PERFORMANCE OF ZERO-SHOT TIME SERIES FOUNDATION MODELS ON CLOUD DATA

William Toner*, **Thomas L. Lee***†, **Artjom Joosen, Rajkarn Singh, Martin Asenov**
Systems Infrastructure Lab, Edinburgh Research Centre, Huawei
william.toner2@huawei.com, T.L.Lee-1@sms.ed.ac.uk

ABSTRACT

Time series foundation models (FMs) have emerged as a popular paradigm for zero-shot multi-domain forecasting. FMs are trained on numerous diverse datasets and claim to be effective forecasters across multiple different time series domains, including cloud data. In this work we investigate this claim, exploring the effectiveness of FMs on *cloud data*. We demonstrate that many well-known FMs fail to generate meaningful or accurate zero-shot forecasts in this setting. We support this claim empirically, showing that FMs are outperformed consistently by simple linear baselines. We also illustrate a number of interesting pathologies, including instances where FMs suddenly output seemingly erratic, random-looking forecasts. Our results suggest a widespread failure of FMs to model cloud data.

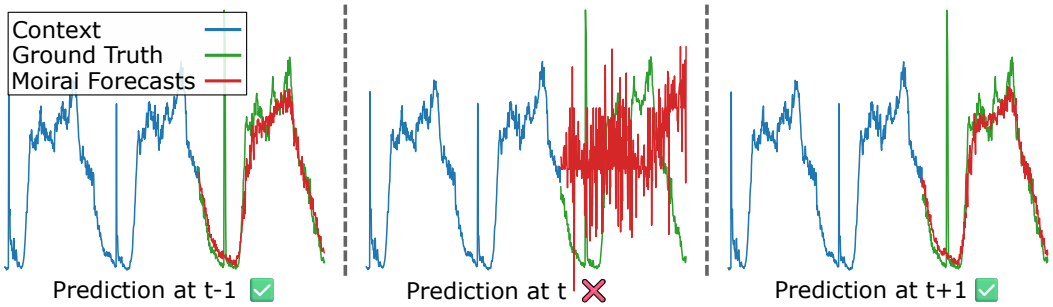

Figure 1: **Demonstration of the pathological behaviour of zero-shot FMs on cloud data:** The figure shows three consecutive forecasts for the Moirai FM on the Huawei Cloud D2 dataset. The forecasts shown are those produced when starting forecasting at the t-1, t, and t+1 time steps, where $t = 7130$. The blue curves show the context Moirai is given to construct the forecast. The plot shows that with only a small change in the context, Moirai's forecasts can change from predicting reasonably (the t-1 and t+1th time steps) to giving inaccurate and chaotic forecasts (the tth time step).

1 INTRODUCTION

Recently, there has been a growing trend in developing foundation models for time series forecasting (Liang et al., 2024; Ye et al., 2024). *Foundation models* (FMs) are neural networks pretrained on large datasets from diverse time series domains. One of the main motivations of FMs is that they claim to generalise zero-shot (i.e., without training data) to a wide range of time series domains, including cloud data (Woo et al., 2024; Li et al., 2024). Time series forecasting is important in cloud settings (Joosen et al., 2023; Diao et al., 2024). For example, forecasting allows cloud providers to scale up resources pre-emptively and reduce costs associated with over-provisioning (Joosen et al., 2024). Furthermore, zero-shot forecasting is needed in cloud settings since it is often the case for cloud data that many time series are created and removed in short periods of time (Darlow et al., 2024). While the cloud setting is a potentially impactful use-case for FMs, it has been, to the best of our knowledge, rarely evaluated in the

*Equal contribution †Work done while an intern at Huawei

FM literature (Ansari et al., 2024; Rasul et al., 2023; Chaudhry et al., 2019). In this work we aim to mend this gap by exploring the question of *how well do zero-shot foundation models generalise to cloud data?*

To examine the behaviour of (zero-shot) FMs on cloud time series, we perform experiments on function demand data drawn from real-world usage. Our results show that FMs perform poorly, being outperformed by the simple baselines of a linear model and a naive seasonal forecaster. This draws into question the claim that FMs can generalise to cloud data out-of-the-box. Instead we find that the particular characteristics of cloud data—such as spikeyness (Joosen et al., 2024; Diao et al., 2024)—can make FMs forecast badly. For instance, Figure 1 illustrates the pathological behaviour of the Moirai FM on typical cloud data. The figure demonstrates a minor change in the time series context leading Moirai to give erratic forecasts. We examine the performance of other FMs and find similar irrational forecasts on cloud data, as shown in Figure 2. Additionally, we show that the best performing FM in our experiments, VisionTS, performs well due to the fact it gives near-identical forecasts to a naive seasonal forecaster. Overall, our results suggest that current FMs are ineffective for forecasting cloud data.

