# Solar, Wind, and Load Forecasting Dataset for MISO, NYISO, and SPP Balancing Areas

Richard Bryce, Cong Feng, Brian Sergi, Ross Ring-Jarvi, Wenqi Zhang, and Bri-Mathias Hodge

*National Renewable Energy Laboratory*

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

**Technical Report**
NREL/TP-5D00-83828.
April 2024

# Solar, Wind, and Load Forecasting Dataset for MISO, NYISO, and SPP Balancing Areas

Richard Bryce, Cong Feng, Brian Sergi, Ross Ring-Jarvi, Wenqi Zhang, and Bri-Mathias Hodge

*National Renewable Energy Laboratory*

**Suggested Citation**
Richard Bryce, Cong Feng, Brian Sergi, Ross Ring-Jarvi, Wenqi Zhang, and Bri-Mathias Hodge. 2023. *Solar PV, Wind Generation, and Load Forecasting Dataset for 2018 across MISO, NYISO, and SPP Balancing Areas*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-5D00-83828. https://www.nrel.gov/docs/fy24osti/83828.pdf.

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

**Technical Report**
NREL/TP-5D00-83828.
April 2024

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

**NOTICE**

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by Advanced Research Projects Agency-Energy under Award, 19/CJ000/07/01. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.



*Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.*

# List of Acronyms

| | |
|---|---|
| ANN | artificial neural network |
| ARPA-E | Advanced Research Projects Agency-Energy |
| BMA | Bayesian Model Averaging |
| BOEM | Bureau of Ocean Management |
| CDF | cumulative distribution function |
| CNN | convolutional neural network |
| CONUS | Contiguous United States |
| CRPS | continuous ranked probability score |
| GHI | Global Horizonal Irradiance |
| ECMWF | European Centre for Medium-Range Weather |
| ERCOT | Electric Reliability Council of Texas |
| IA | Interconnection Agreement |
| ISO | Independent System Operator |
| MBE | mean bias error |
| MISO | Midcontinent System Operator |
| nMAE | normalized mean absolute error |
| NREL | National Renewable Energy Laboratory |
| nRMSE | normalized root-mean-squared error |
| NSRDB | National Solar Radiation Database |
| NWP | Numerical Weather Prediction |
| NYISO | New York System Operator |
| PERFORM | Performance-based Energy Resource Feedback, Optimization, and Risk Management (ARPA-E program) |
| PSM | Physical Solar Model |
| reV | Renewable Energy Potential model |
| SPP | Southwest Power Pool |
| WRF | Weather Research and Forecasting |

# Executive Summary

The Performance-based Energy Resource Feedback, Optimization, and Risk Management (PERFORM) program is an initiative intended to foster "a fundamental shift in grid management rooted in an understanding of asset risk and system risk" (ARPA-E 2020). Launched by the Advanced Research Projects Agency-Energy (ARPA-E), the program supports efforts to incorporate uncertainty in electric power decision making.

In support of PERFORM, the National Renewable Energy Laboratory (NREL) has produced a set of time-coincident forecasts of solar, wind, and load profiles. The time series for these forecasts—which include both deterministic and probabilistic forecasts—and their corresponding actual profile values at high temporal and spatial fidelity. Figure ES1 provides a high-level overview of the process used to generate the actual and forecast profiles in the dataset.

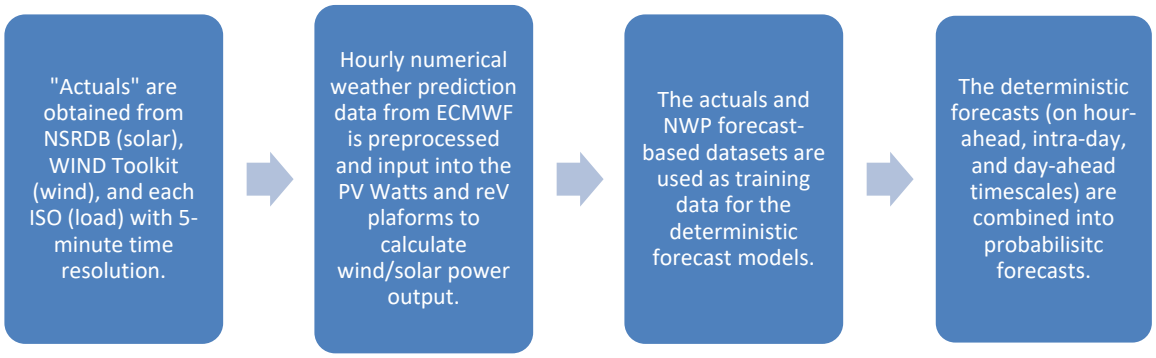



**Figure ES*1*: A schematic of the overall data collection and forecasting process.**

As part of Phase I of the PERFORM effort, NREL created a dataset that consists of one year of time-coincident load, wind, and solar actuals and probabilistic forecasts based on data from the Electric Reliability Council of Texas (ERCOT) (Bryce et al. 2023). In Phase II, NREL developed similar datasets for three other U.S. Independent System Operators (ISO): the Midcontinent Independent System Operator (MISO), the New York Independent System Operator (NYISO), and the Southwest Power Pool (SPP).

Wind and solar profiles are provided for existing facilities as well as planned facilities based on each ISO's interconnection queue. Figure ES1 shows the spatial distribution of solar and wind plants contained in these datasets. For the final PERFORM dataset we include all projects in the queue but add a flag for each project given some indication of its status in the queue, which can help users of the data potentially identify subsets of the projects for their own analyses.

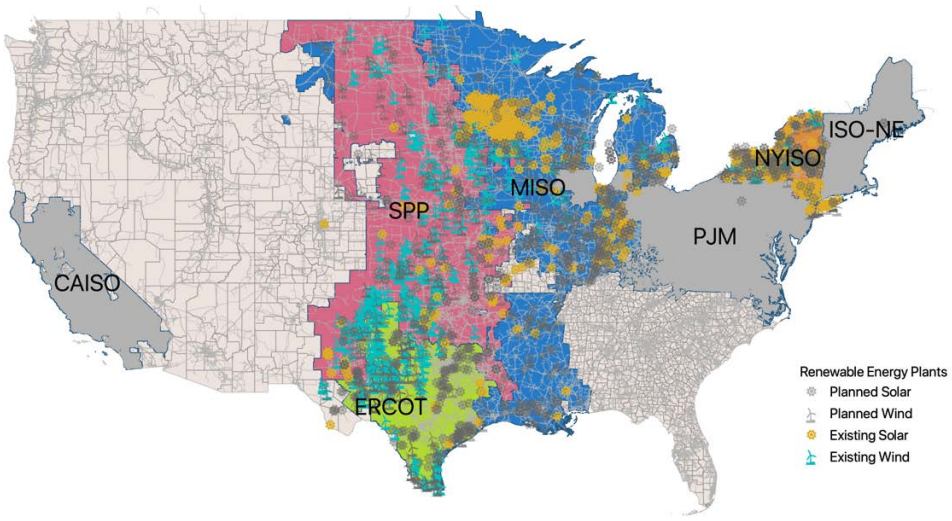



**Figure ES2: Spatial distribution of solar and wind plants contained in these datasets.**

This report describes the process for producing the deterministic and probabilistic forecasts in the Phase II dataset and summarizes some of the metrics used to evaluate the forecasts. The datasets are available for public use, and additional documentation and details on how to access to the data can be found at https://github.com/PERFORM-Forecasts/documentation.

# Table of Contents

# List of Figures

# List of Tables

# 1 Overview of PERFORM Datasets

As part of Phase I of the PERFORM effort, NREL created a dataset that consists of one year of time-coincident load, wind, and solar actuals and probabilistic forecasts based on data from the Electric Reliability Council of Texas (ERCOT) (Bryce et al. 2023). In Phase II, NREL developed similar datasets for three other U.S. Independent System Operators (ISO): the Midcontinent Independent System Operator (MISO), the New York Independent System Operator (NYISO), and the Southwest Power Pool (SPP).

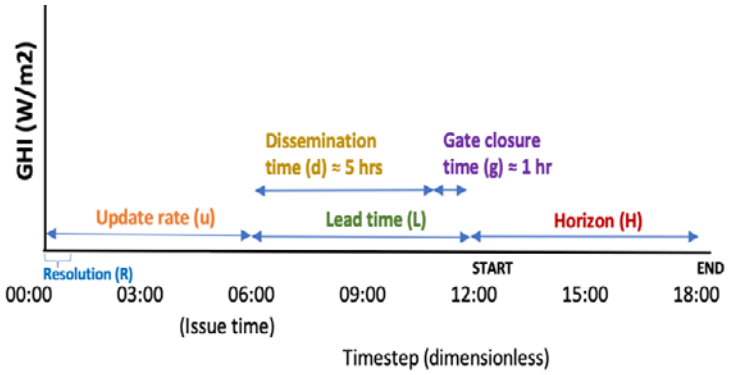
The PERFORM Phase II dataset includes wind, solar, and load time series data for MISO, NYISO, and SPP. The dataset includes both actuals and forecast time series; while actuals are provided for 2018-2019, the forecasts require a year of data for training and are only available for 2019. Solar and wind data have three different spatial-scales, namely site-level, zone-level (defined by the ISO), and system-level (i.e., the balancing area or entire ISO). A summary of the data characteristics is provided in Table 1.

**Table 1: Summary of the actual and forecast datasets provided for MISO, NYISO, and SPP as part of the Phase II PERFORM dataset.**

| | Solar | Wind | Load |
|---|---|---|---|
| **Spatial resolution** | Site-level<br>Zone-level<br>Balancing Area | Site-level<br>Zone-level<br>Balancing Area | Zone-level<br>Balancing Area |
| **Actuals** | 2018-2019 | 2018-2019 | 2018-2019 |
| **Forecasts** | 2019 | 2019 | 2019 |

Both actual and forecasted wind and solar profiles are developed for existing facilities as well as proposed sites based on ISO interconnection queues. Actual time series are based on input meteorological data from the National Solar Radiation Database (NSRDB) and the Weather Research and Forecasting (WRF) model. Forecast time series are developed using forecasts produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Meteorological data from the actual and forecast timeseries serves as input to the renewable energy potential model (reV), which generates wind and solar profiles based on a specified plant configuration. Actual load data is collected from each ISO and is subsequently used with ECMWF forecast data to train a predictive load model. Actuals are provided at 5-min resolution and forecasts are provided either at hourly or 15-min resolution.

Although the forecast time series have similar spatial characteristics as the actual time series, they have more complex temporal characteristics that characterize forecast parameters, such as lead time, horizon, resolution, and update rate. Figure 1 illustrates how these operational parameters are characterized and outlines the assumptions used for the ECMWF forecasts used for this dataset.




| Forecast Run | Lead Time | Horizon | Resolution | Update Rate | Power Percentiles | Method |
|---|---|---|---|---|---|---|
| Day-ahead (Medium-Term) | 11 hours | 48 hours | Hourly | Daily | 1-99 | BMA (using ECMWF ensemble data) |
| Intra-day (Short-Term) | 6 hours | 6 hours | Hourly | 6 hours | 1-99 | BMA (using ECMWF ensemble data) |
| 1 hour-ahead (Very Short-Term) | 1 hour | 2 hours | 15 minutes | Hourly | 1-99 | M3 Machine learning/time series |

**Figure 1: Illustration of the various forecast parameters for the ECMWF forecasts (top) and a description of the parameter assumptions and forecasting method used for the various forecasts in the Phase II dataset (bottom).**

The remainder of this report serves as a reference for the methods used to generate the actuals and forecasts in the PERFORM Phase II dataset, starting with the wind and solar generation profiles (Section 2) and then continuing with the load data (Section 3). The report concludes with an overview of the data set attributes (Section 4).

