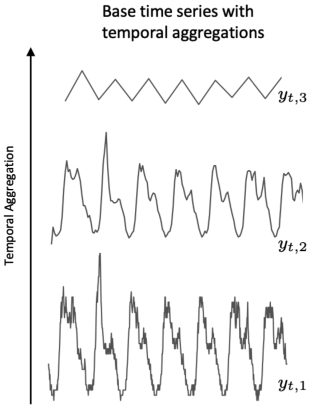
Coherent Probabilistic Forecasting for Temporal Hierarchies

Syama Rangapuram, Shubham Kapoor, Rajbir Singh Nirwan, Pedro Mercado Lopez, Bernie Wang, Tim Januschowski, Michael Bohlke-Schneider

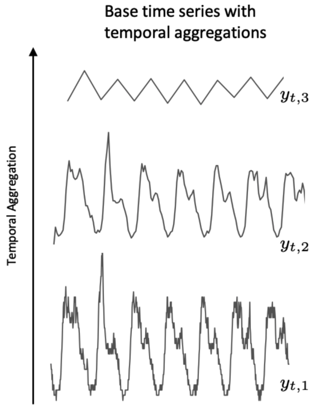
Problem

Given a univariate time series sampled at the observed frequency, generate *coherent*, *probabilistic* forecasts at aggregated time granularities.



Problem

Given a univariate time series sampled at the observed frequency, generate *coherent, probabilistic* forecasts at aggregated time granularities.

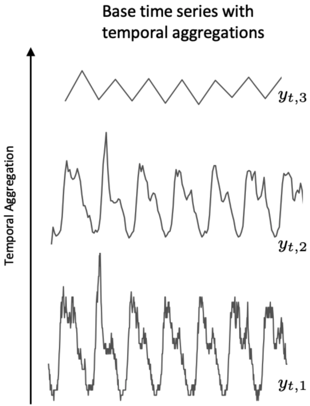


Motivation

- A requirement in practice!
- Improving forecasts of time series sampled at finer frequencies (e.g., 15-minutes or even 1-minute).

Problem

Given a univariate time series sampled at the observed frequency, generate *coherent, probabilistic* forecasts at aggregated time granularities.



Motivation

- A requirement in practice!
- Improving forecasts of time series sampled at finer frequencies (e.g., 15-minutes or even 1-minute).

Challenges

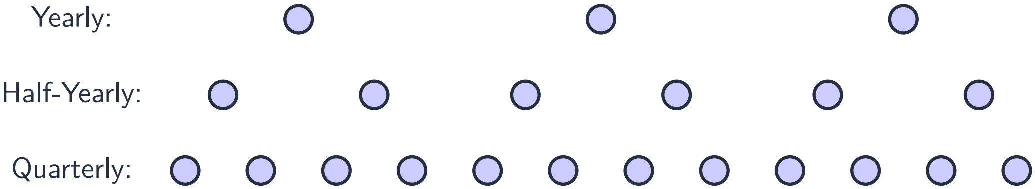
- Produce probabilistic forecasts that are consistent at all granularity levels.
- Handling of non-Gaussian data especially at the finer sampling frequencies.

Outline

- Background on Temporal Hierarchies
- State-of-the-art
- Our Model
- Experiments

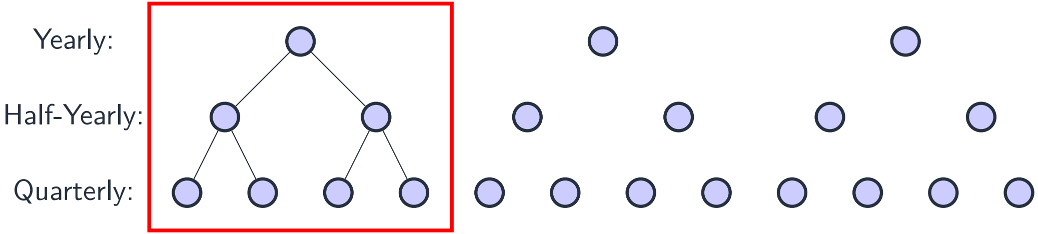
Background: Temporal Hierarchies

Example: Quarterly time series aggregated to half-yearly and yearly time series.



