Supplementary Information: Joint Probabilistic Day-Ahead Energy Forecast for Power System Operations

January 23, 2025

1 Weather Data

1.1 NOAA Operational Model Archive and Distribution System (NOMADS)

RAP and HRRR are operational models. The Rapid Refresh is the continental-scale NOAA hourly-updated assimilation/modeling system operational. RAP covers North America and is comprised primarily of a numerical forecast model and an analysis/assimilation system to initialize that model. RAPv5 implemented from Wed. Dec 2, 2020, at NCEP. RAP model website is `https://rapidrefresh.noaa.gov`.

RAP is complemented by the higher-resolution High-Resolution Rapid Refresh (HRRR) model, which is updated hourly and covers a smaller geographic domain. The HRRR is a NOAA real-time 3 km resolution, hourly updated, cloud-resolving, convection-allowing atmospheric model, initialized by 3 km grids with 3 km radar assimilation. Radar data is assimilated in the HRRR every 15 min over 1 hour adding further detail to that provided by the hourly data assimilation from the 13 km radar-enhanced Rapid Refresh. HRRR model website is `https://rapidrefresh.noaa.gov/hrrr/`.

HRRRv4 provides a 48-hour forecast (i.e. cycle) every 6 hours (00, 06, 12, and 18 UTC) from 2 Dec, 2020. HRRRv3 (available from 12 Jul, 2018) does not have a 48-hour forecast every 6 hours, only a 6-hour forecast every hour. Python API instructions are in `https://mesowest.utah.edu/html/hrrr/`.

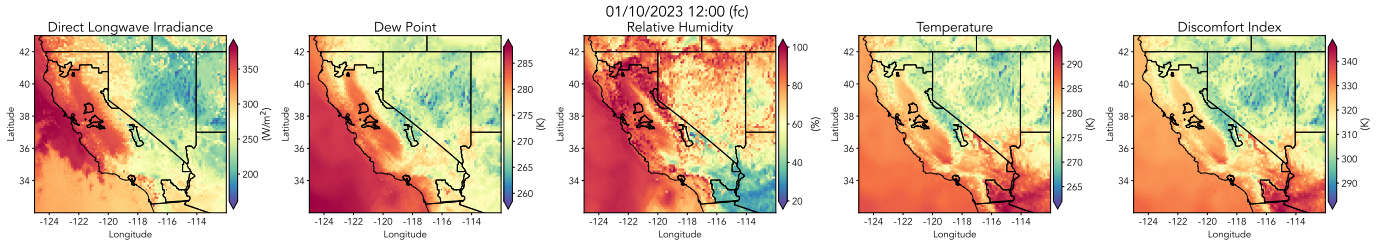

Figure SI1: Weather features in the electric load forecast from the HRRR forecast at 4 pm PTZ for noon the next day. Panels from left to right direct long-ware irradiance (surface), dew point (surface), relative humidity (surface), and discomfort index (2 m). Panels from left to right in the bottom row, relative humidity (2 m), temperature (2 m), and wind speed at 10 m and 80 m.

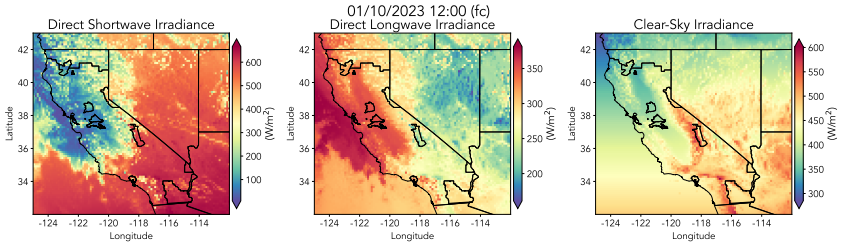

Figure SI2: Weather features in the solar generation forecast from the HRRR forecast at 4 pm PTZ for noon the next day. Panels from left to right direct short-wave irradiance (surface), direct long-wave irradiance (surface), and clear-sky irradiance (surface).

1.2 Wind Velocity Extrapolation at Different Heights

HRRR operation forecast only includes wind components estimated at 10 m and 80 m above sea level. However, research in the literature reports 60 m, 80 m, 100 m, and 120 m as effective heights for a wind operational forecast.

The wind profile can be estimated by applying the Power Law [PH78]. The power low requires knowing the wind velocity components $U_{i,j}$ and $V_{i,j}$ at two different heights. We use the velocity magnitude at 10 m ($W_{10,i,j}$) and 80 m ($W_{80,i,j}$), to calculate the α parameter in the power law,

$$\alpha_{i,j} = \frac{\log W_{80,i,j} - \log W_{10,i,j}}{\log 80 - \log 10}. \tag{1}$$

The wind velocity magnitude at 60 m ($W_{60,i,j}$), 100 m ($W_{100,i,j}$) and 120 m ($W_{120,i,j}$) for each i, j point in the $M \times N$ grid are

$$\begin{aligned} W_{60,i,j} &= W_{10,i,j} \left(\frac{10}{60}\right)^{\alpha_{i,j}}, \\ W_{100,i,j} &= W_{80,i,j} \left(\frac{80}{100}\right)^{\alpha_{i,j}}, \\ W_{120,i,j} &= W_{80,i,j} \left(\frac{80}{120}\right)^{\alpha_{i,j}}. \end{aligned} \tag{2}$$

See the results in the first three panels in Fig. SI3.

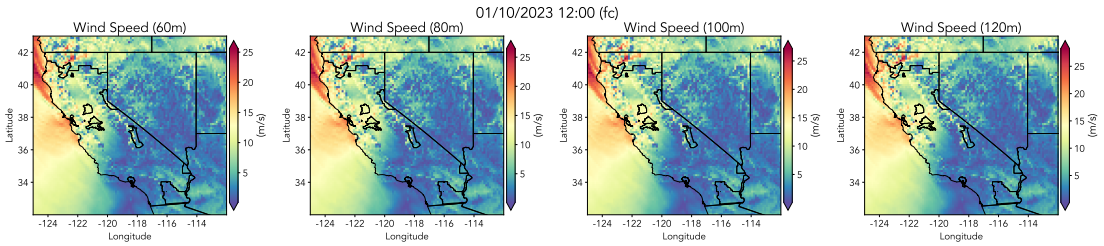

Figure SI3: Panels from left to right, derive wind speed at 60 m, 100 m, 120 m, and clear-sky direct irradiance. The wind speeds were derived wind speed at 10 m and 80 m obtained from HRRR weather forecast at 4 pm PTZ for noon the next day. These weather features are used in the wind generation forecast.

1.3 Apparent Temperature Features

In an operational forecast of load, some metrics have been reported to increase the performances: Discomfort Index (DI) [Tho59] and Wind Chill (WC), HDD, and CDD. The formulas to compute each one of these metrics are here.

- Discomfort Index:
$$DI_{i,j} = (1.8T_{i,j} + 32) + [(0.55 - 0.0055RH_{i,j})(1.8T_{i,j} - 26)] \tag{3}$$
where $T_{i,j}$ (K) is the air temperature, and $RH_{i,j}$ (%) is the relative humidity. $DI_{i,j}$ (°C) is a temperature measure.

- Wind Chill:
$$WC_{i,j} = 13.12 + 0.06215T_{i,j} - 11.37W_{10,i,j}^{0.16} + 0.3965T_{i,j}W_{10,i,j}^{0.16} \tag{4}$$
where $W_{10,i,j}$ (km/h) is the wind velocity magnitude at 10 m, $T_{i,j}$ (K) is the air temperature, and $WC_{i,j}$ (K) is a temperature measure.

- Heating Degree Day (HDD): if $T_{i,j} - 65 < 0$ then $HDD_{i,j} = T_{i,j} - 65$, otherwise $HDD_{i,j} = 0$, where $T_{i,j}$ (°F) is the air temperature.

- Cooling Degree Day (CDD): if $T_{i,j} - 65 > 0$ then $CDD_{i,j} = T_{i,j} - 65$, otherwise $CDD_{i,j} = 0$, where $T_{i,j}$ (°F) is the air temperature. Opposite to HDD.

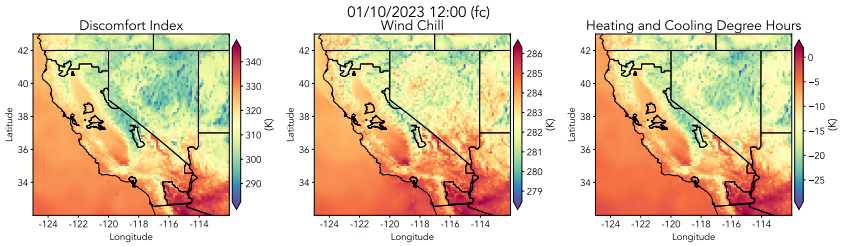

Figure SI4: From left to right, the panels show the discomfort index, wind chill, and heating and cooling degree hours derived from weather features obtained from the HRRR weather forecast at 4 pm PTZ for noon the next day.

2 Spatial Masks

The spatial counts of energy features in a given neighborhood define the points (i.e., pixel) in a grid with weather features, a priori, more informative. For that, consider a point in a grid $\mathbf{x}'_{i,j}$ defined by a pair of spatial coordinates, latitude $\phi_{i,j}$ and longitude $\lambda_{i,j}$. Similarly, the k-th location $\mathbf{x}'_k$ of energy feature $\mathcal{X}$ has latitude ϕ_k and longitude λ_k coordinates.

$$m^{\mathcal{E}}_{i,j} \triangleq \begin{cases} m^{\mathcal{E}}_{i,j} + 1 & \|\mathbf{x}'_{i,j} - \mathbf{x}^{\mathcal{E}}_k\|_2 \geq \sigma \\ m^{\mathcal{E}}_{i,j} & \text{Otherwise.} \end{cases} \qquad \forall k = 1, \ldots, D^{\mathcal{X}}. \tag{5}$$

The parameter $\sigma = 0.25$ controls the neighborhood pixel size used in the counting. The result is spatial density filter $\mathbf{M}^{\mathcal{X}} \in \mathbb{N}^{M \times N}$ for each energy feature $\mathcal{X}$, see in Fig. SI5.

