Schneider-Electric Cold Start Challenge

Summary

Buildings account for 40% of global CO_2 emissions, but are often overlooked in the public consciousness even though they are fertile ground for climate and energy innovation. Forecasting energy consumption in a building due to Heating, Ventilation and Air Conditioning (HVAC) is a vital tool for minimizing energy waste. By knowing how much load to place on these systems for a given outside temperature, we can optimally achieve desired comfort settings without wasteful expenditure.

In practice two situations can arise: either the building under consideration has been instrumented for years, enabling the use of past data to create (and validate) models for the energy consumption of this specific building; or the building is new or only recently instrumented, in which case very little data on this specific building is available. The latter is the situation considered for this “Cold Start” challenge. Forecasts can then be established by analogy with other buildings, considering the little data available for the building of interest.

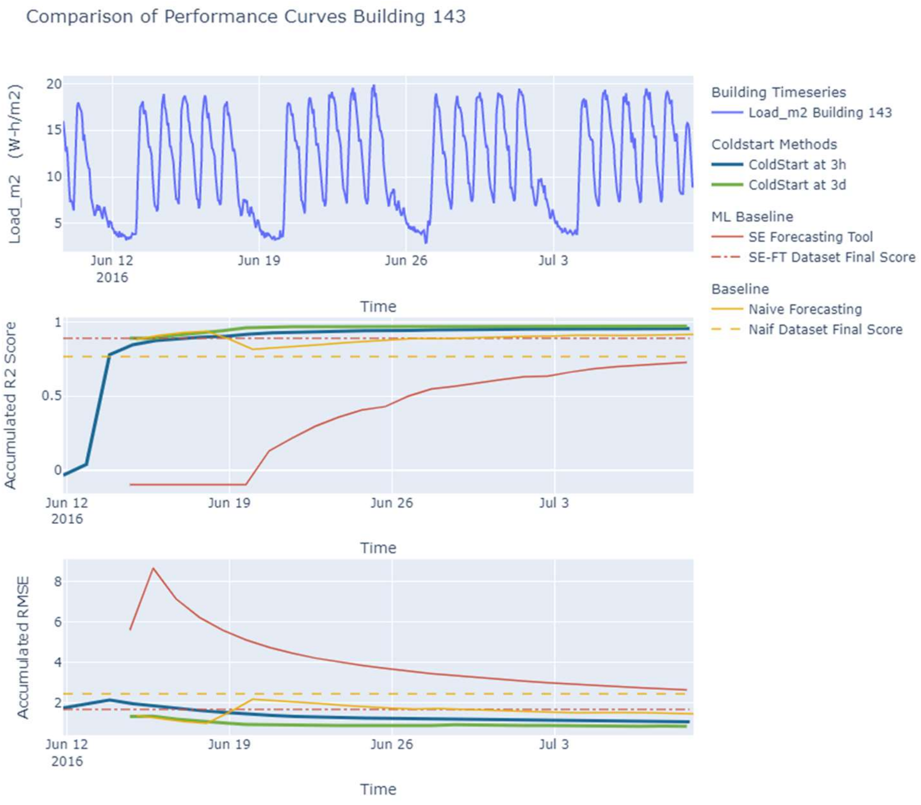


***Figure 1**: Performance of ColdStart Methods vs baselines. **Top**: Energy Consumption for sample building. **Middle**: Accumulated R2 score for chosen method. **Bottom**: Accumulated Root-mean-square for given method. Note how much faster ColdStart achieves above-threshold performance for minimizing energy waste.*

Data Access

Please access the data for this challenge by following these steps:

1. Go to https://shop.exchange.se.com/en-US/apps/39021/forecasting-cold-start-building-energy-consumption
2. Sign up for an account with Schneider Electric Exchange, SE's data sharing platform. Account sign up is free and requires an email address

Access the data by downloading the CSV/JSON/Excel files (all three options contain the same data, just a different file format)Forecasting building energy consumption – Cold Start

Context

Forecasting the energy consumption of a building is essential to optimize its energetic performance both in the short and long terms. Facility managers, utility companies and commissioning projects use consumption forecasts to design and implement energy-savings policies to manage energy storage, optimize the grid's load, and reduce environmental impact. In addition, optimization algorithms sometimes rely on energy consumption forecasts. Therefore, improving forecasts' quality helps get better optimization results.

