<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Solar Energy*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

van der Meer, D., Munkhammar, J., Widén, J. (2018)

Probabilistic forecasting of solar power, electricity consumption and net load:

Investigating the effect of seasons, aggregation and penetration on prediction intervals

Solar Energy, 171: 397-413

<https://doi.org/10.1016/j.solener.2018.06.103>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-362870>

Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals

D.W. van der Meer^{a,*}, J. Munkhammar^a, J. Widén^a

^a*Built Environment Energy Systems Group (BEESG), Division of Solid State Physics, Department of Engineering Sciences, Uppsala University, P.O. Box 534, SE-751 21 Uppsala, Sweden*

Abstract

This paper presents a study into the effect of aggregation of customers and an increasing share of photovoltaic (PV) power in the net load on prediction intervals (PIs) of probabilistic forecasting methods applied to distribution grid customers during winter and spring. These seasons are shown to represent challenging cases due to the increased variability of electricity consumption during winter and the increased variability in PV power production during spring. We employ a dynamic Gaussian process (GP) and quantile regression (QR) to produce probabilistic forecasts on data from 300 de-identified customers in the metropolitan area of Sydney, Australia. In case of the dynamic GP, we also optimize the training window width and show that it produces sharp and reliable PIs with a training set of up to 3 weeks. In case of aggregation, the results indicate that the aggregation of a modest number of PV systems improves both the sharpness and the reliability of PIs due to the smoothing effect, and that this positive effect propagates into the net load forecasts, especially for low levels of aggregation. Finally, we show that increasing the share of PV power in the net load actually increases the sharpness and reliability of PIs for aggregations of 30 and 210 customers, most likely due to the added benefit of the smoothing effect.

Keywords: Probabilistic forecasting, quantile regression, Gaussian process, solar power, electric load, net load

---

1. Introduction

The increasing penetration of photovoltaic (PV) power into the electricity generation mix has at least two consequences that affect one another in an interesting way. Firstly, an increase in the number of installed PV systems in a wider area is accompanied by the smoothing effect, as described by Perez and Hoff (2013). If the data are available, this could allow for more accurate city scale or regional PV power forecasts. In case of insufficient data, one could estimate the output of these "invisible" sites using a representative set of sites, see, e.g., Shaker et al. (2015) for a thorough study. Secondly, however, the increase in the number of installed PV systems or their sizes simultaneously reduces the accuracy of net load forecasts at the distribution feeder or even finer spatial resolution because of the increased variability and ramps, see e.g., Denholm and Margolis (2007); Nguyen et al. (2016). It should be noted that Kaur et al. (2013) showed that penetration did not affect the net load forecast error to the same extent as variability did. However, the reason for their conclusion can mainly be ascribed to the fact that they considered a single PV power farm and looked at the influence of the cloud conditions and the subsequent penetration levels due to these conditions. A logical conclusion

would therefore be to ascribe the forecast error to variability, since very little smoothing occurs and since in this case, the penetration is a direct function of variability.

The installed capacity of PV power is still increasing, around 75 GW of capacity was installed during 2016 (IEA, 2017). In Germany, PV power covered 7.4% of electricity consumption during 2016, which increased to 35% on sunny weekdays (Wirth, 2017). At such levels, challenges such as feeder loading, grid losses and voltage fluctuations occur at in the distribution grid (Aguero and Steffel, 2011; Mohammadi and Mehraeen, 2017; Walling et al., 2008). Accurate forecasts are generally seen as a cost-efficient solution because they allow for e.g., unit commitment, curtailment and demand response, which in turn can be used to reduce operational risks (Kaur et al., 2016b). However, deterministic, or point forecasts, do not express uncertainty, which is important for stakeholders to make informed decisions. Probabilistic forecasting has gained more attention recently, as can be inferred from recent review studies (Hong and Fan, 2016; van der Meer et al., 2018b). The advantage of probabilistic forecasting is that the level of uncertainty can be expressed that accompanies the prediction by means of a probability density function (PDF) or prediction interval (PI), which enables stakeholders to make better informed decisions. The smoothing effect has been shown to improve the accuracy of deterministic forecasts, see e.g., Perez and Hoff (2013). In this paper, we investigate whether this holds in case of

---

*Corresponding author. Tel.: +46 18 471 31 33
Email address: dennis.vandermeer@angstrom.uu.se
(D.W. van der Meer)

probabilistic forecasting.

1.1. Previous work

1.1.1. The effect of aggregation

Variability in solar irradiance, and PV power production by extension, and electricity consumption can be reduced by the well-known smoothing effect. An extensive review by Widén et al. (2015) revealed a series of notable studies into the variability of irradiance and the smoothing effect. For instance, Lave et al. (2012) studied aggregated irradiance measurements of six sites within 3 km of each other on timescales varying from 1s to 1h. They found that on that particular geographical scale and at a time span below 5 min, the ramp rates became uncorrelated and the smoothing effect was substantial. Furthermore, the authors noted that for timescales shorter than 256s, the variance was reduced by a factor of 6 when compared to a single system. On a larger geographical scale where 7 PV plants were separated between 6 and 360 km, Marcos et al. (2012) found that, at a temporal resolution of 1s and distance of 6 km, the power output was already uncorrelated, thus substantially increasing the smoothing effect. However, at lower temporal resolutions, e.g., 30 min as used in this work, the smoothing effect is less pronounced when the same plants are considered, because such a resolution allows similar weather patterns to occur over a wider area (Widén et al., 2015).

In case of electricity consumption forecasting, Hayes et al. (2015) studied what effect aggregation has on the mean absolute percentage error (MAPE) and examined the correlations between the dependent and independent variables. The authors noted that at the low voltage (LV) feeder, i.e., a low level of aggregation, these correlations were significantly lower than at the primary or secondary substations. Consequently, MAPE decreased from 20%-30% at the end user level to 5% at the primary substation. Hayes et al. did not continue to investigate the probabilistic aspect of aggregation. Taieb et al. (2016) used quantile regression (QR) in combination with gradient boosting (GB) to forecast individual and aggregated electricity consumption, although forecasts at intermediate aggregation levels were not performed. The results showed that the prediction intervals (PIs) were substantially wide in case of end user electricity consumption forecasting, due to the high variability of this time series. In contrast, PIs for the aggregated demand of 3639 customers were highly informative and, moreover, accurate. The authors of both aforementioned studies noted that theirs were the first research papers that attempted to quantify the effect of aggregation on forecast accuracy (Hayes et al., 2015) and the second paper to utilize data from individual smart meters to produce probabilistic forecasts (Taieb et al., 2016).

In van der Meer et al. (2017) we showed that aggregation can drastically improve the quality of PIs with aggregations of modest numbers of customers. In this study, we continue that research by, among other things, using

more subsets of the data in order to filter away noise and improve accuracy.

1.1.2. The seasonal effect

Few studies have investigated the effect of seasons on forecast accuracy, and fewer on probabilistic forecast accuracy. In general, such studies have been performed as a side step when, e.g., proposing a new model. For example, Vaz et al. (2016) proposed an artificial neural network (ANN) that utilized spatial and temporal information to produce deterministic forecasts of the PV power output in Utrecht, the Netherlands. The results showed that the forecasts during winter were more accurate in terms of the normalized root mean squared error (nRMSE) than those in summer, and the authors ascribed this to the fact that overcast days are more straightforward to predict. An example of a paper dedicated to study the impact of seasonal variations on deterministic load forecasting was performed by Lusi et al. (2017) in New South Wales, Australia. Here, the authors looked at a wide variety of scenarios where some included dummy variables to indicate the season, and other scenarios included subsets of the data that included only the season of interest. The results showed that, for a single residential load pattern, the scenarios including the dummy variables yielded marginally improved results, whereas the scenarios that included dedicated subsets of the seasons yielded worse forecast accuracy. The authors ascribed the latter result to a lack of robustness to outliers because of a reduced amount of subsets. However, in case of aggregated residential load patterns these negative effects were less pronounced.

