Data Set Sampling and Its Implications on the Heliostat Calibration

Max Pargmann*[a], Moritz Leibauer[b], Daniel Maldonado Quinto[a], Vincent Nettelroth[b], Robert Pitz-Paal[a],
[a]German Aerospace Center (DLR), Lindner Höhe, 51147, Köln, NRW, Germany; [b]Synhelion Germany GmbH, Am Brainergy Park 1, 52428. Jülich, NRW, Germany

ABSTRACT

The ongoing global energy transition towards sustainable and climate-neutral power generation has led to the increasing adoption of concentrated solar tower power plants, relying on heliostats for precise solar tracking. Heliostat calibration, vital for maintaining accurate alignment, traditionally assumes a decrease in accuracy over time due to various factors. However, the impact of data set sampling on reported tracking accuracy has been overlooked. This paper utilizes a kNN (k-Nearest Neighbors) data set sampling approach to investigate data set distribution's impact on model accuracy. Results indicate that conventional time-dependent sampling can lead to an overestimation of reported accuracies. In contrast, the kNN sampling approach demonstrates a strong correlation between model performance and the proximity of test data to training data. Simulations reveal that reported accuracy scores are influenced by the similarity between training and test data sets. The study highlights the critical importance of considering data set distribution when interpreting accuracy scores. The proposed method improves tracking accuracy and offers a dependable metric for evaluating calibration results. It provides valuable insights to enhance heliostat calibration models, advancing precise solar tracking in concentrated solar tower power plants and supporting the global transition towards sustainable energy solutions.

Keywords: Heliostat Calibration, Data Set Sampling, Nearest Neighbors, Data Distribution, Machine Learning

1. INTRODUCTION

The ongoing global energy transition towards sustainable and climate-neutral power generation has prompted the increasing adoption of concentrated solar tower power plants. These systems play a key role in delivering climate-friendly electricity and direct heat for various industrial processes. At the heart of these power plants lie the heliostats - individual mirrors responsible for redirecting sunlight onto a central receiver. The ability of these heliostats to accurately track the sun is critical to achieving high plant efficiency and harnessing the full potential of solar energy. Heliostat calibration, the process of aligning and adjusting these mirrors, is a vital task in solar tower power plants to maintain precise solar tracking. Traditionally, it has been assumed that the accuracy of a heliostat diminishes over time due to various factors such as wear and tear, environmental influences, and slight manufacturing imperfections. Consequently, calibration is frequently performed at regular intervals to ensure continuous precision in heliostat tracking. However, the selection of data for training and testing calibration models and its impact on reported tracking accuracy has been overlooked.

This paper highlights the potential harm of using time-dependent data set sampling and training, which can lead to an overestimation of reported accuracies. Instead, we use a kNN (k-Nearest Neighbors) data set sampling, which was initially introduced in [1] and optimizes the data set distribution. Additionally, we use it as a metric that can assess reported accuracies. The main goal of this method is to improve tracking accuracy and offer a more dependable metric for evaluating calibration results.

