Perceiver Model Ensemble for Solar Power Forecasting: A Winning Solution for ClimateHack.AI 2023-2024

Trevor Yu[1]*, Carter Demars[2]*, Areel Khan[3]*
[1]Systems Design Engineering, University of Waterloo
[2]Mechanical and Mechatronics Engineering, University of Waterloo
[3]Department of Statistics and Actuarial Science, University of Waterloo
{trevor.yu, cdemars, areel.khan}@uwaterloo.ca.ca

Abstract

In this paper, we present Team Waterloo's winning approach for solar power forecasting in ClimateHack.AI 2023-2024, an international machine learning competition. Our model leverages Numerical Weather Prediction (NWP), high-resolution visible (HRV) satellite imagery, and solar panel site metadata to predict photovoltaic (PV) power output over a 4-hour window. Our solution was an ensemble of Perceiver models that used spatial semantic pointers for spatial-temporal encoding, dynamic cropping, and efficient data handling. Our model can provide low-latency, high-accuracy forecasts and achieved a mean absolute error of 0.081 on the competition test set.

1 Introduction

Solar power forecasting is an important task for electricity network operators. It is one of four major areas of energy forecasting, along with wind forecasting, electric load forecasting, and electricity price forecasting [1].

The infrastructure that connects solar panels to the power grid is complex and requires accurate estimates of how much photovoltaic (PV) energy will be available in the future. Unlike non-renewable energy sources, renewable power generation from wind and solar can vary drastically in the short-term. In the case of solar power, cloud cover has a direct and immediate effect on PV output but is very difficult to predict. Energy systems operators (ESOs) must consider the sum of renewable and non-renewable power generation to ensure that electricity supply always matches demand.

To maintain a balanced power grid when facing variable weather conditions, ESOs leverage natural gas turbines to supplement renewable sources during sudden drops in power production. These generators are kept idling because they take several hours to start up from cold, and consequently they produce a significant amount of carbon dioxide [2]. When fluctuations occur in solar power, the natural gas generators, called spinning reserves, ramp up quickly to their capacity. Improving the accuracy of existing solar forecasting models will improve our ability to schedule spinning reserves, thereby reducing our reliance on non-renewable energy sources to satisfy these short-term power deficits.

Open Climate Fix (OCF) is a non-profit product lab funded by Google and NVIDIA that seeks to reduce greenhouse gas emissions as rapidly as possible. OCF places particular emphasis on solar energy forecasting as a research area with high potential for climate impact. In Great Britain, for instance, the National Grid Operator could reduce its carbon emissions by 100 kilotons per year using better PV nowcasting [3]. Globally, this translates to approximately 50 megatons of CO2 per year by 2030.

1.1 ClimateHack.AI 2023-2024

In November 2023, Open Climate Fix announced ClimateHack.AI 2023-2024 (henceforth referred to as ClimateHack), a global machine learning competition for PV forecasting. The goal was to advance the state-of-the-art in site-level solar forecasting over a 4-hour prediction window. The competition targets the United Kingdom as the region of interest due to its cloudy weather, which leads to high variability in PV production. Accurate site-level solar forecasts can be aggregated at the grid supply point level to produce estimates that are directly beneficial to grid operators for the purpose of scheduling reserves. Furthermore, large spinning reserves can take up to four hours to reach their

*All authors contributed equally to the project. Code is available at https://github.com/AreelKhan/waterloo-climatehack.

capacity, matching the forecast horizon of the competition [2].

Since ESOs benefit most from near-real time PV forecasts, designing a low-latency forecasting pipeline that could ingest and output predictions quickly was a high priority. Only data sources that would be available in a live production setting could be utilized.

