Presented at the WCPEC-8, 8th World Conference on Photovoltaic Energy Conversion, 26-30 September 2022, Milan, Italy

IMPROVED BLENDING OF PV POWER FORECASTS IN CASE OF MEASUREMENTS WITH LIMITED RELIABILITY

Wiebke Herzberg[1], Nicolas Holland[1], Tobias Zech[1], Jefferson Bor[1], and Elke Lorenz[1]
[1]Fraunhofer Institute for Solar Energy Systems ISE, Heidenhofstr. 2, 79110 Freiburg, Germany

ABSTRACT: For forecasting the power output of a photovoltaic (PV) power plant, solar irradiance forecasts are an essential input. Forecasts generated from different sources and models such as satellite data, numerical weather models, or irradiance measurements, perform differently depending on the forecast horizon. An optimized forecast for each horizon can be derived by combining several different source forecasts via a machine learning (ML) model to create a resulting 'blended' forecast. The training of an ML blending model requires power generation data of the considered PV power plant as a target variable. Typically, power measurements from the plant are used for this purpose. However, poor quality or limited availability of power measurements will lead to low quality blended forecasts as a result. In this work we consider intra-day forecasts, obtained from blending numerical weather predictions with satellite-derived forecasts, for a large PV plant with a capacity of 1 GW. The plant is frequently subject to curtailment by grid operators, which limits the reliability of its power measurements. This poses a significant challenge for the optimization of the blending model, resulting in a notable underestimation of the power forecasted by the model in high-power situations. We present an approach to improve forecast blending by using satellite derived power as a replacement for measurements in model training. We evaluate the effect of this replacement for two different ML blending models: a Huber linear regressor and a neural network. The performance of the different models is characterized by employing commonly used metrics such as RMSE and BIAS, as well as a distribution-oriented evaluation framework.

Keywords: PV Forecasting, Machine Learning, Satellite

1 INTRODUCTION

Reliable forecasting of PV power production is an important tool for a variety of applications: power grid operators require feed-in forecasts for grid regulation, power plants with integrated storage capabilities utilize production forecasts to optimize their battery usage and energy traders need forecasts when selling energy produced by PV plants. For forecasting the power output of any PV power plant, solar irradiance forecasts are an essential input. Irradiance forecasts can be obtained from different sources and models, such as satellite data, numerical weather models, or irradiance measurements. Each source, and the specific forecast type generated from it, will vary in performance depending on the considered forecast horizon. For example, satellite-based forecasts are viable for short-term horizons of up to several hours, while predictions from numerical weather models are viable for up to several days.

To obtain an optimized forecast for a given location and plant, different input forecasts can be combined through a weighting scheme, resulting in an optimal forecast for each forecast horizon (see, e.g., [1] and references therein). We refer to this process as forecast blending. The blending can be accomplished by a variety of different machine learning models, which may be trained using PV measurements from the power plant. We can express this problem as

$$f(x) = y \tag{1}$$

with a forecast blending model f, power measurements y, and x representing all input features of the model. Depending on the model, input features can consist of irradiance forecasts from different sources, PV power forecasts obtained by converting irradiance forecasts to power through application of a PV simulation, or other ancillary features such as e.g., solar position.

Training of the model is performed on a dataset containing features and corresponding measurements $\{(x_t, y_t)\}_{t\in T}$ spanning a certain time range T. The quality of the forecasts produced by the forecast model very much depends on the quality of the available power measurements used in training. If the quality of the measurements is poor, or their availability is limited, the quality of the resulting blended forecast will be poor as a consequence.

One important issue causing challenging quality and availability of measurements are curtailment events. Curtailment occurs when e.g. grid operators send a signal to artificially reduce the production of a PV plant in order to ensure grid stability. Curtailment will result in PV power measurements which do not reflect the power that could potentially be generated by the plant under the given irradiance conditions. We denote with $\hat{y}$ the actual production of the power plant, which can be different from the potential production y of the plant i.e., for some instances it holds true that $\hat{y} < y$ due to curtailment. While for the applications we mentioned it is usually necessary to forecast the potential output y of the plant, often only the actual production including curtailment $\hat{y}$ is measured and therefore available as a reference for model training. If a power plant is frequently subject to curtailment, the resulting effect on the measurements can cause significant issues in the optimization process of the blending models. In previous works [2] we investigated how different forecast sources can be best combined without the presence of curtailment and in following works we explored ways to handle curtailment by detecting it using a curtailment detection model [3] and adapting model training to the presence of curtailment by tweaking loss functions used in model training [4].

In this work we present an approach to improve forecast blending by introducing satellite derived power as the target variable in model training, thereby replacing the PV power measurements. We investigate this approach for intra-day forecasts up to horizons of 4 hours that are generated for a power plant with 1 GW installed capacity in China. The forecasts are obtained by blending a satellite-based forecast with numerical weather predictions. For the blending model we explore two options, a linear Huber model [5], which possesses a built-in resilience against outliers and a neural network. We evaluate the different models following verification techniques recently illustrated by Yang et. al. [6].

