Short-Term Solar Irradiance Forecasting Using Calibrated Probabilistic Models

Eric Zelikman[*1], **Sharon Zhou**[*1], **Jeremy Irvin**[*1]
Cooper Raterink[1], **Hao Sheng**[1], **Anand Avati**[1]
Jack Kelly[2], **Ram Rajagopal**[1], **Andrew Y. Ng**[†1], **David Gagne**[†3]
[1]Stanford University, [2]Open Climate Fix, [3]National Center for Atmospheric Research
{ezelikman,sharonz,jirvin16,crat,avati}@cs.stanford.edu,haosheng@stanford.edu
jack@openclimatefix.org, ramr@stanford.edu, ang@cs.stanford.edu, dgagne@ucar.edu

Abstract

Advancing probabilistic solar forecasting methods is essential to supporting the integration of solar energy into the electricity grid. In this work, we develop a variety of state-of-the-art probabilistic models for forecasting solar irradiance. We investigate the use of post-hoc calibration techniques for ensuring well-calibrated probabilistic predictions. We train and evaluate the models using public data from seven stations in the SURFRAD network, and demonstrate that the best model, NGBoost, achieves higher performance at an intra-hourly resolution than the best benchmark solar irradiance forecasting model across all stations. Further, we show that NGBoost with CRUDE post-hoc calibration achieves comparable performance to a numerical weather prediction model on hourly-resolution forecasting.

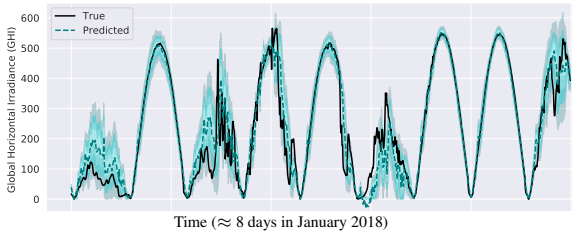

Figure 1: Several days of 30-min horizon NGBoost irradiance forecasts on the Sioux Falls station [1].

1 Introduction

Increasing adoption of renewable energy in the electricity sector is essential to the reduction of anthropogenic greenhouse gas emissions, but significant production increases must occur in order to phase out fossil fuel use [2, 3]. Solar power production has grown dramatically, ranking in the top two electricity-generating capacity additions over the past seven years in the U.S., with 40% of new electric capacity in the grid coming from solar in 2019 [4]. However, due to the high volatility and intermittency of solar energy, solar forecasting methods have become necessary to increase the penetration of solar power into the grid while ensuring power system cost-effectiveness and security [5]. These methods often forecast solar irradiance, which is widely available and correlates strongly with solar photovoltaic (PV) output [6]. While solar forecasting has received extensive attention in the literature [7, 8, 5, 9], it was noted by [10] that the overwhelming majority of solar forecasting methods are not probabilistic. Solar forecasting methods that characterize uncertainty have the potential to aid real-time grid integration of solar energy and help gauge when to deploy new storage [11, 12]. For example, anticipating periods of high uncertainty in solar power production may allow a utility

[*]Equal contribution.
[†]Equal contribution.

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020

to proactively store energy [11] and anticipating periods of low uncertainty can reduce the need for operating reserves, which are often gas-powered [13].

Recent advancements in probabilistic modeling and uncertainty estimation present an opportunity to substantially improve solar irradiance forecasting. The development of probabilistic solar irradiance forecasting models is advancing [14, 15], but most of these methods use numerical weather prediction (NWP) and traditional statistical approaches for producing probabilistic outputs [16, 17]. Moreover, many of the traditional approaches like NWP models are primarily useful on hourly timescales and introduce extensive computational overhead, whereas the best choice for short-term, intra-hourly forecasting has remained open [18]. Deep learning methods for probabilistic solar forecasting have recently begun to emerge, but have not been widely adopted as they have not yet demonstrated superior performance over traditional methods [19]. Furthermore, recent advances in machine learning have developed post-hoc calibration techniques for encouraging well-calibrated predictions [20, 21], but these methods have yet to be used in solar irradiance forecasting [18].

In this work, we develop and validate several probabilistic solar irradiance forecasting models using public data from seven meteorological stations in the U.S. [1]. We compare the performance of the probabilistic models together with several modern post-hoc calibration methods to state-of-the-art benchmarks from [18]. We demonstrate that the best model outperforms the alternatives across all seven stations for intra-hourly forecasting and performs comparably to numerical weather prediction (NWP) models on hourly-resolution forecasting. This work advances solar irradiance forecasting by developing state-of-the-art probabilistic models and associated post-hoc calibration techniques for significant improvements in solar forecasting.

