Eidgenössisches Departement für Umwelt,
Verkehr, Energie und Kommunikation UVEK
Bundesamt für Verkehr BAV
Abteilung Politik
Sektion Direktionsgeschäfte
Mühlestrasse 6, 3063 Ittigen,

Lausanne, 24.11.2023

Progress Report «Project INITIATE»

1 Introduction

This report provides an overview of the progress made in project INITIATE. This project is a research project funded by the funding scheme for railway infrastructure "Bahninfrastrukturfonds" BIF of 2014. In this project, EPFL and SBB work together to improve the detection and prediction of critical conditions in railway power network systems by developing intelligent algorithms for control system data.

The project was initially started at ETH Zurich in 2021 and was transferred to EPFL in March 2022 due to the transfer of the research team from ETHZ to EPFL.

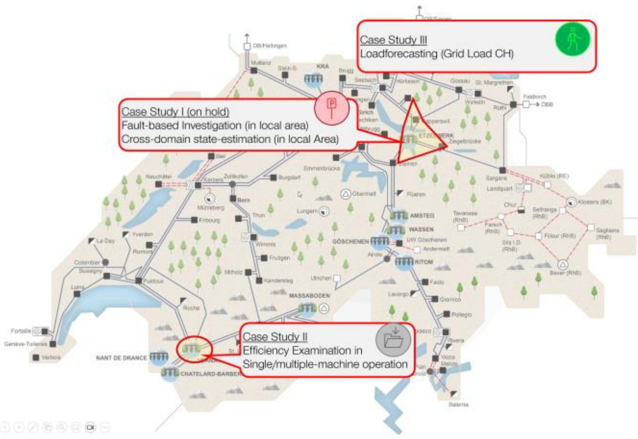

Figure 1: Power grids sites from which the generated data of case study I, II and III originates.

In June 2021, a group of domain experts from different fields collected and rated various case studies as candidates for investigation. The domain experts estimated the highest overall benefits of all case studies as well as immediate benefits if selected as the first case study. Based on these scores, two studies were selected to be the most promising and, thus, were processed first: a case study on state estimation and fault detection from power grid sensors in local areas such as power plants with connected substations (**case 1**) and a case study on hydro-power plant efficiency estimation from hydro and electrical sensors (**case 2**). In 2023, Case Study 1 was temporarily halted to focus on a new case study, load forecasting in traction power grids (**case 3**) based on findings from Case Study 1. Two workgroups were formed to study the different cases, led by Philipp Wenk from SBB's side and Olga Fink from EPFL's side:

- Project Lead: Philipp Wenk and (SBB) Olga Fink (EPFL)
    - Study case 1: state estimation and fault detection from power grid sensors
        - Robert Strietzel (SBB, Collaboration Lead)
        - Raffael Theiler (EPFL, Research Lead)
    - Study case 2: hydro-power plant efficiency estimation
        - Roland Schäfer (SBB, Consultative Participant)
        - Francesco Fusaro (SBB, Collaboration Lead)
        - Mengjie Zhao (EPFL, Research Lead)
    - Study case 3: load forecasting in traction power grids
        - Raffael Theiler (EPFL, Research Lead)
        - Roland Schäfer (SBB, Collaboration Lead)
        - Yvan Tercier (SBB, Passive Participant)
        - Aleksander Lelievre (SBB, Passive Participant)
        - Emma Lacaille (SBB, Passive Participant)

2 Load forecasting in Traction Power Grids (Case 3)

Table 1: Spacetimeformer (STF) forecasting performance on the large dataset compared to the EOB baseline.

	MAE		MAPE		MSE	
	EUB	STF	EUB	STF	EUB	STF
Mo	9,476	9,20	4.32	4.13	145,12	169,73
Di	11,40	10,09	4.64	4.04	208,48	171,22
Mi	10,16	8,77	3.82	3.25	162,14	121,48
Do	10,56	8,70	3.85	3.23	174,81	117,93
Fr	10,43	8,79	3.87	3.29	191,84	128,22
Sa	9,75	9,74	4.18	4.13	157,64	154,96
So	10,04	6,97	5.13	3.59	156,82	78,27