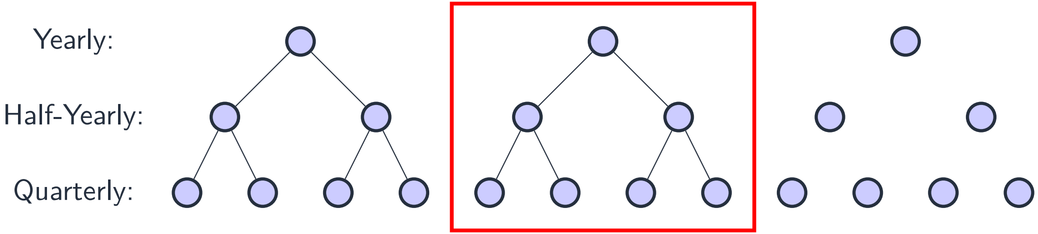
Background: Temporal Hierarchies

Example: Quarterly time series aggregated to half-yearly and yearly time series.

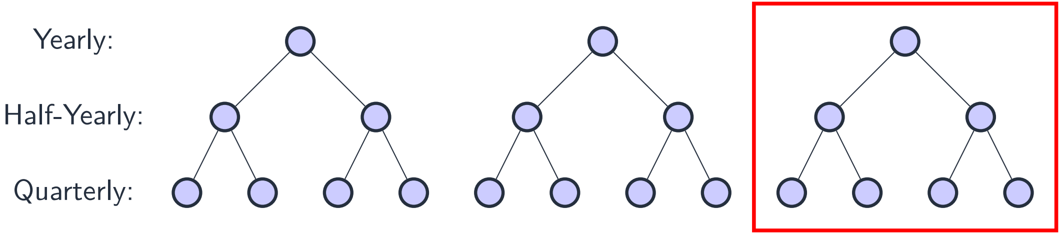


Background: Temporal Hierarchies

Example: Quarterly time series aggregated to half-yearly and yearly time series.



Example: Quarterly time series aggregated to half-yearly and yearly time series.

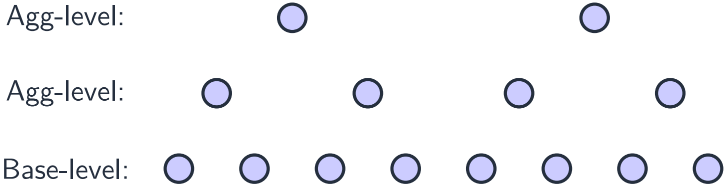


Background: Temporal Hierarchies (Contd.)

Forecasting with Temporal Hierarchies

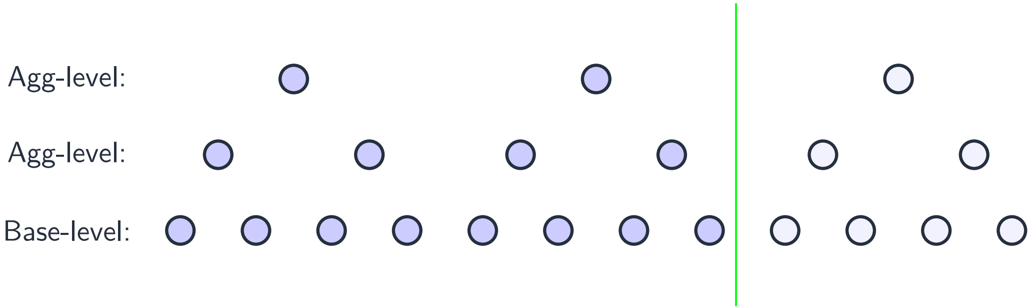
Background: Temporal Hierarchies (Contd.)

Forecasting with Temporal Hierarchies



Forecasting with Temporal Hierarchies

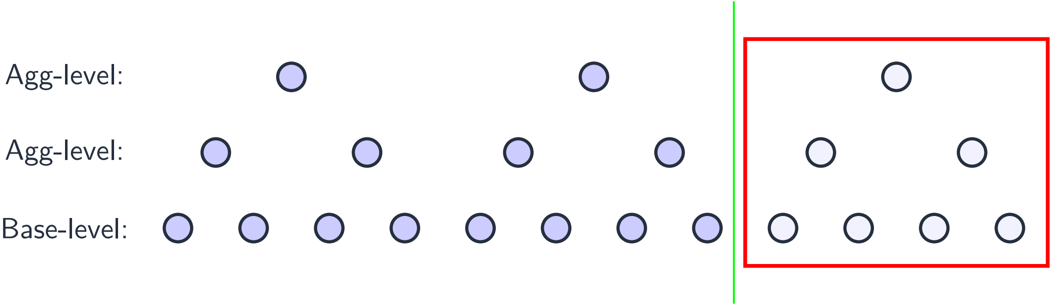
- Forecast horizons are different



Background: Temporal Hierarchies (Contd.)

