machine learning & knowledge extraction

MDPI

Article

Generalisation Bounds of Zero-Shot Economic Forecasting Using Time Series Foundation Models

Jittarin Jetwiriyanon *, Teo Susnjak * and Surangika Ranathunga

School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand; s.ranathunga@massey.ac.nz

* Correspondence: jittarin.jetwiriyanon.1@uni.massey.ac.nz (J.J.); t.susnjak@massey.ac.nz (T.S.)

Abstract

This study investigates the transfer learning capabilities of Time-Series Foundation Models (TSFMs) under the zero-shot setup, to forecast macroeconomic indicators. New TSFMs are continually emerging, offering significant potential to provide ready-trained and accurate forecasting models that generalise across a wide spectrum of domains. However, the transferability of their learning to many domains, especially economics, is not well understood. To that end, we study TSFM's performance profile for economic forecasting, bypassing the need for training bespoke econometric models using extensive training datasets. Our experiments were conducted on a univariate case study dataset, in which we rigorously back-tested three state-of-the-art TSFMs (Chronos, TimeGPT, and Moirai) under data-scarce conditions and structural breaks. Our results demonstrate that appropriately engineered TSFMs can internalise rich economic dynamics, accommodate regime shifts, and deliver well-behaved uncertainty estimates out of the box, while matching and exceeding state-of-the-art multivariate models currently used in this domain. Our findings suggest that, without any fine-tuning and additional multivariate inputs, TSFMs can match or outperform classical models under both stable and volatile economic conditions. However, like all models, they are vulnerable to performance degradation during periods of rapid shocks, though they recover the forecasting accuracy faster than classical models. The findings offer guidance to practitioners on when zero-shot deployments are viable for macroeconomic monitoring and strategic planning.

Keywords: transfer learning; GDP forecasting; time series foundation models; time series forecasting; zero-shot forecasting

Received: 1 August 2025
Revised: 17 September 2025
Accepted: 16 October 2025
Published: 3 November 2025

Citation: Jetwiriyanon, J.; Susnjak, T.; Ranathunga, S. Generalisation Bounds of Zero-Shot Economic Forecasting Using Time-Series Foundation Models. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 135. https://doi.org/10.3390/make7040135

1. Introduction

Macroeconomic indicators, such as gross domestic product (GDP), consumer price inflation, and the unemployment rate, serve as gauges for the economy's direction. These indicators distil vast amounts of data into manageable signals of aggregate demand, supply-side capacity, and financial conditions. Timely and reliable forecasts of gross domestic product (GDP) are essential inputs for policy, finance, and corporate planning worldwide. Governments embed medium-term growth scenarios in budget frameworks, debt-sustainability analyses, and multi-year spending envelopes [1]. Prudential regulators feed GDP paths into system-wide stress tests to gauge the resilience of banks and insurers under adverse macroeconomic conditions [2]. Investors rebalance portfolios when growth expectations shift, treating GDP as a succinct proxy for business-cycle momentum [3]. Multinational firms align production schedules, inventories, and staffing levels with headline

and sectoral GDP projections [4], while credit-rating agencies and development institutions incorporate forward-looking GDP assumptions into sovereign-risk assessments and concessional-finance formulas [5].

However, producing dependable forecasts remains challenging: structural breaks, measurement errors, and sudden shocks can quickly erode model performance, thereby motivating continuous innovation, from classical econometric combinations to modern machine learning approaches [6]. Forecasting GDP accurately across countries presents significant challenges. Firstly, early data releases are frequently revised, sometimes to the extent of reversing the sign of reported growth. This means models must contend with evolving truths rather than a fixed target [7]. Secondly, structural changes, such as those triggered by commodity-price super-cycles, financial crises, natural disasters, and global pandemics, can abruptly disrupt historical relationships and invalidate previously stable parameters [8–11]. Lastly, sector-level GDP series display diverse seasonal patterns and react unevenly to external-demand shocks, hindering the transfer of information from one industry to another [12].

For decades, classical time-series models, such as autoregressive integrated moving-average (ARIMA) frameworks [13] and vector autoregressions (VARs) [14], have been used as forecasting tools for this domain due to their transparency, tractability, and ease of re-estimation. However, their core assumptions of linearity and stable parameters rarely hold in practice, and sudden shifts in policy regimes and technology shocks can all invalidate coefficients calibrated on historical data, causing forecast accuracy to deteriorate rapidly as the horizon lengthens [15]. Empirical surveys have shown that beyond a few quarters ahead, even well-specified and rich multivariate systems seldom outperform simple persistence or random-walk benchmarks [16]. This challenge is further compounded by the constant moving of the goalposts in the form of data revisions, where preliminary GDP releases and other key indicators are often substantially updated, meaning models trained on early vintages chase a moving target and deliver the least reliability precisely when decision-makers most need clarity [17]. Recent studies have explored more sophisticated modelling approaches such as mixed-frequency factor and Bayesian models, which integrate hundreds of monthly indicators to sharpen nowcasts [18], as well as tree-based ensembles and hybrid neural networks capable of discovering non-linear interactions [19]. However, the more recent modelling approaches are accompanied by high overheads in terms of expertise and computation for modest returns in improved accuracies, while being dependent on provision and access to real-time multivariate inputs [20].

Recent advancements in AI have introduced a new class of forecasting tools, namely Time-Series Foundation Models (TSFMs), with the potential to mitigate longstanding challenges in macroeconomic forecasting [21]. TSFMs attempt to leverage transfer learning, which reuses representations learned on a source dataset in order to repurpose it for making forecasts on a target task. TSFMs are pre-trained on millions of heterogeneous time series with knowledge captured in parameters that in theory can then be transferred to improve performance on a different target either with or without model fine-tuning [22]. Under the zero-shot regime, no further fine-tuning is undertaken, and thus the transferability of TSFMs to accurately forecast future values can then be explored in the purest form, thus significantly reducing the modelling resource overheads [23]. These large pre-trained TSFMs, Nixtla's TimeGPT, Amazon's Chronos, and Salesforce's Moirai treat numeric sequences as language tokens and leverage transformer backbones trained on extensive datasets. TimeGPT offers a "plug-and-play" API capability providing zero-shot forecasts without local fine-tuning, while Moirai extends this paradigm to multivariate forecasting with exogenous covariates. Models like Chronos have been used in the literature in domains such as electricity, traffic, and retail data, where they have reported better performances

than tuned statistical baselines with minimal feature engineering [24–26]. Given the largely unexplored capabilities of these models in a macroeconomic context, our study seeks to explore to what degree the current cutting-edge TSFMs can generalise across data vintages, sectors, and structural breaks in economic datasets, under the "out-of-the-box", zero-shot settings. We used macroeconomic data from New Zealand as a case study, which presents an unusually demanding testbed, given the vulnerability of its small economy to external factors, as well as its exposure to commodity-price cycles, natural disasters, and external-demand shocks [27], together with the frequent revisions and thus the uncertainty of macroeconomic indicators estimates [28].

Contribution and Novelty

Macroeconomic forecasting often faces short sample histories, limited data, and tight computational budgets, all of which hinder credible assessments of uncertainty and shocks. We propose a zero-shot transfer learning with pre-trained time-series foundation models (TSFMs) to generate forecasts directly. Because TSFMs encode patterns learned from large heterogeneous datasets, they can generalise to new series and deliver usable predictions without additional training. This lowers the technical burden for applied users in policy settings and, in data-sparse environments, provides a transparent baseline against which the added cost of few-shot adaptation or fine-tuning can be judged. Our contribution is a standardised evaluation harness for macroeconomic time series. The implementation consists of expanding-window validation, consistent data transformations, and comparable accuracy metrics combining error measurements and proper scoring rules for probabilistic performance. It also benchmarks TSFMs against institutional and alternative models under uniform treatment of real-world datasets. This isolates zero-shot capacity, clarifying and enabling comparisons across studies.

While acknowledging that optimal macroeconomic forecasting ideally requires the integration of rich, real-time features capturing key economic indicators as model inputs, this study adopts a distinct approach. We focus specifically on evaluating the pure zero-shot forecasting capabilities of TSFMs, demonstrating true zero-shot transfer learning without incorporating such external features or domain-specific covariates. This deliberate choice is made to establish fundamental performance limits and explore the inherent forecasting potential of these models based solely on their pre-training, thereby establishing baselines of their transfer abilities, given the current paucity of research applying TSFMs, particularly in a zero-shot manner, to the economic domain. The contributions and novelty of this work can be summarised as follows:

- **Empirical benchmark**: We provide the zero-shot evaluation of leading TSFMs (Chronos, Moirai, TimeGPT) against classical econometric baselines from the Reserve Bank of New Zealand (RBNZ) forecasts, covering New Zealand's national GDP and sectoral industries.
- **Performance measurement**: We demonstrate that TSFMs outperform other classical methods across various horizons, including RBNZ's benchmark models, and thus we establish their utility under certain conditions.
- **Operational guidance**: We offer actionable insights for policy analysts by mapping the boundary conditions under which zero-shot TSFMs serve as low-maintenance forecasting tools for practitioners or economists. We also identify scenarios where lightweight classical models remain preferable.

2. Related Works

2.1. Forecasting Difficulty for Macroeconomic Indicators

Forecast accuracy for macroeconomic aggregates is fundamentally constrained by low signal-to-noise ratios. A long tradition of forecast evaluation studies shows that, once the horizon stretches beyond the *nowcast* and the subsequent quarter, point predictions of real GDP growth seldom beat a naive random walk, let alone a purely random-direction guess, when measured against out-of-sample performance [16]. Persistence forecasts, which simply carry forward the latest observed growth rate, provide a standard benchmark—one that even multivariate econometric systems rarely improve upon after the first step in the forecast horizon [29]. Even median private-sector projections at a four-quarter horizon exhibit RMSEs statistically indistinguishable from persistence [30].

This bound tightens whenever rare shocks such as financial crises, pandemics, natural disasters, and geopolitical conflicts create unexpected changes that historical data cannot anticipate. Empirical work documents steep declines in forecast performance during such events [31]; the COVID-19 pandemic, for example, overwhelmed both sophisticated econometric models and advanced machine learning systems because existing training sets contained no historical analogue [32]. This shortcoming has motivated a shift toward non-linear and high-dimensional techniques. Machine learning ensembles (random forests, gradient boosting), support-vector regression, and penalised regressions demonstrate gains by exploiting rich predictor sets [33]. Deep networks extend those gains: LSTMs beat tuned ARIMA baselines in volatile GDP series [34], while residual architectures such as N-BEATS win open forecasting competitions when data are plentiful or creatively augmented [35].

To complement algorithmic advances, researchers now emphasise the timing of data arrival. Mixed-frequency and real-time approaches integrate high-frequency indicators, electronic-card transactions, and daily mobility into quarterly GDP nowcasts [36], providing policymakers near-instant feedback during shocks [37,38]. Model builders also adapt specifications to the country context. In open economies like New Zealand, global commodity prices, foreign demand, and idiosyncratic domestic cycles jointly shape growth dynamics; assessing how well models internalise these influences remains an active line of inquiry [39].

Today, the cutting edge is TSFMs pre-trained on millions of heterogeneous sequences, which promise a further step change. Offering zero-shot and few-shot forecasts with native probabilistic outputs, TSFMs circumvent manual indicator selection and merge deep-learning pattern discovery with classical uncertainty quantification. However, whether this architecture can overcome the persistence benchmark in shock-prone, data-sparse settings such as New Zealand is still an open question and therefore the specific gap that the present study intends to address.

2.2. Modern and Emerging Forecasting Approaches

Economic linear models were the focus of early GDP-forecasting research. Single-equation and small-VAR frameworks inherently assume linear relationships, potentially missing signals within extensive indicator sets. Although Bayesian shrinkage (BVAR) aids in preventing overfitting in larger VARs [40,41], the linear assumption can be a significant limitation.