Timetable-based Energy Forecasting: In the 2022 report, we emphasized the importance of reliable state forecasting for unsupervised anomaly detection. For this reason, we tested both autoregressive and multi-step models, including transformers, for the forecasting of electrical and hydraulic signals in the power grid environment. Although graph neural networks outperformed transformer models in case study 1 by a small margin, we gathered evidence suggesting that a specific spatial-temporal transformer model, the Spacetimeformer (STF), is well-suited for forecasting tasks within power grid environments. This is due to the flexibility of the spatial-temporal attention in STF, which allows us to effectively incorporate timetable-based information within the model. This is highly significant for the task of day-ahead forecasting, as the train scheduling for the next day is usually fixed, and the associated rolling stock weight can be estimated by teams in operation and dispatch. To showcase the potential of spatial-temporal transformers as forecasters in this environment, we tackled the task of total load forecasting on the entire traction power grid of SBB, which is an important source of information for energy trading, grid dispatching and load management applications. At the outset of this reporting period, in the first meeting of 2023 with SBB and BAV, we conveyed the significance of our discovery and its value, leading us to temporarily interrupt Case Study 1 in favor of Case Study 3.

Training Data: For the task of load forecasting for the SBB traction power grid, we collected several datasets consisting of the total grid load, transportation data such as freight transport in gross tonne-kilometers, the total number of trains on the railroad network, and freight tonnage. These features are included separately for passenger and freight traffic. We also consider aggregated weather forecasts consisting of temperature (measured in the morning and afternoon), solar radiation, precipitation (from rain gauges), and a categorical state of the weather (e.g. "rainy" or "sunny"). The training data was increased throughout the project from one (2021-2023Q1) to two and a half years (2020- 2023 Q2). We refer to it as the short and the long dataset.

Baseline Methods: The SBB energy trading department currently maintains a load forecasting model, the "EUB prognosis", to support the decision-making of the trader in the day-ahead energy bidding. We use this model as a baseline to assess the performance of our STF model. The EUB model is a well-tuned linear regression model. Implementation details remain undisclosed due to business confidentiality.

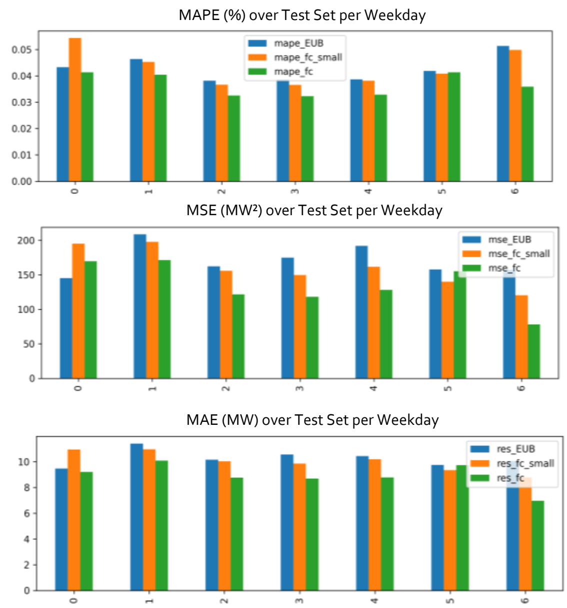

Figure 2: Comparison of the Spacetimeformer (STF) performance trained on the small (fc_small) and large dataset (fc) for each day of the week, starting on Monday (code 0), aggregated over the large test set (calendar weeks 4-30 in 2023). The model generalizes better with more training data and outperforms the EUB forecast on every metric except for MSE on Monday.

Experimental Setup, Challenge Days, and Ablations: To evaluate the performance of the proposed model, three performance evaluations took place. First, the model was trained and validated on 80% of the small dataset and then tested on the remaining 20% (which corresponds to January to April 2023, including the Easter holidays, "small" performance in Figure 1). Second, eight randomly chosen challenge days were selected by SBB of which details such as the date and the performance of the baseline were not shared with EPFL ahead of validation to ensure that no data leaks were biasing the forecast performance. Third, after an initial success, the "large dataset" (as described above) was collected and used for training and testing. Again, the data was split into 80% training and validation set and 20% test dataset. For all testing, the temporal ordering of training, test, and validation data sets were kept, such that, in that order, the validation and then the test indices are higher than the training indices. To test the effectiveness of spatial-temporal attention for timetable-based information, we evaluate the performance of the transformer model without information on the next day (forecasting target) on the dataset. We show the results in

Table 2. Additionally, we tested four different variants of STEMGNN, the best-performing model on the anomaly detection task in case study 1, which did not outperform STF.

