226	0.157	0.642	0.454	0.057	1.616	**0.006**	0.051	0.070	0.020	**0.014**	1.930
Simple Bells (Add)	3.203	2.459	3.182	3.605	3.645	3.483	4.069	3.689	0.392	1.993	**0.055**	0.491	0.641	0.613	**0.035**	4.951
Simple Bells (Mult.)	0.220	0.858	0.826	0.800	0.683	0.473	0.847	0.691	0.177	1.698	**0.004**	0.212	0.206	0.025	**0.019**	1.882
Simple Spikes (Add)	0.010	0.372	0.305	0.364	0.344	0.653	0.823	0.756	0.146	0.644	0.013	0.071	0.231	**0.009**	**0.007**	1.258
Simple Spikes (Mult.)	0.024	0.089	0.092	0.092	0.443	0.581	0.781	0.664	0.453	1.930	**0.004**	0.054	0.458	**0.006**	0.456	1.397
Simple Steps (Add)	**0.023**	0.742	0.510	0.528	0.559	1.653	1.862	1.856	0.422	1.079	0.033	0.337	0.452	0.077	**0.018**	2.577
Simple Steps (Mult.)	0.023	0.113	0.229	0.391	0.671	0.720	1.067	1.014	0.124	2.262	**0.008**	0.076	0.177	**0.007**	0.703	1.808
Single ARP (Add.)	0.601	0.811	0.524	0.623	0.409	0.344	0.823	0.548	0.072	1.200	**0.027**	0.040	0.093	0.035	**0.015**	1.774
Single ARP (Mult.)	0.148	0.593	0.436	0.356	0.236	0.166	0.540	0.418	0.053	1.914	**0.004**	0.048	0.069	0.019	**0.016**	1.877
Single Bells (Add.)	2.675	2.599	3.175	3.870	3.937	3.728	4.079	3.835	0.616	1.945	**0.054**	0.687	0.965	0.346	**0.026**	4.779
Single Bells (Mult.)	0.165	1.122	0.758	0.545	0.609	0.539	0.851	0.684	0.208	1.611	**0.003**	0.214	0.299	0.025	**0.019**	1.989
Single Spikes (Add.)	0.010	0.309	0.347	0.406	0.570	0.653	0.826	0.690	0.112	0.540	**0.002**	0.040	0.235	**0.002**	0.470	1.242
Single Spikes (Mult.)	0.023	0.115	0.147	0.251	0.454	0.575	0.831	0.629	0.449	1.909	**0.002**	0.046	0.453	**0.004**	0.451	1.415
Single Steps (Add.)	0.019	0.514	0.543	0.673	0.614	1.273	1.566	1.501	0.186	0.886	**0.005**	0.064	0.196	0.009	**0.008**	2.281
Single Steps (Mult.)	0.023	0.061	0.196	0.269	0.670	0.735	1.067	0.998	0.127	2.672	**0.004**	0.068	0.192	**0.004**	0.684	1.841
Agg. Relative Score	0.144	0.694	0.714	0.796	0.911	1.000	1.642	1.374	0.257	2.153	**0.013**	0.140	0.341	**0.030**	0.071	2.989
Avg. Rank	5.812	9.625	9.250	9.750	10.000	10.438	14.062	12.375	5.500	13.500	**1.812**	4.438	6.875	**2.688**	4.188	15.688

Table 31: MASE scores per model in 16 **complex** synthetic datasets. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Large)	Moirai (Base)	NHITS	AutoARIMA	TFT	NBEATSx-I	NBEATSx-G	DeepAR	PatchTSTx	Seasonal Naive
Diverse ARP (Add)	1.004	1.282	0.745	0.766	0.779	0.768	2.056	1.848	0.148	2.801	0.317	**0.110**	0.242	0.419	**0.099**	1.947
Diverse ARP (Mult.)	0.500	0.987	0.791	0.783	0.766	0.772	1.928	1.791	0.323	3.091	0.404	**0.314**	0.374	0.370	**0.221**	1.921
Diverse Bells (Add)	3.392	4.066	3.840	3.997	4.008	3.975	4.634	4.713	1.512	2.234	**0.446**	12.561	1.584	2.513	**0.169**	5.563
Diverse Bells (Mult.)	0.925	1.959	1.673	1.720	1.755	1.620	2.834	2.755	1.000	3.453	**0.651**	0.920	1.171	1.462	**0.249**	2.698
Diverse Spikes (Add)	**0.062**	0.294	0.302	0.298	0.295	0.767	1.048	0.999	0.163	1.203	0.077	0.107	0.176	0.132	**0.022**	1.286
Diverse Spikes (Mult.)	**0.173**	0.451	0.471	0.452	0.460	0.691	1.247	1.201	0.516	1.826	**0.121**	0.264	0.537	0.240	0.491	1.427
Diverse Steps (Add)	0.393	0.896	0.889	0.914	0.893	2.099	2.387	2.388	0.409	2.182	0.252	**0.222**	0.490	0.584	**0.056**	2.851
Diverse Steps (Mult.)	0.495	1.054	0.965	1.105	1.119	1.464	2.110	2.195	0.644	3.165	**0.404**	0.581	0.818	1.021	**0.401**	2.329
Noisy ARP (Add)	0.613	0.967	0.793	0.748	0.784	0.713	1.580	1.473	0.477	1.760	0.570	**0.464**	0.483	0.495	**0.459**	1.713
Noisy ARP (Mult.)	0.750	0.884	0.848	0.829	0.799	0.778	1.647	1.619	**0.551**	2.510	0.632	0.552	0.551	0.552	**0.512**	1.665
Noisy Bells (Add)	2.370	3.301	2.144	3.703	3.911	3.328	3.851	3.721	0.867	1.472	0.722	0.805	1.027	**0.714**	**0.436**	4.558
Noisy Bells (Mult.)	**0.693**	1.425	1.383	1.388	1.304	1.360	2.211	2.166	1.105	2.819	1.029	1.187	1.219	0.833	**0.583**	2.155
Noisy Spikes (Add)	0.210	0.344	0.344	0.341	0.345	0.776	0.995	0.974	0.239	0.842	0.241	0.208	0.236	**0.203**	**0.176**	1.194
Noisy Spikes (Mult.)	**0.362**	0.547	0.529	0.566	0.574	0.777	1.281	1.265	0.629	1.717	0.451	0.453	0.660	**0.364**	0.445	1.401
Noisy Steps (Add)	0.524	0.958	0.950	1.050	1.038	1.994	2.030	2.071	0.457	1.479	0.511	**0.381**	0.546	0.417	**0.324**	2.424
Noisy Steps (Mult.)	**0.497**	0.806	0.772	1.027	0.999	1.190	1.597	1.574	0.683	2.123	0.630	0.636	0.714	0.558	**0.462**	2.052
Agg. Relative Score	0.443	0.808	0.722	0.774	0.774	1.000	1.573	1.534	0.417	1.674	**0.324**	0.396	0.471	0.430	**0.201**	1.741
Avg. Rank	4.875	10.062	8.750	9.688	9.875	10.500	14.188	13.562	5.062	14.062	**3.938**	4.250	6.188	4.438	**1.625**	14.938

Table 32: WQL scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.022	**0.016**	0.017	0.016	0.017	0.016	0.019	0.040	0.020	0.020	0.072	0.035	0.016	0.027	**0.015**	0.065
ETT (Hourly)	0.025	0.025	0.023	0.023	**0.023**	**0.022**	0.037	0.076	0.045	0.045	0.026	0.036	0.065	0.063	0.035	0.101
M5	**0.538**	0.557	0.555	0.553	0.552	**0.549**	0.586	0.576	-	0.560	0.617	0.812	0.607	0.604	0.748	0.934
Electricity-BE	0.058	0.049	0.051	0.050	0.050	0.049	0.069	0.075	0.080	0.087	0.078	0.040	**0.036**	0.054	**0.038**	0.071
Electricity-DE	0.455	**0.219**	0.237	0.238	0.284	0.315	0.457	0.414	0.372	0.374	0.235	0.304	**0.195**	0.378	0.295	1.048
Electricity-FR	0.074	0.062	0.065	0.061	0.062	0.060	0.083	0.058	0.061	0.076	0.100	**0.048**	0.061	0.076	**0.041**	0.135
Electricity-NP	0.020	**0.019**	0.021	**0.019**	0.022	0.025	0.039	0.026	0.028	0.026	0.030	0.022	0.027	0.029	0.034	0.037
Electricity-PJM	0.059	0.038	0.037	0.036	0.039	0.039	0.062	0.051	0.075	0.027	0.030	0.030	0.027	**0.027**	**0.024**	0.037
Bull	0.077	**0.062**	**0.060**	0.069	0.070	0.113	0.171	0.171	0.133	0.095	0.104	0.082	0.072	0.073	0.077	0.234
Cockatoo	0.054	0.040	0.046	0.047	0.047	0.044	**0.029**	0.055	0.060	0.055	0.031	0.085	**0.026**	0.091	0.029	0.068
Covid19 (ProEnFo)	0.016	0.011	0.012	0.013	0.012	0.017	0.008	**0.006**	0.018	0.012	0.007	0.008	**0.006**	0.007	0.009	0.014
GFC12	0.064	0.037	0.035	**0.035**	0.036	**0.034**	0.047	0.040	0.047	0.040	0.056	0.037	0.039	0.042	0.042	0.064
GFC14	0.015	**0.013**	0.015	0.015	0.013	0.015	0.019	0.017	0.032	0.016	**0.012**	0.019	0.019	0.022	0.027	0.031
GFC17	0.025	0.019	0.038	0.034	0.038	0.035	0.034	0.031	0.021	**0.018**	0.046	**0.013**	0.021	0.021	0.019	0.031
Hog	0.031	**0.018**	0.019	0.020	0.019	0.024	0.034	0.043	0.037	**0.016**	0.019	0.020	0.020	0.021	0.020	0.022
PDB	0.027	0.017	**0.014**	0.015	0.015	0.015	0.019	0.019	0.020	0.018	0.025	**0.010**	0.017	0.016	0.020	0.030
Spain	0.059	0.038	0.034	0.037	0.036	0.035	0.039	0.054	0.033	0.060	0.046	**0.023**	0.071	0.038	**0.030**	0.057
Rideshare	**0.177**	**0.178**	0.184	0.186	0.198	0.182	0.192	0.194	0.257	0.219	0.190	0.184	0.205	0.190	0.193	0.325
Agg. Relative Score	1.174	**0.878**	0.928	0.929	0.954	1.000	1.204	1.272	1.286	1.091	1.140	0.932	0.939	1.094	**0.926**	1.752
Avg. Rank	9.722	**4.778**	**5.722**	5.889	6.611	6.778	11.056	10.556	11.706	9.389	9.667	6.556	7.056	9.389	6.889	14.000

