HINT: Hierarchical Coherent Networks For Constrained Probability Forecasting

Anonymous Authors[1]

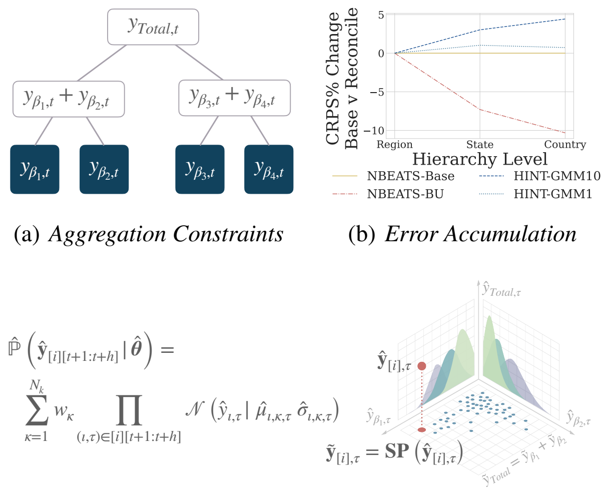
Abstract

Large collections of time series data are commonly organized into hierarchies with different levels of aggregation. We present Hierarchical Coherent Networks (`HINT`), a forecasting framework that adheres to these aggregation constraints. We specialize `HINT` in the task via a multivariate mixture optimized with composite likelihood and made coherent via bootstrap reconciliation. Additionally, we robustify the networks to stark time series scale variations, incorporating normalized feature extraction and recomposition of output scales within their architecture. We demonstrate improved accuracy compared to the existing state-of-the-art. We conduct ablation studies on our model's components and its theoretical foundations. `HINT`'s code is available at this http URL.

1. Introduction

Hierarchical forecasting arises when collections of time series are organized under different aggregation levels. In such scenarios, it is important to ensure the forecasts' *coherence* so that the forecast at disaggregate levels adds-up to the aggregate forecast. In recent years hierarchical reconciliation has become standard across industries, such as supply chain management (Babai et al., 2022; Makridakis et al., 2020b), electricity generation (Nystrup et al., 2020; Ben Taieb et al., 2021), macroeconomics, and tourism management (Eckert et al., 2021; Kourentzes & Athanasopoulos, 2019).

Despite progress in extending neural networks toward hierarchical forecasting, current solutions encounter some limitations like relying on restrictive probabilistic assumptions, enforcing only weak coherence, poor scalability for multivariate approaches, and exhibiting modest accuracy improvements over statistical baselines.


$y_{Total,t}$
$y_{\beta_1,t}+y_{\beta_2,t}$
$y_{\beta_3,t}+y_{\beta_4,t}$
$y_{\beta_1,t}$
$y_{\beta_2,t}$
$y_{\beta_3,t}$
$y_{\beta_4,t}$
CRPS% Change
Base v Reconcile
5
0
−5
−10
Region
State
Country
Hierarchy Level
NBEATS-Base
NBEATS-BU
HINT-GMM10
HINT-GMM1
$\hat{\mathbb{P}}\left(\hat{\mathbf{y}}_{[i][t+1:t+h]}\,|\,\hat{\boldsymbol{\theta}}\right)=\sum_{\kappa=1}^{N_k} w_\kappa \prod_{(\iota,\tau)\in[i][t+1:t+h]} \mathcal{N}\left(\hat{y}_{\iota,\tau}\,|\,\hat{\mu}_{\iota,\kappa,\tau}\;\hat{\sigma}_{\iota,\kappa,\tau}\right)$
$\hat{y}_{Total,\tau}$
$\hat{\mathbf{y}}_{[i],\tau}$
$\hat{y}_{\beta_1,\tau}$
$\hat{y}_{\beta_2,\tau}$
$\tilde{\mathbf{y}}_{[i],\tau}=\mathbf{SP}\left(\hat{\mathbf{y}}_{[i],\tau}\right)$
$\tilde{y}_{Total}=\tilde{y}_{\beta_1}+\tilde{y}_{\beta_2}$

(a) *Aggregation Constraints*

(b) *Error Accumulation*

(c) *Hierarchical Coherent Networks*

Figure 1. (a) Hierarchical forecasting is a multivariate regression problem with aggregation constraints. (b) Summing disaggregated level's forecasts (BU-reconciliation) can accumulate errors. (c) Neural forecast models robust to stark differences in series' scales and a flexible coherent probabilistic output are a solution.

This work addresses the mentioned limitations by introducing the *Hierarchical Coherent Networks* (`HINT`) family of models. Our contributions are summarized below:

(i) **Hierarchical Multivariate Mixture.** Our modular architecture leverages a task-specialized multivariate mixture, optimized with *composite likelihood* and reconciled via *bootstrap sample reconciliation*.

(ii) **Temporal Scale Invariant Networks**. Our residual-learning framework normalizes inputs into the network's non-linearities operating range and recomposes its output scales through a global skip connection, improving accuracy and training robustness.

(iii) **State-of-the-art results** on five hierarchical benchmark datasets: Australian labour, Bay Area lane traffic rates, quarterly and monthly Australian tourist visits, and daily views of Wikipedia articles.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

This paper is structured as follows. Section 2 reviews relevant literature, and introduces notation, Section 3 describes the methodology, Section 4 describes and analyzes our empirical results. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Hierarchical Reconciliation

Classic hierarchical forecasting methods involve a two-stage process in which a set of univariate statistical base forecasts are reconciled. There is a long literature on these reconciliation strategies that include `BottomUp` (Orcutt et al., 1968; Dunn et al., 1976), `TopDown` (Gross & Sohl, 1990; Fliedner, 1999), and more recent optimal reconciliation strategies like `Comb` (Hyndman et al., 2011), `MinTrace` (Wickramasuriya et al., 2019) and `ERM` (Ben Taieb & Koo, 2019).

Probabilistic forecast reconciliation is at the forefront of hierarchical forecasting research. Among the few methods capable of probabilistic coherence, there is `PERMBU` that infuses multivariate dependencies to bottom-level probabilities using copulas (Ben Taieb et al., 2017), `NORMALITY` that constructs a multivariate reconciled distribution building upon Gaussian assumptions (Wickramasuriya, 2023), and `BOOTSTRAP` that generates reconciled forecasts with bootstrap sample reconciliation (Panagiotelis et al., 2023).

2.2. Hierarchical Neural Forecasting

Neural network based methods have gained popularity in forecasting applications, outperforming most alternatives. They have become widespread in industrial forecasting and have consistently performed well in forecasting competitions such as M4 (Makridakis et al., 2020a) and M5 (Makridakis et al., 2020b). As surveys show, in recent years, the academic community has greatly renovated interest in the topic (Benidis et al., 2020).

The literature has permeated into hierarchical forecasting, with contributions like `SHARQ` (Han et al., 2021), `HIRED` (Paria et al., 2021), and `PROFHIT` (Kamarthi et al., 2022) approximate coherent methods using variants of bottom-up aggregation regularization. Fully coherent approaches include `HierE2E` (Rangapuram et al., 2021) a multivariate approach that incorporates `MinTrace`-like reconciliation in the network's optimization, `TDProb` (Das et al., 2022) that learns `TopDown` proportions to probabilistically reconcile univariate base models.

Despite recent progress in extending neural networks toward hierarchical forecasting, existing solutions still face challenges: (i) they rely on restrictive probabilistic assumptions or are not entirely coherent; (ii) multivariate approaches' computational complexity scales poorly; (iii) the accuracy improvements over statistical baselines are still modest.

2.3. Mathematical Notation

A hierarchical time series (HTS) is a multivariate time series under aggregation constraints. We denote the HTS by the vector $\mathbf{y}_{[i],t} = [\mathbf{y}^{\intercal}_{[a],t} \,|\, \mathbf{y}^{\intercal}_{[b],t}]^{\intercal} \in \mathbb{R}^{N_a+N_b}$, for time step t, where $[a]$, $[b]$ denote respectively the aggregate and bottom level indices. The total number of series in the hierarchy is $|[i]| = (N_a + N_b)$. We distinguish between the time indices $[t]$ and forecast indices $\tau \in [t+1 : t+h]$, and hierarchical, bottom and aggregate indexes $\iota \in [i], \beta \in [b]$, $\alpha \in [a]$.

At any time t, the constraints are $\mathbf{y}_{[a],t} = \mathbf{A}_{[a][b]}\mathbf{y}_{[b],t}$ where $\mathbf{A}_{[a][b]}$ denotes the relationship between the bottom-level series to the upper-level series. We can write the HTS as

$$\mathbf{y}_{[i],t} = \mathbf{S}_{[i][b]}\mathbf{y}_{[b],t} \quad \Leftrightarrow \quad \begin{bmatrix}\mathbf{y}_{[a],t}\\ \mathbf{y}_{[b],t}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{[a][b]}\\ \mathbf{I}_{[b][b]}\end{bmatrix} \mathbf{y}_{[b],t} \tag{1}$$

where $\mathbf{S}_{[i][b]}$ and $\mathbf{I}_{[b][b]}$ are the summing and identity matrices. Figure 1(a) is an example with $N_b = 4$ and $N_a = 3$:

$$\begin{aligned}\mathbf{y}_{[a],t} &= [y_{\text{Total},t},\; y_{\beta_1,t} + y_{\beta_2,t},\; y_{\beta_3,t} + y_{\beta_4,t}]^{\intercal}\,,\\ \mathbf{y}_{[b],t} &= [y_{\beta_1,t},\; y_{\beta_2,t},\; y_{\beta_3,t},\; y_{\beta_4,t}]^{\intercal}\,,\end{aligned} \tag{2}$$

where $y_{\text{Total},t} = y_{\beta_1,t} + y_{\beta_2,t} + y_{\beta_3,t} + y_{\beta_4,t}$. The summing matrix associated to Figure 1(a) is given by

$$\mathbf{S}_{[i][b]} = \begin{bmatrix}\mathbf{A}_{\mathrm{[a][b]}}\\ \\ \mathbf{I}_{\mathrm{[b][b]}}\end{bmatrix} = \left[\begin{array}{cccc} 1 & 1 & 1 & 1\\ 1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\\ \hline 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{array}\right].$$

Definition 2.1. (Probabilistic Coherence). Let $(\Omega_{[b]}, \mathcal{F}_{[b]}, \mathbb{P}_{[b]})$ be a probabilistic forecast space, with $\mathcal{F}_{[b]}$ a σ-algebra on $\mathbb{R}^{N_b}$. Let $\mathbf{S}(\cdot) : \Omega_{[b]} \mapsto \Omega_{[i]}$ be the constraints' implied transformation. A coherent probabilistic forecast space $(\Omega_{[i]}, \mathcal{F}_{[i]}, \mathbb{P}_{[i]})$ satisfies:

$$\begin{gathered}\mathbb{P}_{[i]}\left(\mathbf{S}(\mathcal{B})\right) = \mathbb{P}_{[b]}\left(\mathcal{B}\right)\\ \text{for any set } \mathcal{B} \in \mathcal{F}_{[b]} \text{ and image } \mathbf{S}(\mathcal{B}) \in \mathcal{F}_{[i]}\end{gathered} \tag{3}$$

that is, it assigns a zero probability to any set that does not contain any coherent forecasts (Panagiotelis et al., 2023).

Definition 2.2. (Hierarchical Reconciliation). For time t, horizon h, and forecast indexes $\tau \in [t+1 : t+h]$. Reconciliation for point forecasts $\hat{\mathbf{y}}_{[i],\tau}$, is denoted by:

$$\tilde{\mathbf{y}}_{[i],\tau} = \mathbf{S}_{[i][b]}\mathbf{P}_{[b][i]}\hat{\mathbf{y}}_{[i],\tau} = \mathbf{SP}(\hat{\mathbf{y}}_{[i],\tau}) \tag{4}$$

where $\mathbf{P}_{[b][i]}$ is defined by the reconciliation technique . And $\mathbf{SP}(\cdot) : \Omega_{[i]} \mapsto \Omega_{[b]} \mapsto \Omega_{[i]}$ is the reconciliation's implied transformation (Hyndman & Athanasopoulos, 2018).