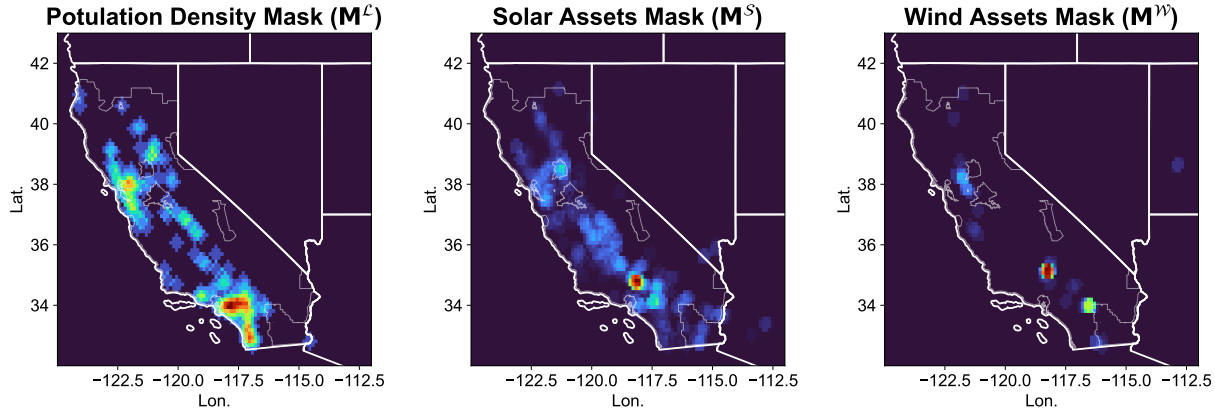

Figure SI5: From left to right, spatial density for electricity demand (i.e. population density) $\mathcal{L}$, solar energy supply $\mathcal{S}$, and wind energy supply $\mathcal{W}$. The fewer counts, the color is darker. The more the counts, the color is more red.

3 Data Processing

The weather features obtained from the NOAA operational forecast are: atmospheric pressure ($\mathrm{AP}_{i,j}$), direct short-wave irradiance ($\mathrm{I}_{i,j}$), direct long-wave irradiance ($\mathrm{R}_{i,j}$), dew point ($\mathrm{D}_{i,j}$), relative humidity ($\mathrm{H}_{i,j}$), air temperature ($\mathrm{T}_{i,j}$), wind velocity components at 10 m ($\mathrm{U}_{10,i,j}$ and $\mathrm{V}_{10,i,j}$) and 80 m ($\mathrm{U}_{80,i,j}$ and $\mathrm{V}_{80,i,j}$).

Other weather features are derived from the NOAA operational forecast. The wind velocity components are transformed to wind velocity magnitudes $\mathrm{W}^{10}_{i,j}$ and $\mathrm{W}^{80}_{i,j}$, and used to derived $\mathrm{W}^{60}_{i,j}$, $\mathrm{W}^{100}_{i,j}$ and $\mathrm{W}^{120}_{i,j}$ applying the power law. The coordinates $\mathbf{x}_{i,j}$ and the elevation $\eta_{i,j}$ (from GMTED2010 dataset) of each point on the grid are used to derive the clear-sky irradiance ($G_{i,j}$). The air temperature ($\mathrm{T}_{i,j}$) and the relative humidity ($\mathrm{H}_{i,j}$) is used to derived the Discomfort Index ($\mathrm{DI}_{i,j}$). Similarly, the air temperature ($\mathrm{T}_{i,j}$) and the wind velocity magnitude ($\mathrm{W}^{10}_{i,j}$) is used to derived the Wind Chill ($\mathrm{WC}_{i,j}$).

A weather feature $X_{i,j,d}$ is defined by two spatial indexes i, j (grid coordinates) and a time index d (operational day). An observed weather feature can be an actual $X_{i,j,d}$ or a prediction $\hat{X}_{i,j,k,t}$ in a 24-hour forecast. The time horizon in the 24-hour forecast is $h = \ell + \rho$, where the day hour is $\rho = \{1, \ldots, 24\}$ and the lead time is $\ell = 8$. After applying the spatial filtering to a weather feature's actual $X_{i,j,d}$ or prediction $\hat{X}_{i,j,k,t}$, the structure has vector form $\mathbf{x}^{\mathcal{X}}_{k,t} = \{\mathrm{X}_{i,j,k,t} \mid m^{\mathcal{X}}_{i,j} > 0\}$ with dimensions $\mathbf{x}^{\mathcal{X}}_{k,t} \in \curvearrowleft^{D^{\mathcal{X}}}$, where $D^{\mathcal{X}} = \sum_{i,j} \mathbb{I}\left(m^{\mathcal{X}}_{i,j} > 0\right)$ and $\mathbb{I}(\cdot)$ is the indicator function.

It applies to $\hat{\mathbf{x}}_{k,t}^{\mathcal{X}}$ in the same manner. The resulting weather feature vectors for each day d are summarized and grouped into different categories:

- **Irradiance features:** clear-sky ($\mathbf{g}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{g}}_{k,t}^{\mathcal{X}}$), direct short-wave ($\mathbf{i}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{i}}_{k,t}^{\mathcal{X}}$), and direct long-wave ($\mathbf{r}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{r}}_{k,t}^{\mathcal{X}}$).
- **Standard weather features:** atmospheric pressure ($\mathbf{p}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{p}}_{k,t}^{\mathcal{X}}$), dew point ($\mathbf{d}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{d}}_{k,t}^{\mathcal{X}}$), relative humidity ($\mathbf{h}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{h}}_{k,t}^{\mathcal{X}}$), and air temperature ($\mathbf{t}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{t}}_{k,t}^{\mathcal{X}}$).
- **Wind velocity magnitude:** 10 m ($\mathbf{w}_{k,t}^{10\mathcal{X}}$ and $\hat{\mathbf{w}}_{k,t}^{10\mathcal{X}}$), 60 m ($\mathbf{w}_{k,t}^{60\mathcal{X}}$ and $\hat{\mathbf{w}}_{k,t}^{60\mathcal{X}}$), 80 m ($\mathbf{w}_{k,t}^{80\mathcal{X}}$ and $\hat{\mathbf{w}}_{k,t}^{80\mathcal{X}}$), 100 m ($\mathbf{w}_{k,t}^{100\mathcal{X}}$ and $\hat{\mathbf{w}}_{k,t}^{100\mathcal{X}}$), and 120 m ($\mathbf{w}_{k,t}^{120\mathcal{X}}$ and $\hat{\mathbf{w}}_{k,t}^{120\mathcal{X}}$).
- **Apparent temperature features:** discomfort index ($\mathbf{s}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{s}}_{k,t}^{\mathcal{X}}$), and wind chill ($\mathbf{c}_{k,t}^{\mathcal{X}}$ and $\hat{\mathbf{c}}_{k,t}^{\mathcal{X}}$).

It is possible to combine the mask from different $\mathcal{X}$ energy features, such as solar $\mathcal{S}$ and wind $\mathcal{W}$, so that $\mathbf{x}_{k,t}^{\mathcal{SW}} = \{\mathrm{X}_{i,j,k,t} \mid m_{i,j}^{\mathcal{S}} > 0 \vee m_{i,j}^{\mathcal{W}} > 0\}$ with dimensions $\mathbf{x}_{k,t}^{\mathcal{SW}} \in \mathbb{R}^{D^{\mathcal{SW}}}$, where $D^{\mathcal{SW}} = \sum_{i,j} \mathbb{I}\left(m_{i,j}^{\mathcal{SW}} > 0\right)$.

