Appendix A - Data Set

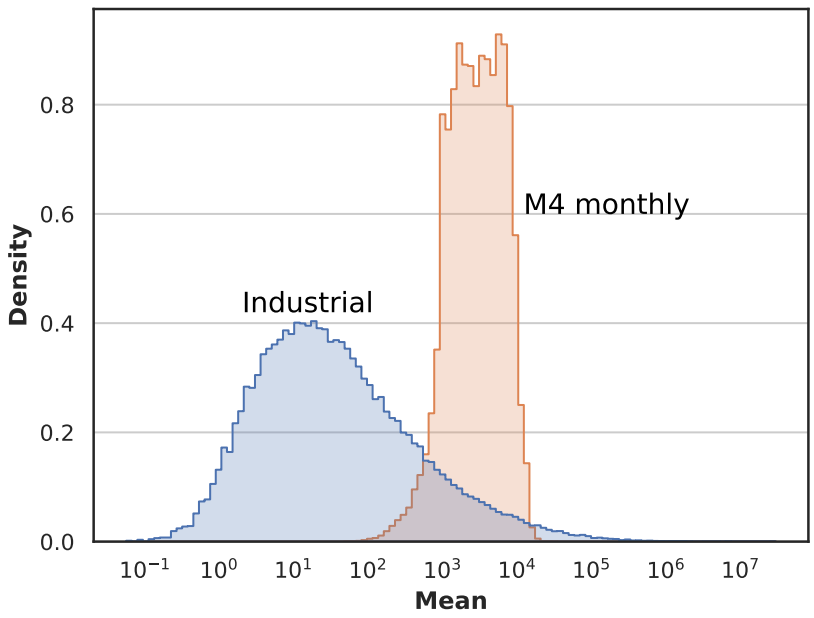


Fig. 1: The histogram displays the distribution density of the mean values of the instances in the industrial data set and the M4 monthly data set. The x-axis is logarithmic.

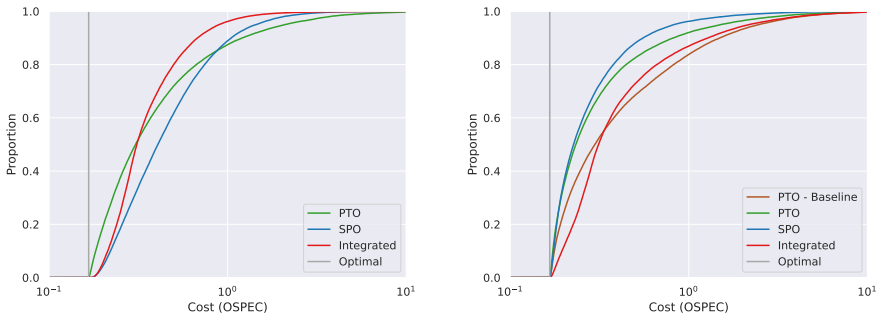
Figure 1 shows that the distribution of mean values of the time series considerably differs between the industrial data set and the M4 monthly data set.

Appendix B - Hyperparameter description

Table 1: Overview of the selected values for different hyperparameters.

Component	Hyperparameter	Value(s)
Optimisation model	Planning horizon T	6
	Order cost o_s	200, 4 000, 80 000
	Storage cost h_s	1
	Backlog cost b_s	20
TimesNet	Top k	5
	Kernels	6
	Encoder layer	2
	Hidden size	64
	Dropout	0.1
	Training epochs	10
	Batch size	32
	Patience	2
	Learning rate	10^{-4}
Time-MoE	Training epochs	3
	Learning rate	10^{-6}

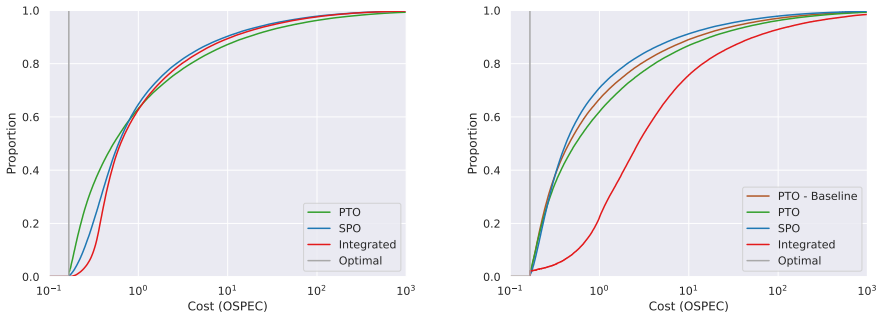
Appendix C - CDF Plots



(a) TimesNet (b) Time-MoE

Fig. 2: CDF plots comparing the different techniques and the optimal case based on the scaled cost (OSPEC) of the different instances in the M4 monthly data set. The order cost were set to 4000 in these experiments and the x-axis is logarithmic.

Figure 2 displays the cumulative distribution functions (CDF) of the scaled cost (OSPEC) resulting from applying the different methods to the instances in the M4 monthly data set. Independent of the utilised prediction model (TimesNet: left, Time-MoE: right), 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 cost are divided by the number of considered periods (six). The other curves, belonging to the experiments where the demand is uncertain, all start with a steep slope before approaching 1.0 with increasing cost 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. Here, 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 2b). While the CDF 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 cost, 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 strategy is preferable.



(a) TimesNet (b) Time-MoE

Fig. 3: CDF plots comparing the different techniques and the optimal case based on the scaled cost (OSPEC) of the different instances in the industrial data set. The x-axis is logarithmic.

The CDF plot regarding the experiments with TimesNet in Figure 3a shows again that PTO dominates across instances with smaller cost, but eventually, DFL 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 CDF of the Integrated learning approach it is noticeable that on the industrial data set 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 CDF compared to not fine-tuning.

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

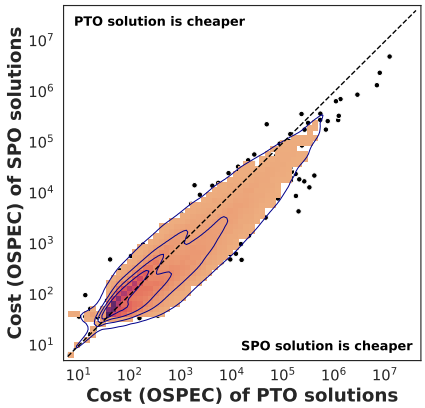


Fig. 4: Scatter plot comparing the PTO and DFL approach based on the induced cost (OSPEC, not scaled). Each point corresponds to one of 1000 instances randomly sampled from the industrial data set and the dashed line represents the identity. All sampled points in the triangle above the dashed identity 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 instances. The heatmap is based on a histogram with 100 bins and a threshold of 5%. Time-MoE was chosen as the prediction model in this case and both axes are logarithmic.