2 Related work and Preliminaries

Zero-Shot FMs Recently, many (zero-shot) FM models have been proposed for time series forecasting. Most of these models follow the same general design principles of: **a)** being a LLM-like transformer model and **b)** being pretrained on a large corpus of time series data, consisting of billions of training points (Rasul et al., 2023; Chen et al., 2024; Liang et al., 2024). For instance, *Chronos* (Ansari et al., 2024), *Moirai* (Woo et al., 2024) and *TimesFM* (Das et al., 2024) are FMs which adhere to both of these design principles. However, some newer FMs are designed differently, using non-transformer architectures and/or not being trained on real-world time series data. For example, TTM (Ekambaram et al., 2024), VisionTS (Chen et al., 2024) and Mamba4Cast (Bhethanabhotla et al., 2024) all use non-transformer architectures. Additionally, Mamba4Cast is trained only on synthetic time series, while VisionTS is not trained on time series at all, using an ImageNet-pretrained masked auto-encoder (He et al., 2022) as its backbone.

Rolling Window Forecasting Here, we describe the time series setting used in this work. A *time series* consists of a sequence of fixed-dimension vectors $\boldsymbol{x}_t$, indexed by the time step t. Each dimension of $\boldsymbol{x}_t$ is called a *channel*. We look at the realistic scenario of *rolling window* forecasting (Nie et al., 2022), which models the deployment stage of a forecast model. In a rolling window setting, at each time step t the forecaster gives a forecast for H time steps into the future, where H is called the *forecast horizon*. To construct the forecast, the forecaster is given a *context*: a look-back window of the L previous time-series values, $\boldsymbol{x}_{t-L},\ldots,\boldsymbol{x}_t$, where we call L the *context length*. To evaluate performance, forecasts are compared against the *ground truth*, $\boldsymbol{x}_{t+1},\ldots,\boldsymbol{x}_{t+H}$, i.e. the actual values the time series takes for the forecasted time steps.

3 Experiments

Experimental setup We evaluate a suite of FMs on cloud time series data. We examine several of the most recent and well know FMs for time series: VisionTS (Chen et al., 2024), TTM (revision 2) (Ekambaram et al., 2024), TimesFM (Das et al., 2024), Chronos (tiny) (Ansari et al., 2024), Moirai (small) (Woo et al., 2024) and Mamba4Cast (Bhethanabhotla et al., 2024). We look at their performance on datasets constructed out of data released on Huawei Cloud function requests,[1] as described in Appendix B.1. We name the datasets D1, D2, D3 and D4, each of which correspond to data from a unique data centre in the Huawei Cloud. We note that, none of the FMs looked at have been pretrained on this data or cloud demand data more generally. As explained in Section 2, we perform experiments using a rolling window evaluation. This is done with a context length of 520 and for three different forecast horizons: 30,96,336. We selected these values as they have been widely used in previous works (Ekambaram et al., 2024; Ansari et al., 2024).

Baselines We compare the FMs against two simple baselines: **a)** A per-channel online *linear model* which is fit using ridge regression (Hamilton, 1994). This model is refit on all available data every 200 time steps. **b)** A *naive seasonal forecaster* which forecasts by copying repeatedly the last seasonal period in the context as its forecast (Ansari et al., 2024). For example, for hourly data with a daily periodicity, the naive seasonal forecaster would predict by taking the last 24 hours and repeating it over the forecast window. We note that the naive seasonal forecaster does not need to have access to any training data. More details on the baselines can be found in Appendix B.2.

[1] Available at https://github.com/sir-lab/time-series-fm-dataset

Table 1: **MASE results of zero-shot FMs forecasts for cloud data compared to the baselines of an online linear model and a naive seasonal forecaster.** The table shows that the baseline methods perform the best across all datasets and forecast horizons (H), demonstrating that currently, zero-shot FMs struggle to perform well on cloud data.