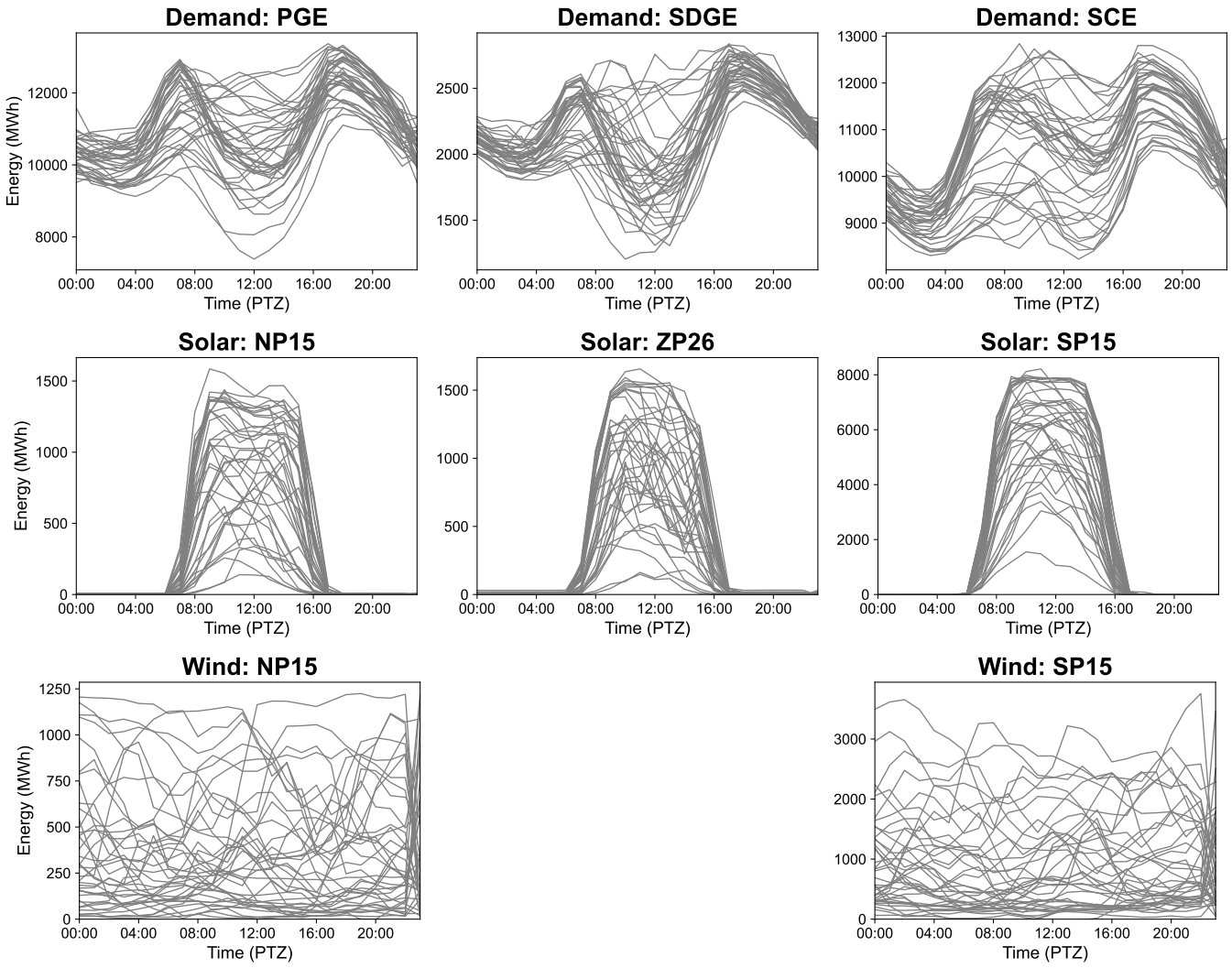

Figure SI6: Top row panels, hourly energy demand on 39 days first days of 2023 from the largest customer-serving utilities in California (PGE, SDGE, and SCE). Middle row panels, hourly solar energy generation at the three different trading hubs operated by CAISO, NP15 (northern California), SP15 (southern California), and ZP26 (central coast). Bottom row panels, hourly wind energy generation at the two CAISO's hubs trading wind energy, NP15 and SP15.

The predictor variables in the 24-hour post-processing forecast are the energy demand or VRE generation. Solar generation (NP15, SP15, and ZP26) and energy demand have 3 spatial components (PGE, SCE, and SDGE) but wind only has 2 (NP15 and SP16). NP15 and PGE are trading hubs and utility in northern California, SP15 and SDGE are

trading hubs and utility in southern California, and ZP26 and SCE are the trading hubs and utility serving parts of the central coast and southern California.

4 Temporal Weather Dissimilitude

The pattern vectors may include temporal features $z_{k,1}$ (year), $z_{k,2}$ (year-day), $z_{k,3}$ (day-hour), $z_{k,4}$ (weekday), $z_{k,5}$ (weekend), $z_{k,6}$ (Holiday) and $z_{k,7}$ (daylight saving time). We consider the year as a continuously increasing variable from the first year in the time series,

$$z_{k,1} = \text{year} - 2019. \tag{6}$$

The idea is to capture the effects in the generation from the cumulative installed capacity and energy demand related to GDP growth.

The cumulative counting in year days count, hours count, and weekday count produces that distance between year days, weekdays, or day hours cannot capture intra-year, intra-week, and intra-day cyclostationary. Cumulative time cannot inform when different day hours (e.g. 23 pm and 1 am) or year days (e.g. Jan 1 and Dec 31) may have similar weather due to climate patterns. The cumulative temporal features are transformed into periodic,

$$\begin{aligned} z_{k,2} &= \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi}{365.25}\cdot \text{year day}\right), \\ z_{k,3} &= \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi}{24}\cdot \text{hour}\right), \\ z_{k,4} &= \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi}{7}\cdot \text{weekday}\right). \end{aligned} \tag{7}$$

In addition, it accounts for effects in the electric load when changing to Daylight Saving Time (DST).

$$\begin{aligned} z_{k,5} &= \begin{cases} 1 & \text{if weekend} \\ -1 & \text{otherwise} \end{cases} \\ z_{k,6} &= \begin{cases} 1 & \text{if Holiday} \\ -1 & \text{otherwise} \end{cases} \\ z_{k,7} &= \begin{cases} 1 & \text{if DST} \\ -1 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

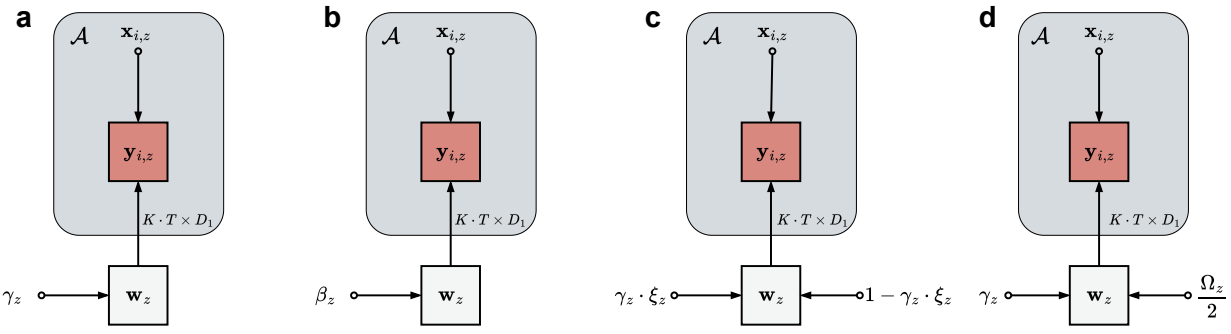

Figure SI7: Diagram of the different sparse learning methods implemented in this research: lasso (a), orthogonal matching pursuit (b), elastic net (c), and group lasso (d). The large gray box represents the reanalyzed dataset ($\mathcal{A}$). The gray box is the learning parameters, and the variables outside the box are the model hyperparameters. The red box is the independent variable. The arrows depict the information flow.

5 Kernel Functions

The kernel functions in this analysis are linear (L), polynomial of order n (P^n), Radial Basis Function (RBF), Rational Quadratic (RQ), and Matérn (M_ν). Their respective functions are,

$$\begin{aligned}
\mathcal{K}_L\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \theta_1 \mathbf{x}_i^\top \mathbf{x}_j + \theta_2, \\
\mathcal{K}_{P^n}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \left(\theta_2 \mathbf{x}_i^\top \mathbf{x}_j + \theta_3\right)^n, \\
\mathcal{K}_{RBF}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \exp\left(-\theta_4 ||\mathbf{x}_i - \mathbf{x}_j||^2\right) + \theta_5, \\
\mathcal{K}_{RQ}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \left(1 + \frac{1}{2\theta_1}\theta_6 ||\mathbf{x}_i - \mathbf{x}_j||^2\right)^{-\theta_7} + \theta_8, \\
\mathcal{K}_{M_\nu}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \cdot \theta_9 ||\mathbf{x}_i - \mathbf{x}_j||^2\right)^\nu K_\nu \left(\sqrt{2\nu} \cdot \theta_9 ||\mathbf{x}_i - \mathbf{x}_j||^2\right) + \theta_{10},
\end{aligned} \tag{9}$$

where $\{\theta_1, \ldots, \theta_{10}\} \in \mathbb{R}^+$, the Matérn order $\nu = \{0.5, 1.5, 2.5\}$, and the polynomial degrees $n = \{2, 3\}$ are hyperparameters [SC04]. $\Gamma(\cdot)$ is the Gamma function, and K_ν is the modified Bessel function of second kind.

6 Scoring Rules

The methods proposed in this investigation are fully probabilistic and provide predictive Probability Density Function (PDF). In addition, the proposed methods can provide ensemble forecasts in a complementary manner. However, it is necessary to apply an appropriate scoring rule to evaluate each approach properly [Win69; MW70]. A scoring rule is a function of a predictive PDF $\hat{f}$, and a realized observation y [BMM21]. The objective of the scoring rule is to perform a *quantitative* evaluation of a forecasting method. The suitable scoring rules are selected from a compilation of proper scoring rules developed for day-ahead resource forecasting [LDP19].

Skill Score (SS). This score measures the improvement of a forecast over a baseline or reference forecast. The SS is a commonly used scoring rule in weather forecasting that is based on another scoring rule. We use the SS to calibrate the location of the predictive mean $\hat{\boldsymbol{\mu}}_{k,t}$ with respect to the baseline point-wise forecast. Root Mean Squared Error (RMSE) is preferable in our task since it weighs large errors heavier than other deterministic metrics (e.g., mean absolute error). The SS is,

$$\begin{aligned}
\mathrm{SS}_{\mathbf{y}}\left(\hat{\mathbf{y}}_{k,t}, \mathbf{y}_{k,t}\right) &= 100 \cdot \left(1 - \frac{\mathrm{RMSE}_{\mathbf{y}}\left(\hat{\mathbf{y}}_{k,t}, \mathbf{y}_{k,t}\right)}{\mathrm{RMSE}_{\text{baseline}}}\right), \\
\mathrm{RMSE}_{\mathbf{y}}\left(\hat{\mathbf{y}}_{k,t}, \mathbf{y}_{k,t}\right) &= \sum_{t=1}^{T} \sqrt{\frac{1}{K} \left\|\hat{\mathbf{y}}_t - \mathbf{y}_t\right\|^2}.
\end{aligned} \tag{10}$$

where $\|\cdot\|$ is the ℓ_2-norm, $\hat{y}_{k,t}$ is a point-wise forecast and $y_{k,t}$ is the actual observation ($\hat{\mathbf{y}}_{k,t} = \hat{\boldsymbol{\mu}}_{k,t}$ in a probabilistic forecast); the index h represent the hour in day d and spatial region t. A positive SS means a performance improvement while a negative one is a reduction, a perfect forecast is 100, and no improvement with respect to the baseline is 0.

