Benchmark Dataset Development and Applications

4-lab project team, Feb 2020, at NREL

LBNL
Tianzhen Hong
thong@lbl.gov

NREL
Lieko Earle
lieko.earle@nrel.gov

ORNL
Piljae Im
imp1@ornl.gov

PNNL
Vikas Chandan
vikas.chandan@pnnl.gov
Project Summary

Timeline:
Start date: 10/1/2019
Planned end date: 9/30/2022

Key Milestones:
1. Data portal launched: 3/30/2020
2. 6 datasets curated and published: 3/31/2021
3. 2 case studies completed: 6/30/2021

Budget:
Total Project $ to Date:
• DOE: $1.98M
• Cost Share: N/A

Total Project $:
• DOE: $3.75M
• Cost Share: N/A

Key Partners:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LBNL</td>
<td>NREL</td>
</tr>
<tr>
<td>ORNL</td>
<td>PNNL</td>
</tr>
</tbody>
</table>

Project Outcome:
The focus of this project is to collect and curate a handful of high-resolution building datasets that have broad applicability to address highest-impact use cases. Main outcomes include:

● 12 high-quality and well-curated building datasets
● Enhanced data tools for metadata representation
● Demonstration of applications of the curated datasets
● A workflow to semi-automate the data curation process for scale up and industry adoption
Challenges

Problem Definition:
Reducing energy use in buildings and optimizing building operations require access to a diverse and integrated set of data. However, it is resource-intensive and often hard to find datasets with adequate data coverage, good data quality, and clear documentation to support research on energy efficiency and demand flexibility technologies for grid-interactive efficient buildings (GEBs).

Therefore, there is a strong need to:

• Collect, curate, and make publicly available high-resolution and high quality measured data from residential and commercial buildings that have broad applicability to a variety of high-impact use cases.
• Determine the level of data resolution required for most effectively optimizing building operations through advances in data analytics and control technologies.
• Provide a common, high-quality benchmark against which competing algorithms can be fairly
Impacts

• The curated datasets are made open access through a data portal to support data-driven research for grid-interactive efficient buildings (GEB) to save energy and improve operations and controls.

• The datasets can be employed for load forecasting and baselining, virtual sensing, building energy modeling, building performance benchmarking (at the whole building and system levels), and non-intrusive load monitoring.

• New and enhanced data tools to represent metadata supporting FAIR (Findable, Accessible, Interoperable and Reusable) principles.

• The project develops and demonstrates a semi-automatic workflow to curate datasets, which enables data contribution from other parties after the project is completed.

• A set of use cases demonstrating how users can utilize these datasets for their unique needs.

The BTO Sensors and Controls Sub-Program [website](#) acknowledges the role of curated datasets in meeting its objectives:

“To complement the portfolio of competitive and interdisciplinary, early-stage R&D projects, testing frameworks and procedures, along with curated datasets are being developed by the sub-program to support the R&D community at-large through common baselines to evaluate performance improvements and encourage additional innovation in advanced analytics and control strategies”
Key Activity 1 - Use Cases Identification & Prioritization

We developed a process to identify and prioritize use cases that would benefit from improved access to high-quality real-world building datasets:

- Identification and refinement of prospective use cases
- Development of a use case scoring method based on the potential benefits of each use case
- An iterative scoring process to prioritize the use cases

We collected use case suggestions from each of the laboratories in the project team. This yielded an initial list of 15 prospective use cases.

Example

**UC#1**
Design, size, or estimate the performance of demand management technologies (e.g., flexible loads, energy storage, supervisory controls) for different applications.

<table>
<thead>
<tr>
<th>Map to Dataset Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy use data</td>
</tr>
<tr>
<td>Indoor environmental data</td>
</tr>
<tr>
<td>Outdoor environmental data</td>
</tr>
<tr>
<td>System and equipment operational data</td>
</tr>
<tr>
<td>Control setting and logic data</td>
</tr>
<tr>
<td>Occulant data</td>
</tr>
<tr>
<td>Design basis data</td>
</tr>
<tr>
<td>Building and system asset data</td>
</tr>
<tr>
<td>Utility and grid signal data</td>
</tr>
<tr>
<td>Onsite energy generation data</td>
</tr>
<tr>
<td>Cyber (IoT) device data</td>
</tr>
<tr>
<td>Dispatchable energy storage data</td>
</tr>
</tbody>
</table>

Framework to describe the data needed to address any particular use case in terms of “bundled” categories.