To harvest that broader information set, the literature turned to large-information factor techniques. Dynamic factor models (DFMs) compress hundreds of macro-financial series into a handful of latent factors that feed simple forecasting equations, delivering substantial accuracy gains [42]. Central banks now view DFMs or their extensions as baseline nowcast engines: FAVARs embed factors inside VAR structures for structural analysis [43], MIDAS regressions link monthly factors to quarterly GDP for real-time

monitoring [44], and a principal-component DFM is documented as the Reserve Bank of New Zealand's benchmark tool [45].

Building on these statistical platforms, institution-specific suites provide operational nowcasts. The Federal Reserve Bank of New York's medium-scale DSGE integrates theory-consistent shocks with factor information for policy analysis [46]. In contrast, the Atlanta Fed's GDPNow decomposes each GDP sub-aggregate via bridge equations and updates almost daily, offering a transparent, additive view of U.S. growth [47]. More recently, attention has shifted to machine learning and hybrid methods that relax linearity and exploit high-dimensional features. Ensemble trees (random forests, gradient boosting) already outperform factor and penalised-regression baselines on Dutch GDP nowcasts [45]. Deep neural networks, especially Long Short-Term Memory (LSTM) models, capture non-linear temporal dependencies and have beaten ARIMA benchmarks in volatile settings [48]. Hybrid ensembles push further by blending economic structure with ML flexibility: weighting forecasts from a time-varying-coefficient DFM and a recurrent neural network by inverse MSE reduces U.S. GDP errors beyond either model alone. At the same time, broader model-averaging strategies remain popular for error reduction [49]. There is clear progression from linear autoregressions to factor-based systems, institutional nowcasting dashboards, and data-hungry ML hybrids, each stage addressing limitations exposed by the last and setting the stage for evaluating emerging foundation model approaches.

2.3. Zero-Shot Transfer Learning for Macroeconomic Forecasting

Zero-shot transfer learning applies a pre-trained model to a novel domain or task without requiring additional training or fine-tuning on new data. In the macroeconomic context, this refers to a model that has been pre-trained on vast heterogeneous datasets of economic indicators, frequencies, and business regimes to forecast fresh and new series that were never part of the training. Historically, this has been a significant challenge for time-series analysis. However, recent research indicates that with appropriate architectures and training methodologies, zero-shot transfer learning is now achievable in this field [50]. This means a model pre-trained on a diverse collection of time series can directly forecast unseen time series without any further adjustments. This approach, also known as zero-shot forecasting, relies on the model's ability to capture universal temporal patterns that transfer effectively to new series. Suppose the characteristics of the new series are adequately represented by the diverse patterns learned during pre-training. In that case, the model can generate reasonable forecasts without specific training for each new series [51]. This enables inference on a new domain that was never seen during training without gradient updates [52].

$$\hat{\mathbf{y}}_{t+1:t+H} = F_{\theta^\star}\left(\mathbf{y}^{\dagger}_{1:t}, \mathbf{x}^{\dagger}_{1:t+H}\right) \tag{1}$$

Equation (1) describes a point of forecast for the next H time steps, using only the past t observations $\mathbf{y}^{\dagger}_{1:t}$ and any covariates $\mathbf{x}^{\dagger}_{1:t+H}$, without any parameter updates, providing the forecast of the distribution over the future H prediction [53].

The feasibility of this approach has significantly increased with the emergence of TSFMs. These are large models trained on heterogeneous time-series data from numerous domains. The benefits of zero-shot transfer learning are substantial; it eliminates the need for task-specific training and avoids the requirement for large target datasets. Furthermore, zero-shot forecasting specifically focuses on predicting future values for new time series by leveraging knowledge from a broad pre-trained model, treating each new time series as a zero-shot task. Recent studies have demonstrated surprisingly strong results for zero-shot forecasting using pre-trained models [54].

2.4. Zero-Shot TSFMs in Economic Forecasting

Early evidence that transformers can act as generic sequence learners came from the Frozen Pre-trained Transformer (FPT) experiment, where a language-model backbone was kept entirely frozen across diverse time-series tasks, showing that self-attention can operate as a domain computation [52]. Follow-up work systematises this line of research by mapping architectures, pre-training objectives, adaptation strategies, and data modalities [55]. On the other hand, architectural variants such as encoder–decoder hybrids [56], sparse-attention blocks [57], and decomposition-style residual paths [58] dominate now with networks of millions of parameters that are pre-trained on a vast heterogeneous collection of time series. These TSFMs are typically transformer-based pre-trained models on massive collections of time-series data, enabling zero-shot capabilities in forecasting under the concept of zero-shot transfer learning, which allows models to generalise to unseen tasks by leveraging knowledge gained from previously seen data [59].

TSFMs are based on the transformer architecture, shown in Figure 1. TSFMs aim for broad transferability with little or no task-specific fine-tuning. A recent survey [60] on TSFMs drew an important distinction between work that pre-trains transformers directly on raw time-series data and work that adapts existing large-language-model (LLM) backbones. Recently, the emergence of several TSFMs illustrates how these architectural ideas are realised at scale:

- Chronos repurposes the T5 language backbone for sequence-to-sequence forecasting, capturing fine-grained temporal dependencies [26].
- Moirai pushes universality further by introducing multi-patch projections that sidestep fixed-frequency constraints and perform well on both sub-hourly energy usage and daily retail sales [25].
- TimeGPT showed that a single globally trained network can forecast across hundreds of public datasets without per-task fine-tuning [24].

Despite these successes, most comparative studies still rely on classic electricity-load (ETT), traffic, and retail (M5) datasets. Even recent transformer baselines are benchmarked on ETT and weather [56]. Consequently, it is not well understood from the literature whether the celebrated efficiency of TSFMs extends to the noisier, lower-sample-size realm of macroeconomic forecasting. Likewise, zero-shot experiments with numeric-token LLMs show promise on synthetic or industrial-sensor data [61,62] but stop short of testing sturdiness under sudden shifts like the COVID-19 shock.

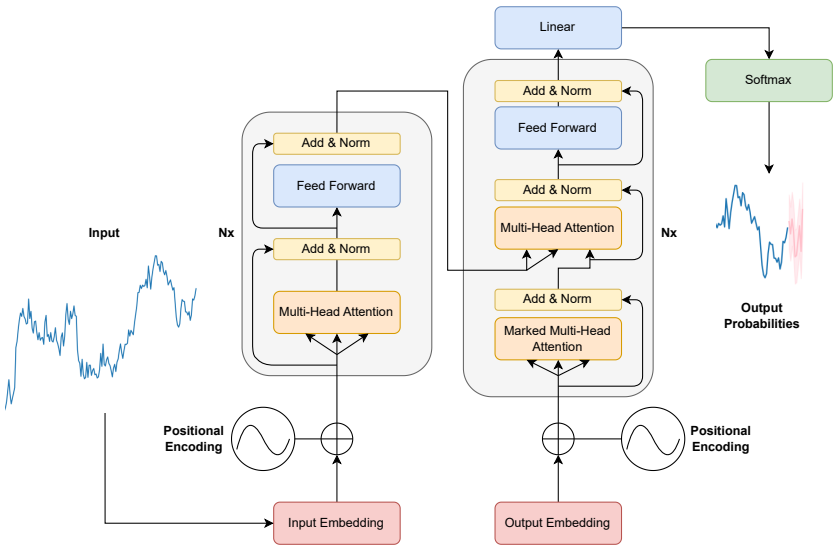

Figure 1. The transformer architecture [63].

2.5. Summary and Research Questions

This study investigates zero-shot forecasting capabilities of TSFMs to achieve zero-shot transfer learning in a macroeconomic forecast, deliberately foregoing any domain-specific features. By directly applying out-of-the-box, pre-trained TSFMs, we establish fundamental performance baselines that reveal how effectively these models can leverage their learned representations to forecast economic time series and perform under major shocks, all without any fine-tuning, and utilise national GDP and industry sectors data as a representative case study for generalisable research questions.

- **RQ1** How effective are state-of-the-art TSFMs for zero-shot univariate forecasting with zero-shot transfer learning of macroeconomic and industry-level time series?
- **RQ2** To what extent do zero-shot TSFM forecasts remain stable when confronted with periods of extreme volatility and significant economic disruption?
- **RQ3** Can zero-shot TSFMs match or surpass the published forecast accuracy of expert judgement models produced by central banks and international agencies?

3. Methodology

3.1. Dataset

The dataset sourced from Stats NZ [64] comprises a continuous quarterly time series of quarterly annual percentage change for four headline sectors: National GDP, Primary Industries, Goods-Producing Industries, and Service Industries, spanning 1999Q3 to 2024Q3. This 26-year horizon captures both routine seasonal rhythms and major unexpected changes for forecasting models. Interpreting the annual-growth metric is straightforward: National GDP aggregates all sectoral industries, Primary Industries show pronounced seasonality, Goods-Producing Industries respond to global demand and investment cycles, and Service Industries mirror domestic consumption and conditions. Tracking these growth rates across the sample reveals regular business-cycle turning points and the sharp dislocations of the 2008 Global Financial Crisis and the 2020–2021 COVID-19 periods that pose particular challenges for traditional forecasting techniques.

RBNZ Operational Dataset

To eliminate look-ahead bias, alongside the public-release series from Stats NZ, we incorporate the real-time GDP forecasts that the Reserve Bank of New Zealand (RBNZ) uses in the quarterly percentage change for overall National GDP forecasts. Specifically, we contrast their [65] gradient boosting with least-squares boosting (LSBoost) and dynamic factor model (DFM) projections.

3.2. Baseline Models

We benchmark TSFMs against four widely used baselines. Persistence and ARIMA serve as standard univariate references for short macroeconomic series and low-resource evaluation; surpassing them indicates incremental value while minimising confounding from intensive hyperparameter tuning. We also include least-squares boosting (LSBoost) and a dynamic factor model (DFM), reflecting state-of-the-art institutional practice at the Reserve Bank of New Zealand (RBNZ), as their implementations rely on proprietary features and data cannot be replicated exactly in public [66–69].

To avoid baseline sprawl and ensure comparability with a zero-shot setting, we deliberately exclude models that typically require long histories and large datasets, such as LSTM and Prophet [70]. Prophet works with high-frequency series with strong seasonal structures but is ill suited to short macroeconomic panels [71,72]. This focused baseline set supports relevant comparisons.

3.2.1. Persistence Model

The persistence forecasting model is given by Equation (2).

$$\hat{y} * t + h = y_t \quad \text{for all } h \geq 1, \tag{2}$$

Here, $\hat{y} * t + h$ is the forecast at time $t + h$, and y_t is the last observed value. Several studies have highlighted the practical utility and inherent limitations of the persistence model. As a prominent example, the impacts of selecting persistence forecasts as baseline references for evaluating forecasting systems were examined. They found that persistence benchmarks substantially influence the assessment of more advanced models, particularly in contexts with strong seasonal patterns [73].

3.2.2. ARIMA Model

The Autoregressive–Integrated–Moving-Average (ARIMA) by Nixtla [74] is an automatic process that employs a stepwise search procedure to identify optimal ARIMA and seasonal orders. This is achieved by combining unit-root tests for stationarity with the minimisation of information criteria. Models form a parsimonious yet expressive class for linear dynamics and remain a statistical baseline in forecasting. The seasonal form, denoted ARIMA$(p, d, q) \times (P, D, Q)$, combines non-seasonal and seasonal operators. The model acts as a dynamic statistical benchmark, adapting to each sector's unique autocorrelation structure and seasonality patterns. It applies differencing to handle trend and drift components and fits candidate models using a maximum search space to maintain computational efficiency.

3.2.3. LSBoost (Least-Squares Boosting)

LSBoost is a gradient boosting method in Algorithm 1, which minimises the squared-error (least-squares) loss, effectively performing a functional gradient-descent step in the space of functions in each boosting round. This machine learning ensemble method sequentially combines a finite set of weak learners to improve predictive performance for regression problems [75]. It works by fitting each new learner to the residual errors of the previous ones.

Algorithm 1: LSBoost algorithm [76].