Table 33: MASE scores per model in real datasets with covariates. Models achieving the **first**, **second**, and third best scores have been highlighted. Scores reported are averaged over three random seeds.

	Full Finetuned	Adapters Only				Pretrained Models			Baselines							
	ChronosX (FF)	ChronosX	ChronosX (OL)	ChronosX (NL)	ChronosX (NL-NR)	Chronos Small	Moirai (Base)	Moirai (Large)	AutoARIMA	TFT	DeepAR	PatchTSTx	NBEATSx-G	NBEATSx-I	NHITS	Seasonal Naive
ETT (15 Min.)	0.338	0.213	0.231	**0.201**	0.216	0.224	0.259	0.513	0.225	0.269	0.790	0.451	**0.208**	0.404	0.239	0.677
ETT (Hourly)	0.354	0.352	0.330	0.327	**0.318**	**0.297**	0.489	0.900	0.469	0.599	0.351	0.463	0.752	0.677	0.443	0.950
M5	**0.907**	0.925	0.923	0.921	0.918	0.919	0.959	0.925	-	**0.911**	0.938	1.038	0.946	0.953	1.326	1.268
Electricity-BE	0.432	0.374	0.387	0.379	0.377	0.380	0.458	0.510	0.306	0.621	0.568	0.292	**0.276**	0.406	**0.276**	0.405
Electricity-DE	1.001	**0.378**	0.472	0.486	0.563	0.636	0.970	0.820	0.699	0.671	0.458	0.625	**0.381**	0.646	0.526	1.582
Electricity-FR	0.676	0.570	0.575	0.561	0.560	0.548	0.801	0.520	0.532	0.691	0.881	**0.451**	0.562	0.628	**0.382**	0.966
Electricity-NP	0.548	**0.518**	0.567	**0.535**	0.595	0.654	1.070	0.674	0.717	0.683	0.769	0.560	0.685	0.682	0.869	0.781
Electricity-PJM	0.389	0.261	0.255	0.246	0.258	0.267	0.372	0.308	0.194	0.155	0.166	0.176	**0.154**	0.162	**0.147**	0.169
Bull	0.830	**0.601**	**0.537**	0.620	0.634	0.990	1.577	1.497	1.248	0.851	0.940	0.783	0.723	0.727	0.769	1.681
Cockatoo	0.911	0.652	0.717	0.730	0.725	0.706	**0.324**	0.797	0.881	0.761	0.442	1.107	**0.308**	1.184	0.471	0.737
Covid19 (ProEnFo)	0.407	0.269	0.289	0.318	0.308	0.411	0.183	**0.141**	0.423	0.330	0.204	0.224	**0.157**	0.174	0.214	0.276
GFC12	1.193	0.738	0.693	0.693	0.709	**0.657**	0.769	0.799	0.718	0.738	1.138	**0.599**	0.709	0.667	0.688	0.925
GFC14	0.499	0.414	0.439	0.434	**0.391**	0.433	0.517	0.501	1.010	0.468	**0.339**	0.524	0.521	0.587	0.693	0.722
GFC17	0.525	0.474	0.775	0.728	0.784	0.732	0.782	0.644	0.535	0.421	0.904	**0.303**	0.486	0.430	**0.413**	0.533
Hog	0.468	0.284	0.286	0.268	0.262	0.302	0.378	0.574	0.720	**0.239**	0.261	0.339	0.324	0.317	0.308	**0.227**
PDB	0.658	0.404	0.348	0.370	0.348	0.355	0.347	0.432	0.497	0.398	0.580	**0.208**	0.352	**0.297**	0.388	0.545
Spain	0.834	0.589	0.425	0.538	0.528	0.454	0.505	0.780	0.390	0.817	0.585	**0.304**	0.850	0.482	**0.357**	0.606
Rideshare	**0.461**	0.467	0.482	0.486	0.528	0.472	0.473	0.476	0.630	0.528	0.475	**0.460**	0.497	0.470	0.469	0.630
Agg. Relative Score	1.227	0.911	0.934	0.943	0.958	1.000	1.123	1.227	1.151	1.062	1.092	0.905	**0.898**	1.016	**0.902**	1.341
Avg. Rank	10.667	**6.056**	6.833	6.556	6.778	7.167	10.611	10.889	10.559	9.556	9.333	6.556	7.056	8.278	**6.500**	12.306

I.7 Probabilistic Forecasts of Chronos Variants

From Figure 17 to 33, we plot the probabilistic forecast of different Chronos variants.