# 2 Renewable Generation Data and Forecasts

This section describes the steps taken to generate the wind and solar generation profiles for the Phase II dataset. The process starts with identifying the relevant wind and solar sites of interest. Actual generation profiles are then created for these sites using wind and solar data and the reV model. For the forecasts data, weather forecasts from ECMWF are translated to generation profiles again using reV.

## 2.1 Wind and Solar Site Selection

### *Overview of site selection process*

The wind and solar sites used for generating profiles include both existing facilities and "planned" sites that are projects under considerations. For existing sites, we use data from the U.S. Wind Turbine Database and the Utility-Scale Solar Database, both maintained by the Lawrence Berkeley National Laboratory (Hoen et al. 2023; LBNL 2022) . For planned sites, data is taken from each ISO's interconnection queue, summarized in Table 2. From the queue data we subset to wind and solar projects with interconnection or commercial in-service dates starting after January 1, 2021.

**Table 2: Summary of interconnection queue data source and download data for each ISO.**

| ISO | Source | Download date |
|---|---|---|
| MISO | https://www.misoenergy.org/planning/generator-interconnection/GI_Queue/ | Feb. 8, 2021 |
| NYISO | https://www.nyiso.com/interconnections | Feb. 3, 2021 |
| SPP | http://opsportal.spp.org/Studies/GIActive | Feb. 8, 2021 |

A few manual adjustments were made to the site metadata, outlined as follows:

- One of the proposed solar facilities in NYISO (Mineral Basin Solar Power) is in Pennsylvania, which was flagged as a potential data processing error. Research on this site confirmed that its physical location is correct and that it will be injecting power into NYISO, so this plant was retained. However, the original load zone identified for power injection was incorrect and was adjusted accordingly.
- The existing SPP sites included 2 solar sites located at Denver International Airport which are not participating in SPP and were therefore dropped them from the dataset.
- The previous metadata used an inconsistent cutoff date for interconnection projects; the sites have been updated to only include planned projects as of January 1, 2021 as previously intended. This resulted in dropping a few sites from NYISO and SPP.
- Previously planned projects that did not include detailed coordinate information were assigned to the county centroid of the project for the purpose of modeling in reV. Although this can result in some plants being placed in bodies of water or other locations that would be unrealistic for development, we believe it is sufficient for realistically modeling the generation of that site.

Table 3 summarizes the number of unique wind and solar plants and the total installed capacity for each ISO. Planned offshore wind capacity for NYISO is broken separately. Note that the count of plants aggregates individual wind turbines associated within a single plant or planned project. The data has been slightly refined to remove outlier plants and duplicate records and address other data cleaning issues; this table thus provides a summary of the final collection of sites processed for wind and solar forecasts. Similarly, Figure 2 provides a visual summary of the total wind and solar capacity of existing and planned projects by ISO for both existing and planned sites, whereas Figure 3 illustrates the spatial distribution of those plants.

**Table 3: Summary of plant count and installed capacity by ISO.**

| | | Count | | | Capacity (MW) | | |
|---|---|---|---|---|---|---|---|
| **ISO** | *Type* | *Existing* | *Planned* | *Total* | *Existing* | *Planned* | *Total* |
| **MISO** | Solar | 376 | 375 | 751 | 1,396 | 56,063 | 57,459 |
| | Wind | 283 | 82 | 365 | 21,434 | 16,341 | 37,775 |
| **NYISO** | Solar | 160 | 138 | 298 | 458 | 11,784 | 12,242 |
| | Wind | 27 | 20 | 47 | 1,984 | 3,714 | 5,698 |
| | Offshore Wind | - | 26 | 26 | - | 27,718 | 27,718 |
| **SPP** | Solar | 47 | 125 | 172 | 347 | 22,702 | 23,048 |
| | Wind | 186 | 103 | 289 | 20,413 | 24,097 | 44,510 |

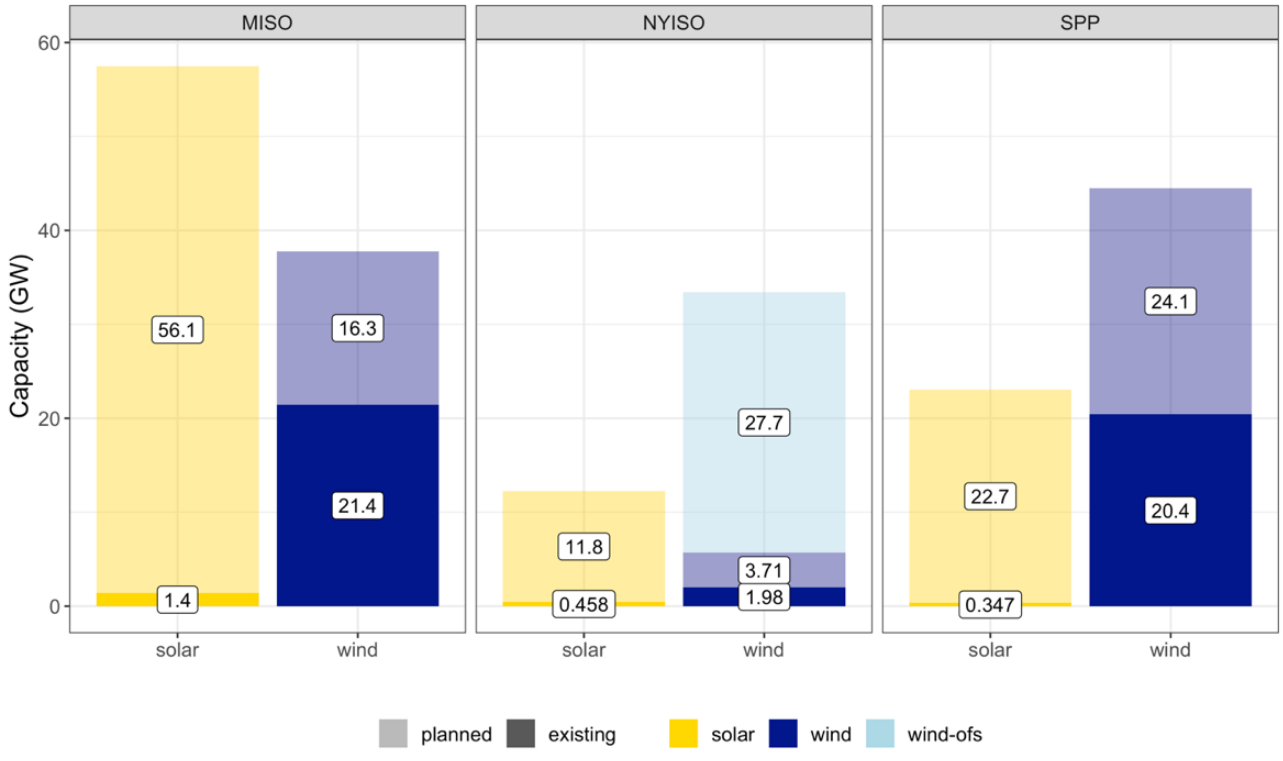



**Figure 2: Summary of existing and planned solar and wind capacity for by ISO.**

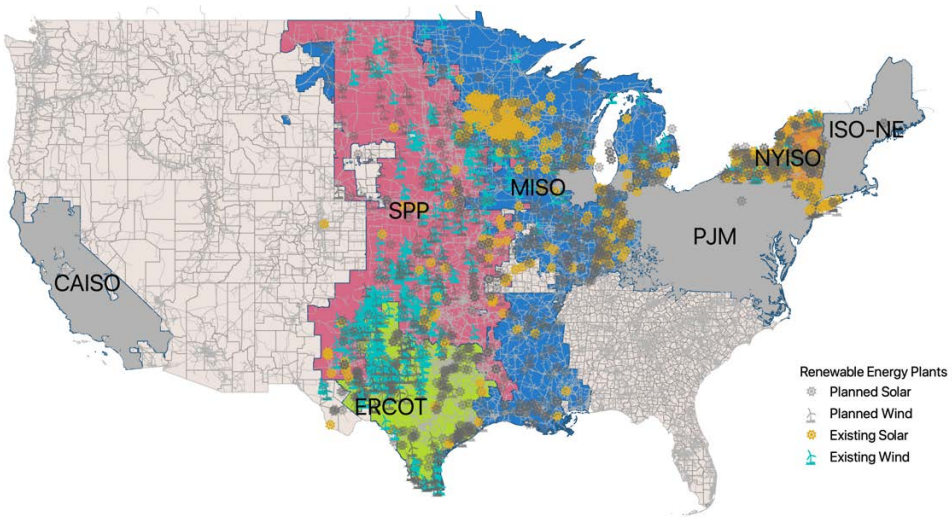



**Figure 3: Spatial distribution of solar and wind plants contained in these datasets.**

### *Treatment of NYISO Offshore Wind Sites*

For the proposed offshore wind sites in NYISO, the location provided in the interconnection queue data typically includes the point of interconnection on land and not the physical location of the turbines. To capture the profiles of the turbines correctly, the locations of these sites were updated. This was done by identifying the proposed Bureau of Ocean Management (BOEM) wind lease zone for each proposed project, either by manual matching using information from the project's website or from lease award information provided by BOEM. A summary of the planned wind sites identified as offshore wind projects and matched to BOEM leasing areas is provided in Table 4. A map depicting the location of existing and planned NYISO sites—including the revised location for offshore projects—is shown in Figure 4.

**Table 4: Summary of NYISO offshore wind projects.**

| Plant name | NYISO Zone | Capacity (MW) | BOEM Lease Area Name | BOEM Lease Area Number |
|---|---|---|---|---|
| **NY Wind - East Garden City** | K | 1272 | Bay State Wind | OCS-A 0500 |
| **NY Wind - Mott Haven** | J | 1272 | Bay State Wind | OCS-A 0500 |
| **NY Wind - Pilgrim** | K | 1276 | Bay State Wind | OCS-A 0500 |
| **NY Wind Gowanus** | J | 1200 | Bay State Wind | OCS-A 0500 |
| **NY Wind Holbrook** | K | 880 | Bay State Wind | OCS-A 0500 |
| **NY Wind Holbrook 2** | K | 1974 | Bay State Wind | OCS-A 0500 |
| **Vineyard Wind I** | K | 1403 | Vineyard Wind 1 | OCS-A 0501 |
| **Vineyard Wind II** | K | 1403 | Vineyard Wind 2 | OCS-A 0501 |
| **EI East Shoreham** | K | 1300 | Empire Wind | OCS-A 0512 |
| **EI Fort Salonga** | K | 1300 | Empire Wind | OCS-A 0512 |
| **EI Glenwood Landing** | K | 1300 | Empire Wind | OCS-A 0512 |
| **EI Melville** | K | 816 | Empire Wind | OCS-A 0512 |
| **EI Oceanside** | K | 1000 | Empire Wind | OCS-A 0512 |
| **EI Oceanside 2** | K | 500 | Empire Wind | OCS-A 0512 |
| **EI Steinway 1** | J | 1300 | Empire Wind | OCS-A 0512 |
| **EI Steinway 2** | J | 1300 | Empire Wind | OCS-A 0512 |
| **EI Sunset Park** | J | 816 | Empire Wind | OCS-A 0512 |
| **South Fork Wind Farm** | K | 96 | South Fork Wind | OCS-A 0517 |
| **South Fork Wind Farm II** | K | 40 | South Fork Wind | OCS-A 0517 |
| **Long Island Offshore Wind** | K | 1200 | Mayflower | OCS-A 0521 |
| **Long Island Offshore Wind Connection** | K | 800 | Mayflower | OCS-A 0521 |
| **New York City Offshore Wind** | J | 1200 | Mayflower | OCS-A 0521 |
| **East Wind 1** | K | 1200 | OW Ocean Winds East | OCS-A 0537 |
| **Atlantic Shores Offshore Wind 7** | K | 880 | Atlantic Shores Offshore Wind Projects | OCS-A 0541 |
| **Atlantic Shores Offshore Wind 8** | J | 880 | Atlantic Shores Offshore Wind Projects | OCS-A 0541 |
| **Atlantic Shores Offshore Wind 9** | J | 880 | Atlantic Shores Offshore Wind Projects | OCS-A 0541 |