Input: Training set $\{(x_i, y_i)\}_{i=1}^{N}$, number of iterations M
Output: Boosted predictor $F_M(x)$
Define loss function:

$$L(y, F) = \frac{(y-F)^2}{2} \quad \text{and let } F_m(x) \text{ be the current regression model.}$$

Initialisation: $F_0(x) \leftarrow \bar{y} \quad (\bar{y} = \frac{1}{N}\sum_i y_i)$
for $m \leftarrow 1$ **to** M **do**
Compute residuals:

$$\tilde{y}_i \leftarrow y_i - F_{m-1}(x_i) \quad (i = 1, \ldots, N)$$

Fit weak learner parameters (ρ_m, α_m) by

$$(\rho_m, \alpha_m) = \arg\min_{\rho,\alpha} \sum_{i=1}^{N} \left[\tilde{y}_i - \rho\, h(x_i; \alpha)\right]^2$$

Update model:

$$F_m(x) \leftarrow F_{m-1}(x) + \rho_m\, h(x; \alpha_m)$$

return $F_M(x)$

3.2.4. Factor Model

Factor models are a class of statistical models designed to explain the co-movement among a large panel of observed variables x_{it} and a smaller set of latent factors F_t, represented by Equation (3).

$$x_{it} = \lambda_i^\top F_t + e_{it}, \qquad i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{3}$$

where $F_t \in \mathbb{R}^r$ with $r \ll N$ captures the common dynamics, λ_i are series-specific, and e_{it} are errors that are allowed weak cross-sectional and serial correlation.

In macroeconomic applications, it is often desirable to model the dynamic evolution of the latent factors. DFMs augment the static framework by Equation (4).

$$F_t = \Phi_1 F_{t-1} + \cdots + \Phi_p F_{t-p} + u_t, \quad u_t \sim \mathcal{N}(0, \Sigma_u), \tag{4}$$

which embeds lead–lag relationships among economic indicators. This model strategy underlies modern nowcasting systems that blend hundreds of monthly and high-frequency indicators into real-time GDP estimates [77].

3.3. Time-Series Foundation Models

We benchmark each foundation model against baseline models, providing comprehensive comparisons. All models' hyperparameter settings were fixed for zero-shot evaluation.

3.3.1. TimeGPT-1 Model

TimeGPT-1 is a large foundation model for time-series forecasting that adapts the transformer architecture to temporal data. The model is trained at scale on a massive and diverse dataset of over 100 billion data points, which includes observations across finance, economics, healthcare, weather, IoT, and web traffic. It learns transferable temporal regularities spanning multiple seasonalities, cycles, trends, and noise regimes. Architecturally, it uses an encoder–decoder transformer with stacked self-attention, residual connections, layer normalisation, and positional encodings. A final projection layer maps decoder states to the forecast horizon, and the learned parameters are frozen to maintain its zero-shot forecasting capability [24]. This is demonstrated by direct comparison to baseline models.

3.3.2. Chronos Model

Chronos is a pre-trained time-series forecasting model for language modelling. Real-valued observations are mean-scaled and uniformly quantised into a fixed numeric vocabulary. A T5-style encoder–decoder transformer is then trained on the tokenised sequences with a next-token cross-entropy objective. At inference, the model conditions on an observed context and generates step-ahead token distributions autoregressively; sampled tokens are subsequently de-quantised to recover forecasts on the original scale. This formulation learns domain-agnostic temporal representations and enables zero-shot application across series without task-specific fine-tuning. We employ Chronos-T5 variants configured with a 4096-token numeric vocabulary, a 512-step context window, and a 64-step prediction horizon. These models are pre-trained at scale on diverse datasets and augmented to support zero-shot baseline comparisons.

The Chronos-T5-small contains approximately 46 million parameters, with 6 encoder and 6 decoder layers, a hidden size of 512, a feed-forward width of 2048, and 8 attention heads. It is suitable for low-memory CPUs and small GPUs, making it decent for establishing quick baselines and for cost-effective fine-tuning, while still maintaining credible zero-shot accuracy. The Chronos-T5-base expands to roughly 200 million parameters, with 12 encoder and 12 decoder layers, a hidden size of 768, a feed-forward width of

3072, and 12 attention heads. This configuration provides a stronger balance between model capacity and predictive accuracy, requiring only moderate additional computational resources, and is suitable for mid-range GPUs seeking enhanced zero-shot performance. The highest-capacity model, Chronos-T5-large, further scales to 710 million parameters, with 24 encoder and 24 decoder layers, a hidden size of 1024, a feed-forward width of 4096, and 16 attention heads [26].

3.3.3. Moirai Model

Moirai is a masked-encoder transformer designed for zero-shot time-series forecasting. It mitigates the quadratic cost of full self-attention through multi-patch embeddings and projection layers, and employs an any-variate attention scheme that flattens multivariate inputs into a single sequence. The model uses rotary positional embeddings and learned variate-bias terms. A mixture-density prediction head is trained via a mixture-distribution likelihood, enabling flexible output distributions. Pre-training on the Large-scale Open Time-Series Archive (LOTSA), which spans 9 application domains, aims to provide the model with patterns that transfer to new series and sampling frequencies. Reported benchmarks indicate competitive performance relative to fully pre-trained TSFMs, supporting cross-domain, cross-frequency generalisation in zero-shot settings. We employ Moirai-1.1-R at three scales and enable multi-patching with 8, 16, 32, 64, and 128 patches and automatic selection. Unless stated otherwise, we use a maximum patch length of 512, a batch size of 32, and the mixture-density prediction head. These choices facilitate direct, zero-shot comparisons against our baseline models without task-specific fine-tuning.

The Moirai-1.1-R-small contains 13.8 million parameters with 6 layers and a hidden size of 384. It is designed to capture both short- and long-term rhythms while remaining lightweight, making it fit for small GPUs and for establishing quick baselines. The Moirai-1.1-R-base expands to 91.4 million parameters with 12 layers and a hidden size of 768. This configuration offers a practical trade-off between accuracy and computational cost, making it suitable for most workloads in mixed-domain forecasting on mid-range GPUs. The Moirai-1.1-R-large further scales to 311 million parameters with 24 layers and a hidden size of 1024. With its higher capacity, it is suitable for capturing complex dependencies. All three models can perform zero-shot forecasting for baseline comparisons [25].

3.4. Model Evaluation

In the forecasting literature, scale-dependent error measures remain fundamental for assessing forecast accuracy.

Mean Absolute Error (MAE) and Mean Squared Error/Root Mean Squared Error (MSE/RMSE) in Equations (6) and (7) are the two widely reported scale-dependent accuracy metrics. MAE specifically offers a direct interpretation as the average error in the original units of the series [78]. Both compare a forecast $\hat{y}_t$ against the observed value y_t over T time steps, but they weigh errors differently. MAE is the arithmetic mean of absolute deviations, squares the residuals before averaging, and then takes the square root. But their error weighting differs. MAE averages the absolute errors $|y_t - \hat{y}_t|$, whereas RMSE averages the squared errors $(y_t - \hat{y}_t)^2$ before taking the square root.

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t| \tag{5}$$

MAE makes it easy to interpret the forecast on average.

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \tag{6}$$

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}. \tag{7}$$

MSE/RMSE, as defined in Equations (6) and (7), squares errors, giving disproportionate weight to large deviations and making the metric sensitive to outliers or significant irregularities. Both share a limitation: they are expressed in the original data units, preventing direct comparison across series with different scales. To address this, Symmetric Mean Absolute Percentage Error (SMAPE) in Equation (8) normalises the absolute error by the average magnitude of the actual and forecast values [79],

$$\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{2\,|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \tag{8}$$

Expressed as a percentage, SMAPE offers direct comparison across series of different units or magnitudes and bounds the error between 0% and 200%. Nevertheless, SMAPE can become unstable if both y_t and $\hat{y}_t$ approach zero simultaneously. Furthermore, arithmetic symmetry does not guarantee true statistical symmetry when distributions are highly skewed. MASE in Equation (9) provides an alternative approach for scale-free comparison. It also enables meaningful benchmarking against a naive baseline by dividing the MAE of the candidate model by the in-sample MAE of a simple seasonal naive forecast [80],

$$\text{MASE} = \frac{\frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|} \tag{9}$$

where m is the seasonal period. This scaling provides a key advantage: MASE is a scale-free error metric. Because the denominator is the MAE of the naive seasonal forecast, the value serves as a direct benchmark. Thus, any MASE < 1 indicates performance superior to the naive baseline, and MASE > 1 denotes inferior accuracy. This scale-free property facilitates widespread adoption in comparative studies of forecasting algorithms.

The Diebold–Mariano (DM) in Equations (10)–(12), introduced by [81], established a general, loss-function-agnostic framework for testing whether two competing forecasts have the same expected predictive accuracy.

$$e_{i,t} = y_t - \hat{y}_{i,t} \tag{10}$$

The forecast error from a model is i, and a loss function is $\ell(\cdot)$; for example, $\ell(e) = e^2$ (squared error) or $\ell(e) = |e|$ (absolute error). Define the loss differential

$$d_t = \ell(e_{1,t}) - \ell(e_{2,t}) \tag{11}$$

and let $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$. For a h-step-ahead forecast, the DM statistic is

$$\text{DM} = \frac{\bar{d}}{\sqrt{\widehat{\text{var}}(d_t)/T}}, \qquad \widehat{\text{var}}(d_t) = \gamma_0 + 2\sum_{k=1}^{h-1}\left(1 - \tfrac{k}{h}\right)\gamma_k, \tag{12}$$

where γ_k represents the sample autocovariance of d_t at lag k. Under suitable regularity conditions, DM is asymptotically distributed as $\mathcal{N}(0,1)$. Consequently, a two-sided z-test provides an asymptotic p-value. The use of the Newey–West (HAC) estimator in the denominator ensures the test's validity even when d_t follows a moving-average MA($h-1$) process, making it applicable interchangeably to MSE/RMSE or MAE comparisons.

3.5. Probabilistic Policy Risk Evaluation

We assess probabilistic forecasts using the Continuous Ranked Probability Score (CRPS) in Equation (13), a strictly proper scoring rule for continuous outcomes that rewards calibrated and sharp predictive distributions [82]. Zero-shot TSFMs produce full predictive distributions, enabling operational risk guidance without task-specific training.

Let F denote the predictive CDF and y the realised outcome. The CRPS is

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 dz \tag{13}$$

where $\mathbb{1}$ is the indicator function. Lower values indicate better probabilistic forecasts. In practice, we compute pointwise CRPS for each forecast–realisation pair and summarise its distribution (mean, median, and percentiles) to inform policy-facing diagnostics.

From the empirical distribution of pointwise CRPS, we construct tail-risk indicators [83–85]. Let Q_p denote the p-th quantile of CRPS:

$$\text{tail\_spread} = Q_{0.95} - Q_{0.50}(Median\_CRPS) \tag{14}$$

$$\text{upper\_tail\_steepness} = Q_{0.95} - Q_{0.80} \tag{15}$$

The tail-spread in Equation (14) measures the gap between typical error and near worst-case error; larger values signal elevated downside risk relative to the median. The upper-tail-steepness in Equation (15) isolates tail escalation independent of the centre; larger values indicate a rapidly thickening tail. Together with overall calibration, these statistics support clearer risk communication and more robust economic decisions [86].

3.6. Zero-Shot Forecasts

TSFMs such as TimeGPT, Chronos, and Moirai that employ zero-shot learning represent a significant advance in predictive analytics. Through large-scale pre-training on a wide variety of time-series datasets, these models develop a generalised understanding of temporal dynamics, capturing regular cycles and anomalous events. Armed with this broad temporal intuition, they can be deployed directly on novel forecasting tasks without requiring extensive domain-specific fine-tuning. At the core of zero-shot forecasting is the idea that once a foundation model has internalised patterns spanning many industries and time scales, it can transfer that knowledge seamlessly to new contexts. This approach significantly reduces the time, computational resources, and specialised expertise normally needed for forecasting solutions, as it eliminates the need for extensive re-training and adaptation for each new task.