Leverage the fact that data collection efforts typically contain several categories of data that are collected together.
# Key Activity 1 - Use Cases Identification & Prioritization

List of 15 use cases (highlighted 5 high-priority use cases)

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td>Design, size, or estimate the performance of demand management technologies (e.g., flexible loads, energy storage, supervisory controls) for different applications.</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Validate methods for modeling how the electrical load of an aggregation of buildings is affected by building-sited energy efficiency or demand management technologies (e.g., HVAC controls, energy storage).</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Evaluate tradeoffs between different levels of weather data collection (e.g., on-site measurement vs. data from miles away) for field studies of weather-sensitive technologies (e.g., space cooling, space heating, thermal energy storage, solar + storage). Sensitivity study on weather data.</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>Validate methods for generating equipment schedule inputs for building modeling applications in which load-shaping outcomes are sensitive to the timing of occupant-driven loads (e.g., plug loads). Method of generating operational schedule for building energy models.</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Use a combination of building time-series data (e.g., electrical load), measured weather data, and forecast data to evaluate the ability of predictive controls to improve the load-shaping capability of building-sited technologies. Development and validation of predictive controls.</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Develop and validate methods (statistical or machine learning) for using indirect observations or proxy data (e.g., WiFi data) to infer occupant count, which can be used in model predictive control.</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Determine potential savings and grid services for smart-zoned residential HVAC systems with room-level zoning and control based on room-by-room occupancy.</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>Understand and quantify human-building interactions to improve building design and operation. (Example: the selection of smart thermostats and the design of operable windows.) Human building interaction.</td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>Fault detection and diagnostics to identify faulty equipment and control, as well as evaluate their impact on energy.</td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>Evaluation of methodologies for development and calibration for different types of building energy models (e.g., blackbox, grey box, high fidelity models, etc.) to understand building operation, controls and performance. Calibration of building energy models.</td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>Use case #11 from original list was merged with use case #5.</td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>Optimization of occupant comfort, health, and safety (IAQ) and minimize cost of operations and/or energy consumption at the same time. Co-optimizing IEQ and energy.</td>
</tr>
<tr>
<td><strong>13</strong></td>
<td>Ensure cyber-secure operation of energy systems in buildings. Identify, diagnose and isolate cyber-attacks on a building control system. Cyber security of building operations.</td>
</tr>
<tr>
<td><strong>14</strong></td>
<td>Develop system-level (e.g., lighting, HVAC) key performance indicators for energy efficiency benchmarking and performance diagnostics. System level performance.</td>
</tr>
<tr>
<td><strong>15</strong></td>
<td>Determine the real-world frequency of energy mismanagement by occupants and the energy savings/load flexibility potential of preventing mismanagement. Identification and prevention of energy mismanagement.</td>
</tr>
</tbody>
</table>

**Demand management technologies**

**Model validation for aggregated Bldgs**

**Sensitivity study on weather data**

**Method of generating operational schedule for building energy models**

**Occupant count prediction**

**Smart-zoned residential HVAC control**

**Human building interaction**

**Fault detection and diagnostics**

**Calibration of building energy models**

**Co-optimizing IEQ and energy**

**Cyber security of building operations**

**System level performance**

**Identification and prevention of energy mismanagement**

---

U.S. DEPARTMENT OF ENERGY  OFFICE OF ENERGY EFFICIENCY & RENEWABLE ENERGY
Key Activity 2 - New Data Collection Efforts

- **Objective**: To design experiments for data collection and develop sensor and instrumentation plan, based on the dataset needs and any existing instrumentation in the buildings.

- **Four buildings (sites) were down selected**: From the 15 potential sites that were identified to cover various DN (Data needs) and UC (Use cases) as well as different building types and climate zones.