The remainder of the paper is structured as following: Section 2 discusses the measurement data, curtailment and the challenges it introduces to the measurements, as well as the dataset of input forecasts used in this work. Section 3 describes our approach to improve the generated blended forecasts, including the satellite data used as replacement in model training and the blending models that were employed. Section 4 compares our results obtained from different models and training data. Section 5 summarizes and concludes.

2 DATASETS

2.1 PV power measurements

The power plant for which we evaluate our approach is a PV plant of 1 GW installed capacity in the Qinghai province of China. Measurements of produced power are available, but the plant is subject to frequent curtailment by grid operators which limits the reliability of the measurements. We work with measurements of the total power plant output resampled to 15-minute mean values. Our dataset includes data from March 2020 to March 2021.

2.2 Curtailment

Curtailment represents the artificial reduction of the power production of a PV plant at times where less power feed-in to the grid is required. The actual information of when curtailment is being applied at the plant and to what extent the power production is curtailed is often not available. In our previous work [3] we developed a curtailment detection method d which aims to predict whether a datapoint $(x_t, \hat{y}_t)$ is subject to curtailment, i.e.,

$$d(\hat{y}) = \begin{cases} 1, & if\ \hat{y} = y \\ 0, & else \end{cases} \qquad (2)$$

The curtailment detection method is based on a comparison of measurements and Active Generation Control (AGC). An example of a day with detected curtailment is shown in Figure 1. The forecast model training is enhanced by replacing the training dataset with a curtailment filtered version $\{(x_t, \hat{y}_t) | d(\hat{y}) = 1\}_{t \epsilon T}$. While we expect some improvement from the application of a curtailment filter to training data of a blending model before the model optimization process, since the number of uncorrelated data points $(x_t, \hat{y}_t)$ will be reduced, it will also introduce new challenges. Applying a curtailment filter will result in many missing data points, especially for periods with potentially large power generation (see Figure 1). Also, due to imperfections in filter performance, the filtered data can still contain curtailed data points as outliers, reducing the quality of the remaining data.

2.3 Model input

To create a blended forecast which is optimized for each horizon, we combine input forecasts from different sources and additional solar position features via a machine learning (ML) model. In this work, we use numerical weather predictions (NWP) from the European Centre for Medium-Range Weather Forecasts (ECMWF), and cloud motion vector forecasts from images of the geostationary Himawari satellite, which is operated by the Japan Meteorological Agency.

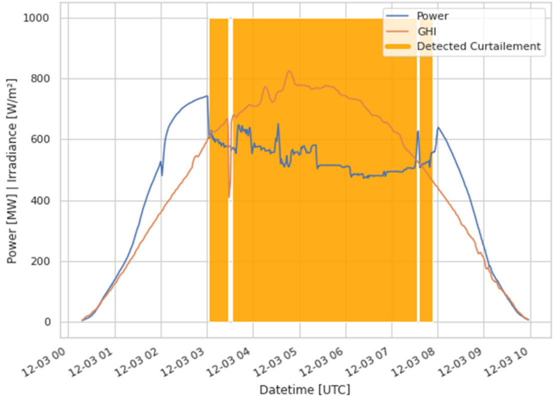

Figure 1: PV power and global horizontal irradiance (GHI) measurements for a day with mostly clear-sky conditions. The highlighted area is labeled as curtailed by the curtailment detection model.

The NWP data with an original resolution of 1 hour is upsampled to time steps of 15 minutes by interpolation of the clear-sky index. The satellite-based forecasts are directly created in 15 minute steps by applying the heliosat method [7] and cloud motion vectors [8] on satellite images. We make use of the deepflow algorithm [9] to compute cloud motion vector fields, using the implementation from the OpenCV library [10]. Both the resampling of NWP irradiance as well as the heliosat method require modelled clear-sky irradiance, for which we use the clear-sky model by Dumortier [11].

To convert these irradiance forecasts to forecasts of PV power, we apply a PV simulation adapted for the specific plant. The simulation consists of several steps: First, global horizontal irradiance (GHI) values are converted into plane of array irradiance using the DIRINT separation model [12] and the transposition model by Perez et. al. [13]. For both models we rely on the implementations in PVLIB [14]. The subsequent steps to derive PV power from plane of array irradiance are implemented inhouse by a parametric model-chain [15]. The simulation also requires an ambient temperature input for which we employ the ECMWF NWP temperature forecasts. A more detailed description of the source forecasts and the PV simulation is available in [4].

2.4 Training and testing set

For the evaluation of the performance of the different blending models, we divide the entire available dataset into a training and a testing set. The models are optimized on the training data, and evaluation is carried out on the testing data. We divided the data along the base times (i.e. forecast creation times), and randomly sampled 67% of forecasts into the training set. The remaining 33% of the forecasts are used for testing.

Training and evaluation for all models is performed on data where curtailment has been removed by applying the filtering method described in Section 2.2.

3 APPROACH AND MODELS

3.1 Blending models

For the blending models we explore two different options, a linear Huber regressor and a Neural Network (NN). The models generate intra-day forecasts with a time resolution of 15 minutes. New forecasts are also generated every 15 minutes. Here, we evaluate the forecasts for horizons up to 4 hours.