It is difficult to define what is considered state-of-the-art in solar forecasting. Photovoltaic power forecasts can be performed at the national level, grid supply point (GSP) level, or site level. Power production can also be forecasted over different time horizons and at different intervals. Forecast errors generally increase with higher spatial resolutions, over longer time horizons, and with shorter forecast frequencies. Moreover, existing solar forecasting models have not been benchmarked on an industry-standard dataset. Error metrics can be easily biased by including more test values in the early morning or evening, when the magnitude of power production is generally lower, yielding lower prediction errors.

Open Climate Fix had previously achieved state-of-the-art results in GSP-level forecasting using the same dataset, achieving 6.34 normalized MAE [4]. This, however, was for 2-hour forecasting at 30-minute intervals, and their models had access to five years of training data. GSP-level solar generation values have significantly less variability than those from individual PV systems, so site-level forecasting is a more challenging task.

This paper outlines Team Waterloo's final submission to the ClimateHack competition. The objective of the competition was to predict a 4-hour forecast of a solar panel site's normalized PV output at 5-minute intervals, given the past hour of historical data and predicted weather data over the forecast window. Model performance was evaluated using mean absolute error (MAE) between model predictions and the true PV site production on a held-out test set that was hidden to the competitors.

2 Solution Formulation

The largest determinant of solar power production is incident solar irradiance—essentially, the amount of sunlight that strikes the surface of the solar panel [1]. This irradiance depends on both weather conditions and calendar variables. Calendar variables, such as the angle of the sun in the sky at a given time of day can be accurately calculated in advance, whereas weather conditions require forecasting.

In our solution, Numerical Weather Prediction (NWP) data generated by the DWD ICON-EU Forecast model [5, 6] was utilized to estimate 38 weather conditions including temperature, precipitation, wind speed, and air pressure. This NWP data is suitable for PV forecasting, even for live inference, since it provides forecasts over extended time horizons. However, as NWP is derived from physical models rather than direct measurements, its accuracy in PV forecasts is limited by spatial and temporal resolution. Currently, NWP data is available at hourly intervals at a 5-km resolution, which poses a challenge for generating precise, site-specific forecasts at a 5-minute frequency.

Satellite imagery provides information on cloud cover surrounding the site. However, only historical imagery can be used to model future cloud cover, and the trajectories of clouds are difficult to predict into the future and are only available in real-time during a live inference setting.

The exact spatial orientation and position of clouds relative to solar sites are critical for PV site-level forecasting. Unlike many computer vision tasks where the presence of a feature matters more than its precise location, determining the location of clouds and weather features relative to the site is essential. Spatial information for each pixel of the image was included to counteract the translation invariance properties of traditional CNN models, which could otherwise limit the ability of our model to make effective use of satellite imagery.

Photovoltaic power generation data is used as the regression target for model training. The previous hour of power generation is known at inference time, to inform the next four hours of prediction.

We use encodings for features like time of day, day of year, solar elevation, and azimuth, that respect the periodicity in these features over time. Additionally, site-specific metrics like average and maximum monthly outputs were computed for each site, allowing the model to learn both site-specific idiosyncrasies and seasonal PV and solar irradiance trends.

The crops of weather and satellite data used for inference were significantly reduced in size. Smaller crop sizes reduced ingestion time, model size and inference time without affecting model performance. Further, using solar azimuth and elevation data, a method henceforth referred to as dynamic cropping, was implemented to select crops directly in line-of-sight between the sun and the solar panel site, capturing only the most relevant regions.

3 Data

3.1 Datasets

The dataset available to the participants of ClimateHack in the evaluation environment was comprised of several

modalities, including satellite imagery at various resolutions and imaging wavelengths, numerical weather prediction (NWP) data, site metadata, and historical PV generation output. The evaluation environment used data from 993 sites from the year 2022. The same modalities were available to the participants for training, but the data represented times between January, 2020 to December, 2021 and covered the same 993 sites.

The PV generation data was collected from live PV systems across the UK from the years 2018 to 2021. Each site had a time-series of the average power generated at the site in watts resampled at 5-minute intervals and normalized as a proportion of the maximum installed capacity. This resulted in values between 0 and 1 for all solar sites provided in the dataset. Each site also had associated metadata, including its latitude and longitude, its orientation relative to due north, the tilt of its panels relative to the horizon, and its maximum installed capacity in kilowatts.