Flexible research platforms - ORNL
Building 59 - LBNL
Lab Homes - PNNL
Bear Creek High School - NREL
<table>
<thead>
<tr>
<th>Building</th>
<th>Flexible research platforms (ORNL)</th>
<th>Building 59 (LBNL)</th>
<th>Lab Homes (PNNL)</th>
<th>Bear Creek High School (NREL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building type</td>
<td>Medium office (Commercial)</td>
<td>Medium office (Commercial)</td>
<td>Residential</td>
<td>Secondary school</td>
</tr>
<tr>
<td>Climate zone</td>
<td>4A</td>
<td>3C</td>
<td>5A</td>
<td>5B</td>
</tr>
<tr>
<td>Location</td>
<td>Oak Ridge, TN</td>
<td>Berkeley, CA</td>
<td>Richland, WA</td>
<td>Lakewood, CO</td>
</tr>
<tr>
<td>Data Needs</td>
<td>1,2,3,6,11,12,13</td>
<td>1,2,3,6,11,12,13</td>
<td>1, 2, 3, 4, 7, 12, 13</td>
<td>1,2,3,6,7,8,11,12,13</td>
</tr>
<tr>
<td>Use Cases</td>
<td>1, 5, 9</td>
<td>1,2,5,6,12,13</td>
<td>1, 2, 5</td>
<td>1,5,6,10,12,14,15</td>
</tr>
<tr>
<td>Expected test/data period</td>
<td>2021 summer to 2021 winter</td>
<td>2019 to 2021</td>
<td>2020 to 2021</td>
<td>2021 to 2022</td>
</tr>
<tr>
<td>Test methods</td>
<td>Normal operation; Thermostat Setback; Pre-cooling; Pre-heating</td>
<td>Normal operation;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key Activity 3 - Data Curation and Organization

- Curation of 6 datasets covering diverse building types, sizes, climate zones and use cases
- Three major steps of data curation (based on FAIR principles)
  - Data preprocessing/cleaning to reduce missing rates and outlier rates
  - Data standardization (metadata of dataset, sensor and meters)
Key Activity 3 - Development and Enhancement of Data Tools

- A review of 24 existing data tools to identify their features and limitations
- A structured “readme” .json file to provide the high-level contextual information of the dataset, and to illustrate application perspectives
- Brick schema extension for representing occupant metadata
Key Activity 4 - Data Portal Development and Dataset Publishing

Search Bar

Search facets customized for Buildings

Dataset download options

Dataset information customized for Buildings
Link: https://bbd.labworks.org/
Registration required to access all features

Key Current Features
- Currently has **6 datasets** from across the US representing multiple climate zones, building types, HVAC system types and footprint (more to be added in FY22)
- Raw as well as processed data available
- Processed data includes
  - **Cleaned data** with gaps filled and outliers removed using ML techniques, with documentation of techniques
  - Data represented using a **custom-built metadata** schema based on extensive literature review
  - System and layout information with hierarchies represented using **BRICK** schema
- Features developed using scoring based prioritization matrix
- Easy navigation using **search facets** customized for the building domain
- **Usage metrics** are tracked

Platform
- Built on existing **mature data management capabilities at PNNL** that support atmospheric, wind, materials discovery, marine energy and other EERE domains
- Built on **FAIR** principles - findability, accessibility, interoperability, and reusability
- Content is driven by team members
- **Scalable**, easy to maintain and deploy
- The attributes maintained on a github repository by a few members closely aware of domain and dataset
- Future capabilities being developed
  - A data processing pipeline that **automatically** ingests data; cleans, curates, and standardizes data based on the metadata schema developed in the project; and provides data access via a “data lake” that can allow users to run custom applications (such as machine learning algorithms)
  - Support for **real-time data streaming** from a data source such as Building Management System, depending on the needs of the users
  - **Enhanced user insights** into time-series data, such as filtering by sensor types and time span without the need to download the entire dataset
Case Study 1 - Model Development and Calibration

- Development of a control-oriented building model using the data from NIST Net Zero Energy Residential Test Facility (NZERTF)
  - Development of a gray box model (i.e., RC model), an ideal tool for optimal control design such as Model Predictive Control (MPC)
  - Prediction of average indoor air temperature (Tin) using the model
  - Model was developed based on available data and best guess for some building characteristics such as roof area, solar absorption rates of envelope and roof, SHGC, etc.

