Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operations (SUMMER-GO): Project Final Report

Stephen Jascourt,[1] Kate Doubleday,[2] Jie Zhang,[3] Cong Feng,[4] Anthony Florita,[4] Bri-Mathias Hodge[2,4]

1 Maxar Technologies
2 University of Colorado Boulder
3 University of Texas at Dallas
4 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-6A40-84147
April 2023

Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operations (SUMMER-GO): Project Final Report

Stephen Jascourt,[1] Kate Doubleday,[2] Jie Zhang,[3] Cong Feng,[4] Anthony Florita,[4] Bri-Mathias Hodge[2,4]

1 Maxar Technologies
2 University of Colorado Boulder
3 University of Texas at Dallas
4 National Renewable Energy Laboratory

Suggested Citation
Jascourt, Stephen, Kate Doubleday, Jie Zhang, Cong Feng, Anthony Florita, and Bri-Mathias Hodge. 2023. *Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operations (SUMMER-GO): Project Final Report*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-6A40-84147.
https://www.nrel.gov/docs/fy23osti/84147.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-6A40-84147
April 2023

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

NOTICE

This work was authored in part 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 U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office Systems Integration. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

List of Acronyms

BMA	Bayesian model averaging
BTM	behind the meter
CAISO	California Independent System Operator
CDF	cumulative distribution function
CRPS	continuous ranked probability score
DNI	direct normal irradiance
DRUC	day-ahead reliability unit commitment
ECMWF	European Centre for Medium-Range Weather Forecasts
EMS	energy management system
ERCOT	Electric Reliability Council of Texas
FTP	File Transfer Protocol
GEFS	Global Ensemble Forecast System
GFS	Global Forecast System
GHI	global horizontal irradiance
HRRR	High Resolution Rapid Refresh
HRUC	hourly reliability unit commitment
HSL	high sustained limit
HTML	HyperText Markup Language
ISO	independent system operators
MAE	mean absolute error
MISO	Midcontinent Independent System Operator
MSE	mean square error
NAM	North American Mesoscale Forecast System
NG-CT	natural gas combustion turbine
nMAE	normalized mean absolute error
NOAA	National Oceanic and Atmospheric Administration
nPL	normalized average pinball loss
NREL	National Renewable Energy Laboratory
nRMSE	normalized mean root mean square error
NSR	non-spinning reserve
NSRDB	National Solar Radiation Database
NSRR	non-spinning reserve requirement
NWP	numerical weather prediction
PDF	Portable Document Format
POA	plane of array
PV	photovoltaic
RAP	Rapid Refresh
RMS	root mean square
RMSE	root mean square error
RTED	real-time economic dispatch
RTO	regional transmission operators
SAM	System Advisor Model
SDG&E	San Diego Gas and Electric Company
SFA	Solar Forecast Arbiter

SREF	Short-Range Ensemble Forecast
SUMMER-GO	Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operation
SURFRAD	Surface Radiation Budget

Executive Summary

The Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operation (SUMMER-GO) project was recently completed through a collaboration among the National Renewable Energy Laboratory, Maxar, the Electric Reliability Council of Texas (ERCOT), the University of Texas at Dallas, the University of California Berkeley, and the University of Colorado Boulder. The project made significant advances in probabilistic solar power forecasting, both through the development of Bayesian model averaging methods for ensemble forecasting and in bringing these and other advancements into practice with Maxar's delivery of operational forecasts to ERCOT. In addition to creating more reliable solar power forecasts, the project developed methods for their utilization in power system operations. These include the development of risk-aware unit commitment and economic dispatch algorithms and methods to reformulate probabilistic forecasts to be used in these power system operational models. Dynamic power system reserve methods were also developed, which have been shown in silico to create economic savings and reliability improvements on an ERCOT-like system as well as financial savings in the ERCOT system through more granular consideration of the uncertainty associated with solar power forecasts. Finally, a situational awareness tool to help grid operators better understand solar power forecast uncertainty in daily operations was developed and extensively vetted.

Table of Contents

List of Figures

List of Tables

Background

The Electric Reliability Council of Texas (ERCOT) manages the transmission grid across most of the state of Texas with only small interconnection ties to other regions; therefore, it is almost islanded and needs to meet most of its own balancing needs. It has a peak demand of approximately 75 GW on the hottest summer days and the record cold winter days. As of late 2021, the weather-dependent renewable capacity of the utility-scale plants operating in ERCOT is approximately half this peak demand, although those resources might not be generating at or near capacity during the peak demand time. Accurate forecasts of wind and solar generation are crucial to reliable grid operations.

Although the capacity of wind power in ERCOT has exceeded 20 GW for approximately 5 years, solar capacity was less than 1 GW when the Solar Uncertainty Management and Mitigation for Exceptional Reliability in Grid Operation (SUMMER-GO) project was proposed, was less than 2 GW only 3 years ago, is now more than 8 GW, and will reach 20 GW in only 2 more years if all projects with signed interconnection agreements and financial security posted complete construction and begin operations. If all projects with signed interconnection agreements but without financial security posted yet were also to be completed, the solar generating capacity would reach 30 GW.

The actual utility-scale solar generation in ERCOT is shown in Figure 1 for the period from January 1, 2017, to September 30, 2021. It shows the daily 5-minute peak of the system-wide total. Each dot is a daily maximum, so the upper part of the envelope corresponds to sunny days, and the envelope dips with more scatter of cloudy days in winter. The exponential increase in the summer generation peaks is obvious. Noteworthy is that the exponential increase became so steep in 2021 that it masked the seasonal fluctuation—the upward trend continued instead of reaching a summer plateau.

At the start of SUMMER-GO, there were 18 solar photovoltaic (PV) units operating. By the end of SUMMER-GO, 72 units were operating and receiving real-time forecasts. The unit AC capacities are shown in Figure 2, and the locations are shown in Figure 3. The newer units are toward the right in Figure 2, showing that, overall, generators have been increasing in AC capacity, with several exceeding 200 MW and most near or exceeding 100 MW.

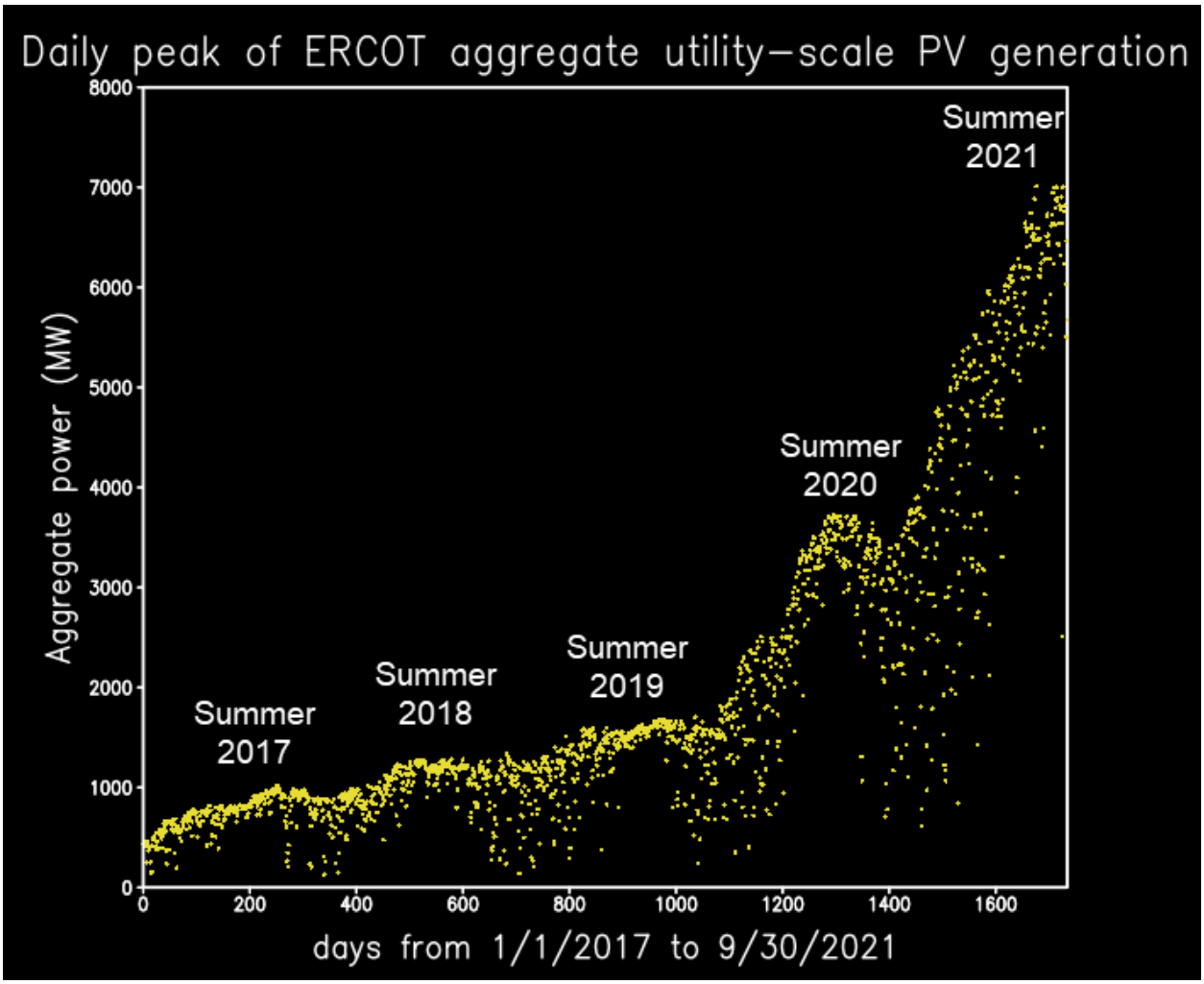


Figure 1. Daily 5-minute peak of aggregate utility-scale solar generation in ERCOT from 2017 to 2021

Most of the solar generating capacity is clustered in west Texas in the vicinity of Fort Stockton. The geographic spread shown in Figure 3 is a recent development. Even now, the capacity is still highly clustered. Figure 4 shows the capacity within 50 km (top) and 100 km (bottom) of a point for each point on the map, depicting the magnitude of this clustering. Figure 5 shows the cluster capacity as a function of the cluster radius in megawatts and as a percentage of the system-wide total solar generating capacity.

Small-scale weather features can have a large impact on the grid because of this clustering. With 2 GW within a 50-km radius and 4 GW within a 100-km radius, a thunderstorm complex confined to west Texas, such as one developing off the Davis Mountains or along the dry line that frequents that area, could rapidly shut down 2 GW to 4 GW of generation. These events can be difficult to predict in general and can be more difficult to predict with precise location and timing, resulting in large forecast errors, affecting grid operations.

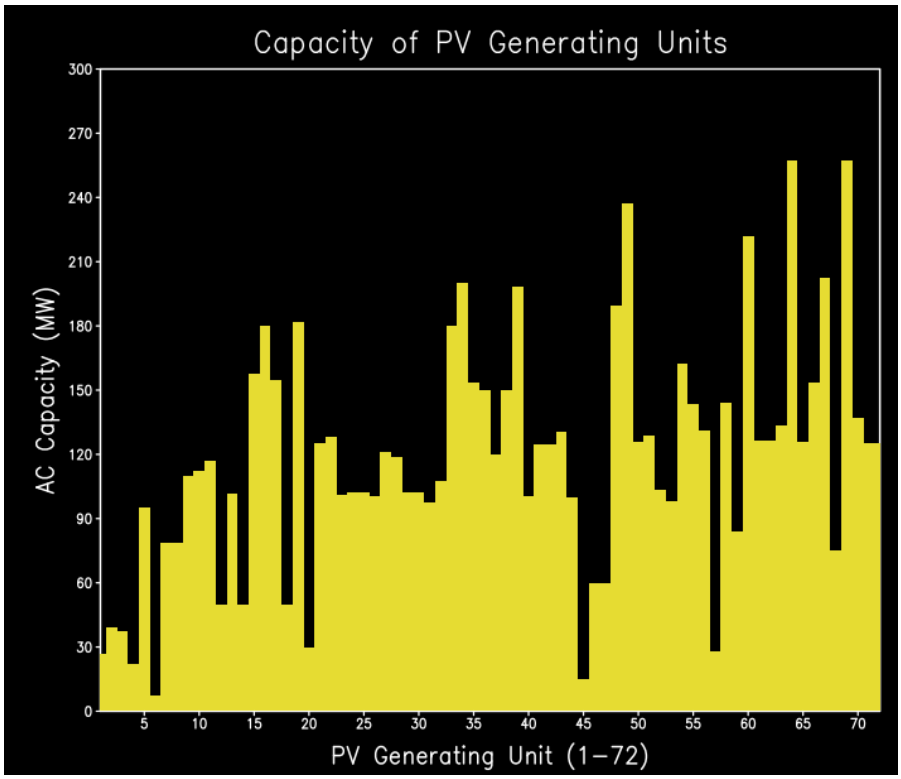


Figure 2. Capacity of utility-scale PV generating units operating as of September 30, 2021

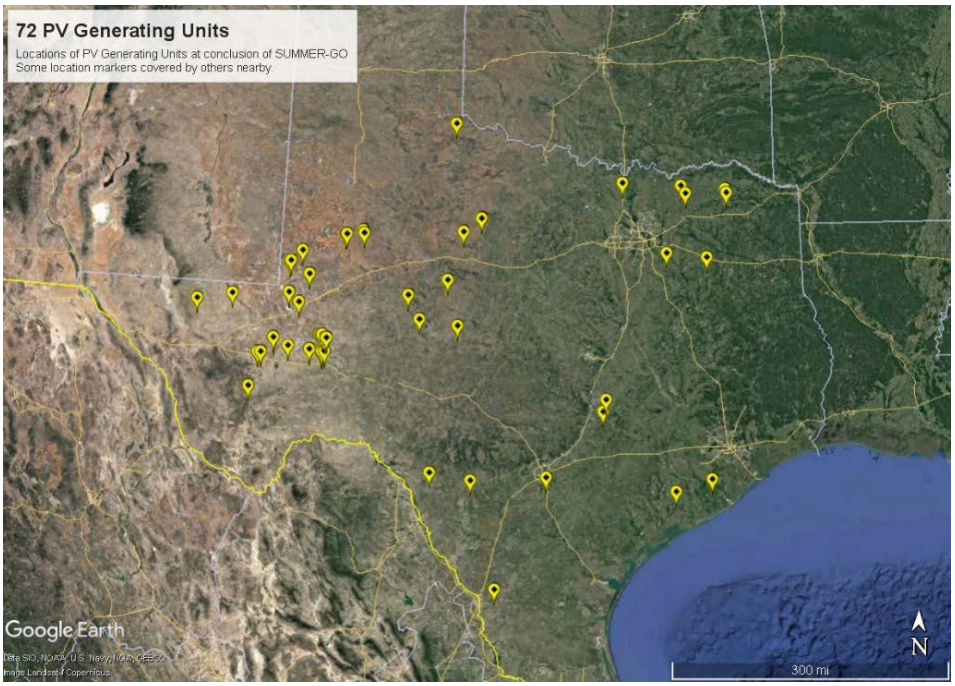


Figure 3. Locations of utility-scale PV generating units operating as of September 30, 2021. Some units are colocated or nearly colocated, resulting in overlapping or hidden location markers.

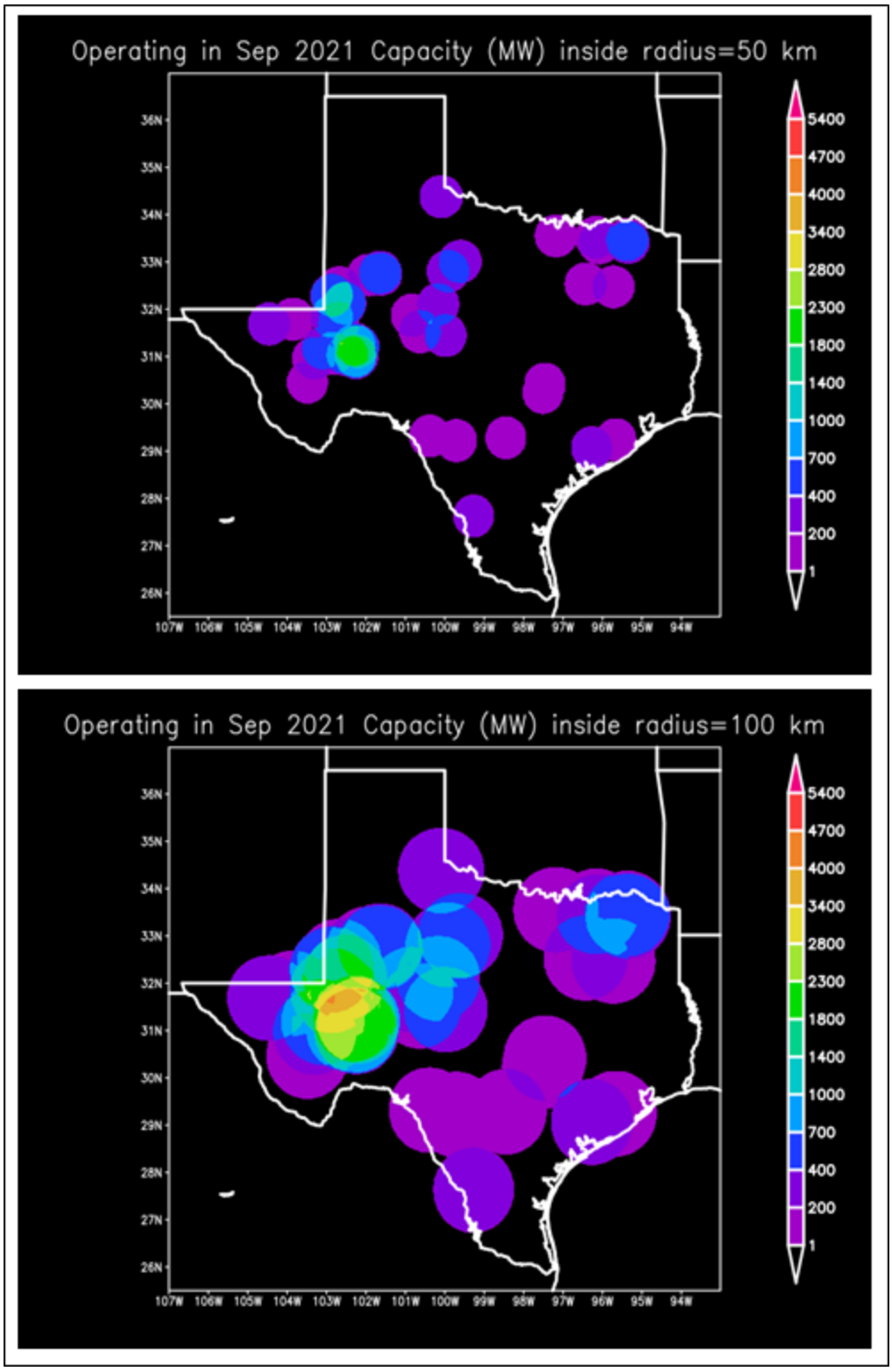


Figure 4. AC capacity of utility-scale solar power plants on the ERCOT grid clustered within 50 km (top) or 100 km (bottom) of each point on the map

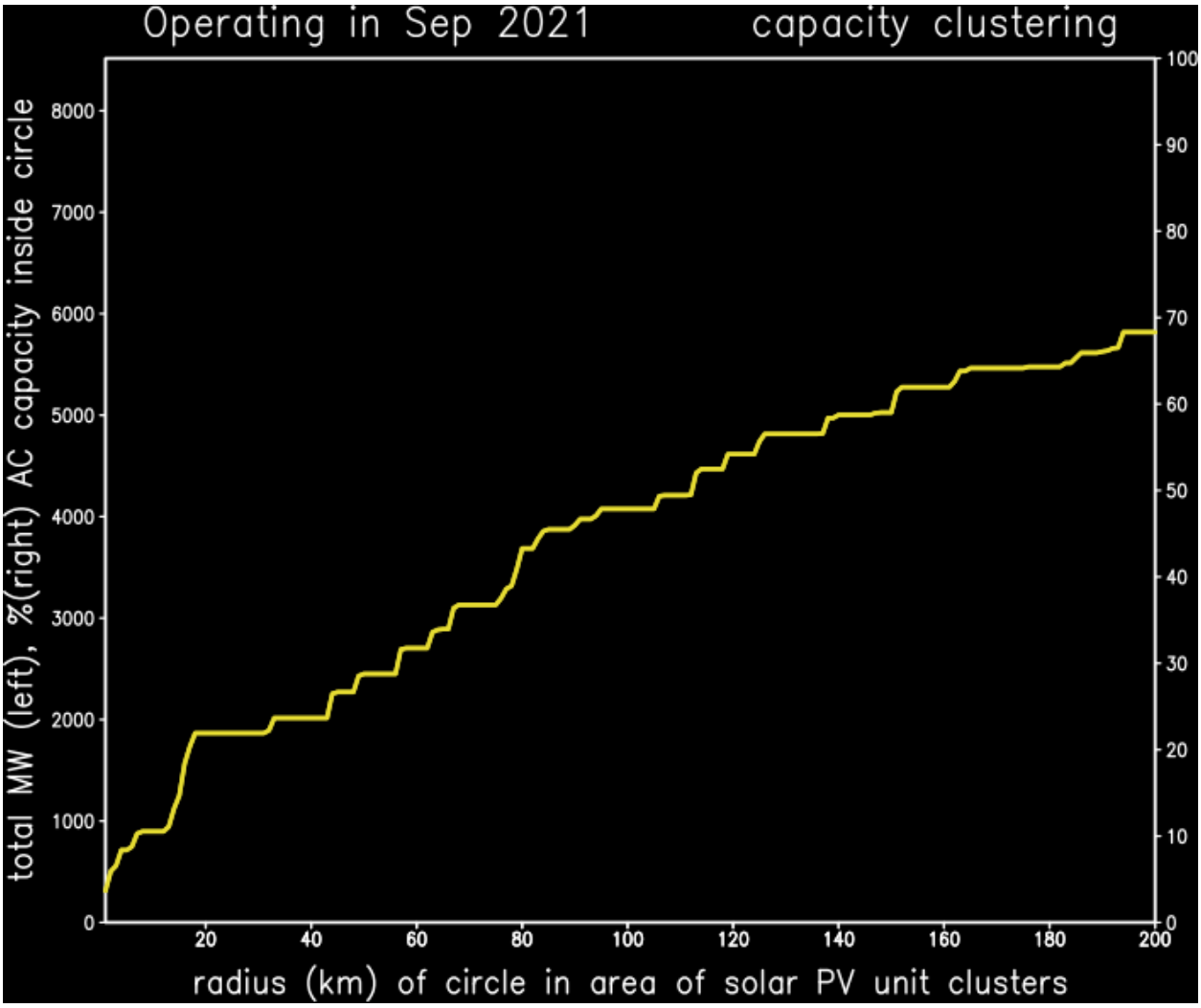


Figure 5. Maximum clustering of capacity as a function of radius. The maxima for all radii are located in west Texas in the vicinity of the maxima shown in Figure 4.

Project Objective

The overarching goal of the project is to integrate advanced probabilistic solar power forecasts into power system operations at multiple timescales to allow greater solar power penetrations while reducing system operating costs and increasing reliability. In the first budget period, the algorithms necessary to achieve this goal were designed and developed. The second budget period focused on validating these algorithms and understanding the value that they can provide to electric power systems. In the third budget period, these algorithms were implemented and tested in ERCOT and ERCOT-like systems to prove their robustness and effectiveness. The project made real-world impacts through the advancement of forecasting and power system modeling techniques and their implementation in ERCOT's system.

Project Results and Discussion

This report breaks the main findings of the project into four research areas, based on different successive tasks conducted throughout the lifetime of the project. The first focuses on the

improvements made to probabilistic solar power forecasting and their implementation in Maxar's operational forecasting system for ERCOT. The second and third research areas detail efforts to use these improved forecasts in operational power system practice in the areas of risk-parity dispatch/unit commitment and dynamic reserves, respectively. Finally, the fourth research area describes the developed forecasting visualization tool.

1 Research Area 1: Solar Forecasting Improvements

Historical Data Set Part 1: Telemetry Observations and Maxar's NWP-Based Hourly Averaged Solar Power Forecasts for 18 Utility-Scale Plants and Limited NWP Ensemble

In November 2018, Maxar delivered an ensemble member forecast data set and matching telemetry observations for the 3-year time period from 2016–2018 for the 18 utility-scale solar power plants operational in ERCOT at that time and delivered an update in early 2019 to complete the year 2018 data. These data were used to develop and test the Bayesian model averaging (BMA) technique of post-processing power forecasts into statistically reliable and accurate probabilistic forecast distributions. Maxar attempted to apply this method to the much bigger "large ensemble," with results shown in what follows. Difficulties were encountered and not resolved during the term of SUMMER-GO, and it is anticipated that better results might be obtained with additional implementation work, although it would also require much more computational resource than the simpler method also tested in SUMMER-GO.

The ensemble member forecast data that Maxar provided is already heavily processed to predict power. The numerical weather prediction (NWP) models predict global horizontal irradiance (GHI). The predicted GHI used for power predictions is time-averaged, in some cases over an output interval of 1 hour or several, or in some cases is instantaneous. The Maxar solar forecast system makes small statistical corrections to the shape and amplitude of the NWP model diurnal irradiance curve; calculates clear-sky direct and diffuse irradiance every 1 minute, accounting for column water vapor and other absorbers and scatterers; converts the model GHI to 1-minute values, assuming a constant clear-sky index over the output interval, and converts this to full-sky direct and diffuse irradiance through a nonlinear relationship; adds stochastic variations to account for the missed temporal variability; transposes these and ground-reflected irradiance into plane-of-array (POA) irradiance for tilted sun-tracking PV panels; and estimates power empirically based on historical data, then averages hourly. Two multivariate empirical power curves are used, resulting in two power forecasts for each POA forecast.

The NWP forecast members contributing to the delivered ensemble forecast set of hourly averaged power are described in Table 1, and the ensemble configuration is described in Table 2. The ensemble members include time lag members from older forecasts that are valid at the same forecast time. Latency also adds to the lead time in the NWP forecasts compared to the lead time in the assembled ensemble set. Lead times from 1 to 48 hours were delivered for the assembled ensemble set, corresponding to longer lead times among the NWP model members. The High Resolution Rapid Refresh (HRRR) model has a shorter maximum lead time, so there are forecast lead times when some or all of those members are absent from the ensemble. Additionally, the ensemble member with the lowest irradiance is duplicated as an additional ensemble member to mitigate the sunny bias of the NWP models. The duplicated NWP member that has lowest hourly averaged GHI varies from one forecast to another, lead time within the forecast, and from one location to another.

Table 1. Configuration of NWP Models Used as Inputs Into Historical Data Set Part 1 Ensemble Collection

Model	Output Grid	Output Maximum Lead Time	Output Interval	Forecast Updates	Time Lag Members in Data Set
ECMWF High-Res Deterministic	0.125°	240 h	3 h	12 h	3
NOAA GFS	0.25°	384 h	3 h	6 h	3
NOAA NAM High-Res Nest	3 km	60 h	1 h	6 h	3
NOAA HRRR	3 km	18 h	15 min	1 h	12

Table 2. Configuration of Delivered Ensemble Set for Historical Data Set Part 1

Source	POA Members	Power Members, Then-Operational Power Curves	Power Members, Updated Power Curves
NWP models	21	42	42
Lowest of NWP models	1	2	2
Total	22	44	44

Additionally, Maxar created and delivered to the National Renewable Energy Laboratory (NREL) weather regime data for the dates of the forecast and ensemble data sets. These data are also described in more detail in the first Maxar annual report on SUMMER-GO. The hope was that a regime-based forecast methodology, perhaps using conditional distributions, could be developed, but that did not happen.

Historical Data Set Part 2: Maxar's NWP-Based Forecasts for 369 Locations for Hourly Averages and 5-Minute Averages for a Large NWP Ensemble

A second, much larger set of ensemble forecast data was delivered to NREL based on a superensemble of the 116 NWP model forecasts listed in Table 3 for the calendar years 2017 and 2018. Some NWP models were individual runs with time lag ensembles, as in the small ensemble set. Others were ensembles of a single NWP system, such as the 51 members from the ensemble set. This is also described in more detail in the first annual report; however, the National Oceanic and Atmospheric Administration (NOAA) Short-Range Ensemble Forecast (SREF) system was removed from the delivery (and thus is not included in Table 3) because it was discovered that all its members only had instantaneous GHI retrieved for 3-hour intervals, which is too infrequent to be representative of the required time averages.

Forecast data were provided for each of the 116 NWP members for the hourly averages of GHI and direct normal irradiance (DNI) out to 48 hours and each 5-minute interval in the first 2 hours using the most recent set of the 116 members, which would have been available, allowing for NWP model latency. The GHI and DNI for these times were based on running the model through the Maxar solar forecast system as described for the smaller ensemble. Additionally, near-surface

wind and temperature were provided for the same hour and 5-minute times through the 2-year period, and Maxar-calculated clear-sky GHI and DNI were provided for reference and scaling.

Table 3. Configuration of NWP Models Used as Inputs Into Historical Data Set Part 2 Large Superensemble Collection

Model	Output Grid	Output Maximum Lead Time	Output Interval	Forecast Updates	Time Lag Members in Data Set	Number of Members per Run
ECMWF High-Res	0.125°	240 h	1 h	6 h	3	1
NOAA GFS	0.25°	384 h	1 h	6 h	3	1
NOAA NAM Nest	3 km	60 h	1 h	6 h	3	1
NOAA HRRR	3 km	18 h	15 min	1 h	15	1
ECMWF Ensembles	1°	360 h	6 h	12 h	1	51
NOAA GEFS	0.5°	384 h	3 h	6 h	1	21
NOAA RAP	13 km	18 h	1 h	1 h	15	1
Canadian Global	0.24°	384 h	3 h	12 h	2	1
Canadian Regional	10 km	48 h	1 h	6 h	3	1

The locations run through the Maxar solar forecast system included the 18 ERCOT solar power plants, 204 prospective solar power plants from the ERCOT interconnection queue, 22 synthetic sites in the ACTIVSg2000 system, 4 Sandia National Laboratories sites with research-quality monitoring data, and 121 zip code centroids used for the analysis of behind-the-meter generation. These 369 locations were selected to meet the needs of all power system analyses being performed in SUMMER-GO.

NREL passed these forecasts through the System Advisor Model (SAM) to synthesize power in a nonproprietary manner consistent for both the then-existing locations and the future high-solar build-out scenario locations. These data were then used by the SUMMER-GO partners to inform grid engineering studies under a high-solar build-out scenario.

Several terabytes of data were compressed and delivered to the NREL Amazon Web Services (AWS) S3 bucket after considerable delays caused by a series of circumstances, starting with waiting to determine the SUMMER-GO revised behind-the-meter project plan, which selected the 121 zip codes used from among the 369 locations, then the data construction being more time-consuming than originally anticipated, compounded by delays in being able to obtain Maxar access to NREL's AWS S3 bucket. Finally, Maxar noticed that a small subset of the NWP superensemble members—North American Mesoscale Forecast System (NAM), Rapid Refresh (RAP), and HRRR—had been run using instantaneous GHI as though they were time-averaged over the preceding model output interval. This caused the calculations to show too high irradiance in the morning, which was truncated to physically reasonable limits, and too low

irradiance in the evening, resulting in the problem mostly manifesting toward the end of the day. Because a small depression of GHI relative to its clear-sky values results in a larger depression of DNI relative to its clear-sky values, DNI was essentially shut down for the last few hours of the day, even on sunny days. This necessitated correcting these calculations for the model subset affected and then, due to the file data structure, reposting the entire, now corrected, data set. That was completed in July 2020.

A. Real-Time Probabilistic Solar Power Forecasts

Four sets of real-time power forecasts were run during SUMMER-GO for each operational ERCOT utility-scale solar generating unit and for the aggregate total of all operational units. Some large plants with several hundred megawatts of AC capacity have more than one grid interconnection, each as a separate generating unit. As noted earlier in this report, the number of operational generating units included in the forecasts increased throughout the project, culminating in more than 70 units by the end.

The four sets of real-time power forecasts are compared in Table 4.

1. The operational official forecast uses the small ensemble set similar to those described in Table 1 and Table 2. The forecast contains only the 50thth percentile and 20th percentile (80% probability of exceedance), and it contains hourly averages out to 1 week, updated hourly. Also, recent observations were used to perform a short-term forecast correction for the first several hours of the forecast.

   The three sets of experimental forecasts contain 99 percentiles to quantitatively describe the forecast distribution in detail to serve as inputs into risk management tools.

2. The first set of probabilistic forecasts uses the medium superensemble set described in Table 5. This includes many more NWP models than the operational system, but for NWP ensemble sets, it uses only the mean from each instead of all the individual members. Also, instead of predicting hourly averages, it predicts 5-minute averages out to 48 hours, and it updates every 5 minutes. When the timing of the real-time deliveries of the 5-minute forecasts into the ERCOT energy management system (EMS) was coordinated, these runs were reduced to only go out to 24 hours to complete faster because the entire 24 hours was updated every 5 minutes. The short-range forecasts for the 5-minute intervals also used smart persistence based on the recent telemetry observations of 5-minute power. The last three observations were used. The smart persistence adjusts the observation based on the panel-sun geometry so that the clear-sky index at the observation time is maintained into the forecast time with some correction for the change in the slant path of the beam through the atmosphere.

3. The second set of probabilistic forecasts uses the large superensemble set described in Table 5, which includes all the individual ensemble members from each NWP ensemble set, a larger time lag ensemble of the hourly updating HRRR and RAP, and up to the last 24 recent 5-minute observations for the short-term forecast. It produces 5-minute averages for the first 2 full hours of the forecast and hourly averages for longer lead times out to 48 hours. It was updated every 15 minutes and later adjusted to update every 5 minutes.

4. The third set of probabilistic forecasts is the same as the second except the probability distribution is determined using the BMA method developed by Doubleday.

The operational hourly forecast is delivered to the ERCOT EMS through a secure web service on a schedule that allows ERCOT time to properly ingest and use the forecast. It was intended to similarly deliver the experimental forecasts, but ERCOT was not able to ingest the additional experimental probabilities during SUMMER-GO and was not able to ingest the 5-minute averages and the 5-minute updates until Phase 3; therefore, the full probability distribution forecasts were made available in real time by File Transfer Protocol (FTP) to both ERCOT and NREL throughout the project. Only the SUMMER-GO-1 forecast in Table 4 was delivered. The intention was to demonstrate the superiority of the SUMMER-GO-2 and SUMMER-GO-3 forecasts and then switch the delivery to one of those, but various complications (described later) did not allow that. The first 2 hours of the SUMMER-GO-1 forecast are now being delivered operationally into the ERCOT EMS. The SUMMER-GO-2 and SUMMER-GO-3 forecasts were run in real time and archived by Maxar for evaluation; they are included in the forecast evaluations provided later in this report.

Table 4. Characteristics of Operational and Three SUMMER-GO Experimental Power Forecasts

Forecast	Start	NWP ensemble set	Percentile Algorithm	Percentiles	Update interval	Lead times
Operational	2016	Small	Old: Maxar	20, 50	60 min	Hours 1–168
SUMMER-GO-1	September 30, 2019	Medium	Old: Maxar	1,2,3,…,99	5 min	Every 5 min to 48 h[a]
SUMMER-GO-2	March 2020	Large	Old: Maxar	1,2,3,…,99	15 min[b]	Every 5 min to 2 full h, then every hour to 48 h
SUMMER-GO-3	August 2020	Large	New: BMA	1,2,3,…,99	15 min	Every 5 min to 2 full h, then every hour to 48 h

[a] Lead times to 24 hours starting March 4, 2021

[b] Update interval 5 minutes starting February 12, 2021

Table 5. NWP Ensemble Sets Used in the Experimental Forecasts

Model	Updates Per Day	Medium Ensemble # Members in Set	Medium Ensemble # Time Lags	Large Ensemble # Members in Set	Large Ensemble # Time Lags	Total NWP Sources/ Day
ECMWF	4	1	3	1	3	4
ECMWF ensembles	4	Mean only	3	51	1	204
HRRR	24	1	3	1	15	24
RAP	24	1	3	1	15	24
GFS	4	1	3	1	3	4
GEFS	4	Mean only	3	30	1	120
SREF – NMMB	4	Mean only	3	13	1	52
NAM	4	1	3	1	3	4
Canadian Global	2	1	3	1	2	2
Canadian Global Ens	2	Mean only	3	20	1	40
Canadian Regional	4	1	3	1	3	4
Canadian Regional Ens	4	Mean only	3	20	1	80
Total NWP		**36 x 2 power curves = 72**		**178 x 2 power curves = 356**		
5-min smart persistence	**288**	**1**	**3**	**1**	**24**	

Trying to maintain all these different code bases and their corresponding databases was challenging. There were times when one or another of the experimental systems were offline due to code changes being worked on, so they have gaps on different days. There were times when the experimental systems were not up to date with the operational system in the list of generating units running. Sometimes new units were not started in the experimental systems. Other times, units were prepared to start in the operational and experimental systems but were not officially ready, so they were not included in the operational system but started right up in the experimental system. These discrepancies impacted the aggregate totals in the experimental systems. Additionally, all systems were intended to use outage or derate schedule information to appropriately reduce the forecast generation, but the different experimental forecasts had that

added and working properly at different times. For the most part, the delivered SUMMER-GO-1 forecasts were handled better than the others on these issues.

The Maxar method of generating percentile distributions, used in the SUMMER-GO-1 and SUMMER-GO-2 forecasts, involves a part determined by the ensemble distribution that ties probabilities to the current weather situation and a part determined based on historical error statistics that allows a wider distribution when the NWP forecasts have insufficient dispersion. The latter scales the error using a distribution-based multiplier that depends on the percentile, which can cause high values to exceed the site capacity. This was automatically truncated, and not a problem; however, it also could exceed any reasonable bounds for a clear-sky condition for the time of day, e.g., below the capacity but above a physically reasonable value. This was not noticed at first and was later corrected.

The forecast error history used in the Maxar method of generating the percentile distributions should correspond to the lead time of the forecast. The error for the first few 5-minute periods of the forecast should be much smaller than that for 5-minute periods with a 1-hour lead time or any of the longer lead time hourly averages and should be from the more skillful weighted superensemble blend rather than individual model members. It was discovered late in SUMMER-GO, however, that the code was actually using errors for the early hours of the forecast, not individual early 5-minute periods, and for an average of errors from each model rather than for the forecast blend. The result is that the SUMMER-GO-1 and SUMMER-GO-2 forecast distributions are not as sharp as the algorithm would make them if it were working properly for any lead time, with a large difference in sharpness for the first few 5-minute intervals. The statistical reliability is also affected by the distribution being too broad.

The smart persistences from recent observations are weighted heavily in the very short-range 5-minute interval experimental forecasts; however, early in the morning, these are not available. These were set to clear-sky values when the experimental systems were first created, and it was not until April 2021 that this assumption was removed and May 2021 when a better methodology of when to not use an observation in the smart persistence was implemented. These improvements were implemented only in the SUMMER-GO-1 forecasts and remained a problem in the morning in the other experimental forecasts.

I discovered that the RAP and SREF forecasts were ingesting instantaneous instead of time-averaged GHI, but it was treating them as time-averaged values, causing those members to yield artificially low generation during the last few hours of the day. This resulted in a forecast distribution shift toward low values in the late afternoon for all lead times, mostly affecting the lower tail of the distribution and widening the forecast distribution. This was fixed for all experimental systems by mid-October 2020.

Outage handling was initially performed as a modifier to the completed forecast distribution and only later during 2020 was correctly applied to each individual superensemble member before creating the forecast distributions.

The forecast distribution for the aggregate totals is made in the same manner as the forecast distribution for an individual unit; however, initially, the aggregate distribution was made by summing the percentiles of each individual unit forecast distribution, which resulted in too broad

of a forecast distribution for the aggregate. (This broader distribution would be correct if the forecasts at all the generating units were perfectly correlated, so the error in this assumption is not very large considering the clustering of capacity in west Texas.) The code structure required to do this properly is surprisingly complex, and the added computational cost is much more than the incremental cost of adding an additional generating unit to the collection of sites.

It proved difficult to implement the BMA algorithm for the probability distributions to the large ensemble containing an order of magnitude more members than the ensemble used to develop the algorithm and employing it into an operational job stream for a larger range of forecast lead times and other operational constraints, including run time and cloud computing cost.

On one hand, the BMA algorithm was able to run and performed excellent bias correction. Achieving real-time runs in an operational environment represents a remarkable research-to-operations transformation—an achievement of considerable significance!

On the other hand, the forecast distribution was too broad and the accuracy poor because the weights iteration would converge near the initial guess regardless of the initial guess, instead of selectively weighting the sources by skill over the training sample. Sufficient time was not available to explore why this was happening and to correct it. Ideas to improve this include alterations to the iteration method to force it to jump around more to explore the phase space of possible coefficients and the *a priori* setting of skill-based weights from some other method before starting the iteration. Also, the BMA method assumes that each deterministic forecast is associated with a probability distribution and assumes that the parameter that controls the width of the beta functions used to define the probability distribution associated with each ensemble member is the same for all ensemble members. Relaxing this assumption is probably necessary to obtain the selective skill impact on the forecast, but it might make the iterations intractable. An idea could be to make some assumption about this parameter based on the error statistics of the individual ensemble member. This might also improve the weights calculation. None of these ideas is guaranteed to work. The further exploration of this has considerable risk of requiring a lot of effort without a certainty of finding success; however, the possibility of being rewarded with a much improved forecast could make the investment worthwhile.

B. Operational Delivery to ERCOT and Use in Real-Time Operations

The original plan was to deliver real-time probabilistic power forecasts for each operational utility-scale solar generating unit and the aggregate total to the ERCOT iTest EMS. The iTest EMS is a testing EMS used as a pathway to seamlessly migrate changes into the production EMS by testing in an equivalent environment without impacting grid operations.

The plan underwent several revisions throughout SUMMER-GO due to ERCOT requirements.

1. ERCOT decided that 99 percentiles were too much data to receive. Instead, they requested files containing only the 50th percentile and 5 other percentiles describing the lower tail of the distribution.
2. ERCOT needed to modify their EMS to be able to handle 5-minute forecast intervals, with 5-minute updates, and the additional percentiles. This major undertaking needed to be coordinated so that it could be used for wind as well as solar generation forecasts.

3. ERCOT faced a critical operational need to manage the large evening down ramp of solar generation coincident with the afternoon/evening peak load times, and therefore the 5-minute forecast for the ERCOT aggregate total had the most immediate operational importance.
4. The iTest EMS became able to handle the ERCOT aggregate 50th percentile 5-minute interval, 5-minute updated forecasts out to 2 hours of lead time in late February 2021.
5. ERCOT's critical operational need for the 5-minute forecasts necessitated delivery to the production EMS to use in operations immediately. The production EMS became able to handle the ERCOT aggregate 50th percentile 5-minute interval, 5-minute updated forecasts out to 2 hours of lead time in late May 2021.
6. ERCOT is still working on changes to the iTest and production EMS to allow it to handle the 5-minute forecasts with additional percentiles and for each individual generating unit as well as to handle the additional percentiles for the hourly forecasts.

The corresponding project impacts were:

1. Maxar generated forecast files in two sets: (1) one containing all 99 percentiles for Maxar and NREL to evaluate and import into any downstream decision support tools and (2) another containing the set of six percentiles ERCOT requested.
2. Because the EMS was not able to ingest the SUMMER-GO forecasts, files from the SUMMER-GO-1 forecasts were made available in real time by FTP for each generating unit and for the aggregate total for all its lead times and intervals, from September 30, 2019, through the end of the project, with a few brief interruptions for code improvements.
3. Maxar evaluated the skill of the ramp forecasts of the aggregate total for the 5-minute intervals. Results are presented subsequently.
4. Maxar coordinated and tested delivery logistics with ERCOT for delivering to iTest the aggregate 50th percentile forecast for the 5-minute intervals with 5-minute updates. Regular deliveries using the same secure web service used for the operational hourly forecasts began on March 4, 2021, and continued through the end of the project.
5. Although beyond the scope for SUMMER-GO, Maxar agreed to deliver the same forecasts to the production EMS as part of SUMMER-GO, using the same secure web service used for the operational hourly forecasts. Further, Maxar hardened the SUMMER-GO-1 forecast code to minimize the possibility of missing any forecasts delivered to ERCOT's operational systems. Deliveries to the production EMS began when the EMS became ready in late May 2021 and had some failure over the Memorial Day weekend. Regular uninterrupted service began on June 1, 2021.
6. ERCOT is still planning to implement probabilistic solar generation forecasts into grid operations in the future, which will be a major accomplishment in managing renewable energy on the transmission grid. Maxar will be ready to provide the necessary forecasts; however, this activity, which was hoped to be part of SUMMER-GO, will not happen until after the end of the project.

The delivery to the production EMS was highly successful. From the start of the uninterrupted delivery to the production EMS on June 1, 2021, through the end of SUMMER-GO on September 30, 2021, ERCOT logged 61 missed 5-minute forecast deliveries. Any time when the forecast did not enter the production EMS was logged as a missed delivery. Some missed deliveries were at night. Misses could have been caused by failures on the forecast side or on the receiving side, by internet problems in between, or by timing problems resulting in delivery outside the expected window. With 288 deliveries expected per day for 122 days in this sample, a total of 61 misses indicates a successful delivery and an ingestion rate of 99.826%. The delivery code already had automatic retries for internet failure and has been modified to use redundancies and other automated retry procedures to address the cloud computing system failures that caused many of the misses; therefore, it is anticipated that there will be extremely few misses in the future—much closer to a 100% delivery rate than the 99.826% delivery rate achieved in SUMMER-GO.

ERCOT began a new era of renewables integration on the grid. ERCOT began using the predicted solar generation ramp rate in its operational dispatch algorithm on June 1, 2021. The predicted ramp rate of the 50th percentile aggregate solar generation as the difference between the first two upcoming 5-minute periods in the forecast is used together with the predicted wind ramp rate, the predicted load ramp rate, and a variety of other factors. The idea is that the aggregate net generation required for dispatch in the next 5-minute period is the amount in the last 5-minute period plus changes due to all factors changing on the grid. Each factor is weighted by a coefficient so that the best fit can be optimized. In this manner, predicted ramps in the aggregate solar (and wind) generation influence the total dispatch of resources on the grid.

This new era of renewables integration on the grid uses the 5-minute interval, 5-minute updating forecasts developed under SUMMER-GO. ERCOT will be requiring commercial procurement of such forecasts for the foreseeable future. This represents an area where U.S. Department of Energy support has successfully transitioned to a sustainable business model to carry the developments forward into a high-renewables future.

Relationship to Decision Support Tools

The goal of SUMMER-GO is to allow increased integration of solar power into the energy mix while reducing system operating costs and increasing reliability. Note that the probabilistic solar generation forecasts are used by decision support and integration tools developed by other SUMMER-GO partners to reach this goal.

One decision support tool, SolarView, is a visualization system that ingests the forecast probabilities and information about the state of the power system, combining this information to alert grid operators in real time about risks. NREL demonstrated this tool to ERCOT.

One integration tool is adaptive reserves. Information about the probability distribution of solar generation in the forecast combined with the load forecast and information about the status of the grid's generating resources of all types and other system information can be used to determine when risk warrants purchasing additional reserves in the day-ahead or a longer lead time frame and when the risk is low enough to purchase less reserves than the current method ERCOT uses. The data Maxar supplied to SUMMER-GO combined with other data were used by project partners to demonstrate increased grid reliability at lower cost using adaptive reserves.

Another integration tool is risk parity economic dispatch. Similar to adaptive reserves, it combines the forecast probability distribution with other information about the system to determine risk, but instead of examining reserves, it examines the more complex matter of dispatch of resources for balancing in the very short term, the upcoming series of 5-minute periods. The data Maxar supplied also contributed to project partner efforts to demonstrate the value of dispatch tools under a high solar penetration scenario to help economically and reliably manage the grid when solar generation reaches a much larger fraction of the total generation.

Given the large fraction of ERCOT generation by wind power, the enormous amount of solar capacity in the interconnection queue, and the history of the exponential growth of installed utility-scale solar on the ERCOT grid, it is only a matter of time before grid operations will turn to tools such as these, which employ probabilistic generation forecasts as one of many inputs to better map risk and more economically maintain grid reliability. And it is only a matter of time that renewables will be large enough in other regions to similarly lead other system operators down this path. SUMMER-GO has contributed to improving awareness and understanding by system operators, facilitating movement down this path toward the future of grid management.

Case Study Summaries

To illustrate the information in the delivered SUMMER-GO-1 probabilistic forecasts, this section summarizes the case study presented in the year 2 annual report, discussing the 5 most salient figures of the 36 shown there. (Please refer to that report for further details.) To compare the information in the three SUMMER-GO experimental forecasts, this section summarizes the case study presented in the Q4 2020 quarterly report.

The first case study shows a clear winter day at a west Texas single-axis solar power plant and the preceding day with morning clouds at that same location and at another west Texas single-axis solar power plant with a different cloud regime. The forecasts and observations are shown as a fraction of AC capacity for each location. There is a double hump in the diurnal cycle of the clear-sky power at the single-axis sites in winter because the sun never gets high enough in the sky to produce angles of incidence on the panels of near 90 degrees, with the highest angles of incidence in the midmorning as the panels tilt toward the east and the afternoon as the panels tilt toward the west.

On the clear-sky day shown in Figure 6, the probability distribution from the SUMMER-GO-1 forecast in the morning through early afternoon is sharp, with only an approximate 15% difference in capacity between the 10th percentile and the 98th percentile, indicating high certainty. Even this range is inflated for the 10-minute lead time shown due to the problem of computing part of the distribution spread based on the historical errors for hour lead times. The clear-sky power, observed power, and day-ahead operational forecast adjusted to 5-minute intervals are shown for comparison. Some curves overlap; thus, only the one plotted on top might be visible. In late afternoon, the lower percentiles abruptly deviate more from the middle to high percentiles. This is likely related to the problem that caused early reduction in the direct beam irradiance in some members, because this case was from before that problem was fixed, and the mid- to upper range of percentiles is unaffected by this issue.

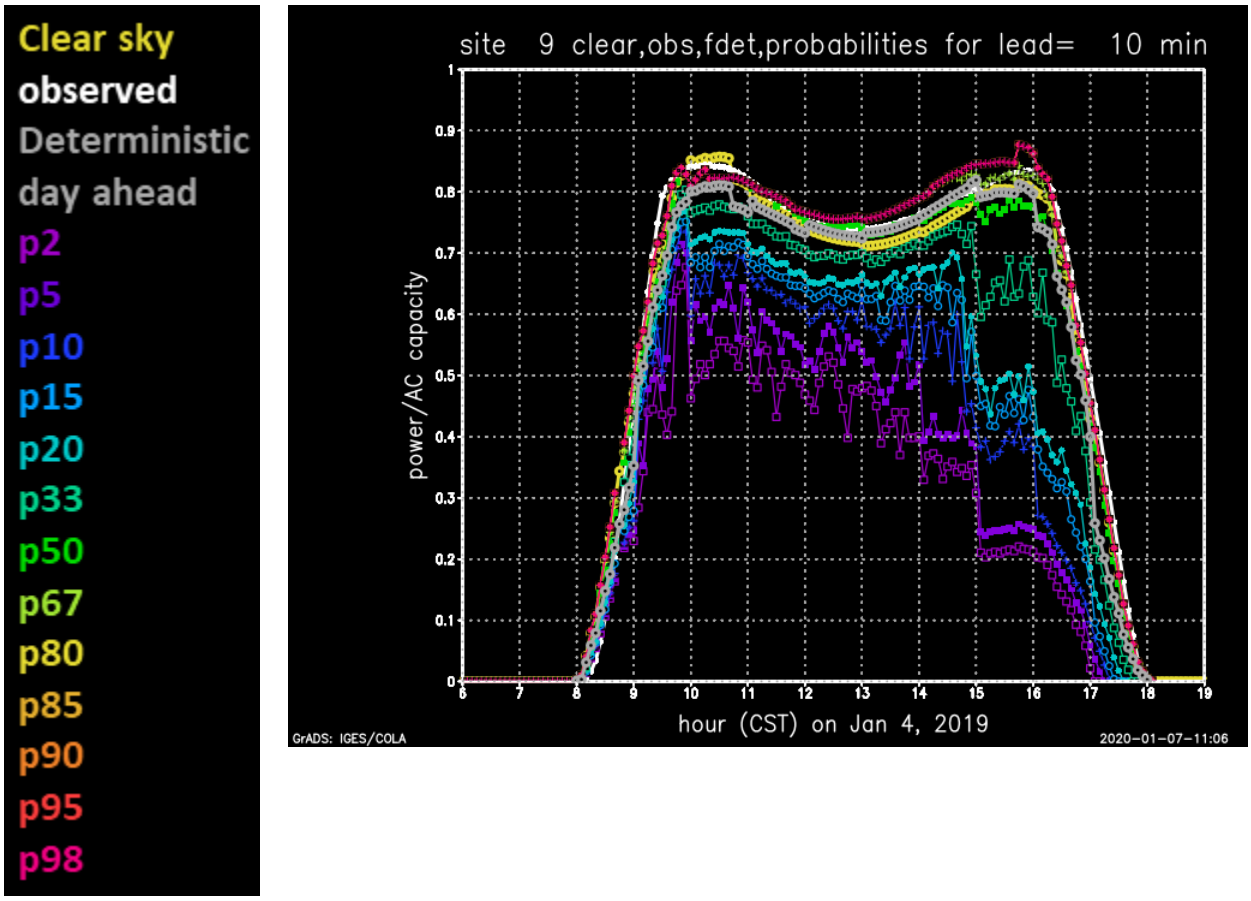


Figure 6. Time series of percentiles for the 10-minute lead time at Site 9 on a sunny day and the observed power, calculated clear-sky power, and day-ahead deterministic forecast. Legend at left

On the cloudy day shown in Figure 7, the probabilistic forecast showed a much wider range, with much less certainty. The observed evolution with thick clouds gradually thinning throughout the day followed by abrupt clearing later in the afternoon was well tracked by the midrange percentile, a bit below the 50th percentile, until the rapid clearing, and then a bit above. The early morning forecast spike was related to the assumption of a clear sky for smart persistence for low sun angles before the observations were considered good enough to use, and this case was from before that assumption was removed. In comparison, the day-ahead forecast was much higher than observed, into the high percentiles of this very short-range forecast.

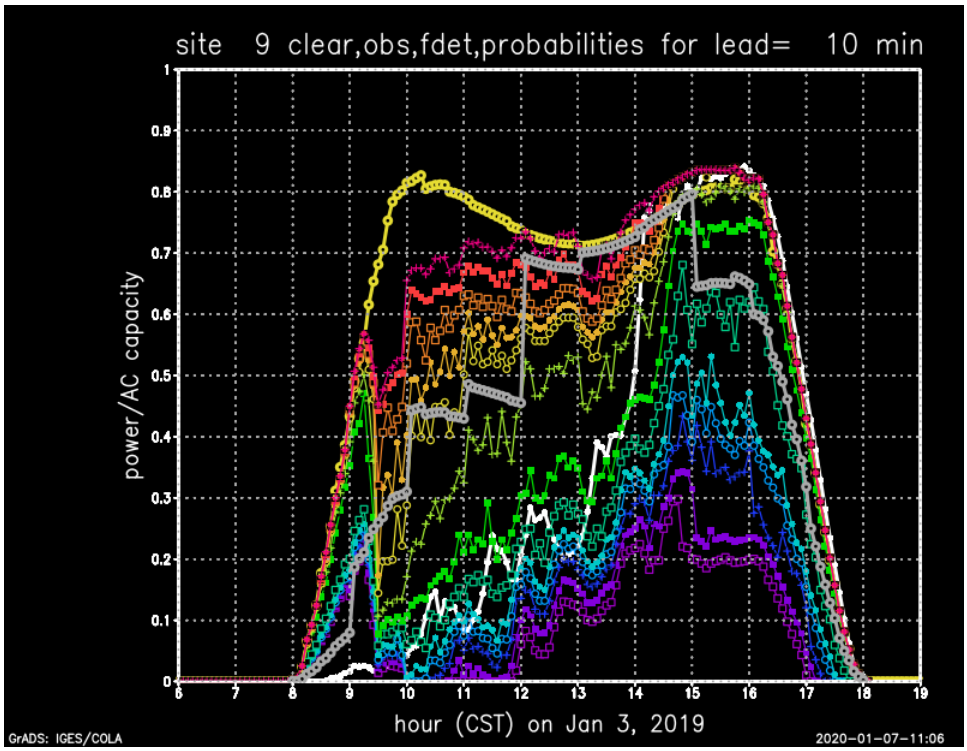


Figure 7. Time series of percentiles for the 10-minute lead time at Site 9 on a day with morning clouds and the observed power, calculated clear-sky power, and day-ahead deterministic forecast. Legend in Figure 6

The cloudy day at the other location, shown in Figure 8, featured high-amplitude fluctuations in the reported power starting around noon, as the cloud regime changed from broadly continuous stratiform low clouds, to scattered, broken, thicker convective clouds. The convective cloud regime resulted in highly variable power ranging from 15% of clear sky to exceeding clear-sky power. The irradiance and power can spike above clear-sky values when the sun is between clouds while reflections off the sides of the clouds reach the array. The convective clouds were much larger than fair weather cumulus although smaller than cumulonimbus (thunderstorms), so they were just the right size for this effect to align with a large solar power plant. A field of small cumulus clouds casting shadows moving across a large solar power plant causes smaller amplitude rapid fluctuations of power because the small fair weather cumulus shadows are typically smaller than these plants, which span hundreds of acres.

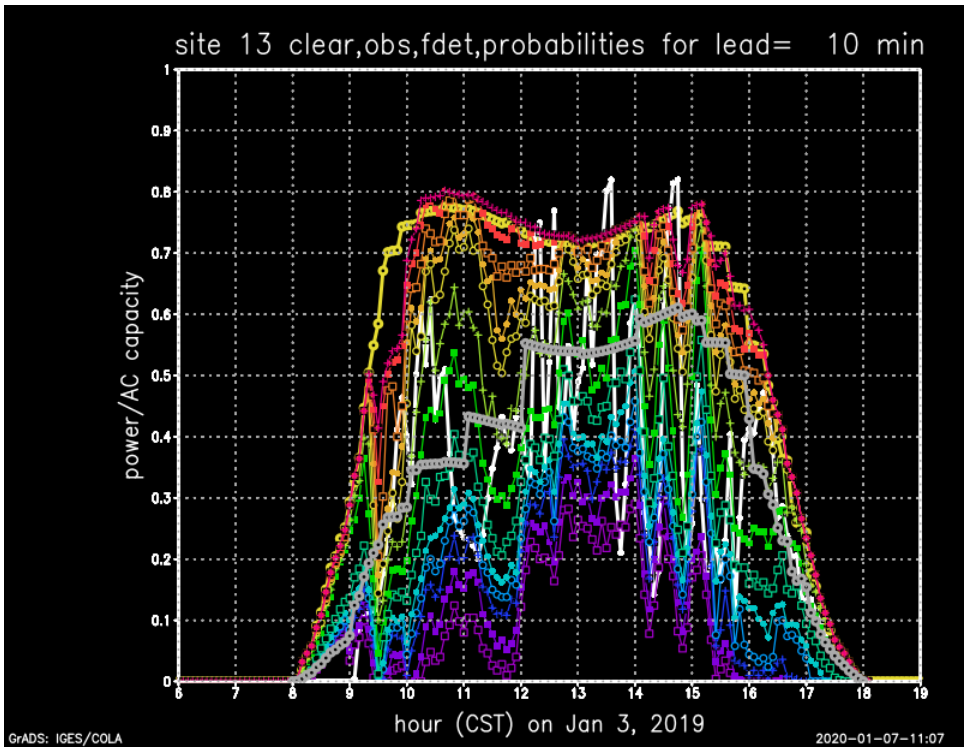


Figure 8. Time series of percentiles for the 10-minute lead time at Site 13 day on a day with variable clouds and the observed power, calculated clear-sky power, and day-ahead deterministic forecast. Legend in Figure 6

The short-term forecast shown in Figure 8 has large-amplitude fluctuations in the midrange percentiles that are sometimes in phase and sometimes out of phase with the observed deep fluctuations. The forecast envelope does a good, albeit imperfect, job of bounding the fluctuations, and the distribution is more uniform, indicating less certainty than at the other location, where there were not high-amplitude fluctuations.

For the day-ahead lead times at the first location, Figure 9 shows that the entire forecast distribution misses the low power in the morning. This example points to the need for fundamental improvements in cloud forecasting in NWP models. Nonetheless, the lower part of the distribution increases throughout the day in tandem with the observations because the forecast was for more certainty of clearing as the day progressed.

The envelope of the day-ahead lead time probabilistic forecasts shown in Figure 10 nicely contains and bounds the observed excursions and shows a diurnal shape to the lower bound (the lower bound is not as low in the middle of the day). The observed excursions will not be predictable at such lead times, but a good probabilistic forecast can provide information about the observed power distribution (e.g., the collection of 5-minute values over a few hours) in cases of high variability.