2 Methods

Data. We use public data from NOAA's Surface Radiation (SURFRAD) network [1], consisting of seven stations throughout the continental United States that measure a variety of meteorological variables. Relative humidity, wind speed, wind direction, air pressure, time of day, solar zenith angle, air temperature, and the five previous irradiance values up to the forecast time are used as input, and daytime global horizontal irradiance (GHI) as output. We compute the clearness index as the ratio of the measured terrestrial GHI values to extraterrestrial radiation [22]. Extraterrestrial radiation estimates are taken from the publicly available CAMS McClear Sky historical clear sky irradiance estimates at the stations [23], as performed in [18]. Using the clearness index allows the model to account for predictable changes in irradiance as a function of deterministic factors and results in a stationary time-series [22]. We use 2016 SURFRAD data to learn model parameters, 2017 data to fit calibration method parameters, and 2018 data to evaluate and compare the models.

Models. We develop four probabilistic models which output a probability distribution over the outcome space instead of a point prediction: a Gaussian process regression model, a neural network with uncertainty based on dropout variation (Dropout Neural Network), a neural network whose predictions parameterize a Gaussian distribution optimized to maximize likelihood (Variational Neural Network), and a decision-tree based model using natural gradient boosting (NGBoost) assuming a Gaussian output distribution [24]. Hyperparameters are provided in the appendix.

Calibration Methods. We explore the use of post-hoc probabilistic calibration methods for encouraging well-calibrated predictions, as they have long been shown to improve the performance of weather models [25]. Post-hoc calibration methods aim to maximize the calibration score (or reliability) and

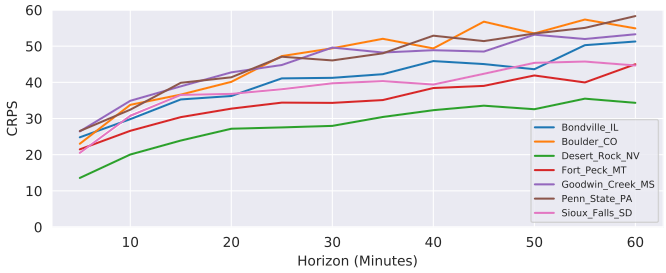

Figure 2: CRPS as a function of forecast horizon for the best model (uncalibrated NGBoost).

Station	Gaussian Process				Dropout Neural Network				Variational Neural Net				NGBoost			
	None	MLE	C	Kul.	*None*	MLE	C	Kul.	*None*	MLE	C	Kul.	*None*	MLE	C	Kul.
Bondville, IL	101.3	53.2	48.5	48.6	48.5	46.0	43.6	44.0	42.0	42.0	41.8	41.9	**40.5**	**40.5**	40.6	40.6
Boulder, CO	110.9	61.7	56.4	56.5	59.3	55.8	53.3	53.9	48.6	48.9	48.3	48.6	**45.9**	46.1	46.0	46.2
Desert Rock, NV	96.6	44.3	35.4	35.7	37.2	40.8	36.1	36.2	31.4	32.5	30.0	30.3	27.9	30.1	**27.8**	28.2
Fort Peck, MT	97.5	50.7	43.6	43.4	41.6	41.9	38.9	39.0	37.9	46.8	37.5	37.6	**34.8**	35.2	35.0	34.9
Goodwin Creek, MS	119.2	59.8	54.7	54.9	57.9	53.3	51.6	51.5	46.9	46.9	46.7	46.9	**44.8**	45.0	**44.8**	45.1
Penn State, PA	111.6	58.8	53.9	53.3	56.5	51.2	49.5	48.0	47.4	47.4	47.3	47.0	**46.0**	46.6	46.1	**46.0**
Sioux Falls, SD	107.2	54.4	49.3	49.5	48.0	46.0	43.4	43.7	43.8	41.8	42.4	43.0	**37.9**	39.1	38.0	38.4

Table 1: **Per-Station Model-Calibration Performance.** Average test CRPS of the probabilistic models with different calibration methods across all seven stations, averaged over each 5-minute horizon from 5 minutes to an hour. The columns correspond to the calibration methods discussed in Section 2, with Kul. for Kuleshov [20] and C for CRUDE [21].