### Schematics of the simplified building thermal network model

#### Dataset:
[https://bbd.labworks.org/ds/bbd/netzeromd](https://bbd.labworks.org/ds/bbd/netzeromd)

<table>
<thead>
<tr>
<th>Model</th>
<th>Time</th>
<th>Training/validation</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>Aug 1 to Aug 7, 2015</td>
<td>Training</td>
<td>0.31°C</td>
</tr>
<tr>
<td></td>
<td>Aug 8 to Aug 14, 2015</td>
<td>Validation</td>
<td>0.95°C</td>
</tr>
<tr>
<td>Improved</td>
<td>Aug 1 to Aug 7, 2015</td>
<td>Training</td>
<td>0.13°C</td>
</tr>
<tr>
<td></td>
<td>Aug 8 to Aug 14, 2015</td>
<td>Validation</td>
<td>0.12°C</td>
</tr>
</tbody>
</table>
Case Study 2: Energy and Occupancy Analytics

- Analysis on 3 research questions using the Building 59 dataset
  - Building energy benchmarking and load shape analysis
  - Prediction of occupant count and its correlation with energy use
  - Impacts of COVID pandemic on occupancy and building performance

Dataset: https://bbd.lblworks.org/ds/bbd/lbnlbldg59

Load Shape - August 5, 2019

Electricity Use vs. Occupant Count (May-July, 2018)

Energy Use - end uses [kWh/day]

Change of HVAC operation (supply air flow rate) during the pandemic
Stakeholder Engagement and Outreach

Convened a Technical Advisory Group (TAG) of industry and academic experts to:
- solicit input on use cases, dataset needs, planned data collection efforts, data tools and repository
- pursue synergistic research opportunities

<table>
<thead>
<tr>
<th>Industry</th>
<th>Academia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay Nesler</td>
<td>Krish Gomatom</td>
</tr>
<tr>
<td>JCI</td>
<td>EPRI</td>
</tr>
<tr>
<td>Zoltan Nagy</td>
<td>Rishee Jain</td>
</tr>
<tr>
<td>UT Austin</td>
<td>Stanford University</td>
</tr>
<tr>
<td>Joel Bender</td>
<td>Sagar Rao</td>
</tr>
<tr>
<td>ASHRAE 223P/Cornell</td>
<td>IBPSA-USA, AEI</td>
</tr>
<tr>
<td>David Eric Schwartz</td>
<td>Goncalo Costa</td>
</tr>
<tr>
<td>Xerox Palo Alto Research Center (PARC)</td>
<td>Robert Bosch, USA</td>
</tr>
<tr>
<td>Hyojin Kim</td>
<td>Michael Blasnik</td>
</tr>
<tr>
<td>New Jersey Institute of Technology (NJIT)</td>
<td>Google (Nest)</td>
</tr>
</tbody>
</table>

Activity Highlights

- **June 2020**: Concierge TAG: One-on-one meetings with TAG members for in-depth discussions
- **November 2020**: Circulated responses to feedback, summarized actions taken
- **March 2021**: All-TAG meeting (Virtual)
- **March 2021**: BTO Sensors and Controls Lab Sharing Workshop
- **November 2021**: A seminar at the ASHRAE BPAC conference
Remaining Project Work

Where are we?

- Use cases identified and mapped with data needs
- Data infrastructure identified and new sensors installed for new data collection
- 6 datasets curated and published on the project data portal
- Demonstrated two use cases using the curated datasets
- Conducted two rounds of TAG meetings
- Engaged communities on related efforts (e.g., IEA Annex 81, ASHRAE 223P)

Plan Ahead (FY22):

- Complete new data collection and publish the datasets
- Streamline and semi-automate the data curation workflow
- Demonstrate another two use cases using the new collected datasets
- Conduct the third round of TAG meetings
- Outreach: publications, seminars, community engagement
- Explore opportunities to continue the data portal as a resource serving the community
Thank You

LBNL
Tianzhen Hong
thong@lbl.gov

NREL
Lieko Earle
lieko.earle@nrel.gov

ORNL
Piljae Im
imp1@ornl.gov

PNNL
Vikas Chandan
vikas.chandan@pnnl.gov
REFERENCE SLIDES
# Project Budget

**Project Budget:** $3.75M  
**Variances:** N/A  
**Cost to Date:** $1.98M  
**Additional Funding:** N/A

## Budget History

<table>
<thead>
<tr>
<th></th>
<th>FY 2020 (past)</th>
<th>FY 2021 (current)</th>
<th>FY 2022 (planned)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOE</td>
<td>$1.15M</td>
<td>DOE</td>
<td>DOE</td>
</tr>
<tr>
<td>Cost-share</td>
<td>N/A</td>
<td>Cost-share</td>
<td>Cost-share</td>
</tr>
<tr>
<td></td>
<td>$1.35M</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$1.25M</td>
<td>N/A</td>
</tr>
</tbody>
</table>
# Project Plan and Schedule

