Large-scale wind generation simulations: Estimating missing technical parameters using Random Forest

Matti Koivisto, Konstantinos Plakas, Poul Sørensen
Department of Wind Energy
Technical University of Denmark
Roskilde, Denmark
mkoi@dtu.dk

***Abstract*—The rapid development of renewable energy sources drives the European power systems toward a green transition. Growing shares of wind power create the need to model the variability in wind generation. For modelling using the reanalysis approach, meteorological data along with the technical parameters of wind power plants (WPPs) are needed. This paper focuses on the technical parameters. Specifically, missing hub height and turbine type data are estimated using the random forest (RF) algorithm. Consequently, a complete onshore WPP dataset with approximately 16000 recordings in Europe is achieved. For the validation of the developed model, wind generation time series comparison for European countries is carried out. The results indicate that especially for countries with a lot of missing technical WPP data, RF shows significant improvements compared to a baseline imputation model. The applicability of the methodology for modelling future scenarios with changing WPP installations is also shown.**

Keywords: Large-scale, random forest, reanalysis, simulation, variability, wind.

I. INTRODUCTION

The increasing amount of installed wind power around the globe creates the need for accurate modelling of wind generation variability. The weather dependent variability in wind power is often modelled using meteorological reanalysis data. This paper shows that in addition to the meteorological data, it is important to consider the technical wind power plant (WPP) parameters in detail to reach more accurate simulations.

The dataset used in this study covers the whole European region, focusing on onshore wind. It consists of approximately 16000 WPPs. However, WPPs with missing values in hub height and turbine type account for a significant share of total installed capacity. Therefore, the objective of this study is to develop a strategy for estimating these missing values. Their impacts on wind generation time series simulation is then assessed.

Previous work shows the importance of accurate technical WPP information in wind time series modelling. Affecting power curves and wind speeds the turbines experience, turbine type and hub height are crucial parameters in large-scale simulation. In, [1] a study comparing the impact of uncertainties both from technical and meteorological side is presented. In [2], regression modelling was used for estimating the missing hub heights.

Machine learning algorithms have been used widely in power forecasting [3], [4], fault detection [5] and condition monitoring of turbines [6]. This paper applies the random forest (RF) algorithm for estimating missing technical WPP parameters. The resulting parameters from the RF model area presented and compared to a baseline model. The baseline model includes simple imputation methods; mean value for the missing hub heights and most frequent type for the missing turbines types.

To compare the RF and baseline model results in large-scale wind generation modelling, time series simulations are compared for European countries. CorRES [7] simulations are carried out using the same underlying meteorological data with both the RF and the baseline WPP datasets, putting the focus to the impacts of different technical WPP parameters. The results indicate that especially for countries with a lot of missing technical data, RF shows significant improvements compared to the baseline model.

Finally, the applicability of the presented methodology for modelling future scenarios is presented. Utilising the modelling validated on measurements, both hub heights and turbine types are changed to assess wind generation in different countries for high hub height and low specific power installations.

II. ESTIMATING THE MISSING TECHNICAL PARAMETERS

This section describes the RF algorithm used in estimating the missing technical WPP parameters. In the case of hub height, a numerical value, a regression model is estimated. For turbine type, a categorical value, a classification problem is solved. The section also describes the data that were used and shows the most important predictors in the different modelling stages.

A. Data

The WPP dataset was obtained from [8] (as it was available in November 2015), including onshore installations on the whole European region. It consists of approximately 16000 recordings and covers 120 GW of installed capacity. In addition, a turbine dataset was acquired [9]. The turbine data gives power curve and other technical information for most of the turbine types in the WPP dataset.

The authors acknowledge support from the NSON-DK (Danish Energy Agency, EUDP, grant 64018-0032; previously ForskEL), OffshoreWake (Danish Energy Agency, EUDP, grant 64017-0017; previously ForskEL) and PSfuture (La Cour Fellowship, DTU Wind Energy) projects.

