Sizing of a short term wind forecasting system

Aurore Dupré

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAX002

Thèse de doctorat

Sizing of a short term wind forecasting system

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École Doctorale de l'Institut Polytechnique de Paris (IP Paris)
Spécialité de doctorat : Météorologie, océanographie physique et physique de l'environnement

Thèse présentée et soutenue à Palaiseau, le 22 janvier 2020, par

AURORE DUPRÉ

Composition du Jury :

Mathilde Mougeot Professeur, ENSIIE	Président
Antoine Rousseau Directeur de recherche, Inria - Montpellier	Rapporteur
Freddy Bouchet Directeur de recherche, ENS - Lyon	Rapporteur
Bénédicte Jourdier Ingénieur de recherche, EDF R&D	Examinateur
Mireille Bossy Directeur de recherche, Inria - Sophia Antipolis	Examinateur
Philippe Drobinski Directeur de recherche - LMD	Directeur de thèse
Jordi Badosa Ingénieur de recherche, LMD	Co-directeur de thèse
Christian Briard Ingénieur, Zephyr ENR	Invité

Dimensionnement d'un système de prévision éolienne à court terme

Résumé – Dans un contexte de réchauffement climatique et de transition énergétique, le développement des énergies renouvelables est indispensable afin de garantir une production d'énergie qui réponde à une demande en croissance constante. Cependant, l'intermittence de ces ressources reste un frein quant à leur pénétration. Avoir accès à des prévisions court terme fiables est essentiel et c'est d'autant plus le cas pour l'éolien qui dépend d'une ressource extrêmement variable.

Les producteurs éoliens Français bénéficient d'une période de rachat obligatoire de leur production de la part d'EDF durant 15 ans. Après cela, ils doivent vendre leur production sur le marché concurrentiel. Pour ce faire ils doivent annoncer à l'avance la quantité d'énergie qu'ils injecteront sur le réseau. En cas de déséquilibre, des pénalités leurs sont imputées. Ainsi, anticiper de manière précise la quantité d'énergie produite permet de maximiser le revenu. En France, l'échéance limite pour vendre son énergie est de 30 minutes. Ainsi, dans cette thèse, plusieurs approches de réduction d'échelle, paramétriques (régression linéaire) et non paramétriques (forêts aléatoires) sont développées, calibrées et évaluées. Les échéances considérées vont donc de 30 min à 3 h. En effet, il est possibe de vendre l'énergie jusqu'à plusieurs heures en avance. Ainsi le modèle de prévision doit être performant de quelques dizaines de minutes jusqu'à quelques heures en avance.

Les méthodes de réduction d'échelle considérées sont très rarement utilisées pour des échéances inférieures à l'heure puisque les modèles numériques sont généralement exécutés toutes les 6 à 12 h. Cependant lorsqu'il s'agit de la prévision du vent, le numérique devient très vite nécessaire. En effet, contrairement à la prévision de l'énergie photovoltaïque, pour laquelle l'utilisation d'images satellites est très courante afin de suivre et d'anticiper le déplacement des nuages, la prévision de l'énergie éolienne et donc du vent se passe difficilement de modélisation. Par ailleurs, l'utilisation de mesures in-situ dans les méthodes de réduction d'échelle, afin de corriger la prévision numérique à l'initialisation, permet un gain de performance significatif. Une comparaison des performances de cette méthode hybride avec les performances des méthodes statistiques classiques pour la prévision de la vitesse du vent à la hauteur du moyeu est réalisée. Le modèle développé surpasse toutes les autres méthodes testées dans cette étude. En particulier l'amélioration par rapport à la méthode de persistance va de 1.5% à 10 min à plus de 30% à 3 h.

Afin de limiter l'accumulation d'erreurs lors du passage de la prévision du vent à la prévision de l'énergie éolienne, une analyse de l'erreur induite par différentes variables météorologiques, comme la direction du vent ou la densité de l'air, est présentée. Dans un premier temps, la prévision ferme par ferme est explorée puis la dimension spatiale est introduite. Tout d'abord, l'information de petite échelle est évaluée au moyen de fermes situées à quelques kilomètres l'une de l'autre. Ensuite l'information grande échelle est étudiée grâce à des fermes situées à environ 200km de distance. Alors que l'utilisation de données d'une ferme proche permet des améliorations dans les prévisions à 10 et 20min, ce n'est pas le cas pour les données des fermes fortement éloignées. En effet, les échéances considérees sont trop courtes pour que les données de parcs si lointains soient pertinentes.

Pour finir, la valeur économique d'un tel système de prévision court terme est explorée. Les différentes étapes du marché de l'électricité sont étudiées et les différentes sources d'incertitude et de variabilité, comme les erreurs de prévision et la volatilité des prix, sont mises en évidence et évaluées. Pour les deux fermes considérées dans cette étude, les résultats montrent que les prévisions court terme permettent une augmentation du revenu annuel entre 4 et 5%.

Sizing of a short term wind forecasting system

Abstract – In a context of global warming and energy transition, the development of renewable energies is essential in order to ensure energy production that meets a constantly growing demand. However, the intermittency of these resources remains a barrier to their penetration. Having access to accurate short term forecasts is essential and especially for wind power, which depends on an extremely variable resource.

French wind power producers benefit from a "obligation to purchase" from EDF for 15 years. After that, they have to sell their production in the competitive market. To do so, they must announce in advance the amount of energy they will inject into the grid. In case of imbalance, they are charged penalties. Thus, accurately anticipating the amount of energy produced helps to maximize the income. In France, the deadline for selling energy is 30 minutes. Thus, in this thesis, several downscaling approaches, parametric (linear regression) and non-parametric (random forests) are developed, calibrated and evaluated. The considered lead times range from 30 min to 3 h. Indeed, it is possible to sell the energy up to several hours in advance. Thus, the forecast model must be efficient from a few tens of minutes to a few hours ahead.

The downscaling methods considered are rarely used for lead times lower than 1 h since numerical models are generally run every 6 to 12 hours. However, when it comes to wind forecasting, numerical modeling becomes necessary. Indeed, unlike photovoltaic energy forecasting, for which the use of satellite images is very common to track and anticipate cloud movement, the forecast of wind energy and speed is difficult to do without modelling. Furthermore, the use of in-situ measurements in downscaling methods to correct the numerical prediction at initialization, allows a significant performance gain. A comparison of the performance of this hybrid method with the performance of traditional statistical methods for wind speed forecasting at hub height is achieved. The developed model overperforms all other methods tested in this study. In particular, the improvement compared to the persistence approach ranges from 1.5% 10 min ahead to more than 30% 3 h ahead.

In order to limit the accumulation of errors in the conversion from wind speed forecast to wind energy forecast, an analysis of the error induced by different meteorological variables, such as wind direction or air density, is presented. First, the forecast at the farm scale is explored and then the spatial dimension is introduced. First, small scale information is assessed using data from wind farms located a few kilometres apart. Then the large scale information is studied using data from wind farms located about 200 km away. While the use of data from a close farm allows improvements for the 10 and 20 min forecasts, this is not the case for data from distant wind farms. Indeed, the considered time scale is too short for data from such distant farms to be relevant.

Finally, the economic value of such a short term forecasting model is explored. The different steps of the electricity market are studied and the different sources of uncertainty and variability, such as forecast errors and price volatility, are identified and assessed. For the two wind farms considered in this study, the results show that the short term forecasts allow an increase in annual income between 4 and 5%.

Remerciements

J'ai passé ces trois années de thèse au sein du Laboratoire de Météorologie Dynamique sur le site de l'École polytechnique à Palaiseau. Si pour des raisons personnelles les débuts ont été difficiles pour moi, je n'aurai jamais pensé, en décembre 2016, que je serai finalement triste de partir. J'ai pris un réel plaisir à travailler au LMD et c'est en grande partie grâce à toutes les personnes que j'ai pu côtoyer.

Tout d'abord je souhaite remercier Freddy Bouchet et Antoine Rousseau pour avoir accepter de rapporter mes travaux de thèse. Merci à Mireille Bossy, Bénédicte Jourdier et Mathilde Mougeot pour avoir fait partie du jury. Enfin, merci à vous tous pour les suggestions et remarques soulevées lors de la soutenance.

Ensuite, je souhaite remercier Christian Briard et plus généralement Zephyr ENR (Valérian, François, Sven, Mme Rübsamen, ...) sans qui cette thèse n'aurait pas eu lieu d'être. J'ai beaucoup apprécié nos échanges. Merci pour votre disponibilité, pour votre confiance et pour toutes ces données qui m'ont été indispensables. Je réalise la chance que j'ai eu d'avoir vu tout l'aspect pratique de ce travail. Je garde un excellent souvenir de toutes nos réunions à Bonneval et je me souviendrai toute ma vie de cette réunion du 11 janvier 2019 durant laquelle nous sommes montés en haut d'une éolienne. Merci pour cette incroyable opportunité.

Pour en revenir au LMD, je souhaite bien entendu remercier mon directeur de thèse Philippe Drobinski et mon co-directeur de thèse Jordi Badosa. Merci également pour votre confiance et votre soutient. Philippe, merci pour ta pédagogie, j'en ai bien eu besoin durant ces trois ans afin d'assimiler tous ces concepts de géophysique, je partais de loin. Merci pour ta disponibilité, je sais bien que ça n'a pas été toujours facile de trouver du temps. Merci pour ton optimisme et ta positivité. Jordi, merci d'avoir été présent quand j'en ai eu besoin, merci pour toutes ces remarques, suggestions, conseils. Je me souviendrais longtemps de ces fameux modèles a, a_{bis}, b, b_{bis}, c, c_{bis}, etc ... Vraiment, merci pour tout.

Plus généralement je souhaite remercier toutes les personnes qui m'ont permis d'accomplir ce travail. Merci Riwal pour tes conseils et pour toutes ces données (je m'excuse par avance car je n'en ai pas encore terminé). Merci Peter pour toutes ces precieuses explications sur le fonctionnement du marché de l'électricité, merci également pour les données et pour le temps que tu m'as consacré. Merci à Mathilde et Aurélie qui, au travers d'un projet étudiant, m'ont donné les clés pour appréhender tous les modèles traités dans cette thèse. Un grand merci à Bastien pour tout le travail que tu as fait et qui a constitué le point de départ de cette thèse. Enfin je souhaite remercier Mireille Bossy. Tout a commencé par le stage de master dont la thèse a été un prolongement direct. Merci pour ces quatres années de travail et merci pour cette année de travail à venir.

Finalement, merci à tous mes collègues doctorants, postdoctorants ou stagiaires au LMD qui m'auront accompagné durant ces trois ans et m'auront permis de découvrir un grand nombre de cultures différentes. Merci à Thibault, Xudong, Olivier, Léo, Artemis, Felipe, Stavros, Nicolas, Bastien, Eivind, Ayat-allah, Trung, Rémy, Antoine, Fuxing, Erik, Namendra, Soheil, Sakina, Alexis, Mathieu, Assia, Miléna, Evangelos, Douglas, ... J'ai une pensée toute particulière pour mes camarades de bureau Fuxing, Léo et Artemis. Un immense merci pour ces années partagées, vous me manquerez. Enfin, un clin d'oeil à mes collègues de la cellule com' du LMD et merci à Mathieu d'avoir repris le flambeau.

Et parce qu'il n'y a pas que le travail dans la vie, j'adresse un immense merci à ma famille. À mes parents tout d'abord sans qui je ne serai pas là. Merci de m'avoir toujours poussé et soutenue dans mes choix, même quand vous n'étiez pas d'accord (je le dis une bonne fois pour toute, maman tu avais raison, j'aurai dû faire S). Merci d'avoir fait en sorte que je puisse faire ces 8 années d'étude sereinement sans avoir à me soucier d'autre chose que de travailler. Merci pour vos concessions, pour votre aide dès que j'en avais/ai besoin, pour votre soutient. Merci d'avoir été là chaque jour pendant ces 3 ans. Je ne pourrais jamais vous remercier pour tout ce que vous avez fait. Ce travail vous est dédié.

Un grand merci également à ma soeur Audrey. Boulinette, à chaque fois que je rentrai à Cohartille, je savais que ça allait être un week end où j'allais pouvoir complètement me déconnecter et changer d'air. Tu sais me faire rire et me faire penser à autre chose comme personne et ça fait maintenant 21 ans que ça dure. Merci pour la personne que tu es devenue, je suis très fière de toi et je te souhaite toute la réussite que tu mérites dans tes études.

Plus généralement, je souhaite remercier toutes les personnes qui, d'une manière ou d'une autre, m'ont soutenu durant ces trois années. Merci à Amélie, Flore, Léa, Milica, Pauline, Lorine, Marco, Laurence, Michel, Nicole, ...

Et puisque comme le veut l'adage, j'ai gardé le meilleur pour la fin, je remercie pour terminer Alexis Gobé. Doubler la taille de ces remerciements ne serait pas suffisant si je devais énumérer toutes les choses que tu as faites pour moi. A commencer par cette thèse qui ne serait définitivement pas de la même qualité sans toi. Merci pour l'aide incommensurable que tu m'as apporté, tu es devenu mon Stack Overflow personnel. Merci pour ton indéfectible soutient durant ces 3 ans, attendre le vendredi soir pour te retrouver n'aura pas été tous les jours faciles. Merci pour ces heures passées à m'écouter me plaindre, à m'écouter répeter ma soutenance, merci pour m'avoir remotivé dans les moments difficiles. Merci d'être à mes côtés depuis maintenant 8 ans.

Contents

1 Introduction **15**
- 1.1 General context . . . 16
  - 1.1.1 Renewable energies in the global energy mix . . . 16
  - 1.1.2 The wind energy sector . . . 18
- 1.2 State of the art in short term forecasting . . . 21
- 1.3 Thesis objectives and context . . . 25
- 1.4 Thesis outline . . . 25

2 Sub-hourly forecasting of wind speed **27**
- 2.1 Introduction . . . 28
- 2.2 Methodology . . . 29
  - 2.2.1 Parametric downscaling approaches . . . 29
  - 2.2.2 Non parametric downscaling approaches . . . 31
  - 2.2.3 Benchmark methods . . . 32
- 2.3 Application at two wind farms . . . 36
  - 2.3.1 Performances for hourly forecasts . . . 37
  - 2.3.2 Performances for sub-hourly forecasts . . . 41
- 2.4 Analysis of the best model . . . 45
- 2.5 Conclusion . . . 47

3 From wind speed to wind power forecast **49**
- 3.1 Introduction . . . 50
  - 3.1.1 Direct approach versus indirect approach . . . 50
  - 3.1.2 Power curve modeling for wind turbines . . . 51
- 3.2 Consideration of wake effect on power output . . . 53
  - 3.2.1 Impact on wind power output . . . 53
  - 3.2.2 Consideration of the wake effect for wind energy modeling . . . 55
  - 3.2.3 Application to wind energy forecasts . . . 56
- 3.3 Air density induced error on wind energy estimation . . . 59
  - 3.3.1 Air density error budget . . . 61
  - 3.3.2 Application to Parc de Bonneval . . . 64
- 3.4 Impact of atmospheric conditions on power output . . . 67
  - 3.4.1 Wind shear . . . 67
  - 3.4.2 Turbulence . . . 68
  - 3.4.3 Atmospheric stability . . . 69

3.5 Performances of wind power forecast . . . 70
3.5.1 Statistical results . . . 70
3.5.2 Forecasts post treatment . . . 71
3.6 Conclusion . . . 72

4 Added value of networking wind farms 77
4.1 Introduction . . . 78
4.2 Added value of small scale information . . . 79
4.2.1 Wind farms location and specificity . . . 79
4.2.2 Improvement of the average wind speed forecast . . . 80
4.2.3 Improvements of the wind turbine downscaling . . . 83
4.3 Added value of large scale information . . . 88
4.3.1 Wind farms location and correlation . . . 88
4.3.2 Application to forecasts . . . 91
4.4 Conclusion . . . 92

5 The economic value of short term forecasting for wind energy 93
5.1 Introduction . . . 94
5.2 Simplified market simulation . . . 95
5.2.1 Electricity market . . . 95
5.2.2 Market simulation . . . 97
5.3 Impact of the short term balancing strategy . . . 101
5.3.1 Impact of the forecasting errors . . . 102
5.3.2 Impact of the price volatility . . . 104
5.4 Performance of the short term forecasting model . . . 106
5.4.1 Balacing fees . . . 106
5.4.2 Added value of short term forecasting model . . . 109
5.5 Conclusion . . . 111

6 Conclusion 115
6.1 Synthesis and main results . . . 116
6.2 Perspectives . . . 118

Appendices 121

A Stochastic Lagrangian approach for wind farm simulation 123
A.1 Introduction . . . 124
A.2 Stochastic Lagrangian Models . . . 125
A.2.1 Numerical analysis of SLM: particle approximation . . . 126
A.2.2 Empirical numerical analysis . . . 129
A.2.3 Particle in mesh method . . . 133
A.3 Wind farm simulation experiment with SDM . . . 135
A.3.1 SDM for atmospheric boundary layer simulation . . . 135
A.3.2 Numerical simulation . . . 139
A.4 Conclusion . . . 145

B List of publications 147

Bibliography **149**

1

Introduction

Contents

1.1 General context 16
1.1.1 Renewable energies in the global energy mix 16
1.1.2 The wind energy sector 18
1.2 State of the art in short term forecasting 21
1.3 Thesis objectives and context 25
1.4 Thesis outline 25

1.1 General context

Humanity is facing a growing challenge: that of energy demand. So far, the majority of our energy is produced from fossil fuels: coal, oil, gas. Early or later on, these reserves will disappear. It is, therefore, necessary to use non-fossil energy sources.

1.1.1 Renewable energies in the global energy mix

Over time, our energy consumption has continued to increase and it exploded over the last century.

In 2000, according to the IEA (International Energy Agency), the total annual electricity consumption per-capita was 2.32 MWh on average worldwide. In 2017, it was up to 3.15 MWh [1]. This value is only an average and does not reflect the wide variations between fossil energy producing countries, industrialized countries, and emerging countries.

Based on a global average consumption, the total energy consumed on earth in one year is around 6.8 billion individuals $\times$ 3.15 MWh = 23695 TWh. This is 67% more than in 2000 and more than twice as much as in 1990. In the coming years, demand will continue to increase. Figure 1.1 shows that if the trend in the coming years remains the same as that observed from 1990 to today, the energy consumption per capita will be around 4.62 MWh in 2050.

Then, if we consider the current trend and add the fact that according to the UN, the world population estimated for 2050 is around 9 billion people, a quick calculation made us assume that the total power consumed on earth will be around 9 billion $\times$ 4.62 MWh = 41580 TWh, which is almost twice the current consumption.

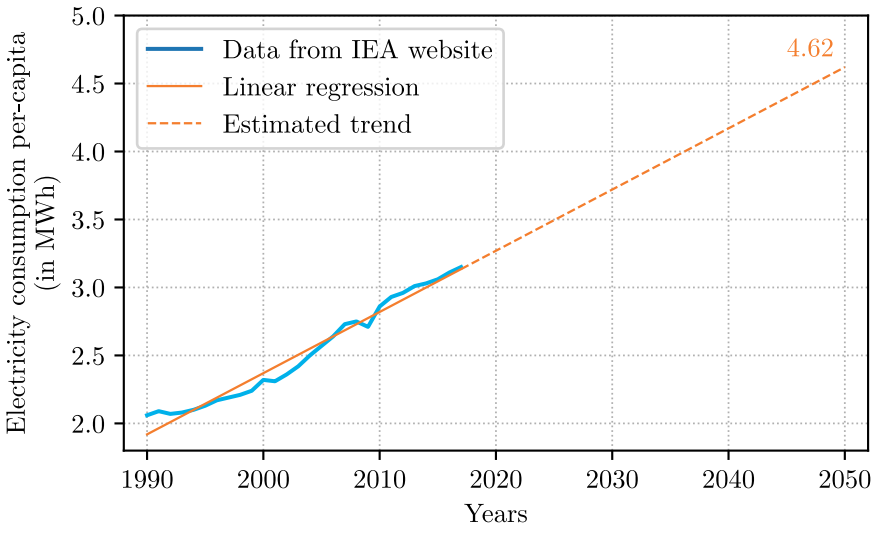


Figure 1.1 | Electricity consumption per-capita worldwide from 1990 to 2017. A linear regression is performed in order to exhibit the trend (solid orange line). Then, using the regression, the trend is estimated up to 2050 (orange dashed line) to retrieve an approximation of the electricity consumption per-capita in 2050.

Then, which resources will be used to provide this energy?

Currently, as shown in figure 1.2, more than 65% of the energy comes from fossil fuels (oil, gas, and coal).

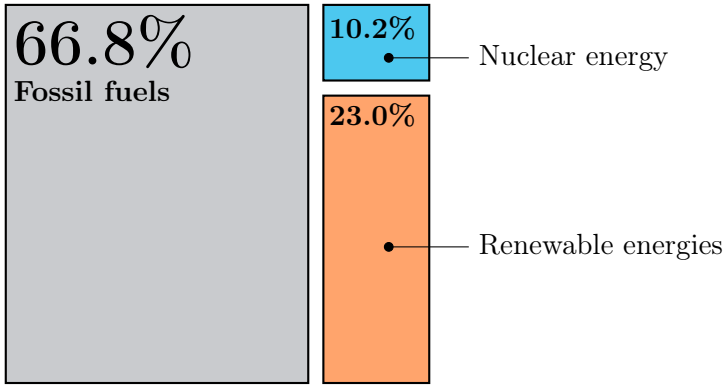


Figure 1.2 | Total world gross electricity production in 2017 [2].

These fossil fuels will remain predominant in the coming decades, particularly for emerging countries. However, in the context of global warming and energy transition, it is essential to find an alternative in our forms of energy production. Indeed in 2018, global energy-related CO_2 emissions increased by 1.7% to a historical record of 33.1 Gt CO_2. While emissions from all fossil fuels increased, the energy sector is responsible for nearly two-thirds of this growth. Figure 1.3 shows the global energy-related CO_2 emissions (in Gt) from 1990 to 2018. Last year's growth rate was the highest since 2013 and 70% higher than the average increase since 2010. This 560 Mt growth is equivalent to the total emissions from international aviation. It is therefore urgent to react if we want to remain below the 1.5°C increase of the Paris Agreement.

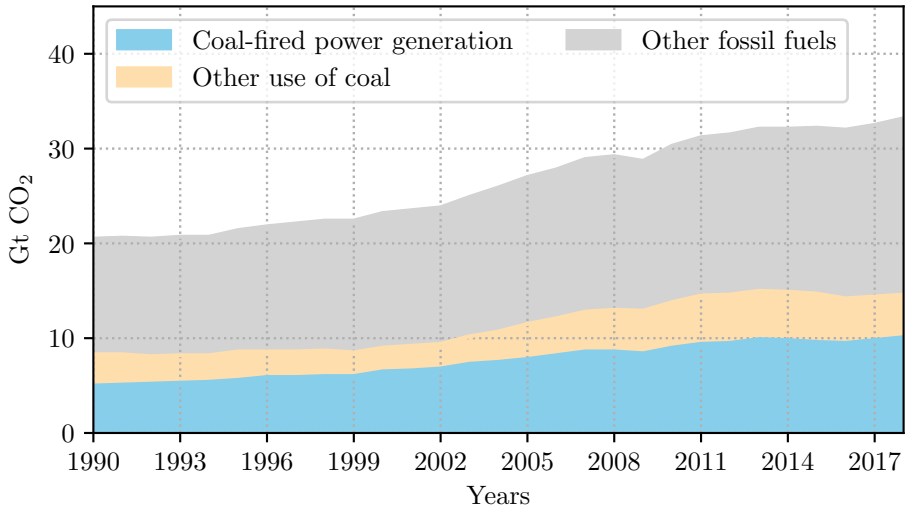


Figure 1.3 | Global energy-related CO_2 emissions (in Gt) from 1990 to 2018. Three sources are distinguished: the coal-fired power generation, the other use of coal, and the other fossil fuels. *Data extracted from IEA website.*

Then, the emergence of renewable energies is a way to reduce the impact of these fossil fuels. Indeed, as shown in figure 1.2, renewable energies accounted for nearly a quarter of global energy production in 2017 [2].

Moreover, in terms of investments, the financing of new renewable energy installations worldwide amounted to $271.8 billion in 2017, with China, Europe, and the United States accounting for nearly 75% of global renewable energy investments. According to [3], the world has invested more than $3000 billion in renewable energy since 2004.

In France, the share of fossil fuels is well below the world average, around 7.2% in 2018. However, it is replaced by nuclear energy, which represents 71.7% of total electricity production in metropolitan France in 2018. Thus, about 21.2% of the production comes from renewable resources, which is slightly less than the share of renewable energy in the world global production [4].

However, it should be specified that the share of renewable energies in the French electricity production mix is rising sharply: it was only 16.4% in 2012. The energy production law sets the objective of increasing this share to 40% by 2030. Among these renewable resources, the first source is hydraulic. Next comes wind energy followed by solar energy, as shown in figure 1.4.

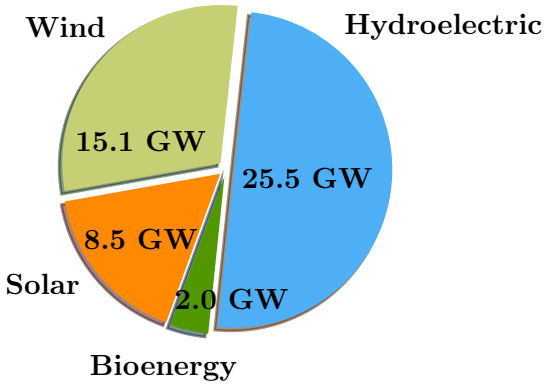


Figure 1.4 | Renewable power capacity as of December 31 2018 in France [4].

1.1.2 The wind energy sector

Apart from hydroelectric, wind energy is the world's leading renewable electricity source far ahead of solar and bioenergy.

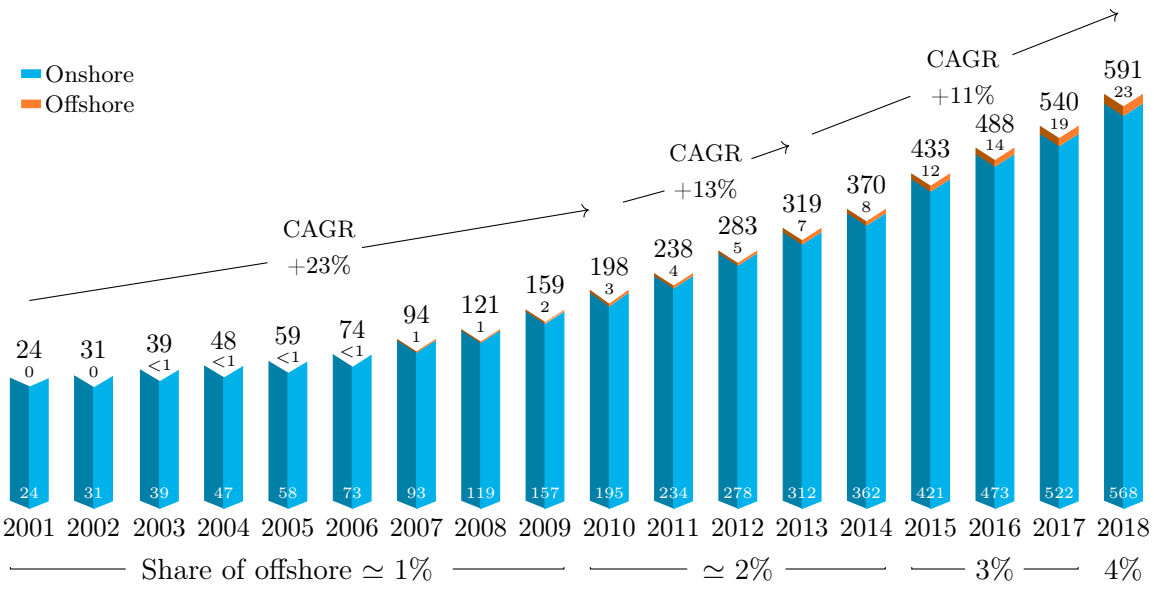
Worldwide Figure 1.5 shows that in 2018, 51.3 GW of wind electricity energy was installed. This is slightly lower than in 2017 from about 4%. Since 2014, new installations have reached 50 GW each year despite fluctuations in some markets [5]. Those new installations bring a cumulative total of installations to nearly 591 GW. For the onshore wind energy market, 46.8 GW was installed, down 4.3% compared to 2017. China and the United States remained the largest onshore markets.

The global offshore market remained stable in 2018 with 4.5 GW of new installations, as in 2017. The total cumulative installations have now reached 23 GW, which represents 4% of the total cumulative installations.

For the past twenty years, there has been a slowdown in this increase, as shown in figure 1.5. Indeed, the Compound Annual Growth Rate (CAGR) (defined in equation (1.1)) goes from +23% for the period 2001-2010 to only +11% over the last six years.

$$CAGR = \left(\frac{A_2}{A_1}\right)^{1/Ny} - 1 \tag{1.1}$$

where A_2 is the final value, A_1 is the initial one, and Ny is the number of year in the period.



Detailed data sheet available in GWEC's members only area

Figure 1.5 | Historic development of total installation in GW for the wind energy sector. *Extracted from GWEC Report 2018, p27.* [5]

Europe has been a leader in the development of wind energy. For instance, Denmark is producing via wind power, the equivalent of 43.4% of its total electricity consumption in 2017. As shown in figure 1.6, this is the only country with the share of wind power in demand higher than 40 %. Moreover, Europe is also known for its dynamism in the development of this energy. This is the second largest region in the world in terms of growth, and there have been 11.7 GW of installations (10.1 GW in the EU) of gross electricity capacity in 2018. Figure 1.6 displays the share of wind power in demand for most European countries. It also shows the new installations (in GW) as well as cumulative installed capacity (in GW) for the five largest wind energy producing countries. We can see that Germany is the first country in terms of cumulative installed capacity with 59 GW installed and with more than 20% of wind power in demand. France is the fourth with 15 GW but with less than 10% of wind power in demand.

In any case, with a total net installed capacity of 189 GW, wind energy remains the second largest form of electricity generation capacity in Europe, even exceeding gas installations by 2019. 2018 was a record year for new wind capacity financed, and 16.7 GW of future projects are under development [5].

In France As shown in figure 1.4 wind power is the second largest renewable energy source in France (after hydroelectric). However, today, electricity production in France still relies heavily on nuclear energy. In 2018, this resource accounted for 71.7% of the electricity produced. As for

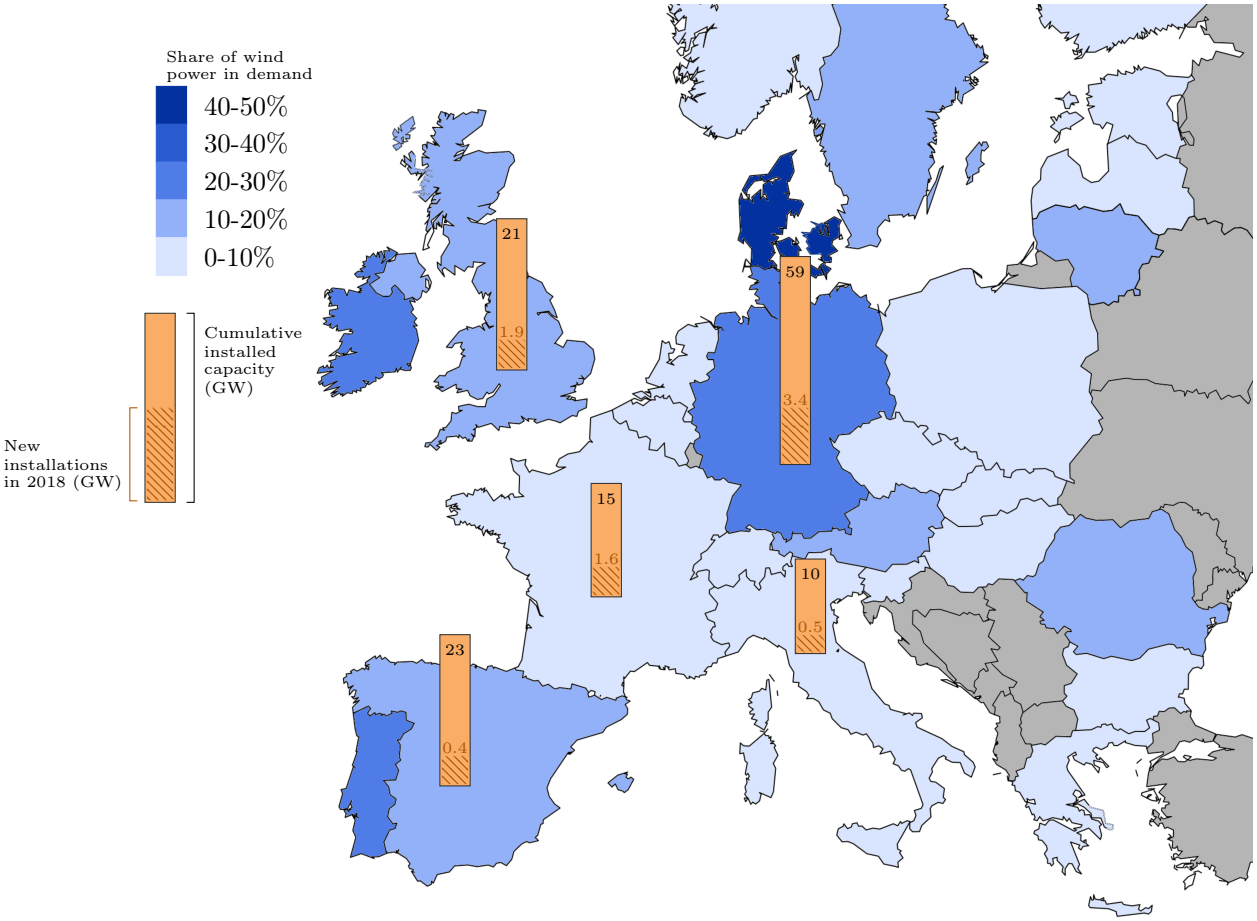


Figure 1.6 | Share of wind power in total electricity demand in 2018 in Europe. The total installed capacity as well as the new installations for 2018 are shown, for the five countries with the largest installed capacity.

fossil fuels (coal, oil or gas), there has been a real drop in production. In 2018 they accounted for 7% of electricity production in France. Faced with this decline, renewable energies are developing considerably, in particular hydroelectric, which produced 12.4% of electricity in 2018. This corresponds to an increase of 25% compared to the previous year, according to RTE (Réseau de Transport d'Électricité), the manager of the public electricity transmission network in France [4]. Wind and solar energies are not to be outdone. They now represent 5.1% and 1.9% of the mix with increases of 15.3% and 11.3%, respectively. Finally, bioenergy is gradually gaining ground. They accounted for 1.8% of production in 2018 [4].

The size and the geographical position of its territory give France the second largest wind energy potential in Europe after Great Britain. The Environment and Energy Control Agency (ADEME for Agence De l'Environnement et de la Maîtrise de l'Énergie) provides a map of the French wind farms: the regularly and strongly windy land areas are located on the western side of the country; it also gives an estimate of the French offshore wind potential: 30000 MWA [6].

Also, France has set ambitious renewable energy development targets in the Energy Transition Law for Green Growth, adopted in August 2015 with 15000 MW in 2018 (15117 MW recorded at the end of 2018) and between 21800 MW and 26000 MW in 2023. Moreover, this law sets France's

production of renewable energy at 40% by 2030. Thus wind energy will see its share in the French electricity mix increase each year.

To ensure that this development takes place in a favorable context, the French government introduced an incentive measure in 2000 and until 2015: the purchase obligation.

In the context of these contracts, EDF or local distribution companies purchase the wind electricity from operators who request it, at a feed-in tariff set by decree.

Under 2008 conditions, contracts for onshore wind power were signed for 15 years. The rate was set in 2008 at 8.2 cts€/kWh for 10 years, then between 2.8 and 8.2 cts€/kWh for 5 years depending on the sites. This tariff is updated each year.

From January 1, 2016, the support for onshore wind power has evolved towards the new remuneration system set up by the Law on Energy Transition for Green Growth. They state that the electricity produced should be sold directly by the producer on the electricity market. The difference between a reference tariff fixed by order and the average market price recorded each month is paid to the producer by EDF. The additional cost incurred by EDF is offset against a contribution called Contribution au Service Public de l'Électricité (CSPE).

In both cases, this period during which the wind producer can sell its electricity at a preferential price lasts only 15 years. At the end of this period, the producer has to sell its electricity on the competitive market. In several ways, this change can cause a significant loss of income.

On this market, electricity is sold the day before at midday for each time slot of the next day (requiring a forecast from +12 h to +35 h). A second market opens at 3 PM the day before. This market is a balancing market. Thus, having access to a more reliable forecast (because shorter term), the producer can correct his sale by buying or selling on this balancing market up to 30 minutes before the delivery date.

However, at any time, the amount of electricity fed into the grid must be equal to the quantity of electricity withdrawn. The balance between production and consumption is ensured in real time by RTE. Thus if the difference contributed to the French total deviation, it will result in a financial penalty for the producer. On the other hand, if the difference has decrease the total French deviation, the producer will receive financial compensation from RTE. However, this financial compensation is, on average, less than what the producer would have received if he could have sold this electricity on the balancing market.

Thus, having access to a reliable short term forecast to limit these gap compensations is essential for the producer. It allows to limit the loss of income due to the end of the feed-in tariffs.

1.2 State of the art in short term forecasting

Many methods are available for wind energy forecasting. They can be classified according to time scales or methodology. The time scale classification of wind energy forecasting methods is quite arbitrary, and differs according to the different descriptions found in the literature. However, in general, four categories can be identified: the very short term, the short term, the medium term, and the long term. For the classification according to methodology, the different studies agree and each of this methodology is generally associated with a specific time scale. Figure 1.7 illustrates this classification.

It shows the performance of the different wind speed forecasting methods depending on the targeted lead time. We can see that persistence approach and statistical methods are preferred for the very short term and short term forecasting while weather models can be used for longer lead times. Finally, climatology remains the best approach for studies ranging from decades to

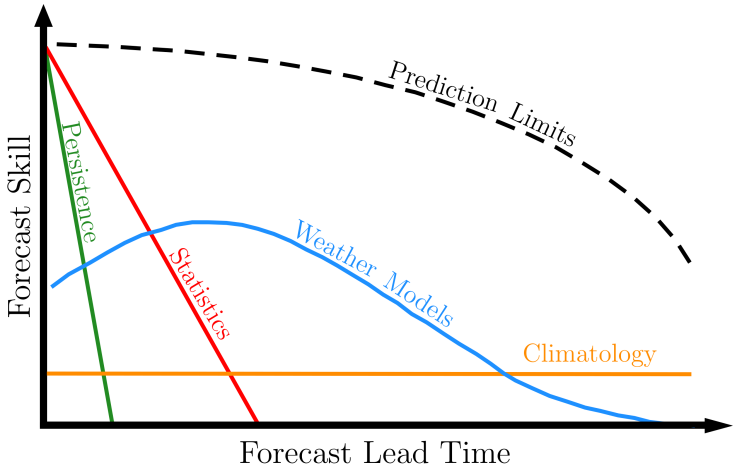


Figure 1.7 | Diagram of the performance of the different wind speed forecasting methods according to time.

centuries. Tables 1.1 sum up and describe these classifications with the approximate range and associated examples. Table 1.1a displays the time classification. For each category, it includes the associated range and the application of the corresponding forecasting methods. Table 1.1b displays the classification according to the methodology. Again, it shows few examples and the typical range for each category. Of course, each category remains indicative, and some models classified as short term, can be used for medium or long term forecasting.

Long term and medium term forecasts Most of the time, long term and medium term forecasts use the same techniques, which are physical models. Physical models are based on the mathematical equations that govern the physics law of the atmosphere. They are generally run on a global or regional space scale and provide, on a coarse grid, forecasts of several physical variables such as temperature, pressure, wind speed, or humidity for example. Usually, models are run once or twice a day due to the difficulty of obtaining information in a short period of time and the high cost involved. That is why they are preferred for long term or medium term forecast. For instance, in [7], Hong evaluates the NCAR (National Center for Atmospheric Research) Mesoscale Model (MM5) on a horizontal grid at a 5 km resolution. The model was run twice a day, and the study focuses on the Taïwan area within a period of two months. He focuses on surface variables and shows that the model tends to overestimate the surface wind speed. In [8], Taylor *et al.* use Weather Ensemble Prediction to predict the wind speed at five wind farms located across UK up to ten days ahead. Instead of producing a single forecast of the most likely weather, a whole set of forecasts is produced. This ensemble is then intended to provide an indication of the range of possible future states of the atmosphere. If statistical models are traditionally categorized for short term forecasts, they can be combined with physical methods. These hybrid methods are then used for all horizons. For instance, in [9], Salcedo-Sanz *et al.* hybridize the MM5 model mentioned above, with an Artificial Neural Network (ANN) to forecast the wind speed two days ahead at a wind farm located in the Southeast of Spain. The outputs of the NWP model are processed by the ANN.

(a) Classification based on time scale

Category	Range	Example of applications
Very short term	Few seconds to few tens of minutes	- Electricity market clearing - Real time grid operations
Short term	Few tens of minutes to few hours	- Economic load dispatch planning
Medium term	Several hours to to few days	- Generator online/offline decisions - Unit commitment decisions
Long term	Few days to one year (or more)	- Maintenance planning - Feasibility study for design of the wind farm

(b) Classification based on methodology

Methodology	Examples	Range
Persistence method	/	Very short and short term
Statistical approaches	- Artificial Neural Network (ANN) - Time series based models	Short term
Physical approach	- Numerical Weather Prediction (NWP) models	Medium and long term
Hybrid methods	- NWP + ANN - Spatial correlation + ANN - NWP + time serie based model	All ranges

Table 1.1 | Classification based on time scale (table 1.1a) and on methodology (table 1.1b) for the wind energy forecasting. In table (a), an approximate range and two examples of applications are shown for each category. In table (b), for each methodology, the associated typical range and few examples are shown.

Very short term forecasts Very short term forecasting is the least represented category in the literature. Typically these forecasts range from few seconds to several minutes and are almost exclusively provided by statistical methods such as time series based model or ANN. For such lead times, they are never combined with NWP models and only use historical in-situ data as input. In [10], Riahy *et al.* use a time series based method to predict the wind speed 1 sec to 5 sec ahead for control system of wind turbines such as changing the pitch angle of the blades. In [11], Potter *et al.* use a model based on neural networks to forecast wind vectors up to 2.5 min ahead in Tasmania, Australia.

Short term forecasts Short term forecasting is the main objective of this thesis. As shown in table 1.1a, short term forecasts usually range from tens of minutes to a few hours. Short term forecasts 1 h ahead are the most studied forecasts in the literature. In this thesis, we will focus on forecasts from 10 min to 3 h ahead.

As shown in table 1.1b, for these time scales statistical methods are the most used. They may or may not be combined with physical methods as in [9]. Statistical approaches aim at finding the relationship between past and future observations. Historical data are used as input, and in the case of hybrid methods, NWP model outputs are also added.

Statistical methods can be divided into two sub-categories: those based on linear time series that are easy to model and inexpensive to implement, and those based on artificial intelligence

methods. These can treat non-linearity but are more complicated to set up and are known to be black-box models. Most of the time, these models are tested against the persistence model, which is the benchmark approach for this time scale. Persistence assumes that wind speed or wind energy at time $t = t_0 + \Delta t$ will be the same as at time t_0.

In [12], Chang uses a back propagation neural network, which consists of feeding the network backward during the training period to tune the parameter more accurately. His goal is the wind power forecasting 10 min ahead. For the best NN, he finds a maximum absolute error of 2.0% and an average absolute error of 0.3%. In [13], Kariniotakis *et al.* develop a NN for wind power time series forecasting up to 2 h ahead. They compare 3 differents NN. Among the 3 models, the one that performs best for the first lead times is also the one that performs worst for the longer lead times and vice versa. For the third model, they find an improvement over persistence around 10%, for the whole period. In [14], Zhao *et al.* investigate a hybrid wind forecasting method. It consists of a NN fed with NWP model outputs. Their goal is the wind power forecasting 1 h ahead at a specific wind farm in China. They find that the Normalized Root Mean Square Error (NRMSE) has a monthly average value of 16.47% which they give as an acceptable value to guide the penetration of wind energy in China.

In [15], Palomares-Salas *et al.* develop a time series based model (ARIMA) used to predict the wind speed. The results are compared with the performance of a back propagation NN. They show that the ARIMA model overperforms the NN model for the short term lead times (10 min, 1 h, 2 h, and 4 h). In [16], Panteri *et al.* also compare NN and time series based model for wind power forecasting at three different wind farms for look-ahead times between 1 h and 12 h. They use a simple configuration for the time series based method (ARX), and they find that this model cannot overperform persistence for the whole period. However, NN is better than the persistence model. In [17], Firat *et al.* develop a model to forecast the hourly wind speed. Their starting point is the time series based model AR. Their implementation of this model leads to an improvement over persistence for the whole period. Moreover, the traditional AR is the model that shows the best improvement for the first lead time (1 h).

While neural network and time series based methods remain the most commonly used and the most studied models for wind speed and wind energy short term forecasting, many other methods have been investigated. For instance, in [18], Alexiadis *et al.* implements a NN based on spatial correlation to forecast the wind speed at six different sites, on islands of the South and Central Aegean Sea in Greece, distant from 52 km to 105 km. Their goal is the wind speed forecasting for the next 1 h, 2 h, and 3 h. The model is tested using data collected over seven years, and its performance are considered satisfactory. In [19], Pinto *et al.* propose a Support Vector Machines (SVM) model for short term wind speed forecasting. Its performance is evaluated and compared with ANN based approaches. SVM models are non-parametric models for classification and regression problems, such as pattern recognition or regression analysis. A case study for predicting wind speed at 5 min intervals is presented. Results show that the best parametrization for the proposed SVM achieves better forecasting results than the ANN based approaches. In [20], Lahouar *et al.* propose a random forest method to build a 1 h ahead wind power forecasting system. Like SVM, random forests are non-parametric models designed for classification and regression problems. The algorithm builds decision trees and performs learning on multiple trees trained on slightly different subsets of data. Results show an improvement of forecast accuracy using the proposed model, as well as an important reduction of the different error criteria compared to classical NN prediction. Finally, in [21], Castellanos *et al.* use a ANFIS model to forecast the hourly wind speed. The Adaptive Neuro-Fuzzy Inference Systems (ANFIS) is a hybrid model between ANN and Fuzzy Inference Systems (FIS). If the ANN part can search for patterns, which gives the

advantage of learning about systems, the FIS part is based on fuzzy logic. It corresponds to a set of fuzzy IF–THEN rules that learns capability to approximate nonlinear functions. Generally, this type of model shows more promising results than a single ANN. In [21], they obtain errors between 25.5% and 32.5% depending on the configuration.

1.3 Thesis objectives and context

The general issue raised in this work is to know if an accurate short term forecasting model can ensure the penetration of wind energy in the electricity grid and if it can compensate for the decrease in income due to the end of the feed-in tariffs period. Only few studies have been done for 30-min ahead, which is the deadline for balancing on the electricity market. In addition, the model must be efficient for these short lead times but also for longer lead times since the balancing market opens several hours before the delivery date. In this context, several questions can be raised:

1. How can the state of the art on wind energy forecasting can be improved for time horizons from few tens of minutes to few hours?
2. What is gained by including available ancillary measurements as input, such as wind direction, wind variability, or temperature?
3. Can wind energy production data from multiple wind farms improve the forecast at a given wind farm?
4. What is the economic value of short term forecasts for a wind energy producer?

This study will be carried out based on the case of a French wind energy producer. Indeed, this work is supported by a private company called Zephyr ENR[1]. The company, created in 2002, is the result of a partnership between an agricultural advisory office in France and a wind power development office in Germany. More precisely, Zephyr ENR is a wind farm development and operation office whose objective is the emergence of participatory projects in which farmers, local residents, and rural people are partners in the development and ownership of wind turbines. Since 2002, six wind farms have been built in France, including 31 wind turbines representing 77 MW and 130000 MWh/year. These wind farms are distributed in the northwest quarter of France and grouped by two, as shown in figure 1.8. Parc de Bonneval (A), Parc de la Renardière (B), and Parc de Beaumont (C) are composed of six 2 MW turbines. Moulin de Pierre (E) is composed of six 3 MW turbines. Parc de la Haute Chèvre (D) is composed of three 2.3 MW turbines and Parc de la Vènerie is composed of four 2.3 MW turbines. For each wind turbine, a substantial amount of data is recorded and transmitted at a frequency of 10 min. We can mention the wind speed, the wind direction, the production, the temperature, and the pitch angle, for example. The first farm has been in service since 2006.

Consequently, in 2021, the feed-in tariffs period will end for this farm, and the company will have to sell its energy on the competitive market.

1.4 Thesis outline

First of all, in chapter 2, a short term wind speed forecasting model is proposed. This model is a hybrid model based on the statistical parametric downscaling of NWP model outputs and

[1] http://www.zephyr-enr.fr

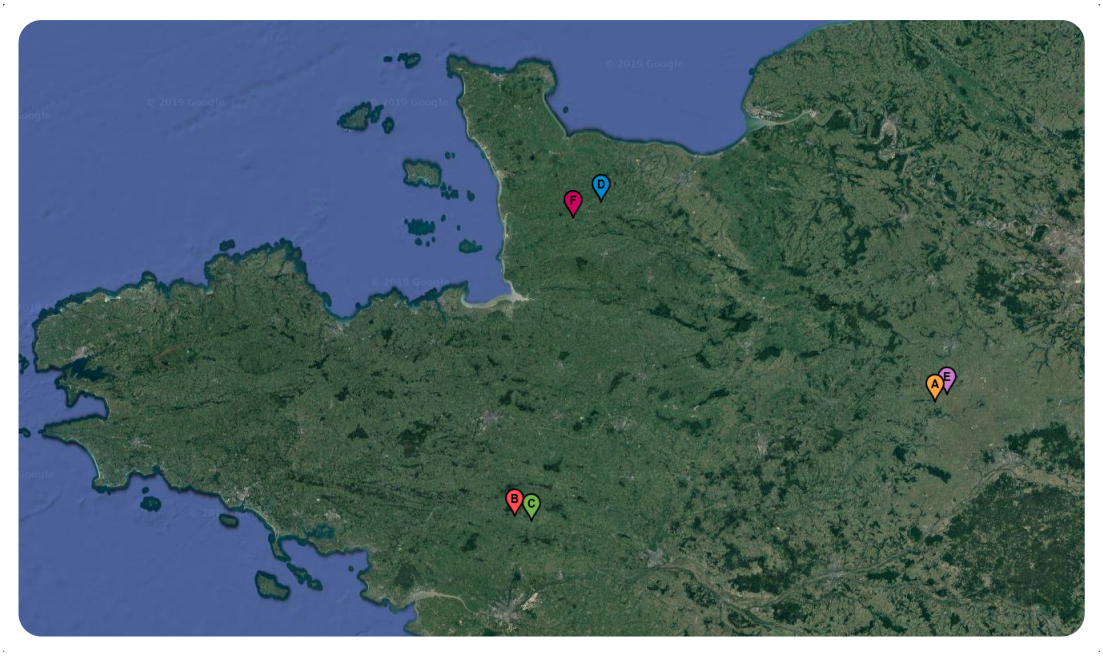


Figure 1.8 | Location of the wind farms owned by Zephyr ENR. There are Parc de Bonneval (A), Parc de la Renardière (B), Parc de Beaumont (C), Parc de la Haute Chèvre (D), Moulin de Pierre (E), and Parc de la Vènerie (F) .

on time series based model. Its performance is compared with the benchmark model, such as the persistence approach, ARMA method, and NN. A non-parametric approach is also used for comparison with a random forest method. The proposed model is evaluated for hourly forecasts from 1 h to 11 h ahead and forecasts at 10 min frequency from 10 min to 3 h ahead.

Then, in chapter 3, the conversion from wind speed forecasts to wind power forecasts is optimized. Using the data collected at the wind turbines, the influence of wind direction and air density on the power output is quantified. Methods to take into account these features are described and evaluated. A sensitivity study on the impact of atmospheric conditions such as wind shear, turbulence, or atmospheric stability is also conducted.

In chapter 4, the fact that the wind farms are grouped by two is tapped. Models based on spatial correlation are explored. The added value of small scale information is first evaluated using two adjacent farms. A few kilometers away, this configuration is adapted for the first lead times that we are interested in, typically from 10 min to 30 min. Then, the impact of large scale information is investigated using data from distant farms. This time, the configuration is more likely to perform well for the longest lead times (>1 h). In both cases, regimes based on the wind direction are distinguished.

Finally, the goal of chapter 5 is to quantify the economic value of a short term forecasting model for a producer. Simulations are performed using in-situ data and data from the actual electricity market. The case where no short term forecasts are available, and the case where a short term forecasting model is used are analyzed, and the different sources of variability are highlighted.

2

Sub-hourly forecasting of wind speed

Contents

2.1 Introduction 28
2.2 Methodology 29
2.2.1 Parametric downscaling approaches 29
2.2.2 Non parametric downscaling approaches 31
2.2.3 Benchmark methods 32
2.3 Application at two wind farms 36
2.3.1 Performances for hourly forecasts 37
2.3.2 Performances for sub-hourly forecasts 41
2.4 Analysis of the best model 45
2.5 Conclusion 47

2.1 Introduction

The intermittency of wind is the main barrier to the development of wind energy. For this resource to establish itself as one of the primary renewable resources on a sustainable basis, it is necessary to allow its penetration on a large scale. However, many difficulties must be overcome to ensure the stability of the system, such as grid operation management, maintenance scheduling, electricity market clearing, for example. Improving wind forecasting is one way to achieve higher penetration of wind power in the electricity system. In the context of climate change and energy transition, this issue is becoming a priority, and over the past two decades, the global energy market is turning increasingly to green energies.

Fortunately, Numerical Weather Prediction (NWP) models have improved significantly over the last 30 years. The forecast skill of the 3-days forecasts for the northern hemisphere rose from 85% to 98.5% between 1981 and 2013 and from 70% to 98.5% for the southern hemisphere [22]. The poor performances in the southern hemisphere was due to a lack of measurement in the region. Even though NWP models perform well for predicting large scale meteorological variables at short term, like mid-tropospheric pressure, they do not perform the same for variables having high variability at small scales, like surface winds. Large scale variables are well understood physically and efficiently modeled numerically, but the variables tied to phenomena occurring on a smaller scale depend more on processes that are not resolved and so parametrized. This leads to significant model errors for variables like surface wind.