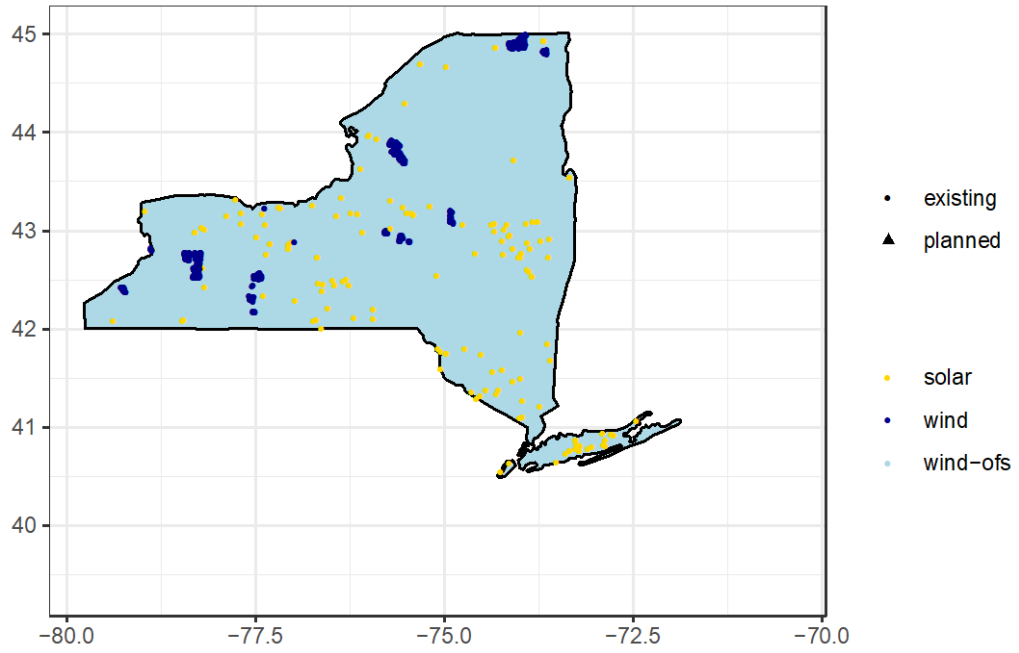



**Figure 4: Depiction of NYISO sites, including revised location for offshore wind sites.**

### *Classification of "likely" proposed projects*

Although the interconnection queues represent possible future projects, many of the projects in the queue may not actually be built. The final update to the site metadata was to include information on the likelihood that planned projects from the interconnection queue would ultimately be built and connected. This addition came at the request of the PERFORM data teams, who noted that the buildouts implied by the full interconnection set would lead to a very large system that may be difficult to manage and to develop into converged power flow cases. For the final PERFORM dataset we include all projects in the queue but add a flag for each project given some indication of its status in the queue, which can help users of the data potentially identify subsets of the projects for their own analyses.

To this end, the data team drew upon research from a Lawrence Berkeley National Laboratory study which categorizes the progress of the interconnection agreement (IA) for each proposed site (Rand et al. 2022). The categorization scheme provide includes the following levels:

- *IA Executed*: an interconnection agreement has been executed for the project.
- *In Progress*: negotiation on an interconnection agreement is underway between the project developer and the ISO.
- *Not Started*: the process for negotiating an interconnection agreement has not yet begun.
- *Unknown*: no data available on the status of any interconnection agreement.

Although these categories are not definitive indicators—some plants with executed IAs may not be built, whereas other plants with no information may be built—in general these categories indicate decreasing likelihood of deployment, with "IA Executed" indicating projects that are most likely to be completed and "Not Started/Unknown" indicating projects that are least likely to be built.

Figure 5 provides a summary of the breakdown of planned projects by interconnection status (with offshore and land-based wind projects summarized together as wind for NYISO). Although the breakdown of projects by IA status varies by ISO, we hope this may provide a useful tool for the PERFORM teams and other users when looking to prioritize projects with the highest likelihood of success.

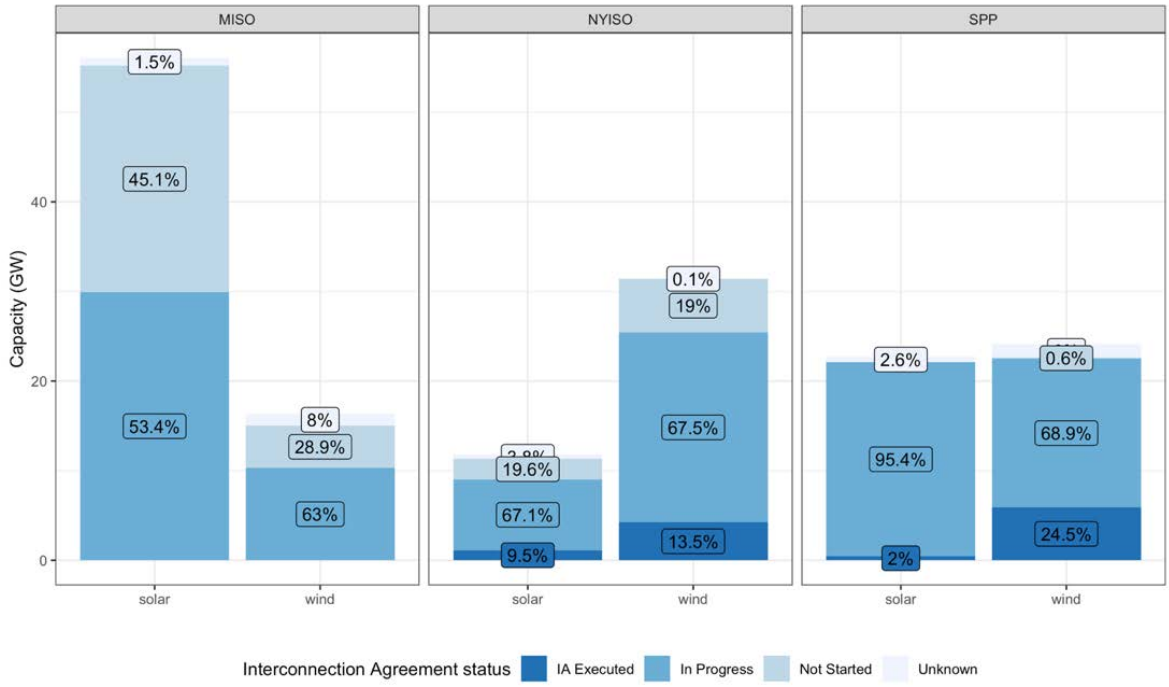



**Figure 5: Summary of planned sites by progress of interconnection agreement.**

## 2.2 Wind and Solar Actuals

To compute the "actual" renewable generation from wind and solar plants in 2018-2019, wind and solar resource data are needed. As with the ERCOT data produced during Phase I, the solar resource is produced using NREL's Physical Solar Model (PSM) and was derived from the 5-min National Solar Radiation Database (NSRDB) dataset that covers the Contiguous United States (CONUS). A detailed description of PSM and the NSRDB can be found in Sengutpa et al. (Sengupta et al. 2022), The average Global Horizontal Irradiance for the continental U.S. (CONUS) is illustrated in Figure 6.


**Figure 6: Average GHI from 2019 NSRDB.**

Wind resource data for 2019 was produced using the Weather Research and Forecasting model (WRF) following a similar procedure to the 2018 wind resource data created during Phase I. The only difference is that for Phase II the WRF was not run for all of CONUS, but instead as two separate models: 1) Covering NYISO and 2) Covering MISO, SPP, and ERCOT. The WRF model was set up using the parameters determined by Optis et al. 2020 and was seeded with the European Centre for Medium-Range Weather Forecasts Reanalysis Version 5 (ERA-5) dataset (Optis et al. 2020). The average 100m wind speed for NYISO and other three ISOs (MISO, SPP, and ERCOT) is illustrated in Figure 7 and Figure 8, respectively.

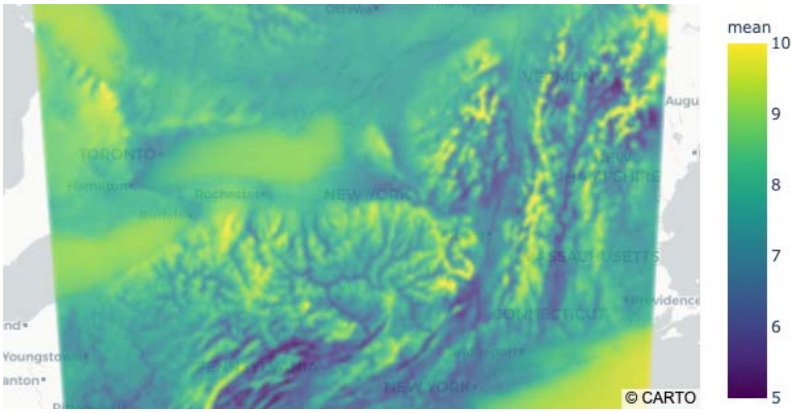



**Figure 7: Average 100m Wind Speed for 2019 over NYISO.**

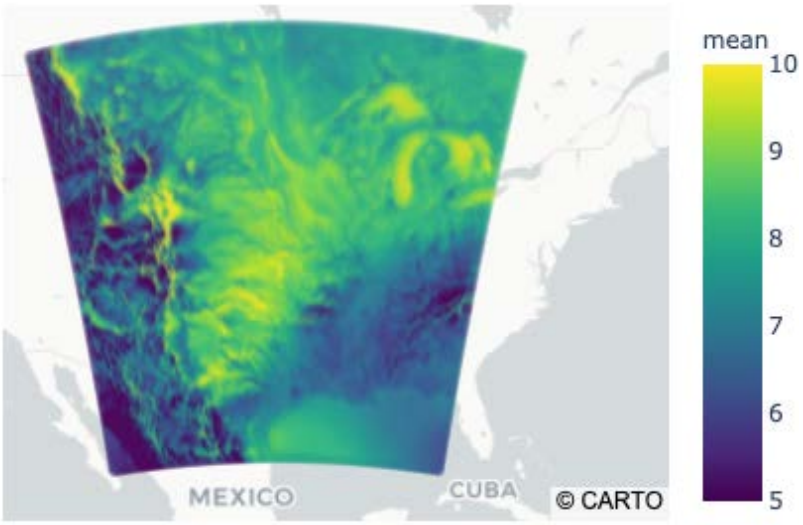



**Figure 8: Average Windspeed at 100m in 2019 over MISO, SPP, and ERCOT.**

To create "actual" generation profiles representative of real power plants using the NSRDB and WRF source data, we perform the following process:

- Run reV for the full spatial extent of an ISO (e.g., all of NYISO) all the way through the reV aggregation step where spatial exclusions will be applied. This step determines the available the technical potential from which the plants are built.
- Parse the locations and technologies of wind and solar plants for the ISO of interest and sample the currently existing plant's technology to create technology specifications for the planned buildouts (e.g., the AC-DC ratio for planned PV installations will be randomly selected from the existing installations).
- Run the reVX plant builder code to assign available wind and solar resource (based on the reV output from #1) to plants based on the actual plant capacity.
- Rerun reV at the final temporal resolution using the plant-specific technology assignments. The reV output from this step will be aggregated to the final plant profiles.
- Rerun the reVX plant builder using the reV generation profiles output from the previous step as input to create the final generation profiles for each plant.