sharpness (or resolution) of a probabilistic model using a held-out calibration dataset [26, 21]. The calibration score penalizes overconfident or underconfident models; for example, one can evaluate the uniformity of the observations' percentiles within their predicted probability distributions [26]. We evaluate three post-hoc probabilistic calibration methods, which we refer to as the Kuleshov method [20], CRUDE [21], and the MLE method inspired by [27]:

- The Kuleshov method is inspired by Platt scaling: the method employs isotonic regression on a calibration dataset, such that a calibrated prediction at a requested percentile p finds the model-predicted percentile $\hat{p}$ which was greater than the requested portion p of the calibration data, and returns the model prediction at $\hat{p}$ [20].
- CRUDE calculates the z-scores of all errors on a calibration dataset $(\hat{\mu}(x) - y)/\hat{\sigma}(x)$, where $\hat{\mu(x)}$ is the predicted mean, y is the observation, and $\hat{\sigma}(x)$ is the predicted standard deviation [21]. Then, to make a prediction at a given percentile, the calibration-set z-score at the percentile is calculated and multiplied by the predicted standard deviation. A constant shift to $\hat{\mu}(x)$ is learned to maximize the calibration score on the calibration set [21].
- The MLE method calculates the constant shift and scale of the predicted distributions derived with maximum likelihood estimation, assuming a Gaussian distribution. This is inspired by [27], which proposed that a simpler maximum likelihood approach was less prone to overfit.

Performance Metrics. All models are trained and calibrated on the clearness index, and evaluated on the irradiance, as in [18]. We primarily evaluate the probabilistic models using the continuous ranked probability score (CRPS), a widely used metric which balances calibration and sharpness [28]. After being shown to be effective for evaluating probabilistic forecasting in other meteorological contexts [29], CRPS has been recently proposed as a standard performance metric for probabilistic irradiance forecasting [30, 18]. [29] presents an intuitive definition of CRPS as the area between each predicted CDF and a step function at the observed value. This requires calibration and rewards sharpness, while being less sensitive to outliers than MLE [29]. We also record the calibration and sharpness metrics discussed in [20] and [21]. A visualization of the calibration curves for the four models using various post-hoc calibration procedures is shown in Figure A1.

3 Results

For intra-hourly forecasting which evaluates each five minute forecast horizon up to an hour, NGBoost attained the best test CRPS scores across all seven stations. For hourly-resolution forecasting, which evaluates each hourly forecast up to six hours, NGBoost performed comparably to the NWP models. Post-hoc calibration improved the test CRPS scores of all models except for intra-hourly NGBoost. As shown in Figure 1, where each color corresponds to an additional 10% interval (10th-90th percentiles), NGBoost quickly responds to changes in uncertainty, performing well under high and low uncertainty. Additionally, as highlighted in Figure 2, there is a steady increase in CRPS as the horizon increases, with stations being consistent in terms of their relative difficulty, e.g., Desert Rock is consistently predictable across methods and horizons, and Penn State and Boulder are more difficult.

Intra-hourly. On the intra-hourly forecasting task, the NGBoost models outperformed the best models in [18] across all stations. The test CRPS scores of all models, with all post-hoc calibration methods and without, are comprehensively reported in Table 1. In [18], the Markov-chain Mixture model (MCM), outperformed several high-quality baselines including smart persistence ensembles and a Gaussian error distribution. The CRPS scores of NGBoost and MCM on each of the stations are shown in Table 2. Performance improvement of NGBoost over MCM ranged from 5% to over 15%.

	CH-P	PeEn	MCM	**NGB**	%Δ
Bondville, IL	92.1	52.8	48.7	**40.5**	-16.8%
Boulder, CO	91.3	61.6	51.6	**45.9**	-11.0%
Desert Rock, NV	47.3	35.2	29.4	**27.9**	-5.1%
Fort Peck, MT	77.0	46.3	39.8	**34.8**	-12.6%
Goodwin Creek, MS	98.4	59.7	52.5	**44.8**	-14.7%
Penn State, PA	98.1	60.0	53.0	**46.0**	-13.2%
Sioux Falls, SD	86.8	47.8	41.0	**37.9**	-7.6%