The model error has several components: part comes from the inadequate representation of physical processes, e.g., uncertainties in the parametrizations used for boundary layer turbulence. Improving parametrizations should reduce this error. Part of the error is numerical error, coming from the discrete representation of a continuous process. Also tied to the limited resolution is the representativity error, which occurs because of the difference of the value over a grid box and the value at a specific point. Downscaling methods such as Model Output Statistics (MOS) are usually used to reduce representativity error [23]. Those models have been developed in the weather forecast for several decades based on NWP model outputs. A statistical relationship is determined between observations and forecasts using past forecasts and corresponding observations and then serves to improve predictions at that observation site.

Downscaling models can be very interesting to get accurate forecasts at a specific location of a wind farm [24]. To do so, different downscaling models and different outputs of NWP models, climate data, or, if relevant, recent surface observations can be used as explanatory variables for the near surface wind speed [25]. Amongst them, markers of large-scale systems (geopotential, pressure fields) and boundary layer stability drivers (surface temperature, boundary layer height, wind, and temperature gradient) can be used [26]. In terms of methodology, several models have already been studied, including linear regression, Support Vector Machine (SVM), or random forests.

However, for hourly and sub-hourly forecasts, downscaling methods are not commonly used because NWP models are only run once or twice a day due to the difficulty of gaining information in a short time and the associated high costs. This usually limits its usefulness to forecasts with lead times longer than 6 hours, at least. Persistence is the reference method for short term and very short term forecasts. It supposes that the wind speed at a particular future time will be the same as it is when the forecast is made. It is extremely accurate for a very short lead time, but its performance degrades rapidly with time. Statistical approaches are also used as a benchmark for short and very short term generally. We can split this category into two sub-categories, which are artificial intelligence methods such as Artificial Neural Network (ANN) using past measurements as explanatory variables and time series models such as Auto-Regressive Moving Average

(ARMA) [27]. ANN models can represent a complex non-linear relationship and extract the dependences between variables through the training process. Statistical methods are based on training with measurements and use differences between the predicted and the actual wind speed to upgrade the model. Both approaches constitute the reference methods for short term forecasts [28]. Usually, ANN models outperform time series models [29] even if some very good time series models can supersede ANN methods [30, 31].

In this chapter, we compare two configurations of downscaling models tested on several wind farms. The first configuration uses explanatory variables available from NWP models, and the second one adds explanatory variables derived from observations. In both cases, we compare the results of linear regression and random forests with persistence methods and with the benchmark methods. The chapter is organized into five sections. The next section describes the data and the different models. In section 2.3, results are analyzed. As there is little literature on sub-hourly time scales, we first (2.3.1) test our methods on hourly time scales, from 1 h to 11 h, which are much more documented. Results of persistence, ARMA, and ANN methods are also shown for comparison with the classical results found in the literature. Then, all methods are applied for sub-hourly forecasts from 10 min to 170 min at a frequency of 10 min (2.3.2). In section 2.4, the best model is analyzed in more detail. In the last section, we discuss the results and conclude.

2.2 Methodology

2.2.1 Parametric downscaling approaches

Downscaling statistical methods have been widely investigated since several decades in order to forecast the wind speed, usually from few to several hours [32, 33, 34]. In this thesis we consider linear regression, a very easy to implement method and numerically low cost [35]. The parametric approach supposes a relation between the target at time t, $\widehat{y}_t$ and the m explanatory variables at time t, $X_{1,t}, ..., X_{m,t}$:

$$\widehat{y}_t = \omega_0 + \sum_{k=1}^{m} \omega_k X_{k,t} + \varepsilon \tag{2.1}$$

where ω_i, $i \in \{0, ..., m\}$, are the model parameters to be estimated by a classical Ordinary Least Squares (OLS) method and ε is the residual following a centered normal distribution $\mathcal{N}(0, \sigma^2)$. The explanatory variables are chosen among ECMWF (European Centre for Medium-Range Weather Forecasts) outputs. It provides global forecasts, climate reanalyses, and specific datasets.In our case, we retrieve the day-ahead hourly forecasts, starting from analysis twice a day, at 00:00 UTC and 12:00 UTC. UTC is the Universal Time Coordinate. However, we would like the downscaling models to be able to start at any time and not only twice a day. In these conditions, we need to dissociate the lead times from ECMWF and the lead times from the downscaling models. For instance, the first lead time $t_0 + 1$ h for the dowscaling models may not be the first lead time for ECMWF. If a forecast from the downscaling models is launched at 05:00 UTC, the first lead time is 06:00 UTC. However, this is the sixth lead time $t_0 + 6$ h for ECMWF as the forecast started at 00:00 UTC. To be sure that this mixing of lead times does not introduce significant errors into the downscaling model, we investigate whether the ECMWF error over the first 12 hours increases significantly or not. Figure 2.1 displays the forecasted error of ECMWF in % (defined later in equation (2.8)) along with persistence (described in 2.2.3). It shows that for the first 12 hours, whether the forecasts start at 00:00 UTC or at 12:00 UTC ('ECMWF sorted') or if the forecast

starts at any time ('ECMWF unsorted') the errors are close. Then hereafter, t_0 does not refer to 00:00 UTC or 12:00 UTC, but it refers to the time when the downscaling models are launched.

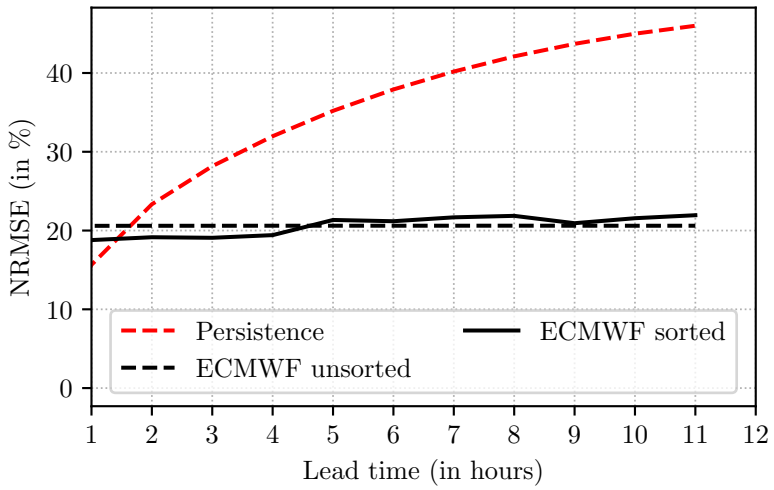


Figure 2.1 | Comparison of ECMWF performances when t_0 refers to 00:00 UTC or 12:00 UTC ('ECMWF sorted') and when t_0 refers to the starting time of the downscaling models forecasts ('ECMWF unsorted'). In the first case, there are two forecasts a day (00:00 UTC or 12:00 UTC), in the second cases these are hourly forecasts. The performance of persistence (described in 2.2.3) is added as reference.

The downscaling model is trained using 47 variables aiming at describing the boundary layer, winds and temperature in the lower troposphere. Tables 2.1, 2.2 and 2.3 show the considered variables. The large scale circulation brings the flow to the given location. The wind speed in altitude, the geopotential height, the vorticity, the flow divergence, or the temperature can be markers of large systems like depressions, fronts, storms, or high pressure systems (table 2.2). At a finer scale, what is happening in the boundary layer is very important to explain the intra-day variations of the wind speed. The state and stability of the boundary layer can be derived from surface variables describing the exchanges inside the layer. Exchanges are driven mostly by temperature gradient and wind shear that develop turbulent flow (table 2.3). Thermodynamical variables like surface, skin, and dew point temperatures and surface heat fluxes can also inform on the stability of the boundary layer, as well as its height and dissipation on its state (table 2.1). The spatial resolution of ECMWF forecasts is of about 16 km (0.125° in latitude and longitude).

Among the explanatory variables, $X_{1,t}, ..., X_{m,t}$, some provide more important information, and some may be correlated. Thus, a stepwise regression (forward selection) is performed to only keep the most important uncorrelated variables [35]. This is an iterative regression, which consists of adding variables from the set of explanatory variables based on the Bayesian Inference Criterion (BIC) [36]. The BIC is based on the likelihood function, and it reduces overfitting by introducing a penalty term for the number of parameter in the model. The model with the lowest BIC is preferred. At each step, a model is built by adding one variable among the remaining ones. The added variable which minimizes the BIC of the model is chosen. The procedure is repeated as long as the BIC decreases.

The linear regression considering all the explanatory variables given in tables 2.1, 2.2 and 2.3 is denoted $\mathrm{LR_A}$ and the linear regression considering a sample of explanatory variables (selected

Altitude (m)	Variable	Unit	Name
10 m / 100 m	Zonal wind speed	m s^{-1}	u
	Meridional wind speed	m s^{-1}	v
2 m	Temperature	K	t
	Dew point temperature	K	dp
Surface	Skin temperature	K	skt
	Mean sea level pressure	Pa	msl
	Surface pressure	Pa	sp
	Surface latent heat flux	J m^{-2}	slhf
	Surface sensible heat flux	J m^{-2}	sshf
-	Boundary layer dissipation	J m^{-2}	bld
	Boundary layer height	m	blh

Table 2.1 | Surface variables

Pressure level (hPa)	Variable	Unit	Name
1000 hPa / 925 hPa / 850 hPa / 700 hPa / 500 hPa	Zonal wind speed	m s^{-1}	u
	Meridional wind speed	m s^{-1}	v
	Geopotential	m^2 s^{-2}	z
	Divergence	s^{-1}	d
	Vorticity	s^{-1}	vo
	Temperature	K	t

Table 2.2 | Altitude variables

Altitude	Variable	Unit	Name
10 m / 100 m	Norm of the wind speed	m s^{-1}	F
10 m to 925 hPa	Wind shear	m s^{-1}	DF
2 m to 925 hPa	Temperature gradient	K	DT

Table 2.3 | Computed variables

by the stepwise regression) is denoted $\mathrm{LR_{SW}}$.

2.2.2 Non parametric downscaling approaches

Non-parametric models do not suppose to advance a specific relation between the variables. Instead, they try to learn this complex link directly from the data itself. As such, they are very flexible. The family of non parametric is quite large. Among others, one may cite the nearest neighbor's rule, the neural network, support vector machines, regression trees, random forests, for example. Regression trees which have the advantage of being easily interpretable, show to be particularly effective when associated with a procedure allowing to reduce their variance as for Random Forest Algorithm. Regression trees are binary trees built by choosing at each step the cut minimizing the intra-node variance, over all explanatory variables $X_{1,t}, ..., X_{m,t}$, and all possible thresholds S_j. More specifically, the intra-node variance is defined by:

$$D(X_k, S_k) = \sum_{X_k < S_k} (y_s - \bar{y}^-)^2 + \sum_{X_k \geq S_k} (y_s - \bar{y}^+)^2 \tag{2.2}$$

where $\bar{y}^-$ (respectively $\bar{y}^+$) denotes the averaged of the target in the area $\{X_k < S_k\}$ (respectively $\{X_k \geq S_k\}$). Then, the selected k_0 variable and associated threshold is given by $(X_{k_0}, S_{k_0}) = \arg\min_{(k,S_k)} D(X_k, S_k)$. The prediction is provided by the value associated to the leaf in which the observations falls.

To reduce variance and avoid over-fitting, it is interesting to generate several bootstrap samples, then fit a tree on every sample and average the predictions, which leads to the so-called Bagging procedure [37]. More precisely, for B bootstrap samples, the predicted value is given by:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}^b \tag{2.3}$$

where $\hat{y}^b$ is the value predicted by the regression tree associated with the $b-$th bootstrap sample. To produce more diversity in the trees to be averaged, an additional random step is introduced in the previous procedure leading to Random Forests. In this case, the best cut is chosen among a smaller subset (corresponding to $\frac{1}{3}m$ variables, where m is the total number of explanatory variable) of randomly chosen variables. The predicted value is the mean of the predictions of the trees, as in equation (2.3). Hereafter, this model is denoted RF. Again, the explanatory variables are retrieved from ECMWF forecasts and given in tables 2.1, 2.2 and 2.3.

Each model introduced above: $\mathrm{LR_A}$, $\mathrm{LR_{SW}}$, and RF has been tested using in-situ measurements collected at three wind farms called Parc de Bonneval, Moulin de Pierre and Parc de la Vènerie. For the three models, two configurations are tested. The first one consists of a classical downscaling using the explanatory variables available from ECMWF outputs only. The second one consists of adding the error between the observed wind speed at time t_0, i.e., when the forecast is launched, and the forecasted wind by ECMWF at time t as an explanatory variable. Hereafter, when the models are trained according to the first configuration, they are denoted $\mathrm{LR_A^{no\text{-}obs}}$, $\mathrm{LR_{SW}^{no\text{-}obs}}$, and $\mathrm{RF^{no\text{-}obs}}$. When the models are trained according to the second configuration, they are denoted $\mathrm{LR_A^{obs}}$, $\mathrm{LR_{SW}^{obs}}$, and $\mathrm{RF^{obs}}$.

In the first case, only one model is fitted. In the second case, a model is fitted at each hour in order to precisely take into account the error between the forecasted wind at time t and the observations at time t_0. For the second configuration, after the variable selection step, between 14 and 21 variables remain, depending on the horizon. In [35], Alonzo *et al.* compare this low cost assimilation to the downscaling without in-situ information. For a 3 h lead-time, they can improve the forecast up to 18% by considering the initial error.

2.2.3 Benchmark methods

For short term predictions, statistical methods are the most used and are always compared to persistence [27]. Persistence assumes that the wind speed at time t will be the same as it was at time t_0. Although this model is very simple, it is, in fact, difficult to beat for look-ahead times from 0 to 4-6 hours. This is due to the fact that changes in the atmosphere take place rather slowly [17]. The statistical approach aims at finding the relationship between past and future observations using measurements (and possibly exogenous variables). They can be split into two

sub-categories: time series based models which are easy to model and cheap to develop an artificial neural network which can deal with non-linearity but which is known as black box model.

Time series models are mainly based on Auto-Regressive Moving Averaged (ARMA) models [38]. An ARMA(p, q) model aims at predicting the wind speed at time t, using a linear combination of the p previous wind speed values, the q previous residuals and potentially m exogenous variables (in that case we define the model as ARMAX). The most sophisticated models are ARIMAX(p, d, q) for Auto-Regressive Integrated Moving Averaged EXogenous. They aim at removing the non-stationarity of the data by applying an initial d-order differencing step as follow:

$$\widehat{y}_t = \sum_{i=1}^{p} \Phi_i \Delta^d y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{m} \beta_k X_{k,t} \tag{2.4}$$

$$\Delta^d y_t = (y_t - y_{t-1}) - \sum_{i=1}^{d-1} (y_{t-i} - y_{t-(i+1)}), \quad d = 1, ..., n \tag{2.5}$$

where y_{t-i} is the observed wind speed at time $t-i$, $\Phi_i, \theta_j, \beta_k$ are the model parameters, Δ^d is the d-order lag operator defined in equation (2.5), ε_{t-j} is the residual at time $t-j$, and $X_{k,t}$ is the k^{th} explanatory variable at time t, which can be an output from NWP.

Artificial neural networks (ANN) are models inspired by biological neural networks. They are based on interconnected groups of nodes, divided into layers. Each connection can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and transmit it to another neuron. Usually, this signal is a set of real number, and the output of each artificial neuron is computed by some non-linear function, called activation function, of a weighted sum of its input. The weights and the activation function are updated through the training process [39, 40]. Those models are very useful in the modeling of complex non-linear relationships and extract dependences between variables.

Choice of the parameters

The choice of the benchmark model parameters is a crucial step. For both hourly and sub-hourly forecasts, we fit several models, and we choose the most efficient one. For the ARMA models, we use the BIC to select the models. For several values of the p and q parameters, we fit the corresponding models, and the model that minimizes the criterion is preferred. Figure 2.2 displays the results for hourly forecasts (a) and sub-hourly forecasts (b) at Parc de Bonneval. For the first one, an ARMA(6,3) appears to be the best model, and for sub-hourly forecasts, it is an ARMA(4,2). The same procedure is applied for Parc de la Vènerie and Moulin de Pierre. At Moulin de Pierre (resp. Parc de la Vènerie), an ARMA(5,3) (resp. ARMA(2,1)) gives the best results for the hourly forecasts. For the sub-hourly forecasts, the selected model is at Moulin de Pierre (resp. Parc de la Vènerie) is an ARMA (3,3) (resp. ARMA(5,4)).

For the ANN, we compute the RMSE, defined in equation (2.6), as a function of the number of layers and the number of neurons per hidden layer. Results are shown for the first lead time. We fix the seed in order to better compare the model's performance and remove the noise due to the stochastic nature of ANN.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\widehat{y}_i - y_i)^2} \tag{2.6}$$

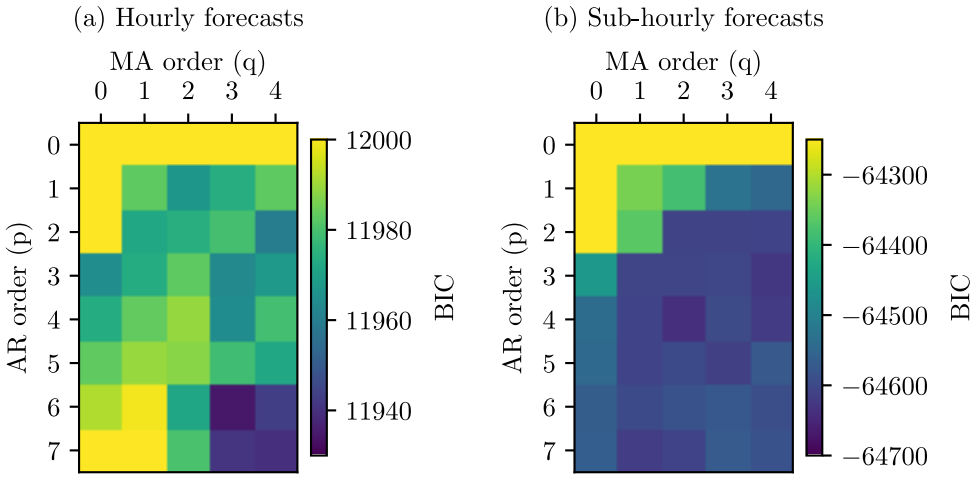


Figure 2.2 | Bayesian Inference Criterion (BIC) of different ARMA(p,q) models depending on the AR order p and the MA order q for the hourly forecasts (a) and the sub-hourly forecasts (b) at Parc de Bonneval.

In equation (2.6), $\widehat{y}_i$ is the i-th wind forecast and y_i is the corresponding observation. N refers to the number of forecasts that have been done to compute the RMSE.

According to figure 2.3, at Parc de Bonneval, the best ANN for the hourly forecasts is a model with 2 layers (one hidden layer and one output layer) with 10 neurons in the hidden layer and the best model for sub-hourly forecasts is an ANN with 4 layers and 10 neurons in the hidden layers. Again, the same procedure is applied for Parc de la Vènerie and Moulin de Pierre. At Moulin de Pierre the same configuration than at Parc de Bonneval is used for the hourly forecasts, and at Parc de la Vènerie, an ANN with 2 layers and 40 neurons in the hidden layer is chosen. For sub-hourly forecasts an ANN with 4 layers and 30 neurons in the hidden layers is chosen at Moulin de Pierre and an ANN with 4 layers and 10 neurons in the hidden layers is chosen at Parc de la Vènerie.

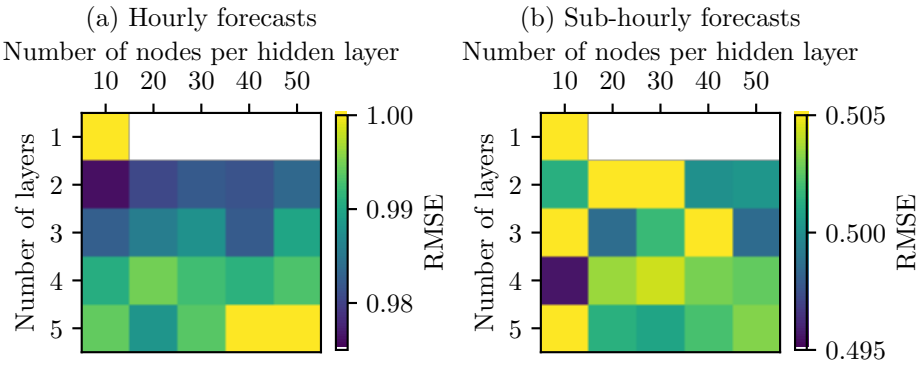


Figure 2.3 | RMSE for the first lead time of different ANN model depending on the number of hidden layers and the number of neurons per hidden layer. The results for hourly forecasts (a) and sub-hourly forecasts (b) are shown.

Use of exogenous variables

Both ARMA and ANN can be used as pure time-series based models or with exogenous variables. Figure 2.4 displays the comparison of the error distributions at Parc de Bonneval for the forecasts at $t_0 + 1$ h and $t_0 + 3$ h between ARMA and ARMAX and between ANN and ANNX. We use only the 100 m wind speed ($F = \sqrt{u^2 + v^2}$) forecasted by ECMWF as exogenous variable. For a lead time of 1 h, the distributions between the pure time-series based models and the models with exogenous variable do not differ significantly. For both ARMA and ANN, the interquartile range (IQR) is slightly reduced by the use of exogenous variables. From 1.63 m s^{-1} to 1.42 m s^{-1} for ARMA and from 1.10 m s^{-1} to 1.07 m s^{-1} for ANN. However, it slightly degrades the bias as it goes from -0.02 for ARMA to -0.03 for ARMAX and from -0.16 for ANN to -0.22 for ANNX. For a lead time of 3 h, the results are the same with a more significant gain. The IQR is reduced from 2.25 for ARMA to 1.71 for ARMAX and from 2.09 for ANN to 1.43 for ANNX. The scope, defined as the difference between the highest and lowest value, is also significantly reduced compared to the lead time of 1 h. At t_0+1 h, the differences start to be visible. However, our goal is sub-hourly forecasts with lead times starting from 10 min. At those time scales, the hourly exogenous variables carry less information than in-situ measurements. Under these conditions, we keep a pure time-series based approach for ARMA and ANN.

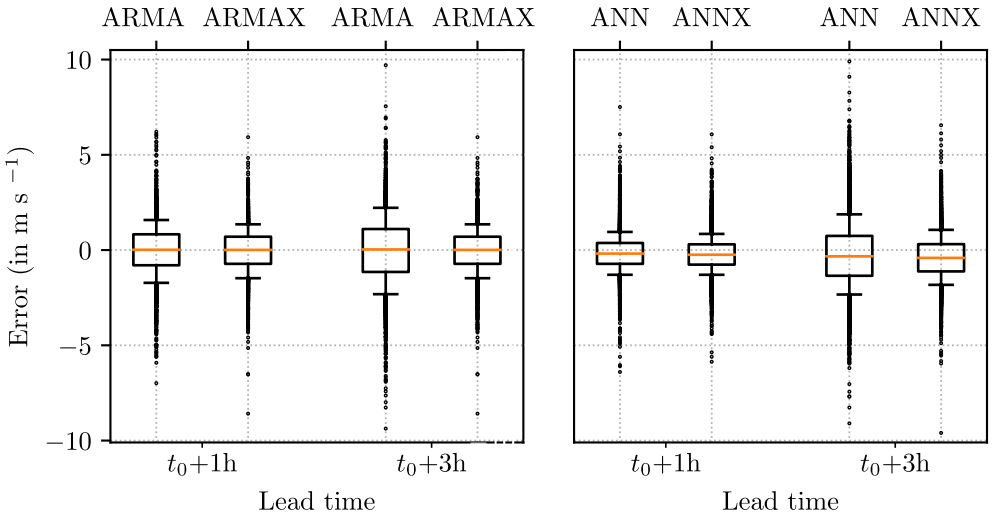


Figure 2.4 | Comparison for Parc de Bonneval between the pure time-series based ARMA (resp. ANN) and a model with the 100 m wind speed forecasted by ECMWF as exogenous variable ARMAX (resp. ANNX). The distribution of the error between the forecasted wind speed (using ARMA, ARMAX, ANN and ANNX) and the measurements are shown at $t_0 + 1$ h and $t_0 + 3$ h.

Data standardization

For both ARMA and ANN, the dataset has to be standardized so as to tend a distribution, close to a Weibull distribution (see 2.5), towards a standardized Gaussian distribution. This transformation allows ARMA and ANN to give optimal results. The standardized dataset is easily obtained by centering and reducing the data, as shown in equation (2.7).

$$Y_{\mathcal{N}(0,1)} = \frac{y - \bar{y}}{\sigma_Y} \tag{2.7}$$

where $\bar{y}$ is the mean and σ_Y is the standard deviation. Figure 2.5 shows the comparison between the two dataset.

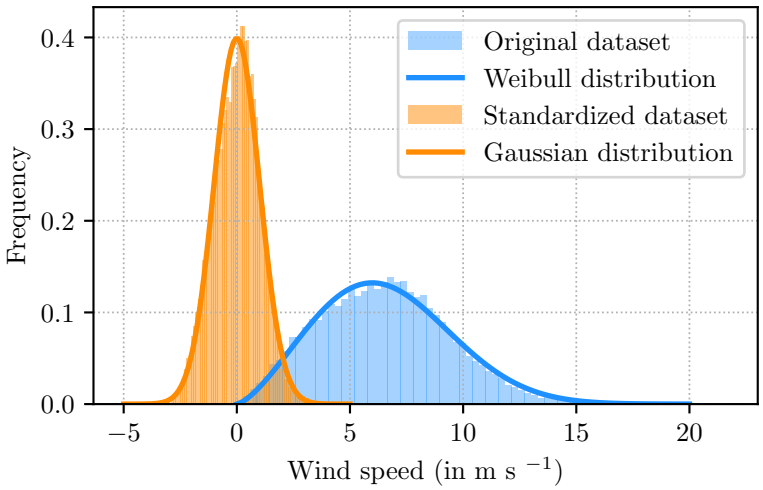


Figure 2.5 | Comparison between the distribution of the original dataset (Weibull distribution) and the standardized dataset (Gaussian distribution)

This transformation is not necessary for the models presented in sections 2.2.1 and 2.2.2. The models are able to learn the statistical relationship using the original dataset.

2.3 Application at two wind farms

To quantify the performance of the models, we used two indicators. The Normalized Root Mean Square Error (NRMSE) defined in equation (2.8), which is often used and facilitates comparisons with classical scores. The second indicator is the improvement over persistence, defined in equation (2.9), that is to say the decrease of the RMSE of the considered model compared to the persistence method. This skill score is referred to as Δ_{RMSE}.

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}}{\bar{Y}} \tag{2.8}$$

$$\Delta_{RMSE} = -\frac{RMSE_{model} - RMSE_{persistence}}{RMSE_{persistence}} \tag{2.9}$$

where, $\widehat{y}_i$ is the i-th wind forecast and y_i is the corresponding observation. N refers to the number of forecasts that have been done to compute the NRMSE, and $\bar{Y}$ is the mean of the observed wind speed over the same sample. By removing the normalization, we obtained the $RMSE$. When $\widehat{y}_i$ is forecasted using persistence, it refers to $RMSE_{persistence}$. When it is forecasted by any other

model, it refers to $RMSE_{model}$, where the model is clearly identified (among ECMWF, $\mathrm{LR}_{\mathrm{A}}^{\text{no-obs}}$, $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$, $\mathrm{RF}^{\text{no-obs}}$, $\mathrm{LR}_{\mathrm{A}}^{\text{obs}}$, $\mathrm{LR}_{\mathrm{SW}}^{\text{obs}}$, RF^{obs}, ARMA and ANN).

In the following, we present the results of the models applied to three different wind farms operated by Zephyr ENR. The first wind farm is located in Bonneval, a small town 100 km South-west of Paris, France (48.20°N and 1.42°E). It is called Parc de Bonneval and it was implemented in 2006. This wind farm is composed by 6 Vestas V80-2 MW turbines of 100 m hub height. The second one, called Moulin de Pierre, is located 5 km from Parc de Bonneval. It is also composed by 6 Vestas of nominal power of 3.3 MW. The second wind farm is located in Normandie, France (49.01°N and -1.16°E). It is called "Parc de la Vènerie" and it was implemented in 2014. This wind farm is composed of 4 Enercon E82-2.3 MW turbines and a hub height of 85 m.

Parc de Bonneval and Moulin de Pierre are located in the middle of the fields with a very flat topography, as shown in figure 2.6a, while Parc de la Vènerie is surrounded by forests with a more rugged topography as shown in figure 2.6b. It makes the environment more complex and, therefore, the forecast more uncertain.

For both wind farms, the explanatory variables are interpolated linearly from the four nearest grid points at the farm location.

(a) Parc de Bonneval **(b)** Parc de la Vènerie

Figure 2.6 | Pictures of Parc de Bonneval (a) and Parc de la Vènerie (b). Moulin de Pierre is similar to Parc de Bonneval.

2.3.1 Performances for hourly forecasts

In this section, the downscaling methods are used for hourly forecasts and tested against the commonly used ANN and ARMA methods. The targeted wind speed is computed by averaging the 10-minutes measurements of the considered wind farms. For each wind farm, spatial averaging is performed by averaging the data of all turbines. Hourly forecasts have been largely studied in the literature, and the results are compared to published reference skill scores. At Parc de Bonneval and Parc de la Vènerie, all the models are trained using hourly averaged of the past observations of years 2015 and 2016. At Moulin de Pierre, a k-fold cross validation is performed using data of 2017.

Figures 2.7, 2.8, and 2.9 display the NRMSE at Parc de Bonneval, Parc de la Vènerie, and Moulin de Pierre depending on the forecast lead times (1 h to 11 h) for persistence, ECMWF forecasts, ARMA and ANN models and for our downscaling methods $\mathrm{LR}_{\mathrm{A}}^{\text{no-obs}}$, $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$, $\mathrm{RF}^{\text{no-obs}}$, $\mathrm{LR}_{\mathrm{A}}^{\text{obs}}$, $\mathrm{LR}_{\mathrm{SW}}^{\text{obs}}$ and RF^{obs}.

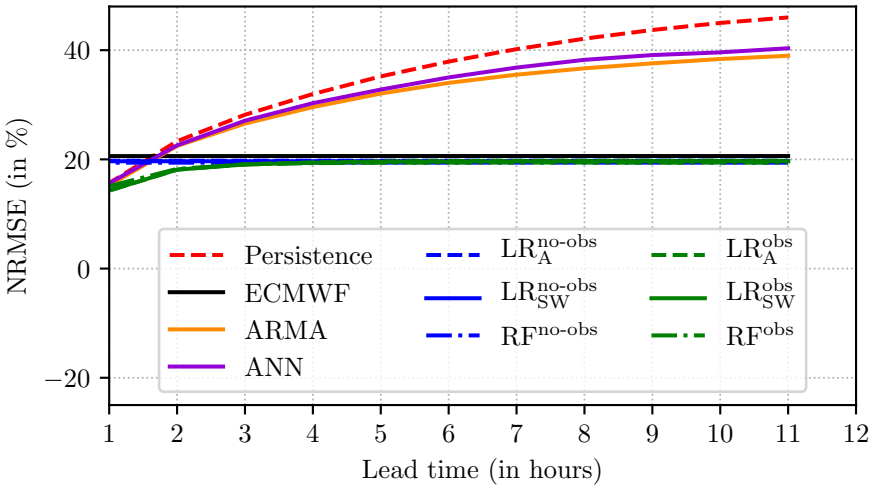


Figure 2.7 | Performances of downscaling statistical models for hourly forecasts from 1 h to 11 h in two configurations against the performances of ECMWF and the benchmark method. $LR_A^{no\text{-}obs}$, $LR_{SW}^{no\text{-}obs}$, and $RF^{no\text{-}obs}$ display the downscaling of explanatory variables from ECWMF outputs only. LR_A^{obs}, LR_{SW}^{obs}, and RF^{obs} show the results when the error between the measurements at t_0 and the 100-m wind speed forecasted by ECMWF at t is adding as explanatory variable. Results of persistence, ANN, and ARMA are added.

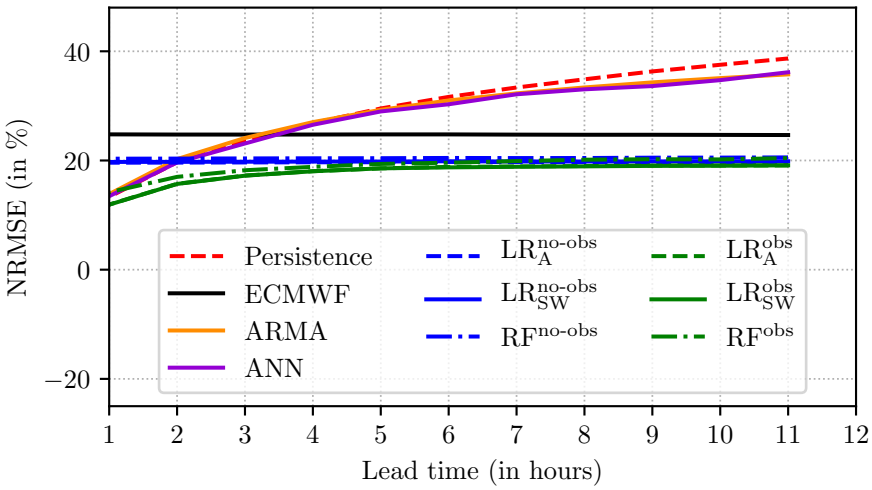


Figure 2.8 | Same than figure 2.7 for Parc de la Vènerie.

Results at Parc de Bonneval and Moulin de Pierre are very similar due to the proximity of the two farms. Consequently, only the results at Parc de Bonneval will be discussed. The main difference with Parc de la Vènerie is the performances of ECMWF. At Parc de Bonneval, the NRMSE is around 20% for all lead times while it is around 25% for Parc de la Vènerie. The flat topography at Parc de Bonneval enables ECMWF to perform well. In this condition, the downscaling leads to minor improvements. However, at Parc de la Vènerie, the improvement over

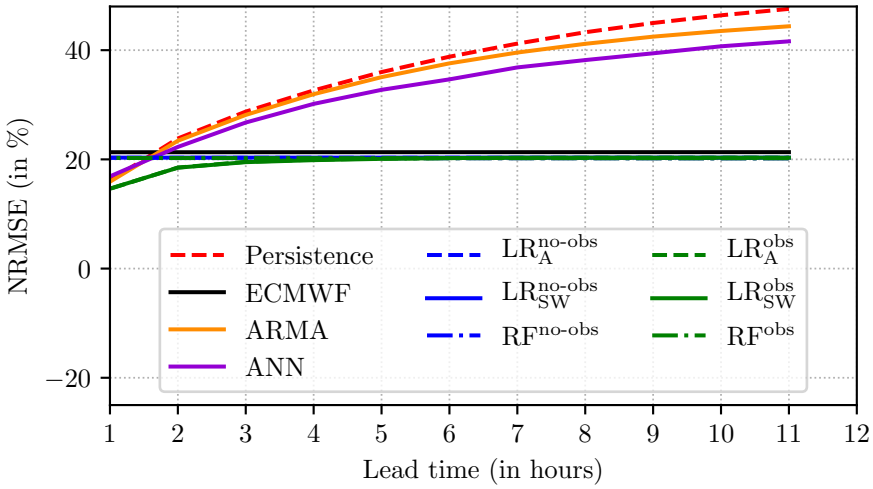


Figure 2.9 | Same than figures 2.7 and 2.8 for Moulin de Pierre.

ECMWF is more significant as its performances are worse. In both cases, ARMA and ANN are close to persistence for the whole period, and as for the downscaling models, the models with observations converge towards the models without observation after few hours.

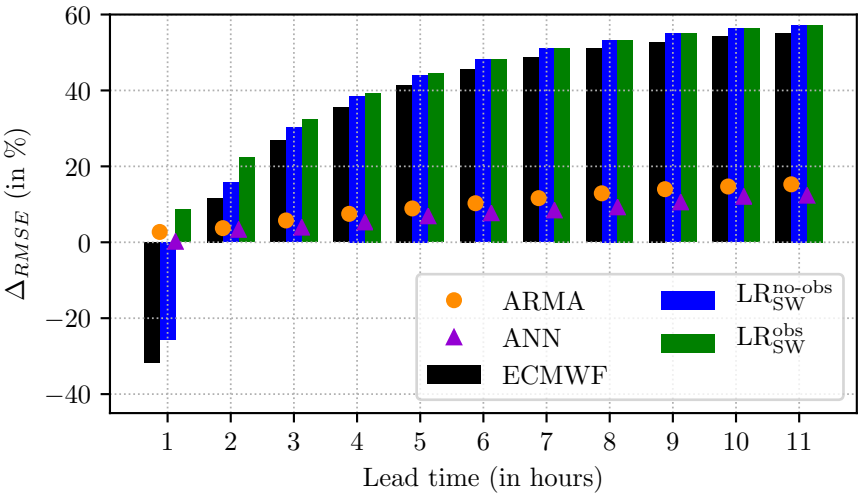


Figure 2.10 | Comparison of the improvements over persistence in percentage for ECMWF forecasts and the best downscaling models from 1 h to 11 h in the two configurations. LR$_{SW}^{no-obs}$ corresponds to the downscaling of explanatory variables from ECWMF outputs only with a variable selection algorithm. LR$_{SW}^{obs}$ shows the results when the error between the measurements at t_0, the time when the forecast is launched, and the 100-m wind speed forecasted by ECMWF at t is added as an explanatory variable. Again with a variable selection algorithm. Improvements of ECMWF, ARMA, and ANN methods are also included. For ECMWF and the downscaling models, the value of the improvement corresponds to the extremity of each bar, while for ARMA and ANN, it corresponds to the center of the circle and triangle, respectively.

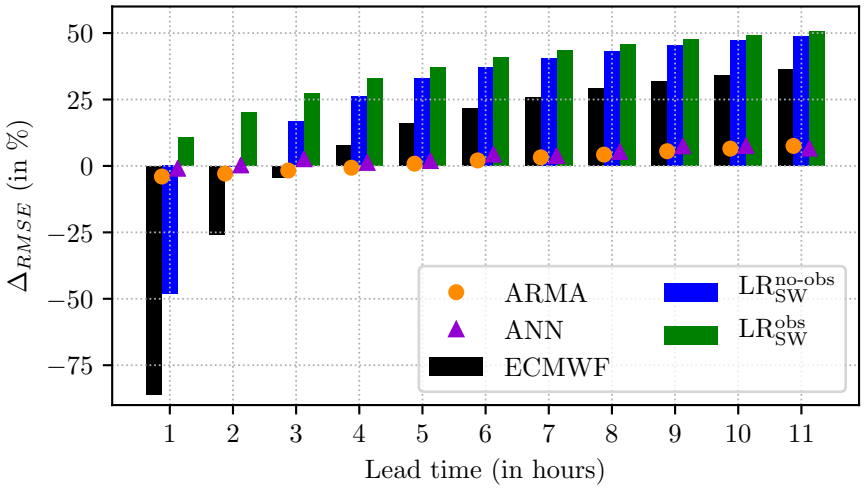


Figure 2.11 | Same that figure 2.10 for Parc de la Vènerie.

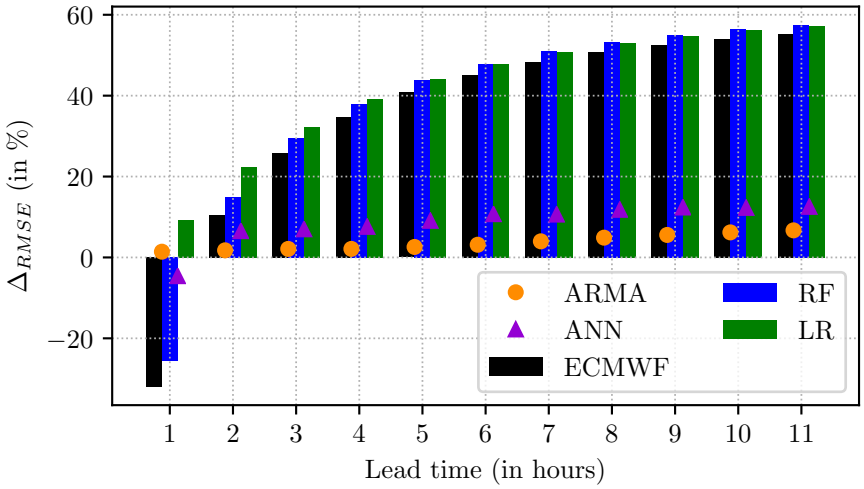


Figure 2.12 | Same that figures 2.10 and 2.11 for Moulin de Pierre.

The improvements over persistence of the benchmark methods ARMA and ANN, ECMWF and of the best downscaling models for each configuration $LR_{SW}^{no\text{-}obs}$ and LR_{SW}^{obs} are displayed in figure 2.10 for Parc de Bonneval, figure 2.11 for Parc de la Vènerie, and figure 2.12 for Moulin de Pierre. One can see that reference methods, ARMA and ANN, perform very similarly for all wind farms. At Parc de la Vènerie, ARMA overperforms persistence from 4 h, and ANN overperforms persistence from 3 h. The maximal improvement is found for the last lead time at $t_0 + 11$ h and is 7.5% for ARMA and 6.5% for ANN. At Parc de Bonneval, the two models overperform persistence at every horizon, and the improvement slightly increases with time from 2.7% for the first hour to 15.3% for the eleventh. The difference between the two wind farms can be explained by the persistence performance, which is better at Parc de la Vènerie. Then, the improvement is lower.

The results at Moulin de Pierre are again very similar to those at Parc de Bonneval, so they will not be further discussed.

Those results are consistent with those found in the literature. For instance, in [41], Torres *et al.*, use ARMA model to predict hourly averaged wind speed 1 h to 10 h lead time for five sites in Spain. They find NRMSE improvement over persistence ranging between 2% and 5% for 1 h ahead and between 12% and 20% for 10 h ahead. In [42], Sfetsos compares the performances of an ARIMA(2,1,2) and an ANN using measurements collected in Crete, Greece. Hourly averaged wind speed forecasts with ANN model overperform persistence by 4.7% while ARIMA overperforms persistence by 2.3%.

Compared to these reference results, $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$ and $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ are significantly better. After the fifth hour, ECMWF, $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$, and $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ are better than persistence by more than 40% for both wind farms. For the first lead time, corresponding to t_0+1 h, $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ performs better than persistence by 8.6% at Parc de Bonneval which is better than ARMA ($\Delta_{RMSE} = 2.7\%$) and ANN ($\Delta_{RMSE} = 0.1\%$). At Parc de la Vènerie $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ performs better than persistence by 10.6% while Δ_{RMSE} = -4.0% for ARMA and Δ_{RMSE} = -1.1% for ANN. At Parc de Bonneval, the improvements remain significantly better than ECMWF and $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$ until the third hour. However, at Parc de la Vènerie ECMWF performances remain significantly worse for all lead times, and the differences between $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$ and $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ are bigger.

The performance shift at t_0+2 h/3 h for Parc de Bonneval and t_0+4 h/5 h for Parc de la Vènerie between the observations based methods and the downscaling methods can easily be explained. For short lead times, an accurate initial state provided by the observations is essential. For the last lead times, the observations no longer constrain the forecast. Thus NWP forecasts, provide the needed information. Moreover, for these lead times, ARMA and ANN models are no longer based on the latest measurements but on previous forecasts. It explains why $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ outperforms all other methods at all lead times.

2.3.2 Performances for sub-hourly forecasts

In this section, we focus on very short term forecasts, which is the principal objective of this thesis. We apply the same methods as in section 2.3.1 to forecast 10 min average winds up to 3 h ahead. In order to retrieve 10-min forecasts, the explanatory variables are linearly interpolated every 10 min. Then, to retrieve the prediction for all hours h at minutes 0, 10, and 20, we apply the model calibrated at hour h. To retrieve the prediction for all hours h at minutes 30, 40, and 50, we apply the model calibrated at hour $h + 1$. However, the calibration leads to an issue with $\mathrm{LR}^{\mathrm{obs}}$. For 10 min and 20 min, $\mathrm{LR}^{\mathrm{obs}}$ is doing exactly the same as persistence. Indeed, the model fitted at time t_0 puts all the weight on the forecasted wind speed by ECMWF and on the initial error. As this model is used at 10 min and 20 min, the results are exactly the results of persistence. To let the model outperforms persistence, one solution is to do a linear regression using only past observations for the first two horizons. Hereafter, $\mathrm{LR}^{\mathrm{obs}}$ denotes a linear regression over past measurements for time 10 min and 20 min and a linear regression over ECMWF outputs and the error at time t_0 for the remaining time as shown in figure 2.13. Same thing for $\mathrm{RF}^{\mathrm{obs}}$, where the explanatory variables for the lead times of 10 min and 20 min are only past measurements. For the reference methods ANN and ARMA, the training is performed directly using the 10-minutes measurements. The procedure applied to choose the models is the same as in section 2.3.1. For the ARMA models, as explained in section 2.2.3, an ARMA(4,2) is used for Parc de Bonneval, an ARMA(3,3) for Moulin de Pierre, and an ARMA(5,4) for Parc de la Vènerie. For the ANN, we fit several models depending on the number of layers and the number of neurons per layer. The

best model is an ANN with 4 layers and 10 neurons per layers for Parc de Bonneval and Parc de la Vènerie and 4 layers with 30 neurons per layers for Moulin de Pierre.

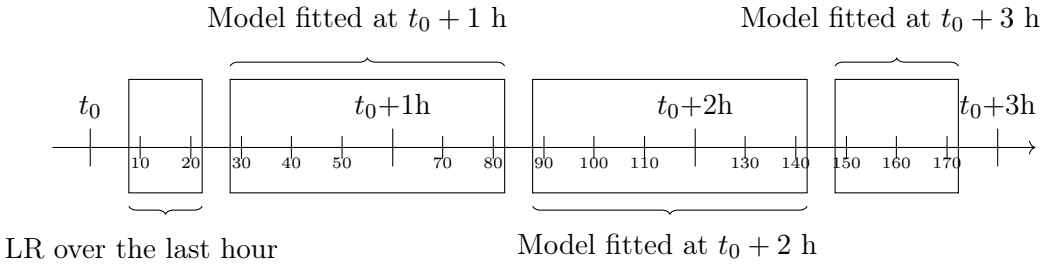


Figure 2.13 | Diagram of downscaling models. Each model is fitted using data at peak hour and then used from -30 min to +20 min.

Figure 2.14 (resp. figure 2.15 and figure 2.16) displays the NRMSE as a function of the time horizon, from 10 min to 170 min, for persistence, ECMWF forecasts, $\text{LR}_{\text{SW}}^{\text{no-obs}}$, and $\text{LR}_{\text{SW}}^{\text{obs}}$ forecasts and reference methods ARMA and ANN at Parc de Bonneval (resp. Parc de la Vènerie and Moulin de Pierre).

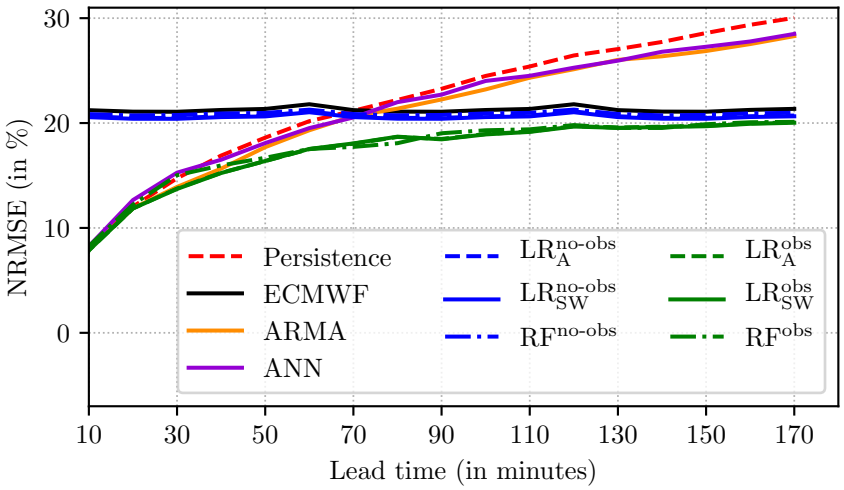


Figure 2.14 | Performances of downscaling statistical models for hourly forecasts from 1 h to 11 h in two configurations against the performances of ECMWF and the benchmark method. $\text{LR}_{\text{A}}^{\text{no-obs}}$, $\text{LR}_{\text{SW}}^{\text{no-obs}}$, and $\text{RF}^{\text{no-obs}}$ display the downscaling of explanatory variables from ECWMF outputs only. $\text{LR}_{\text{A}}^{\text{obs}}$, $\text{LR}_{\text{SW}}^{\text{obs}}$, and RF^{obs} show the results when the error between the measurements at t_0 and the 100-m wind speed forecasted by ECMWF at t is adding as explanatory variable. Results of persistence, ANN, and ARMA are added.

At this time scale, the differences between the models are smaller than for longer lead times, especially at Parc de Bonneval and Moulin de Pierre, but the hierarchy between them remains the same. In all cases, it is hard to distinguish the best model at 10 min and 20 min, but after 30 min, $\text{LR}_{\text{SW}}^{\text{obs}}$ is significantly better. For times between 30 min and 2 h, it provides clearly the best forecasts, with NRMSE less than 20%. For lead times of 2 to 3 h, its performance gradually converges to that of $\text{LR}_{\text{SW}}^{\text{no-obs}}$.

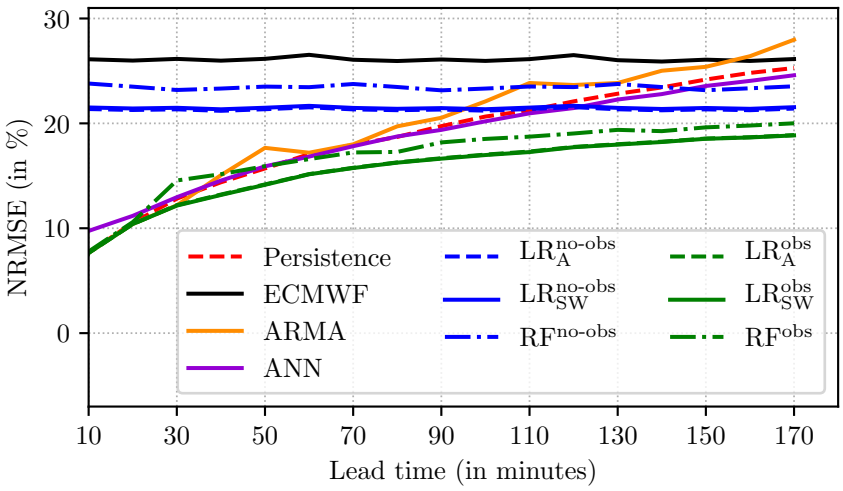


Figure 2.15 | Same than figure 2.14 for Parc de la Vènerie.

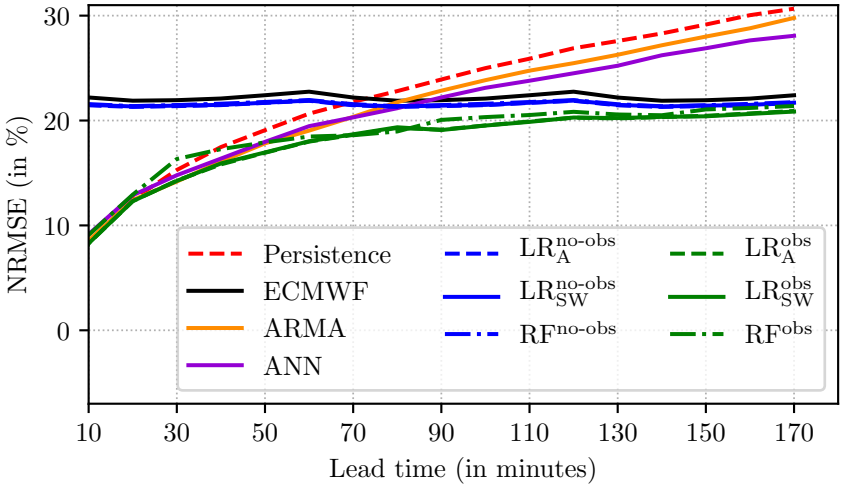


Figure 2.16 | Same than figures 2.14 and 2.15 for Moulin de Pierre.

Figures 2.17, 2.18, and 2.19 are similar to figures 2.10, 2.11, and 2.12 for lead times ranging between 10 min and 170 min. Only LR_{SW}^{obs} overperforms persistence at every horizon. Again it is the model giving the best improvements. The differences with ARMA are not extremely significant for the first lead times, especially at 20 min (1.5% for LR_{SW}^{obs} and 1.3% for ARMA at Parc de Bonneval and 1.2% for LR_{SW}^{obs} and 1.0% for ARMA at Parc de la Vènerie). After 20 min, LR_{SW}^{obs} is by far the best model. The improvement over persistence is 6.9% at 30 min and 33.3% at 170 min at Parc de Bonneval and 5.2% at 30 min and 25.4% at 170 min at Parc de la Vènerie. At Parc de Bonneval ECMWF, $LR_{SW}^{no\text{-}obs}$ and LR_{SW}^{obs} converge with each other with time. ECMWF and $LR_{SW}^{no\text{-}obs}$ start to outperform persistence only from 80 min and 70 min, respectively. However, at Parc de la

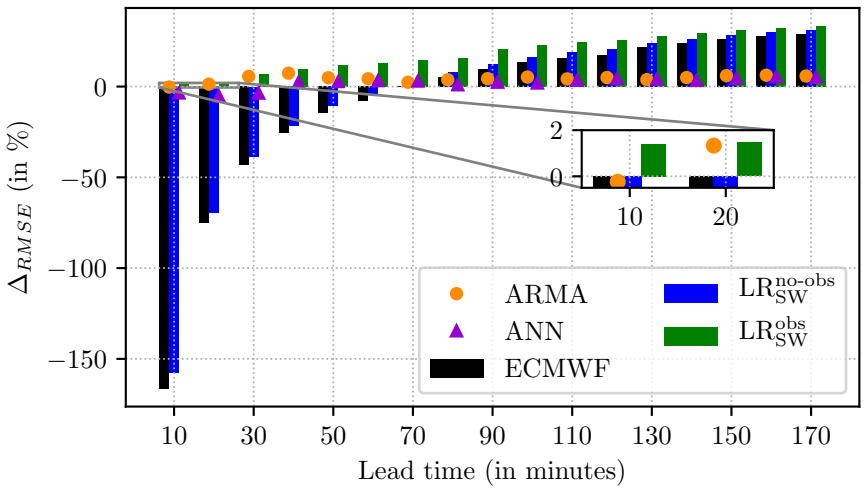


Figure 2.17 | Comparison of the improvements over persistence in percentage for ECMWF forecasts and the best downscaling models from 1 h to 11 h in the two configurations. $LR_{SW}^{no\text{-}obs}$ corresponds to the downscaling of explanatory variables from ECWMF outputs only with a variable selection algorithm. LR_{SW}^{obs} shows the results when the error between the measurements at t_0, the time when the forecast is launched, and the 100-m wind speed forecasted by ECMWF at t is added as an explanatory variable. Again with a variable selection algorithm. Improvements of ECMWF, ARMA, and ANN methods are also included. For ECMWF and the downscaling models, the value of the improvement corresponds to the extremity of each bar, while for ARMA and ANN, it corresponds to the center of the circle and triangle, respectively.