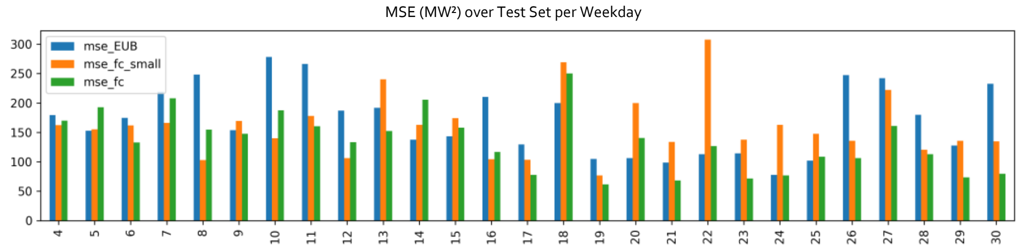

Figure 2: Comparison of the Spacetimeformer (STF) performance trained on the small (fc_small) and large dataset (fc) for each calendar week. The model generalizes better with more training data. We show the MSE because it assigns a higher weight to outliers, which are critical for trading.

Interpretation of the Results: A typical load profile (of day 5), including forecasts from the challenge days, is shown in Figure 4. It emphasizes the dynamic power profile of the traction power grid throughout a single day with the typical two load peaks during the morning and afternoon rush hour that we forecast for this case study. Curve statistics such as the load peak magnitudes strongly differ when comparing workdays with the weekend or with seasonal events such as Easter, requiring the model to generalize well to different seasonal contexts and conditions. Trained on the small dataset, the STF model outperformed the EUB prognosis on 4 out of 8 days in the challenge days (3.95 average MaveE across the eight challenge days compared to 5.20 average MaveE for the EUB prognosis). Furthermore, we validate the performance in Figure 1 per weekday and per calendar week in Figure 2. We also show the tabulated performance of the STF model on the large dataset in Table 1. In the weekday performance analysis, the STF model outperforms the EOB baseline on every weekday except for Monday, in terms of MSE. The comparison to Monday's MAE, where STF performs better, suggests that there may be outliers on Monday that require a detailed evaluation.

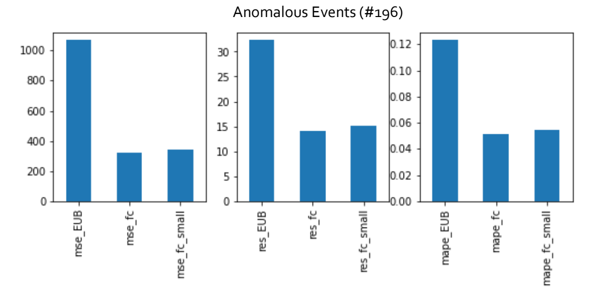

Figure 3: Performance of the STF compared to the EUB forecast for 196 unusual events where the forecast is two standards deviations above the mean MAE. We report MSE, MAE and MAPE.

Model generalization concerning seasonality is showcased in the calendar week performance analysis in Figure 2. Week 18 appears to be an outlier week that is hard to forecast in general. Outside of week 18, the STF model generalizes better compared to the EUB baseline across the first half of 2023, especially during weeks 8-12 and weeks 26-30. The opposite, that the EUB prognosis outperforms the STF model by a similar margin of error does not happen for the large test set, with the closest instances being week 14 and week 18 where the EUB performance remains better compared to STF. Weeks 26 and 27 are during the Swiss summer holidays, highlighting the strong impact of seasonality in this data set. We suggest increasing the dataset size again to have at least one year of test data to test for other seasonal effects. Figures 1 and 2 indicate that the STF model generalization increases when presented with more data from the long dataset (fc vs fc_long), which suggests that the model is well suited for the task and is capable of capturing and generalizing to the patterns in the data. To show the robustness of the STF model towards unusual events, we compare the performance for 196 hourly forecasts where the performance of the EUB prognosis was more than two standard deviations above the mean MAE in Figure 3. In these situations, the STF performance is roughly twice as good as that of the EUB model.