Forecasting with Temporal Hierarchies

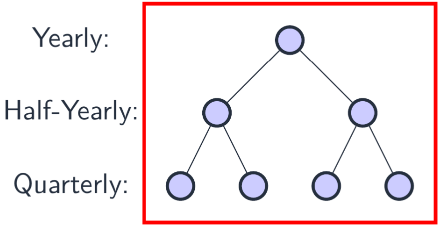
- Forecast horizons are different
- Guarantee forecasts “add-up”!



Aggregation Matrix

Temporal hierarchy satisfies linear aggregation constraints:

$$\mathbf{y} = S\mathbf{b}$$



Aggregation Matrix

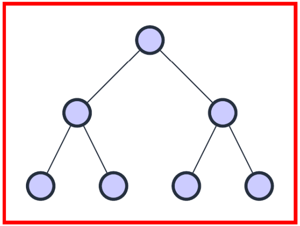
Temporal hierarchy satisfies linear aggregation constraints:

$$\mathbf{y} = S\mathbf{b}$$

Yearly:

Half-Yearly:

Quarterly:



$$\begin{bmatrix} y_T \\ y_{hy_1} \\ y_{hy_2} \\ y_{q_1} \\ y_{q_2} \\ y_{q_3} \\ y_{q_4} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{q_1} \\ y_{q_2} \\ y_{q_3} \\ y_{q_4} \end{bmatrix}$$

Aggregation Matrix

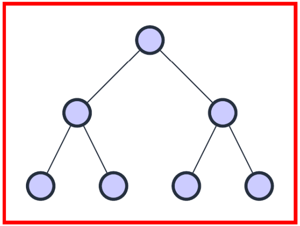
Temporal hierarchy satisfies linear aggregation constraints:

$$\mathbf{y} = S\mathbf{b}$$

Yearly:

Half-Yearly:

Quarterly:



$$\begin{bmatrix} y_T \\ y_{hy_1} \\ y_{hy_2} \\ y_{q_1} \\ y_{q_2} \\ y_{q_3} \\ y_{q_4} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{q_1} \\ y_{q_2} \\ y_{q_3} \\ y_{q_4} \end{bmatrix}$$

Equivalent representation:

$$A\mathbf{y} = \mathbf{0}, \tag{1}$$

where $A := [I_r \mid -S_{\text{sum}}] \in \{0,1\}^{r \times n}$, $\mathbf{0}$ is an r-vector of zeros, and I_r is the $r \times r$ identity.

Related Work

Temporal HIErarchical Forecasting (THIEF) (Athanasopoulos et al., 2017)

- Learn a univariate model for each time granularity *independently*
- Reconcile the mean forecasts $\hat{\mathbf{y}}$ in a post-processing step

Temporal HIErarchical Forecasting (THIEF) (Athanasopoulos et al., 2017)

- Learn a univariate model for each time granularity *independently*
- Reconcile the mean forecasts $\hat{\mathbf{y}}$ in a post-processing step

$$\text{Projection:} \quad \tilde{\mathbf{y}}_t = \arg\min_{\mathbf{y} \in \mathbb{R}^n} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|_D^2$$

$$\text{sb. to:} \quad A\mathbf{y} = 0.$$

Temporal HIErarchical Forecasting (THIEF) (Athanasopoulos et al., 2017)

- Learn a univariate model for each time granularity *independently*
- Reconcile the mean forecasts $\hat{\mathbf{y}}$ in a post-processing step

$$\textcolor{blue}{\text{Projection:}} \quad \tilde{\mathbf{y}}_t = \arg\min_{\mathbf{y} \in \mathbb{R}^n} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|_D^2$$
$$\text{sb. to:} \quad A\mathbf{y} = 0.$$

- D is a symmetric positive definite matrix
  - Identity matrix: `OLS`
  - Diagonal matrix: `Struc`, `MSE`
  - Block diagonal (inverse covariance) matrix: `shr`, `sam`

Temporal HIErarchical Forecasting (THIEF) (Athanasopoulos et al., 2017)

- Learn a univariate model for each time granularity *independently*
- Reconcile the mean forecasts $\hat{\mathbf{y}}$ in a post-processing step

$$\text{Projection:} \quad \tilde{\mathbf{y}}_t = \underset{\mathbf{y} \in \mathbb{R}^n}{\arg\min} \quad \|\mathbf{y} - \hat{\mathbf{y}}\|_D^2$$

$$\text{sb. to:} \quad A\mathbf{y} = 0.$$

- D is a symmetric positive definite matrix
  - Identity matrix: `OLS`
  - Diagonal matrix: `Struc`, `MSE`
  - Block diagonal (inverse covariance) matrix: `shr`, `sam`

Deep Thief (Theodosiou & Kourentzes, 2021): No coherence!

