# EuroSun2026

# Synthetic Operational Data for Surrogate Modelling of Seasonal Thermal Energy Storage


## Summary

Seasonal thermal energy storage (sTES) enables shifting intermittent renewable energy from summer to winter, when heating demand peaks. As district heating systems increasingly employ digital twins and model predictive control, fast and optimization-compatible surrogate models are needed to complement computationally intensive physics-based models. However, data-driven approaches require large and diverse training datasets, which are scarce for sTES systems that undergo only a limited number of annual charge–discharge cycles.

This work investigates the generation of synthetic operational datasets using a calibrated reduced-order sTES model. Historical measurements are analyzed to extract representative monthly charging and discharging regimes, which are used to construct multi-year weather-driven synthetic boundary conditions. Neural network surrogate models are trained and evaluated within both synthetic and measured datasets to assess internal consistency and learnability.

The proposed approach reproduces key seasonal operating characteristics and supports surrogate training. Ongoing work focuses on enhancing temporal realism and evaluating cross-domain transfer between synthetic and measured operation.



*Keywords: Seasonal thermal energy storage, machine learning, neural networks*

*Field of application for the findings: Thermal energy storage*

*Addressed audience: Research, District heating operators*


## 1. Introduction

Modern district heating networks are increasingly employing model predictive control (MPC) to optimally manage energy consumption. Modern district heating networks increasingly employ model predictive control (MPC) to improve operational efficiency and integrate renewable energy sources. Seasonal thermal energy storage (sTES) systems play a key role in enabling long-term energy shifting; however, their large thermal mass and strong stratification dynamics make accurate prediction challenging. Physics-based models provide a detailed and physically consistent representation of sTES behavior and can serve as high-fidelity proxies for real system performance. Nevertheless, such models often involve hundreds to thousands of state variables, leading to high computational cost in optimization frameworks. In addition, they require expert calibration and can be sensitive to changes in system configuration and operating conditions. Data-driven models offer an attractive alternative due to their reduced computational complexity and ability to re-train on new data. However, their performance depends strongly on the availability of representative training datasets. For seasonal storage systems, capturing the full range of operating regimes including charging, discharging, and transitional behavior across multiple seasons requires large volumes of structured data.

This work investigates the generation of statistically representative synthetic boundary condition datasets for sTES systems and evaluates their suitability for surrogate model training.

## 2. Methods

### *2.1 Reference Physics-Based Model*

A validated Modelica implementation of a stratified seasonal pit thermal energy storage (PTES) system based on the model by Bayer et al. (2025) was used to simulate internal tank temperatures. The model represents vertical stratification using a multi-layer discretization and includes diffuser-level mass flow inputs. Model performance was assessed against measured monitoring data from the Dronninglund PTES (Sifnaios et al., 2023), yielding an average RMSE of 2.7 °C across five storage layers, indicating acceptable predictive accuracy for surrogate training purposes. The resulting temperature time series are plotted in Figure 1.

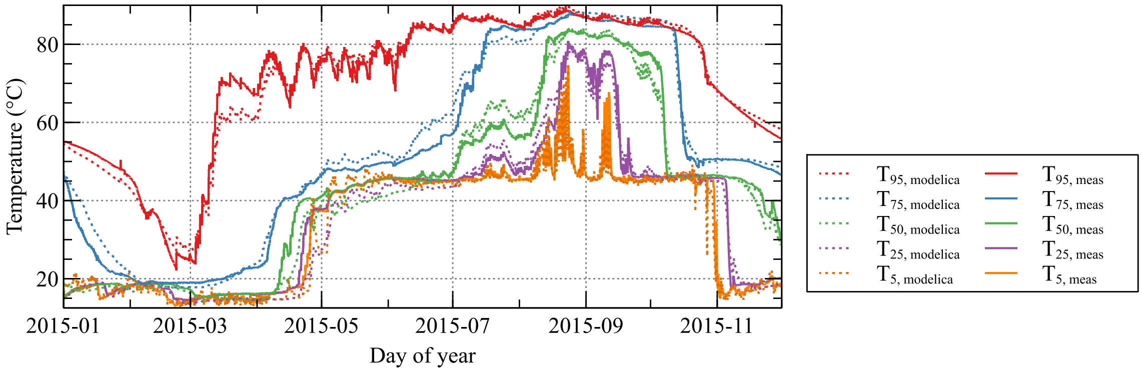


**Figure 1: Comparison of measured ("meas") and modelled ("modelica") stratified layer temperatures**

### *2.2 Synthetic Boundary Condition Generation*

Synthetic boundary conditions were generated using a monthly mode-based clustering approach, also derived from historical operational data from the Dronninglund PTES (Sifnaios et al., 2023). Measured charging and discharging periods were first identified based on net thermal power flow. The dataset was then separated into charging and discharging subsets. For each month, KMeans clustering was applied independently to charging and discharging data using inlet mass flows at the lower, middle, and upper diffusers, outlet mass flows, and inlet temperatures. For each cluster, representative operational modes were defined by the mean inlet and outlet mass flows, mean inlet temperatures, and their occurrence frequency within the month. This procedure resulted in a month- and regime-conditioned library of representative operational modes.

To construct a synthetic multi-year timeline, hourly weather data were obtained using Meteostat (Peters (2004)) for ambient temperature and PVGIS (European Commission JRC (2025)) for global irradiance. Charging periods in the synthetic dataset were determined based on irradiance thresholds, mimicking solar thermal charging of the real Dronninglund PTES. For each synthetic timestep, an operational mode corresponding to the current month and regime (charge or discharge) was sampled probabilistically according to its historical frequency and assigned to the boundary conditions. This approach preserves monthly seasonal operating statistics, relative frequency of operational regimes, and typical diffuser-level flow and temperature magnitudes, while intentionally removing temporal autocorrelation. Figure 2 illustrates the resulting seasonal charging structure for the upper diffuser, comparing synthetic and measured data. Charging boundary conditions for the upper diffuser are shown as hour-of-day versus day-of-year heat maps, with injected mass flow rate on the left and injection temperature on the right for synthetic data (top) and measured monitoring data (bottom).