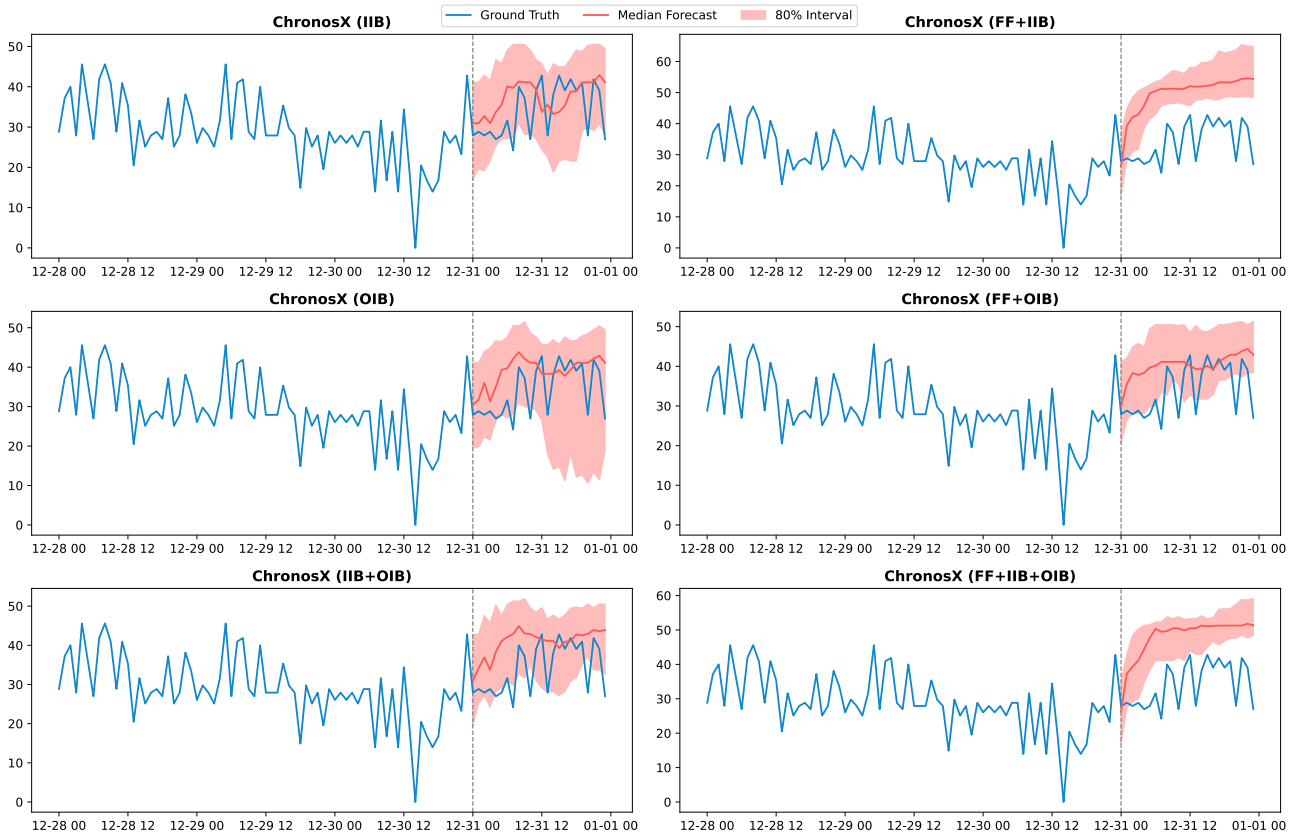


Figure 17: Forecasts on Bull Dataset

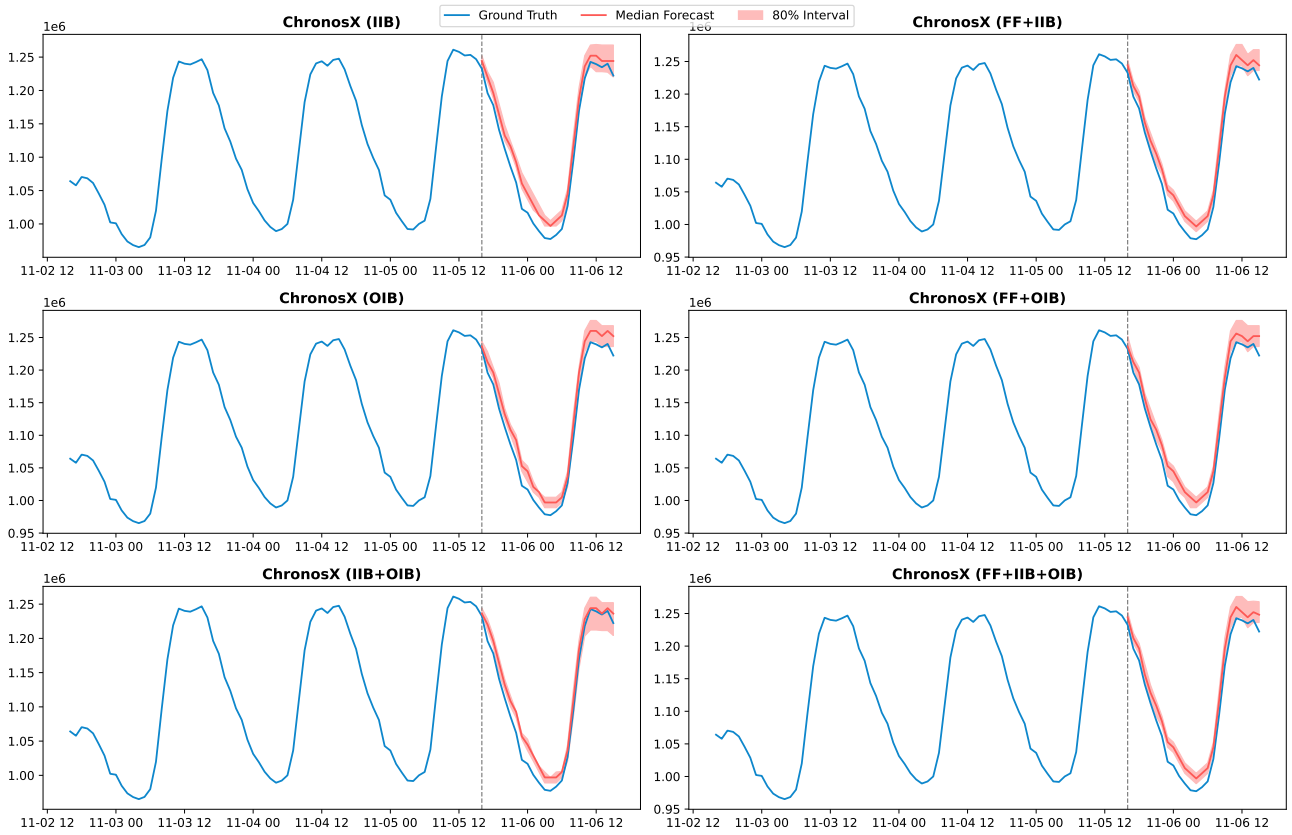


Figure 18: Forecasts on Covid19 Dataset

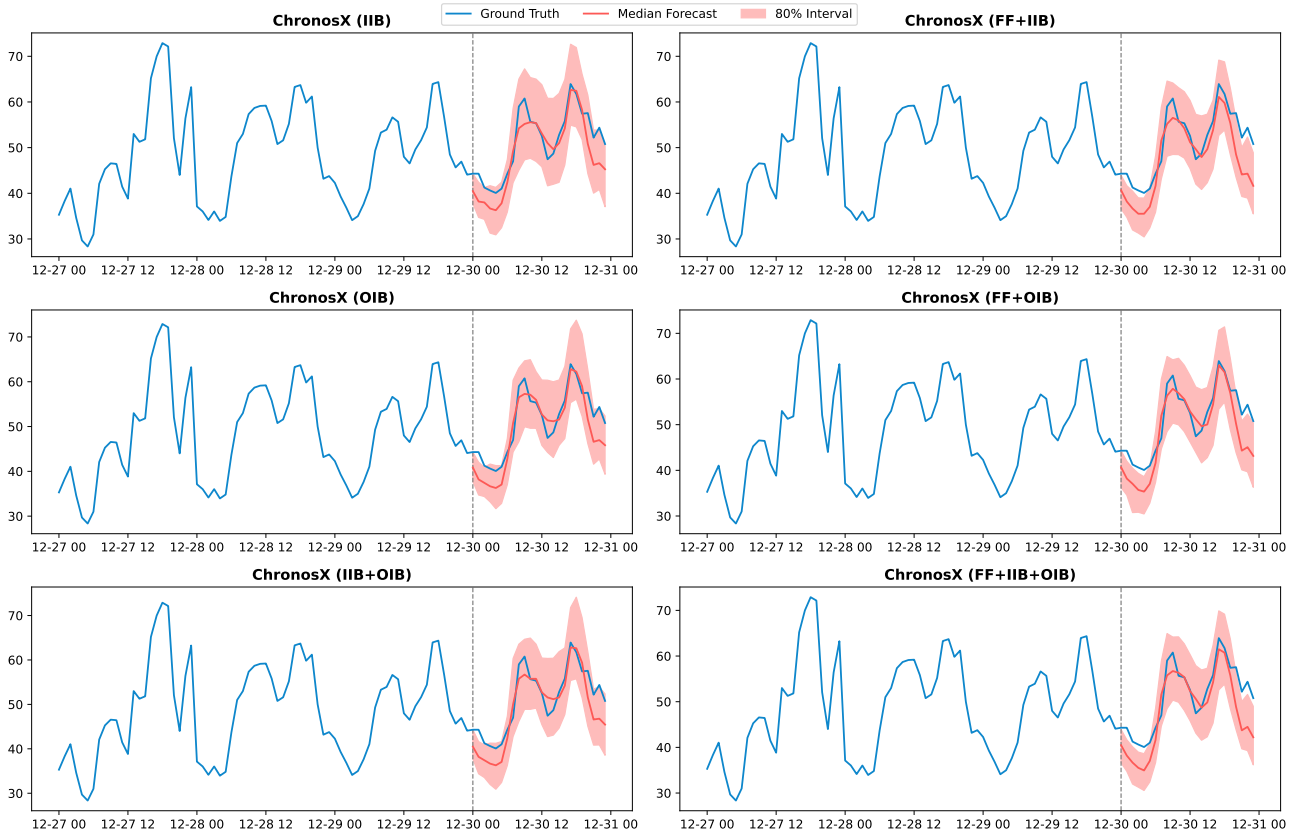


Figure 19: Forecasts on Electricity-BE Dataset

