Supplementary Material

Contents

1 Hyperparameter optimization

Here we give the details of the hyperparameter optimization for our model and its variants/ablations (`Hier-E2E`, `DeepVAR`, `DeepVAR`+). For the competing methods, the hyperparameters are auto-tuned by the corresponding implementations; any hyperparameter setting that is not tuned is reported as a separate model. Note that tuning is done on a separate validation set created in the same way for all methods, as described in the experiments section.

We validated our model over a hyperparameter range for some values and left others at default values set in the `GluonTS` library (Alexandrov et al., 2019). See Table 1 for details of the tuned parameters. Notably the number of layers and the cells of the RNN are kept at their default values: 2 and 40 respectively. `DeepVAR` unrolls the RNN over a short subsequence of the given time series, the length of which is known as `context_length`. This is typically a multiple (denoted `context_length_factor`) of the `prediction_length`. For example, for a `prediction_length` of 8, `context_length_factor`= 2 would give a `context_length` of 16. In the table, two ranges are shown for this hyperparameter; the smaller values were used for datasets with longer prediction horizons (`Tourism`, `Tourism-L`, `Labour`) and the larger values were used for datasets with shorter prediction horizons (`Traffic`, `Wiki`). Batch size of 32 was used for all datasets except for the high-dimensional `Tourism-L` dataset, where batch size was set at 4.

We also employed two forms of the training loss: one that minimizes CRPS loss on the samples (`num_training_samples`=200) directly and one that minimizes negative log-probability (under the Gaussian model) on the parameters of the empirical distribution given by the samples. For larger datasets, we observed that the latter approach offered faster convergence for a small number of samples (`num_training_samples`=50) during training. So, the training loss function is also treated as a hyperparameter of our model with two possibilities, indicated by `num_training_samples` in Table 1.

Parameter	Values
`epochs`	$\{10, 25, 50, 100\}$
`context_length_factor`	$\{2, 3, 4\}$ and $\{15, 25, 40, 60\}$
`warmstart_frac`	$\{0, 0.1\}$
`learning_rate`	1e-3
`batch_size`	$\{4, 32\}$
`num_training_samples`	200 and 50
`num_prediction_samples`	200

Table 1: Hyperparameter values.

We found that "warm-starting" our method by running a base version of `DeepVAR` (maximizing likelihood on distribution parameters directly) for the first 10% of epochs and then completing the rest with our approach improved convergence to an accurate forecast distribution. We included this (to warm-start or not) as a hyperparameter. `warmstart_frac` in Table 1 refers to the fraction of epochs where no sampling of the learned distribution was done during training and instead likelihood loss was optimized directly on the distribution parameters themselves (as in DeepVAR).

We trained different models on a validation split with hyperparameter combinations taken from the parameter grid and selected the best set according to the lowest CRPS score.

2 Aggregation Level-wise Forecast Errors

In order to assess if the gains in the performance are uniform across aggregation levels, we present CRPS scores by level of aggregation. Tables 2 through 6 report the mean CRPS scores computed for time series at each aggregation level for all the datasets considered. For reference we also included the mean CRPS scores computed for all the time series in the hierarchy (same as the ones reported in the main version of the paper) in these tables. The most aggregated level in the hierarchy is Level 1 (i.e., the root of the aggregation tree) and higher level numbers correspond to more disaggregated levels in the hierarchy. As expected, accuracy results generally improve with decreasing level number. However, note that our method achieves performance gains consistently across all aggregation levels unlike some of the state-of-the-art, which trade-off favorable accuracy at the aggregated levels with less favorable accuracy at the disaggregated levels; e.g., see the results of ETS-MinT-ols for the Wiki dataset (Table 5) and ARIMA-MinT-ols for the Tourism-L dataset (Table 6). Also see the table presented in the main version of the paper that summarizes level-wise forecast errors across datasets.