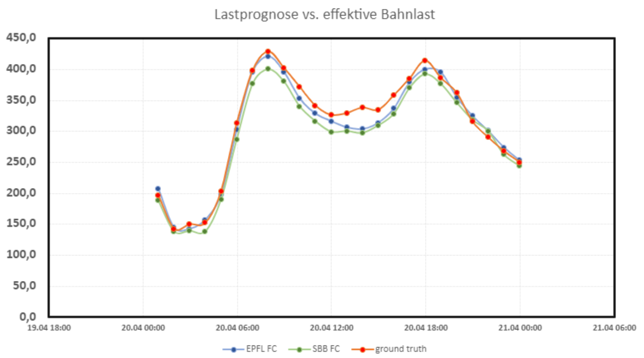

Figure 4: Visual comparison of the EPFL STF forecast (blue) with the SBB EOB baseline forecast and the ground truth (orange) for challenge day 5 around Easter 2023 in MW.

Technical Advantage of the STF: Considering the drawbacks of recurrent neural networks, the original implementation of the transformer model (2017) was targeted at processing long input sequences of natural language. Since then, the transformer has become arguably the most successful model for this task. Shortly thereafter, the model was also adapted to process other sequential data such as time series with minor modifications. However, when the transformer model is applied to multivariate time series, several drawbacks arise. First, because it is a requirement in natural language processing, the attention is permutation invariant. Second, during input tokenization, a preprocessing step, the model projects the input vector at time point *t* in its entirety to an embedding space. Because of these drawbacks, it is difficult for the model to learn complex relationships between the individual time series for multivariate inputs. Empirical studies show that the standard transformer model is inefficient for temporal relations that cannot be captured directly by point-wise semantic correlations. For that reason, we adapt the spatial-temporal attention for case study 3. Within spatial-temporal attention, the model can freely attend to single variables at any given time, translating the multivariate forecasting into a spatial-temporal sequence. Finally, adding a contextual embedding vector to the spatial-temporal sequence enables the integration of covariates such as the timetable-based information of the forecasting target day.

Table 2: Ablation study comparing the effectiveness of adding timetable-based information (TTI) to the Spacetimeformer model, reporting normalized scores.

	NMSE	NSMAPE	NMAE
STF (w/o TTI)	9.35e-4	4.98%	2.17e-2
STF	4.82e-4	3.98%	1.68e-2

3 Efficiency Estimation (Case 2)

Enhanced dynamics modeling with additional nozzle information. Our previous study in 2022 showed that the prediction model encountered challenges when adapting to rapidly varying operating conditions. Therefore, we extended Physics-informed Convolutional Neural Networks (PICNN) to PICNNv3 to incorporate additional nozzle information. This extension not only allowed for a more robust handling of operational dynamics but also enabled the qualitative extraction and assessment of the friction coefficient, μ, as a latent variable for efficiency degradation inspection. In addition, we explored the Encoder-Decoder Long Short-Term Memory (ENCDEC-LSTM) approach, utilizing a reconstruction-based framework rather than the prediction-based framework employed by PICNN. This method showed effectiveness in disentangling single and multiple nozzle operating conditions.

PICNN. The original PICNN model extends a 1D convolutional neural network (1DCNN) to predict the power generated at the next time step based on a sequence of hydraulic and electrical data. Furthermore, it employs another 1DCNN module to infer the nozzle opening. This inferred nozzle opening is then used to compute the blade velocity based on the Pelton geometry and is incorporated as an augmented feature in the second CNN network.

Table 3: Performance comparison between vanilla CNN with nozzle information and PICNNv3 for hydropower prediction.

	Vanilla CNN		PICNNv3	
	2020 (train)	2022 (test)	2020 (train)	2022 (test)
MAE	0.709	1.149	0.447	1.218
RMSE	0.922	1.469	0.590	1.953