				Zero-Shot FMs					
Dataset	H	**Linear**	**Seasonal**	**VisionTS**	**TTM**	**TimesFM**	**Chronos**	**Moirai**	**Mamba4Cast**
D1	30	**1.404**	1.577	1.756	1.596	3.020	2.457	1.972	2.837
	96	1.994	**1.875**	1.966	2.275	5.754	5.622	2.656	5.618
	336	3.175	**2.809**	2.872	3.398	7.225	8.068	3.585	7.565
D2	30	**1.170**	1.231	1.686	1.844	2.772	2.454	2.443	3.207
	96	**1.550**	1.398	1.711	2.396	5.057	4.342	3.642	6.718
	336	2.262	**1.818**	2.107	2.999	6.346	5.725	3.983	13.268
D3	30	**1.082**	1.182	1.510	1.747	2.181	2.037	2.187	2.500
	96	**1.153**	1.179	1.431	1.943	3.466	2.806	2.682	3.994
	336	1.290	**1.236**	1.452	2.128	3.909	3.260	2.573	5.699
D4	30	1.578	**1.349**	1.618	2.618	2.731	2.951	2.843	3.546
	96	1.578	**1.299**	1.424	2.369	2.667	2.852	2.605	3.917
	336	1.695	**1.307**	1.414	2.368	2.657	2.802	2.599	5.228

Metrics We report the MASE of each approach, a common time series evaluation metric (Ansari et al., 2024). MASE is the mean absolute error (MAE) normalized by dividing by the MAE of a naive seasonal forecaster over the context, as detailed in Appendix B.4. This normalisation adjusts for varying scales across channels and over time. *We also report results using RMSSE in Appendix C.4*, which is defined analogously to MASE but with RMSE instead of MAE (Hyndman & Koehler, 2006). Both metrics yield consistent conclusions.

Results The main results of our experiments are displayed in Table 1. The table shows that none of the FMs being evaluated perform as well as either of the simple baselines we compare to. For example, the naive seasonal forecaster performs better than all the FMs across all datasets and forecast horizons. Moreover, the performance difference is often large; for example, the naive seasonal forecaster incurs a MASE typically half that of TimesFM. This suggests that currently zero-shot FMs do not generalise well to cloud data. We also note that alongside the good predictive performance of the baseline methods they are also the most computationally efficient, as demonstrated in Appendix C.1.

3.1 What do Zero-Shot FM Forecasts Look Like?

To analyse further the poor performance of zero-shot FMs in the cloud domain we plot a typical failure case of FMs in Figure 2. The figure displays the forecasts for TimesFM, TTM, Mamaba4Cast and Chronos on the third channel of the D2 dataset—forecasts for Moirai and VisionTS are given in Appendix C.2. The context (shown by the blue curve) is standard of cloud time series, having a strong seasonal pattern and large spikes (Diao et al., 2024), and the ground truth is roughly a repeat of the seasonal pattern. However, despite the regularity of the time series, all of the FMs give poor predictions, appearing to not make use of the clearly periodic behaviour. For example, three of the FMs, in Figure 2, do not identify the seasonal pattern at all; while the other, TTM, gets the rough shape correct but misses the spikes. We note that, in cloud provision settings predicting demand spikes is vital, a failure to do so can degrade service quality.

Overall, the example forecasts in Figures 1 and 2 show that for typical cloud data, zero-shot FM forecasts can fail to resemble the ground truth, even when the ground truth is readily predictable from the context given—as demonstrated by the baseline forecasts given in Appendix C.2. We emphasise that the failures shown are not unique to the displayed channels or time steps, being commonplace in our experiments, evidencing the poor overall performance of FMs on cloud data shown in Table 1.

VisionTS often performs quite well relative to the other FMs in our experiments (see Table 1). While its performance is still worse than the baselines, it is worth exploring why this is. We find in our