3.1.1 Huber regressor

The Huber regressor [5] is a linear regression model with a built-in resilience against outliers, which can be tuned via a hyperparameter. For samples where the deviation between model and measurement falls below a certain threshold, the squared loss is optimized, whereas for samples with deviations above the threshold, the absolute loss is optimized. Thereby, less weight is attributed to samples constituting large deviations i.e., outliers. In this work, the scikit-learn [16] implementation of the Huber model was used and the corresponding hyperparameter ε was kept fix at a value of 1.05.

Three input features were used for this model: the PV simulated values of both, the satellite-derived irradiance forecast and NWP irradiance, and the cosine of the solar zenith angle. A separate model was trained for each forecast horizon. In the training for each horizon, data from a window of a total of 5 horizons around the target horizon were included, to avoid overfitting and to generate a smooth output forecast.

3.1.2 Neural network

In order to compare the linear Huber model with a non-linear model we chose to train a neural network with 2 hidden layers and a 100 neurons per layer. For the implementation we rely on the multi-layer perceptron [17] implemented in scikit-learn [16]. The network was trained using stochastic gradient decent [18] to minimize MSE (see Section 3.3) and the input features for the model were PV simulation values of NWP and satellite-based predictions and the solar zenith angle. In addition to those features we added the forecast horizon as another input feature, allowing the model to make predictions for several horizons in one forward pass of the model.

3.2 Replacement of training data

To avoid model training being affected by curtailment, or potentially other circumstances causing challenging quality and availability of measurements, we introduce a satellite-derived target variable as a replacement for the measurements in model training. The new target variable is obtained from converting satellite-derived irradiance values of our real-time processing chain (forecast horizon of 0 minutes) to PV power by means of a PV simulation. We refer to this as 'nowcast' values. This change in the target variable can be expressed as replacing $\hat{y}$ with simulated nowcast values $s(x_{sat})$. One advantage is that these simulated values are not subject to curtailment, meaning they do not artificially underestimate the true output potential of the power plant. They could also be used to mitigate data gaps which arise from removing curtailed datapoints, this is however not investigated in this work. Possible disadvantages include s being merely an approximation of the potential power plant output $s(x_{sat}) \approx y$.

Figure 2 shows power values from the satellite-derived nowcast against power measurements where curtailed data were removed by applying the curtailment detection method outlined in Section 2.2. The measurements show a reasonable agreement with the nowcast, suggesting that its use as a replacement in model training is worth to investigate. For our evaluation, the blending models were trained once on the actual measurements and once on the satellite-derived nowcast.

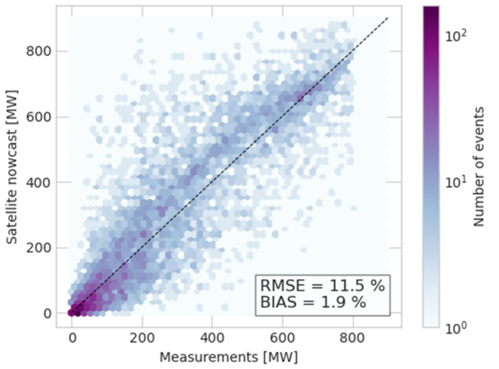

Figure 2: PV power derived from a satellite-based irradiance nowcast against measured power. Only data points for which no curtailment was detected are displayed. Relative RMSE and BIAS are calculated with respect to installed capacity of 1 GW using daylight values only.

3.3 Evaluation metrics

A common metric used in the evaluation of PV power forecasts is the root-mean-square error (RMSE), which is the root of the mean-square error,

$$MSE = \frac{1}{N}\left(\sum_{t}(f(x_t) - \hat{y}_t)^2\right). \tag{3}$$

The RMSE can be complemented by considering the BIAS

$$BIAS = \frac{1}{N}\left(\sum_{t} f(x_t) - \hat{y}_t\right). \tag{4}$$

In addition, to obtain a more in-depth understanding of forecast model performance, a distribution-oriented approach to evaluation is of great value. In this framework, data comprising forecasts generated by the blending models and corresponding measured power can be expressed as a 2-dimensional distribution $p(f,\hat{y})$. This joint distribution $p(f,\hat{y})$ can be related to marginal and conditional distributions via two Murphy-Winkler factorizations [19]

$$p(f,\hat{y}) = p(\hat{y}|f)\,p(f) \tag{5}$$

$$p(f,\hat{y}) = p(f|\hat{y})\,p(\hat{y}) \tag{6}$$

called *calibration-refinement* and *likelihood-base rate* factorization respectively. Following these factorizations, the MSE can also be decomposed in two distinct ways (as presented e.g., in [20], [6])

$$MSE = V(\hat{y}) + E_f[f - E(\hat{y}|f)]^2 - E_f[E(\hat{y}|f) - E(\hat{y})]^2 \tag{7}$$