In practice two situations can arise: either the building under consideration has been instrumented for years, enabling the use of past data to create (and validate) models for the energy consumption of this specific building; or the building is new or only recently instrumented, in which case very little data on this specific building is available. The latter is the situation considered for this "Cold Start" challenge. Forecasts can then be established by analogy with other buildings, considering the little data available for the building of interest.

Objective

The objective is to forecast 24 hours in advance the hourly energy consumption of different buildings based on historical consumption data of other buildings and little data of the buildings for which we want to make forecasts, as well as temperature and other relevant features included in the provided datasets and that will be explained below. Characteristics of the expected forecasts are given in the *Problem Statement* section.

Data

We have gathered data from 91 buildings around the world **(location is not known)**. The quantity of data varies among the buildings. Figure 1 shows the number of data points for each building in the dataset.

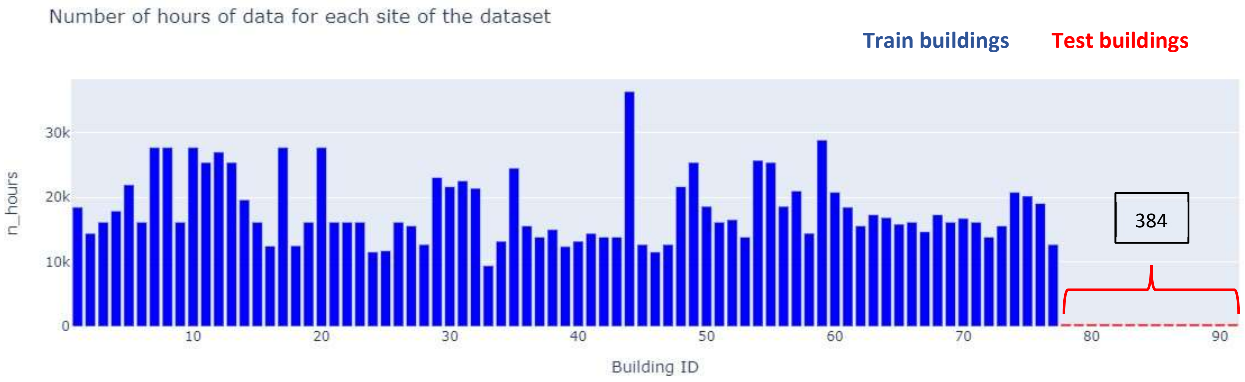


Figure 1. Number of hours for each building of the dataset

77 of these buildings appear only in the training set: the corresponding data can be used to build models.

14 of these buildings appear in the test set: these are the buildings for which forecasts are required. For each of these buildings, 48 actual energy consumption values are provided (equivalent of two days) to enable fine tuning of the models. 336 are to be forecasted. These 336 values are further divided in two halves: the first half is used to establish the public leader board; the second half will remain private and be used solely to detect overfits.

Below are the details of the data contained in the different files provided for the competition.

1. **Train data** - *forecasting-building-energy-consumption-train.csv*
   Contains all the data for training, with features explained below:

   a. **obs_id**: an arbitrary ID.

   b. **site_id**: matches across datasets. An arbitrary ID given to the site.

   c. **timestamp**: hourly timestep under the format YYYY-MM-DD hh:mm:ss. Can be used for extracting additional useful time-related features.

   d. **temperature**: the external temperature outside the building.
   Unit: Celsius degrees

   e. **temperature_-1**: the external temperature, an hour before the observation at a given timestamp.
   Unit: Celsius degrees

   f. **temperature_-2**: the external temperature, two hours before the observation at a given timestamp.
   Unit: Celsius degrees

   g. **load**: the energy consumed during the hour that precedes the timestamp. For example, the load given at timestamp 2021-12-05 10:00:00 is the total energy consumed between 9:00 and 10:00 on December 5, 2021.
   Unit: Watt-hour

   h. **load_-1**: the energy consumption of the building an hour before the observation at a given timestamp.