	CH-P	GAU	NWP	NGB (+*C*)
Bondville, IL	78.1	52.7	**50.8**	53.1 (52.9)
Boulder, CO	75.7	64.2	64.6	**60.3** (**60.4**)
Desert Rock, NV	37.7	42.5	39.2	**36.1** (**35.8**)
Fort Peck, MT	64.8	49.9	48.0	**46.3** (**46.2**)
Goodwin Creek, MS	82.3	58.3	**56.4**	56.9 (56.6)
Penn State, PA	83.4	**55.1**	57.4	58.8 (58.1)
Sioux Falls, SD	74.3	50.6	**49.7**	58.6 (56.6)

Table 2: **Per-Station CRPS Comparison**. The two tables show intra-hourly and hourly-precision results, respectively, with NGBoost denoted as NGB. +*C* denotes NGBoost with CRUDE post-hoc calibration; other calibration methods underperformed CRUDE, and are included in the appendix. The non-NGBoost scores are taken from the best models in [18], which are CH-P, PeEn, and MCM for intra-hourly and CH-P, GAU, and NWP for hourly forecasting. More details on the [18] models are provided in the appendix. [18] also evaluated their models on SURFRAD stations in 2018 during the daytime, with the same horizons and resolution.

Hourly. On the hourly-resolution task, we primarily compared NGBoost with various post-hoc calibrations to numerical weather prediction (NWP) models, as done in [18] due to the lower temporal resolution of NWP models. NGBoost(+C) was the best model on three stations (CO, NV, MT), the NWP model was the best model on another three stations (IL, MS, SD), and the Gaussian error distribution was the best performing model on the other station (PA). The ECMWF control Gaussian (GAU) error distribution was the best performing model on the other station [18]. The NGBoost results on the hourly-resolution forecasting task in Table 2 suggest that this solution can reach performance which is comparable to that of NWP models.

Calibration. Across post-hoc calibration methods, CRUDE and Kuleshov led to more substantial performance improvements overall than the MLE method. Notably, CRUDE and Kuleshov improve the calibration metric of NGBoost at the expense of sharpness, with CRUDE reducing the average calibration error across all stations from 0.040 to 0.031 but worsening the mean sharpness on the clearness index predictions from 0.192 to 0.215 (lower is better). The calibration of NGBoost may result from several factors: the variational neural network is also well calibrated, but likely slightly overfit - the use of the natural gradient by NGBoost is suggested to reduce overfitting [24]; additionally, the under-confidence of the Gaussian process suggests that investigation of alternative priors may be warranted [31].

4 Discussion

We evaluated several probabilistic models for solar irradiance forecasting combined with recent approaches for post-hoc calibration. We found that NGBoost, without post-hoc calibration, outperformed each of the baseline models across all of the stations at the intra-hourly resolution. Additionally, NGBoost achieved higher performance than two NWP model variants across three stations at the hourly-resolution. Our results suggest that NGBoost is an excellent baseline for probabilistic solar irradiance forecasting at both intra-hourly and hourly resolutions.

More sophisticated data and specialized priors could further improve our results, calling for more research on machine learning-based solar irradiance forecasting. The incorporation of satellite imagery with high temporal frequency[3] has the potential to improve intra-hourly uncertainty estimates [32] by tracking the evolution of clouds which are the primary contributor to the variability of solar irradiance [33]. In addition, the priors used in this work to train the various models were Gaussian, but solar irradiance is never negative, and the clearness index is limited to 1; thus, a truncated normal distribution prior with fixed bounds could be more appropriate. Notably, NGBoost supports arbitrary priors as long as the derivative with respect to the output is calculable.

We believe that the development of probabilistic solar forecasting methods will help enable the level of renewable energy adoption that is necessary to phase out fossil fuel use. Furthermore, advancing probabilistic machine learning models and uncertainty estimation has implications to many problems related to climate change beyond solar forecasting [34]. We hope our work helps to motivate further research in applied probabilistic machine learning research, which we believe is key to building technologies for mitigating climate change.

[3]NOAA's GOES-16 and GOES-17 satellites launched in late 2016 and early 2018, respectively.

Acknowledgments

We would like to thank Cooper Elsworth, Kyle Story, and Rose Rustowicz from Descartes Labs for their help in the early directions of this work.

References

[1] J. A. Augustine, J. J. DeLuisi, and C. N. Long, "Surfrad–a national surface radiation budget network for atmospheric research," *Bulletin of the American Meteorological Society*, vol. 81, no. 10, pp. 2341–2358, 2000.