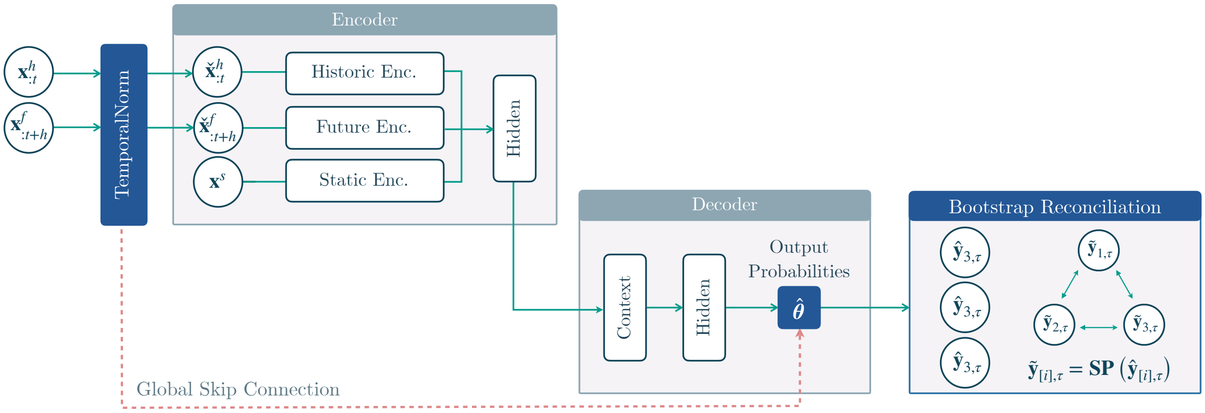

$\mathbf{x}^{h}_{:t}$
$\mathbf{x}^{f}_{:t+h}$
TemporalNorm
Encoder
$\check{\mathbf{x}}^{h}_{:t}$
$\check{\mathbf{x}}^{f}_{:t+h}$
$\mathbf{x}^{s}$
Historic Enc.
Future Enc.
Static Enc.
Hidden
Decoder
Context
Hidden
Output Probabilities
$\hat{\boldsymbol{\theta}}$
Bootstrap Reconciliation
$\hat{\mathbf{y}}_{3,\tau}$
$\hat{\mathbf{y}}_{3,\tau}$
$\hat{\mathbf{y}}_{3,\tau}$
$\tilde{\mathbf{y}}_{1,\tau}$
$\tilde{\mathbf{y}}_{2,\tau}$
$\tilde{\mathbf{y}}_{3,\tau}$
$\tilde{\mathbf{y}}_{[i],\tau} = \mathbf{SP}\left(\hat{\mathbf{y}}_{[i],\tau}\right)$
Global Skip Connection

Figure 2. The `HINT` model family utilizes a specialized multivariate mixture probability output layer that achieves coherency via bootstrap sample reconciliation $\tilde{\mathbf{y}}_{[i],\tau} = \mathbf{SP}(\hat{\mathbf{y}}_{[i],\tau})$. Additionally `HINT` incorporates the `TemporalNorm` module to augment networks with scale-invariance through input normalization and output scale recomposition through a global skip connection.

3. `HINT` Methodology

The `HINT` model family estimates the following conditional probability under coherency constraints from Definition 2.1:

$$\mathbb{P}\left(\mathbf{y}_{[i][t+1:t+h]}\,|\,\boldsymbol{\theta}\right) \tag{5}$$

where $\boldsymbol{\theta}$ depends on $\{\mathbf{x}^{h}_{[i][:t]},\ \mathbf{x}^{f}_{[i][:t+h]},\ \mathbf{x}^{s}_{[i]}\}$ historic, future and static variables. Here we describe our proposed approach, its high-level diagram and main principles of operation are depicted in Figure 2.

3.1. Hierarchical Multivariate Mixture

`HINT` is a highly modular system that supports a wide range of probabilistic outputs. We leverage `HINT`'s flexibility to accommodate a multivariate Gaussian mixture model specialized in hierarchical forecasting. Its conditional forecast distribution is described by:

$$\hat{\mathbb{P}}\left(\mathbf{y}_{[i][t+1:t+h]}\middle|\,\hat{\boldsymbol{\theta}}\right) = \sum_{\kappa=1}^{N_k}\hat{w}_{\kappa}\prod_{(\iota,\tau)\in[i][t+1:t+h]}\mathcal{N}\left(y_{\iota,\tau}\middle|\,\hat{\mu}_{\iota,\kappa,\tau}\ \hat{\sigma}_{\iota,\kappa,\tau}\right) \tag{6}$$

The multivariate mixture has advantageous theoretical properties, proven in Appendix A. It can arbitrarily approximate univariate distributions and describe the series' correlations.

`HINT`'s achieves high computational efficiency, because we optimize it through composite-likelihood (Lindsay, 1988; Varin et al., 2011) of the series in the SGD batches approximating the full joint distribution from Equation (6). To further enhance the prediction's efficiency we opt for architectures using the multi-step forecast strategy (Atiya & Taieb, 2016; Lim et al., 2021; Challu et al., 2023).

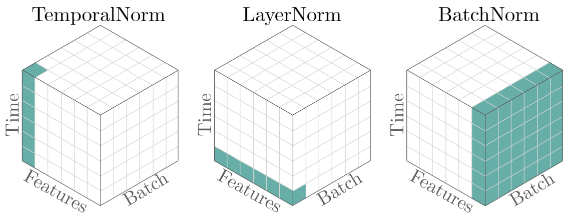


Figure 3. Temporal normalization (left), layer normalization (center) and batch normalization (right). The entries in green show the components used to compute the normalizing statistics.

We ensure the coherence via bootstrap sample reconciliation. Figure 1.c shows it restoring the aggregation constraints into the base samples. Let $\mathcal{H}$ be a coherent forecast set, and $\mathbf{SP}^{-1}(\cdot)$ a reconciliation's inverse image, Appendix A analytically derives the coherent forecast distribution:

$$\tilde{\mathbb{P}}\left(\tilde{\mathbf{y}}_{[i],\tau}\in\mathcal{H}|\,\tilde{\boldsymbol{\theta}}\right) = \hat{\mathbb{P}}\left(\hat{\mathbf{y}}_{[i],\tau}\in\mathbf{SP}^{-1}(\mathcal{H})|\,\hat{\boldsymbol{\theta}}\right) \tag{7}$$

3.2. Temporal Scale Invariant Networks

Neural networks' recent success in forecasting follows the adoption of cross-learning optimization, which enables flexible global models to fit without the risk of overfitting (Smyl, 2019; Semenoglou et al., 2021). However, for hierarchical forecasting applications with natural scale variability, a robustified version of cross-learning optimization is necessary.

The `TemporalNorm` module enhances any architecture to adapt its inputs to the its non-linearities operating range. Its global skip-connection reformulates the task into learning a residual function referenced to the time series level and scale. The residual-learning framework significantly improves the network's accuracy, as Appendix C shows.

Table 1. Empirical evaluation of probabilistic coherent forecasts. Mean scaled continuous ranked probability score (sCRPS), averaged over 10 random seeds, overall hierarchy series. The best result is highlighted (lower measurements are preferred).

† The PROFHIT results differ from (Kamarthi et al., 2022), as the only available implementation suffers from significant numerical instability in its optimization.

* Best performing variant of TopDown (avg. proportions, proportions avg.), and MinTrace (ols, wls, shrinkage) reported. ** The PERMBU/TopDown only available for strictly hierarchical datasets.

		HINT-GMM (Ours)			OTHER		BOOTSTRAP			PERMBU**		
	DATASET	NHITS	TFT	TCN	HierE2E	PROFHIT†	BottomUp	TopDown*	MinTrace*	BottomUp	TopDown*	MinTrace*
sCRPS	Labour	**.0067±.0001**	.0089±.0001	.0120±.0001	.0171±.0003	.2138±.0070	.0078±.0001	.0668±.0000	.0073±.0000	.0077±.0001	.0740±.0001	.0069±.0001
	Traffic	.0589±.0004	.0602±.0004	.0600±.0002	**.0426±.0008**	.1137±.0022	.0736±.0024	.0741±.0012	.0608±.0014	.0849±.0009	.0708±.0008	.0651±.0008
	Tourism	.0666±.0007	.0665±.0004	**.0536±.0004**	.0761±.0007	.1358±.0033	.0682±.0018	.1040±.0014	.0703±.0017	.0649±.0016	.0898±.0012	.0680±.0016
	Tourism-L	**.1176±.0002**	.1354±.0005	.1550±.0006	.1424±.0019	.2139±.0014	.1375±.0013	-	.1313±.0009	-	-	-
	Wiki2	.3625±.0045	**.2447±.0007**	.2918±.0015	.2592±.0031	.4009±.0028	.2894±.0038	.3231±.0037	.2808±.0035	.3920±.0044	.4269±.0036	.3821±.0049
relMSE	Labour	.5802±.0131	.464±.0148	.801±.0637	.8165±0.0353	6.774×10^3	0.5382±.0000	16.8204±.0000	**0.3547±.0000**			
	Traffic	.1212±.0051	.1291±.0036	.1226±.0024	**.0328±0.0019**	0.4536±.0224	0.1392±.0000	0.06144±.0000	0.0744±.0000			
	Tourism	.0898±.0031	.0932±.0018	**.0387±.0007**	.1471±0.0046	0.9745±.0803	0.1002±.0000	0.1919±.0000	0.1235±.0000			
	Tourism-L	**.0577±.0009**	.0834±.0017	.1816±.0031	.2449±0.0096	1.0401±.0296	0.3070±.0000	-	0.1375±.0000			
	Wiki2	1.044±.0531	**.1884±.0012**	.2183±.0036	.6598±0.0249	0.7901±.0384	1.0163±.0000	1.4482±.0000	1.0068±.0000			

4. Experimental Results

Datasets. We follow experimental protocols established in previous research by Rangapuram et al. (2021), Olivares et al. (2023) and Kamarthi et al. (2022). The benchmark datasets are: Monthly Australian Labour (Australian Bureau of Statistics, 2019), SF Bay Area daily Traffic (Dua & Graff, 2017), Quarterly Australian Tourism visits (Tourism Australia, Canberra, 2005), Monthly Australian Tourism-L visits (Tourism Australia, Canberra, 2019), and daily Wiki2 views (Anava et al., 2018). Appendix B includes a detailed dataset's exploration.

Evaluation Metrics. To assess the forecast accuracy of our method, we employ the scaled Continuous Ranked Probability Score (sCRPS; Matheson & Winkler 1976; Makridakis et al. 2022) and the Relative Mean Squared Error (relMSE; Hyndman & Koehler 2006; Olivares et al. 2023).

$$\mathrm{sCRPS}(\mathbb{P}, \mathbf{y}_{[i],\tau}) = \frac{2}{|[i]|} \sum_{i} \frac{\int_0^1 \mathrm{QL}(\mathbb{P}_{i,\tau}, y_{i,\tau})_q dq}{\sum_i |y_{i,\tau}|}$$

$$\mathrm{relMSE}(\mathbf{y}_{[i]}, \hat{\mathbf{y}}_{[i]}, \check{\mathbf{y}}_{[i]}) = \frac{\mathrm{MSE}(\mathbf{y}_{[i]}, \hat{\mathbf{y}}_{[i]})}{\mathrm{MSE}(\mathbf{y}_{[i]}, \check{\mathbf{y}}_{[i]})} \tag{8}$$

where $\mathrm{QL}(\hat{\mathbb{P}}_{i,\tau}, y_{i,\tau})_q$ denotes the q-level quantile loss[1], and $\check{\mathbf{y}}_{[i]}$ is the Naive forecast relative scaler.

Baselines. In our main experiment, we compare the predictions of numerous SoTA probabilistic coherent methods. Neural forecasting baselines include (1) HierE2E (Rangapuram et al., 2021), (2) PROFHIT (Kamarthi et al., 2022), while statistical baselines include variants of (3) BOOTSTRAP (Panagiotelis et al., 2023), and (4) PERMBU probabilistic reconciliation (Ben Taieb et al., 2017) in combination with BottomUp (Orcutt et al., 1968), TopDown (Gross & Sohl, 1990), and MinTrace (Wickramasuriya et al., 2019) mean reconcilers.

HINT configurations. We showcase our methodology's outstanding modularity by augmenting three well-established neural forecast architectures NHITS (Oreshkin et al., 2020; Olivares et al., 2022a; Challu et al., 2023), TFT; (Lim et al., 2021) and TCN (Bai et al., 2018). We report results from HINT best configurations, based on Appendix C.2,C.3,C.1's validation ablation studies. We provide software and hyperparameter details in Appendix D.

4.1. Empirical Results

The HINT achieved the best performance on four datasets, with the Traffic dataset exception, improving sCRPS by 8.1% (extended results in Appendix E). The results show how scale-decoupled optimization and a multivariate joint distribution successfully adapt the latest neural architecture innovations for hierarchical forecasting tasks. Our ablation studies confirm improvement origins in the HINT method. First, the mixture distribution improves performance upon simpler and more constraining probabilistic output assumptions. Second, that scale-decouple optimization is a crucial enabler of cross-learning under stark scale variations.

5. Discussion and Conclusion

We introduced HINT, a new neural network family for hierarchically coherent forecasting. Two highly modular novelties define our approach, an accurate and efficient multivariate mixture optimized with composite likelihood and transformed via bootstrap sample reconciliation; and incorporating normalized feature extraction and recomposition of output scales within the network's architecture. HINT's enhancements can be applied to most neural forecasting architectures enabling their probabilistic coherent forecasts.