The explanatory variables (predictors) used in estimating hub height and turbine type consist of numerical and categorical variables:

- Country in which the farm is located; categorical
- Rated power of the turbine; numerical
- Commissioning year (ComYear) of the power plant; categorical
- Global Wind Atlas (GWA) mean wind speed (GWAValue) [10]; numerical
- Rotor Diameter of the turbine; numerical

When turbine type and power curve are known, cut in and cut out wind speeds and area under the power curve can also be used as predictors [11]. The availability of the different predictors depend on the modelling stage as explained in Section II C. An overview of the different predictors is given in [11].

B. The Random Forest algorithm

RF is an ensemble learning method which can be used both for regression and classification [12]. The method relies on having a large number of models, or trees, that should be as uncorrelated as possible. To achieve uncorrelated trees, RF uses bootstrap aggregation (bagging) and random subspace (feature randomness) methods. The steps of the algorithm are briefly:

1. A bootstrap sample from the original dataset is created. RF allows each tree to sample from the dataset randomly with replacement (so certain recording can be selected twice)
2. A tree is grown using the bootstrap sample. At each node, the optimal split using randomly selected predictors is obtained. This adds additional variation among the trees
3. The steps 1 and 2 are repeated B times (where B is the number of grown trees)

When forecasting with the RF model, the B created trees are combined to form the prediction. In case of a regression problem, an average over the individual predictors is taken. In classification, majority voting is used. The RF algorithm implemented in Matlab was used [13]. More information on how RF is applied in estimating the missing technical parameters is given in [11].

Comparison between linear regression and RF was carried out in [11], and RF gave better results in every test. RF modelling is used in all missing parameter estimation in this paper.

C. The stages in missing parameter estimation

The WPP dataset is divided into four stages depending on the availability of the variables hub height and turbine type, as shown in Figure 1. Different models are implemented at each stage since different list of predictors is available. The stages are:

- Stage 1: Both hub height and turbine type available.
- Stage 2: Turbine type available but hub height missing
- Stage 3: Both hub height and turbine type missing
- Stage 4: Hub height available but turbine type unknown

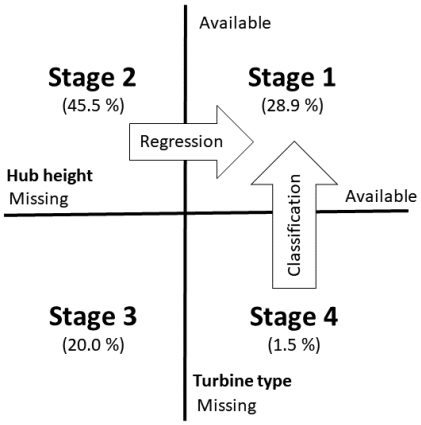

Figure 1. The four stages in estimating the missing technical WPP parameters. For stage 3 data, both regression and classification modelling are required. The percentage values show the shares of installed capacity in each stage; the shares do not sum up to 100 % as some WPPs are modelled using a different approach [11].

The estimation of hub height, a regression problem, is tackled at stages 2 and 3. For the estimation of the turbine type, a multi-class (353 different types) classification problem is solved at stages 3 and 4. RF is used in all stages, resulting in two RF models estimated for stage 3: one for estimating missing hub heights (regression) and one for estimating missing turbine types (classification).

For some recordings which do not belong in any of these aforementioned stages, the missing technical WPP values are taken from a look up table on country level [11].

D. Most important predictors

At stage 1, information on both hub height and turbine type are available; thus, no modelling is required. However, the recordings of this stage are used for the training procedure of the models for the other stages.

Figure 2 shows the predictor importance plot for the stage 2 hub height estimation, where rotor diameter is found to be the most significant predictor. This seems logical, as rotor size and hub height both relate to the physical size of the installation. GWA mean wind speed and country also affect, suggesting that locations with different wind resources tend to have different hub heights and installations in different countries differ in terms of hub height.

The predictor importance plot for the classification problem in stage 3 is shown in Figure 3. It can be seen that rated power of the turbine is the most important explanatory variable. Also the GWA mean wind speed of the installation location affects the estimated turbine type. Most important predictors in the other RF models are presented in [11].