NWP data for the regions surrounding the United Kingdom were taken from the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). The NWP dataset contained predictions for 38 weather variables, such as wind speeds, air pressure, temperature, and cloud cover using the DWD ICON-EU Forecast model [5, 6]. The data presented as a time-series of multi-channel images, with each channel representing a weather variable and each pixel having an associated latitude and longitude coordinate. The NWP data has approximately 5-kilometer spatial resolution and at an hourly temporal resolution.

EUMETSAT had also made available satellite images of the United Kingdom and surrounding regions from the SEVIRI rapid scanning service dataset [7]. Similar to the NWP data, the satellite image data was also presented as a time-series of multi-channel images. The satellite images were captured at 5-minute intervals. The data contained one channel of high-resolution visible (HRV) imagery, which had a spatial resolution of approximately 1 kilometer. There were also 11 non-high-resolution visible (non-HRV) channels, which included visible and infrared wavelenghts. The non-HRV data had a spatial resolution of approximately 3 kilometers. All provided satellite imagery data were scaled between 0 and 1.

The dataset for the competition can be accessed through a Huggingface Datasets repository.

3.2 Data Preprocessing

While the PV generation, HRV satellite, and non-HRV satellite data have been scaled between 0 and 1, the NWP data was not. To address scaling, we added a batch norm layer [8] as the first layer of the model to normalize each weather feature based on the running statistics of the dataset.

To encode timestamps, we encoded the time of year and time of day as separate variables scaled between 0 and 1. The time of year variable was created by dividing the of the year by 365. The time of day variable was created by dividing the minute of the day by 1440.

The apparent solar azimuth and apparent solar elevation variables were computed using the `pvlib` Python package over the history and forecast windows for the location of the solar site. These angles in degrees, along with the solar site orientation and tilt were scaled between 0 and 1 by dividing by 360.

We used the `cartopy` Python package to convert grids in coordinate systems of the satellite and NWP data to latitude and longitude. The continuous latitude and longitude features were processed by further layers in the models, unscaled.

We also included site-specific statistics computed over the historical PV generation data from 2018–2021. We computed the average monthly power output, IQR for power, average intraday variance, 95th and 99th percentile outputs for each site in the dataset.

3.3 Data Splits and Batching

We held out entire days worth of data when constructing the validation dataset. Weather and cloud cover can change significantly overnight when predictions are not being made, so data from separate days is less correlated compared to data from consecutive hours. Creating validation data from temporally correlated examples would be less informative of model performance on competition test data, which was from an entirely different year. We held out 30% of the days from the training data for validation.

Creating training batches from data from different days was not feasible due to the underlying format of the large dataset being slow to read from disk for random points in time. Instead, we chose a random forecast start time between the hours of 10am to 4pm, using 5-minute intervals, and loaded data for the entire region for the corresponding history and forecast window. We then randomly selected 64 solar sites over the UK for that timestamp to create a batch. Due to the large number of possible timestamps and available sites, each consecutive batch is still highly random, even though the items within a batch are temporally correlated.

4 Modelling

4.1 SSP Encoding

To encode space and time, we used a technique called spatial semantic pointers, or SSPs [9]. This method can encode continuous spatial inputs into vectors whose features are random Fourier features [10]. The SSP operator, ϕ, is parameterized by a phase encoding matrix, $\boldsymbol{\theta} \in \mathbb{R}^{d_i \times d_e}$ that projects inputs $\boldsymbol{x} \in \mathbb{R}^{d_i}$ into a phase vector the size of the encoding dimension, d_e, as shown in Equation 1.

$$\phi(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{F}^{-1}\left\{e^{2\pi j(\boldsymbol{x}\boldsymbol{\theta}^\top)}\right\} \tag{1}$$

A desirable property of SSPs is that their dot product similarity is high when input values are spatially close together. With dot product attention used in transformers, SSPs ensure nearby pixels in both space and time will have high attention scores. This is especially important since the data has different spatial and temporal resolutions, and coordinate systems between modalities may not align on a perfect grid. Standard methods of ensuring spatial correlation such as convolution may not be appropriate for this kind of data.