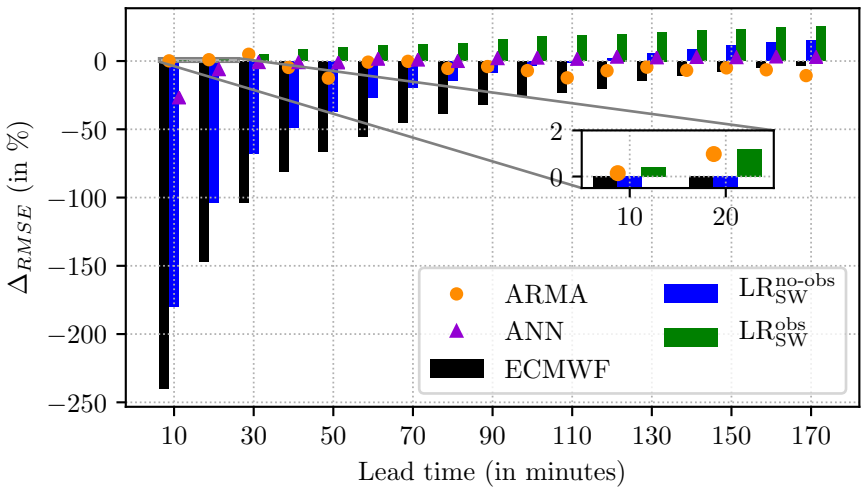


Figure 2.18 | Same than figure 2.17 for Parc de la Vènerie.

Vènerie, ECMWF never improves persistence during the 3 h, and the differences between $LR_{SW}^{no\text{-}obs}$ and LR_{SW}^{obs} remain significant for the whole period. Again, the results for Moulin de Pierre are not

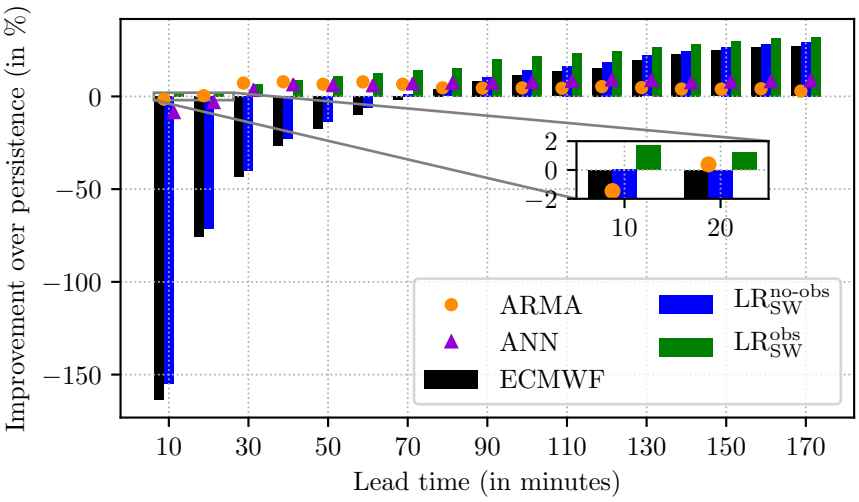


Figure 2.19 | Same than figures 2.17 and 2.18 for Moulin de Pierre.

further discussed due to the similarity with Parc de Bonneval.

2.4 Analysis of the best model

In this section, we analyze more deeply the models for Parc de Bonneval and Parc de la Vènerie. For both wind farms and for hourly and sub-hourly forecasts, LR_{SW}^{obs} is the best model. Figures 2.20 and 2.21 displays the chosen explanatory variables for each models. The name of the explanatory variables are given in tables 2.1, 2.2 and 2.3. For better readability differents colors are used for each models. Figure 2.20 displays the results for Parc de Bonneval and figure 2.21 displays the results for Parc de la Vènerie.

If we focus on the models used for sub-hourly forecasts, that is to say the models at $t_0 + 1$ h, $t_0 + 2$ h and $t_0 + 3$ h, one can see that the number of explanatory variables used at Parc de Bonneval is larger than at Parc de la Vènerie. At Parc de Bonneval, 20 variables are used, while there are 12 different explanatory variables used at Parc de la Vènerie. F100 and the error at t_0 are the main variables, i.e., used by all the models for both farms and with more weight. Then, the three models select indicators of the atmospheric stability and of turbulence such as the wind shear (DF), the surface sensible (sshf) and latent (slhf) heat fluxes. The latter is the transfer of latent heat (the thermal energy released or absorbed by a body during a phase transition without temperature change) with the surface through turbulent dissipation while the surface sensible heat flux is the transfer of heat between the Earth's surface and the atmosphere through the effect of turbulent air motion (this process results in a temperature change) [43]. At Parc de Bonneval F10 and the boundary layer dissipation (bld) are selected by the three models as well as the temperature gradient (DT) and the geopotential at 925 hPa (z925) at Parc de la Vènerie. In general, indicators of the state of the atmosphere, such as wind components at different altitudes and also temperature at different altitudes for Parc de Bonneval, are selected. More occasionally, variables such as vorticity (a measure of the rotation of air in the horizontal [43]), skin temperature

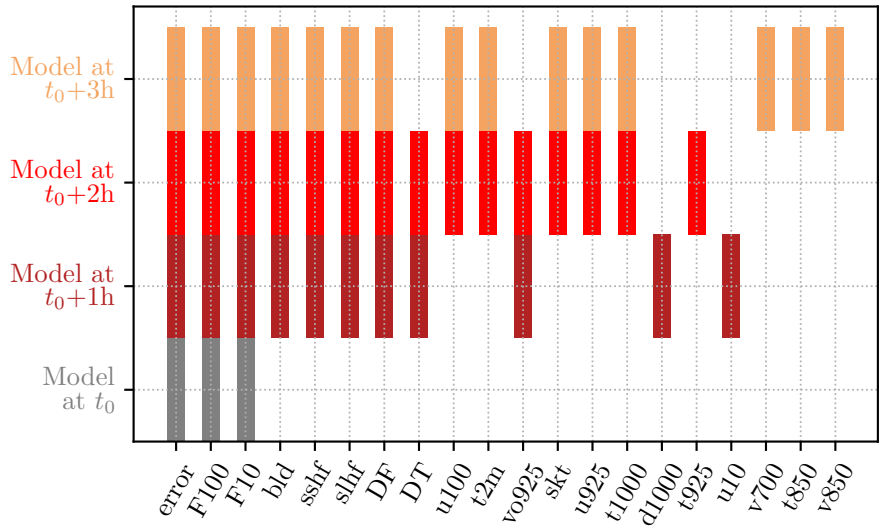


Figure 2.20 | Selected explanatory variables at Parc de Bonneval. The y-axis displays the different models using different colors. The x-axis give the name of the different explanatory variables. Their names can be found in tablees 2.1, 2.2 and 2.3. For each model, everytime an explanatory variable is used, a colored rectangle is drawn.

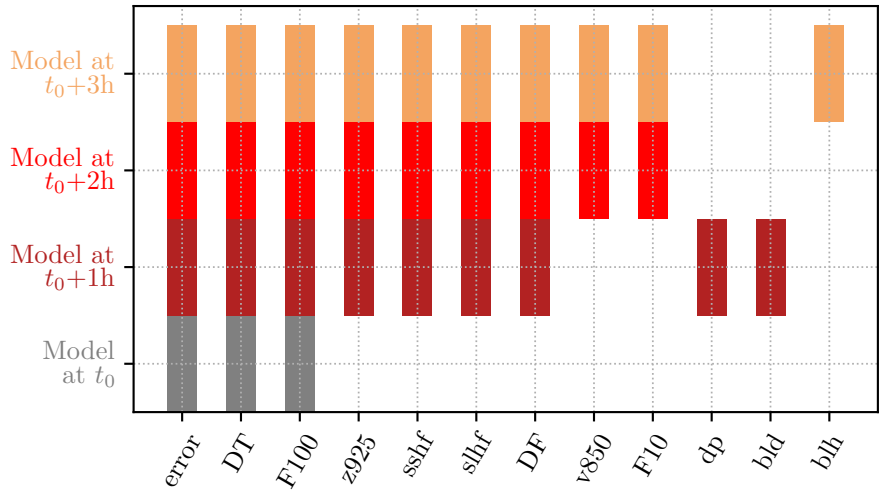


Figure 2.21 | Same as figure 2.20 for Parc de la Vènerie.

(temperature of the surface of the Earth) or dew point temperature (temperature to which the air at 2 m above the surface would have to be cooled for saturation to occur [43]) are selected.

For each model, the process is very similar. The 100 m wind speed forecasted by ECMWF is corrected using the error at t_0. The weight of the error decrease with time. Then, the other explanatory variables are used to update the forecasted values but to a lesser extent as their respective weight is reduced by a factor 10.

Figure 2.22 displays the evolution of the error weight $|\omega_{error}|$ with time for the two wind farms. ω_{error} refers to the parameter associated to the error in equation (2.1). The error is the explanatory variable computed as follows:

$$F^{\text{obs}_{t_0}}_{\text{hub\_height}} - F^{\text{ECMWF}_t}_{100} \tag{2.10}$$

where $F^{\text{obs}_{t_0}}_{\text{hub\_height}}$ is the wind speed measured at the hub height at the wind farm at t_0 that is to say, when the forecast is launched and $F^{\text{ECMWF}_t}_{100}$ is the 100 m wind speed forecasted by ECMWF at time t. The magnitude of the weight $|\omega_{error}|$ associated to this explanatory variable reflects the importance of in-situ information in the forecasting model. In both cases, the weight decreases with time. From 3 h, it is lower than 0.2 for Parc de Bonneval and lower than 0.3 for Parc de la Vènerie. For Parc de Bonneval the error is not selected by the last three models and $|\omega_{error}| = 0$ after $t_0 + 9$ h. At Parc de la Vènerie, this variable carries information for all lead times, and it is higher than 0.1 for the last model at $t_0 + 11$ h. These results are consistent with the results shown on figures 2.10 and 2.11. The difference between $\text{LR}^{\text{no-obs}}_{\text{SW}}$ and $\text{LR}^{\text{obs}}_{\text{SW}}$ remains significant for the whole period at Parc de la Vènerie, while at Parc de Bonneval, the two approaches converge to each other after 4 h.

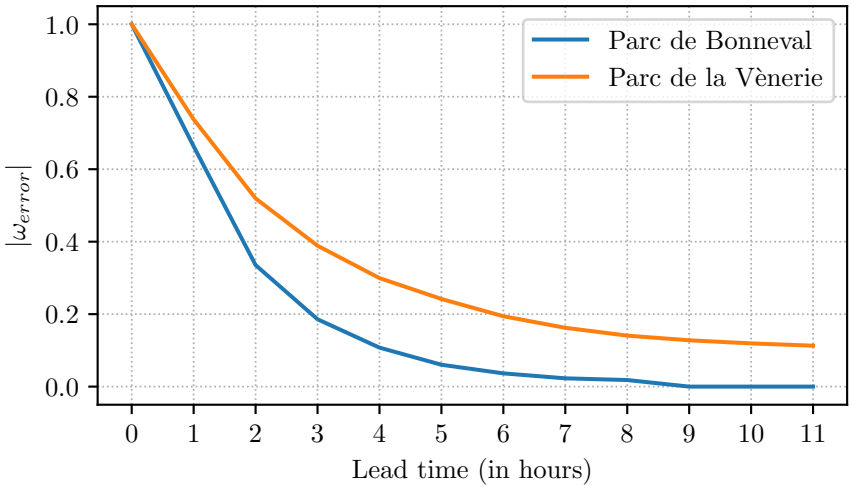


Figure 2.22 | Evolution of the error weight $|\omega_{error}|$ with time. For each model from t_0 to $t_0 + 11$ h the weight assigned to the error calculated by the linear regression $\text{LR}^{\text{obs}}_{\text{SW}}$ is plotted for Parc de Bonneval and Parc de la Vènerie.

2.5 Conclusion

In this chapter, we have developed and tested approaches that combine statistical models and outputs from Numerical Weather Prediction (NWP) model in order to forecast the 100 m wind speed at sub-hourly time scales. All the models are tested on three different wind farms. Traditionally, the main methods used for those time scales are time series based methods using only local observations, while Numerical Weather Prediction (NWP) models are preferred for lead times longer than 6 h at least [44]. However, for the case of the considered wind farms, we have used several

years of data to show that the European Centre for Medium-Range Weather Forecasts (ECMWF) performs well even for short lead times when the topography is smoothed. When it is not the case, a more sophisticated approach is required to obtain acceptable results.

After 80 min, the direct output of ECMWF forecasts gives better results than the classical time series based methods and improves persistence from 5.0% to 28.9% for one of the wind farms. Taking into account those good performances, we have considered parametric and non-parametric approaches to downscale the model outputs at the farm scale. With these downscaling models, we obtain improvements over persistence from 100 min and up to 25.4% at 170 min for Parc de la Vènerie.

In order to have better results for shorter lead times, we have corrected ECMWF forecasts by providing as explanatory variable the error between the forecasted wind speed and the initial measurement. This low cost assimilation lets the linear regression to overperform all other methods. The improvements over the traditional time series based models become significant with time, from 5.3% at 30 min to 30.1% at 170 min at Parc de Bonneval.

Under these conditions, the downscaling model $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ provides more accurate short term forecasts from 10 min to 3 h than the conventional models found in the literature. This is our starting point for the next step, which is the wind energy forecasts at the turbine scale. To do so, several effects have to be taken into account. These crucial steps are described in chapter 3.

3

From wind speed to wind power forecast

Contents

3.1 Introduction 50
3.1.1 Direct approach versus indirect approach 50
3.1.2 Power curve modeling for wind turbines 51
3.2 Consideration of wake effect on power output 53
3.2.1 Impact on wind power output 53
3.2.2 Consideration of the wake effect for wind energy modeling 55
3.2.3 Application to wind energy forecasts 56
3.3 Air density induced error on wind energy estimation 59
3.3.1 Air density error budget 61
3.3.2 Application to Parc de Bonneval 64
3.4 Impact of atmospheric conditions on power output 67
3.4.1 Wind shear 67
3.4.2 Turbulence 68
3.4.3 Atmospheric stability 69
3.5 Performances of wind power forecast 70
3.5.1 Statistical results 70
3.5.2 Forecasts post treatment 71
3.6 Conclusion 72

3.1 Introduction

In recent years, the wind energy sector has soared all over the world. Wind farms are located in more than 90 countries around the world. Nine of these countries have an installed capacity of more than 10 GW, and 30 with more than 1 GW across Europe, Asia, North America, Latin America, and Africa. In 2017, 52.5 GW of new wind power was installed across the globe, bringing total installed capacity up to 539 GW. In France, wind power installation increased by 14.04% in 2017 [45], mainly thanks to the feed-in tariffs. The French leading electricity utility company, Électricité de France (EDF), is under an obligation to purchase green electricity from wind producers for a period of 15 years. After this period, the producers have to sell their electricity on the competitive market. Every day a contract is established between the market and the producer about the quantity of electricity they will inject on the grid. This contract can be updated up to 30 min in advance. Any difference between the contract and the production will be compensated via penalties. This framework prompts the producers to have accurate short term forecasts.

3.1.1 Direct approach versus indirect approach

Concerning wind power forecasting, there are two ways to approach this problem. The direct approach and the indirect approach. The first one is to develop a forecasting model using historical wind power generation and then forecast the wind power directly.

The indirect approach is made in two steps. First, a model is developed to forecast the wind speed, and then this wind speed forecast is converted into wind power forecast by using different methods. This approach introduces an additional step by performing this conversion. Theoretically, the relationship between the wind speed v (in m s $^{-1}$), through swept area A of wind turbine (in m^2), and the wind power p (in W) is:

$$p = \frac{1}{2}\rho A v^3 \tag{3.1}$$

where ρ is the air density (in kg m^{-3}). From equation (3.1), it can be seen that the relationship between wind speed and wind power is cubic. Consequently, small errors in wind speed forecasts can generate significant errors in wind power forecasts. However, the indirect approach is usually less numerically costly, especially for large wind farms. Indeed, it can provide average wind speed forecasts at a specific site, and then these forecasts are converted into individual wind energy forecasts at the turbine scale. With this approach, it is easy to account for situations where turbines are off-line. It is also straightforward to develop the wind farm by adding new turbines. This is less the case with the direct approach. Either one model is required for each turbine, and it is numerically costly, or only one model is fitted to forecast the wind power output at the farm scale, and then the model cannot take into account the cases where one or several turbines are off-line.

In [46], Shi *et al.* compare the performance between direct and indirect ARIMA-based approaches for the wind energy forecasting of a 2 MW turbine 1 h ahead. They find that the direct approach provides more accurate forecasts compared with the indirect one. The NRMSE ranges from 12.6% (in percentage of the nominal power) for the indirect approach to 11.1% for the direct approach. The main reason is related to the power curve, which they use to convert the wind speed into wind energy in the indirect approach. Power curves are functions that give the power output depending on the wind speed only. They are provided by the turbine manufacturer but

give better results if computed using historical data. Power curves only consider the average deterministic relationship between wind speed and power generation, while in reality, the relationship is stochastic and depends on other variables such as air density. This neglected variability may lead to a lower accuracy in predicting wind power generation using the indirect approach. In [47], Hong *et al.* develop an indirect short term wind power forecast approach. They decompose the wind speed into a mean component, and a stochastic component. The mean component is forecasted using polynomial regression, and the stochastic part is forecasted using the SVM-based method. Then, they convert the wind speed forecast into a wind energy forecast using a computed power curve. For the proposed model, they show that indirect forecast performs better than direct power forecast. The NMAE is reduced from 4.1% to 3.8% (in percentage of the installed capacity). New methods are also developed to optimize forecasting by combining the two approaches. For instance, in [48], Bokde *et al.* present a new approach to eliminate the disadvantages of direct and indirect forecasting methods. Their method behaves like a direct-indirect hybrid that does not directly or indirectly predict power. It is based on clustering, and it uses both wind power and wind speed datasets as input to improve accuracy in wind power forecasting. Results reveal that the proposed methodology shows the best performance compared to both direct and indirect approaches.

The direct and indirect approaches have been tested at Parc de Bonneval to forecasts the average power output of the farm. For the indirect approach, we use a computed power curve. That is to say, a power curve fitted using historical wind speed measurements and power outputs. Figure 3.1 compares the performances of a direct $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ and an indirect $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ (defined in chapter 2). From 30 min, the indirect $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ overperforms the direct $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$. Further improvements are expected with better modeling of the power curve.

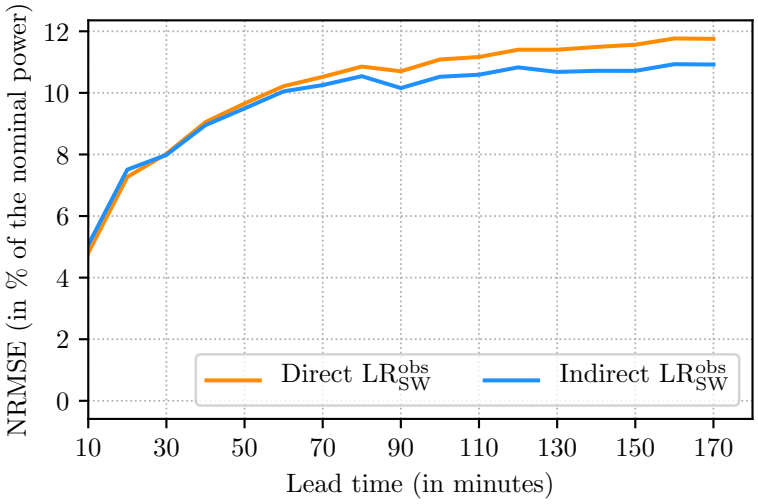


Figure 3.1 | Comparison between a direct $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ and an indirect $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ for the wind energy forecasting at Parc de Bonneval. This model is described in chapter 2. NRMSE in % of the nominal power is shown from $t_0 + 10$ min to $t_0 + 170$ min.

3.1.2 Power curve modeling for wind turbines

Power curve modeling is still a very active research field [49, 50]. Some studies aim at modeling power curves with non-parametric approaches such as smoothing splines [51], which consists of

modeling the power curve with functions defined piecewise by polynomials. Some other studies use neural networks [52]. For instance, in [53], Li *et al.* develop neural networks for each turbine to estimate the wind power output. They use as input, the wind speed and the wind direction. They find that the relative error is considerably reduced by the use of ANN, compared to an estimation using power curves. Monthly results are shown, and for one of the turbines, the error is decreased from 28.8% to 0.8%. Nevertheless, most of the studies focus on parametric approaches, numerically less expensive, such as polynomial, exponential, or cubic power curve approximation [54]. Studies have shown large sensitivity to the empirical estimation method of the wind power, with errors reaching 50% [55], and varying by about 20% between parametric and non-parametric approaches [56]. In this thesis, we model the power curve according to the standard methodology of The Standard International IEC 61400-12-1 [57]. The methodology consists in modeling the power curve by dividing the wind speed dataset into 0.5 m s^{-1} intervals, then the power curve is retrieved by fitting the means of each interval. To illustrate this methodology, we consider the wind speed measured at hub height by anemometers, and averaged over the turbines of the farm. The computed power curve is shown in figure 3.2.

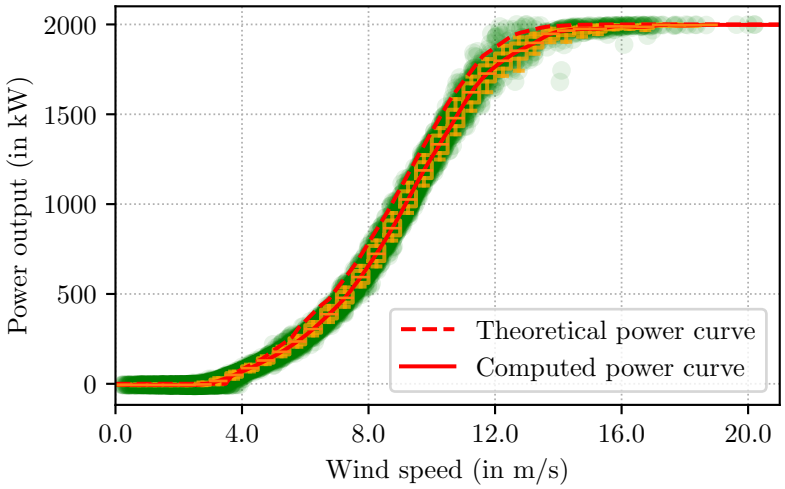


Figure 3.2 | Computed power curve for the turbines at Parc de Bonneval. For each 0.5 m s^{-1} interval, the boxplots of the distribution are shown in orange. The whiskers correspond to the first and the ninth deciles. The means of each interval are fitted in order to retrieve the power curve. The theoretical power curve provided by the turbine manufacturer is also shown by the dashed line.

Again, this power curve only takes into account the relationship between wind speed and power, while several variables impact the power output, such as, for instance, air density. Consequently, for a constant wind speed, the variability of the power output distribution is not negligible. Figure 3.3 illustrates the distribution of the wind energy modeling error. More precisely, it shows the relative error δ_{α_r} given by:

$$\delta_{\alpha_r} = \frac{\hat{y}_i - y_i}{|y_i|} \tag{3.2}$$

where, $\hat{y}_i$ is the i-th estimated power output and y_i is the corresponding measurements. Depending on the wind speed, the first and third quartiles are shown along with the median. First

of all, we can see that there is a tendency to underestimate the power, as more than 75% of the distribution is negative for wind speed around 5 m s^{-1}, and the median remains negative below 11 m s^{-1}. Moreover, the error decreases when the wind speed increases. With stronger wind speed, the power output increases, and as the error is normalized by the power output, δ_{α_r} decreases. The real challenge is about wind speeds between 6 m s^{-1} and 11 m s^{-1}, which correspond to the increasing part of the power curve and where the uncertainty on the power output is the highest. Indeed, for weaker wind speeds, the turbines will barely produce, and for stronger wind speeds, the turbines will reach their nominal power. Consequently, for those extremes, the uncertainty is very low. However, between them, the uncertainty is high. This relative error is inherent to the wind energy modeling. This is the minimum error that can be expected, and it can be referred as to $\varepsilon_{incompressible}$. To this $\varepsilon_{incompressible}$, will be added the forecasting error. Consequently, the lower this $\varepsilon_{incompressible}$, the better the forecast performance can be. In order to decrease it, other variables than wind speed must be taken into account

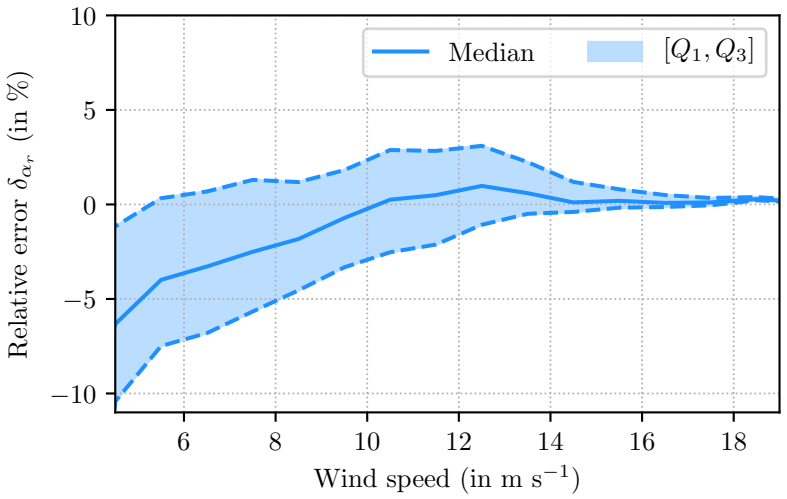


Figure 3.3 | Distribution of the relative error δ_{α_r} defined in equation (3.2). The median and the first and third quartiles (Q_1 and Q_3) are shown depending on the wind speed.

In this chapter, we focus only on Parc de Bonneval. The chapter is organized into five sections. The next section highlights the impact of the wind direction and, more specifically, the wake effect on the power output, and we propose a way to deal with it. Section 3.3 deals with air density, and its impact on wind energy estimation and section 3.4 is dedicated more broadly to atmospheric conditions such as wind shear, turbulence, and atmospheric stability. Finally, in section 3.5, the main steps of the implementation of an operational forecasting system are described. We conclude in section 3.6.

3.2 Consideration of wake effect on power output

3.2.1 Impact on wind power output

After the wind speed, wind direction is one of the most important variables that affects the wind power output. Figure 3.4 displays the characteristics of the wind direction at Parc de Bonneval.

Panel (a) displays the wind direction frequencies, and panel (b) displays the average wind speed depending on wind direction. The wind speed is measured by anemometers located behind the hub. Most of the time, the wind comes from the southwest, which corresponds to the prevailing winds. Moreover, inside this sector, we find the highest winds (around 7 m s^{-1}). More precisely, the sector with the highest frequency is [230°; 235°], and it corresponds to the sector with the highest average wind speed of 7.8 m s^{-1}.

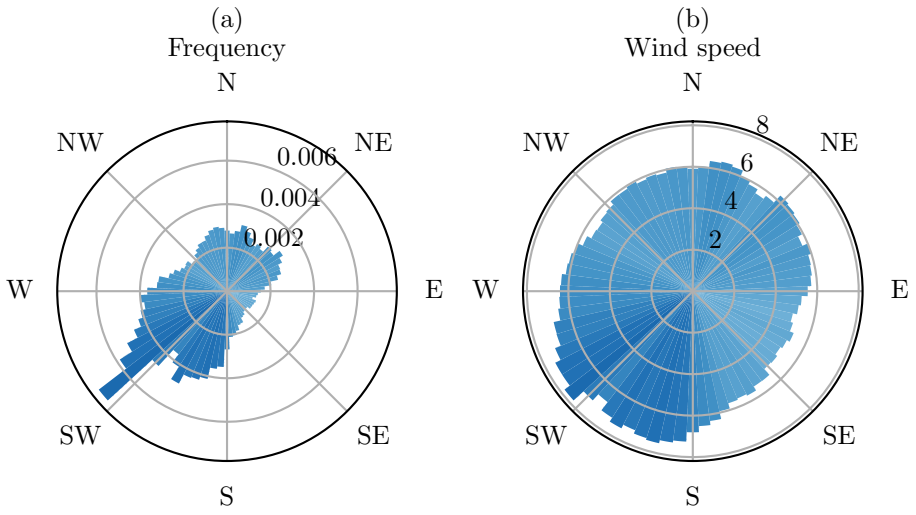


Figure 3.4 | Characteristic of the wind direction at Parc de Bonneval. Panel (a) displays the wind direction frequency and panel (b) displays the average wind speed depending on wind direction. In both cases, sectors of five degrees are considered.

Unfortunately, at Parc de Bonneval, the southwest sector is the most sensitive to the wake effect. As shown in figure 3.5, Parc de Bonneval is composed of six turbines arranged in two lines.

One line is composed of two turbines denoted E1 and E2, at almost 400 m away, and the second line is composed of four turbines denoted E3, E4, E5, and E6 with a distance between 510 m and 525 m between each of them. In the direction of prevailing winds, it is ideal to space the wind turbines between three and nine times the rotor diameter. Because the rotor of the wind turbines is 80 m, the wind turbines must be at least 240 m away and up to 720 m away. They are, therefore, in the right interval but may still be subject to the wake effect. We can see that both lines are lined up under northeast winds and southwest winds. More precisely, the angle formed between the first line (resp. the second line) and the south-north direction is denoted θ_1 (resp. θ_2). $\theta_1 = 72°$, consequently when the wind direction is around 72° or 252°, turbines E1 and E2 are lined up. $\theta_2 = 51°$, so E3, E4, E5, and E6 are lined up for wind directions around 51° and 231°. Prevailing winds are southwest, and they are twice as frequent as other winds (see figure 3.4).

In these conditions, the upstream turbines decrease the flow for the downstream turbines, and consequently, these downstream turbines produce less energy than the upstream turbines for the same wind speed. This is *the wake effect*. Figure 3.6 illustrates this phenomena. It shows time series of wind direction (a), average wind speed measured at the wind farm (b), and power output from the six turbines at Parc de Bonneval (c) for December 22^{nd} 2016 from 05:30 to 10:00 UTC. The red rectangles highlight a period when E3, E4, E5, and E6 are lined up. Between 06:30 and 09:10, the wind direction is around 230°. First of all, we can see that the power output of E1 and E2 is the same. Those two turbines are not lined up. The power output is lower than E3 because

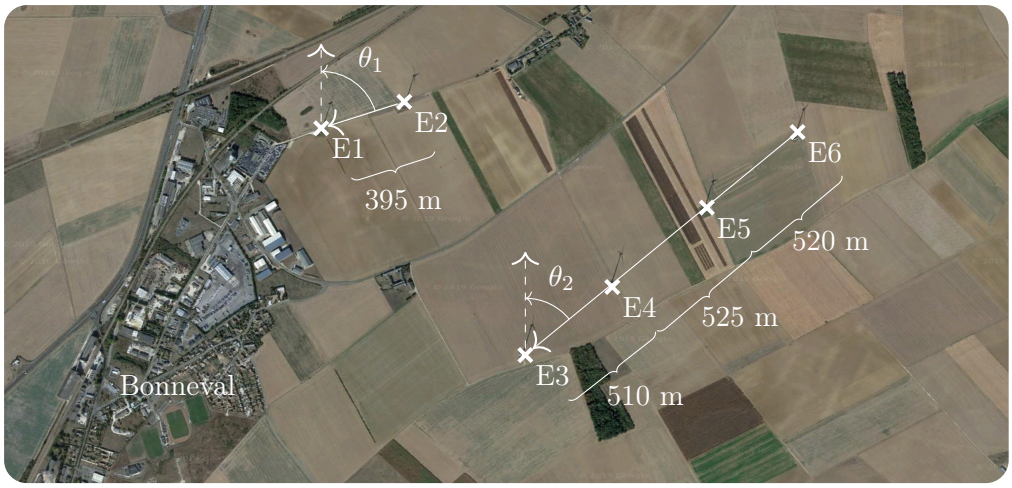


Figure 3.5 | Satellite image of Parc de Bonneval extracted from Google Earth. The white crosses display the turbine location with their respective number. Two lines are visible. One between E1 and E2 (the angle formed between E1-E2 and the south-north direction is denoted θ_1. The second line is formed by the turbines 3, 4, 5 and 6 (the angle between these turbines and the south-north direction is denoted θ_2). The distance between the turbines is added.

E1 and E2 are affected by Bonneval for the prevailing winds. The power output for the second line is proportional to their position. The upstream turbine, E3, is the turbine with the highest power output. Come after E4, E5, and finally E6, which is the most downstream turbine. During this event, around 08:10, E6 produces 48% less than E3 as the power output for E3 is 548 kW while the power output for E6 is 283 kW. This decrease is even more important when E1 and E2 are lined up as the two turbines are closer to each other.

In these conditions, the wake effect can not be neglected as it would lead to a significant overestimation of the wind power output.

3.2.2 Consideration of the wake effect for wind energy modeling

It is crucial to take wind direction into account when estimating and forecasting wind energy. It allows to avoid the overestimation due to the wake effect. To do so, we define S_i, where $i \in [\![1;6]\!]$ is the turbine index as:

- $S_1 = \{\theta_1 \pm 15°\} = [57°; 87°]$
- $S_2 = \{(\theta_1 + 180°) \pm 15°\} = [237°; 257°]$
- $S_3 = \{\theta_2 \pm 15°\} = [36°; 66°]$
- $S_4 = \{\theta_2 \pm 15°\} \cup \{(\theta_2 + 180°) \pm 15°\} = [36°; 66°] \cup [216°; 246°]$
- $S_5 = S_4 = [36°; 66°] \cup [216°; 246°]$
- $S_6 = \{(\theta_2 + 180°) \pm 15°\} = [216°; 246°]$

Those sectors are *the wake sectors*. For each turbine, we compute a power curve using datasets split according to the wake sectors. That is to say, one power curve for wind directions inside the

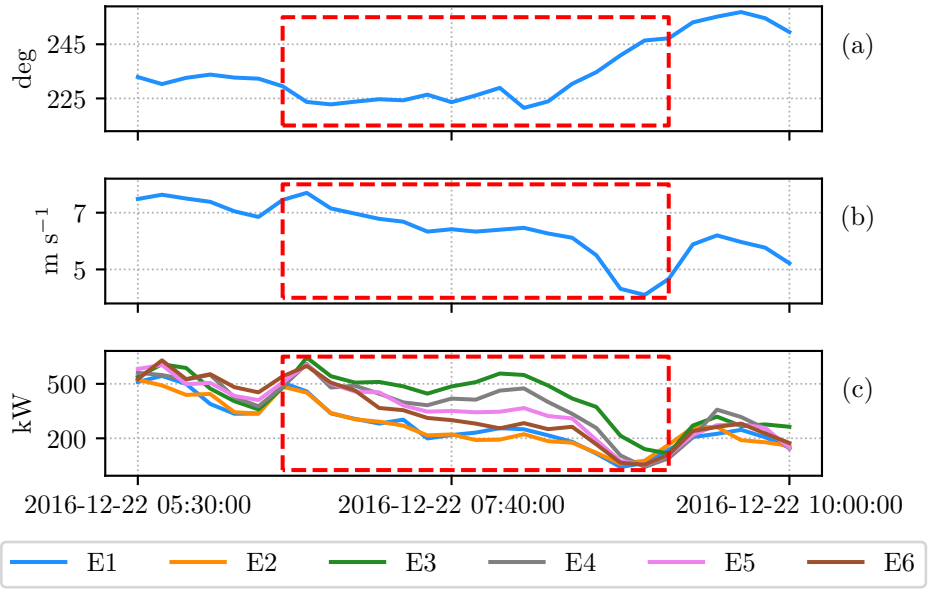


Figure 3.6 | Time series for December 22nd 2016 from 05:30 to 10:00. It displays wind direction at Parc de Bonneval location (a), average wind speed at the farm location (b) and the power output for the six turbines of the farm. The red rectangles highlight the period when the turbines are lined up.

wake sector and one power curve for wind directions outside the wake sector. Figure 3.7 displays the two power curves for each turbine. The difference between the two curves is significant for E1 and E2 only. Indeed, those two turbines are closer to each other than the turbine E3, E4, E5, and E6. Consequently, the wake effect is stronger for E1 and E2. Even if the difference between the two curves seems negligible for E3, E4, E5, and E6. Figure 3.6 shows that the impact on wind energy is not.

Table 3.1 shows the MAE and NRMSE on the wind energy modeling when the wind direction $\in S_i$. The case where only one power curve is used is compared with the case where two power curves are used (that is to say, the case where the power curve is computed using dataset where direction $\in S_i$). The results are shown for the six turbines. For each turbine, the use of a power curve fitted using data in the wake sector only, improves the wind energy modeling in terms of both MAE and NRMSE. The greatest improvements are found for E1 and E2 as these two turbines are the closest to each other. The improvements for E2 are around 41.7% for the MAE and around 39.0% for the NRMSE. It is above 45% for E1 for both MAE and NRMSE. Even for E3, E4, E5, and E6, the improvements are not negligible as they are around 38% for the MAE and 36% for the NRMSE.

3.2.3 Application to wind energy forecasts

With regard to forecasts, the difficulty lies in estimating the wind direction θ over time in order to determine if θ is inside or outside the wake sector. Most of the time, the wind direction is forecasted in the same way as the wind speed. For instance, in [58], Erdem *et al.* use ARMA based method to forecast the tuple of wind speed, and direction 1 h ahead. In [59], Khosravi *et al.*

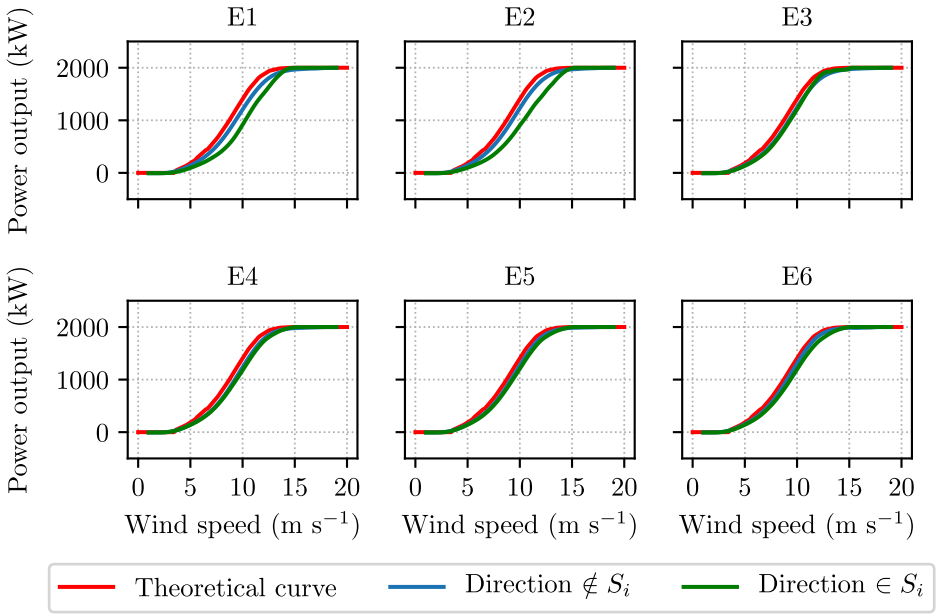


Figure 3.7 | For each turbine of Parc de Bonneval, two power curves are computed. When inside the wake sector (direction $\in S_i$) and one outside (direction $\notin S_i$). The theoretical curve is also added. This is the same for each turbine.

	Number of power curve used	MAE (in %)	NRMSE (in %)
E1	1	7.1	10.9
	2	3.7	6.0
E2	1	7.4	10.7
	2	4.3	6.5
E3	1	6.2	9.5
	2	3.8	5.9
E4	1	5.2	8.1
	2	3.2	5.2
E5	1	4.8	7.5
	2	3.1	4.9
E6	1	4.9	7.6
	2	3.1	4.9

Table 3.1 | Comparison between the case where only one power curve is used for all situation (Number of used power curve = 1) and when two power curves are fitted (Number of used power curve = 2: one for direction $\in S_i$ and one for direction $\notin S_i$). The MAE and NRMSE (in % of the nominal power) between the modeled power and the measured power are shown for the six turbines of Parc de Bonneval.

investigate several non-parametric approaches, such as ANN, SVM, ANFIS, to forecast the wind direction.

In our case, in addition to ARMA and ANN, we also test to forecast the wind direction using

the persistence model and ECMWF forecasts. As shown in the chapter 2, section 2.4, the wind components at 100 m height (u100 and v100) are forecasted by ECMWF. We can use the components to compute the wind direction according to equation (3.3), or we can consider the last measurement at Parc de Bonneval.

$$\begin{cases} u &= V\cos(\theta) \\ v &= V\sin(\theta) \end{cases} \iff \tan(\theta) = \frac{v}{u} \iff \theta = \arctan\left(\frac{v}{u}\right), \quad \theta \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[\tag{3.3}$$

where V is the wind speed: $V = \sqrt{u^2 + v^2}$.

Figure 3.8 displays the mean absolute error (MAE, defined in the next section, in equationation (3.5)b) on the wind direction forecasting depending on the lead time. ECMWF are hourly forecasts, and then the minimum errors are found every full hour and a slight increase due to the interpolation of the components are notice between full hours. In practice, no interpolation should be done, but the closest forecast should be used. On average, the MAE for ECMWF is constant with time, and it is around 20°. When the persistence method is used, the error starts from 8° at 10 min to 30° at 170 min. For ARMA and ANN, the shape is the same as for persistence (increasing with time) but with larger error. Thus, it is better to take, as estimator, the persistence for the first hour and then use the wind components forecasted by ECMWF to estimate the wind direction after the first hour.

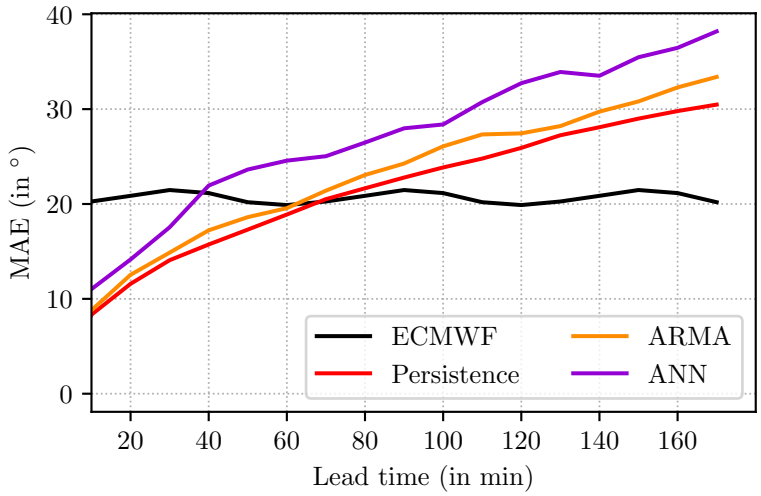


Figure 3.8 | MAE for the wind direction estimation from 10 min to 170 min. Results for ECMWF, persistence, ARMA and ANN are shown.

Figure 3.9 shows the NRMSE in % of the nominal power for each turbine at Parc de Bonneval between forecasted power and measured power. The wind power forecast is provided by $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$, defined in chapter 2. To compute the NRMSE, only the situations where the wind direction is inside the wake sector are considered. For E1, it occurs 5.4% of the time. For E2, it occurs 12.4% of the time. It occurs 6.0% of the time for E3, 13.7% for E4 and E5, and 7% of the time for E6. Consequently, the size of the dataset used to compute the NRMSE is different for each turbine (except for E3 and E4). For each turbine, the use of a specific power curve fitted using data inside the wake sector improves the results compared to the case where only one power curve is fitted.

Again, the differences are more significant for E1 and E2, and few improvements are visible for the turbines of the second line.

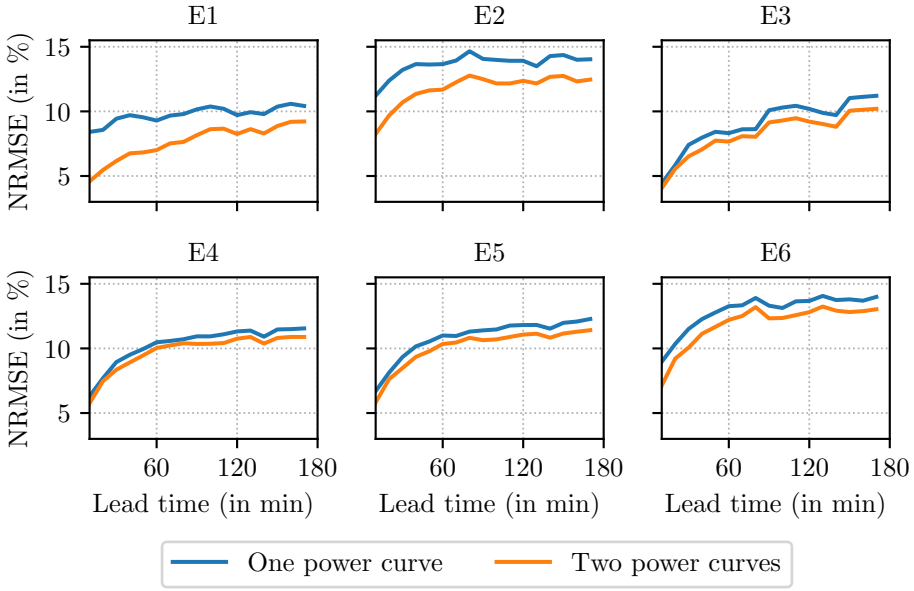


Figure 3.9 | NRMSE (in % of the nominal power) for each turbine of Parc de Bonneval between the forecasted power (provided by $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$) and the measured power from 10 min to 170 min. Results when only one power curve and when two power curves are fitted (one for directions $\in S_i$ and one for direction $\notin S_i$) are shown.

Figure 3.10 displays the improvements from 10 min to 170 min between the case where only one power curve is fitted and the case where two power curves are fitted (one for directions $\in S_i$ and one for direction $\notin S_i$). In other words, it displays the relative difference between the blue curves and the orange curves in figure 3.9. Clearly, the largest improvements are found for E1 and E2 for the whole period with E1 clearly in the lead with an improvement of 52.3% at 10 min and of 11.5% at 170 min. Even for the less affected turbine, E3, the improvements are between 9.7% and 4.9% during the whole period. Generally, improvements decrease with lead time mainly due to the error in the wind direction estimation. At 10 min, the errors in the wind direction leading to the use of the wrong power curve (power curve inside the wake sector while the real direction is outside and inversely) occurs around 4% of the time. It is around 11% of the time for the last lead time. If we could avoid these errors in the wind direction forecasting, the improvements due to the use of two power curve would be around 14% at 170 min.

In any case, even with this uncertainty in the wind direction forecasting, taking into account the wind direction leads to significant improvements for all lead times and for all turbines.

3.3 Air density induced error on wind energy estimation

As seen in the previous section, most of the time, wind energy is computed from the wind speed through a power curve. The theoretical power curve is provided by the wind turbine manufacturer for standard temperature ($T_0 = 15°\mathrm{C}$) and pressure ($P_0 = 1013.25$ hPa). Deriving empirical power

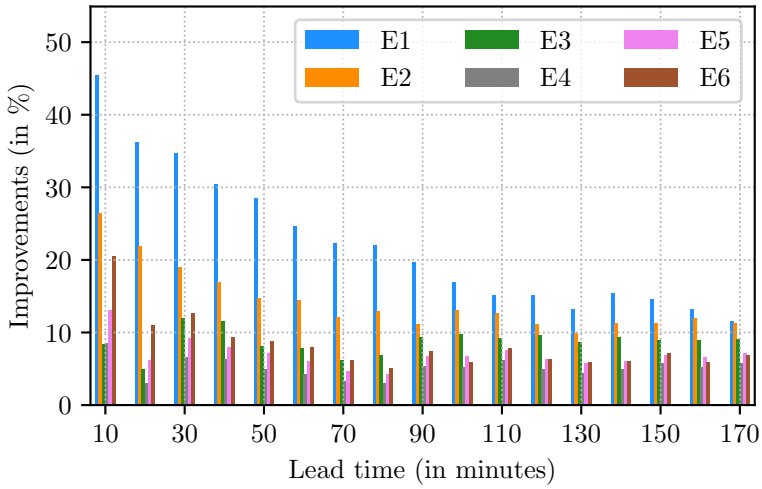


Figure 3.10 | Improvements from 10 min to 170 min between the case where only one power curve is fitted and the case where two power curves are fitted (one for directions $\in S_i$ and one for direction $\notin S_i$). For each lead time, results for the six turbines of Parc de Bonneval are shown.

curve for a given turbine is a key for more accurate wind power estimate and as seen in 3.1.2 there is no lack of method.

However, the power performance of a wind turbine also depends on air density. But most studies neglect it [60]. Its impact is not negligible with an error on wind power estimate, which can be reduced by 20% when temperature correction for air density is accounted for [61]. However, as for the power curve, the sensitivity to the methods used to correct for air density is extremely large, with errors varying by more than 100% depending on the method [62].

An accurate estimate of air density is, therefore, a key to reduce the uncertainty of the wind power forecast. In an operational configuration, various strategies can be adopted to achieve this. Considering default values is clearly the worst strategy, and it is equivalent to ignoring air density variations. The best strategy requires real time temperature and pressure measurements for an a priori empirical derivation of the power curves and an a posteriori method for debiasing locally wind power forecasts. However, wind farms equipped with both sensors are rare. In that case, values from Numerical Weather Prediction models may be a suitable alternative.

This study aims at underscoring the temperature and pressure contributions of the air density computation in order to better take into account the lack of wind farm instrumentation. Section 3.3.1 details the methodology to compute the error budget of the air density with an in-depth analysis of the temperature and pressure contributions. The error budget analysis is performed at a densely instrumented site, and its spatial pattern and sensitivity to the terrain complexity is further investigated using meteorological analysis. Application to wind power estimation is shown at an actual wind farm in section 3.3.2.

3.3.1 Air density error budget

At the SIRTA observatory

To quantify the contributions of temperature and pressure in the air density error budget, we use the large observation dataset from the SIRTA observatory (Site Instrumental de Recherche par Télédétection Atmosphérique), located 20 km South of Paris (France) (48.7°N, 2.2°E, 150 m altitude) [63]. We retrieve surface pressure and temperature at 2 m at 10-minutes frequency from 2015 to 2017 to compute the air density and the different contributions. We compute the air density ρ from the temperature T and pressure P based on the ideal gas law $P = \rho \frac{R}{M} T$ as:

$$\rho = \frac{MP}{RT} = \frac{M}{R}\frac{(P_0 + P')}{(T_0 + T')} = \underbrace{\frac{MP_0}{RT_0}}_{\rho_0}\left(1 + \frac{P'}{P_0}\right)\left(\frac{1}{1 + T'/T_0}\right) \tag{3.4}$$

where $P_0 = 1013.25$ hPa and $T_0 = 288.15$ K are reference values of the standard atmosphere at the Earth's surface. The quantitites P' and T' are the deviations to the reference values, $M = 0.02898$ kg mol^{-1} is the dry air molar mass and $R = 8.31$ J K^{-1} mol^{-1} is the ideal gas constant. To quantify the contributions of the temperature and pressure to the air density error budget, we compute the normalized bias (BIAS), the normalized mean absolute error (MAE) and the normalized root mean square error (NRMSE) as:

$$\text{BIAS} := \frac{1}{\bar{y}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)\right) \tag{3.5a}$$

$$\text{MAE} := \frac{1}{\bar{y}}\left(\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|\right) \tag{3.5b}$$

$$\text{NRMSE} := \frac{1}{\bar{y}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.5c}$$

where, y_i is a measured variable (air density), $\hat{y}_i$ is the computed variable, n is the sample size and $\bar{y}$ is the mean value over the period ranging between 2015 and 2017.

Table 3.2 displays the values of BIAS, MAE, and NRMSE between the measured and computed air density. Two ways of computing it are assessed. The first row corresponds to the values when the pressure is set to its reference value ($P = P_0$). Only temperature deviation is considered (hereafter referred to as "temperature contribution"). The second row corresponds to the values when the temperature is set to its reference value ($T = T_0$). Only pressure deviation is considered (hereafter referred to as "pressure contribution"). The two contributions are evaluated separately because a wind farm may only have access to pressure or temperature measurements. When $P = P_0$ (temperature contribution, upper row), the BIAS and MAE have very similar absolute value (1.34% and 1.37% with respect to the reference density), suggesting a significant negative bias. The NRMSE is of the same order of magnitude. When $T = T_0$ (pressure contribution, lower row), the bias is positive (+1.22% with respect to the reference density). The MAE and NRMSE are more than 1.5 times larger than when $P = P_0$, suggesting that the temperature contribution has a larger weight in the air density error budget. Indeed, despite the averaged relative fluctuations $\frac{P'_{\text{OBS}}}{P_0}$ and $\frac{T'_{\text{OBS}}}{T_0}$ have the same order of magnitude (around 10 hPa over 1000 hPa for the pressure

and 3 K over 300 K for the temperature), the relative standard deviation $\frac{\sigma_P}{P_0}$ and $\frac{\sigma_T}{T_0}$ is around 8.66×10^{-3} and 2.42×10^{-2} respectively. The larger temperature variability causes a larger impact of temperature on the air density error budget.

	BIAS (in %)	MAE (in %)	NRMSE (in %)
$\tilde{\rho} = \rho_0\Big(\frac{1}{1 + T'_{\mathrm{OBS}}/T_0}\Big)$	-1.34	1.37	1.61
$\tilde{\rho} = \rho_0\Big(1 + \frac{P'_{\mathrm{OBS}}}{P_0}\Big)$	1.22	2.22	2.72

Table 3.2 | Bias (BIAS), mean absolute error (MAE) and normalized root mean square error (NRMSE) for air density when the pressure is set to a reference value ($P = P_0 = 1013.25$ hPa) (temperature contribution, upper row) and when the temperature is set to a reference value ($T = T_0 = 288.15$ K) (pressure contribution, lower row). The data used to compute the error indicators are measurements collected at SIRTA observatory, located 20 km South of Paris (France) (48.7°N, 2.2°E, 150 m altitude).