The practical benefits of zero-shot forecasting are most pronounced in settings where conditions change rapidly or data is scarce. For instance, in sectors such as National GDP, Primary Industries, Goods-Producing Industries, and Service Industries, unexpected events and economic shocks can frequently upend historical relationships and data patterns. In such dynamic environments, a zero-shot model can provide immediate, reasonably accurate projections without waiting for new data to accumulate or models to be re-trained. Furthermore, the same pre-trained model can often be utilised for both short-term operational decisions and longer-term strategic planning, all without needing to rebuild the model for each specific forecasting horizon.

3.7. Experiment Pipeline

This empirical experiment analyses the quarterly annual percentage change series for National GDP, Primary Industries, Goods-Producing Industries, and Service Industries as datasets. StatsNZ publishes the economic aggregates used in this analysis. The analysis

employs a long window, spanning from 1999Q3 to 2024Q3, to expose the models to both secular growth phases and major shocks, including the Global Financial Crisis and the 2020–2021 COVID-19 collapse. Using a window expanding approach, accuracy is scored at every point with the evaluation metrics (MAE, RMSE, SMAPE, and MASE).

We design an experimental back-testing pipeline, shown in Figure 2. It compares shift operation persistence and statistical baseline ARIMA models against zero-shot forecasts from TSFMs: TimeGPT-1, Chronos-T5 (small, base, large), and Moirai (small, base, large). We generate one-quarter-ahead forecasts for each model over an expanding training window from 1999Q3 to 2024Q3. The forecasts are produced without any additional fine-tuning; the models were pre-trained before being evaluated on the same error metrics (MAE, RMSE, SMAPE, and MASE). To capture the heterogeneous dynamics of New Zealand's economy, we apply this experiment to four broad production sectors: National GDP, Primary Industries, Goods-Producing Industries, and Service Industries. The selection of these models for the experiments also reveals which classical and TSFMs are strongest with scarce data.

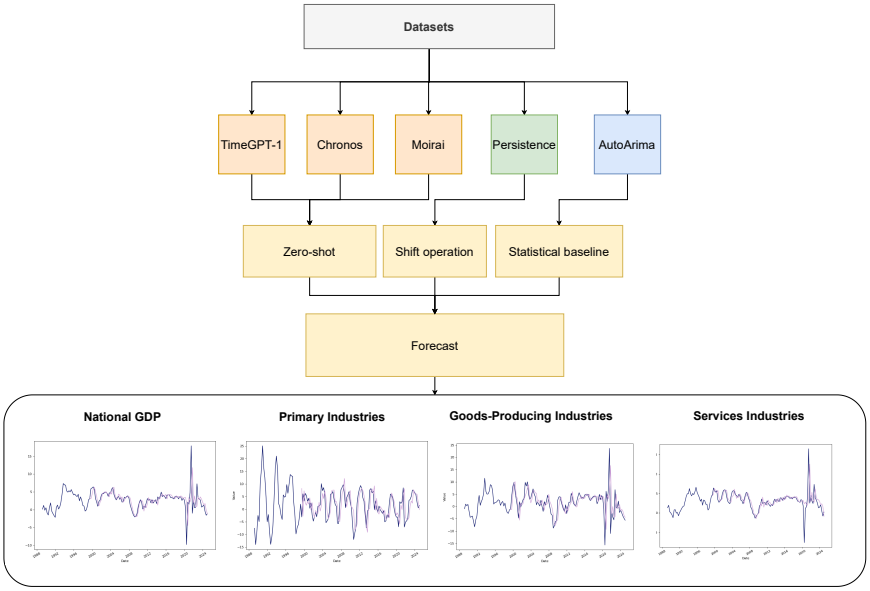

Figure 2. The experiment pipeline.

4. Results

4.1. Analysis of Model Evaluation Results

Table 1 benchmarks nine forecasting models, including seven TSFMs and two classical/statistical models, across four economic sectors: National GDP, Primary Industries, Goods-Producing Industries, and Service Industries. The analysis employs four accuracy metrics (MAE, RMSE, SMAPE, MASE) and four distinct time slices: a 26-year full sample, the calm Pre-COVID-19 period, the COVID-19 shock years, and the Post-COVID-19 rebound. Subsequently, Table 2 details the Diebold–Mariano (DM) test results, comparing TSFMs with Persistence and ARIMA models.

The tables show the performance of the Moirai models, especially Moirai-1.1-R-base (Moirai Base) and Moirai-1.1-R-large (Moirai Large), indicating strong performance. These models halve typical errors relative to traditional baselines. Conversely, orange and red areas concentrate around Persistence and ARIMA in several sectors. TimeGPT-1 and Chronos-t5 models consistently occupy an intermediate position. The accompanying table presents the mean rank of RMSE for each model across four sectors, allowing for simple identification of the best models in this experiment.

Table 1. Forecasting results across four economic sectors in multiple horizons (green = good, yellow = moderate, orange = poor, red = bad).

Model	National GDP				Primary Ind.				Goods-Prod. Ind.				Services Ind.				Mean Rank
	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	RMSE
26-year Past to Present (1999Q3–2024Q3)																	
Persistence	1.44	3.08	0.49	1.01	3.02	4.12	0.88	1.02	3.06	5.54	0.90	1.01	1.18	2.48	0.38	1.00	8.75
Arima	1.49	3.03	0.50	1.05	2.39	3.23	0.76	0.81	2.97	4.63	0.93	0.98	1.15	2.39	0.39	0.98	6.75
TimeGPT-1	1.49	2.80	0.52	1.05	3.32	4.33	0.99	1.12	2.93	4.95	0.91	0.97	1.20	2.27	0.38	1.02	7.50
Chronos-t5-small	1.22	2.31	0.49	0.86	2.70	3.67	0.90	0.91	2.71	4.24	0.94	0.89	1.01	2.00	0.36	0.86	5.25
Chronos-t5-base	1.23	2.26	0.50	0.87	2.76	3.66	0.89	0.93	2.70	4.26	0.92	0.89	0.98	1.88	0.36	0.83	4.75
Chronos-t5-large	1.38	2.98	0.49	0.97	2.78	3.62	0.88	0.94	2.79	4.57	0.94	0.92	1.02	2.03	0.37	0.86	6.00
Moirai-1.1-R-small	0.56	1.49	0.25	0.40	1.17	1.57	0.51	0.40	1.24	3.05	0.43	0.41	0.52	1.31	0.24	0.44	2.75
Moirai-1.1-R-base	0.48	1.04	0.20	0.34	1.19	1.60	0.48	0.40	0.91	1.57	0.44	0.30	0.42	1.18	0.17	0.36	1.75
Moirai-1.1-R-large	0.52	1.43	0.22	0.37	0.93	1.27	0.36	0.32	0.85	1.86	0.36	0.28	0.45	1.13	0.20	0.38	1.50
3-year Pre-COVID-19 (2017Q1–2019Q4)																	
Persistence	0.41	0.50	0.13	0.94	2.29	2.84	0.90	0.95	1.19	1.51	0.39	0.97	0.33	0.38	0.10	0.99	7.25
Arima	0.39	0.45	0.12	0.90	1.80	2.08	0.71	0.75	1.13	1.34	0.41	0.92	0.27	0.33	0.08	0.80	4.25
TimeGPT-1	0.41	0.51	0.12	0.95	2.44	2.99	1.09	1.02	1.02	1.39	0.31	0.83	0.38	0.44	0.11	1.13	7.75
Chronos-t5-small	0.40	0.47	0.13	0.92	1.84	2.37	0.88	0.76	1.22	1.52	0.44	0.99	0.31	0.34	0.09	0.91	6
Chronos-t5-base	0.41	0.55	0.13	0.94	1.84	2.43	0.86	0.76	1.22	1.53	0.48	1.00	0.31	0.35	0.09	0.92	7.5
Chronos-t5-large	0.33	0.40	0.10	0.75	1.96	2.54	0.91	0.82	1.17	1.45	0.40	0.95	0.34	0.38	0.10	1.00	6
Moirai-1.1-R-small	0.26	0.31	0.09	0.60	0.94	1.09	0.54	0.39	0.32	0.39	0.12	0.26	0.23	0.29	0.07	0.67	2.75
Moirai-1.1-R-base	0.23	0.29	0.07	0.52	0.70	0.83	0.28	0.29	0.30	0.40	0.12	0.25	0.11	0.14	0.03	0.33	1.75
Moirai-1.1-R-large	0.15	0.20	0.05	0.34	0.62	0.95	0.36	0.26	0.18	0.25	0.06	0.15	0.14	0.15	0.04	0.40	1.5
3-year During COVID-19 (2020Q1–2022Q4)																	
Persistence	6.42	8.54	1.32	0.94	3.92	6.06	0.75	0.92	11.13	14.52	1.43	0.95	5.11	6.85	0.94	0.94	9
Arima	6.68	8.35	1.27	0.98	3.86	4.83	0.90	0.90	8.73	10.98	1.34	0.75	5.00	6.56	1.07	0.92	6.5
TimeGPT-1	5.57	7.30	1.22	0.82	3.96	5.69	0.94	0.93	9.50	12.22	1.53	0.81	4.44	5.86	0.87	0.82	7.25
Chronos-t5-small	4.70	6.21	1.28	0.69	3.83	5.46	0.89	0.90	8.36	10.28	1.49	0.72	3.66	5.34	0.79	0.67	5.25

Table 1. *Cont.*

Model	National GDP				Primary Ind.				Goods-Prod. Ind.				Services Ind.				Mean Rank
	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	MAE	RMSE	SMAPE	MASE	RMSE
Chronos-t5-base	4.53	5.98	1.25	0.67	3.76	5.38	0.91	0.88	8.60	10.46	1.55	0.74	3.42	4.95	0.81	0.63	4.75
Chronos-t5-large	6.00	8.26	1.28	0.88	3.92	5.32	0.94	0.92	9.24	11.45	1.60	0.79	3.75	5.44	0.79	0.69	6.25
Moirai-1.1-R-small	2.44	4.17	0.66	0.36	1.83	2.21	0.67	0.43	5.98	8.59	1.22	0.51	1.91	3.60	0.38	0.35	3
Moirai-1.1-R-base	2.06	2.83	0.60	0.30	1.40	1.78	0.48	0.33	2.73	3.78	0.64	0.23	1.65	3.26	0.35	0.30	1.5
Moirai-1.1-R-large	2.54	4.04	0.77	0.37	1.17	1.37	0.29	0.27	3.41	5.08	0.76	0.29	1.81	3.15	0.39	0.33	1.5
2-year Post-COVID-19 (2023Q1–2024Q3)																	
Persistence	0.85	1.12	0.71	0.87	1.53	1.81	0.43	0.97	1.71	2.21	0.86	0.88	0.89	1	0.68	0.94	4.25
Arima	1.66	1.96	1.18	1.71	1.53	1.84	0.38	0.97	4.71	5.2	1.73	2.41	1.24	1.48	0.75	1.31	7.75
TimeGPT-1	2.46	3.03	1.21	2.52	2.31	3.01	0.65	1.46	3.85	4.95	1.4	1.97	2.16	2.62	0.98	2.28	8.75
Chronos-t5-small	0.94	1.21	0.82	0.96	1.17	1.41	0.26	0.74	2.13	2.54	1.28	1.09	0.87	1.06	0.66	0.92	6
Chronos-t5-base	0.92	1.27	0.78	0.95	1.59	1.85	0.33	1	2	2.53	1	1.02	0.89	1.02	0.68	0.94	6.5
Chronos-t5-large	0.87	1.2	0.76	0.89	1.9	2.27	0.64	1.2	2.03	2.46	1.12	1.04	0.89	1	0.69	0.94	5.5
Moirai-1.1-R-small	0.52	0.6	0.42	0.54	0.73	0.86	0.23	0.46	0.95	1.18	0.54	0.49	0.45	0.52	0.26	0.47	2
Moirai-1.1-R-base	0.52	0.62	0.47	0.53	0.81	1.18	0.44	0.51	0.57	0.69	0.31	0.29	0.65	0.85	0.38	0.69	2.75
Moirai-1.1-R-large	0.43	0.55	0.47	0.44	0.49	0.55	0.36	0.31	0.58	0.67	0.46	0.3	0.49	0.58	0.52	0.52	1.25

Table 2. Diebold–Mariano two-sided p-values comparing each TSFM with Persistence (p_P) and ARIMA (p_A) across four periods. Bold numbers denote $p < 0.05$.