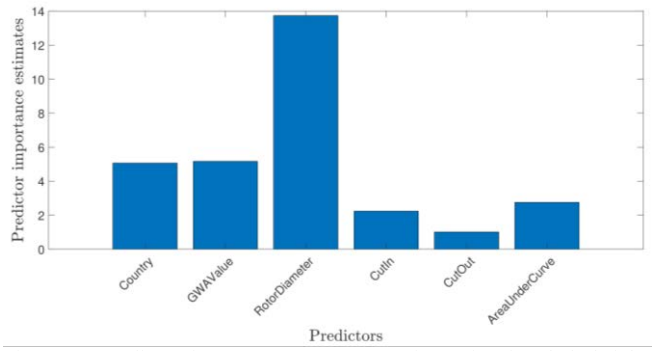

Figure 2. Predictor importance at Stage 2 for estimating hub height (regression); figure from [11].

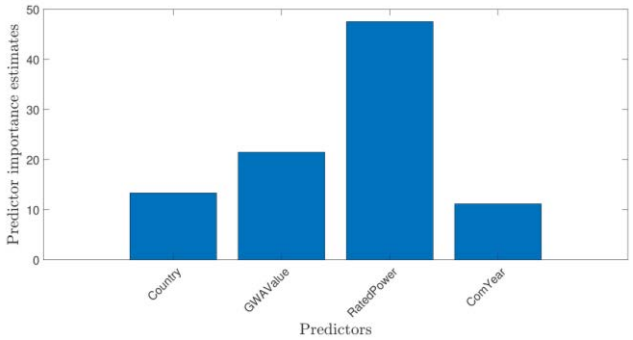

Figure 3. Predictor importance at Stage 3 for estimating turbine type (classification); figure from [11].

III. Resulting Technical Parameters

This section describes the resulting technical WPP parameters when all missing values are estimated. Two models are compared: RF and the baseline. RF modelling is as described in the previous section, whereas the baseline model uses simple imputation techniques.

A. The baseline model

For the baseline model, mean value of the known hub heights is used for the missing hub heights in each country. For turbine type, the most frequent type in each country is assigned to the missing types in that country.

B. Resulting hub heights and specific power values

In Figure 4, effective hub height for Finland, Sweden and Estonia is presented for the two models. Effective hub height refers to weighted average hub height based on installed capacity. RF model in Finland and Sweden shows higher values for the hub heights, while in Estonia the weighted hub height average gets similar values in both models.

In Figure 7, the effective specific power is shown for Finland, Sweden and Estonia. Specific power, the ratio of rated power to the swept area, is a way to quantify numerically the different turbine types, as turbines with lower specific power tend to deliver higher capacity factors (CFs). For Finland, the RF model shows significantly lower effective specific power than the baseline model.

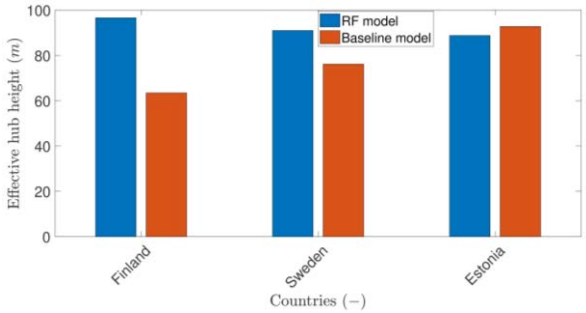

Figure 4. Effective hub height for different countries (blue: RF model, orange: baseline model); figure from [11].

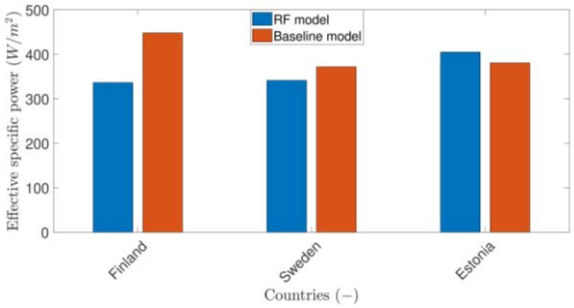

Figure 5. Effective specific power for different countries (blue: RF model, orange: baseline model); figure from [11].