2. THEORETICAL FRAMEWORK

Data set sampling

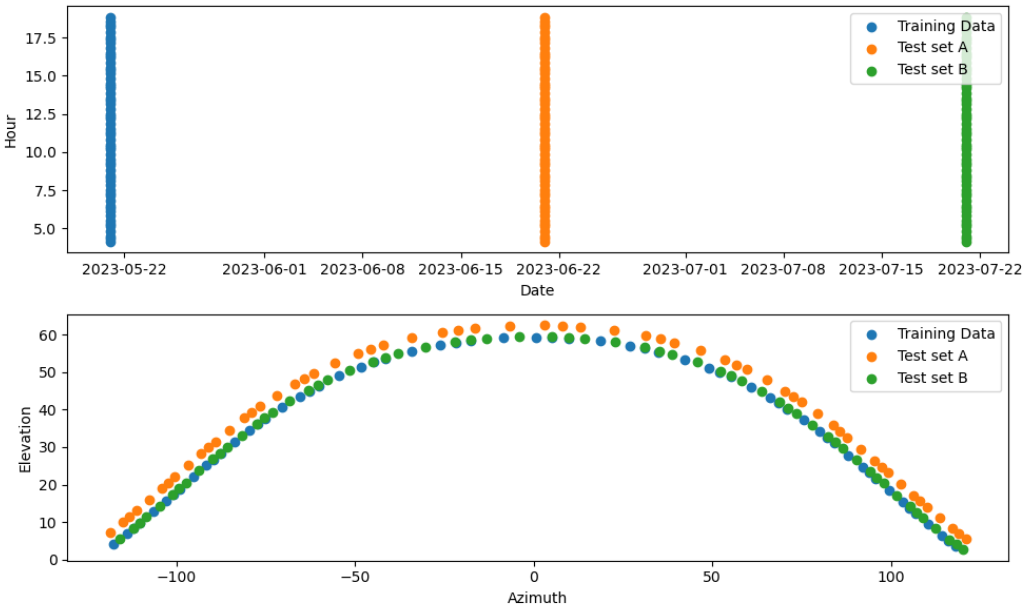

Figure 1. Different visualization of 3 data sets. The training data set is from May, the test data sets from June (A) and July (B). Even though July is further away in time, the sun positions are closer to those of the training dataset.

Data sets play a crucial role in influencing the accuracy of machine learning algorithms, whether it's for linear regression or deep learning [2]. There are two types of data sets: the training data set, which is used to optimize the model, and the test data set, which is used to evaluate the accuracy of the trained model.
The accuracy of a model depends on both the quality of the training data set and how different the test data set is from the data used for training.

The open-loop heliostat calibration, as typically executed in the majority of solar towers [3,4,5,6,7,8], constitutes a dataset-driven optimization targeting the primary function space orientation and aim point. Notably, this optimization is independent to the employed methodology and algorithmic approach. For optimal accuracy, the heliostat should be trained with the widest possible distribution of these input parameters and verified with test data sets including a different distribution.

Published heliostat calibration algorithms usually utilize temporal distribution in their data sets instead of spatially optimized distribution [7,9,10,11]. This may result in heliostat alignments that are nearly identical in both the training and test data sets. Consequently, the reported accuracies often cannot be observed in real-world everyday operation.

To illustrate this, let's consider a training dataset collected in May and two different test datasets from June and July, respectively. According to the conventional time-dependent sampling approach (see Fig. 1 upper panel), one would expect decreasing accuracies over time, with the reported accuracies in test set A being higher than in test set B. However, when we plot the data on the corresponding Azimuth Elevation Plot (see Fix 1, lower panel), a different pattern emerges. We observe that the dataset from July shows close proximity to the training data, while the dataset from June exhibits greater distances. This observation is further validated in subsequent chapters.

These two plots emphasize the significance of considering the range of interpolation and extrapolation required by the model to achieve more reliable accuracy assessment. The range can be quantified by measuring the distance of each sun position from the training data.

