Appendix A - Dataset

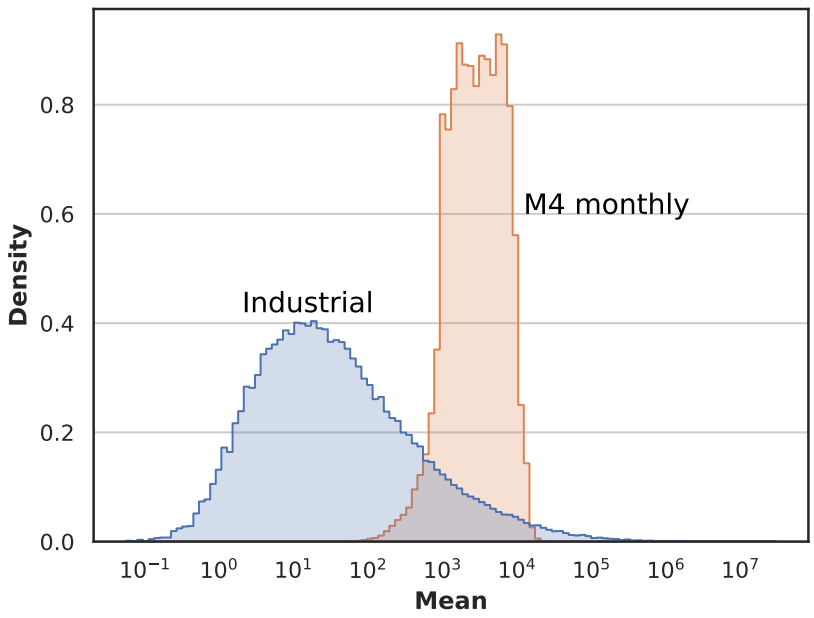


Fig. 1: The histogram displays the distribution density of the mean values of the instances in the industrial dataset and the M4 monthly dataset. The x-axis is logarithmic.

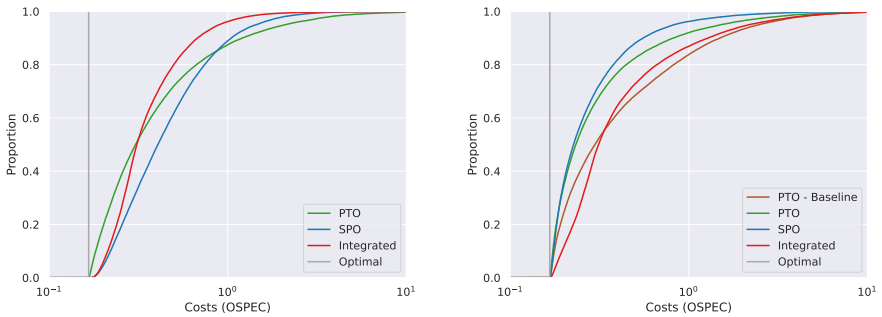
Appendix B - Hyperparameter description

Table 1: Overview over the selected values for different hyperparameters

Component	Hyperparameter	Value(s)
Optimisation model	T	6
	o_s	200, 4 000, 80 000
	h_s	1
	b_s	20
TimesNet	top k	5
	kernels	6
	encoder layer	2
	hidden size	64
	dropout	0.1
	train epochs	10
	batch size	32
	patience	2
	learning rate	1e-4
Time-MoE	training epochs	3
	learning rate	1e-6

Appendix C - CDF Plots

M4



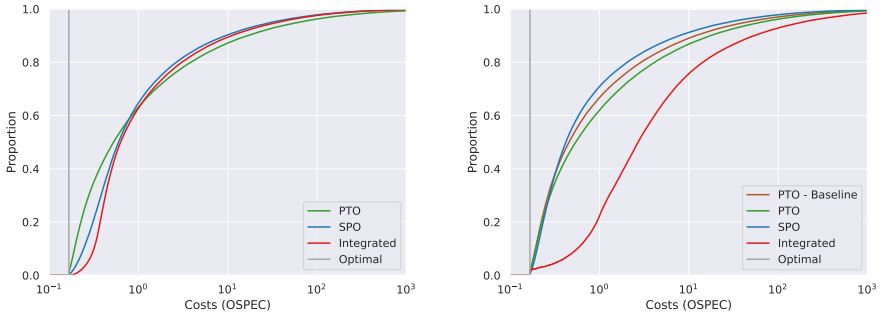
(a) TimesNet (b) Time-MoE

Fig. 2: CDF plots comparing the different techniques and the optimal case based on the scaled costs (OSPEC) of the different instances in the M4 monthly dataset. The order costs were set to 4000.0 in these experiments and the x-axes is logarithmic.