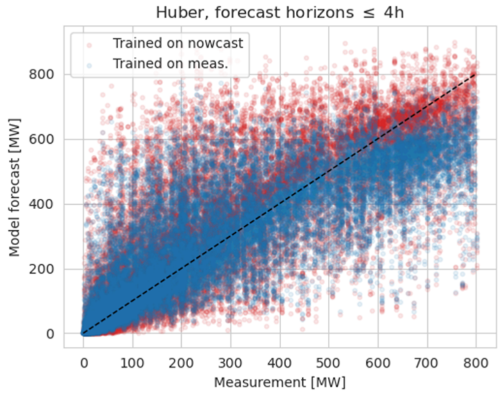

Figure 3: Scatter plot of forecasts from the Huber model for horizons up to 4 hours, trained on measurements (blue) or satellite-derived nowcast (red).

$$MSE = V(f) + E_{\hat{y}}[\hat{y} - E(f|\hat{y})]^2 - E_{\hat{y}}[E(f|\hat{y}) - E(f)]^2 \quad (8)$$

with E indicating the expectation value. In the first decomposition (Eq. (7)), $V(\hat{y})$ denotes the variance of the measurements, while the second and third term can be referred to as *type-1 conditional bias* and *resolution* respectively. In the second decomposition (Eq. (8)), $V(f)$ denotes the variance of the forecast values, while the second and third term can be referred to as *type-2 conditional bias* and *discrimination,* respectively. Note that $V(\hat{y})$ is a fixed property of the dataset, while $V(f)$ is subject to the optimization. When evaluating forecast models, the conditional biases are to be minimized while the resolution and discrimination components are to be maximized.

4 RESULTS

Evaluations shown in this section were all carried out on the testing set.

4.1 Model comparison using RMSE and BIAS

We begin the comparisons by looking into the difference between models trained on power measurements $\hat{y}$ and satellite derived power nowcast values $s(x_{sat})$. Figure 3 and 4 show scatter plots of forecasted vs. measured PV power for the Huber model and NN model respectively. For both ML models, we find that the model trained on measurements has a strong tendency to underestimate large measurements between 600 and 800 MW, which can be explained by those situations being underrepresented in the dataset as well as undetected curtailed values still being present during training. This lowers the correlation between input forecasts and the target variable (the measurements) which in turn tends to pull down the forecasted values towards the mean. Training on the nowcast visibly alleviates this underestimation, the improvement being slightly more noticeable for the Huber model.

However, forecasts from the nowcast-trained models exhibit a higher RMSE and BIAS when evaluated against the measurements, as presented in Figure 5. In part this is to be expected given the relation between measurements and satellite nowcast shown in Figure 2. We note though, that the increase in both metrics is much stronger for the Huber model than for the NN.

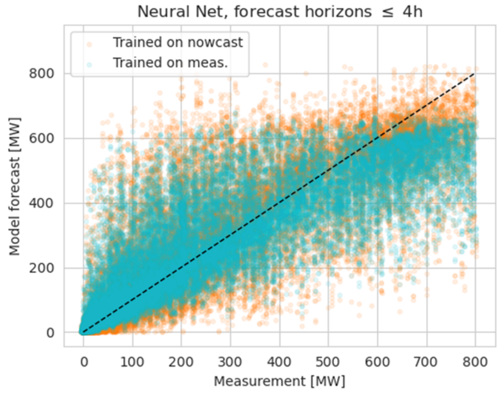

Figure 4: Scatter plot of forecasts from the NN model for horizons up to 4 hours, trained on measurements (cyan) or satellite-derived nowcast (orange).

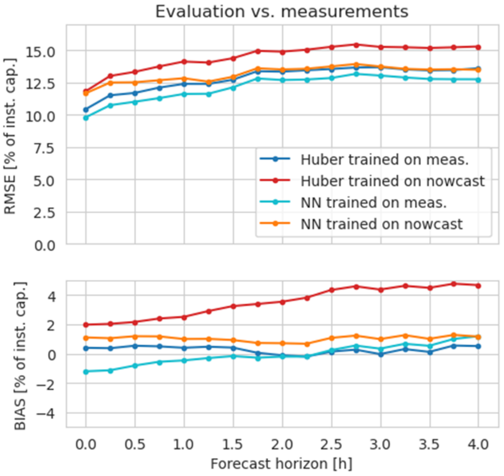

Figure 5: RMSE and BIAS of forecasts generated by different models in dependence of forecast horizon. Values are normalized with respect to installed capacity of 1 GW.

4.2 Model comparison using conditional distributions

In order to further investigate the difference of the measurement-trained and nowcast-trained models we divide the forecasted values into bins and compute conditional distributions following the Murphy Winkler factorizations (Eq. (5) and (6)). In Figure 6, the panels on the left side show conditional distributions binned along the measured and forecasted values for an example horizon of 2 hours for the Huber model. The panels on the right side of Figure 6 show the same for the neural network. We find that the underestimation in the high-power range of 600 to 800 MW of the measurement-trained models (blue and cyan distributions) is very prominent: The distributions in this range are barely present for the forecast binning (Figure 6, upper panels) and the distributions clearly underestimate the measurements for the measurement binning (Figure 6, lower panels). The nowcast-trained models (red and orange distributions) alleviate this underestimation but