PICNNv3. PICNNv3 differs from PICNN as it has access to the actual nozzle opening information. It also contains two CNN modules, where the second module receives the output of the first CNN as its input. The first CNN processes flow rate, nozzle velocity, and nozzle opening positions, inferring the velocity ratio, an essential variable in the theoretical Pelton runner blade efficiency. The second CNN is designed to learn the friction coefficient μ, capturing its temporal evolution alongside other flow rate and pressure measurements. By employing these two CNNs, the model estimates the plant's total power output Pw, total based on the extended efficiency equation for Pelton wheels with multiple nozzles. PICNNv3 provides a unique opportunity compared to the vanilla CNN, in that it can infer μ as the latent variable and utilizes it to decouple the degradation from operational dynamics. Figure 5 shows that when inspecting the difference between the predicted power in the latent space (right), the gradient is much smoother compared to directly comparing the difference in predicted power using PICNN without inference of the friction coefficient μ (left). From the right figure in Figure 5, it can be observed that there is a correlation between the magnitude of μ and the degradation (difference in the predicted power). Specifically, the higher magnitude of the friction coefficient μ is associated with higher degradation. This relationship does not hold for low upper opening regions because nozzle dynamics are not stable. In addition, the results summarized in Table 3 indicate notable differences in the performance metrics of the CNN and PICNNv3 models for hydro efficiency prediction in both the training and testing phases for the years 2020 and 2022. The PICNNv3 model consistently outperforms the vanilla CNN model in MAE and RMSE on the training data from 2020. This suggests that incorporating physics-informed constraints into the neural network provides an effective prior knowledge based on physics that enhances prediction accuracy. Compared to the training set, the test error is higher from both the vanilla CNN and PICNNv3. This error stems not only from the generalization error but also from the mean efficiency degradation of the hydropower plant. The higher test error of PICNNv3 is likely more indicative of the efficiency degradation.

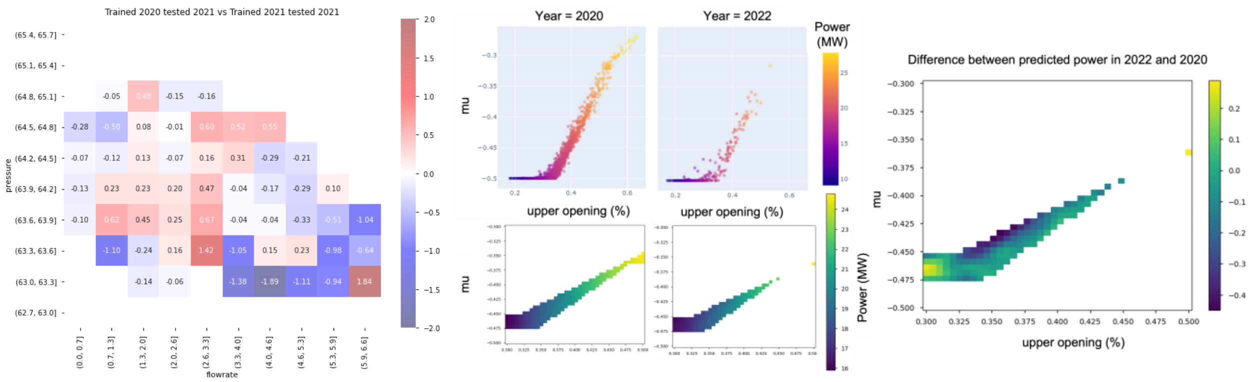

Figure 5: (Left) difference between predicted power in 2021 using the initial PICNN model without actual nozzle data trained with data from 2021 and 2020. Difference indicates efficiency degradation. (Middle) friction coefficient μ of PICNNv3 plotted vs. upper nozzle opening for year 2020 and 2022. (Right) Zoomed in qualitative analysis of mu inspection of PICNNv3. The color bar indicates predicted power in MW.

ENCDEC-LSTM. The ENCDEC-LSTM model is a sequence-to-sequence architecture utilizing LSTM networks for both encoding and decoding phases. The encoder LSTM captures temporal dependencies within the input sequence, passing its final hidden state to initialize the decoder LSTM. The decoder then reconstructs the output sequence through a linear transformation of its hidden states. The results of ENCDEC-LSTM summarized in Table 4 show that including nozzle information leads to a more robust model that is sensitive to the operational state of the hydro powerplant. Figure 6 illustrates that the latent space of the model with nozzle information forms two distinct clusters (left), whereas the latent space without nozzle data forms a single cluster (right). The clusters in the left figure correspond to the single (lighter cluster) and dual-nozzle (darker cluster) operation conditions. In addition, by adding nozzle information, the test error in 2022 is reduced significantly compared to the test error wo. nozzle information. This shows that a model with nozzle information is more generalized to unseen distributions. At the same time, the test error of the model with nozzle information is still higher than that of the train error, this error is likely indicative of the efficiency degradation (note that no physical information is included in this model and therefore generalization might be worse).