However, wind farm operators often lack simultaneous real time temperature and/or pressure measurements at hub height, to compute air density and correct the wind power output accordingly. Meteorological reanalysis, analyses, or even short term forecasts are supposed to be the best 3 D representation of the state of the atmosphere at a given time. We use here the temperature and pressure at 2 m from ERA5 reanalysis to test the added value of the NWP model output when local measurements are missing. ERA5 are reanalysis dataset provided by the European Center for Medium-Range Weather Forecasts (ECMWF). ERA5 provides hourly estimates of a large number of atmospheric, land, and oceanic climate variables. The data cover the Earth on a 30 km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80 km. The grid point nearest to SIRTA (48.75°N, 2.25°E) is located less than 7 km away. In order to have the same time resolution, we compute hourly averaged of the 10-minutes measurements. Table 3.3 displays the error indicators BIAS, MAE, and NRMSE computed by comparing the air density measured at the SIRTA observatory, with that estimated from the temperature and pressure from ERA5 reanalysis. The results are very comparable to those of table 3.2 (see middle and lower rows). Surprisingly, the errors with ERA5 are slightly lower than the errors with the measurements. This can be explained by the chosen reference values, which minimize the deviations in the case of the reanalysis. Indeed, reanalysis generally have less amplitude because they have more difficulty capturing the extremes.

In case only one variable is measured (temperature or pressure), table 3.4 displays the BIAS, MAE and NRMSE using ERA5 for the missing variable. For instance, if the temperature is measured (T_{OBS}) and not the pressure, then the pressure from the model output (P_{NWP}) is used. Conversely, if the pressure is measured (P_{OBS}) and not the temperature, then the temperature from the model output (T_{NWP}) is used. All error indicators (BIAS, MAE and NRMSE) are lower compared to those computed by discarding the missing variable, should it be measured (table 3.2) or obtained from model output (table 3.3).

Spatial pattern

In this section, the impact of both contributions is investigated across France. In both cases, the reference air density is computed from model outputs. Figure 3.11 displays the NRMSE of pressure

	BIAS (in %)	MAE (in %)	NRMSE (in %)
$\tilde{\rho} = \rho_0\Big(\frac{1}{1+T'_{\mathrm{NWP}}/T_0}\Big)\Big(1+\frac{P'_{\mathrm{NWP}}}{P_0}\Big)$	-0.44	0.47	0.58
$\tilde{\rho} = \rho_0\Big(\frac{1}{1+T'_{\mathrm{NWP}}/T_0}\Big)$	-1.29	1.33	1.57
$\tilde{\rho} = \rho_0\Big(1+\frac{P'_{\mathrm{NWP}}}{P_0}\Big)$	0.74	2.06	2.53

Table 3.3 | Same as table 3.2 with data from ERA5 reanalysis at the grid point nearest to SIRTA observatory. The additional upper row compares the air density computed from ERA5 data with the measured air density.

	BIAS (in %)	MAE (in %)	NRMSE (in %)
$\tilde{\rho} = \rho_0\Big(\frac{1}{1+T'_{\mathrm{OBS}}/T_0}\Big)\Big(1+\frac{P'_{\mathrm{NWP}}}{P_0}\Big)$	-0.49	0.49	0.50
$\tilde{\rho} = \rho_0\Big(\frac{1}{1+T'_{\mathrm{NWP}}/T_0}\Big)\Big(1+\frac{P'_{\mathrm{OBS}}}{P_0}\Big)$	0.05	0.27	0.36

Table 3.4 | Same as table 3.2 when measurements at SIRTA observatory and data from ERA5 reanalysis at the grid point nearest to SIRTA observatory are combined to compute the air density.

contribution (left column) and temperature contribution (right column) over France. All data are retrieved from ERA5 reanalysis. The errors are low for most parts of France ($< 4\%$ when $T = T_0$ and $< 5\%$ when $P = P_0$) except in the mountains (up to 6% when $T = T_0$ and up to 30% when $P = P_0$). This is due to the reference values $P_0 = 1013.25$ hPa and $T_0 = 288.15$ K. These are approximations that are not valid anymore at high altitudes.

To overcome this problem, we compute the reference temperature and the pressure, corrected with the altitude according to the International Standard Atmosphere (ISA) as [64] :

$$\tilde{P_0} = P_0\left(1 - \frac{0.0065}{T_0}z\right)^{5.255} \tag{3.6a}$$

$$\tilde{T_0} = T_0 - \frac{6.5}{1000}z \tag{3.6b}$$

with z, the altitude in meters.

After correction, the errors do not exceed 1% in most of France (1.2% in the Alps) when $P = P_0$. When $T = T_0$, the NRMSE is lower than 2.5% except in the Alps where it is around 3.0% as shown in figure 3.12, while it was around 6.0% when $T = T_0$ and around 30% when $P = P_0$ (see figure 3.11). Considering a constant value of temperature introduces larger errors than considering a constant value of pressure. Again, this is due to the higher variability of the temperature. Two different patterns can be distinguished on figure 3.12 depending on the contribution. When the temperature variations are neglected (panel (a)), the errors above sea are very low, around 1%, while the errors above land are around 2.5%. On land, the temperature variations are more important than above the sea. Neglected, it can introduce higher error. When we neglect the

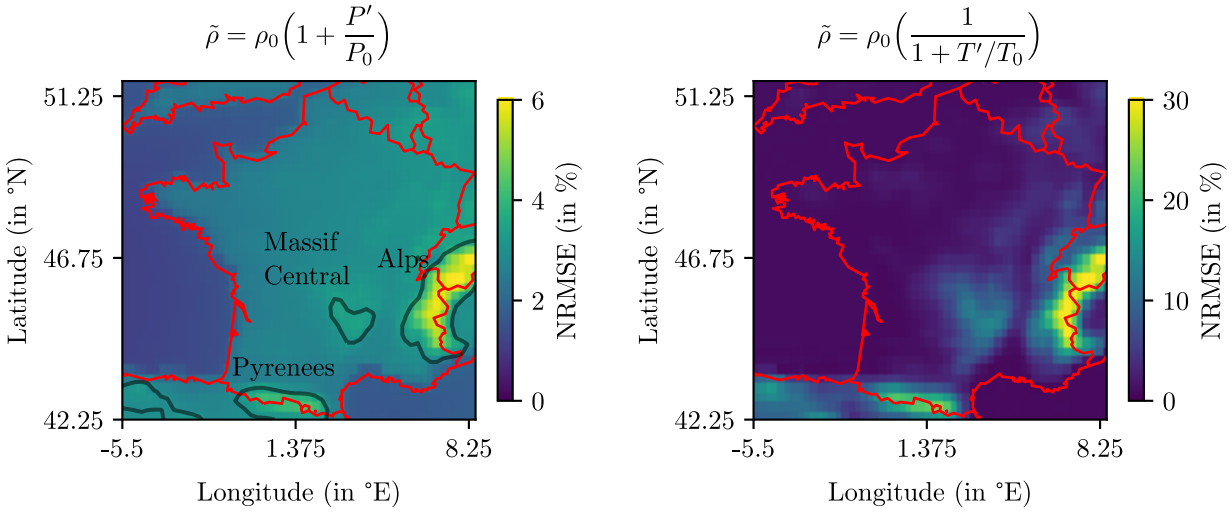


Figure 3.11 | The left column displays the error when $T = T_0$ and the right column displays the error when $P = P_0$. Both figures display the NRMSE in %. The data are retrieved from ERA5 reanalysis.

pressure variations (panel (b)), there is no difference between land and sea, but the errors are proportional to the latitude. Higher errors are found in the north than in the south. This is due to the more frequent storms and the passage of depressions in the north of France.

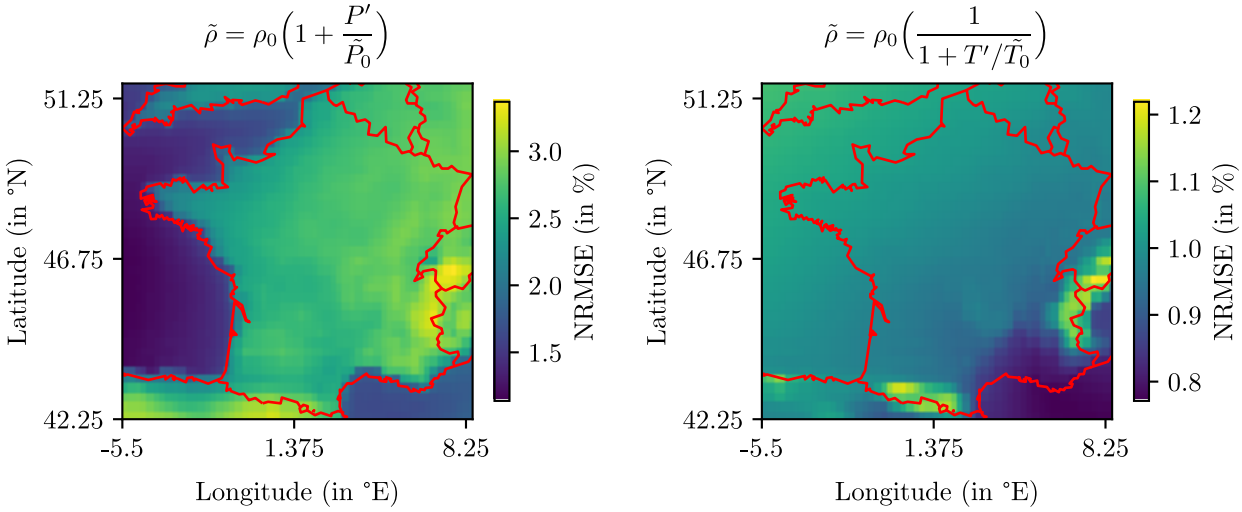


Figure 3.12 | Same as figure 3.11 but in this case the temperature and pressure are corrected according to equation (3.6).

3.3.2 Application to Parc de Bonneval

We consider the data of the Parc de Bonneval (48.20°N, 1.42°E, 135 m altitude). For a reminder, the data are 10-minutes averaged. For consistency, the dataset analyzed here has been collected between 2015 and 2017. The power curve is retrieved by averaging the wind speed and power

measurements at the six turbines, as shown in figure 3.2. The wind speed measurements are binned into 0.5 m s^{-1} intervals. The wind power is averaged in each bin, and the power curve is retrieved by fitting the mean wind power as a function of the mean wind speed.

According to [57], to take into account the air density, the wind speed must be normalized as follows:

$$U_n = U_t \left(\frac{\rho_t}{\rho_0} \right)^{1/3} \tag{3.7}$$

where $\rho_0 = 1.225$ kg m^{-3} is the standard density for which the power curve is given by the wind turbine manufacturer. Applying equation 3.7 is an efficient way to account for the air density. It is less costly than training a dedicated model, and it has already been used for energy potential evaluation [65]. As the power curve is given as a function of the wind speed only, and a reference density, it is necessary to incorporate the density variations in the value of the wind. Figure 3.13 displays different power curves fitted for several temperature intervals. In figure 3.13a, the wind speed is directly taken from the measurements: $U_n = U_t$. For a wind speed of 6.5 m s^{-1}, the power output varies from 356 kW for a temperature between 0°C and 5°C to 298 kW for a temperature between 30°C and 35°C (i.e., 19.5% difference). Figure 3.13b shows how the normalization, given by equation (3.7), corrects the effect of the air density variations. The temperature is retrieved from measurements (at hub height) and the pressure (surface pressure) from ERA5 reanalysis nearest grid point (48.25°N, 1.5°E) around 8 km from Parc de Bonneval. For a wind speed of 6.5 m s^{-1}, the power output varies after air density correction from 346 kW, for a temperature between 0°C and 5°C, to 320 kW for a temperature between 30°C and 35°C (i.e., 8.1% difference). The spread between the power curves is, therefore, highly reduced.

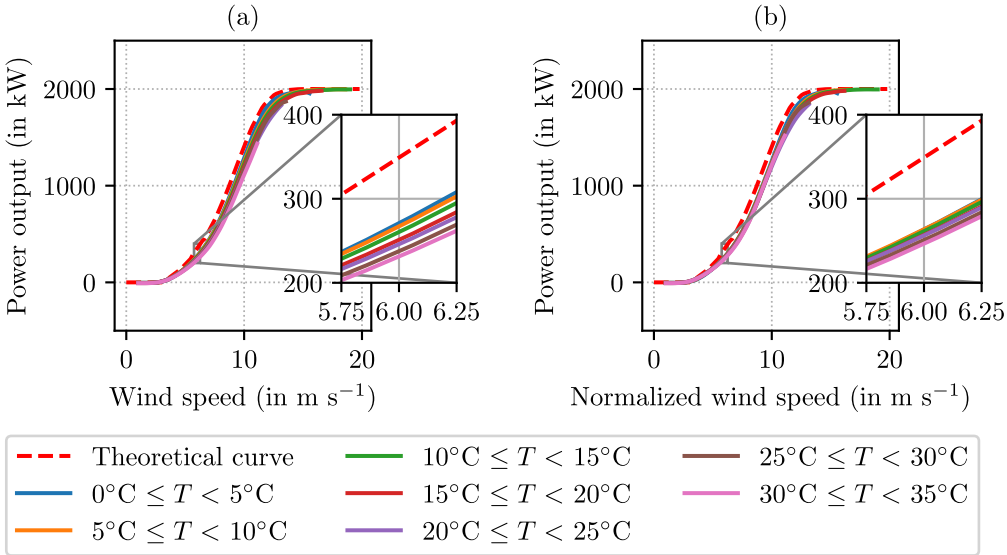


Figure 3.13 | Power curves averaged over the six wind turbines of Parc de Bonneval as a function of temperature ranges, when the wind speed is not corrected for air density variation ($U_n = U_t$) (a) and when it is corrected following equation (3.7) (b). The theoretical power curve provided by the manufacturer is shown in dashed red.

Table 3.5 displays BIAS, MAE and NRMSE between the measured and modeled wind power for the measured wind speed ($U_n = U_t$) and normalized wind speed using equation (3.7). The

normalization is here applied with respect to the nominal power, equal to 2 MW. One can first note that the bias is close to 0. The negligible bias can be explained by the fact that, on average, at this location, the temperature and pressure conditions are close to the reference values. Correcting for air density variation, reduces MAE and NRMSE, as they are indicators quantifying the spread, which is reduced (figure 3.13). However, the improvement is significant but remains low as MAE goes from 0.96% (no normalization) to 0.77% (normalization), and NRMSE from 1.58% (no normalization) to 1.32% (normalization). As explained earlier, the temperature and pressure conditions are close to the reference values (i.e., the mean temperature for this period is around 13°C, and the mean pressure is around 1000 hPa) so the averaged improvement is weak [66].

	BIAS (in %)	MAE (in %)	NRMSE (in %)
No normalization : $U_n = U_t$	0.03	0.96	1.58
$U_n = U_t\left(\left(1+\frac{P'}{P_0}\right)\left(\frac{1}{1+T'/T_0}\right)\right)^{1/3}$	0.02	0.77	1.32

Table 3.5 | BIAS, MAE and NRMSE between the measured and modeled wind power for the measured wind speed ($U_n = U_t$) and normalized wind speed using equation (3.7).

Focusing on more extreme conditions such as temperatures below 5°C or higher than 25°C, improves significantly the impact of the air density correction. Table 3.6 summarizes these results. Compared to table 3.5, the differences between the normalized and measured wind speeds are larger. For instance, the MAE improves by about 20% for all cases (table 3.5) to about 33% for extreme cases only (table 3.6, $T \geq 25$°C). Similarly, NRMSE improves by about 17% in all cases (table 3.5) to about 37% for extreme cases only (table 3.6, $T \geq 25$°C). Those extreme events are not so rare because cold temperatures lower than 5°C (mainly winter months) occur 10.7% of the time and hot temperatures higher than 25°C occur (mainly summer months) 5.5% of the time.

		BIAS (in %)	MAE (in %)	NRMSE (in %)
$T \leq 5$°C	$U_n = U_t$	1.03	1.38	2.12
	$U_n = U_t\left(\left(1+\frac{P'}{P_0}\right)\left(\frac{1}{1+T'/T_0}\right)\right)^{1/3}$	0.31	0.92	1.55
$T \geq 25$°C	$U_n = U_t$	-0.98	1.20	1.89
	$U_n = U_t\left(\left(1+\frac{P'}{P_0}\right)\left(\frac{1}{1+T'/T_0}\right)\right)^{1/3}$	-0.26	0.73	1.19

Table 3.6 | BIAS, MAE and NRMSE between the measured and modeled wind power for temperatures lower than 5°C and higher than 25°C. Comparison between measured wind speed ($U_n = U_t$) and normalized wind speed (normalization using equation (3.7)) is shown.

3.4 Impact of atmospheric conditions on power output

Besides air density and wind direction, which are known for significantly impacting wind turbines performances, some other meteorological variables affect the wind energy output. Among them, we can mention the wind shear, the atmospheric stability, or the turbulence [67].

3.4.1 Wind shear

The wind shear is the variation of wind speed over vertical distance. The more the diameter of the turbines increases, the more important it is to take into account the wind shear [68]. A classical method for determining the wind shear is the wind profile power law. The wind speed at height z can be extrapolated thanks to wind speed measured at height z_0 as in equation (3.8).

$$U_z = U_{z_0}\left(\frac{z}{z_0}\right)^{\alpha} \tag{3.8}$$

where U_z is the wind speed that must be extrapolated at height z, U_{z_0} is the measured wind speed at height z_0 and α is the Hellman's exponent or the shear exponent which depend on the surface roughness. However, the shear exponent can be estimated to quantify the wind shear between the height of z_1 and z_2. From equation (3.8) the shear exponent can be estimated as follow:

$$U_{z_1} = U_{z_2}\left(\frac{z_1}{z_2}\right)^{\alpha} \iff \log\left(\frac{U_{z_1}}{U_{z_2}}\right) = \log\left(\left(\frac{z_1}{z_2}\right)^{\alpha}\right) \iff \alpha = \frac{\log\left(\frac{U_{z_1}}{U_{z_2}}\right)}{\log\left(\frac{z_1}{z_2}\right)} \tag{3.9}$$

Then, to estimate the impact of wind shear on power output, we compute the shear exponent using ECMWF output between 10 m height and 100 m height. The optimal solution would be to compute it between the bottom and top of the rotor using a lidar. Then the wind shear dataset is split into several intervals. First, we considere the low wind shear ($\alpha < 0.1$), then the medium wind shear ($0.1 \leq \alpha < 0.2$) and finally the high wind shear ($0.2 \leq \alpha$). For each bin, the power curve is computed as well as the standard deviation σ_α of the power output in the percentage of the nominal power.

Figure 3.14 displays the different power curve (a) and the evolution of σ_α with the wind speed (b). In both cases, they are computed for each wind shear interval.

The variability among the power curves is not significant. They are computed by fitting the means of each 0.5 m s^{-1} bins, which indicates that the distribution of each bin is symmetric. If we focus on the variability using the standard deviation σ_α, we can see that the maximum deviation is found for wind speed around 10 m s^{-1}, at least for wind shear above 0.1. For low wind shear, only few data were available as it occurs 7.3% of the time. For medium and high wind shear, the standard deviations are close. It is a little bit higher for high wind shear than for medium wind shear, and we can notice small peaks for high wind shear for wind speeds above 12 m s^{-1}. For low wind shear, the standard deviation is weaker than for medium and high wind shears, but there is no data for wind speed above 8 m s^{-1}, and even for wind speeds below, there is very few data. This can be explained by the fact that the wind shear is computed between 10 m height and 100 m height. According to equation (3.8), the wind speed increases very quickly in the first few meters, that is why the wind shear between 10 m and 100 m is most of the time medium or high.

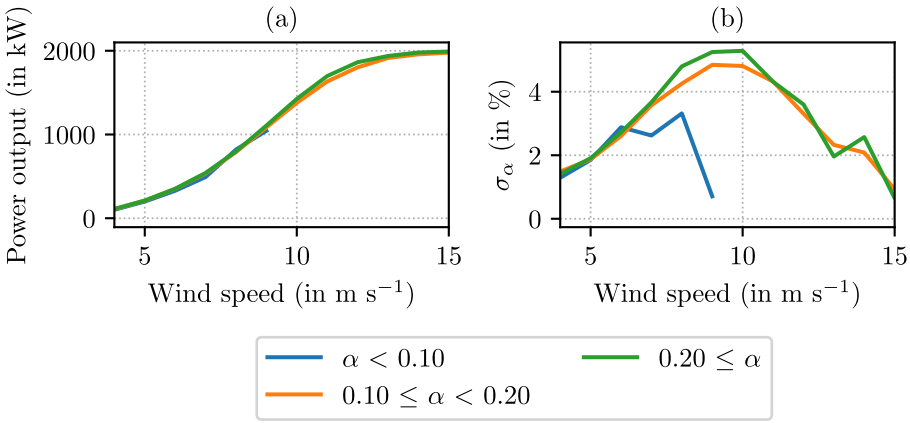


Figure 3.14 | Power curves as a function of shear exponent (3.9) ranges (a). Panel (b) displays the standard deviation σ_α (in % of the nominal power) depending on wind speed, as a function of shear exponent ranges.

Computing the wind shear, from the bottom to the top of the rotor using lidar data, would be a good solution to better estimate the impact of the wind shear on the wind power output. The latter is known to have a significant impact [69].

To take the wind shear into account in the wind power modeling, a simple solution is to compute the average wind speed over the rotor instead of using the measurements at the hub height.

3.4.2 Turbulence

Turbulence is also an important variable that not negligibly influences the power output, as see in [70]. To estimate this influence, the methodology is the same as for the wind shear in the previous section. We compute the turbulence intensity defined in equation (3.10), to split the wind speed and power output sample.

$$\mathrm{TI} = \frac{\sigma_U}{U} \tag{3.10}$$

where U is the 10-minutes average wind speed measurements, and σ_U is the standard deviation of the high-frequency measurements inside the 10 min interval. The turbulence intensity is split in three intervals. Low turbulence ($\mathrm{TI} < 5\%$), medium turbulence ($5\% \leq \mathrm{TI} < 10\%$) and high turbulence ($\mathrm{TI} \leq 10\%$). Figure 3.15 displays the power curves and the standard deviations σ_{TI} computed for each interval of turbulence intensity, and as a function of the wind speed.

More variability should be noted than in figure 3.14. There are again few data for low turbulences. First of all, the measurements are collected at 100 m, and second of all, the anemometer is located on the nacelle behind the rotor. Then, even if an algorithm is provided by the manufacturer to correct the wind speed, there are probably some parts of the rotor induced turbulence which is not smoothed. With regard to the standard deviation, the variations between medium and high turbulence are visible. For wind speed above 13 m s^{-1}, the standard deviation is almost zero for medium turbulence while it is around 1.26% for high turbulence.

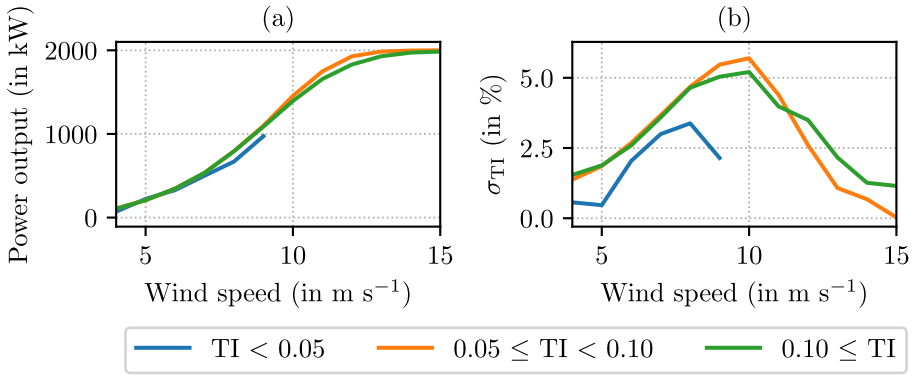


Figure 3.15 | Same as figure 3.14 for the turbulence. The power curve (a) and the standard deviation (b) are computed as a function of the turbulence intensity ranges (3.10).

3.4.3 Atmospheric stability

The atmospheric stability can be estimated through the lapse rate $-\dfrac{dT}{dh}$, where dT is the temperature variation with altitude and dh is the altitude variation. Depending on the value of the lapse rate, the stability of the atmosphere can be deduced as follow:

$$\begin{cases} -\dfrac{dT}{dh} < 6 \text{ K km}^{-1} & \Longrightarrow \quad \text{stable atmospheric conditions} \\ 6 \text{ K km}^{-1} \leq -\dfrac{dT}{dh} < 10 \text{ K km}^{-1} & \Longrightarrow \quad \text{conditionally unstable atmospheric conditions} \\ 10 \text{ K km}^{-1} \leq -\dfrac{dT}{dh} & \Longrightarrow \quad \text{unstable atmospheric conditions} \end{cases} \tag{3.11}$$

The lapse rate is computed between the temperature at 2 m and the temperature measured on the nacelle at 100 m. Figure 3.16 displays the power curves (panel (a)) and the standard deviation $\sigma_{-dT/dh}$ (panel (b)) according to the atmospheric stability.

Again, there is no significant difference between the power curves. However, the standard deviation of the power output depends much more on the atmospheric stability. For stable conditions, the standard deviation is smooth, with a maximum for wind speed around 10 m s^{-1}. For conditionally unstable atmospheric conditions, the standard deviation is less smooth, especially for wind speed above 12 m s^{-1}. Finally, for unstable atmospheric conditions, the standard deviation is significantly higher than for stable and conditionally stable conditions, at least for wind speed above 10 m s^{-1}.

To take into account turbulence and atmospheric stability, a wind normalization can be applied [71] as follow:

$$U_n = U_t \left(1 + 3 \left(\frac{\sigma_U}{U} \right)^2 \right)^{1/3} \tag{3.12}$$

where U_n is the corrected wind, U_t is the measured wind, $\dfrac{\sigma_U}{U}$ is the turbulence intensity TI defined in equation (3.10). In [60], Wagenaar *et al.* found a difference up to 7% between corrected and uncorrected wind.

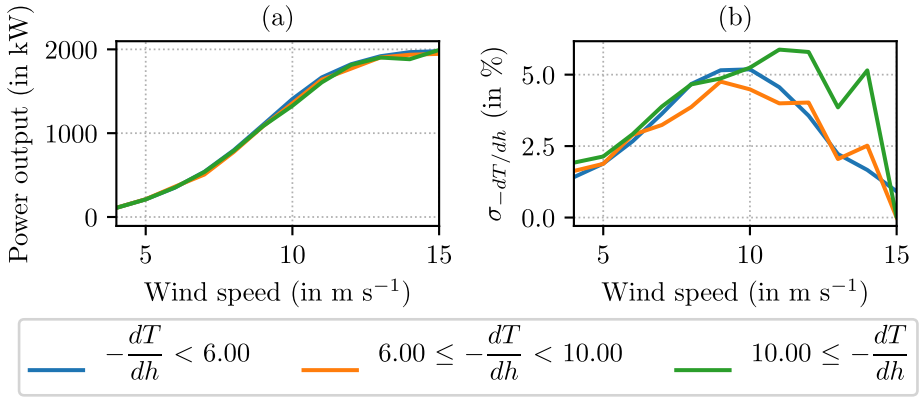


Figure 3.16 | Same as figures 3.14 and 3.15 for atmospheric stability. The power curve (a) and the standard deviation (b) are computed as a function of the atmospheric conidtions stability (3.11).

3.5 Performances of wind power forecast

We used $\text{LR}_{\text{SW}}^{\text{obs}}$, the best model shown in chapter 2, to predict the wind speed from 10 min to 170 min, and then we use the computed power curves that take into account the wind direction and the air density, as shown in sections 3.2 and 3.3.

3.5.1 Statistical results

Figure 3.17 shows the performances of $\text{LR}_{\text{SW}}^{\text{obs}}$ with persistence, $\text{LR}_{\text{SW}}^{\text{no-obs}}$, ECMWF, ARMA, and ANN. Each model forecasts the wind speed, and then the power output is retrieved through the power curve. In these conditions, the hierarchy between the models is respected. $\text{LR}_{\text{SW}}^{\text{obs}}$ is the best model for each lead time, and it is the only model that beats persistence from the first lead time.

Figure 3.18 displays the improvements over persistence for the models cited before. As said previously, only $\text{LR}_{\text{SW}}^{\text{obs}}$ improves persistence from around 1.2% at 10 min to 32.44% at 170 min.

Taking into account the wind direction and the air density allows significant improvements over the year. Figure 3.19 displays those improvements for each turbine depending on the lead times. These improvements are computed over the entire year, that is to say, even when the air density is very close to its standard value and when the wind direction is outside the wake sectors.

Even when not only favorable cases are considered, we found improvements up to 7.1% at 10 min and up to 1.3% at 170 min. The smallest improvements are found for E3, which is very rarely impacted by the wake effect. The highest improvements are found for E1, E2, and E6. The last two are downstream for the prevailing winds, and E1 is also significantly impacted due to the short distance with E2. With regard to air density correction, the improvements are the same for each turbine. In general, the improvements decrease with time due to the errors in the wind direction, temperature, and pressure estimation.

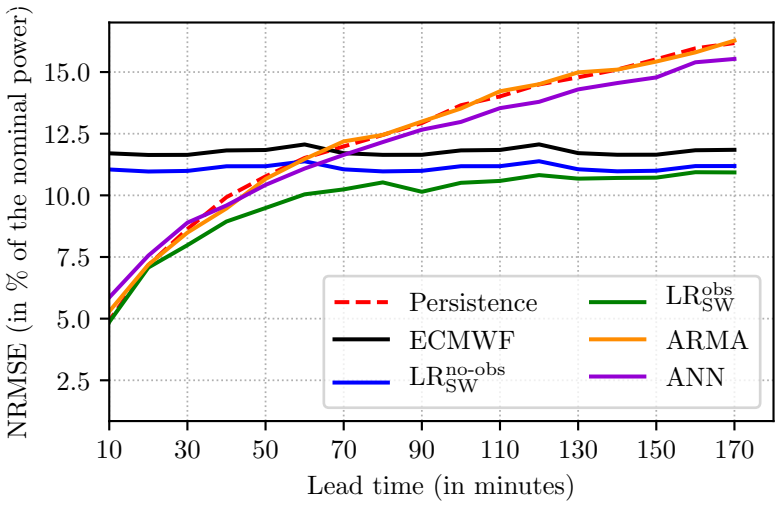


Figure 3.17 | Performances of the different models for sub-hourly forecasting of wind power output from 10 min to 170 min in two configurations (section 2.2.1) against the performances of ECMWF and the benchmark methods (section 2.2.3). The models are exactly the same than chapter 2. The power curve is computed according to the methodology described in section 3.1.2 and it is used to retrieve the power output. The NRMSE is normalized by the nominal power (2000 kW).

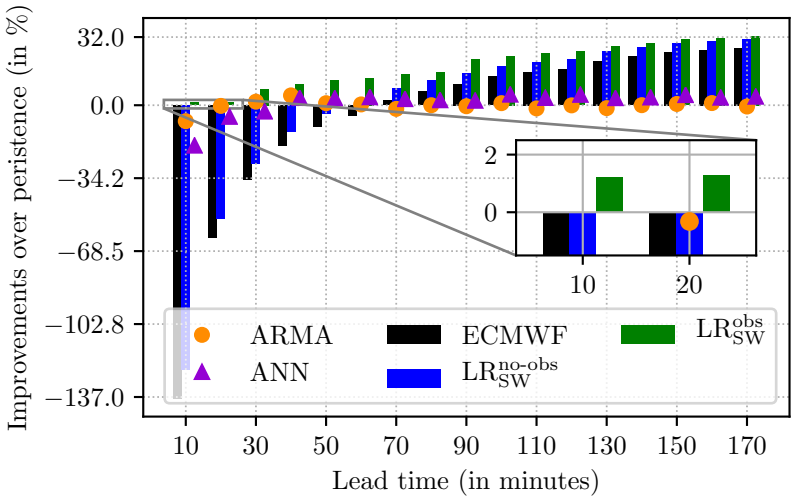


Figure 3.18 | Comparison of the improvements over persistence in percentage for ECMWF forecasts and the best downscaling models: LR$_{SW}^{obs}$ and LR$_{SW}^{no-obs}$ from 10 min to 170 min. Improvements of ECMWF, ARMA, and ANN methods are also included. In every case, the downscaling model is used to forecast the wind speed, and the wind power is retrieved using the power curve.

3.5.2 Forecasts post treatment

Uncertainty and decision making process

Such a nowcasting method should be used for the decision-making process. Therefore, a statistical quantification of the performances is not enough to evaluate the usefulness of the method. Fig-

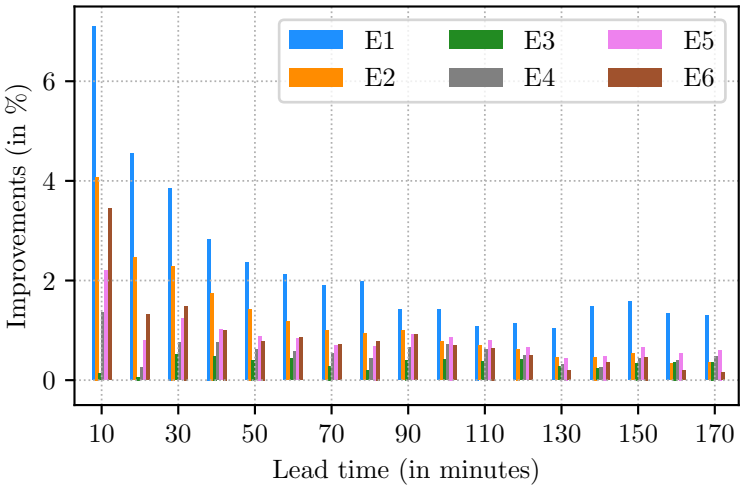


Figure 3.19 | Improvements from 10 min to 170 min for each turbine, between the case when air density and wind direction are taking into account and the case when they are neglected. For each lead time, results for the six turbines of Parc de Bonneval are shown.

ure 3.20 displays forecasted time series, starting from the 10^{th} of December 2019 at 21:00 UTC. A wind power prediction is shown using $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ and a computed power curve. The measurements and confidence intervals are also included.

Three confidence intervals are shown. Each of them is defined depending on the lead time and on the predicted wind speed. For each lead time, we consider wind speed bins of 1 m s^{-1}. For each prediction we compute the difference : $\hat{y}_t - y_t$, where $\widehat{y}_t$ is the forecasted wind speed by $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$, at time t and y_t is the measured wind speed at time t. Those differences are stored in the corresponding bin, depending on $\hat{y}_t$ and t. Using the data of the years 2015 and 2016, we compute for each couple of lead time/wind speed bin, distributions of the error. We compute, for each couple, three intervals: the 10% confidence interval, the 25% confidence interval, and the 50% confidence interval. Then the power output intervals are computed through the power curve.

Smoothing

In order to smooth the forecast due to models changes described in figure 2.13, in chapter 2, a relaxation is set up. For two consecutive lead times, provided by two different models, a correction is done. These lead times are a linear combination of the forecasts provided by the two different models. A higher weight is given to the original model. It allows to reduce the gap due to the change of models as shown in figure 3.21. This leads to an improvement up to 3% for the relevant lead times.

3.6 Conclusion

This chapter highlights the difficulty of converting wind speed forecasts into wind power forecasts. Several external effects without negligible impact on the wind power estimation have to be taken into account. In this chapter, we underscore some of them.

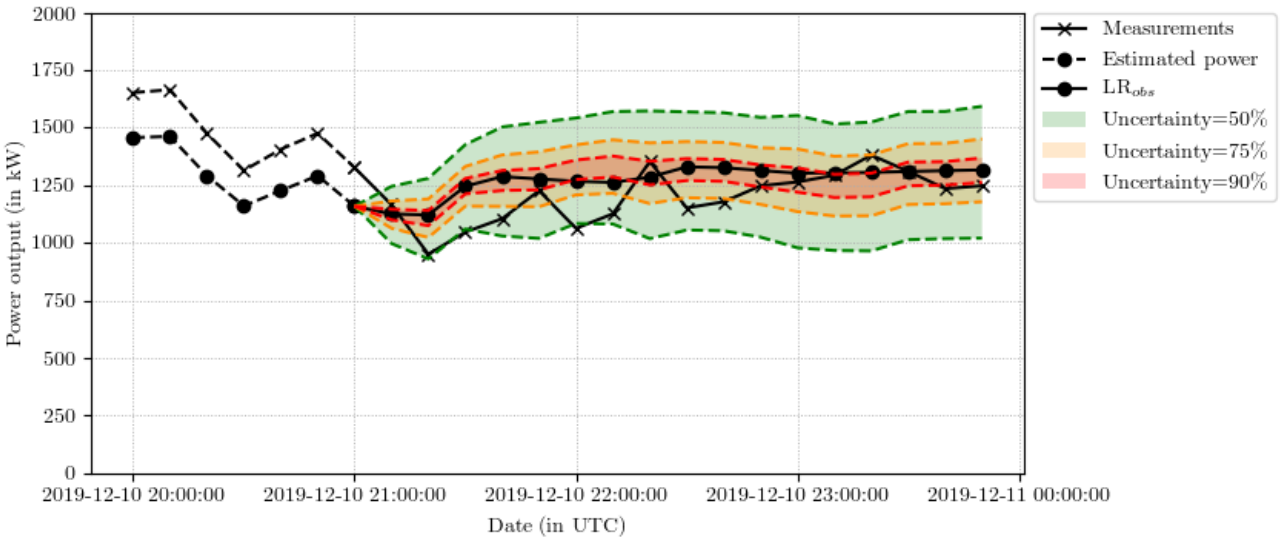


Figure 3.20 | Forecasted time series of wind power starting from the 10^{th} of December 2019 at 21:00 UTC. $\text{LR}_{\text{SW}}^{\text{obs}}$ forecasts are compared with the measurements. First, the wind speed is forecasted by $\text{LR}_{\text{SW}}^{\text{obs}}$, and then the forecasted power is retrieved using the power curve. The 10%, 25%, and 50% confidence intervals are added.

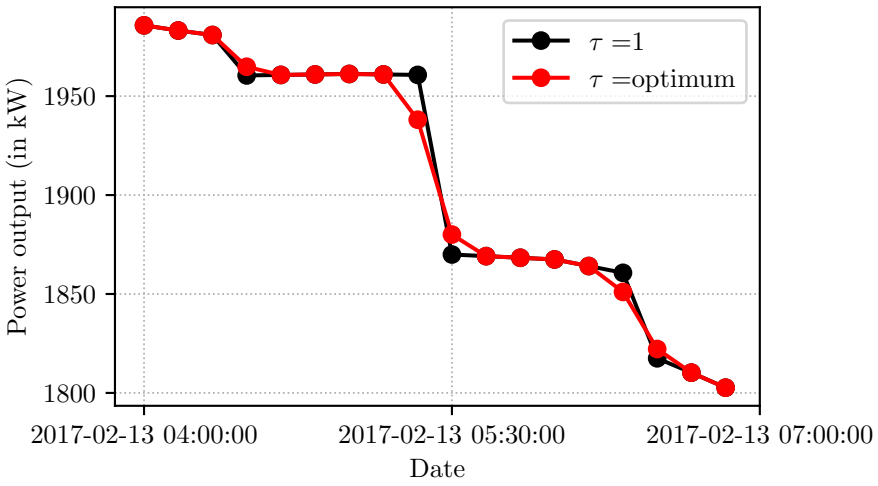


Figure 3.21 | Zoom on a wind power forecasted time serie with and without relaxation to smooth the models change.

First of all, we quantify the impact of wind direction on the power output, and we propose a way to take it into account in the wind power modeling. With turbines between 400 m and 500 m apart from each other, Parc de Bonneval is affected by the wake effect. For the same wind speed, under the southwest or northeast winds, the downstream turbines may produce up to 50% less than the upstream turbines. To deal with this issue, we compute two different power curves for each turbine. One using data for which the turbine is affected by the wake effect, and one with data for which the turbine is not. By doing so, the wind power modeling of the downstream turbines has

been improved up to 48% in terms of MAE and up to 45% in terms of NRMSE. Regarding the forecasts, the main difficulty is to estimate the wind direction over time. For the first lead time, a significant error in this estimation occurs around 4% of the time, while it occurs up to 11% of the time after 170 min. Despite these misestimations, the use of a specific power curve when the turbines are lined up leads to significant improvements in the wind power forecast for every lead times. The highest improvements are found for E1 from 45% at 10 min to 11% at 170 min. Even for E3, which is the least affected turbine, the improvements are between 4% and 9%.

Secondly, we assessed the adding value of the wind normalization to take into account the air density in the wind power modeling using in-situ measurements and meteorological analysis. In state of the art, most of the papers that take into account the air density use non-parametric methods. Those methods are numerically costly. Our parametric method overcomes this issue with also significantly improved results with respect to results that do not account for air density correction. Indeed, this study shows that a correction for air density improves the wind power estimation by more than 15% over the three investigated years (2015 to 2017). In most of the papers that deal with this normalization, there are no skill scores of the improvements due to the normalization but for instance, visual comparison between power curves. Moreover, when skill scores are given, they are given without distinction of the atmospheric conditions and without comparison with the case for which the air density is not considered. The lack of interest in this issue lies in the fact that the overall improvement remains limited, especially in mid-latitudes, where atmospheric conditions are close to the standards. In this study, the usefulness of the air density correction is highlighted by enhancing the situations where atmospheric conditions are far from the standard conditions, and the improvement reaches nearly 40% in those cases (temperatures below 5°C or above 25°C). This study also shows that the temperature is the key variable to account for when correcting for air density as its impact is the largest on the uncertainty of the air density estimation (twice larger than the pressure term). Meteorological analysis (i.e., model-based observations) also have a beneficial impact when one of the key variable (temperature or pressure) or even both variables are not measured. Correction for altitude using a standard atmosphere is the simplest and most efficient way to correct for air density when finer information is not available.

Taking into account the density and the direction decreases the uncertainty and the variability of power output modeling and then of power output forecast. Figure 3.22 shows the reduction of the $\varepsilon_{incompressible}$ which corresponds of the relative error δ_{α_r} defined in equation (3.2) in section 3.1.2. In blue the δ_{α_r} shown in figure 3.3 which correspond to the relative error before accounting for external effects. Orange curves correspond to the relative error after direction and density correction. Taking into account air density and wind direction reduces the median, which is closer to 0. Moreover it reduces the interquartile range (IQR $= Q_3 - Q_1$). For the critical wind speeds between 5 m s^{-1} and 11 m s^{-1} the IQR is reduced between 6% and 34%

In addition to wind direction and air density, some other variables affect wind power output. In future works, it would be worth investigating more deeply their impact on wind power output and solutions to take them into account effectively. Among them, we can mention the wind shear, turbulence, or atmospheric stability.

In order to illustrate the model performances, a case study for a specific time is shown. The wind power forecasted time series is presented. The associated confidence intervals are also displayed. We choose to add the 50%, 25%, and 10% confidence intervals because their range, from 0.20 m s^{-1} to 1.5 m s^{-1}, correspond to the appropriate accuracy for wind energy producers. For instance, a 90% confidence would have been statistically better but too large to be useful.

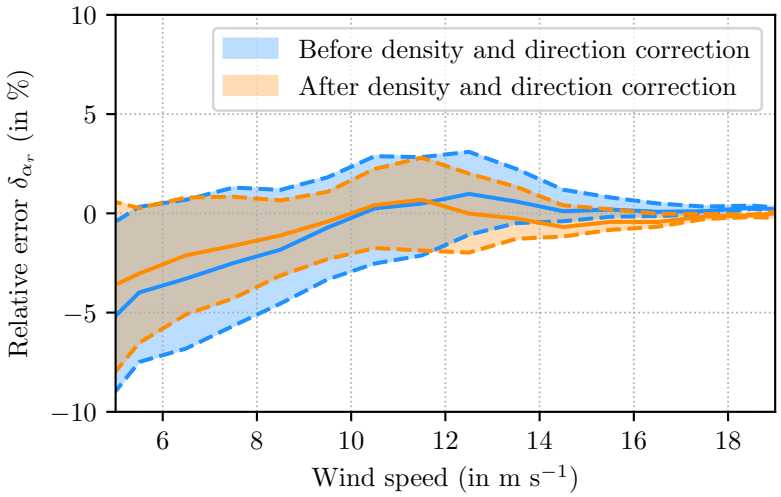


Figure 3.22 | Distribution of the relative error δ_{α_r} defined in equation (3.2) after density and direction correction (orange curves). The median and the first and third quartiles (Q_1 and Q_3) are shown depending on the wind speed. The results of δ_{α_r} before density and direction correction are added (blue curves).

4

Added value of networking wind farms

Contents

4.1 Introduction 78
4.2 Added value of small scale information 79
4.2.1 Wind farms location and specificity 79
4.2.2 Improvement of the average wind speed forecast 80
4.2.3 Improvements of the wind turbine downscaling 83
4.3 Added value of large scale information 88
4.3.1 Wind farms location and correlation 88
4.3.2 Application to forecasts 91
4.4 Conclusion 92

4.1 Introduction

In the attempt to implement the best possible forecasting model, having different wind farms at disposal at varying distances can be an advantage.

Indeed the value of spatial information in wind speed and wind energy forecasting has been thoroughly investigated. Mainly because it is known that changes in the wind might propagate with the wind, and it is possible to use upwind observations to detect the precursors to changes in wind speed at the site of interest. To do so, some simple method might be used. For instance, methods based on spatial correlation have been developed for some time now. In [72], Alexiadis *et al.* build a forecasting model based on cross correlation at neighboring site. They use artificial neural networks to forecast the wind speed from few minutes to several hours ahead. For the shortest lead times, they considered as input of the ANN, measurements at the site of interest, and also at three neighboring sites, from 800 m to 2.7 km. For longer lead times, they considered two distant sites, one at 12 km and one at 39 km. In both cases, the models lead to average errors around 20-40% better than the persistence approach. Another method is introduced in [73]. This is the regime-switching space–time (RST). It consists of identifying atmospheric regimes at the site of interest and fitting conditional predictive models for each regime. For instance, in [73], Gneiting *et al.* distinguish a westerly and an easterly forecast regime based on the wind direction. They use wind speed measurements from meteorological towers located at 39 km and at 146 km from their site of interest. For the 2 h ahead forecast, they find an improvement of the RMSE, for the RST method over persistence, up to 28% for July 2003. In [74], Hering *et al.* improve the RST method introduced in [73] by treating wind direction as a circular variable and including it in the model. Doing that, they improve the forecasts up to 3% in terms of RMSE and up to 4% in terms of MAE, compared to the classical RST method. Based on the previous examples, the addition of spatial information clearly seems to be valuable to forecast the wind speed. However, in [75], Kretzschmar *et al.* refute the use of off-site observations for forecasts of wind speed in Switzerland. In their paper, they evaluate the quality of artificial neural networks for wind speed prediction from 1 h ahead to 24 h ahead. They point out that upwind may refer to distinct geographic locations depending on the atmospheric regime. This last remark highlights one of the difficulties of wind farm networking: global models (statistical or not) that must manage non-systematic situations. Indeed, with respect to statistical models, they are trained over a reasonably long period of time. During this period, many situations occur depending on different atmospheric conditions, and the model must exhibit a general way to express the link between these atmospheric conditions and the targeted variable. If building too many models is counterproductive, a balance must be struck between performance and efficiency.

With regard to Zephyr, they own six wind farms in the northwest quarter of France. The prevailing winds are southwestern, consequently for this chapter, we focus on the forecasting at the two most easterly farms, which are Parc de Bonneval and Moulin de Pierre. They are both located 100 km Southwest of Paris and 5 km from each other. The performance of short term forecasting models for both Parc de Bonneval ans Moulin de Pierre are shown in chapter 2. Given the distance between the farms, it seems to be the optimal configuration for the short term forecast and especially for our first lead times of interest (10 min and 20 min). Two other wind farms, called Parc de la Vènerie and Parc de la Renardière, are around 200 km west of Parc de Bonneval and Moulin de Pierre. Those two farms can be useful for longer lead times ($>$ 2-3 h).

Consequently, in this chapter, we first explore in section 4.2 the added value of small scale information using data from Parc de Bonneval and Moulin de Pierre. We distinguished two regimes based on the wind direction. When the wind comes from Parc de Bonneval to Moulin de Pierre

(wind direction is in [183°,253°]) and when it does not (wind direction is not in [183°,253°]). This small scale information is first used in order to improve the average wind speed forecasts and then in order to improve the downscaling of the wind speed forecast at the turbine scale. In both cases, we focus only on the 10 min lead time and on the 20 min lead time. Each model used has been fully described in chapter 2. In section 4.3, the added value of the large scale information is investigated using data collected at Parc de la Vènerie and Parc de la Renardière. In this section, we explore the large scale information impact on the whole period, from 10 min to 170 min. Finally, we conclude in section 4.4.

4.2 Added value of small scale information

4.2.1 Wind farms location and specificity

Parc de Bonneval and Moulin de Pierre are built around 5 km away from each other. Figure 4.1 shows the location of the two farms and the typical distances between them. The distance between the two nearest wind turbines is around 3.8 km, and the distance between the two most distant is around 6.6 km. Moreover, the average wind speed at Parc de Bonneval is around 6.3 m s^{-1}, and the average wind speed at Moulin de Pierre is around 6.2 m s^{-1}. At this speed, it takes between 10 min, 30 sec, and 18 min, 20 sec to travel the distance between the two farms. In these conditions, one farm can carry information about the other for the first two lead times. Moreover, the angle formed between the two farms and the south-north direction θ is $\theta = 38°$, so Moulin de Pierre may carry information for Parc de Bonneval when the wind direction is around 38°, and conversely, Parc de Bonneval may carry information for Moulin de Pierre when the wind direction is around 218°.

Moulin de Pierre was commissioned at the end of 2016. In this study, we used the data of 2017 as a training period and the data of 2018 as a testing period. In both cases, wind speed and wind direction are collected using anemometers located at the top of the nacelle of each turbine.

There is a bias between Parc de Bonneval and Moulin de Pierre in the wind direction measurements of about 14° and the correlation is around 0.70. However, with regard to the wind speed, the two datasets are very similar. The correlation is up to 0.97, and the bias is only 0.15 m s^{-1}. Figure 4.2 shows an example of a time series collected at the two farms. Panel (a) shows the wind speed measured at the two farms from 19th January 2017 to 20th January 2017 inclusive. Panel (b) shows the wind direction measurements for the same period.

The purpose of this work is to determine if it is possible to improve the short term forecast using spatial information. Regarding the space scale, we focus on the 10 min and 20 min forecasts. As the prevailing winds are southwestern, we investigate the added value of Parc de Bonneval measurements to forecast the wind speed at Moulin de Pierre.

As a reminder, the 10 min and 20 min forecasts are computed using linear regression over the last hour measurements, as explained in chapter 2. From that point, the measurements from another farm can be useful at two different points in the forecasting process. It may carry information to predict the average wind speed more accurately by anticipating rapid changes, for instance. It may also carry information to downscale the average forecast to the turbine scale to improve the consideration of the wake effect, for example.

Figure 4.1 | Satellite image of Parc de Bonneval and Moulin de Pierre extracted from Google Earth. The white crosses display the turbine location at Parc de Bonneval and the white circle at Moulin de Pierre location. Typical distances and angle between the two farms are also shown.

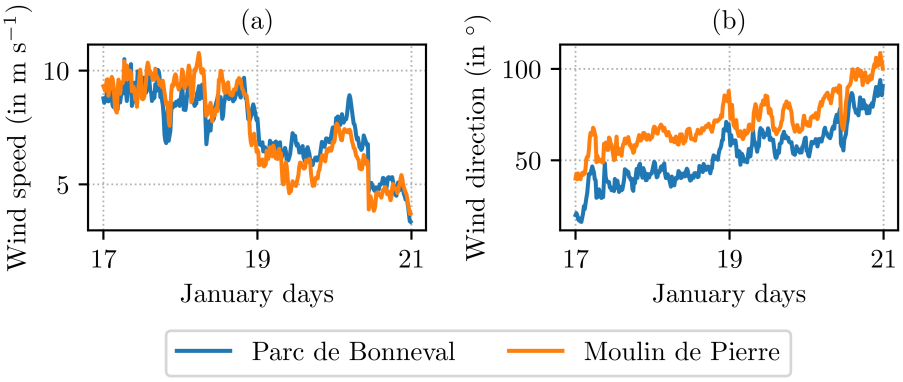


Figure 4.2 | Time series of wind speed (a) and wind direction (b) collected at Parc de Bonneval and Moulin de Pierre. The time series range from the 17$^{\text{th}}$ January 2018 at 00:00 UTC to the 20$^{\text{th}}$ January 2018 at 23:50 UTC.

4.2.2 Improvement of the average wind speed forecast

First of all, we focus on the improvement of the average wind speed forecast at Moulin de Pierre. Then, two cases are distinguished: the case where the wind direction lines up the two farms (we consider a 70° sector), i.e. the wind direction is in [183°,253°] and the case where the wind direction

is not in [183°,253°]. Figure 4.3 displays the distribution of the reference errors. In this case, only the measurements at Moulin de Pierre are used to forecast the wind speed at 10 min and 20 min. Panel (a) only considers data when the wind direction is in [183°,253°] and panel (b) considers the remaining cases, when the wind direction is not in [183°,253°]. The BIAS and the RMSE are also added. These boxplots display the reference errors, that is to say the errors from the original model $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ described in chapter 2.

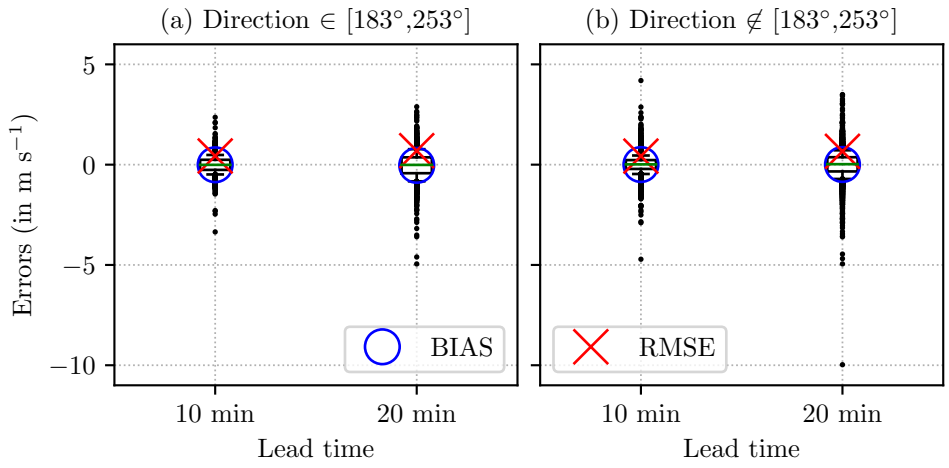


Figure 4.3 | Boxplot of the forecasting errors at 10 min and 20 min when only the measurements collected at Moulin de Pierre are used as explanatory variables. Panel (a) only considers forecasts of the average wind speed when wind direction is in [183°,253°] and panel (b) considers the remaining data, that is to say when the wind direcion is not in [183°,253°].