Unit: Watt-hour

i. **load_-2**: the energy consumption of the building two hours before the observation at a given timestamp.
Unit: Watt-hour

j. **target**: the energy consumption of the building 24h after the observation at a given timestamp. This is the feature we want to predict. For example, the target at timestamp 2021-12-05 10:00:00 is the total energy consumed between 9:00 and 10:00 on December 6, 2021.
Unit: Watt-hour

2. **Test data** - *forecasting-building-energy-consumption-test.csv*
Contains all the data for testing, with features explained below:

a. **obs_id**: an arbitrary ID.
b. **site_id**: matches across datasets. An arbitrary ID given to the site.
c. **timestamp**: hourly timestep under the format YYYY-MM-DD hh:mm:ss. Can be used for extracting additional useful time-related features.
d. **temperature**: the external temperature outside the building.
Unit: Celsius degrees

e. **temperature_-1**: the external temperature, an hour before the observation at a given timestamp.
Unit: Celsius degrees

f. **temperature_-2**: the external temperature, two hours before the observation at a given timestamp.
Unit: Celsius degrees

g. **load**: the energy consumption of the building at a given timestamp.
Unit: Watt-hour

h. **load_-1**: the energy consumption of the building an hour before the observation at a given timestamp.
Unit: Watt-hour

i. **load_-2**: the energy consumption of the building two hours before the observation at a given timestamp.
Unit: Watt-hour

j. **target**: the energy consumption of the building 24h after the observation at a given timestamp. This is the feature we want to predict.
Unit: Watt-hour

[IMPORTANT]: in this cold start forecasting scenario, models can be trained with some buildings' data, and then be applied on new buildings (transfer learning) to predict the energy consumption of the latter. Before doing the inference on a new building, we

provide competitors with 48 new data points (equivalent of 2 days) randomly chosen across the specific building for which the forecasting is to be done. These data points can be used to retrain the models. To keep the challenge simple, it is allowed to use all of these 48 data points right from the start and use the retrained model for all the forecasts of the building (even those which precede the given data points).
Forecasts are expected for 168 * 2 = 336 data points also randomly chosen (eq. 14 days) different from the 48 for which the consumption has been provided. The graph below helps get a better understanding of this process.

