One Year Submillisecond Fast Solar Database: Collection, Investigation, and Application

Yue Cao[1], *Member IEEE*, John A. Magerko III[2], *Member IEEE*, Rodrigo Serna[3], Shibin Qin[4], *Member IEEE*, Robert C. N. Pilawa-Podgurski[5], *Member IEEE*, Philip T. Krein[6], *Fellow IEEE*

***Abstract*—Big data is driving future renewable energy research and, in particular, has implications for fast dynamic control of power electronics. This paper discusses the collection of a full year of rooftop 5 kHz solar data. The database includes open-circuit voltage, short-circuit current, and maximum power information. Data reduction involved remapping of missing intervals, correction of unsynchronized metering, and regularization to go from more than eighteen months of raw data to a consistent full-year database. The results are made available at the full sampling rate. Data analysis shows how much downsampling can be employed without loss of useful information about variability. The downsampled databases are also made available. Given prior work that typically limits sampling rates to 1 Hz or takes fast samples over short intervals, the databases provided here support much more comprehensive examination of photovoltaic energy variability.**

***Index Terms*—photovoltaic data, renewable energy variability, solar, big data, data analysis, fast dynamics, open circuit voltage, short circuit current, maximum power point, energy conversion, power converters**

I. Introduction

Solar energy generators can exhibit rapid power changes. This unpredictability threatens the stability and reliability of the electric grid [1]. However, there are limited data available to quantify actual variability. The objective of this paper is to prepare and provide databases of measured point-source photovoltaic (PV) production, obtained at rates fast enough to capture the full range of operating dynamics. The paper seeks to process the data introduced in [2], [3], [4] into formats suitable for broad use, and to link these data for public access.

To understand PV variability, long-term real-life solar data are needed to identify relevant dynamic behaviors and time scales. Some of the time scales are obvious (diurnal and seasonal details of sun positions), but only fast data can support evaluation of all relevant time scales. The effort described here sets up a solar data collection system and applies a sampling rate intended to be much faster than expected dynamic behaviors. The resulting data are imperfect, so the paper describes the processing methods used to derive usable metrics, and the assumptions made as a complete year-long database is prepared.

A key focus of this work is to quantify PV variability at various time scales and ensure that all possible dynamics are captured. For example, a shadow moving across a 10 cm solar cell at 100 m/s (equivalent to 10 cm/ms) would seem to be an extreme situation; sampling theory constraints imply that sampling rates faster than 500 μs should capture even this case. Previous work has reported dynamic solar data with sampling rates ranging from 1 min [5] to 20 s [6]-[7] and topping out at about 1 Hz [8]. These data would not seem to be fast enough to capture the full possible dynamic range. Here, a sampling rate of 200 μs is used for short-circuit current. The objective is to sample at a rate at least twice as fast as the most rapid plausible dynamics.

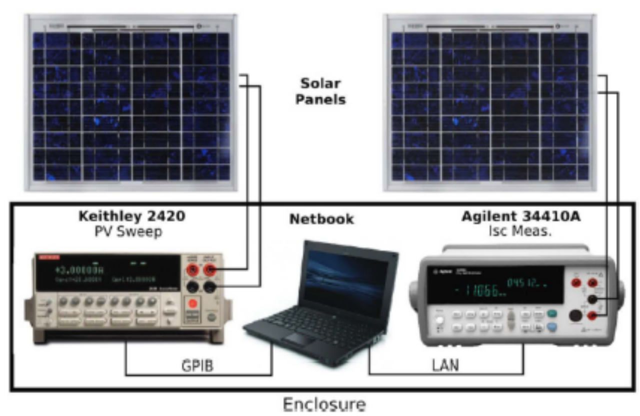

Fig. 1. Solar data acquisition hardware setup [4].

A few recent papers [2], [3], [4] have employed the database presented in this paper. They have explored, for example, how fast data can provide insight about power converter maximum power point tracking (MPPT) update rates, or how to optimize solar production for various weather conditions. Additional publications [9], [10] take advantage of these data for solar-tied energy efficient buildings and energy storage related research. Research utilizing external solar databases has also yielded advanced real-time estimation [11] and automatic detection [12] methods. These papers generally use only a few representative days of solar data, whereas the database offers a full year of results. Given emerging trends of data-driven energy applications and enabling data storage platforms, it is timely to publish comprehensive long-term (one year in this case), measured fast solar data. In addition, down-sampled versions have been prepared to meet a range of user needs.

[1]Yue Cao is with Oregon State University, Corvallis, OR, USA (email: yue.cao@oregonstate.edu). [2]John Magerko III is with EPRI, Knoxville, TN, USA. [3]Rodrigo Serna is with Citadel LLC, New York, NY, USA. [4]Shibin Qin is with Apple Inc., Cupertino, CA, USA. [5]Robert Pilawa-Podgurski is with University of California, Berkeley, CA, USA. [6]Philip Krein is with University of Illinois, Urbana-Champaign, IL, USA. This work was supported in part by the Grainger Center for Electric Machinery and Electromechanics at the University of Illinois, and in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award DE-AR0000217.