Model	1999Q3–2024Q3		Pre-COVID-19		COVID-19		Post-COVID-19	
	p_P	p_A	p_P	p_A	p_P	p_A	p_P	p_A
National GDP								
TimeGPT-1	0.421	0.520	0.897	0.599	0.274	0.456	0.235	0.844
Chronos_Small	0.231	0.072	0.473	0.679	0.252	0.076	0.360	0.484
Chronos_Base	0.190	0.088	0.137	0.218	0.200	0.101	0.332	0.461
Chronos_Large	0.112	0.928	0.082	0.467	0.250	0.889	0.298	0.453
Moirai_Small	**0.048**	**0.017**	**0.033**	0.154	0.087	**0.044**	0.126	0.187
Moirai_Base	**0.037**	**0.011**	**0.045**	0.103	0.055	**0.017**	0.084	0.244
Moirai_Large	0.061	**0.022**	**0.016**	**0.012**	0.104	0.052	0.097	0.225
Primary Industries								
TimeGPT-1	0.275	**0.001**	0.686	0.088	0.119	0.614	0.102	0.156
Chronos_Small	**0.005**	0.054	0.327	0.320	0.277	0.559	0.296	0.866
Chronos_Base	**0.012**	0.057	0.455	0.274	0.389	0.630	0.472	0.656
Chronos_Large	**0.020**	0.067	0.594	0.209	0.298	0.687	0.979	0.414
Moirai_Small	**0.001**	**0.001**	**0.030**	0.076	0.144	**0.035**	0.238	0.300
Moirai_Base	**0.001**	**0.001**	**0.031**	**0.036**	0.142	**0.031**	0.211	0.165
Moirai_Large	**0.001**	**0.001**	**0.017**	0.069	0.135	**0.028**	0.169	**0.045**
Goods-Producing Industries								
TimeGPT-1	0.317	0.534	0.490	0.831	0.268	0.463	0.443	0.978
Chronos_Small	0.174	0.390	0.917	0.258	0.188	0.587	0.596	0.519
Chronos_Base	0.184	0.396	0.855	0.116	0.209	0.635	0.678	0.584
Chronos_Large	0.164	0.901	0.560	0.431	0.178	0.875	0.836	0.622
Moirai_Small	0.080	**0.042**	**0.029**	**0.011**	0.212	0.397	0.323	0.094
Moirai_Base	**0.025**	**0.001**	**0.024**	**0.008**	0.072	**0.032**	0.125	**0.024**
Moirai_Large	**0.021**	**0.002**	**0.024**	**0.008**	0.069	**0.049**	0.118	**0.036**
Service Industries								
TimeGPT-1	0.443	0.614	0.237	0.100	0.273	0.615	0.194	0.701
Chronos_Small	0.257	0.136	0.233	0.790	0.279	0.167	0.362	0.382
Chronos_Base	0.183	0.126	0.278	0.496	0.197	0.188	0.268	0.368
Chronos_Large	0.287	0.137	0.971	0.369	0.315	0.150	0.279	0.371
Moirai_Small	**0.039**	**0.015**	0.233	0.695	0.069	**0.047**	0.090	0.207
Moirai_Base	0.054	**0.039**	**0.005**	**0.020**	0.097	0.121	0.121	0.214
Moirai_Large	**0.040**	**0.025**	**0.013**	**0.029**	0.068	0.074	0.076	0.231

4.1.1. Forecast Analysis During Stable Phases

During the three quiet years leading up to the pandemic (2017Q1–2019Q4), economic patterns were remarkably predictable. This predictability resulted in low forecast errors across all sectors, with evaluation metrics consistently indicating strong accuracy. Traditional forecasting algorithms thrived under these orderly conditions. The Persistence model, which simply projects the most recent value, proved surprisingly effective when conditions remained stable. ARIMA models, particularly adept at detecting seasonal patterns, performed exceptionally well in sectors, aligning with farming and export cycles.

However, the Chronos and Moirai models surpassed these traditional counterparts in four sectors. Their advantage stemmed from their ability to identify and leverage addi-

tional predictable data for incremental improvements. These models effectively captured underlying patterns, potentially related to cyclical behaviour, without overfitting to the period's general stability.

By contrast, the globally trained TimeGPT-1, whose strength normally lies in drawing on vast cross-domain context, had little additional information to exploit in such a calm and peaceful environment and therefore delivered respectable but middling accuracy. In short, the era's predictability rewarded straightforward, transparent methods, leaving more sophisticated architectures only marginal room to demonstrate their advantages.

4.1.2. Forecast Analysis of During Shocks

The COVID-19 shock (2020Q1–2022Q4) ruptured the steady rhythms on which most forecasters relied. Forecast errors spiked across the board, and even algorithms with a reputation for versatility struggled. Tables 1 and 2 show that the pandemic's impact still leaves a measurable track during the COVID-19 period in these two sectors. Some models show error peaks during the COVID-19 pandemic.

These spikes, doubling in some cases, highlight how the extreme, sudden shifts in economic activity during the pandemic strained even the most sophisticated forecasting approaches. This pushed Persistence to the bottom of the rankings. ARIMA also performed less well on National GDP but had fewer errors in Primary Industries, where commodity-driven swings were less violent. Chronos and Moirai proved the most resilient, preserving green scores across sectors. They absorbed the shocks with a level of grace unmatched by simpler rivals, and their attention mechanisms and multi-scale encoders proved capable of rapid model adaptation. In contrast, TimeGPT-1 struggled to digest the unprecedented data, often sliding into the red. The experience made clear the challenges in extreme situations.

The COVID-19 shock affected forecasting accuracy over different periods, as shown in Figures 3 and A1. Showing absolute error units indicates the magnitude of errors, while the percent-change table immediately reveals which models and metrics saw the biggest proportional impacts in the Post-COVID-19 period. This experiment thus provides a clear, reproducible way to assess how sudden shifts occur.

4.1.3. Forecast Analysis Post Instability

Post-COVID-19 (2023Q1–2024Q3) analysis pinpoints how forecasting accuracy shifted by clearly dividing the horizon into two phases, Pre-COVID-19 and Post-COVID-19. Comparing the error metrics across these windows reveals the damage and relative resilience of the models. Differences between models are significantly impacted by the dramatically widened error margins during the crisis. The pandemic's aftershocks echo in the data-generating process, as Post-COVID-19 error metrics are distinctly higher than Pre-COVID-19. However, as the immediate Post-COVID-19 period recedes, forecasting accuracy has rebounded. Most metrics have fallen by two-thirds from their COVID-19 peaks, and "greens" (indicating good performance) re-emerge in every column. Persistence models, often brittle in the face of extreme shocks, now produce errors that are no longer extraordinary, becoming respectable again.

Despite this general recovery, Moirai models consistently deliver leading accuracy across the four sectors. They reclaim leadership and regain dominance by capturing residual asymmetries that persist after the shock, thereby sharpening distinctions among the other models. ARIMA is rarely at the top of the leaderboard but stays within a comfortable margin of the front-runners. TimeGPT-1 continues to trail the specialist models, lacking fine-tuned sensitivity in these domain-specific economic series. Thus, the zero-shot forecasts of Moirai and Chronos models hold the line during these pandemic-induced

spikes. These models demonstrate an ability to handle unprecedented incidents, ruptures, and behavioural swings.

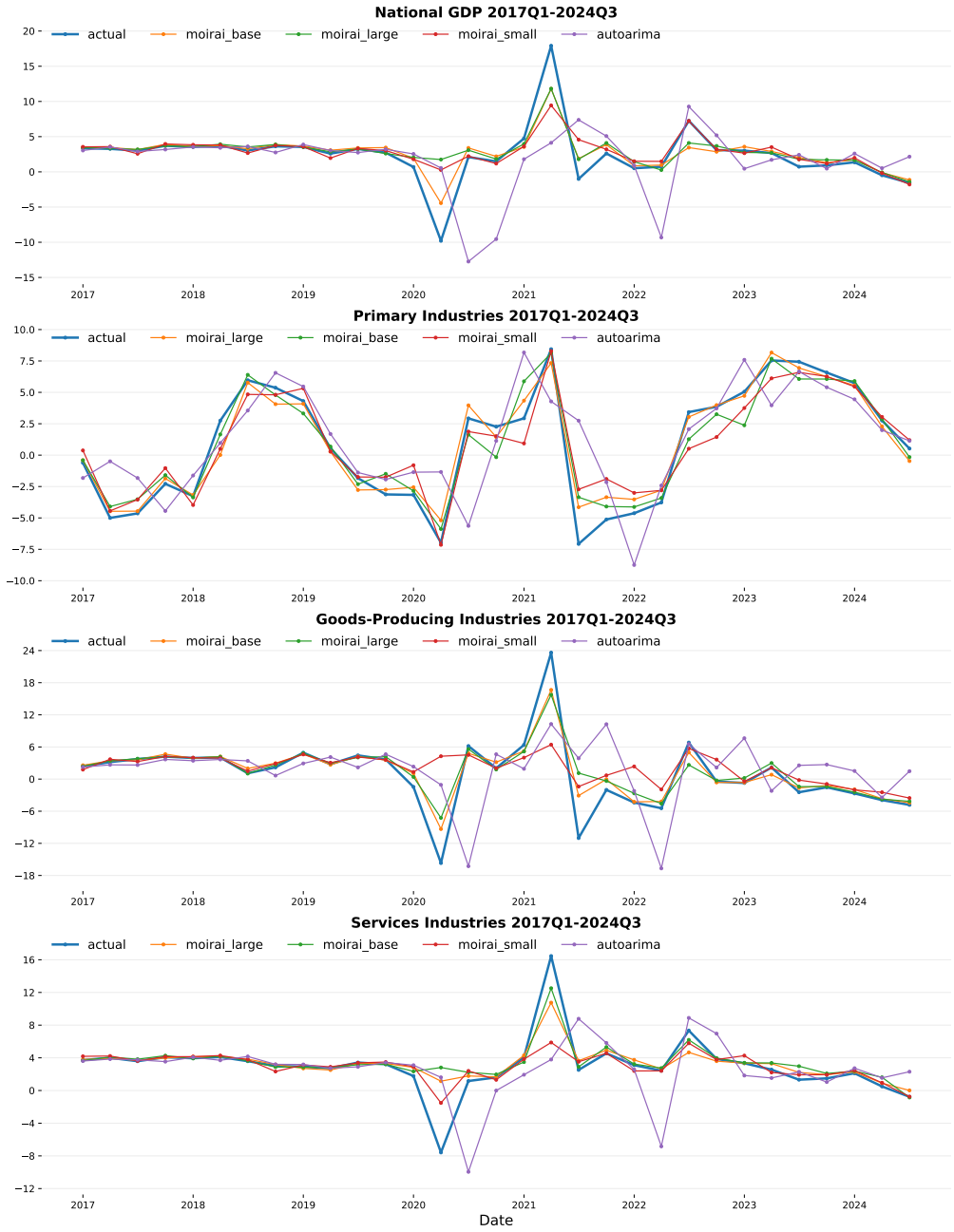

Figure 3. Actual and forecasting values for the quarterly annual percentage change from 2017Q1 to 2024Q4.