As for probabilistic forecasting, there are only a few publications that go into detail as regards the effect of seasons on PIs. Firstly, Scolari et al. (2016) proposed a novel method to predict solar irradiance for very short time horizons, in which the PIs were produced based on quantiles obtained from k -means clusters. The location for their study was Lausanne, Switzerland. They assessed the performance for three seasons, i.e., summer, fall and winter, and noted that the season with highest variability was summer, followed by winter and fall. Consequently, the results showed that their method produced the least sharp PIs during summer, followed by winter and fall, as could be expected by the variability mentioned earlier. In addition, the coverage probability of the PIs produced by the best model remained above the nominal coverage level, indicating that the PIs were valid (Khosravi et al., 2013). Secondly, Davò et al. (2016) used an analog ensemble (AnEn) to predict regional wind power and solar irradiance in Sicily, Italy and Oklahoma, USA, respectively. The AnEn, originally proposed in Delle Monache et al. (2013), combines multiple deterministic forecasts from the past conditional on observations of the current state. The authors showed that the continuous ranked probability score (CRPS, negatively oriented score and formally introduced in Section 2.4) fluctuated substantially as a function of season, with the highest score achieved

during summer and the lowest score during fall. Finally, Wang et al. (2017a) utilized a deep convolutional neural network (DCNN) in combination with quantile regression (QR) to produce probabilistic forecasts of PV power production and reported that CRPS was lowest during winter, and generally highest in summer, where the latter was attributed to the local volatile weather conditions. The location in the study was Flanders, Belgium.

From this literature review, we can conclude that it is reasonable to expect deteriorating performance of the PIs when the variability of the data set increases. Although the papers mentioned here concerned solar power or solar irradiance, it is reasonable to expect similar findings in load forecasting due to the dependence of electricity demand on weather variables, see e.g., Drezga and Rahman (1998); Raza and Khosravi (2015). Moreover, due to climatological differences, it is difficult to predict what season will be least straightforward to predict. Studies into the variability of the data set per season are therefore required, which will be done in Section 2.1.

1.1.3. The effect of penetration

The effect of large scale integration of renewable energy sources (RESs) such as PV power on the power system is an important topic of research because it directly affects the net load, which is the amount of power the utility has to provide or absorb. Several studies exist that study this effect, see e.g., Shaker et al. (2016) for a study into the impact of large scale wind and solar power penetration, who noted that the volatility of the net load could increase up to eight times, making accurate forecasts ever more challenging. Another notable example is the study by Olauson et al. (2016), who investigated the net load variability in Scandinavia with high or full integration of RESs in the power system, and the optimal mix of RESs. They concluded that the optimal mix depends on the time span under consideration, e.g., diurnal or seasonal, and, more importantly, a fully renewable power system is possible with diverse RESs backed up by hydropower.

Various studies investigated net load forecasting with high PV penetration, although only one considered probabilistic forecasting, to the best of our knowledge. Wang et al. (2017b) built their work on the premise that the variable PV power made the net load more challenging to forecast. Therefore, the authors used correlation analysis based on the maximal information coefficient (MIC) to separate net load into three time series, i.e., PV power, load and residual, in order to produce probabilistic forecasts with higher accuracy. They concluded that, for penetration levels up until 20%, separation of net load yielded more accurate forecasts. With regards to the overall performance, it was clearly shown that increasing PV penetration reduced the accuracy of the probabilistic forecasts. However, since the study took place on zonal level data of an independent system operator (ISO), it is challenging to say whether this generalizes to lower levels of the electricity grid.

As for the deterministic studies, Kaur et al. (2013), as mentioned before, showed that the variability of PV power is the main reason behind the error of the load forecast. However, we think their conclusion is incomplete since it appears a single solar farm was used, consequently disregarding the smoothing effect. It is therefore important to study what the effect of distributed generators (DGs) is on the accuracy of net load forecasts, both on the individual and aggregated level. Furthermore, Kaur et al. (2016a) went on to compare two approaches to forecasting net load: an integrated and an additive approach. The former entailed using the PV power forecast as input to the net load forecast, while the latter meant forecasting load and PV power separately and subtracting these to acquire the net load forecast. They found that the integrated approach yielded substantially more accurate results, for varying levels of penetration due to weather factors. However, in both Kaur et al. (2016a, 2013), forecasts were solely deterministic. Finally, Chu et al. (2017) investigated the net load forecasts on four feeders with solar penetration levels ranging from 4% to 58%. Although the forecasts were deterministic, the results clearly showed increasing error for higher levels of penetration.

1.2. Aim and contributions

As a starting remark, it is important to point out that the purpose of this study is not to produce the most accurate forecasts. Considering this, we identify several research gaps from the literature review in Section 1.1. Firstly, although a significant amount of research has focused on the smoothing effect, both for PV and load forecasting, no studies other than van der Meer et al. (2017) have been performed that investigate the effect of aggregation with respect to probabilistic forecasting. In the present study, we aim to continue our earlier work and investigate more thoroughly how aggregation affects PI accuracy by using more samples to reduce noise and by using QR in addition to the Gaussian process (GP) in order to produce both non-parametric and parametric PDFs.

Secondly, to the best of our knowledge, only Wang et al. (2017b) considered the effect of penetration on probabilistic forecasts. However, as mentioned before, the work considered the aggregated net load on ISO level, and it is therefore questionable whether their results generalize to lower levels of the electricity grid on individual and aggregated scale.

Thirdly, our previous work van der Meer et al. (2018a) indicated the necessity for a thorough analysis into the optimal length of training window width M of the dynamic GP, which produced more narrow PIs than the static GP. Although these results may not be generalized directly to other probabilistic models, it should give an indication on the sensitivity with respect to training data length. Moreover, we think that the methodology for finding M can be generalized to other probabilistic models in an online setting that need to handle large amounts of data and regularly update the (hyper-) parameters.

The final aim of this paper is to study the aforementioned during two distinct seasons, i.e., winter and spring, typically recognized by relatively low and high variability, respectively. Therefore, the contributions of this paper are fourfold:

1. We analyze the effect of aggregation on PIs.
2. We study the influence of seasons on PIs.
3. We investigate the effect of increasing penetration of PV power on net load forecast accuracy.
4. We study the optimal training window M for the dynamic GP.

The rest of this paper is organized as follows: Section 2 introduces the methodology, the GP and QR, and the probabilistic forecasting metrics. Section 3 presents the results and the discussion of the results. We finalize the paper with our conclusions in Section 4.

2. Methodology

2.1. Data

We use a publicly available data set that contains residential electricity consumption and rooftop PV power generation of 300 de-identified customers in the Sydney metropolitan area (Ausgrid, 2014). Figure 1a presents a choropleth of the area with the density of customers in the data set based on the postcodes. The time series have been measured by the grid operator on half hourly basis from 1 July 2010 until 30 June 2013, after which we have normalized all time series between 0 and 1 according to $y' = (y - y_{\min}) / (y_{\max} - y_{\min})$, in order to be able to compare CRPS, which is scale dependent. Taieb et al. (2016) took a similar approach but divided the time series by their maximum value, which in our case would result in differently distributed time series since that of net load can be both positive and negative.

2.1.1. Optimal training window width M

In order to find the optimal training window width M , we use data of the year 2011-2012, so as not to use any of the data to investigate the effect of aggregation and penetration. We do this twice, once on the individual level, and once on the aggregated level. For the individual level, we randomly select 15 customers from the list of 54 customers that have high quality data for all three years, according to Ratnam et al. (2017), and normalize each time series individually. In case of the aggregated level, 10 aggregations of 20 customers are randomly selected from the same list, after which each of the 10 aggregations is normalized. Subsequently, we increase the size of M from 1 day to 28 days, and use the learned model to predict the output of the subsequent day in a one-step ahead manner. It should be noted that in case of the dynamic GP, introduced in Section 2.2.1, prediction of the subsequent day in a one-step ahead manner remains the approach. Afterwards, the results are averaged so as to reduce the level of

noise induced by e.g., abnormal behavior of an individual customer. Figure 1b presents the strategy.

It should be noted that Scolari et al. (2016) take a similar approach to determine the optimal training window, although our study was not inspired by theirs. Furthermore, this procedure is only considered for the GP. The reason for this is the computational burden due to matrix inversion, which is computationally expensive as will be explained in Section 2.2. Training the QR model requires solving a linear program, which can be done efficiently as will be explained in Section 2.3.