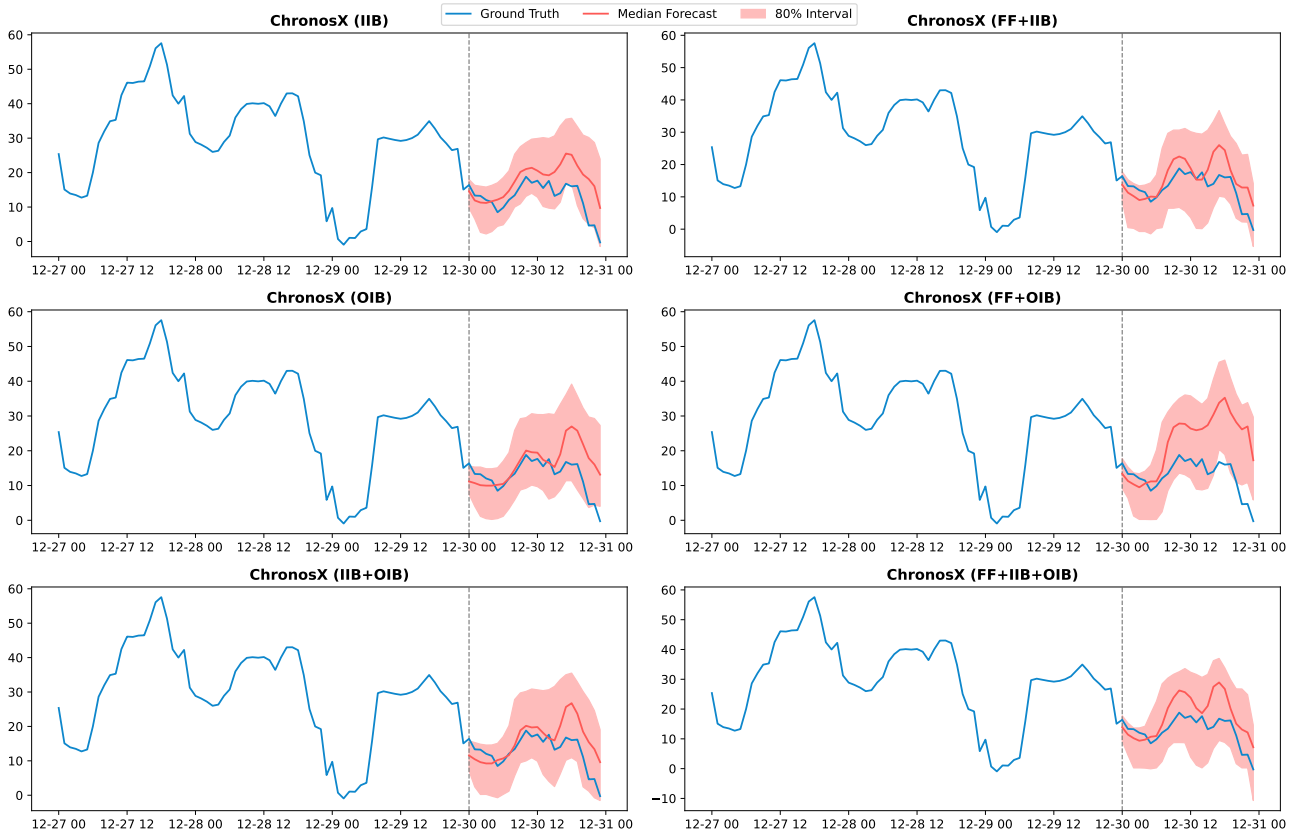


Figure 20: Forecasts on Electricity-DE Dataset

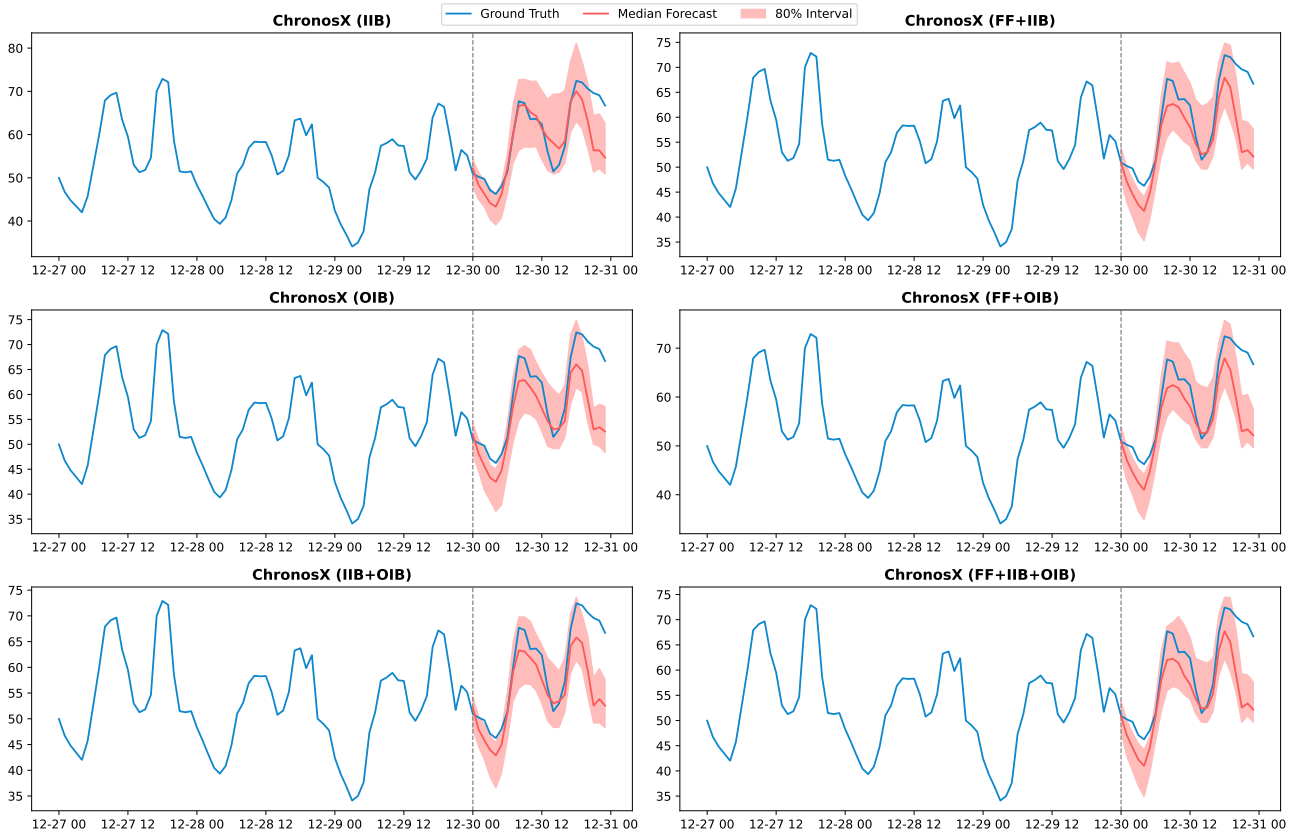


Figure 21: Forecasts on Electricity-FR Dataset

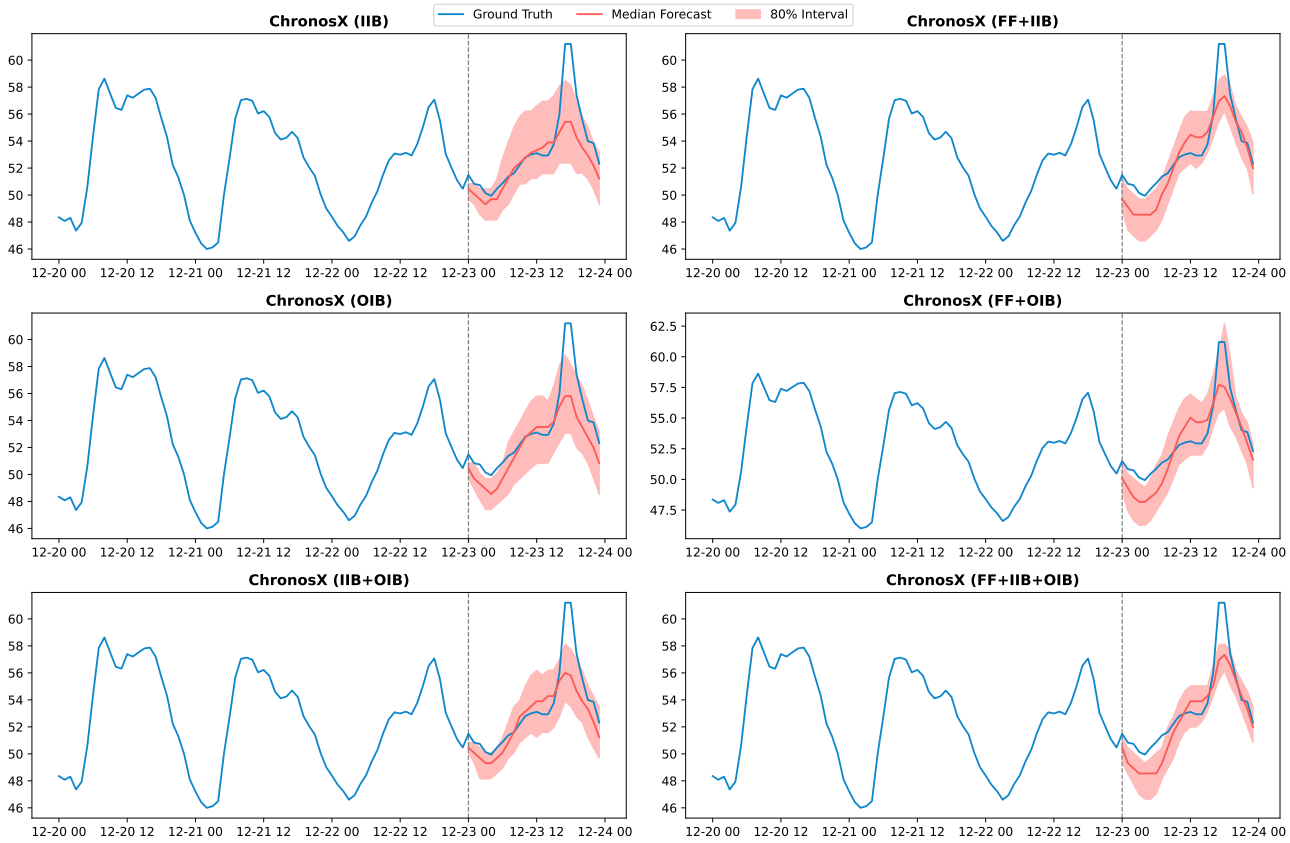