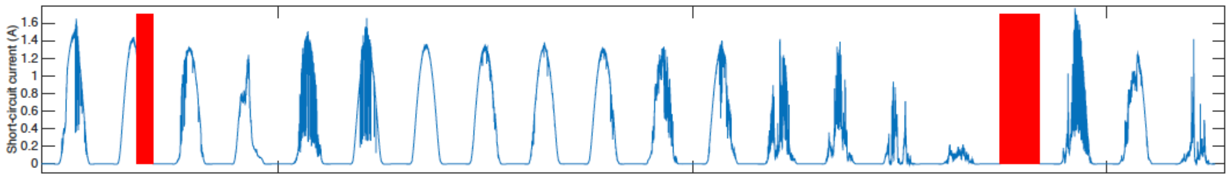

Fig. 2. A sample of consecutive days indicating missing data segments.

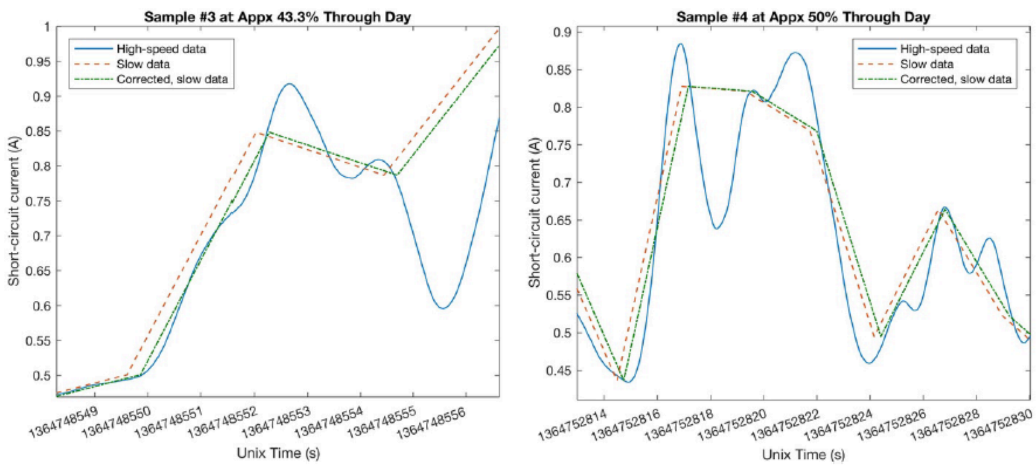

Fig. 3. Two sample windows of data alignment for synchronization.

Fast data collection over intervals of weeks or months offers its own challenges. Software glitches, file access errors, network dropouts and long-term timing synchronization errors lead to various imperfections and inconsistencies. In this particular project, raw data were gathered over an interval of about 18 months. The extra information supports a variety of strategies that yield a consistent full year end result. It is important to discuss the methods and assumptions that have been employed to produce the final databases.

II. Solar Data Acquisition

Raw PV data was collected, as described in [4], over a span of 18 months. Two identical rooftop-mounted PV panels, placed side by side, with nominal ratings of 20 W, were connected to two different meters. One meter was a Keithley 2420, set to perform a sweep across the panel current-voltage (I-V) curve every 2.5-3.9 s. The other was an Agilent 34410A that recorded short-circuit currents on sampling intervals of about 200 µs. The hardware setup is depicted in Fig. 1. The location was a rooftop at approximately north latitude 40.1108°, west longitude 88.2284°, the roof of Everitt Laboratory in Urbana, Illinois. This is about 14 km from the Bondville, Illinois site (site number 725315) in the U.S. national solar database [13]. The sweeps from the Keithley ("slow") meter enable the calculation of open-circuit voltage (V_{OC}), short-circuit current, and maximum power point (MPP) voltage, current, and power (V_{MPP}, I_{MPP}, P_{MPP}). The Agilent ("fast") data provide rapid short-circuit current readings (I_{SC}) that can be used to calculate high-frequency changes in the available power. The Keithley meter records slow short-circuit current data for verification of measurement accuracy against the Agilent meter and as a check of instrument synchronization.

Sampling at 5 kHz might seem excessive, but the objective is a sampling rate that can capture every possible transient without aliasing. Common sources of rapid transient creation might be fast-moving clouds, birds, or even airplanes casting flickering shadows on time scales much less than 1 s. Post processing will be employed to determine the likelihood that aliasing has been avoided. Atmospheric noise and local perturbations have the potential to induce fast dynamics.

III. Data Processing and Clean-up

Typical problems encountered in the raw data are intervals of local missing data, often linked to temporary equipment downtimes, system software updates, or similar system-side glitches. Here, missing data were detected by scanning the raw data for out-of-sequence time stamps. For example, Fig. 2 shows output data over several consecutive days with red boxes indicating missing segments.