2.1.2. Aggregation of power profiles

According to Ratnam et al. (2017), 187 customers have high quality data for 2012-2013. However, the authors specify that their focus is on acquiring high quality data rather than high quantity. For this study, we therefore use less strict constraints in order to obtain a larger data set. More specifically, we remove all customers of whom the PV systems produce less than 0.04 kW over the course of the entire day. Further, we remove customers of whom the PV systems produce more than 0.03 kW during nighttime and finally, we remove customers of whom the PV systems produce less than 0.3 kWh over the course of an entire day. After this process we are left with 240 customers. In order to get a proper insight into what happens during the aggregation process, we use a coarse and a fine aggregation resolution. The former implies aggregating 1 to 240 customers in steps of 30, while the latter implies aggregating 1 to 30 customers in steps of 1. We randomly select customers for each aggregation step without replacement, which means that each step contains unique customers. However, these customers can be drawn again in subsequent aggregation steps because of a limited number of customers.

2.1.3. Penetration of PV power

In order to assess the effect of penetration of PV power into net load, we increase the installed capacity with increments of 10% to 200% for 1, 30 and 210 customers. Subsequently, we subtract the PV power from the electricity consumption to acquire the net load. Each time we perform this calculation, we normalize afterwards in order to be able to compare the results. In order to reduce noise, the cases are repeated 5 times. We do not take 240 customers in the final case because that would not allow us to randomize the customers, although at this level of aggregation we would expect little noise from individual customers. It should be noted that the customers and aggregations thereof used in this section are the same samples as in Sections 2.1.1 and 2.1.2.

2.2. Gaussian process

In this study we use the GP, which is a non-parametric and nonlinear probabilistic model, and is defined as a collection of random variables that has a joint multivariate

Figure 1: In (a), a choropleth map of the Sydney metropolitan area presenting the density of customers in the data set based on the postcodes. For the colored version, the reader is referred to the web version of this article. In (b), a visual aid explaining the procedure elaborated upon in Section 2.1.1 to find optimal training window width M and is inspired by van der Meer et al. (2018a).

Gaussian distribution described by mean μ and covariance matrix $\mathbf{K}$. The GP, and its application in machine learning, have been extensively described by Rasmussen and Williams (2006). In our previous work (van der Meer et al., 2018a) we compared a static versus a dynamic GP, where the former was trained using a fixed amount of training data (6 months), while the latter was trained using 200 data points, slightly over 4 days in case of half hourly data. The dynamic GP had two advantages: Firstly, it produced sharper PIs; and secondly, it was much faster to train and infer the posterior distribution from. Therefore, we disregard the static GP in this study and focus on the dynamic alternative.

2.2.1. Dynamic GP

As stated before, the GP can be described by mean μ and covariance matrix $\mathbf{K}$. The latter represents the relationship between the inputs and can be defined in a time series setting as (Roberts et al., 2012):

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \begin{pmatrix} k(\mathbf{x}_{t-1}, \mathbf{x}_{t-1}) & \cdots & k(\mathbf{x}_{t-1}, \mathbf{x}_{t-N}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_{t-N}, \mathbf{x}_{t-1}) & \cdots & k(\mathbf{x}_{t-N}, \mathbf{x}_{t-N}) \end{pmatrix}, \quad (1)$$

where vector $\mathbf{x}_{t-i} = (y_{t-1-i}, y_{t-2-i}, \dots, y_{t-L-i})$, and where y_{t-i} represents the i th lagged observation and L the maximum lag, such that $\mathbf{X} = (\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \dots, \mathbf{x}_{t-N})^T$. Here, N represents the length of the training data. Covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$ describes the relationship between $\mathbf{x}_i$ and $\mathbf{x}_j$, and is therefore an important design choice. In our previous work (van der Meer et al., 2018a) we tested a variety of (combinations of) covariance functions and found the summation of the squared exponential (SE) and

Matérn 3 to perform best on this data set. It is defined as:

$$k(\mathbf{x}_i, \mathbf{x}_j)(r) = \sigma_f^2 \exp\left(-\frac{r^2}{2\ell^2}\right) + \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\ell}\right) \exp\left(-\frac{\sqrt{3}r}{\ell}\right), \quad (2)$$

where $r = \|\mathbf{x}_i - \mathbf{x}_j\|$, and ℓ and σ_f^2 are hyperparameters representing the characteristic length and amplitude, respectively, to be learned by maximizing the log marginal likelihood, defined as (Rasmussen and Williams, 2006):

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}| - \frac{N}{2} \log 2\pi. \quad (3)$$

In case of a new observation $\mathbf{x}_t$, the cross-covariance between that point and $\mathbf{X}$ can be expressed as:

$$\mathbf{K}(\mathbf{x}_t, \mathbf{X}) = (k(\mathbf{x}_t, \mathbf{x}_{t-1}), k(\mathbf{x}_t, \mathbf{x}_{t-2}), \dots, k(\mathbf{x}_t, \mathbf{x}_{t-N})), \quad (4)$$

and subsequently, the joint distribution including the latest observation can be defined as:

$$p\left(\begin{bmatrix} \mathbf{f} \\ f(\mathbf{x}_t) \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu(\mathbf{X}) \\ \mu(\mathbf{x}_t) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) & \mathbf{K}(\mathbf{X}, \mathbf{x}_t) \\ \mathbf{K}(\mathbf{x}_t, \mathbf{X}) & k(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix}\right), \quad (5)$$

where $\mathbf{f} = (\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-N})^T$. The joint distribution can be used to calculate the posterior distribution. It is formulated as follows:

$$\hat{y}_t = \mu(\mathbf{x}_t) + \mathbf{K}(\mathbf{x}_t, \mathbf{X}) \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1} (\mathbf{f} - \mu(\mathbf{X})), \quad (6)$$

$$\hat{\sigma}_t^2 = k(\mathbf{x}_t, \mathbf{x}_t) - \mathbf{K}(\mathbf{x}_t, \mathbf{X}) \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{K}(\mathbf{X}, \mathbf{x}_t). \quad (7)$$

Equations (3), (6) and (7) are the reasons we prefer to keep N as small as possible. Namely, matrix inversion

costs $\mathcal{O}(N^3)$ in time, which could pose operational issues in an online setting. Moreover, in case of data scarcity, it is highly interesting to have a method that requires a small amount of training data, which is the case for the GP according to Salcedo-Sanz et al. (2014). In order to implement the GP, we use the `GPML` package, which complements Rasmussen and Williams (2006). We refer to van der Meer et al. (2018a) for a complete description of the dynamic GP used in this work.

2.3. Quantile regression

Although the advantages of the GP, e.g., that it is a non-parametric and nonlinear probabilistic model, make it an attractive model to use, it has disadvantages as well. One example is the output of the GP, which is a Gaussian PDF. It has been argued in the literature that this is too restrictive, see e.g., Golestaneh et al. (2016). Therefore, we additionally investigate the topics of this study (outlined in Sections 2.1.1 to 2.1.3) using QR, which was proposed by Koenker and Bassett (1978). The core idea of QR is to extend linear regression (LR), where, instead of ordinary least squares (OLS) minimization, one minimizes the mean absolute error to estimate the conditional median. This can then be extended further to estimate conditional quantiles by using a tilted absolute value function that applies asymmetric weights to positive and negative errors (Koenker and Bassett, 1978). First, similarly as for LR, we establish a linear relationship between the dependent variable y and independent variables $\mathbf{x}$:

$$y_{t,\tau} = \boldsymbol{\beta} \cdot \mathbf{x}_t + \epsilon, \quad (8)$$

where τ represents the quantile probability ($0 < \tau < 1$), $\boldsymbol{\beta}$ the parameters and ϵ the model error. The tilted function, also known as the pinball loss function, is defined as (Koenker and Bassett, 1978):

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if } u \geq 0 \\ (\tau - 1)u & \text{if } u < 0, \end{cases} \quad (9)$$

and can be used in the following minimization problem to find optimal parameters $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}_\tau = \arg \min_{\boldsymbol{\beta}} \sum_{t=1}^T \rho_\tau(y_t - \boldsymbol{\beta} \mathbf{x}_t). \quad (10)$$

Due to the fact that each quantile is estimated independently, quantile crossing could violate the monotonicity property (Taieb et al., 2016):

$$\hat{y}_{\tau_1} \leq \hat{y}_{\tau_2} \quad \forall \tau_1, \tau_2 \text{ so that } \tau_1 \leq \tau_2, \quad (11)$$

which we circumvent by monotone rearranging.