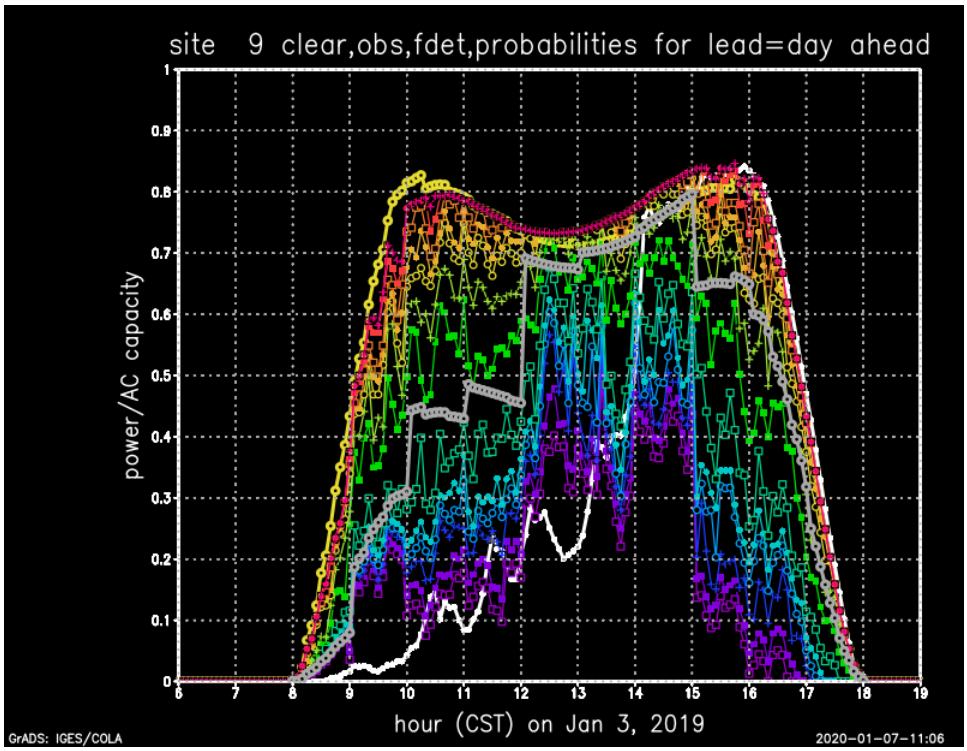


Figure 9. Time series of percentiles for the day-ahead lead time at Site 9 on a day with morning clouds and the observed power, calculated clear-sky power, and day-ahead deterministic forecast. Legend in Figure 6

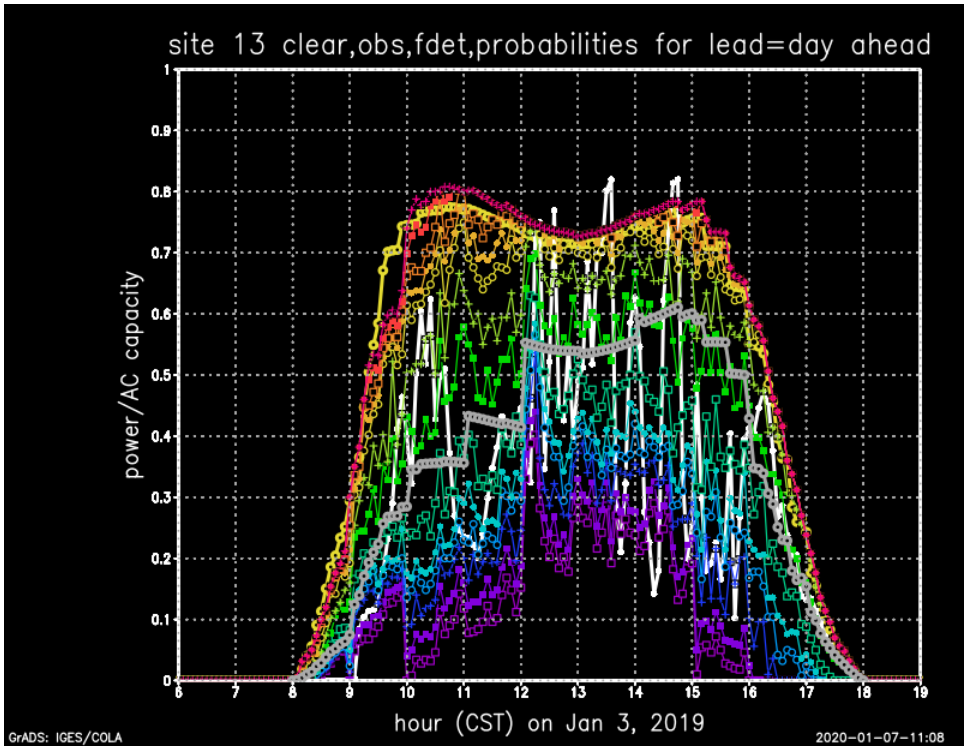


Figure 10. Time series of percentiles for the day-ahead lead time at Site 13 day on a day with variable clouds and the observed power, calculated clear-sky power, and day-ahead deterministic forecast. Legend in Figure 6

The previous case showed forecasts from the SUMMER-GO-1 system during winter. The next case compares forecasts for all three experimental systems using 2 summer days a week apart for comparison. Again, one of the 2 days is sunny, and the other had morning clouds, low generation, and a large morning overforecast error in the operational deterministic forecast.

The previous case showed forecasts at individual generating units. The next case shows forecasts of aggregate generation for 31 generating units, which yielded approximately 2,700 MW of generation during the midday hours on the sunny day. This case precedes correcting the method of calculating percentiles for the aggregate, resulting in forecast distributions that are too broad for all three sets of forecasts.

Observations and probabilistic forecasts for 5-minute intervals with 30-minute lead times are shown in Figure 11 for the sunny day (top; the lower part of the plot is truncated because it shows nothing of interest) and the cloudy day (bottom) for each of the three forecast systems (from left to right): SUMMER-GO-1, SUMMER-GO-2, and SUMMER-GO 3. The forecast time series for all 99 percentiles are plotted with percentile ranges indicated by the color bars, each color bar having a width of 10 percentiles, except the lowest goes only to 5, and the highest is only above 95. For all three forecast systems, the sunny day forecast is much sharper than the cloudy day forecast, showing more certainty for the sunny day, and the upper half of the distribution is sharper than the lower half, especially on the sunny day. The forecasts for the first two systems are very similar, although the second system, with a larger ensemble, centers the cloudy day distribution a little bit better in the late morning to midday, and, overall, it is a bit smoother, possibly due to the contributions from many more sources, including a larger sample of recent observations. The early morning spike in the second system is related to the smart persistence handling issue previously described. The third system, with BMA postprocessing, has a wider forecast distribution on both days, likely attributable to the issues previously discussed. On the sunny day, the observations are also aligned a little toward the high side of the distribution, rather than centered, as in the other forecasts, perhaps because the BMA bias correction accounts for the sunny bias in the models, shifting the center of the distribution a bit lower when the forecast is for sunny conditions.

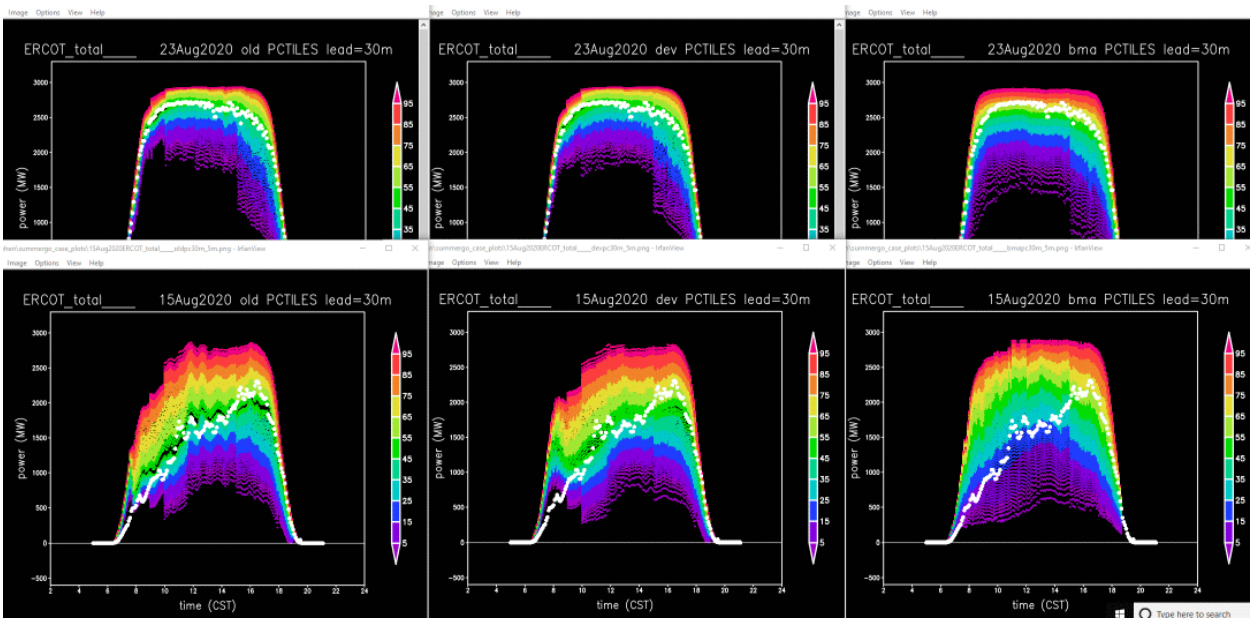


Figure 11. Forecasts of 5-minute averaged power for the 30-minute lead times

The colors are probability bands, and white is observed power. The top row is for a sunny day, and the bottom is for a day with morning clouds and gradual clearing. The forecasts use a medium ensemble (left), a large ensemble (middle), and large ensemble with BMA post-processing (right).

Observations and probabilistic forecasts for day-ahead lead times are shown in Figure 12. The day-ahead forecasts for the large ensemble forecasts (SUMMER-GO-2 and SUMMER-GO-3) have hourly averages at that lead time, and the observations have been averaged to match the forecast averaging times. As in the winter case, the cloudy day has wider forecast distributions than the sunny day in the day-ahead lead time, indicating greater uncertainty for the cloudy day. This is typical on days when the cloud thickness or timing is uncertain. On days with widespread rain and low clouds all day associated with major large-scale storm systems, the distribution can be as small as on sunny days. Similar to the winter case shown above, the morning to midday observations on the cloudy day fall below the day-ahead forecast distribution or in the extreme lower tail of the distribution. As with the shorter lead times, overall, the first two systems give similar probability distribution forecasts, whereas the one with BMA postprocessing has a broader distribution, which allows the lower tail on the cloudy day to encompass the observations at the expense of its distributions generally being too broad. This is a case when all the forecast models at this lead time were much too high, a most difficult case. Typically, even on cloudy days, the observations fall within the envelope of all the predicted forecast distributions.

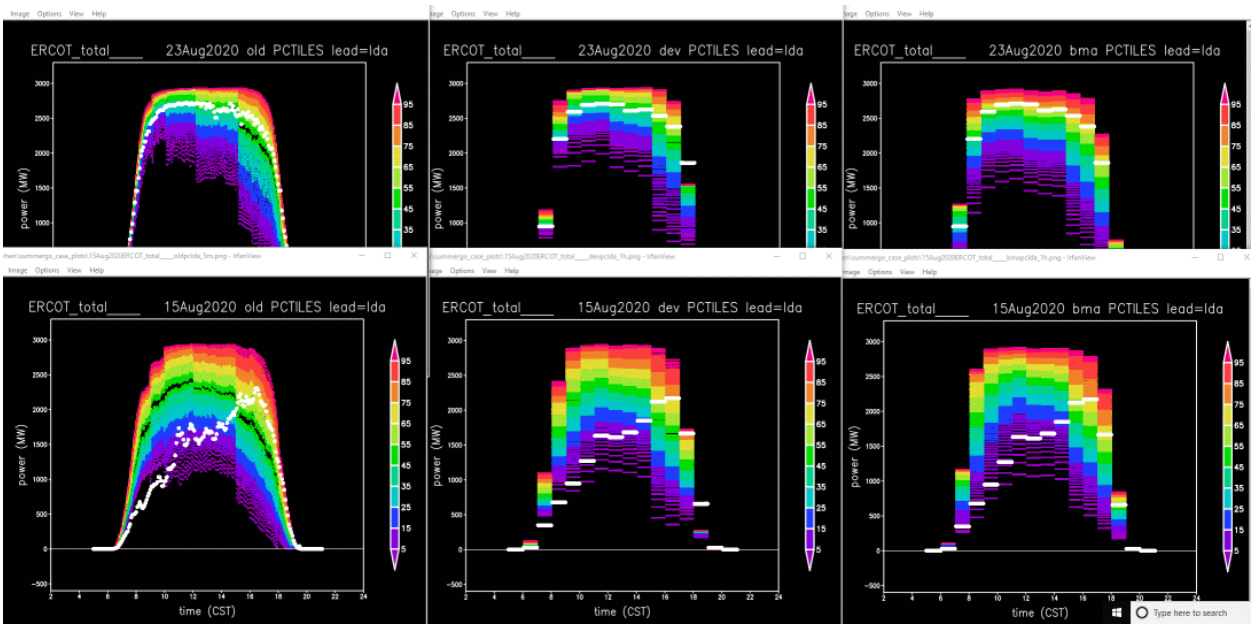


Figure 12. Forecasts of power for the day-ahead lead times, otherwise the same as in Figure 11. The medium ensemble forecasts are for 5-minute intervals, whereas the large ensemble forecasts and corresponding observed power are for hourly averages.

Validation Complications in Experimental Forecast Systems

The observations and the forecasts create challenges for validation. Forecast issues noted in previously include times when different experimental systems had different sets of generating units included in the aggregate due to differences in the startup of new units, times when outages were not handled properly or at all by one or several of the experimental systems, and various other details affecting the forecast distribution. This section demonstrates ways in which outage reporting and observation quality problems combine with the forecast issues to create difficulties to fair and adequate validation.

The fraction of data passing the quality-control procedure sets 1 and 2, as detailed in the first annual report, is shown in Figure 13 for the period from January 1, 2021, through September 30, 2021, for each individual PV generating unit. It shows the percentage of daytime 5-minute periods when both the High Sustained Limit (HSL) and power were flagged and set as missing values or were already missing. Power is the actual generation that can be curtailed. HSL is the same in the absence of curtailment or ERCOT's imputed potential generation that would have occurred without curtailment. The units toward the right of the plot are newer and are missing some observations due to not being fully operational yet for part of this 9-month period. For the established units, a few percentages to approximately 10% of the observations fail the quality-control checks. A few percentages at different times for different units means the aggregate has a large percentage of 5-minute times when it could be missing data. When making hourly averages from the 5-minute averages, a scattering of holes in the 5-minute data lead to a larger fraction of holes in the hourly data. A way around this is to try to patch holes in the 5-minute observations at each location by assuming that the fraction of calculated clear-sky power averaged over surrounding times was the same for the flagged times. The patches then reduce the number of 5-minute periods and hourly averages with incomplete aggregate sums.

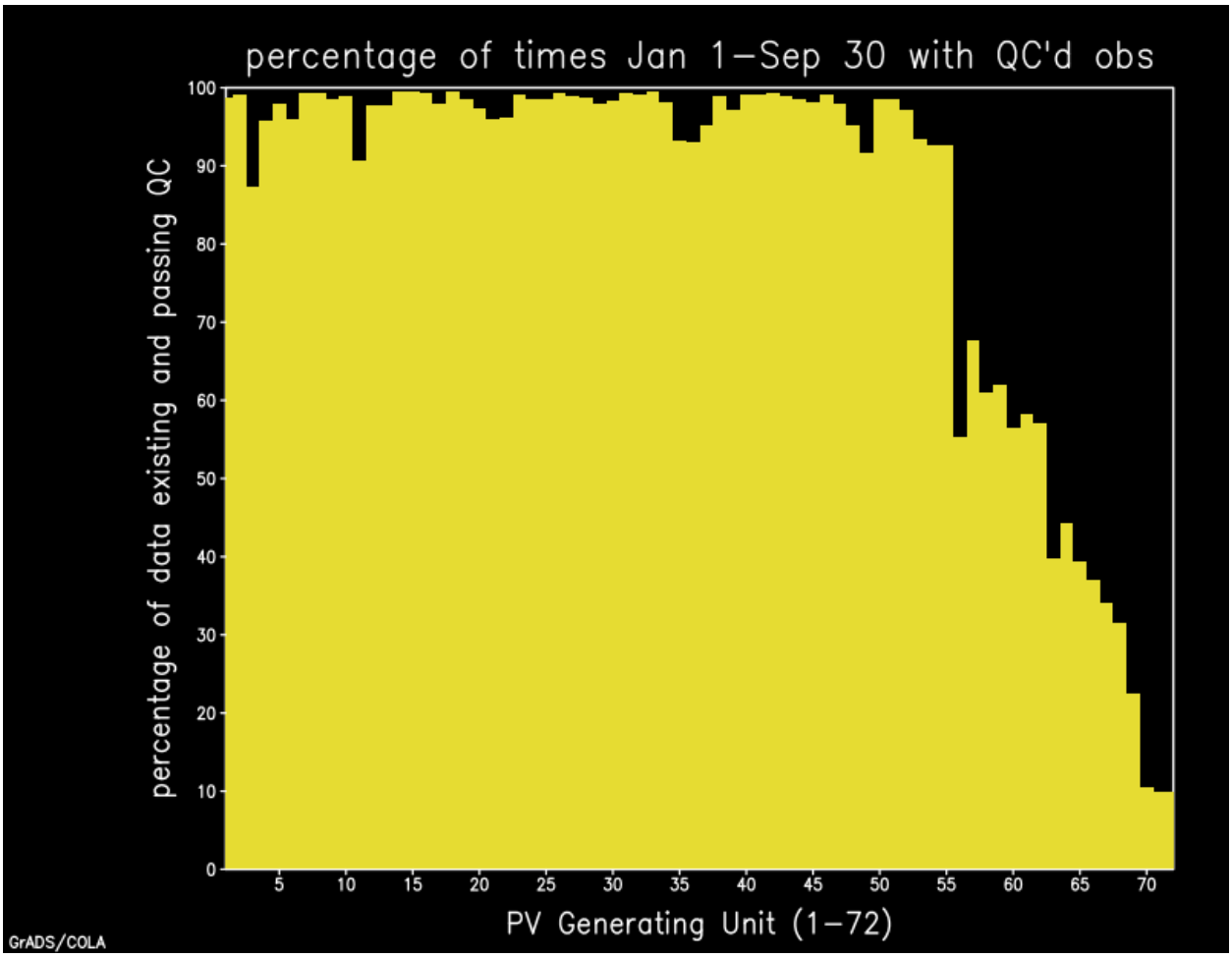


Figure 13. Percentage of 5-minute periods with power data passing the quality-control checks for each PV generating unit for the period from January 1, 2021–September 30, 2021. Small dips are due to bad data, whereas large dips are due to new units not yet being in service for part of the period.

The problem of aggregate sums is more difficult for validating probabilistic forecasts than deterministic forecasts because for deterministic forecasts, a new aggregate forecast can be created to match the passing observation set by summing the forecast over the qualifying units instead of summing over all the operating units; however, the aggregate percentiles are not the sum of the percentile forecasts from the individual units, so this method cannot be used to properly recreate forecast distributions for a subsample of all operating units.

Outages also impact the forecast, partly because the outages are often not reported correctly, so a plant might be operating at a different availability level than reported, or the outage might be reported too late, e.g., after the forecast was created, or the outage might not have been reported at all. Outages also affect some experimental forecasts differently than others or the operational forecasts due to times when they were not properly handled in one or another of the experimental forecasts. For the 9-month period from January 1, 2021, through September 30, 2021, the percentage of 5-minute daytime times not affected by a reported outage is shown in Figure 14 for each generating unit, and an estimate of the associated aggregate derate loss of power, compared to if the derate outage were not in effect, is shown in Figure 15 for the middle hour of the day for

each day in the 9-month period. Figure 15 shows derate losses both in megawatts and as a percentage of capacity. New units have outages or derates for a considerable time upon startup, but even long-established units are reporting outages of approximately or exceeding 10% of the time. For the aggregate derate, the spike near “78” on the x-axis coincides with the now-infamous winter weather event in Texas in February 2021. The large amounts in summer are from a surge in new capacity starting up, so those new units are counted in the aggregate capacity, and their derates that are large at startup are counted toward the aggregate loss.

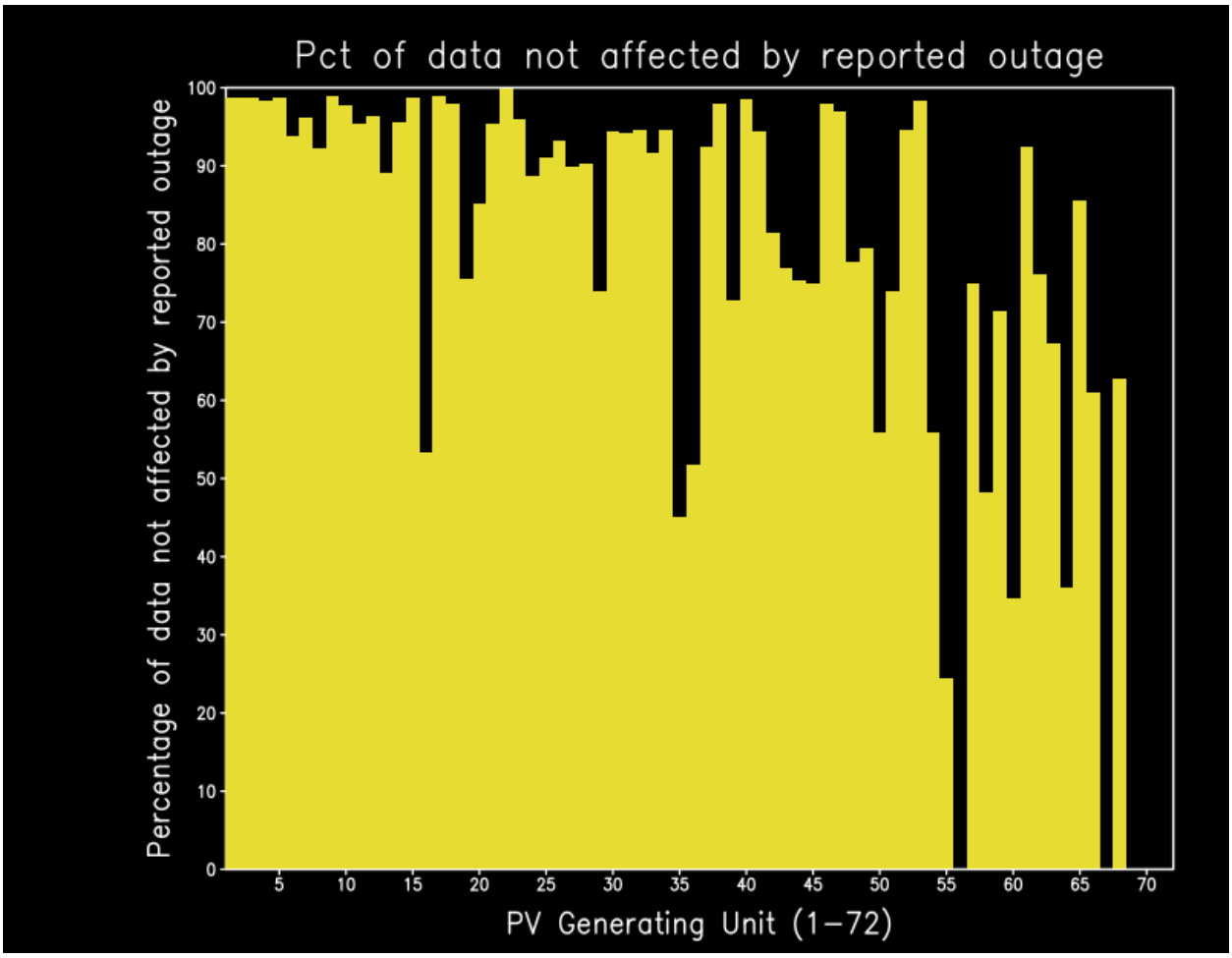


Figure 14. Percentage of 5-minute periods without a derate in effect for each PV generating unit for the period from January 1, 2021–September 30, 2021. New units typically have derates for a while after startup.

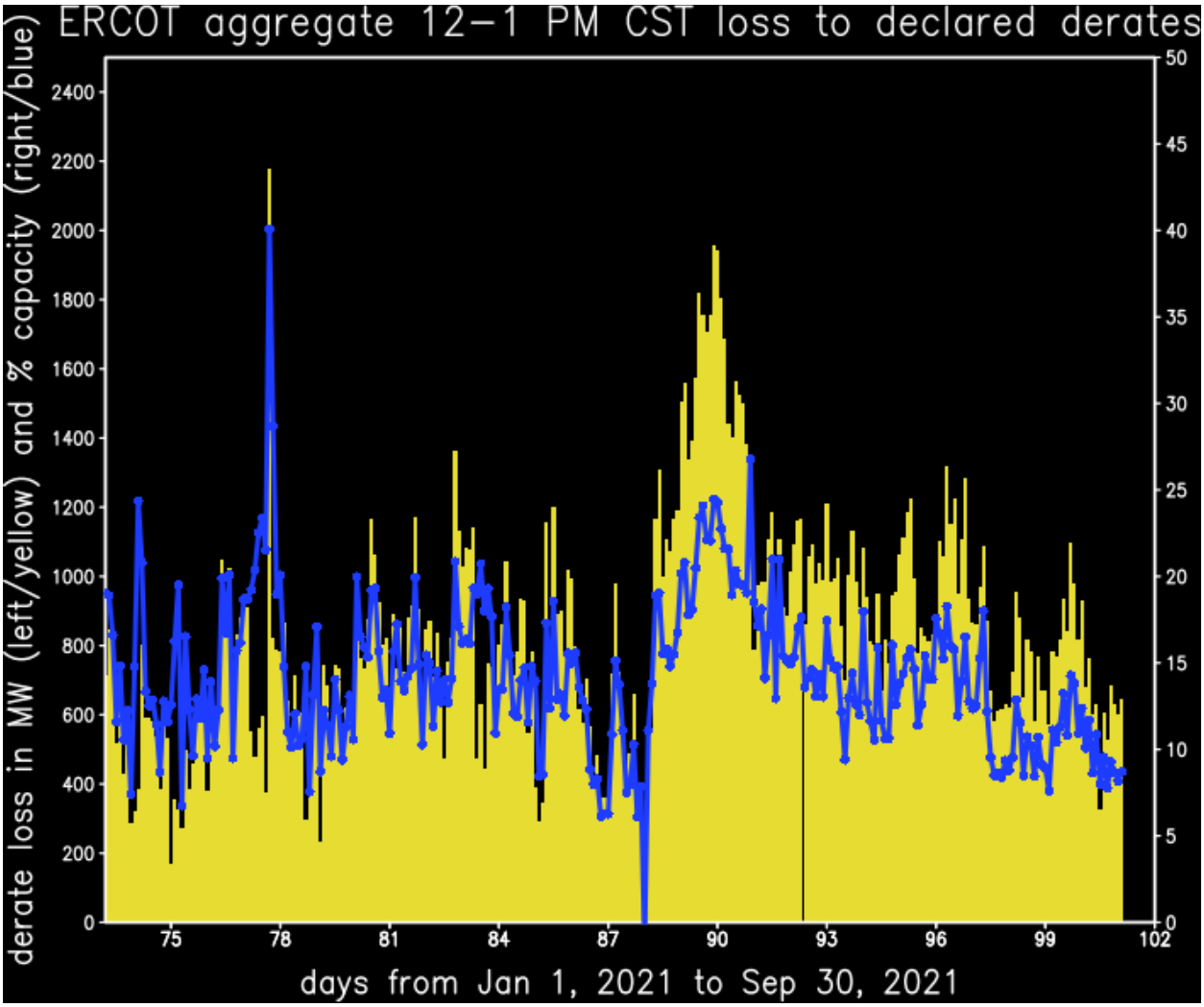


Figure 15. Daily aggregate generation lost to derates for the hour from 12 p.m.–1 p.m. CST

When outage or derate reporting is accurate and properly handled by the forecast, the forecast correctly accounts for it, and data from those times should be included in the validation analysis.

Losses due to the misreporting or non-reporting of outages cause forecast errors due to incorrect information about the availability of the unit rather than due to a poor weather forecast. Estimating such losses from observed data is a more complicated calculation. I calculated the expected power given the reported irradiance and the latest outage or derate information available in time for the day-ahead forecast and accounting for the back-panel temperature effect on efficiency. I calculated this for each 5-minute period for which the irradiance reported survived procedures 1 and 2 of the quality-control checks detailed in the first annual report. I also accounted for some units reporting GHI, whereas most reported POA irradiance. Irradiance reporting has many more problems than power and HSL due to shadowing of the small irradiance sensor, sensor soiling, poor calibration, or calibration drift, among others, but many of these are flagged only in quality-control procedure 3, so they might adversely affect my results. Even checking for only procedures 1 and 2, irradiance was flagged many times. For the times having both quality-controlled irradiance and HSL, I summed the calculated expected power for

the day and the HSL, and I made the ratio of these daily sums. Days showing up with no color indicate insufficient usable data on that day.

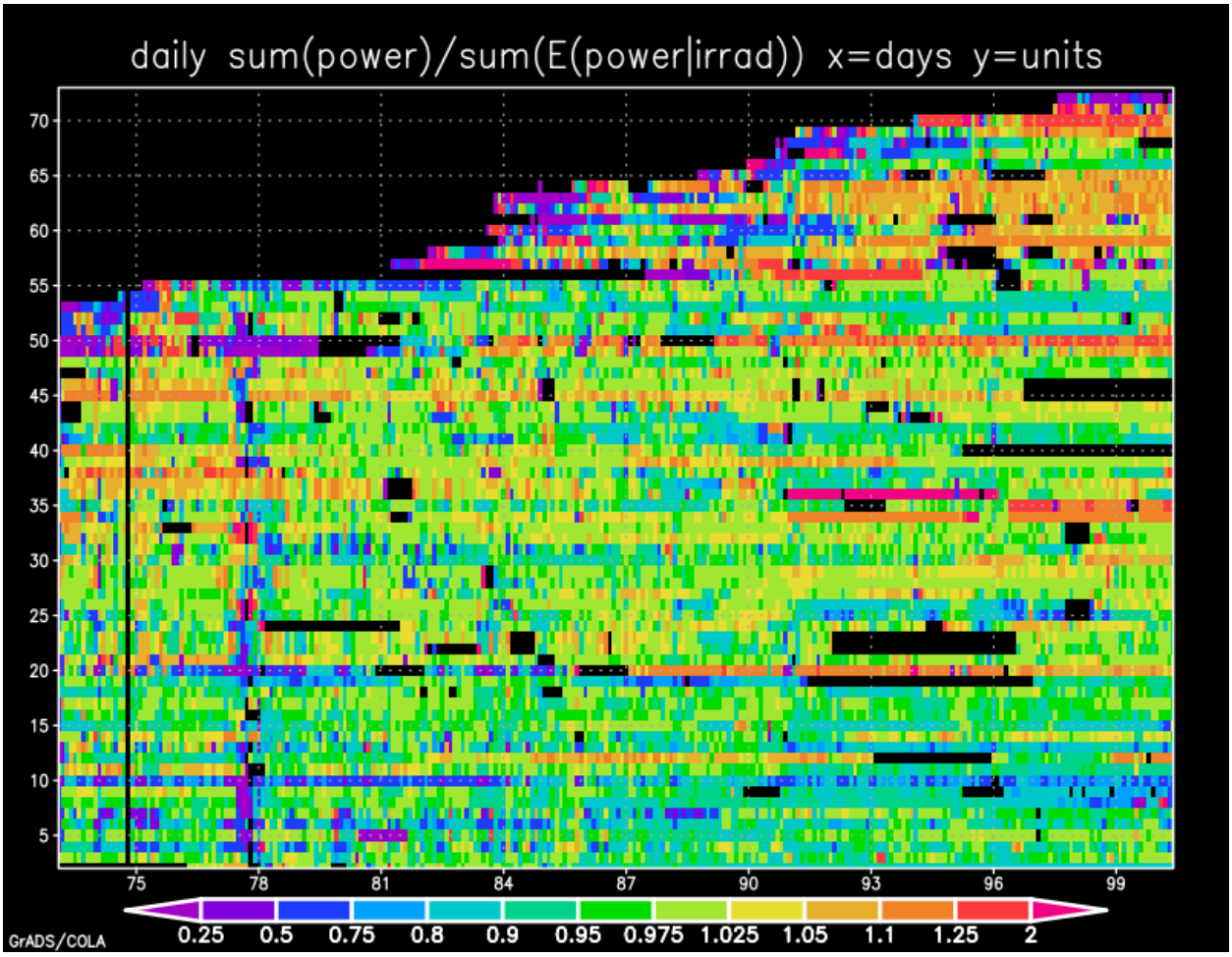


Figure 16. Estimate of missing derate fraction as a daily ratio of reported power to expected power

Expected power is based on the reported irradiance and outage status listed in time for the day-ahead forecast. A ratio <<1 indicates the unit was less available than in the reported status or the irradiance was reported much too high. A ratio >>1 indicates a unit was more available than the reported status or the irradiance was reported much too low. Rows are PV generating units, columns are days from January 1, 2021, on the left to September 30, 2021, on the right.

My calculation results are shown in Figure 16 for each unit as a row for each day from January 1, 2021, on the left to September 30, 2021, on the right. A value close to 1.0 indicates a good match. A value much lower than 1.0 indicates either lower availability of the unit than reported or the irradiance was reported much too high. A value much higher than 1.0 indicates either the unit had higher availability than reported or the irradiance was reported much too low. Some units have low values of this ratio for extended periods, and a few have periods of high values. In particular, unit 10 has low values for almost the whole 9-month period, and it turns out that ERCOT inquiries with the plant operator revealed a preponderance of failing to report derates. ERCOT inquiries turned up a large number of units that have chronically failed to report derates. Also, some units had tracker outages resulting in full availability of the panels but reduced output due to not tracking. These were not reported because there is no current mechanism for reporting this type of information. The larger unit numbers toward the top of the plot started up during the

plotted period and show the typical pattern of low ratios for reporting higher than actual availability, then higher ratios for when they are reporting correctly, but accumulated history from the earlier times therefore underestimates the expected power.

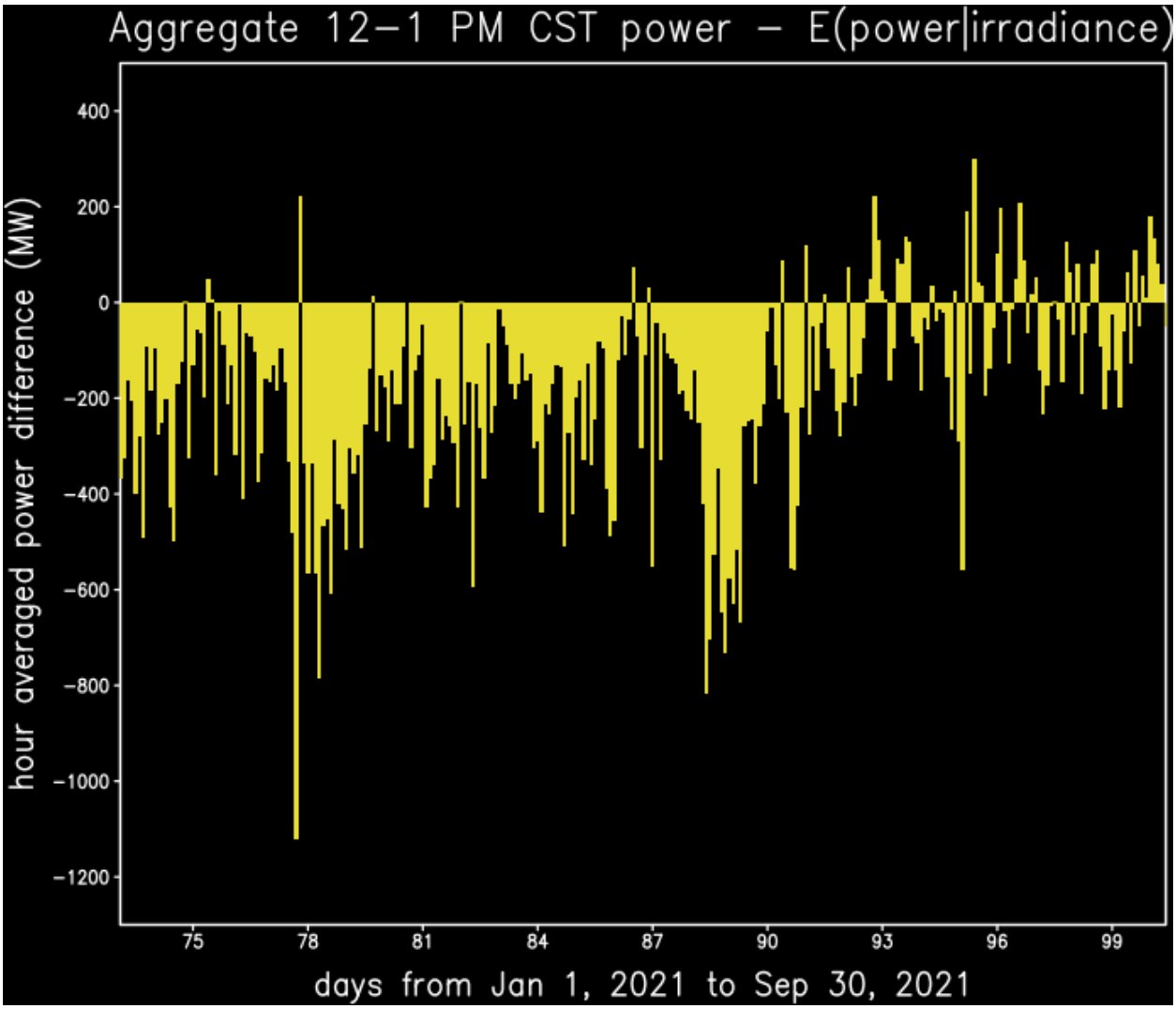


Figure 17. Estimate of effect of missing derates on daily aggregate power for the hour from 12 p.m.–1 p.m. CST

Sum over all PV generating units of power minus expected power given the reported irradiance and outage status. Negative implies missing derates, positive implies derates were overstated, and small values of either sign could be due to inaccurate irradiance reporting or imperfect estimates of power corresponding to the reported irradiance.

I estimated the impact of the missing derates on the aggregate power for the hour in the middle of the day for each day in the 9-month period. These results are shown in Figure 17, which shows an impact of several hundred megawatts for most days, although there is a rather neutral impact in August 2021 and September 2021.

Deterministic forecasts in situations with power lost to failure to report outages or derates face a bias toward overforecasting the generation. This causes balancing problems due to systematically having a shortfall of the predicted generation.

Probabilistic forecasts in situations with power lost to failure to report outages or derates face a distribution shift that affects the statistical reliability of the forecast and therefore its use in downstream applications. For example, if the outage problem caused the forecast distribution to be shifted by one standard deviation of the typical forecast error, and if the error were normally distributed (which it is not, but it is close enough for illustrative purposes), then observations will be smaller than the 50th percentile forecast 84% of the time instead of 50% of the time, and they will be smaller than the 20th percentile forecast 56% of the time instead of 20% of the time. This could have a major impact on the assessment of risk.

In either case (deterministic or probabilistic forecast), it causes forecast errors to show up in the validation that are caused by the information provided to the forecast system rather than errors caused by the forecast system itself. This complicates the skill comparison among different forecast systems.

Validation Methods to Address Complications

To address all these matters, validation statistics were computed in a variety of ways, all using patched observations to fill small holes resulting from removing flagged observations after passing them through the first two quality-control procedures detailed in the first annual report.

- Five-minute observations for the aggregate were patched for times when missing units expected in the aggregate had a total clear-sky power calculated to be less than 10% of the calculated clear-sky total for all units expected in the aggregate. In those cases, the aggregate was assumed to include the missing units using the clear-sky fraction for the total of all the not-missing units.
- Hour observations for the units and the aggregate were then patched for times when 9, 10, or 11 of the 12 5-minute periods in the hour had quality-controlled data, assuming the same clear-sky fraction for the entire hour as was calculated for the available times.

Statistics were then calculated using cleaned and not-cleaned forecasts, where forecasts were "cleaned" by:

- Ensuring monotonic percentiles from high to low (This affected only a few BMA forecasts due to a programming problem. Some low forecasts were calculated as negative, but instead of truncating to zero, used the absolute value; this procedure set those back to zero.)
- Removing times of outages for individual unit validations but not modifying the aggregate, which was already computed as above
- Trimming excess for high percentiles (which affected only forecasts before this was done correctly in the forecast) by limiting to the minimum of: clear-sky power + 5% of capacity, or 1.2 * clear-sky power.

Statistics for the 5-minute forecasts were then calculated for all valid times and separately omitting times when the smart persistence should not have been allowed to be used, such as due to low sun angle and long lead time.

Validation in the Solar Forecast Arbiter (SFA) used the patched quality-controlled observation set as just described but only for uncleaned (original) forecasts and did not remove times when inappropriate smart persistence might have been used.

Validation of the ramp forecasts used the raw, not quality-controlled, HSL and the raw, not cleaned, forecasts, so they would exactly match the delivery against the reported observations. The goal of the ramp validation was to compare what ERCOT is receiving to what ERCOT is reporting. The goal of the other in-house validations was to better compare the various forecasts using the most accurate validation data and mitigating the myriad complicating factors. The goal of the SFA validations was to perform third-party evaluation.

Validation in Solar Forecast Arbiter

The SFA (https://solarforecastarbiter.org) is a web platform developed by the U.S. Department of Energy Solar Forecasting II team awarded under Task Area 1 to provide metrics standardization to enable a consistent methodology for evaluating solar irradiance and solar power forecasts. SFA has a versatile manual user interface to view and generate validation reports, including interactive visualization of the data. It also contains an application programming interface for the upload and download of forecast and observation data; a system for defining the sites, observation data, and forecast data; and a variety of flexible options for the data analysis and validation. SFA, however, was unable to produce validation reports for the aggregate generation of a collection of sites, although it allowed defining the aggregate and importing probabilistic forecasts for the aggregate. Also, SFA cannot manage forecasts with overlapping valid times, such as forecasts for the next 24 hours updated every hour or for the next 48 hours updated daily. Further, SFA has a limited set of probabilistic forecast validation metrics to go along with its more extensive variety of deterministic and categorical metrics, and it tries to get around that by considering categorical or other metrics applied to individual percentiles within a probabilistic forecast.

As part of SUMMER-GO, Maxar was tasked to use SFA to validate some probabilistic power forecasts. Due to the limitations of SFA, more extensive in-house validation was performed, as described in other chapters that follow. For the SFA validation, three sets of probabilistic power forecasts containing 99 percentile values (1, 2, 3, …, 99) were evaluated in SFA:

1. Day-ahead forecasts issued at 2 p.m. CST during standard time (winter), with hourly forecast averages for each hour of the day, which ERCOT uses for market clearing/scheduling for the day-ahead time. Evaluations were during standard time dates from November 2, 2020, to March 14, 2021.
2. The same as (1) but issued at 2 p.m. CDT during daylight saving time (summer), with evaluations covering March 15, 2021, through September 30, 2021 (end of SUMMER-GO)
3. The rolling 10-minute-ahead, 5-minute average power, updated every 5 minutes, with evaluations covering January 1, 2021, to September 30, 2021.

The SFA does not allow combining sets (1) and (2) because they have different initial times, 20:00 UTC in winter and 19:00 UTC in summer.

During the work with SFA, we had a number of phone and email discussions with the SFA team. We found several bugs, which were then fixed, and we discussed various concerns, plans, and strategies.

We decided to upload the SUMMER-GO-2 forecasts to SFA for evaluation, anticipating that the forecasts delivered to ERCOT and NREL would be switching to this set, and to evaluate both 5-minute and hourly intervals (recall that the delivered SUMMER-GO-1 forecasts had only 5-minute intervals output all the way out to 48 hours); however, that switch did not happen as anticipated for a number of reasons related to programming, feasibility, and a lack of improvement over SUMMER-GO-1, particularly for 10-minute lead times. These forecasts were all uploaded using the SFA application programming interface in batches by location, with backfilling to earlier dates for a set of locations and then forward filling in real time. As new PV generating units were added to the ERCOT grid and to the forecast, forecasts for those locations were later backfilled and added going forward in real time.

Unfortunately, the evaluation was affected by outage handling problems and the use of smart persistence problems in this set of experimental forecasts. The HSL observations of potential power reported by ERCOT were manually uploaded after being quality-controlled. This Maxar quality control removed stuck logger values and other artifacts but did not remove inconsistencies between reported irradiance and HSL or power. Many, although not all, of those inconsistencies were due to problems with the irradiance monitors rather than bad values of power or HSL. Because attribution often is not possible in each instance, the removal of HSL values in all cases of inconsistency with irradiance would have significantly reduced the number of observations that could be used. Because those were allowed to remain, there could be some cases where bad HSL reported values were used.

A set of 29 PV generating units were selected for evaluation from the 72 that were operating by the end of SUMMER-GO. A large number were recently added to operation, and thus they did not have much accumulated history to evaluate. Others had a large number of days with derates (partial or total outages) that would have affected the forecast evaluation. The plants that appeared to be most stable in their operations and had data for most of the time periods were selected for evaluation.

The forecasts were evaluated using the continuous ranked probability score (CRPS), which is implemented in SFA using the forecast and observed power in megawatts; thus, the scores are notionally errors in megawatts, with larger score values for plants with larger capacity. CRPS measures a combination of the forecast probability spread (a sharper distribution leads to a better (lower) score if the distribution is centered on the verifying value) and forecast accuracy (a higher (worse) score if the center of the forecast distribution is shifted farther from the verifying value); thus, a high "error" is indicated from either a bad forecast in the deterministic sense that the bulk of the distribution is far from the verifying condition or in a distribution sense that the forecast indicated high uncertainty. The CRPS scores shown here generally range from approximately 5%–10% of capacity, which is similar to the range of mean absolute error (MAE) or root mean square error (RMSE) for deterministic forecasts of large utility-scale solar power plants in this geographic area.

Reports were generated for each of the 29 units using the SFA user interface and downloaded. For the day-ahead forecasts of hourly averages, both Portable Document Format (PDF) files and interactive HyperText Markup Language (HTML) files were readily downloaded. For the 5-minute interval forecasts, the SFA reported generation failed due to the plots of the large number of data points, 288 per day for each of 99 percentiles; however, components of the reports did get

generated. We spent some considerable time obtaining the plots using the application programming interface to retrieve a collection of report objects and using Python to extract the plots from that collection. This convoluted method was tedious. The SFA team responded by modifying the report generation code to skip this feature when the number of data points would be too large, allowing for the generation of PDF reports. This was much more helpful, and I proceeded to create and download the PDF reports for the 5-minute-interval forecasts.

I compiled the 20-page PDF reports for each of the 29 sites and prefaced it with an introductory explanation, yielding three 583-page supplemental volumes: one for the CST day-ahead forecasts, one for the CDT day-ahead forecasts, and one for the rolling 10-minute-ahead forecasts. Also, I compressed and zipped the 58 interactive HTML versions of the reports, creating a 14-MB archive file included as a supplement to this report.

The remainder of this section highlights the key results by showing the average CRPS for each evaluation period by hour of the day for two solar power plants. One is in central Texas, and the other is in west Texas, which are in two different climate regions in SFA. One was recently repowered with bifacial panels and new trackers. The other uses thin film.

Figures 18 and 19 show the time of day CRPS for the 10-minute lead time forecasts at the central Texas bifacial unit and at the west Texas thin film unit. The score as a fraction of capacity is somewhat lower at the sunnier, less variable west Texas unit and shows highest error during the morning and evening ramp periods. The error is larger than expected for such a short lead time, possibly partly because of the inappropriate use of clear-sky values substituting for smart persistence in the early morning; because of input source weighting, which could be optimized better; and because of the larger error in predicting 5-minute intervals than hourly intervals.

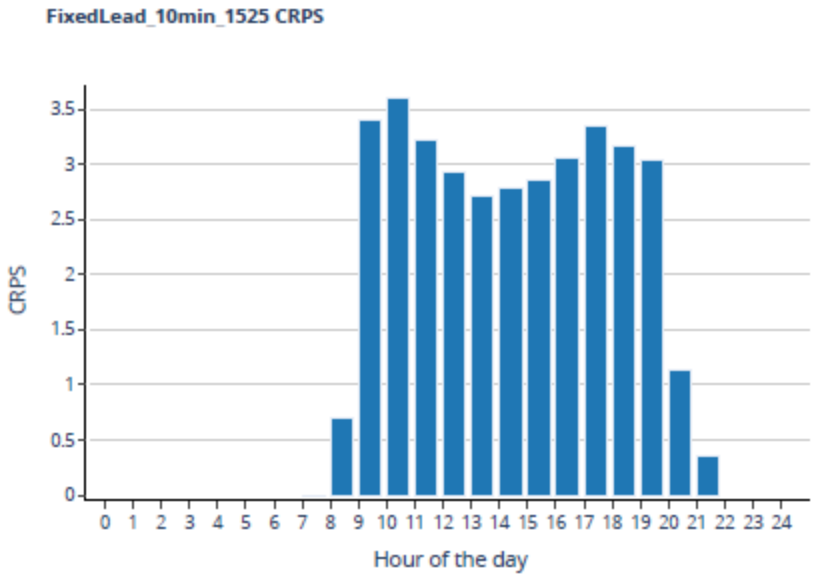


Figure 18. SFA plot of CRPS by hour of the day for the rolling 10-minute lead time forecasts of the 5-minute averaged power for the period from January 1, 2021–September 30, 2021, for a repowered 39-MW AC PV generating unit with new bifacial panels in central Texas

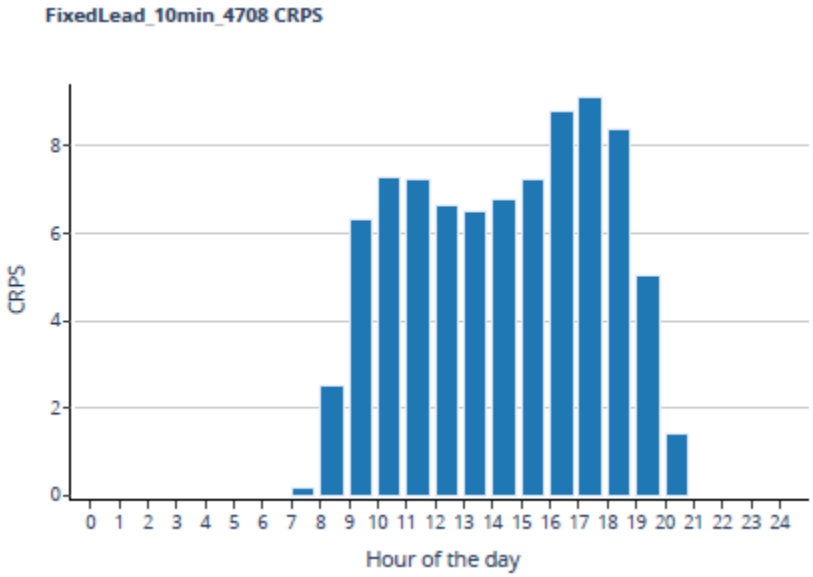


Figure 19. SFA plot of CRPS by hour of the day for the rolling 10-minute lead time forecasts of the 5-minute averaged power for the period from January 1, 2021–September 30, 2021, for a 125-MW AC PV generating unit with thin film panels in west Texas

Figures 20 and 21 show the same for the day-ahead hourly averages for the warm season. They show pronounced peaks during the morning and evening ramp periods, when the trackers are moving the panels and the tilted panels have more of a direct beam component in the POA irradiance, whereas the middle of the day POA is nearly the same as the easier-to-predict GHI.

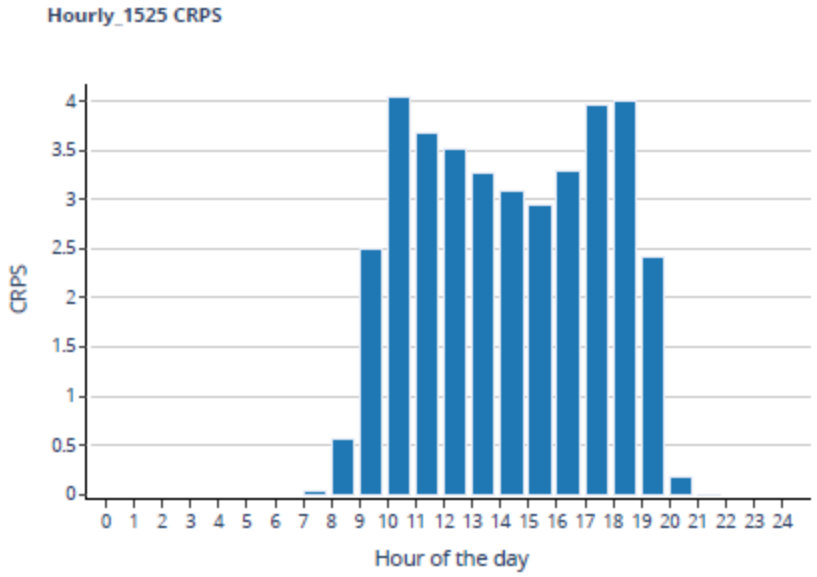


Figure 20. SFA plot of CRPS by hour of the day for the day-ahead forecasts of the hourly averaged power for the period from March 15, 2021–September 30, 2021, for a repowered 39-MW AC PV generating unit with new bifacial panels in central Texas

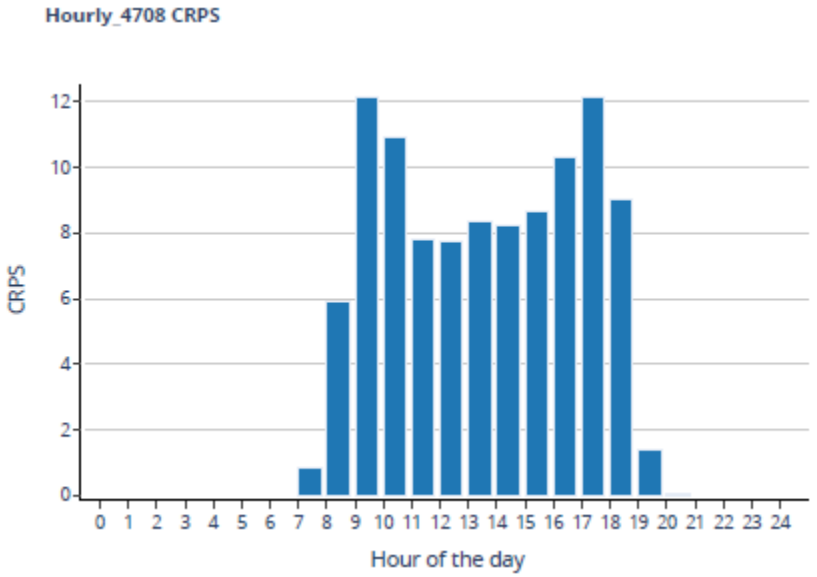


Figure 21. SFA plot of CRPS by hour of the day for the day-ahead forecasts of the hourly averaged power for the period from March 15, 2021–September 30, 2021, for a 125-MW AC PV generating unit with thin film panels in west Texas

Figures 22 and 23 show the same for the winter season. The overall shape is similar to the warm season except for the peculiar error bars at the late hours after the end of the day. The midday CRPS is somewhat smaller in winter due to the smaller envelope of possibilities because the daily peak generation is smaller due to the lower sun angle on these single-axis tracking arrays.

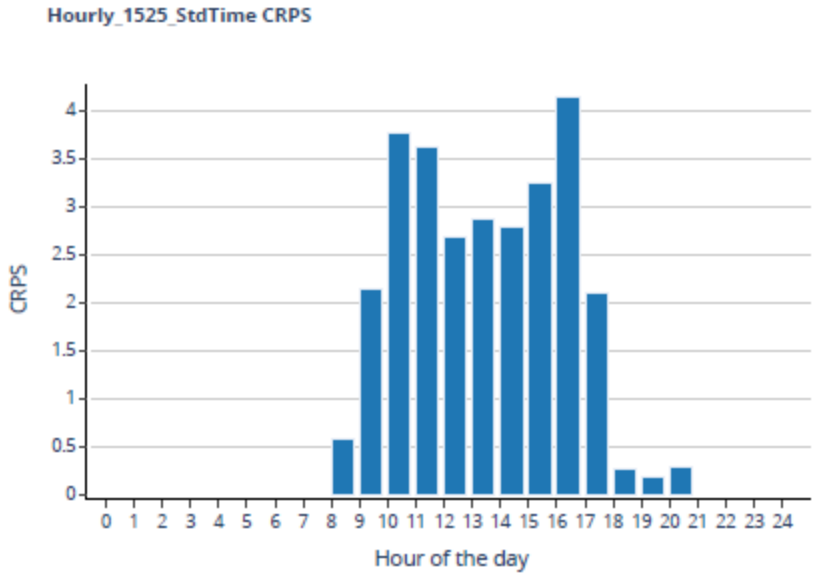


Figure 22. SFA plot of CRPS by hour of the day for the day-ahead forecasts of the hourly averaged power for the period from November 2, 2020–March 14, 2021, for a repowered 39-MW AC PV generating unit with new bifacial panels in central Texas

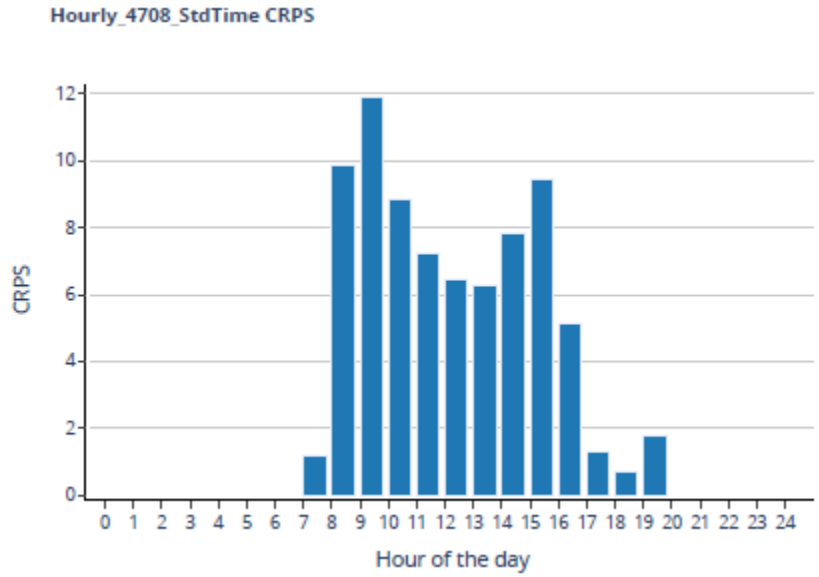


Figure 23. SFA plot of CRPS by hour of the day for the day-ahead forecasts of the hourly averaged power for the period from November 2, 2020–March 14, 2021, for a 125-MW AC PV generating unit with thin film panels in west Texas

Validation of Five-Minute Ramp Rates

On June 1, 2021, ERCOT started a new dispatch algorithm for the upcoming 5-minute period based on system-wide load from the last 5-minute interval and a series of weighted terms involving changes. One change term is the forecast of the 5-minute system-wide solar generation ramp for that upcoming 5-minute period; therefore, skilled evaluation of this ramp rate forecast is paramount.

ERCOT has been ingesting the SUMMER-GO-1 forecasts for this purpose, so those will be the evaluated forecasts. They will be evaluated as received, without any data cleaning.

The raw HSL observations with no quality control were used in the validations to use exactly what was supplied by ERCOT, although that includes some stuck loggers and other artifacts. The aggregate observation is made by summing over all the same units as in the forecast. A particular problem recently uncovered is that sometimes the real-time telemetry fails to load into the feed ERCOT sends, and instead the observations from the previous 5-minute period are sent again, e.g., the entire telemetry set for every parameter at every generating unit is identical to the previous 5-minute period except the time stamp has been advanced to current. This was discovered after all the validations were completed, and the validations have not been rerun. This is especially a problem for validating 5-minute ramp rates because when one of these episodes occurs, the ramp rate will appear to be zero for the ramp ending at the time of the repeated observations, and then it will be double for the next 5-minute period when the next set of actual fresh observations arrives. Some days have none of these duplicates, and some days have as many as 30, with 2 or 3 on many days. Given 144 5-minute daytime intervals on average, this problem affects perhaps a few percentages of the time intervals overall, so hopefully it does not have large impact on the results.

The timing definitions used in the validation are as follows:

- Forecast ramp = second 5-minute average minus first 5-minute average
- Example: Forecast made at 10:15, delivered at 10:18
- *The ramp is the forecast for the 5-minute average period ending at 10:25:00 minus the forecast for the 5-minute average period ending at 10:20:00.*

The observed ramp uses the reported HSL for the matching valid times:

- Example: 5-minute average HSL ending at 10:25 minus 5-minute average HSL ending at 10:20.

The persistence ramp assumes that the ramp rate from the latest telemetry observation pair at the forecast time persists:

- Example: The last two observed times before the dispatch for the 5-minute period beginning at 10:20 are 5-minute periods ending at 10:15 and 10:10.
- *Persistence ramp = HSL for the 5-minute period ending at 10:15 minus the HSL ending at 10:10.*

Note that due to data availability latency, although short, the 10:15 observation might not be available to Maxar's forecast delivered at 10:18, but it would be available to ERCOT in house when the Maxar forecast arrives at 10:18, so this definition of persistence ramp is fair to use in comparison to the forecast because they are two estimates concurrently available to ERCOT.

The validation set included only daylight times when the observed, persistence, and forecast ramps were all available. The aggregate comprised all PV plants that were operational and included in the 5-minute forecast. Some plants were included in the 5-minute forecasts before they became operational, and other plants were not included until they had been operational for a while; in either case, those plants were omitted from the aggregate because either a forecast was

missing or the plant should not have been included. This did not omit much because it generally happened around plant startup, when the plant was not yet fully available anyway.

Error statistics were made for rolling 30-day periods by hour of the day for dates beginning January 1 and rolling forward until ending in early June; thus, each sample contains 30 days x 12 5-minute periods = up to 360 5-minute ramp rate errors. Evaluations were made for the ramp rate errors in (MW/5 minutes) and, to account for the increasing capacity during the 6 months of data, also divided by capacity, e.g., error is (percentage of capacity/5 minutes).

Evaluations were performed for the shortest lead time ramp rate, as in the definition examples. Evaluations were also performed for lead times 5 and 10 minutes later, which found that the Maxar forecast ramp rate errors are only slightly worse for those slightly longer lead times, whereas persistence is much worse; thus, ramp rate forecast improvement over the persistence of the ramp rate increases as lead time increases.

Figures 24–30 show, for each hour of the day, the MAE of the forecast ramp rate for the rolling 30-day periods from January 1 (left) to September 1 (right), thus, 30 days ending from January 30 (left) to September 30 (right). Each plot shows the MAE in megawatts for the forecast ramp (orange) and persistence ramp (dark blue) and the MAE as a percentage of capacity for the forecast ramp (yellow) and persistence ramp (sky blue). The error increases seasonally because the power envelope is wider, and more generation is possible, in late spring and summer than in winter; thus, the errors are also larger. The curves for errors in megawatts are steeper upward with the season than the curves for error as a percentage of capacity because new sites coming online increased capacity. The forecast ramp rate had smaller MAE than a persistence ramp rate for all 30-day periods at all times of day except early morning, and by the last 30-day period, the forecast ramp even in the early morning was better.

The errors as a function of time of day for the 9-month period from January 1, 2021, to September 30, 2021, and for the period when ERCOT had been using the forecast operationally from June 1, 2021, to September 30, 2021, are shown in Figure 31. As before, the blues are persistence, and yellow/orange are the Maxar forecasts.

The upper left panel shows the MAE as a percentage of capacity. It shows that the error has a small peak during the morning ramp, dips during the day, and has a larger peak in the afternoon before or near the start of the evening ramp. The shape is the same for both the persistence error and the Maxar forecast error, with Maxar having a smaller error, except in the early morning. The errors are slightly smaller for the June 1 to September 30 period than for the January 1 to September 30 period for both the Maxar and persistence ramp forecasts.

The other panels in Figure 31 show the median error and percentiles of the upper tail of the error distribution. For example, the 75th percentile of the error distribution is the error such that 3 of 4 forecasts have a smaller error. The ramp rate error is positive if the ramp is too upward in the forecast relative to the observation. An up ramp that is too steep in the forecast results in a positive error, as does a down ramp that is not steep enough in the forecast; thus, positive errors can be associated with overforecast and insufficient power for balancing. The median error is near zero, not much bias. The improvement of the Maxar forecast ramp over the persistence ramp increases for the largest errors (high end of the error tail), which is particularly valuable.