We showcase HINT's accuracy gains in an empirical comparison to statistical and state-of-the-art neural forecast models, improving sCRPS by 8.1% on average across datasets over the second-best alternative. We conclude with an exploration of our framework's theoretical foundations.

[1] We use a Riemann approximation to the sCRPS with the difference dq for quantile intervals of 1 percent.

References

Anava, O., Kuznetsov, V., and (Google Inc. Sponsorship). Web traffic time series forecasting, forecast future traffic to wikipedia pages. Kaggle Competition, 2018. URL https://www.kaggle.com/c/web-traffic-time-series-forecasting/.

Atiya, A. and Taieb, B. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE transactions on neural networks and learning systems*, 27(1): 2162–2388, 2016. URL https://pubmed.ncbi.nlm.nih.gov/25807572/.

Australian Bureau of Statistics. Labour force, australia. Accessed Online, 2019. URL https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6202.0Dec%202019?OpenDocument.

Babai, M. Z., Boylan, J. E., and Rostami-Tabar, B. Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. *International Journal of Production Research*, 60(1):324–348, 2022. doi: 10.1080/00207543.2021.2005268. URL https://doi.org/10.1080/00207543.2021.2005268.

Bai, S., Kolter, J. Z., and Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *Computing Research Repository*, abs/1803.01271, 2018. URL http://arxiv.org/abs/1803.01271.

Ben Taieb, S. and Koo, B. Regularized regression for hierarchical forecasting without unbiasedness conditions. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1337–1347, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330976. URL https://doi.org/10.1145/3292500.3330976.

Ben Taieb, S., Taylor, J. W., and Hyndman, R. J. Coherent probabilistic forecasts for hierarchical time series. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3348–3357. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/taieb17a.html.

Ben Taieb, S., Taylor, J. W., and Hyndman, R. J. Hierarchical probabilistic forecasting of electricity demand with smart meter data. *Journal of the American Statistical Association*, 116(533):27–43, 2021. doi: 10.1080/01621459.2020.1736081. URL https://doi.org/10.1080/01621459.2020.1736081.

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Callot, L., and Januschowski, T. Neural forecasting: Introduction and literature overview. *Computing Research Repository*, 2020.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24, pp. 2546–2554. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler, M., and Dubrawski, A. NHITS: Neural Hierarchical Interpolation for Time Series forecasting. In *The Association for the Advancement of Artificial Intelligence Conference 2023 (AAAI 2023)*, 2023. URL https://arxiv.org/abs/2201.12886.

Das, A., Kong, W., Paria, B., and Sen, R. A deep top-down approach to hierarchically coherent probabilistic forecasting, 2022. URL https://arxiv.org/abs/2204.10414.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dunn, D. M., Williams, W. H., and Dechaine, T. L. Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association*, 71(353): 68–71, 1976.

Eckert, F., Hyndman, R. J., and Panagiotelis, A. Forecasting swiss exports using bayesian forecast reconciliation. *European Journal of Operational Research*, 291(2):693–710, 2021. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2020.09.046. URL https://www.sciencedirect.com/science/article/pii/S037722172030850X.

Fliedner, G. An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers and Operations Research*, 26(10–11):1133–1149, September 1999. ISSN 0305-0548. doi: 10.1016/S0305-0548(99)00017-9. URL https://doi.org/10.1016/S0305-0548(99)00017-9.

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

Gross, C. W. and Sohl, J. E. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, 9(3):233–254, 1990. doi: 10.1002/for.3980090304. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980090304.

Han, X., Dasgupta, S., and Ghosh, J. Simultaneously reconciled quantile forecasting of hierarchically related time series. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 190–198. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/han21a.html.

Hyndman, R. J. and Athanasopoulos, G. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2018. available at https://otexts.com/fpp2/.

Hyndman, R. J. and Khandakar, Y. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, 27(3):1–22, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i03. URL https://www.jstatsoft.org/v027/i03.

Hyndman, R. J. and Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL http://www.sciencedirect.com/science/article/pii/S0169207006000239.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579 – 2589, 2011. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.03.006. URL http://www.sciencedirect.com/science/article/pii/S0167947311000971.

Kamarthi, H., Kong, L., Rodriguez, A., Zhang, C., and Prakash, B. PROFHIT: Probabilistic robust forecasting for hierarchical time-series. *Computing Research Repository*, 06 2022. URL https://arxiv.org/abs/2206.07940.

Kingma, D. P. and Ba, J. ADAM: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations (ICLR), San Diego, 2015.

Kourentzes, N. and Athanasopoulos, G. Cross-temporal coherent forecasts for australian tourism. *Annals of Tourism Research*, 75:393–409, 2019. ISSN 0160-7383. doi: https://doi.org/10.1016/j.annals.2019.02.001. URL https://www.sciencedirect.com/science/article/pii/S0160738319300167.

Lim, B., Arık, S. O., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.03.012. URL https://www.sciencedirect.com/science/article/pii/S0169207021000637.

Lindsay, B. G. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020a. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169207019301128. M4 Competition.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, 10 2020b. URL https://www.researchgate.net/publication/344487258_The_M5_Accuracy_competition_Results_findings_and_conclusions.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., and Winkler, R. L. The m5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4):1365–1385, 2022. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.10.009. URL https://www.sciencedirect.com/science/article/pii/S0169207021001722. Special Issue: M5 competition.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629907.

Nguyen, H. D. and McLachlan, G. J. On approximations via convolution-defined mixture models. *Computing Research Repository*, 2018. URL https://arxiv.org/abs/1611.03974.

Nystrup, P., Lindström, E., Pinson, P., and Madsen, H. Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research*, 280(3):876–888, 2020. doi: 10.1016/j.ejor.2019.07.06. URL https://ideas.repec.org/a/eee/ejores/v280y2020i3p876-888.html.

Olivares, K. G., Challu, C., Marcjasz, G., Weron, R., and Dubrawski, A. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting*, 2022a. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2022.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207022000413.

Olivares, K. G., Garza, F., Luo, D., Challú, C., Mergenthaler, M., Ben Taieb, S., Wickramasuriya, S. L., and Dubrawski, A. HierarchicalForecast: A reference framework for hierarchical forecasting. *Journal of Machine Learning Research, submitted*, abs/2207.03517, 2022b. URL https://arxiv.org/abs/2207.03517.

Olivares, K. G., Meetei, N., Ma, R., Reddy, R., Cao, M., and Dicker, L. Probabilistic hierarchical forecasting with deep poisson mixtures. *International Journal of Forecasting, accepted*, Preprint version available at arXiv:2110.13179, 2023. URL https://arxiv.org/abs/2110.13179.

Orcutt, G. H., Watts, H. W., and Edwards, J. B. Data aggregation and information loss. *The American Economic Review*, 58(4):773–787, 1968. ISSN 00028282. URL http://www.jstor.org/stable/1815532.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., and Hyndman, R. J. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, 306(2):693–706, 2023. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2022.07.040. URL https://www.sciencedirect.com/science/article/pii/S0377221722006087.

Paria, B., Sen, R., Ahmed, A., and Das, A. Hierarchically Regularized Deep Forecasting. In *Submitted to Proceedings of the 39th International Conference on Machine Learning*. PMLR. Working Paper version available at arXiv:2106.07630, 2021.

Paszke et al. Pytorch: An imperative style, high-performance Deep Learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Rangapuram, S. S., Werner, L. D., Benidis, K., Mercado, P., Gasthaus, J., and Januschowski, T. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In Balcan, M. F. and Meila, M. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 06–11 Aug 2021.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., and Assimakopoulos, V. Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, 37(3):1072–1084, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2020.11.009. URL https://www.sciencedirect.com/science/article/pii/S0169207020301850.

Smyl, S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 07 2019. doi: 10.1016/j.ijforecast.2019.03.017.

Titterington, D. M., Afm, S., Smith, A. F., Makov, U., et al. *Statistical analysis of finite mixture distributions*, volume 198. John Wiley & Sons Incorporated, 1985.

Tourism Australia, Canberra. Tourism Research Australia (2005), Travel by Australians. https://www.kaggle.com/luisblanche/quarterly-tourism-in-australia/, 2005.

Tourism Australia, Canberra. Detailed tourism Australia (2005), Travel by Australians, Sep 2019. Accessed at https://robjhyndman.com/publications/hierarchical-tourism/.

Varin, C., Reid, N., and Firth, D. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24309261.

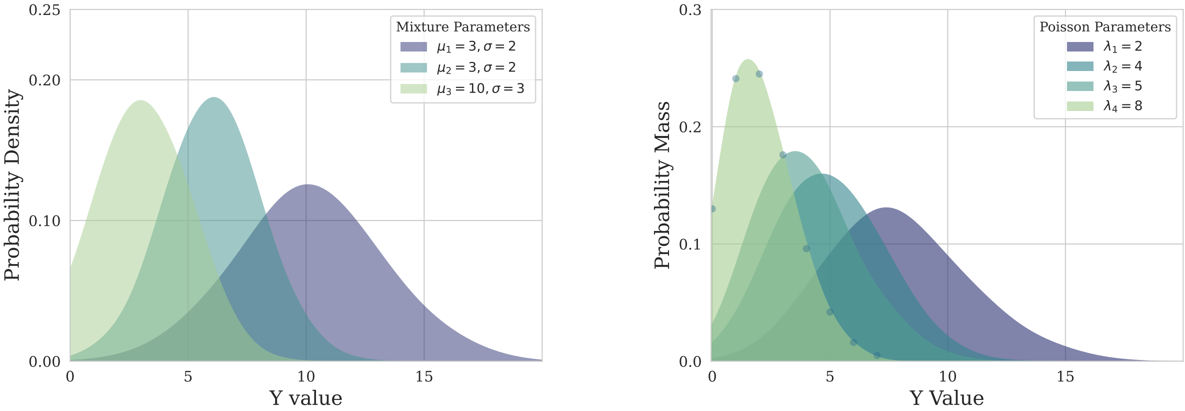
Wickramasuriya, S. L. Probabilistic forecast reconciliation under the Gaussian framework. *Accepted at Journal of Business and Economic Statistics*, 2023.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019. doi: 10.1080/01621459.2018.1448825. URL https://robjhyndman.com/publications/mint/.

Yao, Y., Rosasco, L., and Andrea, C. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. URL https://doi.org/10.1007/s00365-006-0663-2.

A. Hierarchical Multivariate Mixture Properties

A.1. Hierarchical Multivariate Mixture Probability



(a) Gaussian Mixture

(b) Poisson Mixture

Figure 1. HINT's multivariate joint distribution has advantageous properties that make it uniquely suited for hierarchical forecasting. It is highly flexible, capable of efficiently modeling series' relationships, and under minimal restrictions, guarantees probabilistic coherence.

In Section 3, we highlight that HINT boasts a flexible and modular model family that can handle various probabilistic outputs. We use this flexibility to enhance the networks with a multivariate hierarchical mixture density. Specifically, the mixture model describes the joint probability of the hierarchical multivariate time series $\mathbf{Y}_{[i][t+1:t+h]} \in \mathbb{R}^{Ni \times h}$ as follows:

$$\mathbb{P}\left(Y_{[i][t+1:t+h]} = \mathbf{y}_{[i][t+1:t+h]} | \,\hat{\boldsymbol{\theta}}\right) = \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa} \prod_{(\iota,\tau)\in[i][t+1:t+h]} \mathcal{N}\left(y_{\iota,\tau} |\, \hat{\mu}_{\iota,\kappa,\tau}\ \hat{\sigma}_{\iota,\kappa,\tau}\right)$$

where the mixture describes individual series through the location $\hat{\mu}_{\iota,\kappa,\tau}$ and variance parameters $\hat{\sigma}_{\iota,\kappa,\tau}$. For simplicity we denote the combined parameters $\hat{\boldsymbol{\theta}}_{[i][k][t+1:t+h]} = [\hat{\boldsymbol{\mu}}_{[i][k][t+1:t+h]} \mid \hat{\boldsymbol{\sigma}}_{[i][k][t+1:t+h]}]$.

Under reasonable assumptions for the underlying probability, the mixture distribution offers arbitrary approximation guarantees (Titterington et al., 1985; Nguyen & McLachlan, 2018). We can control its flexibility by adjusting the number of components $|[k]| = N_k$. Furthermore, the mixture is not limited to Gaussian components; we can extend it to include discrete variables. Figure A.1 presents an example of its marginal probabilities.