C. Added value of Random Forest

The added value of RF compared to the baseline model can be seen in countries where the missing values greatly influence the effective hub height and the most frequent turbine type. In Figure 6, the difference is visible for hub heights in the cases of Finland and France. In both countries, the missing values are spotted for commissioned WPPs from 2010 onwards, when hub heights are expected to increase significantly. Baseline model, only considering mean of the know hub heights, cannot model the increasing hub heights in Finland. RF, on the other hand, models the tendency of hub heights to increase. In France, some information is available also for the recent years, and therefore the difference between the two models is smaller than in Finland.

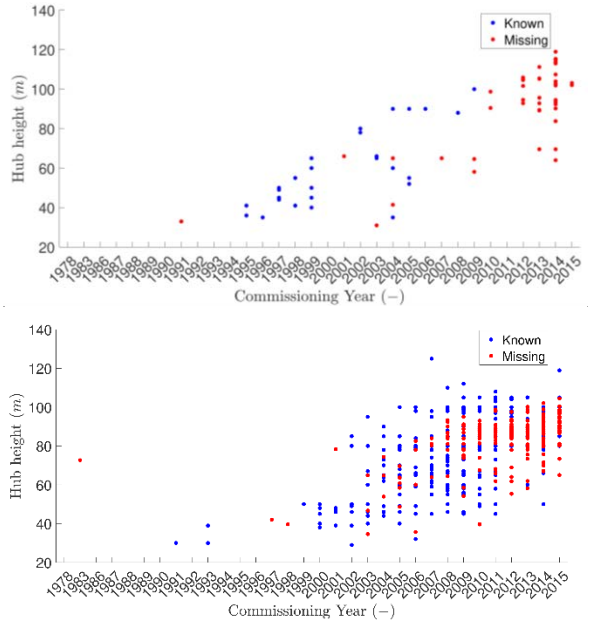

Figure 6. WPP hub heights in Finland (top) and France (bottom) for different commissioning years. The missing values are plotted with the hub height estimated using RF; figure from [11].

IV. LARGE-SCALE SIMULATIONS

This section describes the CorRES tool for simulating large-scale wind generation time series. Simulations with WPP datasets based either on RF or the baseline model are then compared. Finally, modelling of changing WPP installations is demonstrated.

A. *The CorRES model*

The CorRES tool [7] is used to generate the wind generation time series. CorRES combines reanalysis data to the technical parameters of each WPP, estimated either using RF or the baseline model. The combination of microscale and mesoscale wind speeds developed in [14] is utilized when applying the meteorological data. Same meteorological data is used both with the RF and the baseline model. Therefore, the only difference between the two models is the different way the missing technical WPP parameters are estimated.

All CorRES modelling is carried out using the meteorological and the technical WPP data directly; i.e., measured generation data or CFs are not used in calibrating the model in any way.

B. *Capacity factors*

The annual CFs for 2015 are compared for a list of European countries in Table I. Historical data of installed capacity and generation are taken from [15], [16]. For Germany and the UK, Table I includes also CorRES CFs when curtailment is considered. The ratios of curtailment are taken from [17].

Looking at Table I, it can be seen that the RF model gives significantly different CFs compared to the baseline model. In Finland and Sweden, RF clearly outperforms the baseline model. For France, the differences between the models are small. In Germany, the RF models gives CF closer to the measurements; however, the CF is still significantly higher than measured. The UK is the only country where the baseline gives CF closer to measured data compared to the RF model; however, the difference is small.