In each case, the BIAS is very close to zero, and the distribution is symmetric with errors ranging from -5 m s^{-1} to 5 m s^{-1} (there is one exception in panel (b) for the 20 min forecast). The distribution in panel (a) is less widespread. One can note that there is less data in panel (a). Indeed, wind direction in [183°,253°] occurs around 30% of the time.

Figure 4.4 shows the results when the measurements of the previous hour at Parc de Bonneval are also added as explanatory variables. Again, those measurements are obtained by averaging the measurements at the six turbines and the distinction between wind direction being in [183°,253°] (panel (a)) and not being in [183°,253°] (panel (b)) is made. Several differences can be noted but a statistical analysis is necessary in order to better quantify the improvement.

As shown in figure 4.2, the data at the two farms are highly correlated. In these conditions, very few informations can be brought by the data at Parc de Bonneval. This information is all the more difficult to catch since the model used is a parametric one, as it is a linear regression. In chapter 2, we present two types of non-parametric models: the random forest and the artificial neural network. These two approaches can model complex non-linear relationships between datasets. Thus, they may catch more information from correlated data than a linear regression. Figures 4.5 and 4.6 show the distribution of the error between the real wind speed and the forecasted wind speed at 10 min and 20 min ahead when the forecast is provided by a non-parametric model. Figure 4.5 displays the results when random forests are used and figure 4.6 displays the results when neural networks are used.

No significant differences can be found between random forest and linear regression. However, it seems that the neural network degrades the forecasts. In figure 4.6a, there is a positive bias. The

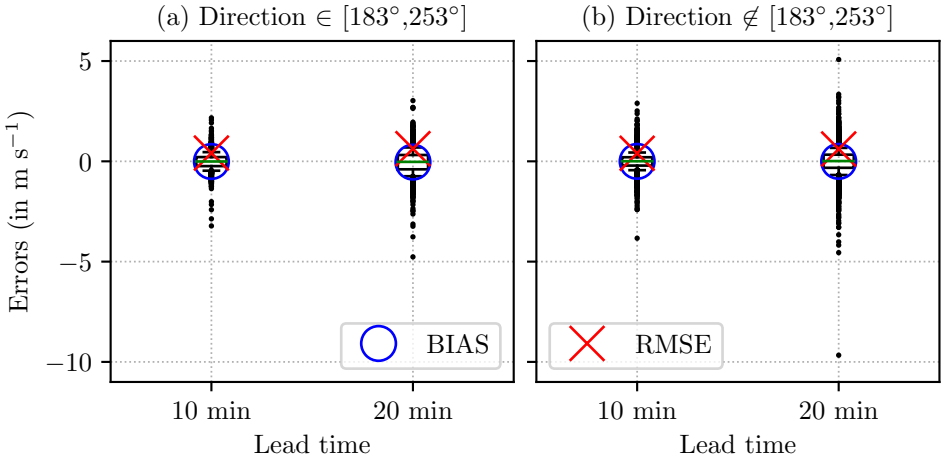


Figure 4.4 | Boxplot of the forecasting errors at 10 min and 20 min when the average wind speed measurements of the previous hour collected at Parc de Bonneval are also added as explanatory variable in the linear regression.

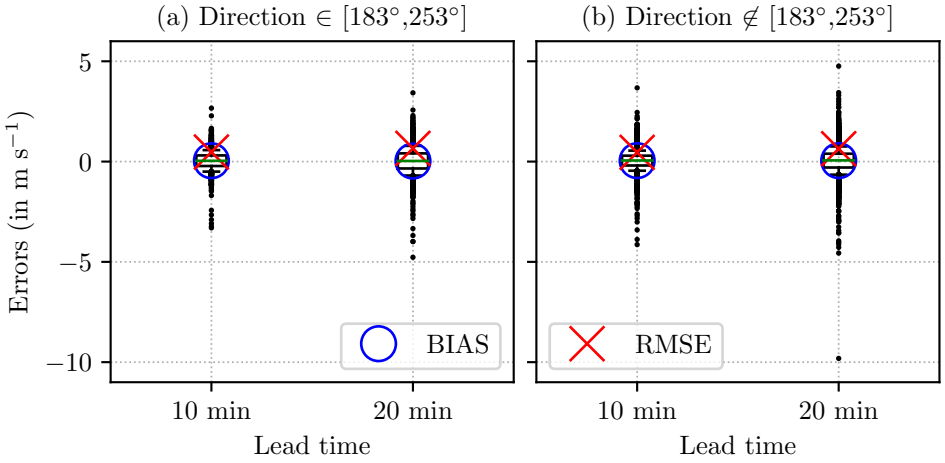


Figure 4.5 | Same as Fig 4.4 but the forecasting model used is a random forest.

distribution is no longer symmetric. Moreover, in both panels, we can see that the distribution is more widespread, at least for the extreme values (1^{st} and 9^{th} decils).

In order to evaluate the performance of the different models more precisely, tables 4.1a and 4.1b group together the characteristics of the different distributions shown in figures 4.3, 4.4, 4.5, and 4.6. The tables compare the means (MEAN), the standard deviations (STD), the interquartile ranges (IQR $= Q_3 - Q_1$) and the scopes (SCOPE = max - min) of the distribution of the error when only the data at Moulin de Pierre are used as explanatory variables (MP) and when the measurements at Parc de Bonneval are added (MP + BO). The linear regression (LR), random forest (RF) and neural network (NN) are considered. Table 4.1a considers the situations where wind direction is in [183°,253°] and table 4.1b considers the situations where the wind direction is not in [183°,253°].

First of all, we can see that the use of Parc de Bonneval measurements leads to some improvements. Most of the maximum improvements are achieved when a linear regression is performed.

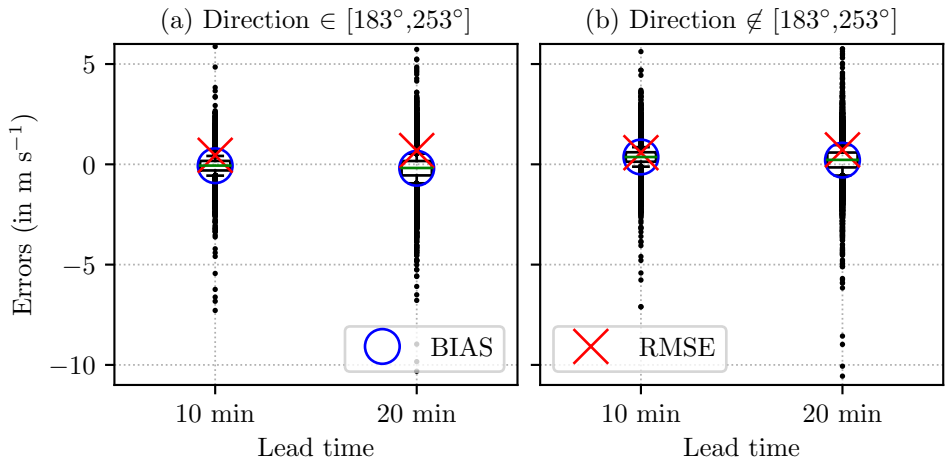


Figure 4.6 | Same as Fig 4.5 but the forecasting model used is a neural network.

Two exceptions can be found: the smallest MEAN for the 20 min forecasts when the wind direction is in [183°,253°] and the smallest SCOPE for the 20 min forecasts when the wind direction is not in [183°,253°] are both found when a random forest is performed. Neural network tends to degrade the forecast, especially when looking at the SCOPE and the MEAN. In terms of STD and IQR, the results are satisfying. In any case, the improvements due to the use of explanatory variable from an upstream wind farm remains limited for the wind speed forecast 10 min and 20 min ahead, but all the models provide accurate forecasts. In general, greater improvements are found when the wind direction is not in [183°,253°], but it can be due to the fact that the dataset is bigger than when the wind direction is in [183°,253°]. The fact that the farms are very close and the data is very correlated limits the amount of new information that can be used by the models, even the non-parametric ones. Moreover, it also limits the time scale. Here, only the 10 and 20 min are considered.

Furthermore, the distinction of cases is based only on the wind direction. Yet it is obvious that when the wind direction is in [183°,253°], it does not necessarily mean that the wind comes from Parc de Bonneval to Moulin de Pierre since the movement of air masses is governed by much more complex phenomena than simple transport. Then, it might be relevant to refine the cases distinction.

4.2.3 Improvements of the wind turbine downscaling

The use of spatial information can also be useful in order to downscale the forecast at the scale of the turbine. In this context, the impact on the forecast of two turbines is investigated. We choose to focus on the two closest turbines in Moulin de Pierre as they are the most sensitive to the wake effect. When the wind direction is in [183°,253°], that is to say when the wind comes from Parc de Bonneval, one of the turbines is the downstream turbine and the other is the upstream turbine. The two cases are discussed in the following. First, the average wind speed is forecasted using average measurements at Moulin de Pierre only. Then, this forecast is used as explanatory variable by a second model as well as the last measurement at 1. the six turbines from Moulin de Pierre; 2. the twelve turbines from Moulin de Pierre and Parc de Bonneval. This workflow is shown in figure 4.7.

(a) Direction $\in [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	**-0.01**	**-0.01**	0.04	0.42	-0.03	-0.04	**0.02**	0.49
STD	0.43	**0.41**	0.47	0.49	0.69	**0.63**	0.65	0.67
IQR	0.50	**0.45**	0.53	0.54	0.79	**0.71**	0.76	0.74
SCOPE	5.72	**5.39**	5.57	11.27	7.84	**7.79**	7.87	17.66

(b) Direction $\notin [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	0.01	**-0.00**	0.05	-0.01	0.01	**-0.00**	0.05	-0.10
STD	0.43	**0.40**	0.45	0.45	0.66	**0.61**	0.64	0.68
IQR	0.44	**0.42**	0.50	0.47	0.71	**0.65**	0.70	0.74
SCOPE	8.91	**6.73**	6.93	13.29	17.02	14.74	**14.15**	18.17

Table 4.1 | Features of the distribution of the forecasting errors on the average wind speed. The tables show the characteristics of the boxplots in the figures 4.3, 4.4, 4.5, and 4.6. The means (MEAN), the standard deviations (STD), the interquartile ranges (IQR) and the scopes (SCOPE) are shown. Four cases are considered, when a linear regression is performed using only the measurements at Moulin de Pierre and when the measurements at Parc de Bonneval are added. We also consider the cases where random forest and neural network are performed using as explanatory variables the measurements at Moulin de Pierre and Parc de Bonneval. Table (a) corresponds to the left panels in the figures, that is to say when the wind direction is in [183°,253°], and table (b) corresponds to the right panels in the figures, that is to say when the wind direction is not in [183°,253°]. For each row, the model that shows the best performances is bolded.

Y_{MP}^{Ei}, $i \in [\![1, 6]\!]$ is the measurements at t_0 of turbine Ei at Moulin de Pierre and Y_{BO}^{Ei}, $i \in [\![1, 6]\!]$ is the measurements at t_0 of Ei at Parc de Bonneval. $\tilde{Y}_{MP}^{t_0+10}$ (resp. $\tilde{Y}_{MP}^{t_0+20}$) is the average wind speed forecasted at $t_0 + 10$ min (resp. $t_0 + 20$ min) using only the past measurements at Moulin de Pierre as explanatory variables. One model is fitted to forecast the wind speed at Ei, $i \in \{1, 2\}$ at 10 min with the explanatory variable framed in green and one model is fitted to forecast the wind speed at Ei, $i \in \{1, 2\}$ at 20 min with the explanatory variable framed in red.

Downstream turbine

First of all, the impact on the downstream turbine is examined. Table 4.2 shows the characteristics of the distribution of the reference errors for the downstream turbine. The reference errors are computed as the difference between the measured wind speed and the wind speed forecasted by a direct linear regression. The target is then the wind speed measured at the turbine and no longer the average wind speed. Again, the case where the wind direction is in [183°,253°] and the case where the wind direction is not in [183°,253°] do not have the same frequency. The wind direction is not in [183°,253°] 70% of the time (this means it is in [183°,253°] 30% of the time).

The errors used for table 4.2 are computed using one model. The explanatory variables are the six last measurements at the turbine, and the target is the wind speed 10 min and 20 min ahead.

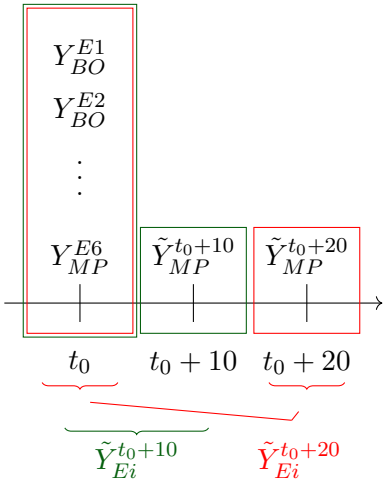


Figure 4.7 | Methodology to construct the models to quantify the added value of small scale information to improve the wind turbine downscaling. For the turbine forecast 10 min ahead, the explanatory variables are: the average wind speed measurements at the two farms for the previous hour and the average forecast provided by $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ at Moulin de Pierre 10 min ahead. For the turbine forecast 20 min ahead, the explanatory variables are: the average wind speed measurements at the two farms for the previous hour and the average forecast provided by $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ at Moulin de Pierre 20 min ahead

	Direction $\in [183°, 253°]$		Direction $\notin [183°, 253°]$	
	10 min	20 min	10 min	20 min
MEAN	-0.01	-0.04	-0.01	-0.01
STD	0.66	0.94	0.62	0.86
IQR	0.72	1.04	0.66	0.94
SCOPE	7.42	9.51	10.78	17.81

Table 4.2 | Features of the distribution of the forecasting errors for the forecast of the wind speed at an upstream turbine. In this case the target of the linear regression is no longer the average wind speed but the wind speed measured at the turbine. The mean (MEAN), the standard deviation (STD), the interquartile range (IQR) and the scope (SCOPE) are shown. Forecasts for wind direction in [183°,253°] (left) and for wind direction not in [183°,253°] (right) are shown.

The features in table 4.2 are the references to be compared with those from the errors distribution obtained by applying the workflow shown in figure 4.7. For this workflow, the characteristics of the distribution of the errors are shown in tables 4.3a and 4.3b. They group together the characteristics of the error distributions depending on the model. In each case, the applied workflow corresponds to the one shown in figure 4.7.

For both 10 min and 20 min ahead, the first columns (MP - LR) correspond to the case where the second model is a linear regression using only the measurements at Moulin de Pierre. The columns corresponding to MP + BO refer to the models where the measurements at Moulin de Pierre and Parc de Bonneval are used as explanatory variables in the second model. Then, as in section 4.2.2, three different approaches are tested for this second model: linear regression (LR), random forest (RF), and neural network (NN). The tables compare the means (MEAN), the standard deviations (STD), the interquartile ranges (IQR), and the scopes (SCOPE). Table 4.3a

considers the situations where the wind direction is in [183°,253°] and table 4.3b considers the situations where the wind direction is not in [183°,253°].

(a) Direction $\notin [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	0.04	0.04	**0.03**	1.27	0.04	**0.03**	**0.03**	1.27
STD	**0.50**	**0.50**	0.54	0.79	**0.53**	**0.53**	0.55	0.83
IQR	0.64	**0.63**	0.64	2.11	0.56	**0.55**	0.59	1.11
SCOPE	**5.28**	**5.28**	5.59	6.33	**5.79**	5.97	6.40	6.59

(b) Direction $\notin [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	**-0.02**	**-0.02**	-0.03	1.03	**-0.00**	**0.00**	**-0.00**	1.01
STD	**0.48**	**0.48**	0.50	0.70	**0.46**	**0.46**	0.48	0.72
IQR	0.51	**0.50**	0.55	0.90	**0.49**	**0.49**	0.53	0.91
SCOPE	7.04	6.91	**6.21**	7.74	7.52	7.58	**6.93**	9.21

Table 4.3 | Features of the distribution of the forecasting errors on the wind speed at the downstream turbine. The tables show the mean (MEAN), the standard deviation (STD), the interquartile range (IQR) and the scope (SCOPE). Table (a) is when the wind direction is in [183°,253°] and table (b) is when the wind direction is not in [183°,253°]. For both 10 min and 20 min ahead, the first columns (MP - LR) correspond to the case where the second model is a linear regression using only the measurements at Moulin de Pierre. The columns corresponding to MP + BO refer to the models where the measurements at Moulin de Pierre and Parc de Bonneval are used as explanatory variables in the second model. Then, as in section 4.2.2, three different approaches are tested for this second model: linear regression (LR), random forest (RF) and neural network (NN). For each time, the model that show the highest improvement compared to table 4.2 is bolded.

These tables lead to the same conclusion as in section 4.2.2. Most of the time, the linear regression that uses data from both farms overperforms the other model. Again, the neural network shows the worst performances. The two linear regressions (with Moulin de Pierre measurements only and when measurements at Parc de Bonneval are added) and the random forest that uses data from the two farms provide very accurate forecasts.

If we now compare the results with those in table 4.2, we can see that the use of a second model to downscale the wind speed forecast at the turbine scale gives better results than the direct forecast. For 10 min ahead, the standard deviation is reduced by 25%, the IQR is reduced by 10%, and the scope is reduced up to 35%. At 20 min, the standard deviation, the IQR, and the SCOPE are reduced by 75%. We can assume that using a second model, filters part of the turbulence.

Upstream turbine

After the downstream turbine, in this section we focused on the possible improvement for the upstream turbine. As in the previous section, table 4.4 shows the features of the distribution of the forecasting errors for the forecast of the wind speed at the upstream turbine. Again, the MEAN,

the STD, the IQR, and the SCOPE are shown. Forecasts when wind direction is in [183°,253°] (left) and when wind direction is not in [183°,253°] (right) are shown. These features will be compared with those computed from the models obtained by applying the workflow shown in figure 4.7.

	Direction $\in [183°, 253°]$		Direction $\notin [183°, 253°]$	
	10 min	20 min	10 min	20 min
MEAN	-0.02	-0.05	0.01	0.00
STD	0.58	0.81	0.59	0.82
IQR	0.65	0.95	0.65	0.91
SCOPE	6.42	9.18	9.16	16.62

Table 4.4 | Same as table 4.2 for the upstream turbine.

As for the downstream turbines, several configurations are tested: linear regressions with measurements at Moulin de Pierre only and when measurements at Parc de Bonneval are added, random forest, and neural network using data at the two farms.

Tables 4.5a and 4.5b sum up the features of the error distribution for each configuration. Again, we can see that the neural network provides the worst forecast. It introduces significant BIAS. For the other configurations, the results are very similar. Linear regressions present slightly better results than random forest, with an advantage for the one using data from the two farms. If we compare with table 4.4, again, the forecasts are significantly improved by the use of a second model. For both the downstream and the upstream turbines, part of the turbulence and so part of the uncertainty is filtered by the first linear regression, and it allows the second model to performs better.

(a) Direction $\notin [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	**0.01**	**0.01**	**0.01**	1.79	0.03	0.02	**0.01**	1.45
STD	**0.47**	**0.47**	0.49	0.88	0.46	**0.45**	0.46	0.85
IQR	**0.65**	**0.65**	0.66	2.72	0.51	**0.50**	0.52	1.14
SCOPE	4.91	4.95	**4.14**	7.12	5.45	5.51	**5.11**	7.75

(b) Direction $\notin [183°, 253°]$

	10 min				20 min			
	MP	MP + BO			MP	MP + BO		
	LR	LR	RF	NN	LR	LR	RF	NN
MEAN	**-0.00**	**-0.00**	0.02	1.49	-0.01	-0.01	**0.00**	1.14
STD	**0.46**	**0.46**	0.50	0.78	**0.43**	**0.43**	0.46	0.74
IQR	**0.51**	**0.51**	0.55	1.04	**0.46**	**0.46**	0.50	0.94
SCOPE	7.40	**7.13**	7.65	8.45	**6.28**	6.47	6.35	9.17

Table 4.5 | Same as tables 4.3 for the upstream turbine.

As in section 4.2.2, we can see that networking wind farm at small scales allows for some improvements for the wind turbines downscaling whether it is a downstream, or an upstream

turbine. In the next part, we test if two distant wind farms can carry some large scale information and lead to further improvements for the wind energy forecasts.

4.3 Added value of large scale information

4.3.1 Wind farms location and correlation

As said previously, Zephyr ENR is the owner of six wind farms spread over the northwest part of France. In these conditions, we can use the data of the distant wind farms in order to improve the forecasts of Parc de Bonneval. In this section, we assess the interest of using the data of two wind farms several hundred kilometers away. The wind farm called Parc de la Vènerie is located 210 km northwest of Parc de Bonneval, and 220 km southwest of Parc de Bonneval is the Parc de la Renardière, as shown in figure 4.8. For reminder, Parc de la Vènerie is composed of 4 Enercon E82-2.3 MW turbines and a hub height of 85 m. Parc de la Renardière is composed of 6 Senvion MM92-2 MW turbines and a hub height of 100 m. Parc de la Vènerie was implemented in 2014, and Parc de la Renardière was implemented in 2009. For each farms, the 10 min data of 2015 are used as training period and the data of 2016 are used as a testing period.

Figure 4.8 | Satelite image of the north of France which shows the location of Parc de Bonneval, Parc de la Renardière and Parc de la Vènerie. The map is extracted from Google Earth.

Both Parc de la Vènerie and Parc de la Renardière are distant by more than 200 km from Parc de Bonneval, and at a wind speed of 6 m s^{-1}, it would take around 9 h to travel from one farm to another. However, in this part, the information we want to include in the model is not the same as in section 4.2. More than wind propagation, here the information is based on general atmospheric conditions. Typically, on the pressure variations that induce a circulation. That is why, in this section, we also retrieved the forecasts from ECMWF at Parc de la Renardière and Parc de la Vènerie, as done in chapter 2. But first of all, we want to know if a significant correlation between the farms according to a systematic lag can be found. Figure 4.9 displays an example of cross correlation between Parc de Bonneval and Parc de la Vènerie (panel (a)) and between Parc de Bonneval and Parc de la Renardière (panel (b)). The cross correlation is computed as follow:

$$r = \frac{\sum_{n=1}^{N} (Y_n^1 - \bar{Y}^1)(Y_{n+k}^2 - \bar{Y}^2)}{\sqrt{\sum_{n=1}^{N} (Y_n^1 - \bar{Y}^1)^2} \sqrt{\sum_{n=1}^{N} (Y_{n+k}^2 - \bar{Y}^2)^2}} \tag{4.1}$$

where N is the sample size, Y_n^1 is the *n-th* measurement at Parc de Bonneval and $\bar{Y}^1$ is the sample averaged. k is the lag which goes from -25 h to 25 h. Then, Y_{n+k}^2 is the $(n+k)$-*th* measurement at Parc de la Vènerie or at Parc de la Renardière and $\bar{Y}^2$) is the sample averaged at Parc de la Renardière or Parc de la Vènerie.

Figure 4.9 shows an illustration of the cross correlation. For both Parc de la Vènerie and Parc de la Renardière $k = 0$ corresponds to the 28$^{\text{th}}$ of July 2015 at 13:00 UTC. For Parc de la Vènerie (panel (a)), the optimal lag is $k = 0$ with a correlation around 0.8. A strong anticorrelation is also found (around -0.8) for k=14. For Parc de la Renardière, the strongest correlation is found for $k = -3$, and it is around 0.7. A negative lag would suggest that Parc de Bonneval carries information for Parc de la Renardière and not the other way. No significant anticorrelation is found in this case.

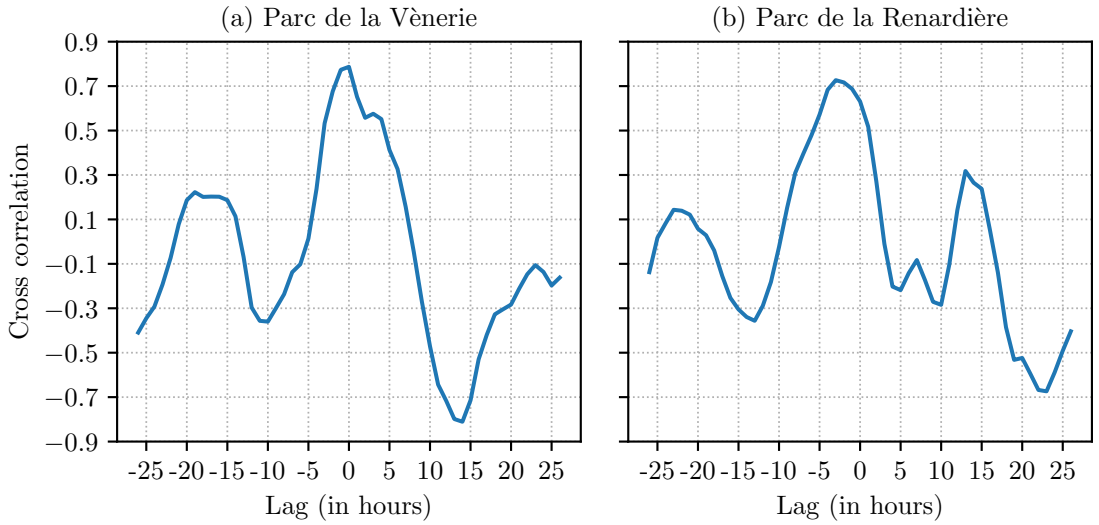
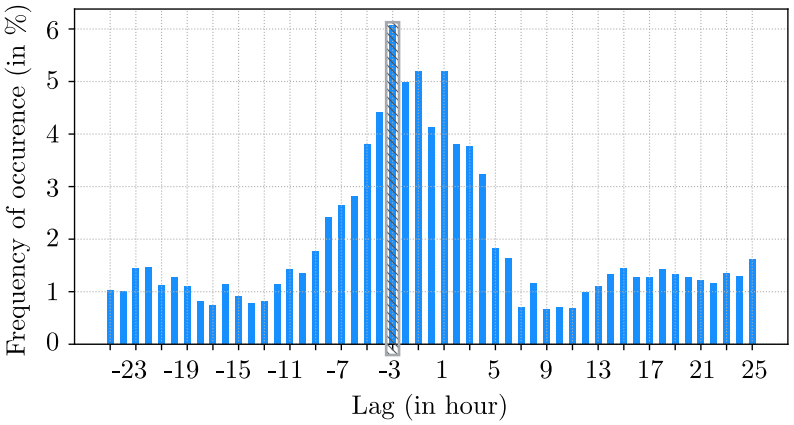


Figure 4.9 | Cross correlation between Parc de Bonneval and Parc de la Vènerie (panel (a)) and between Parc de Bonneval and Parc de la Renardière (panel (b)). The lag k goes from -25 h to 25 h and $k = 0$ corresponds to the 28$^{\text{th}}$ of July 2015 at 13:00 UTC.

From that point, we compute the frequency of occurrence of the maximal lag over the years 2015 and 2016. For the data of those years, for a value of $k \in [\![-25, 25]\!]$, we compute the cross correlation between Parc de Bonneval and Parc de la Renardière and between Parc de Bonneval and Parc de la Vènerie. For each case, we keep the lag corresponding to the maximum correlation. Then, figure 4.10 displays for each hour, the number of times it is the optimal lag. That is to say, the number of times it corresponds to a maximum correlation. The results are shown in figure 4.10a for Parc de la Vènerie, and in figure 4.10b for Parc de la Renardière. If we could expect an optimal lag between 5 h and 10 h, the most frequent lag is actually -3 h for Parc de la Vènerie and 1 h for Parc de la Renardière. This positive optimal lag implies that predictability would be in the

direction from Parc de Bonneval to Parc de la Renardière instead of from Parc de la Renardière to Parc de Bonneval. This is counter-intuitive since easterly winds only represent 24% of cases while westerly winds represent more than 40% of cases.

(a) Parc de la Vènerie



(b) Parc de la Renardière

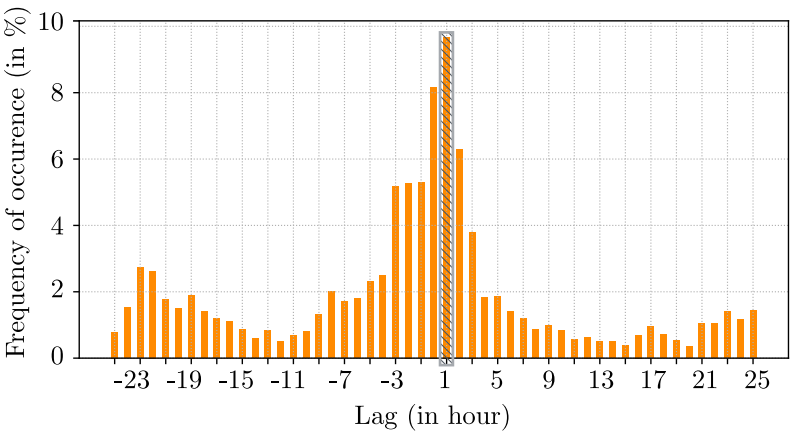


Figure 4.10 | For each lag $k \in [\![-25, 25]\!]$, in hour, the number of times it has been the optimal lag is shown in percentage. In other words, it displays the number of time each lag has been the maximum correlation between the sample at Parc de Bonneval and the sample at Parc de la Vènerie (panel (a)) and between Parc de Bonneval and Parc de la Renardière (panel (b)). The data at Parc de Bonneval are fixed and the data at Parc de la Vènerie and Parc de la Renardière are lagged by k h. In both cases, for more visibility the optimal lag is framed and hatched in grey.

This clearly suggests that Parc de la Renardière and Parc de la Vènerie do not have a predictive value on Parc de Bonneval. Indeed, between Parc de Bonneval and Parc de la Renardière, even if the optimal lag is positive, it is only 1 h while the two farms are more than 200 km apart. In these conditions, it is clear that the use of the data at Parc de la Renardière or at Parc de la Vènerie to correct the forecasts at Parc de Bonneval is unlikely to be relevant. At least, in a systematic way. There are some specific meteorological conditions, as the passage of a front, for instance, that can be anticipated using this large scale informations.

4.3.2 Application to forecasts

To conclude this part, the addition of large scale information in the forecasts is briefly investigated. We use data from Parc de la Renardière and Parc de la Vènerie. Figure 4.11 shows the results. As for the previous parts, random forest, neural network, and linear regression are investigated (dashed lines). For comparison, we add the results, shown in chapter 2, for the $\text{LR}_{\text{SW}}^{\text{obs}}$ model, where only data from Parc de Bonneval are used as explanatory variables. The NRMSE (in %) is shown depending on lead times ranging from 10 min to 170 min.

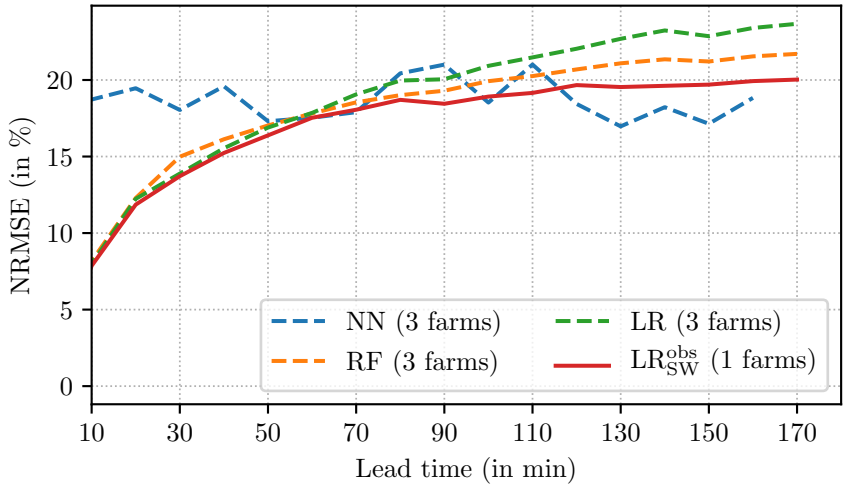


Figure 4.11 | Performances of linear regression (LR), random forest (RF), and neural network (NN) for wind speed forecasting when the data from Parc de la Renardière and Parc de la Vènerie are added as explanatory variables. For comparison, we add the results, shown in chapter 2, for the $\text{LR}_{\text{SW}}^{\text{obs}}$ model, where only data from Parc de Bonneval are used as explanatory variables. The NRMSE (in %) is shown depending on lead times ranging from 10 min to 170 min.

In this simplified configuration, the addition of Parc de la Renardière and Parc de la Vènerie data (measurements and ECMWF forecasts) leads to a degradation (except for the neural network for the last lead times). The performances of the neural network are quite different from those shown in chapter 2 because, in this case, explanatory variables from ECMWF outputs are used as exogenous variables. It explains the poor performances for the first lead time. However, it seems to performs well after 2 h. For the linear regression and the random forest, the use of data from Parc de la Vènerie and Parc de la Renardière, disrupts the model by introducing unnecessary data to which the model assigns a low but not zero weight.

These data, used systematically as explanatory variables is not the right way to include large scale information. Again, in this case, this large scale information is not significant enough, and it cannot improve the forecasts in a systematic way. To go further, it would be interesting to investigate more deeply some specific cases where the large scale information is relevant. The identification of several regimes is a key step, which is probably more crucial than for the small scale information. From these regimes, it would be necessary to determine how much information could be brought to the model and how.

4.4 Conclusion

In this chapter, we assess the added value of networking wind farms. More precisely, the added value of small scale and large scale information are investigated in order to improve the short term wind speed forecasts (from 10 min to 170 min).

First of all, regarding the small scales, we use the data of Parc de Bonneval located around 5 km away from Moulin de Pierre. For the 10 min and 20 min forecasts, it seems to be the right distance to carry some information in advance. Especially when the wind comes from Parc de Bonneval, which are the prevailing winds. In this part, we highlight two different issues for which networking wind farms could be useful. On the one hand, we use data from Parc de Bonneval in order to improve the average wind speed forecasts. We compared the performances of linear regressions, random forest, and neural network. Two cases are distinguished: when the wind comes from Parc de Bonneval (i.e., wind direction is in [183°,253°]) and when it does not. In both cases, the use of Parc de Bonneval data improves the forecast at Moulin de Pierre. On the other hand, we evaluate the impact of Parc de Bonneval data to downscale the forecast at the turbine scale. To do so, we build several models such as linear regression, random forest, and neural network. Each model takes as input the average forecast at the farm scale and the measurements at the turbines (only Moulin de Pierre or Moulin de Pierre and Parc de Bonneval). Then, using these models, we downscale the average forecast at the turbine scale. Again, some improvements can be observed by the use of Parc de Bonneval data. Moreover, the use of a second model allows significant improvements compared with the direct forecast at the turbine scale.

After the small scales, the impact of large scale information is investigated. For this purpose, we use the data from two distant farms (more than 200 km away from Parc de Bonneval), called Parc de la Renardière and Parc de la Vènerie. In this part, we try to identify an optimal lag, but without success. The optimal lags that emerge do not show a strong potential for predictability from one farm to another, at least for the method that is considered. This could be explained by the use of an overly simplistic method, which takes into account only the wind direction. If the predictability exists, it would surely be necessary to add other conditions or variables than direction, such as the spatial distribution of the wind (e.g., to capture depressions, anticyclone). Despite this, we add large scale data in the model to see the impact on the forecasts. As expected, it degrades the forecasts for the whole period except for the neural network with exogenous variables, which provide the best forecasts after 2 h.

In the first case, the two farms seem too close and are too correlated, but in the second case, they are too distant. A farm located between 50 km and 20 km would probably be the best solution to add relevant information for our lead times. Moreover, the main issue of this study is that the farms would certainly not provide relevant information all the time. Therefore, their systematic inclusion would not be the right approach. Then, it should not be included as an explanatory variable, as done in this chapter. A thorough selection of regimes is a key for an efficient networking of wind farms such as Zhu *et al.* in [76] who use a RST model to forecast the wind speed that allows the regimes to vary with the wind direction and according to the diurnal and seasonal patterns, hence avoiding a subjective choice of regimes.

5

The economic value of short term forecasting for wind energy

Contents

5.1 Introduction 94
5.2 Simplified market simulation 95
5.2.1 Electricity market 95
5.2.2 Market simulation 97
5.3 Impact of the short term balancing strategy 101
5.3.1 Impact of the forecasting errors 102
5.3.2 Impact of the price volatility 104
5.4 Performance of the short term forecasting model 106
5.4.1 Balacing fees 106
5.4.2 Added value of short term forecasting model 109
5.5 Conclusion 111

5.1 Introduction

Due to the variation in wind energy production caused by unpredictable changes in wind speed, producers who are part of a liberalized electricity market are exposed to penalties related to the costs of regulating the grid.

The central aspect of these liberalized markets is that participants must make offers in advance on a spot market. On such market, electricity purchases and sales are notified for the next day. Producers and consumers announce once every day, at 12:00, their offers regarding quantities and prices for the next day from 00:00 to 23:00. Thus, offers are given 12 h to 35 h before the delivery. Based on these offers, market prices are determined by demand and supply on an hourly basis for the following 24 h. Then, producers are charged for any imbalance. Imbalance is defined as the difference between real production and the previous day's offer. These balancing penalties are defined afterward, depending on the cost of the grid regulation [77]. Studies have shown that medium term forecasts can be used to enhance the value of wind energy production. For instance, in [78], Roulston *et al.* compare the use of ECMWF-based forecast and climatology forecast to bid on the spot market in the UK. They found that with the ECMWF-based forecast, the daily income is higher than with climatology on 60% of the days, and weekly income is higher on 80% of the weeks. In [79], Pinson *et al.* show that providing medium term forecasts with information on their uncertainty can be the basis for defining advanced strategies for participation in the Dutch market. In [80], Barthelmie *et al.* show that for a wind farm of 12 MW (which correspond to Parc de Bonneval) using a forecasting model, is profitable as long as the price of this model does not exceed 500000£ or around 550000€. As for [78], their study is based on the UK market.

From this point on, readers might question the usefulness of short term forecasts. However, in the above market description, a crucial step has been omitted. Between the two steps previously described, a stage during which the producer can adjust its original offer is possible via the intraday market. In practice, the intraday market is not a market whose prices are governed by supply and demand, but it is an order book. It gathers all sales and purchase orders in real time. The five best offers and the five best demands are visible from the others in order to be able to position their order being fully informed and avoid that it is not executed, or executed at a bad price. Thus, as soon as two offers match, they can be executed. This order book opens the day before at 15:00 and closes 30 min before each hourly delivery date. A few hours to 30 min are the typical lead times of the forecast model considered in this thesis. In order to minimize imbalance penalties, the balancing via intraday is the target of this chapter. In [81], Fabbri *et al.* model the prediction errors through a probability density function that represents the accuracy of the model. Production hourly deviations and associated trading costs from the Spanish market are also calculated. Considering three study cases, they show that the error prediction costs can reach as much as 10% of the total generator energy incomes. In [82], Usaola *et al.* show that revenues can be increased with a short term wind energy forecast model, even if the latter is of medium accuracy. They use a model based on in-situ measurements and NWP outputs to forecast the wind energy on an hourly basis. Using the rules of the Spanish electricity market, they show that with such a forecasting model, the decrease in income due to forecasting errors is 7.5% compared with the case where the forecast is perfect. However, the reduction is 9.5% if a persistence model is used and 10% if no model is used. In [83], Matevosyan *et al.* present a method to minimize imbalance costs. They develop a model based on imbalance price, to simulate the Nordic power market (Norway, Finland, Sweden, and Denmark), and they combine it with wind energy forecasts to build a stochastic optimization model to generate optimal wind power production bids for intraday. For January 2003, they show

that the income from using this model is 700€ higher than the income from the case where the bid is based on a wind energy forecast only.

Such economic quantification of the value of the short term forecast is another way to evaluate its performance. Indeed, although forecast accuracy is the main objective of forecasters, their users are more interested in maximizing revenue from the use of predictions [84].

Then, the goal of this chapter is to quantify the economic value of a short term forecasting model for a producer. All the tests are conducted using the data at two specific wind farms and are based on the French electricity market rules. We again use data from Parc de Bonneval and Parc de la Vènerie. First, we briefly describe the electricity market and present an example of a simplified simulation of the French electricity market in section 5.2. In section 5.3, various scenarios are compared: the case where no short term forecast is available, the case where a perfect short term forecast is available, and the case where a realistic short term forecast is available. For the latter scenario, the short term models presented in chapter 2 are used. Then, the impact of those three scenarios is analyzed, and the different sources of variability are highlighted. Finally, in section 5.4, the results are combined to exhibit the scenario that maximizes the income. The economic value of having access to short term forecasts is also quantified.

It is essential to specify that in practice, it is not up to the producer himself to go to the market, but the latter goes through an aggregator. That said, the methodology of this study remains valid, it would be necessary to transpose it to the aggregator scale, and only the numbers presented would vary.

5.2 Simplified market simulation

5.2.1 Electricity market

As of December 31, 2017, more than 32 million individual sites and more than 2 million professional and industrial sites were connected to the electricity grid in France. In 2000, the French market moved from a national monopoly market, owned by the French utility company EDF (Électricité De France) to a competitive European market. Due to its number of customers (individuals and companies), the electricity market is one of the largest in Europe. However, EDF still has 80% of market share among individuals, some ten years after liberalization.

In France in 2018, 71.7% of the electricity produced came from nuclear power, 21.2% from renewable energy sources (mainly hydroelectric power generation: 12.4% and wind power: 5.1%) [4]. In terms of consumption, renewable energies cover 22.7% of needs in France [85]. To achieve the objectives set by the Paris Agreement, France has set up a system called a *purchase obligation*. This system allows renewable energy producers to sell their electricity at a regulated price to EDF during a limited period. For wind energy producers, this period lasts 15 years. After this period, they have to sell their electricity on the competitive market. This market is organized in three steps. Figure 5.1 describes those 3 steps.

First of all, the producer has to sell his electricity on the *day ahead* market at 12:00 the day before the delivery date for each slot of the next day. Then, the producer has the possibility to balance his sale via the *intraday* market from the day before at 15:00 and until 30 min before the delivery date. Finally, the differences between the electricity sold and the electricity produced are balanced via *imbalance settlement* [77]. Those three steps are described below.

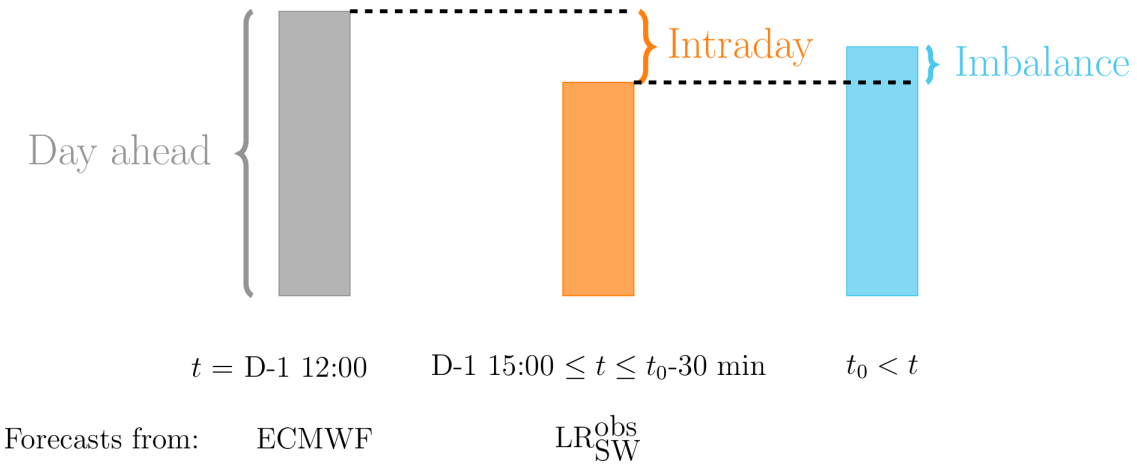


Figure 5.1 | Electricity market organization for a delivery time at t_0. The market is divided in 3 steps. The sale on the day ahead market, at 12:00 the day before the delivery date, for each slot of the next day. The balancing by the producer on the intraday market up to 30 min before the delivery date and the balancing from RTE after the electricity delivery (imbalance).

Day ahead market

The first step in the short term electricity market is the sale on the day ahead market. Unlike the other two, the day ahead market is a real market where prices are governed by supply and demand. This market opened every day at 12:00 for the next day. At this moment, the producer has to sell a medium term forecast of wind energy production for each hourly slot of the next day. At this moment, he needs wind energy forecasts from +12 h to +35 h.

Intraday

The intraday market is a balancing market. It opens the day before at 15:00, and transactions are possible up to 30 min before the delivery date. Using short term forecasts, the producer can update the medium term forecast used to sell the energy on the day ahead market. If too much electricity has been sold, the producer can buy the difference on intraday. On the contrary, if not enough electricity has been sold on day ahead, he can sell the surplus on intraday. This market is actually a book order that includes the demands and supplies of the market participants. Each participant can assign one or more orders, depending on his forecasts. It is a priori more advantageous to sell the production on the day ahead market if it can be well forecasted. It can be tempting to balance the production on intraday at the last moment in order to have access to a forecast that is as accurate as possible. However, price volatility increases significantly as the delivery date approaches. Consequently, it is more risky to wait to balance.

Imbalance

Once the electricity is delivered, the differences between the sold production and the real production is financially compensated. These prices are set by the French electricity transmission system operator (RTE) according to the cost of the balancing actions to balance the French electricity system. If too much electricity has been delivered by the producer, the difference is refunded at a price generally lower than the price the producer would have had by selling on intraday. On the

contrary, if not enough electricity has been delivered, the producer has to buy the difference at a price generally higher than the price the producer would have had by buying on intraday.

5.2.2 Market simulation

From this point, only the price data are missing in order to make the first simulation of the market. For the day ahead market, the hourly prices are available on the Epex Spot website[1] from April 22nd, 2005 to now. Epex Spot is a European electricity exchange. It gathers and manages the energy transactions of France, Germany, Austria, and Switzerland, as well as the intraday market. In this continuous market, market members' orders are entered in the order book continuously. As soon as two orders are compatible, they are executed. On the Epex Spot website, for each delivery date, the highest price, the lowest price and the last price of the transactions are available. However, we have access to the entire book order for the year 2015. From here, we are able to retrieve each transaction (demand or supply) and its associated price. Finally, the balancing prices are computed by RTE, depending on the balancing cost. On the RTE website, the balancing prices are available from 2003 to now. As we retrieve the intraday prices only for 2015, we focus on this year in this chapter.

First of all, for sale on the day ahead market, we need wind energy forecasts from 12 h to 35 h. ECMWF forecasts are hourly forecasts up to 4 days. However, for the preliminary study we only have access to the first 24 h of the hourly forecasts. Consequently, for each day at 12:00, we use the ECMWF forecast to retrieve the wind speed (with forecasts of more than 24 h, we should use the ECMWF forecast at 00:00 because each forecast is available around 7 h after the lauching date). Then, using the power curve, we retrieve the wind energy forecast from 12 h ahead to 24 h ahead. For the last 11 h, we consider as a prediction the last value from ECMWF, i.e. the forecast 24 h ahead. An intermediate step is necessary for the medium term forecasting at Parc de la Vènerie. Indeed, unlike Parc de Bonneval where the turbines are at 100 m height as well as the ECMWF forecasts, at Parc de la Vènerie, the turbines are at 85 m height. Those 15 m are essential and will lead to an overestimation of the production. In these conditions, it will be necessary to buy this excedent on intraday, which constitutes an avoidable loss of income. Since ECMWF forecasts wind components at 10 m and 100 m, it is possible to interpolate the wind speed at 85 m. To do so, we use a wind profile power law defined in equation (5.1):

$$U_z = cz^{\alpha} \tag{5.1}$$

where U_z is the wind speed at altitude z and where c and α are parameters that have to be estimated. The law is fitted using the average wind profile. Here, the data from 2015 are used. From equation 5.1, α can be easily calculated through the equation (5.2):

$$\alpha = \frac{\log(\overline{U_{10}}) - \log(\overline{U_{100}})}{\log(10) - \log(100)} \tag{5.2}$$

where $\overline{U_{10}}$ (resp. $\overline{U_{100}}$) is the average wind speed at 10 m (resp. 100 m) forecasted by ECMWF over the year 2015. From this point, it is easy to determine the wind speed at 85 m from ECMWF wind speed forecast and the coefficient α:

$$\widehat{U}_{85} = U_{100}\left(\frac{85}{100}\right)^{\alpha} \tag{5.3}$$

[1] https://www.epexspot.com/fr/donnees_de_marche/dayaheadfixing

Figure 5.2 shows the NRMSE (panel (a)) and the BIAS (panel (b)), in % of the installed capacity, between the real wind energy and the wind energy forcasted using ECMWF forecast. Results for both Parc de Bonneval and Parc de la Vènerie for the year 2015 are shown, depending on the lead time. For Parc de la Vènerie, we add in dashed line the performances of ECMWF without the height correction. When the altitude correction is applied to the forecasts, the NRMSE is slightly decreased, and the BIAS is significantly reduced.

If we now focus on the solid lines, we can see that they are noisy. This can be explained by the lack of forecasts (one per day). For the NRMSE, the trend is clearly visible. From 12 h to 24 h, where the forecasted wind by ECMWF is used, the errors are constant, slightly below 15%. However, from 24 h to 35 h, where a naïve estimate is used, the errors overgrow. For the BIAS, it is positive, around 2% for the whole period at Parc de la Vènerie while it tends to be negative at Parc de Bonneval, and it falls below 5% after 30 h.

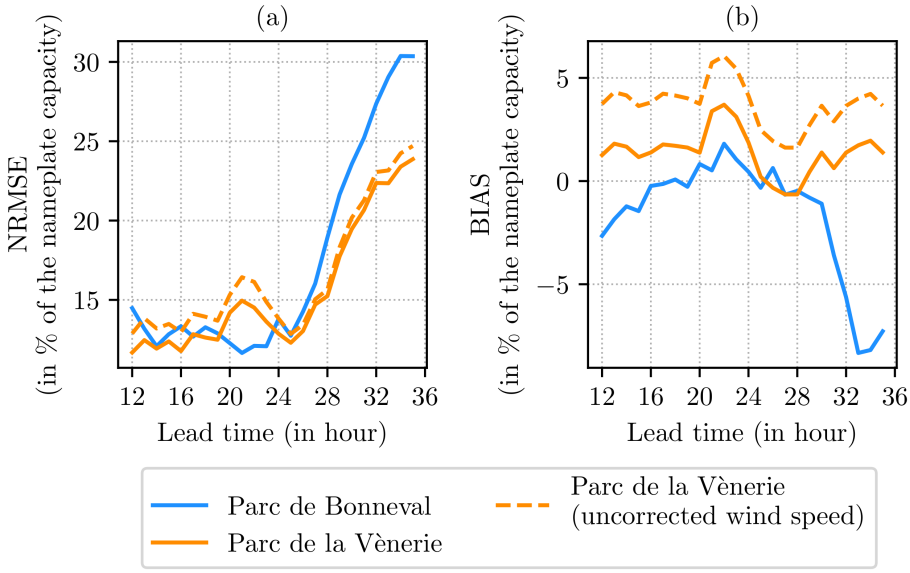


Figure 5.2 | Errors between ECMWF forecasts and the real production from 12 h to 24 h. Panel (a) shows the NRMSE in % of the installed capacity, and panel (b) shows the bias in % of the installed capacity. For the forecasts, we retrieve the wind speed forecasted by ECMWF, and using the power curve, we compute the wind energy forecasts. As ECMWF data are hourly forecast up to 24 h, for the forecasts from 24 h to 35 h, we use the last value forecasted by ECMWF as an estimator. Errors for Parc de Bonneval and Parc de la Vènerie are shown over the year 2015 at a rate of one forecast per day.

Now, figure 5.3 shows the monthly BIAS for the first lead time (12 h ahead) and the last lead time (35 h ahead). Panel (a) shows the results at Parc de Bonneval, and panel (b) shows the results at Parc de la Vènerie. First of all, for both panels, there is no significant differences in the trend between the first and the last lead time. However, in terms of amplitude, we can see that the BIAS is significantly higher at Parc de Bonneval for the last lead time. Moreover, for this farm, there is a clear seasonal cycle. The BIAS tends to be positive in winter while it is negative in summer. In Parc de la Vènerie, there is no seasonal cycle or significant difference between the first and the last lead time. In both cases, the BIAS tends to be positive.

In any case, the objective of this study is the economic value of the short term forecasts. Consequently, the income from the day ahead market is used as an approximative estimator, but

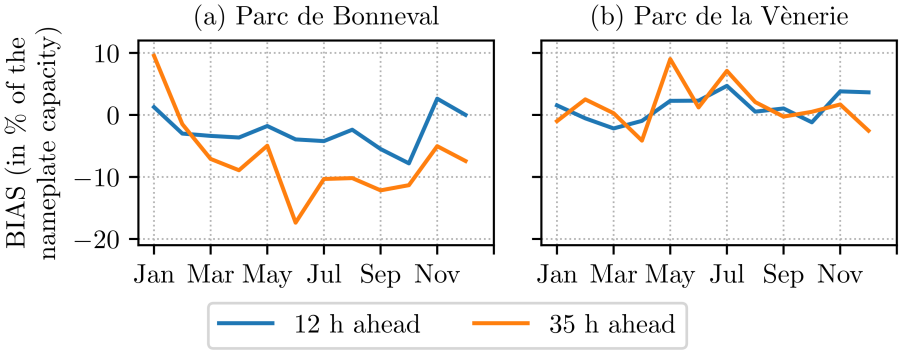


Figure 5.3 | Bias ECMWF forecasts and the real production 12 h ahead and 35 h ahead. The bias is monthly averaged, and it is in % of the installed capacity. Panel (a) shows the monthly bias at Parc de Bonneval, and panel (b) shows the monthly bias at Parc de la Vènerie.

its real value is not an important point. However, the second step is the crucial issue of this work. Indeed, now we have to compute the income from the intraday and balancing markets. As said previously, for each delivery date, we have access to every transaction made on intraday and to their associated price. Then when balancing, a purchase or sale price corresponding to an offer available at that time will be considered.