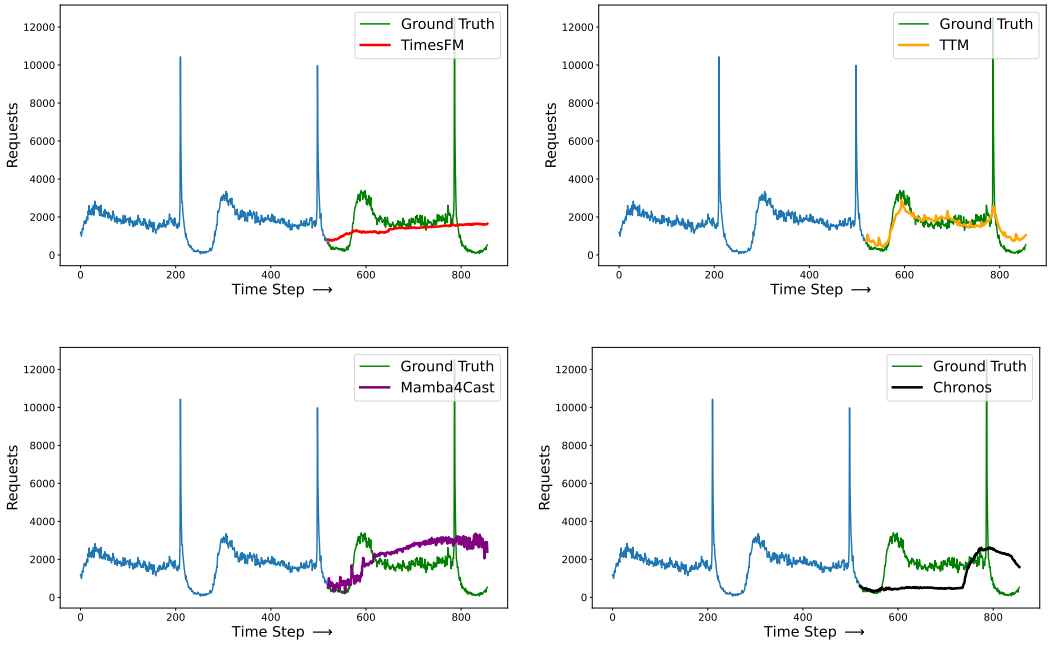

Figure 2: **FM forecasts at the same time-step ($t = 6030$) for the third channel of the Huawei Cloud D2 dataset.** The plots shows the forecasts for TimesFM, TTM, Mamba4Cast and Chronos FMs. None of the FMs forecasts are accurate. TimesFM, Mamba4Cast and Chronos do not identify the seasonal pattern. While, TTM performs better than the rest but does not forecasts the periodic spike in demand, incurring a large inaccuracy at that point.

explorations that VisionTS behaves very similarly to a naive seasonal forecaster in most cases, simply coping the context as its forecast. An example of this behaviour is shown in Figure 3. Since most channels are relatively periodic this strategy performs well, largely explaining the good performance of this model. To analyse this further, we computed the normalised (zero-lag) cross-correlation between the forecasts of VisionTS and the naive seasonal forecaster on D1 *obtaining a median correlation of 0.9992*.

4 Conclusions

In this work we examined the performance of zero-shot time series foundation models (FMs) in the cloud domain. While time series FMs claim to generalise to many domains including cloud (Woo et al., 2024), we find that FMs perform worse than the simple baselines of an online linear model and naive seasonal forecaster on cloud data. We also present evidence of pathological behaviours of FMs, demonstrating they can produce chaotic (as in Figure 1) or illogical (as in Figure 2) forecasts. Finally, we show that the best performing FM, VisionTS, performs well due to giving very similar forecasts to a naive seasonal forecaster. The results suggest that the particular characteristics of cloud data—such as spikiness (Diao et al., 2024)—can make FMs produce bad zero-shot forecasts. This questions their current applicability to cloud forecasting problems.

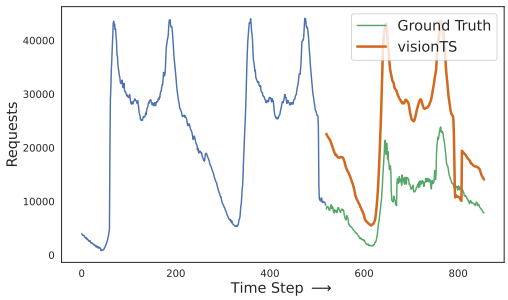

Figure 3: **An example forecast for VisionTS, demonstrating that it is roughly equivalent to a naive seasonal forecaster for cloud data.** Because the VisionTS forecast is roughly the same as the last seasonal period in the context, it is very similar to the naive seasonal forecast.

ACKNOWLEDGMENTS

We would like to kindly thank Luke Darlow for his help and Thomas L. Lee would like to thank his PhD supervisor, Amos Storkey, for his general guidance on research practice.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: a Benchmark for General Time Series Forecasting Model Evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle Maddix Robinson, Hao Wang, Michael Mahoney, Kari Torkkola, Andrew Wilson, Michael Bohlke-Schneider, and Yuyang (Bernie) Wang. Chronos: Learning the Language of Time Series. *Transactions of Machine Learning Research*, 2024.

Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. Mamba4Cast: Efficient Zero-Shot Time Series Forecasting with State Space Models. In *Proceedings of the 2024 NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On Tiny Episodic Memories in Continual Learning. *arXiv preprint arXiv:1902.10486*, 2019.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual Masked Autoencoders Are Free-lunch Zero-shot Time Series Forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

Luke Darlow, Qiwen Deng, Ahmed Hassan, Martin Asenov, Rajkarn Singh, Artjom Joosen, Adam Barker, and Amos Storkey. Dam: Towards a Foundation Model for Time Series Forecasting. In *Proceeding of the Twelfth International Conference on Learning Representations*, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A Decoder-only Foundation Model for Time-series Forecasting. In *Proceedings of the Forty-first International Conference On Machine Learning*, 2024.