Logarithmic Score (LogS). This scoring rule evaluates the negative predictive log probability. In the case of a Normal distribution the LogS is defined as,

$$\begin{aligned}
\mathrm{LogS}_{\mathbf{y}}\left(\hat{f}_{k,t}, \mathbf{y}_{k,t}\right) &= -\log \hat{f}_{k,t}\left(\mathbf{y}_{k,t}\right) \\
&= -\log p\left(\mathbf{y}_{k,t} \middle| \hat{\boldsymbol{\mu}}_{k,t}, \hat{\boldsymbol{\Sigma}}_{k,t}\right) \\
&= \frac{D}{2}\log 2\pi + \frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}_{k,t}\right| + \left(\mathbf{y}_{k,t} - \hat{\boldsymbol{\mu}}_{k,t}\right)^\top \hat{\boldsymbol{\Sigma}}_{k,t}^{-1}\left(\mathbf{y}_{k,t} - \hat{\boldsymbol{\mu}}_{k,t}\right) \\
&\propto \frac{1}{2}\log\left|\hat{\boldsymbol{\Sigma}}_{k,t}\right| + \left(\mathbf{y}_{k,t} - \hat{\boldsymbol{\mu}}_{k,t}\right)^\top \hat{\boldsymbol{\Sigma}}_{k,t}^{-1}\left(\mathbf{y}_{k,t} - \hat{\boldsymbol{\mu}}_{k,t}\right)
\end{aligned} \tag{11}$$

where $\hat{f}_{k,t}$ is the predictive PDF for hour h in day d; $\mathbf{y}_{k,t}$ is the realized observation and $\hat{\boldsymbol{\mu}}_{k,t}$, and $\hat{\boldsymbol{\Sigma}}_{k,t}$ are the mean and covariance parameters in a multivariate normal distribution $\hat{f}_{k,t} \triangleq \mathcal{N}(\hat{\boldsymbol{\mu}}_{k,t}, \hat{\boldsymbol{\Sigma}}_{k,t})$. This scoring rule penalizes unlikely observation (i.e., observation in the tails of PDF) harder than other scoring rules [BMM21]. However, the approach taken in this investigation is based on Bayesian inference, and, in this context, this score may be the most appropriate [Bos+22]. The lower $\mathrm{IS}_{\mathbf{y}}$ is the better it is the forecast.

Energy Score (ES). This scoring rule is the multivariate generalization of the Continuous Rank Probability Score [GR07]. IS is accessible in forecasting methods that provide a predictive density. In this context, ES is ideal for ensemble forecasts generated by an approach based on the Monte Carlo Markov chain. ES is defined as,

$$\begin{aligned}\mathrm{ES}_{\mathbf{y}}\left(\hat{F}_{k,t}, \mathbf{y}_{k,t}\right) &= \int_{-\infty}^{\infty} \left(\hat{F}_{k,t}\left(\mathbf{u}\right) - \mathbb{I}\left(\mathbf{u} \geq \mathbf{y}_{k,t}\right)\right)^2 \mathrm{d}\mathbf{u} \\ &= \frac{1}{M}\sum_{j=1}^{M} \|\hat{\mathbf{y}}_{k,t,j} - \mathbf{y}_{k,t}\| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} \|\hat{\mathbf{y}}_{k,t,j} - \hat{\mathbf{y}}_{k,t,k}\|,\end{aligned} \tag{12}$$

where $\hat{F}_{k,t}$ is the predictive CDF for hour h in day d; $\|\cdot\|$ is the ℓ_2-norm, and M is the number of forecasts in the ensamble. It is important to mention that the ensembles are drawn for the predictive probability that depends on $\hat{\mathbf{y}}_{h-1,d}$ the ensemble drawn for the previous hour,

$$\hat{\mathbf{y}}_{k,t,j} \sim p\left(\hat{\mathbf{y}}_{k,t} \middle| \hat{\boldsymbol{\mu}}_{k,t}, \hat{\boldsymbol{\Sigma}}_{k,t}, \hat{\mathbf{y}}_{h-1,d,j}, \ldots, \hat{\mathbf{y}}_{1,d,j}\right). \tag{13}$$

ES measures the squared distance between $\hat{\mathbf{y}}_{k,t,j}$ an ensemble forecast and $\mathbf{y}_{k,t}$ the realized observation, and is considered numerically more robust than the LogS because is based on the CDF instead of the PDF. The lower $\mathrm{ES}_{\mathbf{y}}$ is the better a model represents the correlation between the horizons of the ensemble scenarios.

Variogram Score (VS^p). Another scoring rule suitable for multivariate ensemble forecasts is the VS of order p [BMM21]. This scoring rule, also named structure function [SH15], properly detects the correlation structure in multivariate predictions. It is defined as,

$$\mathrm{VS}^p_{\mathbf{y}}\left(\hat{F}_{k,t}, \mathbf{y}_{k,t}\right) = \sum_{t=1}^{T}\sum_{t'=1}^{T}\left(|y_{h,d,t} - y_{h,d,t'}|^p - \frac{1}{M}\sum_{j=1}^{M}|\hat{y}_{h,d,t,j} - \hat{y}_{h,d,t',j}|^p\right)^2 \tag{14}$$

where $|\cdot|$ denotes the absolute value. The parameter p is used to transform the distribution of absolute differences and improve the separability. We use $p = 0.5$ as it was found the most adequate choice for Gaussian distributions [SH15]. The lower $\mathrm{VS}^{0.5}_{\mathbf{y}}$ is the better.

Interval Score (IS). This score is important for quantile prediction. However, we can use it to evaluate the accuracy of different predictive intervals in a fully probabilistic forecast. A lower IS indicates that a model approximates more accurately a predictive interval $1 - \alpha$,

$$\begin{aligned}\mathrm{IS}_{1-\alpha}\left(\hat{f}_{k,t}, \mathbf{y}_{k,t}, z, \alpha\right) = 2z\hat{\boldsymbol{\sigma}}_{k,t} &+ \frac{2}{\alpha}\left(\hat{\boldsymbol{\mu}}_{k,t} - z\hat{\boldsymbol{\sigma}}_{k,t} - \mathbf{y}_{k,t}\right)\mathbb{I}\left(\mathbf{y}_{k,t} < \hat{\boldsymbol{\mu}}_{k,t} - z\hat{\boldsymbol{\sigma}}_{k,t}\right) \\ &+ \frac{2}{\alpha}\left(\mathbf{y}_{k,t} - \hat{\boldsymbol{\mu}}_{k,t} + z\hat{\boldsymbol{\sigma}}_{k,t}\right)\mathbb{I}\left(\mathbf{y}_{k,t} > \hat{\boldsymbol{\mu}}_{k,t} + z\hat{\boldsymbol{\sigma}}_{k,t}\right),\end{aligned} \tag{15}$$

where the indicator function $\mathbb{I}(\cdot)$ is counts the number of samples outside the prediction interval; $\hat{\boldsymbol{\mu}}_{k,t}$ and $\hat{\boldsymbol{\sigma}}_{k,t}$ are the predictive mean and standard divination; α is the p-value and z is the z-score of a given prediction interval. For instance, a 95% prediction interval has $\alpha = 0.05$ and $z = 1.959$. The lower $\mathrm{IS}_{1-\alpha}$ is the better. The forecast is rewarded in narrow intervals but penalized when an observation is outside.

In addition to the above scores there are quatile-based scoring rules and diagnosis histograms [LDP19]. However, the proposed methods are not based on quantile regression. The forecast is obtained from a fully probabilistic model and provides a predictive density. For that reason, we consider quantile and interval scoring rules are not appropriate in this investigation.

References

[Tho59] Earl Crabill Thom. "The discomfort index". In: *Weatherwise* 12.2 (1959), pp. 57–61.

[Win69] Robert L Winkler. "Scoring rules and the evaluation of probability assessors". In: *Journal of the American Statistical Association* 64.327 (1969), pp. 1073–1078.

[MW70] Allan H Murphy and Robert L Winkler. "Scoring rules in probability assessment and evaluation". In: *Acta psychologica* 34 (1970), pp. 273–286.

[PH78] Ernest W Peterson and Joseph P Hennessey Jr. "On the use of power laws for estimates of wind power potential". In: *Journal of Applied Meteorology and Climatology* 17.3 (1978), pp. 390–394.

[SC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press, 2004. ISBN: 0521813972.

[GR07] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[SH15] Michael Scheuerer and Thomas M Hamill. "Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities". In: *Monthly Weather Review* 143.4 (2015), pp. 1321–1334.

[LDP19] Philippe Lauret, Mathieu David, and Pierre Pinson. "Verification of solar irradiance probabilistic forecasts". In: *Solar Energy* 194 (2019), pp. 254–271.

[BMM21] Mathias Blicher Bjerregård, Jan Kloppenborg Møller, and Henrik Madsen. "An introduction to multivariate probabilistic forecast evaluation". In: *Energy and AI* 4 (2021), p. 100058.

[Bos+22] Nikos I Bosse et al. "Evaluating forecasts with scoring utils in R". In: *arXiv preprint arXiv:2205.07090* (2022).