From this point, we consider 3 case studies:

1. The case where we do not have access to short term forecast. In this case, the difference between the production sold on the day ahead market and the real production is directly balanced a posteriori via imbalance.
2. The case where we have access to a perfect short term forecast. Here, we can anticipate the difference between the sold production and the real one, thanks to the short term forecast. This quantity is balanced on intraday. As the short term forecast is perfect, there is no difference between the real production and the already sold production.
3. Finally, we consider the realistic case with a realistic short term forecast. As for case 2, we balance the difference between the production sold on the day ahead market and the production forecasted by a short term forecasting model on intraday. However, in this case, as the short term forecast is no longer perfect, the quantity already sold and the real production is not the same. Consequently, this difference is balanced a posteriori via imbalance. In this case, we consider four sub-cases corresponding to four different short-term forecasting models. Each model has been described in chapter 2. There are the best downscaling model $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$, its equivalence without adding the error at t_0 $\mathrm{LR}_{\mathrm{SW}}^{\text{no-obs}}$, ECMWF forecast, and the persistence. In each case, it is assumed that the balancing on intraday is made 30 min before the delivery date.

The results of the different case studies for 2015 are shown in figure 5.4. Panel (a) shows the results for Parc de Bonneval, and panel (b) shows the results for Parc de la Vènerie. In both cases, the best scenario, i.e. when the income is the highest, is the second one (perfect short term forecast). For the first test case, (no short term forecast), it is the case where the income is the lowest at Parc de Bonneval. At Parc de la Vènerie, this scenario is extremely close to the realistic forecast where the model is ECMWF. As shown in chapter 2, ECMWF performs very poorly at

Parc de la Vènerie. The forecasts are provided at 100 m, and even with the correction shown in equation 5.3, there is still a risk of overestimating production. With such a significant bias, balancing via imbalance consists mainly of buying the lack of energy. This results in such a low income.

For the other cases, the hierarchy between the models is consistent with chapter 2 in both sites. The more efficient the model, the higher the income. The best being LR_{SW}^{obs} and the worst being ECMWF. The difference in the amounts between the two wind farms can be explained by their respective production. Even if the turbines in Parc de la Vènerie have a nominal power of 2300 kW while the turbines in Parc de Bonneval have a nominal power of 2000 kW, Parc de Bonneval is composed of six turbines while Parc de la Vènerie is composed of four turbines. Consequently, the total production from Parc de la Vènerie is weaker than that of Parc de Bonneval. The same is also valid for the income.

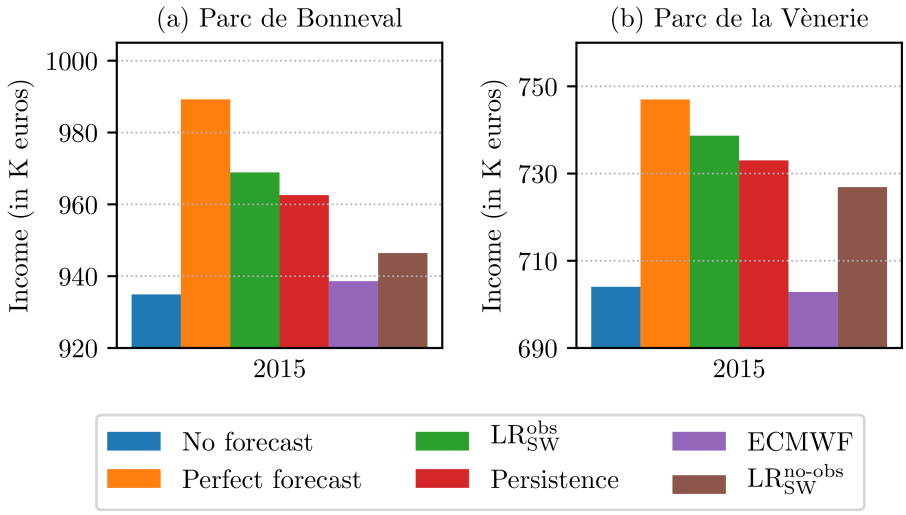


Figure 5.4 | Global income for 2015 for Parc de Bonneval (panel (a)) and Parc de la Vènerie (panel (b)). For both panels, the three cases described above are shown. In blue is the case where no short term forecast is available (case 1). In orange is the case where the short term forecast is perfect (case 2). The four last bars correspond to the third case, where the short term forecast is a realistic one. The four subcases correspond to different short term forecasting models all introduced in chapter 2. There are LR_{SW}^{obs}, persistence, ECMWF, and LR_{SW}^{no-obs}. The scale is different from one figure to another.

What interests us is not the amount of income but rather the difference between each. If we focus on the best forecasting model LR_{SW}^{obs}, for Parc de Bonneval, the difference between the realistic case and the *no forecast* case is higher than 30000€, and the difference with the perfect case is around -20000€. For Parc de la Vènerie, the difference with the *no forecast* case is around 35000€, and with the perfect case, it is less than -10000€.

Each global income can be divided into three parts. Each of them corresponds to a market, and the global income shown in figure 5.4 is the sum of the three. As said previously, the income from the day ahead market corresponds to a medium term forecast, and that is not the purpose of this work. However, the incomes from intraday and imbalance depend on the short term forecast. If the income on intraday is mainly driven by the medium term forecasts since it is a correction of this forecast, the time when we balance might be optimized.

5.3 Impact of the short term balancing strategy

The objective of this section is to investigate the impact of the balancing time on intraday. Figure 5.5 shows the cumulative incomes at Parc de Bonneval in 2015 for four different balancing times (30 min, 1 h, 2 h, and 3 h before the delivery date). Although it is possible to balance from the previous day at 15:00 the objective here is to quantify the contribution of a short term forecast model. Moreover, there is much less order available during the first opening hours. Since the maximum lead time of LR_{SW}^{obs} is 3 h, we do not consider any previous lead time.

Figure 5.5a shows the results at Parc de Bonneval when the short term forecast is provided by LR_{SW}^{obs} while figure 5.5b shows the results at Parc de Bonneval when the short term forecast is perfect that is to say when it is equal to the real production. In both cases, the general trend is an increase throughout the year. To be more precise, we can see that the income tends to decrease during winter and then increase during summer. This can be explained by the sign of the BIAS shown in figure 5.3a. The positive BIAS in summer implies that the balancing mainly consists of buying a lack of energy while the negative BIAS in summer implies that the balancing consists of selling a surplus of energy. In any case, the variability during the year is very high. Small increases and decreases (gains and losses) occur over the year. So that, when the forecast is provided by LR_{SW}^{obs}, the maximum income is reached the 3rd of November between 03:00 and 11:00 depending on the balancing time. For the perfect forecast, it is reached the same day at 03:00. In general, the income trend in Parc de Bonneval is rather intuitive.

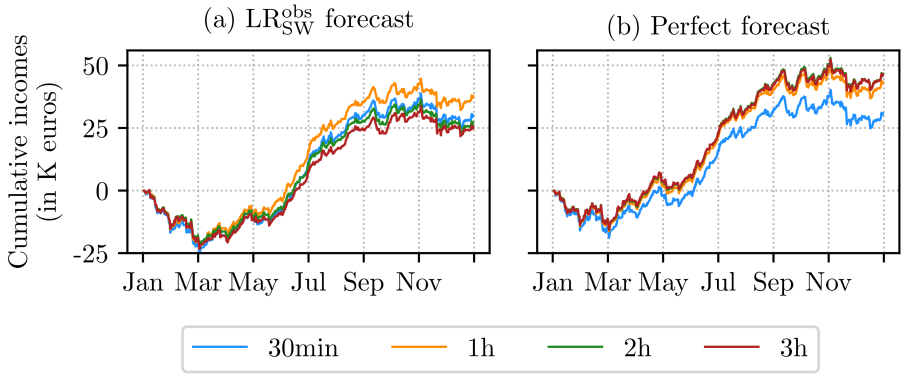


Figure 5.5 | Cumulative incomes at Parc de Bonneval in 2015 for four different balancing times (30 min, 1 h, 2 h, and 3 h before the delivery date). Panel (a) displays the results when the short term forecast is provided by LR_{SW}^{obs}, and panel (b) displays the results when the short term forecast is perfect that is to say when it is equal to the real production.

However, if we now look at Parc de la Vènerie, the behavior is totally different. As for figure 5.5, figure 5.6 shows the cumulative incomes depending on the balancing time over the year 2015 at Parc de la Vènerie. Again, panel (a) displays the incomes when the short term forecast is provided by LR_{SW}^{obs}, and panel (b) shows the incomes when the short term forecast is perfect. In this case, we note a constant decrease over the year, both for LR_{SW}^{obs} forecast and for the perfect forecast. Again, the positive BIAS in figure 5.3b explains the trend.

If we put these two aspects aside for a moment, there are several remarks, similar to the two farms, that emerge. If we first look at the two panels (a), in both cases, the best scenario is when the balancing occurs 1 h before the delivery date. Then, the gaps between the other three scenarios remain small. If we now look at the two panels (b), we can see that in both cases, when

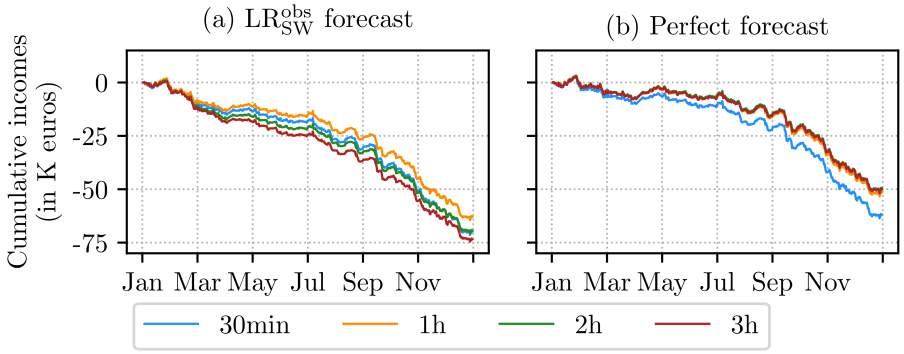


Figure 5.6 | Same than figure 5.5 for Parc de la Vènerie.

the balancing occurs 30 min before the delivery date, the income is minimized. For the other three scenarios, the differences are barely noticeable. However, for these panels, (b) it is important to remind that the forecast is the same for all scenarios (this is a perfect forecast corresponding to the real production), then the differences are only due to price variability.

Thus, we understand here the difficulty of analyzing these cumulative incomes and the impact of the balancing strategy. Indeed, the price volatility and the difference between the two forecasts (medium and short term) can explain part of the incomes.

5.3.1 Impact of the forecasting errors

First of all, we focus on the forecasting errors. Here, this is not about the model's performance, obtained by comparing the forecasted and the real production. The amount of energy which is sold or bought on intraday, is the difference between the amount already sold on day ahead the day before (obtained from a medium term forecast), and the amount forecasted by the short term forecast. Hereafter, this quantity will be noted $\varepsilon^\star$ and it is defined in equation (5.4):

$$\varepsilon^\star = P_{short-term} - P_{already-sold} \tag{5.4}$$

where, $P_{short-term}$ is the production in MWh forecasted by the short term forecast model and $P_{already-sold}$ is the energy in MWh that has already been sold on the day ahead market. Then, when $\varepsilon^\star > 0$, this means that it is necessary to sell the surplus and when $\varepsilon^\star < 0$ it is necessary to buy the lack.

This $\varepsilon^\star$ defines the general trend of the cumulative incomes shown in figures 5.5 and 5.6. If it tends to be negative, it is therefore necessary to buy a lack of energy most of the time. This case results in a cumulative income that tends to be negative. On the contrary, if it tends to be positive, it means that, most of the time, it is necessary to sell a surplus of energy, and it results in a cumulative income that tends to be positive. Table 5.1 displays the percentage value of $\varepsilon^\star$ exceeding several thresholds. Four thresholds are considered: $\varepsilon^\star < -2$ MWh (the highly negative errors), $\varepsilon^\star < 0$ MWh (the negative errors), $\varepsilon^\star > 0$ MWh (the positive errors) and $\varepsilon^\star > 2$ MWh (the highly positive errors). Results are shown for Parc de Bonneval in table 5.1a and for Parc de la Vènerie in table 5.1b. In both cases, the first four rows displays the results for the four balancing strategy (30 min, 1 h, 2 h, and 3 h before the delivery date) when the forecast is provided by LR$^{obs}_{SW}$. The last row displays the results when the forecast is provided by the perfect forecast.

First of all, in Parc de Bonneval, we can see that there are about 20% more positive errors than negative errors. In addition, highly positive errors are about 2% more frequent than highly negative errors. This is the case for all balancing scenarios, and also for the perfect forecast. Thus, in most cases, the balancing is done by selling an excess on intraday. We expect to have a positive final income from intraday, and this is the case, as shown in figure 5.5.

If we now look at the results for Parc de la Vènerie in table 5.1b, we can see that the share of negative errors and positive errors is more or less the same. However, the share of highly negative errors is almost 5% higher than the share of highly positive errors. Here, in most cases, the balancing consists of buying a lack of energy on intraday. In this condition, we expect to have a negative income. Again we can see in figure 5.6 than this is confirmed.

In both cases, we can see that the balancing scenario barely impacts the percentage, and the income trend is mainly driven by $\varepsilon^\star$.

(a) $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ for Parc de Bonneval

	$\varepsilon^\star < -2$ MWh occurrence rate (in %)	$\varepsilon^\star < 0$ MWh occurrence rate (in %)	$\varepsilon^\star > 0$ MWh occurrence rate (in %)	$\varepsilon^\star > 2$ MWh occurrence rate (in %)
30 min	8.1	40.2	54.0	11.3
1 h	8.1	39.6	54.7	11.1
2 h	7.5	39.9	54.4	9.3
3 h	7.2	40.7	53.7	9.2
Perfect forecast	9.7	40.8	53.6	13.2

(b) $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ for Parc de la Vènerie

	$\varepsilon^\star < -2$ MWh occurrence rate (in %)	$\varepsilon^\star < 0$ MWh occurrence rate (in %)	$\varepsilon^\star > 0$ MWh occurrence rate (in %)	$\varepsilon^\star > 2$ MWh occurrence rate (in %)
30 min	7.7	50.8	49.2	3.7
1 h	7.8	51.4	48.5	3.5
2 h	7.2	52.6	47.3	2.7
3 h	7.0	53.2	46.7	2.3
Perfect forecast	8.4	49.6	50.1	4.9

Table 5.1 | Occurrence rate of the medium term error $\varepsilon^\star$, defined in equation (5.4), above or below several thresholds. $\varepsilon^\star$ is defined as the difference between the production forecasted by the short term model and the production already sold on the day ahead market. Percentage for the highly negative errors ($\varepsilon^\star < -2$ MWh), the negative errors ($\varepsilon^\star < 0$ MWh), the positive errors ($\varepsilon^\star > 0$ MWh) and the highly positive errors ($\varepsilon^\star > 2$ MWh) are shown. Table 5.1a shows the results for Parc de Bonneval and table 5.1b shows the results for Parc de la Vènerie. In both cases the results are shown for four balancing times (30 min, 1 h, 2 h and 3 h before the delivery date) for the $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ forecasts. The last row corresponds to the perfect forecast (which is the same regardless the lead time).

These results make sense regarding figure 5.2. Panel (b) of the figure shows that the bias between the medium term forecast and the measurement, tends to be positive at Parc de la Vènerie (meaning that the measurement is lower than the forecast and leading to an underestimation or a lack). This is the opposite at Parc de Bonneval with a bias that tends to be negative, especially after 30 h, where it is strongly negative (meaning that the measurement is higher than the forecast

leading to a surplus). Even if $\varepsilon^\star$ is defined as the difference between the medium term forecast and the short term forecast and not the real production, the short term forecast is relatively close to the measurements with a very low bias, then we could expect a similar trend.

One way to increase the income and more precisely to make the final income from intraday at Parc de la Vènerie positive, would be to play on the quantity sold on the day ahead market. Usually, too much energy is sold on the day ahead market, which results in buying a lack of energy on intraday. This leads to a negative final income from intraday. For instance, if only 95% or 90% of the medium term forecast is sold on the day ahead market, it can decrease the number of cases where there is a lack, and increase the number of cases where there is a surplus. Then the intraday final income would be positive. Table 5.2 sums up these results. According to figure 5.6, the best balancing scenario for $\mathrm{LR}_{\mathrm{SW}}^{\mathrm{obs}}$ forecast at Parc de la Vènerie is 1 h before the delivery date. We then considered this scenario and compared both the income from intraday and from day ahead.

As a reference, the first row corresponds to the case where the entire quantity predicted is sold on day ahead. Then, the results of the cases where 95%, 90%, 85%, and 80% of the prediction is sold on day ahead, are shown. The final income from intraday (second column) and also the final income from day ahead (third column) are displayed. We can see that even with only 95% of the forecast sold on the day ahead market, the final intraday income starts to be positive. It is up to more than 200000€ when only 80% of the forecast is sold. However, this leads to a significant decrease in the final income from day ahead as the sold quantity is smaller. Moreover, if we add the two incomes, we see that despite the fact that the intraday income is negative, selling the entire forecast on day ahead remains the scenario that maximizes the total income.

Quantity sold on day ahead (in % of the medium term forecast)	Intraday income (in €)	Day ahead income (in €)
100	-62861	808557
95	9444	718850
90	79366	630703
85	146631	544636
80	210186	462035

Table 5.2 | Final incomes from day ahead and intraday depending on the quantity sold on the day ahead market. Several cases are considered: when the entire forecast is sold on day ahead, when 95% is sold, when 90% is sold, when 85% is sold and when 80% is sold. Results for Parc de la Vènerie are shown.

5.3.2 Impact of the price volatility

In finance, the price volatility is just the degree of variation of the price. It is associated with the risk. When volatility is low, prices fluctuate slightly around an average value. In this situation, the most likely scenario is, therefore, to obtain an average price, and the associated risk is low. On the other hand, when volatility is high, there is a strong variability around the mean. Here it is possible to obtain very good prices (low purchase price and high sale price) just as it is possible to obtain very bad prices (high purchase price and low sale price). In such situations, the risk associated with high volatility is therefore high.

Generally speaking, the closer the delivery date is, the more the volatility increases. To illustrate this, we pick a random day, and we consider a delivery date at 18:00 during peak hours. Then, the balancing market is open, from the day before at 15:00, to this day at 17:30. If we now look at the

standard deviation of sale prices from the opening of the market to midnight, it is 0.53€/MWh. But the standard deviation of prices between 14:00 and 17:30 is 3.55€/MWh. There is a factor of seven. This volatility is also explained by the number of transactions. The closer is the delivery date, the more they are. For example, for the computation of the first standard deviation, there are 63 transactions over a period of 9 hours, but for the second one, there are 1044 transactions over a period of 3 hours and a half.

Another way to illustrate this volatility, is no longer to look at the standard deviation of transactions for a given delivery date, but to look at the standard deviation of prices over the year, depending on the balancing time. As said previously, for the intraday prices, we consider a purchase or sale price available at the time we want to balance. Then for a given balancing time, we can retrieve a time series of selling and buying prices over the year corresponding to transactions available at this balancing time for each delivery date. Table 5.3 shows the standard deviations of those time series for each considered balancing time. The left column is for selling prices, and the right column is for buying prices.

What we can see is that again, the closer the delivery date is, the higher the standard deviation. It goes from 12.2€/MWh 3 h before the delivery date to 14.5€/MWh 30 min before the delivery for the selling prices, and from 12.4€/MWh to 16.4€/MWh for the buying prices. Moreover the increase accelerates as the delivery date approaches. Between 3 h and 2 h, there is an increase of 0.30€/MWh for the selling prices and of 0.40€/MWh for the buying prices. But between 1 h and 30 min, there is an increase of 1.5€/MWh for the selling prices and 2.70€/MWh for the buying prices.

The difference in magnitude with the standard deviations calculated on transactions for the same delivery date, can be explained easily. For a given delivery date, the orders tend to be aligned with each other. For example, if a seller proposes a much higher price than others, he has very little chance of finding a buyer. Thus the standard deviation will be in the order of a few cents to a few euros. On the other hand, if we look at the standard deviation of prices over the year, we are now interested in something else. The inter-annual variability of prices is much larger since it is affected by peak and off-peak hours on a daily basis or, on a seasonal basis, with significant demand in winter due to heating, for instance. Thus, this standard deviation will be more in the order of about ten euros.

Balancing time	$\sigma_{\text{Selling prices}}$ (in €/MWh)	$\sigma_{\text{Buying prices}}$ (in €/MWh)
30 min	14.5	16.4
1 h	13.0	13.7
2 h	12.5	12.8
3 h	12.2	12.4

Table 5.3 | Standard deviation σ of intraday prices for each considered balancing times (30 min, 1 h, 2 h and 3 h before the delivery date). The standard deviations are computed using the prices over 2015. The left column corresponds to the selling prices and the right column corresponds to the buying prices.

Thus, the later the balancing is, the higher the associated risk. This can result in a higher gain than what we would have had by balancing earlier, just as it can result in a much lower gain.

The effect of this volatility can be seen in panels (b) of figures 5.5 and 5.6. Indeed, they illustrate the cumulative incomes at Parc de Bonneval and Parc de la Vènerie when the short term forecast is provided by a perfect forecast. Then, the forecast is the same for all balancing scenarios.

The amount of energy which is sold or bought is the same for each scenario. Then the differences between the scenarios are only due to the price volatility. The impact of the significantly higher volatility for a balancing 30 min before the delivery date is obvious. For both Parc de Bonneval and Parc de la Vènerie, this scenario leads to a significant decrease in the final income from intraday. At Parc de Bonneval, this difference is around 5100€, and at Parc de la Vènerie, it is around 7700€. For the three other scenarios, the differences are very small, especially at Parc de la Vènerie, where they are barely visible (few hundred euros). At Parc de Bonneval, we can distinguish the balancing scenario corresponding to 1 h before the delivery date that stands out very slightly compared to the other two. But nothing compared with the balancing 30 min before the delivery date.

To conclude this section, it is obvious that if we set aside the forecast errors that decrease as the delivery date approaches, and focus only on the price volatility, a balancing 30 min before the delivery date is the worst scenario. A balancing 2 or 3 hours before the delivery date, seems to be a way to maximize the final income from intraday. However, the income from imbalance also plays an important role in the balancing strategy.

5.4 Performance of the short term forecasting model

As seen above, the income from intraday reflects the performance of the medium term model more than that of the short term model. In order to quantify the economic value of a short term model, it is, therefore, preferable to focus on income from imbalance.

5.4.1 Balacing fees

Table 5.4 shows the mean bias between the short term forecast and real production depending on the lead time. In other words, it is the average quantity that remains to be balanced via imbalance depending on when the balancing has been done on intraday. Results are shown at Parc de Bonneval (second column) and Parc de la Vènerie (third column) for the usual balancing times (30 min, 1 h, 2 h, and 3 h). For both farms, the average amount of energy remaining when a short term forecast is not available is also shown (last row).

The first thing we notice is that the bias is positive and very small for each balancing time at both farms. It is slightly higher at Parc de la Vènerie than at Parc de Bonneval. Moreover, as expected, the closer the delivery date, the lower the bias. Indeed, as shown in chapter 2, the shorter the horizon, the more efficient the forecast model is. In every case, there is a slight tendency to underestimate the production.

The last row shows the results for the scenario where no short term forecast is available. In this case, the quantity that has to be balanced correspond to the medium term error, that is why the magnitude is larger than for the other scenarios. It is positive at Parc de Bonneval and negative at Parc de la Vènerie. Again, the fact that the wind turbines measure 85 m at Parc de la Vènerie while the forecasts from ECMWF are provided at 100 m can explain this tendency of overestimation at Parc de la Vènerie (even if a power law correction is applied to these forecasts).

With table 5.4, one might think that the income from imbalance would tend to be positive since most often balancing consists of selling an excess of production. However, it is not that simple, and two reasons can explain this.

First of all, if we look at the data, one can see that the difference between a sale penalty and a purchase penalty is much larger on imbalance than the difference between the intraday sale price and the intraday purchase price. If we consider the average annual intraday selling price for a 1 h balancing, it is about 31.6€/MWh. The average annual buying price is about 34.0€/MWh. There

Balancing time	BIAS (in MWh)	
	Parc de Bonneval	Parc de la Vènerie
30 min	0.004	0.04
1 h	0.01	0.05
2 h	0.07	0.09
3 h	0.08	0.11
No short term forecast	0.18	-0.13

Table 5.4 | Average energy to be balanced (in MWh) via imbalance depending on the balancing time. Results for Parc de Bonneval and Parc de la Vènerie are shown.

is only a few euros difference. However, the average annual selling penalty on imbalance is about 32.4€/MWh, and the average annual buying penalty on imbalance is about 44.8€/MWh. In this case, we have more than 10€/MWh difference. Therefore, even with zero bias, we would have a negative imbalance income trend because the purchase penalties have much more impact than the selling penalties.

From that point, the second reason can be understood from table 5.5. This is the same as table 5.1 but for the short term forecast. For the four balancing scenarios considered (30 min, 1 h, 2 h and 3 h before the delivery date), it shows the percentage of highly negative errors (< -2 MWh), of negative errors (< 0 MWh), of positive errors (> 0 MWh) and of highly positve errors (> 2 MWh). Table 5.5a shows the results for Parc de Bonneval while table 5.5b shows the results for Parc de la Vènerie.

First, if we look at the results in table 5.5a at Parc de Bonneval, we can see that the share of positive and negative errors is very similar for all balancing times. In these conditions, the positive errors cannot compensate for the huge difference between selling and buying prices. Again, the share of large errors increase when the balancing is done early. When it is done 30 min before the delivery date, the share of highly negative errors is higher than that of highly positive errors, but their share remains very low. When the balancing is done 1 h before the delivery date, the share of highly negative and highly positive errors is very similar, but the share of negative error is higher by almost 1% than those of positive error. That, combined with the difference between buying and selling prices, suggests that this scenario will be one of the worst. Finally, when the balancing is done 2 h or 3 h before the delivery date, the share of highly positive errors is significantly higher than that of highly negative errors.

If we now look at the results in table 5.5b at Parc de la Vènerie, it would appear that the income from imbalance is slightly positive. For this farm, the share of positive errors is significantly higher than that of the negative errors. As expected, the proportion of large errors increases when the balancing is done early. Moreover, the share of highly positive errors is higher than that of highly negative errors when balancing is done 2 h or 3 h before the delivery date. We observe the opposite when the balancing occurs 30 min or 1 h before the delivery date, but the difference is less significant. However, the share of positive errors is sufficiently higher than that of negative errors to compensate.

These results suppose that the income from imbalance is lower at Parc de Bonneval than at Parc de la Vènerie. Figure 5.7 confirms it. It shows the cumulative income from imbalance for Parc de Bonneval (figure 5.7a) and for Parc de la Vènerie (figure 5.7b) for 2015. Again, the four balancing scenarios (3 h, 2 h, 1 h, and 30 min before the delivery date) are shown. As supposed, the income at Parc de Bonneval is lower than at Parc de la Vènerie. At Parc de Bonneval, the

(a) LR_{SW}^{obs} for Parc de Bonneval

	< -2 MWh occurrence rate (in %)	< 0 MWh occurrence rate (in %)	> 0 MWh occurrence rate (in %)	> 2 MWh occurrence rate (in %)
30 min	1.0	46.2	46.3	0.4
1 h	2.2	46.8	46.1	2.1
2 h	3.7	47.6	47.0	5.5
3 h	4.2	47.4	47.5	6.5
No short term forecast	9.7	40.8	53.6	13.2

(b) LR_{SW}^{obs} for Parc de la Vènerie

	< -2 MWh occurrence rate (in %)	< 0 MWh occurrence rate (in %)	> 0 MWh occurrence rate (in %)	> 2 MWh occurrence rate (in %)
30 min	0.3	43.3	56.5	0.0
1 h	0.4	45.6	54.2	0.3
2 h	0.8	45.7	54.2	1.5
3 h	1.1	45.7	54.3	2.0
No short term forecast	8.4	49.6	50.3	4.9

Table 5.5 | Occurrence rate of the short term error defined as the difference between the real production and the production forecasted by LR_{SW}^{obs}. Percentage for the highly negative errors (< -2 MWh), the negative errors (< 0 MWh), the positive errors (> 0 MWh) and the highly positive errors (> 2 MWh) are shown. Table 5.5a shows the results for Parc de Bonneval while table 5.5b shows the results for Parc de la Vènerie. In both cases the results are shown for four balancing times (30 min, 1 h, 2 h and 3 h before the delivery date). The last row correspond to the case where no short term forecast is available.

worst scenario is when the balancing is done 1 h before the delivery date (around 5000€ lower than the other scenario). At Parc de la Vènerie, the four scenarios are close to each other.

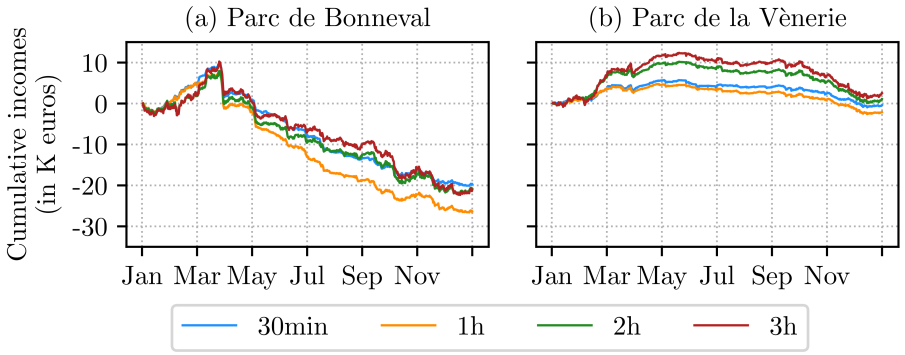


Figure 5.7 | Cumulative incomes from imbalance at Parc de Bonneval (panel (a)) and at Parc de la Vènerie (panel (b)) in 2015 for four different balancing times (30 min, 1 h, 2 h and 3 h before the delivery date).

We can see that the sign of the short term model bias has a significant impact given the

difference between the selling and the buying penalties. However, it is important to recall that the goal here is not to maximize the income but to ensure that it tends towards 0. Indeed, even if it is positive thanks to the sales penalty obtained by selling a surplus of energy, the sale of this energy on the intraday market would certainly pay more. For instance, at Parc de la Vènerie, where the income from imbalance is close to 0, the average income is -0.07€ when the balancing is done 30 min before the delivery date. If this difference could have been balanced via intraday, it would have brought in 0.56€. At Parc de Bonneval, where the income from imbalance is negative, it cost on average, -40.6€ to balance. If the balancing could have been done on the intraday market, it would have cost only -29.9€ (which makes a difference of nearly 94000€ over the year).

5.4.2 Added value of short term forecasting model

Total income

From this point on, it is clear that the balancing strategy is less important than having access to a short term forecast. Table 5.6 shows the total income (day ahead + intraday + imbalance) for 2015 at Parc de Bonneval (second column) and at Parc de la Vènerie (third column) depending on the balancing time. The case where no short term forecast is available is added (last row). Basically, the difference between the scenarios is between 1000€ and 5000€. However, the difference between having access and not having access to a short term forecast is around 35000€.

	Total income at Parc de Bonneval (in € rounded to the thousand)	Total income at Parc de la Vènerie (in € rounded to the thousand)
30 min	969000	739000
1 h	970000	744000
2 h	965000	740000
3 h	963000	738000
No forecast	935000	704000

Table 5.6 | Total income at Parc de Bonneval (second column) and at Parc de la Vènerie (third column) for 2015 depending on the balancing strategy (3 h, 2 h, 1 h or 30 min before the delivery date). For both farms, the total income of the case where no short term forecast is available is added (last row). All the incomes are rounded to the thousand.

Monthly income

Another way to quantify the economic value of short term forecasting model is to look at the monthly income. That is to say, the total income averaged over a month. Unlike the total income over a year, with the monthly income, it is possible to catch the seasonal variability. This cycle has two sources. First, the price cycle: more demand in winter because of the heating so higher prices. There is also a peak in summer, for instance, because of the cooling, but the peak is less significant. Secondly, the seasonal variability of wind and production. More wind in winter means more production, and less wind in summer means less production.

Figure 5.8 displays monthly income at Parc de Bonneval (panel (a)) and Parc de la Vènerie (panel (b)) depending on the balancing strategy (30 min, 1 h, 2 h or 3 h before the delivery date). In both cases, seasonal variability is clearly visible. The income is significantly higher in winter for both farms due to higher production. Indeed, the correlation between average monthly production

and average monthly income is significant. It is 0.93 at both Parc de Bonneval and Parc de la Vènerie. However, the difference between the balancing strategy is negligible. It is impossible to distinguish the different curves.

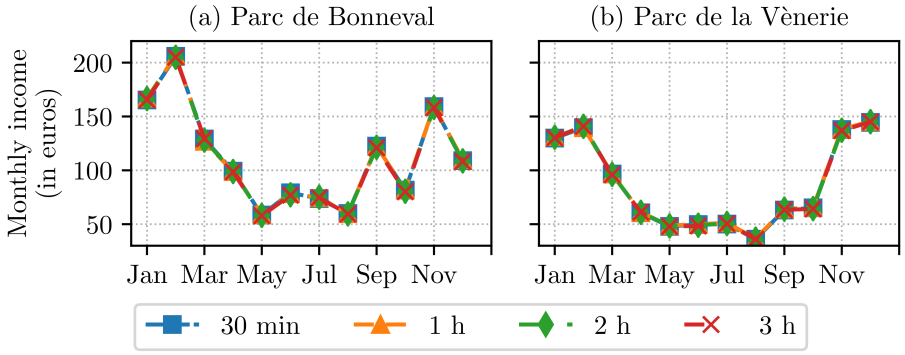


Figure 5.8 | Monthly income depending on the balancing strategy (30 min, 1 h, 2 h or 3 h before the delivery date). The monthly income is computed as the average income over each month. Results are shown for Parc de Bonneval (panel (a)) and for Parc de la Vènerie (panel (b)) for 2015.

To look at things in more detail, table 5.7 shows, for both farms and for each month, the balancing strategy that maximizes the monthly average income.

Month	Parc de Bonneval	Parc de la Vènerie
January	1 h	1 h
February	2 h	1 h
March	1 h	3 h
April	3 h	30 min
May	1 h	1 h
June	1 h	30 min
July	1 h	2 h
August	1 h	30 min
September	1 h	30 min
October	1 h	1 h
November	1 h	1 h
December	1 h	1 h

Table 5.7 | Balancing strategy that maximizes the monthly average income. Results are shown for each months and for both farms.

If the results at Parc de Bonneval are totally conclusive (balancing 1 h before the delivery date maximizes the income ten months out of twelve), this is not as clear for Parc de la Vènerie. When balancing is done 1 h before the delivery date, it leads to a maximization of the income for six months out of twelve. The second scenario that emerges is the balancing 30 min before the delivery date. Indeed it maximizes the income four months out of twelve, especially during the summer off-peak period (April, June, August, and September). The next step would be to allow a dynamic balancing strategy during the year, in particular for Parc de la Vènerie. In any case, even if we do

not have an apparent emergence of a strategy to Parc de la Vènerie, it is essential to specify that the differences between the two are of the order of 1€ per month.

To conclude this section, figure 5.9 (resp. figure 5.10) shows the monthly average income at Parc de Bonneval (resp. Parc de la Vènerie) framed by the first and the third quartiles of the incomes over each month. Results are shown for the four considered balancing times (3 h, 2 h, 1 h, and 30 min before the delivery date). First of all, for both farms, no difference can be seen between the four balancing scenarios. However, what we can see is that for both farms, the distribution is not symmetric, especially in summer, where the third quartile is closer to the mean, than the first quartiles. Moreover, the seasonal variability visible in the monthly average income is also visible with the variability. Again, for both farms, it is more important in winter than in summer.

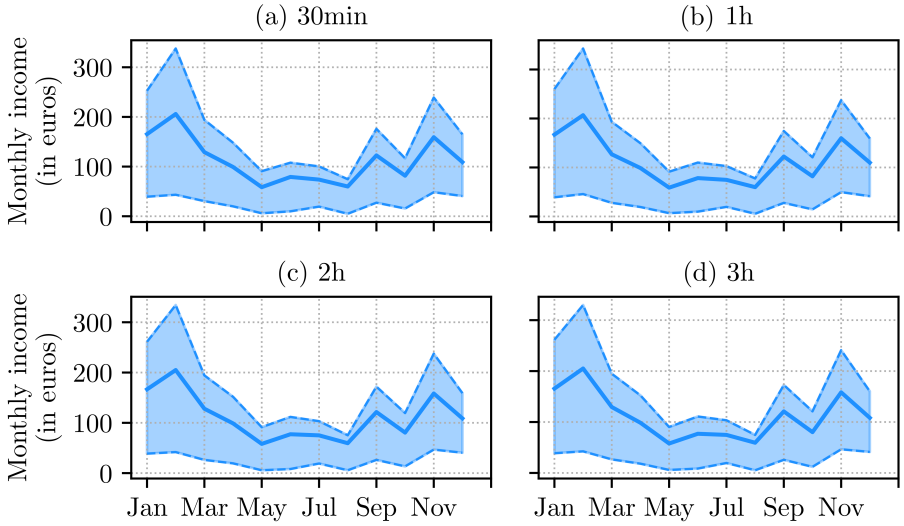


Figure 5.9 | Monthly average income and quartiles (first and third) computed over each month for the four balancing times: 3 h, 2 h, 1 h, and 30 min before the delivery date.

This can be explained by the highest variability in the production in winter than in summer. Figure 5.11 shows the monthly production for Parc de Bonneval (panel (b)) and Parc de la Vènerie (panel (c)). The monthly price from day ahead is also shown (panel (a)). For each case, the means are framed by the first and the third quartiles. We can see that for the price from day ahead (that represents almost 90% of the total income), there is no difference in the variability in summer or winter. However, the shape of the variability in panel (b) and (c) are very close to the those in figures 5.9, and 5.10. The variability of the production in winter is more important than in summer (about 50%). Then, the seasonal variability in the monthly income is driven by the seasonal variability in the production.

5.5 Conclusion

This chapter aims at quantifying the economic value of a short term forecasting model for a wind energy producer. To do so, the electricity market is simulated using real price data and production data from Parc de Bonneval and Parc de la Vènerie for 2015.

The three steps of the electricity market are studied. First of all, we use ECMWF forecasts from 12 h ahead to 35 h ahead to sell the wind energy production on the day ahead market. Then, short

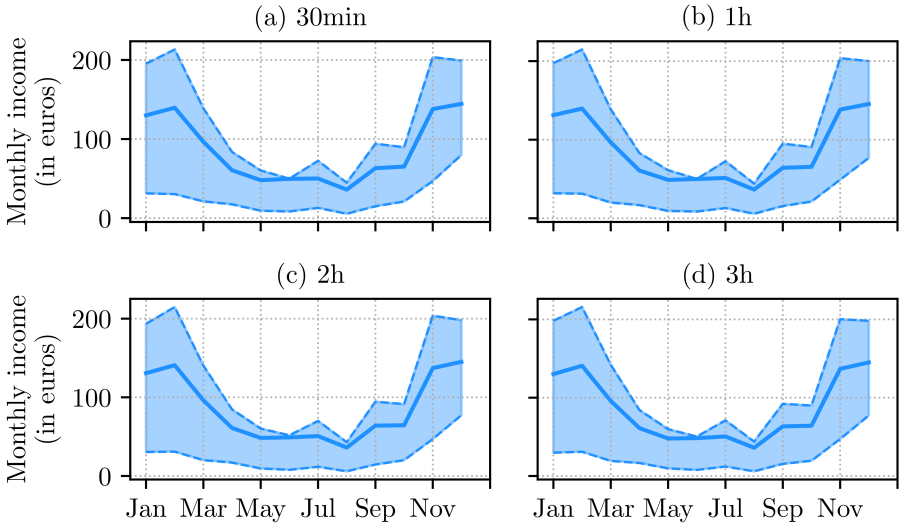


Figure 5.10 | Same than 5.9 for Parc de la Vènerie.

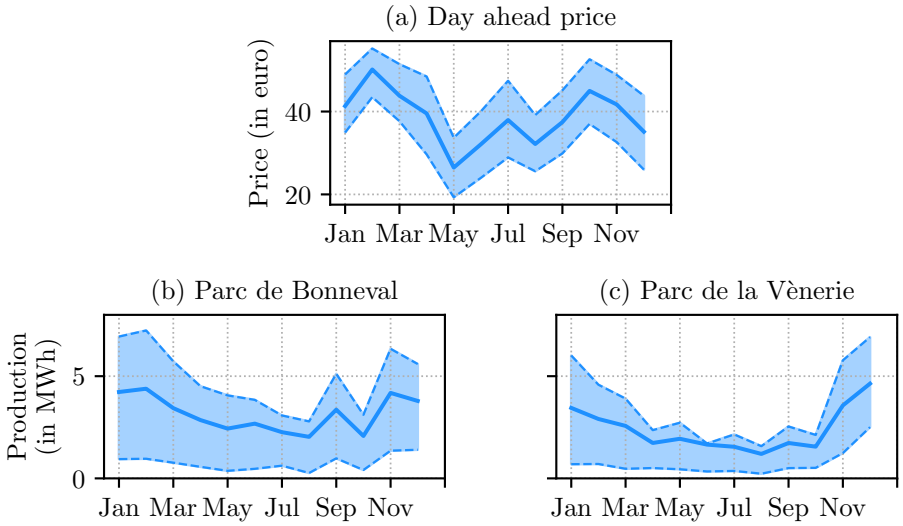


Figure 5.11 | Monthly production for Parc de Bonneval (panel (b)) and Parc de la Vènerie (panel (c)). The monthly price from day ahead is also shown (panel (a)). For each case, the means are framed by the first and the third quartiles.

term forecasting models are used for balancing on the intraday market. This market is opened up to 30 min before the delivery date. Then four balancing strategies are investigated: 3 h, 2 h, 1 h, and 30 min before the delivery date. We compare the results with the case where a perfect short term forecast is used. We show that the income from balancing via intraday is mainly driven by ECMWF forecasts. At Parc de Bonneval, these forecasts tend to underestimate the real production. This means that most of the time, the balancing consists of selling the surplus of energy. Thus, regardless of when the balancing is done, the overall income from intraday is positive. However, at Parc de la Vènerie ECMWF forecasts tend to overestimate the real production. In this case, the balancing consists mainly of buying a lack of energy, and in these conditions, the total income from

intraday is negative. In addition, price volatility is identified as a significant source of variability in total income. Indeed, volatility increases as the delivery date approaches, then the later the balancing is done, the bigger the risk taken. Typically, the results show that it is better to balance 1 h before the delivery date using an imperfect forecast model, rather than 30 min before, using a perfect forecast.

Finally, the last step of the simulation of the electricity market highlights the impact of the performance of a short term forecast model. After the delivery date, any imbalance is compensated through balancing penalties. The case where no short term forecast is available, is used as comparison. Again, the sign of the bias drives the income trend. This is especially the case as there is a significant gap between the price of the purchase penalties and the price of the sale penalties. Thus even a zero bias induces a negative total income. This is the case, for example, at Parc de Bonneval.

It is possible to estimate the economic added value of a short term forecast model. At Parc de Bonneval, which with its six wind turbines of 2 MW produces about 27500 MWh annually, the gain due to the use of a short term forecast model is about 35000€ per year. This represents more than 4% of the total annual income of the farm. At Parc de la Vènerie, the annual amount of energy produced by the four turbines of 2.3 MW of the farm is up to 21000 MWh, and the gain due to the use of the short term model is up to 40000€. This is 5% of the annual income at the farm.

To conclude, this chapter also shows the importance of the metric used in the evaluation of a forecasting system. The NRMSE is very often used to evaluate a model, but in terms of the economic impact, results show that the most important metric is the BIAS. Indeed, if positive errors are compensated by negative errors, their amplitude does not matter. This is all the more true since the gap between purchase prices and sales prices is small. Finally, it also shows the importance of an unbiased day-ahead forecast.

6

Conclusion

Contents

6.1 Synthesis and main results 116
6.2 Perspectives 118

6.1 Synthesis and main results

This thesis aims at sizing a short term wind energy forecasting system. In the context of climate change and energy transition, renewable energies are increasingly on the rise, and the need for producers to have access to forecasts is becoming more and more important. In a liberal electricity market, as in Europe, short term forecasts are essential for producers in order to balance the amount of energy sold. Indeed, 24 h before the delivery date, the producer sells his energy on the spot market. Then, thanks to accurate short term forecasts, they have the possibility to balance the quantity sold, by buying or selling, up to 30 min before the delivery date. In this context, four questions were raised, and this work aims to answer them.

How can the state of the art on wind energy forecasting can be improved for time horizons from few tens of minutes to few hours? This question is answered in two steps. In chapter 2, downscaling models for wind speed forecasting are developed, calibrated, and evaluated. Parametric models such as linear regressions and non-parametric models such as random forests are investigated. Numerical Weather Prediction model outputs from ECMWF are used as explanatory variables. Moreover, forecast errors are taken into account as explanatory variables to tune the model from the initialization. Models are tested on three different wind farms. The first two, called Parc de Bonneval and Moulin de Pierre, are very close to each other, located 100 km Southwest of Paris, and have very smooth topography. The third one, called Parc de la Vènerie, is located more to the west and has a rough topography. For each case, the hybrid configuration allows our models to overperform some of the state of the art models studied in this work such as persistence, time serie based method (ARMA), and artificial neural network (ANN). More precisely, the linear regression with variables selection algorithm provides the most accurate forecasts for the whole period (from 10 min to 3 h). The improvement over the persistence method ranges from 1.5%, 10 min ahead, to 33%, 3 h ahead for Parc de Bonneval and Moulin de Pierre, and it ranges from 0.4%, 10 min ahead, to 25.4%, 3 h ahead for Parc de la Vènerie.

Once the wind speed forecasting model is well calibrated and gives better results than state of the art, the second step is to convert these wind speed forecasts into wind energy forecasts. This crucial step is studied in chapter 3. To do so, we use a computed power curve. The power curve gives the wind power as a function of the wind speed. Although wind speed is the most important variable for wind power estimation, other variables can be taken into account to reduce the variability associated with this conversion. In this chapter, the impact of several variables, such as wind direction, or air density, is quantified. Methods to take them into account are presented. This leads to better estimations. For example, we show that the wind power forecast improvement ranges between 7.1%, 10 min ahead, and 1.3%, 3 h ahead, thanks to the more accurate conversion between wind speed and wind power forecasts.

What is gained by including available ancillary measurements as input, such as wind direction, wind variability, or temperature? This question is answered in chapter 3. As explained previously, this chapter deals with the conversion of wind speed forecasts to wind energy forecasts. The impact of several meteorological variables is investigated using in-situ real time data. This study focuses mainly on the wind direction and air density.

Regarding the wind direction, it is linked to the wake effect: when turbines are lined up, the upstream turbines decrease the flow for the downstream turbines, which produce less energy. To take it into account, several power curves are fitted to estimate the wind energy. One power curve

using only data for which the wind direction causes the wind turbine to be exposed to the wake effect, and one using the rest of the data. Taking this into account leads to an improvement in the wind energy estimation up to 45% when the turbines are lined up.

For the air density, its impact may seem less crucial, especially at mid-latitudes, where the pressure and temperature are close to the standards. Over the whole year, the improvement due to the air density accounting is up to 16%. The density is taken into account by normalizing the wind speed, before computing the wind energy through the power curve. It leads to improvement up to 40% when the atmospheric conditions (pressure and temperature) are far from the standards (temperature higher than 25°C or lower than 5°C). Such conditions occur more than 15% of the time.

Thus, considering both the wind direction and the air density leads to a reduction of the error inherent to the wind power estimation up to 30%.

Can wind energy production data from multiple wind farms improve the forecast at a given wind farm? The third question raises the issue of spatial correlation. To answer this question, a study on the impact of small scale and large scale information is conducted in chapter 4. The use of data from upstream wind farms, when the wind direction is in the correct sector, is studied for wind speed forecasting at the downstream farm.

To assess the added value of small scale information, we use data from Moulin de Pierre and Parc de Bonneval. Parc de Bonneval is located 5 km Southwest from Moulin de Pierre. At those two farms, the prevailing winds are southwest. Consequently, the data from Parc de Bonneval are used in the forecasting model fitted using data from Moulin de Pierre. First of all, measurements of the upstream farm (here Parc de Bonneval) are added as explanatory variables of a linear regression, a neural network, and a random forest to improve the wind speed forecasts at Moulin de Pierre. Second of all, these measurements are used to downscale the wind speed forecast from the farm scale to the turbine scale thanks to a second model (again linear regression, neural network, or random forest). In both cases, the use of data from Parc de Bonneval leads to significant improvement. Moreover, the linear regression provides the best results. The forecasting errors 10 min and 20 min ahead, are reduced up to 10% for the wind speed forecast at the farm scale and up to 45% for the downscaling at the turbine scale.

After that, the added value of large scale information is investigated using data from two distant wind farms called Parc de la Vènerie and Parc de la Renardière. These two farms are located 200 km West of Parc de Bonneval and Moulin de Pierre. Numerical Weather Prediction model outputs from ECMWF, and the last measurements at the location of Parc de la Renardière and Parc de la Vènerie are added as explanatory variables of a linear regression, a random forest, and a neural network. These models are compared with the best forecasting model at Parc de Bonneval, presented in chapter 2, which is the linear regression. Given the distance, the data from the two other farms do not provide relevant information, and the model that uses only data from Parc de Bonneval remains the best forecasting model, at least for the two first hours. For the last hour, the neural network gives better results.

What is the economic value of short term forecasts for a wind energy producer? An economic quantification of the value of short term forecasts is another way to evaluate the model's performance. Although forecast accuracy is the main objective of forecasters, their users are more interested in maximizing revenue from the use of predictions. The usefulness of short term forecasts is then quantified in chapter 5.

The electricity market can be split into three steps. The spot market where participants sell their energy the day before. The intra-day market, where they have the possibility to balance the sold energy by buying or selling. Finally, any imbalance between the sold energy and the real production is balanced through fines. These three steps are simulated using real price data and production data from the Parc de Bonneval and Parc de la Vènerie.

The study mainly deals with the balancing market that depends on the short term forecast. The quantity traded on this market depends both on the forecasting errors and on the price volatility. These two uncertainty sources are examined separately. To do so, we consider three test cases: when no short term forecasts are available, when perfect short term forecasts are available, and when realistic short term forecasts are available. Results show that price volatility plays a significant role. For instance, at both farms, it is better to balance 1 h before the delivery date using a realistic forecasting model than 30 min before the delivery date using a perfect forecast model. Other than that, the shape of the income from the balancing market is driven by the forecasting bias. When it is negative, the balancing consists mainly of selling a surplus of energy. Consequently, the final income tends to be positive, like at Parc de Bonneval. However, when the bias is positive, the balancing consists mainly of buying a lack of energy. In this case, the final income tends to be negative, like at Parc de la Vènerie.

In terms of total income, the use of a short term forecasting model allows an increase between 3.7% and 5.4% depending on the wind farms. It can be expected that this increase, which is already significant, will be more important for a larger wind farm. Moreover, monthly income varies widely throughout the year. From 50000€ in May, it can rise to more than 200000€ in January for one of the wind farms. This is related to the variability of production. More wind in winter means more production, and therefore, higher income.

6.2 Perspectives

All these results raise interesting perspectives.

Use of other available data Although a large number of data were used throughout this thesis, other available data remain to be used to improve the results further.

First, in the spatial correlation study, we see that data from two nearby wind farms improve the forecasts at 10 and 20 min ahead. For longer lead times, up to three hours, the use of data from very distant farms (200 km) is investigated, but no significant results are observed. Nevertheless, we have now the possibility of using data from two farms about 20 km away. Given the considered time scale, this distance appears to be a good intermediate for providing information from the upstream to the downstream farm. The application of the methods, presented in chapter 4, to the data from these two farms, would be a good way to complete the study.

Secondly, we have at our disposal very high frequency data. Those data are recorded directly at the wind turbines every 1 sec since several months. Studies show the importance of wind variability in energy estimation [86]. Therefore, it would be interesting to use those data to investigate other approaches to forecast the wind variability. First, variance predictors could be identified. Once a variance forecasting model is calibrated, the correlation between wind variance and the performance of the wind energy forecasting model could be investigated.

Forecasting uncertainty As shown in chapter 3, an estimation of the uncertainty inherent to the forecasts is calculated using confidence intervals. Thus, there are several intervals framing the

forecast in which real production has a certain probability of being. These intervals are statistically calculated over a training period of two years and depend only on the wind speed. Further work on forecast uncertainty, by improving the error model, would be a way to complete the existing model. To do so, it would be interesting to identify explanatory variables that could be good predictors of the forecast error. Parameters, such as the atmospheric stability, and the wind shear (mentioned in chapter 3), or the variability observed over the previous hour, could be considered.

More realistic simulation of the electricity market Chapter 5 presents a simulation of the electricity market in which the wind energy producer sells its energy. In practice, access to the electricity market is restricted to so-called Balance Responsible Entities. They are operators who have contractually committed themselves to finance the cost of the differences observed a posteriori between the electricity injected and the electricity consumed within a balanced perimeter. At a producer level, the Balance Responsible Entities are aggregators. The aggregators are the intermediary between electricity producers and the electricity market. It is the company that, after buying the production from a partner installation, sells it, either directly to customers, or on the market. In any case, the methodology applied in the chapter 5 remains valid. However, the numbers mentioned are not realistic since they must be transposed to the aggregator level.

Now we have data on six wind farms over several years. Consequently, it would be interesting to reproduce this study by aggregating the data from the six farms. First of all, this would make the study more realistic, at least in terms of the amounts mentioned. In addition, it would allow the impact of aggregation to be precisely quantified. In particular, to see if aggregating data from several wind farms reduces the forecast errors (differences from one farm can be offset by another) and thus, increases the final income.

Appendices

A

Stochastic Lagrangian approach for wind farm simulation

Contents

A.1 Introduction 124
A.2 Stochastic Lagrangian Models 125
A.2.1 Numerical analysis of SLM: particle approximation 126
A.2.2 Empirical numerical analysis 129
A.2.3 Particle in mesh method 133
A.3 Wind farm simulation experiment with SDM 135
A.3.1 SDM for atmospheric boundary layer simulation 135
A.3.2 Numerical simulation 139
A.4 Conclusion 145

This chapter has been published under the following reference:

Mireille Bossy, Aurore Dupré, Philippe Drobinski, Laurent Violeau and Christian Briard: Stochastic Lagrangian Approach for Wind Farm Simulation, *Proceedings of Forecasting and Risk Management for Renewable Energy, Paris, J une 7-9*, 23-44, 2017

Abstract We present a stochastic Lagrangian approach for atmospheric boundary layer simulation. Based on a turbulent-fluid-particle model, a stochastic Lagrangian particle approach could be an advantageous alternative for some applications, in particular in the context of downscaling simulation and wind farm simulation. This paper presents two recent advances in this direction, first the analysis of an optimal rate of convergence result for the particle approximation method that grounds the space discretisation of the Lagrangian model, and second a preliminary illustration of our methodology based on the simulation of a Zephyr ENR wind farm of six turbines.