[2] P. A. Owusu and S. Asumadu-Sarkodie, "A review of renewable energy sources, sustainability issues and climate change mitigation," *Cogent Engineering*, vol. 3, no. 1, p. 1167990, 2016.

[3] D. Elzinga, S. Bennett, D. Best, K. Burnard, P. Cazzola, D. D'Ambrosio, J. Dulac, A. Fernandez Pales, C. Hood, M. LaFrance, *et al.*, "Energy technology perspectives 2015: mobilising innovation to accelerate climate action," *Paris: International Energy Agency*, 2015.

[4] S. E. I. Association and W. M. P. . Renewables, "U.s. solar market insight," tech. rep., Solar Energy Industries Association, Washington, D.C., 2020.

[5] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

[6] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in pv output power forecast," *Solar Energy*, vol. 136, pp. 125–144, 2016.

[7] P. Mathiesen and J. Kleissl, "Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states," *Solar Energy*, vol. 85, no. 5, pp. 967–977, 2011.

[8] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in energy and combustion science*, vol. 39, no. 6, pp. 535–576, 2013.

[9] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif, "A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization," *Renewable and Sustainable Energy Reviews*, vol. 124, p. 109792, 2020.

[10] D. Yang, "A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ropes)," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 2, p. 022701, 2019.

[11] S. E. Haupt, M. G. Casado, M. Davidson, J. Dobschinski, P. Du, M. Lange, T. Miller, C. Mohrlen, A. Motley, R. Pestana, *et al.*, "The use of probabilistic forecasts: Applying them in theory and practice," *IEEE Power and Energy Magazine*, vol. 17, no. 6, pp. 46–57, 2019.

[12] S. Taheri, V. Kekatos, and S. Veeramachaneni, "Energy storage sizing in presence of uncertainty," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1–5, IEEE, 2019.

[13] J. Wu, A. Botterud, A. Mills, Z. Zhou, B.-M. Hodge, and M. Heaney, "Integrating solar pv (photovoltaics) in utility system operations: Analytical framework and arizona case study," *Energy*, vol. 85, pp. 1–9, 2015.

[14] D. W. Van der Meer, J. Widén, and J. Munkhammar, "Review on probabilistic forecasting of photovoltaic power production and electricity consumption," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1484–1512, 2018.

[15] B. Li and J. Zhang, "A review on the integration of probabilistic solar forecasting in power systems," *Solar Energy*, vol. 207, pp. 777–795, 2020.

[16] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied energy*, vol. 157, pp. 95–110, 2015.

[17] P. Lauret, M. David, and H. T. Pedro, "Probabilistic solar forecasting using quantile regression models," *energies*, vol. 10, no. 10, p. 1591, 2017.

[18] K. Doubleday, V. V. S. Hernandez, and B.-M. Hodge, "Benchmark probabilistic solar forecasts: Characteristics and recommendations," *Solar Energy*, vol. 206, pp. 52–67, 2020.

[19] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management*, vol. 198, p. 111799, 2019.

[20] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *International Conference on Machine Learning*, pp. 2796–2804, 2018.

[21] E. Zelikman, C. Healy, S. Zhou, and A. Avati, "CRUDE: Calibrating Regression Uncertainty Distributions Empirically," *Workshop on Uncertainty & Robustness in Deep Learning of the International Conference on Machine Learning (ICML) 2020*, May 2020.

[22] T. Kato, "Prediction of photovoltaic power generation output and network operation," in *Integration of Distributed Energy Resources in Power Systems*, pp. 77–108, Elsevier, 2016.

[23] C. Granier, S. Darras, H. D. van der Gon, D. Jana, N. Elguindi, G. Bo, G. Michael, G. Marc, J.-P. Jalkanen, J. Kuenen, *et al.*, "The copernicus atmosphere monitoring service global and regional emissions (april 2019 version)," 2019.

[24] T. Duan, A. Avati, D. Y. Ding, S. Basu, A. Y. Ng, and A. Schuler, "Ngboost: Natural gradient boosting for probabilistic prediction," *arXiv preprint arXiv:1910.03225*, 2019.

[25] T. R. Stewart, K. F. Heideman, W. R. Moninger, and P. Reagan-Cirincione, "Effects of improved information on the components of skill in weather forecasting," *Organizational behavior and human decision processes*, vol. 53, no. 2, pp. 107–134, 1992.