7 Supporting Figures

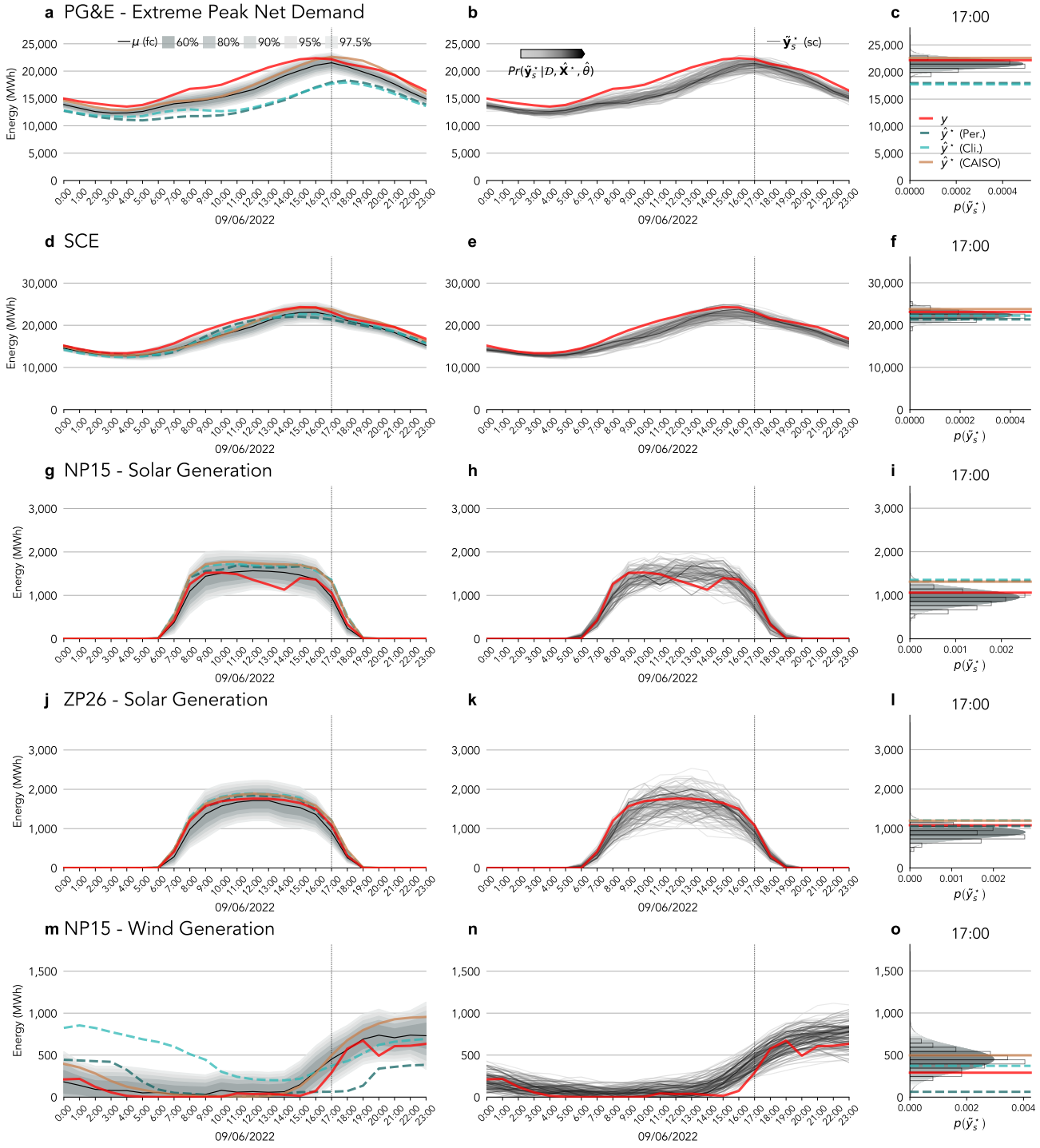

Figure SI8: Independent probabilistic day-ahead forecasts during the extreme peak demand event in Sep. 6., 2022. Probabilistic day-ahead solar generation forecast for PG&E (a, b, and c) and SCE (d, e, and f). Probabilistic day-ahead solar generation forecast at NP15 (g, h, and i) and ZP26 (j, k, and l). Probabilistic day-ahead wind generation forecast at NP15 (m, n, and o). (a, d, g, j, and m) Probabilistic forecast predictive mean $\boldsymbol{\mu}^{\star}$ compared to the actual $y^{\star}$, and the baselines $\hat{\mathbf{y}}^{\star}$ (persistence, climatology, and CAISO). The represented predictive intervals (60%, 80%, 90%, 95%, and 97.5%) have a color gradient from dark (60%) to light gray (97.5%). (b, e, h, k, and n) dashed lines mark detailed hours in (c, f, i, l, and o). The color gradient in the predictive scenarios represents the probability $\Pr(\tilde{\mathbf{y}}_s^{\star} \mid \mathcal{D}, \hat{\Theta}, \hat{\mathbf{X}}^{\star})$ of the s^{th} scenario ($\hat{y}_s^{\star}$) of electricity demand ($\mathcal{L}$), solar ($\mathcal{S}$) or wind ($\mathcal{W}$) generation (darker means higher probability).

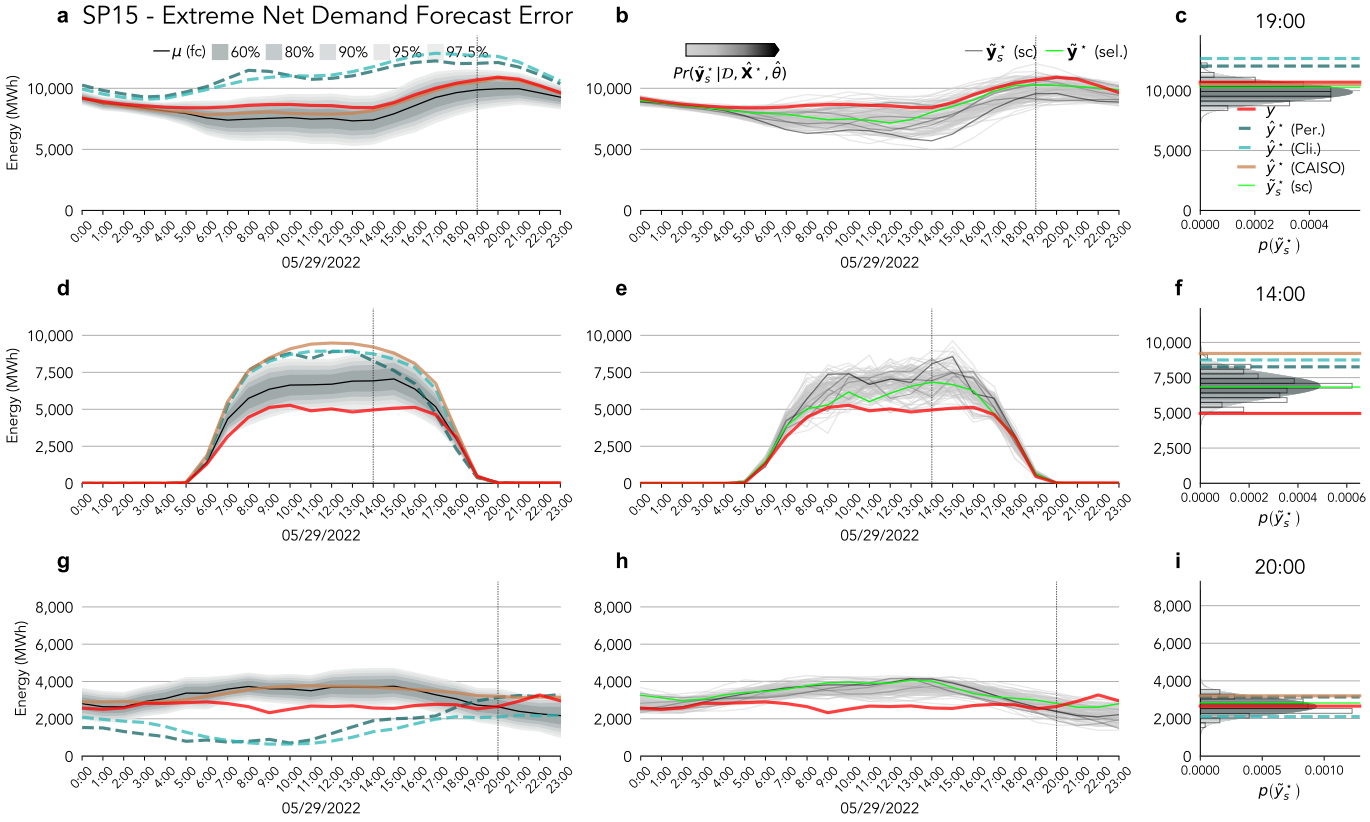

Figure SI9: (a, d, g) Joint forecast predictive mean $\boldsymbol{\mu}^\star$, actual $\mathbf{y}^\star$ and the baseline forecasts $\hat{\mathbf{y}}^\star$ (persistence, climatology, and CAISO) of SCE electricity demand (a), solar (d), and wind generation (g) at SP15. The lines are joint scenarios drawn from the predictive density function (the gray color intensity represents their probability). The highlighted scenario is a joint draw of electricity demand (b), solar generation (e), and wind generation (h). Density function details (c, f, i) are of the hours (marked by vertical dashed lines) with the largest error between the predictive mean $\mu_t^\star$ and the actual $y_t^\star$.