Examples of power output profiles from this process for 5 solar and 5 wind plants in NYISO are shown in Figure 9 and Figure 10.

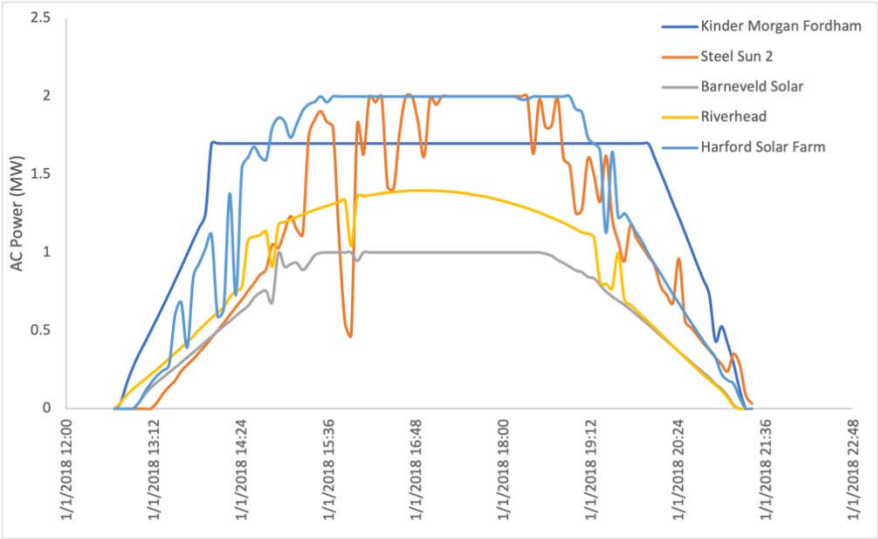



**Figure 9: AC power generation for five solar plants in the NYISO region for 1 day.**

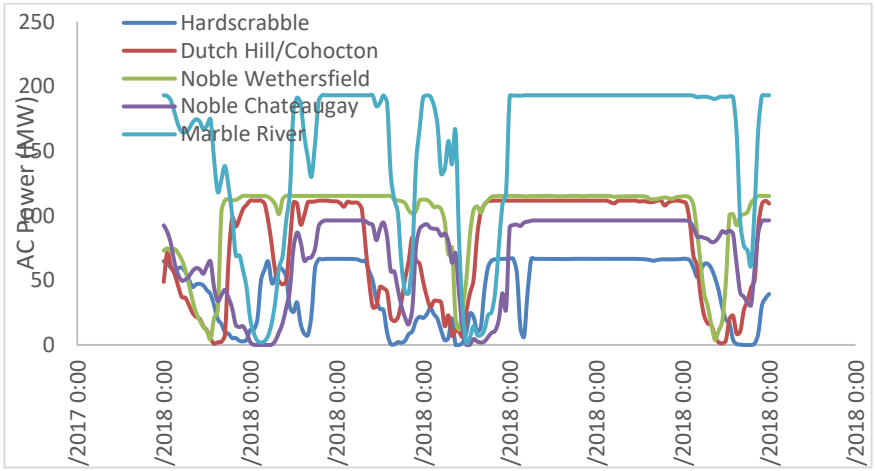



**Figure 10: AC power generation for five wind plants in the NYISO region for 1 day.**

## 2.3 Wind and Solar Forecasts

### *Deterministic Forecasts*

To generate the deterministic forecast, we rely on the European Centre for Medium-Range Weather Forecasts (ECMWF) weather forecast data, made up of 51 ensemble members. The ECMWF's Ensemble Prediction System (EPS) represents uncertainty in initial conditions by creating a set of 50 forecasts (the perturbed ensemble) starting from slightly different states that are close, but not identical, to our best estimate of the initial state of the atmosphere (the control group). Each forecast is based on a model which is close, but not identical, to their best estimate of the model equations, thus representing the influence of model uncertainties on forecast error. The divergence, or spread, of the control plus 50 perturbed forecasts gives an estimate of the uncertainty of the prediction on that day.

Two years (i.e., 2018 and 2019) of day-ahead and intra-day forecasts with an hourly resolution were downloaded. Table 5 shows the list of ECMWF meteorological parameters, which serve as

input into the reV model for solar/wind power modeling. The ECMWF forecasts of NYISO, MISO, and SPP were downloaded in a single regional request, as shown in Figure 11. The ECMWF download requests were optimized to minimize the total time required (queueing time + downloading time).

**Table 5: ECMWF parameters.**

| Name | Category | Units |
|---|---|---|
| **Direct solar radiation** | Irradiance | J m-2 |
| **Surface solar radiation downwards** | Irradiance | J m-2 |
| **Total sky direct solar radiation at surface** | Irradiance | J m-2 |
| **Clear-sky direct solar radiation at surface** | Irradiance (Clearsky) | J m-2 |
| **10 metre U wind component** | Wind | m s-1 |
| **10 metre V wind component** | Wind | m s-1 |
| **100 metre U wind component** | Wind | m s-1 |
| **100 metre V wind component** | Wind | m s-1 |
| **Surface pressure** | Pressure | hPa |
| **2 metre temperature** | Temperature | K |
| **2 metre dewpoint temperature** | Temperature | K |

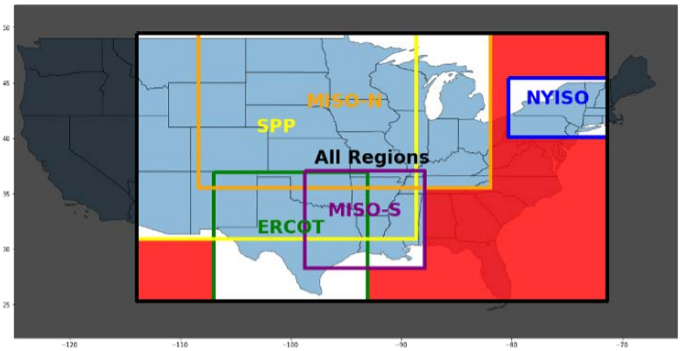



**Figure 11: Bounding boxes for the four ISOs.**

We generated forecasted plant power generation profiles using the ECMWF solar and wind resource data for MISO, SPP, and NYISO. The process is identical to the steps outlined above to create the "actuals" power generation profiles but instead of using the NSRDB and WRF data we use the ECMWF solar and wind forecast data. The result is 153 forecast profiles to accompany the "actual" profiles: intraday, 1-day-ahead, and 2-day-ahead forecasts for the control model and 50 perturbed forecast models (3+3x50=153).

The format of the forecast profile outputs is the same as the actuals (one profile for every plant) but with a different time index to account for the different temporal resolution of the source ECMWF data. Examples are shown below of solar and wind profiles at a single plant comparing the actuals to several of the forecasts. Not all 153 forecast profiles are shown in a single plot for

clarity. As we can see from both plots, the deterministic forecasts are biased and tend to underestimate the power. The probabilistic forecasts we generate in the following section will post process the deterministic forecasts to mitigate both bias and under-dispersion issues.

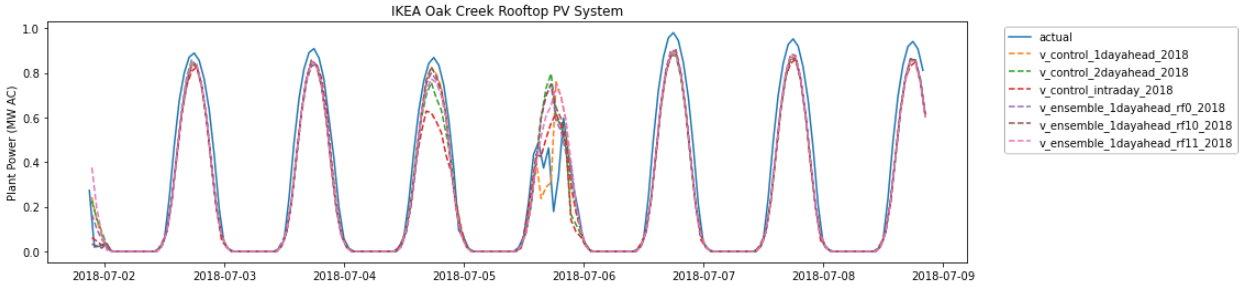



**Figure 12: PV power generation (actuals and forecasts) for the IKEA Oak Creek Rooftop PV system in the MISO domain.**

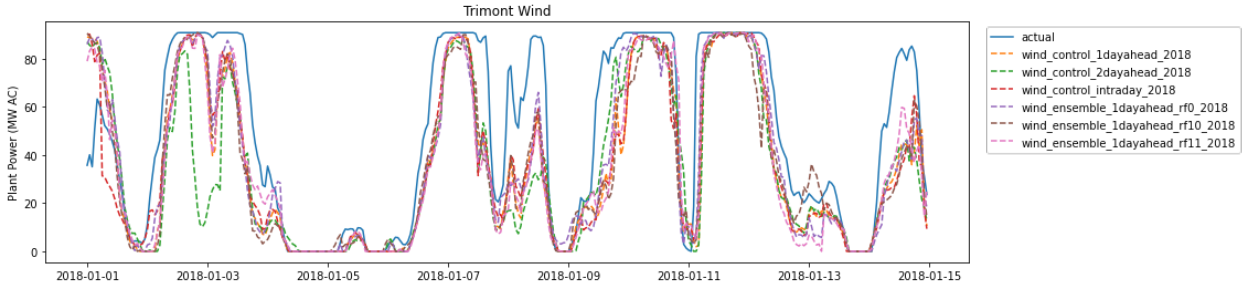



**Figure 13: Wind power generation (actuals and forecasts) for the Trimont Wind Plant in the MISO domain.**

## *Probabilistic Forecasts*

Based on the ECMWF-derived solar and wind power forecasts, the Bayesian Model Averaging (BMA) and the machine learning-based multi-model (M3) methods are used to generate probabilistic solar and wind forecasts at intraday/day-ahead, and hour-ahead timescales, respectively. This section describes each of those approaches in turn.

For the intraday and day-ahead probabilistic solar and wind power forecasting we use Bayesian Model Averaging (BMA) approach. BMA is a kernel dressing technique that applies a probability density to each member of a numerical weather prediction (NWP) ensemble, with each member dressed in a mixture model: each model includes a discrete component forecasting power clipped at the inverter rating plus a continuous kernel for outputs that are less than the rated maximum (Doubleday et al. 2021). As is appropriate for this method, we gathered perturbed forecast data and control forecast NWP ensemble data from ECMWF. These NWP ensembles are then post-processed with BMA to address weaknesses and to smooth the ensemble from a discrete set of points to a full cumulative distribution function (CDF), mitigating bias and under-dispersion typically found in these ensembles.

The intra-day and day-ahead forecasting errors and scores use BMA with Beta kernels for solar power forecasting. Similar to the solar power forecasting, we apply BMA to the intraday and day-ahead probabilistic wind power forecasting. The only difference is that Gaussian kernels work better than Beta kernels for wind power forecasting.