Figure 22: Forecasts on Electricity-NP Dataset

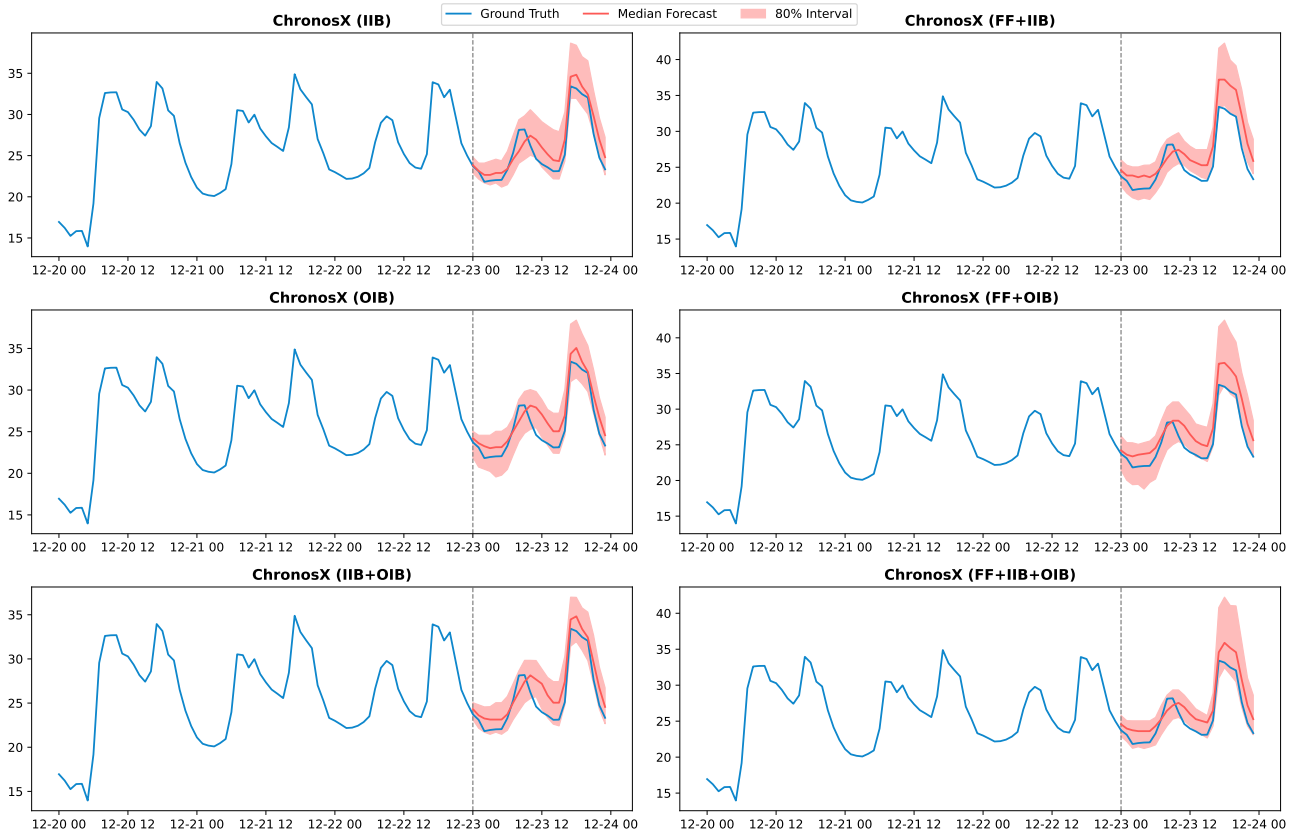


Figure 23: Forecasts on Electricity-PJM Dataset

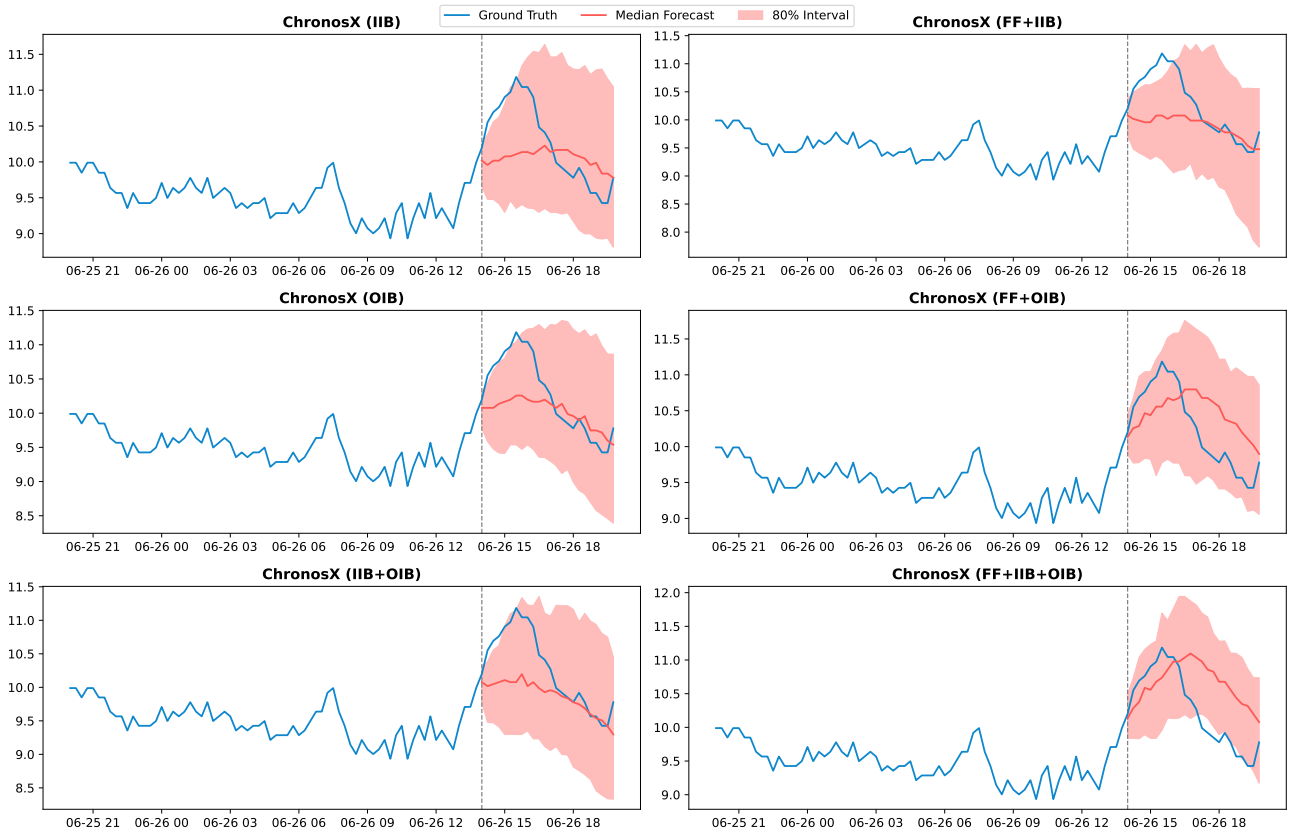


Figure 24: Forecasts on ETT (15 Min.) Dataset