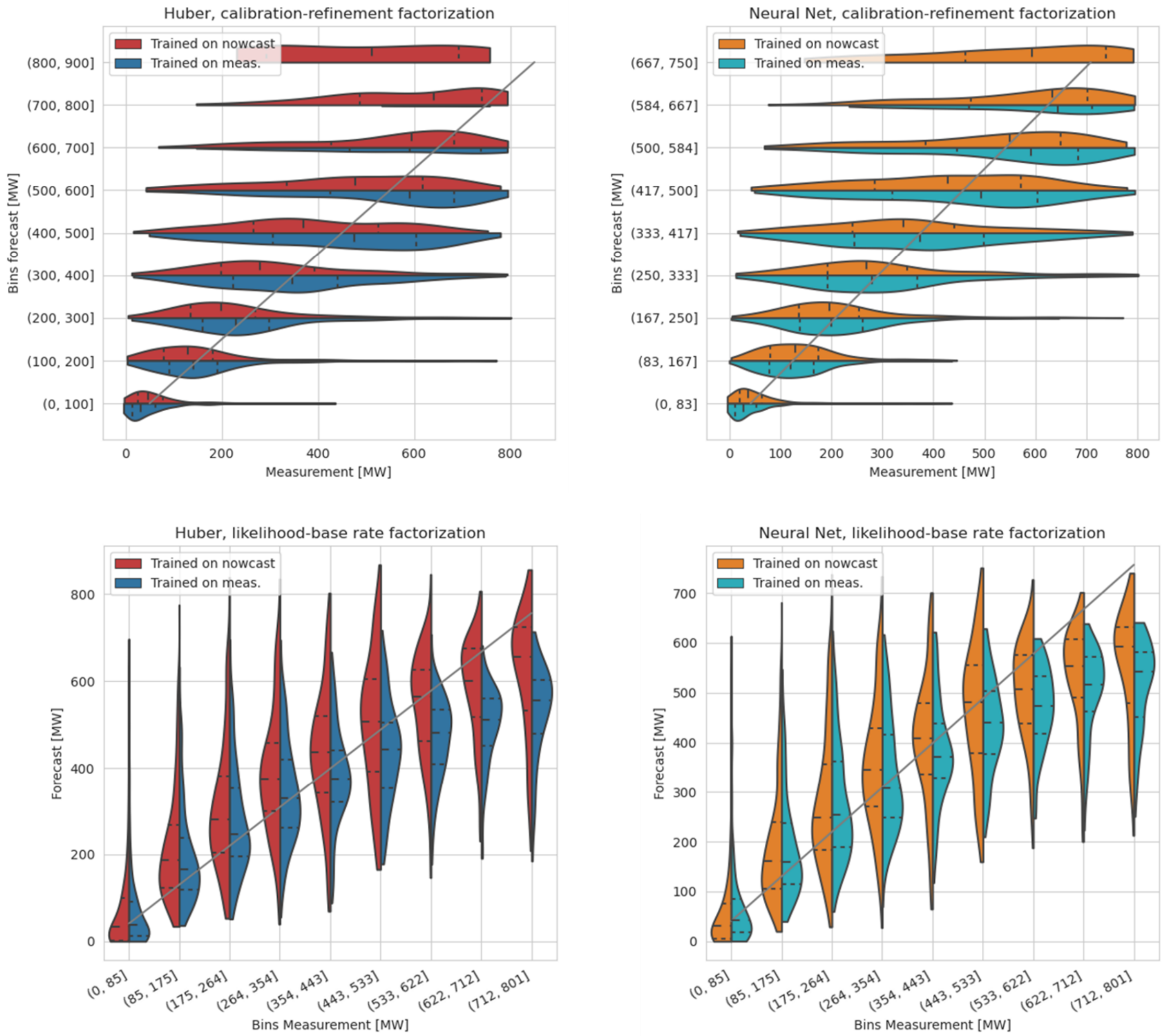

Figure 6: Comparison of conditional distributions for Huber models (left panels) and NN models (right panels) trained on either measurements or nowcast, for an example forecast horizon of 2h. Upper panels: distributions conditioned on forecasts $p(\hat{y}|f)$. Lower panels: distributions conditioned on measurements $p(f|\hat{y})$.

also introduce an overestimation in the low to intermediate power region of 100 to 300 MW, which is particularly strong for the Huber model. Apart from that, it is interesting to note that the measurement distributions for the upper most forecast bins, which are only visible for the nowcast-trained models (see Figure 6, upper panels) tend towards a uniform distribution. This highlights again that the measurements have almost no correlation to forecast values in that region.

4.3 Comparison using MSE factorization

Lastly, we compare the different models with respect to the components of the MSE factorizations given in Eq. (7) and Eq. (8). The upper left panel of Figure 7 shows the three components of Eq. (7), the variance of measurements, the type-1 conditional bias, and the resolution, for all four investigated models and with increasing forecast horizon. The respective MSE components are of similar magnitude for all four models, with the contributions from variance and resolution being much larger than the type-1 conditional bias.

We observe that training on the nowcast (red and orange lines) tends to decrease the resolution and increase the type-1 bias for both the Huber and NN model, meaning both MSE components become slightly worse when evaluated against measurements. This results in the overall increased RMSE as seen in Figure 5.