We generate probabilistic forecasts from 1st to 99th percentiles for NYISO, MISO and SPP. The overall forecasting performance for selected solar sites and wind sites in NYISO, MISO and SPP is shown in Table 6 - Table 11. The forecasts are scored on these metrics using a range of metrics, including normalized root-mean-squared error (nRMSE), normalized mean absolute error (nMAE), mean bias error (MBE)—where a negative MBE indicates predictions are generally lower than actuals and a positive MBE indications prediction are higher than actuals (the 50th percentile is used to calculate deterministic forecast error metrics). Finally, we also apply the continuous ranked probability score (CRPS)[1]. Based on the metrics, the probabilistic forecasts are reliable while our forecasts are accurate from the deterministic perspective.

**Table 6: Forecasting accuracy for selected NYISO solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| Baer_Road_CSG | 8.12 | 4.85 | 0.095 | 3.01 |
| Barneveld_Solar | 8.02 | 4.93 | 0.028 | 2.99 |
| Brookside_Solar | 7.68 | 4.82 | 3.015 | 3.21 |
| Harford_Solar_Farm | 7.89 | 4.50 | 0.098 | 3.34 |
| Hobart_William_Smith_College_Gates_Rd | 6.69 | 3.58 | 0.006 | 2.48 |
| Kinder_Morgan_Fordham | 7.15 | 3.76 | 0.078 | 2.88 |
| Madison_County | 7.88 | 4.62 | 0.092 | 3.25 |
| Minisink_Solar_1_LLC | 7.95 | 4.65 | 0.085 | 3.28 |

**Table 7: Forecasting accuracy for selected MISO solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| FastSun_14_CSG | 8.25 | 4.89 | 0.039 | 3.57 |
| GRE_Marshan_Solar | 8.34 | 4.87 | 0.050 | 3.15 |
| IKEA_Oak_Creek_Rooftop_PV_System | 8.36 | 4.98 | 0.068 | 3.18 |
| McLeod_Community_Solar_One_LLC_CS | 7.89 | 4.68 | 0.055 | 3.08 |
| Paynesville_Community_Solar | 8.05 | 4.85 | 0.223 | 3.16 |
| Staunton | 7.57 | 4.24 | 0.164 | 2.98 |
| TCLP_Solar_Phase_1 | 8.26 | 4.56 | 0.052 | 3.15 |
| Western_Michigan_Solar_Gardens | 8.31 | 4.75 | 0.058 | 3.22 |

[1] CRPS is a metric to quantify the difference of two distributions, $CRPS(F,x) = \int_{-\infty}^{\infty}(F(y) - \mathbb{1}\ (y-x)\ )^2 dy$

**Table 8: Forecasting accuracy for selected SPP solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| FastSun_14_CSG | 8.12 | 4.25 | 0.0312 | 3.05 |
| GRE_Marshan_Solar | 7.95 | 3.89 | 0.0305 | 2.86 |
| IKEA_Oak_Creek_rooftop_PV_System | 7.67 | 4.02 | 0.0386 | 2.95 |
| McLeod_Community_Solar_One_LLC_CSG | 7.21 | 3.85 | 0.0289 | 2.76 |
| Paynesville_Community_Solar | 7.85 | 4.09 | 0.1524 | 2.98 |
| Staunton | 7.08 | 3.61 | 0.0968 | 2.58 |
| TCLP_Solar_Phase_1 | 7.42 | 3.86 | 0.0295 | 2.86 |
| Western_Michigan_Solar_Gardens | 7.92 | 3.97 | 0.0313 | 2.97 |

**Table 9: Forecasting accuracy for selected NYISO wind sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| Hardscrabble | 6.15 | 4.12 | 1.25 | 2.85 |
| Dutch Hill/Cohocton | 8.12 | 4.32 | 2.94 | 3.12 |
| Noble Wethersfield | 7.43 | 4.14 | 2.89 | 2.96 |
| Noble Chateaugay | 6.87 | 4.05 | 2.45 | 2.93 |
| Marble River | 7.96 | 3.98 | 5.21 | 2.98 |
| Steel Winds II | 6.50 | 3.82 | 0.58 | 2.56 |
| Noble Clinton | 7.24 | 4.18 | 3.24 | 2.95 |
| Zotos | 3.25 | 2.85 | 0.05 | 2.15 |

**Table 10: Forecasting accuracy for selected MISO wind sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| Highland I | 4.23 | 3.58 | 3.40 | 2.94 |
| Cross Winds Energy Park III | 4.15 | 3.45 | 0.58 | 2.86 |
| Fenton | 4.20 | 3.48 | 1.25 | 3.05 |
| Pleasant Valley | 4.18 | 3.50 | 1.28 | 3.10 |
| Velva Wind Farm | 3.95 | 3.27 | 0.32 | 2.79 |
| Redwood Falls (SMMPA) | 3.85 | 3.21 | 0.08 | 2.65 |
| Pomeroy | 4.12 | 3.80 | 3.56 | 3.13 |
| Trimont Wind | 4.05 | 3.78 | 3.12 | 3.08 |

**Table 11: Forecasting accuracy for selected SPP wind sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| Marshall Wind Energy | 5.26 | 4.27 | 1.28 | 2.25 |
| Creston Ridge | 5.14 | 4.23 | 0.29 | 2.38 |
| Oak Tree | 4.98 | 4.05 | 0.95 | 2.56 |
| Buffalo Dunes | 5.05 | 4.15 | 3.17 | 3.58 |
| Red Hills | 5.12 | 4.18 | 1.45 | 2.19 |
| Buckeye | 4.85 | 3.98 | 2.84 | 3.08 |
| Grant Wind | 4.96 | 4.02 | 2.24 | 2.99 |
| Little Elk | 5.02 | 4.13 | 1.24 | 2.06 |

For the hour-ahead forecasts we use the machine learning-based multi-model (M3) forecasting framework, which is a two-step data-driven methodology that provides both deterministic and probabilistic forecasts for very short-term wind, solar, and load forecasting (Feng et al. 2017; Feng and Zhang 2020; Feng et al. 2019). The M3 method showed superior performance than single algorithm machine learning models regarding the accuracy and robustness. The M3 method generates deterministic forecasts in the first step with a two-layer machine learning ensemble algorithm, which serves as the input to the pinball loss optimization-based predictive distribution model to generate probabilistic forecasts in the second step.

Specifically, in the deterministic forecasting step, a collection of machine learning models are used in the first layer, including three artificial neural networks (ANNs) with backpropagation, three support vector regression (SVR) models with different kernels, three gradient boosting machine (GBM) models with three different distribution functions, and a random forest (RF) model. These models generate independent deterministic forecasts, which will be ensembled in the second layer by another machine learning model to generate final deterministic forecasts. The ensemble model is expected to provide more accurate and robust deterministic forecasts than single-algorithm models.

Next, in the probabilistic forecasting step, a genetic algorithm is used to optimize the standard deviation of the predictive distributions, sigma, given the mean value (assumed to be deterministic forecast from the first step). An SVR surrogate model is first trained based on deterministic forecasts and sigma values, which is used to estimate pseudo sigma to generate the quantiles.

This approach was used to generate $1^{st}$-$99^{th}$ quantiles for site-level plants, zonal-level, and ISO-level solar and wind power forecasts for 2019, with the models trained using 2018 data. The predictors include calendar features and power lags, as well as ECMWF intra-day power forecasts. Table 12 - Table 14 list the overall forecasting evaluation metrics for selected solar sites in NYISO, MISO, and SPP. The forecasts represent state-of-the-art accuracy from both deterministic and probabilistic perspectives. Figure 14 - Figure 16 demonstrate the reliability of probabilistic forecasts for selected sites in NYISO, MISO, and SPP. From the reliability diagram, it is observed that the probabilistic forecasts are reliable, especially considering the varying characteristics of site-level solar power time series.

**Table 12: Forecasting accuracy for selected NYISO solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| Baer_Road_CSG | 7.03 | 3.55 | 0.0500 | 2.51 |
| Barneveld_Solar | 7.03 | 3.59 | 0.0253 | 2.51 |
| Brookside_Solar | 6.43 | 3.45 | 2.4404 | 2.62 |
| Harford_Solar_Farm | 6.97 | 3.51 | 0.0509 | 2.48 |
| Hobart_William_Smith_College_Gates_Rd | 5.62 | 2.82 | 0.0030 | 2.02 |
| Kinder_Morgan_Fordham | 6.20 | 2.98 | 0.0336 | 2.15 |
| Madison_County | 7.02 | 3.55 | 0.0458 | 2.50 |
| Minisink_Solar_1_LLC | 6.73 | 3.29 | 0.0400 | 2.37 |
| RIT_Henrietta_Solar_1_LLC | 6.99 | 3.50 | 0.0400 | 2.46 |
| Riverhead | 7.38 | 3.59 | 0.0366 | 2.54 |
| Steel_Sun_2_2303_III_4_Hamburg_Tpke | 7.06 | 3.52 | 0.0531 | 2.46 |
| Villa_Roma_Rd_4_CSG | 7.57 | 3.86 | 0.0572 | 2.74 |
| White_Creek_Solar | 6.54 | 3.49 | 3.3744 | 2.81 |

**Table 13: Forecasting accuracy for selected MISO solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| FastSun_14_CSG | 7.84 | 4.13 | 0.0284 | 2.91 |
| GRE_Marshan_Solar | 7.31 | 3.68 | 0.0268 | 2.57 |
| IKEA_Oak_Creek_Rooftop_PV_System | 7.58 | 3.80 | 0.0345 | 2.66 |
| McLeod_Community_Solar_One_LLC_CSG | 6.97 | 3.56 | 0.0252 | 2.48 |
| Paynesville_Community_Solar | 7.50 | 3.75 | 0.1237 | 2.64 |
| Staunton | 6.82 | 3.46 | 0.0826 | 2.46 |
| TCLP_Solar_Phase_1 | 7.23 | 3.60 | 0.0261 | 2.52 |
| Western_Michigan_Solar_Gardens | 7.62 | 3.76 | 0.0285 | 2.65 |

**Table 14: Forecasting accuracy for selected SPP solar sites.**

| Site | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| FastSun_14_CSG | 7.84 | 4.13 | 0.0284 | 2.91 |
| GRE_Marshan_Solar | 7.31 | 3.68 | 0.0268 | 2.57 |
| IKEA_Oak_Creek_Rooftop_PV_System | 7.58 | 3.80 | 0.0345 | 2.66 |
| McLeod_Community_Solar_One_LLC_CSG | 6.97 | 3.56 | 0.0252 | 2.48 |
| Paynesville_Community_Solar | 7.50 | 3.75 | 0.1237 | 2.64 |
| Staunton | 6.82 | 3.46 | 0.0826 | 2.46 |
| TCLP_Solar_Phase_1 | 7.23 | 3.60 | 0.0261 | 2.52 |
| Western_Michigan_Solar_Gardens | 7.62 | 3.76 | 0.0285 | 2.65 |