TABLE I. ONSHORE WIND CAPACITY FACTORS IN 2015

Country	RF (considering curtailment)	Baseline (considering curtailment)	IRENA	ENTSO-E
Germany	0.26 (0.25)	0.29 (0.28)	0.20	0.22
Italy	0.20	0.23	0.19	0.19
Denmark	0.29	0.29	0.29	
France	0.28	0.28	0.25	0.25
UK	0.33 (0.31)	0.32 (0.30)	0.30	
Sweden	0.28	0.23	0.34	0.33
Finland	0.31	0.18	0.32	0.32
Estonia	0.28	0.29	0.28	0.26

The values in brackets show resulting CFs if curtailment is considered (Germany and the UK only). As ENTSO-E data does not differentiate between onshore and offshore, values for Denmark and the UK, with significant offshore wind share, are not given. German ENTSO-E numbers are given as the share of offshore in 2015 was modest.

C. *Generation probability distributions*

Historical hourly wind generation time series are taken from [18], utilising mostly data from ENTSO-E. For Germany, the given standardized generation (profile) data is used directly. For the other countries, the generation data is divided by installed capacity data from [15] (considering onshore wind only) or [16] depending on which source gives annual energy generation closer to [18].

Probability density function (PDF) estimates of generation are shown for example countries in Figure 7, Figure 8 and Figure 9. For Sweden and Finland, the RF models shows PDFs closer to measured data. For Denmark, the PDF of RF and the baseline model are similar.

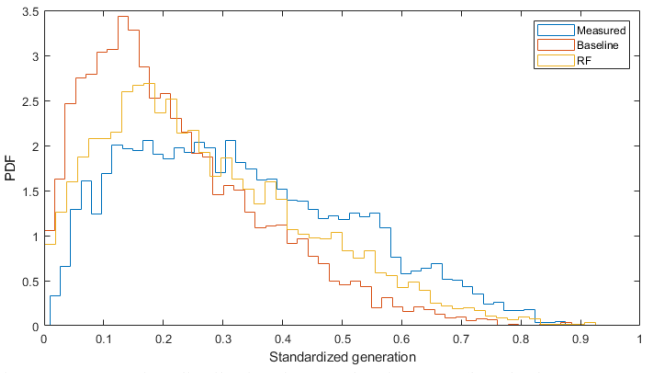

Figure 7. Generation distribution for Sweden for 2015 (hourly data). Mean values are: Measured 0.33, baseline 0.22 and RF 0.27 (means for the models are slightly different that the CFs in Table I because the measured time series had some missing data and these time steps were removed also from the simulated time series).

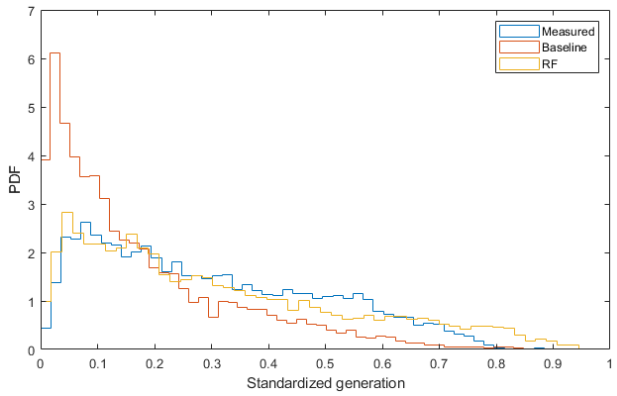

Figure 8. Generation distribution for Finland for 2015 (hourly data). Mean values are: Measured 0.30, baseline 0.18 and RF 0.31.

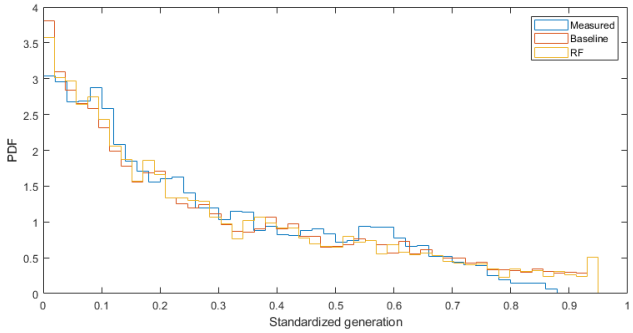

Figure 9. Generation distribution for Denmark for 2015 (hourly data). Mean values are: Measured 0.27, baseline 0.29 and RF 0.28.