An Example on kNN Sampling

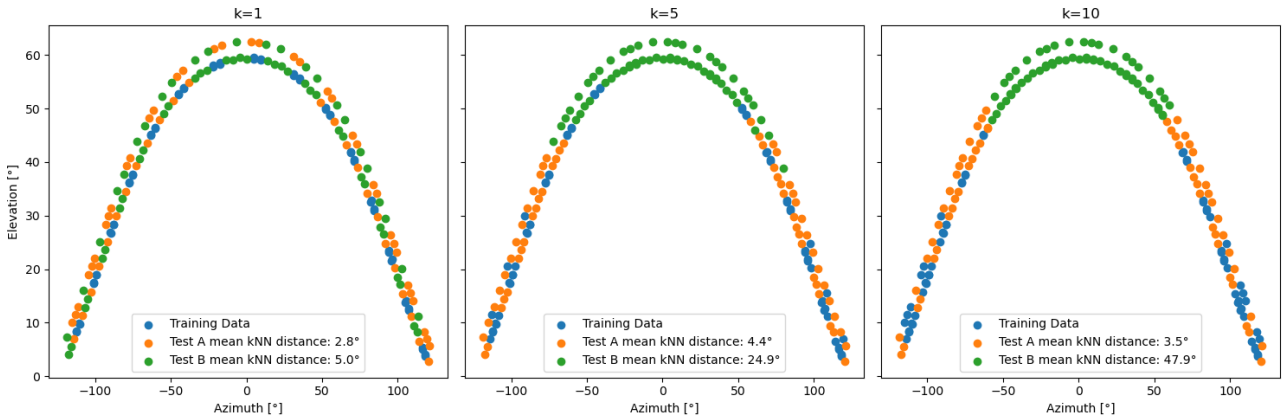

Figure 2. The figures show again the data set of May, June and July, but not separated in time but using the kNN metric with the values k=1, 5 and 10. The distances to the training data set increase with increasing k. Test data set A is always closer to the training than test data set B.

The kNN data set sampling method calculates the nearest neighbor distance between calibration points in the training data set and those in the test data set(s), using the included sun positions represented by Euler angles (Azimuth, Elevation) as relevancy scores. The value of k represents the number of nearest neighbors considered in the scoring process.

Continuing with the example shown in Fig. 1, using a 1-NN (k=1) metric, the mean distance between the training data set and test set A would be 4.2°, while the mean distance to test set B would be 2.8°. Clearly, test set B is more straightforward to predict, as its distances are closer to the training data.

This observation will be further validated in a subsequent chapter.

Furthermore, the kNN metric enables the creation of sub data sets with optimized data set distribution, as depicted in Fig. 2 using the same data set as in the previous chapter. The plot shows the k=1, 5, and 10 distance metrics. To create these sub datasets, the complete data set is initially sorted based on the kNN distance of each data point to every other data point in the set (excluding the trivial one with distance 0). The third with the smallest kNN distance is designated as the training data set. The remaining data points are then sorted by their kNN distance to the training data set. The sorted data set is split in half, with the data points having the highest distances to the training set becoming Test set B and the second highest distances forming Test set A.

As seen in Fig. 2, this data sampling strategy ensures that the distances in test set B are consistently larger than those in test set A. Depending on the number of nearest neighbors considered (k), the distribution of distances changes. While the distances in test set A remain relatively constant, the distances in test set B increase rapidly.

The kNN sampling method sorts the data based on sun distances, resulting in training with closely clustered data points. This approach makes extrapolation and interpolation more challenging. The test sets A and B have large distances, making alignment prediction in these sets even more difficult. As a result, the prediction task becomes more demanding

than it would be in daily operation. Therefore, the kNN data set sampling can be viewed as a conservative and reliable estimation of the heliostats' accuracy.

Moreover, the kNN distance can also be employed for creating the data set when calibrating heliostats in the field. Each time the calibration routine is done, the heliostat with the largest distance to the current sun position is selected for calibration. Additionally, a minimum kNN distance can be applied to ensure a balanced data distribution and achieve optimal data sets.

The Heliostat model

In this chapter, we present the heliostat model used to investigate the impact of data set distribution on the model's accuracy. For this purpose, we utilize two distinct models.

The first model is a static model, commonly employed in literature [3,4,5,8]. It follows the Rigid-Body Alignment Model approach, which captures the ideal kinematics of the heliostat along with constant deviation factors. This model represents the heliostat's alignment based on actuator positions and predicts the corresponding concentrator plane's normal orientation. We optimize seven parameters, including the position (3 parameters) and orientation of the first and second axes (2 parameters each).