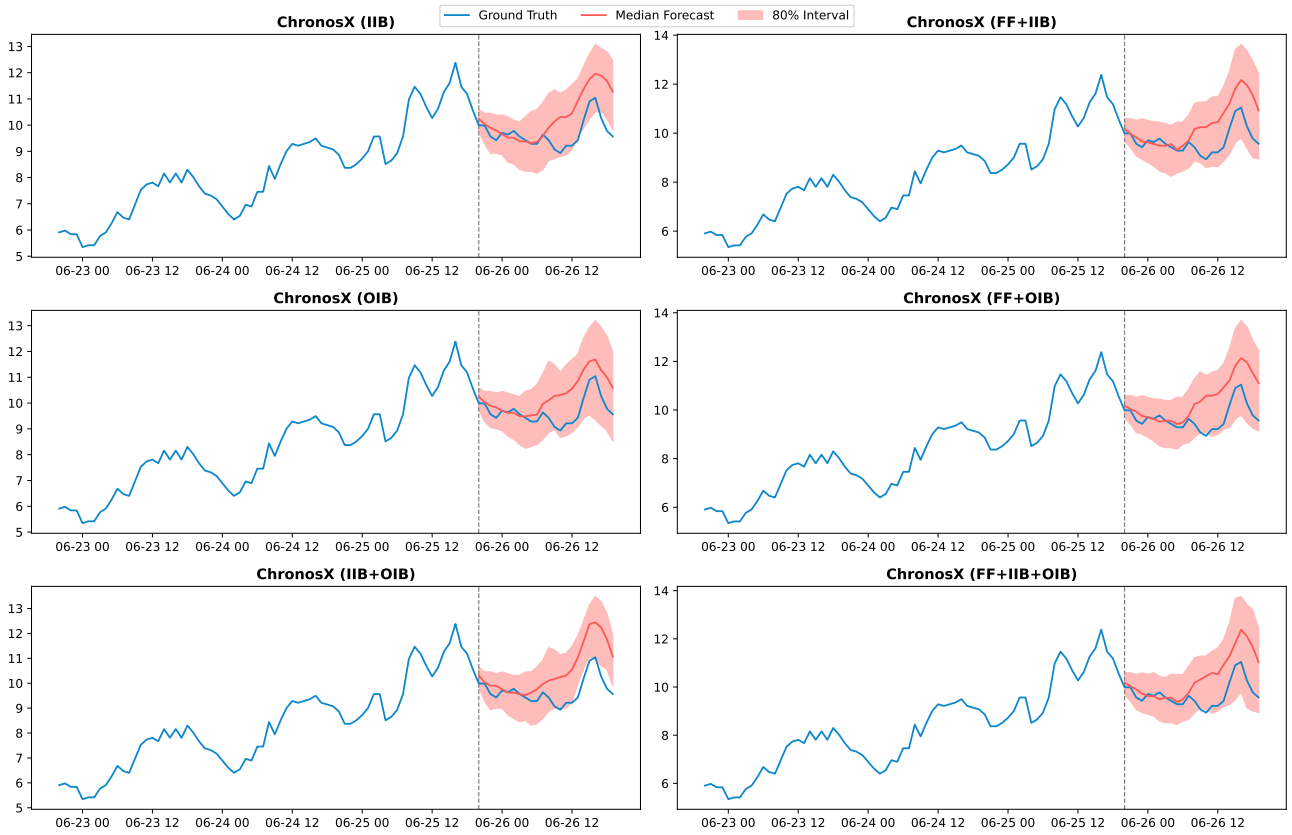


Figure 25: Forecasts on ETT (Hourly) Dataset

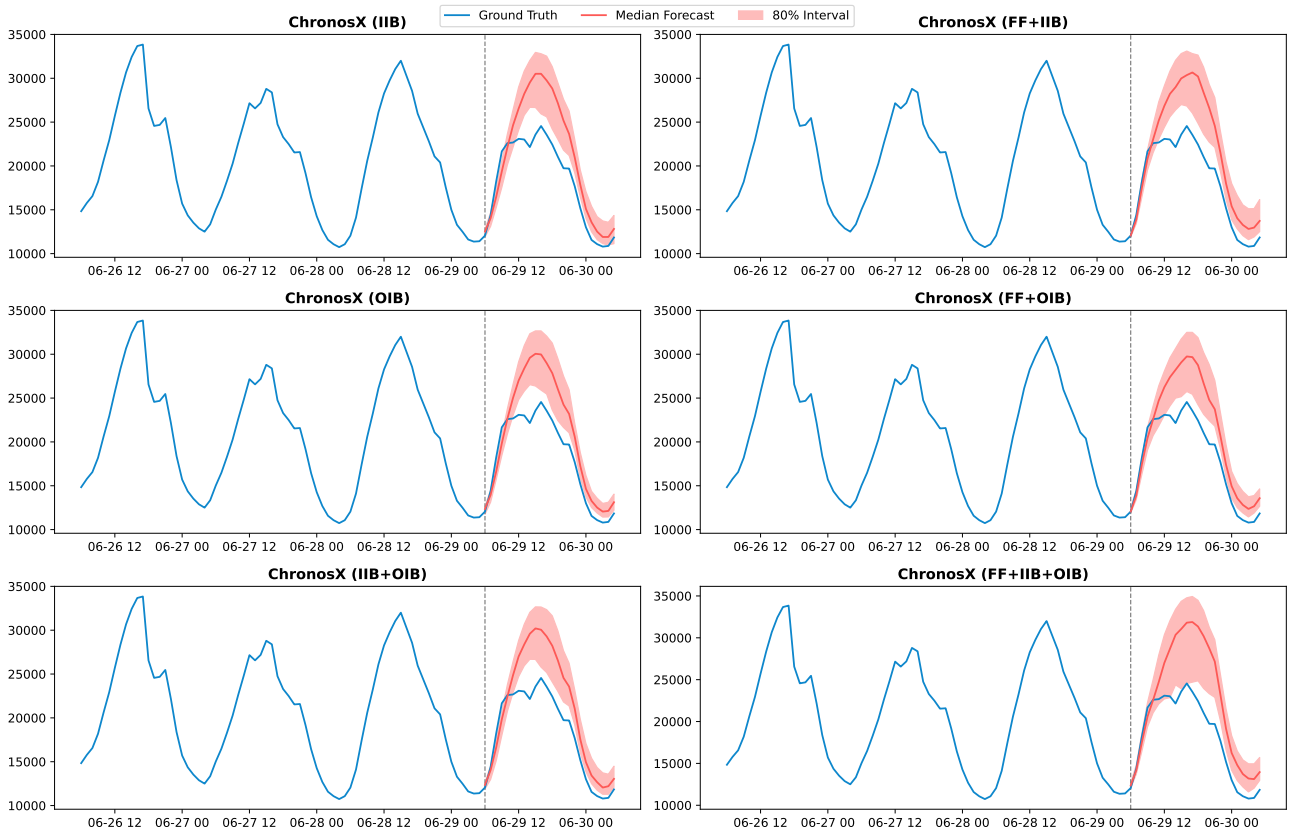


Figure 26: Forecasts on GEF12 Dataset

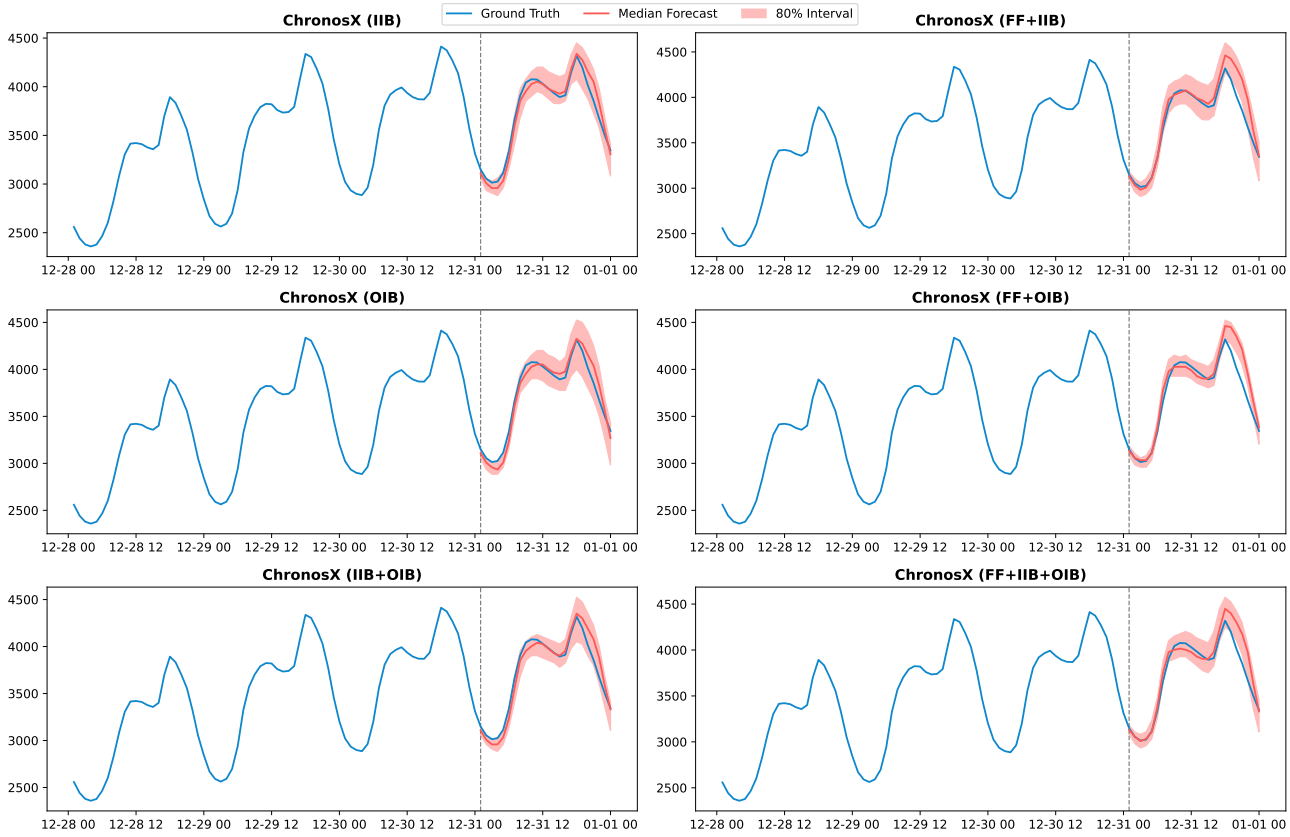


Figure 27: Forecasts on GEF14 Dataset