Yanlei Diao, Dominik Horn, Andreas Kipf, Oleksandr Shchur, Ines Benito, Wenjian Dong, Davide Pagano, Pascal Pfeil, Vikram Nathan, Balakrishnan Narayanaswamy, et al. Forecasting Algorithms for Intelligent Resource Scaling: an Experimental Analysis. In *Proceedings of the 2024 ACM Symposium On Cloud Computing*, pp. 126–143, 2024.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H Nguyen, Wesley M Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/few-shot Forecasting of Multivariate Time Series. *CoRR*, 2024.

James D. Hamilton. *Time series analysis*. Princeton university press, 1994.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Rob J Hyndman and Anne B Koehler. Another Look At Measures of Forecast Accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

Artjom Joosen, Ahmed Hassan, Martin Asenov, Rajkarn Singh, Luke Darlow, Jianfeng Wang, and Adam Barker. How Does It Function? Characterizing Long-term Trends in Production Serverless Workloads. In *Proceedings of the 2023 ACM Symposium On Cloud Computing*, pp. 443–458, 2023.

Artjom Joosen, Ahmed Hassan, Martin Asenov, Rajkarn Singh, Luke Darlow, Jianfeng Wang, Qiwen Deng, and Adam Barker. Serverless Cold Starts and Where to Find Them. *arXiv preprint arXiv:2410.06145*, 2024.

Thomas L Lee, William Toner, Rajkarn Singh, Artjom Joosem, and Martin Asenov. Lightweight Online Adaption for Time Series Foundation Model Forecasts. *arXiv preprint arXiv:2502.12920*, 2025.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, et al. Foundts: Comprehensive and Unified Benchmarking of Foundation Models for Time Series Forecasting. *arXiv preprint arXiv:2410.11802*, 2024.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation Models for Time Series Analysis: a Tutorial and Survey. In *Proceedings of the 30th ACM Sigkdd Conference On Knowledge Discovery and Data Mining*, pp. 6555–6565, 2024.

Shaohui Ma and Robert Fildes. *Large-Scale Time Series Forecasting with Meta-Learning*, pp. 221–250. 2023.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Llanguage Models: A Survey. *arXiv preprint arXiv:2402.06196*, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A Time Series is Worth 64 Words: Long-Term Forecasting with Transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards Foundation Models for Time Series Forecasting. In *R0-fomo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

William Toner and Luke Nicholas Darlow. An Analysis of Linear Time Series Forecasting Models. In *Proceeding of the Forty-first International Conference On Machine Learning*, 2024.

Barry Payne Welford. Note On a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419–420, 1962.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Forty-First International Conference On Machine Learning*, 2024.

Jiexia Ye, Weiqi Zhang, Ke Yi, Yongzi Yu, Ziyue Li, Jia Li, and Fugee Tsung. A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Mode. *arXiv preprint arXiv:2405.02358*, 2024.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

A GENERALIZATION IN ZERO-SHOT TIME SERIES FORECASTING

While in the main text we have aimed to show if and how zero-shot FMs fail to forecast cloud data here we examine the question of why. To do this we ask more generally how much can we hope to generalize in time series forecasting? In both the vision and text domain it has been shown that zero-shot FMs have good generalization capabilities (Minaee et al., 2024; Zhou et al., 2022) and hence it can be natural to suggest by using zero-shot FMs in the time domain we can achieve similar generalization capabilities. However, there are reasons to question this belief, mainly due to less constrained nature of the continuation of a time series versus the continuation of an image (i.e. inpainting) or text generation. For instance, given the same context there maybe many valid forecasts based on the time series domain the data was generated from. More concretely, in cloud data there are spikes which happen at regular intervals which should be forecasted but in other domains spikes could be caused by noise and so should not appear in the forecast. This suggests that as the claims of zero-shot time series FM generalization expands it is quickly apparent that there is no free lunch. Hence, claims of widespread generalization ability or generalization to untested domains for zero-shot forecasts should be inspected carefully.

Additionally, the above suggest that to create well performing FMs for time series data there needs to be a mechanism for *conditioning* on previous data from the time series being forecasted or from similar time series. This can be straightforwardly achieved by finetuning methods. However, this can be compute and time intensive and requires access to a significant amount of offline data. So, in real-world scenarios where time series data is generated online, shifts in distribution and compute/training time is constrained, there is a need to look at new conditioning mechanisms. This has already been looked at in some works (Lee et al., 2025; Ma & Fildes, 2023) but there is scope to do more in this area which could realize the goal of general time series FMs.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 CONSTRUCTION OF THE HUAWEI CLOUD DATASETS

Table 2: **Dataset statistics and the seasonalities used for each dataset in the computation of MASE.**