A fundamental advantage for PV data is that there is a long daily interval with zero power. This means that missing data can be corrected by inserting data from a different year, on a complete day basis. Here, whole days are used even when partial data are available on existing days. This avoids introducing sudden changes or inconsistent weather pattern contrasts. Substitution was made according to the following rules:

1) Use a complete day if available from the same date in a different year – 4 days in this category;

2) If this is not possible, use a complete day from a date equidistant from the winter or summer solstice – 3 days in this category;

3) If neither is available, use a complete day with similar historic weather patterns and temperatures close to the original date. In this case, a scaling factor from the substitute is chosen to account for annual changes in solar irradiance if the problem day is mostly missing, or scaled to match the original data if valid partial data are present – 5 days in this category.

Notice that none of these rules involves interpolation or extrapolation. The final result is indeed a full year of actual measured data, with sequenced time stamps, preserving the data and all of its dynamics.

After missing data substitutions, data synchronization is required because the time stamps on slow and fast meters are not fully aligned. One challenge is incompatible file length with varying numbers of data points per file. For example, the fast short-circuit current data are recorded in segments of about one minute. The actual rates vary between 5000 Hz to 5250 Hz and total time stamps span between 56 and 60 s for each file. Therefore, synchronizing to a constant 5 kHz sampling rate causes time offset accumulation. To remedy this, the computer program that scans the data utilizes its master clock to increment with mean period T_{mean} between data samples as determined by

$$T_{mean} = \frac{t_{end} - t_{start}}{\#\ of\ samples\ in\ file} \tag{1}$$

$$max\left[\sum_{t=m}^{t=n-m} I_{SC_{slow}}(t) I_{SC_{fast}}(t - t_{offset})\right], -m < t_{offset} < m \tag{2}$$

Since it is expected that the short-circuit current will be the same in both the fast and slow data at a given instant, it is reasonable to adjust the slow meter time stamp to better match the time stamp of the fast meter. This is accomplished by maximizing the correlation between the two current measurements for each day. The timestamp offset for the slow meter will be the value of t_{offset} that maximizes (2), where n is the total number of points in a day, and m is the maximum sample offset to be considered. Two example windows of fast meter data, slow meter data, and timestamp-corrected slow meter data are shown in Fig. 3. These samples depict results after alignment of data from the two separate meters.

Following data substitution and time synchronization, the 5 kHz data for short-circuit current are in an immediately usable form. The slow I-V curve data require processing to obtain short-circuit current, open-circuit voltage, MPP current, MPP voltage, and MPP power values. Slow short-circuit current data consists of three data points near 0 V, indicated by the square symbols in Fig. 4. Open-circuit voltages are obtained from the sweeps that cross the voltage axis, shown as the triangle symbols. In Fig. 4, the MPP region is taken from among 100 points across the knee of the I-V curve for MPP voltage, current, and power calculations. However, as can be seen in Fig. 5, the measurements contain a combination of high-frequency fluctuations and measurement noise. Simply picking the raw point with the peak value tends to yield misleading and inaccurate MPP values. To alleviate this, a 4th-order polynomial fit is applied to each sweep for smoothing. Implementing a 4th-order fit instead of the 2nd-order polynomial used in [4] increases the regression coefficient from R^2 = 0.95 to R^2 = 0.995 for a typical MPP sweep. Rather than solving for the peak algebraically, it is computationally efficient to evaluate the polynomial function at 500 equally spaced points over the same range of voltages as the original MPP region and then select the maximum from this discretized set.

There are data instances when the polynomial keeps increasing or decreasing monotonically because the slow meter sweep has missed the MPP for some reason. In such cases, the local maximum, usually an endpoint, is chosen as the MPP. An example is provided in Fig. 6, which shows three consecutive MPP sweeps. In this case, the middle sweep has failed to span the peak power value. This failure is likely due to a sudden drop in irradiance following a previously increasing trend. In this example, the MPP power at the specific moment in time is taken to be the power at the left end of the middle curve.

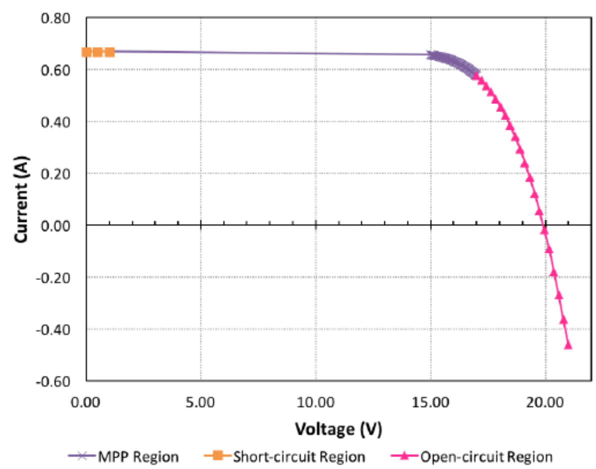

Fig. 4. I-V sweep showing short-circuit current, open-circuit voltage, and MPP regions.