On the other hand, the Dynamic Model acts as a surrogate model to approximate the heliostat's dynamic behavior and deviations in a realistic scenario. Dynamic effect can be caused by orientation dependent deformation or environmental influences such as wind. However, for this study, we do not model these influences precisely. Instead, we employ a function to introduce random perturbations to the heliostat's axes, thereby simulating an unknown heliostat behavior.

The Training

The training process involves both models, starting with data generation using the Dynamic Model. Subsequently, the Static Model is trained using the data generated by the Dynamic Model. The Static Model represents the ideal kinematics of the heliostat with the same 7 constant optimizable parameters as the Dynamic Model but without perturbations. The training process utilizes a defined loss function, the mean squared error, to measure the discrepancy between the predicted output of the Static Model and the data provided by the Dynamic Model. The objective is to minimize this loss through optimization, using the ADAM optimizer. During training, the model iteratively feeds the data, computes predictions, calculates the loss, and updates parameters to refine the model's performance. This process is repeated for a certain number of epochs to optimize the model's parameters. Once training is complete, the trained Static Model is evaluated on a separate test data set to assess its generalization ability to new data.

For the data set, we use the same one as in the previous sections. We employ both time-dependent sampling as well as 1, 5, and 10 nearest neighbor (NN) sampling methods.

3. RESULTS

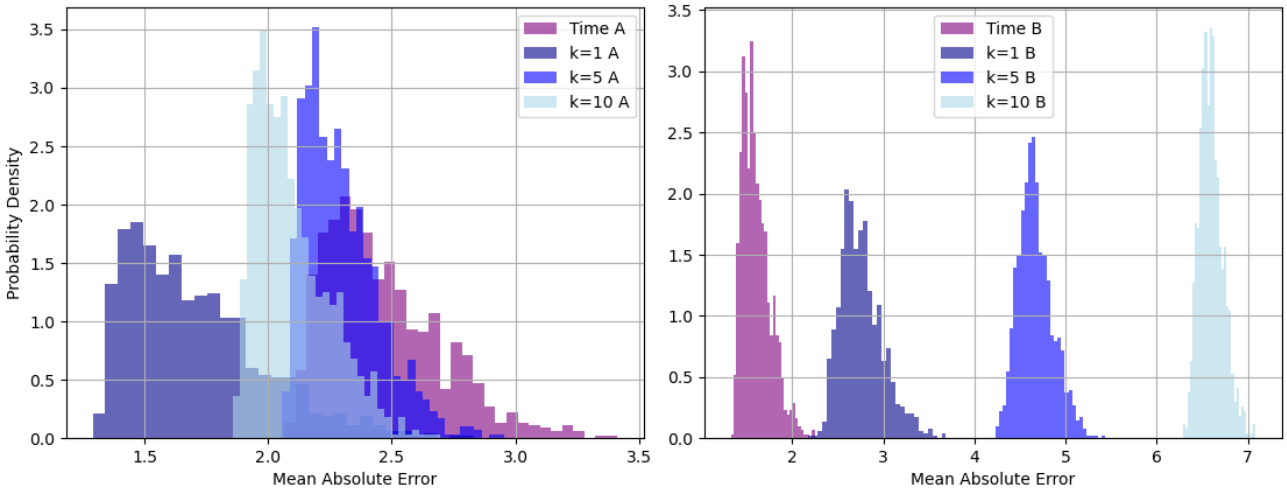

Figure 3. Predictions of the same model trained on different data set distributions. Each training had randomly selected starting conditions, resulting in different local optima and thus different prediction accuracies. The mean values of the resulting Gaussian curves clearly sort with the kNN distance.

Fig. 3 shows the deviation between the heliostat alignments in the test data sets and the prediction as the mean absolute error (MAE). The left panel shows the results from test set A, the right from test set B. The models where trained 100 times starting with different initializations, which results in scattering around a mean value in other words a gaussian normal distribution. All models show a prediction width of approximately 1mrad. Moreover the distribution becomes increasingly skewed in closer proximity to zero.