Figure 2 displays the cumulative distribution functions (cdf) of the scaled costs (OSPEC) resulting from applying the different methods to the instances in the M4 monthly dataset. Independent of the utilised prediction model (TimesNet: top, Time-MoE: bottom), the cdf of the optimal solutions is a constant function at one over six, because the cost in case of a given instance is scaled with the corresponding optimal cost value, before the costs are divided by the number of considered periods (six). The other curves, belonging to the experiments where the demand is uncertain, all are starting with a steep slope before approaching 1.0 with increasing costs and decreasing slope. In case TimesNet is deployed as prediction model, it can be seen that PTO finds very cheap solutions for more instances than SPO and Integrated, but is inferior to the two techniques, if the proportions below a higher cost value are considered. Hereby, the Integrated approach proves to be better than the SPO approach as its cdf is closer to the optimal curve. If fine-tuned versions of Time-MoE however are utilised as prediction models, the Integrated approach is inferior to the PTO and SPO approaches (see Figure 2 bottom). This can be explained by the fact that in the pretraining of Time-MoE the incorporation of a mapping from demands to orders into the model was not considered.
While the cumulative distribution function of the PTO approach used for fine-tuning is closer to the optimal curve than the cdf of the PTO baseline and approximately overlays with the cdf of the SPO approach for small true costs, it

is clearly visible that the cdf of the SPO technique approaches 1.0 faster. This means that fine-tuning the Time-MoE foundation model in a PTO fashion can improve the final decision quality compared to the omission of fine-tuning, but that fine-tuning with a DFL startegy is preferable.

Industrial Data



(a) TimesNet (b) Time-MoE

Fig. 3: CDF plots comparing the different techniques and the optimal case based on the scaled costs (OSPEC) of the different instances in the industrial dataset. Note the logarithmic x-axis.

The CDF plot regarding the experiments with TimesNet in Figure 3 shows again that SPO can solve less instances than PTO with scaled costs below 0.5, but eventually slightly converges faster towards 1.0 than PTO. In case Time-MoE is deployed as prediction model, the SPO curve is, especially for higher cost values, again closer to the optimal curve than the PTO curve.
With regard to the cumulative distribution function of the Integrated learning approach it is noticeable that on the industrial dataset for both prediction models the curve is below the curve of the SPO method. Moreover, it is visible that the fine-tuning of Time-MoE in a prediction-focused manner does not lead to a better cumulative distribution function over the instances in the industrial dataset.

Appendix D - Instance-based comparison of DFF and prediction focused fine-tuning

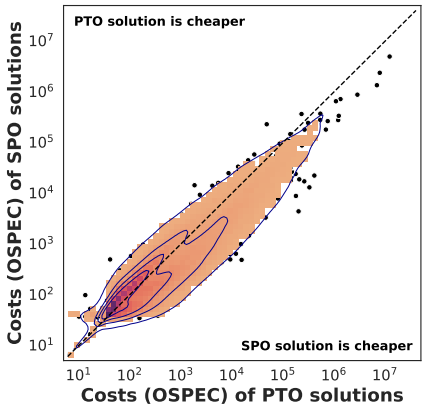


Fig. 4: Bivariate plot comparing the PTO and DFL approach based on the induced costs (OSPEC, not scaled). Each point corresponds to one of 1 000 instances sampled from the industrial dataset and the dashed line represents the identity. All sampled points in the triangle above the dashed identity line, represent instances in which DFF leads to more expensive solutions than prediction-focused fine-tuning. For the lower right triangle vice versa holds. The blue lines display the contours of a two dimensional kernel density estimation over all instance. The heatmap is based on a histogram with 100 bins and a threshold of 5%. Time-MoE was chosen as prediction model in this case and both axes are logarithmic.