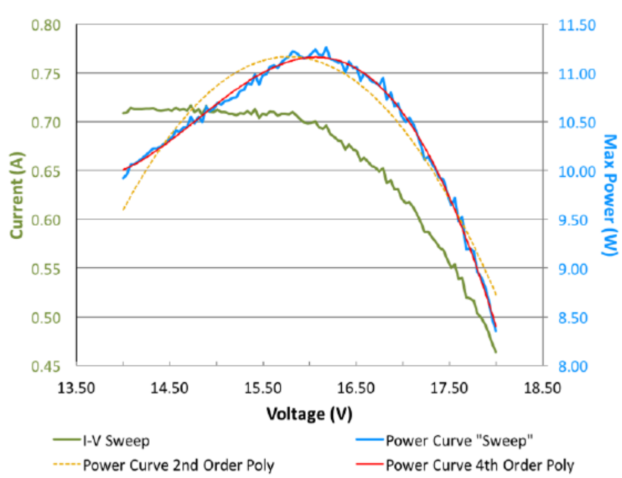

Fig. 5. I-V sweep for MPP and polynomial least squares fits showing max power curves.

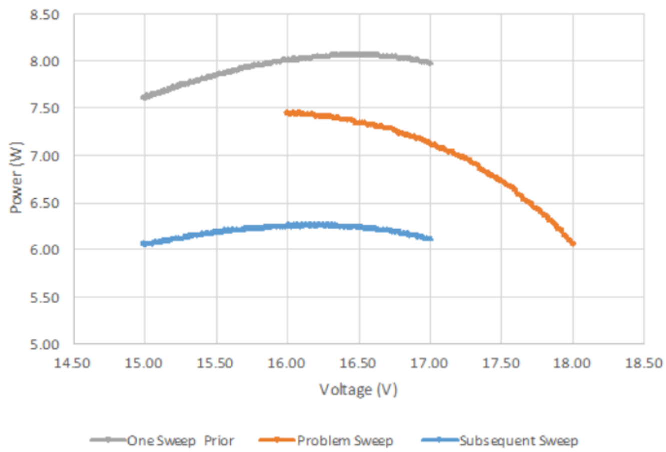

Fig. 6. Three consecutive MPP power curve sweeps with the middle sweep missing the MPP.

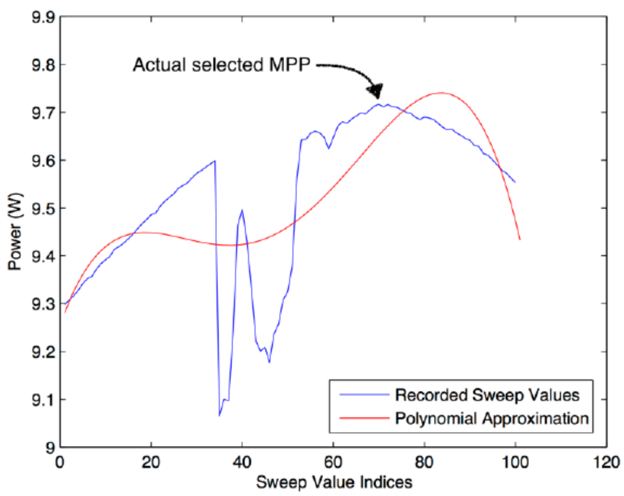

Fig. 7. Rapid transient during MPP sweep showing poor fit polynomial approximation.

Sometimes the fluctuations are so fast that within a single MPP sweep of the slow meter, the polynomial approximation generates multiple peaks. Fig. 7 shows such a scenario, where the polynomial has produced a poor approximation of the raw data. In cases like this one, the maximum is taken directly from the raw sweep data rather than from the polynomial fit.

IV. What Sampling Rate Is Fast Enough?

The discussion of this section starts with evaluation of whether a 5 kHz sampling rate is sufficiently fast to capture all meaningful PV system dynamics, followed by investigation of down-sampling rates that would preserve dynamic information in a compressed database. To determine whether the sampling rate is sufficient, one metric is to show that the data are smooth, do not contain discontinuities, and in general can be represented sufficiently at a bandwidth of about 1 kHz. To explore this, one sample day with an especially high variability is closely investigated in Fig. 8. On this day, the solar panels experienced intermittent cloud coverage, resulting in rapid ramp rates as seen in the top image of Fig. 8. The subsequent images depict subsets of data to demonstrate that during the most variable moments, all dynamics are captured. The bottom image, selected as an interval of extremely fast variation, still contains about 100,000 data points. There are many points even during the quickest variations, and the data provide a smooth re-creation of the actual irradiance change.