The accuracy of the heliostat models exhibits a strong correlation with the k-nearest neighbors (kNN) distance, indicating a clear relationship between the model's performance and the proximity of the test data to the training data.

To further validate our findings, we conducted an additional experiment where the model was trained exclusively on data from February 21 and then tested on data from March to June. The results, depicted in Figure 4, demonstrate a stepwise rightward shift of the Gaussian curve. While the predictions in March still remain highly accurate, the accuracy sharply declines in the following month. Subsequently, the accuracy reduces proportionally with the increase in kNN distance. The significant step observed in the results is likely attributable to the simplicity of our model.

This observation emphasizes the critical importance of considering the distance between the training and test data sets when interpreting reported accuracy scores. The model's accuracy is inherently linked to the similarity between the training and test data, underscoring the sensitivity of the model's predictions to data distribution.

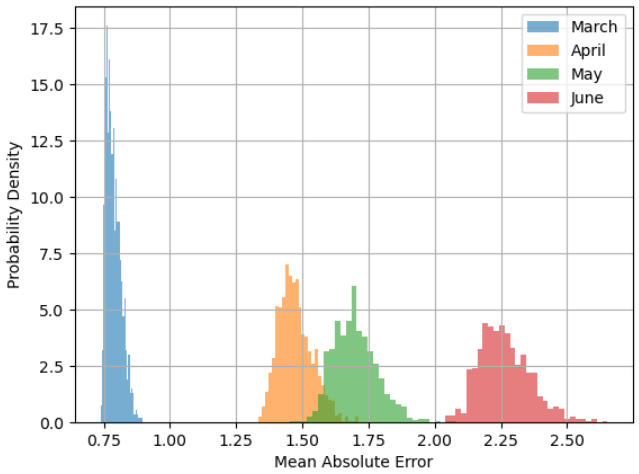

Figure 4. The model was trained using a dataset from February and tested with data from subsequent months. Despite not being explicitly time dependent, the model seems to show a temporal pattern. However, the error increases nearly linear as the kNN (k-Nearest Neighbors) distance grows.

4. DISCUSSION

Our findings have significant implications for the development and evaluation of heliostat models. The strong correlation between accuracy and kNN distance emphasizes the critical importance of carefully selecting training data to achieve optimal model performance. The observed patterns in the distribution of accuracies offer valuable insights into the models' behavior and present potential opportunities for enhancing their robustness under various data distributions. Notably, our study demonstrates that the commonly used time-dependent data set sampling may not accurately reflect the model's behavior in actual power plant operation.

We acknowledge the limitations of our study, including the constrained use of the Rigid-Body Alignment Model and the specific choice of the Dynamic Model's perturbation function. Future research endeavors may explore more advanced modeling techniques and dynamic perturbations to enhance the model's representation of diverse scenarios. Additionally, it is essential to examine the extent of time dependency in different heliostat types, as its influence may be small enough to be neglected in the calibration routine.

5. CONCLUSION

In this study, we focused on the calibration of heliostats in solar tower power plants to ensure precise solar tracking. Traditionally, it has been assumed that the accuracy of heliostats diminishes over time due to factors such as wear and tear, environmental influences, and slight manufacturing imperfections. Consequently, calibration is frequently performed at regular intervals to maintain precise solar tracking. However, the selection of data for training and testing calibration models and its impact on reported tracking accuracy has been overlooked.

Our investigation revealed the potential harm of using time-dependent data set sampling and training, which can lead to an overestimation of reported accuracies. To address this, we showed on a kNN (k-Nearest Neighbors) data set sampling approach, previously introduced in [1], which optimizes the data set distribution. Additionally, we used it as a metric to ensure the quality of reported accuracies. The main goal of our method was to improve tracking accuracy and offer a more dependable metric for evaluating calibration results.