Our Method: An End-to-End Model

Motivation:

- Can we utilize/exploit patterns seen at the aggregate level for fitting models at disaggregated levels?
- Can we fit models directly on the reconciled forecasts instead of adjusting the base forecasts after-the-fact (after fitting independent forecasting models)?

Our Method: An End-to-End Model

Motivation:

- Can we utilize/exploit patterns seen at the aggregate level for fitting models at disaggregated levels?
- Can we fit models directly on the reconciled forecasts instead of adjusting the base forecasts after-the-fact (after fitting independent forecasting models)?

Challenges:

- How to include the reconciliation step as a part of the overall model and how to optimize it in automatic fashion?
- How to produce coherent probabilistic forecasts?

Recap: DeepAR

- AR(p):

$$P(z_t | z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid \sum_{k=1}^{p} \theta_k z_{t-k}, \sigma^2\right)$$

Recap: DeepAR

- AR(p):

$$P(z_t|z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid \sum_{k=1}^{p} \theta_k z_{t-k}, \sigma^2\right)$$

- Non-linear AR:

$$P(z_t|z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid f(z_{t-p:t-1}; \boldsymbol{\theta}), \sigma^2\right)$$

Recap: DeepAR

- AR(p):

$$P(z_t|z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid \sum_{k=1}^{p} \theta_k z_{t-k}, \sigma^2\right)$$

- Non-linear AR:

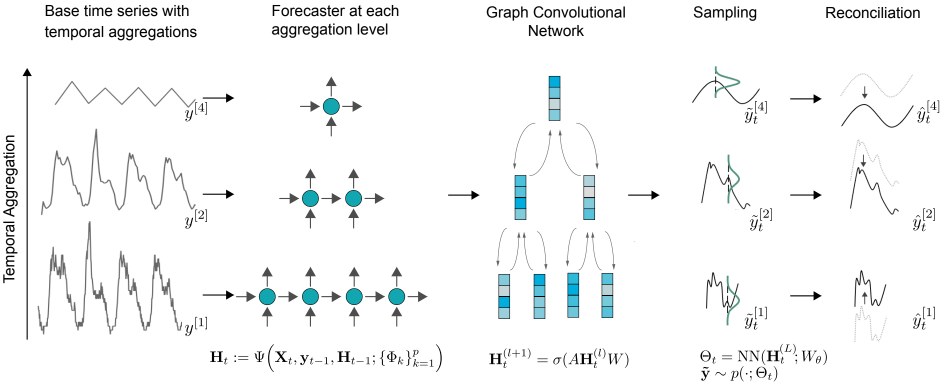
$$P(z_t|z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid f(z_{t-p:t-1};\boldsymbol{\theta}), \sigma^2\right)$$

- Recurrent non-linear AR (DeepAR):

$$P(z_t|z_{t-p:t-1}) = \mathcal{N}\left(z_t \mid \phi(\mathbf{h}_t;\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})\right)$$
$$\mathbf{h}_t = \psi(\mathbf{h}_{t-1}, z_{t-1};\boldsymbol{\theta})$$

End-to-End Model for Temporal Hierarchies

Extension of our previous work (Rangapuram et al. 2021) on hierarchical forecasting:



Model Details

GNN Layer:

- Pass the message at each node to its children
- Pass the messages from each node to its parent
- Keep the original message

Model Details

GNN Layer:

- Pass the message at each node to its children
- Pass the messages from each node to its parent
- Keep the original message

Messages exchanged for L rounds

$$f(\mathbf{H}_t, A) = \sigma(A\mathbf{H}_t W).$$

$$A := (A_{\text{acc.}} + A_{\text{dist.}} + A_{\text{ret.}})/3,$$

$$A_{\text{acc.}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_{\text{distr.}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Model Details (Contd.)