Conditional Independence: A key consequence of the hierarchical mixture probability in Equation (A.1), is the assumption that the modeled series $\mathbf{y}_{[i][t+1:t+h]}$ are conditionally independent given the latent parameters $\hat{\boldsymbol{\mu}}_{[i][k][t+1:t+h]}$ and $\hat{\boldsymbol{\sigma}}_{[i][k][t+1:t+h]}$. That is for any series and horizons $(\iota, \tau) \neq (\iota', \tau')$, $(\iota, \tau), (\iota', \tau') \in [b][t+1:t+h]$ and $\kappa \in [k]$:

$$\mathbb{P}(Y_{\iota,\tau}, Y_{\iota',\tau'} | \hat{\theta}_{\iota,\kappa,\tau}, \hat{\theta}_{\iota',\kappa,\tau'}) = \mathbb{P}(Y_{\iota,\tau} | \hat{\theta}_{\iota,\kappa,\tau}) \hat{\mathbb{P}}(Y_{\iota',\tau'}, | \hat{\theta}_{\iota',\kappa,\tau'}) \tag{9}$$

Computational Efficiency: To handle large-scale data scenarios, we explicitly avoid using a multivariate covariance matrix, which has an $\mathcal{O}(N_i^2)$ complexity. Instead, we rely on the mixture latent variables κ and its associated weights $\hat{\mathbf{w}}_{[k]} \in [0,1]^{N_k}$, $\hat{\mathbf{w}}_{[k]} \geq 0$ and $\sum_{\kappa=1}^{N_k} \hat{w}_{\kappa} = 1$ to model the series correlations. We show the relationship between the mixture components and the covariance in Appendix A.3. Instead of relying on the Markov assumption, we have adopted a joint multi-step forecasting approach that can significantly enhance the computational efficiency of our algorithm. By making predictions in a single forward pass, we can avoid the need for recurrent computations.

In the following subsections, we delve into the properties of the hierarchical mixture, such as the analytic version of its implied marginal probability, the relationship between its covariance and the number of mixture components, the bootstrap sample reconciled probability, and the optimization strategies that make it well-suited for large-scale applications. The proofs are inspired on previous work by (Olivares et al., 2023), generalized and extended.

A.2. Marginal Distributions

We define the joint distribution of all hierarchical time series in Equation (A.1). By integrating the joint probability on the remaining series and time indices, we can obtain the marginal distribution for a single future horizon $\tau \in [t+1:t+h]$ and series $\iota \in [i]$. We express the resulting marginal distribution as follows:

$$\mathbb{P}(Y_{\iota,\tau} = y_{\iota,\tau} |\, \hat{\boldsymbol{\theta}}) = \sum_{\kappa=1}^{N_k} w_{\kappa} \mathcal{N}\left(y_{\iota,\tau} |\, \hat{\mu}_{\iota,\kappa,\tau}\ \hat{\sigma}_{\iota,\kappa,\tau}\right) \tag{10}$$

Proof.

$$\begin{aligned}
\mathbb{P}(Y_{\iota,\tau} = y_{\iota,\tau} |\, \hat{\boldsymbol{\theta}}) &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \mathbb{P}\left(Y_{[i][t+1:t+h]} = \mathbf{y}_{[i][t+1:t+h]} |\, \hat{\boldsymbol{\theta}}\right) \delta y_{\iota',\tau'} \setminus \delta y_{\iota,\tau} \\
&= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa}\ \mathbb{P}(y_{\iota,\tau} | \hat{\theta}_{\iota,\kappa,\tau}) \times \prod_{(\iota',\tau') \in [i][t+1:t+h] \setminus (\iota,\tau)} \sum_{y_{\iota',\tau'}} \mathbb{P}(y_{\iota',\tau'} | \hat{\theta}_{\iota',\kappa,\tau'}) \delta y_{\iota',\tau'} \\
&= \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa}\ \mathbb{P}(y_{\iota,\tau} | \hat{\theta}_{\iota,\kappa,\tau}) \times \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \prod_{(\iota',\tau') \in [i][t+1:t+h] \setminus (\iota,\tau)} \sum_{y_{\iota',\tau'}} \mathbb{P}(y_{\iota',\tau'} | \hat{\theta}_{\iota',\kappa,\tau'}) \delta y_{\iota',\tau'} \\
&= \sum_{\kappa=1}^{N_k} w_{\kappa}\ \mathbb{P}(y_{\iota,\tau} | \hat{\theta}_{\iota,\kappa,\tau}) = \sum_{\kappa=1}^{N_k} w_{\kappa} \mathcal{N}\left(y_{\iota,\tau} |\, \hat{\mu}_{\iota,\kappa,\tau}\ \hat{\sigma}_{\iota,\kappa,\tau}\right)
\end{aligned}$$

□

By removing all other time series and forecast horizons from the joint probability product, the conditional independence expressed in Equation A.1 can efficiently generate forecast distributions for individual variables.

A.3. Efficient Covariance Matrix Low-Rank Approximation

Due to the computational challenges of estimating high-dimensional covariance matrices, existing multivariate methods are limited in their ability to handle a large number of series. To overcome this challenge, our method utilizes a low-rank covariance structure implied by the latent variables of the mixture probability, thereby avoiding the need to compute the covariance matrix explicitly. By doing so, we significantly reduce the number of parameters and enable the modeling of time-varying correlations across millions of time series.

Let a multivariate random variable $\mathbf{Y}_{[i][t+1:t+h]} \in \mathbb{R}^{Ni \times h}$ distribution be described by the mixture from Equation (A.1), the non-diagonal terms of its implied covariance is the following $N_k - 1$ rank matrix:

$$\mathrm{Cov}(\mathbf{Y}_{[i],\tau}) = \sum_{\kappa=1}^{N_k} \hat{\mathbf{w}}_{\kappa} (\hat{\boldsymbol{\mu}}_{[i],\kappa,\tau} - \bar{\boldsymbol{\mu}}_{[i],\tau})(\hat{\boldsymbol{\mu}}_{[i],\kappa,\tau} - \bar{\boldsymbol{\mu}}_{[i],\tau})^{\intercal} \in \mathbb{R}^{N_i \times N_i} \tag{11}$$

Proof. We will start by showing that for a pair of series the covariance function is given by:

$$\mathrm{Cov}(Y_{\iota,\tau}, Y_{\iota',\tau'}) = \overline{\sigma}_{\iota,\tau} \mathbb{1}(\iota = \iota') \mathbb{1}(\tau = \tau') + \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa} \left(\hat{\mu}_{\iota,\kappa,\tau} - \overline{\mu}_{\iota,\tau}\right) \left(\hat{\mu}_{\iota',\kappa,\tau'} - \overline{\mu}_{\iota',\tau'}\right) \tag{12}$$

By the law of total covariance:

$$\mathrm{Cov}(Y_{\iota,\tau}, Y_{\iota',\tau'}) = \mathbb{E}\left[\mathrm{Cov}(Y_{\iota,\tau}, Y_{\iota',\tau'} | \hat{\theta}_{\iota,\kappa,\tau}, \hat{\theta}_{\iota',\kappa,\tau'})\right] + \mathrm{Cov}\left(\mathrm{E}\left[Y_{\iota,\tau} | \hat{\theta}_{\iota,\kappa,\tau}\right], \mathbb{E}\left[Y_{\iota',\tau'} | \hat{\theta}_{\iota',\kappa,\tau'}\right]\right)$$

Using the conditional independence from Equation (9). We can rewrite conditional covariance expectation:

$$\begin{aligned}\mathbb{E}\left[\mathrm{Cov}(Y_{\iota,\tau}, Y_{\iota',\tau'}|\hat{\theta}_{\iota,\kappa,\tau}, \hat{\theta}_{\iota',\kappa,\tau'})\right] &= \mathbb{E}\left[\mathrm{Var}(Y_{\iota,\tau}|\hat{\theta}_{\iota,\kappa,\tau})\right] \mathbb{1}(\iota = \iota')\mathbb{1}(\tau = \tau') \\ &= \mathbb{E}\left[\hat{\sigma}_{\iota,\kappa,\tau}\right] \mathbb{1}(\iota = \iota')\mathbb{1}(\tau = \tau') \\ &= \overline{\sigma}_{\iota,\tau}\mathbb{1}(\iota = \iota')\mathbb{1}(\tau = \tau')\end{aligned}$$

where $\overline{\sigma}_{\iota,\tau} = \mathbb{E}\left[\hat{\sigma}_{\iota,\kappa,\tau}\right] = \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa}\hat{\sigma}_{\iota,\kappa,\tau}$.

In the second term, because the conditional distributions are Normal we have

$$\mathrm{E}\left[Y_{\iota,\tau}|\hat{\theta}_{\iota,\kappa,\tau}\right] = \hat{\mu}_{\iota,\kappa,\tau} \quad \text{and} \quad \mathrm{E}\left[Y_{\iota',\tau'}|\hat{\theta}_{\iota',\kappa,\tau'}\right] = \hat{\mu}_{\iota',\kappa,\tau'}$$

Which implies

$$\mathrm{Cov}\left(\mathbb{E}\left[Y_{\iota,\tau}|\hat{\theta}_{\iota,\kappa,\tau}\right], \mathbb{E}\left[Y_{\iota',\tau'}|\hat{\theta}_{\iota',\kappa,\tau'}\right]\right) = \sum_{\kappa=1}^{N_k} \hat{w}_{\kappa}\left(\hat{\mu}_{\iota,\kappa,\tau} - \bar{\mu}_{\iota,\tau}\right)\left(\hat{\mu}_{\iota',\kappa,\tau'} - \bar{\mu}_{\iota',\tau}\right)$$

Combining the two partial results we recover the pair-wise covariance formula in Equation (12), which can be easily extended to the multivariate case from Equation (11). The rank of the matrix can be infered by observing that Equation (11) is the sum of N_k vectors centered around their means. □

A.4. Bootstrap Reconciled Probabilities

Let $(\Omega_{[i]}, \mathcal{F}_{[i]}, \hat{\mathbb{P}}(\cdot\,|\,\hat{\boldsymbol{\theta}}))$ be a probabilistic forecast space, with $\mathcal{F}_{[i]}$ a σ-algebra on $\mathbb{R}^{N_i}$. Let a hierarchical reconciliation transformation be denoted by $\mathbf{SP}(\cdot) : \Omega_{[i]} \mapsto \Omega_{[b]} \mapsto \Omega_{[i]}$. Consider $\hat{\mathbf{y}}^{s}_{[i],\tau}$, $s = 1, \ldots, S$ samples drawn from an unconstrained base probability $\hat{\mathbb{P}}(\cdot\,|\,\hat{\boldsymbol{\theta}})$, and the transformed samples $\tilde{\mathbf{y}}^{s}_{[i],\tau} = \mathbf{SP}\left(\hat{\mathbf{y}}^{s}_{[i],\tau}\right)$.

The probability distribution of the reconciled samples is given by:

$$\tilde{\mathbb{P}}\left(\tilde{\mathbf{y}}_{[i],\tau} \in \mathcal{H}|\,\tilde{\boldsymbol{\theta}}\right) = \hat{\mathbb{P}}\left(\hat{\mathbf{y}}_{[i],\tau} \in \mathbf{SP}^{-1}(\mathcal{H})|\,\hat{\boldsymbol{\theta}}\right) \tag{13}$$

with $\mathcal{H}$ be a coherent forecast measurable set, and $\mathbf{SP}^{-1}(\cdot)$ the reconciliation's inverse image.

Proof. This proof makes only minor modifications to the arguments presented in (Panagiotelis et al., 2023).

$$\begin{aligned}\tilde{\mathbb{P}}\left(\tilde{\mathbf{y}}_{[i],\tau} \in \mathcal{H}|\,\tilde{\boldsymbol{\theta}}\right) &= \lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S}\mathbb{1}\{\tilde{\mathbf{y}}_{[i],\tau} \in \mathcal{H}\} \\ &= \lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S}\mathbb{1}\{\hat{\mathbf{y}}_{[i],\tau} \in \mathbf{SP}^{-1}(\mathcal{H})\} \\ &= \hat{\mathbb{P}}\left(\hat{\mathbf{y}}_{[i],\tau} \in \mathbf{SP}^{-1}(\mathcal{H})|\,\hat{\boldsymbol{\theta}}\right)\end{aligned} \tag{14}$$

The first and final equalities follow from the weak law of large numbers, as by definition the indicator functions are independent, identically distributed with a finite mean and variance. The second equality follows from the definition of the inverse image $\mathbf{SP}^{-1}(\mathcal{H}) = \{\hat{y}_{[i],\tau}\,|\,\mathbf{SP}(\hat{y}_{[i],\tau}) \in \mathcal{H}\}$. □

A general analytic reconciled probability is derived in Appendix A.5. It is worth noting that the bootstrap reconciliation induces a tradeoff between reduced inference speed and the requirement for knowledge of the reconciled parameters $\tilde{\boldsymbol{\theta}}$.

A.5. Analytical Reconciled Probabilities

We use the bootstrap sample reconciliation technique (Panagiotelis et al., 2023) to ensure the probabilistic coherence of `HINT`. This technique can restore hierarchical aggregation constraints to base samples, regardless of their original distribution. It enhances `HINT`'s modularity by ensuring its probabilistic coherence on a wide range of base probabilities, including non-parametric ones, without requiring any modifications to the original algorithm. In this section we show how a reconciled probability can be recovered analytically through change of variables and marginalization.