A.1 Introduction

The stakes of the simulation of wind farm production are growing with the development of renewable energies. The various time scales involved (from wind potential evaluation, to short-term production forecast), the mix of various constraints on existing sites or on new projects are all issues where numerical simulations can bring quantified answers.

Although some computational fluid dynamics models, together with wind turbine models, and software are already established in this sector of activity (see eg. Sørensen [87], Niayifar and Porté-Agel [88], and the references cited therein), the question of how to enrich and refine a wind simulation (from a meteorological forecast, or from a larger scale information, eventually combined with measurements) remains largely open. This is particularly true at the scale of a wind farm, regarding the production estimation of a given site, wind turbine by wind turbine. Among various existing approaches for wind farm simulation we can distinguish

- wind extrapolation methods, and parametrization of wake effect for real-time simulation response,
- fluid and structure interaction models for wake computations, with often laminar flow hypothesis and rather simple terrain description,
- Large eddy simulation (LES) models for turbulent flows, including turbine contribution forces related to actuator disc modeling.

The turbulent nature of the atmospheric boundary layer (ABL) contributes to the uncertainty of the wind energy estimation. This has to be taken into account in the modeling approach when assessing the wind power production. This paper is devoted to a downscaling approach that typically aims to compute the wind at a refined scale in the ABL, from a coarse wind computation obtained with a mesoscale meteorological solver. This is the purpose of the Stochastic Downscaling Model (SDM) presented here.

The main features of SDM reside in the choice of a fully Lagrangian viewpoint for the turbulent flow modeling. This is allowed by stochastic Lagrangian modeling (SLM) approaches that adopt the viewpoint of a fluid-particle dynamics in a flow. Such methods are computationally inexpensive when one need to refine the spatial scale. This is a main advantage of the SDM approach, as particles methods are free of numerical constraints (such as the Courant Friedrichs Lewy condition that imposes a limit to the size of the time step for the convergence of many explicit time-marching numerical methods).

The developpment of SDM is a collaborative long term task (see [89, 90, 91] for detailed presentation), that addresses jointly mathematical and modeling issues with the elaboration of a numerical solver. It is an interdisciplinary work involving disciplines such as stochastic analysis and numerical analysis for the design and the optimal use of the Lagrangian particle solver, physics of the ABL for the calibration and validation of SDM equations and boundary conditions, and engineering for the Lagrangian adaptation of actuator disk model for the turbine wake effect.

This paper presents two recent advances in these directions:

- Section A.2 is dedicated to the convergence rate analysis of the stochastic particle algorithm used in SDM. We analyse the convergence rate with numericals experiments and check its adequacy with the theoretical optimal rate of convergence result obtained in [92] for the particle approximation method that grounds the SDM numerical algorithm.

- Section A.3 presents some first SDM simulation, by computing the wind energy production of an existing wind farm: the Parc de Bonneval operated by Zephyr ENR. With the initial and boundary conditions generated from the MERRA reanalysis, we evaluate SDM result against measurements collected at the wind farm. This numerical experiment is representative of the SDM capabilities to refine the spatial scale of the wind computation up to the scale of the wind farm: starting from the MERRA wind profile computed on a horizontal grid of 60 km by 60 km, SDM is refining the wind computation on a spatial grid of 40 m by 60 m, during a computational time interval of 24 hours.

A.2 Stochastic Lagrangian Models

Lagrangian approaches for turbulent flow are already well established for turbulent subgrid-scale modeling. This refers to the representation of the small-scales of the flow that cannot be adequately resolved solely on a computational mesh. In the context of atmospheric flow, the so-called Lagrangian Particle Dispersion Models (LPDM) are widely used for the analysis of air pollutants dispersion (see e.g. Stohl [93] and the references therein). Such method adopts perspective of a 'air parcel' by tracking a number of fictitious particles (with position X_t) released into a flow field:

$$dX_t = \overline{U}(t, X_t)dt + u(t)dt \tag{A.1}$$

where $u(t)$ is a random fluctuation of the mean velocity $\overline{U}$, given for example by a LES computation. The velocity fluctuation is modeled with stochastic differential equation (SDE) of various degrees of complexity according to the involved representations, but generally starting from the simplest Langevin model:

$$du(t) = -\frac{u(t)}{T}dt + \sqrt{C_0\varepsilon(t, X_t)}dW_t \tag{A.2}$$

where the stochastic (or fast) part of the motion is described by the 3-dimensional Brownian motion W, amplified with the turbulent pseudo dissipation of the flow ε. Stochastic description of particles in turbulent flows are also well established in the case of disperse two-phase flows and may concern many other applications (see e.g. Minier [94]).

The SDM methodology also makes use of the air parcel viewpoint. But now the mean velocity (in the particle velocity dynamics (A.1)) is not given any more but has to be computed as a statistical mean velocity $\langle \mathrm{U} \rangle$ by solving locally a Lagrangian probability density function (PDF) model. This approach relies on the so-called fluid particle approach developed in the seminal work of S. Pope ([95], see also [96] and the references therein). In this approach, a fluid-particle, or virtual fluid parcel with a position, an instantaneous velocity and a temperature state $(\mathrm{X}_t, \mathrm{U}_t, \theta_t)$ is described as the solution of a stochastic differential equation (SDE), generically of the form:

$$\begin{aligned} d\mathrm{X}_t =& \mathrm{U}_t dt, \\ d\mathrm{U}_t =& -\frac{1}{\varrho}\nabla_x \langle \mathscr{P} \rangle (t, \mathrm{X}_t)dt - G(t, X_t)\left(\mathrm{U}_t - \langle \mathrm{U} \rangle (t, \mathrm{X}_t)\right)dt \\ &+ F_t dt + \sqrt{C(t, \mathrm{X}_t)\varepsilon(t, \mathrm{X}_t)}dW_t, \\ d\theta_t =& D_1(t, \mathrm{X}_t, \theta_t)dt + D_2(t, \mathrm{X}_t, \theta_t)d\widetilde{W}_t. \end{aligned} \tag{A.3}$$

$(W, \widetilde{W})$ is a $(3d \times 1d)$-Brownian motion. From a SDE like (A.3), it is always possible to write (at least formally) the partial differential equation (PDE) of its density function, and from that

to recover the dynamics of the associated velocity field. (A.3) is in the just enough detailed form that allows to recognize/intensify the corresponding coefficients in a given targeted Navier Stokes equation combined with a chosen turbulence modeling (we refer the reader to [90] for details). Except for the mean gradient pressure term $-\frac{1}{\varrho}\nabla_x\langle\mathscr{P}\rangle$, the choice of the coefficients in the right-hand side of (A.3) corresponds to the choice of the turbulence closure. In particular, the chosen coefficients and forces in (A.3) for SDM in the ABL are described in Section A.3.1.

All computational approaches in turbulence modeling are focused on the computation of the Eulerian statistical average of the velocity and of other associated quantities. This averaging operator is classically represented by the $\langle \mathrm{U}\rangle$ in Reynolds-averaged Navier-Stokes (RANS) approaches, by $\widetilde{\mathrm{U}}$ or $\overline{\mathrm{U}}$ in LES approaches. In SDM, the Eulerian average is recovered as the probabilistic conditional expectation[1] of the particle velocity U_t, knowing that its position X_t is at point x. Denoting $\mathbb{P}$ the probability of the model (A.3), provided with expectation symbol $\mathbb{E}$, the mathematical definition of Eulerian average in SDM is:

$$\langle \mathrm{U}\rangle(t,\mathrm{x}) := \mathbb{E}\left[\mathrm{U}_t|\mathrm{X}_t=\mathrm{x}\right], \tag{A.4}$$

More generally, for any integrable function f, we set:

$$\langle f(\mathrm{U},\theta)\rangle(t,\mathrm{x}) := \mathbb{E}\left[f(\mathrm{U}_t,\theta_t)|\mathrm{X}_t=\mathrm{x}\right]. \tag{A.5}$$

Equivalently, in term of PDF approach (see [97] for further details), denoting $\gamma(t,\cdot,\cdot,\cdot)$ the probability density law of the random variable $(\mathrm{X}_t,\mathrm{U}_t,\theta_t)$, and $\rho(t,x)=\int_{3\times\mathbb{R}}\gamma(t,x,u,\theta)d\theta dx$ the renormalized mass, the statistical average also writes:

$$\langle f(U,\theta)\rangle(\mathrm{x},t)=\frac{\int_{\mathbb{R}^3\times\mathbb{R}} f(u,\theta)\gamma(t,x,u,\theta)dud\theta}{\rho(t,x)}.$$

Thus, the coefficients of the stochastic equation (A.3) are (function of, or derivatives of) statistical averages $\langle u^{(i)}\rangle$, $\langle u^{(i)}u^{(j)}\rangle$, defined as in (A.5). Here and in the sequel, we make use of the notation $\mathrm{U}_t=(u_t^{(1)},u_t^{(2)},u_t^{(3)})$.

A.2.1 Numerical analysis of SLM: particle approximation

Solution of nonlinear SDE, with coefficients depending on expectations of the unknowns, can be constructed (under some appropriated regularity hypotheses) as the mean field limit of a linear system of N-interacting particles, as N tends to infinity. Such particle approximation principle is at the basis of the SDM numerical method (see e.g. [98] for an introductory review). We detail this principle in the simplified prototype equation

$$\begin{aligned}\mathrm{X}_t &= \mathrm{X}_0+\int_0^t \mathrm{U}_s ds\\ \mathrm{U}_t &= \mathrm{U}_0+\int_0^t \mathbb{E}[b(\mathrm{U}_s)|X_s]ds+\sigma W_t,\end{aligned} \tag{A.6}$$

preferably to the complex model (A.3). In this section, we adopt a formal mathematical viewpoint to analyze numerical algorithms, and $u\mapsto b(u)$ in (A.6) is any generic function that can play role of the mean velocity field ($x\mapsto\mathbb{E}[b(\mathrm{U}_t)|\mathrm{X}_t=x]=\langle U\rangle(t,x)$), or turbulent kinetic energy,

[1] We consider here only the case of constant mass density flow, for the sake of clarity.

or more complex quantities appearing in the SDM model in (A.20), but the resulting algorithm remains similar.

Particle approximation for the solution of (A.6) relies on a statistical estimator for the conditional expectation function $x \mapsto \mathbb{E}[b(\mathrm{U}_t)|\mathrm{X}_t = x]$. Typically, an conditional estimator uses local averaging estimates on the N-particle set $(X_t^i, U_t^i, i = 1, \ldots, N, t \in [0, T])$:

$$\mathbb{E}[b(\mathrm{U}_t)|\mathrm{X}_t = x] \quad \text{is approximated by} \quad \sum_{i=1}^{N} \mathcal{W}_{N,i}(x) b(U_t^i). \tag{A.7}$$

Propositions for the weights $\mathcal{W}_{N,i}(x)$ are mainly of two kinds: the Nadaraya-Watson kernel estimator relies on a choice of a kernel function $K_\epsilon(x) = K(\frac{x}{\epsilon})$:

$$\mathcal{W}_{N,i}(x) = \frac{K_\epsilon(x - X^i)}{\sum_{j=1}^{N} K_\epsilon(x - X^j)}, \tag{A.8}$$

while partitioning (or mesh) estimator relies on a given M-partition $\mathcal{P}_M = \{\mathcal{B}_{M,1}, \mathcal{B}_{M,2}, \ldots, \mathcal{B}_{M,M}\}$ (or a mesh) of the space domain:

$$\mathcal{W}_{N,i}(x) = \frac{\mathbb{1}_{\{X^i \in \mathcal{B}_{M,j}\}}}{\sum_{k=1}^{N} \mathbb{1}_{\{X^k \in \mathcal{B}_{M,j}\}}}, \quad \text{for } x \in \mathcal{B}_{M,j}. \tag{A.9}$$

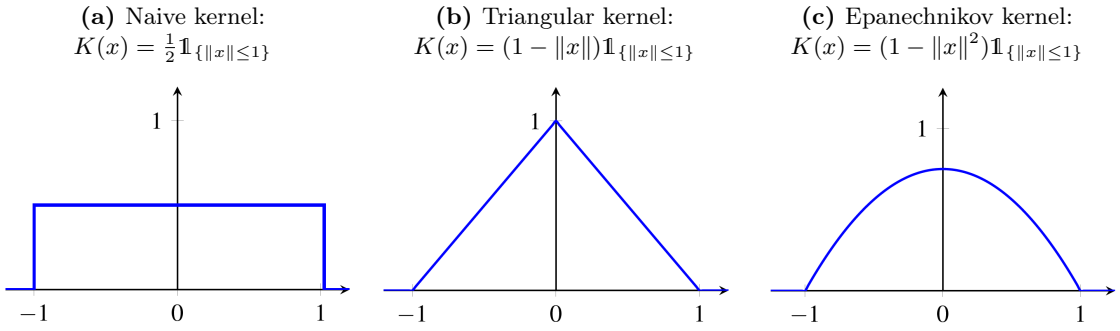


Figure A.1 | Some examples of normalized kernel functions K.

It is worth noting that the algorithm complexity of a particle system based on kernel estimator is up to $\mathcal{O}(N^2)$ whereas the partitioning estimator version is up to $\mathcal{O}(N)$ (see also Section A.2.3). We retained this last solution for SDM together with some refinement of Particle-in-cell (PIC) technics (see further details in [90]).

The convergence and precision of a particle-based numerical algorithm for solving (A.6) is driven by N the number of particles to simulate and ε the characteristic size of the partition or the characteristic size of the support of the kernel K when it is applied on particles. In [92], Bossy and Violeau prove the theoretical rate of convergence for the particle approximation of the solution of (A.6). This result gives a relationship between the two parameters N and ε in order to achieve the optimal reduction of the error (or bias). This is the first mathematical result of this kind and to make the difficulty of the mathematical analysis more affordable, the boundary conditions are assumed periodic for simplicity. In a periodic box or torus domain equal to $\mathcal{D} = [0, 1]^d$, the Lagrangian model in (A.6) becomes:

$$\begin{aligned}
\mathrm{X}_t =& \Big[\mathrm{X}_0 + \int_0^t \mathrm{U}_s\, ds\Big] \mod 1 \\
\mathrm{U}_t =& \mathrm{U}_0 + \int_0^t B[\mathrm{X}_s;\rho_s]\, ds + W_t, \quad \text{and } \rho_t \text{ is the density law of } (\mathrm{X}_t, \mathrm{U}_t),
\end{aligned} \tag{A.10}$$

where, we have written $\mathbb{E}[b(\mathrm{U}_t)|\mathrm{X}_t]$ with its equivalent mathematical form $B[\mathrm{X}_t;\rho_t]$, for $(x,\gamma) \mapsto B[x;\gamma]$ defined for all probability density function γ by:

$$B[x;\gamma] = \frac{\int_{\mathbb{R}^d} b(v)\gamma(x,v)\, dv}{\int_{\mathbb{R}^d} \gamma(x,y) dv} \mathbb{1}_{\{\int_{\mathbb{R}^d} \gamma(x,y)dv>0\}}.$$

The associated particle approximation system $((X^{i,N}, U^{i,N}), N \geq 1)$ is defined as the solution of:

$$\begin{aligned}
X_t^{i,N} &= \Big[X_0^i + \int_0^t U_s^{i,N} ds\Big] \mod 1, \\
U_t^{i,N} &= U_0^i + \int_0^t B_\varepsilon[X_s^{i,N}; \bar{\mu}_s^{N,\varepsilon}]\, ds + W_t^i, \\
\bar{\mu}_t^{N,\varepsilon} &= \frac{1}{N}\sum_{j=1}^N \delta_{\{(X_t^{j,N}, U_t^{j,N})\}} \text{ is the particles empirical measure}
\end{aligned} \tag{A.11}$$

where the kernel regression version B_ε of B, given by the approximation (A.7),(A.8), is defined for all density γ by:

$$B_\varepsilon[x;\gamma] := \frac{\int_{[0,1]^d\times\mathbb{R}^d} b(v) K_\varepsilon(x-y)\, \gamma(y,v) dy\, dv}{\int_{[0,1]^d\times\mathbb{R}^d} K_\varepsilon(x-y)\, \gamma(y,v) dy\, dv} \mathbb{1}_{\{\int_{\mathbb{R}^d} \gamma(x,y)dv>0\}}.$$

The $(W_t^i, t \leq T, 1 \leq i \leq N)$ are independent Brownian motions valued in $\mathbb{R}^d$, and independent from the initial variables $(X_0^i, U_0^i, 1 \leq i \leq N)$, independent, identically distributed with initial law ρ_0. The nonlinear model (A.10) is thus approximated with the linear system (A.11) (of dimension $2dN$), easy to discretize in time with the help of a time-discretisation Euler scheme (see below (A.16)). This algorithm is at the basis of the so-called Stochastic Lagrangian numerical algorithm (see e.g. Pope [99] and for the SDM method [90]).

The theoretical convergence analysis

In the algorithm (A.11), conditional expectation $\mathbb{E}[f(U_t)|X_t = x]$, for $f = b$, and more generally for any f measurable bounded on $\mathcal{D}$, is approximated by

$$x \mapsto F_\varepsilon[x; \bar{\mu}_t^{\varepsilon,N}] := \frac{\sum_{j=1}^N f(U_t^{j,N}) K_\epsilon(x - X_t^{j,N})}{\sum_{j=1}^N K_\epsilon(x - X_t^{j,N})}$$

the corresponding kernel approximation function, where $\bar{\mu}_t^{\varepsilon,N}$ is the empirical measure of particles as in (A.11). A pertinent criterion for the evaluation of the algorithm (A.11) is then the measure of the mean error on the conditional expectation used all along the time loop:

$$\mathbb{E}\Big|\mathbb{E}[f(U_t)|X_t = x] - F_\varepsilon[x; \bar{\mu}_t^{\varepsilon,N}]\Big|. \tag{A.12}$$

We reduce this error function by its L^1-norm on $\mathcal{D}$ weighted by the particles position distribution ρ_t), by considering:

$$\text{Error}_{L^1_{\rho_t}(\mathcal{D})} := \int_{\mathcal{D}} \mathbb{E}\Big|\mathbb{E}[f(U_t)|X_t = x] - F_\varepsilon[x; \bar{\mu}_t^{\varepsilon,N}]\Big|\rho_t(x)dx. \tag{A.13}$$

Theorem A.2.1 *(see Bossy Violeau [92]) Assume the following:*

(i) f and b are smooth and bounded functions with bounded derivatives

(ii) the kernel K is positive and bounded, with compact support in $\{x; \|x\| \leq 1\}$

(iii) the initial density law ρ_0 is smooth and bounded below by a constant $\zeta > 0$.

Then for any $T > 0$, $1 < p < 1 + \frac{1}{1+3d}$ and $c > 0$, there exists a constant C such that for all $\epsilon > 0$ and $N > 1$ satisfying $(\epsilon^{(d+2)} N^{\frac{1}{p}})^{-1} \leq c$, we have for all $1 \leq i \leq N$,

$$Error_{L^1_{\rho_T}(\mathcal{D})} \leq C\Big(\epsilon + \frac{1}{\epsilon^{(d+1)}N} + \frac{1}{\epsilon^{(d+1)p}N} + \frac{1}{(\epsilon^d N)^{\frac{1}{p}}} + \frac{1}{\epsilon^{\frac{dp}{2}}\sqrt{N}}\Big). \tag{A.14}$$

The optimal rate of convergence is achieved for the choice $N = \epsilon^{-(d+2)p}$ and

$$Error_{L^1_{\rho_T}(\mathcal{D})} \leq C N^{-\frac{1}{(d+2)p}}. \tag{A.15}$$

Notice that p can be chosen almost equal to one. The global error given in (A.14) is a combination of several sources of approximations. First, the $\mathcal{O}(\epsilon)$ term corresponds to the smoothing error for F. The $\mathcal{O}(\epsilon^{-\frac{dp}{2}}\sqrt{N}^{-1})$ term is the Monte Carlo variance contribution to the error, next $\mathcal{O}((\epsilon^d N)^{-\frac{1}{p}})$ is the error due to the replacement of the law ρ_t by the empirical measure $\bar{\mu}_t^{N,\varepsilon}$. There is also the approximation due to the replacement of the position of the exact process as the location where the conditioned expectation is computed by the position of a numerical particle. This is a part of the statistical error, (the use of the Nadaraya Watson estimator to compute the expectation) in $\mathcal{O}(\epsilon + \frac{1}{\epsilon^{d+1}N} + \frac{1}{\epsilon^{(d+1)p}N})$.

A.2.2 Empirical numerical analysis

In this section, we measure and analyse the effective convergence of the algorithm with numerical experiments in order to verify and illustrate that the claimed convergence rate in Theorem A.2.1 is optimal. For both computational time reason and clarity of the presented graphs, we limit our experiments to $d = 2$, (the wind farm simulation presented in Section A.3.1 is a fully 3 dimensional case).

Numerical experiments proceed using an Euler scheme. We decompose the time interval $[0, T]$ into M time steps of length $\Delta t := \frac{T}{M}$ and we introduce the time discretization of the interacting particle process:

$$\begin{cases} X_t^{i,N,\Delta t} &= \Big[X_0^i + \displaystyle\int_0^t U_{\eta(s)}^{i,N,\Delta t} ds\Big] \mod 1 \\ U_t^{i,N,\Delta t} &= U_0^i + \displaystyle\int_0^t B_\varepsilon[X_{\eta(s)}^{i,N,\Delta t}; \bar{\mu}_{\eta(s)}^{N,\varepsilon,\Delta t}]ds + W_t^i, \quad \bar{\mu}_t^{N,\varepsilon,\Delta t} = \dfrac{1}{N}\displaystyle\sum_{j=1}^N \delta_{(X_t^{j,N,\Delta t}, U_t^{j,N,\Delta t})} \end{cases} \tag{A.16}$$

for all $1 \leq i \leq N$ and $t \in [0, T]$ where $\eta(t) := \Delta t \left\lfloor \frac{t}{\Delta t} \right\rfloor$ is the Δt-step time function. For all time step $k\Delta t$, $0 \leq k \leq M$, each random variable $(X^{i,N,\Delta t}_{(k+1)\Delta t}, U^{i,N,\Delta t}_{(k+1)\Delta t})$ is computed from the values of all the variables $(X^{j,N,\Delta t}_{k\Delta t}, U^{j,N,\Delta t}_{k\Delta t})$, $1 \leq j \leq N$.

This algorithm has a total complexity of order $\mathcal{O}(M)\mathcal{O}(N^2)$. The major drawback of the kernel estimator method used here lies on the computation of the drift at any point x that requires a loop over all the N particles, even if they do not contribute to the final result. As we already mention, for this reason, we preferably use the alternative particle-mesh algorithm for SDM.

The test case description

We introduce some nontrival behavior in the model (A.10) by adding a potential function $P(x, y)$ that models an external, but static in time, pressure force as

$$P(x, y) = \frac{1}{2\pi} \cos(2\pi x) \sin(2\pi y) - \frac{1}{2} x, \quad \text{for all } (x, y) \text{ in } \mathcal{D} = [0, 1]^2.$$

The drift $(x, u, \gamma) \mapsto B[x, u; \gamma]$ is a mean reverting term such as:

$$B[x, u; \gamma] = \frac{\int_{\mathbb{R}^d} (v - 2u)\gamma(x, v)\, dv}{\int_{\mathbb{R}^d} \gamma(x, v)\, dv} \quad \text{for all } (x, u) \text{ in } \mathcal{D} \times \mathbb{R}^2 \text{ and all } \gamma \text{ in } \mathcal{P}(\mathcal{D} \times \mathbb{R}^2)$$

with, for all (x, u) in $\mathcal{D} \times \mathbb{R}^d$:

$$B[x, u; \rho_t] = \mathbb{E}[\mathrm{U}_t | \mathrm{X}_t = x] - 2u, \text{ when } \rho_t \text{ is the density of } (\mathrm{X}_t, \mathrm{U}_t).$$

We solve for $t \leq T = 2$,

$$\begin{cases} \mathrm{X}_t &= \left[\mathrm{X}_0 + \displaystyle\int_0^t \mathrm{U}_s ds\right] \mod 1 \\ U_t &= \mathrm{U}_0 - \displaystyle\int_0^t \nabla P(\mathrm{X}_s)\, ds + \int_0^t B[\mathrm{X}_s, \mathrm{U}_s; \rho_s]\, ds + W_t, \quad \rho_t \text{ is the density of } (\mathrm{X}_t, \mathrm{U}_t) \end{cases}$$

The initial distribution ρ_0 of $(\mathrm{X}_0, \mathrm{U}_0)$ is such that X_0 has a Gaussian distribution on $\mathbb{T}^d$ with variance σ^2 (i.e. $X_0 = \sigma Z \mod 1$, $\sigma^2 = 0.3$) and U_0 is a centered Gaussian random variable independent from X_0, with variance $\nu^2 = 1$. On Figure A.2, we represent the time evolution of the *particles mass* density $\rho_t(x) = \int_{\mathbb{R}^2} \rho_t(x, u) du$ of the process X_t distributed in the torus (plot (a)), as well as the turbulent kinetic $tke(t, x) = \frac{1}{2}\mathbb{E}[(U_t - \mathbb{E}[U_t | X_t = x])^2 | X_t = x]$ (plot (b)). We can observe that the density is clearly non uniform in space, and we expect this should put some stress on the estimation of the mean fields in low density areas.

Moreover, although starting from a Gaussian distribution, the density quickly converges in time to a stationary state and this allows to fix the final time to $T = 2$ for all the error analysis simulations, with $M = 128$ time steps. The kernel regression is performed with the Epanechnikov kernel (see Figure A.1-(c)) and $\varepsilon = \frac{1}{16}$.

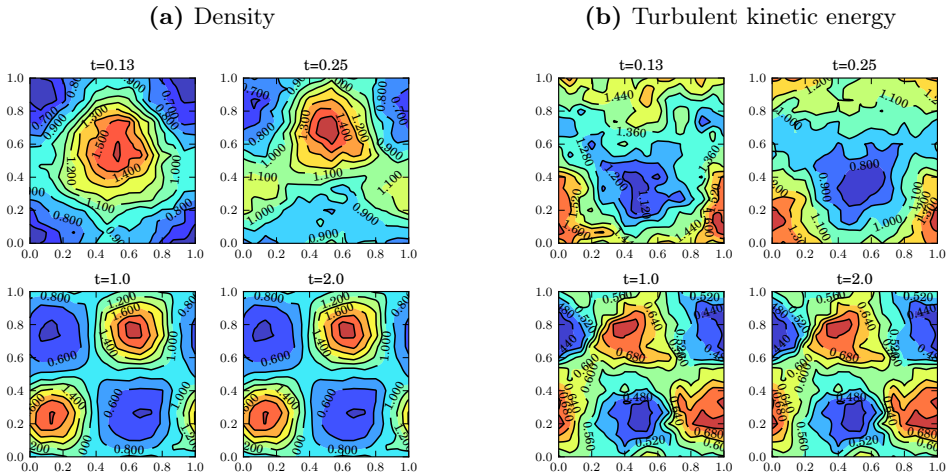


Figure A.2 | Evolution of the density and TKE for (X_t, U_t), $[N = 10^5, \varepsilon = 16^{-1}]$.

Expected L^1 error of the kernel method

We focus our attention on the *expected* L^1 *error* defined in (A.13). In order to estimate this quantity, we need to proceed with some approximations on the integral. In the following, we write $\pi^{\Delta x}(g)$ for the spline-interpolated function g on a grid with mesh size Δx. The reference numerical solution for $\mathbb{E}[f(U_T)|X_T = x]$ is approximated by the splined mean fields defined by:

$$\overline{F_{\underline{\varepsilon}}\big[x; \bar{\mu}_T^{\underline{\varepsilon},\bar{N}}\big]}^{\Delta x} := \pi^{\Delta x}(\overline{F_{\underline{\varepsilon}}\big[\cdot; \bar{\mu}_T^{\underline{\varepsilon},\bar{N}}\big]})(x) \tag{A.17}$$

for a large number of particles $\bar{N}$ and a sufficiently small window parameter $\underline{\varepsilon}$. The numerical approximation is also splined to ease the integration step:

$$F_{\varepsilon}^{\Delta x}[x; \bar{\mu}_T^{\varepsilon,N}] := \pi^{\Delta x}\big(F_{\varepsilon}[\cdot; \bar{\mu}_T^{\varepsilon,N}](x)\big) \tag{A.18}$$

The reference mass density $\rho_T(x)$ is also estimated by using the Monte Carlo mean of kernel density estimation:

$$\bar{\rho}_T(x) := \frac{1}{N_{\text{mc}}} \sum_{k=1}^{N_{\text{mc}}} \frac{1}{\bar{N}} \sum_{j=1}^{\bar{N}} K_{\underline{\varepsilon}}(x - X_T^{j,\underline{\varepsilon},\bar{N}}(\omega_k)), \quad \text{and} \quad \bar{\rho}_T^{\Delta x}(x) := \pi^{\Delta x}(\bar{\rho}_T) \tag{A.19}$$

where the ω_k represent N_{mc} independent realizations of the simulation. The computation of the integral of splined functions can be carried out very precisely over regular grids with the help of numerical libraries. All that remains is to evaluate the expected splined L^1 error by means of a Monte Carlo simulation:

$$\text{Error}_{L^1_{\rho_T}(\mathcal{D})} \sim \frac{1}{N_{mc}} \sum_{k=1}^{N_{mc}} \int_{\mathcal{D}} \Big|\overline{F_{\underline{\varepsilon}}\big[x; \bar{\mu}_T^{\underline{\varepsilon},\bar{N}}\big]}^{\Delta x} - F_{\varepsilon}^{\Delta x}[x; \bar{\mu}_T^{\varepsilon,N}(\omega_k)]\Big| \bar{\rho}_T^{\Delta x}(x)\, dx$$

In Figure A.3, we plot the expected L^1 error calculated as above as a function of the window parameter ε for different total number of particles N: for each choice of N, we observe that the error is first decreasing with the value of ε (from right to the left) toward a minimum value, but next start to increase with two small values of ε : this is the effect of the competition between the

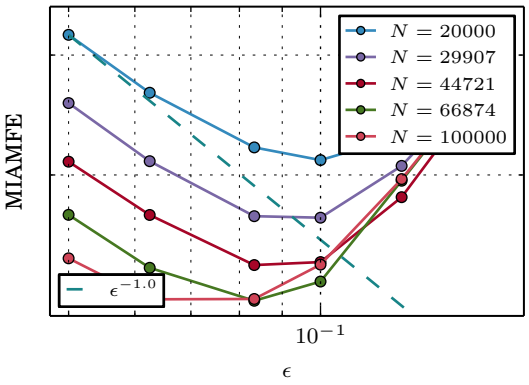


Figure A.3 | L^1 error as a function of ε for different number of particles N.

terms ε and $\frac{1}{\varepsilon^\alpha}$ in the bias formula (A.14). This is effect is delayed by choosing larger values of N who reduces the variance in the computation. We can also notice that the asymptotic slope of the error when ε tends to zero is very close to -1 for a log-log scale (represented with a blue dashed line). We expect the error to behave like $\mathcal{O}(\varepsilon + \frac{C}{\varepsilon^3 N} + \frac{C}{\varepsilon\sqrt{N}})$.

Then, it seems reasonable to infer that the term of order $\mathcal{O}(\frac{1}{\varepsilon\sqrt{N}})$ related to the variance of the stochastic integral in the model dominates the L^1 error.

Recall, however, that our theoretical analysis of the error is valid under the constraint $\frac{1}{\varepsilon^{d+2}N^{1/p}} \leq c$, for some positive constant c, so we cannot rigorously extend the bound to an asymptotic analysis when ε decreases to zero.

Finally, we can observe that the slope of L^1 is bounded by one when N is sufficiently large and ε becomes large. This is in complete agreement with the bounds in Theorem A.2.1 although this figure does not explain the relative contribution of the smoothing error and the kernel estimation error in the total L^1 error.

We can also consider the expected L^1 error as a function of $\frac{N}{\varepsilon^d}$, as in Figure A.4.

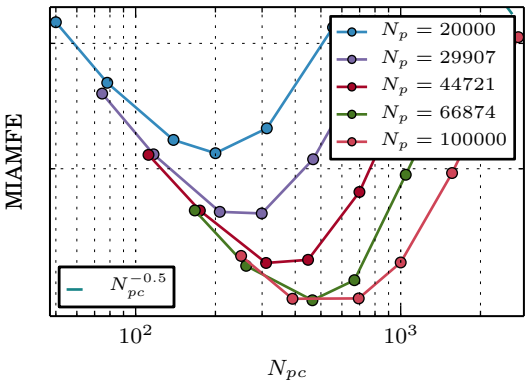


Figure A.4 | L^1 error as a function of ε for different densities of particles $\frac{N}{\varepsilon^d}$.

Note that $\frac{N}{\varepsilon^d}$ loosely represents the number of particles in interaction with a given particle (for compact support kernel functions) and is often referred to as the number of particle "per cell"

(denoted N_{pc}), especially in the case of partitioning estimates. N_p here denotes the total number of particles. This figure A.4 illustrates the concept of bias-variance trade-off and its relation with the number of particle per cell: for a given small number of particle per cell (compared to the optimal number of particle per cell), we can observe that the L^1 error is almost independent of the absolute value of ε. This clearly shows that the variance is directly related to the number of particles used in the computation of the estimator. On the contrary, when the number of particle per cell becomes large and the bias dominates, the L^1 error becomes smaller with ε.

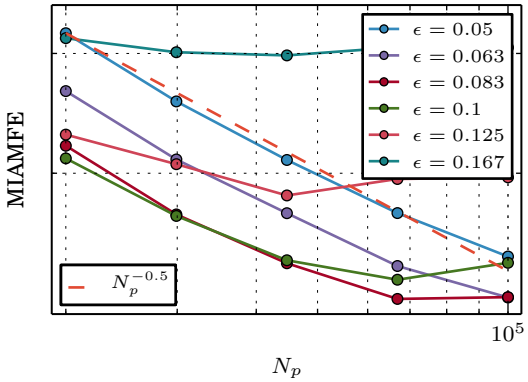


Figure A.5 | L^1 error as a function of the total number of particles, for different value of ε.

The convergence of the error with respect to the number of particles N ($= N_p$) can be observed in Figure A.5. When ε is sufficiently small, we notice as expected a convergence of order $\mathcal{O}(\frac{1}{\sqrt{N}})$, related to the reduction of the variance component of the error. On the other hand, when ε is large, increasing the number of particle does not reduce the error as the bias dominates.

Given this bias-variance trade-off, one may be interested in finding the optimal value of ε that minimizes the expected L^1 error for a given number of particles. From the simulations we ran for different couples (ε, N) of parameters, we plot the surface of the error in Figure A.6 (left). We can then plot the curve of optimal ε as a function of the number of particles which is very close to $\frac{1}{\varepsilon^3}$ (for $d = 2$). This result is in-line with what we expected from Theorem A.2.1 where the optimal value of window size is given by $N^{-\frac{1}{d+2}}$.

Moreover, if we plot the error associated with the optimal couple of parameters as a function of ε, we can observe the optimal experimental rate of convergence of the algorithm.

The theoretical optimal error (A.15) in Theorem A.2.1, is of order $\mathcal{O}(N^{-\frac{1}{4p}})$, with p close to 1, while in Figure A.6 (right), we observe a rate of order close to $-\frac{1}{4}$ to $-\frac{1}{3}$. Theoretical and observed convergence rates are here in a very good adequacy.

A.2.3 Particle in mesh method

We end this section with some experiments on the particle-mesh version of the algorithm. The principle of the Particle-Mesh methods is to aggregate the N scattered data points $(X^i, f(U^i))$, for $1 \leq i \leq N$ onto a regular mesh covering the simulation domain $\mathcal{D}$, thus reducing the size of the data set to the number of nodes in the mesh. The mean field is evaluated from the mesh charges at each particle position using standard regression techniques as in (A.7),(A.9). If we design the charge assignment and the force interpolation operation such that they can be performed in constant

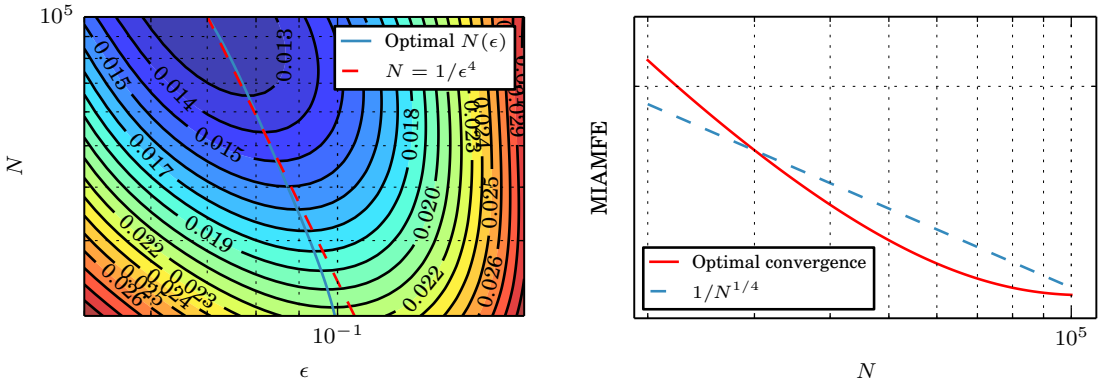


Figure A.6 | L^1 error as a function ε and N. and optimal rate of convergence for the L^1 error.

time for each particle, the Particle-Mesh algorithm has a $\mathcal{O}(M)\mathcal{O}(N)$ complexity, i.e. it has linear complexity with respect to the total number of particles. This is a tremendous improvements over the previous kernel regression method, and the speed-up is not only theoretical but is actually achieved in practical simulations.

The drawback of this approach is that it introduces new sources of numerical errors, and unlike classical particle computer simulations, increasing the number of nodes in the mesh does not necessarily reduce the error if the total number of particles is left unchanged. Moreover, refining the mesh increases the computational cost, so it is particularly important to be able to reduce the errors for a given mesh size in order to achieve the best compromise between quality and computational cost. In this regard, we will consider three charge assignment and interpolation functions that are designed to be optimal according to smoothness and spatial localization of errors criteria: the Nearest Grid Point (NGP), the Cloud in Cell (CIC), and the triangular Shaped Cloud (TSC) (see Figure A.7 for details).

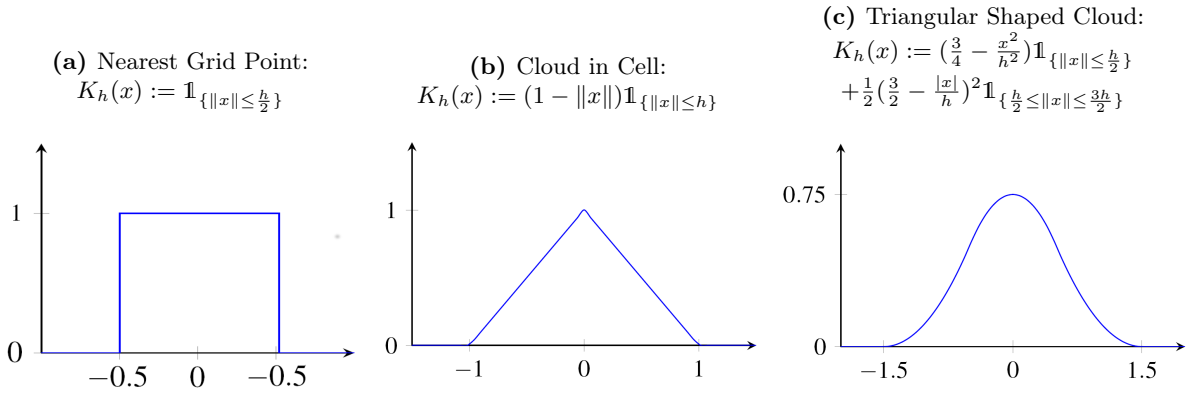


Figure A.7 | Charge assignment functions (from left to right: NGP, CIC, TSC).

Charge assignment

Consider a mesh of cell size h (also called window size). Let x_i be the position of the i-th node. Then the charge c_i and the charge density d_i assigned at node i are defined by:

$$c_i := \frac{1}{N}\sum_{j=1}^{N} f(U^j)K_h(x_i - X^j), \quad d_i := \frac{1}{N}\sum_{j=1}^{N} K_h(x_i - X^j)$$

where K is a charge assignment function. By definition of c_i and d_i the ratio $\frac{c_i}{d_i}$ is simply the kernel regression estimate at the node point x_i as in (A.7):

$$\mathbb{E}[b(\mathrm{U}_t)|\mathrm{X}_t = x] \sim \frac{c_i}{d_i} = \frac{\frac{1}{N}\sum_{j=1}^{N} f(U^j)K_h(x_i - X^j)}{\frac{1}{N}\sum_{j=1}^{N} K_h(x_i - X^j)}$$

The computation of the mesh charge values can be performed efficiently in $\mathcal{O}(N)$ with an outer loop on the particles and the use of a mesh localization procedure that makes it possible to loop only on the nodes charged by a given particle.

Of course, it is important that the localization of the particle in the mesh and the computation of the list of nodes charged by the particle be performed in constant time. In practice, the lists of neighbor cells are computed once and for all (in linear time) at the beginning of the procedure to speed up the execution of the algorithm.

In Figure A.8, we measured the influence of the regularity order of the charge assignment function K_h. Aside from the smoothing aspect of the obtained velocity field, we can observe a gap between the error produced by the partitioning estimates (corresponding to NGP assignment charge) and the higher order CIC or TSC functions, and CIC appears to be a good compromise between the error level and the ease of implementation.

A.3 Wind farm simulation experiment with SDM

Our SDM model has been evaluated against measurements collected at a wind farm located in Bonneval, a small town 100 km Southwest of Paris, France (at 48.20°N and 1.42°E). The wind farm is operated by Zephyr ENR, a private company managing five other wind farms. The Bonneval wind farm, called Parc de Bonneval, has been implemented in 2006 and is composed of six wind turbines, each with a power rated of 2.0 MW. In order to evaluate the SDM simulations with the data collected at Parc de Bonneval, wind turbines have been numerically integrated in SDM,based on an actuator disk model. This model allows the simulation of the dynamical effect of the presence of wind turbines, in the form of trailing wakes, as well as the computation of the wind energy production.

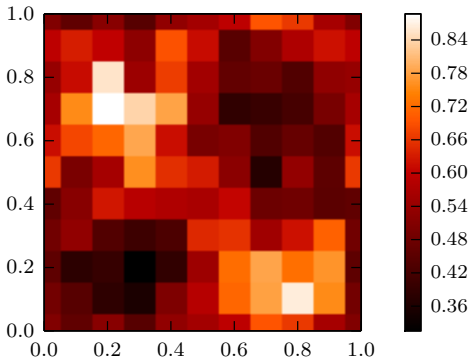
A.3.1 SDM for atmospheric boundary layer simulation

We run SDM for the winter day of December 22th 2016, with the equation (A.3) configured for the case of the neutral atmosphere hypothesis. Here and in the sequel we denote by

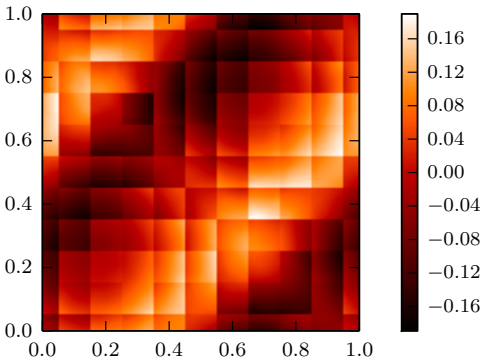
$$\mathrm{U}_t = (u_t^{(1)}, u_t^{(2)}, u_t^{(3)}) = (u_t, v_t, w_t)$$

the velocity components (with numbering or with letters, depending on how it is convenient in the equations), and for the components of the instantaneous turbulent velocity:

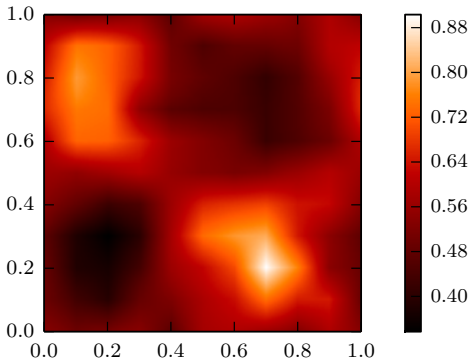
(a) Velocity norm for the NGP scheme



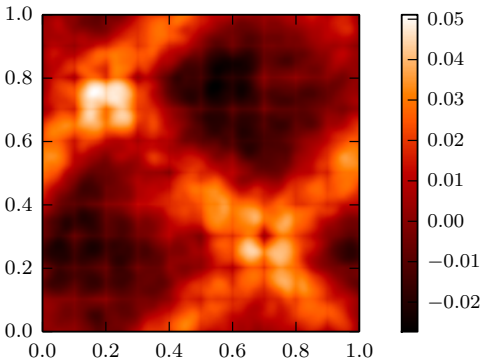
(b) Velocity error for the NGP scheme



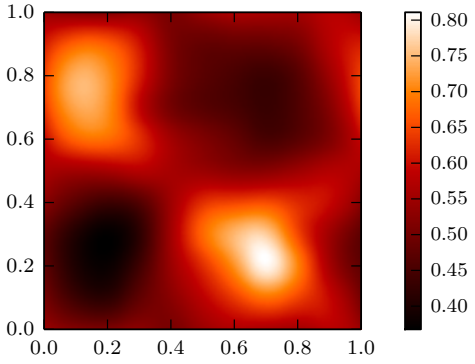
(c) Velocity norm for the CIC scheme



(d) Velocity error for the CIC scheme



(e) Velocity norm for the TSC scheme



(f) Velocity error for the TSC scheme

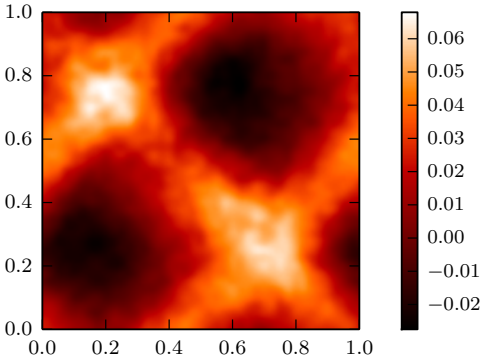


Figure A.8 | Velocity norm and average error for the NGP, CiC and TSC schemes

$$\mathrm{U}_t - \langle \mathrm{U} \rangle(t, \mathrm{X}_t) = (u_t'^{(1)}, u_t'^{(2)}, u_t'^{(3)}) = (u_t', v_t', w_t')$$

In order to elaborate the SDM model, we start from the General Langenvin model introduced by Pope [97]:

$$\begin{cases} dX_t &= \mathrm{U}_t dt, \quad \text{with } \mathrm{U}_t = (\mathrm{u}_t^{(i)}, i = 1, 2, 3) \text{ and } \mathbf{u}_i'(t) = \mathrm{u}_t^{(i)} - \langle \mathrm{u}_t^{(i)} \rangle \\ d\mathrm{u}_t^{(i)} &= -\partial_{x_i} \langle \mathscr{P} \rangle (t, X_t) dt + \left(\sum_j G_{ij} \left(\mathrm{u}_t^{(j)} - \langle \mathrm{u}^{(j)} \rangle \right) \right) (t, X_t) dt + \sqrt{C_0 \varepsilon(t, X_t)} dB_t^{(i)} \end{cases} \tag{A.20}$$

As a stand-alone PDF method, all the Eulerian statistical means needed by the SDM model in (A.20) are computed within the simulation. In the ABL, we pay great attention to the modeling of the ground effects. We incorporate to SDM a model for the effect of the wall blocking of normal velocity component (following [100], see also [91] for details). For the wind farm simulation, we further incorporate a model for the effect of pressure reflection from the surface (by adapting the Durbin elliptic relaxation method [101]). This model refinement mainly impacts the form of the (G_{ij}) relaxation tensor we use in (A.20). We shortly describe (G_{ij}), decomposing the tensor in this common basic diagonal relaxation term $\frac{1}{2}\frac{\varepsilon}{tke}$ and the more complex γ_{ij} part, decomposed itself in its near wall part $\gamma_{ij}^{\text{wall}}$ and its internal flow part $\gamma_{ij}^{\text{homogeneous}}$:

$$G_{ij}(t,x) = -\gamma_{ij}(t,x) - \frac{1}{2}\frac{\varepsilon(t,x)}{tke(t,x)}\delta_{ij}, \qquad \text{with} \quad C_0\varepsilon(t,x) = \frac{2}{3}\sum_{i,j}(\gamma_{ij})\langle u_i' u_j' \rangle(t,x)$$

$$\text{and} \quad \gamma_{ij}(t,x) = (1 - \alpha(t,x)tke(t,x))\gamma_{ij}^{\text{wall}}(t,x) + \alpha(t,x)tke(t,x)\ \gamma_{ij}^{\text{homogeneous}}(t,x)$$

$$-\gamma_{ij}^{\text{homogeneous}} = -\frac{1}{2}(C_R - 1)\frac{\varepsilon}{tke}\delta_{ij} + C_2\frac{\partial\langle \mathrm{u}^{(i)} \rangle}{\partial x_j}, \qquad \text{and} \ -\gamma_{ij}^{\text{wall}} = -7.5\frac{\varepsilon}{tke}n_i n_j$$

where n is the wall-normal unit vector. The coefficients C_0 and C_2 have to satisfy some realizability constraints (see [102], [103]). The elliptic blending coefficient α (that balances $\gamma_{ij}^{\text{wall}}$ and $\gamma_{ij}^{\text{homogeneous}}$) solves near the ground the Poisson equation:

$$L^2\nabla^2\alpha - \alpha = -\frac{1}{tke}$$

where L is a length scale defined as a maximum of the turbulent scale and the scale connected with dissipative eddies.

Finally, we make use of the Lagrangian methodology to easily introduce complex terrain description in SDM: when a fluid-particle meets the ground during the simulation, according to the wall-boundary condition, we perform a reflection of it velocity, according to the friction velocity computed as:

$$u_*(t,x) = \kappa\frac{\sqrt{\langle u \rangle^2(t,x) + \langle v \rangle^2(t,x)}}{\log(x^{(3)}/z_0(x))}$$

where the roughness length z_0 may vary with the surface terrain.

Lagrangian actuator disk model

SDM method allows some fluid and structure interaction modeling, in particular when the structure are porous objects like actuator disk models for turbine.

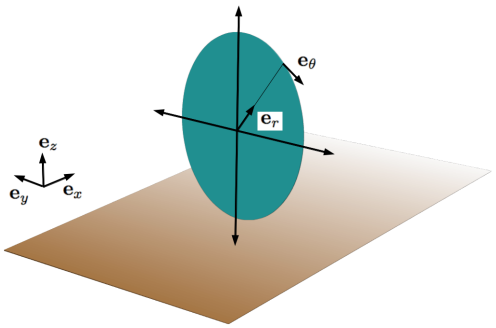
The SDM approach could be used with various actuator disk modelling options (see [91] and the references therein). In the actuator disc approach, each mill is represented as an immersed surface which concentrates all the forces exerted by the mill on the flow. In the SDM context, the presence of wind mills is taken into account thanks to an additional force f that represents the body forces that the blades exert on the flow. This force term is incorporated in the SDEs that govern the movement of the particles. To this end, Equation (A.20), which governs the time evolution of the velocity $\mathrm{U}_t = (u_t, v_t, w_t)$ of a particle, is modified as follows:

$$\begin{aligned} d\mathrm{U}_t = &-\frac{1}{\rho}\nabla_x\langle\mathscr{P}\rangle(t,\mathrm{X}_t)dt + f\,(t,\mathrm{X}_t,\mathrm{U}_t)\,dt \\ &- G(t,\mathrm{X}_t)\Big(\mathrm{U}_t - \langle\mathrm{U}\rangle(t,\mathrm{X}_t)\Big)dt + C(t,\mathrm{X}_t)dW_t \end{aligned} \tag{A.21}$$

where the term $f(t,x,U)$ represents the body forces of the turbine seen by the particle at point x with velocity U. We refer to [91] for a detailed discussion on the turbine force terms implementation in the Lagrangian context (including nacelle and mast forces).

For the simulation of the Parc de Bonneval wind farm presented hereafter, we have chosen a rather basic non rotating uniformly loaded actuator disc model. Such model can be easily parametrized with the characteristic data of thrust coefficient C_T and power coefficient C_p, provided by the turbine manufacturer, and varying with the dynamics of the inflow wind at the turbine.

(a) Local coordinates



(b) The cylinder $\mathcal{C}$

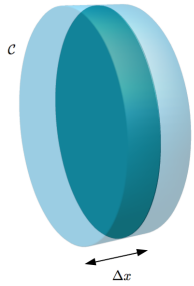


Figure A.9 | Non rotating uniformly loaded actuator disc model. (a) The local reference frame at the actuator disc of the turbine, using cylindrical coordinates; (b) The cylinder $\mathcal{C}$ that extends the actuator disc. Mill forces are applied to particles that lie inside.

We describe the force f, using the local reference frame of cylindrical coordinates centered at the hub of the turbine, with basis vectors $\mathbf{e}_x$, $\mathbf{e}_r$ and $\mathbf{e}_\theta$ as shown in Figure A.9a. Assuming that the flow moves along the positive direction of the x axis, and that the turbine's main axis is aligned with the x axis, so that it faces the wind directly, the total thrust force exerted by the turbine is formally given by (see e.g. [87])

$$F_x = -\frac{1}{2}\rho A C_T U_\infty^2 \mathbf{e}_x$$

where U_∞ is the unperturbed velocity far upstream from the turbine's location, A is the surface area of the turbine's disc, ρ is the density of air, and C_T is a dimensionless, flow dependent parameter called the *thrust coefficient*. As in Réthoré et al. [104], the local velocity magnitude U_D is used instead of U_∞ and the thrust force expression in SDM becomes

$$F_x = -\frac{1}{2}\rho A C_T U_D^2 \mathbf{e}_x \quad \text{with} \quad U_D(t) = \mathbb{E}[\mathrm{U}_t^2 | \mathrm{X}_t \in D] \tag{A.22}$$

In order to adapt this thrust force model to particles, the disc is extended to a cylinder $\mathcal{C}$ of length Δx and mass $\rho A \Delta x$ (see Figure A.9b). The force per unit mass inside region $\mathcal{C}$, and to include in (A.21), is then given by:

$$f(t,x) = -\frac{1}{\Delta x} C_T U_D^2(t) \mathbb{1}_{\{x \in \mathcal{C}\}} \mathbf{e}_x \tag{A.23}$$

The available power is computed following the same idea:

$$P(t) = \frac{1}{2}\rho A C_p U_D^3(t).$$

A.3.2 Numerical simulation

Numerical setup

The modeled domain is a 3D box, with flat ground surface and a variable roughness length inferred from Google-Earth and lookup tables of roughness lengths for typical types of land-use. Four different roughness lengths have been used with respect to the land-use pattern shown in the Figure A.10. The roughness length varies between 0.01 and 0.4 m. The characteristics of the numerical domain of the simulation and of the turbines are summarized in Table A.1.