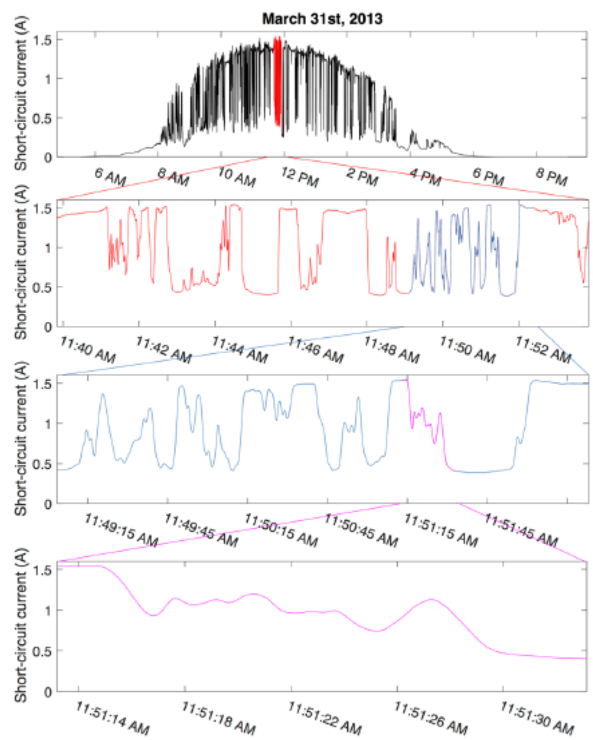

Fig. 8. A sample day containing rapid transients and subsequent zoom-in windows.

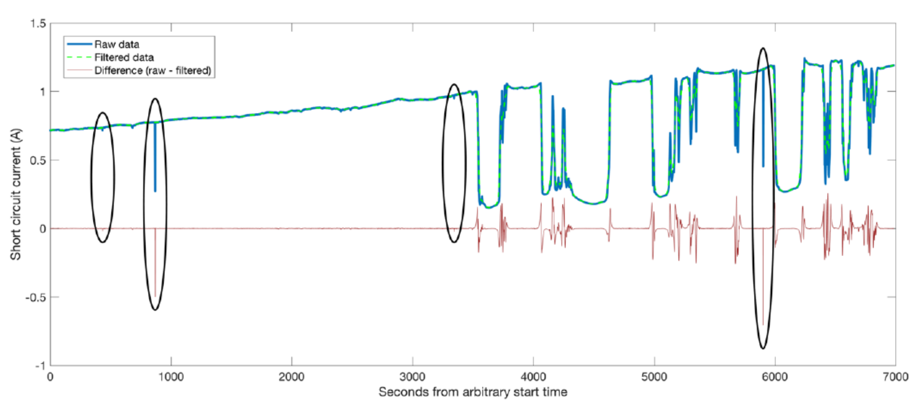

Fig. 9. Sample solar data containing multiple rapid dips in power output (circled).

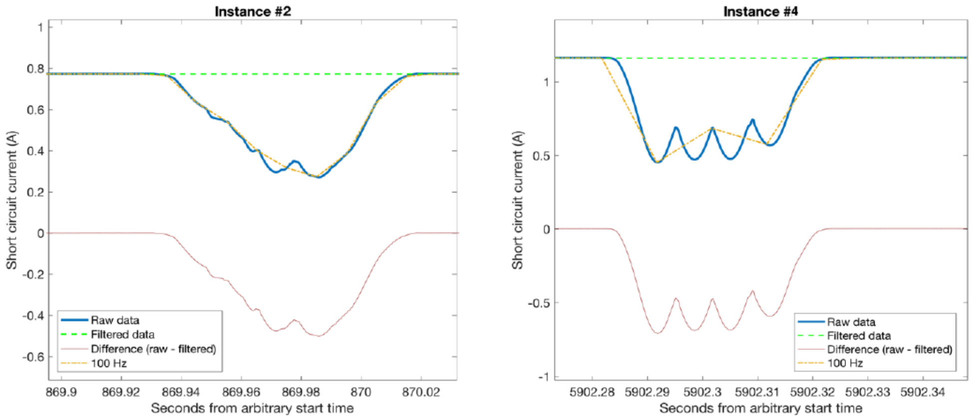

Fig. 10. Zoomed-in views of two most significant circled instances.