D. Spatiotemporal correlations

The previous subsections showed that the RF and baseline model differ significantly in terms of CF and generation PDF in some countries. For spatiotemporal dependencies, only modest differences were observed between the models. Figure 10 shows autocorrelation functions (ACFs) for Finland for the measured data and the models. Even though Finland shows significant differences in CF between the two models in Table I, the ACFs are similar; however, RF is closer to the measured data.

Spatial correlations between selected countries are shown in Tables II, III and IV for the measured data and the models. Some correlations are closer to the measurements in the RF and some in the baseline model; however, on average the spatial correlations in the two models are similar.

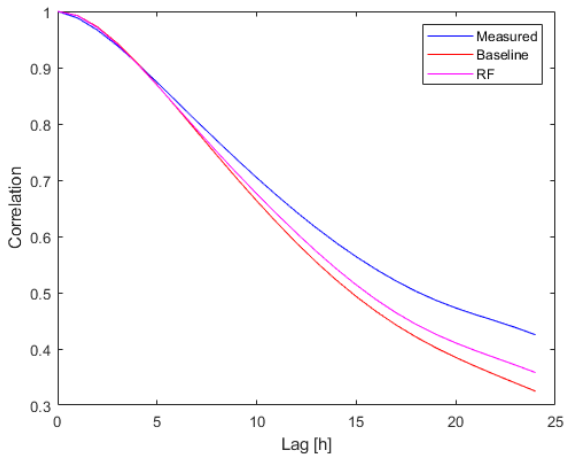

Figure 10. ACFs for the measured data and the RF and baseline models for Finland for 2015.

TABLE II. SPATIAL CORRELATIONS: MEASURED DATA

	DE	DK	SE	FI	EE	FR
Germany (DE)		0.64	0.34	0.15	0.25	0.59
Denmark (DK)	0.64		0.60	0.19	0.32	0.25
Sweden (SE)	0.34	0.60		0.59	0.57	0.23
Finland (FI)	0.15	0.19	0.59		0.59	0.22
Estonia (EE)	0.25	0.32	0.57	0.59		0.20
France (FR)	0.59	0.25	0.23	0.22	0.20	

Coloring shows high correlations in red and low correlations in green.

TABLE III. SPATIAL CORRELATIONS: RF MODEL

	DE	DK	SE	FI	EE	FR
Germany (DE)		0.64	0.33	0.13	0.22	0.57
Denmark (DK)	0.64		0.56	0.13	0.25	0.23
Sweden (SE)	0.33	0.56		0.54	0.52	0.21
Finland (FI)	0.13	0.13	0.54		0.58	0.16
Estonia (EE)	0.22	0.25	0.52	0.58		0.19
France (FR)	0.57	0.23	0.21	0.16	0.19	

Coloring shows high correlations in red and low correlations in green.

TABLE IV. SPATIAL CORRELATIONS: BASELINE MODEL

	DE	DK	SE	FI	EE	FR
Germany (DE)		0.64	0.34	0.11	0.22	0.57
Denmark (DK)	0.64		0.58	0.13	0.24	0.23
Sweden (SE)	0.34	0.58		0.54	0.53	0.23
Finland (FI)	0.11	0.13	0.54		0.60	0.19
Estonia (EE)	0.22	0.24	0.53	0.60		0.19
France (FR)	0.57	0.23	0.23	0.19	0.19	

Coloring shows high correlations in red and low correlations in green.

E. Modelling changing installations

The capability of CorRES to model scenarios with changing WPP installations is demonstrated with two example scenarios. For selected countries, all onshore WPP installations are replaced by different hub height and specific power installations; i.e., repowering with a single turbine type and hub height is modelled. In scenario A, all installations are assumed to be replaced by 100 m hub height V90/2000 turbines. This scenario thus considers relatively high hub height installations with a widely used turbine type of 313 W/m^2 specific power. Scenario B considers 120 m hub height V126/3000 turbines, thus modelling high hub height installations with a low specific power of 240 W/m^2. It needs to be noted that the suitability of the turbines for

specific locations is not considered in these example simulations. Power curves are taken from [9].