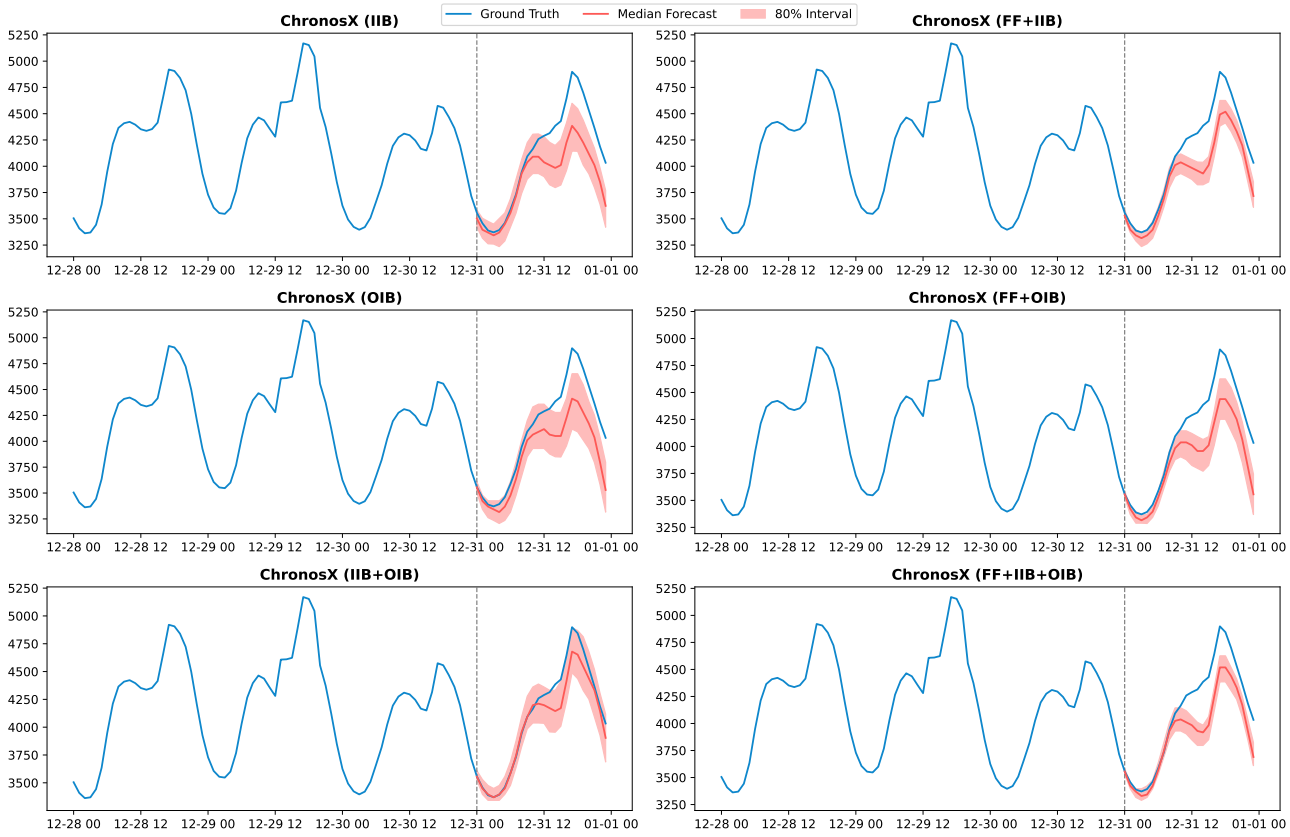


Figure 28: Forecasts on GEF17 Dataset

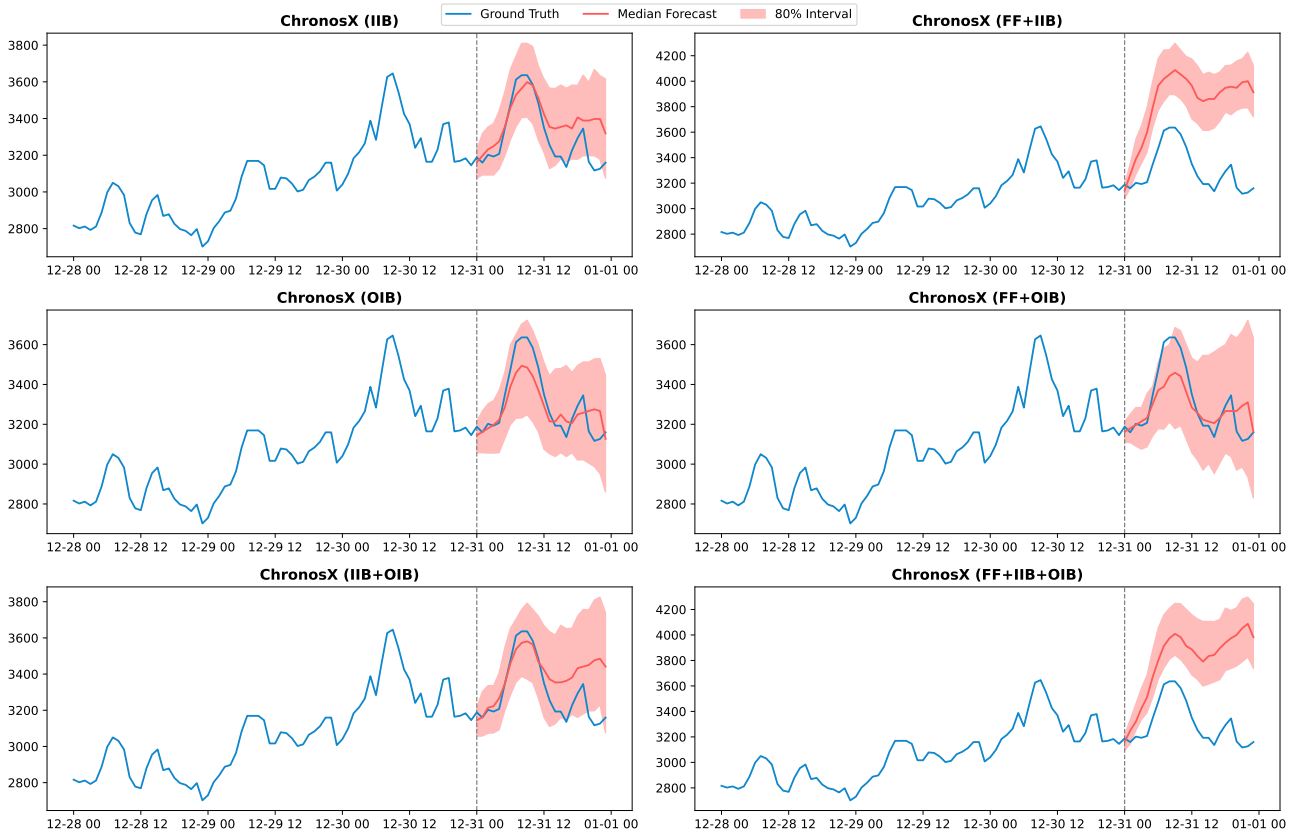


Figure 29: Forecasts on Hog Dataset

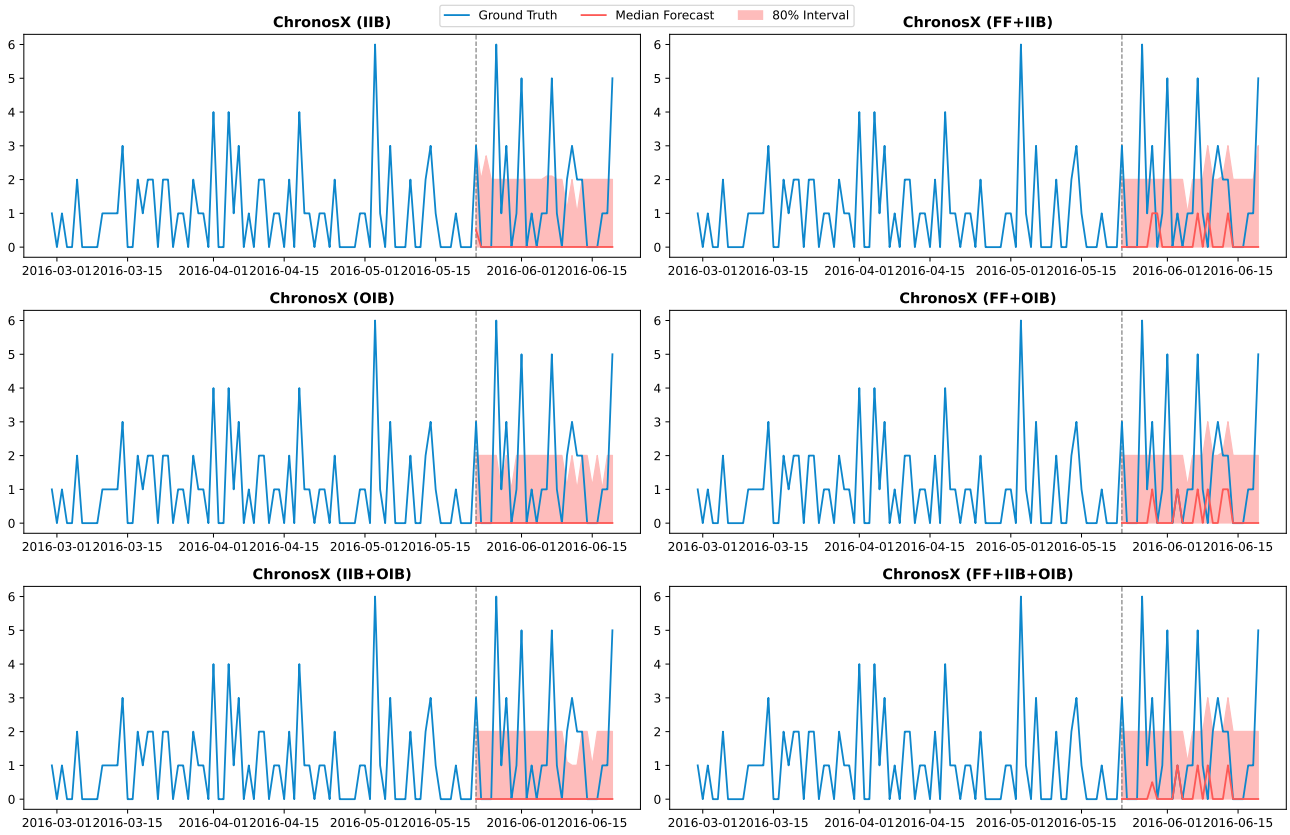


Figure 30: Forecasts on M5 Dataset