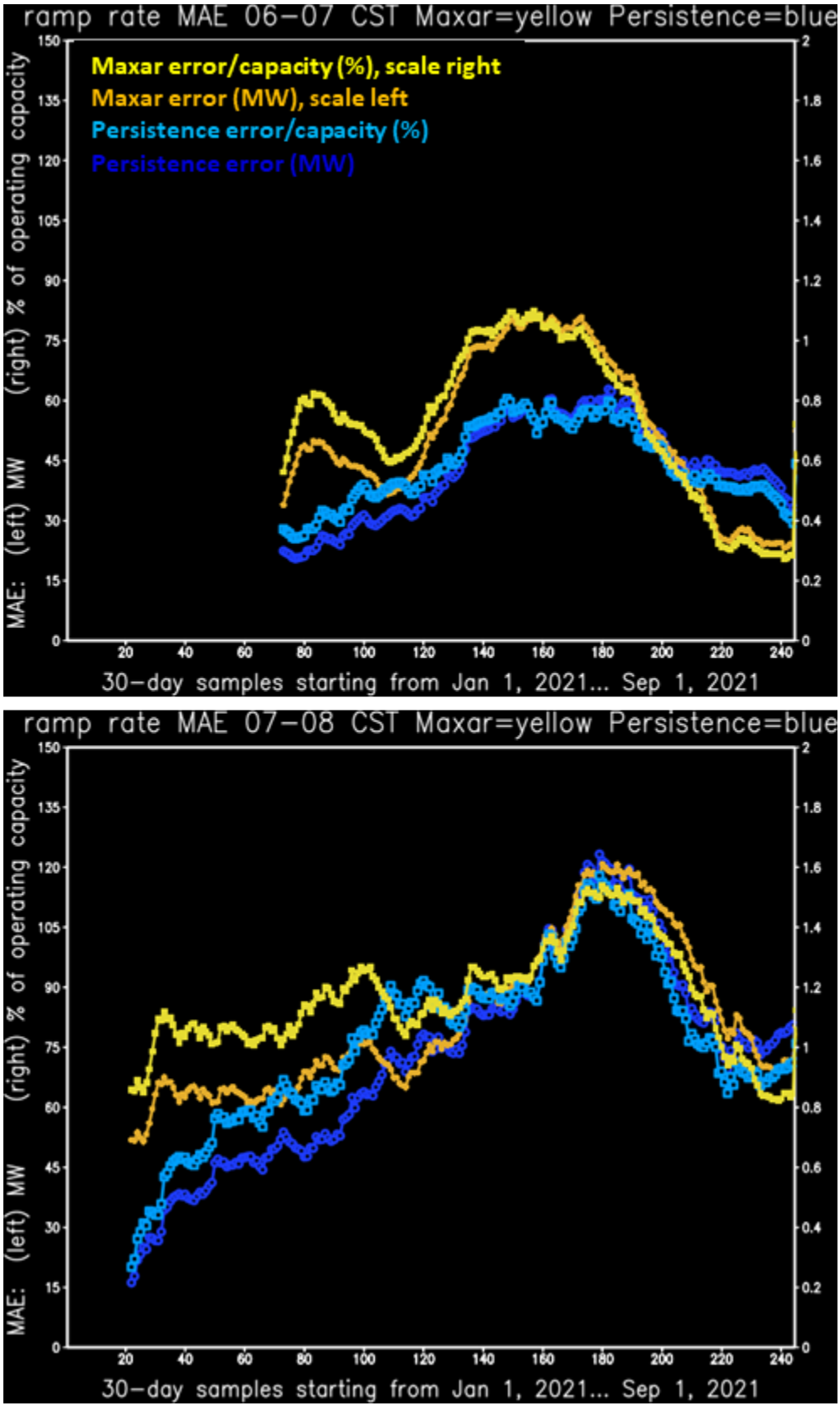


Figure 24. Rolling 30-day MAE of early morning 5-minute ramp forecasts with lead time 10 minutes

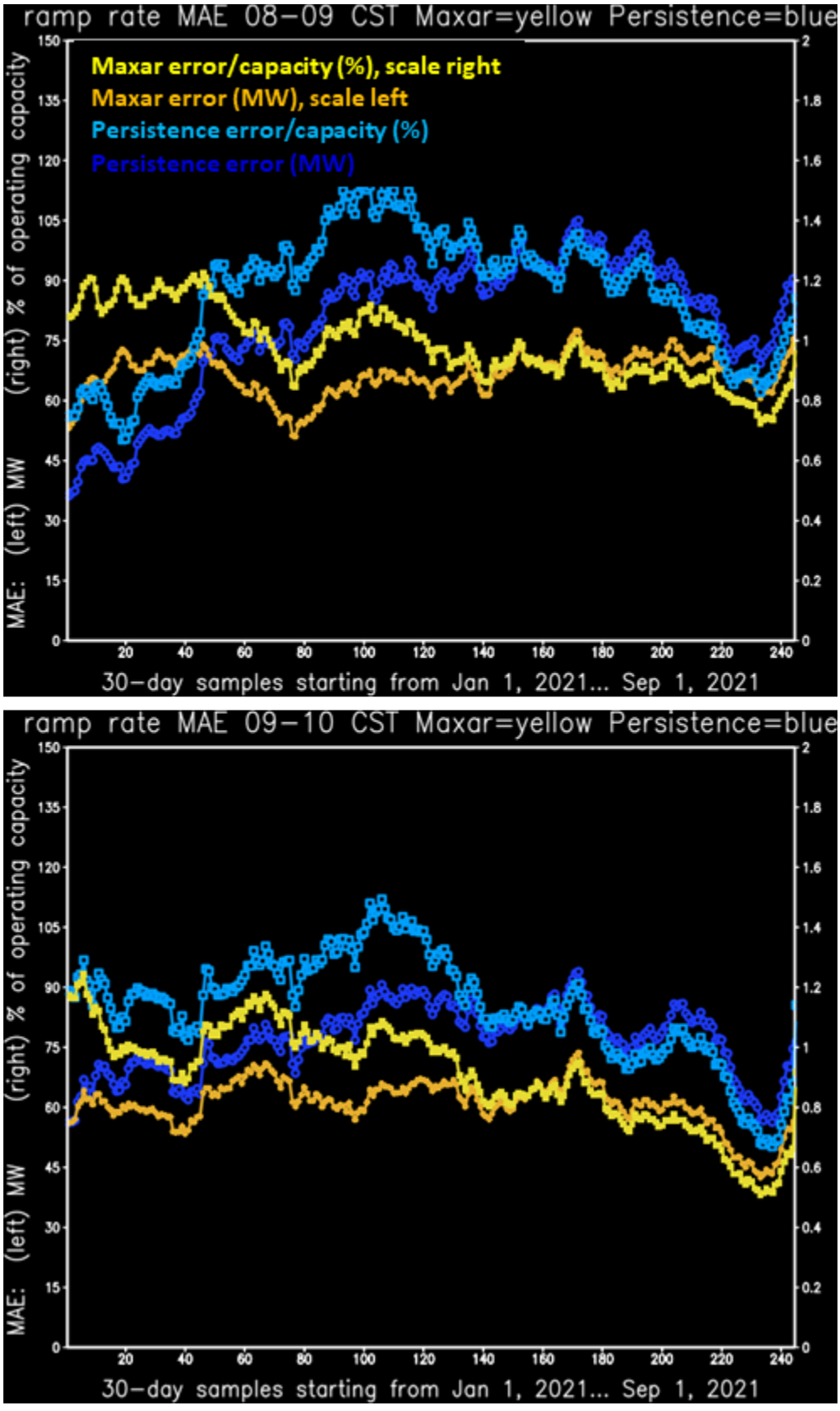


Figure 25. As in Figure 24 but for the hours ending at 9 a.m. and 10 a.m. CST

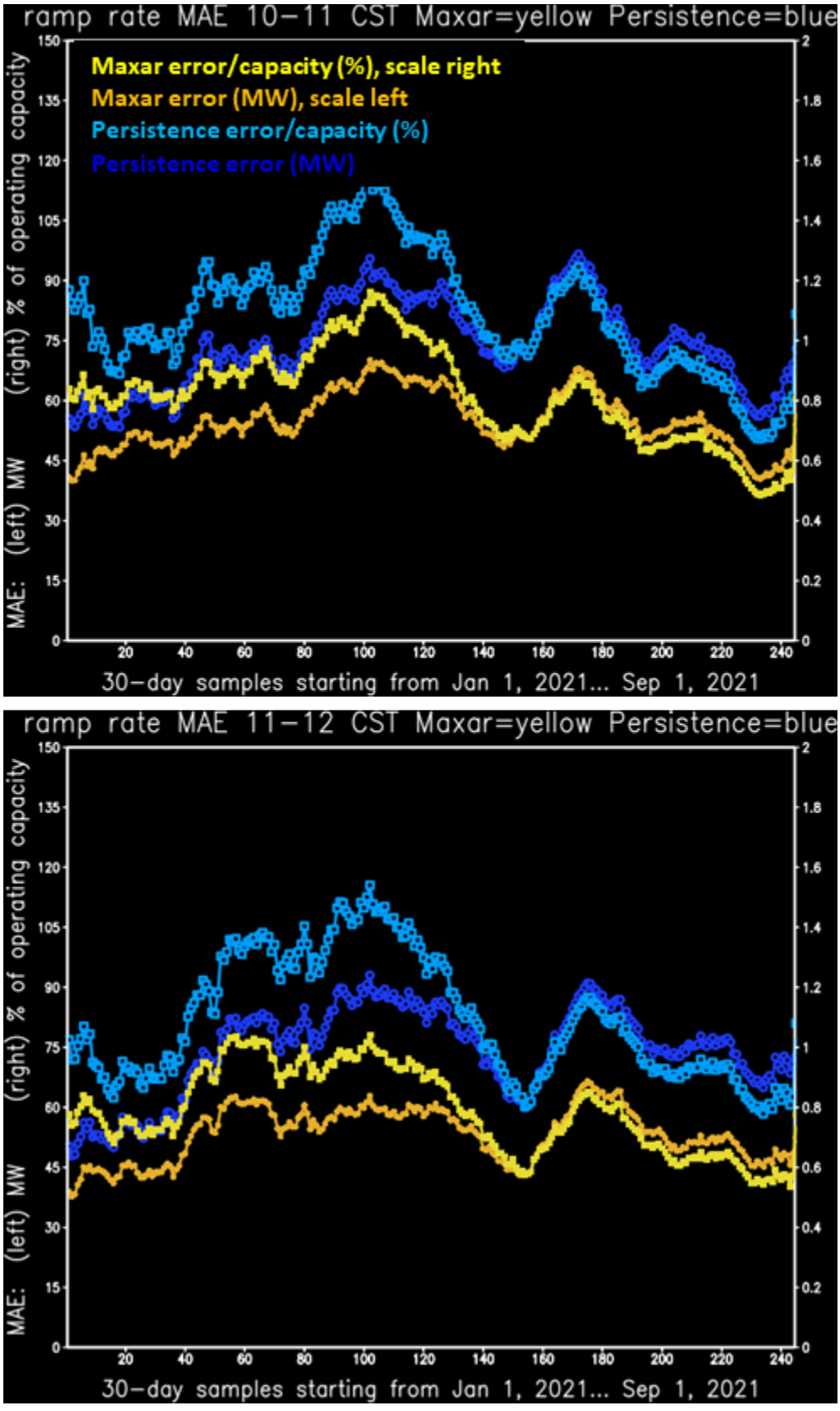


Figure 26. As in Figure 24 but for the hours ending at 11 a.m. and 12 p.m. CST

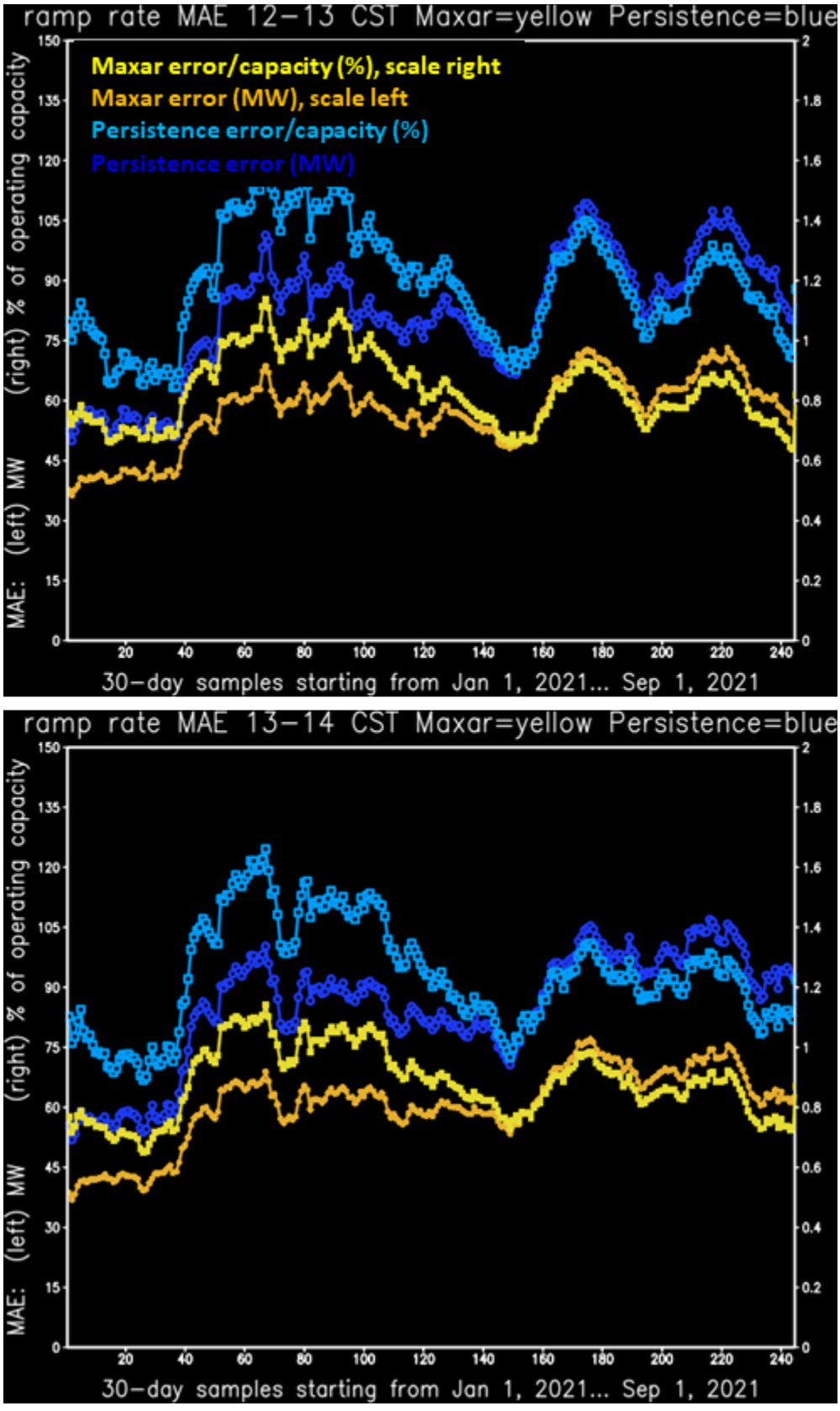


Figure 27. As in Figure 24 but for the hours ending at 1 p.m. and 2 p.m. CST

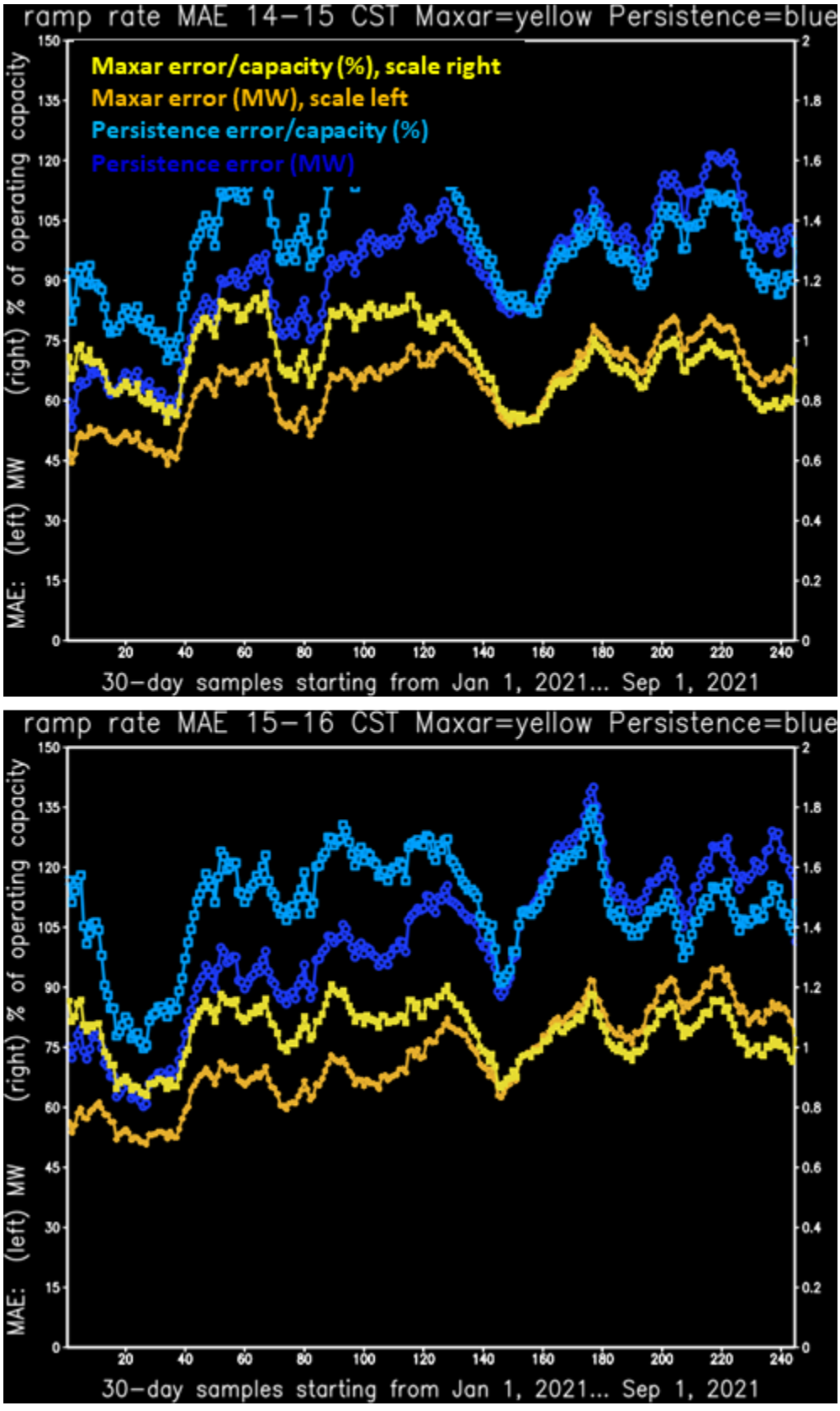


Figure 28. As in Figure 24 but for the hours ending at 3 p.m. and 4 p.m. CST

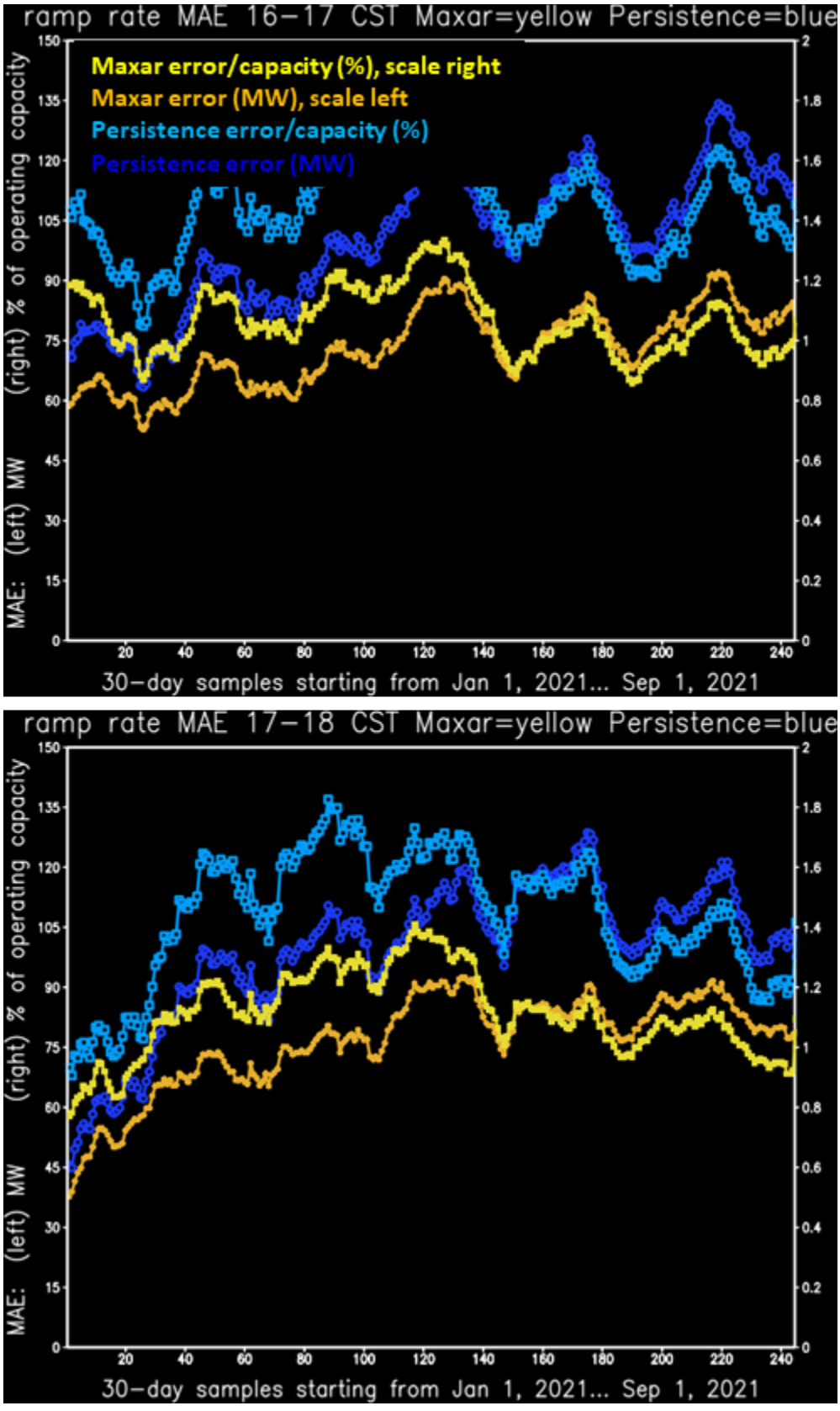


Figure 29. As in Figure 24 but for the hours ending at 5 p.m. and 6 p.m. CST

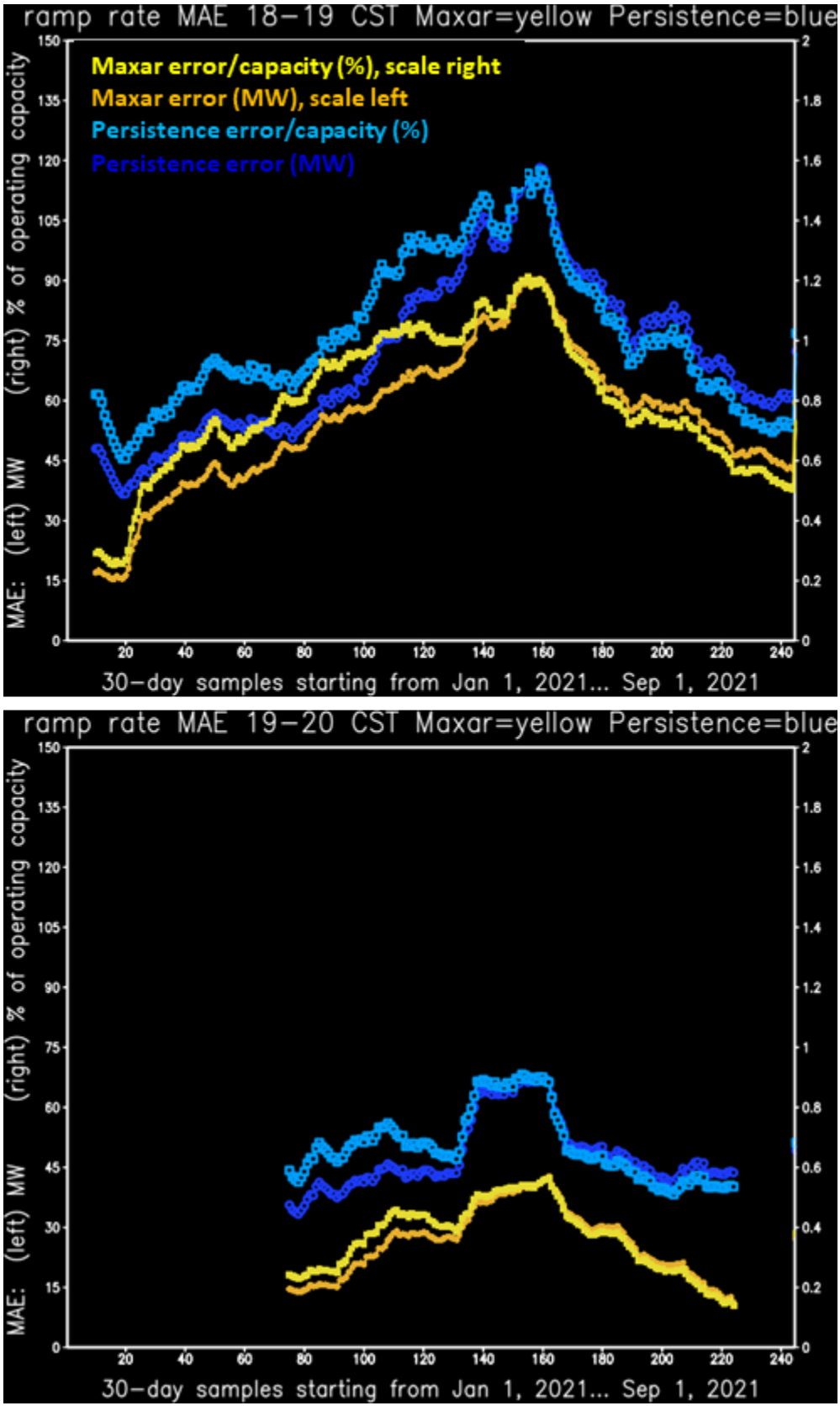


Figure 30. As in Figure 24 but for the hours ending at 7 p.m. and 8 p.m. CST

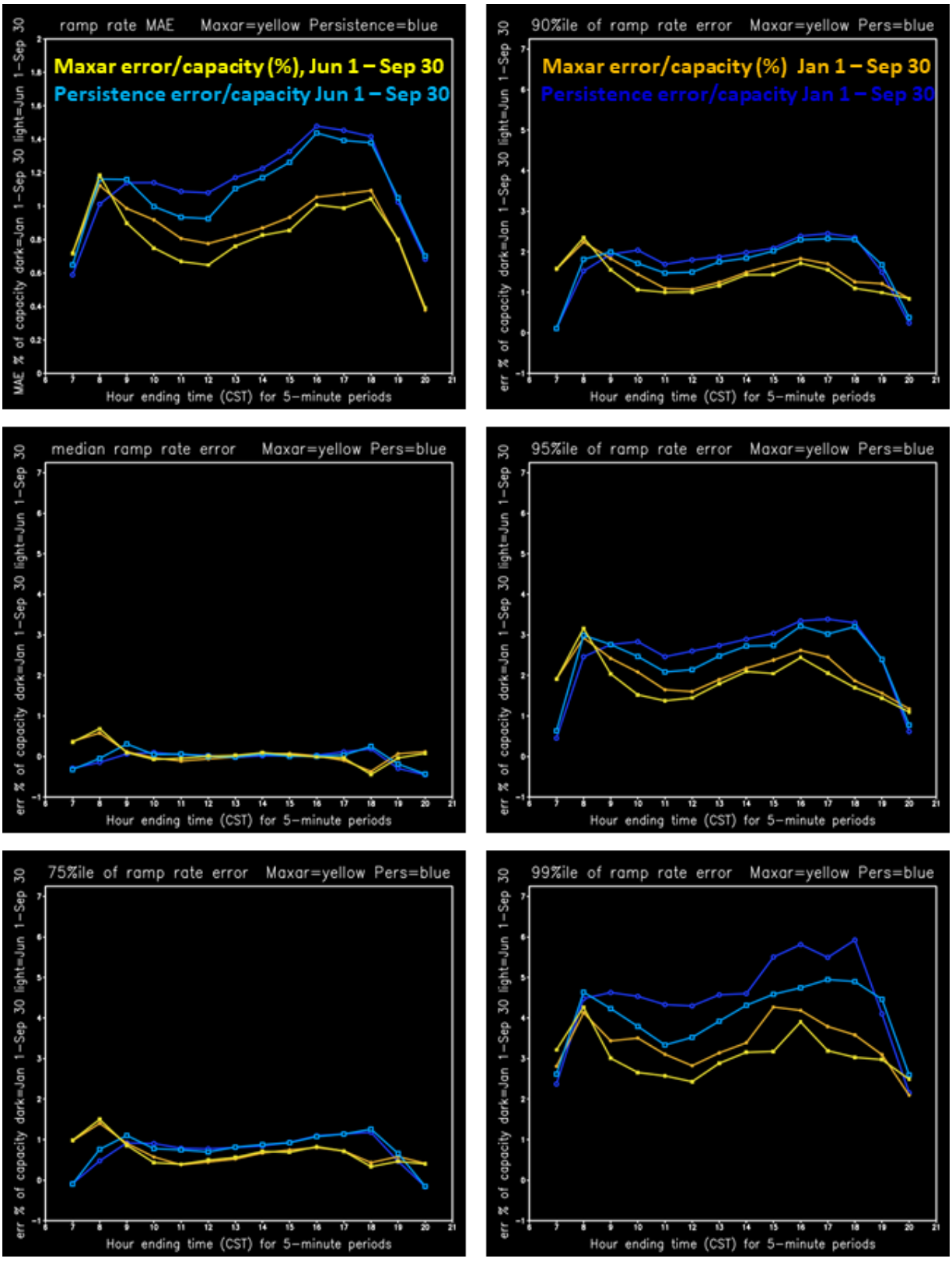


Figure 31. Error versus time of day for the 5-minute ramp forecast with the 10-minute lead time. The upper left is MAE, and the other panels are the percentiles of the error distribution from the median to the upper tail of the error.

The ramp rate errors in Figure 31 quantify how often we can expect errors of a particular magnitude. The 95th percentile of the 5-minute ramp rate error for the Maxar forecast is approximately 2% of capacity during the afternoon. That means that 1 in 20 forecasts in late afternoon will have an error magnitude of 200 MW or more in only 5 minutes for a system with 10 GW of solar capacity.

To view the skill of the Maxar forecast relative to the persistence forecast, a standard skill score is shown in Figure 32, similarly as a function of the time of day. The skill score is defined as:

$$\text{Skill score} = 100\% * [\text{MSE(forecast)}-\text{MSE(reference)}] / [\text{MSE(perfect)}-\text{MSE(reference)}]$$

$$= 100\% * (1 - [\text{MSE(forecast)}/\text{MSE(persistence)}])$$

where mean square error (MSE) is the usual metric, and MSE(perfect) is zero. The skill score indicates the percentage improvement in MSE of the forecast compared to the reference. The Maxar forecast ramp rate reduces the MSE by 50% compared to the persistence of the ramp rate observed at the most recent time available when the forecast arrives.

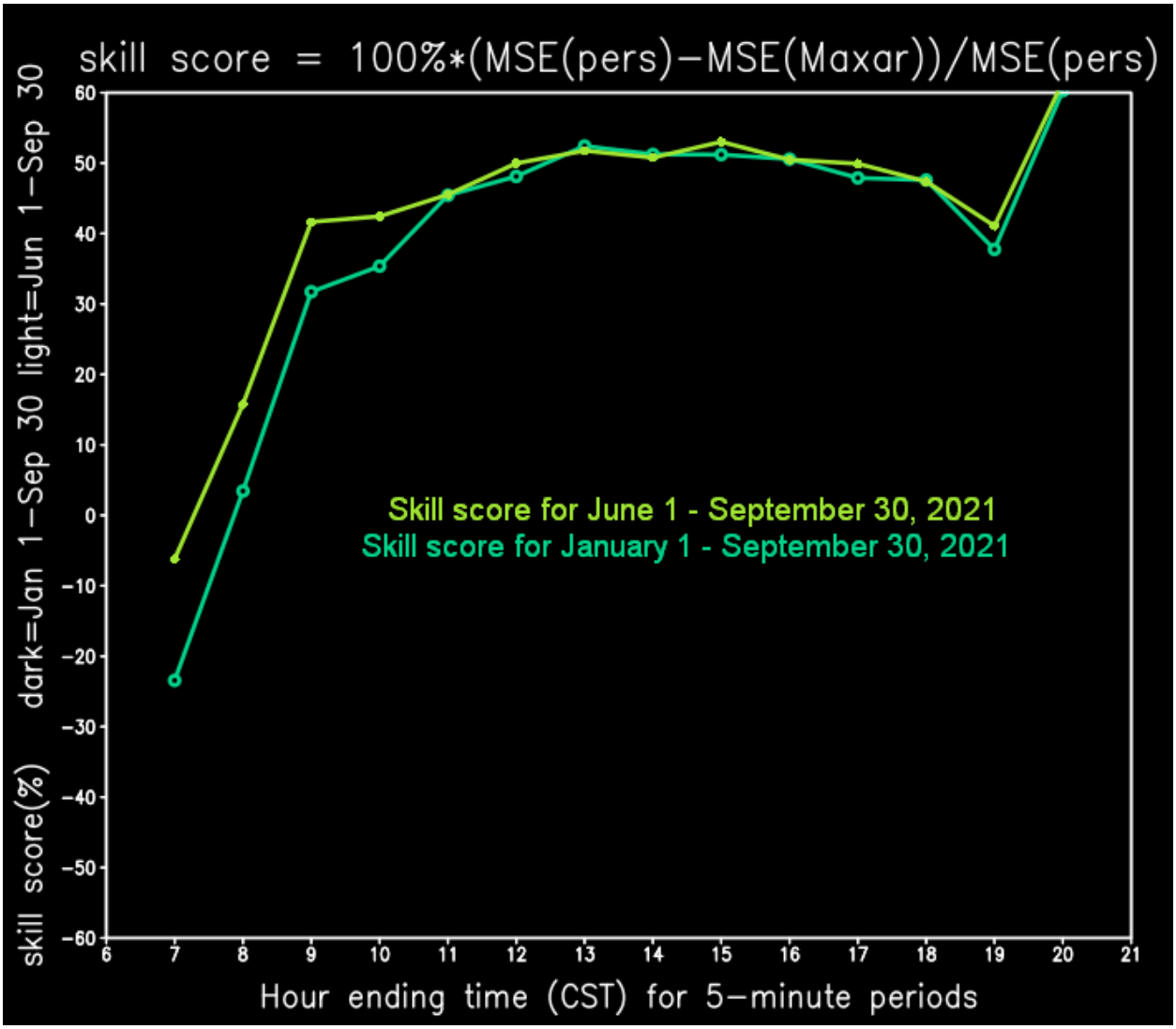


Figure 32. Skill score as a function of the time of day for the Maxar 5-minute ramp forecasts with the 10-minute lead time. The reference forecast is the persistence of the observed ramp rate for the 5 minutes ending at the forecast creation time.

Probabilistic Forecast Validation

The three sets of probabilistic forecasts were evaluated to compare performance. To the extent possible, they were also compared to the operational forecast. This was done by making the operational forecast into a probability forecast that has a cumulative distribution function that is zero for power up to the predicted value and 1 (100%) for power greater than or equal to the predicted value.

Forecasts for the 5-minute averages were evaluated against each other and validated against the 5-minute averaged observations, and, likewise, forecasts for the hourly averages were evaluated against each other and validated against the hourly averaged observations. The SUMMER-GO-1 delivered forecasts only had 5-minute averages, even out to 48 hours, so it could not be compared among the hourly averages. The forecast samples contained all initial and lead times for which all the forecast sets being compared had a forecast, matching sets of forecasts, and for

which daytime observations are available. The forecast samples for various evaluations were then stratified by lead time and time of day. SUMMER-GO-3 forecasts that use the BMA algorithm for assigning probabilities were found to perform poorly for the 5-minute short-range forecasts due to the insufficient skill-based weighting of the more skillful smart persistence forecasts and therefore were not included in the comparisons for the 5-minute forecasts.

The forecast and observed power were first scaled by AC capacity so that all evaluations were performed for power as a fraction of capacity. In this manner, results could be easily interpreted and readily compared for all plants and the aggregate. The observations were quality-controlled using the first and second procedures detailed in the first year annual report, including removing stuck logger events.

To avoid the complications with different experimental forecasts having various problems that were fixed in others at different times, such as with outage handling, the forecasts were all cleaned, and for the 5-minute forecasts that use smart persistence among the superensemble collection, validation times when the smart persistence assumed clear skies instead of using observed values were omitted. The validation period for all statistics shown was January 1, 2021–September 30, 2021, which is after the instantaneous GHI from the RAP and SREF started being handled properly (so the evaluated distributions were not affected by the member forecasts being mistakenly much too low toward the end of the day).

The probabilistic forecasts were evaluated using the CRPS, which is a summary metric that rewards forecasts that have a narrow distribution of probability density as a function of power instead of a wide distribution, and the peak in the probability density is accurately placed. Also, statistical reliability was evaluated, which measures how well the probabilities match against the frequency of an event. For example, a 20th percentile forecast should exceed the observations 20% of the time. The reliability plot shows how often it actually exceeded the observations. Finally, sharpness was evaluated, which measures how narrow or wide the probability distribution is, i.e., it measures how much certainty is indicated by the probabilistic forecast.

Continuous Ranked Probability Score

The CRPS is computed as:

$$\text{CRPS} = \int_0^1 \left(prob(forecast) - prob(observation)\right)^2 dP$$

where *P* is the power as a fraction of AC capacity, the integral is bounded from zero power to 1 (power = AC capacity), and the probabilities are cumulative probability density functions, which means that for each P, it is the probability that the value is smaller than P. Then, prob(observation) = 0 for P < observed value and 1 for P >= observed value. The CRPS is then averaged over the whole forecast, shown as observation pairs in the evaluation sample.

The CRPS of a perfect forecast is zero. Because P goes from 0 to 1 and the probability goes from 0 to 1, the worst forecast has the probability distribution of the forecast minus the observed equal to 1 for all values of P, and then the worst forecast has a CRPS of 1; thus, the range of possibilities is from 0 (best) to 1 (worst). A deterministic forecast is a probabilistic forecast with a cumulative distribution function (CDF) of zero for P < forecast and 1 for P >= forecast, which

is the same as the absolute error, so the average CRPS for a set of deterministic forecasts is the MAE. Because the prob(observation) is 0 or 1 for each value of P, the integral is minimized, and score is best (lowest) when the CDF is small for P < observed and high for P > observed, which means the probability distribution is narrow. Wide probability distributions are penalized. CRPS rewards probabilistic forecasts that are both accurate and sharp. Overall metric scores should be similar to MAE scores, so if 0.1 (e.g., 10% of AC capacity) is a typical MAE, a comparable probabilistic forecast will have a CRPS of approximately 0.1.

CRPS scores as a function of lead time are shown in figures 33–36 for the aggregate and various PV generating units selected for illustrating key points. The left side shows the 5-minute intervals for lead times from 5 to 120 minutes. The right side shows the hourly averages for lead times from 3 to 48 hours (hours 1 and 2 were made only for 5-minute intervals).

The top panel in Figure 33 shows the results for the aggregate. There was a problem with the aggregate in the BMA forecasts (SUMMER-GO-3) that did not appear for the individual units and was not resolved during the project, which explains why the CRPS score is so much worse than for the others. The infinitely sharp operational deterministic forecast converted into a CDF was also accurate and scored by far the best for all lead hours, although part of the problem for the SUMMER-GO-2 (large ensemble) experimental forecasts was that in some cases some new power plants were started before they were supposed to be, in some cases for times when no outage yet existed but the site was not yet running, and in other cases they were not started when they were supposed to be. I do not yet have an explanation for the oscillatory behavior of the CRPS for the 5-minute forecasts. The forecasts for the aggregate generally scored better (lower) for all lead times and forecasts than for the individual units for the same lead times and forecasts (aside from the BMA forecasts).

The bottom panel in Figure 33 shows results for a fixed-tilt PV plant in sunny west Texas. Typically, the operational system performs better for that unit than for most other units, and that did happen in this evaluation. I have no explanation for the very high (bad) score for the hourly SUMMER-GO-2 forecasts.

Figure 34 shows results for a unit that was running for long periods near 10% to 20% below the expected generation given the irradiance. We recently learned that the unit was not fully available but did not report its derates. The errors without controlling for these unannounced derates were large. The BMA algorithm self-scaled to the recent data, keeping up with the bias caused by these unannounced averages, yielding much better error scores.

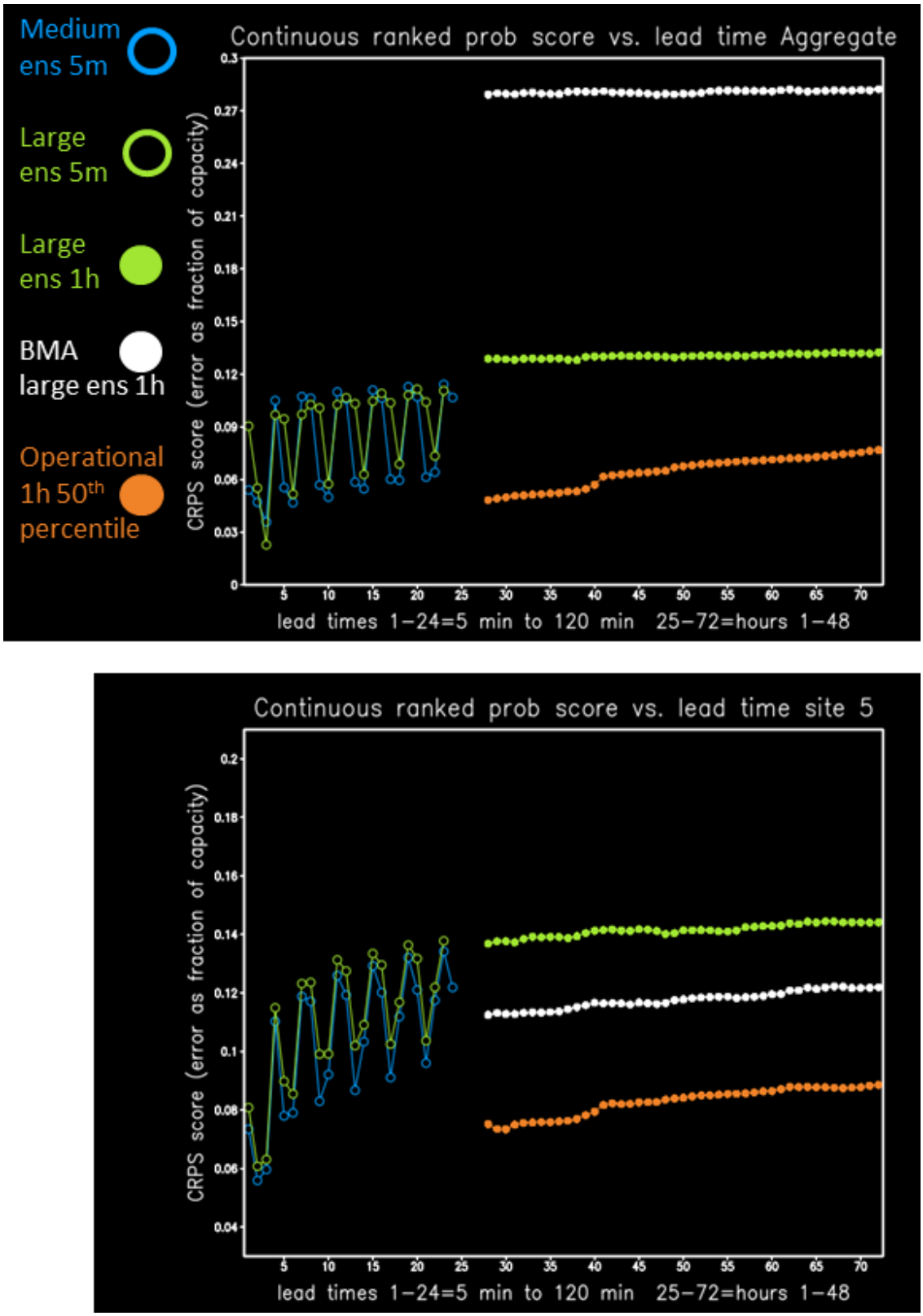


Figure 33. CRPS versus lead time for the 5-minute interval forecasts for the leads from 5–120 minutes (left) and the hourly averages for the leads from 3–48 hours (right) for the aggregate (top) and a west Texas fixed-tilt PV plant (bottom)

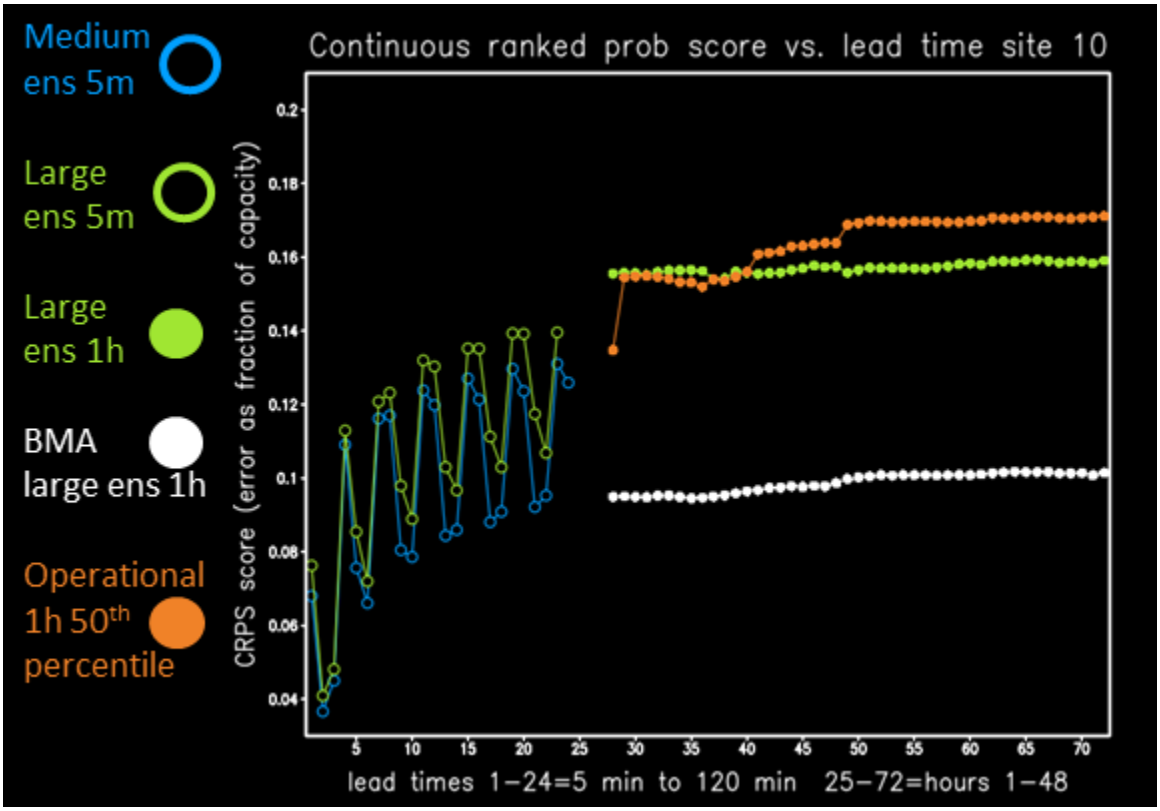


Figure 34. As in Figure 33 but for a dual-axis site with unreported derates introducing forecast bias. The BMA forecast successfully corrects for the bias, resulting in a much lower error than the other forecasts, which simply assume that the reported unit status is correct.

The effect of weather regime differences is shown in Figure 35. The upper panel shows scores for a sunny location in west Texas, whereas the bottom is for a site along the Gulf Coast where clouds are much more abundant and highly variable from minute to minute as well as day to day. Both sites use single-axis trackers. The 5-minute forecasts struggled with the high-frequency, high-amplitude variability in the cumulus regime at the coastal location because optically thick, small clouds cover and uncover the sun. The BMA forecasts struggled with the day-to-day variability at the coastal location while performing very well for the more consistent west Texas location. The poor performance of the SUMMER-GO-2 forecast at this location is disappointing.

Scores averaged among established sites are shown in the top panel of Figure 36, and scores averaged over newer startups are shown in the bottom panel. New starting-up units typically go through a period of a week to sometimes several months when the announced and actual derates differ widely. Sometimes derates are unannounced or the unit is not producing even near the derate value even under full sun, and this behavior can fluctuate within an hour or at different times of the same day as engineering tests are performed while the unit is delivering some power to the grid. It is not surprising that error scores show up as higher (worse) at these startup sites on average than at the more well established sites and that the BMA algorithm does poorly due to the inconsistent bias created by the engineering tests. For the established units, the BMA forecasts outperformed SUMMER-GO-2 for the hourly time intervals at all lead times, whereas the operational system still outperformed the experimental systems.

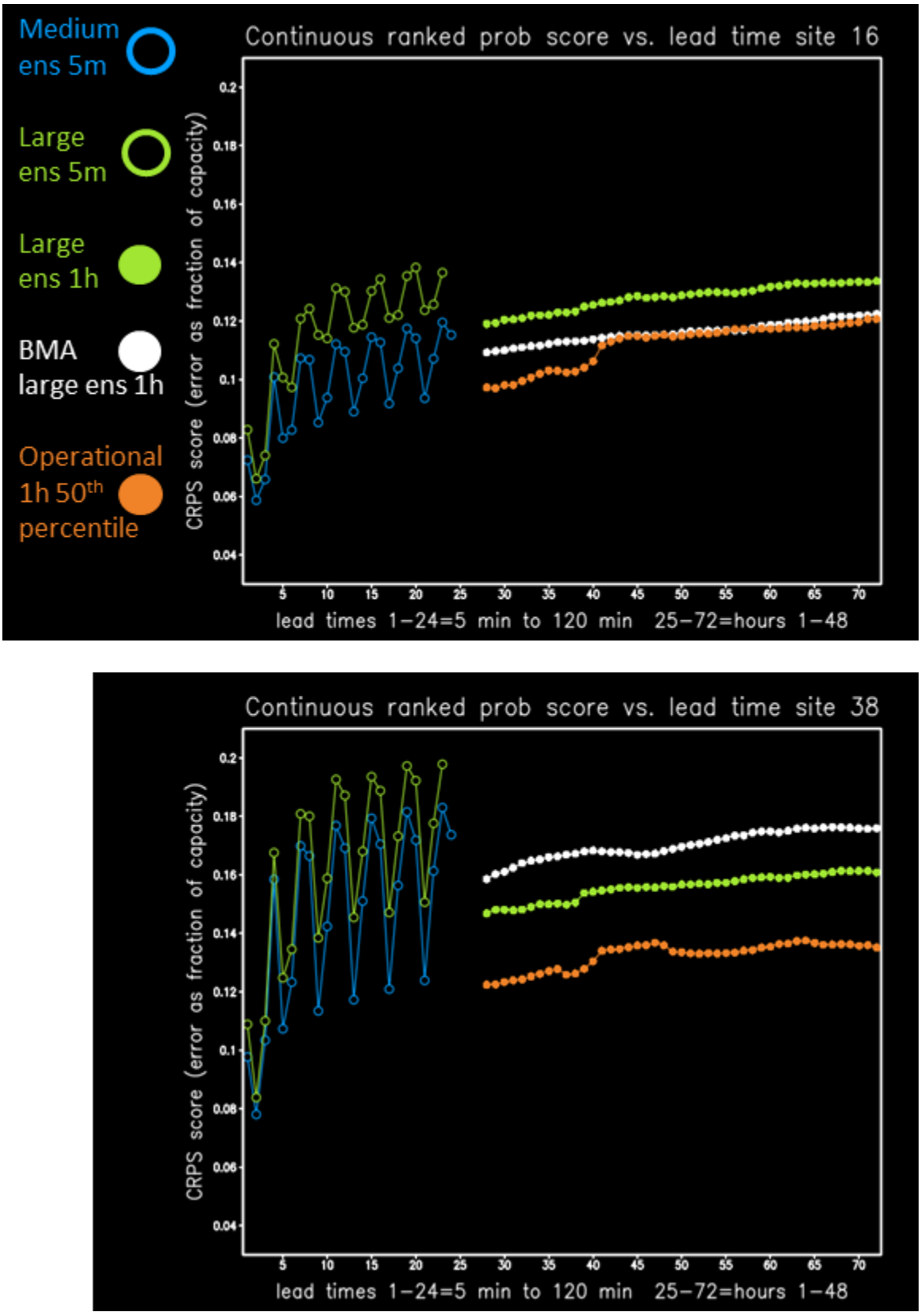


Figure 35. As in Figure 33 but for a single-axis site in sunny west Texas (top) and a single-axis site in east Texas with abundant, highly variable clouds (bottom)

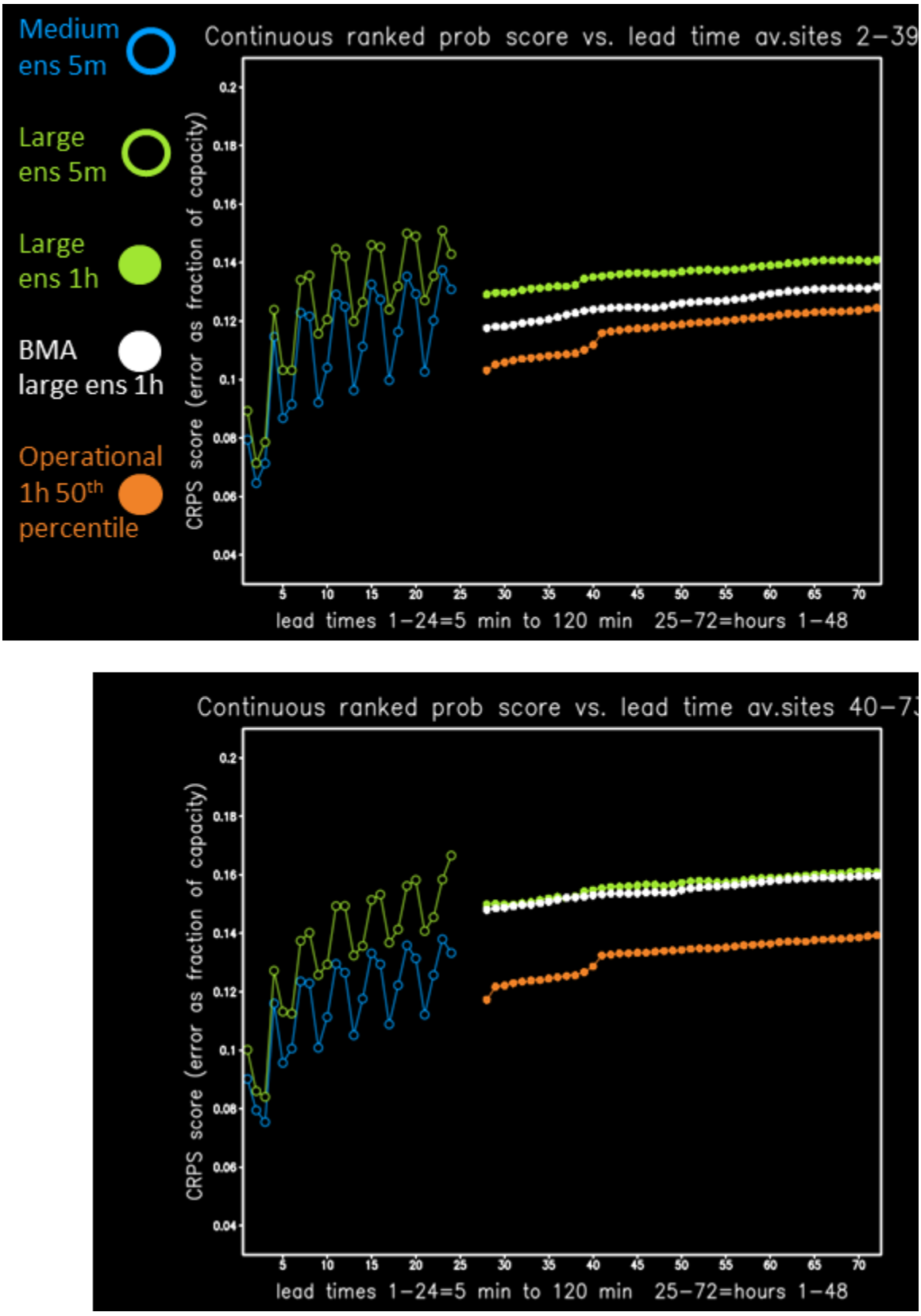


Figure 36. As in Figure 33 but averaged over 38 well-established generating units (top) and averaged over 34 more recently started units (bottom)

All the forecasts show rather flat increases of error with lead time, so forecasts were combined for all lead times and stratified by time of day to examine the diurnal shape of the error. The next set of plots shows these results for CRPS as a function of the time of day for all 5-minute interval lead times out to 120 minutes and that are valid any time within the hour and separately for all hourly lead times out to 48 hours. The same locations are shown in the same order.

Overall, the CRPS scores are highest (worst) near the peaks in the morning toward the end of the morning up ramp and in the afternoon near the start of the evening down ramp and are not as high in the middle of the day. Most experimental forecasts at most sites scored better (lower) than the operational deterministic forecast during the middle of the day but were much worse (higher) at both ends of the day. This requires some investigation. It is encouraging that the experimental systems performed so well for most of the day. If the problems in the mornings and evenings could be identified and fixed, the experimental systems skill as measured by the CRPS would surpass that of the single 50th percentile from the operational system.

Scores for the aggregate (Figure 37, top panel) are relatively flatter across the day than scores for many of the individual plants. The 5-minute interval scores are better for the SUMMER-GO-1 medium ensemble than for the SUMMER-GO-2 large ensemble.

Scores for the fixed-tilt system (Figure 37, bottom panel) have a single hump instead of being broad with a double hump, as for the operational forecast. This is expected for a fixed-tilt array. It is surprising that the experimental forecasts instead have a high CRPS at the start and end of the day for the fixed-tilt array.

For the site with persistent unannounced derates (Figure 38), the BMA scores look good and about how they should look for any site, whereas others are all thrown off by the unannounced derates.

Comparing the single-axis tracking array in sunny west Texas (Figure 39, top panel) to the single-axis tracking array in the variably cloudy regime (Figure 39, bottom panel), the shape of the scores were similar, just shifted to a higher error at the more variably cloudy location, which has a more difficult forecast. In both, the hourly scores for SUMMER-GO-2 outperformed the operational and BMA systems through most of the day but not in the morning and evening. The same holds for the average of the established sites and the average of the newer sites (Figure 40). Both have a similar shape, but the error is a bit higher for the newer sites, likely due to startup engineering tests causing inconsistent irradiance-power relationships due to incomplete derate information, and the SUMMER-GO-2 outperformed the other hourly averaged forecasts for most of the day but not at the start and end of the day.

The daily variations in the CRPS, which are caused by daily variations in the cloud regime and weather predictability, are shown in figures 41–43 for all the same locations in the same order, except the average over the newer locations is not shown. There is a gap for May 3–19 when the SUMMER-GO-2 large ensemble system was undergoing some code changes and was not running. The other forecasts were running, but the scores were only calculated and plotted when all the forecasts were available for fair comparisons. After May 19, the SUMMER-GO-2 hourly averaged forecasts showed a much higher error in the aggregate (Figure 41, top panel) because apparently those changes introduced some problems with the outage handling. Its CRPS was also

higher after the gap at the fixed-tilt site (Figure 41, bottom panel); the reason is unknown. Similarly, after the gap, the SUMMER-GO-2 5-minute forecast was higher at the site with the unannounced long-term partial derates, whereas it was the best forecast before the code changes.

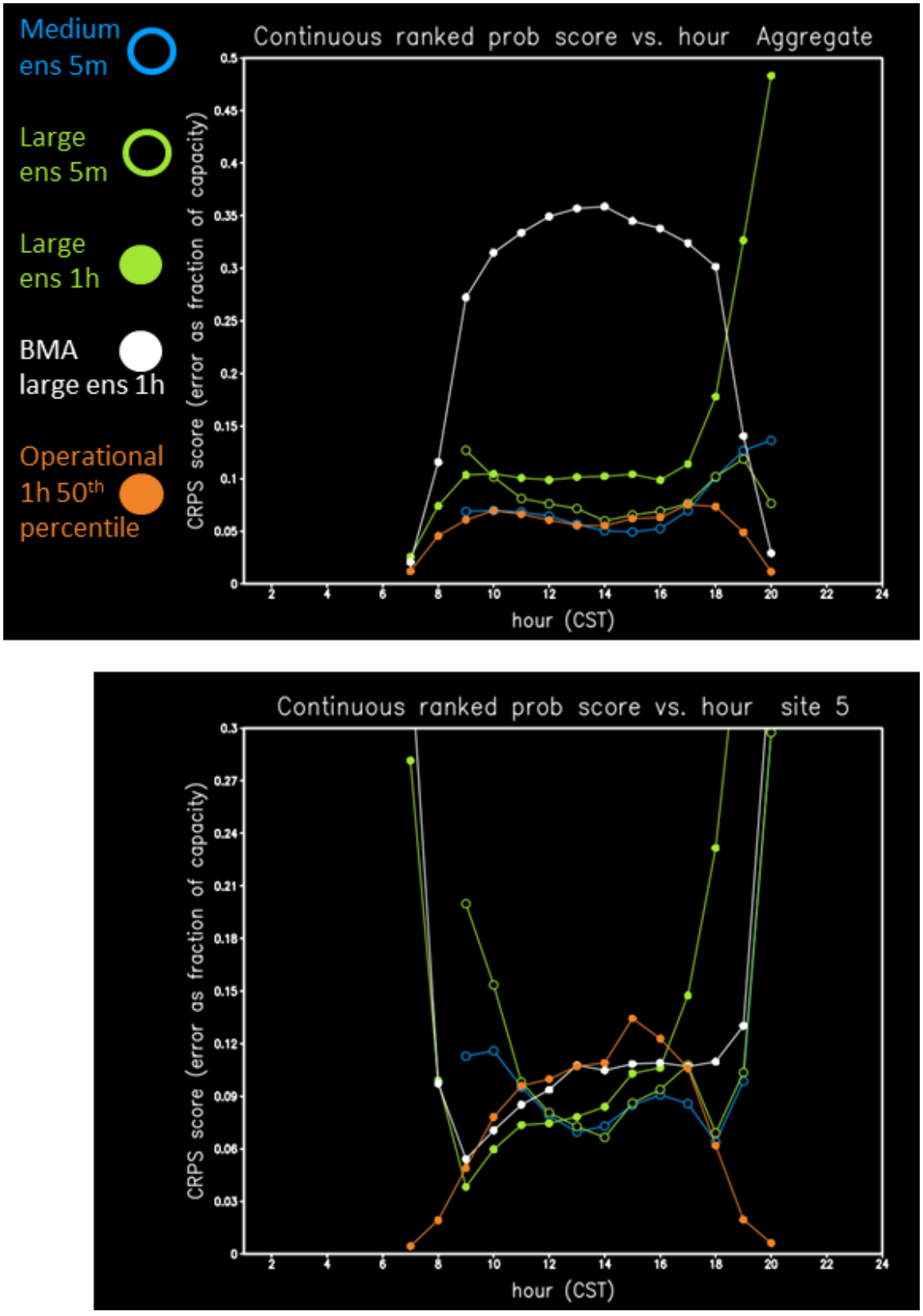


Figure 37. CRPS versus hour of the day for the 5-minute interval forecasts for the leads from 5–120 minutes and the hourly averages for the leads from 3–48 hours for the aggregate (top) and a west Texas fixed tilt PV plant (bottom)

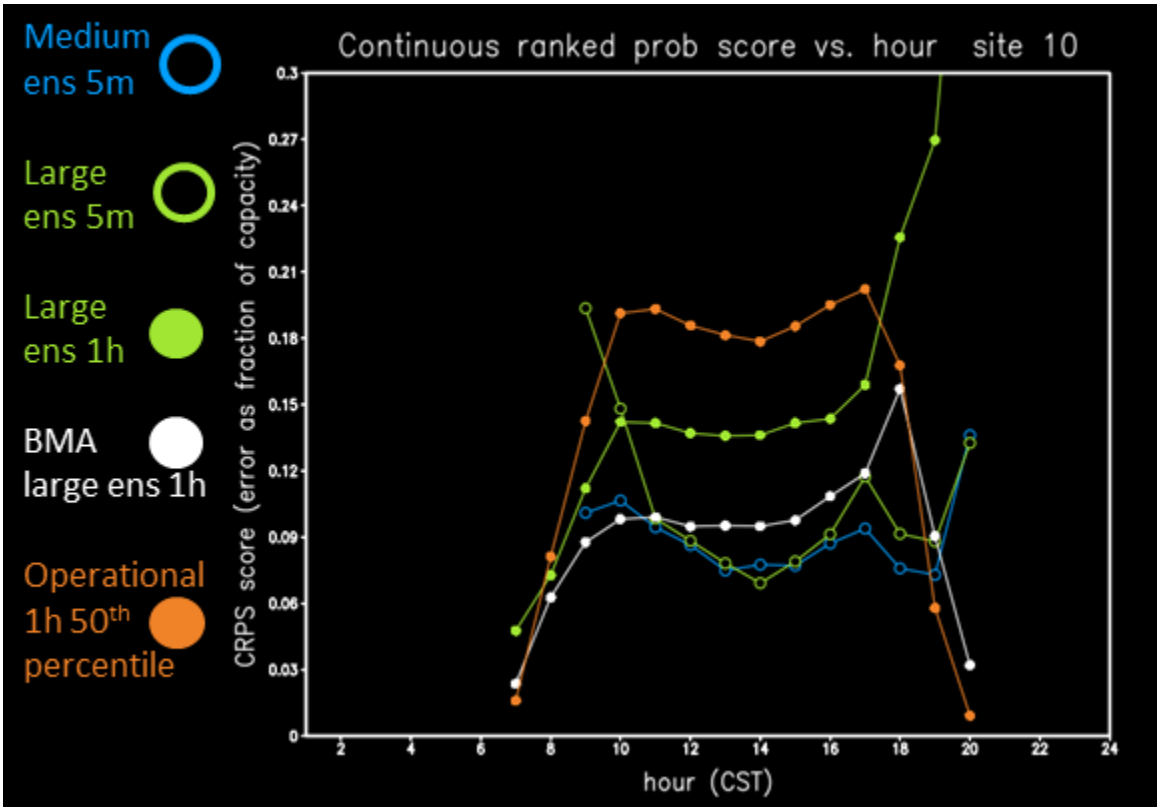


Figure 38. As in Figure 37 but for a dual-axis site with unreported derates introducing forecast bias. The BMA forecast successfully corrects for the bias, resulting in a much lower error than the other forecasts from 3–48 hours, which simply assume the reported unit status is correct.