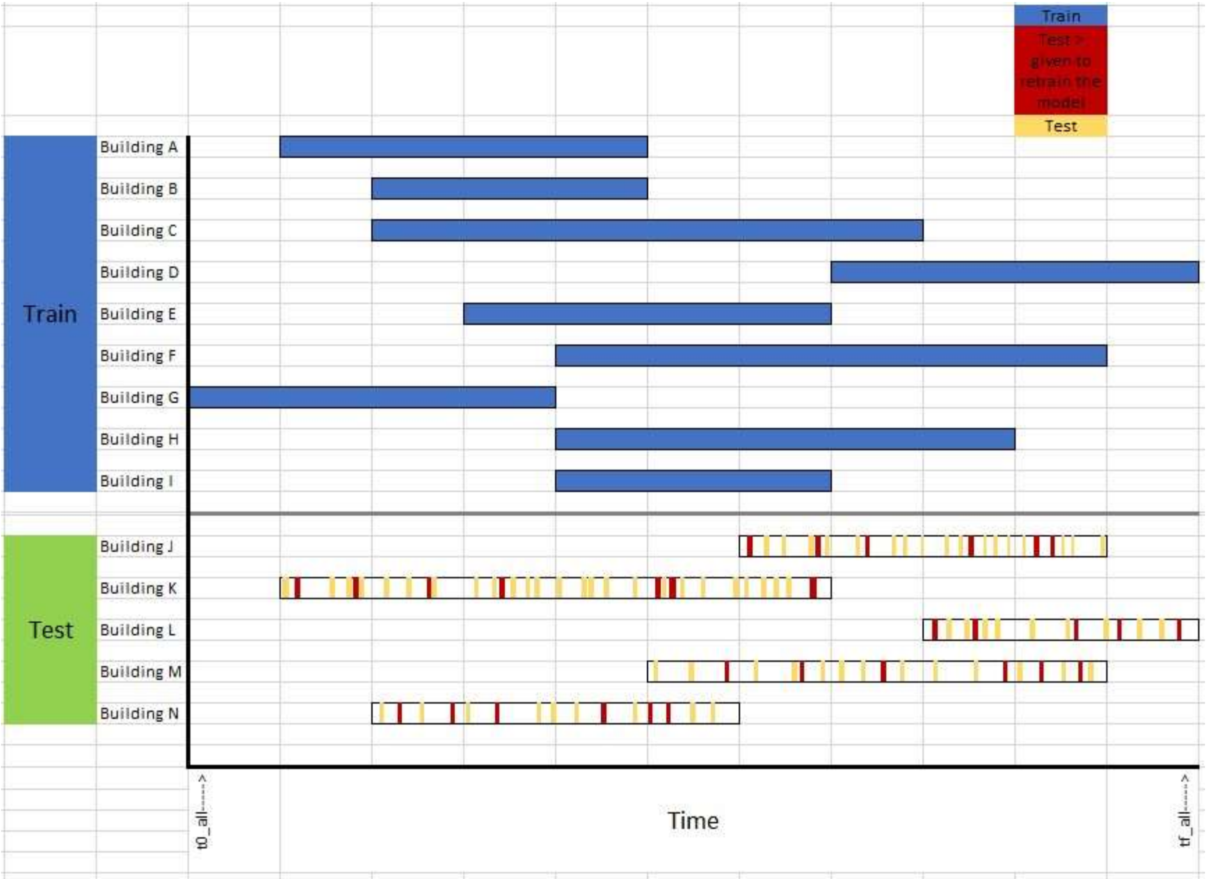
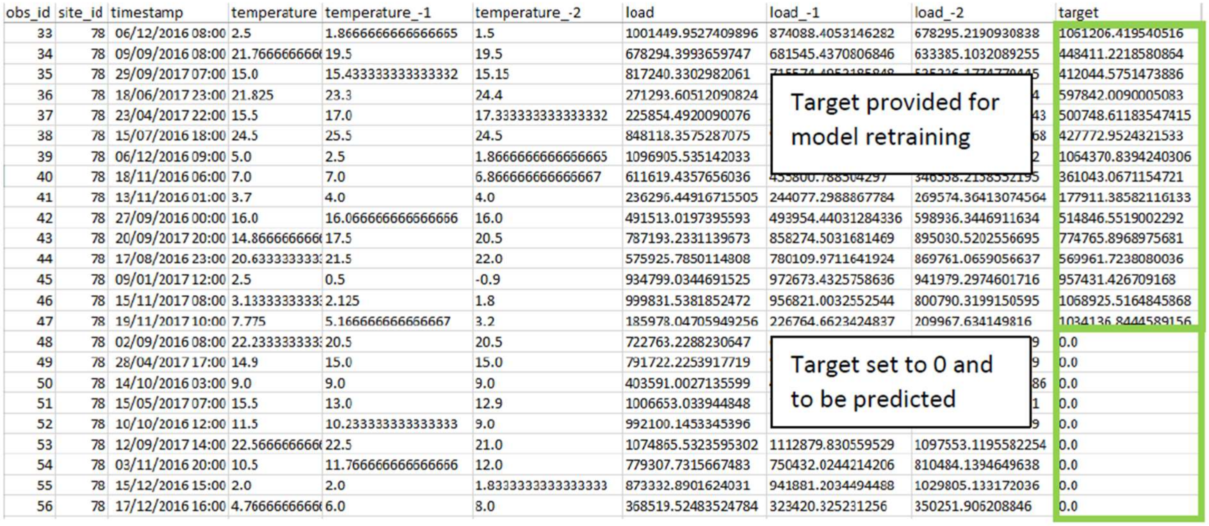