The results demonstrated a significant influence of data set distribution on the accuracy of machine learning algorithms used in heliostat calibration. The accuracy of the models showed a strong correlation with the k-nearest neighbors (kNN) distance, highlighting the relationship between the model's performance and the proximity of the test data to the training data. Published heliostat calibration algorithms that utilize temporal distribution in their data sets may yield misleadingly high reported accuracies that do not align with real-world everyday operation.
Moreover, a fundamental question arises regarding the potential influence of time on the accuracy of heliostats or whether the apparent shift in sun positions was erroneously attributed to time dependence. Unfortunately, our simulated data does not allow us to directly investigate this aspect, as our simulated heliostat is not time-dependent. Nevertheless, our results demonstrate that even differences in sun positions can be misconstrued as time dependency.

Based on our findings, we recommend further explorations into heliostat calibration data point relevance scoring approaches for future research. Specifically, investigating dataset balancing techniques with k-nearest neighbor solar angle distances as relevance scores holds promise and could advance the field significantly.

In conclusion, the calibration of heliostats in solar tower power plants is critical for maintaining precise solar tracking, ensuring high plant efficiency, and effective utilization of solar energy. Our used kNN data set sampling approach and the introduced metric provide a more accurate and reliable method for evaluating calibration results.

REFERENCES

[1] M. Pargmann, M. Leibauer, V. Nettelroth, D. Maldonado Quinto, R. Pitz-Paal, It is Not About Time - A New Standard for Open-Loop Heliostat Calibration Methods, preprint, In Review, 2023. URL: https://www.researchsquare.com/article/rs-2898838/v1. doi:10.21203/rs.3.rs-2898838/v1.

[2] Olson, David L. "Data set balancing." Chinese Academy of Sciences Symposium on Data Mining and Knowledge Management. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[3] J. C. Sattler, M. Röger, P. Schwarzbözl, R. Buck, A. Macke, C. Raeder, J. Göttsche, Review of heliostat calibration and tracking control methods, Solar Energy 207 (2020) 110–132. URL: https://linkinghub.elsevier.com/retrieve/pii/S0038092X20306447. doi:10.1016/j.solener.2020.06.030.

[4] F. Gross, M. Geiger, R. Buck, A universal heliostat control system, in: AIP Conference Proceedings, volume 1850, AIP Publishing, 2017.

[5] K. W. Stone, Automatic heliostat track alignment method, 1986. US Patent 4,564,275.

[6] M. P. Sarr, A. Thiam, B. Dieng, ANFIS and ANN models to predict heliostat tracking errors, Heliyon 9 (2023) e12804. URL: https://www.sciencedirect.com/science/article/pii/S2405844023000117 doi:10.1016/j.heliyon.2023.e12804.

[7] M. Pargmann, D. Maldonado Quinto, P. Schwarzbözl, R. Pitz-Paal, High accuracy data-driven heliostat calibration and state prediction with pretrained deep neural networks, Solar Energy 218 (2021) 48-56. URL: https://www.sciencedirect.com/science/article/pii/S0038092X21000621. doi:10.1016/j.solener.2021.01.046.

[8] R. Baheti, P. Scott, Design of self-calibrating controllers for heliostats in a solar power plant, IEEE Transactions on Automatic Control 25 (1980) 1091–1097. URL: http://ieeexplore.ieee.org/document/1102527/. doi:10.1109/TAC.1980.1102527.

[9] J. Armendariz, C. Ortega-Estrada, F. Mar-Luna, E. Cesaretti, Dual-axis solar tracking controller based on fuzzy rules emulated networks and astronomical yearbook records, in: Proceedings of the world congress on engineering, volume 1, 2013, pp. 3–5.421

[10] S. Khalsa, C. Ho, C. Andraka, An automated method to correct heliostat tracking errors, Proceedings of SolarPACES (2011) 20–23.

[11] E. Smith, C. Ho, Field demonstration of an automated heliostat tracking correction method, Energy Procedia 49 (2014) 2201–2210