The lower left panel of Figure 7 shows the three components of the MSE decomposition given in Eq. (8), the variance of the forecasts, type-2 conditional bias, and discrimination, for the investigated models and with increasing forecast horizon. We can see more prominent differences between the models, particularly in the variance of forecast values and the discrimination. It is evident that training on the nowcast lowers the type-2 conditional bias for both the Huber and NN model and additionally increases the discrimination, which are in general favorable qualities. The lower type-2 conditional bias reflects forecast distributions which are better centered on the diagonal (see Figure 6, lower panels) and a higher discrimination means forecasts will differ more strongly for different measurement situations. Taken together, this means we obtain a better mapping of the forecasts to the potential range of power measurements. However, the significantly larger variance of the forecasts outweighs the favorable trends in type-2 conditional bias

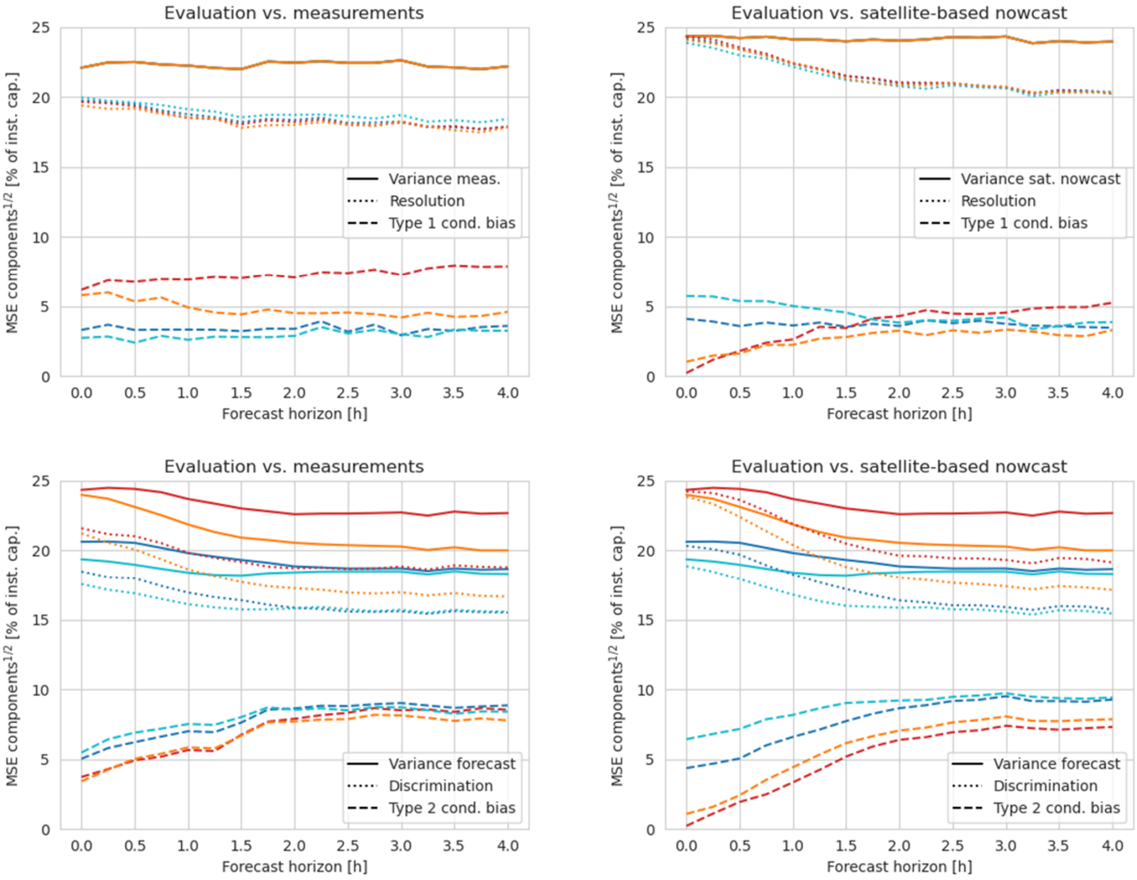

Figure 7: Square root of the components of the MSE factorization given in Eq. (7) (upper panels) and Eq. (8) (lower panels) normalized with respect to installed capacity for measurement-trained Huber (blue) and NN (cyan) models, as well as nowcast-trained Huber (red) and NN (orange) models in dependence of forecast horizon. Left panels: Calculation with respect to measurements. Right panels: Calculation with respect to satellite-derived nowcast.

and discrimination and leads to the overall increased RMSE for both nowcast-trained models, when evaluated against measurements, as displayed in Figure 5.

It should be noted here that J. Moskaitis points out in [20] that a comparison of discrimination between two models is only fair if their variances are similar, because forecast variance and discrimination are related. In their work they therefore split the MSE only into two terms, the conditional bias and a so-called shape term, which comprises both variance and discrimination. Here, following Yang et al [6] we chose to present the full decomposition into three terms to allow a more detailed discussion.