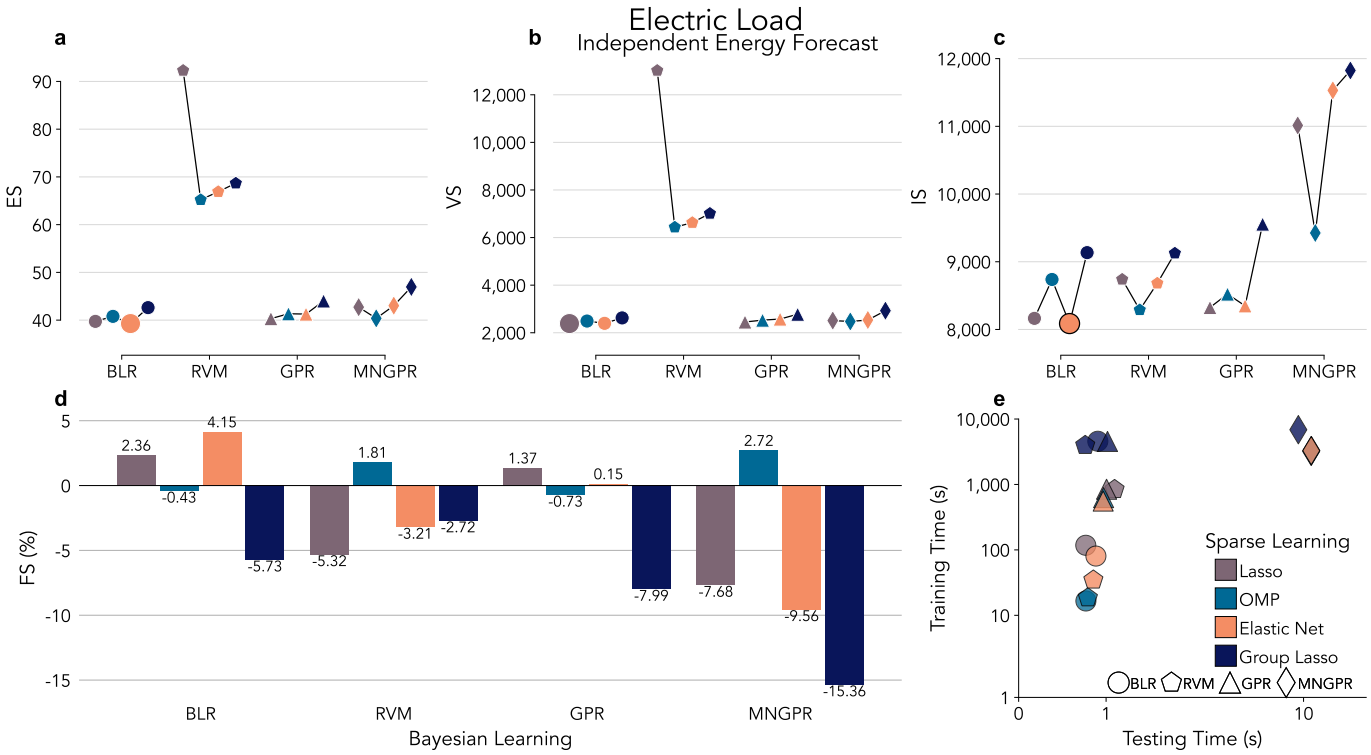

Figure SI10: Independent electric load day-ahead probabilistic forecast model section by lower Ignorance Score (IS). (a) Energy Score (ES), (b) Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$), and (c) IS. (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

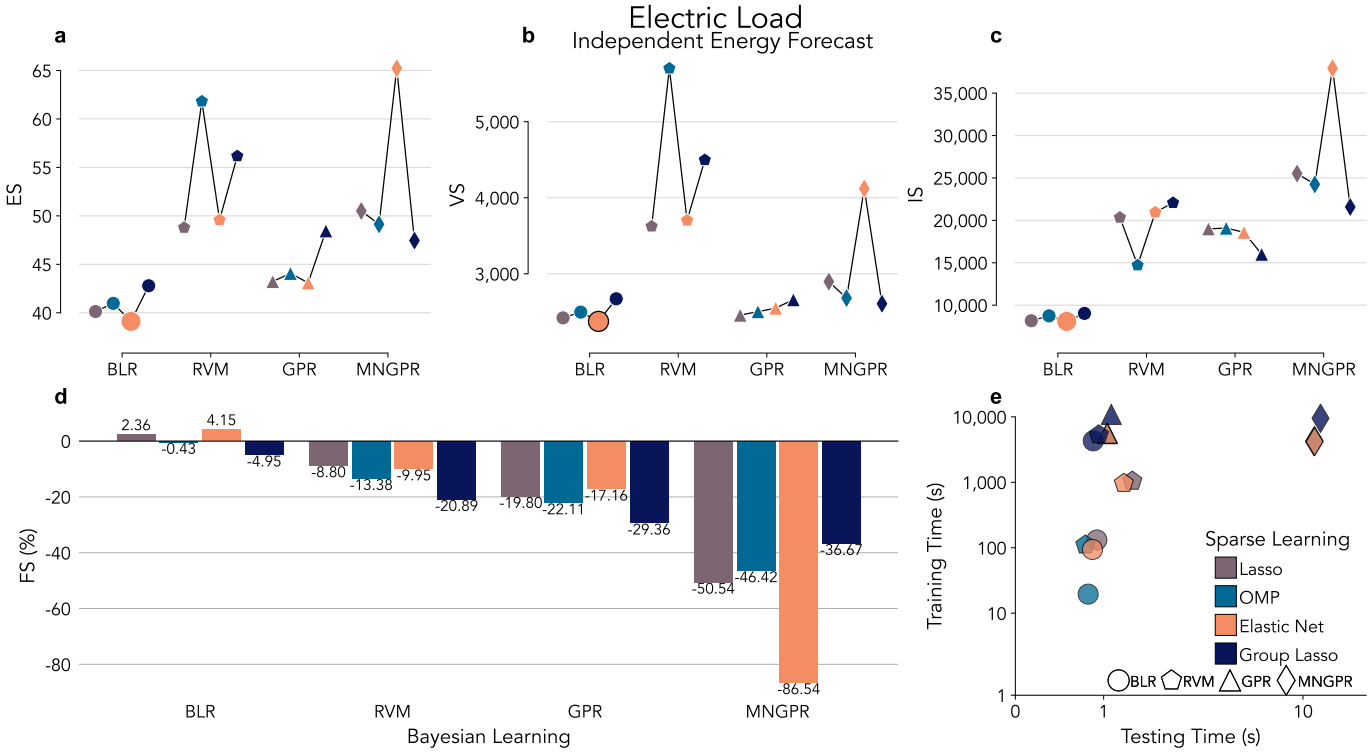

Figure SI11: Independent electric load day-ahead probabilistic forecast model section by lower Variogram Score with $p = 0.5$ ($\text{VS}^{p=0.5}$). (a) Energy Score (ES), (b) $\text{VS}^{0.5}$, and (c) Ignorance Score (IS). (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

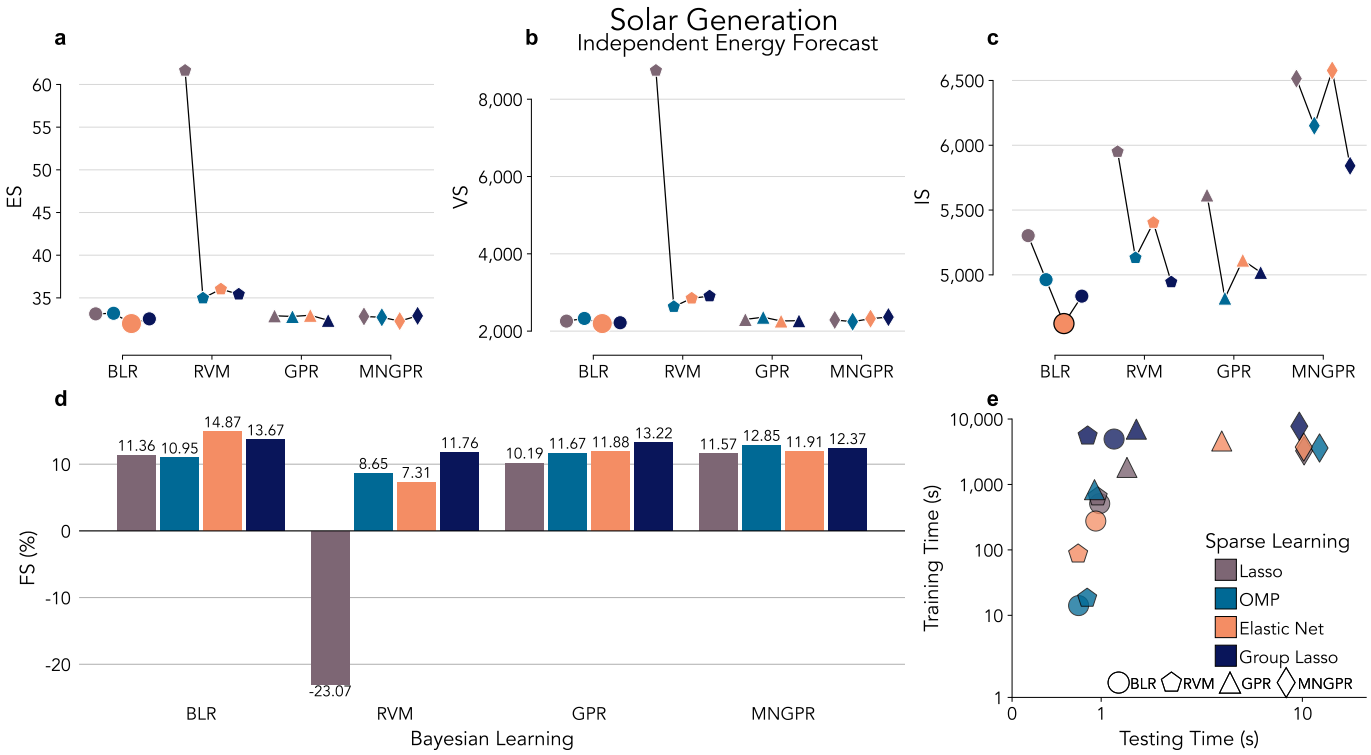

Figure SI12: Independent solar energy generation day-ahead probabilistic forecast model section by lower Ignorance Score (IS). (a) Energy Score (ES), (b) Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$), and (c) IS. (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