As can be seen from its formulation, the QR model is a linear and parametric model that produces a non-parametric PDF. It should be noted that more advanced techniques exist that utilize QR, such as QR forests (QRFs)

(Meinshausen, 2006) or QR neural networks (QRNNs) (Canon, 2011), and these will be the focal point of future work. In terms of model configuration, the QR will be identical to the GP, i.e., the same number of predictors will be used. However, as stated before, training and inference time is less of an issue with QR and the model will therefore be trained on all seasons of that particular year, except the test season of interest. As a final note, we use the `quantreg` package in R developed by Koenker (2016) to implement the QR model.

2.4. Performance metrics

2.4.1. Quantitative metrics

Accurate and meaningful PIs have two features, i.e., they are narrow and cover many of the observations. The former implies *informative* PIs that have high sharpness, which can be quantified by the PI normalized average width (PINAW), which is a negatively oriented score, i.e., a low score is better. The latter pertains to the *reliability* of PIs, which can be quantified by the PI coverage probability (PICP) and is a positively oriented score, i.e., a high score is better. As will become clear from their respective formulations, these counterbalance each other, but are both important when assessing the quality of the PIs. After all, high PICP can easily be achieved by similarly high PINAW, and extremely low PINAW most likely results in low PICP.

The PINAW is formulated as follows (Quan et al., 2014):

$$\text{PINAW} = \frac{1}{TR} \sum_{t=1}^T (U_t - L_t), \quad (12)$$

where T represents the length of the test sample, L_t and U_t represent the lower- and upper-bound of the PIs, respectively, and R represents a normalization factor, defined as the maximum minus the minimum of the test sample.

The PICP is defined as (Quan et al., 2014):

$$\text{PICP} = \frac{1}{T} \sum_{t=1}^T \epsilon_t, \quad (13)$$

where $\epsilon_t = \mathbb{1}_{\{y_t \in [L_t, U_t]\}}$. As mentioned before, PICP expresses the reliability of the PIs. According to Khosravi et al. (2013), PICP should be higher than the nominal coverage level η , which is 80% in this study.

As a final metric, we employ the CRPS, which is a negatively oriented score. This metric has the advantages that it quantifies both the reliability and informativeness of PIs, and that it reduces to the mean absolute error (MAE) in case of a deterministic forecast (Gneiting and Katzfuss, 2014). Consequently, it has the same units as the forecast (Gneiting and Katzfuss, 2014). The CRPS can be formulated as follows (Gneiting and Katzfuss, 2014):

$$\text{CRPS}(F, y) = \mathbb{E}_F |Y - y| - \frac{1}{2} \mathbb{E}_F |Y - Y'|, \quad (14)$$

where F is the cumulative distribution function (CDF) of the probabilistic forecast and Y and Y' are independent samples drawn from F . In order to calculate CRPS, we use the package `SpecsVerification` by Siegert (2017).

2.4.2. Visual verification

The reliability diagram is a graphical verification tool that allows one to assess the reliability of a probabilistic forecast in a straightforward manner. The core concept of the reliability diagram in the case of continuous variables is that the observed frequency is plotted against the quantile probabilities. Ideally, forecasts are close to the diagonal, here indicated by the nominal quantile probabilities τ . This can be calculated by counting the number of occurrences a quantile forecast overpredicts and calculating the average over the test period (Pinson et al., 2010):

$$\hat{a}_\tau = \frac{1}{T} \sum_{t=1}^T \mathbb{1}_{\{y_t < \hat{y}_{t,\tau}\}} \quad \forall \tau. \quad (15)$$

However, deviations from the diagonal can occur due to a limited amount of forecast and observation pairs, despite being reliable (Pinson et al., 2010). Therefore, it is important to present the likelihood of the observed frequencies at each quantile as well, which can be done using consistency resampling to produce consistency bars (Bröcker and Smith, 2007). Consistency resampling is a process that is repeated N times, where we first create surrogate observations $\tilde{y}_{t,i}$ using the predictive distribution at time t and repetition i . If we assume $\hat{F}_t(Y_t) \sim U[0,1]$, the series of predictive densities is reliable and we can produce surrogate observations via inverse transformation, i.e., $\hat{F}_t^{-1}(u_i) = \tilde{y}_{t,i}$, where u_i is a draw from $U[0,1]$. Subsequently, it is possible to find the distribution of $\hat{a}_{\tau,i}$ for $i = 1, \dots, N$ of which we can infer the range of the consistency bars at each quantile probability τ . A probabilistic forecast is said to be reliable if the reliability curve falls within the consistency bars. It should be noted that this approach is similar to that described by Pinson et al. (2010), except we do not account for serial correlation, which would effectively increase the size of the consistency bars.

3. Results and discussion

3.1. Data analysis

3.1.1. Variance analysis

Figures 2a and 2b display boxplots of the variances of the normalized time series as a function of season for individual customers and aggregated customers, respectively. The figures indicate two important premises on which this study builds. Firstly, winter and spring appear to provide the extreme cases in terms of variance for all quantities considered, i.e., electricity consumption (EC), net demand (ND, i.e., net load) and PV power production (PV). For

example, PV has least variability during the winter for individual customers and aggregated customers, while it is high during spring for both cases. Conversely, variability of electricity consumption is in both cases high during winter, whereas it is relatively low during the spring. Therefore, we use only the winter and spring seasons in our further analysis for brevity. Secondly, the figures clearly show that there is more spread in the results for the individual customers. In order to reduce noise in the results, it is important to use more samples in case of individual customers (cf. Section 2.1.1) and look at the effect of increasing the number of customers in both a coarse and fine manner (cf. Section 2.1.2). Finally, it is important to stress that the variances reported in (a) and (b) cannot be compared among each other, and that the variances can only be compared within their respective figures. The reason for this is that the time series have been normalized between 0 and 1 separately.

Figures 2c and 2d present the time series for three levels of aggregation of 5 samples from the data set during winter and spring, respectively. These figures aim to show the impact that aggregation has on the variability of the time series.

3.1.2. Aggregation

Figures 2c and 2d present the effect aggregation has on electricity consumption, PV power production and net load during winter and spring, respectively. It should be noted that all time series are normalized separately, except for Figures 2c and 2d. The figures clearly indicate the significant impact of aggregation on the variability. In case of the individual customers, electricity consumption appears to be almost entirely stochastic, whereas daily patterns become exceedingly clear when we combine 30 and 210 customers. A similar observation can be made for the PV power, and it is therefore reasonable to expect that aggregating a relatively small number of customers can improve PI accuracy.

3.1.3. Penetration

Figure 3a presents a histogram of the penetration levels in logarithm scale achieved by the original PV systems, defined as the peak PV production as a fraction of peak electricity consumption (Hoke et al., 2012) in orange, and the peak PV production as a fraction of the electricity consumption at the moment the peak power was produced by the PV system in blue. The figure shows that most of the customers have penetration levels around or below 25% using the original definition, but that this can increase dramatically. Figure 3b gives an insight into the effect of an increasing PV system size on the net load. The ramps increase for both levels of aggregation and it therefore seems reasonable to expect that increasing the penetration of PV power in the net load reduces PI accuracy.

Table 1: Selected training window widths M (in days) for the quantities EC, PV and ND, classified based on season and aggregation level.

Season	Aggregation	EC	PV	ND
Winter	Individual	14	8	16
	Aggregated	14	12	8
Spring	Individual	14	8	21
	Aggregated	6	8	9

3.2. Optimal training window width M

In this section, we briefly mention the results of the method described in Section 2.1.1 by means of Table 1, which presents all training window widths M . The full analysis hereof is presented in Appendix A. Overall, the results show that the dynamic GP can achieve satisfactory with training data amounting up to 21 days, making it an interesting method that can be used in case of data scarcity.

3.3. The effect of aggregation on PIs

In this section, we investigate the effect of aggregation on PIs using a fine and coarse resolution, i.e., aggregating 1 to 30 customers with increments of 1, and aggregating 1 to 240 customers with increments of 30. Figures 4a to 4l present the results of PINAW, PICP and CRPS as a function of the aggregation levels. Furthermore, Figures 5a to 5f present the reliability diagrams and Figures 6a to 6l present conventional forecast plots during a day in winter and spring for the first individual and aggregated customers of the random set of aggregations. The remainder of this subsection is divided in order to investigate in detail the difference between the fine and coarse aggregation resolution, and the performance of the GP and QR models. Finally, we conclude this subsection with Figure 7 that allows the reader to quickly recapitulate the results of this section. It is perhaps helpful to remind the reader that we use a nominal coverage level of 80% in this study, implying that seek to produce PIs with a coverage probability over 80%.