	Level / Method	1 (root)	2	3	4	All
	ARIMA-NaiveBU	0.0437	0.0441	0.0447	0.0489	0.0453
	ETS-NaiveBU	0.0416	0.0418	0.0421	0.0471	0.0432
	ARIMA-MinT-shr	0.0453	0.0455	0.0459	0.0499	0.0467
	ARIMA-MinT-ols	0.0448	0.0450	0.0455	0.0499	0.0463
	ETS-MinT-shr	0.0440	0.0442	0.0444	0.0492	0.0455
	ETS-MinT-ols	0.0445	0.0447	0.0448	0.0495	0.0459
	ARIMA-ERM	[0.0365]	[0.0379]	0.0391	0.0459	0.0399
	ETS-ERM	0.0409	0.0437	0.0452	0.0525	0.0456
	PERMBU-MinT	0.0406±0.0002	0.0389±0.0002	[0.0382±0.0002]	[0.0397±0.0003]	0.0393±0.0002
	Hier-E2E (Ours)	**0.0311±0.0122**	**0.0336±0.0089**	**0.0336±0.0082**	**0.0378±0.0060**	**0.0340±0.0088**
ablation study	DeepVAR	0.0352±0.0079	0.0374±0.0051	0.0383±0.0038	0.0417±0.0038	0.0382±0.0045
ablation study	DeepVAR+	0.0416±0.0094	0.0437±0.0078	0.0432±0.0076	0.0448±0.0066	0.0433±0.0079

Table 2: Labour: Mean CRPS scores (lower is better) computed for time series at each aggregation level, averaged over 5 runs. For reference, in the last column, we also include the mean CRPS scores computed for all the time series in the hierarchy (same as the ones reported in the main version of the paper). The best result is highlighted in **boldface**, while the best result among the state-of-the-art (without the proposed method Hier-E2E and its variants) is highlighted in [boxes]. Among the competing methods (without Hier-E2E and its variants), both PERMBU-MinT and ARIMA-ERM perform consistently better by achieving the best result for 2 out of 4 levels; since PERMBU-MinT has better overall (last column) CRPS score than ARIMA-ERM, its result is included in the summary table presented in the main version of the paper.

	Level / Method	1 (root)	2	3	4	All
	ARIMA-NaiveBU	0.0471	0.0471	0.0480	0.1812	0.0808
	ETS-NaiveBU	0.0128	0.0128	0.0351	0.2053	0.0665
	ARIMA-MinT-shr	0.0466	0.0466	0.0466	0.1682	0.0770
	ARIMA-MinT-ols	0.0852	0.0852	0.0852	0.1905	0.1116
	ETS-MinT-shr	0.0601	0.0601	0.0601	0.2050	0.0963
	ETS-MinT-ols	0.0765	0.0765	0.0765	0.2145	0.1110
	ARIMA-ERM	**0.0087**	**0.0112**	0.0255	0.1410	0.0466
	ETS-ERM	0.0828	0.0828	0.0828	0.1624	0.1027
	PERMBU-MinT	0.0331±0.0085	0.0341±0.0081	0.0417±0.0061	0.1621±0.0027	0.0677±0.0061
	Hier-E2E (Ours)	0.0184±0.0091	0.0181±0.0086	0.0223±0.0072	**0.0914±0.0024**	**0.0376±0.0060**
ablation study	DeepVAR	0.0225±0.0109	0.0204±0.0044	**0.0190±0.0031**	0.0982±0.0012	0.0400±0.0026
ablation study	DeepVAR+	0.0250±0.0082	0.0244±0.0063	0.0259± 0.0054	0.0982± 0.0017	0.0434±0.0049

Table 3: `Traffic`: Mean CRPS scores (lower is better) computed for time series at each aggregation level, averaged over 5 runs. For reference, in the last column, we also include the mean CRPS scores computed for all the time series in the hierarchy (same as the ones reported in the main version of the paper). The best result is highlighted in **boldface**, while the best result among the state-of-the-art (without the proposed method `Hier-E2E` and its variants) is highlighted in boxes. Among the competing methods (without `Hier-E2E` and its variants), `ARIMA-ERM` performs consistently better by achieving the best result for all 4 levels; hence its result is included in the summary table presented in the main version of the paper.