Because SSP encodings can apply to any continuous-valued input data, we used them to transform continuous metadata features into $d-$dimensional vectors. We applied SSPs to the spatial pixel encoding and all continuous metadata encodings. We used a special formulation of the $\boldsymbol{\theta}$ phase vector for encoding time of day, time of year, solar elevation and solar azimuth that maintains their cyclical representations.

4.2 Perceiver Backbone

The backbone of our architecture was based on the Perceiver transformer architecture [11, 12]. The first layer of the Perceiver uses a cross-attention layer between a latent array of random vectors and an input array. This allows for a long input sequence length to be compressed into a shorter latent sequence to process through a deep transformer encoder.

For the transformer encoder backbone, we used a pre-normalization architecture [13], no bias on any linear layers, and swish activation function [14] in feed-forward network (FFN) layers. We used a hidden size $d = 64$, 8 attention heads, a FFN expansion factor of 4, and 20 transformer layers. Overall, the model had 1.1 million parameters.

4.3 Model Input Format

We formatted our latent array to represent the forecast window and any information relating to the historical PV production. We had each sequence element in the latent array represent a 5-minute interval in the historical and prediction window. Each element was composed of a sum of spatial and temporal encodings for the site over the forecast window, along with the solar positions, since this is easily computed over the forecast window. The historical window also included encodings for the historical PV generation. To this sequence, we concatenated encodings of the static site metadata (tilt, orientation, maximum capacity, site PV statistics) and some random vectors. After all input transformations, the latent array has a shape of $56 \times d$. The predictions made on the 48 elements of the final hidden states that represented the forecast window are used as the model outputs. The components of the latent array and its outputs are shown in Figure 1.

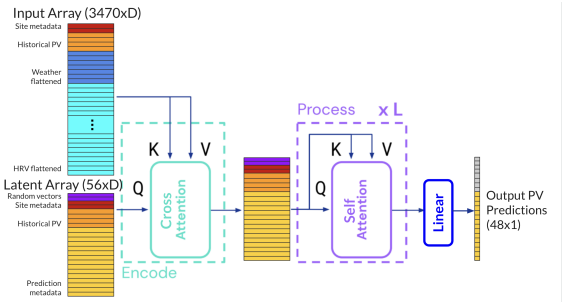


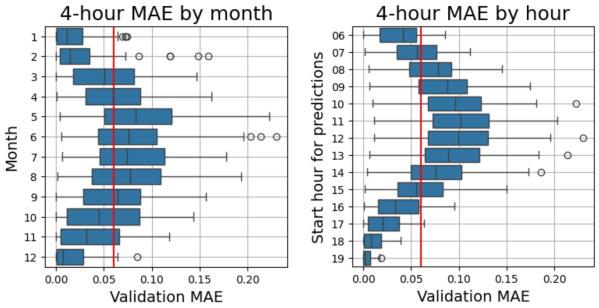
Figure 1: Perceiver architecture diagram.

We formatted the input array to represent the larger image modalities. HRV and NWP data were cropped to 16×16 pixels and 8×8 pixels respectively, centered on the site. A linear projection from the modality channel dimension to the hidden dimension was then applied. Each tensor had a per-pixel spatial encoding and a temporal encoding added and broadcasted along the appropriate dimensions. The resulting tensors were flattened and concatenated with representations for the historical PV and static site metadata as described previously. After all input transformations, the input array has a shape of $3470 \times d$. The components of the input array are shown in Figure 1.