As the database was explored for the fastest dynamics, the most likely explanations for infrequent rapid changes were flickering shadows of birds or airplanes. Large birds such as Canada geese are active in the area, and their fast-moving shadows have been observed to cross the panels. To help identify narrow spikes, a high-pass Butterworth filter was applied to the raw short-circuit current data, as shown in Fig. 9. Four power dips of interest are circled there. Zooming in on these circled regions, as in Fig. 10, reveals fine details of these transients. Instance #2 on the left, for example, lasts about 70 ms, and therefore contains about 350 samples. Instance #4 on the right most likely captures the details as a fast-moving small shadow crosses the four columns of solar cells on the panel. This transient lasts about 35 ms and includes about 175 samples. The inferred motion is about 10 cm in 6 ms, or about 17 m/s. Even these extreme dynamics have bandwidth below 1 kHz. Fig. 10 shows what will happen with 100 Hz downsampling for these two fast changes. The 70 ms interval is covered well. The 35 ms interval still yields a PV energy source estimate very close to the actual value, although it cannot track the fast column-by-column changes in the raw data.

The discussion so far provides evidence that a sampling rate of 5 kHz captures realistic PV panel dynamics up to the fastest changes. The results offer, for the first time, a basis for determining suitable PV sampling rates. Can the database be downsampled to achieve compression without loss of information? Signal content was evaluated with fast Fourier transforms (FFT) for various test days to confirm the frequency content. Results for three days are shown in Fig. 11. In these traces, small peaks can be seen at 60 Hz and at 180 Hz, indicating that power line interference is detectable. Otherwise, the FFT results suggest that measurement accuracy and other factors yield a noise floor that is crossed very close to 100 Hz. This is an indication that 200 µs sampling is indeed capturing all dynamics up to the resolution limits of the meters. The FFT results presented suggest that, other than the grid harmonic at 180 Hz, frequency content above 100 Hz is below one part in 10 million. At these scales, frequency content is effectively negligible, limited by measurement accuracy.

Downsampled data imposes a certain risk of error. In this case, the objective is to represent the amount of solar energy that can be captured by a PV system given a sufficiently fast MPPT control. To determine such missed energy, and based on the rate results above, an MPP update rate of 200 µs is taken as a baseline, equivalent to an unlimited continuous update control. A hypothetical MPPT process is then applied to the data. The amount of energy missed at each time step is summed according to

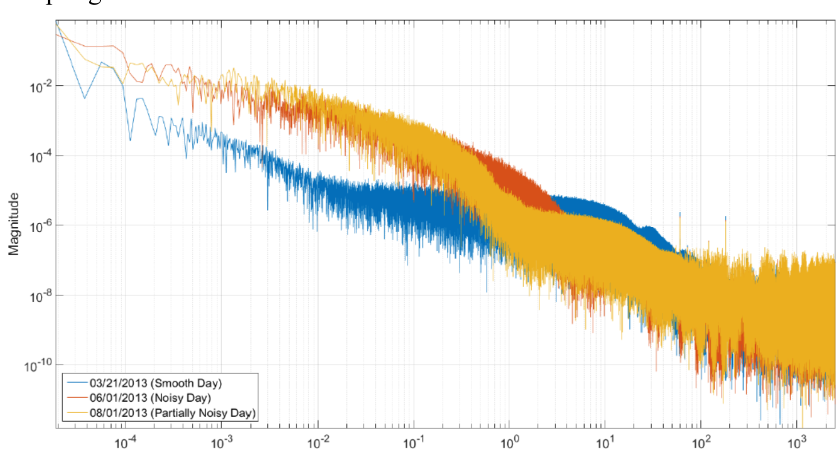

Fig. 11. Full-day FFT of 5 kHz short-circuit current for three different days.

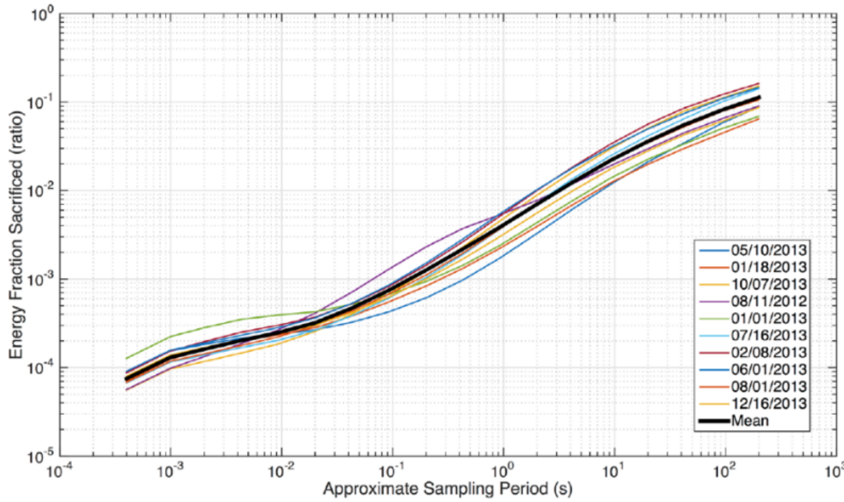

Fig. 12. Modeled energy sacrifice of ten 10-day samples with varying MPPT update rates and mean in bolded black. Entries in the legend represent the last day in each 10-day series.

$$\boxed{Energy\ Missed = \sum_{m=1}^{S_T/S_U} V_{MPP_k} \cdot \sum_{n=1}^{S_U} \left| \left(I_{SC_n} \cdot I_{SF} \right) - I_{MPP_m} \right|} \tag{3}$$

where I_{SF} is a scaling factor between I_{MPP} and I_{SC} [3], S_T is the total number of samples in the selected data segment, and S_U is the number of samples per MPPT control update. This is normalized to the result at 5 kHz to represent the missed energy fraction.