3.3.1. Fine aggregation resolution

Figures 4a to 4f present the results of the fine aggregation resolution. Of these figures, Figures 4a to 4c present the results of the GP, whereas Figures 4d to 4f present the results of the QR model. Interestingly, Figures 4a and 4d indicate that aggregating the electricity consumption does not substantially improve the forecast accuracy, especially in case of QR. While there is a clear improvement in sharpness of the forecast of the GP during both seasons, the accompanied steady decrease in coverage probability causes the CRPS to only marginally improve over the aggregation levels. However, it is important to note that the empirical coverage of the PIs is higher than the nominal coverage

level. This is not the case for QR, which tends to be overconfident during winter. Since this can be ascribed to the season (cf. Figures 4d to 4f and 4j to 4l) it could be that even though 11 months of training data have been used, using less training data but from a similar period in the preceding year improves its accuracy. As a final note, both the GP and the QR models clearly have lower accuracy during winter than during spring, which is when the variance is higher as well, as was shown in Figure 2b.

The reliability diagrams presented in Figures 5a and 5d show that, in case of the GP, aggregation substantially improves the reliability of the predictive density. The shape for the lower levels of aggregation indicates that the model is not confident enough: lower quantiles frequently predict lower than the observed values whereas the higher quantiles frequently predict higher than the observation (Wilks, 2006). This implies that the PIs produced by the GP can effectively be narrower, e.g., by optimizing the hyperparameters such that PINAW is minimized, subject to the reliability. In case of QR, we observe that there exists a systematic bias in the model during winter, most likely due to a lack of training data of the same time period. Interestingly, aggregation does not seem to affect the reliability during spring and the reason for this is unclear.

In case of PV power forecasting, Figures 4b and 4e display that the smoothing effect has a substantial impact on the PIs. However, it is not possible to see where the PV systems are located due to the de-identified data. Especially in case of the GP, we observe a substantial gain in both sharpness and reliability for both seasons when aggregating a modest number of PV systems. In case of QR, the PINAW decreases when increasing the aggregation level but at the cost of deterioration of the PICP, which results in an empirical coverage that is lower than the nominal one. In both cases, CRPS is lower during winter than during spring due to lower variability. It is important to note here that CRPS does not penalize PICP being lower than the nominal coverage level because the score takes the entire CDF into consideration, and therefore it remains lower during winter than during spring.

Figures 5b and 5e present the reliability diagrams for the probabilistic PV power forecasts. In case of the GP, we note that the effect of aggregation on the predictive density is not as pronounced as indicated by Figures 4b and 4h. This might be due to the combination of decreasing PINAW and increasing PICP as shown in the aforementioned figures, i.e., the predictive CDF sharpens but not enough to ensure that it becomes as narrow as the CDF of the observations. Figure 5e shows that during spring, the QR model is biased and aggregation does not improve this. However, during winter, we can observe that aggregation improves the reliability, especially when aggregating from 1 to 30 customers, as was also shown in Figure 4e.

Finally, Figures 4c and 4f present the effect of aggregation on PIs for net load. In both cases, we observe that the accuracy during winter is lower than during spring, which was also the case for the electricity consumption forecasts.

This, and the development of the scores over the aggregations, suggest that the accuracy of net load forecasts is largely dependent on the variability of the electricity consumption for increasing levels of aggregation, as we noted in the previous section. The reason for this is most likely the significant contribution of the smoothing effect in case of PV power, since the aggregations taken here are simply samples of the individual customers. However, this does not imply that the effect of combining PV power and electricity consumption is trivial. To the contrary, the figures show that all scores of the net load forecasts have improved over those of the electricity consumption forecasts, indicating that including an aggregated PV power profile, i.e., a smoothed profile, improves predictability. The reason for this is most likely that, according to Figures 4b and 4e, aggregation has the most pronounced effect in case of PV power production. In addition, Figure A.10c showed that on an individual level, combining electricity consumption and PV power yielded worse forecast performance, while Figure A.10f indicated that on an aggregated level, the net load forecasts were sharper than those of electricity demand. Figures 4c and 4f corroborate the latter results, since the PINAW of the net load forecasts has improved over the PINAW of the electricity consumption forecasts. Interestingly, Shaker et al. (2016) came to a similar conclusion, albeit not in the case of forecasting but in the case of estimating regional PV power production of unobserved sites. They noted that by increasing the capacity of a sub region, the estimation accuracy was improved due to increased aggregation and the subsequent smoothing.

Figures 5c and 5f again show the reliability of the entire predictive density. Similar as was the case for electricity consumption, we can observe here that aggregation has a positive effect on the entire predictive density of the GP. Again, we can see as well that reliability during winter is lower. Moreover, Figure 5c confirms that combining electricity consumption and PV power production improves the accuracy of probabilistic forecasts when compared to the separate forecasts. The reliability diagram of the QR model partly corroborates this, specifically in case of the forecasts of a single customer. However, the bias already present in the electricity consumption forecasts for 210 customers shows an increase for higher levels of aggregation during winter.

3.3.2. Coarse aggregation resolution

Figures 4g to 4l present the results of the coarse aggregation resolution. While in the previous section we noted that aggregating electricity consumption did not substantially improve the forecast accuracy, Figures 4g and 4j indicate that aggregation on a larger scale definitely improves the sharpness of the PIs, both for the GP and QR. Furthermore, in case of the GP, there is a clear decrease in coverage probability as a consequence of the first aggregation step, caused by the increased sharpness of the PIs. The most likely reason for this is that the GP is not trained to optimize sharpness subject to the coverage probability,

and the drastic reduction in PINAW results in a reduction of PICP due to the still variable nature of an aggregation of 30 customers. This is corroborated by the results from the following aggregation steps, where the PINAW improves further while the PICP remains fairly constant. In contrast, the coverage probability of the QR PIs remains fairly constant through the aggregation steps, which was also shown in the previous section. Finally, in both cases performance during spring is better although the difference is smaller and converges towards each other, owing to the aggregation steps.

In the case of the PV power forecasts, Figures 4h and 4k indicate that the first aggregation step is the most useful in terms of improving the accuracy of the PIs, especially in terms of the sharpness and, by extension, CRPS. The reason that CRPS improves substantially as well is that, unlike in case of the electricity consumption forecasts, the coverage probability does not decrease drastically, if it decreases at all. This supports our previous observation, namely that the first aggregation step has a substantial impact on these time series. Furthermore, there is a clear difference in performance between the seasons, albeit the opposite between the two models. More specifically, the GP performs substantially better during winter than during spring due to the lower variability, whereas the QR model performs better, in terms of PICP, during spring. However, the former is what one should expect, whereas the latter is a returning phenomenon in this study for QR, and might be improved by using less training data from a similar period in the preceding year, as mentioned before.

Finally, Figures 4i and 4l present the results of the net load forecasts. In these figures, the positive impact of combining electricity consumption and PV power time series becomes clear also for higher levels of aggregation. In case of the GP, the sharpness of the PIs has improved over the sharpness of the electricity consumption forecasts, while the coverage probability has improved during the spring and remained similar during winter. We can observe a similar pattern for the QR model, where the inclusion of PV power improves the sharpness of the net load forecasts over the sharpness of the electricity consumption forecasts during both seasons. During spring, this is accompanied by a substantial improvement in PICP as well. From these observations we infer that including the PV power into the net load during periods with high variability and aided by the smoothing effect, can reduce the higher error of the PV power forecasts, as observed in Figures 4b, 4e, 4h and 4k.

Figures 6a to 6l present conventional forecast plots during a day in spring and a day in winter of both models for three levels of aggregation. These figures show that both models display difficulty predicting peaks using the 80% nominal coverage level. Furthermore, we observe that especially the width of the PIs decreases with an increasing level of aggregation, while the coverage probability remains satisfactory. In addition, both models exhibit more uncertainty during noon in case of PV power production and the evening in case of electricity consumption and the net

load, expressed by the increase in PI width during these periods.