Lemma. Consider the classic reconciliation approach where the entire hierarchy's forecasts are combined into reconciled bottom-level forecasts using a composition of linear transformations $\mathbf{SP}(\cdot) = \mathbf{S}_{\mathbf{[i][b]}}\mathbf{P}_{\mathbf{[b][i]}}(\cdot)$. The reconciled probability for the new bottom-level series is given by:

$$\tilde{\mathbb{P}}_{[b]}\left(\mathbf{b}\right) = |\mathbf{P}^*| \int \mathbb{P}_{[i]}\left(\mathbf{P}_\perp \mathbf{a} + \mathbf{P}^- \mathbf{b}\right) \delta \mathbf{a} \tag{15}$$

where $\hat{\mathbb{P}}\left(\cdot\right)$ is the unconstrained base forecast distribution, $\mathbf{P}_\perp \in \mathbb{R}^{N_i \times N_a}$, $\mathbf{P}^- \in \mathbb{R}^{N_i \times N_b}$ are the orthogonal complement of $\mathbf{P}_{[i][b]}$ and its Moore-Penrose inverse; matrix $\mathbf{P}^* = [\mathbf{P}_\perp \,|\, \mathbf{P}^-]$, and bottom level $\mathbf{b}$ and aggregate level $\mathbf{a}$ vectors are obtained through the following variable change:

$$\hat{\mathbf{y}}_{[i]} = \mathbf{P}^* \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \tag{16}$$

Proof. Using the change of multivariate change of variables theorem, and properties of the Jacobian of a linear mapping:

$$\tilde{\mathbb{P}}(\mathbf{a}, \mathbf{b}) = \left| \det \left[\frac{d\mathbf{P}^*(\mathbf{z})}{d\mathbf{z}} \Big|_{\mathbf{z}=(\mathbf{a},\mathbf{b})} \right] \right| \mathbb{P}\Big(\mathbf{P}^* \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \Big) = |\mathbf{P}^*| \mathbb{P}_{[i]}\left(\mathbf{P}_\perp \mathbf{a} + \mathbf{P}^- \mathbf{b}\right) \tag{17}$$

Marginalizing $\mathbf{a}$ we obtain the reconciled probability for the bottom level series. □

Theorem. Consider the classic reconciliation approach where the entire hierarchy's forecasts are combined into reconciled bottom-level forecasts using a composition of linear transformations $\mathbf{SP}(\cdot) = \mathbf{S}_{\mathbf{[i][b]}}\mathbf{P}_{\mathbf{[b][i]}}(\cdot)$. The reconciled probability for the entire hierarchical series is given by:

$$\tilde{\mathbb{P}}\left(\mathbf{y}_{[i]}\right) = |\mathbf{S}^*| \tilde{\mathbb{P}}_{[b]}\left(\mathbf{S}^- \mathbf{y}_{[i]}\right) \mathbb{1}\{\mathbf{y}_{[i]} \in \mathcal{H}\} \tag{18}$$

where $\tilde{\mathbb{P}}_{[b]}(\cdot)$ is the reconciled bottom forecast distribution, $\mathbb{1}(\tilde{\mathbf{y}}_{[i]} \in \mathcal{H})$ indicates if realization belongs in the N_b-dimensional hierarchically coherent subspace $\mathcal{H}$, $\mathbf{S}^- \in \mathbb{R}^{N_b \times N_i}$ is $\mathbf{S}_{[i][b]}$ Moore-Penrose inverse and $\mathbf{S}_\perp \in \mathbb{R}^{N_i \times N_a}$ its orthogonal complement.

Proof. This proof follows closely that provided in (Panagiotelis et al., 2023). Given bottom level forecast distribution from the Lemma, one can create a degenerate distribution for the entire hierarchy by adding additional dimensions $\mathbf{u} \in \mathbb{R}^{N_a}$.

$$\tilde{\mathbb{P}}_{[i]}(\mathbf{u}, \mathbf{b}) = \tilde{\mathbb{P}}_{[b]}\left(\mathbf{b}\right) \mathbb{1}\{\mathbf{u} = \mathbf{0}\} \tag{19}$$

Let $\mathbf{S} = \mathbf{S}_{[i][b]}$ and $\mathbf{S}_\perp^\intercal$ its orthogonal complement, and $\mathbf{S}^-$ and $\mathbf{S}_\perp^-$ the respective Moore-Penrose inverses. Using the following change of variables $\mathbf{b} = \mathbf{S}^- \mathbf{y}_{[i]}$ and $\mathbf{u} = \mathbf{S}_\perp^\intercal \mathbf{y}_{[i]}$ we obtain

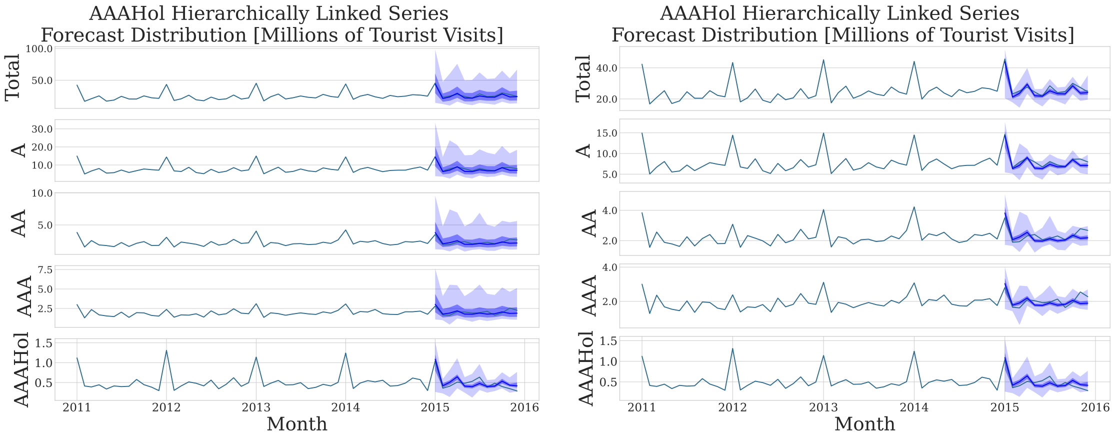
$$\mathbf{y}_{[i]} = \left[\mathbf{S}_\perp^- \,|\, \mathbf{S}\right] \begin{bmatrix} \mathbf{u} \\ \mathbf{b} \end{bmatrix} \qquad \Longleftrightarrow \qquad \mathbf{S}^* \mathbf{y}_{[i]} = \begin{bmatrix} \mathbf{S}_\perp^\intercal \\ \mathbf{S}^- \end{bmatrix} \mathbf{y}_{[i]} = \begin{bmatrix} \mathbf{u} \\ \mathbf{b} \end{bmatrix} \tag{20}$$

$$\tilde{\mathbb{P}}\left(\mathbf{y}_{[i]}\right) = |\mathbf{S}^*| \tilde{\mathbb{P}}_{[b]}\left(\mathbf{S}_\perp^\intercal \mathbf{0} + \mathbf{S}^- \mathbf{y}_{[i]}\right) \mathbb{1}\{\mathbf{S}_\perp^\intercal \mathbf{y}_{[i]} = \mathbf{0}\} = |\mathbf{S}^*| \tilde{\mathbb{P}}_{[b]}\left(\mathbf{S}^- \mathbf{y}_{[i]}\right) \mathbb{1}\{\mathbf{y}_{[i]} \in \mathcal{H}\} \tag{21}$$

By definition of the orthogonal complement if $\mathbf{S}_\perp^\intercal \mathbf{y}_{[i]} = \mathbf{0}$, that means that $\mathbf{y}_{[i]} \in \mathrm{span}(\mathbf{S})$, that matches the definition of the hierarchically coherent subspace $\mathcal{H}$.

□

As mentioned earlier, the analytical version of reconciled probabilities can provide highly efficient inference times, depending on the properties of the reconciliation and the base forecast distributions. For instance, recent studies have utilized Gaussian distributions (Panagiotelis et al., 2023; Wickramasuriya, 2023), and Poisson Mixtures (Olivares et al., 2022b).



(a) `HINT-BottomUp` Univariate

(b) `HINT-BottomUp` Multivariate

Figure 2. Estimation methods' comparison. The first row displays the total Australian tourist visits, followed by rows showing the North-South Wales state visits (A), visits in the metropolitan area of New South Wales (AA), visits to Sydney (AAA), and holiday visits to Sydney. Light and dark blue represent the forecast distributions and 99% and 75% prediction intervals. Modeling the series' correlations can play an important role in the reconciled forecast distributions sharpness.

A.6. `HINT` Parameter Estimation

Maximum Likelihood Estimation

To estimate the parameters of `HINT`, we can use maximum likelihood estimation for the multivariate probability as shown in Equation (6). Specifically, we denote as $\boldsymbol{\omega}$ the neural network parameters that condition the the probabilistic output layer parameters. Then, we express the negative log-likelihood function as follows:

$$\mathcal{L}(\boldsymbol{\omega}) = -\log\left[\sum_{\kappa=1}^{N_k}\hat{w}_\kappa(\boldsymbol{\omega}) \prod_{(\iota,\tau)\in[i][t+1:t+h]} \left(\frac{1}{\hat{\sigma}_{\iota,\kappa,\tau}(\boldsymbol{\omega})\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y_{\iota,\kappa,\tau}-\hat{\mu}_{\iota,\kappa,\tau}(\boldsymbol{\omega})}{\hat{\sigma}_{\iota,\kappa,\tau}(\boldsymbol{\omega})}\right)^2\right\}\right)\right] \tag{22}$$

Although standard maximum likelihood estimation can model relationships between multiple time series across the forecast horizon, a scalability challenge arises when the number of series and forecast horizon increase significantly. Since its computation requires access to the entire multivariate series, MLE can become computationally intensive and time-consuming. For this reason, MLE is only effective for hierarchical time series with a small number of series.

Maximum Composite Likelihood Estimation

Composite likelihood provides a computationally efficient alternative to maximum likelihood estimation for optimizing the parameters of `HINT`. Unlike MLE, which computes the whole multivariate likelihood, composite likelihood decomposes the hierarchical variable high-dimensional space support into sub-spaces and optimizes the weighted product of the subspaces' marginal likelihood. When defining the sub-spaces in composite likelihood, the probabilistic model is restricted to learning relationships within each sub-space while assuming independence across non-overlapping sub-spaces. These sub-spaces can be defined based on the user's application needs. For instance, they can be guided by the geographic proximity of the time series data. In order to simplify the `HINT` algorithm, we randomly assign each series to the sub-spaces defined by the stochastic gradient batches. Let $\mathcal{B} = \{[b_i]\}$ be time-series SGD batches, then `HINT`'s negative log composite likelihood is:

$$\mathcal{CL}(\boldsymbol{\omega}) = -\sum_{[b_i]\in\mathcal{B}}\log\left[\sum_{\kappa=1}^{N_k}\hat{w}_\kappa(\boldsymbol{\omega}) \prod_{(\iota,\tau)\in[b_i][t+1:t+h]} \left(\frac{1}{\hat{\sigma}_{\iota,\kappa,\tau}(\boldsymbol{\omega})\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y_{\iota\tau}-\hat{\mu}_{\iota,\kappa,\tau}(\boldsymbol{\omega})}{\hat{\sigma}_{\iota,\kappa,\tau}(\boldsymbol{\omega})}\right)^2\right\}\right)\right] \tag{23}$$

For the composite likelihood's independence assumptions, sub-spaces can be defined as each series forecast, leading to the univariate estimation approach. Figure 2 compares univariate and composite likelihood estimation.

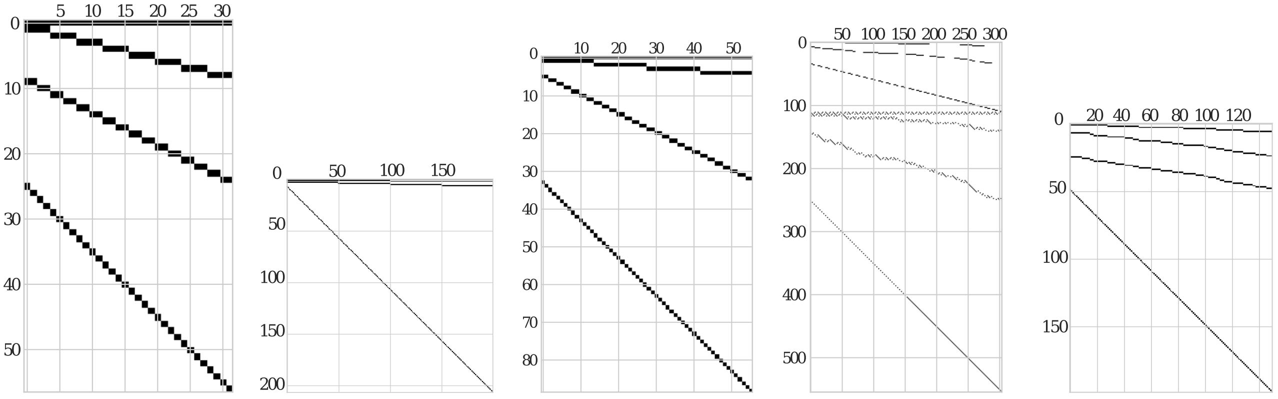
B. Hierarchical Dataset's Exploration

In this Appendix we complement the description of the benchmark datasets from Section 4.