The resulting CFs in the scenarios are shown in Table V. For comparison, also the 2015 WPP installation scenario was simulated using the same meteorological data. For Finland and Sweden, with high hub height installations already in the 2015 fleet, the change in CF to scenario A is modest. Denmark and the UK see CFs around 0.4 when the scenario A installations are modelled; a significant increase from the CFs with 2015 installations. When considering scenario B, with low specific power and high hub height installations, Sweden, Finland, Estonia and Germany all reach CFs of around 0.4. In scenario B, Denmark and the UK show offshore-level CFs of around 0.5; however, it needs to be noted that the suitability of the scenario B turbine type for high wind speed locations was not considered.

TABLE V. CAPACITY FACTORS FOR DIFFERENT SCENARIOS

Country	2015 installations	Scenario A: HH 100 m, SP 313 W/m^2	Scenario B: HH 120 m, SP 240 W/m^2
Germany	0.25	0.31	0.42
Denmark	0.24	0.38	0.49
The UK	0.29	0.40	0.51
Sweden	0.24	0.28	0.39
Finland	0.27	0.28	0.39
Estonia	0.25	0.31	0.42

The CFs are based on hourly CorRES simulations using meteorological data from 1982 to 2018. HH = hub height, SP = specific power. Note: suitability of the modelled turbine type in scenarios A and B for specific locations is not considered in the simulations.

V. DISCUSSION

The results suggest that modelling the technical WPP parameters using machine learning can improve the accuracy of large-scale wind generation simulations, especially in countries where missing data influence the effective hub height and most frequent turbine type significantly. However, even the RF model showed some inaccuracies compared to measured data.

For Sweden, although RF showed more accurate results compared to the baseline, significant deviation from the measured data was observed for CF and generation PDF. The Swedish WPP dataset has a total installed capacity of 3.5 GW by the end of 2014, although historical data shows 4.8 GW [15]. This difference can affect the CorRES simulations significantly, as many WPPs are not included in the simulations.

Even though curtailment is considered for Germany, the RF model shows too high CF compared to the measured data. Onshore wind CF of 0.227 is reported in [19] for 2015 for Germany. This is slightly closer to the simulated CF of 0.25 (considering curtailment) than the IRENA (0.20) or ENTSO-E (0.22) numbers in Table I. However, the simulated CF is too high, and the reason for this will be studied in future work.

Although CFs are quite well modelled for Denmark and Finland, more accurate modelling of the tails of the generation PDFs in Figure 8 and Figure 9 will be focused on in future work.

A fixed availability of 95 % is considered in all simulations. However, availability can be expected to depend, e.g., on commissioning year. Thus, more detailed availability modelling will be considered in future work. Wake modelling will also be considered.

VI. CONCLUSIONS

This paper has shown the importance of accurate modelling of the technical WPP parameters in large-scale wind generation simulations. Applying the RF algorithm, missing hub height and turbine type data were estimated for WPPs covering the whole of Europe. As a result, a complete dataset of approximately 16000 WPPs is available for pan-European wind generation simulations.

When comparing the historical CF and generation time series data, the RF model outperformed a baseline model in almost every analysed country. CFs and generation PDFs were more accurately modelled when RF was used in estimating the missing technical parameters. The added value of the RF algorithm was shown especially for countries with missing values on recently commissioned WPPs. For these countries, the baseline model was not able to capture the development of the industry (higher hub heights and lower specific power). Spatiotemporal dependencies were also compared for the RF and baseline model; however, no significant differences were observed between the models.

Finally, the applicability of the CorRES tool to simulate future scenarios with changing WPP installations was demonstrated. Using the modelling validated on measurements, both hub heights and turbine types were changed to assess wind generation with high hub height and low specific power installations. The studied scenarios showed significantly higher CFs compared to the installations effective in 2015.