4.4 Complementary evaluation with respect to the satellite-derived nowcast

Due to the limited reliability of the power measurements, an evaluation of the forecasts based solely on these measurements is of course inherently problematic. Therefore, we complement the measurement-based evaluation with an evaluation against the satellite-based nowcast which is free of curtailment and was used to replace the measurements in model training. In this evaluation a lower RMSE of the nowcast-trained models compared to the measurement-trained models can be expected, since the evaluation is carried out with respect to the target variable used in training. Figure 8 shows the RMSE and BIAS for all models with respect to the satellite-derived nowcast. Indeed, we see an improved RMSE and BIAS of the nowcast-trained models, except for horizons >1.5 hours for the Huber model, where an increased BIAS manifests, leading also to an increased RMSE (this can occur since the Huber model does not optimize MSE in training). For a horizon of 0 minutes, we can observe the Huber model having an RMSE of zero, which is due to the fact that the nowcast values are an input feature for the models and that the Huber model was trained separately for each horizon. The NN does not drop to an RMSE of zero since several horizons of the model are trained together in one pass.

The upper right panel of Figure 7 shows the MSE components of Eq. (7) calculated with respect to the satellite-derived nowcast. There is an overall higher variance of the nowcast values compared to the variance of measurements, which is accompanied by a higher resolution for all four models. As opposed to the evaluation against measurements in the upper left panel, we see here that the type-1 conditional bias is improved for the nowcast-trained models, apart from the Huber model for horizons larger than 1.5 hours, exhibiting an increased type-1 bias. Since the difference between the variance and resolution is smaller or similar here than in the evaluation against measurements this gives together with the improved type-1 bias the overall smaller RMSE for the nowcast-trained models compared to their measurement-trained counterparts (see Figure 8). The

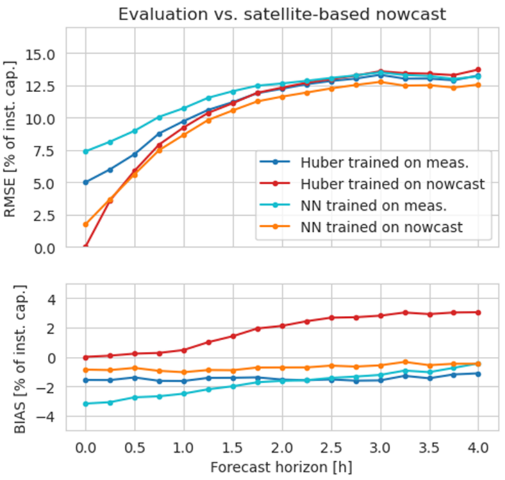

Figure 8: RMSE and BIAS of forecasts generated by different models in dependence of forecast horizon, calculated with respect to the satellite-derived nowcast. Values are normalized with respect to installed capacity of 1 GW.

exception being again the later horizons of the Huber model, where the type-1 bias was not improved.

In the lower right panel of Figure 7 the components of the MSE decomposition given in Eq. (8) are shown. The improvement of the discrimination and conditional type-2 bias components of the nowcast-trained models compared to the measurement-trained models, which were already discernable in the measurement-based evaluation (lower left panel), are again present, and here outweigh the increased variance of the nowcast-trained models, leading to the overall smaller RMSE (see Figure 8). These results emphasize, although to some degree expected, the improvements obtained from the nowcast-trained models.

5 SUMMARY AND CONCLUSION

For a dataset with measurements of limited reliability, we explored the replacement of measurements by a satellite derived PV power nowcast in model training to improve the results of forecast blending. Two types of blending models were investigated, a linear Huber model and a neural network. For both models we compared results after training the model on measurements and on a satellite-derived nowcast by considering standard metrics such as RMSE and BIAS and a distribution-oriented evaluation framework.

Due to the lack of reliable ground truth data, it is difficult to assess which model has the best performance. Evaluating a model against the variable it was trained on is generally expected to give a better RMSE (assuming training is based on MSE optimization). Nowcast-trained models for short horizons have the additional advantage of having an input (the satellite-derived forecast) that is highly correlated to their target variable, allowing their RMSE to reach even lower values when calculated with respect to the nowcast. When evaluating against measurements, the observed higher RMSE of the nowcast-trained models compared to the measurement-trained models is therefore expected and based on the differences between measurements and nowcast. First there is the inherent difference between satellite data and ground-based measurements, next there is the difference between a PV power simulation and actual power measurements. These two sources of error would be present in any model trained on satellite and evaluated against measurement data. The third component in our case are the effects of curtailment, which further lower the correlation between measurements and satellite-derived nowcast. Because, even though a model for curtailment detection was applied, some undetected curtailed values remain in the data set. In turn, when evaluating against the nowcast, as expected the nowcast-trained models exhibit a lower RMSE than their measurement-trained counterparts (an exception being larger horizons of the Huber model, which can be explained by Huber models not optimizing the MSE). Therefore, from these results alone we cannot find a definitive indicator of forecast quality.