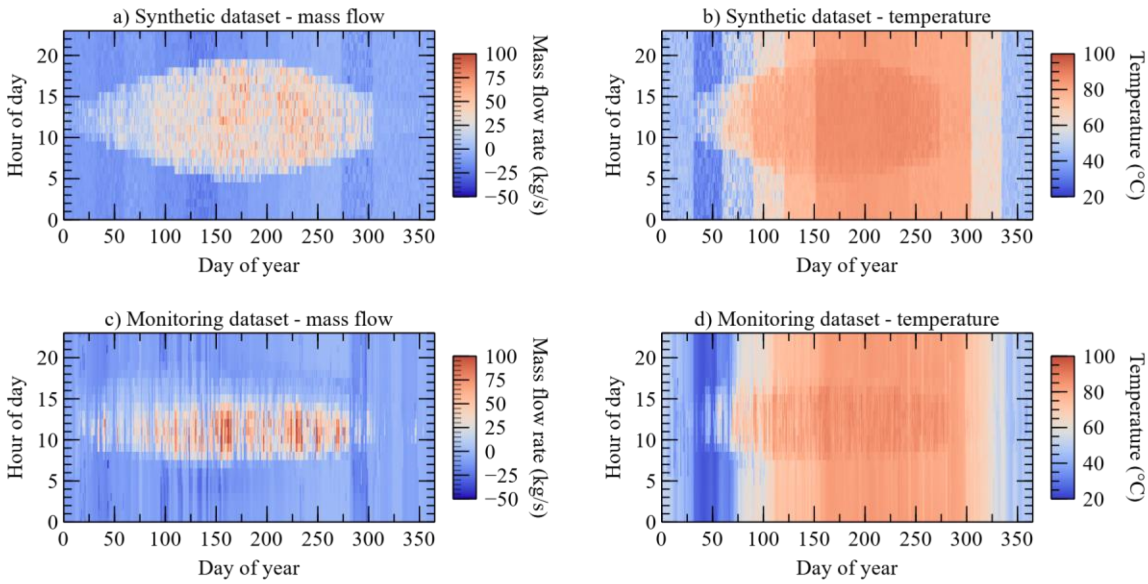


**Figure 2: Synthetic (top) and measured (bottom) charging profile for upper diffuser in one year. Injected mass flow rate (left) and injection temperature (right)**

### *2.3 Surrogate Model Training*

As an initial assessment of dataset suitability, neural network-based surrogate models were trained to predict layer temperatures in a timestep-by-timestep fashion. Input features included diffuser mass flows, inlet temperatures, ambient temperature, and lagged storage temperatures using 30-day and 60-day history windows.

Separate experiments were conducted using train-test splits within the synthetic dataset and within the measured dataset. The objective was to evaluate internal consistency and learnability of each dataset.

## 3. Results

The synthetic boundary condition dataset reproduces the principal seasonal charging structure observed in measured data, including concentrated summer charging periods, characteristic daily operating windows, and comparable temperature and mass flow magnitudes.

For five storage layer temperatures, the neural network achieved a synthetic-to-synthetic RMSE of 0.754 °C for a 30-day forecasting horizon and a measured-to-measured RMSE of 2.5 °C for the same forecasting horizon. The lower error in the synthetic case reflects the structured and statistically smoothed nature of the generated dataset. The measured dataset exhibits greater variability and dynamic complexity, resulting in higher prediction error. These results indicate that the synthetic dataset is internally consistent and supports further surrogate model training.

## 4. Conclusions and Outlook

A month and regime conditioned synthetic boundary condition generator for seasonal thermal energy storage systems was developed and statistically validated against measured operational data. The proposed clustering-based approach preserves seasonal operating structure and representative diffuser-level flow and temperature magnitudes while enabling the generation of multi-year datasets driven by weather inputs. Surrogate models trained and evaluated within the synthetic dataset demonstrate stable and internally consistent predictive performance, confirming the structural coherence of the generated data.

However, the synthetic generator intentionally removes temporal autocorrelation and state-dependent operational dynamics. While this simplification facilitates seasonal trends, it may limit representativeness of real short-term system behavior. Ongoing work therefore focuses on systematically investigating cross-domain transfer, specifically training surrogate models on synthetic data and evaluating performance on measured operational data. Additional efforts include introducing controlled stochastic perturbations within clusters and modeling temporal transitions between operational modes. These developments aim to improve realism while maintaining controllability of the synthetic dataset for surrogate model development and optimization studies.

## 5. References

Bayer, M., Meister, C., Schuetz, P., Villasmil, W., Walter, H. and Dahash, A., 2025. Development of a reduced-order dynamic model for large-scale seasonal thermal energy storage applications. Energy, 333, p.137379.

Sifnaios, I., Gauthier, G., Trier, D., Fan, J. and Jensen, A.R., 2023. Dronninglund water pit thermal energy storage dataset. Solar Energy, 251, pp.68-76.

Peters, T. (2024). *Meteostat Python Library*. Available at: https://meteostat.net

European Commission, Joint Research Centre (JRC). (2025). PVGIS – Photovoltaic Geographical Information System. Available at: https://re.jrc.ec.europa.eu/pvgis/.