The initial and boundary conditions are generated from the MERRA reanalysis with a 3-hourly time sampling [105]. All MERRA fields are provided on the same 5/8 degree longitude by 1/2 degree latitude grid. The data used to extract initial and boundary conditions are those of the closest grid point located at 25 km Southwest of Parc de Bonneval (48°N and 1.25°E). The vertical mesh has 72 pressure levels but only the first three levels from the surface up to 970 hPa (about 400 m) are used. The pressure level coordinates are converted into altitude coordinates using the surface pressure from the MERRA reanalysis. The wind components are then interpolated onto the refined grid of SDM. The time step of the SDM simulation is 5 s. The profiles extracted from the MERRA reanalysis at the closest grid point are therefore interpolated linearly in time with a 5 s time sampling.

Case study description

Parc de Bonneval is composed of six turbines of type Vestas V80-2.0 MW, each named by its number from 1 to 6 in Figure A.10a. The simulated study-case corresponds to the 22^{th} December 2016, a winter day, allowing neutral atmosphere approximation, and chosen for its typical wind events, producing wake effects. Figure A.11 displays the time evolution of the measured wind direction, wind speed and wind energy production at the 6 turbines. The wind speed and direction are measured directly at Parc de Bonneval by anemometers located on the hub of each turbine. The wind energy production is also provided directly from the generator.

Those time series are used to evaluate SDM model performance, with a sampling period of 10 minutes.

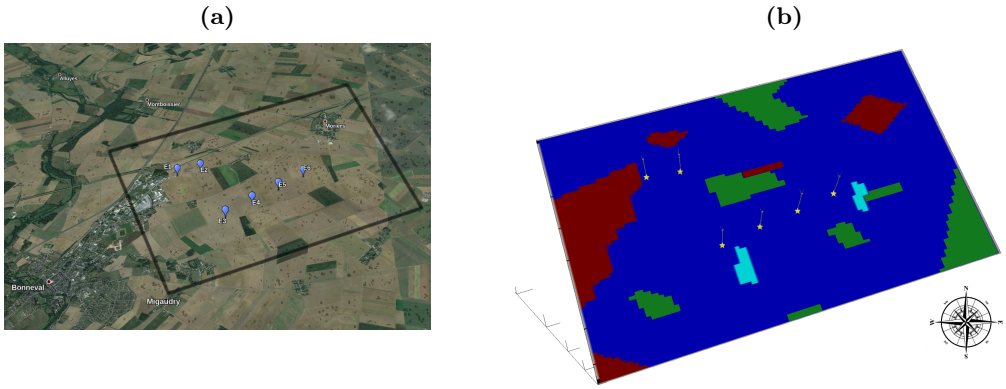


Figure A.10 | (a) Aerial view of the Parc de Bonneval from Google-Earth; (b) Aerial view of the simulated wind farm. The pattern define the roughness length. Blue part represents farmland (0.04 m), red are small town (0.4 m), green are uncut grass (0.01 m), cyan are small forest (0.15 m). Yellow stars represent the turbines.

(a) Configuration of the simulation

Simulation parameters	
Domain size x	3000 m
Domain size y	4787 m
Domain size z	408 m
75 cells in x	$\Delta x = 40$ m
75 cells in y	$\Delta y = 63.83$ m
85 cells in z	$\Delta z = 4.8$ m
Particles per cell	80
Final time is 24 h	Time step is 5 s

(b) Parameters of the mill

Mill configuration	
Hub height	100 m
Radius	40 m
Nacelle radius	4 m
Rotational speed	1.75 rad s^{-1}

Table A.1 | Main parameters of the simulation.

The chosen episode is characterized by a strong wind blowing until 5:00 local time (LT). Between 5:00 and 16:00 LT, the wind speed weakens from 10 m s^{-1} to 2 m s^{-1}. It increases again up to 6 m s^{-1} and decreases down to less than 2 m s^{-1} in 2 hours. As a consequence, the turbines production vary from 0 to almost the turbine nominal power of 2 MW during this day. Moreover, the wind shifts progressively from the South to the North. According to the position of the turbines (see Figure A.10), a wind direction around 230° lines up the turbines 3 to 6, and a direction around 250° lines up the turbines 1 and 2. We mainly chose this particular episode of December 22th, as it contains such wind event, happening between 7:00 and 9:00 LT. Indeed we can observe the wake effect in Figure A.11. The phenomenon decreases the production downstream by 50%.

Results

Figure A.12 displays the time evolution of the simulated wind direction, wind speed and wind energy production at the 6 turbines. It can be directly compared to Figure A.11. The time variability is well reproduced with a slightly increasing wind speed between 0:00 and 3:30 LT and a constant wind direction. The wind speed increases between 8 and 9.2 m s^{-1}. The simulated wind speed is slightly weaker than the measured wind speed which remains constant and equal to 10 m s^{-1} over this period of time. Such underestimation is caused by the initial and boundary

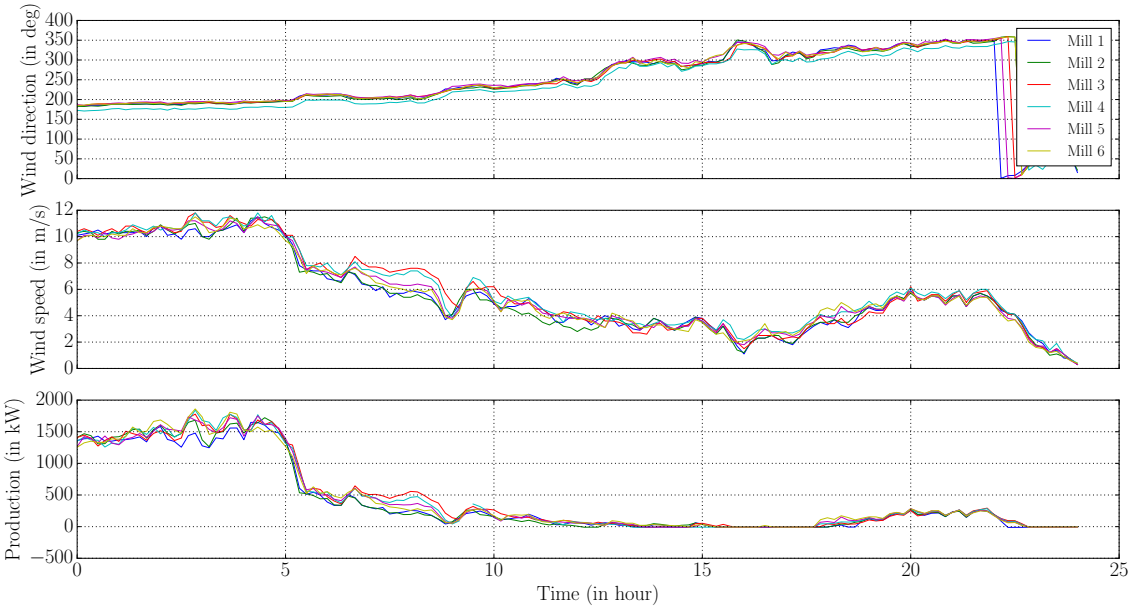


Figure A.11 | Time evolution of Parc de Bonneval measurments during the 22^{th} December 2016

conditions from MERRA reanalysis which provide a weaker wind speed at the hub height. The wind direction is also slightly biased by about 10°. The simulated wind speed then decreases at a similar rate than the observed wind speed. The short increase of the wind speed followed by a fast decrease between 15:00 and 23:00 LT is underestimated in the simulation as the wind speed peaks at about 3.4 m s^{-1} in the simulation versus 6 m s^{-1} in the measurements. The bias in wind direction disappears after 8:00 LT. Finally, we observe that the high frequency variability is much too smooth in the simulated mean velocity. We mainly impute this phenomenon to the combination of low frequency data set for the initial and boundary conditions, with the small size of the numerical domain, that induces a strong forcing by the lateral inflow boundary conditions.

However, as shown in Figure A.13, the intrinsic variability contains in the model is representative of the observations variability. Figure A.13 displays the evolution of the norm of the turbulent part of the wind $U' = U - \langle U \rangle$ between 6:30 and 9:30 LT, when turbines 3, 4, 5 and 6 are lined up. During the wake alignment period, computed and measured turbulent velocity norms are displayed at a forefront turbine (turbine 3), and at a downstream turbine (turbine 5). To this end, in SDM, we have extracted a realization of the turbulent part of the velocity, by randomly picking-up every 10 minutes, one particle velocity at the neighborhood of the rotors. Hourly moving means are computed and subtracted to its instantaneous velocity. We proceed similarly with the measured velocity.

In both case, the variability around the downstream turbine is higher than the variability around the forefront turbine. Moreover, the variability of the turbulent velocity computed in SDM is higher than the one measured at Parc de Bonneval. This can be explained by the way the instantaneous velocity is retrieved. For SDM we used an instantaneous velocity at 5 s frequency picked every 10 min. For Parc de Bonneval, the velocity measured by anemometers is at a high frequency, but then it as been averaged over 10 min. This time averaging decrease the variability in the observations.

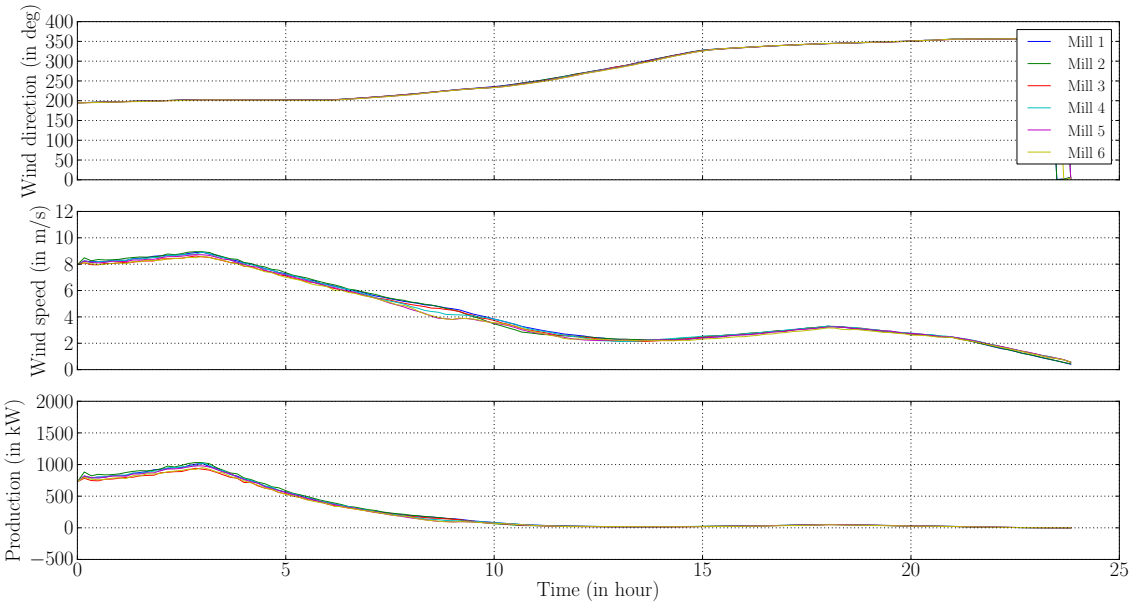


Figure A.12 | Time evolution of SDM results for the 22^{th} December 2016

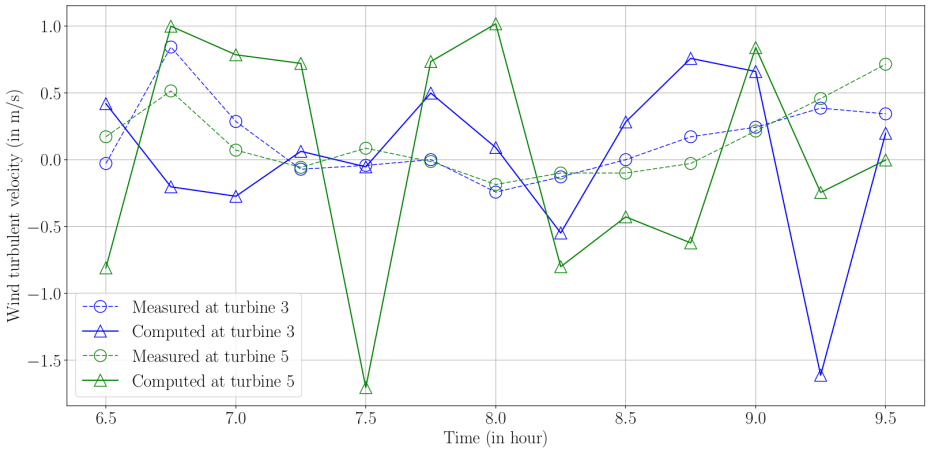


Figure A.13 | Evolution of the wind turbulent velocity between 6:30 and 9:30 LT, when turbines 3, 4, 5 and 6 are lined up. Blue curves display the velocity for turbine 3 (upstream) and green curves display the velocity for turbine 5 (downstream). Dotted line with circles are measured at Parc de Bonneval and solid line with triangles are computed in SDM.

Wake effect. Going back to Figures A.12 and A.11, we observe that the wake effect is well reproduced in the simulation between 7:00 and 12:00 LT. The magnitude is underestimated but the sheltering effect by the forefront turbines is clearly visible. The difference of wind speed between the forefront turbines and those located downstream is about 1-1.5 m s^{-1} in the simulation against 2 m s^{-1} in the measurements. Figure A.14 displays a zoom between 6:00 and 13:00 LT of the measured and simulated wind direction, wind speed and wind energy production. In detail, the measured wind speed and energy production displays a continuously decrease between the forefront turbines and the most downstream turbines. At Parc de Bonneval, we can distinguish two groups of wind turbines. The forefront turbine 3 with turbines 4, 5 and 6 downstream in the wake between 6:30 and 9:00 LT and forefront turbine 1 with turbine 2 downstream in the wake between 10:00 and 12:00 LT. The simulation displays a similar behavior with however significant differences. Between 6:30 and 9:00 LT, wind speed and energy production at turbines 1 and 2 are similar to wind speed and energy production simulated at turbine 3, and turbines 4, 5 and 6 are in the wake of turbine 3 as observed. Between 10:00 and 12:00 LT, the simulated wind speed and energy production varies as observed at the locations of the wind turbines with however a weaker difference between the forefront and the trailing wind turbines.

Figure A.15 shows surface views of the simulated turbulent kinetic energy at the hub height (100 m) at different times (0:20, 8:00 and 11:00 LT). At this altitude the main source of turbulence is due to the interaction with the turbines. Figure A.15a displays the turbulent kinetic energy pattern 20 minutes after the beginning of the simulation at 00:20 LT. At this time the turbines are not lined up and they all produce the same energy. Figure A.15b is similar as Figure A.15a at 8:00 LT. At this time, the wind direction is around 220°. Consequently, the turbines 3, 4, 5 and 6 are lined-up. Figure A.15b displays the sheltering effect by the forefront wind turbine and the turbulence generated in its wake. At 11:00 LT (see Figure A.15c), the wind veers so that turbine 1 creates a wake which reaches turbine 2.

To summarize the performance of the simulation against the measurements, Table A.2 displays skill scores: the Normalized Root Mean Square Error (NRMSE) and the MAE (Mean Absolute Error) defined by

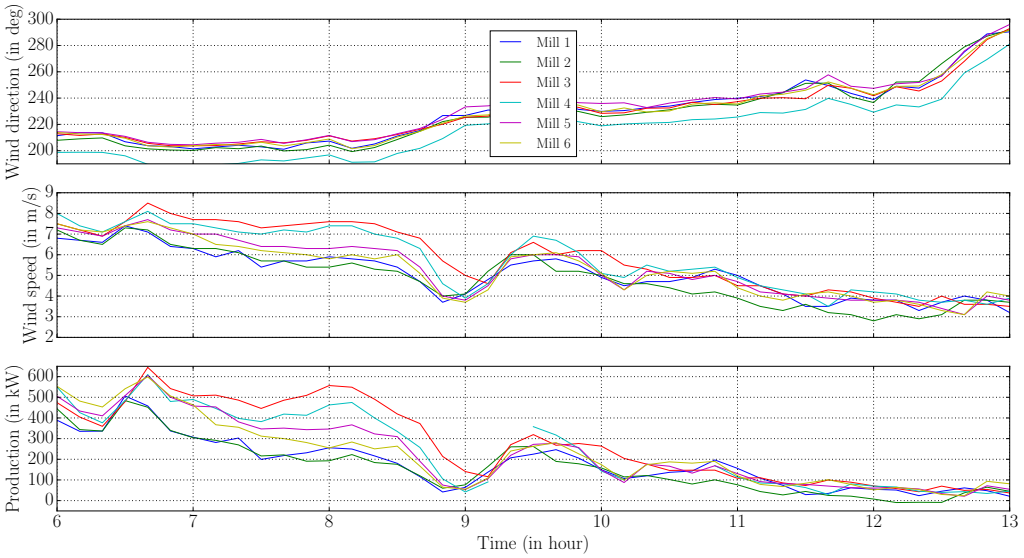
$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2}}{y_{max} - y_{min}}, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{y_i} - y_i|. \tag{A.24}$$

N is the number of measurements. It is equal to 145 (one measurement every 10 minutes from the 22^{th} December 2016 00:00 LT to the 23^{th} December 2016 00:00 LT). We make use of the same number of simulated data saved at the same time. y is the measured wind speed and $\widehat{y}$ is the simulated wind speed.

Table A.2 shows a systematic bias of 1.5 m s^{-1} between the simulation and the measurements, while the NRMSE range varies between about 14.5 to 17%. This is in part due to the initial and lateral boundary conditions from MERRA reanalysis.

Figure A.16 shows vertical profiles of the wind at different times and locations. Both panels display one profile forefront and one profile downstream, at 8:00 (left) and at 11:00 LT (right). The profiles displaying a continuously increasing wind speed (blue curves) correspond to forefront profiles. They are taken at the same location, in front of the turbines and far from their interaction in the middle of the domain. As a consequence, it displays the upstream vertical wind. At 8:00 LT (Figure A.16a), the profile displaying a strong wind speed decreased between 60 and 150 m height (green curve) is extracting downstream turbine 6. This decrease is due to the forefront turbines

(a) Observed wake effect



(b) Simulated wake effect

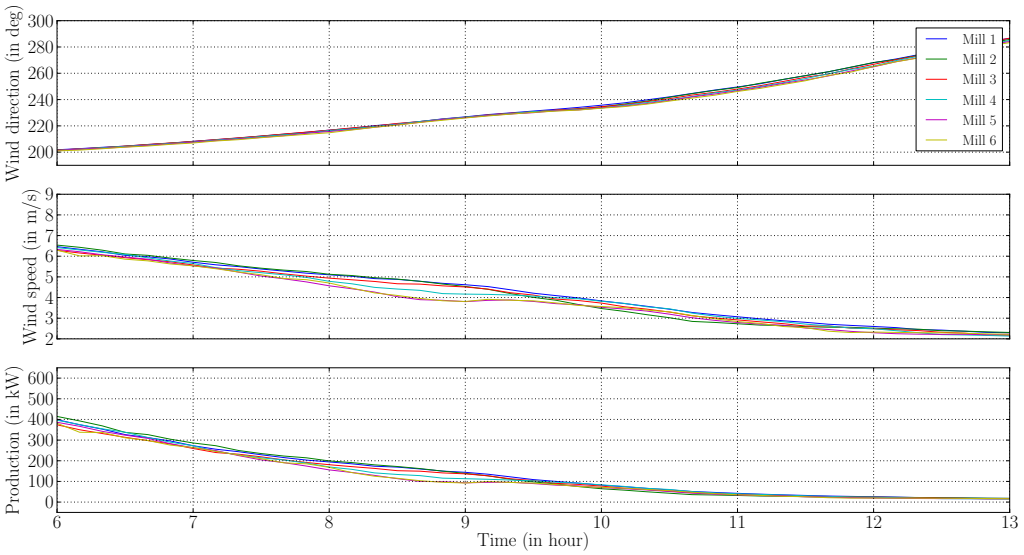
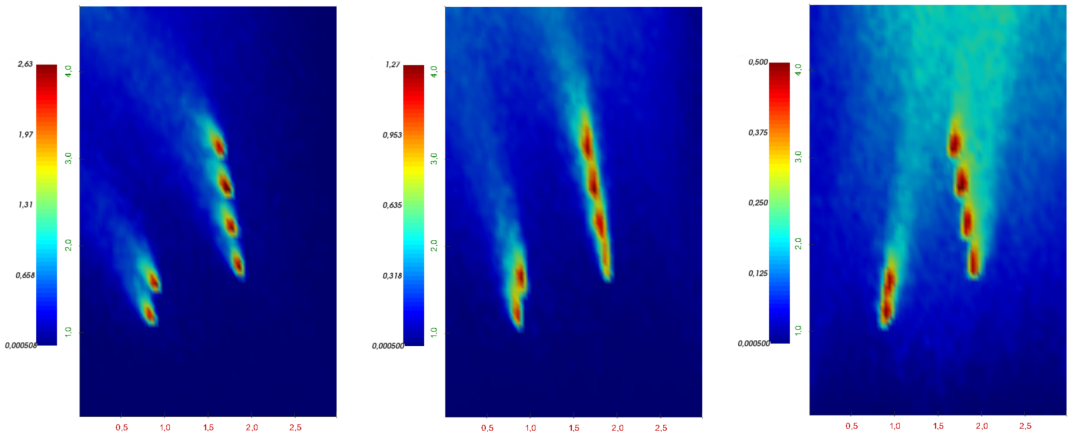


Figure A.14 | Zoom between 6:00 and 12:00 LT

which disrupt the flow and slowdown the wind in front of the downstream turbines. Indeed, at 8:00 LT, turbines 3, 4, 5 and 6 are lined up. At 11:00 LT (Figure A.16b) the green profile is extracting downstream turbine 2. At this time, turbines 1 and 2 are lined up and this is why the wind speed downstream the turbine 2 is slowed by turbine 1. In both case, the interaction with the turbines decreases the wind speed from 2 m s^{-1} maximum at 80 m and 120 m height (just under and above the hub). This figure well describes the wake effect.



(a) at 00:20 LT **(b)** at 8:00 LT **(c)** at 11:00 LT

Figure A.15 | Surface view at hub height (100 m) at different times. The three panels show the turbulent kinetic energy.

	NRMSE (in %)	MAE (in m/s)
Turbine 1	14.57	1.369
Turbine 2	14.56	1.334
Turbine 3	15.88	1.578
Turbine 4	16.83	1.681
Turbine 5	14.92	1.455
Turbine 6	14.71	1.425

Table A.2 | Indicator of the deviation between the simulated wind $\widehat{y}$ and the observed wind y over the six turbines.

A.4 Conclusion

In this paper, we have presented some first numerical experiments obtained from the SDM numerical approach, for a wind farm simulation in condition of use, and we have compared the obtained result with the reality of measures at the turbines.

We have also presented some numerical analysis and experiments that evidence the way the numerical algorithm for SDM is converging.

Some other experiments of wind farm simulation are in preparation, with improvements both in the model and in the description of initial and boundary condition. The objectives are to perform better and reduce the bias against measure, but also to illustrate the ability of SDM to compute not only the mean velocity, but also the local distribution of the turbulent wind, who takes part in the uncertainty of wind power production.

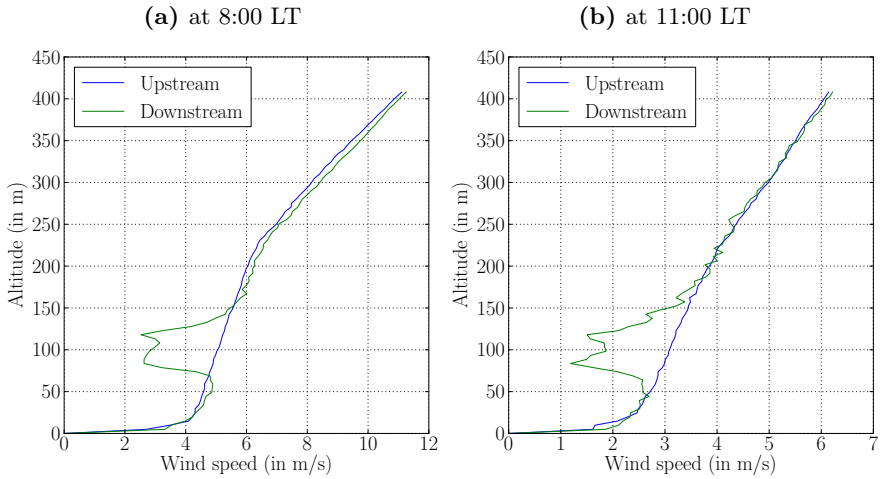


Figure A.16 | Vertical profils taken at different time and place. (a) is taken when the turbines 3, 4, 5 and 6 are lined up; (b) is taken when the turbines 1 and 2 are lined up.

B

List of publications

Publications in peer-reviewed journal as main author

[1] Aurore Dupré, Philippe Drobinski, Bastien Alonzo, Jordi Badosa, Christian Briard and Riwal Plougonven: Sub-hourly Forecasting of Wind Speed and Wind Energy, *Renewable Energy*, 145:2373-2379, 2020, DOI: https://doi.org/10.1016/j.renene.2019.07.161

[2] Aurore Dupré, Philippe Drobinski, Jordi Badosa, Christian Briard and Peter Tankov: The economic value of short term forecasting for wind energy, *Energies - Special Issue related to Wind and Solar Energy*, in preparation

Publications in peer-reviewed journal as co-author

[1] Bastien Alonzo, Riwal Plougonven, Mathilde Mougeot, Aurélie Fishcer Aurore Dupré and Philippe Drobinski: From Numerical Weather Prediction Outputs to Accurate Local Wind Speed: Statistical Modeling and Forecasts, *Proceedings of Forecasting and Risk Management for Renewable Energy, Paris, June 7-9*, 23-44, 2017

[2] Naveen Goutham, Bastien Alonzo, Aurore Dupré, Riwal Plougonven, Rebeca Doctors, Lishan Liao, Mathilde Mougeot, Aurélie Fischer, Philippe Drobinski: Using Machine Learning Methods to Improve Surface Wind from the Outputs of a Numerical Weather Prediction Model, *Boundary Layer Meteorology*, accepted

[3] Mireille Bossy, Aurore Dupré, Philippe Drobinski, Laurent Violeau and Christian Briard: Stochastic Lagrangian Approach for Wind Farm Simulation, *Proceedings of Forecasting and Risk Management for Renewable Energy, Paris, June 7-9*, 23-44, 2017

Oral communications

[1] Aurore Dupré, Badosa Jordi, Philippe Drobinski and Riwal Plougonven: Comparison of Wind Speed Measurement With Sodar, Lidars and Sonic Anemometer at SIRTA, *POSTER presented for the 15ème Journée du SIRTA, École Polytechnique, Palaiseau, 29, June 2017*

[2] Aurore Dupré, Bastien Alonzo, Badosa Jordi, Philippe Drobinski and Riwal Plougonven: The Value of Explanatory Variables Derived From Observations in Downscaling Statistical Methods for Short Term Wind Forecasting, *POSTER presented for the TREND-X / MISTIGRID workshop, École Polytechnique, Palaiseau, 11-12, December 2017*

[3] Aurore Dupré: Sizing of a Short Term Wind Forecasting System, *Oral presentation for the 17ème Journée du SIRTA, École Polytechnique, Palaiseau, 5, July 2019*

[4] Aurore Dupré, Bastien Alonzo, Badosa Jordi, Philippe Drobinski and Riwal Plougonven: Sub-Hourly Forecasting of Wind Speed and Wind Energy, *POSTER presented for the Climate Change and Energy Transition on The Mediterranean workshop, École Polytechnique, Palaiseau, 21-22, December 2019*

[5] Aurore Dupré, Badosa Jordi, Philippe Drobinski and Peter Tankov: The Economic Value of Short Term Forecasting for Wind Energy, *POSTER presented for the Climate Change and Energy Transition on The Mediterranean workshop, École Polytechnique, Palaiseau, 121-22, December 2019*

Bibliography

[1] Internation Energy Agency (IEA) Statistics. *https://www.iea.org/statistics/*.

[2] Internation Energy Agency (IEA). Electricity information: Overview. *https://webstore.iea.org/electricity-information-2019*, 2019, Technical Report.

[3] REN21. Renewables 2018 Global Status Report. *Renewable Energy Policy Network for the 21st century*, 2018, Technical Report.

[4] Réseau de transport d'électricité (RTE). Bilan éléctrique 2018. *https://www.rte-france.com/sites/default/files/be_pdf_2018v3.pdf*, 2018, Technical Report.

[5] Global Wind Energy Council. Global Wind Report. *https://gwec.net/members-area-market-intelligence/reports/*, 2018, Technical Report.

[6] ADEME. Dans l'air du temps, l'énergie éolienne. *Guide ADEME de l'énergie éolienne*, https://ademe.typepad.fr/files/guide_ademe_energie_eolienne.pdf, 2011.

[7] Jing-Shan Hong. Evaluation of the High-Resolution Model Forecasts over the Taiwan Area during GIMEX. *Weather and Forecasting*, 18:836–846, 2003.

[8] J. W. Taylor, P. E. McSharry, and R. Buizza. Wind Power Density Forecasting Using Ensemble Predictions and Time Series Models. *IEEE Transactions on Energy Conversion*, 24(3):775–782, 2009.

[9] Sancho Salcedo-Sanz, Ángel M. Pérez-Bellido, Emilio G. Ortiz-García, Antonio Portilla-Figueras, Luis Prieto, and Daniel Paredes. Hybridizing the Fifth Generation Mesoscale Model with Artificial Neural Networks for Short-Term Wind Speed Prediction. *Renewable Energy*, 34(6):1451–1457, 2009.

[10] G. H. Riahy and M. Abedi. Short Term Wind Speed Forecasting for Wind Turbine Applications Using Linear Prediction Method. *Renewable Energy*, 33(1):35–41, 2008.

[11] Cameron W. Potter and Michael Negnevitsky. Very Short-Term Wind Forecasting for Tasmanian Power Generation. *IEEE Transactions on Power Systems*, 21(2):965–972, 2006.

[12] Wen-Yeau Chang. Application of Back Propagation Neural Network for Wind Power Generation Forecasting. *Interational Journal of Digital Content Technology and its Application*, 7:502–509, 2013.

[13] G. N. Kariniotakis, G. S. Stavrakakis, and E. F. Nogaret. Wind Power Forecasting Using Advanced Neural Networks Models. *IEEE Transactions on Energy Conversion*, 11(4):762–767, 1996.

[14] Pan Zhao, Jiangfeng Wang, Junrong Xia, Yiping Dai, Yingxin Sheng, and Jie Yue. Performance Evaluation and Accuracy Enhancement of a Day-Ahead Wind Power Forecasting System in China. *Renewable Energy*, 43:234–241, 2012.

[15] J. C. Palomares-Salas, J. J. G. De La Rosa, J. G. Ramiro, J. Melgar, A. Aguera, and A. Moreno. ARIMA vs. Neural Networks for Wind Speed Forecasting. *Proceeding of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Hong-Kong - China, May 11-13*, pages 129–133, 2009.

[16] Evaluation of Two Simple Wind Power Forecasting Models. *Proceedings of the European Wind Energy Conference (EWEC), Brussels - Belgium, March 31 - April 3*, 2008.

[17] Umut Firat, Seref Naci Engin, Aysin Baytan Ertuzun, and Murat Saraclar. Wind Speed Forecasting Based on Second Order Blind Identification and Autoregressive Model. *Proceedings of the 9th International Conference on Machine Learning and Applications, Washington - USA, December 12-14*, pages 618–621, 2010.

[18] M. C. Alexiadis, P. S. Dokopoulos, H. S. Sahsamanoglou, and I. M. Manousaridis. Short-Term Forecasting of Wind Speed and Related Electrical Power. *Solar Energy*, 63(1):61–68, 1998.

[19] Tiago Pinto, Sérgio Ramos, Tiago Sousa, and Zita Vale. Short-Term Wind Speed Forecasting Using Support Vector Machines. *Proceedings of IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments (CIDUE), Orlando - USA, December 9-12*, pages 40–46, 2014.

[20] A. Lahouar and J. Ben Hadj Slama. Hour-Ahead Wind Power Forecast Based on Random Forests. *Renewable Energy*, 109:529–541, 2017.

[21] Fernando Castellanos and Nickel James. Average Hourly Wind Speed Forecasting with ANFIS, 2009.

[22] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The Quiet Revolution of Numerical Weather Predicion. *Nature*, 525:47–55, 2015.

[23] Harry R. Glahn and Dale A. Lowry. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, 11:1203–1211, 1972.

[24] Nathalie S. Wagenbrenner, Jason M. Forthofer, Brian K. Lamb, Kyle S. Shannon, and Bret W. Butler. Downscaling Surface Wind Prediction From Numerical Weather Prediction Models in Complex Terrain With WindNinja. *Atmospheric Chemistry and Physics*, 16:5229–5241, 2016.

[25] R. L. Wilby, M. L. Wigley, D. Conway, P. D. Jones, Hewitson B. C., J. Main, and D. S. Wilks. Statistical Downscaling of General Circulation Model Output: A Comparison of Methods. *Water Ressources Research*, 34:2995–3008, 1998.

[26] T. Salameh, P. Drobinski, M. Vrac, and P. Naveau. Statistical Downscaling of Near-Surface Wind Over Complex Terrain in Southern France. *Meteorology and Atmospheric Physics*, 103:253–265, 2009.

[27] Saurabh S. Soman, Hamidreza Zareipour, Om Malik, and Paras Mandal. A Review of Wind Power and Wind Speed Forecasting Methods With Different Time Horizons. *Proceedings of the 2010 North American Power Symposium, Arlington - USA, September 26-28*, pages 1–8, 2010.

[28] Wen-Yeau Chang. A Literature Review of Wind Forecasting Methods. *Journal of Power and Energy Engineering*, 2:161–168, 2014.

[29] A. More and M. C. Deo. Forecasting Wind With Neural Network. *Marine Structures*, 16:35–49, 2003.

[30] Erasmo Cadenas and Wilfrido Rivera. Wind Speed Forecasting in the South Coast of Oaxaca, México. *Renewable Energy*, 32:2116–2128, 2007.

[31] Erasm Cadenas, Oscar Alfredo Jaramillo, and Wilfrido Rivera. Analysis and Forecasting of Wind Velocity in Chetumal, Quintana Roo, Using the Single Exponential Smoothing Method. *Renewable Energy*, 35:925–930, 2010.

[32] R. L. Wilby and C. W. Dawson. The Statistical DownScaling Model: Insight From One Decade of Application. *International Journal of Climatology*, 33:1707–1719, 2013.

[33] Michaël Zamo, Liliane Bel, Olivier Mestre, and Joël Stein. Improved Grided Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression. *Weather and Forecasting*, 31:1929–1945, 2016.

[34] Robert J. Davy, Milton J. Woods, Christopher J. Russell, and Peter A. Coppin. Statistical Downscaling of Wind Variability from Meteorological Fields. *Boundary Layer Meteorology*, 135:161–175, 2010.

[35] Bastien Alonzo, Riwal Plougonven, Mathilde Mougeot, Aurélie Fischer, Aurore Dupré, and Philippe Drobinski. From Numerical Weather Prediction Outputs to Accurate Local Wind Speed: Statistical Modeling and Forecasts. *Proceedings of Forecasting and Risk Management for Renewable Energy, Paris - France, June 7-9*, pages 23–44, 2017.

[36] Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[37] Leo Breiman. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[38] Barbara G. Brown, Richard W. Katz, and Allan H. Murphy. Time Series Models to Simulate and Forecast Wind Speed and Wind Power. *Journal of Applied Meteorology*, 23:1184–1195, 1984.

[39] Coryn A. L. Bailer-Jones, Ranjan Gupta, and Harinder P. Singh. An Introduction to Artificial Neural Network. *Automated Data Analysis in Astronomy*, pages 51–68, 2001.

[40] Coryn A. L. Bailer-Jones, David J. C. MacKay, and Philip J. Withers. A Recurrent Neural Network for Modelling Dynamical Systems. *Network: Computation in Neural Systems*, 9:531–548, 1998.

[41] J. L. Torres, A. Garcia, M. De Blas, and A. De Fransisco. Forecast of Hourly Averaged Wind Speed with ARMA Models in Navarre. *Solar Energy*, 79:65–77, 2005.

[42] A. Sfetsos. A Novel Approach for the Forecasting of Mean Hourly Wind Speed Time Series. *Renewable Energy*, 27:163–174, 2002.

[43] IFS Documentation - Cy43r1. ECMWF. *Part IV: Physical Processes*, 2016.

[44] J. Jung and R. P. Broadwater. Current Status and Future Advances for Wind Speed and Power Forecasting. *Renewable and Sustainable Energy*, 31:762–777, 2014.

[45] Global Wind Energy Council. Global Wind Report. *https://gwec.net/members-area-market-intelligence/reports/*, 2017, Technical Report.

[46] Jing Shi, Xuili Qu, and Songato Zeng. Short-Term Wind Power Generation Forecasting: Direct Versus Indirect Arima-based Approaches. *International Journal of Green Energy*, 8(1):100–112, 2011.

[47] D. Y. Hong, T. Y. Ji, L. L. Zhang, M. S. Li, and Q. H. Wu. An Indirect Short-Term Wind Power Forecast Approach with Multi-Variable Inputs. *Proceedings of IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia), Melbourne - Australia, November 28 - December 1*, pages 793–798, 2016.

[48] Neeraj Bokde, Andrés Feijóo, Daniel Villanueva, and Kishore Kulat. A Novel and Alternative Approach for Direct and Indirect Wind Power Prediction Methods. *Energies*, 11(11), 2018.

[49] M Lydia, S. Suresh Kumar, A. Immanuel Selvakumar, and G. Edwin Prem Kumar. A Comprehensive Review on Wind Turbine Power Curve Modeling Techniques. *Renewable and Sustainable Energy Reviews*, 30:452–460, 2014.

[50] Vinay Thapar, Gayatri Agnihotri, and Vinod Krishna Sethi. Critical Analysis of Methods for Mathematical Modeling of Wind Turbines. *Renewable Energy*, 36:3166–3177, 2011.

[51] Fatih O. Hocaoğlu, Ömer N. Gerek, and Mehmet Kurban. A Novel Hybrid (wind-photovoltaic) System Sizing Procedure. *Solar Energy*, 83:2019–2028, 2009.

[52] Antonino Marvuglia and Antonio Messineo. Monitoring of Wind Farms' Power Curves Using Machine Learning Techniques. *Applied Energy*, 98:574–583, 2012.

[53] Shuhui Li, Donald C. Wunsch, Edgar A. O'Hair, and Michael G. Giesselmann. Using Neural Network to Estimate Wind Turbine Power Generation. *Transactions on Energy Conversion*, 16(3):276–282, 2001.

[54] C. Carrillo, A. F. Obando Montaño, J. Cidrás, and E. Díaz-Dorado. Review of Power Curve Modelling for Wind Turbines. *Renewable and Sustainable Enregy Reviews*, pages 572–581, 2013.

[55] M. Lydia, A. Immanuel Selvakumar, S. Suresh Kumar, and G. Edwin Prem Kumar. Advanced Algorithms for Wind Turbine Power Curve Modeling. *IEEE Transactions on Sustainable Energy*, 4:827–835, 2013.

[56] Shahab Shokrzadeh, Mohammad Jafari Jozani, and Eric Bibeau. Wind Turbine Power Curve Modeling Using Advanced Parametric and Nonparametric Methods. *IEEE Transactions on Sustainable Energy*, 5:1262–1269, 2014.

[57] International Standard IEC 61400-12-1. *Part 12-1: Power performance measurements of electricity producing wind turbines, technical report*, pages 1–556, 2005.

[58] Ergin Erdem and Jing Shi. ARMA Based Approaches for Forecasting the Tuple of Wind Speed and Direction. *Applied Energy*, 88(4):1405–1414, 2011.

[59] A. Khosravi, R. N. N. Koury, L. Machado, and J. J. G. Pabon. Prediction of Wind Speed and Wind Direction Using Artificial Neural Network, Support Vector Regression and Adaptive Neuro-Fuzzy Inference System. *Sustainable Energy Technologies and Assessments*, 25:146–160, 2018.

[60] J. W. Wagenaar and P. J. Eecen. Dependence of Power Performance on Atmospheric Conditions and Possible Corrections. *Technical Report, ECN-M–11-033, March*, pages 1–12, 2011.

[61] Aurélie Fischer, Lucie Montuelle, Mathilde Mougeot, and Dominique Picard. Statistical Learning for Wind Power: a Modeling and Stability Study Towards Forecasting. *Wind Energy*, 20:2037–2047, 2017.

[62] Francis Pelletier, Christian Masson, and Antoine Tahan. Wind Turbine Power Curve Modelling Using Artificial Neural Network. *Renewable Energy*, 89:207–214, 2016.

[63] M. Haeffelin, L. Barthès, and C. Boitel et al. SIRTA, A Ground-Based Atmospheric Observatory for Cloud and Aerosol Research. *Annales Geophysicae*, 23:253–275, 2005.

[64] International Organization for Standardization ISO 2533:1975. *Standard Atmosphere, technical report*, pages 1–108, 1975.

[65] A. W. Dhamouni, M. Ben Salah, F. Askri, Kerkeni C., and S. Ben Nasrallah. Assessment of Wind Energy Potential and Optimal Electricity Generation in Borj-Cedria, Tunisia. *Renewable and Sustainable Energy Reviews*, 15:815–820, 2011.

[66] A. S. Ahmed Shata and R. Hanitsch. Evaluation of Wind Energy Potential and Electricity Generation on the Coast of Mediterranean Sea in Egypt. *Renewable Energy*, 31:1183–1202, 2006.

[67] Radian Belu and Darko Koracin. Effects of Complex Wind Regimes and Meteorlogical Parameters on Wind Turbine Performances. *Proceedings AWEA Windpower, Cleaveland - USA, May 29-31*, 2012.

[68] Warit Werapun, Yutthana Tirawanichakul, and Jompob Waewsak. Wind Shear Coefficients and their Effect on Energy Production. *Energy Procedia*, 138:1061–1066, 2017.

[69] A. Honrubia, A. Vigueras-Rodríguez, E. Gómez Lázaro, and D. Rodríguez-Sánchez. The Influence of Wind Shear in Wind Turbine Power Estimation. *Proceedings EWEC, Warsaw - Poland, April 20-23*, 2012.

[70] Dennis L. Elliot and Jack B. Cadogan. Effects of Wind Shear and Turbulence on Wind Turbine Power Curves. *Proceedings European Community Wind Energy, Madrid - Spain, September 10-14*, 1990.

[71] Ioannis Antoniou, Rozenn Wagner, Søren M. Pedersen, Uwe Paulsen, Helge A. Madsen, Hans E. Jørgensen, Kenneth Thomsen, Peder Enevoldsen, and Leo Thesberg. Influence of Wind Characteristics on Turbine Performance. *Proceeding EWEC, Milano - Italy, May 7-10*, 2007.

[72] M. C. Alexiadis, P. S. Dokopoulos, and H. S. Sahsamanoglou. Wind Speed and Power Forecasting Based on Spatial Correlation Models. *IEEE Transactions on Energy Conversion*, 14(3):836–842, 1999. doi: 10.1109/60.790962.

[73] Tilmann Gneiting, Kristin Larson, Kenneth Westrick, Marc G. Genton, and Eric Aldrich. Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space-Time Method. *Journal of the American Statistical Association*, 101(475):968–979, 2006. doi: 10.1198/016214506000000456.

[74] Amanda S. Hering and Marc G. Genton. Powering Up With Space-Time Wind Forecasting. *Journal of the American Statistical Association*, 105(489):92–104, 2010. doi: 10.1198/jasa.2009.ap08117.

[75] Ralf Kretzschmar, Pierre Eckert, and Danierl Cattani. Neural Network Classifiers for Local Wind Prediction. *Journal of Applied Meteorology*, 43(5):727–738, 2004. doi: 10.1175/2057.1.

[76] Xinxin Zhu, Marc G. Genton, Yingzhong Gu, and Le Xie. Space-time wind speed forecasting for improved power system dispatch. *TEST*, 23(1):1–25, 2014. doi: 10.1007/s11749-014-0351-0.

[77] Wind Power and a Liberalised North European Electricity Exchange. *Proceedings of European Wind Energy Conference (EWEC), Nice - France, March 1-5*, pages 379–382, 1999.

[78] M. S. Roulston, D. T. Kaplan, J. Hardenberg, and L. A. Smith. Using Medium-Range Weather Forecasts to Improve the Value of Wind Energy Production. *Renewable Energy*, 28(4):585–602, 2003.

[79] Pierre Pinson, Christophe Chevallier, and George N. Kariniotakis. Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power. *IEEE Transactions on Power Systems*, 22(3):1148–1156, 2007.

[80] R. J. Barthelmie, F. Murray, and S. C. Pryor. The Economic Benefit of Short Term Forecasting for Wind Energy in the UK Electricity Market. *Energy Policy*, 36(5):1687–1696, 2008.

[81] Alberto Fabbri, Tomás Gómez San Román, Juan Rivier Abbad, and Víctor Méndez Quezada. Assessment of the Cost Associated With Wind Generation Prediction Errors in a Liberalized Electricity Market. *IEEE Transactions on Power Systems*, 20(3):1440–1446, 2005.

[82] Julio Usaoloa, Oswaldo Ravelo, Gerardo González, Fernando Soto, Carmen Dávila, and Belén Díaz–Guerra. Benefits for Wind Energy in Electricity Markets from Using Short Term Wind Power Prediction Tools; a Simulation Study. *Wind Engineering*, 28(1):119–128, 2004.

[83] J. Matevosyan and L. Söder. Minimization of Imbalance Cost Trading Wind Power on the Short Term Power Market. *Proceedings of IEEE Russia PowerTech, St Petersburg - Russia, June 27-30*, pages 1–7, 2005.

[84] Allan H. Murphy. What Is a Good Forecast? A Essay on the Nature of Goodness in weather Forecasting. *Weather and Forecasting*, 8:281–293, 1993.

[85] Réseau de transport d'électricité (RTE). Panorama de l'Électricité Renouvelable en 2018. *https://www.rte-france.com/sites/default/files/panoramat4-2018-hd.pdf*, 2018, Technical Report.

[86] Claire L. Vincent and Pierre-Julien Trombe. 8 - Forecasting Intrahourly Variability of Wind Generation. *Renewable Energy Forecasting - Woodhead Publishing Series in Energy*, pages 219–233, 2017.

[87] J.N. Sørensen. Aerodynamic Aspects of Wind Energy Conversion. *Annual Review of Fluid Mechanics*, 43(1):427–448, 2011.

[88] A. Niayifar and F. Porté-Agel. Analytical Modeling of Wind Farms: A New Approach for Power Prediction. *Energies*, 9(9, 741), 2016.

[89] F. Bernardin, M. Bossy, C. Chauvin, P. Drobinski, A. Rousseau, and T. Salameh. Stochastic Downscaling Methods : Application to Wind Refinement. *Stoch. Environ. Res Risk. Assess*, 23(6):851–859, 2009.

[90] F. Bernardin, M. Bossy, C. Chauvin, J.-F. Jabir, and A. Rousseau. Stochastic Lagrangian Method for Downscaling Problems in Computational Fluid Dynamics. *ESAIM: M2AN*, 44(5):885–920, 2010.

[91] M. Bossy, J. Espina, J. Morice, C. Paris, and A. Rousseau. Modeling the wind circulation around mills with a Lagrangian stochastic approach. *SMAI-Journal of Computational Mathematics*, 2:177–214, 2016.

[92] M. Bossy and Violeau L. Optimal rate of convergence of particle approximation for conditional McKean-Vlasov kinetic processes. *arxiv preprint*, (–), 2018.

[93] A. Stohl. Computation, accuracy and applications of trajectories—A review and bibliography. *Atmospheric Environment*, 32(6):947 – 966, 1998.

[94] J.-P. Minier. Statistical descriptions of polydisperse turbulent two-phase flows. *Physics Reports*, 665(Supplement C):1 – 122, 2016. Statistical descriptions of polydisperse turbulent two-phase flows.

[95] S. B. Pope. Lagrangian PDF methods for turbulent flows. In *Annual review of fluid mechanics, Vol. 26*, pages 23–63. Annual Reviews, Palo Alto, CA, 1994.

[96] J.-P. Minier, S. Chibbaro, and S. B. Pope. Guidelines for the formulation of Lagrangian stochastic models for particle simulations of single-phase and dispersed two-phase turbulent flows. *Physics of Fluids*, 26(11):113303, 2014.

[97] S. B. Pope. *Turbulent flows*. Cambridge University Press, Cambridge, 2000.

[98] M. Bossy. Some stochastic particle methods for nonlinear parabolic PDEs. In *GRIP—Research Group on Particle Interactions*, volume 15 of *ESAIM Proc.*, pages 18–57. EDP Sci., Les Ulis, 2005.

[99] S. B. Pope. Particle Method for Turbulent Flows: tegration of Stochastic Model Equations . *Journal of Computational Physics*, 117(2):332 – 349, 1995.

[100] M. Waclawczyk, J. Pozorski, and J.-P. Minier. Probability density function computation of turbulent flows with a new near-wall model. *Physics of Fluids*, 16(5):1410–1422, 2004.

[101] P. A. Durbin. A Reynolds stress model for near-wall turbulence. *Journal of Fluid Mechanics*, 249:465–498, 1993.

[102] P.-A. Durbin and C.-G. Speziale. Realizability of second-moment closure via stochastic analysis. *J. Fluid Mech.*, 280:395–407, 1994.

[103] S. B. Pope. Lagrangian pdf methods for turbulent flows. *Annu. Rev. Fluid Mech.*, 26:23–63, 1994.

[104] P.E. Réthoré, N.N. Sørensen, A. Bechmann, and F. Zahle. Study of the atmospheric wake turbulence of a CFD actuator disc model. In *Proceedings of European Wind Energy Conference*, Marseille, France, 2009. 16-19 March.

[105] W. McCarty, L Coy, R. Gelano, A. Huang, Merkova D., Smith E. B., M. Sienkiewicz, and K. Wargan. MERRA-2 Input Observations: Summary and Assessment. *NASA Technical Report Series on Global Modeling and Data Assimilation*, 46, 2016.

ECOLE DOCTORALE

Titre : Dimensionnement d'un système de prévision éolienne à court terme

Mots clés : Prévision court terme, Énergie éolienne, Réduction d'échelle, Valeur économique

Résumé : Dans un contexte de réchauffement climatique et de transition énergétique, le développement des énergies renouvelables est indispensable afin de garantir une production d'énergie qui réponde à une demande en croissance constante. Les producteurs éoliens Français bénéficient d'une période de rachat obligatoire de leur production de la part d'EDF durant 15 ans. Après cela, ils doivent vendre leur production sur le marché concurrentiel. Pour ce faire ils doivent annoncer à l'avance la quantité d'énergie qu'ils injecteront sur le réseau. En cas de déséquilibre, des pénalités leurs sont imputées. En France, l'échéance limite pour vendre son énergie est de 30 minutes. Ainsi, dans cette thèse, plusieurs approches de réduction d'échelle, paramétriques (régression linéaire) et non paramétriques (forêts aléatoires) sont développées, calibrées et évaluées. Les échéances considérées vont de 30 min à 3 h. Les méthodes de réduction d'échelle considérées sont très rarement utilisées pour des échéances inférieures à l'heure puisque les modèles numériques sont généralement exécutés toutes les 6 à 12 h. L'utilisation de mesures in-situ dans les méthodes de réduction d'échelle, afin de corriger la prévision numérique à l'initialisation, permet un gain de performance significatif. Comparé avec les méthodes statistiques classiques pour la prévision court terme, l'amélioration par rapport à la méthode de persistance va de 1.5% à 10 min à plus de 30% à 3 h. Afin de limiter l'accumulation d'erreurs lors du passage de la prévision du vent à la prévision de la puissance éolienne, une analyse de l'erreur induite par différentes variables météorologiques, comme la direction du vent ou la densité de l'air, également est présentée. Dans un premier temps, la prévision ferme par ferme est explorée puis la dimension spatiale est introduite. Pour finir, la valeur économique d'un tel système de prévision court terme est explorée. Les différentes étapes du marché de l'électricité sont étudiées et les différentes sources d'incertitude et de variabilité, comme les erreurs de prévision et la volatilité des prix, sont mises en évidence et évaluées. Pour les deux fermes considérées dans cette étude, les résultats montrent que les prévisions court terme permettent une augmentation du revenu annuel entre 4 et 5%.

Title : Sizing of a short term wind forecasting system

Keywords : Short term forecasting, Wind energy, Downscaling, Economic value

Abstract : In a context of global warming and energy transition, the development of renewable energies is essential in order to ensure energy production that meets a constantly growing demand. French wind power producers benefit from a "obligation to purchase" from EDF for 15 years. After that, they have to sell their production in the competitive market. To do so, they must announce in advance the amount of energy they will inject into the grid. In case of imbalance, they are charged penalties. In France, the deadline for selling energy is 30 minutes. Thus, in this thesis, several downscaling approaches, parametric (linear regression) and non-parametric (random forests) are developed, calibrated and evaluated. The considered lead times range from 30 min to 3 h. The downscaling methods considered are rarely used for lead times lower than 1 h since numerical models are generally run every 6 to 12 hours. The use of in-situ measurements in downscaling methods to correct the numerical prediction at initialization, allows a significant performance gain. Compared to traditional statistical methods for short term forecasting, the improvement compared to the persistence method ranges from 1.5%, 10 min ahead, to more than 30%, 3 h ahead. In order to limit the accumulation of errors in the conversion from wind speed forecast to wind power forecast, an analysis of the error induced by different meteorological variables, such as wind direction or air density, is presented. First, the forecast at the farm scale is explored and then the spatial dimension is introduced. Finally, the economic value of such a short term forecasting model is explored. The different steps of the electricity market are studied and the different sources of uncertainty and variability, such as forecast errors and price volatility, are identified and assessed. For the two wind farms considered in this study, the results show that the short term forecasts allow an increase in annual income between 4 and 5%.

Institut Polytechnique de Paris
91120 Palaiseau, France