Table A1. Summary, of experiment hierarchical datasets.

Dataset	Total	Aggregate	Bottom	Frequency.	h	Levels
Labour	57	25	32	Month	8	4
Traffic	207	7	200	Daily	7	3
Tourism	89	33	56	Quarterly	4	4
Tourism-L	555	175	76 / 304	Month	12	4/5
Wiki2	199	49	150	Daily	7	5

`Labour` reports monthly Australian employment from February 1978 to December 2020. It contains a structure built by the labour categories (Australian Bureau of Statistics, 2019). `Traffic` measures the occupancy of 963 traffic lanes in the Bay Area, the data is grouped into a year of daily observations and organized into a 207 hierarchical structure (Dua & Graff, 2017). `Tourism` consists of 89 Australian location quarterly visits series; it covers from 1998 to 2006. Several studies have used this dataset in the past (Tourism Australia, Canberra, 2005). `Tourism-L` summarizes an Australian visitor survey managed by the Tourism Research Australia, the dataset contains 555 monthly series from 1998 to 2016, and it is organized into geographic and purpose of travel (Tourism Australia, Canberra, 2019). `Wiki2` contains the daily views of 145,000 Wikipedia articles from July 2015 to December 2016. The dataset is filtered and processed into 150 bottom series and 49 aggregate series (Anava et al., 2018). Figure 4 shows each dataset's most aggregated series along with its training methodology partition. Figure 3 shows each dataset's hierarchical aggregation constraint matrices.



(a) Labour (b) Traffic (c) TourismS (d) TourismL (e) Wiki2

Figure 3. Dataset's hierarchical constraints. (a) `Labour` groups 32 occupation series by gender and geography. (b) `Traffic` groups 200 highways' occupancy series into quarters, halves and total. (c) `Tourism` groups 56 quarterly Australian tourist visits by geographic levels. (d) `Tourism-L` groups its 555 monthly Australian regional visit series, into a combination travel purpose, zones, states and country geographical aggregations. (e) `Wiki2` groups 150 daily visits to Wikipedia articles by language and article categorical taxonomy.

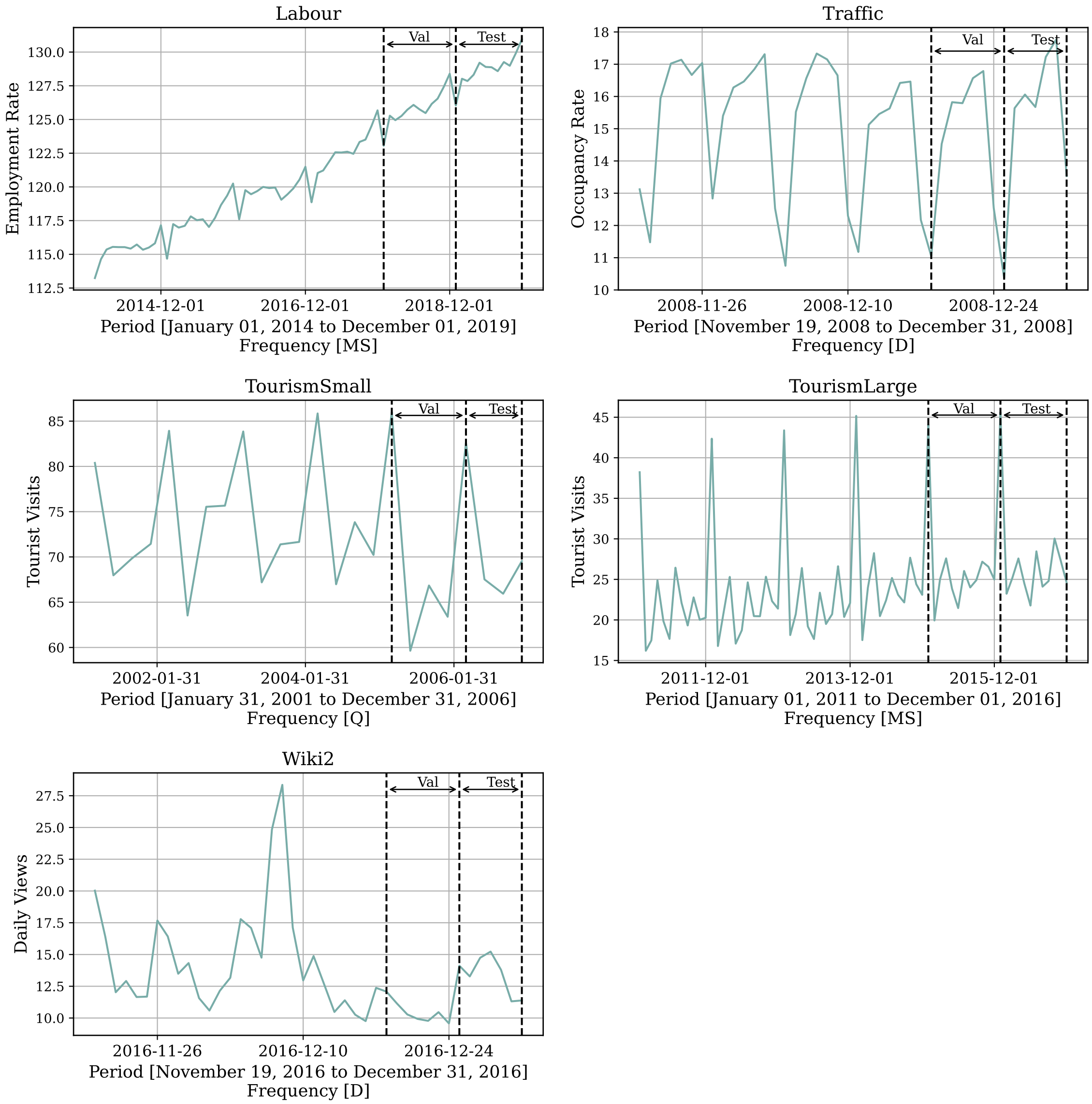


Figure 4. Datasets' partition into train, validation, and test sets used in our experiments. All use the last horizon window as defined in Table A1 (marked by the second dotted line), and the previous window preceding the test set as validation (between the first and second dotted lines). Validation provides the signal for hyperparameter optimization.

C. Ablation Studies

In this Appendix, we perform ablation studies on the validation set of five hierarchical datasets `Labour`, `Traffic`, `Tourism`, `Tourism-L`, and `Wiki2`. For these experiments, we change minimally the `HINT` settings defined in Table A4, removing the Temporal Normalization, varying the number of Mixture components, and exploring different hierarchical reconciliation strategies to understand their contribution to the performance of the method.

C.1. Scaled Decoupled Optimization

In Section 3.2, we introduced `HINT`'s scale decouple optimization strategy along with the Temporal Normalization transform. Here we study the effects of different temporal normalization strategies on the forecast accuracy performance of the model, measured with the overall sCRPS. For simplicity consider a network with only temporal input $\mathbf{x}^{h}_{[i][:t][c]}$, with $[i]$ batch, $[:t]$ time, and $[c]$ feature channel indexes, we consider normalization transformations that follow the general form:

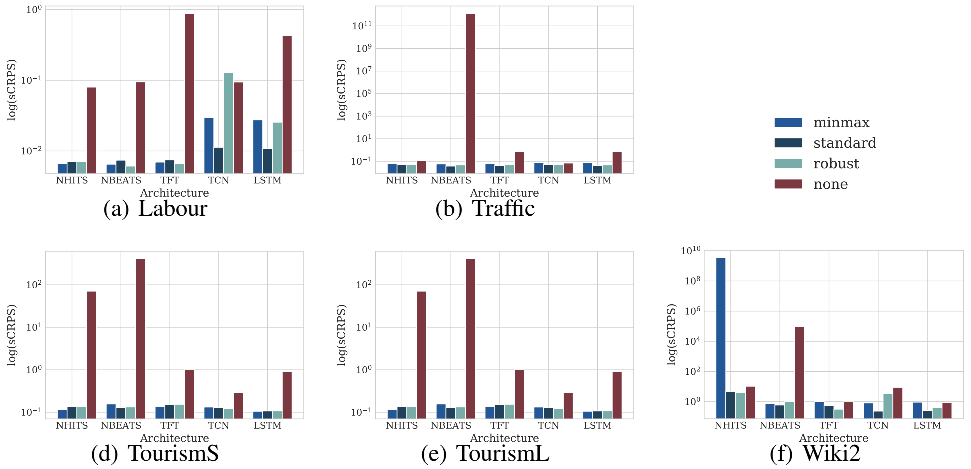
$$\begin{aligned}\check{\mathbf{x}}^{h}_{[i][:t+h][c]} &= \mathrm{TemporalNorm}(\mathbf{x}^{h}_{[i][:t+h][c]}) = \frac{\mathbf{x}^{h}_{[i][:t][c]} - \mathbf{a}}{\mathbf{b}} \\ \hat{\theta}(\check{\mathbf{x}}^{h}_{[i][:t+h][c]}) &= \mathrm{TemporalNorm}^{-1}(\boldsymbol{\omega}(\check{\mathbf{x}}_{[i][:t+h][c]})) = \mathbf{b}\boldsymbol{\omega}_{[i][t+h]} + \mathbf{a}\end{aligned} \tag{24}$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{(N_a+N_b)\times N_c}$ is the shift and the scale of the historic inputs. In this experiment we augment `NHITS`, `NBEATS`, `TFT`, `TCN`, and `LSTM` with `TemporalNorm` and measure its effects using four different types of normalization schemes:

$$\text{minmax}: \frac{(\mathbf{x}^{h}_{[i][:t][c]} - \min(\mathbf{x}^{h}_{[i][:t][c]})_{[i][c]})}{(\max(\mathbf{x}^{h}_{[i][:t][c]})_{[i][c]} - \min(\mathbf{x}^{h}_{[i][:t][c]})_{[i][c]})} \qquad \text{standard}: \frac{(\mathbf{x}^{h}_{[i][:t][c]} - \bar{\mathbf{X}}_{[i][c]})}{\hat{\sigma}_{[i][c]}}$$

$$\text{robust}: \frac{(\mathbf{x}^{h}_{[i][:t][c]} - \mathrm{median}(\mathbf{x}^{h}_{[i][:t][c]}))}{\mathrm{mad}(\mathbf{x}^{h}_{[i][:t][c]})_{[i][c]}} \qquad \text{None}: \mathrm{Identity}(\mathbf{x}^{h}_{[i][:t][c]})$$

Figure C.1 shows that incorporating scale-decoupled optimization into models significantly improves their accuracy across various datasets and architectures. In contrast, the performance improvements of different temporal normalization types are varied. In cases where signals are smooth, the type of normalization used makes little difference. However, the robust normalization strategy using median shift and mad scale is preferred when signals are noisy or volatile. We chose robust Temporal Normalization as the default for all subsequent experiments based on these observations.



(a) Labour (b) Traffic (d) TourismS (e) TourismL (f) Wiki2

Figure 5. Validation *scaled Continuous Ranked Probability Score* (sCRPS) curves on five different hierarchical datasets. We show the accuracy for `NHITS`,`NBEATS`, `TFT`, `TCN`, and `LSTM` with and without scaled-decoupled optimization. Scale-Decouple Optimization offers substantial accuracy improvements across architectures and datasets, with its *robust* variant being preferred with noisy datasets.

C.2. Mixture Size Exploration

As mentioned in Section 3.1 and proven in Appendix A, our multivariate mixture probability model is capable of capturing the relationships among the hierarchical series and its expressivity directly determined by the number of components in the Gaussian mixture. Here we conduct a study to compare the accuracy effects of different mixture sizes. We vary the number of components monotonically and follow the model's performance. An `NHITS` model configuration is automatically selected as defined from Table A4.