**D = Accomplished AOP Deliverable/Milestone**  
**G/N = Accomplished AOP Go/No-Go**

<table>
<thead>
<tr>
<th>Task</th>
<th>Deliverable/Milestone</th>
</tr>
</thead>
</table>
| 1. Existing data collection and curation (M1-21) | Schematic mapping data sources to use cases: Describe mapping of the data sources (existing and/or new datasets) to the identified use cases (M3)  
Existing datasets curated: Collect and curate at least 4 existing datasets from the identified existing datasets. Curation scripts, curated datasets and dataset description documents reviewed with the project team and technical advisors. (M2a) | D  
D  
D  
G/N |
| 2. New Data Collection (M1-33) | Report on data collection sites: Evaluate and finalize the facilities for the new data collection for the identified use cases (M6)  
Go/No-Go: Data collection plan, instrumentation design for selected buildings completed, and data representation requirements finalized (M12)  
SMART Milestone: Complete installations and start the collection (M24)  
Summary reports for all datasets collected in Task 1 (M33) | D  
G/N  
D  
D |
| 3. Data tools and database (M1-30) | Propose enhancements to data representation tools and development of requirements of data warehouse tools: Built upon the reviews done in FY19, propose enhancements to data representation tools; develop requirements of data warehouse tools (M9)  
Metadata map for all the curated datasets (M24)  
SMART Milestone: Data portal with all datasets collected and curated in Tasks 1 & 2 uploaded (M30) | D  
D  
D  
D |
| 4. Demonstration of four case studies (M1-35) | SMART Milestone: Technical reports on four case studies using datasets collected in this project (two on M21, two on M36) | D |
| 5. Engagement and dissemination (M1-36) | Establish the Technical Advisory Group and organize three meetings  
BPAC Seminar proposals and presentation slides | D  
D  
D |

### Budget Year 1 (Fiscal Year 2020)

<table>
<thead>
<tr>
<th>CY 2019</th>
<th>Oct Nov Dec Jan Feb Mar Apr May June Jul Aug Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
</tbody>
</table>

### Budget Year 2 (Fiscal Year 2021)

<table>
<thead>
<tr>
<th>CY 2020</th>
<th>Oct Nov Dec Jan Feb Mar Apr May June Jul Aug Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
</tbody>
</table>

### Budget Year 3 (Fiscal Year 2022)

<table>
<thead>
<tr>
<th>CY 2021</th>
<th>Oct Nov Dec Jan Feb Mar Apr May June Jul Aug Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
</tbody>
</table>
Development and Enhancement of Data Tools

- A review on 24 existing data tools to identify their features, applications, and limitations

Categorizing of 24 existing data tools

Key features of 24 existing data tools


https://doi.org/10.1016/j.rser.2021.111224
Feature identification and prioritization

- 40 features were identified belonging to categories such as collection of data, curating datasets, metadata, data discovery, security and download.
- Each feature was scored by a multi-lab team on 2 parameters: ease of implementation and value to users.
- A composite score was developed for each feature which was used to determine its priority: Very High, High, Medium, Low, Trivial.
- FY20 Focus: Very High, and High priority features.
- FY21 Focus (ongoing): Medium priority features.

**Very High priority features**
- Explicit public license
- Contact for dataset (distribution list)
- Citation (if open) DOI
- Climate zone in metadata
- Data resolution
- Access and Authentication
- Secure web portal

**High priority features**
- Search through metadata
- Measurements in metadata
- Keyword based filtering of datasets to allow users find datasets with desired attributes
- Adding google analytics to capture metrics such as page visits, most viewed datasets, and downloads
- Processing data into standard form
- Data cleaning, run manually
- Downloadable data format

**Medium priority features**
- A link to the document that provides information about the structure/standards documentation
- Clean/curated metadata
- A link to the physical layout of the buildings for each dataset (if appropriate).
- A link to the document describing how data was cleaned before sharing the data with users
- Documentation of our curation/cleaning
- Dataset usage example/recipe
- Raw data downloadable
- Subset data
- Periodic batch collection/processing (cleaning and standardizing)
- Make Data curation/cleaning scripts available to users
- Transforming data resolutions (regridding)
- Dataset time coverage information
- Marked data quality for different time periods
Selecting Fields for the portal

- A second prioritization matrix was developed by multi-lab team to identify the appropriate level for each field: project page, dataset page, or search queries.