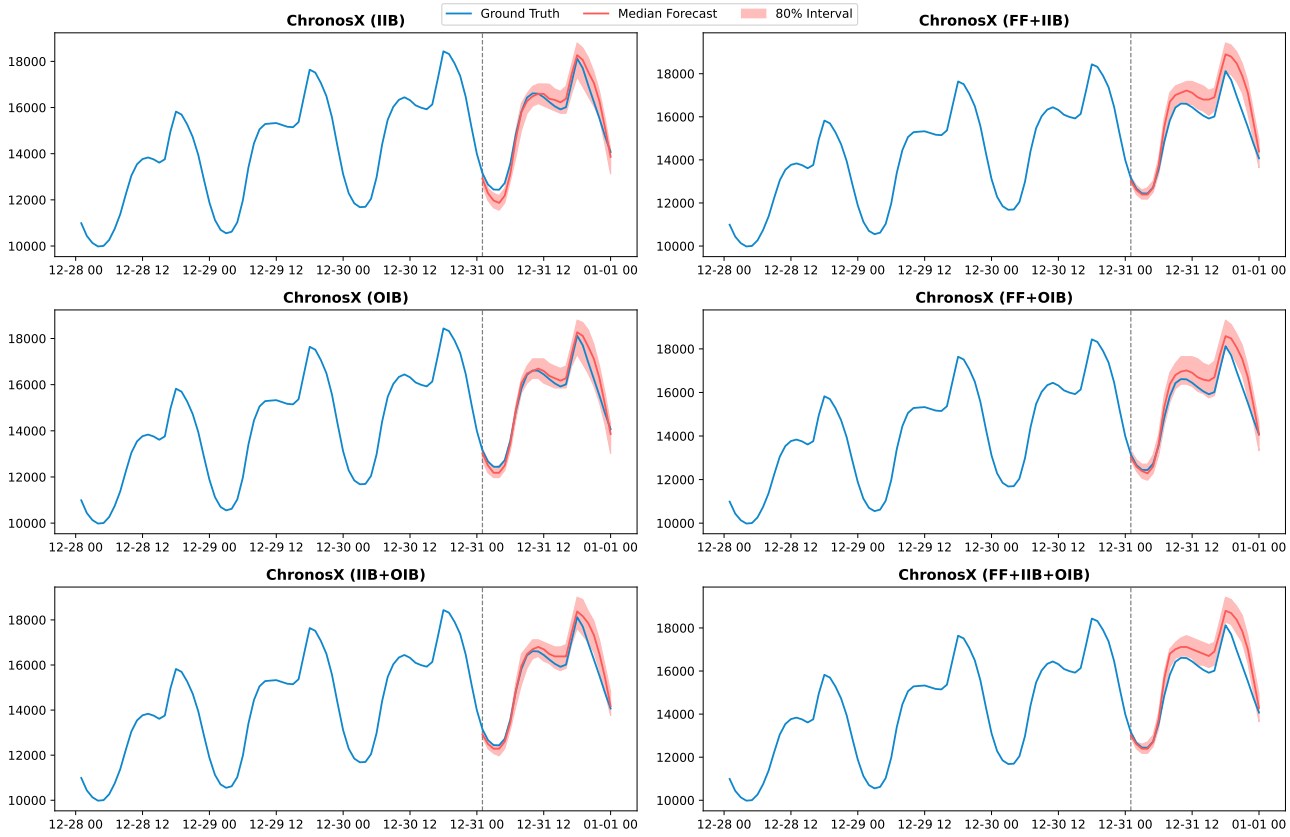


Figure 31: Forecasts on PDB Dataset

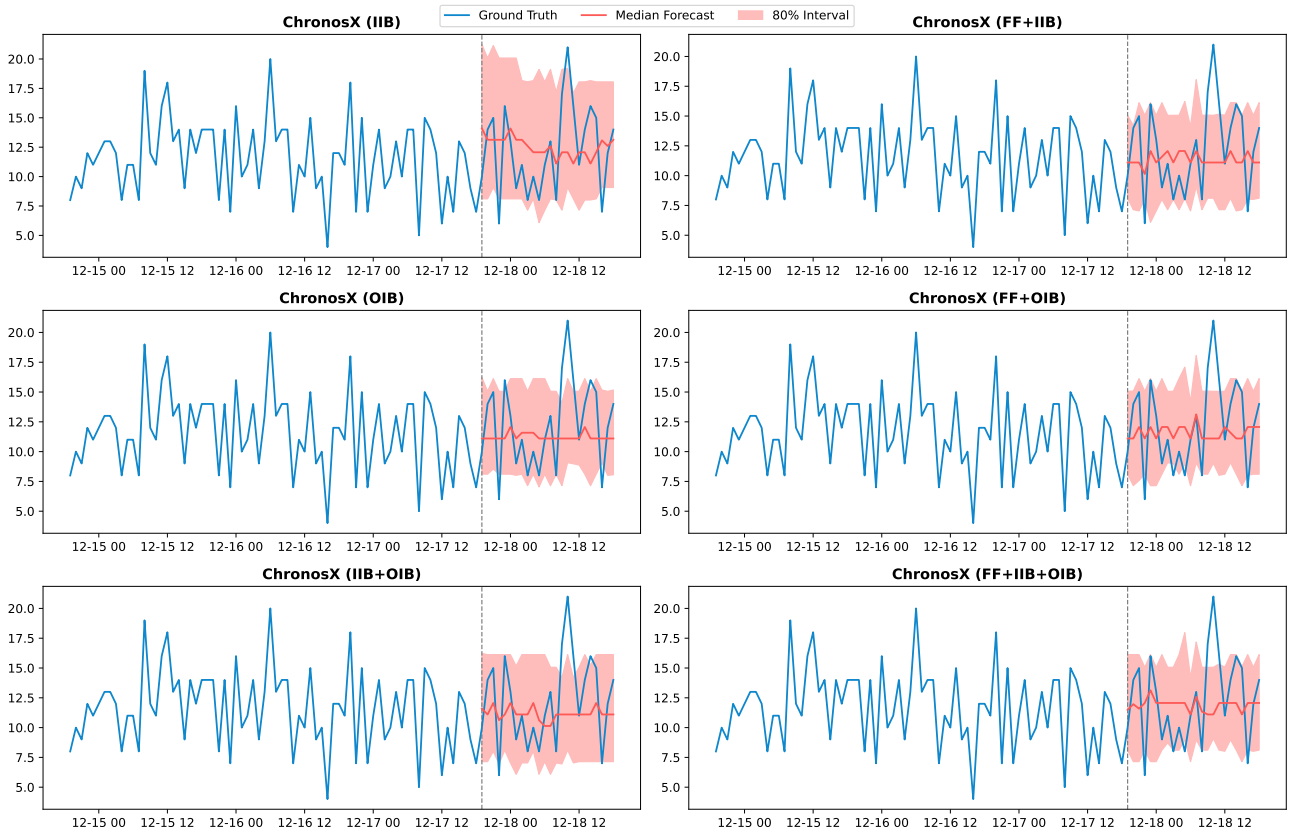


Figure 32: Forecasts on Rideshare Dataset

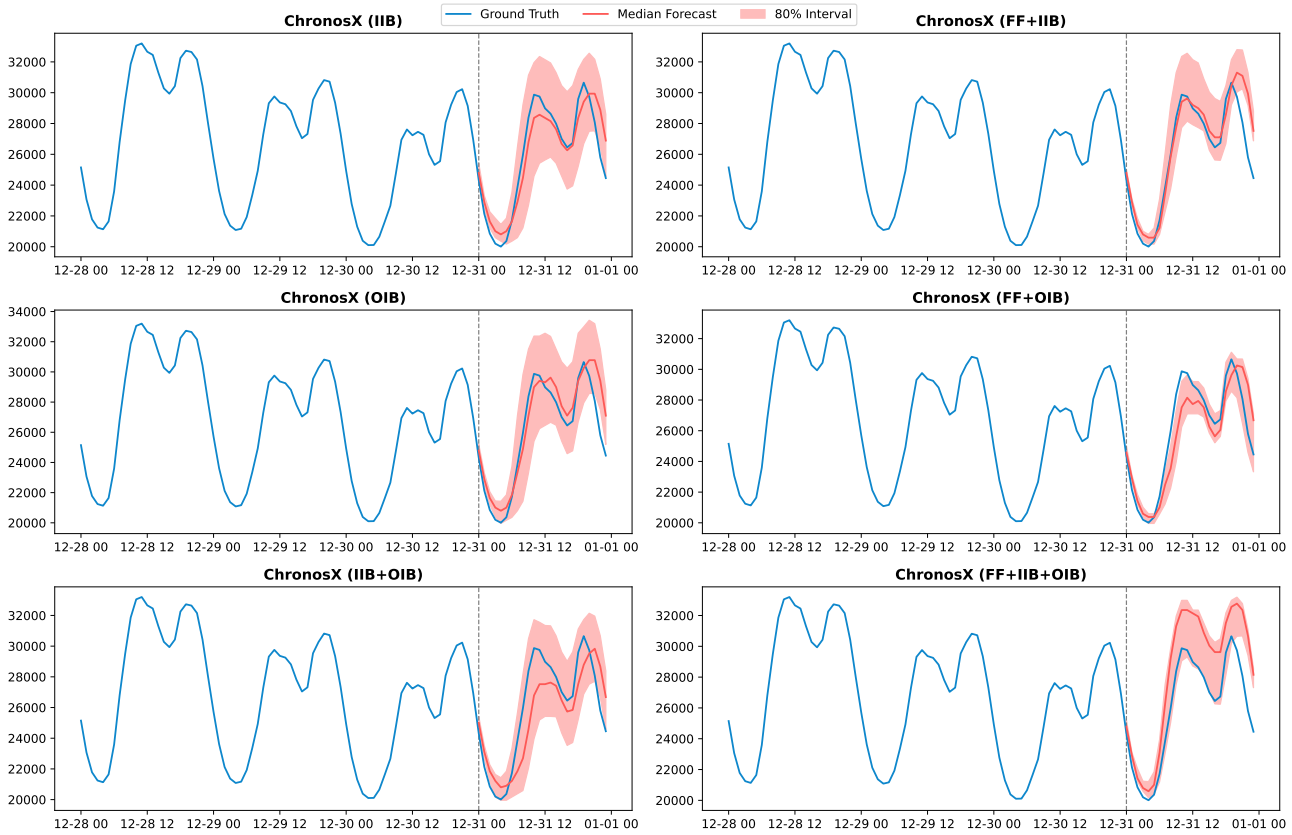


Figure 33: Forecasts on Spanish Dataset