4.1.4. Summary

Overall, classical and statistical baseline models, such as persistence and ARIMA, exhibit increased errors across most metrics. In contrast, TSFMs demonstrate considerably greater resilience. Their MAE and MASE remain close to zero, and RMSE frequently shows improvement, suggesting that extensive pre-training enables them to adapt to new levels. TimeGPT-1 is notable for significant error increases, particularly concerning National GDP. The Moirai models represent a middle ground; they are better at capturing scale shifts with lower RMSE but still experience increases in MAE and SMAPE. This analysis was conducted over 26 years across four sectors, evaluating performance using multiple error metrics (MAE, RMSE, SMAPE, MASE).

Moirai-1.1-R-base (Moirai Base) and Moirai-1.1-R-large (Moirai Large) demonstrate strong consistency with low mean ranks of RMSE across four sectors over time. The COVID-19 period, reflecting a difficult time for models adapting to sudden structural shifts, presented a significant challenge. As markets begin to normalise Post-COVID-19, the differences in handling volatility become clearer.

The regression analysis in Figure 4 highlights this: Moirai Base appears to cope with macroeconomic turbulence, such as the COVID-19 shock, more robustly than Moirai Large. To quantify this, we drew 200 random sub-periods from the 1999Q3–2024Q4 sample (each window containing at least 10 quarters), computed the variance of realised GDP growth for every window, and paired it with the model's out-of-sample RMSE. Plotting variance (x-axis) against RMSE (y-axis) therefore yields 200 (variance, RMSE) points for each model. A linear regression fitted to these clouds reveals a markedly flatter slope for Moirai Base than for Moirai Large, indicating that the large model's error climbs much faster as volatility increases. In practical terms, Moirai Base maintains accuracy when conditions become volatile, whereas Moirai Large deteriorates more sharply, making the base model the more advantageous choice during periods of economic instability.

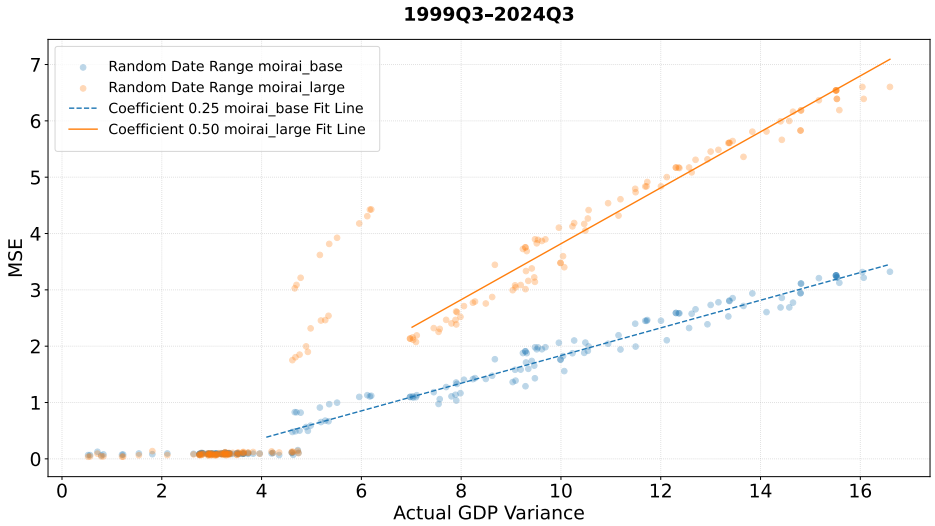

Figure 4. A regression plot of actual GDP variance against MSE over 26 years.

4.2. Forecast Benchmarking Against State of the Art

We compare our best model, Moirai Base, with the RBNZ's LSboost and Factor models, using the quarterly percentage change for overall National GDP forecasts.

We extracted forecasts from the RBNZ's LSboost (Gradient boosting) and Factor model, observing that they are quarterly percentage changes. Consequently, we converted the national GDP data from a quarterly annual percentage change to a quarterly percentage change to match the forecast format. For comparison in Figure 5, we also used Moirai Base to forecast quarterly percentage change values. The table includes Diebold–Mariano two-sided tests computed against an ARIMA model for the period 2009Q1–2019Q1, evaluating the RBNZ's published LSboost and Factor GDP nowcasts [65]. Table 3 presents forecast accuracy comparisons based on RMSE and Diebold–Mariano (DM) tests.

Moirai Base yields the lowest RMSE, indicating the best point-forecast performance overall. The DM statistics show that Moirai's errors are significantly smaller than those of a benchmark ARIMA model. Still, the differences from the RBNZ's LSBoost and Factor models are not statistically significant at conventional levels. Thus, while Moirai demonstrably outperforms ARIMA, it performs on par with LSBoost and Factor, an important finding that highlights their comparable predictive accuracy.

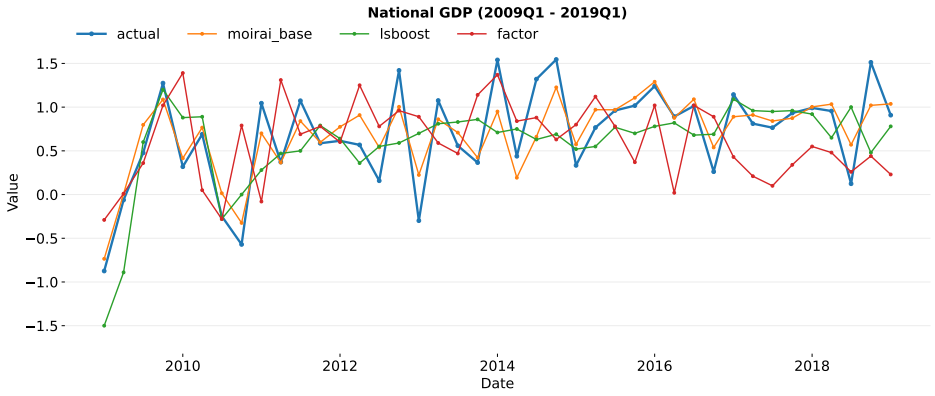

Figure 5. Comparisons of the Moirai Base Model with LSboost and Factor Model using Reserve Bank Nowcasts.

Table 3. Diebold–Mariano test *p*-values comparing with models of RBNZ.

	Comparison	
Model Comparison	**RMSE**	***p*-Value**
Moirai Base	0.2692	0.0014
LSBoost [65]	0.4876	0.2389
Factor [65]	0.6328	0.4936

4.3. Probabilistic Evaluation

We evaluate full predictive distributions for each forecast origin and horizon of national GDP to support policy-relevant interpretation. Table 4 reports mean, median, and percentile summaries of CRPS, alongside tail-risk diagnostics. Because CRPS is a strictly proper scoring rule, lower values indicate better overall probabilistic performance combining calibration with sharpness. Tail metrics make adverse-scenario behaviour explicit: tail-spread (p95-median-CRPS) and upper-tail steepness (p95–p80) indicate how quickly uncertainty escalates.

Table 4. CRPS and tail-risk results of national GDP 1999Q3–2024Q3 (green = good, yellow = moderate, red = bad).

	CRPS						Tail		
Model	**Mean**	**Median**	**p95**	**p90**	**p85**	**p80**	**Spread**	**(Upper) Steepness**	**Mean Rank**
Moirai_Base	0.48	0.24	1.32	0.95	0.59	0.47	1.08	0.85	1
Moirai_Large	0.52	0.21	1.36	0.93	0.59	0.47	1.14	0.88	2
Moirai_Small	0.56	0.27	1.13	0.76	0.61	0.52	0.86	0.61	3
Chronos_Small	1.22	0.71	3.18	2.17	1.63	1.51	2.46	1.68	4
Chronos_Base	1.23	0.76	3.17	2.05	1.69	1.46	2.41	1.71	5
Chronos_Large	1.37	0.75	3.28	2.01	1.80	1.39	2.53	1.88	6
Persistence	1.43	0.74	4.17	2.10	1.88	1.33	3.43	2.84	7
Arima	1.48	0.73	8.39	2.47	1.85	1.46	7.65	6.93	8
TimeGPT-1	1.49	0.88	5.28	2.52	2.12	1.68	4.40	3.60	9

In our experiments, the Moirai family performs best overall. Moirai-Large attains the lowest mean and median CRPS, followed by Moirai-Base and Moirai-Small. The Chronos variants form a middle tier, with Chronos-Small outperforming Chronos-Base and Chronos-Large. Classical baselines (AutoARIMA, Persistence) perform worst, indicating weaker

reliability of their predictive distributions. From a policy standpoint, Moirai-Base offers the strongest balance of accuracy and cost, while Moirai-Large exhibits the most stable tails for risk-sensitive applications. Chronos models may be acceptable where moderately heavier tails are tolerable. Reporting both CRPS and tail-risk diagnostics (tail-spread, upper-tail steepness) should be standard practice; elevated tail-spread warrants caution, whereas stable tails support greater trust in predictive intervals.

All distributional metrics are computed recursively in an expanding-window, real-time setting to reflect operational conditions and reduce evaluation bias. Where relevant, differences in scores should be accompanied by uncertainty assessments to substantiate model rankings.

5. Discussion

5.1. TSFM Zero-Shot Effectiveness in Macroeconomic Forecasting

Based on **RQ1**, empirical studies reveal a clear hierarchy among state-of-the-art TSFMs in a zero-shot setting. These TSFMs represent promising steps toward universal forecasting. Empirically, Moirai currently demonstrates the most consistent accuracy on macro and industry series, exhibiting the lowest scaled errors and highest ranks. This is likely attributable to task-specific design and massive pre-training [25]. Chronos also performs strongly, especially on monthly and other high-frequency series matching state-of-the-art zero-shot accuracy with only minor configuration tweaks [26]. By contrast, although TimeGPT offers broad cross-domain versatility, it trails leading alternatives on several specialised economic benchmarks [24]. Both models, however, can significantly narrow this gap when they are fine-tuned or adapted in context with domain-specific economic and industry datasets [87]. Overall, these models generalise better across domains than traditional local models. However, it is important to be aware of their limitations and select model size and type accordingly. Because zero-shot forecasts rely solely on patterns learned during pre-training, they cannot adapt when the data-generating process shifts to a new regime not represented in the training corpus.

Regarding the research question, we can address that TSFMs are indeed a powerful new forecasting tool whose zero-shot forecasts can rival and even surpass classical and deep-learning baselines. Nevertheless, a limitation arises when evaluating a zero-shot forecast on the latest data vintage while a baseline was trained on older or different data. In such cases, the comparison involves forecasting different targets, making meticulous vintage control and strict alignment of transformations crucial to avoid a mismatch.

5.2. Effectiveness of TSFMs in Zero-Shot Forecasting During Shocks

For **RQ2**, we observe the following: TSFMs are not immune to the effects of abrupt market changes; they allocate additional representational capacity to capture sudden spikes and crash patterns once these appear in the training data stream. In experiments, this manifests as the zero-shot forecast overshooting immediately after a shock subsides, followed by a glide path before error metrics converge to their pre-crisis baselines. Pre-trained models exhibit an even stronger lag because their universal embeddings adapt more slowly to new domain-specific data, such as the shifts seen during the pandemic. Consequently, Post-COVID-19 forecasts inherit a systematic bias, and their prediction intervals remain excessively wide. This phenomenon has also been documented in earlier crisis episodes, where models struggled to distinguish between statistical noise and lasting shifts with higher penetration. Nevertheless, zero-shot forecasts are not flawless. The initial quarters following an unprecedented shock often exhibit residual biases in predictions, especially when the disruption's characteristics diverge significantly from the model's learned priors or involve novel policy responses outside the training distribution. In such

cases, light post-event calibration can be beneficial. It is important to recognise that zero-shot does not equal shock-proof. Because no local fine-tuning occurs, TSFMs can carry a small residual bias in the few post-shock quarters, potentially overreacting if the current crisis deviates significantly from patterns seen in pre-training.

Addressing the research question, the findings suggest that one cannot rely on the zero-shot forecast alone. While TSFMs offer a theoretical advantage in leveraging cross-domain patterns and adaptive attention, modest fine-tuning or bias correction using new data can markedly improve forecast accuracy over the long run without eroding the model's broad generalisation capabilities. This highlights the importance of modest adaptations to effectively handle novel extreme events.