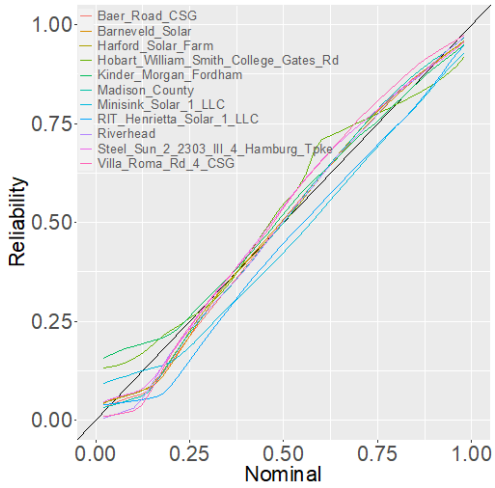



**Figure 14: Probabilistic forecast reliability diagram for selected NYISO solar sites.**

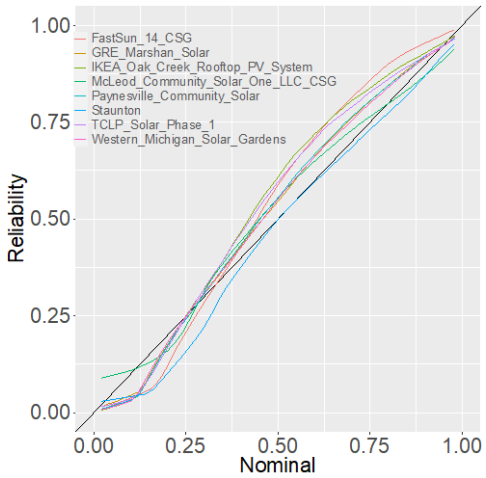


**Figure 15: Probabilistic forecast reliability diagram for selected MISO solar sites.**

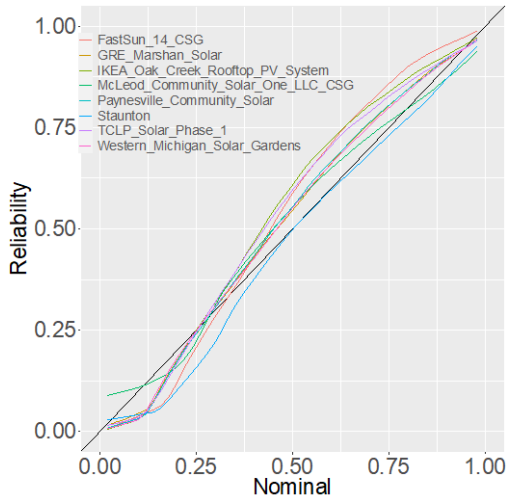



**Figure 16: Probabilistic forecast reliability diagram for selected SPP solar sites.**

# 3 Load Data and Forecasts

This section describes the data sources and processing steps for the load actuals and the load forecasts.

## 3.1 Load Actuals

### *NYISO*

Historical 5-min load data is publicly available from NYISO online (NYISO 2021). Data was downloaded by load zone region from 2010-2020 and then down selected to 2018-2019. Table 15 below provides summary statistics on the 5-min load data.

Approximately 0.04% of the 5-min time periods from 2018-2019 are missing from the raw NYISO 5-minute data that is available for download. The missing data gaps generally occur for an hour at a time (i.e., 12 consecutive 5 min periods are missing) and occur for all the load zones. An example of one of missing data gap is illustrated in Figure 17 below, whereas Figure 18 depicts which intervals are missing data.

**Table 15: Summary statistics on 5-min load data collected from NYISO for 2018-2019.**

| Zone | Missing (%) | Min | Mean | Max | SD |
|---|---|---|---|---|---|
| CAPITL | 0.044 | 830 | 1,357 | 2,452 | 245 |
| CENTRL | 0.044 | 1,187 | 1,832 | 3,104 | 287 |
| DUNWOD | 0.044 | 206 | 668 | 1,427 | 157 |
| GENESE | 0.044 | 717 | 1,114 | 2,078 | 202 |
| HUD VL | 0.044 | 538 | 1,078 | 2,286 | 242 |
| LONGIL | 0.044 | 1,430 | 2,248 | 5,472 | 654 |
| MHK VL | 0.044 | 489 | 900 | 1,436 | 162 |
| MILLWD | 0.044 | 95 | 311 | 685 | 79 |
| N.Y.C. | 0.044 | 3,753 | 5,902 | 11,110 | 1,265 |
| NORTH | 0.044 | 200 | 533 | 882 | 66 |
| WEST | 0.044 | 1,157 | 1,740 | 2,802 | 244 |
| Total NYISO | 0.044 | 11,817 | 17,725 | 31,943 | 3,349 |

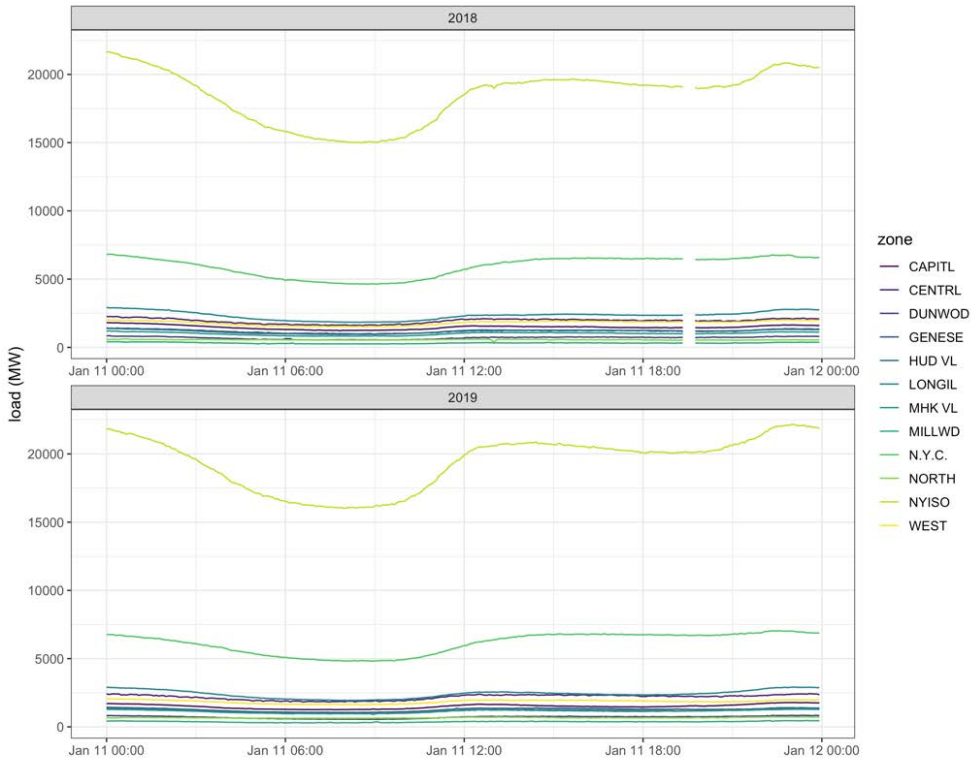



**Figure 17: Illustration of the gaps in the published NYISO load for an example time window in the 1st week of January in 2018 and 2019.**

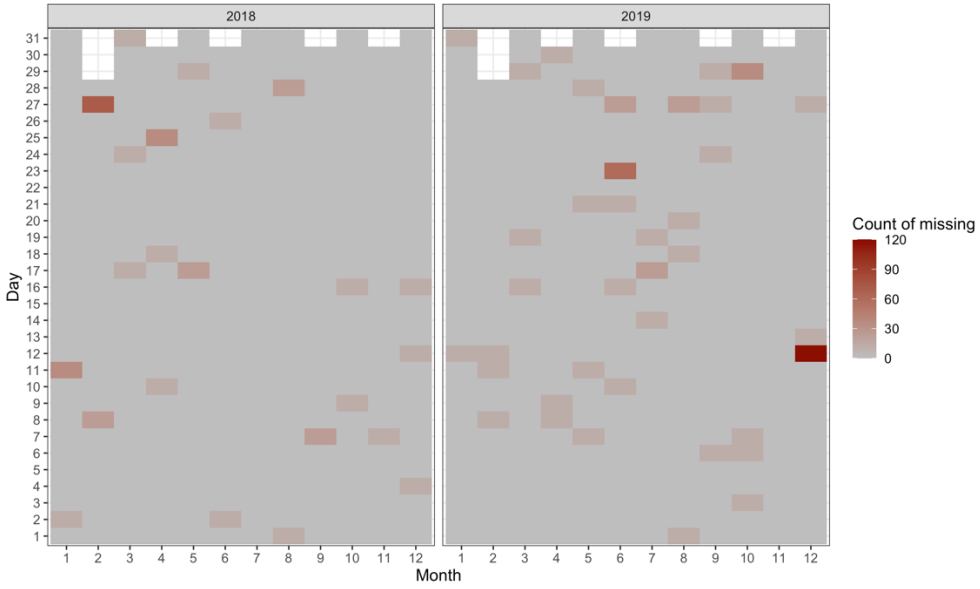



**Figure 18: Heatmap of the intervals of missing data for NYISO.**

To fill in these gaps, NREL used a method consisting of artificial neural network (ANN) ensemble and linear interpolation. While simple linear interpolation would readily fill the missing data, this approach would lack realistic variability. Instead, an ANN Ensemble comprised of 20 ANN models was developed and used to fill the missing data with realistic modelled data, combined with the linear interpolation to reduce model variability and address

biasing. It was determined that an ensemble comprised of 20 ANNs produced sufficient variability among the model outputs, as can be seen in Figure 19 below. Each ANM is trained on the same set of data comprised of ECMWF metrological intraday forecasts from three cells within each NYISO load zone, temporal variables, and the available 5-minute resolution load historical data collected during the years of 2018 and 2019. Different training seeds are applied to set up the initial weights of the ANN models, resulting in variability among the 20 ensemble ANNs. All missing time points within the 5-minute load data are filled using the modeled load data.

Each member model of the ensemble is structured as an ANN with a single hidden layer. An input layer is comprised of a row matrix of normalized scalar values corresponding to the mean of 51 distinct ECMWF ensemble members for intraday forecasts GHI and dry-bulb temperature from three coordinates within each load zone. The hourly intraday forecasts were interpolated to a 5-minute resolution to allow for training of a 5-minute resolution surjective model of the load data. In addition to these weather forecasts, minute of day, day of week, month of year, and year were including as model predictors.