	Level / Method	1 (root)	2	3	4	All
	ARIMA-NaiveBU	0.0588	0.0945	0.1366	0.1653	0.1138
	ETS-NaiveBU	0.0545	0.0809	0.1194	0.1483	0.1008
	ARIMA-MinT-shr	0.0625	0.0989	0.1395	0.1677	0.1171
	ARIMA-MinT-ols	0.0619	0.1018	0.1419	0.1723	0.1195
	ETS-MinT-shr	0.0592	0.0793	0.1202	0.1466	0.1013
	ETS-MinT-ols	0.0597	0.0748	0.1200	0.1461	0.1002
	ARIMA-ERM	0.2201	0.3905	0.8121	0.9321	0.5887
	ETS-ERM	1.4397	1.9941	2.8494	3.2190	2.3755
	PERMBU-MinT	0.0472±0.0012	**0.0605±0.0006**	**0.0903±0.0006**	**0.1106±0.0005**	**0.0771±0.0001**
	Hier-E2E (Ours)	**0.0402±0.0040**	0.0658±0.0084	0.1053±0.0053	0.1223±0.0039	0.0834±0.0052
ablation study	DeepVAR	0.0519±0.0057	0.0755±0.0011	0.1134±0.0049	0.1294±0.0060	0.0925±0.0022
ablation study	DeepVAR+	0.0508±0.0085	0.0750±0.0066	0.1180±0.0053	0.1393±0.0048	0.0958±0.0062

Table 4: `Tourism`: Mean CRPS scores (lower is better) computed for time series at each aggregation level, averaged over 5 runs. For reference, in the last column, we also include the mean CRPS scores computed for all the time series in the hierarchy (same as the ones reported in the main version of the paper). The best result is highlighted in **boldface**, while the best result among the state-of-the-art (without the proposed method `Hier-E2E` and its variants) is highlighted in boxes. Among the competing methods (without `Hier-E2E` and its variants), `PERMBU-MinT` performs consistently better by achieving the best result for all 4 levels; hence its result is included in the summary table presented in the main version of the paper.

	Level / Method	1 (root)	2	3	4	5	All
	ARIMA-NaiveBU	0.1897	0.2790	0.4111	0.4117	0.5943	0.3772
	ETS-NaiveBU	0.3410	0.3863	0.4631	0.5051	0.6410	0.4673
	ARIMA-MinT-shr	0.0801	0.1384	0.2558	0.2951	0.4642	0.2467
	ARIMA-MinT-ols	0.1079	0.1743	0.2857	0.3253	0.4979	0.2782
	ETS-MinT-shr	0.2180	0.2666	0.3451	0.3880	0.5936	0.3622
	ETS-MinT-ols	**0.0234**	0.1456	0.2616	0.3138	0.6065	0.2702
	ARIMA-ERM	0.0788	0.1236	0.2346	0.2758	0.3902	0.2206
	ETS-ERM	0.1558	0.1614	**0.2010**	**0.2399**	**0.3506**	0.2217
	PERMBU-MinT	0.0791±0.0171	0.1575±0.0132	0.2778±0.0385	0.3138±0.0383	0.5776±0.0552	0.2812±0.0240
	Hier-E2E (Ours)	0.0419±0.0285	**0.1045±0.0151**	0.2292±0.0108	0.2716±0.0091	0.3720±0.0150	**0.2038±0.0110**
ablation study	DeepVAR	0.0905±0.0323	0.1418±0.0249	0.2597±0.0150	0.2886±0.0112	0.3664±0.0068	0.2294±0.0158
	DeepVAR+	0.0755±0.0165	0.1289±0.0171	0.2583±0.0281	0.3108±0.0298	0.4460±0.0271	0.2439±0.0224