Dataset	Time-Step Period	Seasonality	#Channels	#Time Steps	Selected channels
D1	5 mins	288	17	8640	[365,917,919, 920,921,926, 929,1016,1289, 1235,976,504, 745,889,923, 958,961]
D2	5 mins	288	18	8640	[965,1089,1175, 945,450,1017, 1346,1017,1346, 1132,932,909, 1478,1465,949, 1499,975,1213]
D3	5 mins	288	7	8640	[2310,2214,2223, 129,2472,2461, 2471]
D4	5 mins	288	8	8640	[449,444,435, 487,447,438, 455,426]

To look at how well zero-shot FMs perform on cloud time series we construct a cloud dataset out of publicly accessible data from the Huawei Cloud.[2] This data consists of function requests from Huawei's

[2]The full Huawei Cloud data release is available at https://github.com/sir-lab/data-release

serverless cloud system. The data is divided into the function requests across the data centres of five regions in China. As described in more detail in Joosen et al. (2024), the data is typical of the cloud domain having large spikes and strong periodicity (Diao et al., 2024; Darlow et al., 2024; Aksu et al., 2024). From this large store of data we curate four manageable datasets from the first two and last two regions, respectively, calling them D1, D2, D3 and D4. We disregard the third region which is much smaller than the other region's datasets[3] and is generally much sparser. From the regions we selected, we downsample the number of channels by selecting ones which have a large amount of activity. We ignore channels which are near-identical copies of other selected channels, have very low activity and/or are highly random, making them unsuitable for forecasting. The indices of the selected channels are shown in the *selected channels* column of Table 2. After channel selection, the data is cleaned by filling missing values with zeros. It is then downsampled to 5-minute intervals by summing the values within each interval. Importantly, the selection of the channels and other dataset construction decisions was finalised *before* studying any forecast models and the dataset was not subsequently altered in any way when studying the performance of the FMs. We have released this dataset at https://github.com/sir-lab/time-series-fm-dataset and an overview of the statistics of each dataset is shown in Table 2.

B.2 Details of Baselines

The baselines we employ are simple, and we have included the main details in the text. However, there are three additional implementation details to mention: **a)** For numerical stability, we standardise the data when fitting the online linear model. This is achieved by normalising the data with a running standard deviation for each channel. The running standard deviation is updated online using Welford's algorithm (Welford, 1962). At inference time no data standardisation is required—as $\alpha W\left(\frac{\boldsymbol{x}}{\alpha}\right)=W\boldsymbol{x}$ for any α and where W is a matrix (Toner & Darlow, 2024). **b)** The ridge regression regularisation coefficient used when fitting the online linear model is set to 0.5 for all experiments. **c)** Before the linear model is fit for the first time it will give forecasts based on how it was initialised, in this work we initialise it to be the naive seasonal forecaster.

B.3 Details of FMs

We have aimed to used the same hyperparameters/settings as the original works for each FM looked at. However, both TTM and TimesFM were trained for a context length of 512 and therefore when using them we drop the first 8 values of each 520-long context. Also, for TTM the currently released models only predict to a horizon length of 96. Therefore, to generate the forecasts for the horizon length of 336 we auto-regressively feed-in the constructed forecast, as done in the paper proposing TTM (Ekambaram et al., 2024).

B.4 Metrics

The Mean Absolute Scaled Error (MASE) (Hyndman & Koehler, 2006) is for a given (per-channel) forecast $\hat{\boldsymbol{y}}$, context $\boldsymbol{x}$ and ground truth forecast $\boldsymbol{y}$ calculated as

$$\text{MASE}(\hat{\boldsymbol{y}},\boldsymbol{y},\boldsymbol{x})=\frac{L-S}{T}\frac{\sum_{i=1}^{T}|\hat{y}_i-y_i|}{\sum_{i=1}^{H-S}|x_i-x_{i+S}|},$$

where S represents the seasonality. As explained in the main text, this the MAE normalised using the MAE of the naive seasonal forecaster computed on the context.

The Root Mean Squared Scaled Error (RMSSE) metric (Hyndman & Koehler, 2006) is computed in the same way as MASE but with MAE replaced with MSE and with the whole formula rooted:

$$\text{RMSSE}(\hat{\boldsymbol{y}},\boldsymbol{y},\boldsymbol{x})=\sqrt{\frac{L-S}{T}\frac{\sum_{i=1}^{T}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{H-S}(x_i-x_{i+S})^2}}.$$

[3]E.g., region one has 2487 functions while Region three has 237 functions

C Additional Experiments

C.1 Computational Cost

Table 3: **Run-time of the online linear model and a naive seasonal forecaster baselines compared to the most efficient zero-shot FM, TTM.** We perform the experiments on two Intel(R) Xeon(R) Platinum 8168 CPUs. The table shows that alongside the baselines being the most accurate forecasters, they are also the most computationally efficient.