[26] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 243–268, 2007.

[27] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *arXiv preprint arXiv:1905.11659*, 2019.

[28] G. Candille and O. Talagrand, "Evaluation of probabilistic prediction systems for a scalar variable," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 131, no. 609, pp. 2131–2150, 2005.

[29] A. Avati, T. Duan, S. Zhou, K. Jung, N. H. Shah, and A. Y. Ng, "Countdown regression: sharp and calibrated survival predictions," in *Uncertainty in Artificial Intelligence*, pp. 145–155, PMLR, 2020.

[30] P. Lauret, M. David, and P. Pinson, "Verification of solar irradiance probabilistic forecasts," *Solar Energy*, vol. 194, pp. 254–271, 2019.

[31] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of Mathematical Psychology*, vol. 85, pp. 1–16, 2018.

[32] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, "History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining," *Solar Energy*, vol. 168, pp. 60–101, 2018.

[33] D. P. Larson, M. Li, and C. F. Coimbra, "Scope: Spectral cloud optical property estimation using real-time goes-r longwave imagery," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 2, p. 026501, 2020.

[34] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, *et al.*, "Tackling climate change with machine learning," *arXiv preprint arXiv:1906.05433*, 2019.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[36] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," 2015.

[37] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "Gpytorch: blackbox matrix-matrix gaussian process inference with gpu acceleration," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7587–7597, Curran Associates Inc., 2018.

[38] J. Munkhammar, D. van der Meer, and J. Widén, "Probabilistic forecasting of high-resolution clear-sky index time-series using a markov-chain mixture distribution model," *Solar Energy*, vol. 184, pp. 688–695, 2019.

Appendix

Hyperparameters

Neural Network. We train our neural network models using the Adam optimizer [35] with a learning rate of 1e-4 and for the variational neural network, a weight decay of 1e-2. We use three 1024-wide hidden layers for the dropout-uncertainty neural network and three 256-wide hidden layers for the variational neural network.

Gaussian Process. We train our Gaussian process using the method proposed in [36], using GPyTorch [37], under a Gaussian probability distribution, with 1000 batches of 1000 points each, with early stopping after 10 batches without improvement.

NGBoost. For short term forecasting, we train NGBoost with the default hyperparameters and 2000 estimators. For hourly forceasting, we also use a mini-batch fraction of 0.5.

Evaluation Details. Each model was trained on each time horizon 10 times, and then calibrated and tested on random samples of $2,000$ points for each year. Only daytime data was evaluated (based on times when clear-sky irradiance was positive, as in [18]), but data from before sunrise was used for early-morning forecasts. On average, there are about $52,000$ daylight data points per station per year for intra-hourly forecasting, varying across stations and years by less than $\pm 1,000$.

Models from Doubleday [18]

We compare to scores of the following models from [18], though [18] includes further elaboration on each model:

- **CH-P.** CH-P corresponds to the complete-history persistence ensemble. It uses the historical distribution of clearness index values at a particular time of day at a particular station.
- **PeEn.** PeEn corresponds to the persistence ensemble. Unlike the complete-history persistence ensemble, it uses only the most recent examples. For the intra-hourly case, this uses the clearness index values of the past two hours.
- **MCM.** The Markov-chain mixture (MCM) model was first proposed in [38]. MCM attempts to model implicit states and their associated transition probabilities.
- **NWP.** The numerical weather prediction (NWP) model used in [18] was an ensemble of the 51 predictions by the European Centre for Medium-Range Weather Forecasts (ECMWF) predictions (the control forecast and the 50 perturbed forecasts).
- **GAU.** The Gaussian error distribution for the hourly forecasts exclusively uses the ECMWF unperturbed control forecast, alongside a standard deviation derived from the distribution of errors for the same time of day over the course of a year. It uses a double-truncated Gaussian distribution to exclude negative or very positive clearness index values.

Calibration Curves

We include some example calibration curves to help visualize the impact of post-hoc calibration on different models.

