Pearl PV CKAN Data Challenge:
Artificially Faulty Solar PV Datasets for Development and Benchmarking of Fault Detection Algorithms

The development of fault detection algorithms is a very active field in solar PV research. To properly train, test, and validate these algorithms, it is necessary that researchers have access to a comprehensive database of realistic PV monitoring datasets, including a priori labelled faults. However, here we often face a chicken/egg situation, as the only way to determine if individual datapoints in PV monitoring datasets are faulty, is to apply some fault detection algorithm. Hence, sometimes researchers make use of datasets where typical PV system faults are introduced artificially. The mean by which these faults are introduced is however not trivial, furthermore, if individual research groups undertake this activity by themselves, the potential for collaboration and training, testing, and validating the developed algorithms on a broad selection of PV monitoring datasets is left unused.

Hence, we would like to suggest a data challenge within the framework of Pearl PV, to address this issue. For this challenge, we ask you for your participation! Participants are asked to submit:

- A PV monitoring dataset, spanning at least **ONE** full year of operation
- Time resolution: **at least** hourly
- Datasets contains between **2.5% and 7.5%** of faulty datapoints, that are artificially/manually introduced by manually or programmatically **changing the real, measured datapoints** to values that reflect a failure, and **chosen from these possible faults**[1]:
    - Module level faults
        - Soiling (10% power loss, range: 1%-20%)
        - Improperly installed module (5% power loss, range: 1%-100%)
        - Glass breakage (10% power loss, range: 10%-50%)
    - Inverter level faults
        - Inverter not operating (100% power loss)
        - Inverter error message (25% power loss, range: 25%-100%)
        - Inverter fan failure and overheating (20% power loss, range: 10%-100%)
- All faults should be labelled, including the type of fault (e.g. Module – Soiling; Inverter – Not Operating)
- As module faults propagate through the module string, dataset should include sufficient metadata to account for this
- Participants should shortly describe the approach they used to generate the faults (e.g. duration, severity, power loss applied)
- By contributing a dataset, participants gain access to the datasets of other participants

The challenge is to use the datasets to develop, train, test, and validate your fault detection algorithms. At the end of the challenge, we will summarize the results in a (joint) publication, containing a ranking and comparison of the fault detection algorithms.

Timeline:

- Participants express their interest (deadline: 23 dec 2021)
- Participants submit/upload their dataset (deadline: 15 jan 2022)
- Participants develop, train, test and validate their algorithms
- Participants submit results (deadline: 28 Feb 2022)

For questions, please contact Atse Louwen (atse.louwen@eurac.edu)

[1] These are the top-3 faults, for modules and inverters, in terms of missing production as identified in the project “Solar Bankability” (www.solarbankability.org)