3.4. The effect of penetration on PIs

In this section we investigate the effect on the PIs of the net load forecasts when the PV system size increases with 10% increments to 200%. Figures 8a to 8f present the results in terms of PINAW, PICP and CRPS, while Figures 9a to 9f present the reliability diagrams for three penetration levels of PV power in the net load: 110%, 130% and 200%. In these figures, it is interesting to note the differences between the reliability diagrams here and Figures 5c and 5f, which is due to the selection of different random samples of customers.

At an individual level, it becomes clear that increasing the PV system size only slightly deteriorates the coverage probability in case of the GP during both seasons, whereas in case of QR, the coverage probability marginally improves and deteriorates during spring and winter, respectively. These results, although less pronounced, are similar to those in Appendix A.1, where it was found that combining PV and electricity consumption on the individual level reduced the forecast accuracy. Taking into consideration the results from Sections 3.3.1 and 3.3.2, the reason for the reduced accuracy is most likely the absence of the smoothing effect in case of PV power. Figures 9a and 9d confirm that the level of penetration has little influence on the reliability of the predictive density.

In Figures 8b and 8e, we can observe the effect of increasing the PV system size in case of 30 aggregated customers. However, in contrast to the results reported in (Chu et al., 2017; Kaur et al., 2016a; Wang et al., 2017b), our results indicate that the sharpness increases as the system size increases, while the coverage probability remains fairly constant. The most likely explanation for these results is the added smoothing effect, which was substantial when aggregating the output of 1 to 30 PV systems, as shown in Figures 2c and 2d. This effect propagates positively into the accuracy of net load forecasts, as shown in Section 3.3.1. The most likely explanation for the discrepancy between the results reported here and those in the aforementioned studies is the scale. The aforementioned studies considered data sets in which peak loads of 1939 MW (Wang et al., 2017b), 2.5 MW (Kaur et al., 2016a) and 2.60 MW - 6.39 MW (Chu et al., 2017) occurred, compared to 111.46 kW on the aggregated level of 30 customers. Figures 9b and 9e show that the reliability slightly improves with increasing penetration level in case of the former, and remains similar in case of the latter.

Finally, we can observe that the positive effect of increasing the PV system size continues in Figures 8c and 8f, where, especially in case of the GP, the smoothing effect improves the net load forecasts during the spring, when variability was more substantial as shown in Figures 2b and 2d. The aforementioned results are consistent with those found in Sections 3.3.1 and 3.3.2. This implies that in case of substantial smoothing, increasing the PV power

penetration in the electricity mix in electricity distribution grids can have a positive influence on probabilistic forecasts. Figures 9c and 9f indicate that the penetration level hardly influences the reliability of the forecasts during either season.

4. Conclusion

In this work we investigated the effect of the aggregation of time series and an increasing share of photovoltaic (PV) power in the net load on prediction intervals (PIs) in local electricity distribution grids. This study was performed during two seasons, namely winter and spring on the Southern Hemisphere, where the former season represented a challenging case for electricity consumption, while the latter represented a challenging case for PV power production. The PIs were produced using two models with distinct features: quantile regression (QR) as a linear and parametric model, capable of producing a non-parametric probability density function (PDF), and the Gaussian process (GP), a nonlinear and non-parametric model that produces a parametric PDF (Gaussian). For the latter, we built on previous work to find the optimal training window width depending on the aggregation level and season. The most important conclusion of that component of this work was that the dynamic GP was capable of producing PIs with high accuracy with very limited amounts of training data, making it an ideal method in case of data scarcity.

When studying the effect of aggregation on PIs, the results indicated that a substantial gain in both the PI normalized average width (PINAW) and the PI coverage probability (PICP) can be realized during both seasons when aggregating a modest number of PV systems, most likely due to the smoothing effect. However, the reliability diagrams also indicated that the benefit of the smoothing effect can only be fully exploited when the sharpness is maximized subject to reliability, something that requires a specialized learning technique. In addition, we noted that the positive effect that smoothing has on PV power propagated into the net load forecasts, especially for low levels of aggregation. This is an interesting result because during periods with high variability of PV power output, i.e., spring, when the accuracy of the PV power forecast tended to be lower, forecasting net load directly resulted in higher accuracy than forecasts of electricity consumption alone. Moreover, it also suggested that a bottom-up approach with small aggregations of customers can be highly beneficial for the accuracy. Furthermore, higher levels of aggregation showed that the performance of PIs increased further.

Finally, we increased the share of PV power in the net load to assess the effect on the PIs. While the reliability of the PIs decreased slightly at the individual level, the results showed that at higher levels of aggregation, increasing the share of PV power actually improved the sharpness of

the PIs. The explanation for this is most likely the smoothing effect that, as stated before, had a positive effect on the net load forecasts. This result deviates from other studies in the literature, which is most likely due to the scale difference between those studies and this one. On the local distribution grid, the impact of aggregation is more pronounced than on higher levels of the electricity grid, and the inclusion of PV power during variable periods can enhance net load forecasts.

Acknowledgments

This work was financially supported by SamsPEL 2016 - 2020 in the project "Development and evaluation of forecasting models for solar power and electricity use over space and time", financed primarily by the Swedish Energy Agency.