Figure 2. Cold Start Process
NB : the number of bars for each building is arbitrary.



obs_id	site_id	timestamp	temperature	temperature_-1	temperature_-2	load	load_-1	load_-2	target
33	78	06/12/2016 08:00	2.5	1.8666666666666665	1.5	1001449.9527409896	874088.4053146282	678295.2190930838	1061206.419540516
34	78	09/09/2016 08:00	21.766666666(	19.5	19.5	678294.3993659747	681545.4370806846	633385.1032089255	448411.2218580864
35	78	29/09/2017 07:00	15.0	15.433333333333332	15.15	817240.3302982061	715574.4053185848	535336.1774770445	412044.5751473886
36	78	18/06/2017 23:00	21.825	23.3	24.4	271293.60512090824	[illegible]	[illegible]	597842.0090005083
37	78	23/04/2017 22:00	15.5	17.0	17.333333333333332	225854.4920090076	[illegible]	[illegible]	500748.61183547415
38	78	15/07/2016 18:00	24.5	25.5	24.5	848118.3575287075	[illegible]	[illegible]	427772.9524321533
39	78	06/12/2016 09:00	5.0	2.5	1.8666666666666665	1096905.535142033	[illegible]	[illegible]	1064370.8394240306
40	78	18/11/2016 06:00	7.0	7.0	6.866666666666667	611619.4357656036	455800.788504297	346558.2158552195	361043.0671154721
41	78	13/11/2016 01:00	3.7	4.0	4.0	236296.44916715505	244077.2988867784	269574.36413074564	177911.38582116133
42	78	27/09/2016 00:00	16.0	16.066666666666666	16.0	491513.0197395593	493954.44031284336	598936.3446911634	514846.5519002292
43	78	20/09/2017 20:00	14.866666666(	17.5	20.5	787193.2331139673	858274.5031681469	895030.5202556695	774765.8968975681
44	78	17/08/2016 23:00	20.63333333:	21.5	22.0	575925.7850114808	780109.9711641924	869761.0659056637	569961.7238080036
45	78	09/01/2017 12:00	2.5	0.5	-0.9	934799.0344691525	972673.4325758636	941979.2974601716	957431.426709168
46	78	15/11/2017 08:00	3.13333333333:	2.125	1.8	999831.5381852472	956821.0032552544	800790.3199150595	1068925.5164845868
47	78	19/11/2017 10:00	7.775	5.166666666666667	3.2	185978.04705949256	226764.6623424837	209967.634149816	1034136.8444589156
48	78	02/09/2016 08:00	22.23333333:	20.5	20.5	722763.2288230647	[illegible]	[illegible]	0.0
49	78	28/04/2017 17:00	14.9	15.0	15.0	791722.2253917719	[illegible]	[illegible]	0.0
50	78	14/10/2016 03:00	9.0	9.0	9.0	403591.0027135599	[illegible]	[illegible]	0.0
51	78	15/05/2017 07:00	15.5	13.0	12.9	1006653.033944848	[illegible]	[illegible]	0.0
52	78	10/10/2016 12:00	11.5	10.233333333333333	9.0	992100.1453345396	[illegible]	[illegible]	0.0
53	78	12/09/2017 14:00	22.566666666(	22.5	21.0	1074865.5323595302	1112879.830559529	1097553.1195582254	0.0
54	78	03/11/2016 20:00	10.5	11.766666666666666	12.0	779307.7315667483	750432.0244214206	810484.1394649638	0.0
55	78	15/12/2016 15:00	2.0	2.0	1.8333333333333333	873332.8901624031	941881.2034494488	1029805.133172036	0.0
56	78	17/12/2016 16:00	4.766666666(	6.0	8.0	368519.52483524784	323420.325231256	350251.906208846	0.0

Target provided for model retraining

Target set to 0 and to be predicted

Figure 3. Extract from the test dataset, the 0 target values are the ones to replace by the forecasts

3. Submission format - *forecasting-building-energy-consumption-submission-format.csv* This file must be filled in with forecasted values, output of the models: in other words, the instances where target feature is set to 0 in the test dataset. It is then used to compute the score to get the leader board.

