data and case studies from wind energy research and industrial practices are used in this book. Readers who may benefit from reading this book include practitioners in the wind industry who look for data science solutions and faculty members and students who may be interested in the research of data science for wind energy in departments such as industrial and systems engineering, statistics, and power engineering.

There are a few books on renewable energy forecasting [117], which overlap, to a certain degree, with the content of Part I. A topic related to wind energy but left out in the book is about grid integration, for which interested readers can refer to the book by Morales et al. [148].

1.2.2 Note for Instructors

This book can be used as the textbook for a stand-alone course, with the course title the same as or similar to the title of this book. It can also be used to as a reference book that provides supplementary materials for certain segments of either a data science course (supplementing wind energy application examples) or a power engineering course (supplementing data science methods). These courses can come from the offerings of a broad set of departments, including Industrial Engineering, Electrical Engineering, Statistics, Aerospace Engineering, or Computer Science.

We recommend that the first chapter be read before later chapters are covered. The three parts after the first chapter are more or less independent of each other. It does not matter in which sequence the three parts are read or taught. Within each part, however, we recommend following the order of the chapters. It will take two semesters to teach the whole book. One can, nevertheless, sample one or two chapters from each part to form the basis for a one-semester course.

Most of the examples are solved using the `R` programming language, while some are solved using the `MATLAB`® programming language. At the end of a chapter, acronyms and abbreviations used in that chapter are summarized and explained in the Glossary section.

1.2.3 Datasets Used in the Book

In this book, the following datasets are used:

1. `Wind Time Series Dataset`. This dataset comes from a single turbine on an inland wind farm. The dataset covers the duration of one year, but data at some of the time instances are missing. Two time resolutions are included in the dataset: the 10-min data and the hourly data; the latter is the further average of the former. For each temporal resolution, the data is arranged in three columns. The first column is the time stamp, the second column is the wind speed, and the third column is the wind power.

2. `Wind Spatial Dataset`. This dataset comes from ten turbines in an offshore wind farm. Only the hourly wind speed data are included. The duration of the data covers two months. The longitudinal and latitudinal coordinates of each turbine are given, but those coordinates are shifted by an arbitrary constant, so that the actual locations of these turbines are protected. The relative positions of the turbines, however, remain truthful to the physical layout. The data is arranged in the following fashion. Under the header row, the next two rows are the coordinates of each turbine. The third row under the header is purposely left blank. From the fourth row onwards are the wind speed data. The first column is the time stamp. Columns 2-11 are the wind speed values measured in meters per second.

3. `Wind Spatio-Temporal Dataset1`. This dataset comprises the average and standard deviation of wind speed, collected from 120 turbines in an inland wind farm, for the years of 2009 and 2010. Missing data in the original dataset are imputed by using the iterative singular value decomposition [139]. Two data files are associated with each year—one contains the hourly average wind speed, used in Eq. 3.18, and the other contains the hourly standard deviation of wind speed, used in Eq. 3.25. The naming convention makes it clear which year a file is associated with and whether it is for the average speed (`Ave`) or for the standard deviation (`Stdev`). The data arrangement in these four files is as follows—the columns are the 120 turbines and the rows are times, starting from 12 a.m. on January 1 of a respective year as the first data row, followed by the subsequent hours in that year. The fifth file in this dataset contains the coordinates of the 120 turbines. To protect the wind farm's identity, the coordinates have been transformed by an undisclosed mapping, so that their absolute values are no longer meaningful but the turbine-to-turbine relative distances are maintained.

4. `Wind Spatio-Temporal Dataset2`. The data used in this study consists of one year of spatio-temporal measurements at 200 randomly selected turbines on a flat terrain inland wind farm, between 2010 and 2011. The data consists of turbine-specific hourly wind speeds measured by the anenometers mounted on each turbine. In addition, one year of hourly wind speed and direction measurements are available at three met masts on the same wind farm. Columns `B` through `OK` are the wind speed and wind power associated with each turbine, followed by Columns `OL` through `OQ`, which are for wind speed and wind direction associated with each mast. The coordinates of the turbines and masts are listed in the top rows, preceding the wind speed, direction, and power data. The coordinates are shifted by a constant, so that while the relative positions of the turbines and the met masts remain faithful to the actual layout, their true geographic information is kept confidential. This anemometer

network provides a coverage of a spatial resolution of one mile and a temporal resolution of one hour.

5. `Inland Wind Farm Dataset1` and `Offshore Wind Farm Dataset1`. Data included in these two datasets are generated from six wind turbines and three met masts and are arranged in six files, each of which is associated with a turbine. The six turbines are named WT1 through WT6, respectively. The layout of the turbines and the met masts is shown in Fig. 5.6. On the offshore wind farm, all seven environmental variables as mentioned above are available, namely $\boldsymbol{x} = (V, D, \rho, H, I, S_a, S_b)$, whereas on the inland wind farm, the humidity measurements are not available, nor is the above-hub wind shear, meaning that $\boldsymbol{x} = (V, D, \rho, I, S_b)$. Variables in $\boldsymbol{x}$ were measured by sensors on the met mast, whereas y was measured at the wind turbines. Each met mast has two wind turbines associated with it, meaning that the $\boldsymbol{x}$'s measured at a met mast are paired with the y's of two associated turbines. For WT1 and WT2, the data were collected from July 30, 2010 through July 31, 2011 and for WT3 and WT4, the data were collected from April 29, 2010 through April 30, 2011. For WT5 and WT6, the data were collected from January 1, 2009 through December 31, 2009.