During our `Labour`, `Traffic`, and `Tourism-L` experiments, we observed a bias-variance trade-off relationship between the number of mixture components and the validation sCRPS. If we set the number of components to 1, the sCRPS score is the highest, but as the number of components increases, the sCRPS improves. However, if there are too many components, the CRPS score worsens again after a certain point. Our theory is that if the mixture has too few components, the probability does not have enough parameters to accurately depict the data's correlations; this leads to a high bias and reflects in a poor sCRPS. On the other hand, if the mixture has too many components, the model becomes too complicated and quickly overfits the training data, resulting in high variance and poor performance on the validation data. The results of this ablation experiment show a clear benefit of using a flexible multivariate mixture distribution, contrasting it to the simpler approach of using a single component (Gaussian/Poisson regression). We explain these improvements from the flexibility of the mixture approach that operates as a Kernel density estimation and can arbitrarily approximate a wide variety of target distributions.

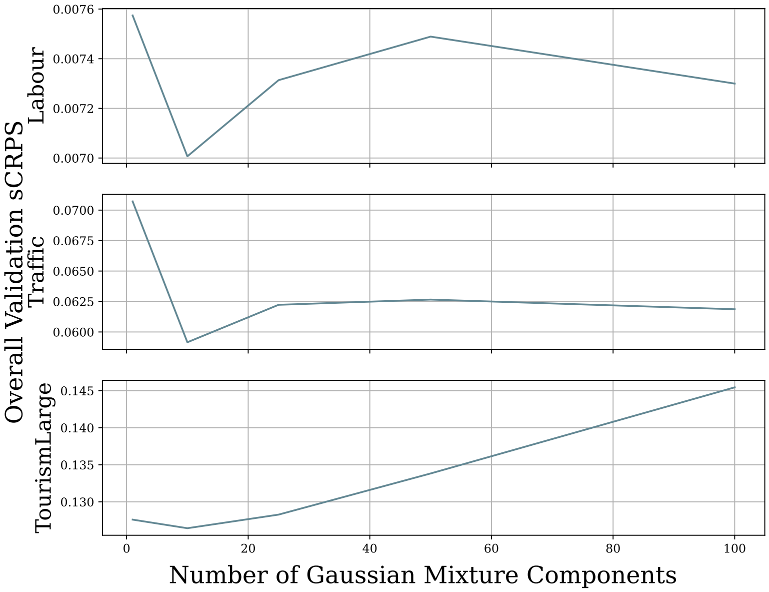


Figure 6. Validation *scaled Continuous Ranked Probability Score* (sCRPS) for `Labour`, `Traffic`, and `Tourism-L`. We show performance curves for `NHITS` as a function of the number of Mixture Components. We observe a bias-variance tradeoff, where initially, the sCRPS decreases as the number of components increases and reaches an optimal value at K=10 components. From there, after that, we see the sCRPS worsening, thus giving us the classic U-shaped tradeoff pattern.

C.3. Probabilistic Reconciliation Exploration

As mentioned in Section 3.1 and proven in Appendix A, the flexibility of the multivariate mixture probability model is compatible with most probabilistic reconciliation techniques. In this ablation study, we compare the accuracy effect of different reconciliation strategies. The reconciliation strategies considered are `BottomUp` (Orcutt et al., 1968; Dunn et al., 1976), `TopDown` (Gross & Sohl, 1990; Fliedner, 1999), and `MinTrace` (Hyndman et al., 2011; Wickramasuriya et al., 2019) variants. We describe them in detail below.

Consider the base forecasts $\hat{\mathbf{y}}_{[i],\tau} \in \mathbb{R}^{N_a+N_b}$, a reconciliation process uses a matrix $\mathbf{P}_{[b][i]} \in \mathbb{R}^{N_b \times (N_a+N_b)}$ that collapses the original base forecasts into bottom-level forecast that are later aggregated for the upper levels of the hierarchy into the reconciled forecasts $\tilde{\mathbf{y}}_{[i],\tau}$. Here we use the convenient representation of the reconciliation strategies introduced in Section 2.

$$\tilde{\mathbf{y}}_{[i],\tau} = \mathbf{S}_{[i][b]}\mathbf{P}_{[b][i]}\hat{\mathbf{y}}_{[i],\tau} = \mathbf{SP}\left(\hat{\mathbf{y}}_{[i],\tau}\right)$$

Bottom-Up. The most basic hierarchical reconciliation consists of simply aggregating the bottom-level base forecasts $\hat{\mathbf{y}}_{[b],\tau}$. By construction, it satisfies the hierarchical aggregation constraints. Here the reconciliation matrix is given by:

$$\mathbf{P}_{[b][i]} = [\mathbf{0}_{[\mathrm{b}][\mathrm{a}]} \mid \mathbf{I}_{[\mathrm{b}][\mathrm{b}]}] \tag{25}$$

Top-Down. The `TopDown` strategy distributes an aggregate level forecast into the bottom-level forecasts using proportions $\mathbf{p}_{[b]}$. Proportions can be historical values, or they can be forecasted. Its reconciliation matrix is given by:

$$\mathbf{P}_{[b][i]} = [\mathbf{p}_{[\mathrm{b}]} \mid \mathbf{0}_{[\mathrm{b}][\mathrm{a,b}\ -1]}] \tag{26}$$

MinTrace. Newer reconciliation strategies use all the information available throughout the hierarchy optimally. In particular, the `MinTrace` reconciliation is proven to be the optimizer of a mean squares error objective that transforms base predictions into hierarchically coherent predictions under an unbiasedness assumption. Its reconciliation matrix is given by:

$$\mathbf{P}_{[b][i]} = \left(\mathbf{S}^{\intercal}\hat{\mathbf{\Sigma}}_{\tau}\mathbf{S}\right)^{-1}\mathbf{S}^{\intercal}\hat{\mathbf{\Sigma}}_{\tau}^{-1} \tag{27}$$

We summarize the ablation study results for the different reconciliation strategies in Table A2. We report the overall sCRPS across five datasets for the `NHITS`, `TCN`, and `TFT` architectures. We obtain the probabilistic predictions using bootstrap (Panagiotelis et al., 2023). We observe clear advantages from adopting novel reconciliation techniques such as `MinTrace` as it improves accuracy over `BottomUp` by 20 to 30 percent margins across well-established neural forecast architectures. We find that post processing reconciliation is capable of improving complex end-to-end approaches that integrate the hierarchical constraints into the training procedure Rangapuram et al. 2021; Kamarthi et al. 2022; Han et al. 2021. Based on the results of these ablation studies, we conducted the main experiments of this work with the `MinTrace` and the `BottomUp` reconciliation techniques. To include `TopDown` as an alternative we need to robustify its implementation.

Table A2. Empirical evaluation of probabilistic coherent forecasts. Mean scaled continuous ranked probability score (sCRPS), averaged over 10 random seeds, at each aggregation level. The best result is highlighted (lower measurements are preferred).

[*] The `TopDown` reconciliation is only available for strictly hierarchical datasets.

	DATASET	Base	MinTrace-ols	MinTrace-wls	TopDown-ap*	TopDown-pa*	BottomUp
NHITS	Labour	0.0082	0.0082±0.0001	0.0084±0.0001	0.0092±0.0001	0.0091±0.0000	0.0094±0.0001
	Traffic	0.0629	0.0635±0.0011	0.0643±0.0010	0.0651±0.0010	0.0650±0.0013	0.0660±0.0008
	Tourism	0.0791	0.0806±0.0011	0.0771±0.0014	0.0920±0.0010	0.0913±0.0014	0.0756±0.0012
	Tourism-L	0.1274	0.1281±0.0004	0.1261±0.0006	-	-	0.1351±0.0005
	Wiki2	1.4531	1.3165±0.0302	1.8399±0.0904	0.5165±0.0159	0.5178±0.0088	3.3351±0.1690
TCN	Labour	0.0213	0.0243±0.0004	0.0202±0.0003	0.0237±0.0003	0.0237±0.0004	0.0187±0.0004
	Traffic	0.0566	0.0569±0.0007	0.0577±0.0007	0.0605±0.0010	0.0605±0.0009	0.0623±0.0008
	Tourism	0.0664	0.0660±0.0009	0.0641±0.0011	0.0803±0.0012	0.0807±0.0015	0.0678±0.0018
	Tourism-L	0.1632	0.1638±0.0010	0.1640±0.0008	-	-	0.1677±0.0007
	Wiki2	2.7345	2.0942±0.0539	2.9305±0.1247	1.6188±0.0374	1.6232±0.0476	4.2341±0.1338
TFT	Labour	0.0073	0.0071±0.0001	0.0074±0.0000	0.0084±0.0001	0.0084±0.0001	0.0087±0.0001
	Traffic	0.0632	0.0641±0.0007	0.0638±0.0008	0.0650±0.0015	0.0646±0.0011	0.0658±0.0009
	Tourism	0.0944	0.1015±0.0010	0.0895±0.0007	0.0834±0.0008	0.0832±0.0006	0.0922±0.0013
	Tourism-L	0.1360	0.1364±0.0007	0.1362±0.0008	-	-	0.1442±0.0010
	Wiki2	0.2609	0.2608+0.0010	0.2641±0.0025	0.2560±0.0016	0.2560±0.0013	0.2755±0.0041

D. Software and Training Methodology

D.1. Hyperparameters and Training Methodology

Table A3. HINT fixed hyperparameters.

HYPERPARAMETER	FIXED VALUES		
Architecture	NHITS	TFT	TCN
Activation	ReLU	ReLU	ReLU
Encoder units	256	256	256
Encoder layers*	4	3	4
Encoder type	MLP	LSTM	Conv1D
Train Objective	Comp.Lik.	Comp.Lik.	Comp.Lik.

Table A4. HINT optimized hyperparameters.

HYPERPARAMETER	CONSIDERED VALUES
Initial learning rate.	{1e-3,5e-4,1e-4}
Number of learning rate decays.	{None, 3}
Training steps.	{.5e3, 1e3, 1.5e3, 2e3, 2.5e3, 3e3}
Input size multiplier (L=m*H).	$m \in \{2, 3, 4\}$
Reconciliation strategy.	{BottomUp, MinTraceOLS, MinTraceWLS }

Training HINT and the benchmark models involves dividing the data into training, validation, early stopping, and test sets, as shown in Figure 4. The training set consists of the observations before the last two horizon windows; validation is the window between the train and test sets, with test being the last window. The model's performance on the validation set guides the exploration of the hyperparameter space (HYPEROPT, Bergstra et al. 2011). During the recalibration phase, we retrain the models to incorporate new information before being tested.

We followed a standard two-stage approach for hyperparameter selection. In the first stage, based on validation ablation studies from Appendix C, we fixed the architecture and the probability distribution to be estimated; Table A3 describes the hyperparameters. Then, in the second stage, we optimized the training procedure of the architecture, optimally exploring the space defined in Table A4 with HYPEROPT. This approach allowed us to explore the hyperparameter space while keeping it computationally tractable. It also demonstrated the HINT's robustness, broad applicability, and potential to achieve high accuracy with only slight adjustments.

We train HINT to maximize the composite likelihood from Equation (23) using the ADAM (Kingma & Ba, 2014) stochastic gradient algorithm. An early stopping strategy (Yao et al., 2007) is employed to halt training if there is no improvement in loss on the validation set.

D.2. Software Implementation

All statistical baselines use StatsForecast's AutoARIMA (Hyndman & Khandakar, 2008; Garza et al., 2022) and HierarchicalForecast's reconciliation methods implementations (Olivares et al., 2022b). We created a unified Python re implementation of various widely-used hierarchical forecasting techniques that we make publicly available in the HierarchicalForecast library (Olivares et al., 2022b). The shared implementation allows us to standardize the comparison of the methods, controlling for experimental details and guaranteeing the quality of statistical baselines. The code is publicly available in a dedicated repository to support reproducibility and related research.

Regarding the hierarchical neural forecast baselines, HierE2E (Rangapuram et al., 2021) is available in the GluonTS library, while PROFHIT (Kamarthi et al., 2022) is available in a PROFHIT dedicated repository. As mentioned earlier the only available implementation for PROFHIT suffers from significant numerical instability in its optimization. We use the optimal configurations reported in HierE2E and PROFHIT repositories.

The HINT model family is implemented in PyTorch (Paszke et al., 2019) and can be run on both CPUs and GPUs. We have made the HINT source code available, along with all the experiments in the following HINT dedicated repository.

E. Extended Main Results

Table 1 reports the overall probabilistic and point forecast accuracy complying with the page restrictions. In this section, we present accuracy measurements for different hierarchy levels of aggregation. The results follow the same training and hyperparameter selection methodologies described in Appendix D. The top row of each panel reports the overall sCRPS or relMSE. One can observe that the forecast errors increase when moving from aggregate levels toward disaggregate levels.

HINT improves on the second-best alternative overall sCRPS by an average of 8.14% across datasets. With specific improvements of 8.2% on Labour, 17.4% on Tourism, 10.9% on Tourism-L, and 15.5% on Wiki2. Regarding Traffic we observed HierE2E outperforming HINT by -11.3% due to the clear Granger causalities in the dataset that merit the use of HierE2E's VAR approach. The results for the relative MSE are highly correlated.