Reconciliation:

- Euclidean projection of each sample onto the coherent space (`OLS`)
- Enforcing non-negativity: alternating projection onto coherent space and the non-negative orthant (Dykstra's method)
- Output of the model are coherent samples

Model Details (Contd.)

Reconciliation:

- Euclidean projection of each sample onto the coherent space (`OLS`)
- Enforcing non-negativity: alternating projection onto coherent space and the non-negative orthant (Dykstra's method)
- Output of the model are coherent samples

Training: CRPS (sum of quantile losses at all the possible quantile levels of the sample forecast):

$$\mathrm{CRPS}\left(y_j^{[k]}, \{\hat{y}_j^{[k]}\}\right) = \sum_{s_i \in \{\hat{y}_j^{[k]}\}} \Lambda_{\alpha_i}(y_j^{[k]}, s_i), \tag{2}$$

where α_i is the quantile level of the sample s_i.

Model Details (Contd.)

Reconciliation:

- Euclidean projection of each sample onto the coherent space (`OLS`)
- Enforcing non-negativity: alternating projection onto coherent space and the non-negative orthant (Dykstra's method)
- Output of the model are coherent samples

Training: CRPS (sum of quantile losses at all the possible quantile levels of the sample forecast):

$$\mathrm{CRPS}\left(y_j^{[k]}, \{\hat{y}_j^{[k]}\}\right) = \sum_{s_i \in \{\hat{y}_j^{[k]}\}} \Lambda_{\alpha_i}(y_j^{[k]}, s_i), \tag{2}$$

where α_i is the quantile level of the sample s_i.

Differentiability:

- Sampling (reparameterization trick)
- Projection (matrix-vector multiplication).

Datasets:

DATASET	NO. TIME SERIES	HIERARCHY	τ	NO. ROLLS	FREQ
TAXI-1MIN	185	[60, 30, 1]	180	8	1-MIN
TAXI-5MIN	185	[12, 6, 1]	144	2	5-MIN
ELEC-15MIN	319	[4, 2, 1]	96	1	15-MIN
SOLAR-1H	137	[8, 1]	24	7	1-HOUR
EXCHANGERATE-1D	8	[5, 1]	30	5	1-BUSINESS DAY

Experiments

Datasets:

DATASET	NO. TIME SERIES	HIERARCHY	τ	NO. ROLLS	FREQ
TAXI-1MIN	185	[60, 30, 1]	180	8	1-MIN
TAXI-5MIN	185	[12, 6, 1]	144	2	5-MIN
ELEC-15MIN	319	[4, 2, 1]	96	1	15-MIN
SOLAR-1H	137	[8, 1]	24	7	1-HOUR
EXCHANGERATE-1D	8	[5, 1]	30	5	1-BUSINESS DAY

Methods:

- Base models: No reconciliation!
- Thief-variants: forecaster + reconciliation combinations
- DeepThief, `LogSparse`, `TimeGrad`
- Ours: CoP-DeepAR

Quantitative Results (Bottom-Level)

Scaled CRPS: Weighted mean quantile loss for the bottom time series (finest granularity)

	Taxi-1min	Taxi-5min	Elec-15min	Solar-1h	ExchangeRate-1d
ARIMA	-	0.594	0.139	0.521	0.008
Thief-ARIMA-MSE	-	-	0.135	0.502	0.008
Thief-ARIMA-OLS	-	-	*0.130*	*0.480*	0.008
Thief-ARIMA-Struc	-	-	0.132	0.485	0.008
ETS	0.671	0.882	0.366	0.606	0.008
Thief-ETS-MSE	0.661	0.854	0.303	0.577	0.008
Thief-ETS-OLS	0.650	0.850	*0.273*	*0.561*	0.008
Thief-ETS-Struc	*0.624*	*0.832*	0.278	0.562	0.008
Theta	0.649	0.973	0.212	1.083	**0.007**
Thief-Theta-MSE	0.636	0.942	0.201	*0.751*	**0.007**
Thief-Theta-OLS	0.621	0.937	*0.172*	0.786	0.008
Thief-Theta-Struc	*0.598*	*0.930*	0.181	0.827	**0.007**
DeepAR	0.447 ± 0.047	0.688 ± 0.089	0.127 ± 0.009	0.364 ± 0.008	0.016 ± 0.015
COPDeepAR	$\mathbf{0.327 \pm 0.007}$	$\mathbf{0.374 \pm 0.011}$	$\mathbf{0.114 \pm 0.004}$	$\mathbf{0.353 \pm 0.003}$	$\mathit{0.011 \pm 0.003}$