Likewise, SUMMER-GO-2 5-minute forecasts were the best performers at the sunny west Texas single-axis site (Figure 42, bottom panel) and the average over 38 long-established sites (Figure 43, bottom panel) before the May 3–19 gap, whereas the SUMMER-GO-1 5-minute forecasts were best after that.

The BMA forecasts (SUMMER-GO-3 hourly averages) had many exceedingly high error days at the highly variable coastal location (Figure 43, top panel). It appears this algorithm locks into a bias regime, and when the cloud regime changes, the reference sample it uses from recent days is not representative of the forecast situation, and it misdirects the forecast refinement. This problem occurs more often when large swings from day to day or over several days in cloudiness and associated power generation occur more often.

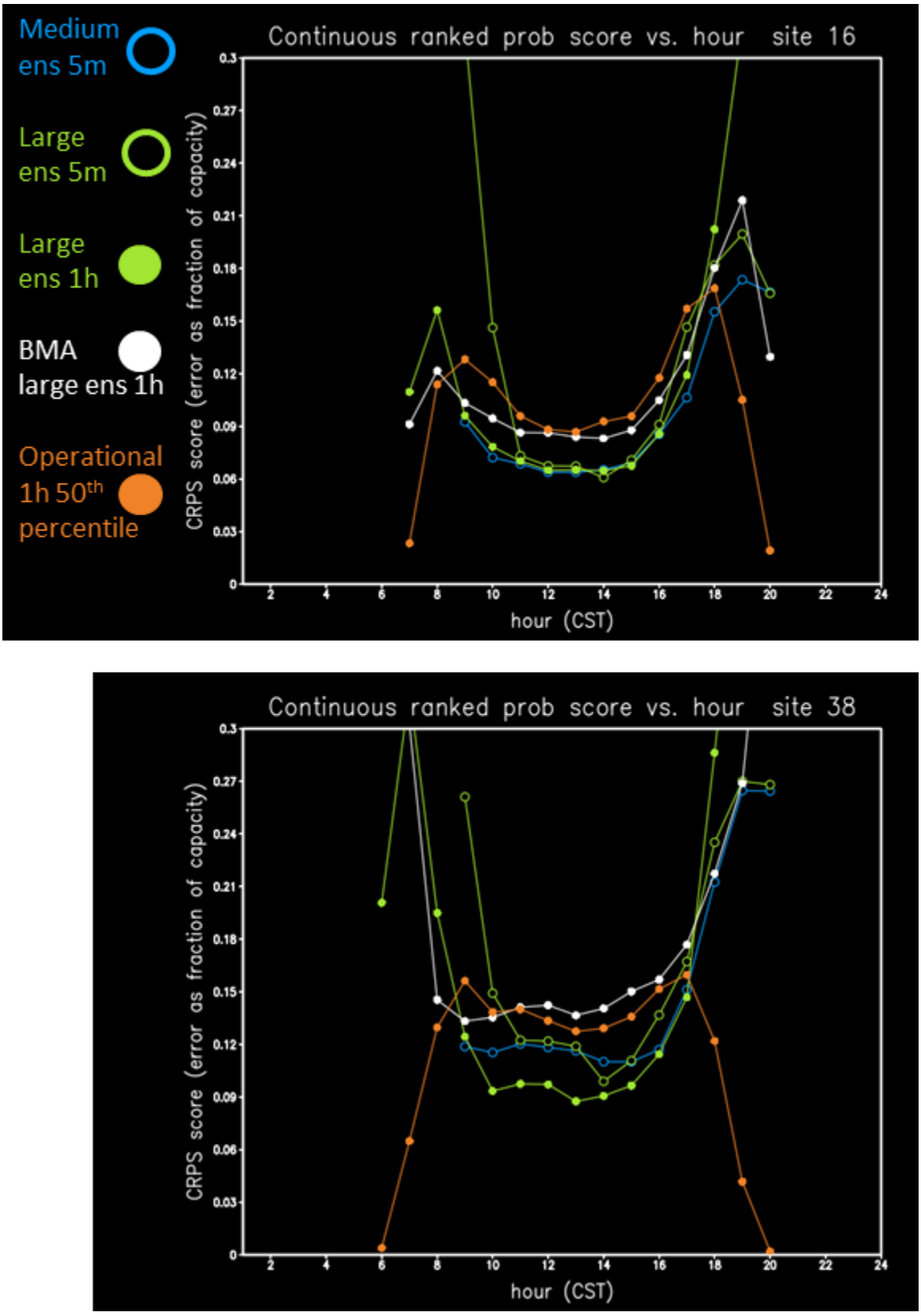


Figure 39. As in Figure 37 but for a single-axis site in sunny west Texas (top) and a single-axis site in east Texas with abundant highly variable clouds (bottom)

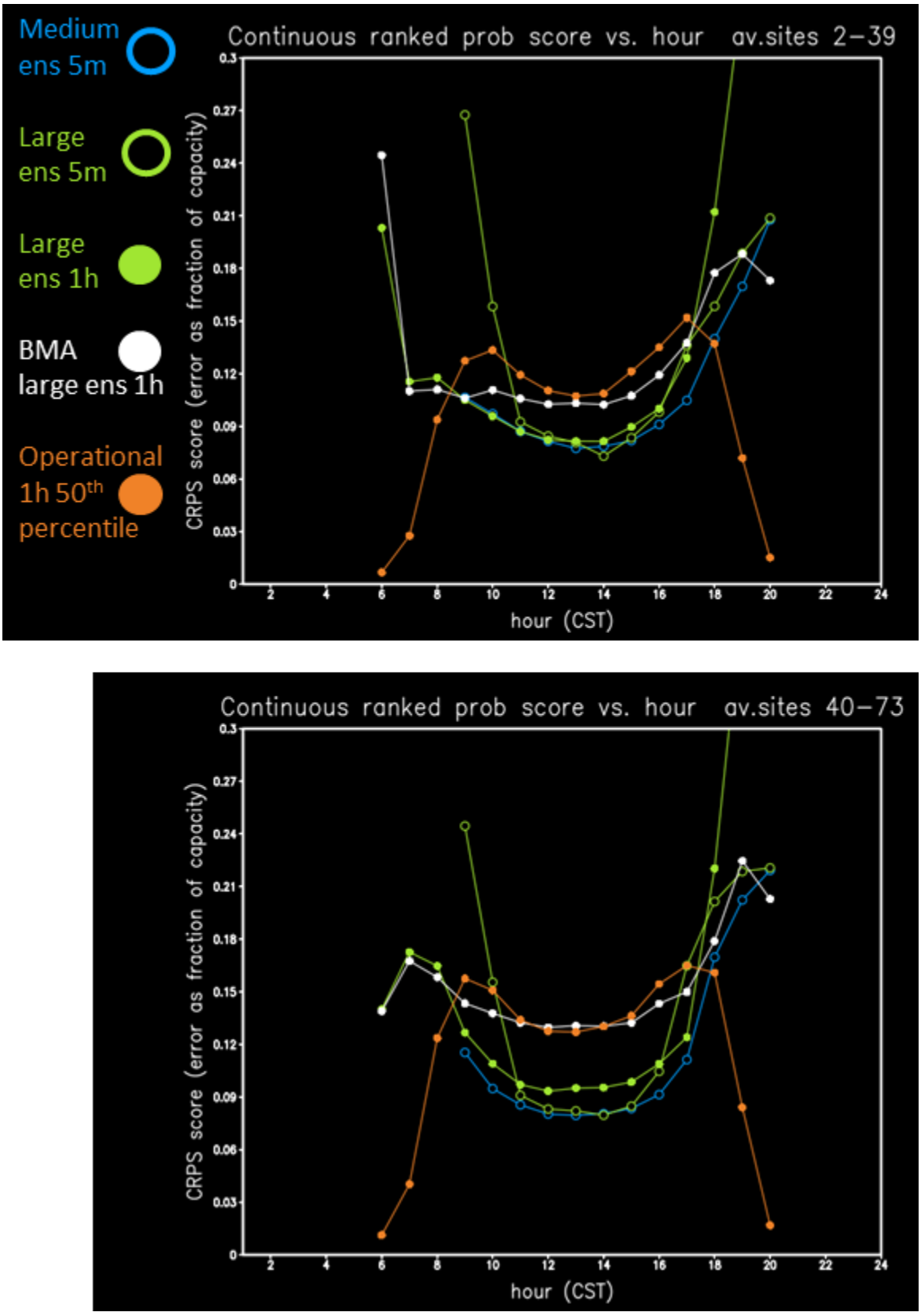


Figure 40. As in Figure 37 but averaged over 38 well-established generating units (top) and averaged over 34 more recently started units (bottom)

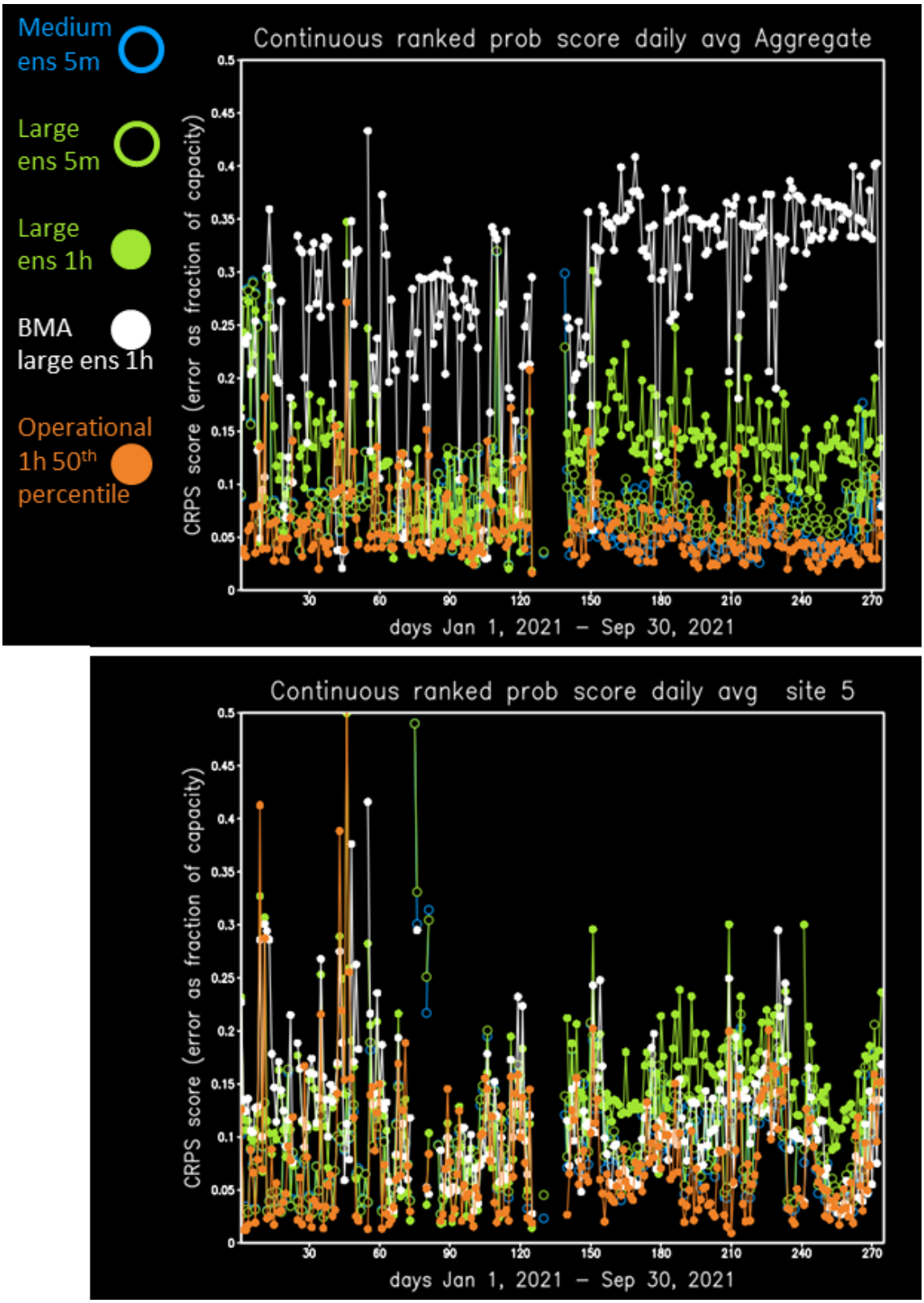


Figure 41. Daily average CRPS for the 5-minute interval forecasts for the leads from 5–120 minutes and the hourly averages for the leads from 3–48 hours for the aggregate (top) and a west Texas fixed-tilt PV plant (bottom)

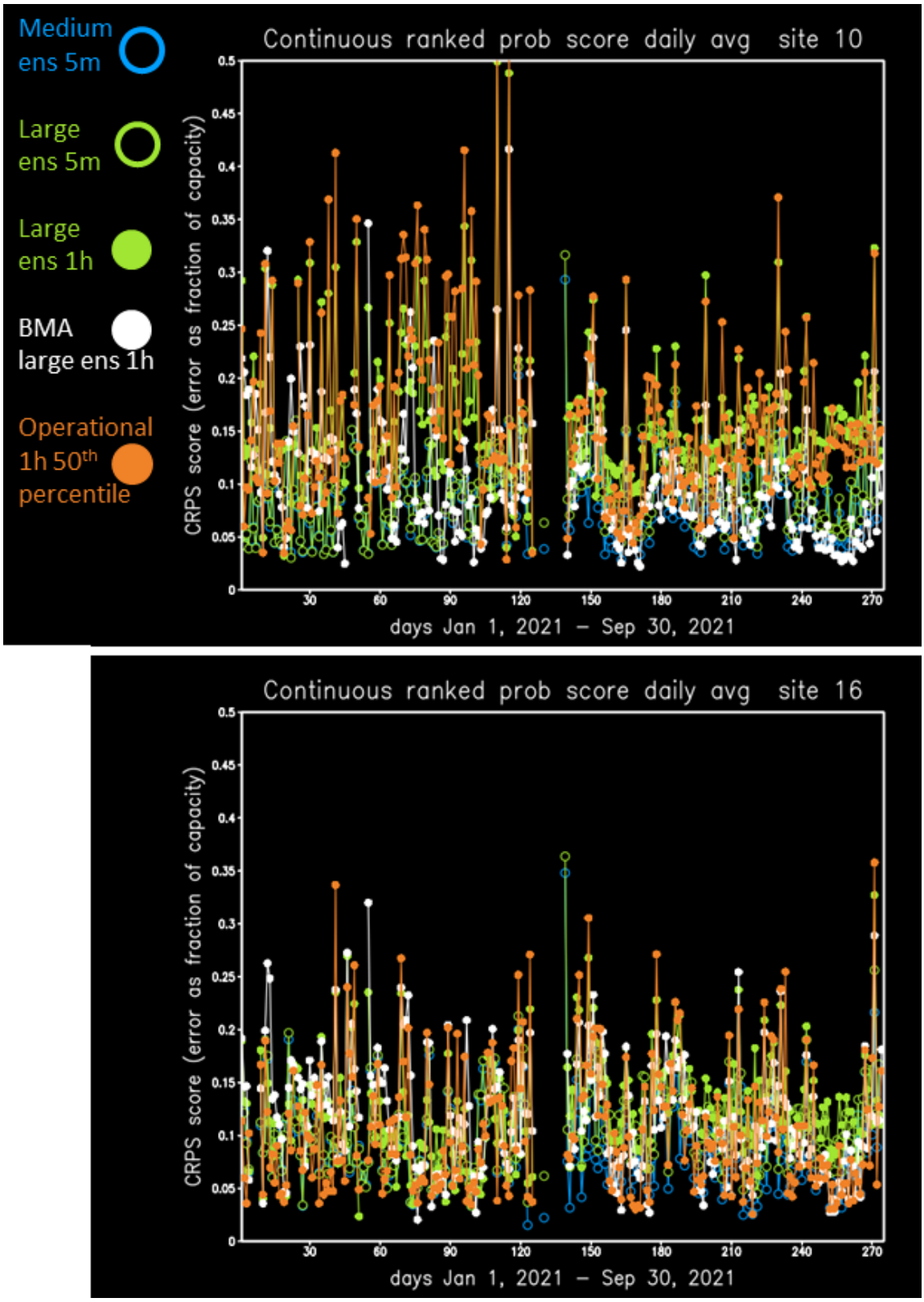


Figure 42. As in Figure 41 but for a dual-axis site with unreported derates introducing forecast bias (top) and for a single-axis site in sunny west Texas (bottom)

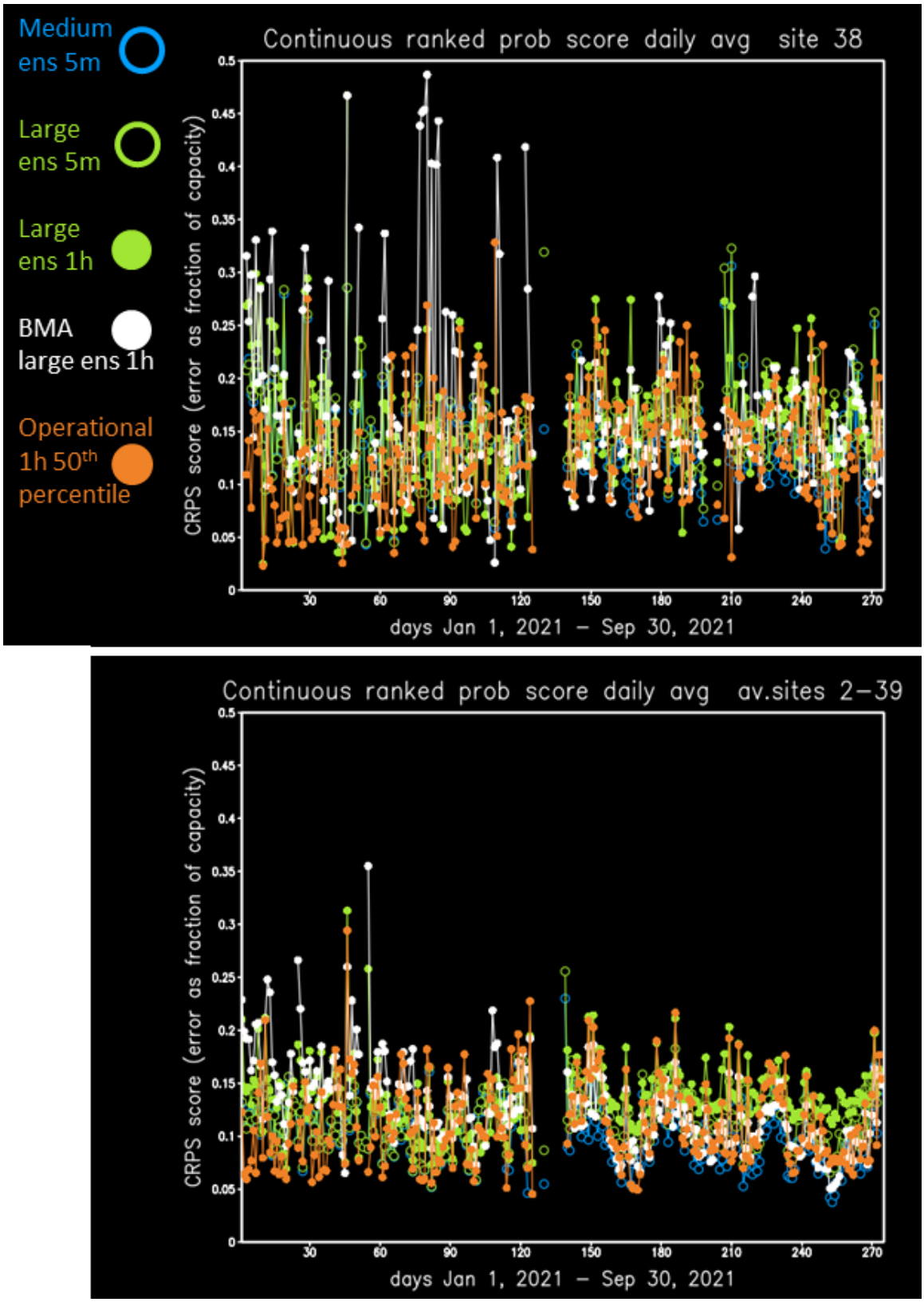


Figure 43. As in Figure 41 but for a single-axis site in east Texas with abundant highly variable clouds (top) and the average score of 38 well-established generating units (bottom)

Reliability

A reliability diagram shows the relationship between the forecast probability of an event and the observed frequency of the event. For power forecasts, the forecast percentile on the abscissa is the forecast percentage of how often the observed power should be less than or equal to this forecast value. For example, the 75th percentile forecast should have observations lower than this forecast value on 75% of occasions. The ordinate is the observed frequency that the power was less than or equal to the forecast value for that percentile. If the observed frequency for the 75th percentile is 90%, it means that the observations were lower 90% of the time, which means that the forecast 75th percentile was often too high. A perfect forecast lines up on the diagonal line where the observed frequency * 100% = forecast percentile. A reliability curve above the diagonal indicates that too many of the observations were less than the predicted power for that percentile, so the forecast values were too high. A reliability curve below the diagonal indicates that the forecast values were too low for that percentile.

An S-shaped reliability curve with too low observed frequency for low percentiles and too high observed frequency for high percentiles indicates that the forecast values at high percentiles were too high and the forecast values at low percentiles were too low, which means that the forecast distribution is too wide, e.g., the forecast shows too much uncertainty. These forecasts poorly distinguish outcomes, and most any outcome will be within the forecast range.

A flat reliability curve with too high observed frequency for low percentiles and too low observed frequency for high percentiles indicates that the forecast values at high percentiles were too low and the forecast values at low percentiles were too high, so the forecast distribution is squashed to a narrow range, e.g., the forecast shows too much certainty. These forecasts clearly distinguish outcomes, but often the actual result falls outside the forecast range.

Even perfect reliability, however, does not make a perfectly useful forecast. Climatology will have perfect reliability because it represents the observed distribution over a large sample.

Reliability curves are shown in figures 44–48 for the forecast valid times in the morning, defined as intervals ending by 10:00 CST; midday, defined as intervals starting at or after 10:00 CST and ending by 15:00 CST; and evening, defined as intervals starting at or after 15:00 CST. The forecast samples for the 5-minute forecasts include all lead times out to 120 minutes (e.g., 5, 10, 15, …, 120 minutes). The forecast samples for the hourly forecasts include all forecast hours out to 48 hours. Reliability is not shown for the deterministic forecasts converted into step function cumulative distributions because it is only a 50th percentile forecast, and the other percentiles are not defined from it.

The reliability diagrams for the aggregate are shown in Figure 44. The low values for the BMA hourly averages were due to a bug that caused very low forecasts only in the aggregate, which did not affect the forecasts of the individual generating units. Conversely, the high arcing curves for the SUMMER-GO-2 hourly averages indicate a high bias, which might have been due to problems with the startup of new sites, although it looks much better for the evening time period. The 5-minute forecasts are more tied to the smart persistence, which projects the most recent several observations forward along a diurnal trajectory, so there should be less overall bias. The evening curves for SUMMER-GO-2 5-minute forecasts look very good, although all the 5-minute forecasts show a bit of an S-shaped profile, indicating that the forecast range is too broad.

This is expected from the bug in defining one of the width parameters for percentiles away from the 50th percentile, as described in Chapter C. This problem affects only the 5-minute forecasts for SUMMER-GO-1 and SUMMER-GO-2.

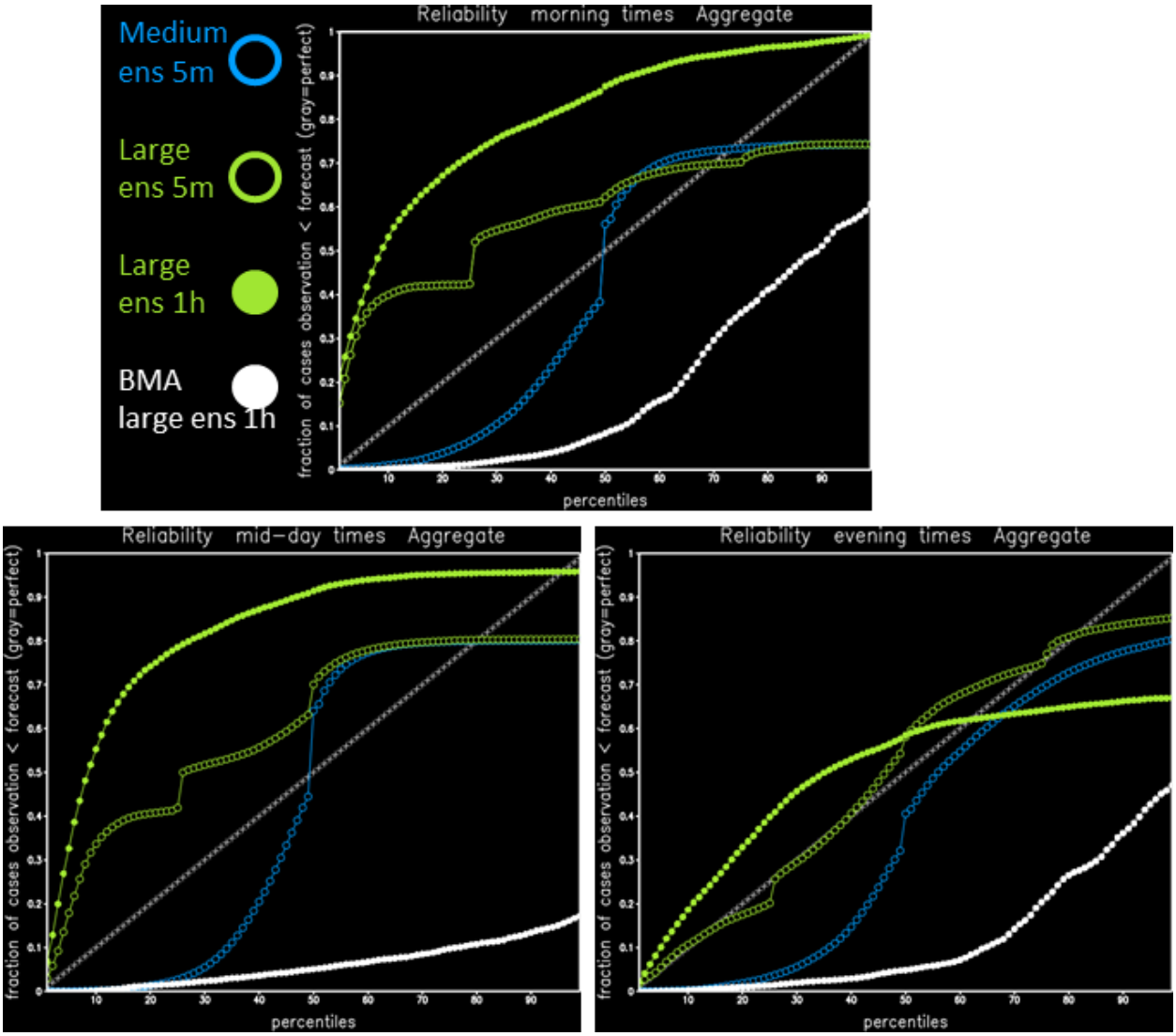


Figure 44. Reliability diagram for probabilistic forecasts of the aggregate at different times of day

For each of the 99 percentiles on the abscissa, the ordinate value of the curve is the fraction of the observations less than the forecast. A perfectly reliable forecast will coincide with the gray diagonal line so that percentiles correspond to 1 frequency of exceedance. Times shown are morning (top, forecast valid intervals ending by 10:00 CST), midday (bottom left, forecast valid intervals starting at or after 10:00 CST and ending by 15:00 CST), and evening (bottom right, forecast intervals starting at or after 15:00 CST).

The reliability diagrams for a single-axis tracking site near Austin are shown in Figure 45. The 5-minute forecasts show a low bias but good slope. The hourly SUMMER-GO-2 forecasts show nearly perfect reliability in the midday times and are very good in the morning! The BMA forecasts also show good reliability, especially in the morning, and they are quite good at other times.

The reliability diagrams for another single-axis tracking plant in a more consistently sunny location in west Texas are shown in Figure 46. Results show nearly perfect reliability in the morning and midday for the BMA forecasts and are very good for all others at those times, whereas forecasts of the middle and high percentiles have a low bias in the evening time period for all forecasts.

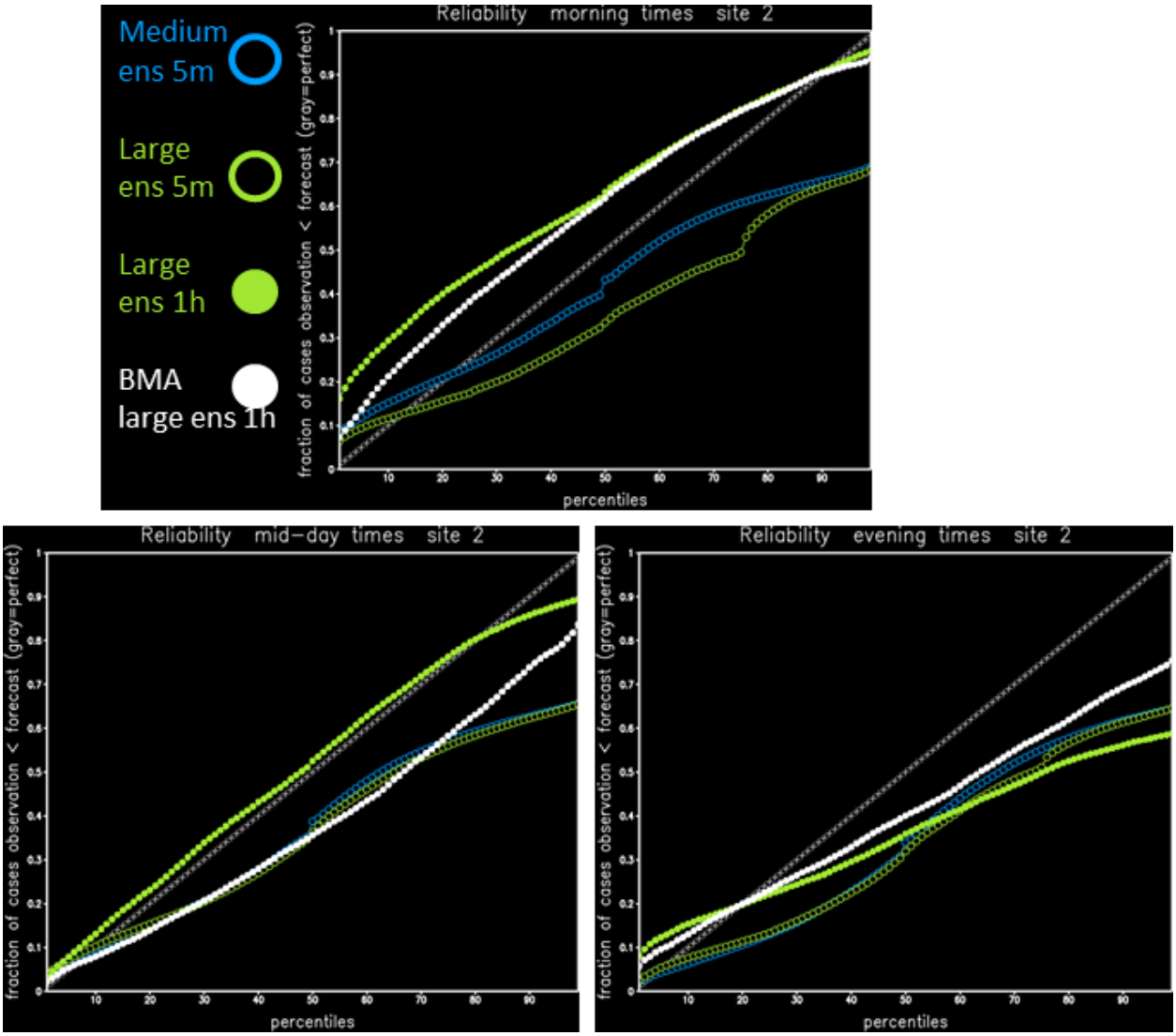


Figure 45. As in Figure 44 but for a single-axis site in central Texas

The average reliability curves for the individual unit forecasts for long-established units are shown in Figure 47. The 5-minute forecasts have the S shape, which is indicative of forecast distributions that are too wide, as discussed for the aggregate. The hourly BMA forecasts have nearly perfect reliability in the morning and are pretty good with a bit of a low bias in the high percentile forecast values at other times of day. The SUMMER-GO-2 hourly forecasts have pretty good reliability in the morning and are nearly perfect in the middle of the day.

The average reliability curves for the individual unit forecasts for new units are shown in Figure 48. The mornings show low bias as though the forecasts are ramping up too slowly, particularly

at the high end of the forecast distributions. This could indicate a problem in the startup procedures setting up a power curve for new locations without history. The power curves are based on data from similarly configured units with sufficient history, which might be using older technologies that might have less optimized tracking in the morning, such as different speeds of trackers, different overnight stow positions the trackers start in at sunrise, and different tracker algorithms. The evening plots show a similar story, possibly with a similar explanation. The midday plots are more similar to those of the well-established sites.

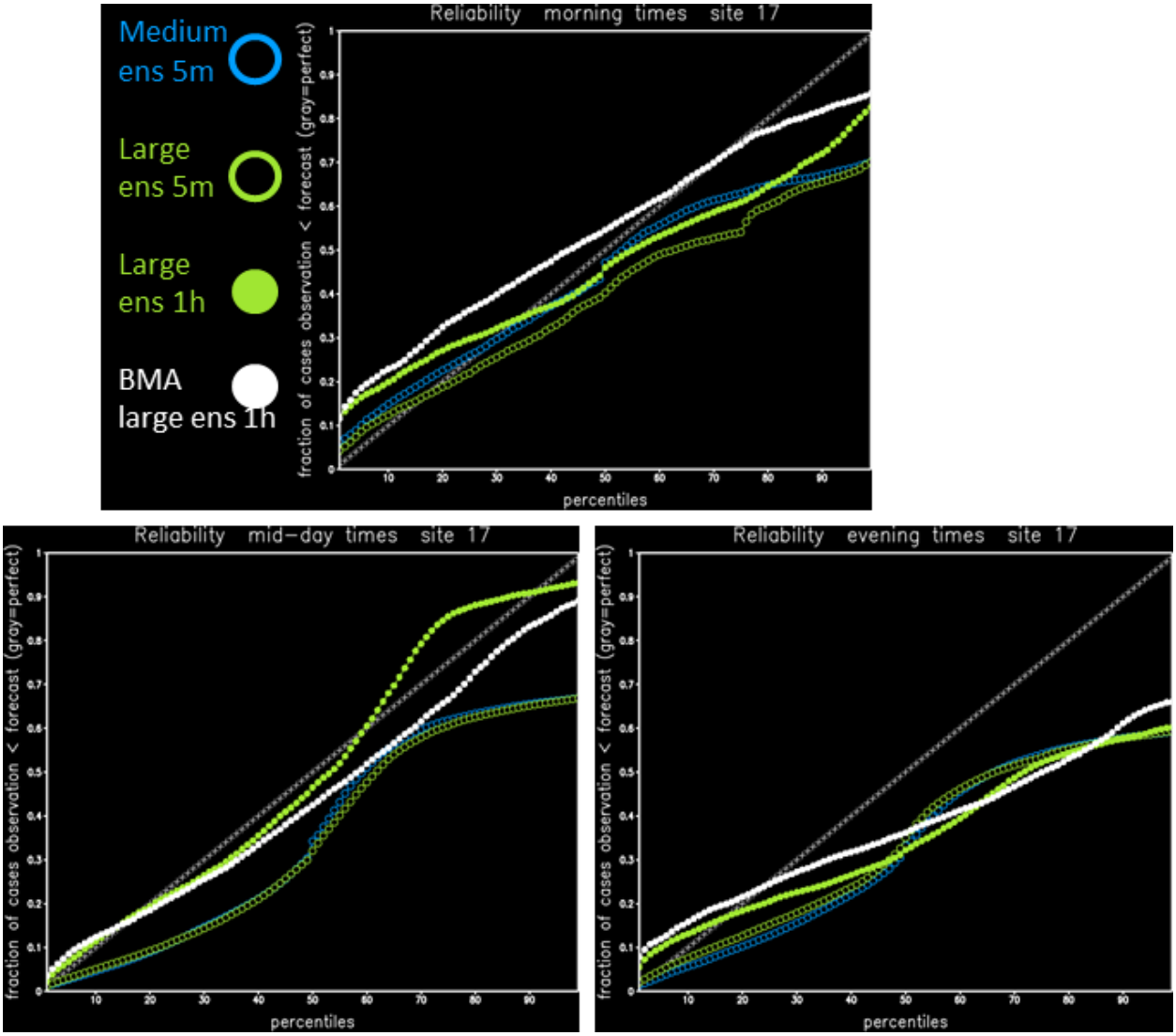


Figure 46. As in Figure 45 but for a single-axis site in west Texas

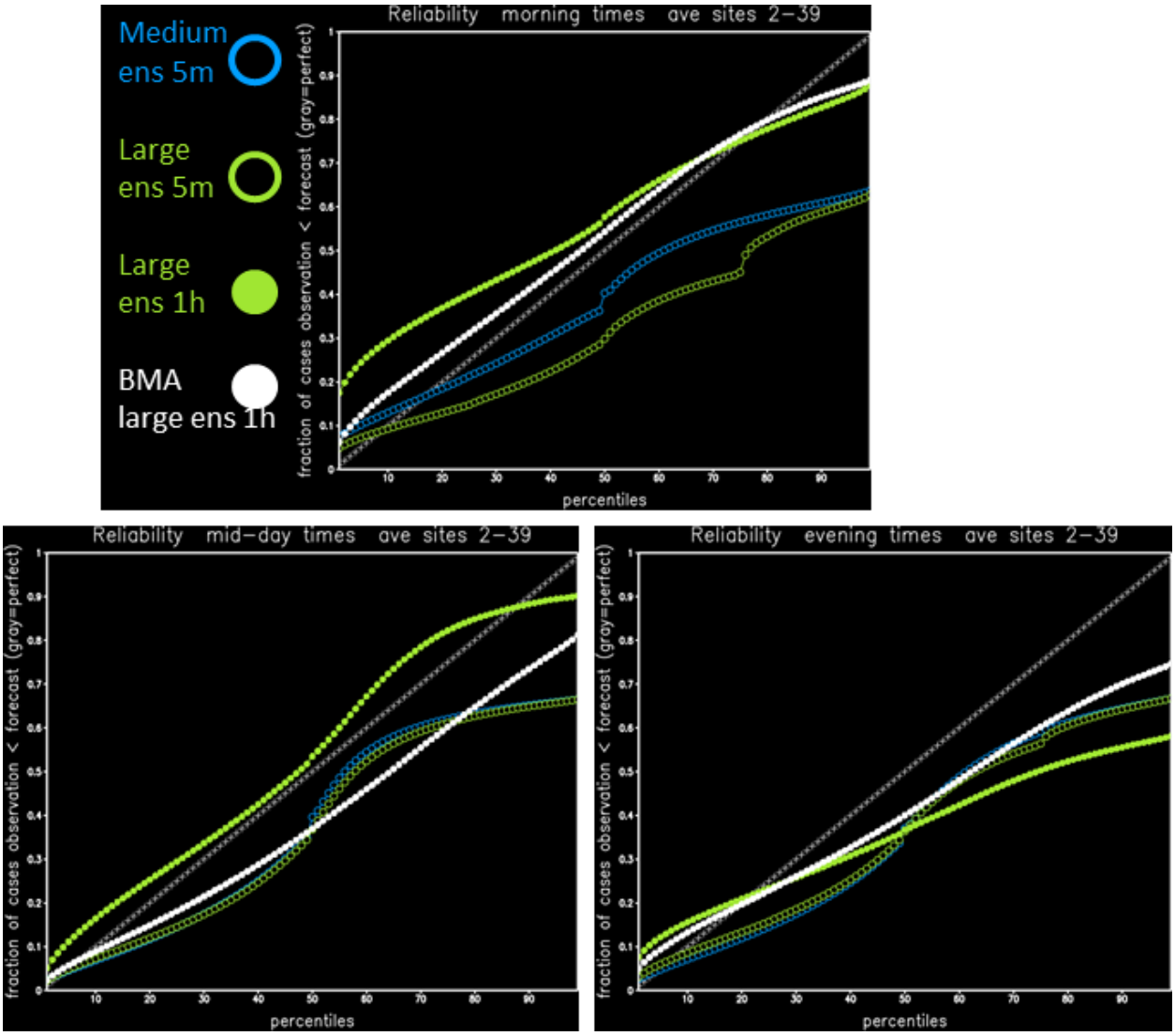


Figure 47. As in Figure 45 but for the average non-exceedance frequency for 38 well-established generating units

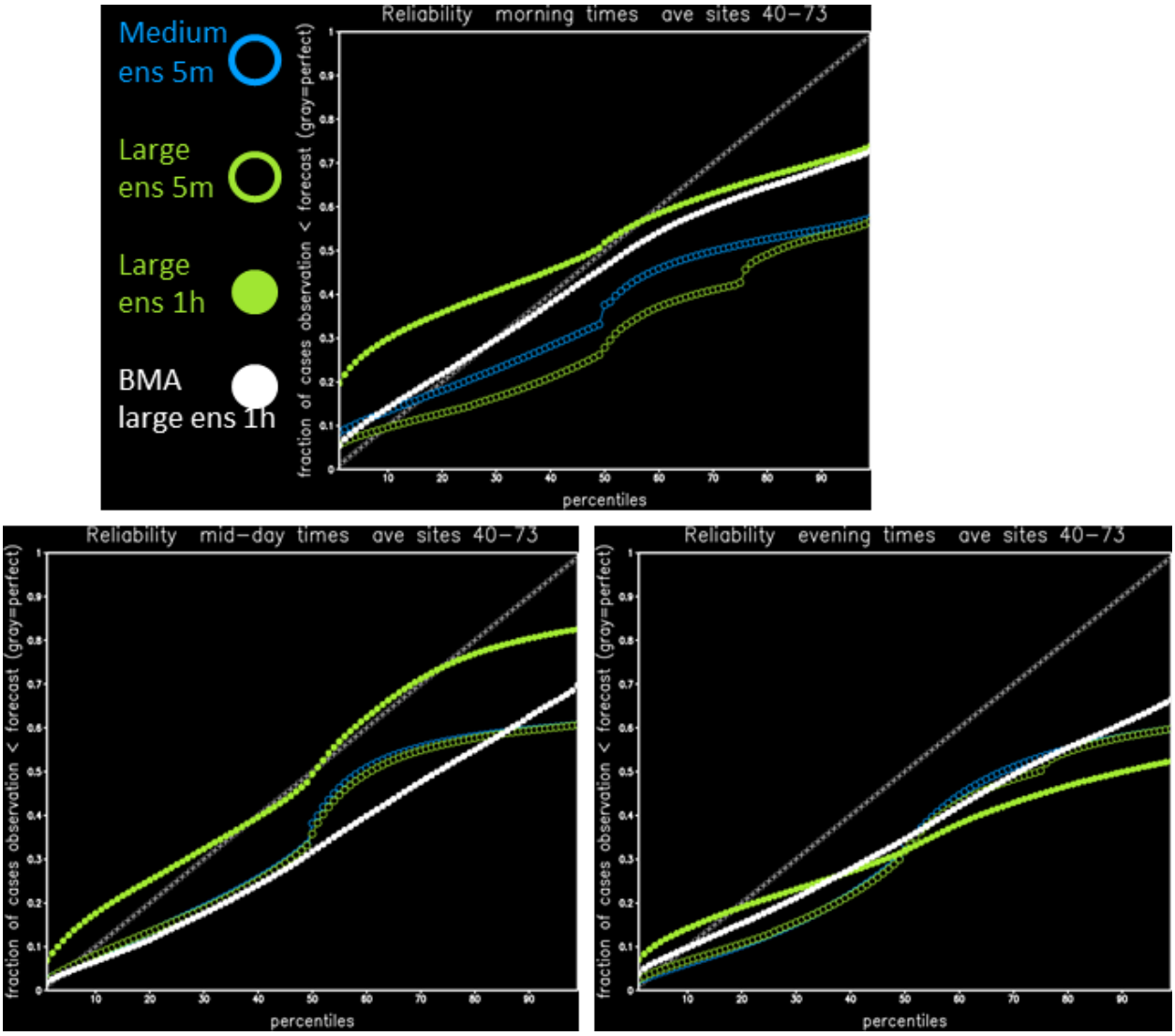


Figure 48. As in Figure 45 but averaged over 34 more recently started units

Sharpness

The sharpness is defined as the difference in the forecast across a range of percentiles. The plots shown use three ranges: 75th–25th percentiles (interquartile range), 90th–10th percentile, and 95th–5th percentile (90% confidence interval width). The higher the range or sharpness, the less sharp the distribution is. A wide range of forecast values across the distribution indicates a high uncertainty in the forecast. A narrow range indicates the forecast has high certainty. Sharpness should not be confused with accuracy. The forecast can show high certainty and the observed condition can be outside this range, e.g., the forecast was not accurate. The CRPS metric combines aspects of sharpness and accuracy to prefer forecasts with a small range—high certainty—but that are accurate. An accurate forecast that has a wide distribution might suffer a poor CRPS, which could be improved by rescaling the distribution with care to preserve reliability.

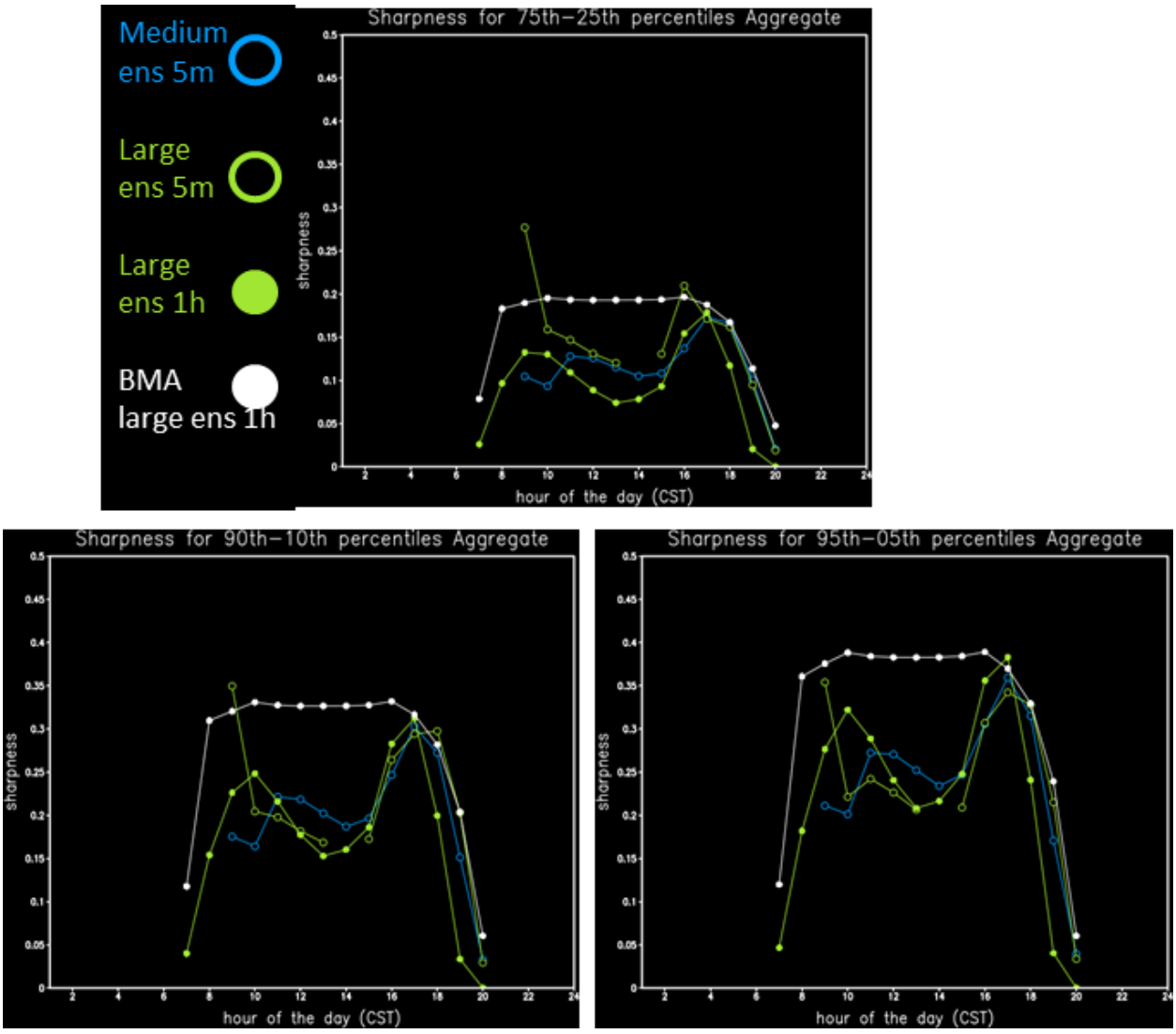


Figure 49. Sharpness as a function of time of day for probabilistic forecasts of the aggregate

Sharpness is shown as the difference between the 75th–25th percentiles (interquartile range, top), 90th–10th percentiles (bottom left), and 95th–5th percentiles (bottom right). Smaller values indicate a sharper probability distribution.

For the sharpness calculation, forecasts were stratified by the hour containing the forecast valid time to obtain profiles of sharpness by time of day. The forecast samples contained all lead times, 5-minute lead times to 120 minutes, and hourly lead times to 48 hours.

Sharpness for the aggregate forecasts is shown in Figure 49. It shows that the 5-minute forecasts have a distribution as wide as the hourly SUMMER-GO-2 forecasts, which is due to using hourly errors in scaling the 5-minute forecast distribution width. The distributions are wider during the shoulders of the day than during the middle of the day, which corresponds to when forecast errors are higher, as discussed in other chapters and with the CRPS discussion above. The interquartile range is, on average, approximately 10% of the aggregate capacity, whereas the 90% confidence interval is approximately 20% of the capacity midday to 30% peak on the shoulders of the day. That means the tail of the forecast distribution is quite wide, showing medium confidence near a central value but a much wider range in the 90% confidence interval.

For 5-minute intervals, interestingly, the larger ensemble has a smaller width of 80% and 90% confidence intervals than the smaller ensemble, indicating more input sources, which might have more extremes and could have more packing toward the middle, giving more confidence in the forecast.

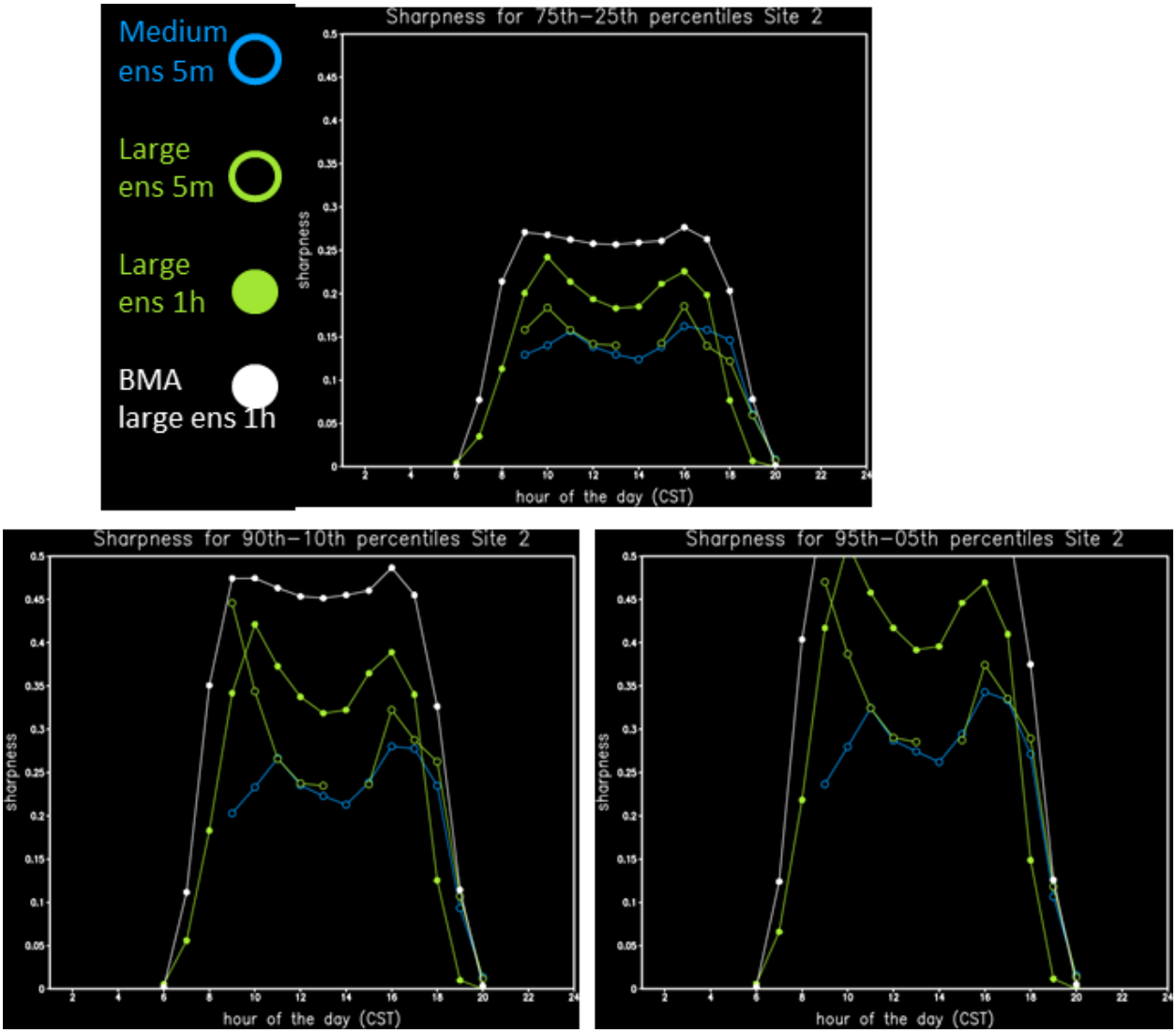


Figure 50. As in Figure 49 but for a single-axis site in central Texas

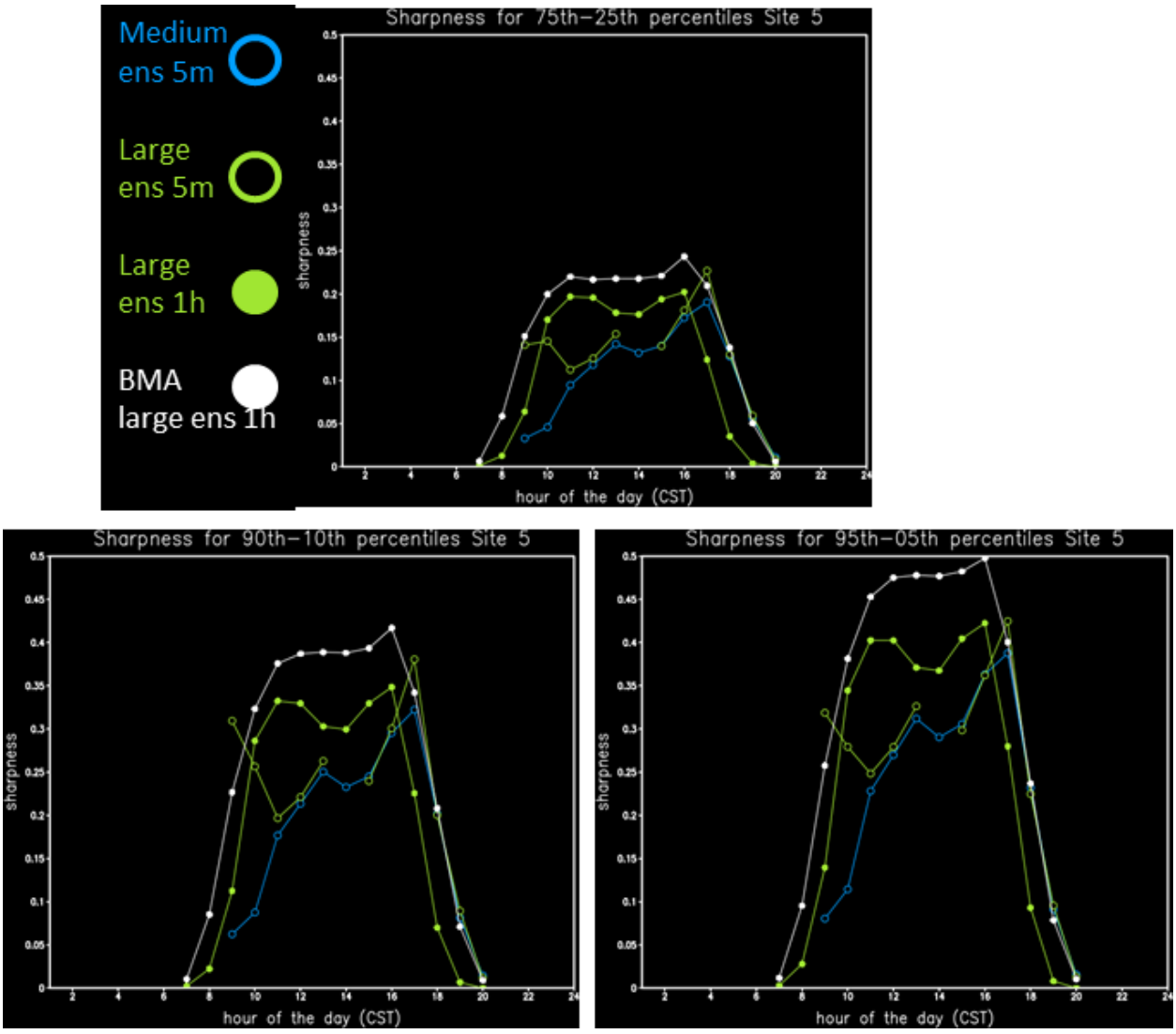


Figure 51. As in Figure 49 but for a fixed-tilt site in west Texas

Comparing the single-axis tracking site near Austin (Figure 50) with the more consistently sunny single-axis tracking site in west Texas (Figure 51), the forecast ranges are wider (less certainty) for all the forecasts for the Austin location. This is a typical illustration of weather dependency. It also happens on a daily basis. On sunny days, the forecast typically shows high certainty in a narrow range of values close to a clear-sky forecast. On days with uncertain weather conditions, such as uncertainty in the timing of a cloud deck clearing during the day or uncertainty due to the expectation of variable cloudiness throughout the day, the range is larger. With more days of higher uncertainty in Austin than in west Texas, the average range is higher in Austin.

The average ranges in the forecasts for the 38 individual units that have been well established are shown in Figure 52. As with the others, the average shows the widest ranges, the least certainty, on the shoulders of the day. And as with the others, the average shows wider ranges for the BMA forecasts, which are likely due to a problem in the skill weighting of contributing forecast members that is not sufficiently selective. And there appears to be some problem with the SUMMER-GO-2 5-minute forecasts in the morning.