In general, conventional comparisons of summary error metrics like RMSE or BIAS are not suitable to highlight different aspects of forecast quality. In our case they do not convey the fact that large power values are underrepresented in the forecasts. Therefore, we investigated the forecasts with a distribution-oriented approach, where different attributes of forecasts can be distinguished. We observe that the nowcast-trained models have lower type-2 conditional biases and larger discriminations, in both the evaluation against measurements and against the nowcast. This corresponds to a better mapping of forecasts to the range of power values and reflects the visible alleviation of the underestimation in the high-power region. When evaluated against measurements, this partial improvement in conditional type-2 bias and discrimination is mitigated by a strongly increased variance, leading to an overall higher RMSE of the nowcast-trained models. When evaluated against the satellite-derived nowcast, the drawbacks of disproportionally large forecast variances disappear (except for horizons >1.5 hours of the Huber model), while the improvements in discrimination and type-2 conditional bias become more pronounced, thus leading to an overall lower RMSE of the nowcast-trained models compared to the measurement-trained ones.

Though we cannot claim that evaluation with MSE decompositions do directly lead to “best-performing” model recommendations, the detailed evaluations give additional insight into model performance. The results presented in this work suggest it worthwhile to explore the use of satellite based nowcast-trained models. These models have shown to overcome the underestimation for large PV power values that occur when the measurement data used in training is of limited reliability. Which kind of model (e.g., NN, Huber or other options) is most suitable may depend on the individual use-case and input data. In the work presented here, the Huber model alleviates the underestimation in the high-power range of 600 to 800 MW more strongly than the NN, but this comes at the cost of also introducing a stronger overestimation in low to intermediate power values around 100 to 300 MW. The neural network (which is MSE optimized) produces more balanced results and with its nonlinear nature additionally might have more flexibility in adaption to the data. It exhibits the lowest RMSE in the evaluation based on the nowcast for almost all horizons.

Further enhancements in the generated forecasts can be expected from a fine tuning of the models or improving the satellite-derived PV power data. The latter could be achieved for example by improving the underlying satellite method and PV simulation, or by adaption of the satellite-derived data to measurements through postprocessing (e.g., bias or trend removal). The more accurate we can infer PV power from satellite data, the more accurate will the models be when compared to actual measurements. Furthermore, using satellite based nowcasts for model training holds the potential to mitigate data gaps that arise from removing curtailed datapoints, which is expected to be an additional advantage.

6 REFERENCES

[1] Sengupta, M., Habte, A., Wilbert, S., Gueymard, C., Remund, J. (2021). Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition.

[2] Holland, N., et al. (2019). Solar and PV forecasting for large PV power plants using numerical weather models, satellite data and ground measurements, 2019 IEEE 46th Photovoltaic Specialists Conference.

[3] Holland, N., et al. (2020). Trainable curtailment detection for the enhancement of PV power forecasting, 2020 PVSEC-30 & GPVC.

[4] Holland, N., et al. (2021). Combination of physics based simulation and Machine Learning for PV power forecasting of large power plants, Proceedings of the 38th European Photovoltaic Solar Energy Conference and Exhibition.

[5] Huber, P. J. (1981). Robust Statistics, John Wiley and Sons, New York.

[6] Yang, D., et al. (2020). Verification of deterministic solar forecasts, Solar Energy, Vol. 210, Pages 20-37.

[7] Hammer, A., Heinemann, D., Hoyer, C., Kuhlemann, R., Lorenz, E., Müller, R., Beyer, H.G. (2003). Solar energy assessment using remote sensing technologies, Remote Sensing of Environment, 86(3): 423–32.

[8] Kühnert, J., Lorenz, E., Heinemann, D. (2013). Satellite-Based Irradiance and Power Forecasting for the German Energy Market, Solar Energy Forecasting and Resource Assessment, Elsevier.

[9] Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C. (2013). Deepflow: Large displacement optical flow with deep matching, ICCV – IEEE International Conference on Computer Vision, pages 1385–1392.

[10] Bradski, G. (2000). The opencv library, Dr. Dobb's Journal of Software Tools.

[11] Dumortier, D. (1995). Modelling global and diffuse horizontal irradiances under cloudless skies with different turbidities, Daylight II, jou2-ct92- 0144, final report vol. 2. Tech. rep., CNRSENTPE.

[12] Perez, R., Ineichen, P., Maxwell, E., Seals, R., Zelenka, A. (1992). Dynamic Global-to-Direct Irradiance Conversion Models. ASHRAE Transactions-Research Series, pp. 354-369.

[13] Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R. (1990). Modeling daylight availability and irradiance components from direct and global irradiance. Solar Energy 44 (5), 271-289.

[14] Holmgren, W.F., Hansen, C.W., Mikofski, M.A. (2018). pvlib python: a python package for modeling solar energy systems, Journal of Open Source Software, 3(29), 884.

[15] Müller, B., et al. (2015). Yield predictions for photovoltaic power plants: empirical validation, recent advances and remaining uncertainties, Progress in Photovoltaics: Research and Applications: published online

[16] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

[17] Hinton, G.E. (1989). Connectionist learning procedures, Artificial intelligence 40.1, 185-234.

[18] Kingma, D., and Ba, J. (2014). Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[19] Murphy, A. H., & Winkler, R. L. (1987). A General Framework for Forecast Verification, Monthly Weather Review, 115(7), 1330-1338.

[20] Moskaitis, J.R. (2008). A Case Study of Deterministic Forecast Verification: Tropical Cyclone Intensity, Weather and Forecasting, 23(6), 1195-1220.