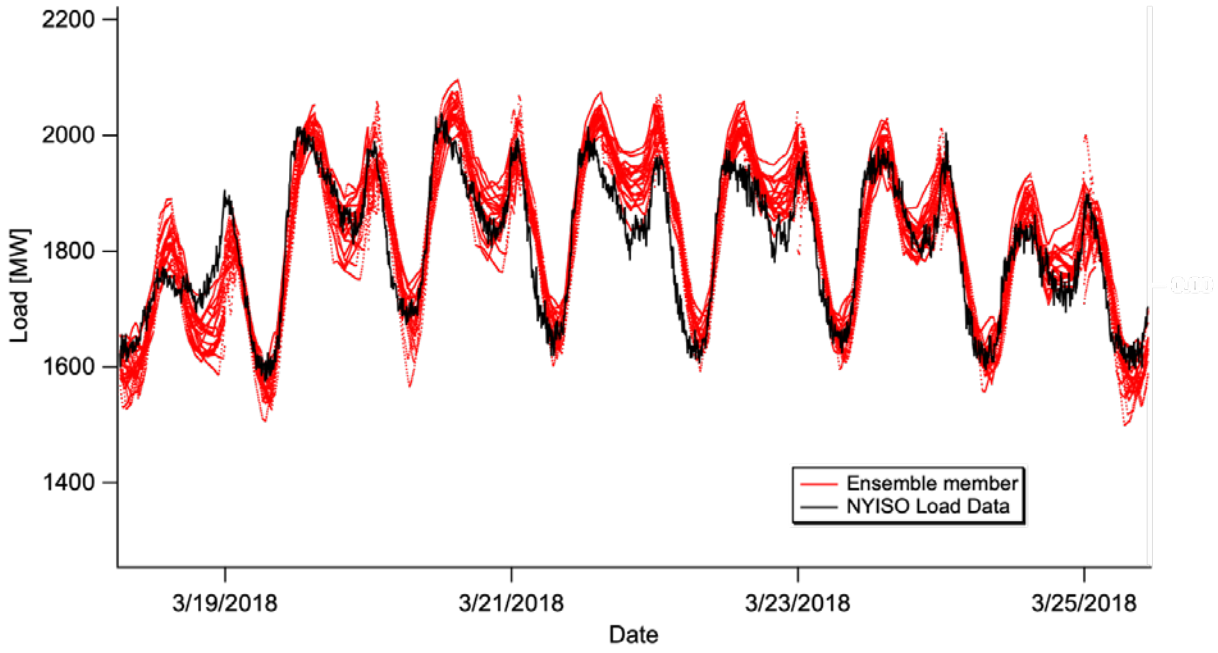



**Figure 19: ANN ensemble members and NYISO historic load data.**

Each ensemble member is trained using the MLPRegressor function in the scikit-learn Python library. Training hyperparameters were tuned to ensure that each ensemble member featured a correlation coefficient of greater than 90%. The hyper-parameters of the training algorithm are tabulated below:

**Table *16*: Hyper-parameters of the ANN training algorithm.**

| Parameter | Number of hidden nodes | Activation function | Solver | Alpha | Initial learning rate | Power | Maximum iterations | Shuffle | Random State | Tolerance | Learning rate method |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Value | 100 | 'relu' | 'adam' | 0.0001 | 0.001 | 0.5 | 2000 | 'True' | 'None' | 0.00001 | 'adaptive' |

The final resulting model output represented a weighted average of the mean model output from the ANN ensemble and a simple linear interpolation of the 5-minute load data. Figure 20 below shows the historic load data as well as model output from each member of the ANN ensemble. The ANN Ensemble mean and the linear interpolation are given equal weights because, without ground sensors, we cannot claim that the ANN ensemble is any closer to reality than we can assert that a direct linear interpolation is any closer to reality. The direct average between a linear interpolation and the ANN Ensemble mean provides realistic variability and reasonable values about the linear interpolation. This attenuated model featured variability larger than that of the historic load data whereas the simple linearly interpolated load data contained none of the load variability that is present in the historical load data. Thus, the attenuated model consisting of the weighted average of the ANN Ensemble and the linear interpolated load features more reasonable load variability than either method individually, as shown in Figure 20.

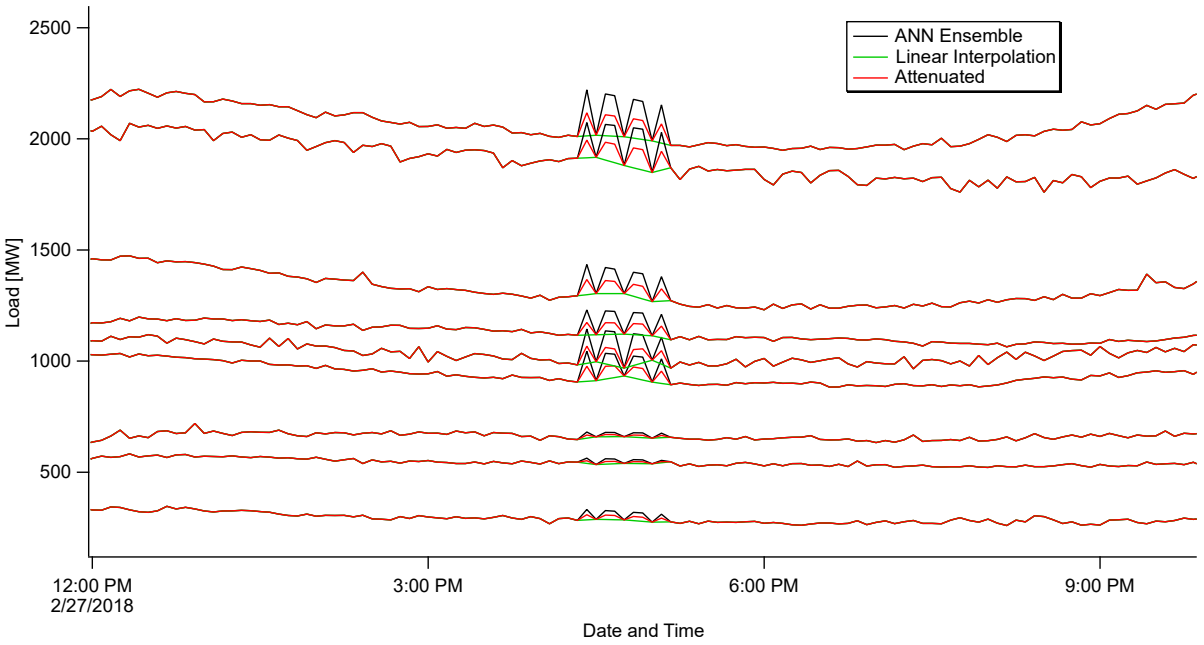



**Figure 20: Filled load data using ANN Ensemble, linear interpolation, and final attenuated results.**

## *MISO and SPP*

Neither MISO nor SPP have publicly available 5-min load data at the load zone level, although SPP does make ISO-wide 5-min load available. Accordingly, we developed synthetic load actuals by zone for these two ISOs by downscaling hourly data to 5-min resolution (SPP 2021; MISO

2021). For this purpose, we apply a downscaling method previously developed for downscaling solar profiles that relies on modeling Cholesky factors of autocovariance matrices (Zhang et al. 2022). The approach includes a nonstationary and non-Gaussian moving average model for the purposes of stochastic temporal load downscaling from hourly resolution to 5-min resolution.

The model developed with this method is a correlated mixture of Logistic random variables whose coefficients vary diurnally and seasonally. As direct calculations are inaccessible, we introduce an estimation approach exploiting empirical Cholesky factors and decorrelated residuals. The development, testing, and validation of the method was initially done on ERCOT and NYISO datasets since those ISOs have both 5-minute and hourly resolution data available. The model is then applied to the MISO and SPP hourly data to downscale to 5-min resolution.

Figure 21 illustrates the result of the downscaling method by comparing the heatmaps of the hourly and the downscaled MISO BA-level data from 2018 to 2019. In each heatmap, x-axis shows time of the day and y-axis represents day of the year. We can see that our stochastic downscaling method not only captures the diurnal and seasonal pattern shown in the hourly heatmap but also the interaction between diurnal and seasonal variability. This is related to the way we handle diurnal and seasonal variations in both mean and covariance structure in our downscaling framework.

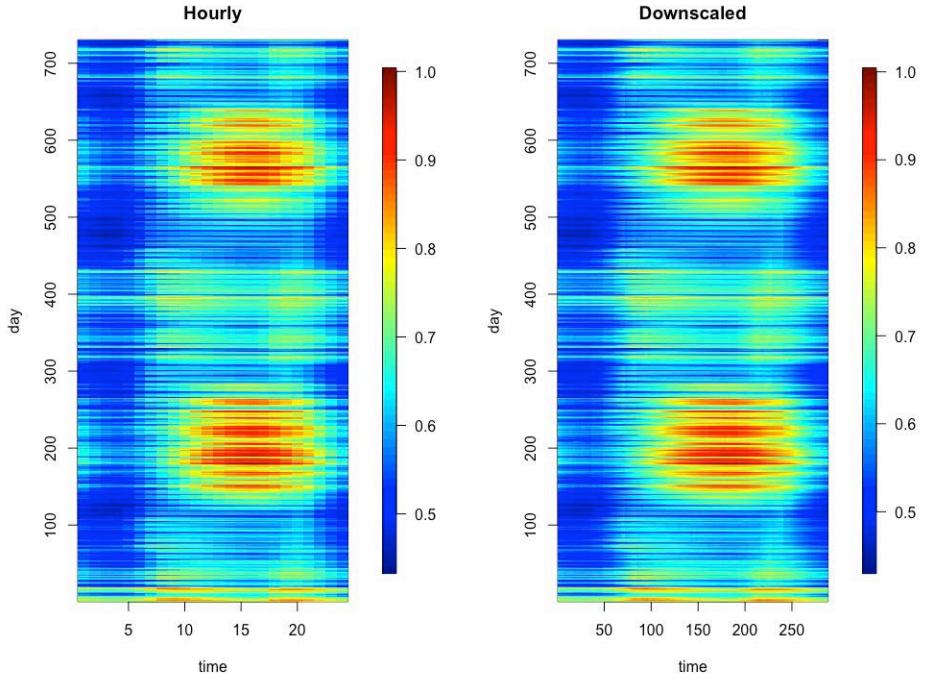



**Figure 21: Heatmaps of the hourly and downscaled MISO BA-level data.**

As an important validation, the reliability plot in Figure 22 contains boxplots of empirical coverage against nominal coverage for the NYISO load data, comparing the downscaled 5-min data from the model with the actual historical 5-min load data. Each coverage percentage is based on 1000 downscaled ensembles, which is repeated 100 times as represented by the boxplot. A perfectly calibrated simulation would follow the identity line; as can be seen, the median of each boxplot is

generally close to the identity line with slight overdispersion around medium nominal levels. This plot indicates that our model is well calibrated.

For MISO, we use the model trained on the NYISO load data to downscale the historical hourly load into 5-min load data. In the case of SPP, the ISO does make 5-min historical load data available for the entire ISO. Accordingly, we train the model on the SPP-wide 5-min load data and then use that model to generate 5-min load data for each of the load zones.

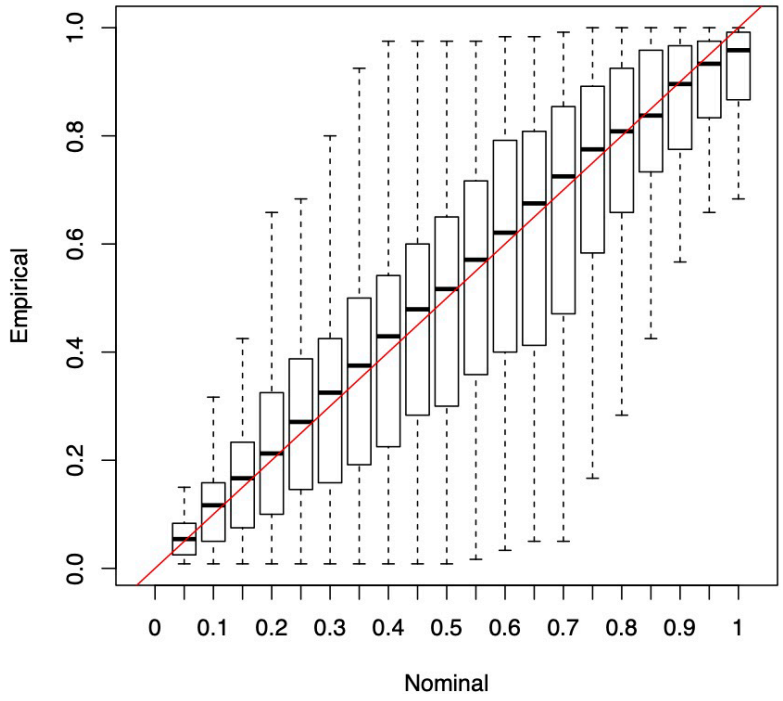



**Figure 22: Reliability plot for NYISO.**

## 3.2 Load Forecasts

### *Deterministic Load Forecasts*

The deterministic load forecasts are generated using a convolutional neural network (CNN), a supervised machine learning technique originally developed for images but that has been shown to work well for timeseries problems (Zhao et al. 2017). The CNN is used to generate a filter that is passed over input features, with the filter extracting patterns from the relevant features.