References

- Aguero, J.R., Steffel, S.J., 2011. Integration challenges of photovoltaic distributed generation on power distribution systems, in: 2011 IEEE Power Energy Soc. Gen. Meet., pp. 1–6. doi:10.1109/PES.2011.6039097.
- Ausgrid, 2014. Solar home electricity data. URL: <https://www.ausgrid.com.au/Common/About-us/Corporate-information/Data-to-share/Solar-home-electricity-data.aspx>. online; accessed 31 August 2017.
- Bröcker, J., Smith, L.A., 2007. Increasing the reliability of reliability diagrams. *Wea. Forecasting* 22, 651–661. doi:10.1175/WAF993.1.
- Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.* 37, 1277–1284. doi:10.1016/j.cageo.2010.07.005.
- Chu, Y., Pedro, H.T., Kaur, A., Kleissl, J., Coimbra, C.F., 2017. Net load forecasts for solar-integrated operational grid feeders. *Sol. Energy* 158, 236–246. doi:10.1016/j.solener.2017.09.052.
- Davò, F., Alessandrini, S., Sperati, S., Delle Monache, L., Airoldi, D., Vespucci, M.T., 2016. Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. *Sol. Energy* 134, 327–338. doi:10.1016/j.solener.2016.04.049.
- Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic Weather Prediction with an Analog Ensemble. *Mon. Weather Rev.* 141, 3498–3516. doi:10.1175/MWR-D-12-00281.1.
- Denholm, P., Margolis, R.M., 2007. Evaluating the limits of solar photovoltaics (PV) in electric power systems utilizing energy storage and other enabling technologies. *Energy Policy* 35, 4424–4433. doi:10.1016/j.enpol.2007.03.004.
- Drezga, I., Rahman, S., 1998. Input variable selection for ANN-based short-term load forecasting. *IEEE Trans. Power Syst.* 13, 1238–1244. doi:10.1109/59.736244.
- Gneiting, T., Katzfuss, M., 2014. Probabilistic Forecasting. *Annu. Rev. Stat. Its Appl.* 1, 125–151. doi:10.1146/annurev-statistics-062713-085831.
- Golestaneh, F., Pinson, P., Gooi, H.B., 2016. Very Short-Term Non-parametric Probabilistic Forecasting of Renewable Energy Generation; With Application to Solar Energy. *Power Syst. IEEE Trans.* PP, 1–14. doi:10.1109/TPWRS.2015.2502423.
- Hayes, B., Gruber, J., Prodanovic, M., 2015. Short-Term Load Forecasting at the local level using smart meter data. *IEEE Eindhoven PowerTech*, 1–6. doi:10.1109/PTC.2015.7232358.
- Hoke, A., Butler, R., Hambrick, J., Kroposki, B., 2012. Steady-State Analysis of Maximum Photovoltaic Penetration Levels on Typical Distribution Feeders. *IEEE Trans. Sustain. Energy* 4, 350–357. doi:10.1109/TSTE.2012.2225115.
- Hong, T., Fan, S., 2016. Probabilistic Electric Load Forecasting: A Tutorial Review. *Int. J. Forecast.* 32, 914–938. doi:10.1016/j.ijforecast.2015.11.011.
- IEA, 2017. 2016 Snapshot of Global Photovoltaic Markets. Technical Report. doi:ISBN978-3-906042-42-8.
- Kaur, A., Nonnenmacher, L., Coimbra, C.F., 2016a. Net load forecasting for high renewable energy penetration grids. *Energy* 114, 1073–1084. doi:10.1016/j.energy.2016.08.067.
- Kaur, A., Nonnenmacher, L., Pedro, H.T., Coimbra, C.F., 2016b. Benefits of solar forecasting for energy imbalance markets. *Renew. Energy* 86, 819–830. doi:10.1016/j.renene.2015.09.011.
- Kaur, A., Pedro, H.T., Coimbra, C.F., 2013. Impact of onsite solar generation on system load demand forecast. *Energy Convers. Manag.* 75, 701–709. doi:10.1016/j.enconman.2013.08.026.
- Khosravi, A., Nahavandi, S., Creighton, D., 2013. Prediction intervals for short-term wind farm power generation forecasts. *IEEE Trans. Sustain. Energy* 4, 602–610. doi:10.1109/TSTE.2012.2232944.
- Koenker, R., 2016. quantreg: Quantile Regression in R, version 5.35. URL: <https://CRAN.R-project.org/package=quantreg>.
- Koenker, R., Bassett, G., 1978. Regression Quantiles. *Econometrica* 46, 33–50.
- Lave, M., Kleissl, J., Arias-Castro, E., 2012. High-frequency irradiance fluctuations and geographic smoothing. *Sol. Energy* 86, 2190–2199. doi:10.1016/j.solener.2011.06.031.
- Lusis, P., Khalilpour, K.R., Andrew, L., Liebman, A., 2017. Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Appl. Energy* 205, 654–669. doi:10.1016/j.apenergy.2017.07.114.
- Marcos, J., Marroyo, L., Lorenzo, E., García, M., 2012. Smoothing of PV power fluctuations by geographical dispersion. *Prog. Photovoltaics Res. Appl.* 20, 226–237. doi:10.1002/pip.1127.
- van der Meer, D., Shepero, M., Svensson, A., Widén, J., Munkhammar, J., 2018a. Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes. *Appl. Energy* 213, 195–207. doi:10.1016/j.apenergy.2017.12.104.
- van der Meer, D., Widén, J., Munkhammar, J., 2017. Investigating the Effect of Aggregation on Prediction Intervals in Case of Solar Power, Electricity Consumption and Net Demand Forecasting, in: 7th Sol. Integr. Work., Berlin.
- van der Meer, D., Widén, J., Munkhammar, J., 2018b. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renew. Sustain. Energy Rev.* 81, 1484–1512. doi:10.1016/j.rser.2017.05.212.
- Meinshausen, N., 2006. Quantile Regression Forests. *J. Mach. Learn. Res.* 7, 983–999. doi:10.1111/j.1541-0420.2010.01521.x.
- Mohammadi, P., Mehraeen, S., 2017. Challenges of PV Integration in Low-Voltage Secondary Networks. *IEEE Trans. Power Deliv.* 32, 525–535. doi:10.1109/TPWRD.2016.2556692.
- Nguyen, A., Velay, M., Schoene, J., Zhiglov, V., Kurtz, B., Murray, K., Torre, B., Kleissl, J., 2016. High PV penetration impacts on five local distribution networks using high resolution solar resource assessment with sky imager and quasi-steady state distribution system simulations. *Sol. Energy* 132, 221–235. doi:10.1016/j.solener.2016.03.019.
- Olauson, J., Ayob, M.N., Bergkvist, M., Carpmann, N., Castellucci, V., Goude, A., Lingfors, D., Waters, R., Widén, J., 2016. Net load variability in Nordic countries with a highly or fully renewable power system. *Nat. Energy* 1, 1–8. doi:10.1038/ENERGY.2016.175.
- Perez, R., Hoff, T.E., 2013. Chapter 6 - Solar Resource Variability. In: Kleissl J, editor. *Solar energy forecasting and resource assessment*. Boston: Academic Press.
- Pinson, P., McSharry, P., Madsen, H., 2010. Reliability diagrams for non-parametric density forecasts of continuous variables: Accounting for serial correlation. *Q. J. R. Meteorol. Soc.* 136, 77–90. doi:10.1002/qj.559.
- Quan, H., Srinivasan, D., Khosravi, A., 2014. Short-term load and wind power forecasting using neural network-based prediction intervals. *IEEE Trans. Neural Networks Learn. Syst.* 25, 303–315.

doi:10.1109/TNNLS.2013.2276053.

- Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press.
- Ratnam, E.L., Weller, S.R., Kellett, C.M., Alan, T., 2017. Residential load and rooftop PV generation : an Australian distribution network dataset. *Int. J. Sustain. Energy*, 787–806doi:10.1080/14786451.2015.1100196.
- Raza, M.Q., Khosravi, A., 2015. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renew. Sustain. Energy Rev.* 50, 1352–1372. doi:10.1016/j.rser.2015.04.065.
- Roberts, S., Osborne, M., Ebdon, M., Reece, S., Gibson, N., Aigrain, S., 2012. Gaussian processes for time-series modelling. *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.* 371.
- Salcedo-Sanz, S., Casanova-Mateo, C., Munoz-Mari, J., Camps-Valls, G., 2014. Prediction of Daily Global Solar Irradiation Using Temporal Gaussian Processes. *IEEE Geosci. Remote Sens. Lett.* 11, 1936–1940.
- Scolari, E., Sossan, F., Paolone, M., 2016. Irradiance prediction intervals for PV stochastic generation in microgrid applications. *Sol. Energy* 139, 116–129. doi:10.1016/j.solener.2016.09.030.
- Shaker, H., Zareipour, H., Wood, D., 2015. A Data-Driven Approach for Estimating the Power Generation of Invisible Solar Sites. *IEEE Trans. Smart Grid PP*, 1–11. doi:10.1109/TSG.2015.2502140.
- Shaker, H., Zareipour, H., Wood, D., 2016. Impacts of large-scale wind and solar power integration on California’s net electrical load. *Renew. Sustain. Energy Rev.* 58, 761–774. doi:10.1016/j.rser.2017.05.212.
- Siebert, S., 2017. SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate in R, version 0.5-2. URL: <https://CRAN.R-project.org/package=SpecsVerification>.
- Taieb, S.B., Huser, R., Hyndman, R.J., Genton, M.G., 2016. Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression. *IEEE Trans. Smart Grid* 7, 2448–2455. doi:10.1109/TSG.2016.2527820.
- Vaz, A.G.R., Elsinga, B., van Sark, W.G.J.H.M., Brito, M.C., 2016. An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in Utrecht, the Netherlands. *Renew. Energy* 85, 631–641. doi:10.1016/j.renene.2015.06.061.
- Walling, R.A., Saint, R., Dugan, R.C., Burke, J., Kojovic, L.A., 2008. Summary of distributed resources impact on power delivery systems. *IEEE Trans. Power Deliv.* 23, 1636–1644. doi:10.1109/TPWRD.2007.909115.
- Wang, H., Yi, H., Peng, J., Wang, G., Liu, Y., Jiang, H., Liu, W., 2017a. Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional neural network. *Energy Convers. Manag.* 153, 409–422. doi:10.1016/j.enconman.2017.10.008.
- Wang, Y., Zhang, N., Chen, Q., Kirschen, D.S., Li, P., Xia, Q., 2017b. Data-Driven Probabilistic Net Load Forecasting with High Penetration of Behind-the-Meter PV. *IEEE Trans. Power Syst.*, 1–10doi:10.1109/TPWRS.2017.2762599.
- Widén, J., Carpmann, N., Castellucci, V., Lingfors, D., Olauson, J., Remouit, F., Bergkvist, M., Grabbe, M., Waters, R., 2015. Variability assessment and forecasting of renewables: A review for solar, wind, wave and tidal resources. *Renew. Sustain. Energy Rev.* 44, 356–375. doi:10.1016/j.rser.2014.12.019.
- Wilks, D., 2006. Chapter 7 - Forecast Verification. *Statistical methods in atmospheric sciences*. San Diego: Academic Press.
- Wirth, H., 2017. Recent facts about photovoltaics in Germany. Technical Report. Fraunhofer ISE. URL: <https://www.ise.fraunhofer.de/>.