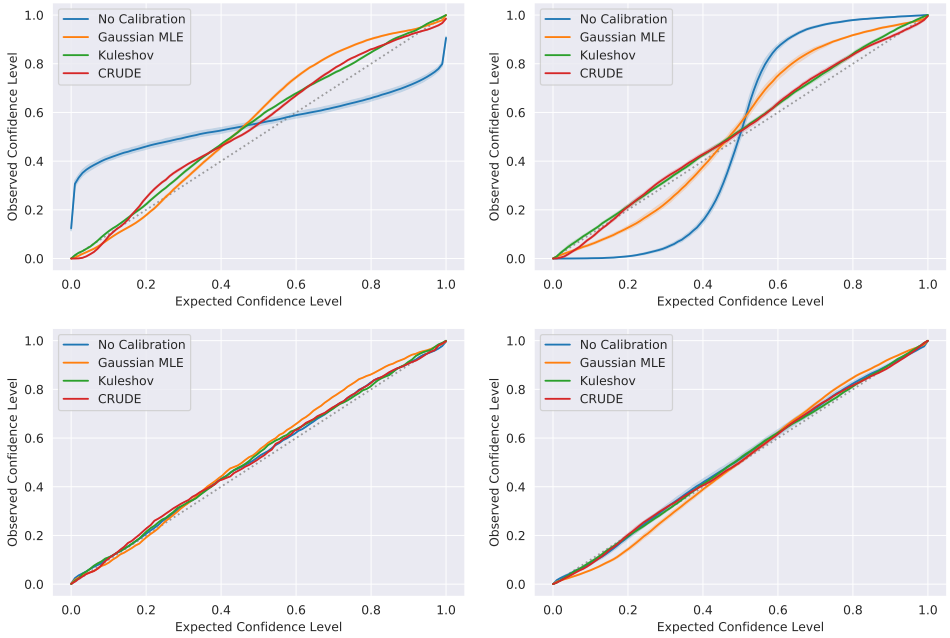

Figure A1: Calibration curves for the dropout neural network (top left), the Gaussian process (top right), NGBoost (bottom left), and the variational neural network (bottom right) on the Penn State, PA SURFRAD station with a 30 minute horizon.

Hourly Forecasting with NGBoost

	None	MLE	CRUDE	Kuleshov
Bondville, IL	53.1	53.8	**52.9**	53.0
Boulder, CO	**60.3**	68.0	60.4	61.7
Desert Rock, NV	36.1	37.3	**35.8**	37.0
Fort Peck, MT	46.3	46.3	**46.2**	46.9
Goodwin Creek, MS	56.9	56.8	**56.6**	57.8
Penn State, PA	58.8	**58.1**	**58.1**	58.3
Sioux Falls, SD	58.6	57.8	**56.6**	57.7

Table A1: **Hourly NBoost forecast CRPS by station**. A comparison of the CRPS of each calibration method applied to NGBoost on each station.

As mentioned in the main text, CRUDE consistently outperformed the other calibration methods. Curiously, while there was no significant positive impact by calibration on NGBoost for intra-hourly forecasting, there was a notable improvement in the hourly forecasting case. It is possible that there is less yearly variation in this longer timescale, and thus calibration is more effective. This suggests that training and calibrating over a period of more years would improve performance. Note that the difference between the intra-hourly curve and the hourly curve at a horizon of one hour comes from aggregation: the intra-hourly metric evalautes the 5-minute average CRPS, while the hourly metric predicts the one-hour average CRPS.

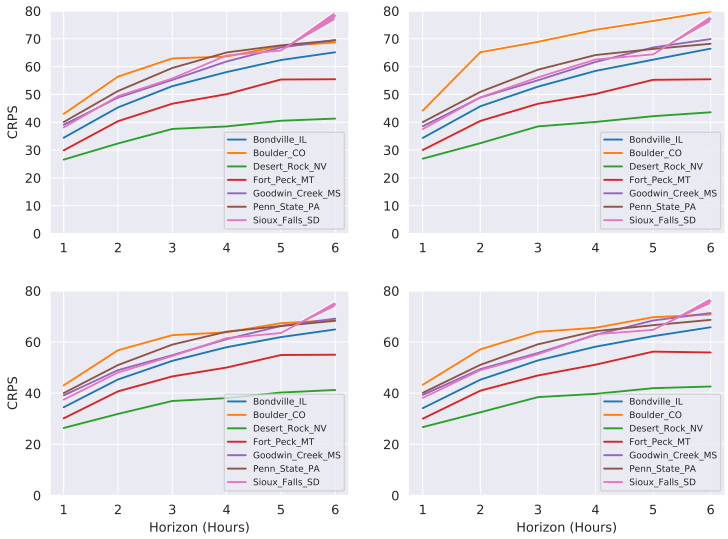

Figure A2: Performance of various hourly NGBoost calibration methods over all stations. These plots also show the variance in CRPS scores over each of a model's 10 evaluations at a station and horizon - notably this $\pm 1\sigma$ is generally not visible for NGBoost.