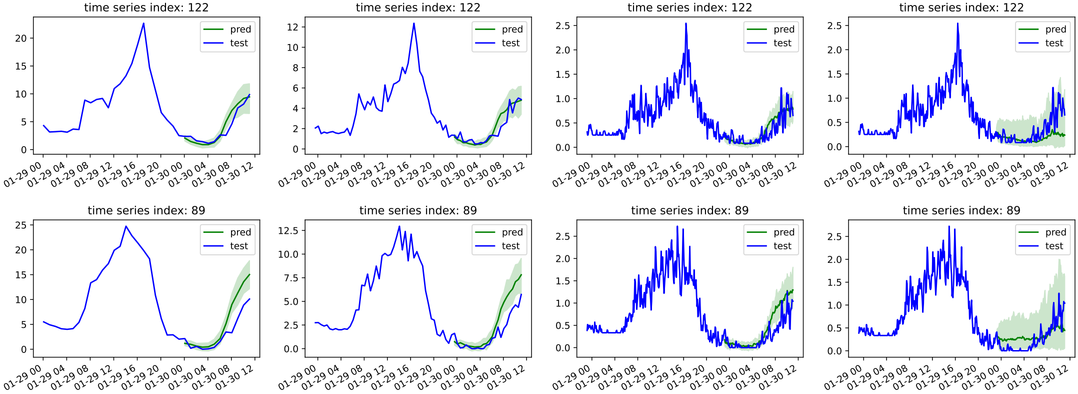
Quantitative Results (All Levels)

Weighted mean quantile losses at different temporal granularities

Dataset	Level	COPDeepAR	Best of Thief Variants
Taxi-1min	1 hour	$\mathbf{0.235 \pm 0.008}$	0.295 (Theta-Struc)
	30 min	$\mathbf{0.263 \pm 0.007}$	0.334 (Theta-Struc)
	1 min	$\mathbf{0.327 \pm 0.007}$	0.598 (Theta-Struc)
Taxi-5min	1 hour	$\mathbf{0.307 \pm 0.014}$	**0.307** (ETS-Struc)
	30 min	$\mathbf{0.330 \pm 0.013}$	0.437 (ETS-Struc)
	5 min	$\mathbf{0.374 \pm 0.011}$	0.832 (ETS-Struc)
Elec-15min	1 hour	$\mathbf{0.106 \pm 0.006}$	0.120 (ARIMA-OLS)
	30 min	$\mathbf{0.109 \pm 0.005}$	0.123 (ARIMA-OLS)
	15 min	$\mathbf{0.114 \pm 0.004}$	0.130 (ARIMA-OLS)
Solar-1H	8 hour	$\mathbf{0.343 \pm 0.008}$	0.399 (ARIMA-OLS)
	1 hour	$\mathbf{0.353 \pm 0.003}$	0.480 (ARIMA-OLS)
ExchangeRate-1D	1 week	0.011 ± 0.003	**0.007** (ARIMA-NaiveBU/MSE)
	1 day	0.011 ± 0.003	**0.007** (ARIMA-NaiveBU/MSE)

Forecast Visualization

Taxi-5min: First three plots in each row are forecasts of COP-DeepAR at 3 levels; the rightmost one is that of DeepAR at the bottom level.



Conclusions

- First End-to-End model that produces coherent, probabilistic forecasts for temporal hierarchies.
- Further evidence that reconciliation helps especially for high frequency time series.
- Can use any other univariate forecasting models (amenable to auto-differentiation).
- Can replace the Gaussian output with other density functions (student-t) or even a quantile function (via splines) to handle non-Gaussian data! Sampling still does not pose problem for the auto-differentiation.

Conclusions

- First End-to-End model that produces coherent, probabilistic forecasts for temporal hierarchies.
- Further evidence that reconciliation helps especially for high frequency time series.
- Can use any other univariate forecasting models (amenable to auto-differentiation).
- Can replace the Gaussian output with other density functions (student-t) or even a quantile function (via splines) to handle non-Gaussian data! Sampling still does not pose problem for the auto-differentiation.
- Extensions
  - Cross-sectional, cross-temporal hierarchies.
  - Different Forecaster at each granularity: e.g., simple models at coarser granularity and DeepAR at finer granularities.