The CNN is trained on load actuals and forecasts from 2018 to produce forecasts for 2019. In addition to the load data, the CNN is given a range of relevant input features, including dry bulb temperature, humidity, day of the week, year, hour, month, and holiday. For the input weather features the CNN is provided both forecast data taken from the ECMWF forecasts and actuals based on data from three weather stations in each load zone.

### *Probabilistic Load Forecasts*

Machine learning training has some inherent variability depending on initial conditions which can affect forecast accuracy. To address this and develop a probabilistic load forecast, a total of 100

CNN models were trained to map the weather forecasts to load for each load zone and ISO. Each of the 100 CNN models were initialized with different seeds. A Gaussian distribution was fit for each timestamp based on the 100 forecasts from the CNN models to generate load forecast quantiles.

The probabilistic forecasts were scored based on sharpness—how well the deterministic forecasts agree—and reliability—how well the ensemble agrees with the observed values. The forecasts are scored on these metrics using a range of metrics, including normalized root-mean-squared error (nRMSE), normalized mean absolute error (nMAE), mean bias error (MBE)—where a negative MBE indicates predictions are generally lower than actuals and a positive MBE indications prediction are higher than actuals. Finally, we apply the continuous ranked probability score (CRPS) to measure both reliability and sharpness for each probabilistic forecast. In Table 17, the overall forecasting performance based on intraday load forecasts for all NYISO zones and the ISO as a whole is shown as a representative example for all forecasting horizons.

**Table 17: Forecasting accuracy based on intraday forecasts for all NYISO zones and BA.**

| Zone | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| CAPTIL | 3.58 | 2.85 | -15.21 | 2.00 |
| CENTRL | 3.39 | 2.65 | -25.17 | 1.87 |
| DUNWOOD | 3.60 | 2.70 | 1.16 | 1.93 |
| GENESE | 3.45 | 2.61 | -10.43 | 1.85 |
| HUD VL | 3.66 | 2.81 | -12.64 | 1.99 |
| LONGIL | 3.60 | 2.72 | -28.85 | 1.97 |
| MHK VL | 3.91 | 3.11 | -1.48 | 2.21 |
| MILLWD | 4.28 | 3.26 | 1.95 | 2.39 |
| N.Y.C. | 2.99 | 2.26 | -47.85 | 1.61 |
| NORTH | 3.47 | 2.75 | 9.95 | 1.97 |
| WEST | 3.52 | 2.73 | -46.31 | 1.93 |
| ALL NYISO | 2.90 | 2.30 | -179.17 | 1.64 |

**Table 18: Forecasting accuracy based on intraday forecasts for all MISO zones and BA.**

| Zone | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| LRZ1 | 2.42 | 1.77 | 80.46 | 1.31 |
| LRZ2_7 | 2.29 | 1.51 | 239.04 | 1.17 |
| LRZ3_5 | 2.55 | 1.86 | 41.71 | 1.36 |
| LRZ4 | 2.86 | 2.06 | 23.61 | 1.50 |
| LRZ6 | 3.09 | 2.20 | 28.39 | 1.62 |
| LRZ8_9_10 | 2.89 | 2.08 | 114.41 | 1.52 |
| ALL MISO | 1.95 | 1.45 | 595.48 | 1.11 |

**Table19: Forecasting accuracy based on intraday forecasts for all SPP zones and BA.**

| Zone | nRMSE [%] | nMAE [%] | MBE [MW] | nCRPS [%] |
|---|---|---|---|---|
| CSWS | 3.38 | 2.41 | 38.80 | 1.76 |
| EDE | 4.75 | 3.23 | 1.39 | 2.46 |
| GRDA | 4.78 | 3.83 | -16.13 | 2.84 |
| INDN | 4.01 | 2.67 | 1.77 | 2.02 |
| KACY | 3.66 | 2.67 | 5.28 | 1.96 |
| KCPL | 3.93 | 2.73 | 26.39 | 2.04 |
| LES | 3.63 | 2.50 | 3.11 | 1.88 |
| MPS | 4.20 | 2.91 | 5.75 | 2.20 |
| NPPD | 2.78 | 2.08 | 9.44 | 1.50 |
| OKGE | 3.52 | 2.41 | -13.58 | 1.78 |
| OPPD | 3.31 | 2.18 | 1.58 | 1.63 |
| SECI | 2.84 | 2.16 | -5.72 | 1.59 |
| SPRM | 4.35 | 2.76 | 3.74 | 2.14 |
| SPS | 2.31 | 1.71 | -21.00 | 1.27 |
| WAUE | 2.50 | 1.93 | -49.07 | 1.40 |
| WFEC | 3.84 | 2.76 | -4.66 | 2.02 |
| WR | 3.33 | 2.35 | 2.92 | 1.72 |
| ALL SPP | 2.38 | 1.79 | -116.69 | 1.30 |

# 4 Dataset Structure and Access

Below we outline the final dataset structure for Phase I and Phase II. Documentation is available for the datasets at https://github.com/PERFORM-Forecasts/documentation. Note that all timeseries data are provided in UTC.

## 4.1 Phase I: The ARPA-E PERFORM ERCOT Dataset

The ERCOT dataset has one year (2018) renewable generation and load actuals and probabilistic forecasts. This data is provided at various spatial (i.e., site-level, zone-level, and system-level) and temporal scales (i.e., day-ahead, intra-day, and intra-hour). Specifically, data is provided for 125 existing wind sites, 22 existing solar sites, 139 proposed wind sites, and 204 proposed solar sites. The following variables are provided for ERCOT:

- Meta data (coordinates, capacity, and other configuration data):
  - 125 actual wind sites
  - 139 proposed wind sites
  - 22 actual solar sites
  - 204 proposed solar sites
- Actual generation profiles for 2017-2018 (power [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)
  - Load (zone-level and system-level)
- Probabilistic forecasting generation profiles for 2018 (power [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)
  - Load (zone-level and system-level)
- ECMWF deterministic weather forecasts for 2017-2018 (day-ahead and intra-day):
  - Day-ahead control member forecasts
  - Intra-day control member forecasts
- ECMWF generation deterministic forecasts for 2017-2018 (day-ahead and intra-day, [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)

## 4.2 Phase II: The ARPA-E PERFORM NYISO, MISO, and SPP Datasets

As with ERCOT, the datasets for the other three ISOs (NYISO, MISO, and SPP) are stored in hdf5 files and uploaded to the AWS repository. These datasets are structured comparably to the ERCOT dataset and include the following features:

- Meta data (coordinates, capacity, and other configuration data)
- Actual generation profile for 2018-2019 (power [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)
  - Load (zone-level and system-level)
- Probabilistic forecasting data for 2019 (power [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)
  - Load (zone-level and system-level)
- ECMWF deterministic weather forecasts for each ISO for 2018-2019 (day-ahead and intra-day):
  - Day-ahead control member forecasts
  - Intra-day control member forecasts
- ECMWF generation deterministic forecasts for 2018-2019 (day-ahead and intra-day, [MW]):
  - Wind power (site-level, zone-level, and system-level)
  - Solar power (site-level, zone-level, and system-level)

# 5 References


ARPA-E. 2020. “ARPA-E’s PERFORM Program.” Arpa-e.Energy.Gov. February 7, 2020. http://arpa-e.energy.gov/technologies/programs/perform.

Bryce, Richard, Grant Buster, Kate Doubleday, Cong Feng, Ross Ring-Jarvi, Michael Rossol, Flora Zhang, and Bri-Mathias Hodge. 2023. “Solar PV, Wind Generation, and Load Forecasting Dataset for ERCOT 2018: Performance-Based Energy Resource Feedback, Optimization, and Risk Management (P.E.R.F.O.R.M.).” NREL/TP-5D00-79498, 1972698, MainId:33724. https://doi.org/10.2172/1972698.

Doubleday, Kate, Stephen Jascourt, William Kleiber, and Bri-Mathias Hodge. 2021. “Probabilistic Solar Power Forecasting Using Bayesian Model Averaging.” *IEEE Transactions on Sustainable Energy* 12 (1): 325–37. https://doi.org/10.1109/TSTE.2020.2993524.

Feng, Cong, Mingjian Cui, Bri-Mathias Hodge, Siyuan Lu, Hendrik F. Hamann, and Jie Zhang. 2019. “Unsupervised Clustering-Based Short-Term Solar Forecasting.” *IEEE Transactions on Sustainable Energy* 10 (4): 2174–85. https://doi.org/10.1109/TSTE.2018.2881531.

Feng, Cong, Mingjian Cui, Bri-Mathias Hodge, and Jie Zhang. 2017. “A Data-Driven Multi-Model Methodology with Deep Feature Selection for Short-Term Wind Forecasting.” *Applied Energy* 190 (March): 1245–57. https://doi.org/10.1016/j.apenergy.2017.01.043.

Feng, Cong, and Jie Zhang. 2020. “Assessment of Aggregation Strategies for Machine-Learning Based Short-Term Load Forecasting.” *Electric Power Systems Research* 184 (July): 106304. https://doi.org/10.1016/j.epsr.2020.106304.

Hoen, Ben, James Diffendorfer, Joseph Rand, Louisa Kramer, Chris Garrity, and Hannah Hunt. 2023. “US Wind Turbine Database.” 2023. https://doi.org/10.5066/F7TX3DN0.

LBNL. 2022. “Utility-Scale Solar Database.” 2022. https://emp.lbl.gov/utility-scale-solar.

MISO. 2021. “MISO Market Data.” 2021. https://www.misoenergy.org/markets-and-operations/real-time--market-data/market-reports/#t=10&p=0&s=MarketReportPublished&sd=desc.

NYISO. 2021. “Load Data.” 2021. https://www.nyiso.com/load-data.

Optis, Michael, Oleksa Rybchuk, Nicola Bodini, Michael Rossol, and Walter Musial. 2020. “Offshore Wind Resource Assessment for the California Pacific Outer Continental Shelf (2020).” NREL/TP-5000-77642, 1677466, MainId:29568. https://doi.org/10.2172/1677466.

Rand, Joseph, Ryan Wiser, Will Gorman, Dev Millstein, Joachim Seel, Seongeun Jeong, and Dana Robson. 2022. “Queued Up: Characteristics of Power Plants Seeking Transmission Interconnection As of the End of 2021.” 2022. https://emp.lbl.gov/publications/queued-characteristics-power-plants-0.

Sengupta, Manajit, Yu Xie, Aron Habte, Grant Buster, Galen Maclaurin, Paul Edwards, Haiku Sky, Mike Bannister, and Evan Rosenlieb. 2022. “The National Solar Radiation Database (NSRDB) Final Report: Fiscal Years 2019-2021.” *Renewable Energy*.

SPP. 2021. “Hourly Load.” 2021. https://portal.spp.org/pages/hourly-load.

Zhang, Wenqi, William Kleiber, Bri-Mathias Hodge, and Barry Mather. 2022. “A Nonstationary and Non-Gaussian Moving Average Model for Solar Irradiance.” *Environmetrics* 33 (3): e2712. https://doi.org/10.1002/env.2712.

Zhao, Bendong, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. 2017. "Convolutional Neural Networks for Time Series Classification." *Journal of Systems Engineering and Electronics* 28 (1): 162–69. https://doi.org/10.21629/JSEE.2017.01.18.