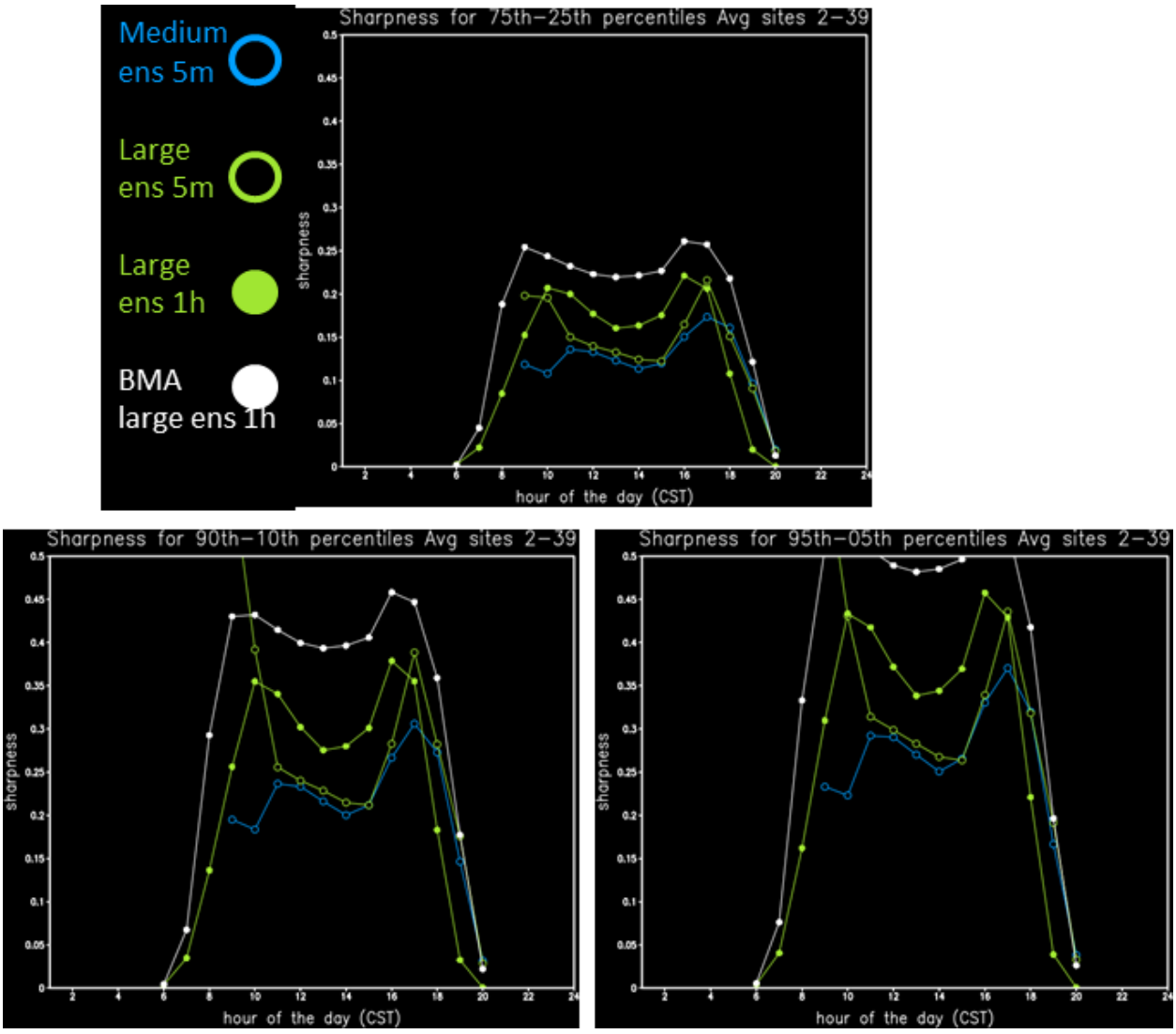


Figure 52. As in Figure 49 but for the average sharpness for 38 well-established generating units

C. Deterministic Forecast Validation

The forecast evaluation was performed in the same manner as described for the probabilistic forecast validation, except that the metrics are MAE and bias of the 50th percentile forecast. Bias is defined as forecast minus observation, so an overforecast has a positive sign on its bias. The hourly average 50th percentile for the SUMMER-GO-1 medium ensemble forecasts was made by summing the 50th percentile forecasts from the 5-minute intervals. This method would not be correct for any other percentiles, so probabilistic forecasts for SUMMER-GO-1 could not be synthesized to hourly averages; thus, the deterministic forecast evaluation compares SUMMER-GO-1 and SUMMER-GO-2 for 5-minute forecasts and compares all four forecasts, operational and all three experiments, for the hourly averages. As with the probabilistic forecasts, lead times of 5, 10, …, 120 minutes are combined in the forecast sample for the forecasts of 5-minute averages, and lead times out to 48 hours are combined in the sample for the forecasts of hourly averages. As with the probabilistic forecast evaluation, only times with matching sets of forecasts are included so that all forecasts have the same weather sample (separately for the 5-

minute and the hourly averages)—if any forecast was missing, that time was not used in the evaluation.

The forecasts are stratified into the same morning, midday, and evening times of day as with the reliability evaluation: Morning is defined as intervals ending by 10:00 CST, midday is defined as intervals starting at or after 10:00 CST and ending by 15:00 CST, and evening is defined as intervals starting at or after 15:00 CST.

Results for the aggregate are shown in Figure 53. MAE and bias are best in the SUMMER-GO-1 5-minute forecasts, which were the delivered forecasts. Among hourly forecasts, the operational forecast was best, but this is affected by the problems in the collection of sites used with the startup of new units in the experimental forecasts. The BMA forecasts, as noted earlier, had some bizarre problem with the aggregate that did not affect the forecasts for the individual units, and therefore it has an extreme low bias of 25%–40% of capacity.

For the easiest forecast, the fixed-tilt site in west Texas, results are shown in Figure 54. MAE is nearly identical for the hourly forecasts from the operational, SUMMER-GO-1, and SUMMER-GO-2 forecasts, whereas the BMA has a larger error; however, BMA had the smallest bias magnitude among the hourly averaged forecasts, whereas the other two experimental forecasts had larger midday high biases than the operational forecast.

Scores for the location with persistent undeclared derates are shown in Figure 55. As noted in the probabilistic forecast evaluation, the BMA method corrects for this, producing by far the best forecasts. The other hourly averaged forecasts have a large positive (overforecast) bias. The 5-minute forecasts that heavily use recent observations also do not have this bias, with the SUMMER-GO-1 5-minute averages forecast having an almost perfect near zero bias and by far the best MAE averaged across the day, although the SUMMER-GO-2 forecast had a slightly better midday MAE.

Comparing forecasts for a single-axis unit in sunny west Texas (Figure 56) to forecasts for a single-axis unit in east Texas, where the climate has more abundant and variable clouds (Figure 57), the MAE is higher for all forecasts at the cloudier, more variable location, and otherwise the story is the similar. The SUMMER-GO-1 and SUMMER-GO-2 hourly averages have lower MAE than the operational forecast in the morning and midday, whereas SUMMER-GO-2 is a little higher in the evening time period. The BMA forecasts correct well for the midday high bias at the sunny location yet have higher MAE, whereas its MAE and bias are poor at the variably cloudy location. The SUMMER-GO-2 5-minute forecasts score almost exactly the same as the SUMMER-GO-1 forecasts in the midday times but have a large magnitude low bias in the morning and evening in both locations.

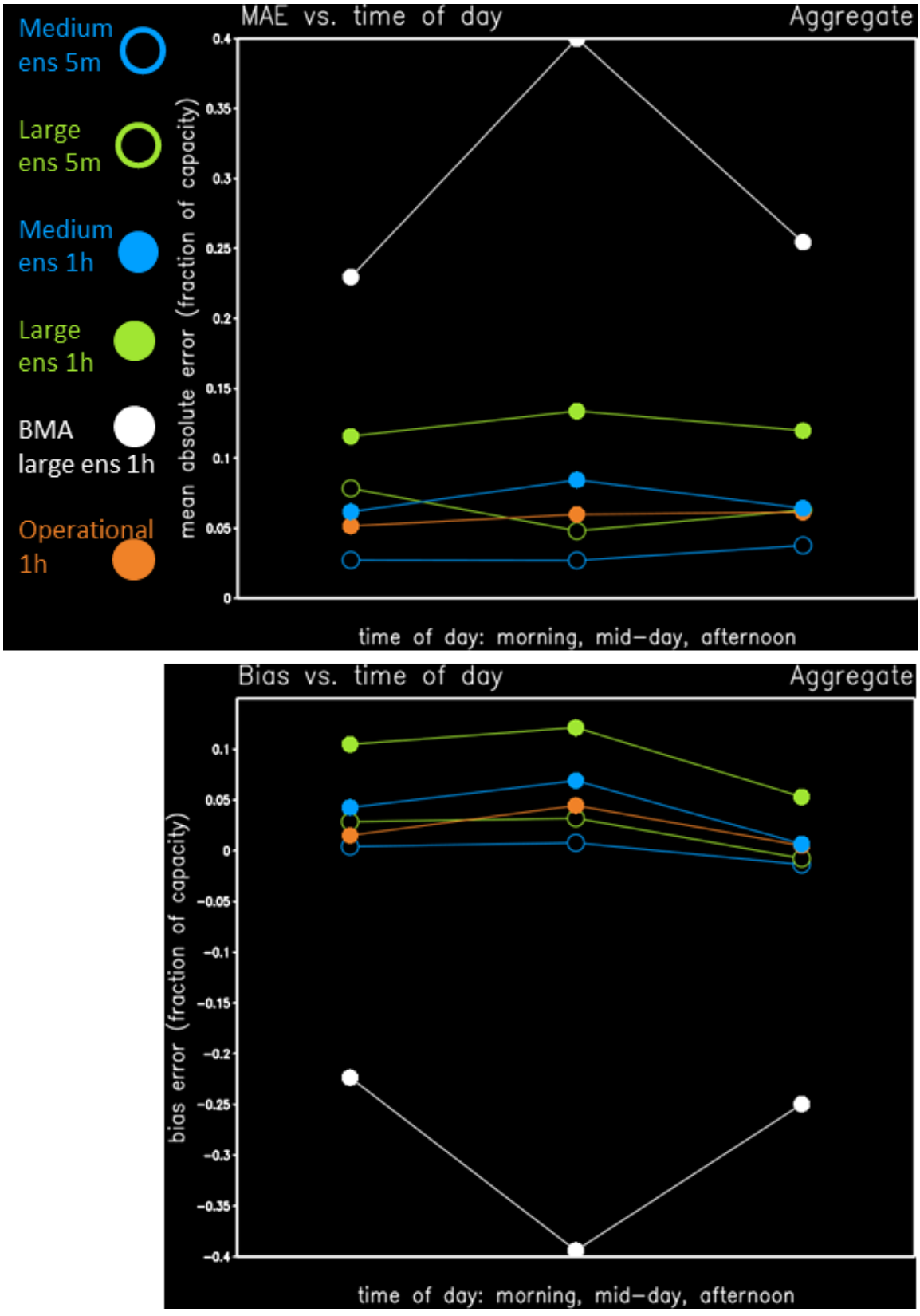
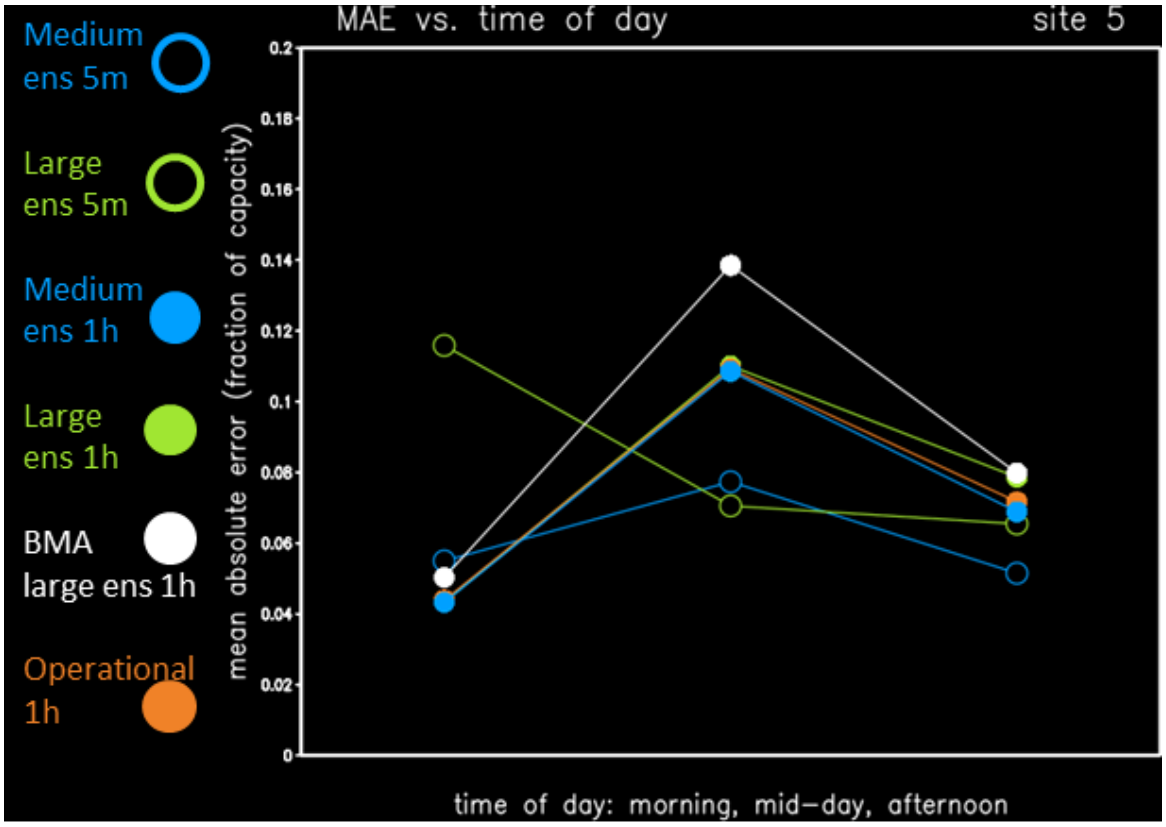


Figure 53. Mean absolute error and bias for morning, midday, and afternoon times (as defined in Figure 44) for the aggregate



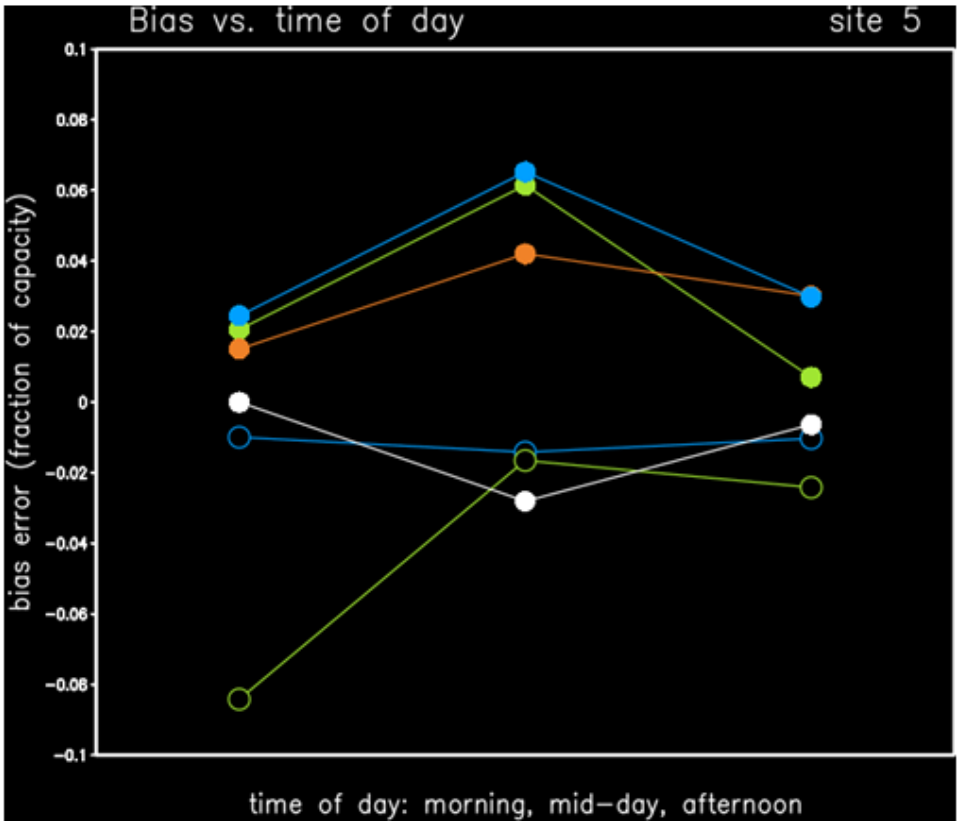


Figure 54. As in Figure 53 but for a fixed-tilt site in west Texas

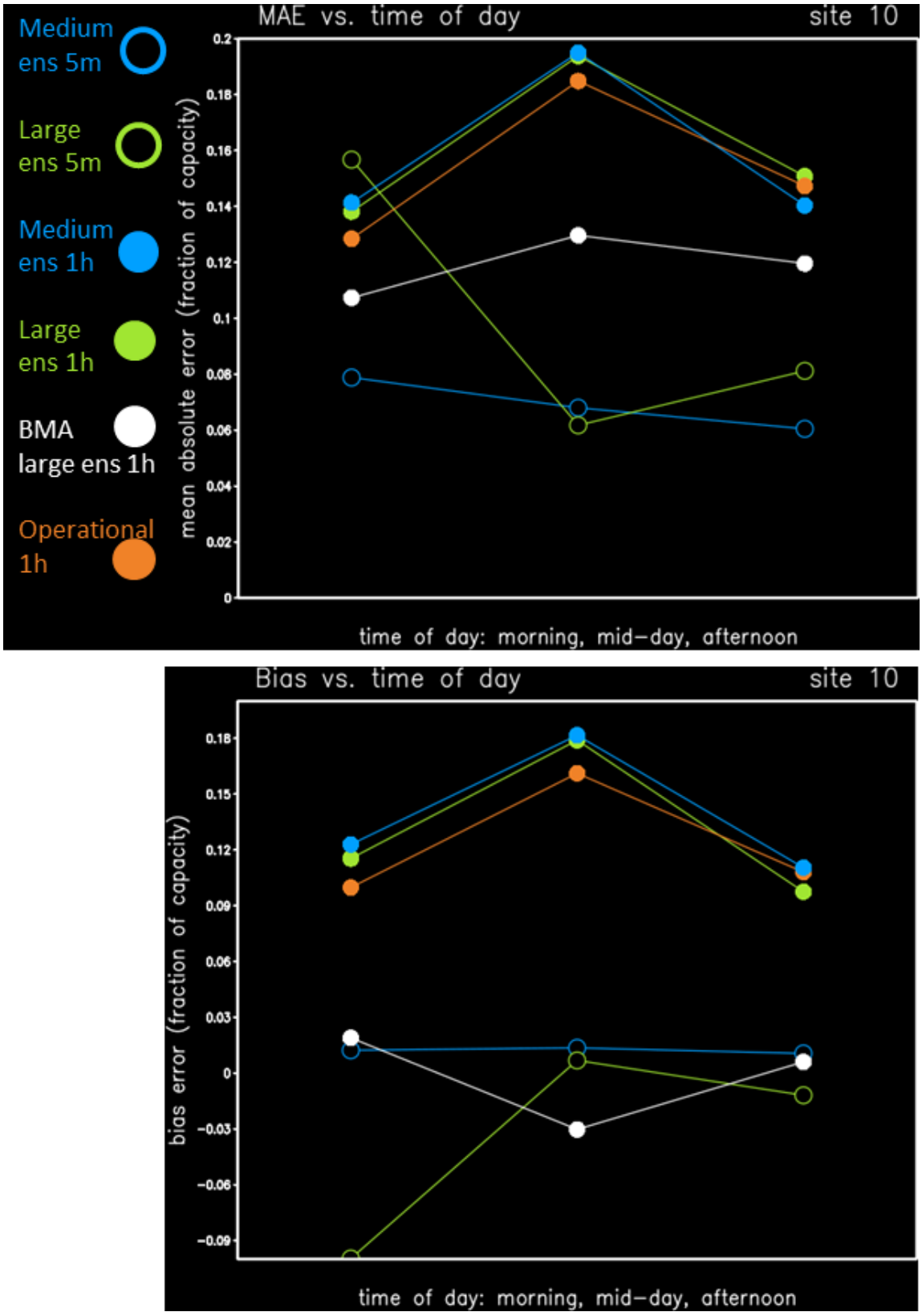
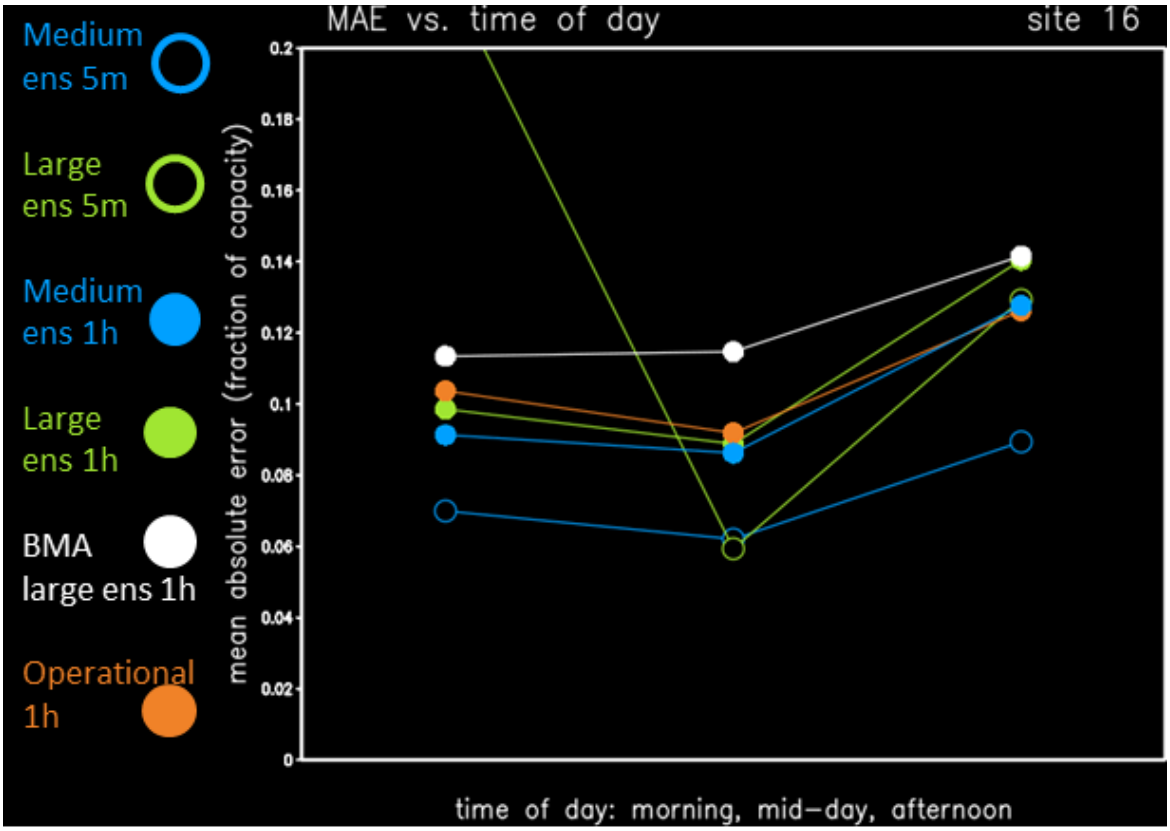


Figure 55. As in Figure 53 but for the single-axis site, which had unreported derates causing forecast bias



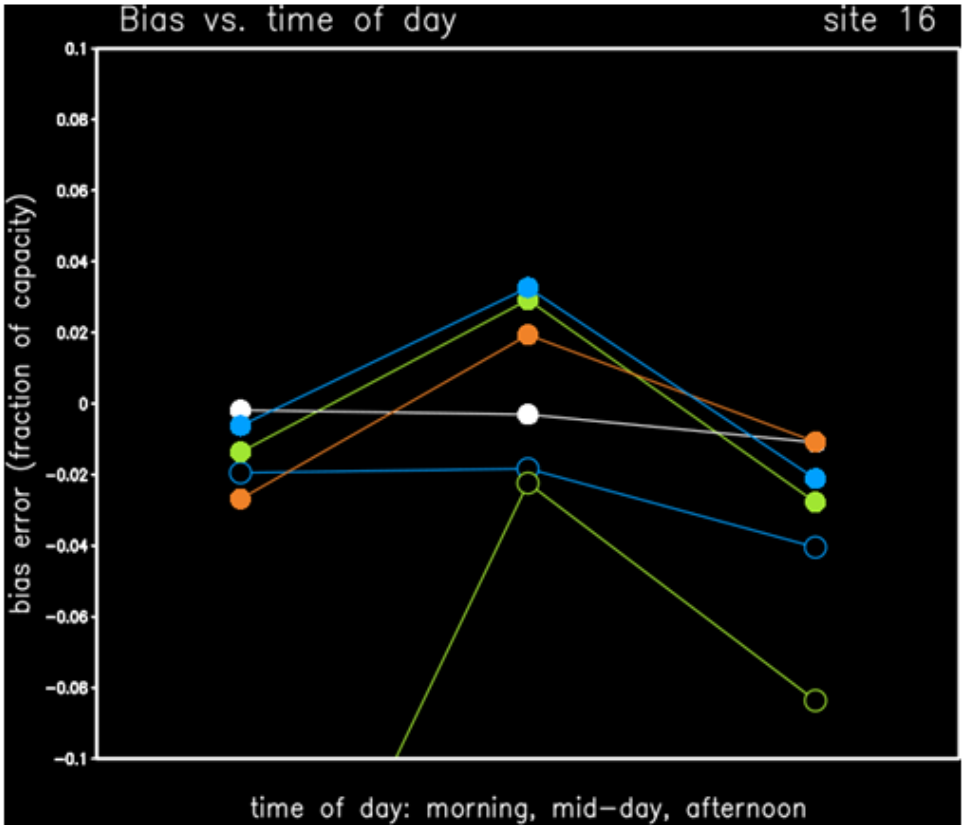


Figure 56. As in Figure 53 but for a single-axis site in sunny west Texas

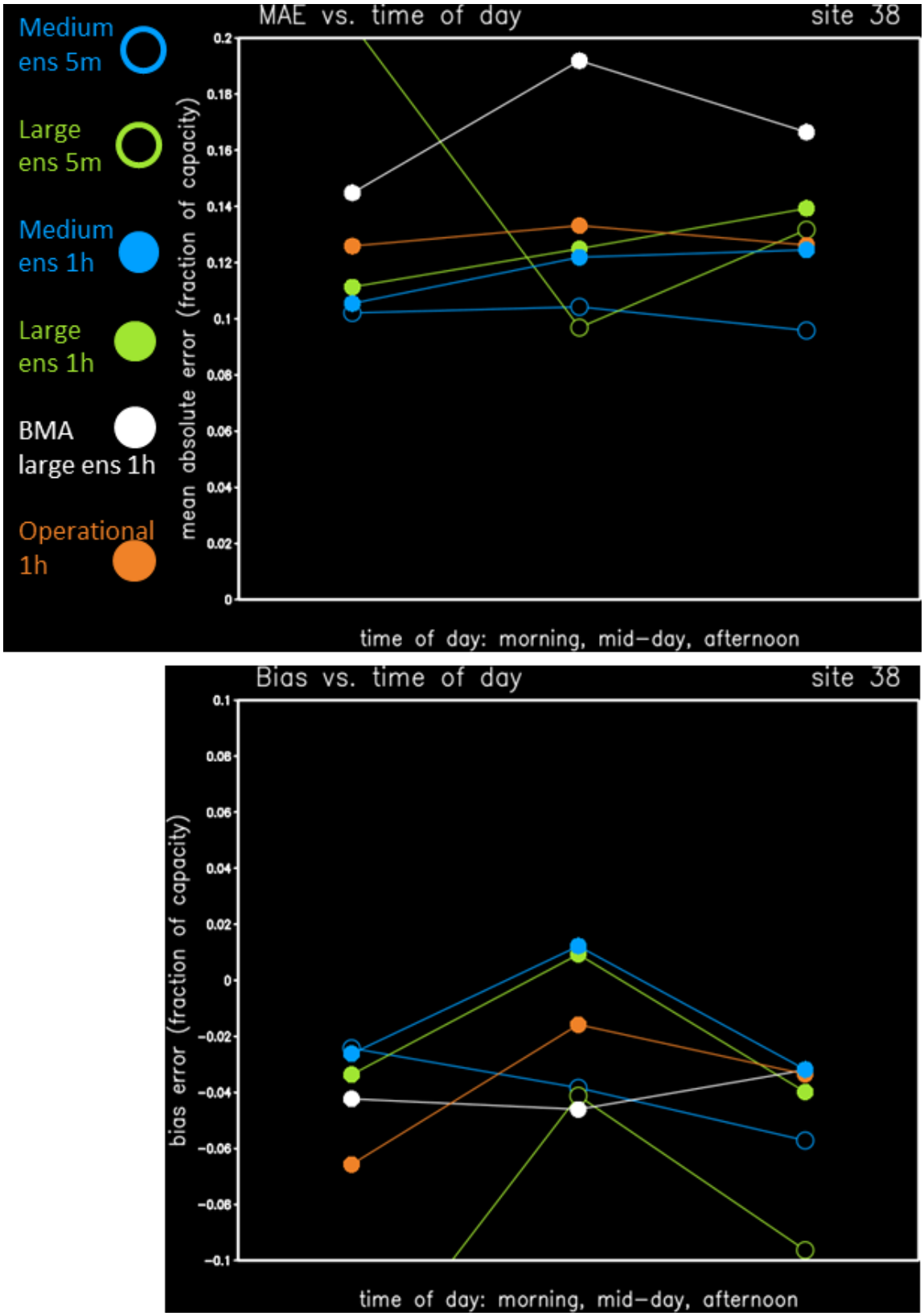
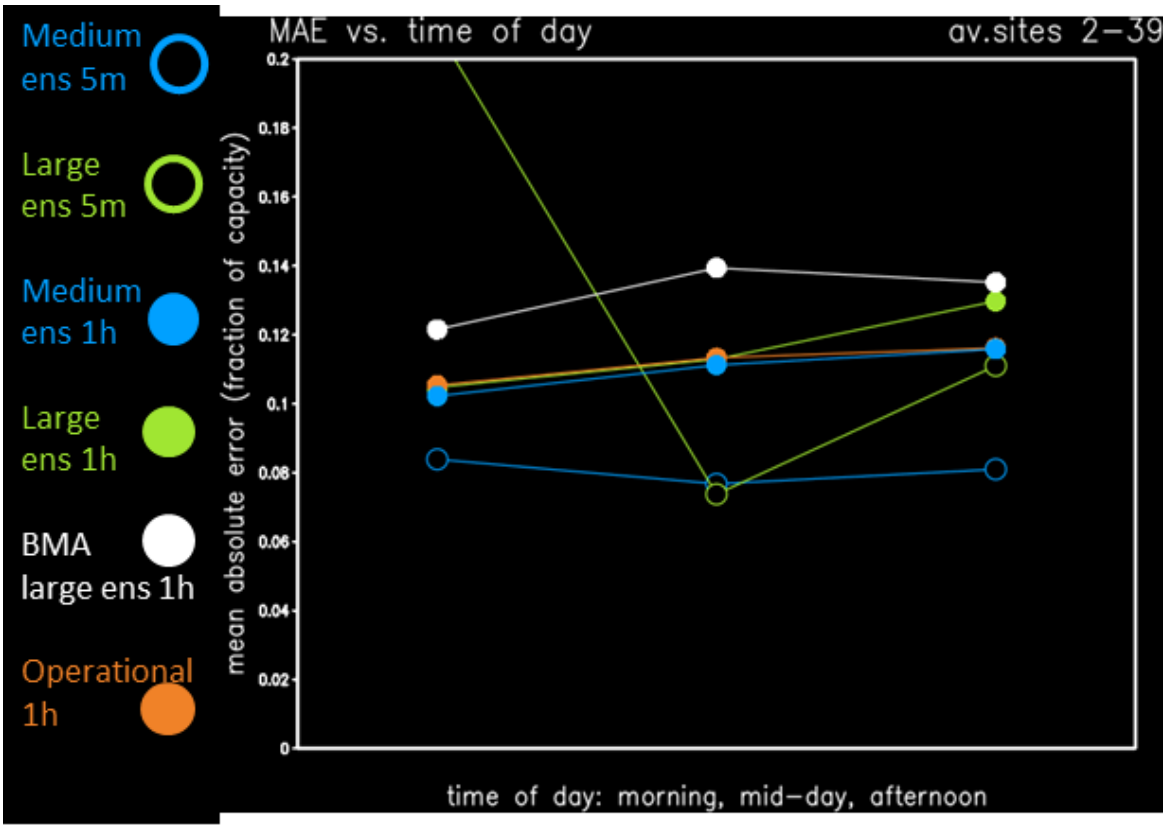


Figure 57. As in Figure 53 but for a single-axis site in east Texas with abundant, highly variable clouds

The story is the same for comparing the average scores for the long-established individual units (Figure 58) and the average scores for the newer units (Figure 59). The newer units have a larger error magnitude for all forecasts and poorer bias correction by the BMA method than for the more established units. The hourly averaged SUMMER-GO-1 and SUMMER-GO-2 forecasts have just slightly lower MAE than the operational forecast in mornings and midday hours, whereas SUMMER-GO-2 is a bit worse in the evening hours. The 5-minute forecasts have nearly identical scores for the midday hours, whereas the SUMMER-GO-2 5-minute forecasts have a large magnitude low bias in mornings and evenings.



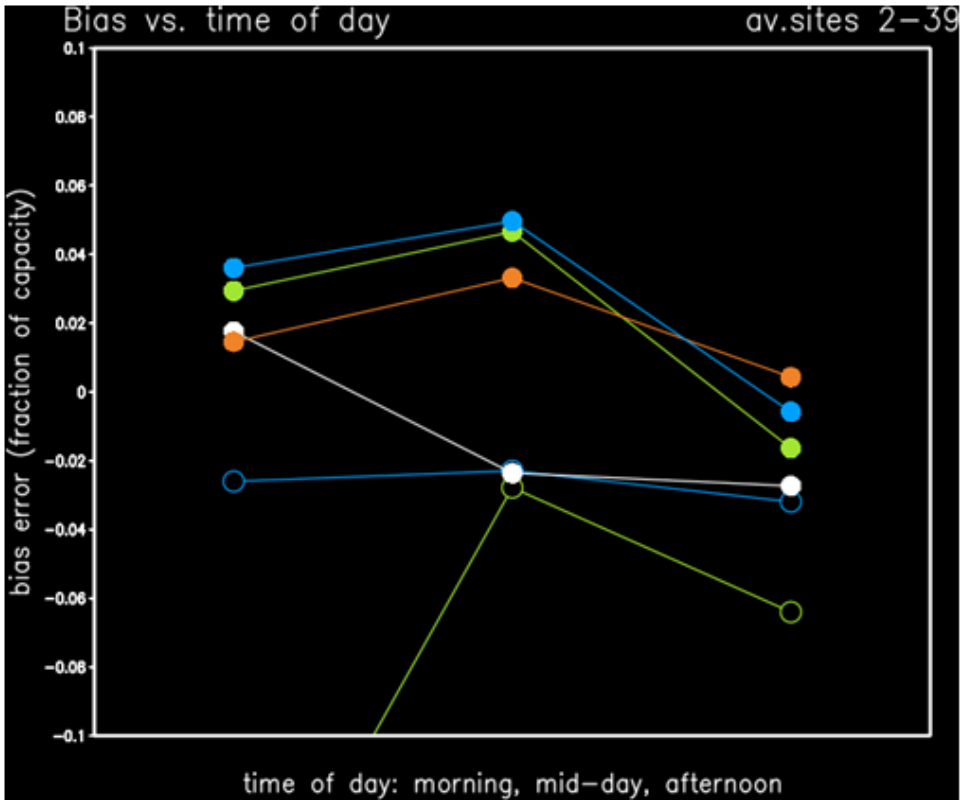


Figure 58. As in Figure 53 but for bias and MAE averaged over 38 well-established generating units

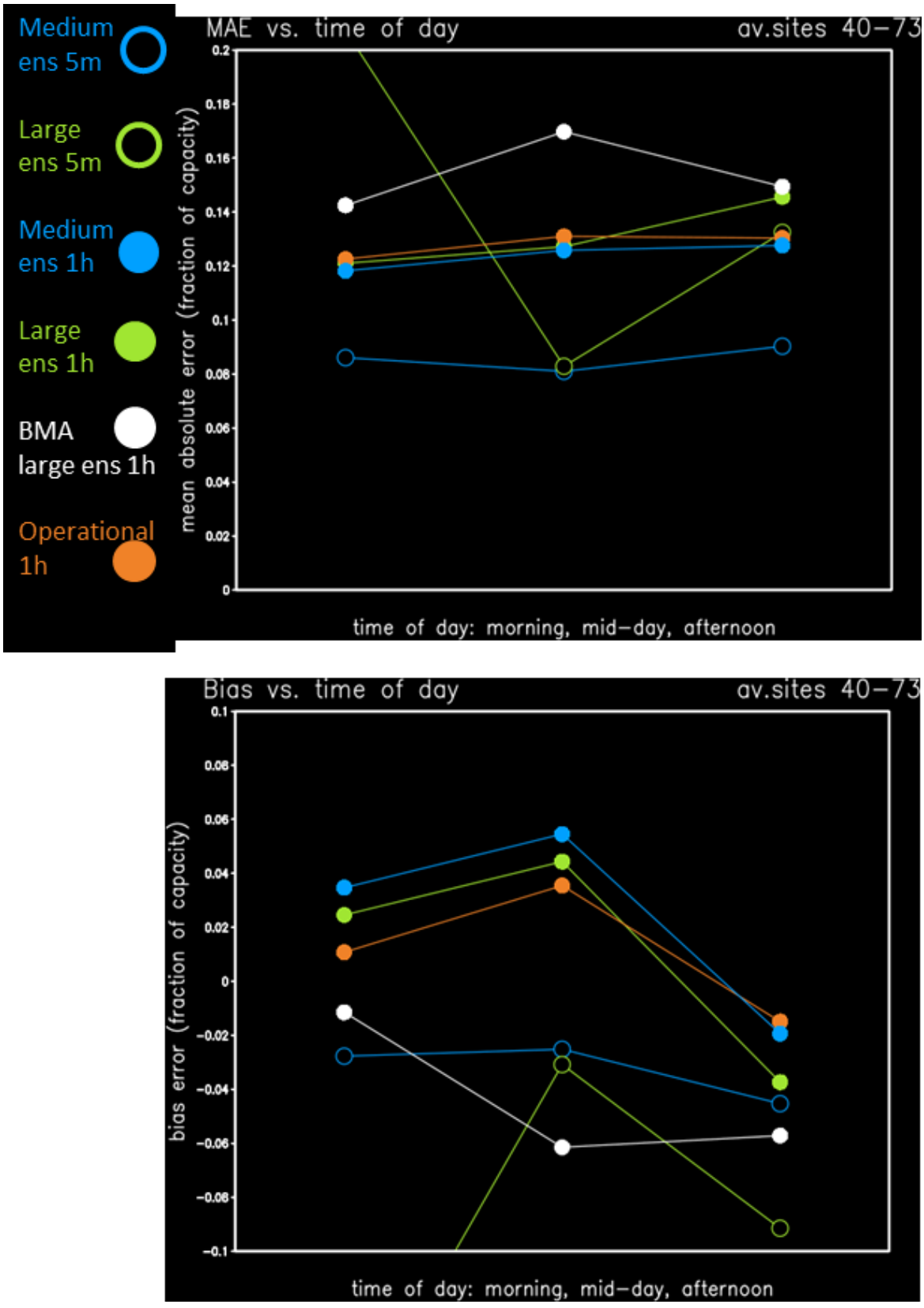


Figure 59. As in Figure 53 but for MAE and bias averaged over 34 more recently started units

Behind-The-Meter Solar Power Forecasts

Purpose and Overview

The behind-the-meter (BTM) PV generation analysis of the SUMMER-GO project examines the possibility of forecasts of BTM generation improving the load forecast in the San Diego Gas & Electric Company (SDG&E) load region of the California Independent System Operator (CAISO). This region was selected because it has a high penetration of solar power generation on the distribution grid, which impacts the electric demand required to be supplied from the transmission grid, which is what the load forecast must predict.

To have some validation estimate for BTM generation, sampled inverter data were procured from Genscape, which is now part of Wood Mackenzie. These sampled data were upscaled to the installed BTM estimates in individual zip codes and then upscaled to the SDG&E load zone aggregate. A detailed description of the upscaling methodology and analyses of those results are presented in the previous report, *Maxar Report on Examination of Escapee Behind-the-Meter Sampled Data for Use in SUMMER-GO*, by Stephen Jascourt, February 2021. The metadata used as a basis for upscaling were compiled by Maxar into daily zip code capacities and further into groupings by tilt and azimuth or tracking, derived from individual installations listed in the California Energy Commission (CEC) net-metered interconnection list.

The Genscape BTM generation data offer an element of ground truthing for the unknown BTM generation estimates; however, the Genscape data have incomplete coverage. To fill this gap, NREL created irradiance-based estimates of BTM generation using irradiances from the National Solar Radiation Database (NSRDB) and the same zip code-level daily metadata Maxar provided. Maxar evaluated these estimates compared to the Genscape-based estimates and used the comparisons to make bias corrections to the NSRDB-based estimates for the gaps in the Genscape coverage, resulting in a merged best estimate of BTM generation.

For forecasts, NREL found that the CAISO load forecasts for the SDG&E load zone for the historical period of interest are available for rolling 15-minute periods starting 45 minutes after the forecast time. BTM generation forecasts were also examined for the matching time periods. Maxar supplied irradiance forecasts from a suite of 116 NWP models. NREL converted those to BTM generation in the same manner as the NSRDB irradiances were converted to BTM generation, using the Maxar-provided metadata for PV installation characteristics in each zip code. Maxar bias-corrected these BTM generation forecasts using comparisons of forecasts over the preceding 30 days against the merged BTM generation estimates and compared the uncorrected and bias-corrected forecasts. Maxar also used the merged best estimate to create "smart persistence" forecasts that can be combined with the NWP model forecasts to make improved BTM forecasts for the lead times of interest.

This report describes in detail the process of making the merged best estimates and the model and smart persistence forecasts and compares the results with and without the bias corrections to note the improvements achieved by using the procured data.

Maxar provided the merged best estimate of BTM generation, the bias-corrected NWP forecasts, and the smart persistence forecasts to NREL. Follow-up steps are for NREL to convert the NWP

model and smart persistence BTM generation forecasts into probabilistic forecasts and to analyze the forecasts compared to errors in the load forecasts.

Observational Basis: Sampled Metered Data

The procured data contain samples of 15-minute average power for each of 86 zip codes in the region of interest for each 15-minute period in the complete calendar years 2017 and 2018. Samples were not available for 35 other zip codes in the region of interest. For each zip code, for each 15-minute period, the zip code aggregated inverter generation (AC power output) of all monitored samples was provided, along with the aggregated DC installed capacity of the sampled installations. These data were further stratified by DC installation size: 0–20 kW, 20–50 kW, 50–250 kW, and >250 kW. I estimated that the 86 zip codes contained between 85%–95% of the total aggregate power of all 121 zip codes because the missing 35 zip codes were generally those with smaller installed capacity; however, the samples within the 86 zip codes were only a small fraction of the total, resulting in metered samples totaling only 6%–8% of the aggregate power of all 121 zip codes. Despite the small percentage of sampled power, the number of installations in this region even in 2017 was so large that the samples represent an estimated more than 5,000 individual installations across the 86 zip codes, or an average of roughly 60 samples per reported zip code. My analysis presented in that report showed that the spatiotemporal coherence and structure of the resulting estimates appeared to be consistent with the known meteorology and physiographic features, including a lower generation-to-capacity ratio along the narrow coastal plain, increasing ratio with altitude, and increasing behind bluffs that partially block marine intrusions associated with stratocumulus clouds.

Daily metadata for each zip code for each date in the 2-year period of interest were created from the March 31, 2019, version of the CEC's net-metered interconnected list. The list contained start dates for each of the >100,000 installations in the 121 zip codes. The total aggregate across the 121 zip codes increased from 693 MW on January 1, 2017, at the start of the period of interest to 1,034 MW on December 31, 2018, at the end of the period of interest. I tallied the daily AC and DC capacity in each zip code.

The sampled data for each size bin were upscaled into a zip code total power for each 15-minute period as follows:

$$\text{Powerzip} = \text{DCCEC} * \text{sum}(\text{DCCEC}(i)*\text{PowerGen}(i)/\text{DCGen}(i)) / \text{sum}(\text{DCGen}(i))$$

where *Powerzip* is power upscaled to the zip code, *DCCEC* is the CEC zip code installed DC capacity, *DCCEC*(*i*) is the CEC zip code DC capacity for size bin *i*, and *DCGen*(*i*) is the Genscape DC capacity for size bin i and the sums are made only for the size bins having nonzero capacity. One zip code had Genscape data only for size bins showing zero in the CEC list and no Genscape data for the size bins having nonzero capacity in the CEC list, so that zip code had to be totaled using:

$$\text{Powerzip} = \text{DCCEC} * \text{sum}(\text{PowerGen}(i)) / \text{sum}(\text{DCGen}(i)).$$

Merged Observation-Based "Truth" Data Set

D. NSRDB-Based Estimates

The NSRDB has direct normal and diffuse irradiance at approximately 4-km spatial resolution every 30 minutes for the region and period of interest. These irradiances represent snapshots primarily based on GOES observations using a physical retrieval method.

NREL created BTM generation estimates from these NSRDB irradiances at the zip code centroid latitude and longitude for each of the 121 zip codes by passing irradiances through SAM. SAM requires inputs on the tilt and orientation of the solar panels, AC and DC capacity, and temperature and wind speed, as well as assumptions about overall power losses. Maxar supplied the temperature, the wind speed, and the installation metadata for each zip code. Maxar supplied daily metadata as described in the previous section, derived from the CEC interconnected net-meter list, further broken into more detail by providing the AC totals into fixed tilt and tracking, and for fixed tilt, by breaking it into combinations of 5 tilt ranges and 7 azimuth ranges for a total of 35 combinations of tilt and azimuth in each zip code. The CEC data contained the installation configuration data for many locations, but some were missing this information or listed it as "multiple" or "mixed." The missing or ambiguous configurations were randomly assigned based on the frequency of each in the collection of sites containing this information (across all 121 zip codes, not only in the zip code of the site being assigned, to ensure a large sample for the frequencies used in the random assignment). For each zip code, NREL passed each solar panel configuration through SAM using the same zip irradiance time series, then weighted it by the AC power for each configuration, and summed to make the zip code total.

Maxar converted the NREL NSRDB-based BTM generation estimates in each zip code into 15-minute time averages to time match the metered samples, the forecast, and the load data by assuming linear trends in time between the 30-minute snapshots as follows:

	hh:00	hh:15	hh:30	hh:45	hh+1:00
Instantaneous	A		B		C
Time averaged	(3A+B)/4	(3B+A)/4	(3B+C)/4		(3C+B)/4

These results, however, also had to be time-shifted by 15 minutes to align with the Genscape results, perhaps owing to the satellite scan times for the latitude around 32°N being offset from the 30-minute time stamps. To determine the best time alignment, the upscaled sample estimates and the NSRDB-based 15-minute time averages were shifted from UTC to PST so that each daytime period would be in the middle of the day, and sunny days were compared so that the shapes of the diurnal curves would be similar, only minimally affected by clouds. The 730-day time series was divided into 73 10-day periods, and the highest observed (estimate from the metered samples) day was selected as the sunniest day in each 10-day period. The daily results for the 73 sunny days were scaled by the daily maxima (e.g., NSRDB daily time series/NSRDB daily maximum versus upscaled samples daily time series/upscaled samples daily maximum); thus, the shapes of the daily curves were compared without being affected by amplitude or bias differences. Difference statistics were computed across the 73 sunny days for each of the 86 zip codes having data for both metered samples and NSRDB-based generation estimates, and the same was done by shifting the NSRDB-based series by 15, 30, 45, and 60 minutes earlier and

later for 9 comparisons (4 earlier, no shift, 4 later). After the 15-minute shift was found to be optimal and the NSRDB data were shifted by 15 minutes, the comparison was repeated, and the results are shown in Figure 60. The figure shows the RMS difference as a fraction of capacity for each zip code (x-axis) for each time shift (y-axis) for each of the 86 zip codes. The other 35 zip codes are the blank (black background) vertical stripes. This clearly shows that now the minimum difference in the shape requires no further shift, and the difference is approximately symmetric for shifts earlier and later. This result holds for almost every zip code individually.

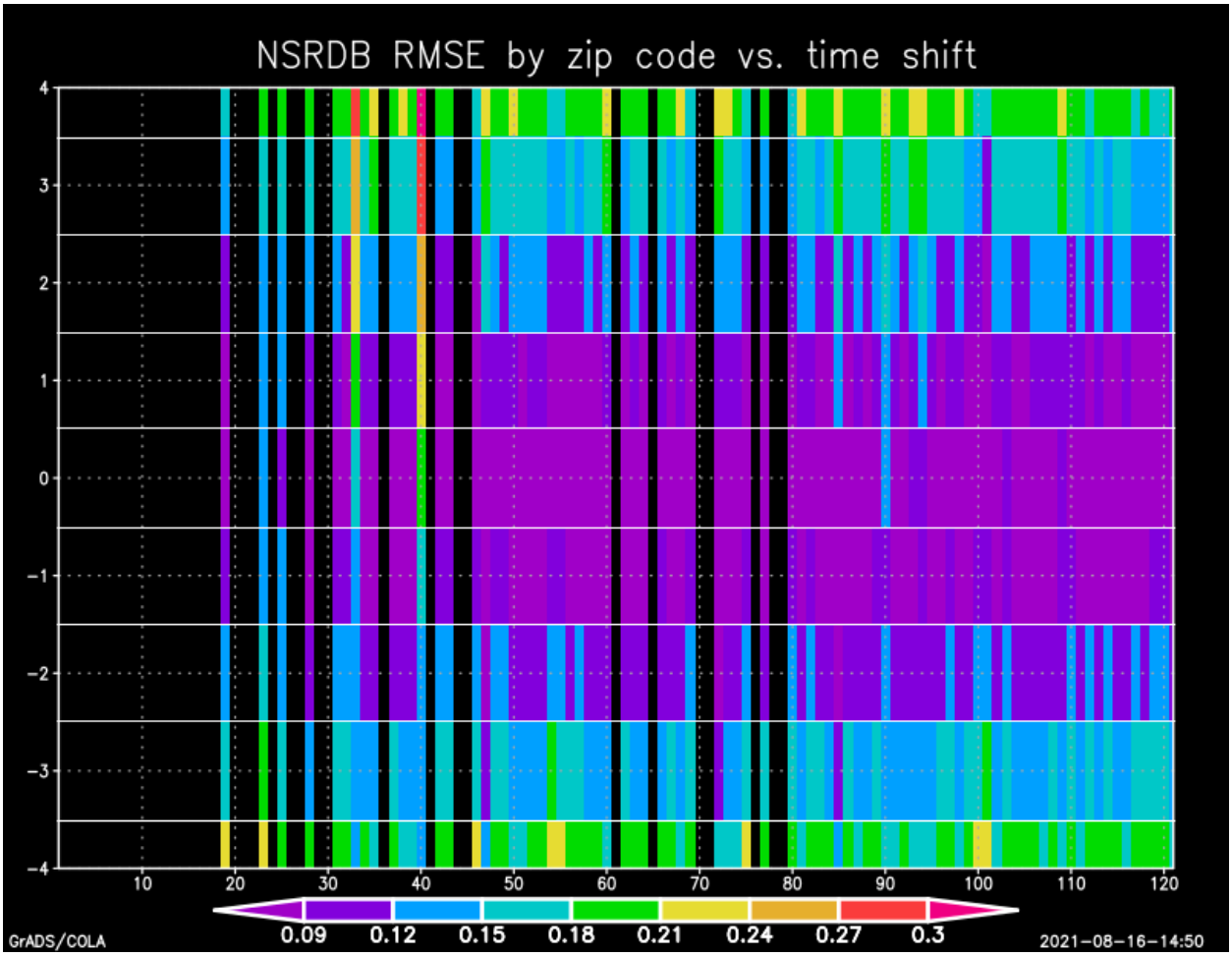


Figure 60. RMS differences as explained in text for each zip code (horizontal axis) for the indicated number of 15-minute time shifts (vertical axis) after one 15-minute time shift was already applied

E. Comparison Between NSRDB-Based and Sample-Based Estimates

Several comparisons were made to evaluate the NSRDB-based estimates for the 86 zip codes containing sample-based estimates. The daily bias of the NSRDB-based estimates, shown in Figure 61 (bias = difference in daily daytime means as a fraction of AC capacity) is close to zero in many zip codes, although mostly on the high side (between 0 and 0.1), indicating an overestimate using the NSRDB approach; however, some zip codes are consistently low, and others are often high. Daily RMS differences, shown in Figure 62, show large day-to-day differences that are consistent across most zip codes, indicating that those days have discrepancies, whereas other days have good agreement. An example of the differences in one

zip code, the zip code with largest installed capacity, is shown in Figure 63, indicating the largest differences between midmorning and midday and the smallest differences in the late afternoon, when morning clouds have usually dissipated, and in the periods near sunrise and sunset, when the power is low regardless. Figure 64 shows the same for the aggregate over the 86 zip codes but with a smaller magnitude of differences, in this case as a fraction of daily maximum rather than of capacity.

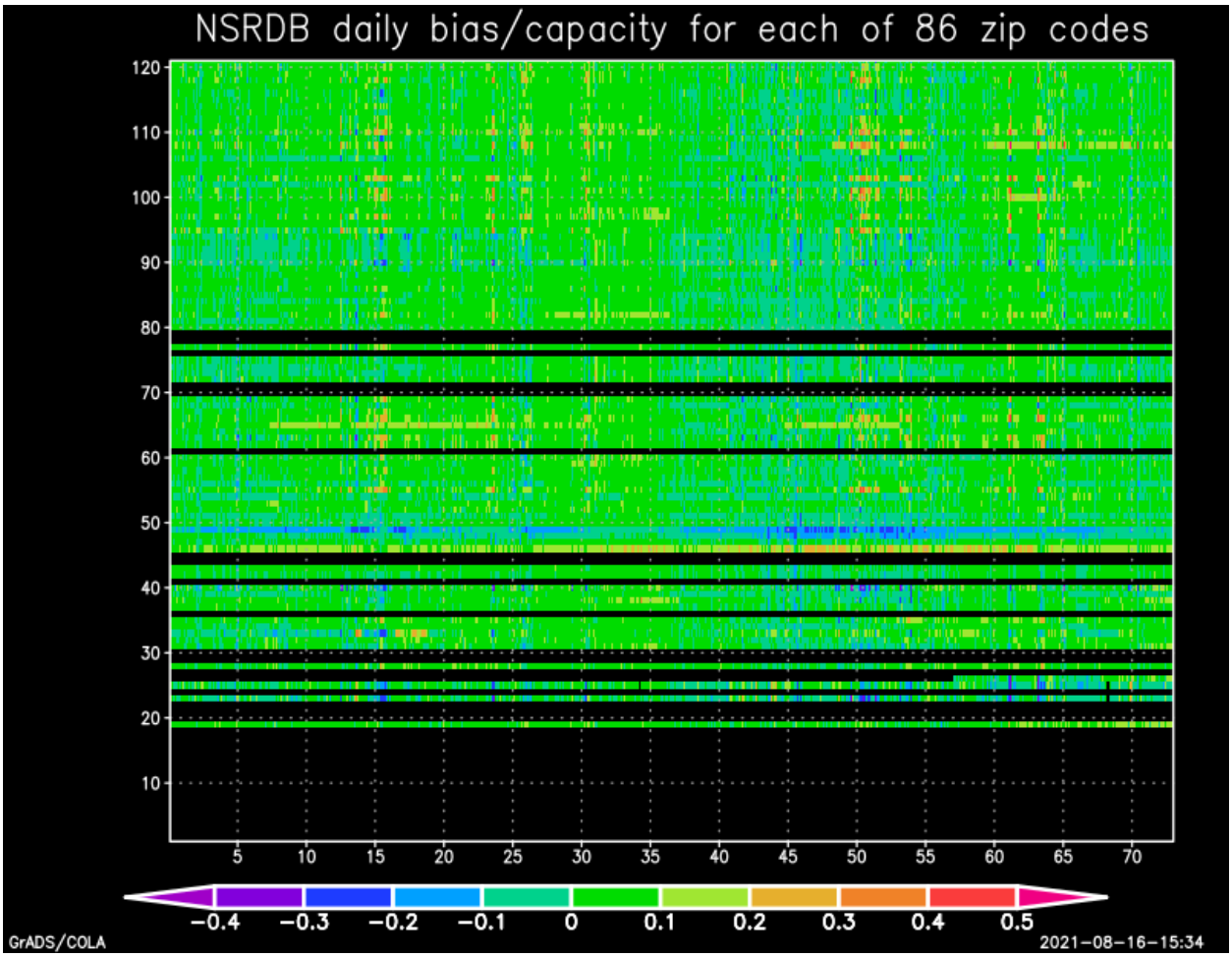


Figure 61. Daily bias (daytime mean of NSRDB minus daytime mean of sample-based estimate) as a fraction of capacity for each zip code (vertical axis, horizontal stripe per zip code) for each of the 730 days (horizontal axis, each vertical stripe is 1 day)

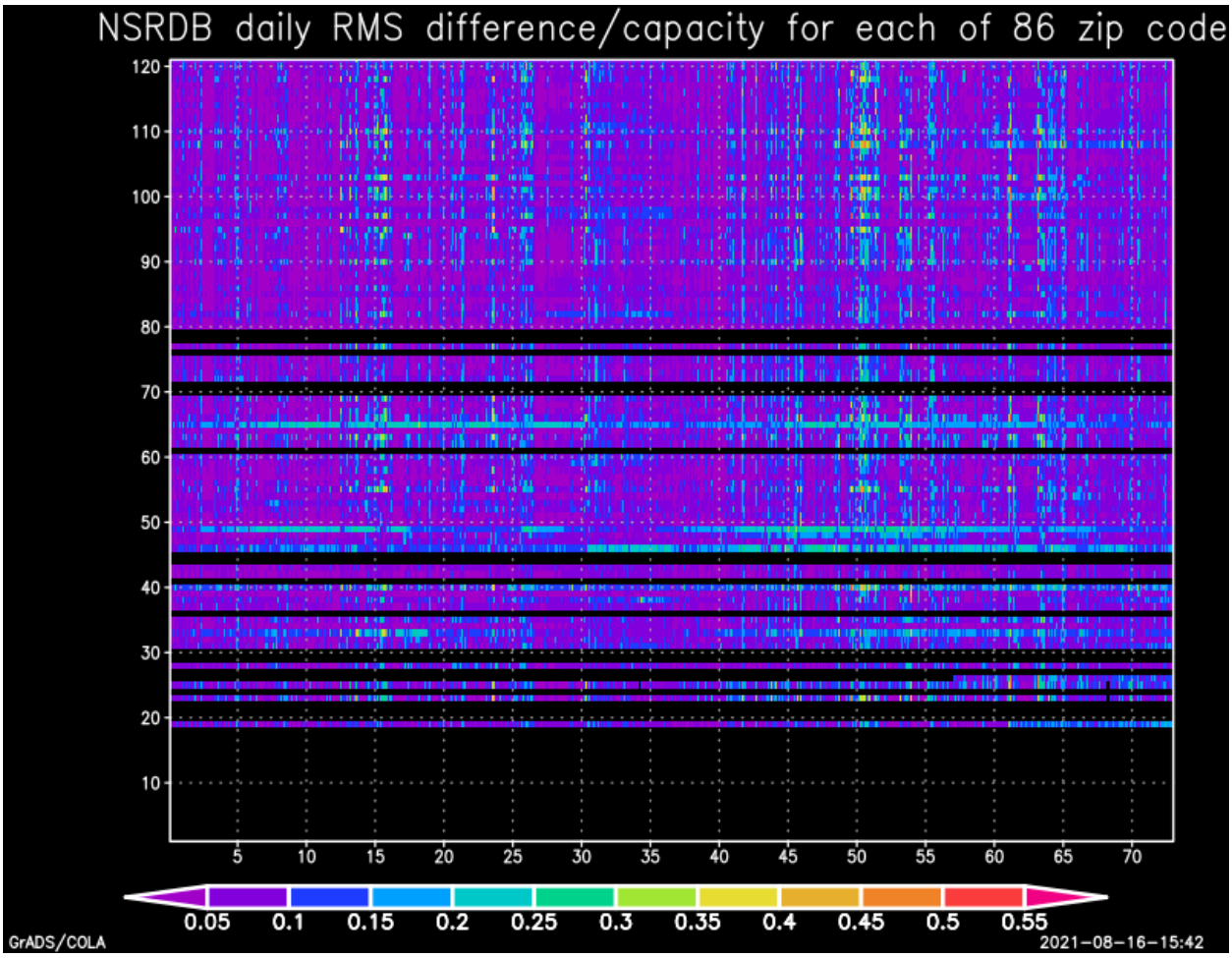


Figure 62. Daily RMS difference (daytime RMS difference of NSRDB and sample-based estimates) as a fraction of capacity for each zip code (vertical axis, horizontal stripe per zip code) for each of the 730 days (horizontal axis, each vertical stripe is 1 day)

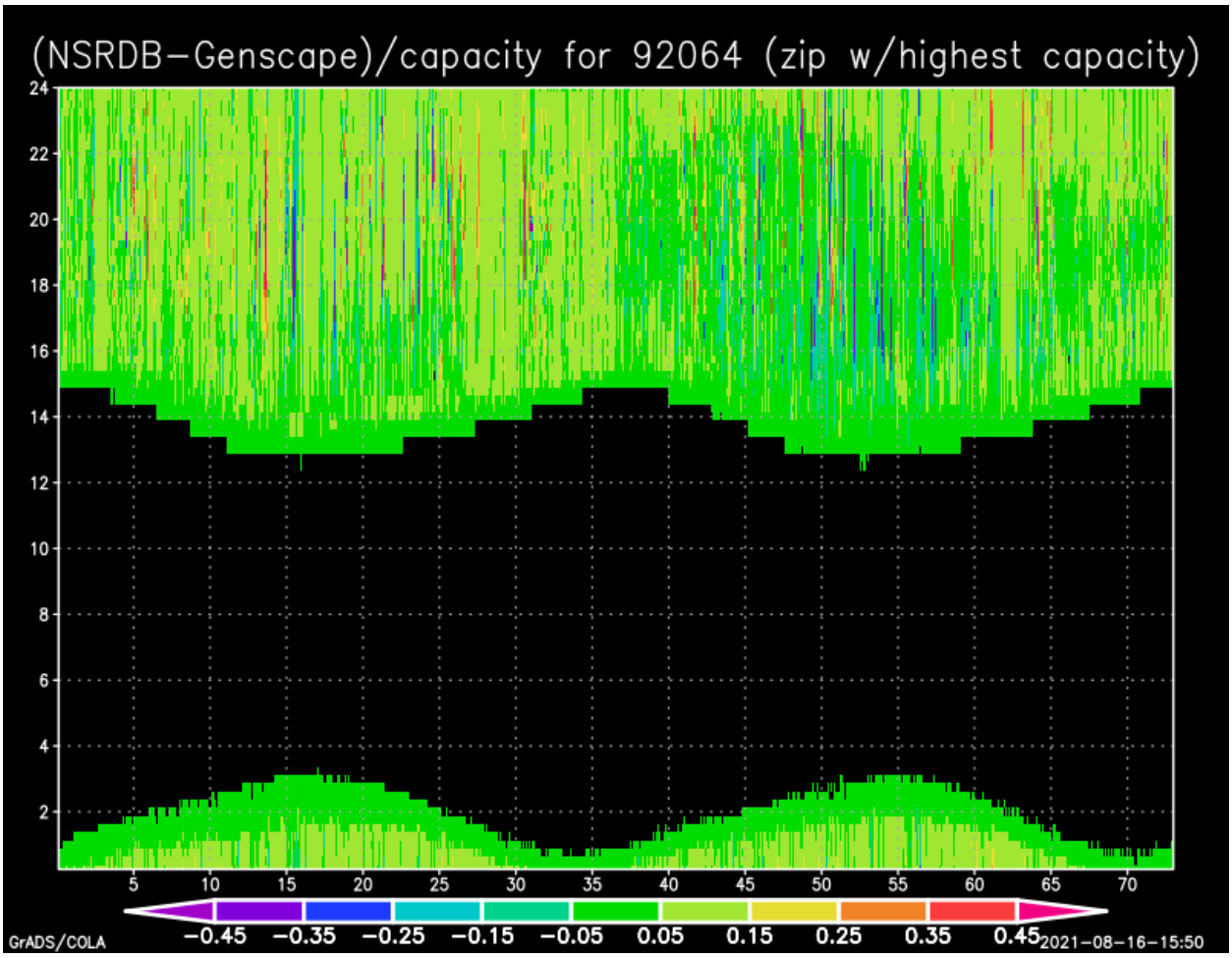


Figure 63. Differences for a single zip code for each day (horizontal axis, vertical stripe is 1 day) and each 15-minute period during the day (vertical axis, labels are hours UTC, horizontal stripe is one 15-minute time period for all days)

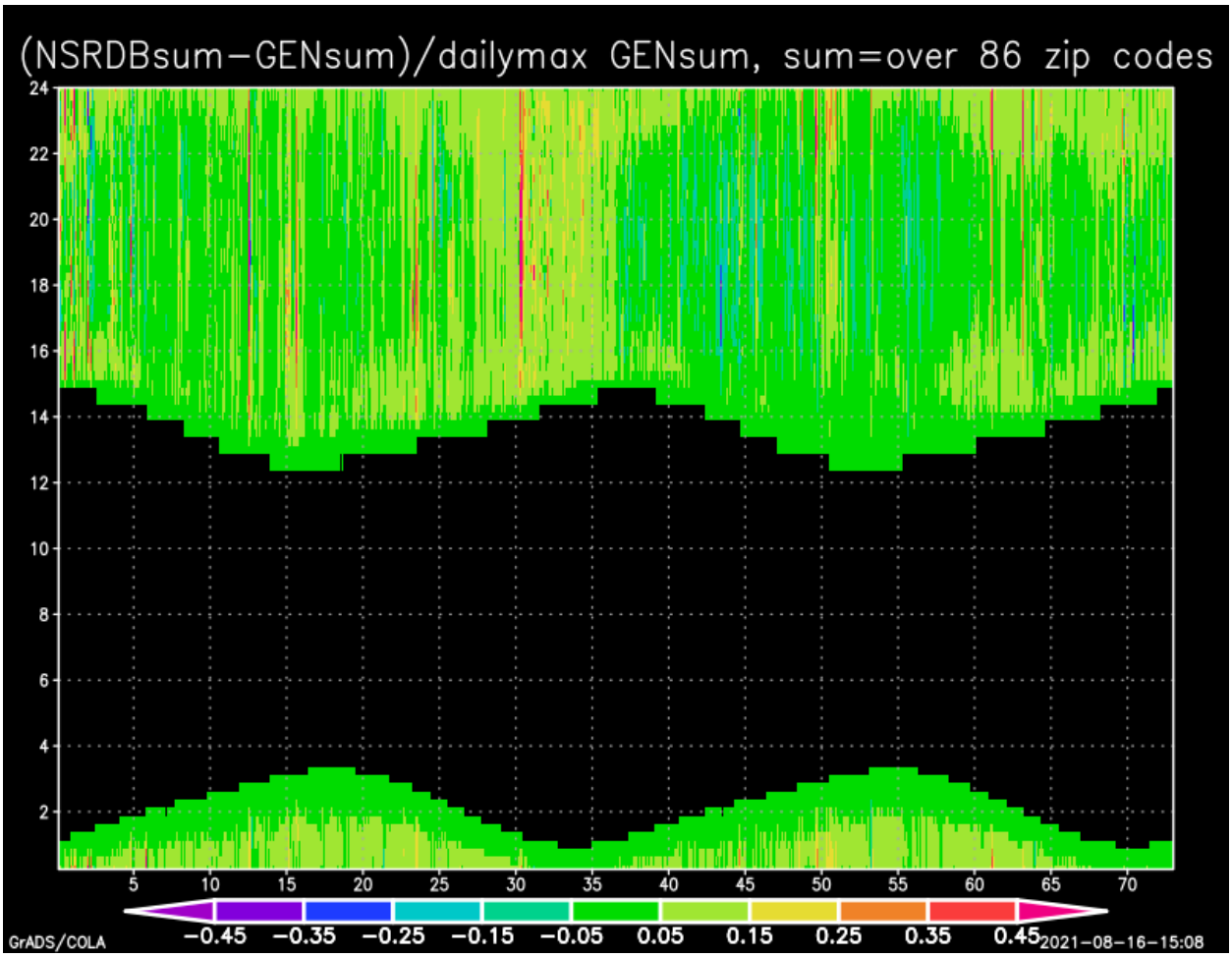


Figure 64. Differences for the sum of 86 zip codes for each day (horizontal axis, vertical stripe is 1 day) and each 15-minute period during the day (vertical axis, labels are hours UTC, horizontal stripe is one 15-minute time period for all days)

Grouping the differences by hour of the day for each month and combining the same month in 2017 and 2018 (Figure 65) reveals the largest variability of the differences in May and June, and it is also large in September, whereas the smallest variability of differences is in December, January, and April. Variability is largest in the middle of the days because the envelope of constrained values is narrower at other times of day, although it is skewed slightly toward larger in mornings than afternoons.