Fig. 12 summarizes the results of this calculation for ten different 10-day periods. The mean is shown in bold. For example, if the MPPT control updates at 100 Hz, about 1 part in 4000 of the available energy will not be captured compared to the baseline case. This is about 63 mW for a 250 W panel. It is likely to be an overestimate, since the noise floor suggests that data resolution is no better than one part in 10,000. Indeed, uncertainty in power converter ripple makes it unlikely that *any* MPPT control can extract as much as 99% of the available incident radiation. A 1 Hz update rate sacrifices an average of about 1 part in 250 in terms of energy production, about 16 times as much as a 100 Hz update rate. An MPPT controlled converter updating at 100 Hz should be able to capture an extra 1 W on average from a 250 W panel compared to a 1 Hz update rate. Fig. 12 can be used to analyze, in general terms, the economic opportunity cost associated with increased update rates. An update rate of 1 Hz yields a mean energy fraction sacrifice of 0.41%, and 100 Hz update yields a sacrifice of 0.025%. The extra effort of a 100 Hz update rate can be valued at about 0.38% of the value of panel energy production.

A general result from the FFT results and from Fig. 12 is that downsampling to 100 Hz will capture almost all the dynamics of interest. It should be possible to use a downsampled 100 Hz version of the database (a compression of 50:1) to analyze and compare control strategies for energy capture and grid interfacing without introducing uncertainty higher than about ±0.02%.

V. Data Archives

The process of gathering raw PV data at time intervals of about 200 μs and post-processing to recover a full year of measured data at this rate have been discussed. A further day-by-day downsampled version, converted to 10 ms intervals, has been discussed. The end results are two databases, available for public use:

1. The original, raw, data, including both fast (short circuit, 200 μs) and slow (sweep, 2.5-3.9 s) information for 18 months. These show intervals of missing points, but are provided to allow potential users to reproduce any new work.
2. Data provided at consistent 10 ms (short circuit) and 2.5-3.9 s (sweep) time bases for 365 days, after being corrected and recovered. The dataset includes short-circuit current, open-circuit voltage, and MPP voltage, current, and power extracted.

The database is available as open access at DOI: 10.21227/5s8f-ha92, hosted by IEEE DataPort. The DataPort page includes detailed download and usage instructions. Sample screenshots of the data archive, after download and extraction, are shown in Fig. 13 and Fig. 14.

VI. Conclusion

A point-based PV dataset was gathered at a sampling rate of approximately 5 kHz. It has been confirmed that this rate is fast enough to capture even the most extreme dynamics of a PV system. The raw data have been post processed in a manner that yields a full year of measured data. It has been shown that downsampling to rates as slow as 100 Hz can be performed with minimal loss of fidelity. The measured data and downsampled 100 Hz database are made available for public use.

```
1100.txt - Notepad
File  Edit  Format  View  Help
#2013/03/08 11:00:00.00  (# = Original, ! = Substituted)
Date/time recording started:
2013/03/08 11:00:00.00
Time,         HS_Isc [A]
11:00:00.00,  1.329948E+00
11:00:00.01,  1.330081E+00
11:00:00.02,  1.330182E+00
11:00:00.03,  1.330255E+00
11:00:00.04,  1.330242E+00
11:00:00.05,  1.330221E+00
11:00:00.06,  1.330208E+00
11:00:00.07,  1.330234E+00
11:00:00.08,  1.330276E+00
11:00:00.09,  1.330276E+00
11:00:00.10,  1.330255E+00
11:00:00.11,  1.330153E+00
11:00:00.12,  1.330055E+00
11:00:00.13,  1.329961E+00
11:00:00.14,  1.329974E+00
11:00:00.15,  1.329940E+00
11:00:00.16,  1.329813E+00
```