Through the construction of the latent and input arrays, the site location and time of forecast serve as queries against different locations and times from HRV and NWP data in the input array. This is where the SSP encodings help attention scores select relevant input pixels for each site and time in the latent array.

4.4 Model Training

The model was implemented using PyTorch. All layers of the model were trained with the AdamW [15, 16] optimizer with $\beta_1 = 0.95, \beta_2 = 0.99, \lambda = 0.001$. The learning rate was scheduled using the OneCycleLR pol-



(a) MAE by month of year (b) MAE by forecast start time

Figure 2: Best individual model performance on validation dataset. The red line represents the average error across all examples.

icy [17] with a peak learning rate of 5e-4 with 30% warm-up steps and a cosine decay to 1% the peak learning rate. Training used a batch size of 64 over 15 epochs using L1 loss between the model outputs and regression targets.

5 Results

The best performing individual model had a MAE score of 0.08253 on the competition test set and a MAE score of 0.06505 on our validation dataset.

We suspected the model's performance depends on the variability of the solar output, so we further examined our model's predictions on the validation dataset during different time periods. Figure 2a shows the models prediction error is highly variable during the summer months, where days are longer and the potential for PV generation is higher. However, the average error was relatively low during the winter months where PV output was consistently lower. We observed a similar pattern in Figure 2b, where forecast periods that occurred during midday had a greater variability in error, while forecast periods that occur during the evening had lower error as the PV output was nearly zero for most of the window.

Based on the success of our individual model, we also tried ensembling variations of this model recipe. We trained the same model recipe but with different seeds and a smaller validation set. We also trained models with data from the summer months, where power production was most variable. Our best-performing ensemble model used a weighted average of four models, with one model from the summer months. This model had a MAE score of 0.08105 on the competition test set. In particular, ensembling was possible because an individual model only had 1.1 million parameters, compared to other competitors whose models had tens of millions of parameters each.

For inference metrics, our model completed inference on 1000 sites in 2.5 seconds on an RTX 3060 GPU. This is well within the 5-minute run-time required to make predictions at the required interval.

6 Discussion

Overall, our model achieved the lowest error on the competition test set, while being one of the smallest deep learning models presented in the competition. For reference, the next-best team had a model based on two ResNeXt-50 [18] that separately processed non-HRV and NWP data. Their model had a total of more than 50 million parameters and a test set MAE score of 0.08209. Our ensemble model was 11$\times$ smaller and achieved better performance.

Further analysis of model outputs revealed that the model tended to favour smooth predictions of future power output. This could indicate that the resolution of available inputs is a limiting factor in the model's available to predict short-term fluctuations in power output. We expect the quality of solar power forecasts to increase greatly with the quality and granularity of weather models.

During the development of our model and training pipeline, we found that using small crops and temporally-correlated batching allowed us to avoid loading large tensors into memory from disk. Many competitors experienced challenges loading data fast enough to efficiently train deep learning models without precomputing all batches.

The most promising avenue for improving accuracy using existing modelling techniques is the inclusion of new weather data sources, such as aerosol forecasts. Aerosols such as smog and dust can block visible light from reaching the surface of the planet, reducing the potential for solar power generation. Furthermore, other aerosols can undergo chemical and physical processes in the atmosphere that can either hinder or facilitate the formation of clouds. Aerosol forecasts could be used in a similar manner to NWP data.

For grid scheduling purposes, point forecasts of solar output do not provide full context for energy systems operators. If current models could be adapted to produce probabilistic forecasts to predict a distribution of possible outcomes, then grid operators could better understand the likelihood of different solar generation outcomes. One possible avenue for exploration would be conformal prediction, a model-agnostic technique that can be applied in postprocessing to assign prediction intervals to point predictions.

7 Conclusion

Solar forecasting is a difficult but necessary task to ensure a smooth transition to renewable energy production. The physics of electricity grids dictate that supply and demand must be equal, so variability in solar energy production due to cloud cover and other facts must be accounted for in grid scheduling. Under the current paradigm, spinning reserve turbines are used in cases where solar power generation suddenly drops, producing carbon emissions that can be mitigated by improved solar energy forecasts.