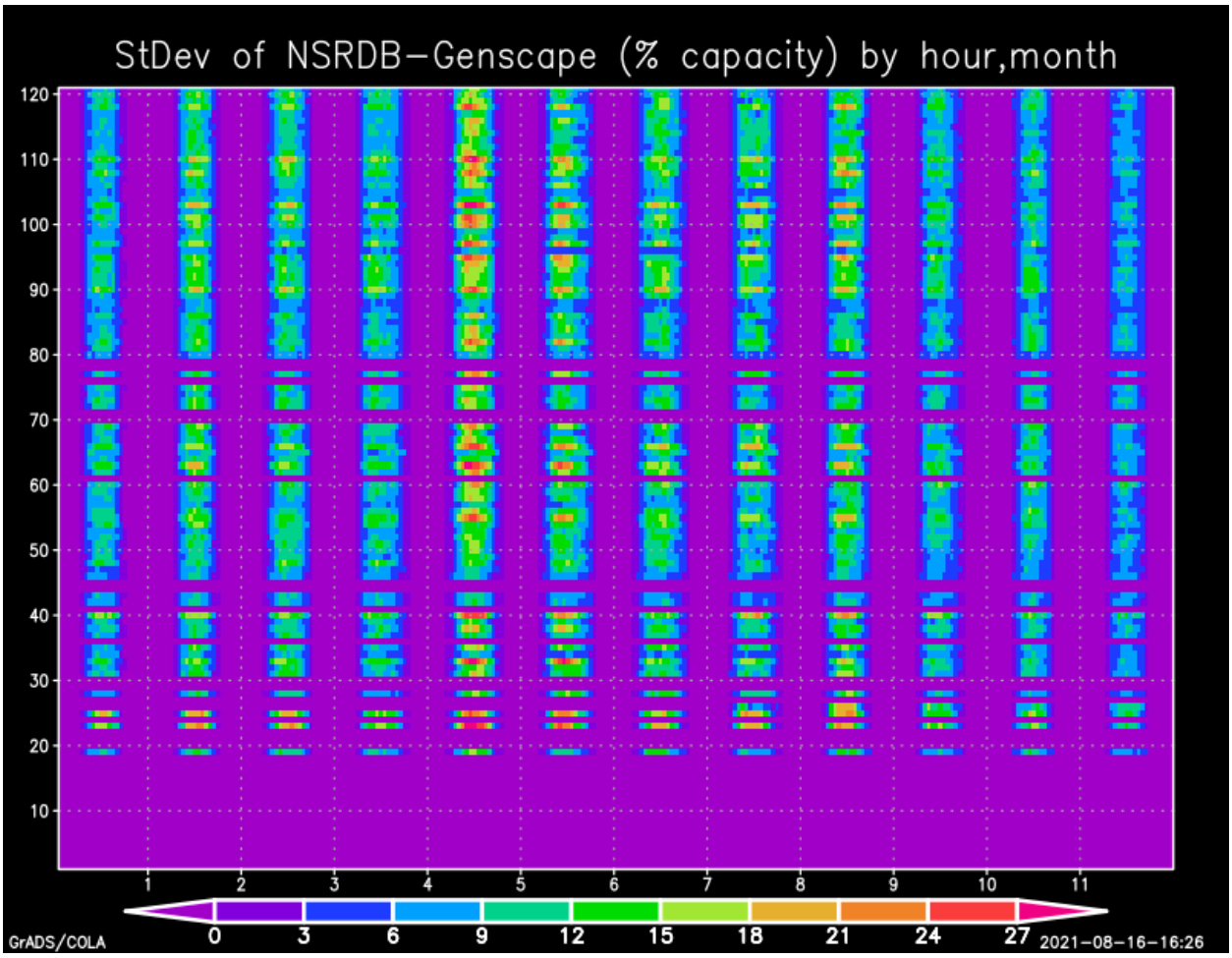


Figure 65. Variability of differences, as measured by the standard deviation of the differences within each hour of the day for each month of the year

The horizontal axis is labeled by month, with each plotted square an hour of the day in PST. Each horizontal stripe, the vertical axis, is a different zip code. Zero values (purple) indicate no data due to night or one of the 35 zip codes not sampled.

F. Bias Correction of NSRDB-Based Estimates

The NSRDB-based estimates were bias-corrected to match the sample-based estimates by multiplying by the ratio:

$$A = \text{sum(sample-based estimates)/sum(NSRDB-based estimates)}$$

where the sum is applied over times within the same hour of the day in the same month of the year, combining the months in 2017 and 2018. For a 30-day month during daytime hours, that would be 60 days x four 15-minute periods = 240 time periods in each sum.

The resulting A values are shown in Figure 66 and are plotted in the maps in the appendix (separate document) in the same manner as the climatology appendix maps of the sample-based estimates of power/capacity in the previous report.

The *A* are small in the morning and evening for almost all months for almost all zip codes. This might be due to shading at a low sun angle not accounted for by SAM, which does not allow for time-dependent or sun angle-dependent losses such as shading. There also are both small and large *A* ratios in the mornings in May and June, possibly indicating some problems with the NSRDB-based irradiances during these more volatile times. And a few zip codes show a large *A* ratio most hours of most months, perhaps indicating something unusual about the installation in those zip codes. In Figure 67, the *A* ratio for an individual zip code, the one with the most installed capacity, shows the small *A* in the mornings and evenings, with *A* peaking above 0.9 in the middle of the day each month. Although *A* indicates the overall bias by hour and month, the variability of the NSRDB-based estimates minus the sample-based estimates by hour and month is shown by the green curve. This variability is larger in May than any other month, with June the next largest.

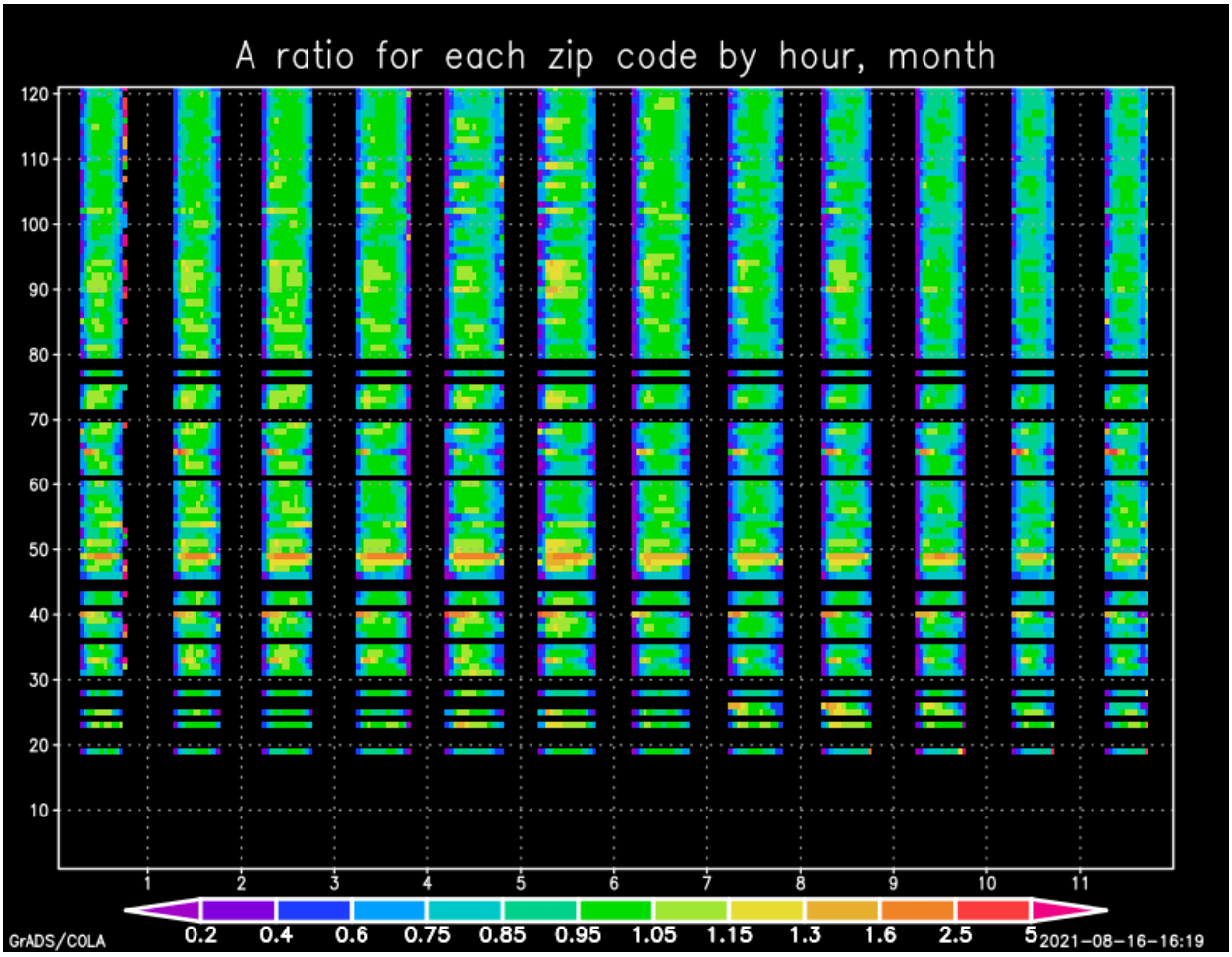


Figure 66. A ratio (defined in the text) for each hour of the day for each month of the year

The horizontal axis is labeled by month, with each plotted square an hour of the day in PST. Each horizontal stripe, the vertical axis, is a different zip code. Missing values (black) indicate no data due to night or one of the 35 zip codes not sampled.

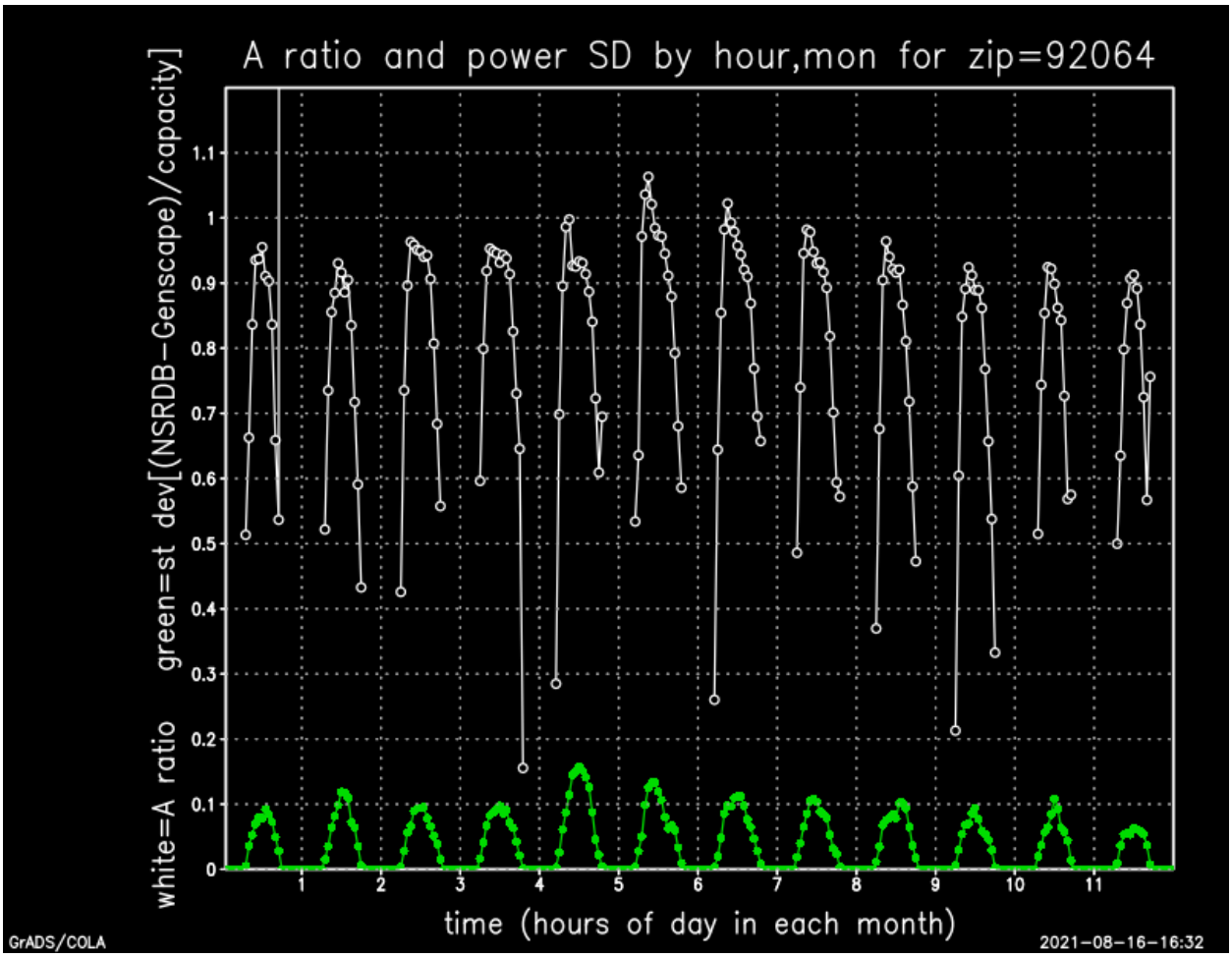


Figure 67. A ratio (defined in the text) for each hour of the day for each month of the year

The horizontal axis is labeled by month, with each plotted square an hour of the day in PST. Each horizontal stripe, the vertical axis, is a different zip code. Zero values (purple) indicate no data due to night or one of the 35 zip codes not sampled.

The difficulty is that the purpose is to correct the NSRDB-based estimates for the 35 zip codes not having sampled data; therefore, the *A* ratio needs to be imputed. Three imputation strategies were tested using cross-validation, in which the imputation is applied to a zip code without using data from that zip code and then compared against the *A* ratio calculated directly for that zip code.

The imputation strategies evaluated were:

1. Median of the five nearest neighbors with data, constraining the "neighbors" to zip codes within an elevation range 2*e(imputed zip code)/3–50 <e(neighbor) <1.5*(e(imputed zip code)+50), where *e* is the elevation in meters, but the elevation constraint applies only on the upslope from the coast inland, west of a line from 32.6°N 116.4°W to 33.6°N 117.4°W
2. Nearest neighbor (no constraints) after omitting zip codes with *A* more than three standard deviations from the mean for the same hour, month combination

3. Average of all zip codes after omitting zip codes with *A* more than three standard deviations from the mean for the same hour, month combination.

The cross-validation results, shown in Figure 68, indicate a large error in the morning and evening bias-correction *A* values and also larger errors in May and June, with a midday RMSE of only approximately 0.05 for other months.

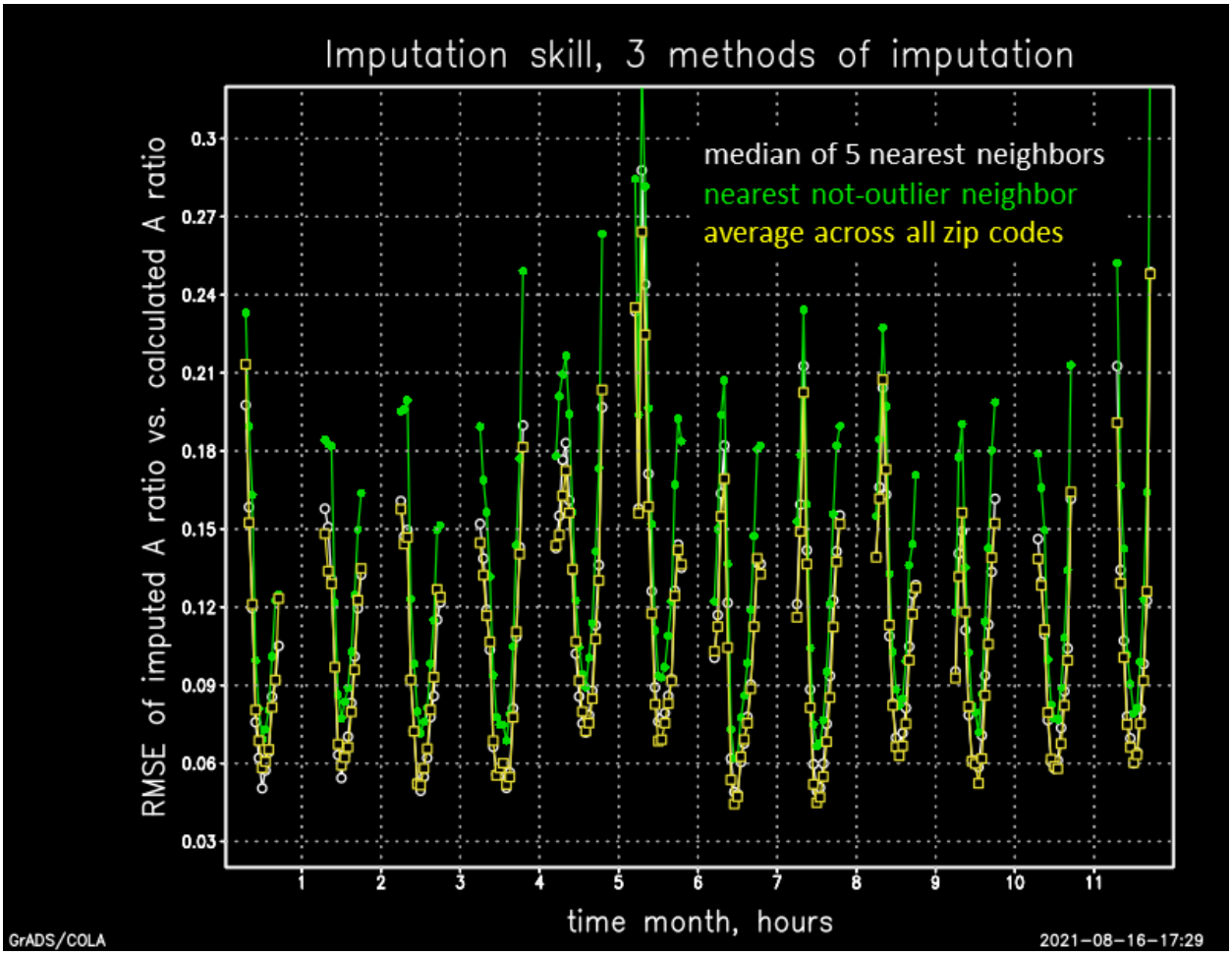


Figure 68. Cross-validation results for the *A* ratio, where the horizontal axis is the hour of the day for each month of the year. Gaps are the night hours.

Taking it one step further, the corrected power estimates were computed from the bias-corrected *A* values. The large *A* differences in morning and evening do not result in large power differences due to the small generation at those times. For power, the root mean square (RMS) difference of the NSRDB-based estimate and the sample-based estimate was computed for no correction and for applying the *A* values calculated directly and imputed using each of the three methods. This RMS difference is considered the "error," using the sample-based estimate as the "truth." Then, the error should be smallest for the power estimates that were bias-corrected using the directly calculated *A* rather than the imputed *A*. To better visualize the differences in skill, the differences in the power error as a fraction of capacity are shown for one zip code, with the largest installed capacity, in Figure 69 and for the aggregate in Figure 70. All three imputation methods are clearly superior to not making a bias correction. For the individual zip code, the

imputations have the highest error relative to using the directly calculated *A* ratio during the afternoons, and the method (1) is the best method during the difficult months of May and June. For the aggregate power of the 86 zip codes, all three methods have a similar skill, and the skill is as good as using the directly calculated *A* for some months.

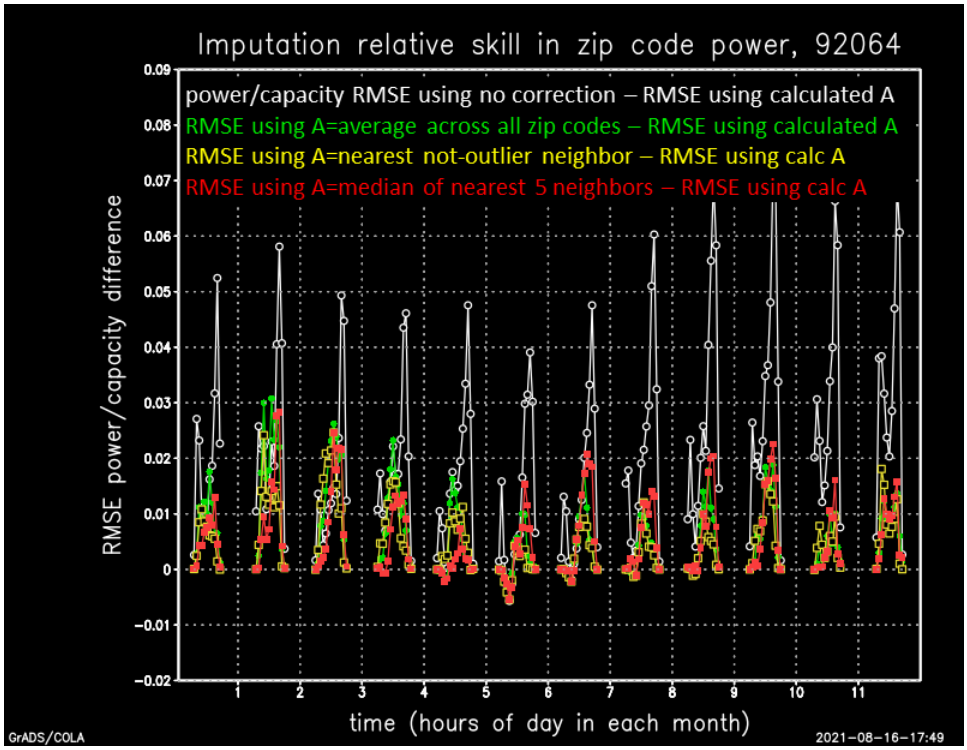


Figure 69. Difference in RMSE of power/capacity between using the imputed *A* and the calculated *A* for one zip code for three different imputation methods and no corrections for each hour by month

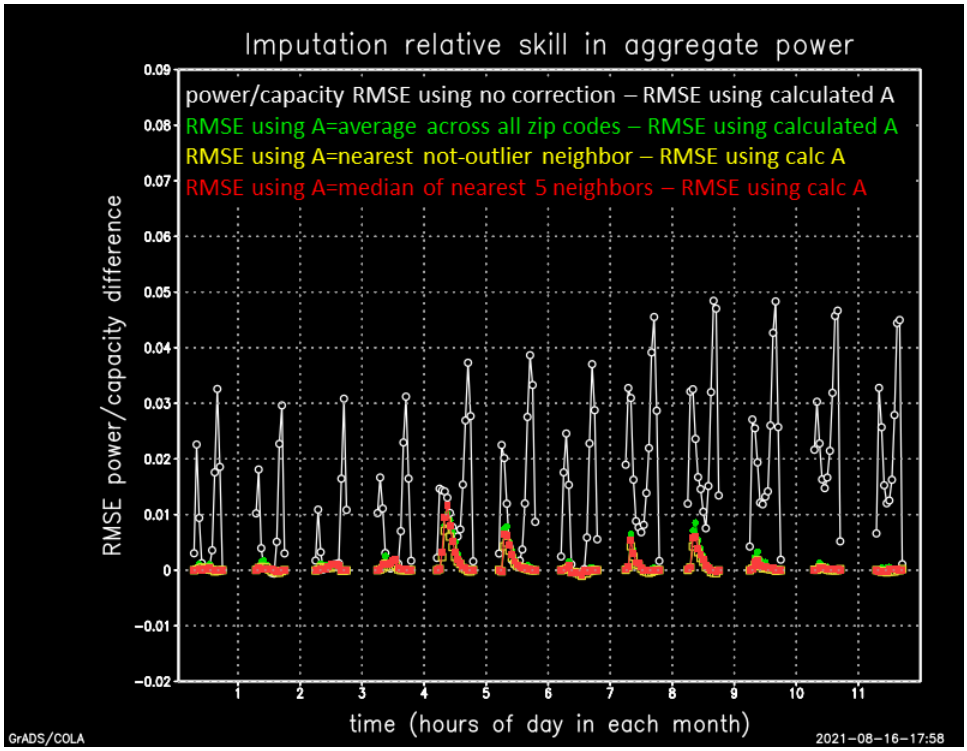


Figure 70. As in Figure 69 except for the aggregate power of 86 zip codes rather than for one zip code

Based on these results, method (1) was selected to have the most spatial fidelity during the volatile spring months when marine stratocumulus clouds have the most impact and the most day-to-day variability, although for the aggregate, it might not make much difference which method is used as long as bias corrections are made by month and time of day.

G. Merged Observation-Based Data Set

The sample-based data for the 86 zip codes were assumed to be the best estimate, and the other 35 zip codes used the NSRDB-based estimates bias-corrected using imputation method (1) for the *A* ratios for each hour of the day for each month of the year. The merged observation data set using the metered sample changes, and presumably improves, the NSRDB-based estimates of the aggregate BTM generation with RMS differences of 6% of capacity in May and up to 8% in the afternoons in September (Figure 71). RMS differences are more than twice as large in individual zip codes on average, including those with the most installed capacity, and even much larger than that in some zip codes.

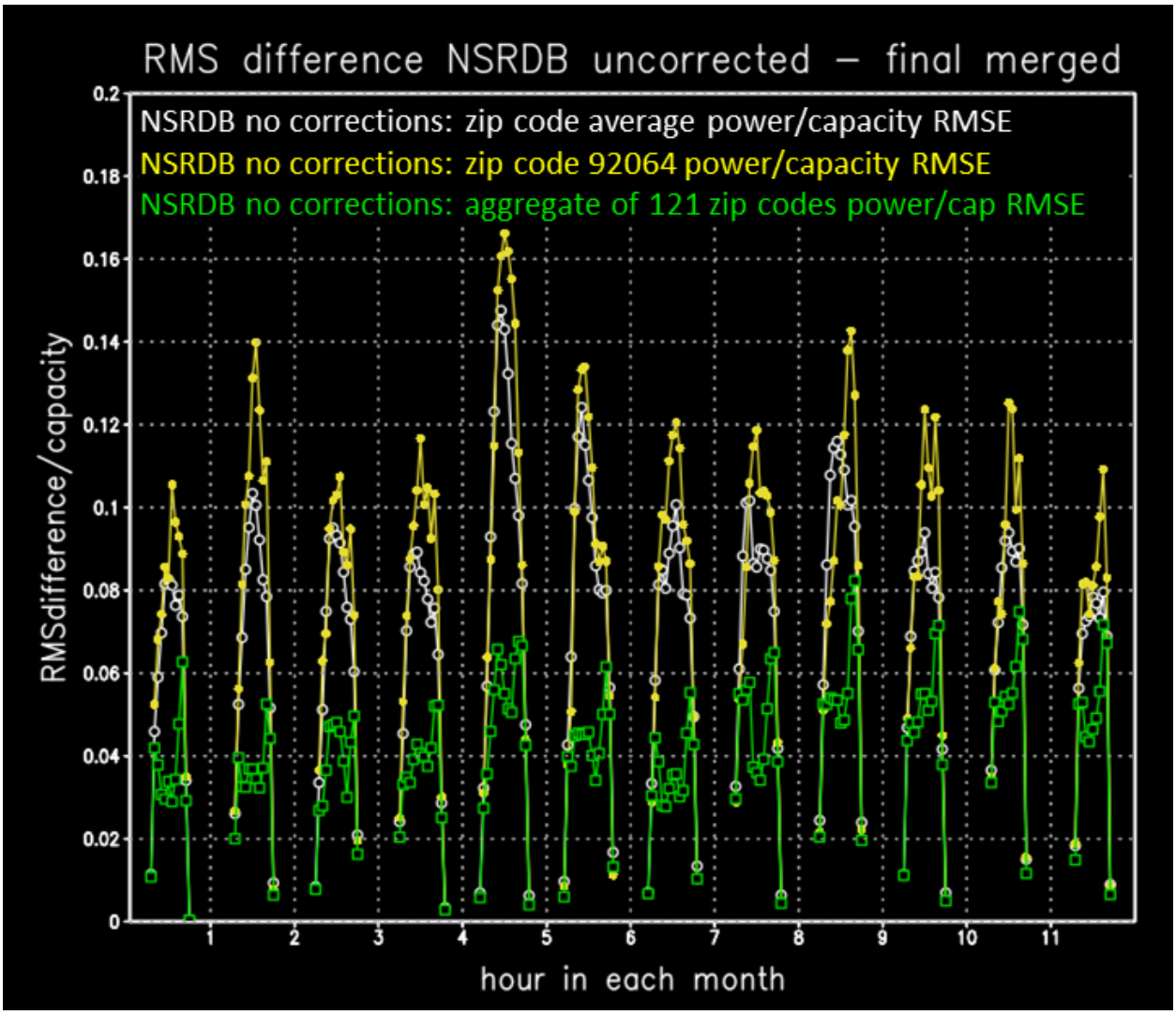


Figure 71. RMS differences between the final merged BTM generation estimate and the uncorrected NSRDB-based estimate for the aggregate (green), the average of individual zip codes (white), and the largest zip code (yellow) as a fraction of capacity

Forecasts and Bias Correction

Two types of forecasts are considered:

1. NWP model forecasts
2. "Smart persistence" forecasts.

These forecasts are for lead times matching the load forecast lead times. NREL stated that the publicly available historical CAISO load forecasts for the SDG&E load zone are found to be for rolling 15-minute periods beginning 45 minutes after the forecast time. The forecasts here are created to match that timing.

NWP Model Data Preparation

The NWP forecasts are bias-corrected similarly to the NSRDB-based estimates of actual BTM generation, except the forecasts can only be corrected looking backward in time, to emulate what would be possible in a forecast environment.

Maxar created historical superensemble sets of direct normal and diffuse irradiance forecasts for each of the 121 zip code centroids for 5-minute time averages for lead times to 2 hours for the 2-year period of interest for 116 NWP models. These were created by the Maxar solar forecast system, which sharpens the time resolution of the NWP time averages by assuming the same clear-sky fraction for the longer time averages in the NWP model output and multiplying by the Maxar-calculated clear-sky irradiance based on the sun position every 1 minute, the weather-dependent and sun-angle-dependent aerosol scattering, the water vapor and ozone absorption, among other factors. The Maxar system parses the GHI into a direct beam and diffuse part, which follows the REST2 clear-sky model under clear conditions and deviates from that based on calculations derived from years of Surface Radiation Budget (SURFRAD) observations. Stochastic perturbations conditional on the clear-sky fraction are added separately to each superensemble member to reproduce temporal variability consistent with the predicted weather.

The forecast ensemble sets were assembled based on the availability of the models considering their latency, which arises as the NWP forecast systems wait for observation collection; perform the computationally intensive data assimilation to create their initial conditions; and then integrate the equations of motion for a fluid on a rotating sphere along with the change of water phase physics, the surface layer energy exchange, the radiative transfer, the turbulence, the cloud microphysics, and other physical processes; and then the output is run through postprocessing routines to convert it to the output parameters on the output grid. Latency varies by model set from approximately 1–2 hours to more than 8 hours. The forecast ensembles provided used the following models, in this order of appearing in the data set (thus, the plots are in this order):

- 1–3 = NOAA Global Forecast System (GFS), last 3 runs
- 4–24 = 21 ensemble members from the latest run of the NOAA Global Ensemble Forecast System (GEFS)
- 25–27 = ECMWF, last 3 runs
- 28–78 = 51 ensemble members from the latest run of the ECMWF ensemble forecasts
- 79–81 = NOAA NAM, last 3 runs
- 82–96 = NOAA RAP, last 15 runs (new forecasts every 1 hour)
- 97–99 = Canadian regional high-resolution model, last 3 runs
- 100–114 = NOAA HRRR, last 15 runs (new forecasts every 1 hour)
- 115–116 = Canadian global model, last 2 runs (new forecasts every 12 hours).

NREL created 15-minute time-averaged model-based generation for each of the 121 zip codes for the desired forecast period by averaging the three 5-minute time-averaged irradiances provided by Maxar and then passing them through SAM using the same metadata and assumptions as for the NSRDB irradiances.

The time-shift check was performed on the 116-member ensemble mean forecasts of the BTM generation in the same manner as for the NSRDB-based estimates. It revealed that the data times

were consistent with the sample-based BTM generation estimate times; thus, there was no need to time shift.

The data passed back to Maxar had the model sequence out of order. If each superensemble member were to be bias-corrected separately from the members, then that does not matter; however, it was of interest to bias correct using the same corrections for all like members, yielding 9 sets of bias correction factors instead of 116 because the forecast bias characteristics should be the same within each set.

To untangle the model sequence, the following diagnostics were run for the model intercomparisons:

- Bias ratio *A* was computed for each hour of the day for each month of the year as for the NSRDB-based estimates but using the NSRDB-based estimates as the "actuals" because they had the same SAM assumptions as the model forecasts
- Model correlations (e.g., model A versus model B for A = 1 to 116 and B = 1 to 116) of the time series (hour, month) of the *A* ratio for each zip code, then averaged across all zip codes (compares bias characteristics between each pair of models)
- Same but for aggregate instead of zip codes
- RMS difference in the *A* ratios between models
- RMS difference in the power time series for the models

Most discerning was the model matrix of the RMS difference in aggregate power, Figure 72, and after untangling to restore the order, Figure 73. The colors are the RMS differences between the models on the horizontal and vertical axes, with the lower values indicating a closer relationship (the differences are smaller). The diagonal is black because I omitted plotting correlations of 1.0 for each model against itself. The white lines identify the model sets as listed above, so the lowest values should be aligned inside the white boxes, and for the HRRR and RAP time lag ensemble sets, the lowest values are around the diagonals.

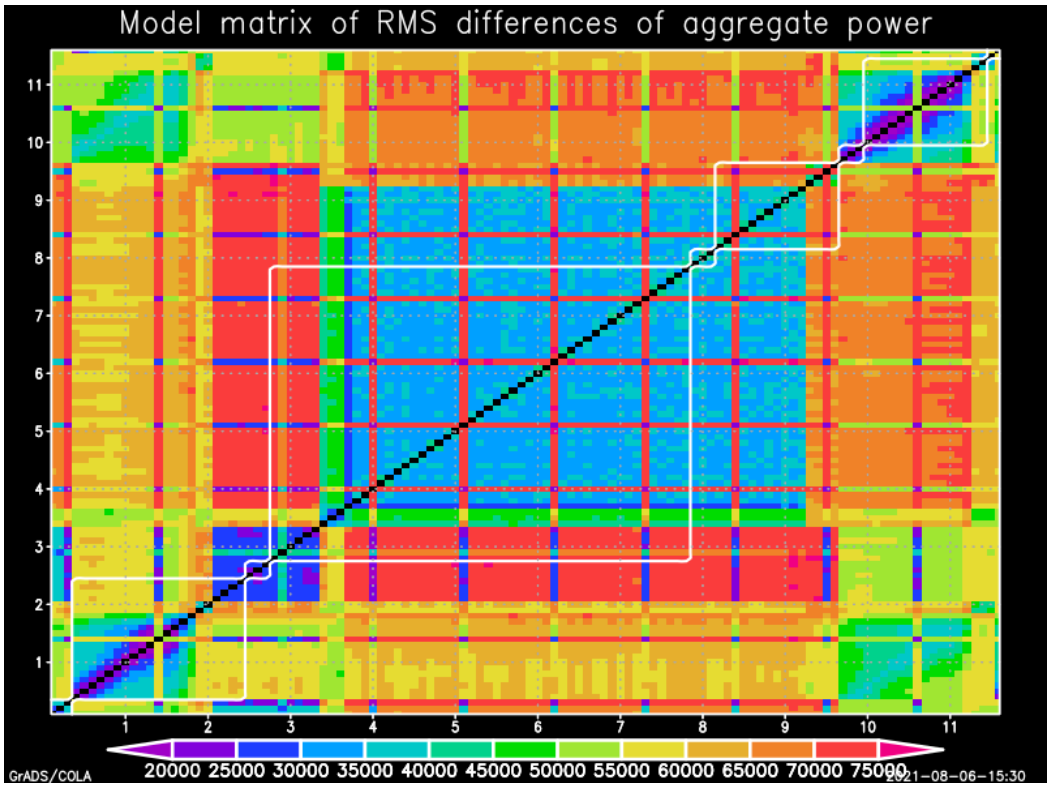


Figure 72. Model matrix of differences as described in the text, model order in the files from NREL

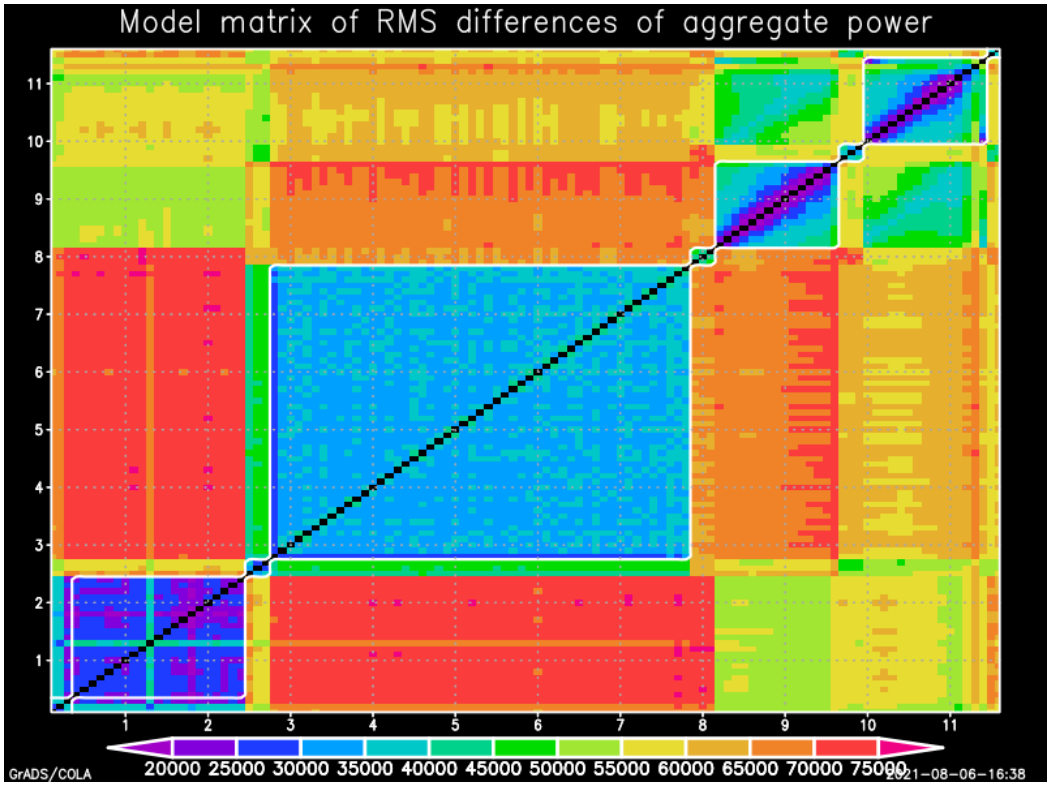


Figure 73. Same as Figure 72 but model order as restored to the list in the text

Model Bias Correction

The model bias correction is similar to the NSRDB bias correction except the collection period is all in the past and covers a wider time-of-day window. Each 15-minute valid time period uses the bias data collection over the preceding and following four 15-minute periods plus the same 15-minute period from the most recent N previous days. For example, the July 15 forecast for the interval from 18:00 UTC–18:15 UTC would collect data from 15-minute periods from 17:00 UTC–17:15 UTC, 17:15 UTC–17:30 UTC, … , 19:00 UTC–19:15 UTC on N days, the last of which is July 14. The corrected forecast is the sum of the merged observation data set for those times and dates divided by the model-based estimate for the same collection of times and dates.

The model bias correction is done separately for each model or model set.

Three variations of methods were evaluated:

1. Various look-back lengths—N = 2, 3, 5, 10, 20, 30, and 60 days—and no corrections
2. Make collection sums separately for each individual superensemble member or make a smaller number of large collections containing all members of the same model set.
3. Make corrections separately for each zip code and make the aggregate corrected values as the sum of the zip code corrected values or make the corrections for the aggregate directly based on the aggregate itself without separately correcting by zip code (then the corrected aggregate might not exactly equal the sum of the corrected zip codes).

Evaluations showed that the differences are small for all combinations involving N = 10, 20, 30, and 60. Figure 74 shows the RMSE as a fraction of capacity averaged over all 116 models as a function of N. It shows the mean of the RMSE, not the RMSE of the ensemble mean. The upper curves are for the average RMSE for the individual zip codes, and the lower are for the aggregate of all 121 zip codes, indicating that the forecasts for the aggregate have a better RMSE by more than 3% of capacity. In all cases, these summary statistics show a higher error for no corrections than for any of the correction strategies for any number of days, N. The two zip code-level curves show comparison number (2), indicating slightly better skill (smaller RMSE) when using the same corrections for all members in the same model set (e.g., all ECMWF ensemble members use the same corrections as each other). That also holds true for the aggregate, but the differences are so small that they are hard to see. The aggregate also has very slightly better results for the direct bias correction rather than as a sum of the bias-corrected zip codes. Figure 75 shows how that breaks down by model for the individual zip codes, and Figure 76 shows that for the aggregate. The lines for 10, 20, 30, and 60 days are on top of each other and are best for all models, although the amount of improvement over 2, 3, or 5 days or no correction varies by model and is less for the aggregate than for the individual zip codes. Those are averaged across all hours and months. Figures 76 and 77 show the same by hour and month, averaged across models. They show the familiar pattern of the largest errors toward the middle of the day, although the peaks in the highest error months of May, June, and September are in late morning.

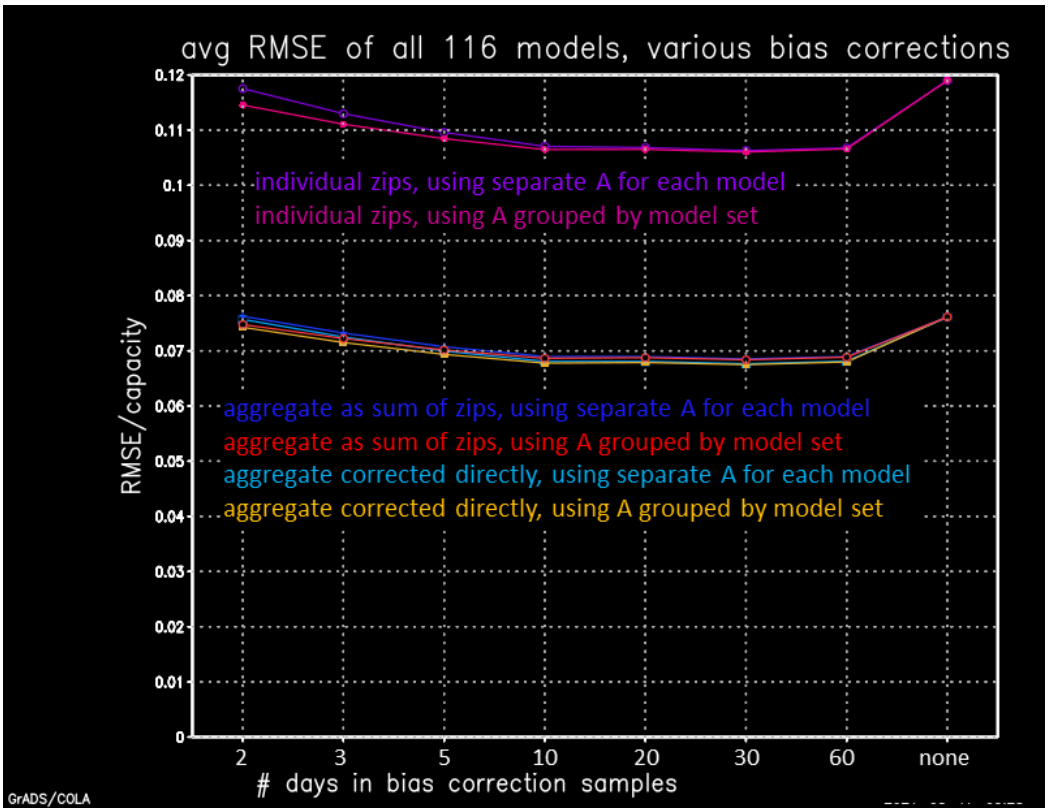


Figure 74. Average of RMSE for all 116 models using various bias-correction strategies

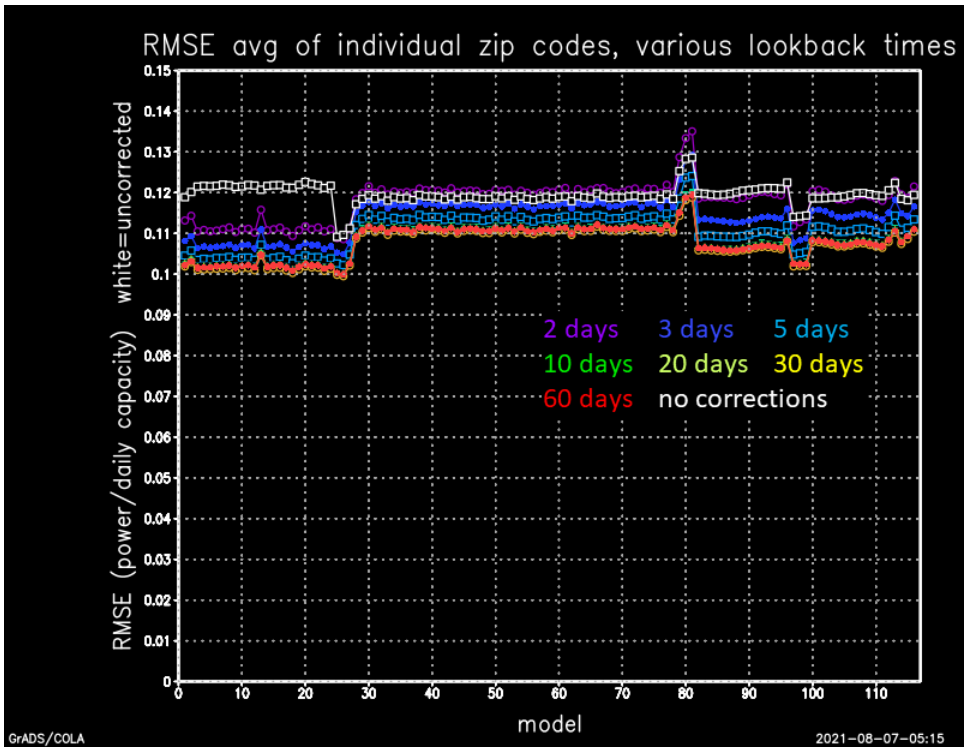


Figure 75. RMSE averaged over hours, months, and individual zip codes as a function of the model number for various N days of collections used in the bias correction

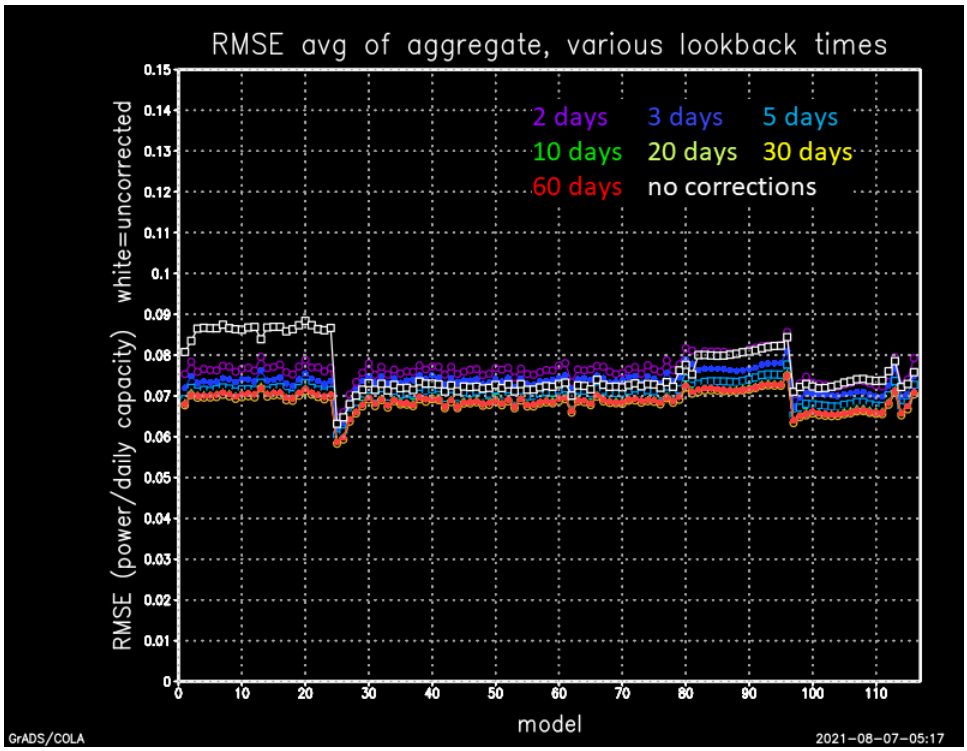


Figure 76. As in Figure 75 except for the RMSE of the 121 zip code aggregate

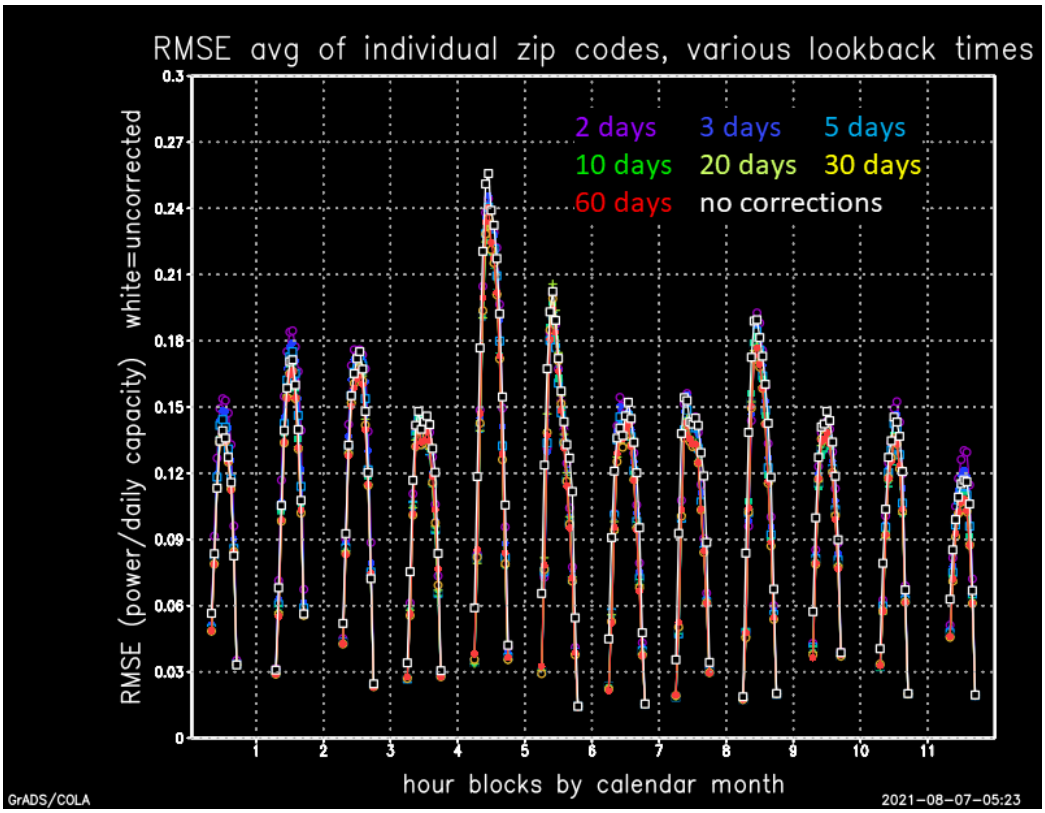


Figure 77. RMSE averaged across models and zip codes for each hour in each month for various N

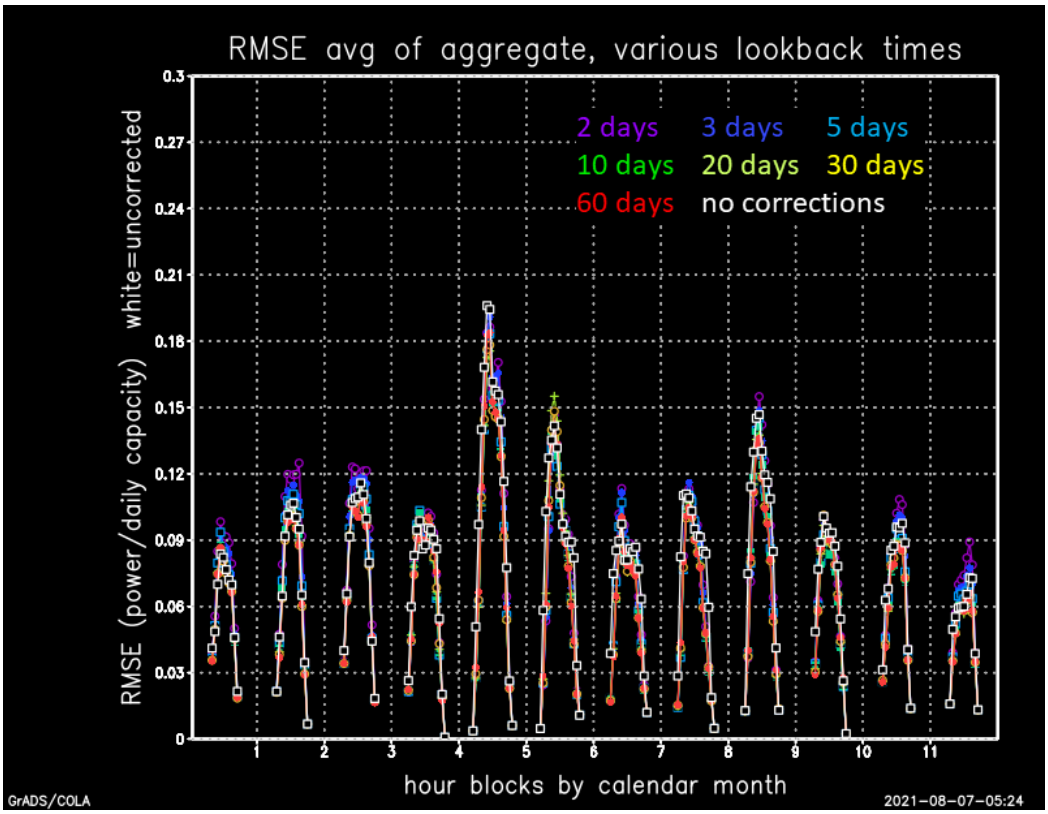


Figure 78. RMSE of aggregate averaged across models for each hour in each month for various N

Viewing the model bias by hour and month, shown in Figure 70, reveals more nuanced issues that can inform the decision on the best number of days, N, to use for the bias correction. Without the bias correction, the models have a large bias in May all day. This bias is still high in late morning with a large N because it was not present in April, so a long look-back period will wash out the signal coming from May. September also has a large high bias in the models, but any look-back up to 30 days is still good for September because it is not that different from August. Indeed, all the months from May through November have a large bias spike in the morning, which is well taken care of by the bias correction for most look-back periods, except for May, which has long look-backs. Because May appears to be a critical time of high forecast risk, and because the best overall forecasts for all months combined was earlier shown to be about the same using look-back periods of 10, 20, 30, and 60 days, it looks like 10 days is the best choice.

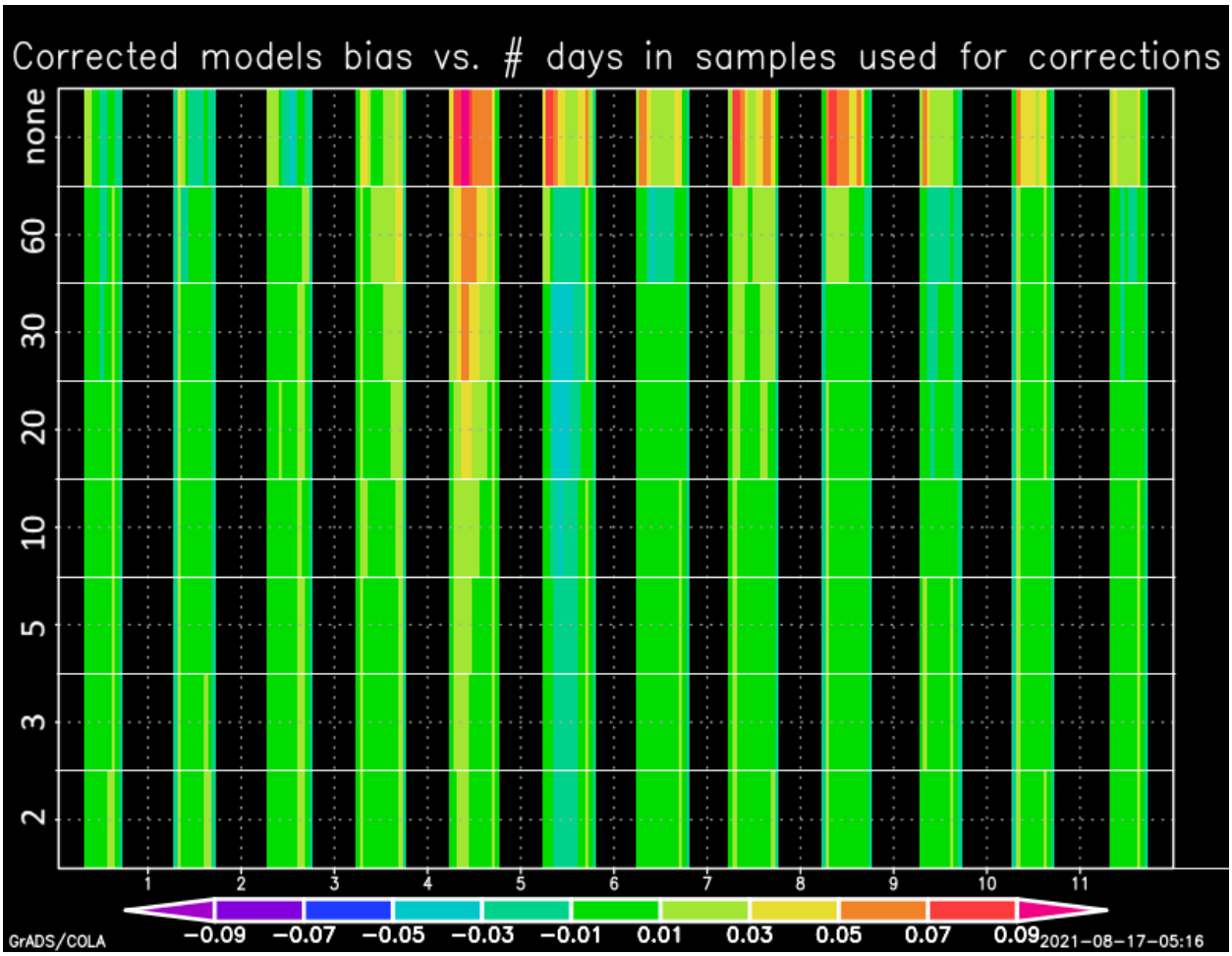


Figure 79. Model bias averaged across model sets and zip codes for each hour of the day in each month for various values of number of days, N, in correction calculations

The model bias corrections were performed with N = 10 days using the same correction factors for all models in a set for the nine sets listed above. Corrections were performed for each zip code, and the aggregate was formed from the corrected zip code forecasts. The resulting skill levels are shown in Figure 80. The improvement over the uncorrected model forecasts has spikes of the largest improvement during the morning and evening ramp times for many months.