Fig. 13. Screenshot of the downsampled short circuit current data archive text file from an example date and time.

```
SWEEP_DATA.txt - Notepad
File  Edit  Format  View  Help
#2013/03/08 04:00:00.00  (# = Original, ! = Substituted)
Date/time recording started:
2013/03/08 04:00:00.00
Time,        Pmpp [W],    Vmpp [V],    Impp [A],    Isc [A],     Voc [V],
11:00:01.23, 2.139090E+01, 1.723000E+01, 1.241492E+00, 1.322571E+00, 1.700000E+01
11:00:03.64, 2.137304E+01, 1.724200E+01, 1.239592E+00, 1.324139E+00, 1.700000E+01
11:00:06.04, 2.136026E+01, 1.723000E+01, 1.239714E+00, 1.322258E+00, 1.700000E+01
11:00:08.44, 2.139274E+01, 1.723400E+01, 1.241310E+00, 1.324139E+00, 1.700000E+01
11:00:10.84, 2.139901E+01, 1.723800E+01, 1.241386E+00, 1.324139E+00, 1.700000E+01
11:00:13.24, 2.136298E+01, 1.723600E+01, 1.239439E+00, 1.323196E+00, 1.700000E+01
11:00:15.64, 2.135393E+01, 1.724400E+01, 1.238340E+00, 1.321789E+00, 1.700000E+01
11:00:18.04, 2.136422E+01, 1.724200E+01, 1.239080E+00, 1.322102E+00, 1.700000E+01
11:00:20.44, 2.133783E+01, 1.725200E+01, 1.236832E+00, 1.321632E+00, 1.700000E+01
11:00:22.84, 2.136962E+01, 1.726800E+01, 1.237527E+00, 1.319286E+00, 1.700000E+01
11:00:25.25, 2.137963E+01, 1.728400E+01, 1.236961E+00, 1.319599E+00, 1.700000E+01
11:00:27.66, 2.139935E+01, 1.728000E+01, 1.238388E+00, 1.322102E+00, 1.700000E+01
11:00:30.07, 2.141586E+01, 1.726800E+01, 1.240205E+00, 1.321789E+00, 1.700000E+01
11:00:32.49, 2.140238E+01, 1.727400E+01, 1.238994E+00, 1.321789E+00, 1.700000E+01
```

Fig. 14. Screenshot of the sweep data archive text file from an example date and time.

REFERENCES

[1] S. Shivashankar, S. Mekhilef, H. Mokhlis, and M. Karimi, "Mitigating methods of power fluctuation of photovoltaic (PV) sources – A review," *Renew. Sustain. Energy Rev.*, vol. 59, pp. 1170–1184, Jun. 2016.

[2] S. Qin, R. Serna, and R.C.N. Pilawa-Podgurski, "A data-driven approach to the design of photovoltaic maximum power point tracking techniques using field transient data," in *Proc. IEEE Energy Conversion Congress and Expo (ECCE)*, 2015, pp. 6629-6634.

[3] J. A. Magerko, Y. Cao, and P. T. Krein, "Quantifying photovoltaic fluctuation with 5 kHz data: implications for energy loss via maximum power point trackers," in *Proc. IEEE Power and Energy Conference at Illinois (PECI)*, 2016.

[4] R. J. Serna, B. J. Pierquet, J. Santiago, and R. C. N. Pilawa-Podgurski, "Field measurements of transient effects in photovoltaic panels and its importance in the design of maximum power point trackers," in *Proc. IEEE Applied Power Electronics Conf.*, 2013, pp. 3005-3010.

[5] R. A. Gansler, S. A. Klein, and W. A. Beckman, "Investigation of minute solar radiation data," *Sol. Energy*, vol. 55, no. 1, pp. 21–27, 1995.

[6] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker, and T. Hoff, "Parameterization of site-specific short-term irradiance variability," *Sol. Energy*, vol. 85, no. 7, pp. 1343–1353, Jul. 2011.

[7] T. E. Hoff and R. Perez, "Quantifying PV power output variability," *Sol. Energy*, vol. 84, no. 10, pp. 1782– 1793, Oct. 2010.

[8] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, and C. O. Pedersen, "EnergyPlus: New capabilities in a whole-building energy simulation program," *Seventh Int. IBPSA Conf.*, pp. 51–58, 2001.

[9] J. A. Magerko, "Alternative methods for mitigating natural photovoltaic variability: Dynamic HVAC load compensation and curtailed PV power," M.S. thesis, Department of Electrical and Computer Engineering, Univ. of Illinois, Urbana, IL, 2016.

[10] Y. Cao, J. A. Magerko, T. Navidi, and P. T. Krein, "Power electronics implementation of dynamic thermal inertia to offset stochastic solar resources in low-energy buildings," *IEEE Journal Emerging and Selected Topics in Power Electronics*, vol. 4, no. 4, pp. 1430-1441, 2016.

[11] K. Kamono and Y. Ueda, "Real-time estimation of areal photovoltaic power using weather and power flow data," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 754-762, 2018.

[12] B. H. Ellis, M. Deceglie, and A. Jain, "Automatic Detection of Clear-Sky Periods From Irradiance Data," *IEEE Journal of Photovoltaics*, vol. 9, no. 4, pp. 998-1005, 2019.

[13] S. Wilcox et al., "National solar radiation database 1991-2005 update: user's manual," National Renewable Energy Laboratory, Golden, Colorado, Rep. no. NREL/TP-581-41364, April 2007.