5.3. Zero-Shot TSFMs vs. Domain-Specific Models for Macroeconomic Forecasting

RQ3 sought to investigate whether fully zero-shot TSFMs can compete with, or even replace, the bespoke multivariate systems employed by central banks. Diebold–Mariano tests on real-time New Zealand GDP forecasts demonstrate that Moirai-1.1-R-base is statistically superior to an ARIMA benchmark. However, it is only comparable to the RBNZ LSBoost ensemble and dynamic factor model, not significantly better. Consequently, while the RBNZ's suite retains a narrow statistical edge, the performance gap has effectively closed.

The decisive factor in this comparison is information breadth. LSBoost and the factor model incorporate hundreds of carefully curated covariates, including business-confidence surveys, commodity prices, export receipts, and high-frequency trackers. These capture cross-sectional signals that a univariate zero-shot model inherently lacks. Despite this handicap, Moirai's relative success stems from its architecture and pre-training strategy: a hierarchical transformer trained on millions of heterogeneous series. This training enables it to learn universal priors for seasonality, sudden shifts, and global disturbances. These priors allowed Moirai to maintain low errors through the COVID-19 shock periods, during which many traditional models required re-estimation.

Addressing the research question, the demonstration that TSFMs can match the RBNZ's sophisticated multivariate machinery emphasises the value for agencies with limited analytical resources. Furthermore, because Moirai's performance remains stable during sudden shocks, embedding its forecasts in early-warning dashboards would bolster risk monitoring and help policymakers respond more swiftly to economic turning points. In economies where external shocks propagate quickly, such robustness is particularly valuable for setting prudent fiscal buffers, calibrating macro-prudential tools, and stress-testing contingency plans.

5.4. Policy Interpretation and Trust

Policymakers should interpret probabilistic forecasts as distribution-based paths conditional on the information available at release. Narrow predictive bands indicate high confidence suitable for routine planning, whereas wider bands shift attention to low-probability, high-impact risks such as budget stress, output gaps, and debt dynamics. Probability-of-exceedance statements for policy-relevant thresholds, such as the chance that GDP grows next quarter or that inflation exceeds the target band, are more actionable than symmetric intervals.

Table 4 in Section 4.3 shows these ideas via a colour code. Green (trusted): mean-CRPS and median-CRPS are in the top quartile among peers (lower is better), and upper-tail-steepness is low. Yellow (caution): tail-spread materially exceeds the central interval; add scenarios and qualify confidence in communications. Red (escalate): both average loss and tail metrics are high; widen contingency ranges and initiate senior review.

To aid interpretation, translate metrics into risk statements tied to actions. For example, a Moirai-large run with low median-CRPS and mild upper-tail-steepness supports communicating central intervals and leaving contingency allocations unchanged. By contrast, an AutoARIMA run with pronounced upper-tail-steepness warrants heightened stress monitoring and additional scenarios.

Trust is built through auditable checks that pair CRPS summaries with distributional snapshots around economic turns. When the preferred model leads without tail spikes during shocks, uncertainty bands and tail-risk statistics become operational when mapped to explicit thresholds, supported by transparent evidence that decision-makers can readily interpret.

5.5. Limitations and Future Works

Although our back-tests suggest that TSFMs set a competitive baseline for zero-shot GDP forecasting, those performances are contingent on the stability of the underlying time series and economic environment. This study also acknowledges the possibility that time series from our experimental domain may have been included in the training datasets for the various TSFMs, which may have influenced the accuracy results. A further limitation is that our study did not consider multivariate capabilities of TSFMs, which we leave to future work.

Furthermore, we plan on expanding our future work by moving beyond zero-shot evaluation to explore a fine-tuning approach for re-training models with plausible exogenous shocks and exploring modern baselines and benchmarks. This expansion will also involve extending evaluations to other economies and time periods for accuracy improvement, and applying a quarterly intercept correction to trim forecast errors without compromising cross-domain generality. Lastly, we will report full predictive distributions via density forecasts and coverage metrics for assessing whether these models deliver calibrated probabilities as well as sharp point forecasts.

6. Conclusions

This research delivers one of the first systematic zero-shot evaluations of TSFMs for macroeconomic indicator forecasting. Without any fine-tuning, several TSFM variants, including TimeGPT, Chronos, and Moirai, were tasked with predicting the quarterly year-on-year growth rates of four headline series: National GDP, Primary Industries, Goods-Producing Industries, and Service Industries. Each model operated in a purely univariate setting, eliminating bespoke econometric specifications and long local training histories. Point forecasts and predictive interval calibration were benchmarked against classical/statistical baselines and operationalised models from large institutions.

The Moirai variants outperformed Persistence and ARIMA across several horizons and matched industry benchmarks. Crucially, TSFMs remained broadly resilient during the COVID-19 structural break, yet the findings revealed small and systematic post-shock biases that suggest a need for multivariate extensions that incorporate other indicators. These results show that carefully engineered TSFMs can internalise rich economic dynamics and deliver actionable forecasts for macroeconomic monitoring and strategic planning. At the same time, clear directions exist for future work targeting multivariate integrations, fine-tuning, ensemble blending, and rigorous stress-testing across additional crisis scenarios.

Author Contributions: Conceptualization, T.S. and J.J.; Methodology, J.J.; Software, J.J.; Validation, J.J. and S.R.; Resources, J.J.; Writing—original draft, J.J.; Writing—review & editing, T.S., S.R. and J.J.; Supervision, T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The resources for this research are openly available in the GitHub repository at https://github.com/JittarinJet/Generalisation-Bounds-of-Zero-Shot-Economic-Forecasting-using-Time-Series-Foundation-Models (accessed on 19 September 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Figure A1 shows the quarterly annual percentage change in the forecasting accuracy of other models for National GDP, Primary Industries, Goods-Producing Industries, and Service Industries over the period 2017Q1–2024Q3.

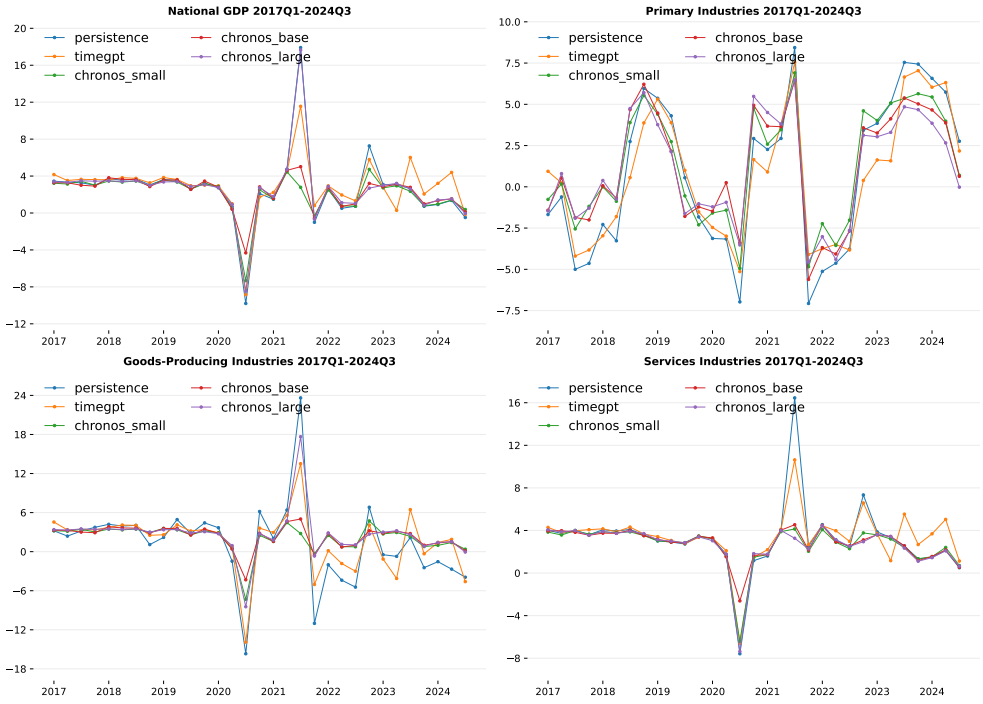

Figure A1. Forecasting accuracy results: quarterly annual percentage change by remaining models (2017Q1–2024Q3).

References

1. International Monetary Fund. *World Economic Outlook: Policy Pivot, Rising Threats*; IMF: Washington, DC, USA, 2024.
2. Borio, C.; Drehmann, M.; Tsatsaronis, K. Stress–Testing Macro Stress Testing: Does It Live Up to Expectations? *J. Financ. Stab.* **2014**, *12*, 3–15. [CrossRef]
3. Bloom, N. Fluctuations in Uncertainty. *J. Econ. Perspect.* **2014**, *28*, 153–176. [CrossRef]
4. OECD. *OECD Economic Outlook, Volume 2023 Issue 2*; OECD Publishing: Paris, France, 2023. [CrossRef]
5. S&P Global Ratings. Sovereign Rating Methodology. Credit FAQ. 2017. Available online: https://enterprise.press/wp-content/uploads/2017/05/Sovereign-Rating-Methodology.pdf (accessed on 22 May 2025).
6. Clark, T.E.; West, K.D. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *J. Econom.* **2007**, *138*, 291–311. [CrossRef]
7. Croushore, D.; Stark, T. A Real-Time Data Set for Macroeconomists. *J. Econom.* **2001**, *105*, 111–130. [CrossRef]
8. Perron, P. The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis. *Econometrica* **1989**, *57*, 1361–1401. [CrossRef]
9. Hamilton, J.D. Understanding Crude Oil Prices. *Energy J.* **2009**, *30*, 179–206. [CrossRef]
10. Cavallo, E.; Noy, I. Natural Disasters and the Economy—A Survey. *Int. Rev. Environ. Resour. Econ.* **2011**, *5*, 63–102. [CrossRef]
11. Jorda, O.; Singh, S.R.; Taylor, A.M. Longer-Run Economic Consequences of Pandemics. *Rev. Econ. Stat.* **2022**, *104*, 166–175. [CrossRef]
12. Marcellino, M. Sectoral Aggregation in Multivariate Time-Series Models. *Int. J. Forecast.* **2005**, *21*, 277–291.
13. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*, revised ed.; Holden-Day: San Francisco, CA, USA, 1976.
14. Sims, C.A. Macroeconomics and Reality. *Econometrica* **1980**, *48*, 1–48. [CrossRef]