Table 5: Wiki: Mean CRPS scores (lower is better) computed for time series at each aggregation level, averaged over 5 runs. For reference, in the last column, we also include the mean CRPS scores computed for all the time series in the hierarchy (same as the ones reported in the main version of the paper). The best result is highlighted in **boldface**, while the best result among the state-of-the-art (without the proposed method Hier-E2E and its variants) is highlighted in boxes. Among the competing methods (without Hier-E2E and its variants), ETS-ERM performs consistently better by achieving the best result for 3 out of 5 levels; hence its result is included in the summary table presented in the main version of the paper.

	Level / Method	1 (root)	2 (geo.)	3 (geo.)	4 (geo.)	2 (trav.)	3 (trav.)	4 (trav.)	5 (trav.)	All
	ARIMA-NaiveBU	0.0818	0.1015	0.1569	0.2106	0.1016	0.1564	0.2479	0.3364	0.1741
	ETS-NaiveBU	0.0802	0.0989	0.1561	0.2058	0.0927	0.1484	0.2408	0.3291	0.1690
	ARIMA-MinT-shr	0.0438	**0.0816**	0.1433	0.2036	**0.0830**	0.1479	0.2437	0.3406	0.1609
	ARIMA-MinT-ols	**0.0394**	0.0825	0.1500	0.2164	0.1056	0.1642	0.2610	0.3638	0.1729
	ETS-MinT-shr	0.0505	0.0902	0.1501	0.2024	0.0890	0.1439	0.2415	0.3343	0.1627
	ETS-MinT-ols	0.0484	0.0897	0.1542	0.2102	0.0891	0.1455	0.2499	0.3473	0.1668
	ARIMA-ERM	0.2546	0.3756	0.4947	0.6354	0.3620	0.5368	0.7974	1.0511	0.5635
	ETS-ERM	0.1161	0.3231	0.4684	0.6143	0.2622	0.4853	0.7741	1.0209	0.5080
	PERMBU-MinT	—	—	—	—	—	—	—	—	—
	Hier-E2E (Ours)	0.0810±0.0053	0.1030±0.0030	**0.1361±0.0024**	0.1752±0.0026	0.1027±0.0062	**0.1403±0.0047**	**0.2050±0.0028**	**0.2727±0.0017**	**0.1520±0.0032**
ablation study	DeepVAR	0.1029±0.0188	0.1076±0.0119	0.1407±0.0081	**0.1741±0.0066**	0.1100±0.0139	0.1485±0.0099	0.2078±0.0076	0.2731±0.0066	0.1581±0.0102
	DeepVAR+	0.1214±0.0360	0.1364±0.0299	0.1713±0.0243	0.2079±0.0215	0.1370±0.0289	0.1776±0.0221	0.2435±0.0170	0.3108±0.0164	0.1882±0.0242

Table 6: Tourism-L: Mean CRPS scores (lower is better) computed for time series at each aggregation level, averaged over 5 runs. Tourism-L is a grouped dataset that contains two hierarchies sharing a common root: one for geographic divisions with 4 levels and 76 bottom series and one for purpose-of-travel with 5 levels and 304 bottom series. For reference, in the last column, we also include the mean CRPS scores computed for all the time series in both hierarchies (same as the ones reported in the main version of the paper). The best result is highlighted in **boldface**, while the best result among the state-of-the-art (without the proposed method Hier-E2E and its variants) is highlighted in boxes. Among the competing methods (without Hier-E2E and its variants), ARIMA-MinT-shr performs consistently better by achieving the best result for 3 out of 8 levels; hence its result is included in the summary table presented in the main version of the paper.

References

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J., et al. GluonTS: Probabilistic Time Series Models in Python. *JMLR*, 2019.