ACKNOWLEDGMENT

The authors acknowledge support from the NSON-DK (Danish Energy Agency, EUDP, grant 64018-0032; previously ForskEL), OffshoreWake (Danish Energy Agency, EUDP, grant 64017-0017; previously ForskEL) and PSfuture (La Cour Fellowship, DTU Wind Energy) projects.

REFERENCES

[1] F. Monforti, I. Gonzalez-Aparicio, "Comparing the impact of uncertainties on technical and meteorological parameters in wind power time series modelling in the European Union", *Applied energy*, vol. 206, pp. 439-450, November 2017.

[2] M. Koivisto, P. Maule, P. Sørensen, L. Galdikas, N. Cutululis, S Biondi. "Large-scale wind generation simulations: From the analysis of current installations to modelling the future", *Journal of Physics: Conference Series*, vol. 1102, No. 1, p. 012034, October 2018.

[3] A. Dolara, A. Gandelli, F. Grimaccia, S. Leva, M. Mussetta, "Weather-based machine learning technique for Day-Ahead wind power forecasting", *IEEE 6th international conference on renewable energy research and applications*, pp. 206-209, San Diego, USA, November 2017.

[4] G. Li, J Shi. "On comparing three artificial neural networks for wind speed forecasting", *Applied Energy*, vol 87, pp 2313-2320, July 2010.

[5] A. Bakri, E. Koumir, M. Boumhidi. "Extreme learning machine for fault detection and isolation in wind turbine", *International Conference on Electrical and Information Technologies*, pp. 174-179. Tangiers, Marocco, May 2016.

[6] M. Zekveld, G.P. Hancke "Vibration Condition Monitoring Using Machine Learning", *44th Annual Conference of the IEEE Industrial Electronics Society*, pp.4742-4747, Washigton, USA, October 2018.

[7] M. Koivisto, K. Das, F. Guo, P. Sørensen, E. Nuño, N. Cutululis, P. Maule, "Using time series simulation tool for assessing the effects of variable renewable energy generation on power and energy systems", *WIREs Energy and Environment*, vol. 8, no. 3, e329, May/June 2019.

[8] The Wind Power, Onshore wind farm database: Available at: https://www.thewindpower.net/ (accessed on 15 Nov 2015).

[9] The Wind Power, Turbine and power curve database: Available at: https://www.thewindpower.net/ (accessed on 16 Jan 2018).

[10] Global Wind Atlas, DTU Wind Energy in partnership with the World Bank Group, utilizing data provided by Vortex: https://globalwindatlas.info (accessed 1 July 2019).

[11] K. Plakas, "Large-scale bottom-up wind generation time series simulation with missing data", Master Thesis, DTU Wind Energy, June 2019.

[12] L. Breiman, "Machine Learning", Publisher: Kluwer Academic Publishers, vol. 45, no. 1, pp 5-32, October 2001.

[13] TreeBagger class in Matlab 2019: https://www.mathworks.com/help/stats/treebagger.html

[14] E. Ellmann, "Application of Global Wind Atlas microscale data in large-scale wind generation time series simulation", Master Thesis, DTU Wind Energy, July 2019.

[15] IRENA Renewable Energy Statistics Report 2019. Available at: https://www.irena.org/publications/2019/Jul/Renewable-energy-statistics-2019 (accessed 1 July 2019).

[16] ENTSOE Yearly Statistics and Adequacy Retrospect. Available at: https://docstore.entsoe.eu/publications/statistics/yearly-statistics-and-adequacy-retrospect/Pages/default.aspx (accessed 1 July 2019).

[17] M. Joos, I. Staffell, "Short-term integration costs of variable renewable energy: Wind curtailment and balancing in Britain and Germany", *Renewable and Sustainable Energy Reviews*, vol 86, pp. 45-65, April 2018.

[18] Open Power System Data. 2019. *Data Package Time Series*. Version 2019-06-06 https://doi.org/10.25832/time_series/2019-06-05. (Primary data from various sources, for a complete list see URL) (accessed 1 July 2019).

[19] IEA Wind TCP Annual Report 2015. Available at: https://community.ieawind.org/publications/ar?page=1 (accessed 1 July 2019).