Appendix A. Optimal training window width M

In this section we select the training window width M for the dynamic GP. It is important to note that the

selection depends on various factors, e.g., interpretation of the PI accuracy by means of the scores and the amount of training data. It is therefore a designer choice that is supported by the results of the procedure explained in Section 2.1.1.

Appendix A.1. Individual level

Figures A.10a to A.10f present the various scores as a function of training window width M . More specifically, Figures A.10a to A.10c present the results for individual customers, whereas Figures A.10d to A.10f present the results for aggregated customers, for two seasons.

Figure A.10a shows that, for individual customers, increasing window the width M has a slight positive effect on the PICP during the winter, while the average width remains constant. During the spring, we observe that increasing M has a positive effect on the sharpness of the PIs, while the coverage probability decreases somewhat. This is reflected in the CRP score, which increases with M during spring and decreases with M during winter. This can be explained using Figure 2a, where it was shown that variance during winter is higher than during spring. Because of the higher variance, it is preferable to consider more past data. Conversely, in case of zero variance, a simple persistence model would suffice because the previous time step would contain all required information. Therefore, we select $M = 14$ in both seasons, mainly because of the fact that improvements for including more days are marginal and because training and inference time increases cubic when including more days. It is interesting to point to the larger spread in the individual results during winter than during spring, indicating the challenging task of forecasting individual electricity consumption during periods with high variability.

Figure A.10b shows rather different results when compared to electricity consumption. More specifically, the overall accuracy during winter is higher than during spring, which is in accordance with Figure 2a. The reason for this is, however, similar as previously mentioned, i.e., M can be smaller because of lower variability in the weather during winter. During highly variable periods, it is generally preferable to increase M in order to properly learn the behavior of the power produced by the PV system. This is reflected in the CRPS plot, which, albeit marginally, steadily decreases as a function of M . Because of this marginal improvement, we select $M = 8$ during spring in order to reduce training and inference time. Regarding the winter season, it is clear that the minimum PINAW and CRPS lie at $M = 8$, which is where PICP is maximum. Similarly as with Figure A.10a, the results display more spread for the season with highest variability, in this case spring.

Figure A.10c presents an interesting case because the variability of the net load depends on both electricity consumption and PV power production. Figure 2a showed that during winter, the variance of the net load is actually slightly lower than that of electricity consumption while

during spring, despite the significantly higher variance of PV power, the variance of net load was only marginally higher than that of electricity consumption. We can observe the consequences of this in Figure A.10c. Firstly, we see that the forecasts are worse in terms of PINAW and PICP during spring than during winter, most likely due to the increased variability of the PV time series and higher output of the PV systems. In addition, there is also more spread in the individual results during spring. Secondly, comparing PINAW and PICP of the net load and electricity consumption forecasts, we can observe that these scores are worse for net load during spring, again indicating that the net load time series has been negatively affected by the highly variable PV time series. During winter, i.e., lower PV power variability, the accuracy of the net load forecasts has slightly improved over those of the electricity consumption forecasts, most likely due to the inclusion of PV power. Because of the complicated nature of the time series during both seasons, it seems fitting to select more training days. More specific, we select $M = 16$ and $M = 21$ during winter and spring, respectively. It is interesting to point out that these results, although not in absolute values, bare more resemblance to those of the PV power forecasts than those of the electricity consumption forecasts. This potentially indicates that, on an individual level, the net load is subject to substantial influence by the PV power.

Appendix A.2. Aggregated level

Figures A.10d to A.10f present the results for 10 aggregations of 20 customers, and it is immediately clear that the spread between the aggregations is far lower than in Figures A.10a to A.10c, as indicated by Figure 2b, which showed that the spread of the variances is low. This is unmistakably the effect of aggregation, which will be further investigated in the following section. Figure A.10d shows a clear deterioration for PINAW when M increases to 7 days and again when M increases to 12 days. Therefore, we take $M = 6$ during spring. During the winter, increasing M improves PINAW but simultaneously deteriorates PICP. In order to reduce training and inference time, we select $M = 14$ as training window width. Finally, it is interesting to note the contrast in CRP scores during winter and summer, clearly representing the increased difficulty of electricity consumption during winter.

The results in Figure A.10e show a similar trend as those in Figure A.10b, i.e., that PV power forecasts tend to have higher accuracy during a period with lower variability (winter). In addition, we can observe that PICP reaches its maximum value for $M = 16$, while PINAW has a minimum at $M = 12$, similar to CRPS. During spring, we observe a slight deterioration in PICP and CRPS when M increases, which is most likely caused by the increased variability in this period. In order to reduce training and inference time on one hand, but maintain a reasonable amount of training data on the other hand, we select $M = 8$.

Figure A.10f presents the results for the net load forecasts. While in case of individual customers, the development of the net load results arguably resembled that of the PV forecast results, for the aggregated case the development of the forecast results bare more resemblance to the electricity consumption forecast results. However, it is clear that the results show substantial improvement over those in Figure A.10d. Thus, despite the observation that the development of the results are closely related to those of the electricity consumption forecasts, there is a clear and positive effect of the inclusion of PV power, most likely due to the smoothing effect. As for optimal training window width M , the sudden jump in PINAW for $M > 11$ indicates that incorporating too many days is counterproductive. Therefore, we take $M = 9$ during spring. In contrast, the steady decline in PINAW during winter shows that increasing M has a positive effect on the sharpness, but causes the PIs to be overconfident for larger values of M , as expressed by a deteriorating reliability, i.e., PICP. Consequently, we select $M = 8$ during winter.

Figure 2: Boxplot representation of the the normalized variances in the data set during different seasons for (a) individual customers and (b) aggregated customers. Here, electricity consumption is abbreviated as EC, net load as ND and photovoltaic as PV. Furthermore, (c) and (d) present example time series for three levels of aggregation of 5 samples from the non-normalized data set during winter and spring, respectively.

Figure 3: In (a), a histogram presenting the penetration levels on a logarithmic scale of the original PV systems using two definitions as defined in Section 2.1.3. In (b), a visual aid presenting the effect PV system sizes on the normalized net load power profile for an individual and aggregated set of customers, as explained in Section 2.1.3.

Figure 4: Results in terms of PINAW, PICP and CRPS as a function of aggregation levels for individual and aggregated customers. In this figure, (a), (b), (c), (g), (h) and (i) are the result of the GP, while (d), (e), (f), (j), (k) and (l) are the result of QR. In addition, (a) - (f) represent the fine resolution with a 1 customer increment, whereas (g) - (l) represent the coarse resolution with a 30 customer increment. The columns from left to right concern electricity consumption, PV power production and net load. The light lines represent the scores of the individual customers, whereas the bold lines represent the averaged scores.

Figure 5: Reliability diagrams for a single customer and aggregations of 30 and 210 customers for both the GP and QR during winter and spring. In addition, the diagrams feature the consistency bars that have been moved horizontally in order to improve readability. The reliability diagrams shown here are those for the first individual and aggregated customers of the random set of aggregations.

Figure 6: Probabilistic forecasts of the GP and QR during a day in winter and spring, where the PIs with nominal coverage level of 80% are depicted using the shaded colors. The effect that aggregation has on the PIs during the two seasons becomes clear, in particular in terms of the width and especially for PV power, as was shown in Figures 4a to 4l.

Figure 7: Overview of the results discussed in Section 3.3 where the color indicates whether the result concerns electricity consumption (EC), PV power production (PV) or net load (ND).

Figure 8: The effect of increasing PV system size and its effect on the PIs of net load forecasts. The left column shows this effect on an individual level, the middle column on an aggregation level of 30 customers and the right column on an aggregation level of 210 customers. The top row shows the results of the GP, whereas the bottom row shows the results of the QR model.

Figure 9: Reliability diagrams for three levels of aggregation, i.e., a single customer, 30 customers and 210 customers for three levels of penetration levels of PV power in the net load: 110%, 130% and 200%. Note the difference between the reliability diagrams here and Figures 5c and 5f, which is due to the selection of different random samples of customers.

Figure A.10: Results in terms of PINAW, PICP and CRPS scores as a function of training window width M for individual and aggregated electricity consumption (a) and (d), PV power production (b) and (e), and net load (c) and (f), respectively. The light lines represent the scores of the individual customers, whereas the bold lines represent the averaged scores.