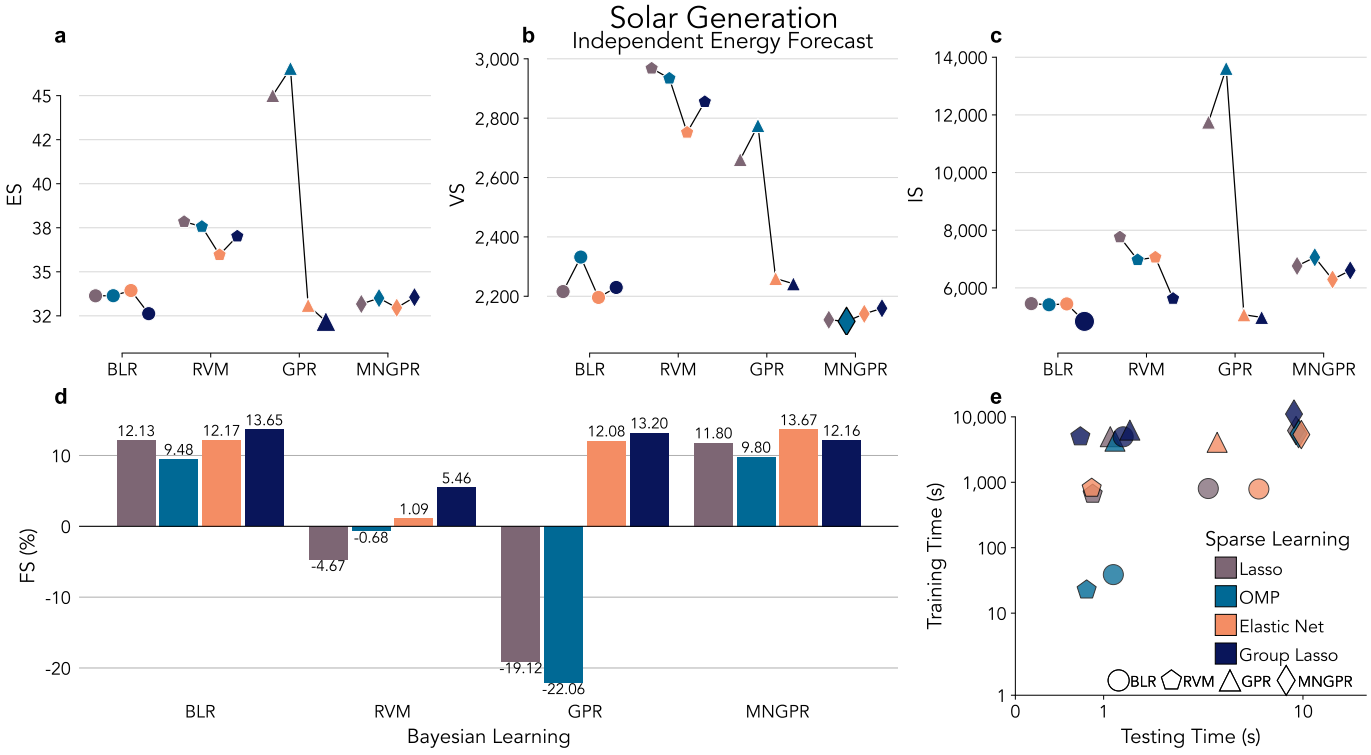

Figure SI13: Independent solar energy generation day-ahead probabilistic forecast model section by lower Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$). (a) Energy Score (ES), (b) $\text{VS}^{0.5}$, and (c) Ignorance Score (IS). (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

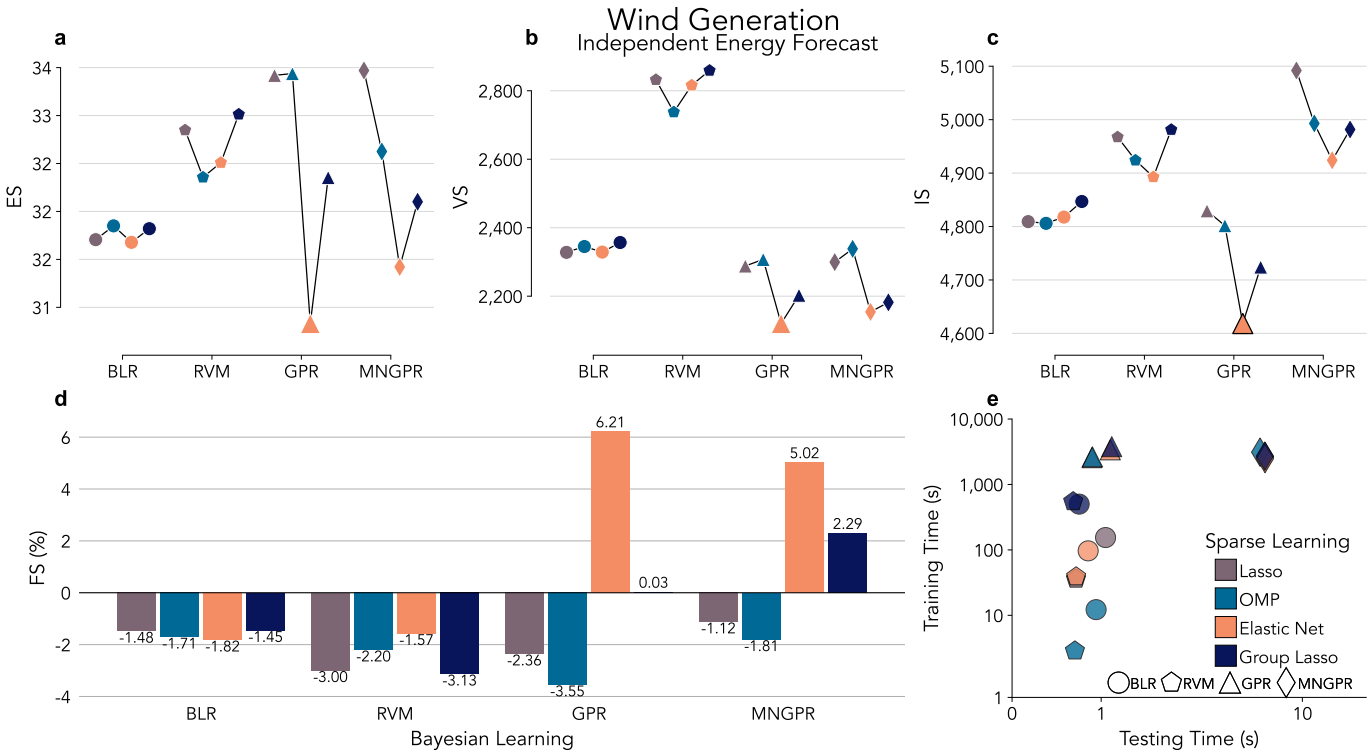

Figure SI14: Independent wind energy generation day-ahead probabilistic forecast model section by lower Ignorance Score (IS). (a) Energy Score (ES), (b) Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$), and (c) IS. (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

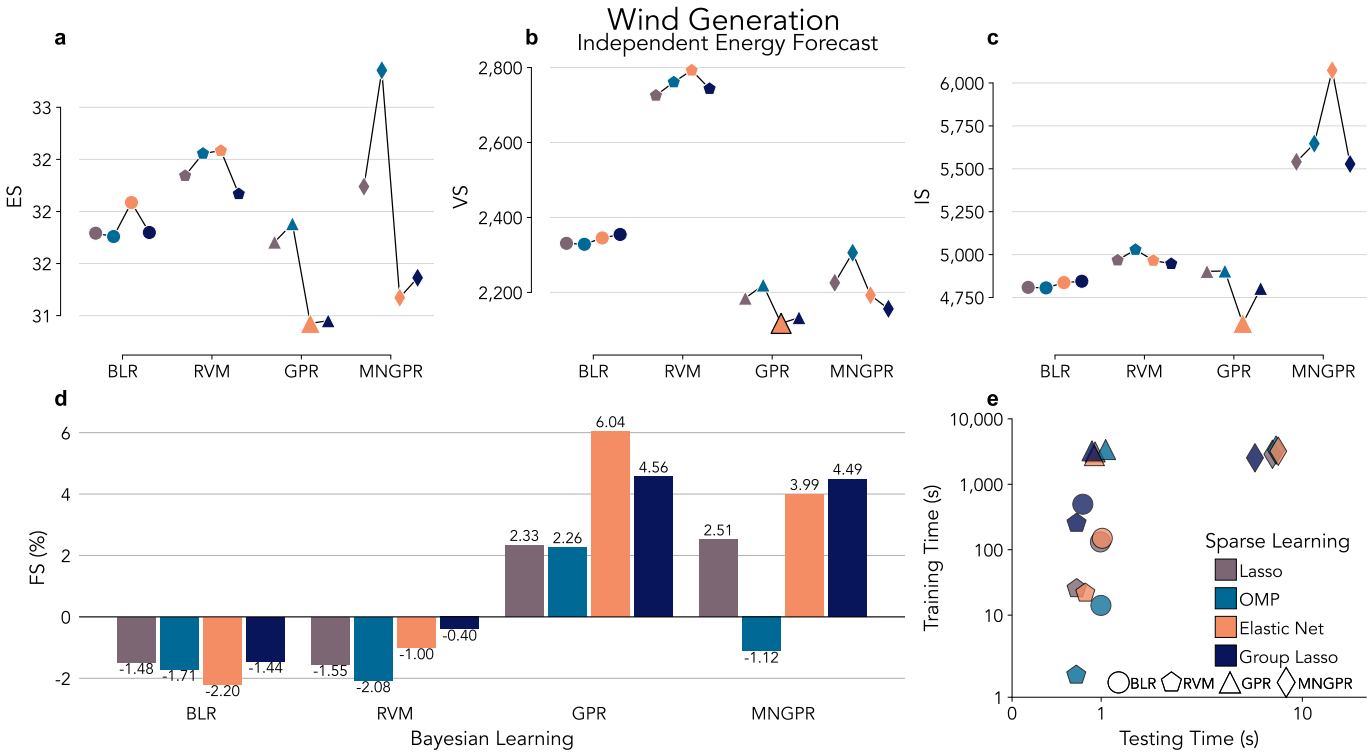

Figure SI15: Independent wind energy generation day-ahead probabilistic forecast model section by lower Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$). (a) Energy Score (ES), (b) $\text{VS}^{0.5}$, and (c) Ignorance Score (IS). (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