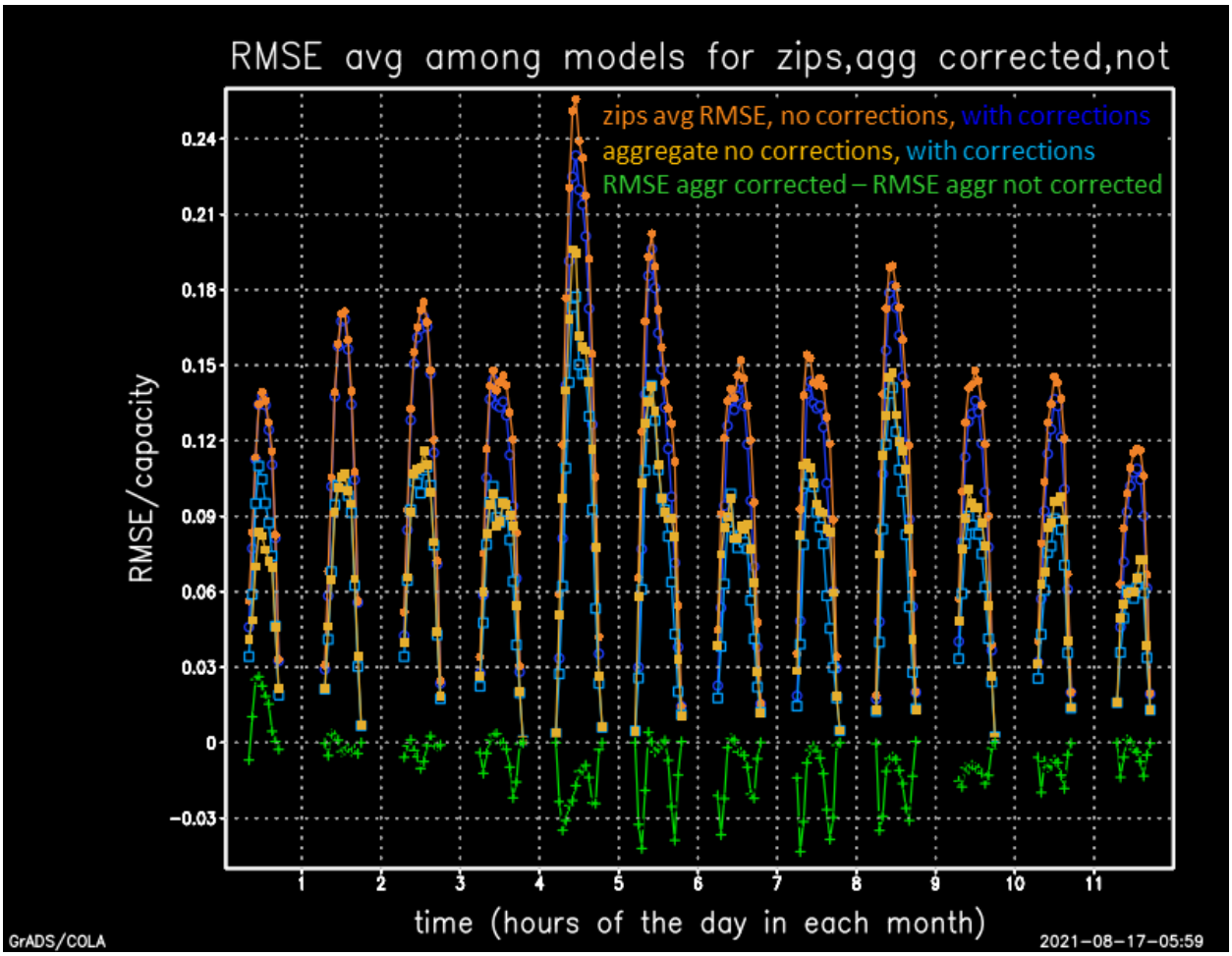


Figure 80. RMSE/capacity averaged over the 116 models for the aggregate and averaged over the 121 zip codes with and without bias correction and the improvement (green, negative is improvement) for the aggregate versus no correction

Smart Persistence

Smart persistence represents an extrapolation of current conditions, which can be useful for a short-term forecast and can complement NWP model forecasts that consider the expected evolution of the atmosphere and associated irradiance but do not have current information about the PV system. A collection of recent observations can be extrapolated to yield a collection of smart persistence forecasts that can complement a collection of NWP-based forecasts to yield a probabilistic forecast. The key to making persistence "smart" is following the expected time dependence of the system. The way we do that here is simplistic: It follows a diurnal curve based on the daily shape of the maximum at each time interval over the last 30 days and then assumes that today's observed fraction of that remains constant. To avoid excessive values, we bound the ratio *f*:

- $f = \min(1, \text{obs(today at time t)}/\text{obsmax30(time t)})$
- smart persistence = f * obsmax30(forecast valid time).

where *obsmax30* is the maximum generation in the zip code over the previous 30 days for each 15-minute interval and is taken for the same time of day as the observation in the equation for *f*,

and then it is taken for the forecast valid time in the equation for the smart persistence. The smart persistence for the aggregate is the sum of the smart persistence for each of the 121 zip codes. The "observation" uses the merged validation data set based on upscaled samples for 86 zip codes and bias-corrected NSRDB-based estimates for the other 35 zip codes.

Because we are interested in a forecast for the fourth 15-minute interval after the present and because we might not have the latest observation available in time for the forecast, the last observation available would be five 15-minute intervals before the valid time, so we extrapolate the observation ahead by five 15-minute intervals. To generate a collection, we can extrapolate ahead six intervals, seven intervals, etc., but not too many because performance will quickly degrade as the observation gets too old. For this evaluation effort, I prepared a collection of eight smart persistence forecasts. For example, if the forecast was issued at 20:00 UTC, it needs to be valid for the time interval from 20:45 UTC–21:00 UTC, and the most recently available observation time to deliver that forecast would have been from the 19:30 UTC–19:45 UTC interval. So, we go back 2 hours to make a collection of 8, using ending times of 17:45, 18:00, 18:15, 18:30, 18:45, 19:00, 19:15, and 19:30.

The accuracy of the smart persistence is poor when the observation is based on a low sun angle with cos(solar zenith angle) smaller than approximately 0.1, which corresponds to roughly the first and last hour of the day. To make matters simple, the first and last hours of daylight are not included. The effect of omitting these is shown in Figure 81 and Figure 82, depicting the eligible smart persistence forecasts of the youngest and oldest ages in our collection of eight. The first usable smart persistence in the morning, as it ages to become the eighth in the collection, is now being used for a time 3 hours later than its observation, which is 4 hours after sunrise.

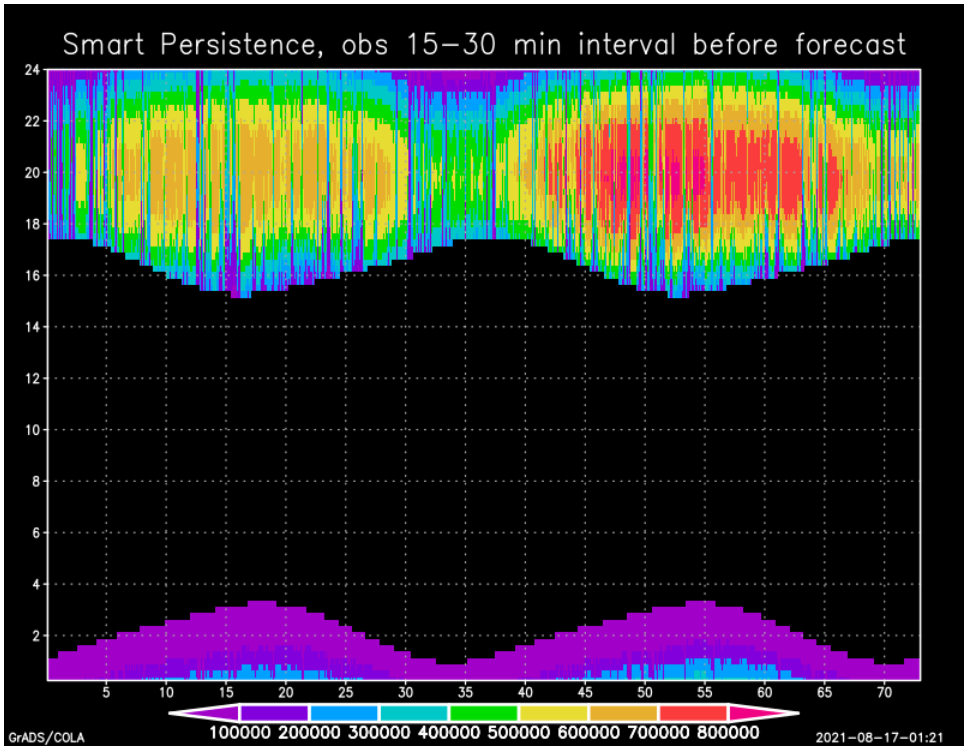


Figure 81. Smart persistence forecast of the 121 zip code aggregate from latest available observation, omitting observations at times of low sun angle

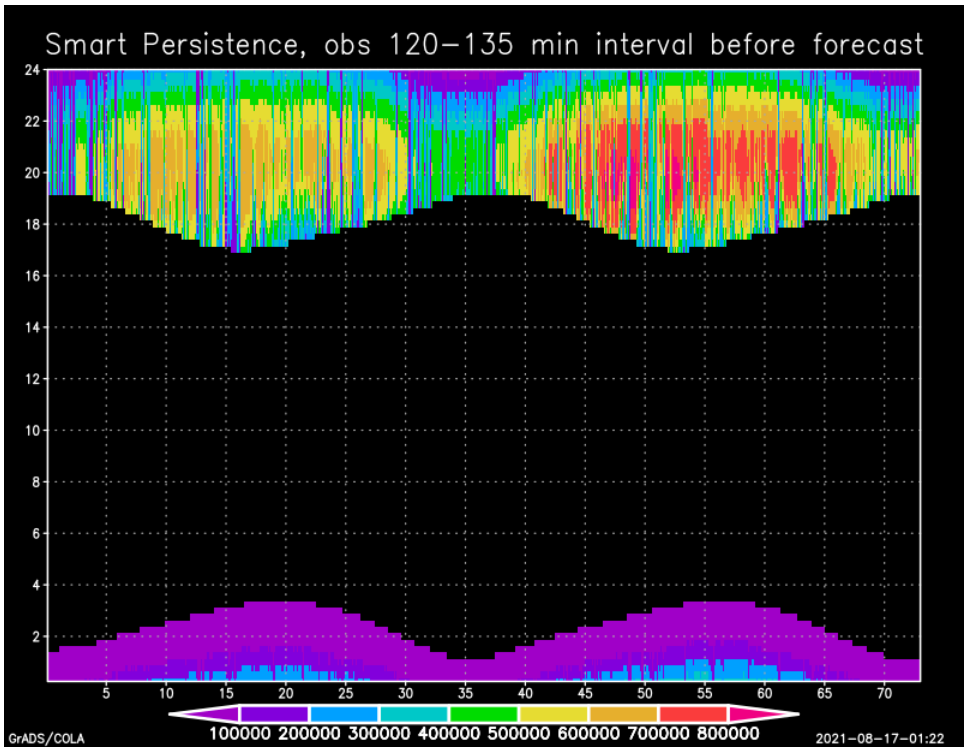


Figure 82. Smart persistence forecast of the 121 zip code aggregate from the eighth latest available observation, omitting observations at times of low sun angle, which means it becomes available later in the morning than the one shown in Figure 81

My smart persistence is too simple and not very smart, resulting in a substantial late-morning low bias and afternoon small high bias, as shown in Figure 83; however, we can make interesting combinations of them and combine them with model forecasts. I tried using the most recent four smart persistence forecasts to make a trend by subtracting the average of the third and fourth from the average of the most recent and second most recent, and I added that trend to the most recent. Then I combined this with the NWP-based forecasts by averaging this result with an unweighted average of the 116-member NWP ensemble collection. Figure 84 shows the daily cycle of the MAE of the aggregate across all 730 days for the individual models, the unweighted ensemble average, and the combined trended smart persistence and ensemble average. The combined forecast reduces the ensemble mean MAE by more than 30%, and the error decreases throughout the day. Unfortunately, it is not available early in the day, when the error steeply ramps up.

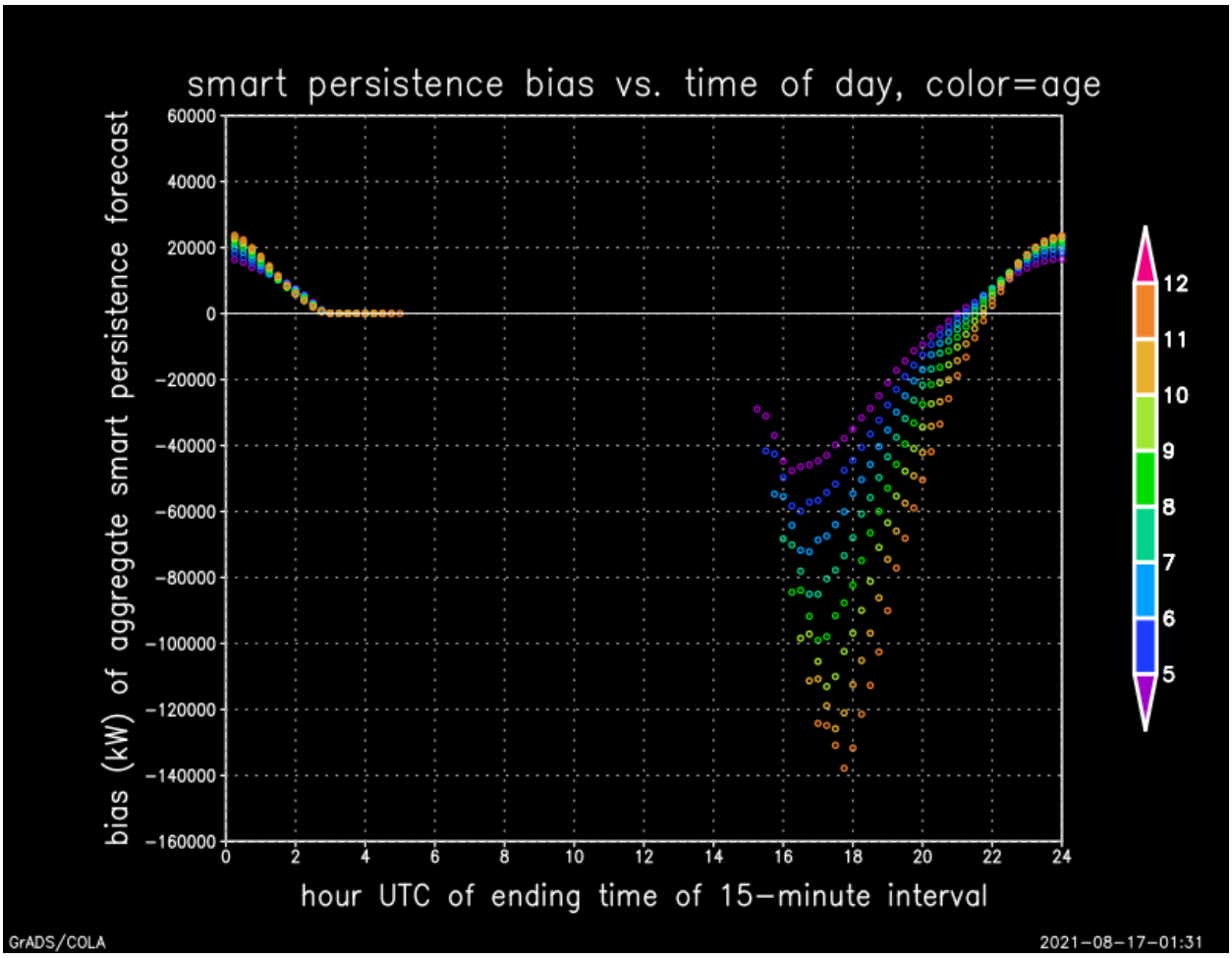


Figure 83. Bias of aggregate smart persistence for the set of 8 with valid times from 5 to 12 intervals after the observation time

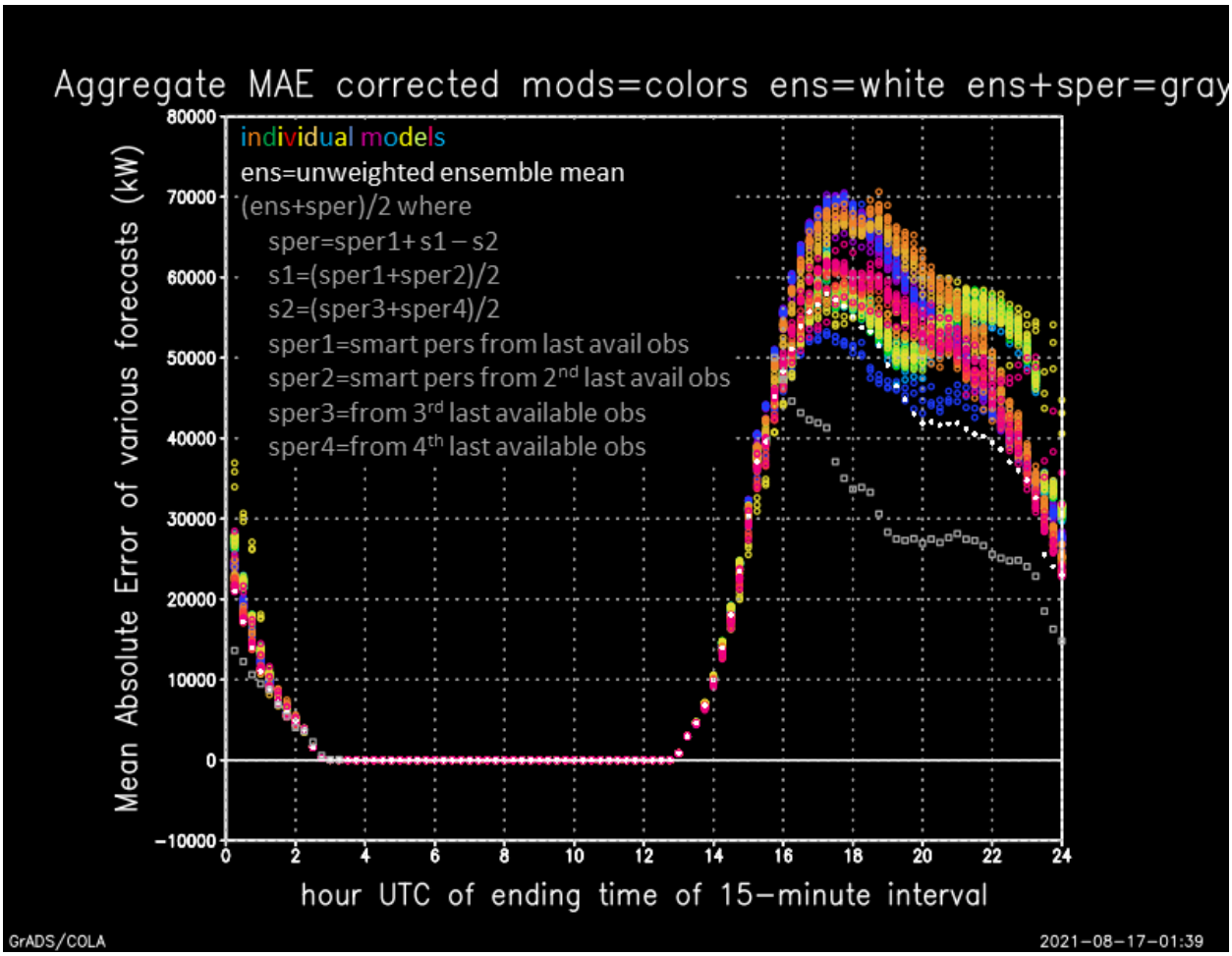


Figure 84. Mean absolute error by 15-minute block of the day for the 116 individual NWP model-based BTM aggregates over 121 zip codes (colors), the unweighted ensemble mean (white), and the combination of ensemble mean and trended smart persistence (gray)

Conclusion

BTM metered samples comprising approximately 6%–8% of the aggregate load zone BTM generation while sampling from zip codes containing approximately 85%–95% of the aggregate load zone BTM generation were acquired, analyzed, and used in producing estimates of the load zone aggregate BTM generation of the SDG&E load zone in CAISO. These BTM generation estimates were further used to bias-correct NWP model forecasts and to generate smart persistence forecasts.

The BTM metered samples appear to have improved over the NSRDB-estimated BTM load zone aggregate generation estimates in the morning and evening ramp periods by an approximate 6% RMS difference as a fraction of capacity all day in May and during the late afternoon ramp-down times in most other months.

The BTM metered samples as a fraction of capacity in each zip code exhibited spatial coherence and spatial structure showing lower capacity fractions at times and locations more exposed to marine stratocumulus incursions. The comparisons against the NSRDB-based estimates

primarily show time-of-day variations, likely due to limitations in using SAM to convert the NSRDB-based irradiance to power—specifically, not being able to account for shading. This is a hypothesis based on the largest bias ratio differences from 1.0 at low sun angles, but there are days when the discrepancies occur in the middle of the day, so there might also be limitations in the NSRDB irradiances contributing as well, or other factors that are not examined.

Cross-validation results show that several methods can be skillful for improving the NSRDB-based estimates to fill gaps in the coverage of the metered samples. For the load zone aggregate, although it is not for individual zip codes, it is just as effective to use a spatially uniform bias ratio to adjust the NSRDB-based estimates, rather than using a more localized, spatially varying approach.

The merged best BTM generation estimate was also used for forecast improvements, focusing on short lead time forecasts of the 45- to 60-minute period after a power forecast is issued. This lead time was used to match the availability of the load forecast data for the load zone examined here. Note that the lead time for the NWP models used as input into a power forecast is actually hours longer because the models are run only a while after the initial time to ingest observations of the atmosphere, and then there is computing time to assimilate the observations, integrate the forecast, generate the output, and deliver the output.

The merged best BTM generation estimate was used to bias-correct forecasts. Again, multiple methods were tested, and comparisons found not much overall difference between them. Bias-correcting using look-back periods of 10 days worked best in volatile periods of changing cloud regime when longer look-back periods were misdirected, but otherwise longer look-back periods performed as good or better. For NWP ensemble sets in which the ensemble was created by running the same NWP model configuration with different initial conditions, the error characteristics of all members should be the same. Combining the power forecasts based on all of those members into a single sample for calculating the bias correction and then applying the same bias correction to each of those members produced a small improvement over treating each separately. Bias-correcting the aggregate directly, e.g., relating the forecasts of aggregate generation to the estimated actual aggregate generation, yielded small improvement over bias-correcting the forecasts by zip code and then making the aggregate forecast from the sum of the bias-corrected forecasts for the generation in the individual zip codes; however, that small improvement comes with allowing the forecast of the aggregate generation to differ from the sum of the forecasts of the generation in the individual zip codes.

Even bias-corrected, the individual models had RMS differences from the merged best estimates of approximately 6%–10% of capacity for the aggregate during the middle of the day, higher in May, and approximately or exceeding 15% during the middle of the day for the individual zip codes.

The merged best BTM generation estimate was also used to create a collection of smart persistence forecasts that might, in combination with the NWP superensemble, be able to improve forecast probability distributions compared to those based on the NWP ensemble alone. Figure 84 shows that for a deterministic forecast, an unweighted ensemble mean of bias-corrected NWP members does as well or better than the best-performing individual members,

and when combined with a linear combination of the smart persistence forecasts, the error is reduced by approximately one-third compared to the ensemble mean itself.

2 Research Area 2: Adaptive Reserves

Introduction

Power systems are subject to inherent uncertainty, mostly due to the variability of system load as well as the uncertainty associated with generation. Even though forecasting techniques can partially mitigate the uncertainty, forecasting errors cannot be avoided, and system components will always fail; therefore, power system operators maintain ancillary services through operating reserve to ensure the reliability of the grid. This operating reserve is provided to compensate for a number of different factors, including regulation, forecasting errors, equipment outages, and local area protection.

Different power markets offer their own ancillary services, where definitions and requirements of reserve are distinctive (Zhou, Levin, and Conzelmann 2016). For example, ERCOT requires generation units that provide spinning reserve to respond within a shorter time (within the first few minutes) than other regional transmission operators (RTOs) or independent system operators (ISOs). The Midcontinent Independent System Operator (MISO) provides two separate products for both spinning reserve and non-spinning reserve (NSR), which is different from most other ISOs/RTOs. Generally, all ISOs/RTOs maintain at least three types of reserves: regulation reserve, spinning reserve, and non-spinning reserve. Regulation reserve is usually used to correct the system area control error, whereas the others are used for contingency events. Compared to the first two categories of reserves, non-spinning reserve is less studied in the literature. Traditionally, the non-spinning reserve requirement (NSRR) is often assumed to be the capacity of the largest generator (Wood, Wollenberg, and Sheblé 2013). This simple method fails to reflect the increased variability and uncertainty owing to the addition of non-dispatchable resources. With the increasing penetration of renewable energy in power systems, the reserve scheduling problem becomes more challenging. Most ISOs/RTOs estimate reserve empirically from historical data. For example, CAISO's requirement for contingency reserves is equal to the largest credible contingency or 5% of the load served by hydropower generation and 7% of the load served by thermal generation (CAISO 2019). ISO New England requires the sum of 10-minute spinning reserve and 10-minute non-spinning reserve to be at least equal to the capacity of the largest single system contingency multiplied by a contingency reserve adjustment factor in the most recent operating quarter (ISO-NE 2019). In contrast, the contingency reserve requirement in the MISO is set to 2,000 MW, of which approximately 50% is spinning reserve, and the rest is supplemental reserve (MISO 2013).

Most reserve requirement determination algorithms are performed months, or even years, ahead of the operating time period, and thus fail to consider information about the potential uncertainty that is available closer to the operating hour. Recently, several works have improved the reserve requirement determination by considering renewable power uncertainty, load forecasting errors, and control area imbalance in the historical data (Maurer, Krahl, and Weber 2009). The consideration of system uncertainty (in terms of renewable generation, load, or net load) as one random variable neglects the heterogeneity of randomness over time. For example, Wang et al. (2015) developed a distributionally robust optimization method that determined the reserve by accounting for the wind power forecasting error distribution; however, the method exhausted all the error distribution scenarios, which might not be located between the different types of wind power distributions that could occur at the forecasting timescale in real-world operations.

Similarly, in Ortega-Vazquez and Kirschen (2008), the net load uncertainty information was quantified from the historical forecasts, which was then used to estimate the spinning reserve requirements. Recently, probabilistic methods have increasingly been used by system operators to account for wind or solar uncertainties (Matos and Bessa 2011; Ortega-Vazquez and Kirschen 2009; Holttinen et al. 2012). Although these methods can reduce the amount of reserve, their impacts on system economics and reliability must be examined in a real-world power market setting, which is usually conducted by using market simulation models (Cui et al. 2017; Cui and Zhang 2018; Wu et al. 2017).

In this paper, a data-driven method is developed to determine the adaptive NSRR in ERCOT. The method balances the practical possibilities in daily operations and the utilization of an advanced uncertainty quantification method. Specifically, we follow the ERCOT NSRR determination procedure and estimate the adaptive NSRR based on probabilistic net load forecasting. Three dimensions of adaptiveness are introduced by the method to dynamically update the NSRR by using more recent forecasts with updated uncertainty information. To validate the performance of the proposed adaptive NSRR estimation method, a set of probabilistic net load forecasts is first generated by state-of-the art physics-informed machine learning methods following the real-world ERCOT market requirements. Second, the adaptive NSRR is calculated from the probabilistic net load forecasts and evaluated statistically. Third, a multi-timescale market simulation model is used to accurately simulate the operation of thermal generators over varying timescales, and its applicability is demonstrated using a synthetic network that is built on the footprint of the ERCOT power grid. Finally, we examine the economic impacts of different NSRR levels and reveal the trade-off between the flexibility of online and offline units.

Methodology

In this section, we first introduce the ERCOT NSRR method as a baseline for comparing current techniques to those developed in this work. Then, the overall framework of the adaptive NSRR determination method is developed, followed by the description of key components of the method, including net load forecasting and adaptive percentile thresholding.

The ERCOT NSRR

ERCOT determines ancillary service requirements annually and posts them by December 20 for the entire coming year (i.e., 1 year ahead). The requirements can be updated 1 day ahead or even closer to real time, if needed. In the ERCOT system, the non-spinning reserve service consists of generation resources that are capable of being ramped to a specified output level within 30 minutes or load resources that are capable of being interrupted within 30 minutes and that are capable of running (or being interrupted) at a specified output level for at least 1 hour (ERCOT 2019).

The daily non-spinning reserve profile has six distinctive values, each of which indicates the requirement for a 4-hour block (e.g., 2 a.m.–6 a.m.). Data that fall within the time block from the same month of the previous 3 years are used to determine one specific value, which also means that the daily profile is consistent within every month. Specifically, the hourly *net load uncertainty* data are used to determine the NSRR. The net load uncertainty is defined as:

$$\bar{y}_{nl} \coloneqq |\hat{y}_{nl} - y_{nl}| \qquad (1)$$

where $\hat{y}_{nl}$ and y_{nl} indicate the net load forecast and the actual net load vectors, respectively.

The net load and net load forecast are defined as: $y_{nl} := y_l - y_w - y_s$ and $\hat{y}_{nl} := \hat{y}_l - \hat{y}_w - \hat{y}_s$, where y_l, y_w, and y_s are the actual ERCOT system-wide load, wind power generation, and solar power generation, respectively; and $\hat{y}_l$, $\hat{y}_w$, and $\hat{y}_s$ are forecasts of the ERCOT system-wide load, wind power generation, and solar power generation, respectively.

The percentile of net load uncertainty from the same month of the previous 3 years is set as the sum of the NSRR and the regulation-up reserve requirement for one 4-hour block per month:

$$RR_{nlr,h} \coloneqq P_{PC,\bar{y}_{nl,h}} = f^{-1}_{\bar{y}_{nl,h}}(PC) \qquad (2)$$

where $P_{PC,\bar{y}_{nl,h}}$ is the PCth percentile of the net load uncertainty; PC ranges from 70–95 based on the hourly net load ramp in ERCOT; f is the probability density function; and h is the hour index.

Finally, the NSRR is calculated by subtracting the average regulation-up reserve of the same 4-hour block:

$$RR_{nsp,h} = RR_{nlr,h} - RR_{regup,h} \qquad (3)$$

where RR_{regup} is the regulation-up reserve.

There are a few areas where the current ERCOT method could potentially be improved: (1) The current method lacks an updating mechanism. Even though the reserve profiles can be adjusted, the decision relies solely on the experience of the operator. (2) The timeline of the reserve determination is not consistent with the daily operation schedule. (3) The reserve levels are mainly based on historical data; thus, the current and future system uncertainty is not considered.

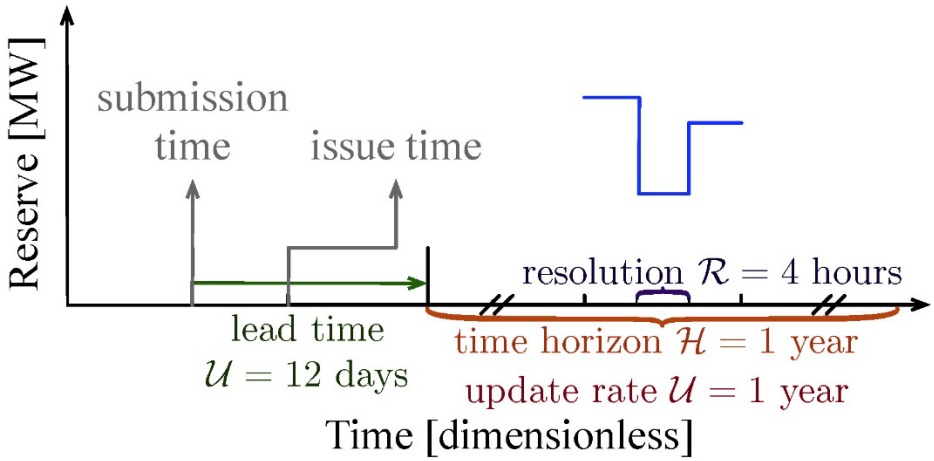


Figure 85. The reserve determination timeline

The Adaptive NSRR Algorithm Framework

To ensure the practical value, the developed method follows the procedure of determining the NSRRs in ERCOT's daily operations. Four time-related requirements need to be considered: lead time (L), time horizon (H), time resolution (R), and update rate (U) (Yang, Wu, and Kleissl 2019). As shown in Figure 85, the lead time indicates the difference between the reserve posting time and the time of the first reserve value. The time horizon is the span of the reserve profile generated at each issue time. The resolution is the interval between two distinctive reserve values. The update rate is the interval between two issue times. In current ERCOT practice, the lead time is 12 days, the time horizon is 1 year, the time resolution is 4 hours, and the update rate is 1 year. In this work, improvements of the developed method are made by introducing three-dimensional adaptiveness into the ERCOT procedure: (1) adaptive time resolution, (2) adaptive update rate/time horizon, and (3) adaptive percentile thresholds. In more detail, the developed method calculates hourly NSRR instead of every 4 hours because the 4-hour block requirement is not a technical constraint. Second, in addition to the 1-year-ahead scheduling for the entire year, NSRRs are adaptively adjusted 1 week ahead every week, 1 day-ahead every day, and 1 hour ahead every hour. In this procedure, probabilistic load, wind, and solar forecasts with three different time horizons are used to provide uncertainty information, which is used to adjust the percentile threshold.

The overall process of the developed method is shown in Figure 86. Based on the update rate/time horizon, the method contains four major steps, which provide NSRRs with (L12 days, H1 year, R1 hour, U1 year), (L1 week, H1 week, R1 hour, U1 week), (L1 day, H1 day, R1 hour, U1day), and (L1 hour, H1 hour, R1 hour, U1 hour). The procedure of Step 1 is similar to that described in Section II-A, except for the different resolution (i.e., 2 hour or 1 hour). In steps 2–4, probabilistic net load forecasting with different time requirements is used to predict the future system uncertainty. A novel two-step probabilistic forecasting method is adopted in these three steps that dynamically quantifies the net load uncertainty at every forecasting step. Finally, the percentile threshold is adaptively determined based on the probabilistic forecasts and is used to estimate the non-spinning reserve.

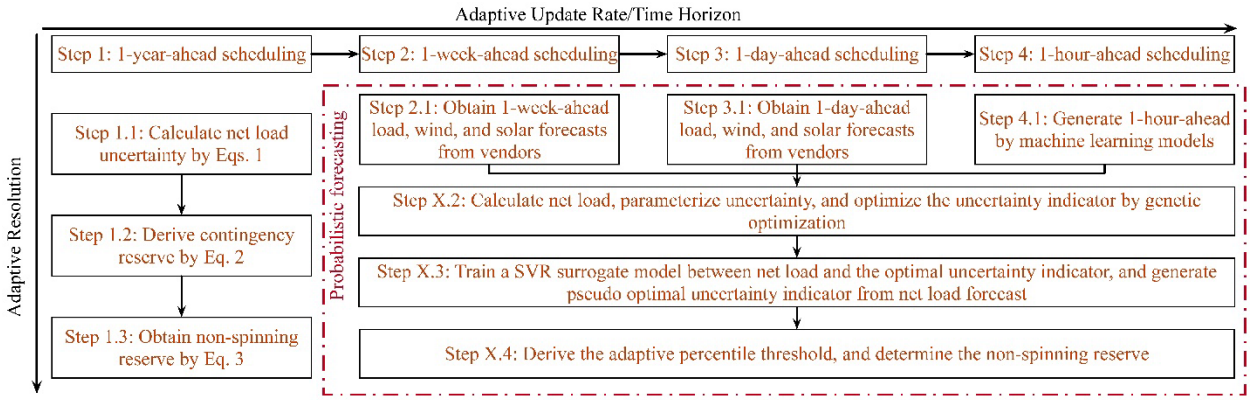


Figure 86. Overall framework of the developed method for NSRR estimation

Net Load Forecasting

The key component in steps 2–4 is net load forecasting, which measures the system uncertainty and therefore directly impacts the NSRRs. In the developed method, a two-step probabilistic

method is used to produce the probabilistic net load forecasts, which is shown in the red dashed box in Figure 86. In the first step, the deterministic net load forecasts are obtained from vendors or produced by a machine learning method (i.e., random forest) (Feng, Sun, and Zhang 2019). Specifically, the base estimators in the random forest are classification and regression trees: ntrees = 1,000, and mtry = 5. Then, the uncertainty is added to the deterministic forecasts by an optimal uncertainty indicator in the second step (steps X.3 and X.4). The procedure for probabilistic net load forecasting is described as follows:

1. *Step X.1*: Obtain load, wind power generation, and solar power generation from vendors for short-term forecasts or from the random forest forecasting model for very short-term forecasts.
2. *Step X.2*: Calculate the net load and parameterize the uncertainty of the net load in terms of μ and σ, where μ is assumed to be the net load value itself. Then, the quantile, q, and its corresponding pinball loss, L_q, are derived and expressed by q and σ:

$$F_t\left(y_{nl,t}\middle|\mu_t,\sigma_t\right) = F_t(\sigma_t) \qquad (4)$$

$$Q_t(p) = F_t^{-1}(p) = F_t^{-1}(p,\sigma_t) \qquad (5)$$

$$L_{q,t}(q,\sigma_t) = \{q - H\left(y_t - Q_t(q)\right)\}\{y_t - Q_t(q)\} \qquad (6)$$

where t is a time index, which means the predictive distribution differs at different forecasting times; $F(\cdot)$ and $F^{-1}(\cdot)$ are a cumulative distribution function and its corresponding inverse function, respectively; $Q(\cdot)$ is the quantile function; p and q are the probability and a quantile, respectively; and $H(\cdot)$ is the Heaviside step function. Next, the net load uncertainty indicator, σ (the only unknown parameter), at each forecasting time step is optimized by minimizing the average pinball loss of all quantiles with a genetic algorithm:

$$\sigma_t^* = \arg\min \frac{1}{Nq}\sum_{q=1}^{Nq} Lq,t(\sigma_t) \quad (7)$$

$$\text{s.t.} \quad \zeta_1 < \sigma_{q,t} < \zeta_2$$

where σ_t^* is the optimized standard deviation; $Nq = 99$ is the number of quantiles; and ζ_1 and ζ_2 are the lower and upper bounds of σ which are 0.001 and 1, respectively.

3. *Step X.3*: A support vector regression surrogate model, Ψ , is constructed to fit the actual net load and σ^* set $\{\mathbf{y}_{nl},\sigma^*\}$ in the training stage, which is used to generate unknown pseudo standard deviations, $\hat{\sigma}^*$, in the forecasting stage.

The random forest model and support vector regression model were empirically selected based on experience (Feng, Sun, and Zhang 2019; Feng et al. 2017; Sun et al. 2019). Note that the focus of this paper is to develop a NSRR determination method that mitigates the uncertainties associated with forecasting errors rather than reducing forecasting errors by building the most accurate forecasting model.

Adaptive Percentile Thresholding

Different from most probabilistic forecasting that takes the forecast target variable as a random variable, the developed probabilistic net load forecasting considers the forecast at each time stamp as a random variable; therefore, the pseudo standard deviation at each forecasting time stamp, σ_t^*, is an adaptive indicator of the future system uncertainty. The pseudo standard deviation is used to determine the adaptive percentile threshold:

$$PC_t = \frac{(\sigma_t^* - \sigma_{min}^*)(PC_{max} - PC_{min})}{\sigma_{max}^* - \sigma_{min}^*} + PC_{min} \quad (8)$$

where σ_{max}^* and σ_{min}^* are the maximal and minimal optimized standard deviation, respectively; and PC_{max} and PC_{min} are the maximal and minimal percentile threshold, respectively. Then, the net load percentile is obtained by plugging PC_t into (2). Finally, the non-spinning reserve is derived from (3).

Multi-Timescale Market Simulation Tool for NSRR Validation

This section introduces the data set and forecasts (including both historical net load uncertainty and future net load uncertainty) for the case studies. We also describe the reserve requirement profiles determined by different steps and the sensitivity analysis of the reserve requirements for validation. Most importantly, a multi-timescale market simulation model, which includes both day-ahead and real-time ERCOT markets, is developed to validate our NSRR algorithms.

Data Description

In this research, the ERCOT hourly system-wide load, wind, solar data, and their short-term forecasts (up to 1 week ahead) are used for the case studies. The data span from January 1, 2017, to December 31, 2018. The data in 2017 are used to determine the reserve requirement profiles in 2018. Ideally, the previous 3 years of data should be used for the reserve requirement calculations; however, only 1 year of data is used in the process due to the data limitations. And we assume that its impacts on the different methods are equivalent. In the case studies, the current ERCOT NSRR determination method (denoted as ERCOT) is considered as the baseline method, which is compared to results from the four steps described in Section II. Specifically, two reserve resolution cases, i.e., 2 hour and 1 hour (denoted as Step 2-2 and Step 2-1), are tested in Step 2, and three forecasting update rates/time horizons, i.e., 1 week ahead, 1 day ahead, and 1 hour ahead (denoted as 7DA, 1DA, and 1HA), are analyzed in steps 2–4.

To simulate real-world power systems, we use a well-developed and publicly available synthetic network developed by (Birchfield et al. 2017), which is built on the footprint of ERCOT. The forecasts of load, solar, and wind power are made by following real-world schedules, i.e., the day-ahead forecasts are made at noon on the previous day, the hour-ahead forecasts are made 1 hour ahead, and we use the realized data as the real-time economic dispatch (RTED) inputs. Specifically, the day-ahead forecasts from ERCOT are directly used. A machine learning-based multi-model (M3) forecasting methodology has been proven to be effective in wind (Feng et al. 2017), solar (Feng et al. 2018), and load (Feng, Sun, and Zhang 2019) forecasting, and we use it to generate hour-ahead forecasts. Detailed temporal characteristics of the forecasts used in this study are summarized in Table 6.

Table 6. Temporal Characteristics of Model Inputs

Model	Frequency	Refresh	Horizon	Forecasts Used
DRUC	Daily	Hourly	24 hour	Day ahead
HRUC	Hourly	5 min	60 min	Hour ahead
RTED	Every 5 min	5 min	15 min	Real time

We assume that the load-shedding cost is $9,000/MWh, according to the value of load loss used by ERCOT (Surendran et al. 2016). We use $2,000/MWh, the low system-wide offer cap provided by the Public Utility Commission of Texas (2018), as the penalty price for both responsive and non-responsive reserve shortages. We use $5,500/MWh, the average of the low system-wide offer cap and the value of load loss used, as the penalty price for regulating up/down reserve shortages. The penalty prices are selected such that responsive reserves are curtailed ahead of regulating reserve, and load will be the last to be curtailed (ERCOT 2020).

Multi-Timescale Unit Commitment/Economic Dispatch Model in ERCOT System

Power markets are typically operated over multiple timescales, ranging from day ahead to real time. Like many power markets in the United States, ERCOT has a day-ahead market and a real-time market. The day-ahead reliability unit commitment (DRUC) runs 1 day previous to the operating hour by taking the day-ahead forecasts of load and renewable power productions, and it commits slow thermal units. In the operating day, hourly reliability unit commitment (HRUC) models are run 60 minutes before the operating hour based on the latest system states and forecasts to fine-tune the commitment statuses of other fast thermal units. After all thermal units are committed, ERCOT runs the RTED every 5 minutes to determine the desired generation resource output levels. Energy and ancillary services are co-optimized in the unit commitment and economic dispatch models.

The timeline of our simulation is shown in Figure 87. We simulate the ERCOT market process by following its temporal characteristics: The DRUC runs once per day and consists of 24 1-hour intervals, the HRUC runs every 1 hour and consists of 12 5-minute intervals, and the RTED runs every 5 minutes and covers three 5-minute intervals. Note that at the end of each hour, the RTED horizons are limited by the numbers of remaining 5-minute intervals in that hour, and this is because all thermal generators must be committed in an RTED run, whereas the previous RTUC model can provide commitment statuses only up to the last 5-minute interval of this hour. In summary, 1 day's simulation includes 1 DRUC run, 24 HRUC runs, and 288 RTED runs.

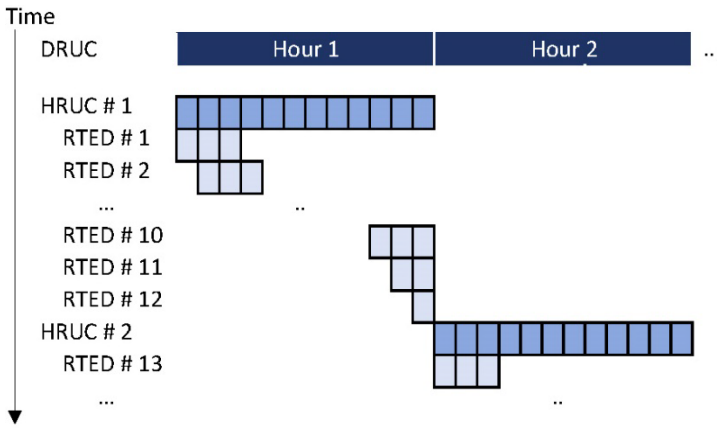


Figure 87. Timeline of the multi-timescale simulation

Note that the DRUC covers 24 hours, but only the first 2 hours are shown here.

Initial power generation levels ($p_{g,0}$) and commitment statuses ($v_{g,0}$) must be given to the DRUC, HRUC # 1, and RTED #1. To initialize the model, we run a DC optimal power flow model by using the forecasts of load and power generation from non-thermal units in the first DRUC interval. In addition, we assume that only the first interval of each RTED model is financially binding, and the initial power generation levels and commitment statuses of all models after DRUC, HRUC # 1, and RTED #1 are obtained from the binding interval of the RTED run immediately preceding that model.

The unit commitment model is formulated as a mixed-integer program by following Li et al. (2020) and Li, Feng, and Zhang (2021), and the economic dispatch model is derived by fixing all commitment variables in the unit commitment model. The objective of the model is to minimize the total costs, which include actual production costs, fixed costs, and penalty terms.

System-level load balance: We assume during each interval that the power level ramps up or down linearly; therefore, the energy production during interval t is obtained by the mean of $p_{g,t}$ and $p_{g,t-1}$, and the system-level load balance is:

$$\sum_{g\in G}\frac{p_{g,t}+p_{g,t-1}}{2}+\sum_{b\in B^L}\delta_{b,t}=\sum_{b\in B^L}D_{b,t}\,,\forall t\in T \qquad (9)$$

Maximum and minimum power generation: Note that for $g \in \boldsymbol{G}$, $P^F_{g,t}$ is given by forecasts:

$$\underline{p}_{g,t}\le p_{g,t}\le \bar{p}_{g,t},\forall g\in G \qquad (10a)$$

$$\bar{p}_{g,t}\le P^F_{g,t},\forall g\in \boldsymbol{G^N} \qquad (10b)$$

Active power flow: $\forall \ell \in \mathbf{L}$:

$$-\Lambda_l\le\sum_{b\in B}SF_{l,b}\left(\sum_{g\in G_b}p_{g,t}-\delta_{b,t}-D_{b,t}\right)\le\Lambda_l \qquad (11)$$

Regulating up/down reserve requirements: $\forall t \in \mathbf{T}$:

$$\sum_{g\in G^T}r^{RU}_{g,t}+\delta^{RU}_t=RQ^{RU}_t \qquad (12a)$$

$$\sum_{g\in G^T}r^{RD}_{g,t}+\delta^{RD}_t=RQ^{RD}_t \qquad (12b)$$

NSRRs: ∀t ∈ **T**

$$\sum_{g\in G^T} r_{g,t}^{NSR} + \delta_t^{NSR} = RQ_t^{NSR} \quad (13)$$

Cost of thermal generators: ∀g ∈ $\mathbf{G}^{\mathrm{T}}$, t ∈ **T**:

$$z_{g,t}^{P} = C_g^0 \cdot v_{g,t} + C_g \cdot p_{g,t} \quad (14a)$$

$$z_{g,t}^{SU} \geq C_g^{SU}\left(v_{g,t} - v_{g,t-1}\right) \quad (14b)$$

$$z_{g,t}^{SD} \geq C_g^{SD}\left(v_{g,t-1} - v_{g,t}\right) \quad (14c)$$

Penalty costs associated with load curtailment and reserve shortage: ∀t ∈ **T:**

$$z_t^{\Delta} = C^{\Delta L}\sum_{b\in B^L}\delta_{b,t}^{\xi} + C^{\Delta R}(\delta_t^{RU} + \delta_t^{RD} + \delta_t^{NSR}) \quad (15)$$

Objective function:

$$\min z = z_t^{\Delta} + \sum_{g\in G^T, t\in T}\left(z_{g,t}^{P} + z_{g,t}^{SU} + z_{g,t}^{SD}\right) \quad (16)$$

Constraint (9) ensures the load balance at each bus. Equation (10a) allows all units to operate within their output limits, whereas the outputs of the renewable units are limited by the power forecasts in (10b). The power flow on all transmission lines is limited by (11) using shift factors. Reserve requirements are met in (12)–(13), where the regulating reserve requirements in the upward and downward directions are enforced in (12a) and (12b), respectively. The NSRRs are enforced in (13). The costs of the thermal generators consist of production costs and fixed costs. Typically, production costs of thermal generators are represented by quadratic functions; here, for simplicity, we use linear functions to approximate them, as shown in (14a). The fixed costs are given by (14b)–(14c), which include startup and shutdown costs. Equation (15) calculates total curtailment costs by summing load curtailment costs and reserve curtailment costs. Finally, the objective function in the deterministic run is given in (16) by summing all costs.

Note that the model in (9)–(16) does not constrain the maximum/minimum available power or the maximum/minimum available reserve from the thermal generators, which we discuss in the coming subsection.

Model Enhancement

1. *Representation of non-spinning reserve:* To qualify as an ancillary service provider, a generating resource must meet market-specific time response requirements, which indicate the maximum time the generating resource can take before it ramps up/down to the scheduled amount of reserve (Zhou et al. 2016; Jorgenson and Denholm 2019). Based on ERCOT's nodal protocol, the time limits for regulating reserve, spinning reserve, and non-spinning reserve to be deployed are 5, 10, and 30 minutes, respectively (ERCOT 2020). Typically, such time response requirements are enforced in market simulation models by ramp rate and maximum available power constraints. For example, the available upward regulating reserve from an online thermal unit, $g \in \mathbf{G}^{TC}$, during time $t \in \mathbf{T}$, is subject to the following constraints:

$$r_{g,t}^{RD} \leq \frac{5}{60} \cdot R_g^D \qquad (17a)$$

$$r_{g,t}^{RD} \leq p_{g,t} - \underline{p}_{g,t} \qquad (17b)$$

Similarly, the available downward regulating reserve from a thermal generator, $g \in \mathbf{G}^{TC}$, is subject to:

$$r_{g,t}^{RD} \leq \frac{5}{60} \cdot R_g^D \qquad (18a)$$

$$r_{g,t}^{RD} \leq p_{g,t} - \underline{p}_{g,t} \qquad (18b)$$

Unlike regulating reserve and spinning reserve, non-spinning reserve is typically provided by offline thermal units in ERCOT, i.e., $g \in \mathbf{G}^{TU}$; therefore, the available non-spinning reserve from an offline thermal unit, g, must be subject to the following constraint:

$$r_{g,t}^{NS} \leq \left(1 - v_{g,t}\right) \cdot P_g^{NS} \qquad (19)$$

where P_g^{NS} denotes the available non-spinning reserve from unit g and is given by the following equation:

$$P_g^{NS} = \begin{cases} 0, \ T^{SU} > \ 0.5 \ h \\ P_g^{min} \ + \ (0.5 \ - \ T^{SU}) \ R_g^U \ , \ T^{SU} \leq \ 0.5 \ h \end{cases} \qquad (20)$$

2. *Multi-timescale representation of thermal generators:* A critical challenge in multi-timescale market simulation is the commitment status of a thermal generator. Typically, the day-ahead market is used to determine the commitment statuses of slow-starting units, which are given by the solution of the DRUC. Once committed, these units will go online in the operating day and follow their committed schedules. The remaining thermal generators, which are usually medium- to fast-starting units, will be committed in the real-time market, which runs HRUC to determine the schedules. The economic dispatch models, which run in real time, do not commit any units but are used to produce a least-cost dispatch of online resources and calculate locational marginal prices. Once a commitment is made, the following models must start up or shut down the unit by following the committed schedules; therefore, the HRUC model must distinguish the committed units from the uncommitted units. In our study, we use $\mathbf{G}^{TU}$ and $\mathbf{G}^{TC}$ to represent the set of uncommitted and committed units, respectively. Note that we assume $\mathbf{G}^{TC} = \emptyset$ in the DURC model because all units are uncommitted in the day-ahead market. Similarly, we assume $\mathbf{G}^{TU} = \emptyset$ in all RTED models because all units are committed. We follow Arroyo and Conejo (2004) to model the startup/shutdown trajectories of a thermal unit if its startup/shutdown time spans more than one interval—i.e., we treat it as an "extended" unit ($g \in \mathbf{G}^{TUE}$). In addition, we treat a thermal unit as an "instant" unit ($g \in \mathbf{G}^{TUI}$) if its startup/shutdown time is less than one interval, and we follow the formulations in Carrión and Arroyo (2006) to describe its behaviors. In summary, the set of thermal generators ($\mathbf{G}^{T}$) is further divided into three subsets—i.e., committed units ($\mathbf{G}^{TC}$), instant units ($\mathbf{G}^{TUI}$), and extended units ($\mathbf{G}^{TUE}$)—and each subset will be modeled differently. The maximum ($\bar{p}_{g,t}$) and minimum (p) available power of a thermal generator are constrained, $\underline{p}_{g,t}$, according to the subset to which it belongs.

Committed units ($\mathbf{G}^{TC}$): Once a thermal unit is committed, its power generation can range from P_g^{max} to P_g^{min}, i.e., its upper and lower bounds of power production, respectively. In addition, we calculate the dispatch limits of a thermal unit if this unit is scheduled to shut down such that the power generation schedules will not violate its commitment schedules. The $\bar{p}_{g,t}$ and $p_{g,t}$ are jointly determined by their upper and lower bounds of power production and their dispatch limits.

Instant units ($\mathbf{G}^{TUI}$): The maximum and minimum available power of an instant unit is determined jointly by its ramp rates and upper/lower bounds of power production:

$$\bar{p}_{g,t} - p_{g,t-1} \le R_g^U \cdot v_{g,t-1} + R_g^{SU}\left(v_{g,t} - v_{g,t-1}\right) + P_g^{max}\left(1 - v_{g,t}\right) \forall g \in G^{TUI}, t \in T \quad (21a)$$

$$\bar{p}_{g,t} \le P_g^{max} \cdot v_{g,t+1} + R_g^{SD}\left(v_{g,t} - v_{g,t+1}\right), \forall g \in G^{TUI}, t \in T \quad (21b)$$

$$p_{g,t-1} - p_{g,t} \le R_g^D \cdot v_{g,t} + R_g^{SD}\left(v_{g,t-1} - v_{g,t}\right) + P_g^{max}\left(1 - v_{g,t-1}\right), \forall g \in G^{TUI}, t \in T \quad (21c)$$

where (21a) and (21b) limit the maximum available power of the thermal generators by accounting for ramp rates, and (21c) limits the power generation during a ramp-down or shutdown process.

Extended units ($\mathbf{G}^{TUE}$): The startup/shutdown trajectory of $g \in \mathbf{G}^{TUE}$ can span multiple intervals, and $\bar{p}_{g,t}$ and $p_{g,t}$ are jointly determined by their commitment status, ramp rates, and startup/shutdown indicators. We follow Arroyo and Conejo (2004) to formulate these constraints, and these equations are omitted for brevity.

Results and Discussion

This section comprehensively discusses the validation results, including the forecasting evaluation, the NSRR statistical validation, the market simulation results, the thermal units' behavior, and the economic impacts of the developed NSRR method.

Forecasting Evaluation

Two sets of system-wide net load forecasts are used in the NSRR determination process: historical forecasts and real-time forecasts. Both forecasts are critical to the NSRRs. The forecasting normalized mean absolute percentage error, normalized mean absolute error (nMAE), and normalized mean root mean square error (nRMSE) are used to evaluate the deterministic forecasts. The probabilistic net load forecasting is evaluated by the normalized average pinball loss (nPL) (Feng, Sun, and Zhang 2019).

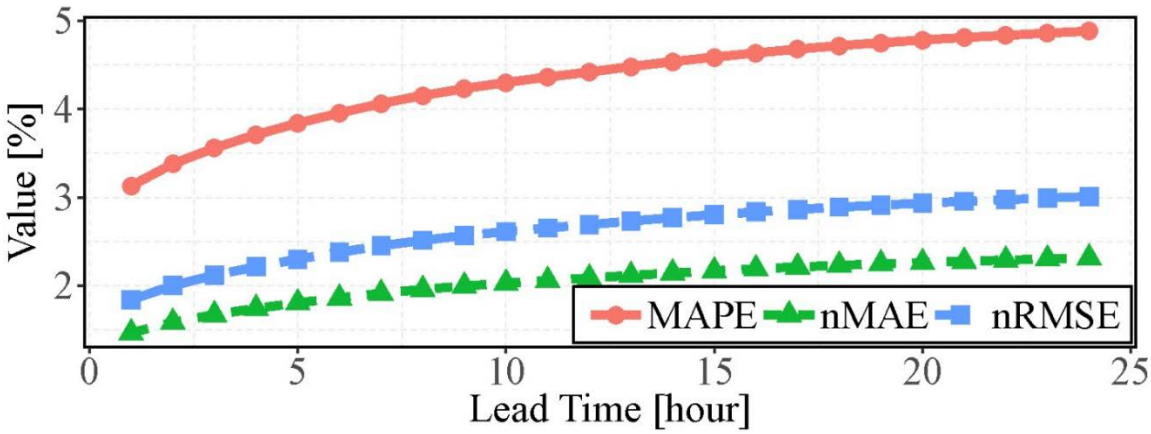


Figure 88. Forecasting error versus lead time

The deterministic forecasting errors of the ERCOT 2017 net load with respect to lead time are shown in Figure 88. It is observed that the forecasting errors increase with the increasing lead time. The impact of increasing forecasting errors will be analyzed in what follows. The real-time forecasting results are listed in Table 7. Similar to the historical forecasting, the real-time forecasting errors increase with the lead time for both deterministic and probabilistic forecasts.

Table 7. Forecasting Result Evaluation Metrics [%]

Step	Deterministic Forecasting			Probabilistic Forecasting
	nMAE	nRMSE	MAAPE	nPL
2	4.05	5.42	8.69	1.52
3	2.37	3.04	5.04	0.94
4	1.55	1.98	3.37	0.67

Adaptive NSRR Results

1. NSRR comparison: NSRRs determined by the current ERCOT method and four steps of the developed method are calculated based on 3-hour-ahead net load forecasting uncertainty in the historical data (results with other lead time forecasts will be studied later). The minimal percentile threshold is set as 0.7 in this section (other thresholds will be discussed in the sensitivity analysis). The average hourly reserve requirements are listed in Table 8. Note that the calculated NSRR is different from the real ERCOT NSRR because only 1 year of data is used and the net load uncertainty forecasting time horizon is short term instead of mid-term. It is observed that by changing the update rate in Step 2, the hourly NSRR is decreased by 6.11%. The hourly NSRR is reduced by 13.93%, 21.10%, and 50.62%, respectively, by the adaptive percentile thresholding based on forecasts with 7DA, 1DA, and 1HA time horizons.

   The hourly NSRRs in each month and each hour are shown in figures 89 and 90, respectively. The reserve reductions of the 1DA scheduling and 1HA scheduling are consistent and significant in all hours and months. On the contrary, 7DA scheduling might estimate larger reserve requirements than the baseline method due to the larger

forecasting errors. It is also found that the NSRRs with adaptive scheduling adopt a diurnal pattern, which reflects the daily pattern in the net load uncertainty.

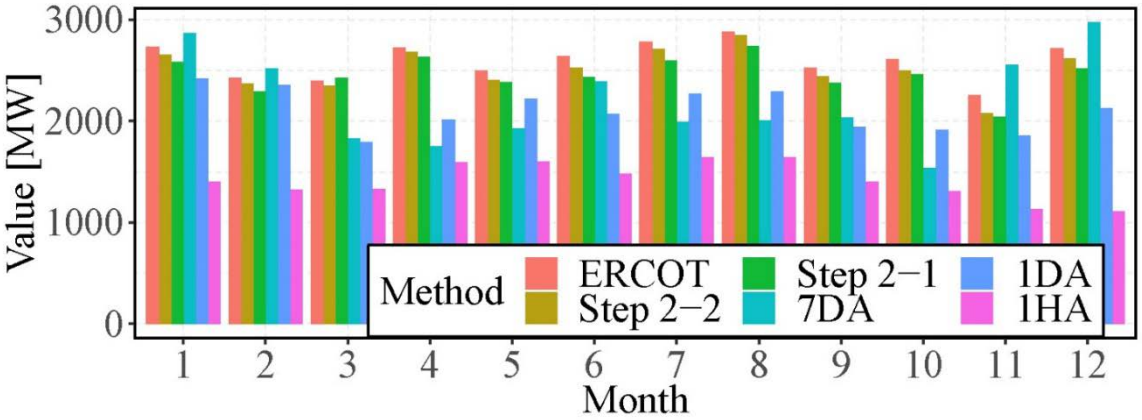


Figure 89. NSRR in each month

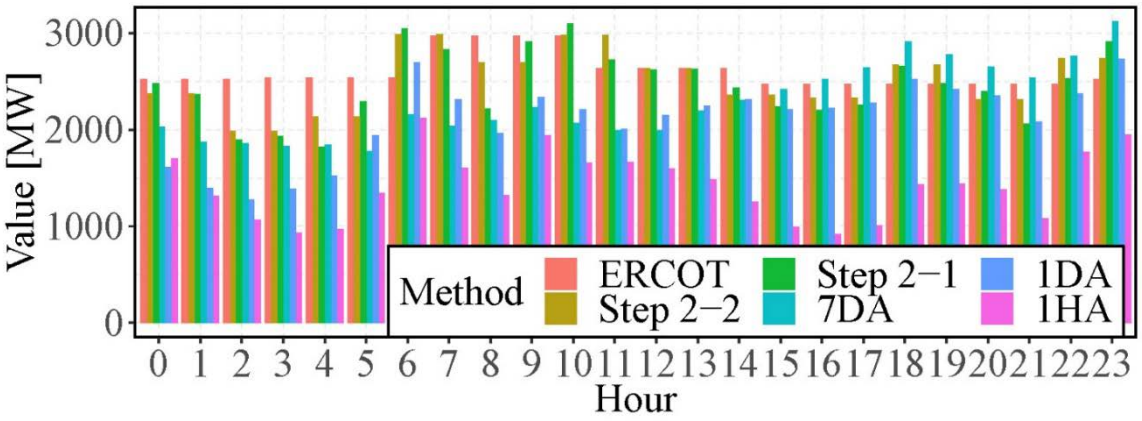


Figure 90. NSRR in each hour

Table 8. Hourly Average Non-Spinning Reserve [MW]

Method	Notes	Hourly Average
ERCOT	The current ERCOT method	2353.87
Step 1 (R2 hour)	Result from Step 1 with a 2-hour resolution	2268.05
Step 1 (R1 hour)	Result from Step 1 with a 1-hour resolution	2210.16
Step 2	Weekly updated reserve	2026.09
Step 3	Daily updated reserve	1857.32
Step 4	Hourly updated reserve	1165.33

2. Sensitivity analysis: NSRRs estimated by the six models are all based on the historical net load uncertainty (previous 1-year forecasting errors), which are directly affected by the forecasting lead time and time horizon. Sensitivity of the NSRRs with respect to the forecasting time horizon is shown in Figure 91. It is observed that the NSRR increases with the time horizon of the historical forecasts. The lower bound of the percentile in (8) also impacts the NSRR, which is shown in Figure 92. These two parameters can be

adjusted based on the ISO's need, which makes the developed method flexible to different requirements. For example, in seasons with more outages or larger forecasting errors, a larger lower percentile bound can be used in the method to hold more NSRRs.