Calibration and Sharpness

Calibration

	Gaussian Process				Dropout Neural Network				Variational Neural Net				NGBoost			
	None	MLE	C	Kul.	None	MLE	C	Kul.	None	MLE	C	Kul.	None	MLE	C	Kul.
Bondville, IL	0.19	0.08	0.01	0.03	0.15	0.07	0.02	0.04	0.04	0.05	0.02	0.03	0.04	0.05	0.02	0.03
Boulder, CO	0.18	0.09	0.03	0.04	0.14	0.07	0.03	0.05	0.05	0.07	0.04	0.05	0.04	0.06	0.03	0.03
Desert Rock,NV	0.21	0.13	0.04	0.04	0.10	0.11	0.03	0.04	0.10	0.12	0.04	0.04	0.05	0.11	0.03	0.04
Fort Peck, MT	0.19	0.10	0.02	0.03	0.13	0.09	0.03	0.04	0.05	0.13	0.02	0.02	0.03	0.05	0.02	0.03
Goodwin Creek, MS	0.19	0.08	0.02	0.03	0.15	0.07	0.02	0.03	0.05	0.05	0.03	0.03	0.03	0.05	0.03	0.04
Penn State, PA	0.19	0.08	0.02	0.03	0.17	0.08	0.04	0.04	0.04	0.05	0.03	0.03	0.02	0.05	0.03	0.03
Sioux Falls, SD	0.19	0.08	0.04	0.05	0.19	0.10	0.09	0.09	0.12	0.07	0.10	0.10	0.07	0.08	0.07	0.08

Table A2: **Per-Station Calibration Comparison**. We include a comparison of the stations and their calibration scores. Lower scores correspond to better calibration. The abbreviations correspond to those used in Table 1.

Sharpness

	Gaussian Process				Dropout Neural Network				Variational Neural Net				NGBoost			
	None	MLE	C	Kul.	None	MLE	C	Kul.	None	MLE	C	Kul.	None	MLE	C	Kul.
Bondville, IL	0.99	0.36	0.28	0.35	0.06	0.25	0.23	0.25	0.21	0.24	0.22	0.24	0.20	0.23	0.21	0.22
Boulder, CO	1.02	0.39	0.31	0.39	0.06	0.27	0.25	0.26	0.25	0.30	0.25	0.30	0.21	0.25	0.22	0.25
Desert Rock,NV	0.91	0.31	0.25	0.30	0.07	0.21	0.20	0.21	0.22	0.26	0.21	0.26	0.16	0.27	0.19	0.27
Fort Peck, MT	1.07	0.42	0.29	0.39	0.07	0.27	0.26	0.26	0.26	0.60	0.19	0.24	0.19	0.24	0.21	0.24
Goodwin Creek, MS	1.10	0.38	0.30	0.37	0.06	0.26	0.25	0.26	0.23	0.25	0.23	0.24	0.20	0.22	0.21	0.22
Penn State, PA	1.07	0.40	0.31	0.40	0.06	0.27	0.25	0.27	0.23	0.27	0.24	0.27	0.20	0.24	0.22	0.23
Sioux Falls, SD	1.08	0.39	0.30	0.38	0.07	0.26	0.26	0.26	0.20	0.30	0.23	0.28	0.19	0.27	0.23	0.27

Table A3: **Per-Station Sharpness Comparison**. We include a comparison of the stations and their sharpness scores. We report sharpness in terms of clearness index. Note that sharpness is only meaningful under a well-calibrated model. Lower scores correspond to sharper models. The abbreviations also correspond to those used in Table 1.

As discussed in the main text, we analyzed and compared sharpness and calibration metrics, based on [20, 21]. In general, NGBoost is both slightly better calibrated and sharper than the variational neural network model, which corresponds to their CRPS performance. Generally, with post-hoc calibration, the variational neural network and NGBoost were sharper than the Gaussian process and dropout-uncertainty based neural network. We also see that the Gaussian process model is underconfident without post-hoc calibration, while the dropout-uncertainty neural network is overconfident.