15. Fildes, R.; Stekler, H. The State of Macroeconomic Forecasting. *J. Macroecon.* **2002**, *24*, 435–468. [CrossRef]
16. D'Agostino, A.; Giannone, D.; Surico, P. (Un)Predictability and Macroeconomic Stability. *SSRN Electron. J.* **2006** [CrossRef]
17. Jacobs, J.P.A.M.; van Norden, S. Why Do Revisions to GDP and Inflation Agree? *J. Monet. Econ.* **2011**, *58*, 450–465. [CrossRef]
18. Carriero, A.; Clark, T.E.; Marcellino, M. *Nowcasting Tail Risks to Economic Activity with Many Indicators*; Working Paper 20-13R2, Revised 22 September 2020; Federal Reserve Bank of Cleveland: Cleveland, OH, USA, 2020. [CrossRef]
19. Maccarrone, G.; Morelli, G.; Spadaccini, S. GDP Forecasting: Machine Learning, Linear or Autoregression? *Front. Artif. Intell.* **2021**, *4*, 757864. [CrossRef]
20. Long, X.; Bui, Q.; Oktavian, G.; Schmidt, D.F.; Bergmeir, C.; Godahewa, R.; Lee, S.P.; Zhao, K.; Condylis, P. Scalable Probabilistic Forecasting in Retail with Gradient Boosted Trees: A Practitioner's Approach. *arxiv* **2023**, arXiv:2311.00993. [CrossRef]
21. Goel, A.; Pasricha, P.; Kanniainen, J. Time-Series Foundation AI Model for Value-at-Risk Forecasting. *arXiv* **2025**, arXiv:2410.11773.
22. Germán-Morales, M.; Rivera-Rivas, A.; del Jesus Díaz, M.; Carmona, C. Transfer Learning with Foundational Models for Time Series Forecasting using Low-Rank Adaptations. *Inf. Fusion* **2025**, *123*, 103247. [CrossRef]
23. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2022**, arXiv:2108.07258.
24. Garza, A.; Challu, C.; Mergenthaler-Canseco, M. TimeGPT-1. *arXiv* **2024**, arXiv:2310.03589.
25. Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; Sahoo, D. Unified Training of Universal Time Series Forecasting Transformers. *arXiv* **2024**, arXiv:2402.02592. [CrossRef]
26. Ansari, A.F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Arango, S.P.; Kapoor, S.; et al. Chronos: Learning the Language of Time Series. *arXiv* **2024**, arXiv:2403.07815. [CrossRef]
27. McKenzie, S. *How Vulnerable Is New Zealand to Economic Shocks in Its Major Trading Partners?* Analytical Note 24/04; New Zealand Treasury: Wellington, New Zealand, 2024.
28. Gordon, M. First Impressions: Forthcoming Revisions to Lift NZ GDP Growth. *Economics*, 27 November 2024. Available online: https://www.westpaciq.com.au/economics/2024/11/nz-first-impressions-gdp-revisions-november-2024 (accessed on 1 August 2025).
29. Hartigan, L.; Rosewall, T. *Nowcasting Quarterly GDP Growth During the COVID-19 Crisis Using a Monthly Activity Indicator*; Research Discussion Paper 2024/04; Reserve Bank of Australia: Sydney, Australia, 2024.
30. Edge, R.M.; Rudd, J.B. Real-Time Properties of the Federal Reserve's Output Gap. *Rev. Econ. Stat.* **2016**, *98*, 785–791. [CrossRef]
31. Castle, J.L.; Fawcett, N.W.P.; Hendry, D.F. Forecasting Breaks and Forecasting during Breaks. In *The Oxford Handbook of Economic Forecasting*; Clements, M.P., Hendry, D.F., Eds.; Oxford University Press: Oxford, UK, 2011; pp. 315–349. [CrossRef]
32. Lewis, D.J.; Mertens, K.; Stock, J.H.; Trivedi, M. Measuring Real Activity Using a Weekly Economic Index. *J. Appl. Econom.* **2022**, *37*, 667–687. [CrossRef]
33. Rossi, T.; Guhathakurta, S. Machine Learning Methods for Capturing Nonlinear Relationships in Travel Behavior Research: A Review. *Travel Behav. Soc.* **2023**, *32*, 100–116.
34. Oancea, B.; Simionescu, M. Improving Quarterly GDP Forecasts Using Long Short-Term Memory Networks: An Application for Romania. *Electronics* **2024**, *13*, 4918. [CrossRef]
35. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. *arXiv* **2020**, arXiv:1905.10437. [CrossRef]
36. Susnjak, T.; Schumacher, C. Nowcasting: Towards Real-Time GDP Prediction. Technical Report, GDP Live Technical Report. 2018. Available online: https://gdp-live.s3-ap-southeast-2.amazonaws.com/GDP_Live_Working_Paper.pdf (accessed on 18 October 2025).
37. Giannone, D.; Reichlin, L.; Small, D. Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *J. Bus. Econ. Stat.* **2008**, *26*, 464–480. [CrossRef]
38. Herculano, M.C. *A Monthly Financial Conditions Index for New Zealand*; Discussion Paper DP2022-01; Reserve Bank of New Zealand: Wellington, New Zealand, 2022.
39. Galt, D. *New Zealand's Economic Growth*; Treasury Working Paper 00/09; New Zealand Treasury: Wellington, New Zealand, 2000.
40. Wu, Y.; Zhou, X. VAR Models: Estimation, Inferences, and Applications. In *Handbook of Financial Econometrics and Statistics*; Lee, C., Lee, J.C., Eds.; Springer: New York, NY, USA, 2015; pp. 2077–2091. [CrossRef]
41. Litterman, R.B. Forecasting with Bayesian Vector Autoregressions—Five Years of Experience. *J. Bus. Econ. Stat.* **1986**, *4*, 25–38. [CrossRef]
42. Stock, J.H.; Watson, M.W. Macroeconomic Forecasting Using Diffusion Indexes. *J. Bus. Econ. Stat.* **2002**, *20*, 147–162. [CrossRef]
43. Bernanke, B.S.; Boivin, J.; Eliasz, P. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Q. J. Econ.* **2005**, *120*, 387–422.
44. Ghysels, E.; Sinko, A.; Valkanov, R. MIDAS Regressions: Further Results and New Directions. *Econom. Rev.* **2007**, *26*, 53–90. [CrossRef]

45. Kant, D.; Pick, A.; de Winter, J. *Nowcasting GDP Using Machine Learning Methods*; DNB Working Paper 790; De Nederlandsche Bank: Amsterdam, The Netherlands, 2024.
46. Del Negro, M.; Giannoni, M.P.; Schorfheide, F. *The FRBNY DSGE Model: Description and Forecasting Performance*; Staff Report 674; Federal Reserve Bank of New York: New York, NY, USA, 2014.
47. Higgins, P. *GDPNow: A Model for GDP "Nowcasting"*; Working Paper 2014-07; Federal Reserve Bank of Atlanta: Atlanta, GA, USA, 2014.
48. Oancea, B.; Simionescu, M. GDP Forecasting with Long Short-Term Memory Networks: Evidence from Romania. *Econ. Comput. Econ. Cybern. Stud. Res.* **2024**, *58*, 101–118.
49. Longo, L.; Riccaboni, M.; Rungi, A. A Neural Network Ensemble Approach for GDP Forecasting. *Econ. Model.* **2021**, *104*, 105657. [CrossRef]
50. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. Meta-Learning Framework with Applications to Zero-Shot Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 9242–9250. [CrossRef]
51. Dooley, S.; Khurana, G.S.; Mohapatra, C.; Naidu, S.; White, C. ForecastPFN: Synthetically-Trained Zero-Shot Forecasting. *arXiv* **2023**, arXiv:2311.01933.
52. Zhou, T.; Niu, P.; Wang, X.; Sun, L.; Jin, R. One Fits All: Power General Time Series Analysis by Pretrained LM. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.
53. Auer, A.; Parthipan, R.; Mercado, P.; Ansari, A.F.; Stella, L.; Wang, B.; Bohlke-Schneider, M.; Rangapuram, S.S. Zero-Shot Time Series Forecasting with Covariates via In-Context Learning. *arXiv* **2025**, arXiv:2506.03128.
54. Xiao, C.; Zhou, J.; Xiao, Y.; Lu, X.; Zhang, L.; Xiong, H. TimeFound: A Foundation Model for Time Series Forecasting. *arXiv* **2025**, arXiv:2503.04118. [CrossRef]
55. Liang, Y.; Wen, H.; Nie, Y.; Jiang, Y.; Jin, M.; Song, D.; Pan, S.; Wen, Q. Foundation Models for Time Series Analysis: A Tutorial and Survey. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), Barcelona, Spain, 25–29 August 2024; ACM: New York, NY, USA, 2024; pp. 6555–6565. [CrossRef]
56. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *arXiv* **2021**, arXiv:2012.07436. [CrossRef]
57. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
58. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *arXiv* **2022**, arXiv:2106.13008.
59. Yoon, M.; Palowitch, J.; Zelle, D.; Hu, Z.; Salakhutdinov, R.; Perozzi, B. Zero-shot Transfer Learning within a Heterogeneous Graph via Knowledge Transfer Networks. *arXiv* **2022**, arXiv:2203.02018. [CrossRef]
60. Ye, J.; Zhang, W.; Yi, K.; Yu, Y.; Li, Z.; Li, J.; Tsung, F. A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Model. *arXiv* **2024**, arXiv:2405.02358. [CrossRef]
61. Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J.; Shi, X.; Chen, P.; Liang, Y.; Li, Y.; Pan, S.; et al. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv* **2024**, arXiv:2310.01728. [CrossRef]
62. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
64. Statistics New Zealand. Gross Domestic Product: December 2024 Quarter—Visualisation Data. 2025. Available online: https://www.stats.govt.nz/assets/Uploads/Gross-domestic-product/Gross-domestic-product-December-2024-quarter/Download-data/gross-domestic-product-december-2024-quarter-visualisation.csv (accessed on 18 October 2025).
65. Richardson, A.; van Florenstein Mulder, T.; Vehbi, T. Nowcasting GDP Using Machine-Learning Algorithms: A Real-Time Assessment. *Int. J. Forecast.* **2021**, *37*, 941–948. [CrossRef]
66. Bayarmagnai, G. *Nowcasting New Zealand GDP Using a Dynamic Factor Model*; Analytical Note AN2025/01; Reserve Bank of New Zealand: Wellington, New Zealand, 2025.
67. Arro-Cannarsa, M.; Scheufele, R. *Nowcasting GDP: What Are the Gains from Machine Learning Algorithms?* SNB Working Papers 2024-06; Swiss National Bank: Zürich, Switzerland, 2024.
68. Tenorio, J.; Perez, W. Monthly GDP nowcasting with Machine Learning and Unstructured Data. *arXiv* **2024**, arXiv:2402.04165. [CrossRef]
69. Supriyatna, P.; Prastyo, D.; Akbar, M. Application of the dynamic factor model on nowcasting sectoral economic growth with high-frequency data. *Media Stat.* **2024**, *17*, 128–139. [CrossRef]
70. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2021**, *379*, 20200209. [CrossRef] [PubMed]
71. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]

72. Cascaldi-Garcia, D.; Luciani, M.; Modugno, M. *Lessons from Nowcasting GDP Across the World*; Technical Report 1385; Board of Governors of the Federal Reserve System: Washington, DC, USA, 2023. [CrossRef]
73. Barnes, J.; Barnes, M. The Role of Persistence Models in Forecast Evaluation. *J. Forecast.* **2020**, *35*, 123–135.
74. Garza, A.; Mergenthaler Canseco, M.; Challú, C.; Olivares, K.G. *StatsForecast: Lightning-Fast Forecasting with Statistical and Econometric Models*; PyCon: Salt Lake City, UT, USA, 2022.
75. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
76. Alajmi, M.S.; Almeshal, A.M. Least Squares Boosting Ensemble and Quantum-Behaved Particle Swarm Optimization for Predicting the Surface Roughness in Face Milling Process of Aluminum Material. *Appl. Sci.* **2021**, *11*, 2126. [CrossRef]
77. Forni, M.; Hallin, M.; Lippi, M.; Reichlin, L. The generalized dynamic factor model consistency and rates. *J. Econom.* **2004**, *119*, 231–255. [CrossRef]
78. Chatfield, C. *Time-Series Forecasting*, revised ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2000.
79. Kim, S.; Kim, H. A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts. *Int. J. Forecast.* **2016**, *32*, 669–679. [CrossRef]
80. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
81. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
82. Waghmare, K.; Ziegel, J. Proper scoring rules for estimation and forecast evaluation. *arXiv* **2025**, arXiv:2504.01781. [CrossRef]
83. van der Meer, D.; Pinson, P.; Camal, S.; Kariniotakis, G. CRPS-based online learning for nonlinear probabilistic forecast combination. *Int. J. Forecast.* **2024**, *40*, 1449–1466. [CrossRef]
84. Wang, S.; Wang, Q.; Lu, H.; Zhang, D.; Xing, Q.; Wang, J. Learning about tail risk: Machine learning and combination with regularization in market risk management. *Omega* **2025**, *133*, 103249. [CrossRef]
85. Taillardat, M.; Fougères, A.L.; Naveau, P.; de Fondeville, R. Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *Int. J. Forecast.* **2023**, *39*, 1448–1459. [CrossRef]
86. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
87. Das, A.; Faw, M.; Sen, R.; Zhou, Y. In-Context Fine-Tuning for Time-Series Foundation Models. *arXiv* **2024**, arXiv:2410.24087.