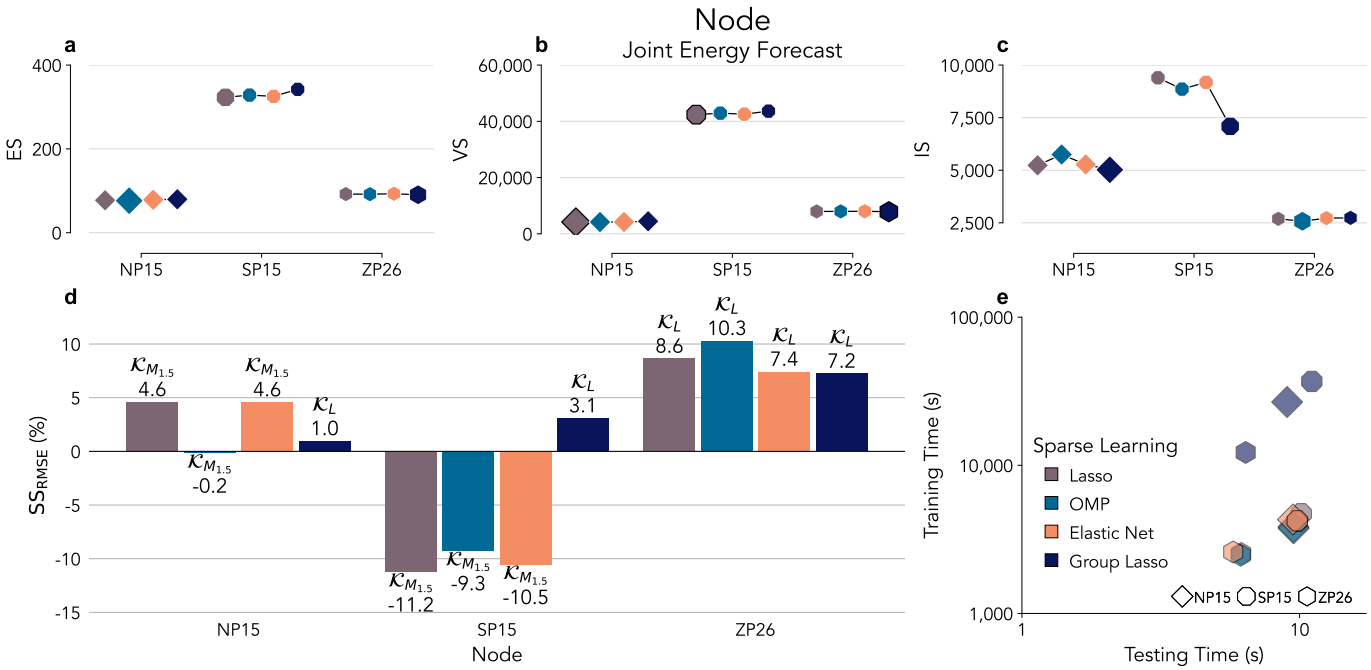

Figure SI16: Joint multi-resource day-ahead probabilistic forecast model section by lower Energy Score (ES). (a) ES, (b) Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$), and (c) IS. (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

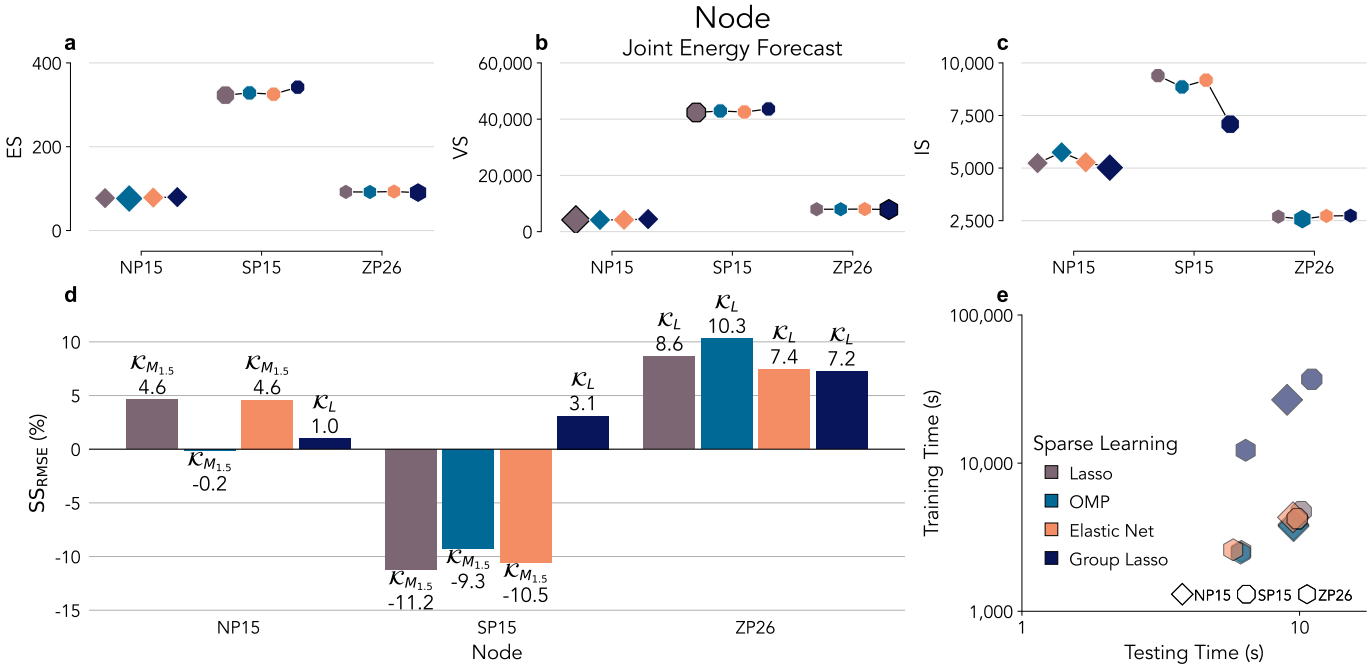

Figure SI17: Joint multi-resource day-ahead probabilistic forecast model section by lower Variogram Score with $p = 0.5$ ($\text{VS}^{0.5}$). (a) Energy Score (ES), (b) $\text{VS}^{0.5}$, and (c) Ignorance Score (IS). (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.

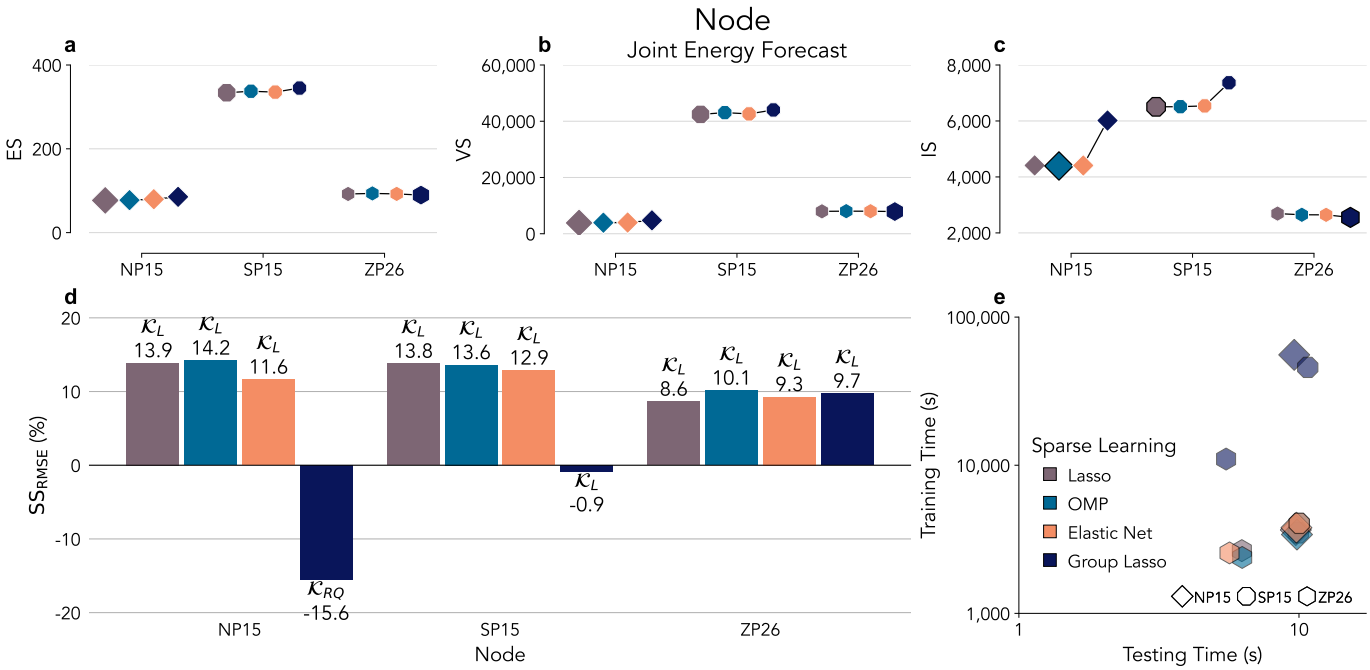

Figure SI18: Joint multi-resource day-ahead probabilistic forecast model section by lower Ignorance Score (IS). (a) Energy Score (ES), (b) Variogram Score with $p = 0.5$ ($VS^{0.5}$), and (c) IS. (d) Skill Score based on RMSE (SS_{RMSE}) relative to CAISO's forecast, and (e) training and testing computing time.