Table 4: Performance Metrics for the ENCDEC-LSTM Model with and without nozzle information.

	ENCDEC-LSTM wo. nozzle		ENCDEC-LSTM w. nozzle	
	2020 (train)	2022 (test)	2020 (train)	2022 (test)
MAE	0.433	2.184	0.395	0.906
RMSE	0.597	1.551	0.521	1.217

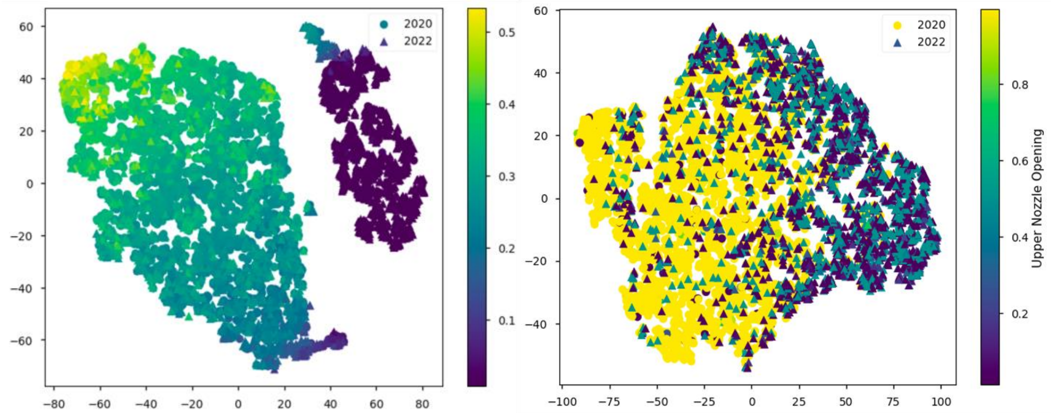

Figure 6: Latent space inspection of ENCDEC-LSTM. (Left) with nozzle information, (right) without nozzle information. Color indicates upper nozzle opening (%). The lower nozzle is always in operation, while the upper nozzle is activated when the machine is in dual nozzle operation.

Conclusion. The additional nozzle study has shown that the additional nozzle information can significantly impact the prediction and reconstruction performance. Introducing physical priors into the model serves as an effective regularization technique during training. PICNNv3 allows for the extraction of important system parameters that are indicative of efficiency degradation, such as the friction coefficient μ . Comparing model performance in the latent space of the friction coefficient μ is easier and qualitatively offers more insight into efficiency degradation. Moreover, reconstruction-based methods, as evidenced by their lower training and test error compared to prediction-based methods, have been demonstrated to be better suited for efficiency degradation estimation.

Future work. With the inclusion of nozzle information in the previously investigated models, we have finalized the efficiency case study. A comprehensive report detailing all findings from 2021-2023 has been summarized and is submitted alongside this document. The code and analysis in this case study have been shared and adopted by SBB. Future work from SBB's side could benefit from integrating physics-informed neural networks with reconstruction-based frameworks (PICNN + reconstruction) to potentially improve both accuracy and interpretability. Additionally, the exploration of the latent space μ as a degradation indicator presents a promising research direction. Efforts should also be directed towards generalizing both PICNN and reconstruction models to mixed machine operations, specifically, to implement a Graph Neural Networks (GNN) based reconstruction model with physics priors.

4. Outlook

Case study 3

In 2023, our research shows that the Spacetimeformer transformer model (STF) outperformed all other forecast methods currently used by the SBB energy trading division. A known limitation of time series transformers (such as STF) remains their restricted ability to implicitly break down trends and seasonality within the data. We therefore rely on manual trend decomposition which leads to reduced performance after unusual events when the decomposition is temporarily biased. It should be investigated why the trend decomposition is still necessary if we provide timetable-related covariates to the model that should contain some of that information. Furthermore, the STF model would benefit greatly from additional data. Therefore, the data set should be expanded to incorporate further expert knowledge such as the variation of energy consumption per gross ton-kilometer in regional (roughly 30 Wh/BTkm) and long-distance transport (roughly 20 Wh/BTkm) which is currently missing. Other data such as information on known unusual events such as national holidays could further improve the performance. In the upcoming period, attention should be primarily directed towards these details which could greatly improve performance. The detailed results of case study 3 will be summarized in a publication in early 2024.