		Run-Time (seconds)		
Dataset	H	**Linear**	**Seasonal**	**TTM**
D1	30	30	20	52
	96	33	22	53
	336	39	28	202
D2	30	32	22	50
	96	34	23	44
	336	41	29	213
D3	30	12	8	17
	96	13	9	17
	336	16	11	73
D4	30	15	10	25
	96	16	10	18
	336	18	13	65

C.2 Baseline and Other FM Forecasts for Figure 2

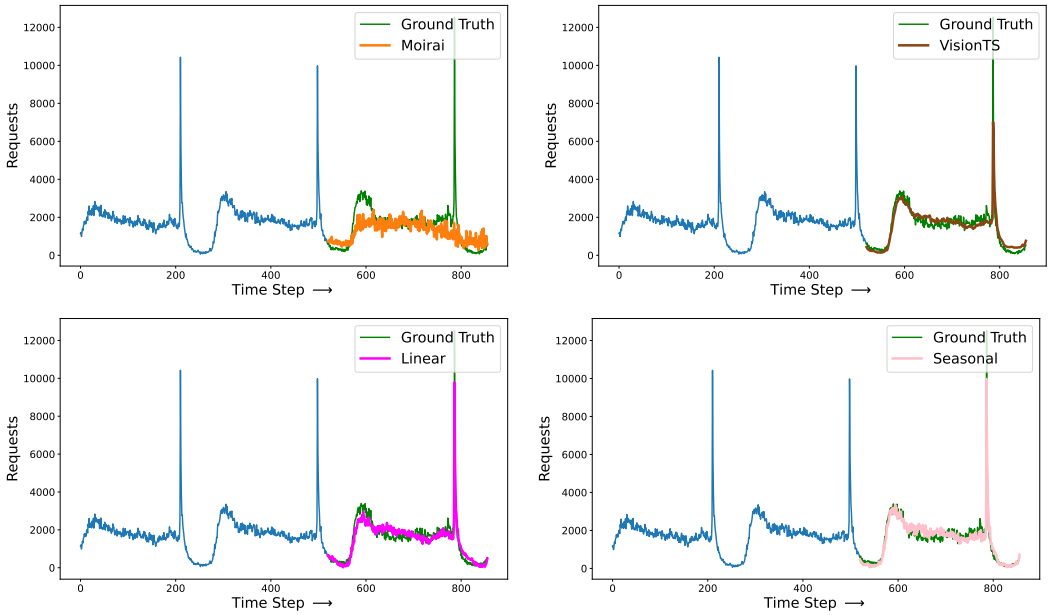

Figure 4: **FM and baseline forecasts at the same time-step ($t = 6030$) for the third channel of the Huawei Cloud D2 dataset.** These plots show the forecasts for FMs not shown in Figure 2: Moirai and VisionTS. It also shows the forecasts for the two baseline models: the online linear and naive seasonal forecasters. As in Figure 2 the FMs do not perform as well as the baselines. Moirai gives a poor forecast and does not predict the spike. While, as discussed in Section3.1, VisionTS gives a similar—but dampened and so worse—forecast to the naive seasonal forecaster. This is in contrast to the two baseline methods which predict accurately, importantly forecasting the spike in the data.