6. `Inland Wind Farm Dataset2` and `Offshore Wind Farm Dataset2`. The wind turbine data in these two datasets include observations during the first four years of the turbines' operations. The inland turbine data are from 2008 to 2011, whereas the offshore data are from 2007 to 2010. The measurements for the inland wind farm include the same $\boldsymbol{x}$'s as in the `Inland Wind Farm Dataset1` and those for the offshore wind farm include the same $\boldsymbol{x}$'s as in the `Offshore Wind Farm Dataset1`. Most of the environmental measurements $\boldsymbol{x}$ are taken from the met mast closest to the turbine, with the exception of wind speed and turbulence intensity which are measured on the wind turbine. The mast measurements are used either because some variables are only measured at the mast (such as air pressure and ambient temperature, which are used to calculate air density) or because the mast measurements are considered more reliable (such as wind direction).

7. `Turbine Upgrade Dataset`. This dataset includes two sets, corresponding, respectively, to an actual vortex generator installation and an artificial pitch angle adjustment. Two pairs of wind turbines from the same inland wind farm, as used in Chapter 5, are chosen to provide the data, each pair consisting of two wind turbines, together with a nearby met mast. The turbine that undergoes an upgrade in a pair is referred to as the *experimental turbine*, the *reference turbine*, or the *test turbine*, whereas the one that does not have the upgrade is referred to as the *control turbine*. In both pairs, the test turbine and the control turbine

are practically identical and were put into service at the same time. This wind farm is on a reasonably flat terrain.

The power output, y, is measured on individual turbines, whereas the environmental variables in $\boldsymbol{x}$ (i.e., the weather covariates) are measured by sensors at the nearby mast. For this dataset, there are five variables in $\boldsymbol{x}$ and they are the same as those in the `Inland Wind Farm Dataset1`. For the vortex generator installation pair, there are 14 months' worth of data in the period before the upgrade and around eight weeks of data after the upgrade. For the pitch angle adjustment pair, there are about eight months of data before the upgrade and eight and a half weeks after the upgrade.

Note that the pitch angle adjustment is not physically carried out, but rather simulated on the respective test turbine. The following data modification is done to the test turbine data. The actual test turbine data, including both power production data and environmental measurements, are taken from the actual turbine pair operation. Then, the power production from the designated test turbine on the range of wind speed over 9 m/s is increased by 5%, namely multiplied by a factor of 1.05, while all other variables are kept the same. No data modification of any kind is done to the data affiliated with the control turbine in the pitch angle adjustment pair.

The third column of a respective dataset is the `upgrade status` variable, of which a zero means the test turbine is not modified yet, while a one means that the test turbine is modified. The `upgrade status` has no impact on the control turbine, as the control turbine remains unmodified throughout. The vortex generator installation takes effect on June 20, 2011, and the pitch angle adjustment takes effect on April 25, 2011.

8. `Wake Effect Dataset`. This dataset includes data from six pairs of wind turbines (or, 12 wind turbines in total) and three met masts. The turbine pairs are chosen such that no other turbines except the pair are located within 10 times the turbine's rotor diameter. Such arrangement is to find a pair of turbines that are free of other turbines' wake, so that the wake analysis result can be reasonably attributed to the wake of its pair turbine. The operational data for the six pairs of turbines are taken during roughly a yearlong period between 2010 and 2011. The datasets include wind power output, wind speed, wind direction, air pressure, and temperature, of which air pressure and temperature data are used to calculate air density. The wind power outputs and wind speeds are measured on the turbine, and all other variables are measured at the met masts. The data from Mast 1 are associated with the data for Turbine Pairs 1 and 2, Mast 2 with Pairs 3 and 4, and Mast 3 with Pairs 5 and 6. Fig. 8.6 shows the relative locations of the six pairs of turbines and three met masts.

9. `Turbine Bending Moment Dataset`. This dataset includes two parts. The first part is three sets of physically measured blade-root flapwise bending moments on three respective turbines, courtesy of Risø-DTU (Technical University of Denmark) [180]. The basic characteristics of the three turbines can be found in Table 10.1. These datasets include three columns. The first column is the 10-min average wind speed, the second column is the standard deviation of wind speed within a 10-min block, and the third column is the maximum bending moment, in the unit of MN-m, recorded in a 10-min block. The second part of the dataset is the simulated load data used in Section 10.6.5. This part has two sets. The first set is the training data that has 1,000 observations and is used to fit an extreme load model. The second set is the test data that consists of 100 subsets, each of which has 100,000 observations. In other words, the second dataset for testing has a total of 10,000,000 observations, which are used to verify the extreme load extrapolation made by a respective model. Both simulated datasets have two columns: the first is the 10-min average wind speed and the second is the maximum bending moment in the corresponding 10-min block. While all other datasets are saved in CSV file format, this simulated test dataset is saved in a text file format, due to its large size. The data simulation procedure is explained in Section 10.6.5.

10. `Simulated Bending Moment Dataset`. This dataset includes two sets. One set has 600 data records, corresponding to the training set referred to in Section 11.4.1, whereas the other set has 10,000 data records, which are used to produce Fig. 11.1. Each set has three columns of data (other than the serial number). The first column is the wind speed, simulated using a Rayleigh distribution, and the second and third columns are, respectively, the simulated flapwise and edgewise bending moments, in the unit of kN-m. The flapwise and edgewise bending moments are simulated from TurbSim [112] and FAST [113], following the procedure discussed in [149]. TurbSim and FAST are simulators developed at the National Renewable Energy Laboratory (NREL) of the United States.

GLOSSARY

CSV: Comma-separated values Excel file format

DTU: Technical University of Denmark

NREL: National Renewable Energy Laboratory

PTC: Production tax credit

SCADA: Supervisory control and data acquisition

STEM: Science, technology, engineering, and mathematics

US: United States of America