Team Waterloo was ultimately awarded the first place prize in ClimateHack.AI 2023-2024 for their Perceiver-ensemble solution.

Acknowledgments

We would like to acknowledge the hard work of the ClimateHack.AI 2023-2024 organizers, including Jeremy and the DOXA AI competition platform team, Open Climate Fix for providing the dataset and challenge, and the competition sponsors.

References

[1] D. Yang, W. Wang, C. A. Gueymard, T. Hong, J. Kleissl, J. Huang, M. J. Perez, R. Perez, J. M. Bright, X. Xia, D. van der Meer, and I. M. Peters, "A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112348, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032122002593

[2] R. Tipton and D. Travers, "Nowcasting: How OCF will reduce carbon emissions with solar forecasts | Open Climate Fix Blog," Nov. 2022. [Online]. Available: https://www.openclimatefix.org/post/nowcasting-how-ocf-will-reduce-carbon-emissions-with-solar-forecasts

[3] J. Taylor, J. Leloux, L. M. H. Hall, A. M. Everard, J. Briggs, and A. Buckley, "Performance of Distributed PV in the UK: A Statistical Analysis of Over 7000 Systems," in *31st European Photovoltaic Solar Energy Conference and Exhibition*, Hamburg, Germany, Oct. 2015.

[4] "Solar PV Nowcasting Using Deep Learning," Open Climate Fix, Research Report NIA21-WP1-1, Dec. 2021. [Online]. Available: https://drive.google.com/file/d/1sDKZ8WEJlTNa5oyonbNl2xGyZ7GLXKtQ/view?usp=embed_facebook

[5] D. Reinert, F. Prill, H. Frank, M. Denhard, M. Baldauf, C. Schraff, C. Gebhardt, C. Marsigli, J. Förstner, G. Zängl, and L. Schlemmer, "DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System," Deutscher Wetterdienst, Tech. Rep. Version 2.3.1, Apr. 2024. [Online]. Available: https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/icon/icon_dbbeschr_aktuell.pdf;jsessionid=500E6FB85783C6DF6215D621FD1E05F9.live11042?view=nasPublication&nn=495490

[6] J. Bieker, S. Cotton, and O. C. Fix, "DWD ICON-EU Forecast." [Online]. Available: https://huggingface.co/datasets/openclimatefix/dwd-icon-eu

[7] E. O. for the Exploitation of Meteorological Satellites, "Rapid Scan High Rate SEVIRI Level 1.5 Image Data - MSG," Mar. 2009. [Online]. Available: https://navigator.eumetsat.int/product/EO:EUM:DAT:MSG:MSG15-RSS

[8] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Mar. 2015, arXiv:1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167

[9] B. Komer, T. C. Stewart, A. R. Voelker, and C. Eliasmith, "A neural representation of continuous space using fractional binding."

[10] P. M. Furlong and C. Eliasmith, "Fractional Binding in Vector Symbolic Architectures as Quasi-Probability Statements."

[11] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver: General Perception with Iterative Attention," Jun. 2021, arXiv:2103.03206 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2103.03206

[12] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," Mar. 2022, arXiv:2107.14795 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2107.14795

[13] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On Layer Normalization in the Transformer Architecture," Jun. 2020, arXiv:2002.04745 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2002.04745

[14] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," Oct. 2017, arXiv:1710.05941 [cs]. [Online]. Available: http://arxiv.org/abs/1710.05941

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[16] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 2019, arXiv:1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101

[17] L. N. Smith and N. Topin, "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates," May 2018, arXiv:1708.07120. [Online]. Available: http://arxiv.org/abs/1708.07120

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," Apr. 2017, arXiv:1611.05431. [Online]. Available: http://arxiv.org/abs/1611.05431