Table A5. Empirical evaluation of probabilistic coherent forecasts. Mean scaled continuous ranked probability score (sCRPS) and mean relative squared error (relMSE), averaged over 10 random seeds, at each aggregation level. The best result is highlighted.

† The PROFHIT results differ from (Kamarthi et al., 2022), as the only available implementation suffers from significant numerical instability in its optimization.

* Best performing variant of TopDown (avg. proportions, proportions avg.), and MinTrace (ols, wls, shrinkage) reported. ** The PERMBU/TopDown only available for strictly hierarchical datasets.

			HINT-GMM (Ours)			OTHER		BOOTSTRAP			PERMBU**		
	Dataset	Level	NHITS	TFT	TCN	HierE2E	PROFHIT†	BottomUp	TopDown*	MinTrace*	BottomUp	TopDown*	MinTrace*
sCRPS	Labour	Overall	**.0067±.0000**	.0089±.0000	.0120±.0001	.0171±.0003	.2138±.0007	.0078±.0001	.0668±.0000	.0073±.0000	.0077±.0001	.0623±.0001	.0069±.0001
		Country	**.0017±.0000**	.0038±.0001	.0071±.0001	.0052±.0003	.2097±.0038	.0021±.0001	.0012±.0001	.0013±.0001	.0026±.0001	.0014±.0001	.0016±.0001
		Region	**.0047±.0001**	.0067±.0000	.0103±.0001	.0181±.0003	.2150±.0028	.0058±.0001	.0458±.0000	.0045±.0001	.0060±.0001	.0420±.0001	.0045±.0001
		Region/Gender	**.0075±.0000**	.0097±.0000	.0128±.0001	.0188±.0003	.2161±.0012	.0088±.0001	.0798±.0001	.0087±.0001	.0084±.0001	.0739±.0001	.0078±.0001
		Region/Gender/Empl.	**.0128±.0000**	.0152±.0001	.0179±.0000	.0262±.0004	.2142±.0027	.0145±.0001	.1403±.0000	.0148±.0001	.0137±.0001	.1320±.0002	.0137±.0001
	Traffic	Overall	.0589±.0004	.0602±.0004	.0600±.0002	**.0426±.0008**	.1137±.0022	.0736±.0024	.0741±.0012	.0608±.0014	.0849±.0009	.0708±.0008	.0651±.0008
		Level1	.0340±.0008	.0357±.0006	.0353±.0003	**.0276±.0011**	.0899±.0059	.0468±.0031	.0301±.0020	.0299±.0020	.0651±.0012	.0373±.0011	.0367±.0011
		Level2	.0347±.0006	.0357±.0005	.0359±.0002	**.0287±.0009**	.0879±.0034	.0483±.0030	.0329±.0017	.0323±.0017	.0622±.0013	.0367±.0009	.0357±.0008
		Level3	.0392±.0005	.0403±.0007	.0391±.0004	**.0297±.0009**	.0926±.0032	.0530±.0025	.0360±.0013	.0385±.0014	.0614±.0010	.0383±.0009	.0405±.0010
		Level4	.1275±.0002	.1290±.0003	.1295±.0003	**.0845±.0003**	.1842±.0014	.1463±.0017	.1975±.0017	.1424±.0015	.1507±.0004	.1709±.0010	.1473±.0004
	Tourism	Overall	.0666±.0007	.0665±.0004	**.0536±.0004**	.0761±.0007	.1358±.0033	.0682±.0018	.1040±.0014	.0703±.0017	.0649±.0016	.0898±.0012	.0680±.0016
		Country	.0233±.0009	.0157±.0004	**.0147±.0004**	.0400±.0009	.0941±.0151	.0290±.0028	.0333±.0025	.0335±.0026	.0267±.0023	.0329±.0021	.0333±.0025
		Purpose	.0513±.0007	.0468±.0006	**.0360±.0004**	.0609±.0012	.1300±.0069	.0490±.0027	.0782±.0017	.0507±.0023	.0450±.0017	.0697±.0021	.0497±.0018
		State/Purpose	.0851±.0007	.0891±.0007	**.0709±.0006**	.0914±.0008	.1323±.0076	.0828±.0016	.1399±.0010	.0845±.0016	.0793±.0014	.1176±.0013	.0806±.0014
		Region/Purpose	.1068±.0008	.1143±.0006	**.0929±.0006**	.1122±.0007	.1867±.0031	.1118±.0012	.1646±.0010	.1124±.0013	.1087±.0017	.1390±.0014	.1085±.0016
	Tourism-L	Overall	**.1176±.0002**	.1354±.0005	.1550±.0006	.1424±.0019	.2139±.0014	.1375±.0013	-	.1313±.0009	-	-	-
		Country	**.0325±.0006**	.0493±.0008	.0714±.0011	.0698±.0029	.1353±.0090	.0622±.0026	-	.0471±.0018	-	-	-
		State	**.0606±.0006**	.0704±.0006	.0955±.0005	.0936±.0019	.1610±.0020	.0820±.0019	-	.0723±.0011	-	-	-
		Zone	**.1025±.0004**	.1164±.0007	.1352±.0006	.1260±.0017	.1893±.0034	.1207±.0010	-	.1143±.0007	-	-	-
		Region	**.1457±.0003**	.1597±.0003	.1797±.0007	.1653±.0016	.2277±.0022	.1646±.0007	-	.1591±.0006	-	-	-
		Purpose	**.0706±.0006**	.0868±.0008	.1122±.0009	.0996±.0028	.1845±.0071	.0788±.0018	-	.0723±.0014	-	-	-
		State/Purpose	**.1088±.0003**	.1252±.0007	.1508±.0007	.1317±.0021	.2160±.0031	.1268±.0017	-	.1243±.0014	-	-	-
		Zone/Purpose	**.1772±.0003**	.1989±.0005	.2147±.0006	.1926±.0015	.2679±.0019	.1949±.0010	-	.1919±.0008	-	-	-
		Region/Purpose	**.2426±.0005**	.2766±.0005	.2804±.0005	.2606±.0017	.3296±.0010	.2698±.0008	-	.2694±.0006	-	-	-
	Wiki2	Overall	.3625±.0045	**.2447±.0007**	.2918±.0015	.2592±.0031	.4009±.0028	.2894±.0038	.3231±.0037	.2808±.0035	.3920±.0044	.4269±.0036	.3821±.0049
		World	.3715±.0118	**.1247±.0016**	.1209±.0028	.1007±.0046	.1244±.0085	.1796±.0069	.1777±.0084	.1793±.0067	.1777±.0125	.1945±.0109	.1801±.0123
		Country	.3512±.0059	**.1805±.0011**	.1935±.0021	.1963±.0037	.2775±.0141	.2392±.0047	.2437±.0058	.2232±.0043	.2778±.0073	.3036±.0029	.2684±.0066
		Access	.3461±.0020	**.2546±.0010**	.3124±.0022	.2784±.0038	.4405±.0034	.2966±.0032	.3379±.0026	.2781±.0028	.4196±.0059	.4621±.0071	.4006±.0059
		Agent	.3529±.0023	**.2699±.0010**	.3345±.0018	.2900±.0043	.4526±.0084	.3036±.0033	.3427±.0026	.2855±.0029	.4255±.0058	.4669±.0070	.4073±.0060
		Topic	.3905±.0021	**.3938±.0016**	.4975±.0016	.4307±.0039	.7094±.0109	.4282±.0038	.5134±.0037	.4379±.0029	.6595±.0060	.7074±.0049	.6540±.0058
relMSE	Labour	Overall	.5802±.0131	1.4644±.0148	2.8013±.0637	.5667±.0265	6.774×10^3	.5382±.0000	16.8204±.0000	**.3547±.0000**			
		Country	.1032±.0130	.8613±.0171	2.6597±.0724	.1536±.0284	9.424×10^2	.2362±.0000	.0542±.0000	**.0729±.0000**			
		Region	.6502±.0276	1.6251±.0231	2.8514±.1074	1.1486±.0413	6.635×10^2	.8281±.0000	14.6118±.0000	**.3740±.0000**			
		Region/Gender	1.6870±.0241	3.0355±.0271	3.6567±.0730	1.1206±.0395	4.113×10^2	.9021±.0000	35.6038±.0000	**.7519±.0000**			
		Region/Gender/Empl.	1.8103±.0131	2.7969±.0270	2.7274±.0389	1.4491±.0361	1.664×10^2	.8069±.0000	5.6047±.0000	**.8041±.0000**			
	Traffic	Overall	.1212±.0051	.1291±.0036	.1226±.0024	**.0340±.0051**	.4536±.0224	.1394±.0000	.0614±.0000	.0744±.0000			
		Level1	.1004±.0057	.1073±.0036	.1047±.0023	**.0253±.0057**	.4591±.0413	.1296±.0000	.0491±.0000	.0634±.0000			
		Level2	.1156±.0051	.1230±.0034	.1195±.0027	**.0302±.0050**	.4291±.0336	.1342±.0000	.0625±.0000	.0690±.0000			
		Level3	.1602±.0041	.1736±.0079	.1510±.0028	**.0529±.0038**	.4612±.0240	.1582±.0000	.0730±.0000	.0958±.0000			
		Level4	.8893±.0042	.8879±.0033	.8032±.0047	**.4206±.0048**	.7709±.0081	.6457±.0000	.6525±.0000	.6194±.0000			
	Tourism	Overall	.0898±.0031	.0932±.0018	**.0387±.0007**	.1471±.0046	.9745±.0803	.1002±.0000	.1919±.0000	.1235±.0000			
		Country	.0577±.0036	.0257±.0015	**.0200±.0013**	.1821±.0094	1.2240±.1474	.0841±.0000	.1328±.0000	.1233±.0000			
		Purpose	.0945±.0027	.1197±.0026	**.0342±.0009**	.1038±.0040	.8208±.0487	.0778±.0000	.1669±.0000	.0957±.0000			
		State/Purpose	.1409±.0049	.1725±.0020	**.0750±.0020**	.1550±.0032	.8511±.0489	.1563±.0000	.3482±.0000	.1620±.0000			
		Region/Purpose	.1525±.0051	.2022±.0030	**.0928±.0015**	.1772±.0027	.7227±.0942	.2000±.0000	.3628±.0000	.2007±.0000			
	Tourism-L	Overall	**.0577±.0009**	.0834±.0017	.1816±.0031	.2449±.0096	1.0401±.0296	.3070±.0000	-	.1375±.0000			
		Country	**.0336±.0009**	.0477±.0022	.1545±.0032	.2918±.0194	1.3473±.1430	.4399±.0000	-	.1268±.0000			
		State	**.0598±.0012**	.0716±.0017	.1767±.0031	.2850±.0107	1.0854±.0358	.3504±.0000	-	.1564±.0000			
		Zone	**.1263±.0019**	.1656±.0018	.2659±.0051	.3620±.0087	1.0821±.0397	.3950±.0000	-	.2664±.0000			
		Region	**.1777±.0024**	.2214±.0016	.3213±.0062	.3594±.0055	.9447±.0300	.3996±.0000	-	.3211±.0000			
		Purpose	**.0478±.0006**	.0872±.0024	.1721±.0029	.1581±.0086	.8257±.0758	.1624±.0000	-	.0759±.0000			
		State/Purpose	**.0752±.0009**	.1131±.0020	.2054±.0035	.1884±.0055	.7839±.0491	.1860±.0000	-	.1332±.0000			
		Zone/Purpose	**.1638±.0017**	.2025±.0013	.2959±.0048	.2872±.0049	.8178±.0187	.2932±.0000	-	.2550±.0000			
		Region/Purpose	**.2296±.0016**	.2772±.0014	.3628±.0052	.3360±.0047	.8013±.0215	.3661±.0000	-	.3464±.0000			
	Wiki2	Overall	1.0445±.0531	**.1884±.0012**	.2183±.0036	.6598±.0249	.7901±.0384	1.0163±.0000	1.4482±.0000	1.0068±.0000			
		World	3.2761±.1790	**.1955±.0037**	.2220±.0094	.2738±.0301	.3274±.0798	.9245±.0000	1.6135±.0000	.9883±.0000			
		Country	.7750±.0356	**.1648±.0015**	.1841±.0048	.7427±.0430	.9150±.0886	1.0204±.0000	1.3529±.0000	1.0252±.0000			
		Access	.4384±.0226	**.1921±.0012**	.2365±.0032	.9575±.0331	1.1538±.0429	1.1267±.0000	1.4159±.0000	1.0267±.0000			
		Agent	.4308±.0222	**.1972±.0010**	.2367±.0021	.9384±.0381	1.1312±.0695	1.1008±.0000	1.3562±.0000	1.0047±.0000			
		Topic	.2640±.0191	**.1923±.0010**	.2122±.0015	1.0305±.0145	1.1881±.0492	1.0603±.0000	1.2282±.0000	1.0182±.0000			