   Features are:

   a. **obs_id**: an arbitrary ID.
   b. **site_id**: matches across datasets. An arbitrary ID given to the site.
   c. **timestamp**: hourly timestep under the format YYYY-MM-DD hh:mm:ss.
   d. **leaderboard_target**: in ["public", "private"], represents the type of evaluation of the prediction. Competitors get their score from the public leader board but are also evaluated on a private one to avoid over-training of the model. This feature represents no importance to the competitor but is used by the company to compute the leader board.
   e. **target**: the energy consumption of the building 24h after the observation at a given timestamp. It is set to 0 and asked to the candidates to fill in this column with their predictions.
      Unit: Watt-hour

4. Metadata - *forecasting-building-energy-consumption-meta.csv* Contains metadata for all 91 buildings. Metadata are:

   a. **obs_id**: an arbitrary ID.
   b. **site_id**: matches across datasets. An arbitrary ID given to the site.
   c. **surface**: surface of the building
      Unit: m²

d. **base_temperature**: base temperature inside the building (setpoint).
Unit: Celsius degrees

e. **[DAY]_is_day_off**: DAY being one of [monday, tuesday, wednesday, thursday, friday, saturday, sunday]
Boolean variable being True if DAY is a non-working day, False otherwise.

5. Holidays - *forecasting-building-energy-consumption-holidays.csv* Contains holidays data for all buildings. Features are:
   a. **obs_id**: an arbitrary ID.
   b. **site_id**: matches across datasets. An arbitrary ID given to the site.
   c. **holiday**: name of the public holiday.
   d. **date**: date of the public holiday in the country where the building is located. Format of the date is YYYY-MM-DD.

Problem Statement

The problem here focuses on cold start energy forecasting. The competitor is asked to predict the energy consumption of various buildings that have not been seen during the training phase. Cold start forecasting aims at predicting the energy consumption of a building without gathering much information on this building.

The data is split across the buildings (c.f. Figure 2). Training buildings have as much data as possible, but test buildings contain the equivalent of 16 days of data, among which 14 days to predict and 2 days to fine tune the model, randomly obtained from samples across the target building. To be clear, we have randomly chosen 384 data points (16 days equivalent) across each test building's dataset to avoid any possibility of reconstructing the original time-series.

The 14 days equivalent of forecasting are split in two for evaluation: an equivalent of 7 days is used to get the public score the candidates will see while competing. For each data point, the difference between the provided forecast and the real consumption value will be weighted by the inverse of the surface of the building. This provides an error in kWh/m^2. The leader board will report the Mean Square Error (MSE) in kWh/m^2. The others 7 days equivalent will also be used to compute the MSE in kWh/m^2 to establish a private leader board. This will ensure competitors do not over-train their models to achieve good results only on the "public" 7 days predictions. The private leader board will not be shared. The characteristics for the forecasting are listed below:

- **Input window**: Data is provided for an input window of 3h: temperature and load are lagged in the dataset.
  Precision on the input window: for instance, if we have Monday data at 01:00, 02:00 and 03:00, these represent the energy consumptions between 00:00 and 01:00, between 01:00 and 02:00, and between 02:00 and 03:00. These can be used to forecast the energy consumption of Tuesday at 03:00, i.e., between 02:00 and 03:00.

- **Time horizon**: The prediction horizon is 24h. We do not supervise the predictions for the next 23 hours, but only the 24th one.
- **Forecast update frequency**: Hourly forecast.
- **Metric for leader board** – Mean Square Error (MSE) in kWh/m^2.