C.3 ADDITIONAL FORECASTS

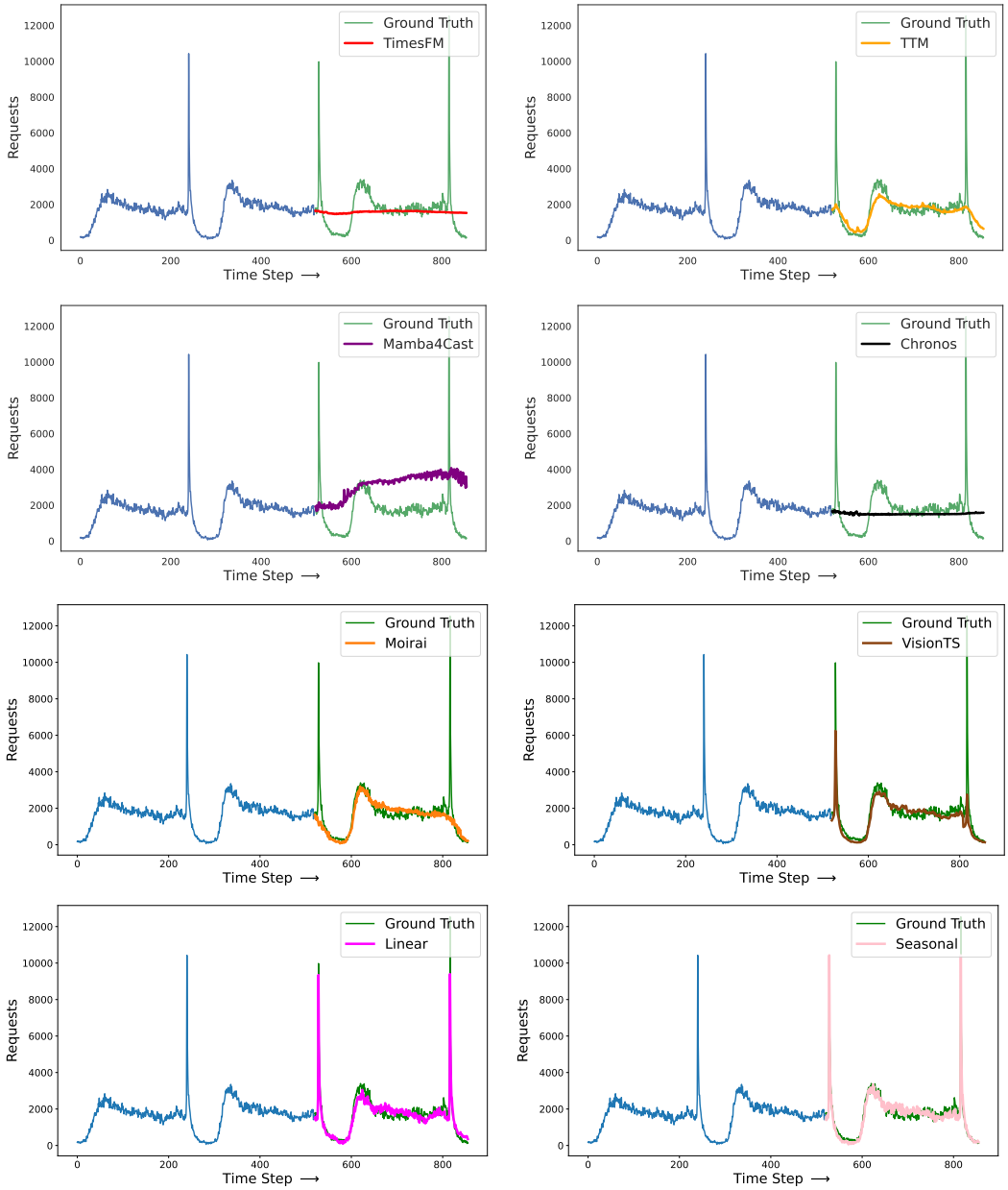

Figure 5: **FM forecasts at the same time-step ($t = 6000$) for the third channel of the Huawei Cloud D2 dataset.** The figure shows an additional set of forecasts for the FMs and the baselines. These forecasts occur 30 time steps before the ones shown in Figure 2. The figure shows that the FMs suffer from the same failures as shown in Figure 2, giving evidence that these problems occur frequently. None of the FMs perform as well as the simple baselines.

C.4 RMSSE RESULTS

Table 4: **RMSSE results of zero-shot FMs forecasts for cloud data compared to the baselines of an online linear model and a naive seasonal forecaster.** The table shows, as for MASE, that the baseline methods perform the best across all datasets and forecast horizons (H), demonstrating that currently zero-shot FMs struggle to perform well for cloud data.

				Zero-Shot FMs					
Dataset	H	**Linear**	**Seasonal**	**VisionTS**	**TTM**	**TimesFM**	**Chronos**	**Moirai**	**Mamba4Cast**
D1	30	**1.061**	1.577	1.756	1.596	3.020	2.457	1.972	2.837
	96	1.583	**1.557**	1.596	1.814	4.664	4.551	2.188	4.419
	336	2.627	**2.433**	2.463	2.905	5.990	6.685	3.145	5.950
D2	30	**0.977**	1.017	1.599	2.001	2.722	2.509	2.486	3.006
	96	1.358	**1.258**	1.927	2.782	4.819	4.398	3.782	6.137
	336	2.233	**1.908**	2.789	3.927	6.416	6.114	4.699	10.569
D3	30	**1.012**	1.098	1.380	1.608	1.968	1.841	1.995	2.108
	96	**1.200**	1.237	1.473	1.971	3.232	2.757	2.626	3.481
	336	1.484	**1.457**	1.660	2.374	4.003	3.602	2.908	4.526
D4	30	1.081	**1.053**	1.181	1.816	2.015	2.095	2.019	2.166
	96	1.194	**1.175**	1.233	1.978	2.410	2.474	2.321	2.654
	336	1.400	**1.328**	1.414	2.188	2.693	2.837	2.595	3.185