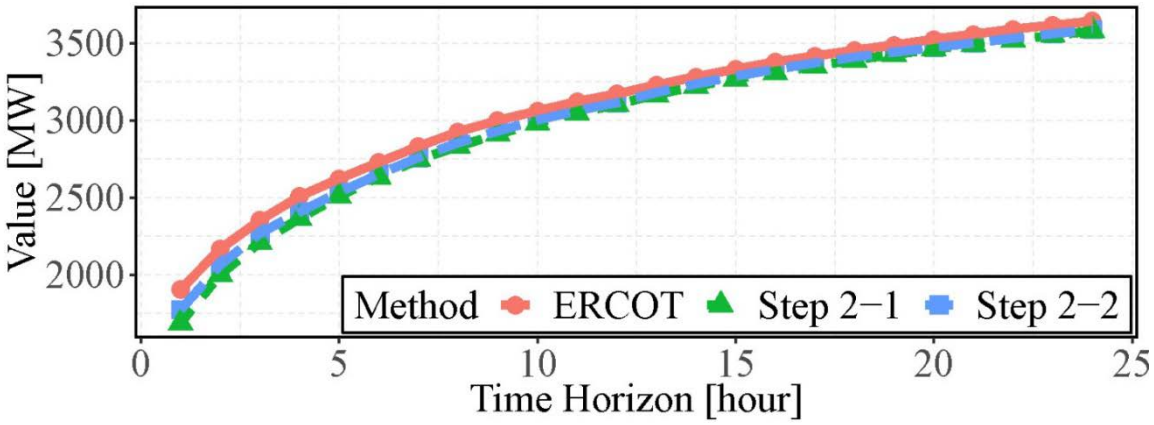


Figure 91. NSRR versus time horizon of historical forecasts

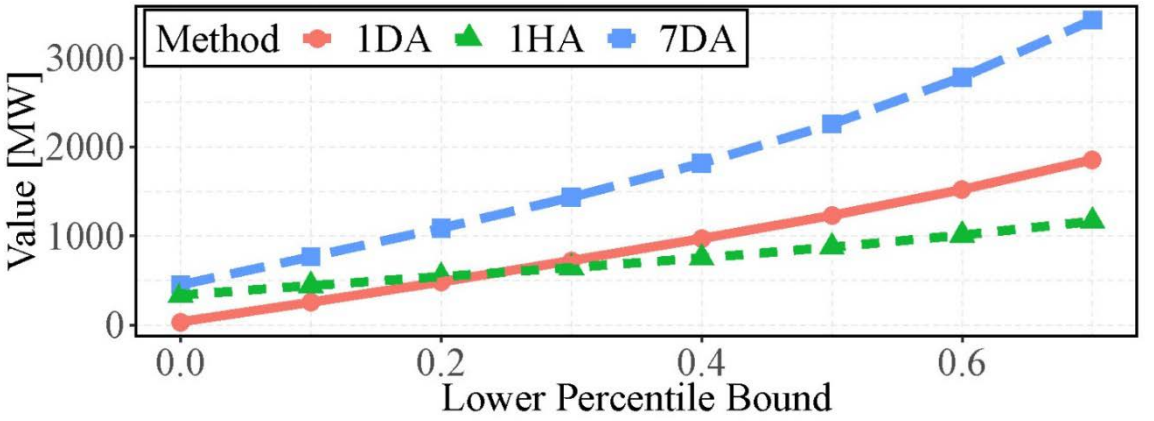


Figure 92. NSRR versus lower percentile bound

Market Simulation Results

To demonstrate our model's capability of simulating the allocation of non-spinning reserve and to examine the economic impacts of different NSRRs, we use two sets of NSRRs: baseline and dynamic. The baseline method is used by ERCOT in their daily operations. This method determines the non-spinning reserve requirement using the 70th to 95th percentile of hourly net load uncertainty from the same month of the previous 3 years (ERCOT 2019). The dynamic method is developed in this work to incorporate the probabilistic net load forecasts (Feng et al. 2020). As shown in Figure 93, the NSRRs in the baseline are greater than those in the dynamic case during most time intervals. We also introduce another case where non-spinning reserve constraints are not enforced, i.e., no non-spinning reserve requirement, and we denote it as the "No NSR" case. We use 2% of demand as the regulating reserve requirement. Note that we assume a constant reserve margin for the regulating reserve to focus on the NSRRs.

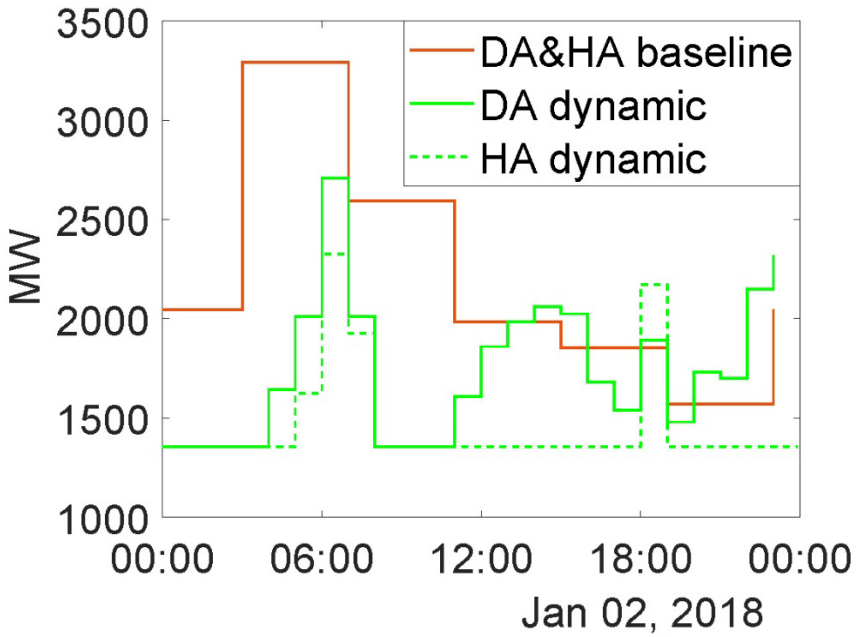


Figure 93. Requirements of NSRRs in both the baseline and the dynamic case

Profiles of Thermal Generators

We first examine the power generation schedules of the thermal units. Figure 94 shows the scheduled power generation levels of two natural gas-fired units in both the day-ahead and real-time markets. We demonstrate the startup/shutdown trajectories by examining two units: the gas unit in Figure 94a is a slow unit that takes 5 hours to start up completely, whereas the unit in Figure 94b is a fast unit that takes 15 minutes to start up completely. The slow unit is scheduled to start at 1 a.m. in the DRUC model, and the RTED profile shows that it follows the DRUC schedule and ramps up to its minimum online capacity, *Pmin*, at exactly 6 a.m. In contrast, although the fast unit is scheduled to remain offline throughout the day in the DRUC model, the RTUC model updates its schedule by starting up at 0 a.m. and ramping up to *Pmin* after three RTED intervals. In addition, it shuts down at 1:40 a.m., completes the shutdown process at 1:55 a.m., and starts up again at around 2 a.m.

Figure 94 also shows the schedule updates owing to forecast errors. The slow unit is scheduled to ramp up to approximately 130 MW at 7 a.m. in the day-ahead schedule; however, the real-time load is significantly lower, and it dispatches only approximately 55 MW in the RTED schedule.

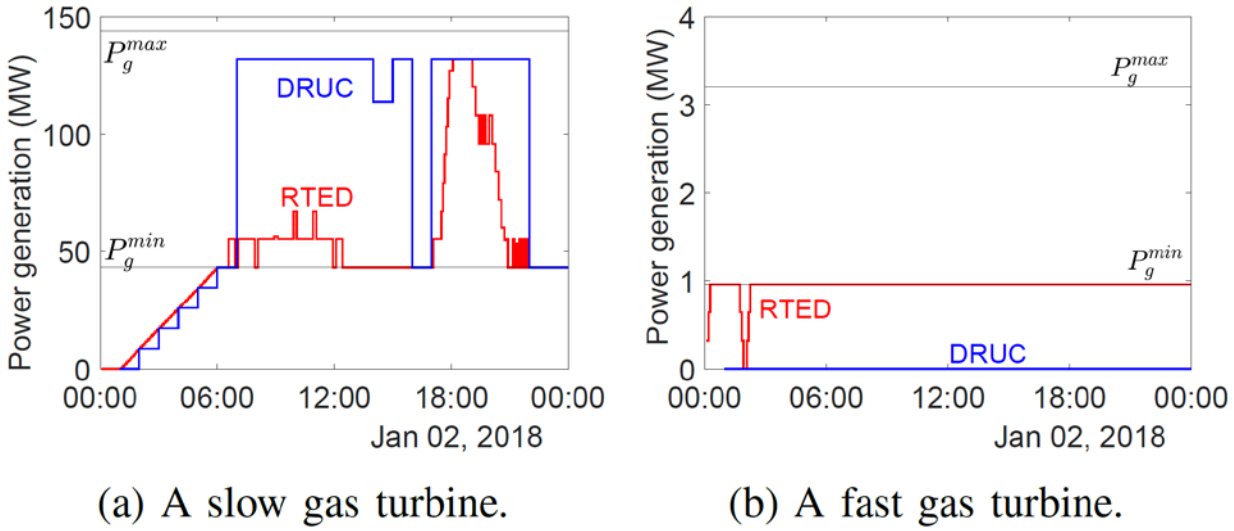


Figure 94. Scheduled power generation of gas turbines in the DURC and RTED models

Note that TSU = 5 h in (a) and TSU = 15 min in (b).

The Impact of NSRRs on Thermal Units

In this study, we have requirements for both regulating reserve and non-spinning reserve, which are provided by different units in compliance with ERCOT operating protocols: regulating reserve must be provided by online units (vg,t = 1), whereas non-spinning reserve can only be provided by offline units (vg,t = 0). By examining the numbers and capacities of online and offline units, we can understand the effects of different NSRRs on system flexibility and adequacy. Figures 95a and 95b show the numbers and capacities of online and offline units, respectively. Note that Figure 95b includes only the number of offline units that qualify as non-spinning reserve resources, i.e., TSU ≤ 30 min.

As shown in Figure 95a, the adoption of non-spinning reserve significantly reduces the number of online units. In the "No NSR" case, there are more than 350 units online throughout most of the day, whereas the baseline only has fewer than 300 units online. In addition, the size of the NSRRs has direct impacts on the numbers and capacities of the online units: The comparison between the baseline and the dynamic case suggests that the dynamic case has both greater numbers and capacities of online units as a result of the reduced NSRRs.

Although the online capacity is typically positively correlated with the number of online units, Figure 95a shows otherwise. For example, despite that the "No NSR" case has the greatest number of online units throughout the day, it has the least online capacities, and the dynamic case has 2.4 GW of additional online capacity compared to the other two cases. In addition, the online capacities of the "No NSR" case and the baseline are almost identical after 4 a.m., although the latter has significantly fewer online units. This implies that the adoption of non-spinning reserve constraints can affect the types of online thermal units. Specifically, the qualified non-spinning reserve units are typically small but fast natural gas-fired gas turbines, which are kept offline in the non-spinning reserve enforced cases. Consequently, the two non-spinning reserve enforced cases use large units to meet the load balancing requirements. This can explain the smaller numbers of online units but greater online capacities in both non-spinning reserve enforced cases.

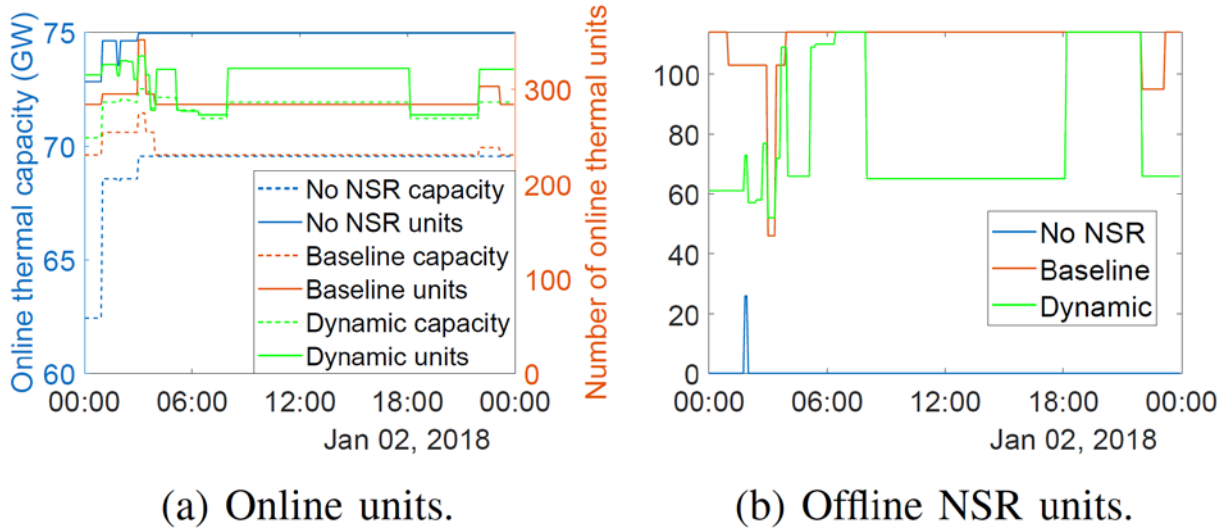


Figure 95. (a) Capacities and numbers of online units. (b) Numbers of offline non-spinning reserve units

The numbers of offline non-spinning reserve units are directly relevant to the NSRRs, and we further examine them in Figure 95b. As shown in Figure 95b, when there are no NSRRs, all qualified non-spinning reserve units are online in the "No NSR" case. The baseline keeps more non-spinning reserve units offline as a result of its greater NSRR values. In addition, the dynamic case starts up and shuts down approximately 60 non-spinning reserve units following the changes in the NSRR values during early morning and late afternoon. As shown in Figure 93, the peaks of the numbers of offline non-spinning reserve units overlap with the peaks of the NSRRs in the dynamic case, which occur in the early morning (6 a.m.) and late afternoon (6 p.m.). These non-spinning reserve units can follow such schedules because they are very flexible units that can start up or shut down within half an hour; however, the frequent startups and shutdowns can incur additional costs, therefore negatively impacting system economics.

Economic Insights Discussion

The costs are tabulated by category in Table 9. The "No NSR" case presents the least total cost because the NSRRs are not enforced, and because in minimization problems, additional constraints can only lead to greater objective values; however, this case would likely cause unacceptable levels of reliability in real power system operations. In contrast, both the baseline and the dynamic cases show greater total costs as a result of the NSRRs. We further demonstrate the economic impacts of the NSRRs by examining each cost term. The production cost primarily comes from the fuel and maintenance costs of online units and therefore are driven by the online capacities. Figure 96 shows the production costs during all binding RTED intervals, and a comparison with the capacities of online units in Figure 95a indicates that the dynamic case has the greatest production costs because of the greatest online capacities. The fixed cost, on the other hand, is mainly driven by the startup and shutdown of the thermal units. As discussed in the previous section, the more frequent startup and shutdown of non-spinning reserve units in the non-spinning reserve enforced cases results in greater fixed costs: The fixed cost in the dynamic and baseline cases are 40% and 20% greater than in the "No NSR" case, respectively.

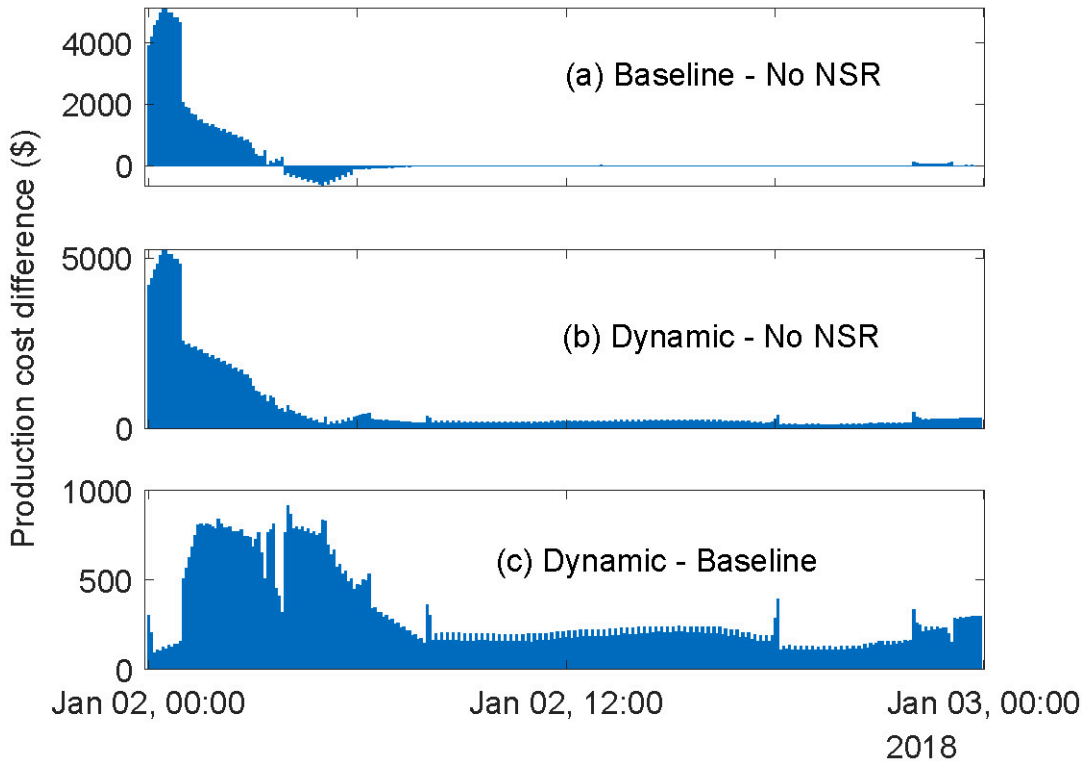


Figure 96. Differences in production costs as a function of time between: (a) the baseline and the "No NSR" case, (b) the dynamic case and the "No NSR" case, and (c) the dynamic case and the baseline

Table 9. Cost Breakdown of All Cases [$]

	No NSR	Baseline	Dynamic
Production	21,141,243	21,221,403	21,309,697
Fixed	63,888	75,967	91,441
Load curtailment	0	0	0
Reserve shortage	0	94,714	33,205
Total cost	21,205,131	21,392,083	21,434,343

Summary

The NSRR determination method we developed showed flexible adaptiveness by incorporating updated system uncertainty while keeping the easily implementable practices. Statistical validation showed that the developed determination method could significantly reduce the NSRR. Additionally, we developed a realistic multi-timescale market simulation tool based on the ERCOT system to validate the NSRR algorithm.

Results showed that the NSRR algorithm improved the contingency handling capability without adding significant costs by keeping more flexible units online.

The reserve shortage costs in both non-spinning reserve enforced cases come from a shortage of regulating reserve. This implies a reduction of flexibility in the online units as a result of fewer

flexible units online. As discussed in the previous section, the online units in both non-spinning reserve enforced cases consist of large but less flexible units, whereas regulating reserve requires a 5-minute response time; therefore, although both non-spinning reserve enforced units have greater online capacities, their capabilities of responding to rapid net load changes are compromised. In most power markets, the regulating reserve is used to mitigate small fluctuations in net load in real time, whereas the non-spinning reserve is intended to help the system recover from unanticipated contingencies (Zhou et al. 2016).

The comparison between the "No NSR" case and the non-spinning reserve enforced cases implies that the capabilities of dealing with contingencies are improved by increasing the risks of violating the requirements of regulating reserve. In addition, the comparison between the baseline and the dynamic case indicates that the dynamic NSRRs can reduce the shortage of regulating reserve. This can be explained by the fact that the dynamic case requires fewer non-spinning reserve units, which leaves more flexible units online toward fulfilling the regulation reserve requirements.

Conclusion

This project developed a data-driven method for the ERCOT system NSRR determination. The method follows the ERCOT procedure to determine the NSRR, which made the developed method easy to implement in practice. In addition, the method considers the future system uncertainty by relying on probabilistic net load forecasts. Three dimensions of adaptiveness were introduced in the developed method to dynamically update the NSRR with higher temporal resolutions and varying confidence thresholds. Case studies show that the developed method significantly reduces the NSRR while keeping flexibilities to adjust the NSRR reduction. The estimated NSRR was validated with simulations of multi-timescale power system operations by using a synthetic power system that is built on the footprint of the ERCOT system. The results suggest that although the NSRRs can improve the system's capability of recovering from contingencies, this improvement was made available by keeping more flexible units offline. In addition, our results highlight the trade-off between the flexibility of the online units and the offline units, which implies that market operators should consider the overall system reliability by jointly determining the requirements of the operating reserve and contingency reserve.

3 Research Area 3: Risk-Parity Dispatch Algorithms

A potential unit commitment formulation using CvaR has been developed in this project to address ramp risks under high solar penetrations. The formulation was just one of the CvaR-based formulations the project team simulated on the high solar scenario case study system in an attempt to develop a viable risk-aware formulation using CvaR. These experiments highlighted the practical and unresolved issues with using a risk metric within operational optimizations and illustrated that a classic stochastic formulation with no risk term is more reliably interpretable and not prone to the unintended consequences that a CvaR-based approach can incur.

At its heart, one fundamental challenge of a CvaR-based formulation is selecting an appropriate function to insert into the risk equation. This function should represent a future risk and emulate possible future costs that might be incurred. In power systems, this cost might be future recourse actions required to address the risky outcome. Selecting a function that results in desirable and reasonable outcomes has proven to be exceptionally difficult. Attempts have included minimizing the risk of curtailment (Zhou, Levin, and Conzelmann 2016) or the risk of price volatility (Wood, Wollenberg, and Sheblé 2013). Within the scope of power system risks, however, renewable curtailment is not worth focusing on—curtailment is one of the easiest and cheapest flexibility options, and curtailment can actually provide value through ancillary services, as will be a focus here. Meanwhile, trying to limit price volatility can logically incur some of the unintended consequences that appear to be common to CvaR formulations. If the stochastic approach by itself already determines the lowest cost for the least-expensive and most-expensive forecast scenarios, then adding a CvaR term to restrict the range of cost outcomes necessarily makes the least-expensive scenario more expensive. The cost of the most-expensive scenario cannot be further reduced. This has the likely unintended consequence of making most scenarios more expensive, which is not a desirable outcome.

This discussion on forecast "scenarios" also points to one more unresolved drawback of the CvaR approach for power system applications. The original appeal of a risk-limiting formulation was as a workaround to the conservativeness of robust unit commitment and the computational burden of stochastic unit commitment. The idea was to make use of a risk-metric calculated from a forecast probability density function at each time point, t. The probability density function is one of two typical ways of representing forecast data, but it is not the common data format for power systems applications. The more common data format is as a set of trajectories over time, where the temporal dependence is clearly defined. There is an incompatibility with using the probability density function format for representing various time-linked constraints from time-step t to $t + 1$, including thermal generator ramping constraints and battery state-of-charge constraints—the probability density function format does not inherently include temporal linkages to enable the easy formulation of these constraints in optimization software. For another part of the project, the team is attempting a workaround using an "M matrix" that represents the transition probabilities from one quantile to another over time, but this is the first attempt to reconcile the probability density function or quantile format with power systems equations. Instead, CvaR implementations in the literature have extended the stochastic formulation using the sampled-data approximation of CvaR over a set of $\mathcal{J}$ trajectories, which can be easily used to formulate time-linked constraints:

$$CVaR = \min_{\beta,z} \beta + \frac{1}{|\mathcal{J}|(1-\alpha)} \sum_{j \in \mathcal{J}} z_j$$

This approach enables CvaR-based formulations for power systems, but note that it therefore does nothing to resolve the computational hindrances of stochastic unit commitment. Instead, it is an extension or hybrid variation on stochastic unit commitment and can suffer from the same computational issues, potentially requiring scenario-reduction techniques or minimizing its usefulness for tight operational time frames. Because the unit commitment does not require a very rapid solve time, a stochastic approach is still viable, but these practical drawbacks are important to consider.

To continue the discussion from above, a new CvaR formulation was attempted this quarter, which found unintended consequences similar to some of the cases discussed from the literature; therefore, rather than focusing on the CvaR formulation itself, our research focus turned toward the value of a classic stochastic unit commitment formulation (no CvaR term) under advanced operating regimes likely under high solar penetrations. These simulations reinforce the value that the probabilistic solar forecasts provide to reduce cost and ensure system reliability, and that has become the focus of this research area.

To begin, three unit commitment formulations were compared: a baseline deterministic unit commitment using a solar point forecast, a classic stochastic approach using 31 solar trajectories, and a variant of the stochastic formulation using CvaR. Two additional flexibility measures were considered that interact with the day-ahead scheduling stage under high solar penetrations: battery storage for energy arbitrage and ancillary service provisions by solar PV (abbreviated as PV-AS). The CvaR variant specifically responds to PV-AS to model the risk that non-spinning reserve will be needed to spinning reserve in instances where solar is unable to meet its PV-AS offers in real time. Cost and reliability impacts are assessed by feeding the hourly resolution outcomes of one of the three unit commitment formulations into a deterministic economic dispatch stage with 5-minute resolution.

First, we show the results of the deterministic and stochastic unit commitment comparisons, focusing on the impacts of using probabilistic day-ahead forecasts, then we conclude by demonstrating the difficulties of the CvaR approach.

Deterministic and Stochastic Formulation Results

This section presents the case study results, investigating the benefits of the stochastic formulation to support battery storage and PV-AS scheduling as well as interactions between the two flexibility measures. As the results show, the day-ahead forecast temporal resolution as well as uncertainty will both play important roles. To distinguish between the impacts of resolution and uncertainty, we focus on the results for the 2 days shown in Figure 97. The first day, April 15, is a clear-sky day with a “near-perfect” day-ahead point forecast. Although there is some slight variance in the solar trajectories around the up and down ramps, the average realization tracks the mean point forecast (Figure 97(a)).

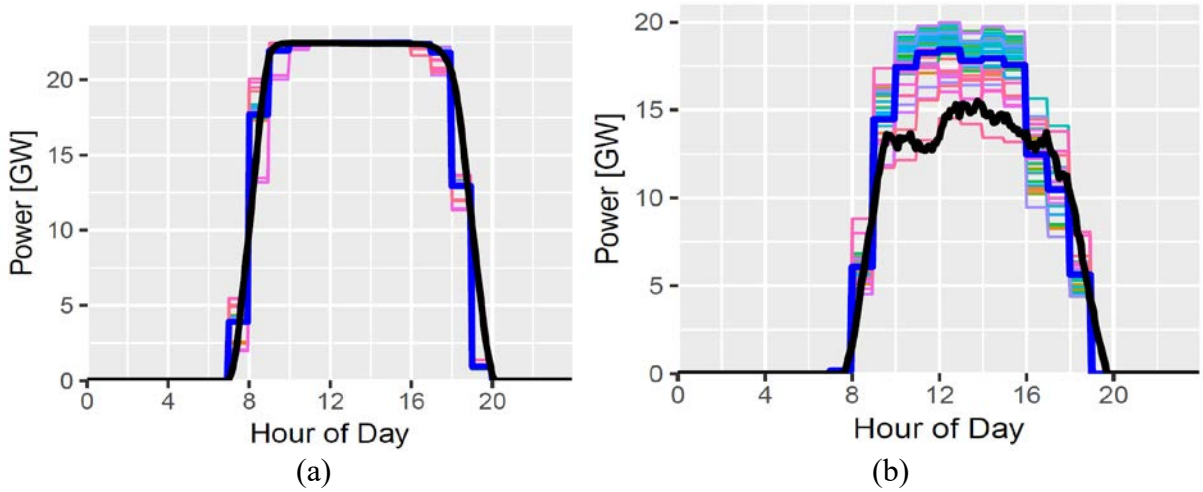


Figure 97. Forecasts and realizations for (a) April 15, a day with clear skies and a "near-perfect" day-ahead hourly resolution point forecast, and (b) March 15, a cloudy day with high uncertainty in the day-ahead forecast

The blue line is the hourly resolution point forecast derived from the trajectory set, and the black line is the realization.

By considering this day, we can identify the impacts of the forecast resolution. The second day, March 15, has high forecast uncertainty (Fig 97(b)). In addition to having a range of midday uncertainty spanning 5 GW–7 GW, the 5-minute realization also falls in the "low tail" of the day-ahead trajectory set. That is, most of the day-ahead trajectories overforecast solar output, and the point forecast used in the baseline deterministic formulation overforecasts by 3 GW–5 GW. Because solar underforecasts can be more easily managed through curtailment, this overforecast is an example of one of the more challenging situations that power system operators will face under high solar shares.

To begin, we assess the interactions of battery storage and PV-AS using the baseline deterministic formulations on the "near-perfect" forecast day. Two questions are addressed: How do the chosen flexibility measures interact, and what is the impact of the hourly resolution in the day-ahead forecast at these very high solar shares?

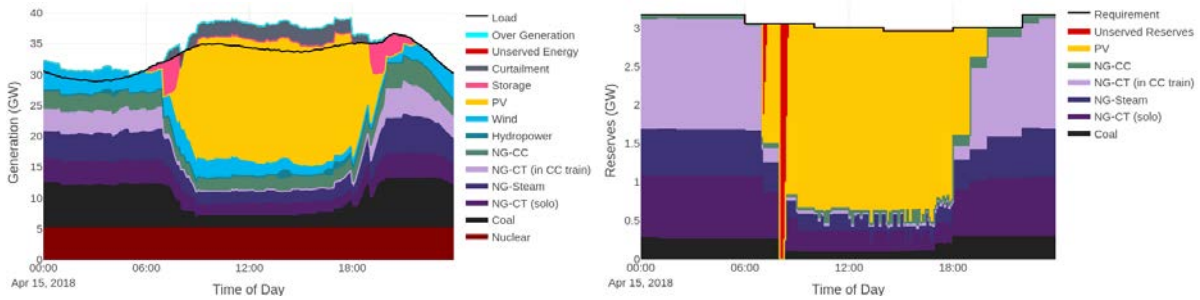
Table 10. April 15 Economic Dispatch Cost Comparison Under Different Flexibility Combinations

Formulation	BS	PV-AS	SU ($)	SD ($)	No-load ($)	Variable ($)	Total ($)
Deterministic	✓	X	351,730	4,598	5,261,119	1,429,416	7,046,863
Deterministic	X	✓	500,059	47,240	3,737,403	1,649,186	5,933,888
Deterministic	✓	✓	418,034	45,615	3,572,416	1,584,444	5,620,509

As more flexibility options are incorporated into the power system, they can be complementary, competitive or conflicting, or neutral. In all likelihood, the type of interaction will also depend on the specifics of the system in question. For battery storage and PV-AS, a few possible interactions can be imagined. During times of solar curtailment, there could be competition for whether curtailed power is best stored for later use in the energy market or moved into the current ancillary service market. Alternatively, the level of curtailment might be driven up to

provide both. This interaction will, of course, depend on several relative costs, including the value of battery peak-shaving and the value of displacing thermal generators in both the reserve and energy markets. For these two flexibility measures, the answer appears to be that their interaction depends on the situation.

Table 10 lists the cost impacts for April 15—including the startup (SU), shutdown (SD), no-load, and variable costs of the thermal fleet—for three scenarios: battery storage (BS) alone, PV-AS alone, and both flexibility measures together. With these relative levels of storage and renewable capacity, the two measures are nominally complementary. When considered individually, PV-AS offers more cost savings than battery storage for this case study, but using both flexibility options together reduces the total cost another 5% past the PV-AS scenario alone. This is shown visually in Figure 98, which shows the fuel stack plot from the economic dispatch stage as well as an analogous stack plot of the generator types used to meet the spinning reserve requirement. The system is putting the available renewable curtailment to good use by both charging the battery storage for evening peak-shaving as well as providing upward spinning and regulation reserve (not shown), which displaces expensive natural gas combustion turbine (NG-CT) plants.

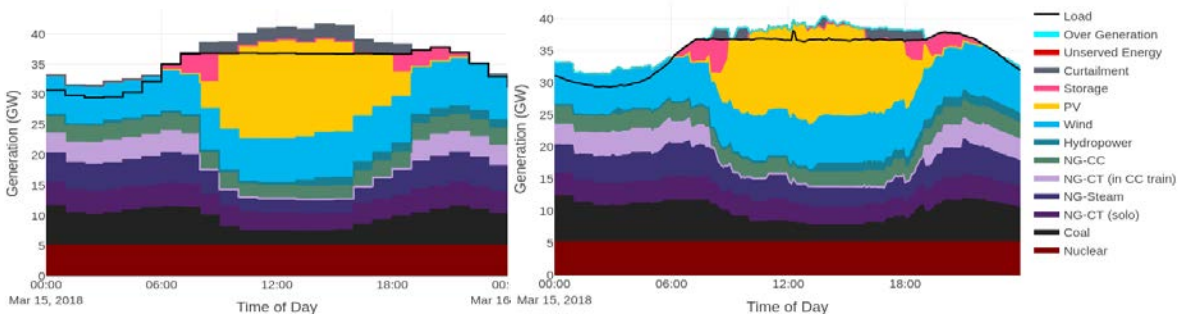


(a) (b)

Figure 98. Fuel and spinning reserve stack plots from the economic dispatch on April 15, following the baseline deterministic unit commitment

Solar power shown above the load line in (a) implies battery charging, whereas curtailment represents not only wasted power but also the upward margin designated to participate in upward ancillary services (regulation up and spinning reserve). The system is receiving the benefits of both flexibility measures: peak-shaving by battery storage and displacement of the expensive NC-GT generators by PV-AS, although there are reliability impacts at the morning solar ramp due to the hourly forecast resolution in the unit commitment.

There is a major caveat, however, in that allowing PV-AS causes reliability issues during the morning solar ramp, resulting in unserved reserve and load. These occur an hour apart due to the change in temporal resolution from the unit commitment to economic dispatch stages. At the start of each hour during the sun-up ramp, there are moments of transition when the solar 5-minutely realization transitions from significantly above the previous hourly average forecast to significantly below the next hour's average forecast (in this case, from 4 GW–7 GW). These transitions can cause reliability impacts, as shown here, as a logical consequence of the limited information available to the day-ahead stage with hourly resolution. Note that the NWP models that currently underlie operational day-ahead forecasts are generally available no finer than hourly resolution; thus, this is a prime opportunity for additional workarounds, whether administrative rules to limit the amount of solar reserves in the first and last hours of the day or an intraday recommitment preceding each ramp using intra-hourly resolution forecasts.

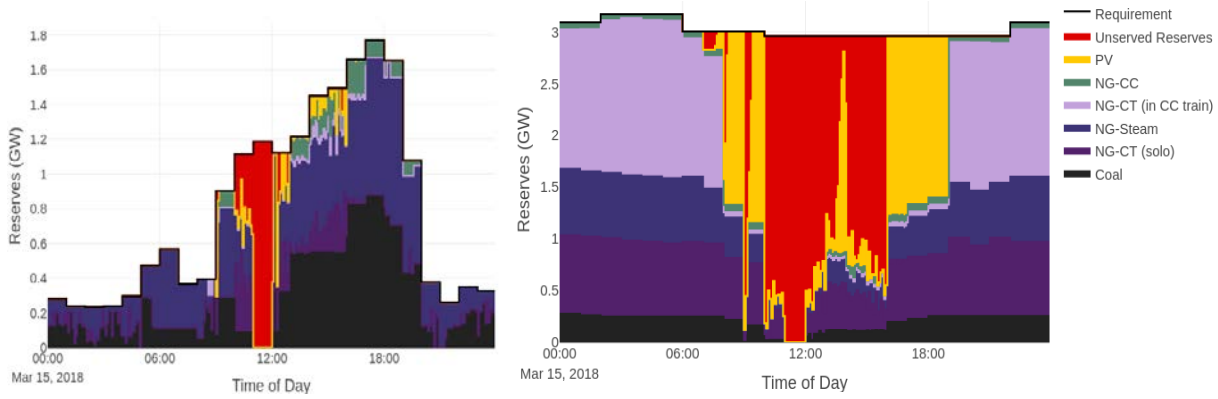


(a) (b)

Figure 99. Results of the (a) day-ahead unit commitment and (b) economic dispatch for March 15

On this day when the realization in the economic dispatch fell far below the day-ahead point forecast, the system had far less solar power and thus solar curtailment than expected. The battery charging targets (i.e., energy market) were prioritized over the solar provisions in the ancillary service markets using whatever excess solar power was available. Coal, natural gas steam (NG-Steam), and NG-CT plants holding upward reserves were instead deployed to serve load and meet the battery charging targets, especially in the hour before noon.

Next, we add a layer of forecast uncertainty by considering March 15 (Figure 97(b)), which has dramatic impacts on the outcomes. The unit commitment and economic dispatch energy market results are illustrated in Figure 99, and Figure 100 shows the resulting upward reserve provisions in the economic dispatch. In this case, when the forecasted solar power fails to materialize, the additional flexibility provisions meant to support solar integration instead conflict and put the system at further risk. At midday, a few interactions can be observed in Figure 99(b). First, the day-ahead battery storage charging commitments increase the total load. This charging load was intended to be provided by PV, but the coal, natural gas steam, and NG-CT generators that were holding upward reserves instead ramp to meet the load.This effectively simulates the upward reserve deployment, although a full simulation of the reserve deployment process is not part of the current framework. Figure 100 shows that this deployment then leaves the two upward reserves categories entirely unserved, leaving the system vulnerable to contingencies and any additional fluctuations in load or renewables.It bears repeating, however, that this simulation does not include an intraday recommitment stage or out-of-market or emergency actions that would remedy the situation closer to real time.



(a) (b)

Figure 100. Results of the upward ancillary service provisions in the economic dispatch on March 15: (a) regulation up and (b) spinning reserve

Widespread unserved reserves are the result of two actions: deployment of the thermal generators holding reserves and the preferential enforcement of battery charging targets over the reserve provisions.

Although this situation was instigated by forecast error in the day-ahead stage, one obvious administrative step would be to halt all battery charging in real time to effectively reduce the load. The battery energy targets serve a clear purpose, though, and such emergency actions might put the system in jeopardy a few hours later if insufficient peak-shaving energy is stored. Operationally, ISOs will likely need a logical framework of thresholds and triggers to balance both charging targets and reserve provisions under extreme forecast errors such as these.

This issue provides clear motivation to advance the day-ahead unit commitment stage with probabilistic solar forecasts that represent the forecast uncertainty. The stochastic unit commitment, achieved through advanced market structures with access to improved information, allows the day-ahead unit commitment to hedge against a range of outcomes well in advance of needing emergency recourse actions to remedy situations like those in Figure 100.

We now switch from the deterministic to the stochastic unit commitment formulation to assess the impact of ingesting probabilistic day-ahead solar forecasts and modifying market structures to grant flexibility around the solar uncertainty. This formulation optimizes the thermal commitments to minimize the average cost over all 31 day-ahead solar forecast trajectories, allowing unique thermal set points to adjust to the solar output within each scenario. For the high-uncertainty day, March 15, Figure 101 illustrates the economic dispatch outcome. Compared to the deterministic version in Figure 99(b), it is evident that the stochastic unit commitment retains a significant amount of natural gas steam turbine commitments as a hedge to compensate for potentially low solar output. Although the unique thermal set points lend more flexibility to the power balance equation, the stochastic formulation does not allow any additional compensating measures in the reserve requirement equations; therefore, all day-ahead scenarios result in identical reserve provisions, meaning that the reserve provisions are naturally set by the worse-case (lowest) scenario or a combination of scenarios over the time horizon. The reserve requirements are thus decided according to a robust mentality. Given the outputs of the stochastic day-ahead unit commitment, the results of the deterministic economic dispatch stage in figures 101 and 101 show clear improvement over the deterministic unit commitment formulation in Figure 101. Natural gas steam plants are selected to provide more of both regulation up and spinning reserve, and the system is prepared to rely on a lower upward reserve provision from solar.

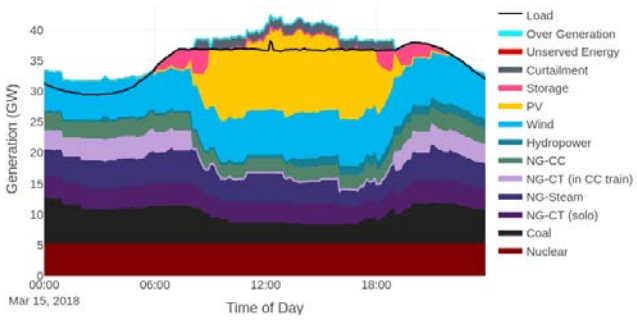


Figure 101. Results of the economic dispatch following the *stochastic* day-ahead unit commitment for March 15

The thermal generators are allowed to adjust up and down to accommodate the difference in solar power in each trajectory, in this case hedging against the possibility of low solar output. The fuel stack shows that by retaining more NG-CT and steam generators compared to 9(b) , there is sufficient solar curtailment at almost all times to serve the needs of both battery charging and the upward reserve markets.

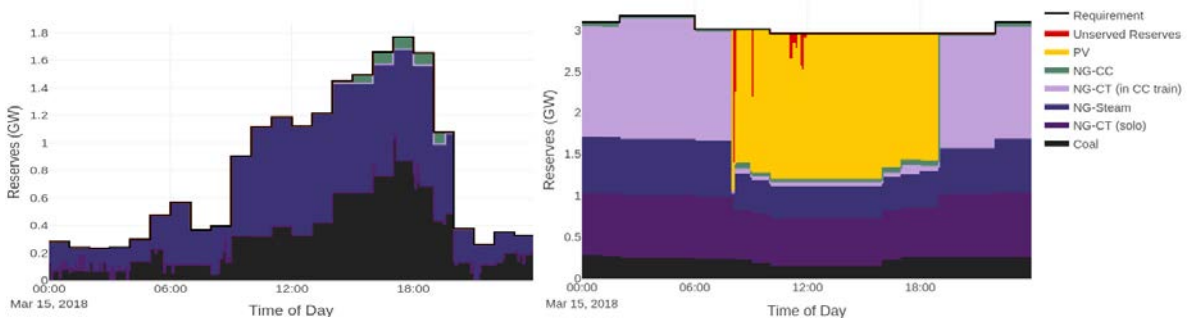


Figure 102. Results of the economic dispatch ancillary service provisions for March 15, following a day-ahead unit commitment with the stochastic formulation

Figure (a) shows the regulation up, and (b) shows the spinning reserve provisions, respectively. In contrast with Figure 100, the stochastic unit commitment results in much more successful reserve provisions. Unserved spinning reserve from 8 a.m.–9 a.m. is due to the forecast resolution, and unserved reserve from 11 a.m.–12 p.m. is due to the forecast error, compared to even the lowest day-ahead solar trajectory (Figure 07(b)).

Table 11. March 15 Economic Dispatch Cost Comparison Under Different Day-Ahead Unit Commitment Formulations

Formulation	BS	PV-AS	SU ($)	SD ($)	No-load ($)	Variable ($)	Total ($)
Deterministic	X	X	171,787	5,202	5,173,505	1,146,771	6,497,265
Deterministic	✓	✓	276,446	36,727	3,788,269	1,360,665	5,462,107
Stochastic	✓	✓	304,750	37,414	3,909,293	1,417,612	5,669,069

Rather than having widespread reserve shortage as well as unmet load impacts, the reserve shortage is reduced to the same difficult morning transitions due to the hourly forecast resolution as well as a much-reduced spinning reserve shortage just before noon. Referencing Figure 97(b), this hour corresponds to the worst forecast errors in the day, when the 5-minute realizations drop below even the lowest day-ahead forecast trajectory, a full 5 GW below the mean of the trajectory set. This is an important reiteration of the fundamental principle that, like the deterministic formulation, the stochastic approach is only as good as the available forecasts.

In summary, the stochastic formulation, using probabilistic day-ahead solar forecasts and additional market flexibility around the thermal generator set points, resolves most of the reliability concerns on this “low-tail” forecast day. Table 11 breaks down the cost estimates for these two formulations with battery storage and PV-AS as well as for the base case with no flexibility measures. Compared to the base case with no added flexibility, the stochastic version with BS + PV-AS achieves a cost reduction of 13%—namely, by decreasing no-load cost while increasing startup, shutdown, and variable costs by lesser amounts. As expected, the stochastic version with BS + PV-AS incurs higher costs than the deterministic version with BS + PV-AS because it retains more thermal generation to hedge against low solar outcomes; the total cost is 3.8% higher, but with much better performance in terms of successfully supplying ancillary services in real time. With the stochastic formulation, 24% of the day’s spinning reserve

requirement is provided by solar PV while incurring 1.5% unserved spinning reserve. Though still unresolved by the combination of flexibility measures considered here, this can be viewed as a prime candidate for layering in additional flexibility options, such as allowing battery storage to hold back some capacity to provide spinning reserve, reshuffling existing reserve categories, as discussed, or conducting an intr-day recommitment with updated, higher-resolution forecasts.

CVaR Formulation Results

The CVaR variation builds on the stochastic unit commitment formulation but relaxes the spinning reserve requirement through a "supplemental reserve" term that emulates non-spinning reserve that would be deployed if solar spinning reserve, $SPIN_{S,j}(t)$, go unserved in real time:

$$\sum_{g \in \mathcal{G}} SPIN_g(t) + SPIN_{S,j}(t) \geq SPIN(t) - \sum_{g \in \mathcal{G}} SUPP_{g,j}(t)$$

A penalty term, $C^{RR}CVaR$, is then added to the overall cost function to minimize these supplemental reserves, where CVaR is calculated using auxiliary variables, z_j, for $j \in J$ (ISO-NE 2019):

$$CVaR = \min_{\beta,z} \beta + \frac{1}{|\mathcal{J}|(1-\alpha)} \sum_{j \in \mathcal{J}} z_j$$

$$Subject\ to\colon\ z_j \geq 0,\ \forall j \in \mathcal{J}$$

$$z_j \geq \sum_{t \in \mathcal{T}} \left(\sum_{g \in \mathcal{G}} SUPP_{g,j}(t) \right) - \beta,\ \forall j \in \mathcal{J}$$

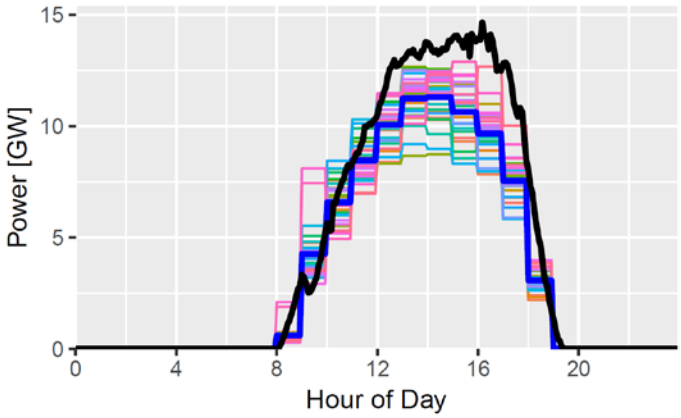


Figure 103. Forecasts and realizations for February 18, a day with a significant underforecast

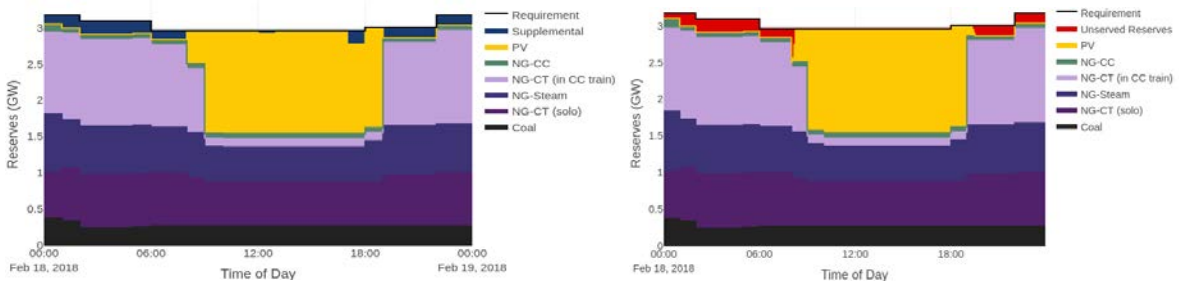
The blue line is the hourly resolution point forecast derived from the trajectory set, and the black line is the realization.

Given that the CVaR formulation relaxes the spinning reserve constraint, it can be expected to have the most value on solar underforecast days, when relaxing the spinning reserve constraint might offer some cost savings. These simulations are conducted on February 18, a day with a consistent afternoon underforecast (Figure 103).The CVaR formulation is parameterized by the level, α, and penalty cost, C^{RR}, selected here as 0.9 and $50/MW.

Table 12. February 18 Economic Dispatch Cost Consumption Under Different Day-Ahead Unit Commitment Formulations

Formulation	BS	PV-AS	SU ($)	SD ($)	No-load ($)	Variable ($)	Total ($)
Deterministic	✓	✓	325,886	70,824	4,540,943	1,861,560	6,799,214
Stochastic	✓	✓	333,171	69,673	4,557,121	1,862,731	6,822,696
CVaR Stochastic	✓	✓	327,672	70,673	4,556,996	1,885,194	6,840,535

The CVaR formulation allows supplemental reserves at night only (C^{RR} = \$50/MW, α = 0.9).



(a) (b)

Figure 104. (a) Spinning reserve provisions for one scenario in the day-ahead unit commitment using the CVaR formulation for February 18, showing an unintended consequence of supplemental reserve provisions reducing the overall requirement at night, and (b) the resulting spinning reserve provisions in the economic dispatch, in which those provisions go unserved (i.e., must be supplied by non-spinning reserves).

The spinning reserve results in Figure 104 show an unintended consequence. In the day-ahead unit commitment stage, the supplemental reserves are consistently being assigned at night to reduce the overall spinning reserve requirement. At this penalty cost, there is more value in reducing the opportunity cost incurred by the thermal units to participate in reserves rather than the energy market, as opposed to managing solar uncertainty during the daytime. This results in a consistent under-commitment, followed by unserved reserves in the economic dispatch stage. Similar results are shown with penalty costs ranging from \$10/MW to \$100/MW, tapering at approximately \$200/MW.

Next, a modified constraint is explored that only allows the supplemental reserve relaxation during the day, when any of the solar trajectory values ($\widehat{p_j^S}(t)$) is greater than zero. Table 12 shows the economic dispatch costs using this modified day-ahead unit commitment compared to both the stochastic and the deterministic formulations. The results are nearly indistinguishable. The CVaR formulation cost is 0.3% higher than the stochastic version, and all formulations experience a moment of unserved spinning reserve at the hourly transition in the morning ramp. In this case, the CvaR formulation does not offer cost benefits; in fact, it results in marginally higher costs while also adding complexity and potential difficulty in parameter tuning. Practically, the solar underforecast days simply do not have the disproportionate impact on system operations that overforecast days do. Because, of course, system operators do not know in advance if a day will be an under- or overforecast, any attempt to hedge toward cost savings in the event of an underforecast can necessarily worsen reliability (and incur more cost) if an overforecast materializes instead; therefore, given these practicalities, the implementation

difficulties, and the marginal performance of the CvaR formulation, we emphasize the classic stochastic approach as the preferred probabilistic method.

4 Research Area 4: Situational Awareness Tool: SolarView

This project developed a visualization platform to provide probabilistic forecasting information in an easy-to-understand form for power system operators to understand the current and future status of solar power generation. This effort expanded on a platform already developed at NREL for wind, WindView, to provide information on probabilistic forecasts that is directly applicable to operators. SolarView development benefited from the capabilities already enabled under WindView, including visualization of the forecast on a site-by-site or aggregate basis, real-time updates of forecast information, identification and alert of events, and visualization of site metadata and location relative to the transmission network.

Two major expansion efforts were made: expanding the platform's capabilities to include solar-specific information and improving the platform's usability. To distinguish SolarView from WindView, the platform was rebranded with solar imagery, likely including warmer colors. In terms of capabilities, the SolarView map was updated to illustrate daytime/nighttime as the sun moves across the region; see Figure 105. The platform also shows clear-sky information, along with the probabilistic forecasts. The temporal resolution was increased from hourly to 5 minutes, enabling visualization of a 5-minute forecast for the first 2 hours, followed by the hourly forecast up to 24 hours ahead.

SolarView also includes new documentation and platform updates to improve its usability. The platform was updated to allow more customization for each user's needs. For example, the parameters that initiate certain alerts and warnings are customizable and persistent. Additionally, the installation process and software usage was documented for an average computer user, without the need for an information technology department to provide support. To get user feedback on the platform, the project team scheduled multiple webinars with power system operators from ERCOT throughout the project and made significant changes to the final product based on the feedback received. The ERCOT Planning Operations team also conducted an in-depth trial of SolarView during a 1-month period, resulting in significant improvements to the user interface. SolarView has been released as open source, and the project team spent significant time to have it be adopted by at least one grid operator and multiple forecasting vendors and commercial traders.

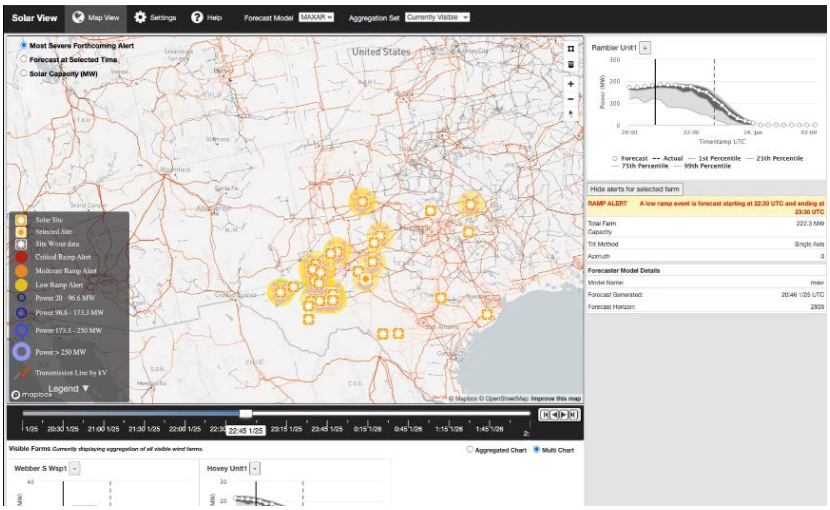


Figure 105. SolarView user interface with an example of real-world data for Texas. This includes aggregated probabilistic forecasts as well as the ability to zoom into individual plants.

Inventions, Patents, Publications, and Other Results

Publications

Burcin Cakir Erdener, Cong Feng, Kate Doubleday, Anthony Florita, Bri-Mathias Hodge: "A review of behind-the-meter solar forecasting," **Renewable and Sustainable Energy Reviews**, Vol. 160, 2022.

Kate Doubleday, Jose Daniel Lara, Bri-Mathias Hodge: "Investigation of Stochastic Unit Commitment to Enable Advanced Flexibility Measures for High Shares of Solar PV," **Applied Energy**, Vol. 321, 2022**.**

Jose Daniel Lara, Oscar Dowson, Kate Doubleday, Bri-Mathias Hodge, Duncan Callaway: "A Multi-Stage Stochastic Risk Assessment with Markovian Representation of Renewable Power," **IEEE Transactions on Sustainable Energy**, Vol. 13, Iss. 1, 2022.

Kate Doubleday, Stephen Jascourt, William Kleiber, Bri-Mathias Hodge: "Probabilistic Solar Power Forecasting Using Bayesian Model Averaging," **IEEE Transactions on Sustainable Energy**, Vol. 12, Iss. 1, 2021.

Kate Doubleday, Vanessa Van Scyoc Hernandez, Bri-Mathias Hodge: "Benchmark Probabilistic Solar Forecasts: Characteristics and Recommendations," **Solar Energy**, Vol. 206, 2020.

Jeffrey Sward, Jose Daniel Lara, Kate Doubleday, Bri-Mathias Hodge: "Online Updating of A Markovian Forecast Representation," Accepted for the **XXII Power Systems Computation Conference (PSCC)**, *June 27–July 1, 2022, Porto, Portugal,* to appear in **Electric Power Systems Research.**

Jose Daniel Lara, Rodrigo Henriquez-Auba, Duncan Callaway, Bri-Mathias Hodge: "AGC Simulation Model for Large Renewable Energy Penetration Studies," **52nd North American Power Symposium (NAPS 2020),** *April 11– 14, 2021, Tempe, AZ.*

Cong Feng, Mucun Sun, Jie Zhang, Kate Doubleday, Bri-Mathias Hodge, Pengwei Du: "A Data-driven Method for Adaptive Reserve Requirements Estimation via Probabilistic Net Load Forecasting," **IEEE Power & Energy Society General Meeting**, *August 2–6, 2020, Montreal, Canada.*

References

Arroyo, J. M., and A. J. Conejo. 2004. "Modeling of Start-Up and Shut-Down Power Trajectories of Thermal Units." *IEEE Transactions on Power Systems* 19 (3): 1562–1568.

Birchfield, A. B., T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye. 2017. "Grid Structural Characteristics as Validation Criteria for Synthetic Networks," *IEEE Transactions on Power Systems* 32 (4): 3258–3265.

California Independent System Operator (CAISO). 2019. *Business Practice Manuals*. http://www.caiso.com/rules/Pages/BusinessPracticeManuals/Default.aspx.

Carrión, M., and J. M. Arroyo. 2006. "A Computationally Efficient Mixed-Integer Linear Formulation for the Thermal Unit Commitment Problem." *IEEE Transactions on Power Systems* 21 (3): 1371–1378.

Cui, M., and J. Zhang. 2018. "Estimating Ramping Requirements with Solar Friendly Flexible Ramping Product in Multi-Timescale Power System Operations." *Applied Energy* 225 (2018): 27–41.

Cui, M., J. Zhang, H. Wu, and B. Hodge. 2017. "Wind-Friendly Flexible Ramping Product Design in Multi-Timescale Power System Operations." *IEEE Transactions on Sustainable Energy* 8 (3): 1064–1075.

Electric Reliability Council of Texas (ERCOT). 2019. *2019 Methodology for Determining Minimum Ancillary Service Requirements*. http://www.ercot.com/content/wcm/key documents lists/137978/9 2019 Methodology for Determining Minimum Ancillary Service Requirements.pdf.

———. 2020. *ERCOT Nodal Protocols*. http://www.ercot.com/mktrules/nprotocols/current. [24]

Feng, C., M. Cui, B.-M. Hodge, and J. Zhang. 2017. "A Data-Driven Multi-Model methodology with Deep Feature Selection for Short-Term Wind Forecasting." *Applied Energy* 190: 1245–1257.

Feng, C., M. Cui, B.-M. Hodge, S. Lu, H. F. Hamann, and J. Zhang. 2018. "Unsupervised Clustering-Based Short-Term Solar Forecasting." *IEEE Transactions on Sustainable Energy* 10 (4): 2174–2185.

Feng, C., M. Sun, and J. Zhang. 2019. "Reinforced Deterministic and Probabilistic Load Forecasting via Q-Learning Dynamic Model Selection." *IEEE Transactions on Smart Grid*.

Feng, C., M. Sun, J. Zhang, K. Doubleday, B.-M. Hodge, and P. Du. 2020. "A Data-Driven Method for Adaptive Reserve Requirement Estimation via Probabilistic Net Load Forecasting." Presented at the 2020 IEEE Power & Energy Society General Meeting (PES GM): 1–5.

Holttinen, H., M. Milligan, E. Ela, N. Menemenlis, J. Dobschinski, B. Rawn, R. J. Bessa, D. Flynn, E. Gomez-Lazaro, and N. K. Detlefsen. 2012. "Methodologies to Determine Operating Reserves due to Increased Wind Power." *IEEE Transactions on Sustainable Energy* 3 (4): 713–723.

ISO New England (ISO-NE). 2019. *ISO New England Operating Procedure No.8 Operating Reserve and Regulation*. https://www.iso-ne.com/markets-operations/markets/reserves/.

Jorgenson, J. L., and P. L. Denholm. 2019. *Modeling Primary Frequency Response for Grid Studies*. Golden, CO: National Renewable National Laboratory.

Li, B., C. Feng, and J. Zhang. 2021. "Multi-Timescale Simulation of Non-Spinning Reserve in Wholesale Electricity Markets." Presented at the 2021 IEEE Green Technologies Conference (GreenTech): 520–527.

Li, B., et al. 2020. "A Clustering-Based Scenario Generation Framework for Power Market Simulation with Wind Integration." *Journal of Renewable and Sustainable Energy* 12 (3): 036301.

Matos, M. A., and R. J. Bessa. 2011. "Setting the Operating Reserve Using Probabilistic Wind Power Forecasts." *IEEE Transactions on Power Systems* 26 (2): 594–603.

Maurer, C., S. Krahl, and H. Weber. 2009. "Dimensioning of Secondary and Tertiary Control Reserve by Probabilistic Methods." *European Transactions on Electrical Power* 19 (4): 544–552.

Midcontinent Independent System Operator (MISO). 2013. *Multi-Faceted Solution for Managing Flexibility with High Penetration of Renewable Resources*. https://www.ferc.gov/CalendarFiles/20140411130433-T1-A%20-%20Navid.pdf.

Ortega-Vazquez, M. A., and D. S. Kirschen. 2008. "Estimating the Spinning Reserve Requirements in Systems with Significant Wind Power Generation Penetration." *IEEE Transactions on Power Systems* 24 (1): 114–124.

Ortega-Vazquez, M. A., and D. S. Kirschen. 2009. "Estimating the Spinning Reserve Requirements in Systems with Significant Wind Power Generation Penetration." *IEEE Transactions on Power Systems* 24 (1): 114–124.

Public Utility Commission of Texas. 2018. "Electric Substantive Rules – Chapter 25 Rules." https://www.puc.texas.gov/agency/rulesnlaws/subrules/electric/25.505/25.505.pdf.

Sun, M., C. Feng, E. K. Chartan, B.-M. Hodge, and J. Zhang. 2019. "A Two-Step Short-Term Probabilistic Wind Forecasting Methodology Based on Predictive Distribution Optimization." *Applied Energy* 238: 1497–1505.

Surendran, R., W. Hogan, H. Hui, and C.-N. Yu. 2016. *Scarcity Pricing in ERCOT*. https://www.ferc.gov/CalendarFiles/20160629114652-3-FERC2016ScarcityPricing ERCOT ResmiSurendran.pdf.

Wang, Z., Q. Bian, H. Xin, and D. Gan. 2015. "A Distributionally Robust Co-Ordinated Reserve Scheduling Model Considering CVaR-Based Wind Power Reserve Requirements." *IEEE Transactions on Sustainable Energy* 7 (2): 625–636.

Wood, A. J., B. F. Wollenberg, and G. B. Sheblé. 2013. *Power Generation, Operation, and Control*. John Wiley & Sons.

Wu, H., I. Krad, A. Florita, B. Hodge, E. Ibanez, J. Zhang, and E. Ela. 2017. "Stochastic Multi-Timescale Power System Operations with Variable Wind Generation." *IEEE Transactions on Power Systems* 32 (5): 3325–3337.

Yang, D., E. Wu, and J. Kleissl. 2019. "Operational Solar Forecasting for the Real-Time Market." *International Journal of Forecasting*.

Zhou, Z., et al. 2016. *Survey of U.S. Ancillary Services Markets*. Argonne, IL: Argonne National Laboratory.

Zhou, Z., T. Levin, and G. Conzelmann. 2016. *Survey of U.S. Ancillary Services Markets*. Argonne, IL: Argonne National Laboratory.