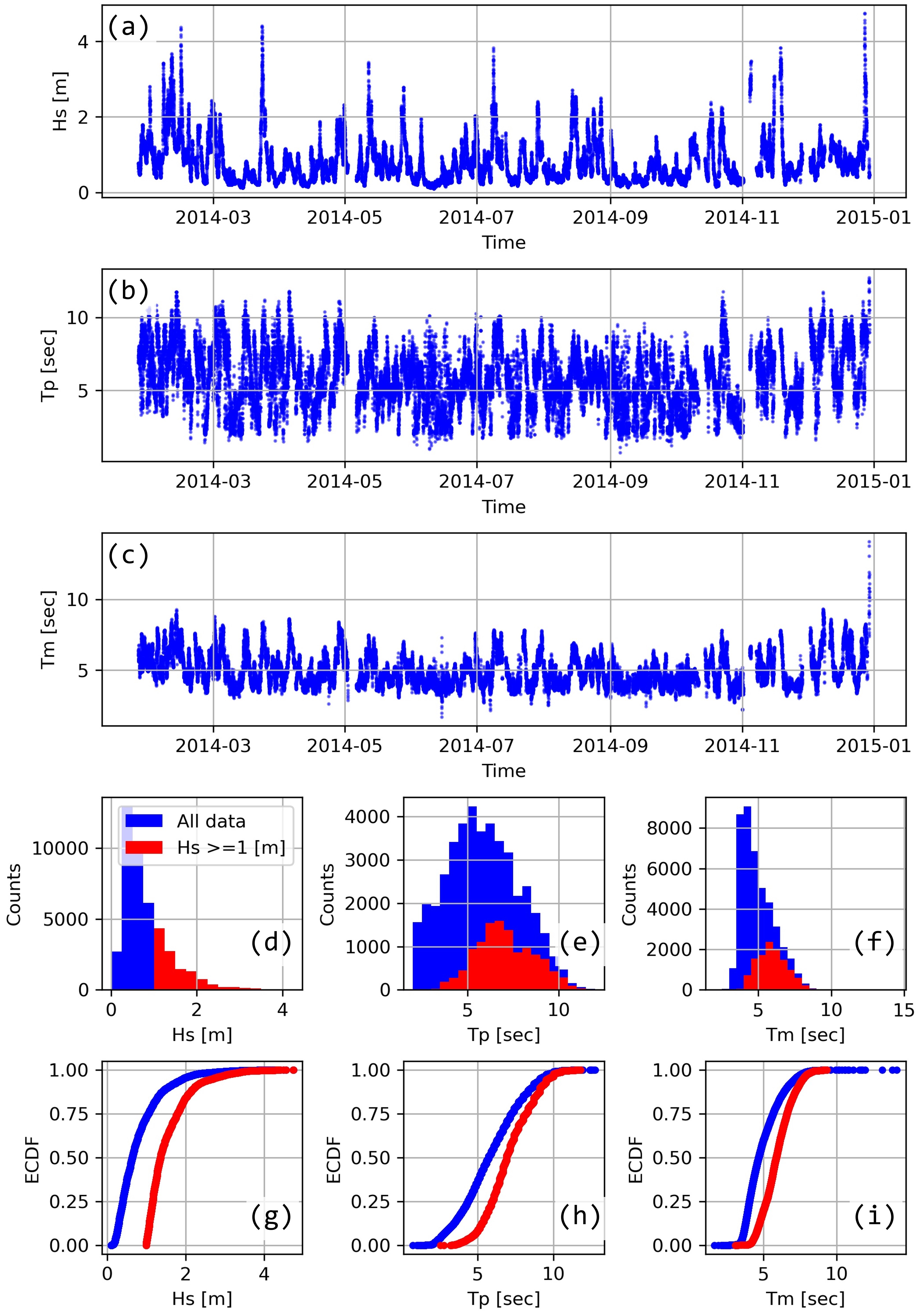
# Computing systems

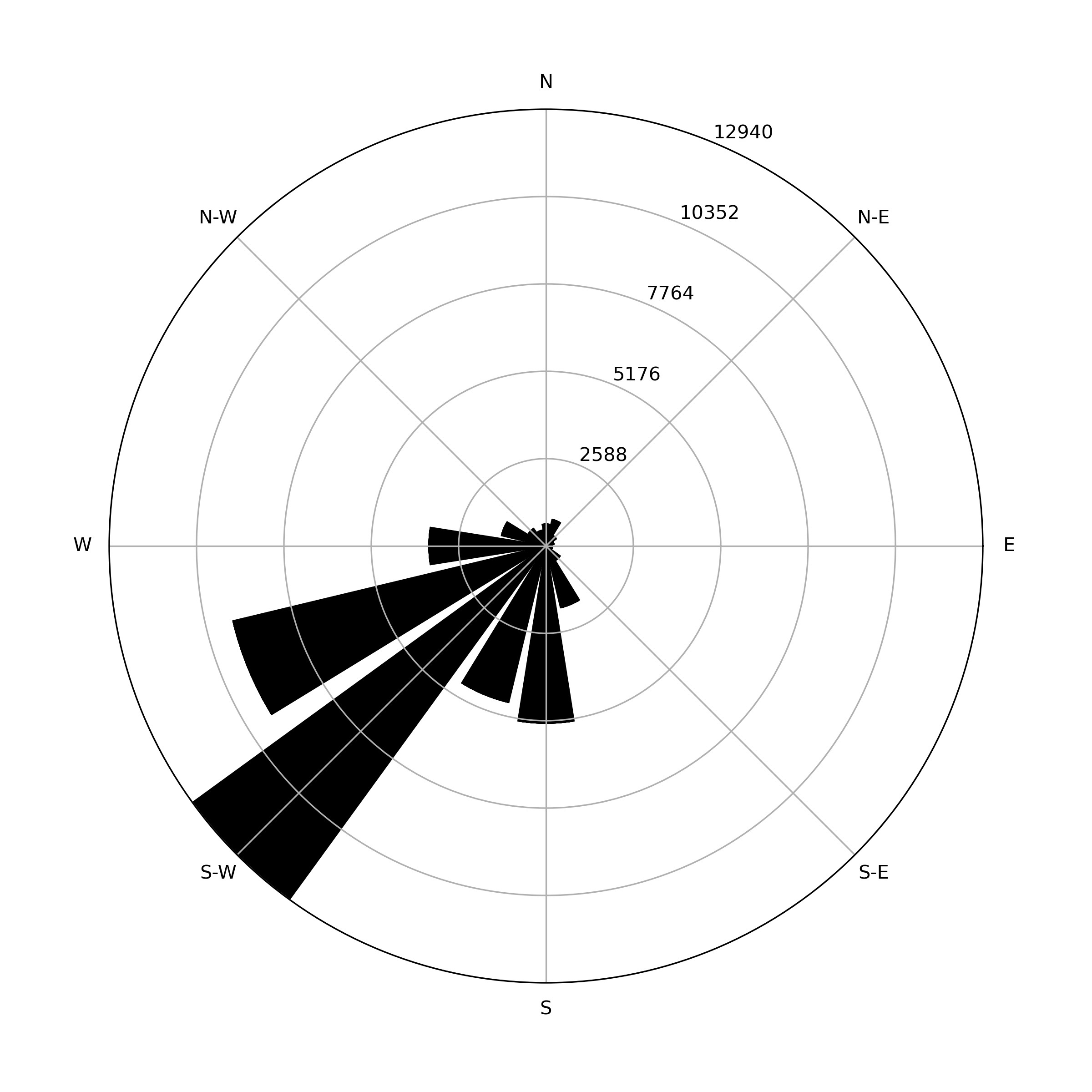
The computational power needed for training the ML algorithms depends on the type of algorithms and on the amount of data. For our study, high performance computing was not used but a Graphical Processing Unit (GPU) was necessary for training deep learning algorithms. Therefore, an Intel i5‑3450 (4 cores, 3.1 GHz) with 16 GB RAM memory featuring a GeForce GTX 1050 Ti graphics card was sufficient to train the ML algorithms. Once the algorithms were trained, they could run in real-time on a less performing, but cheaper, server computer also responsible for data storage. Specifically, we used an Intel NUC8i5BEH with an Intel Core i5‑8259U Processor and 8 GB RAM.

# Exploratory training-data analysis

In this section, we provide some insights into the dataset we used and the results of some relevant analysis. The dataset for training the ML algorithms was made of simultaneous buoy records and micro-seismic PSD records. The size of the buoy dataset and the number of missing records are reported in Suppl. Table 1. The values of Hs, Tp, and Tm recorded by the buoy are shown in Suppl. Figure 1, where a subset, with Hs>1 m, is shown in red in histograms and ECDF plots. It is important to notice that high values of Hs are less frequent, impacting both training and validation results.



Suppl. Figure . Time plots (a, b, c), histograms (d, e, f), and ECDFs (g, h, i) of the sea waves variables measured by the buoy (43114 records). The records with Hs>1 m are displayed in red in the histograms and the ECDFs.



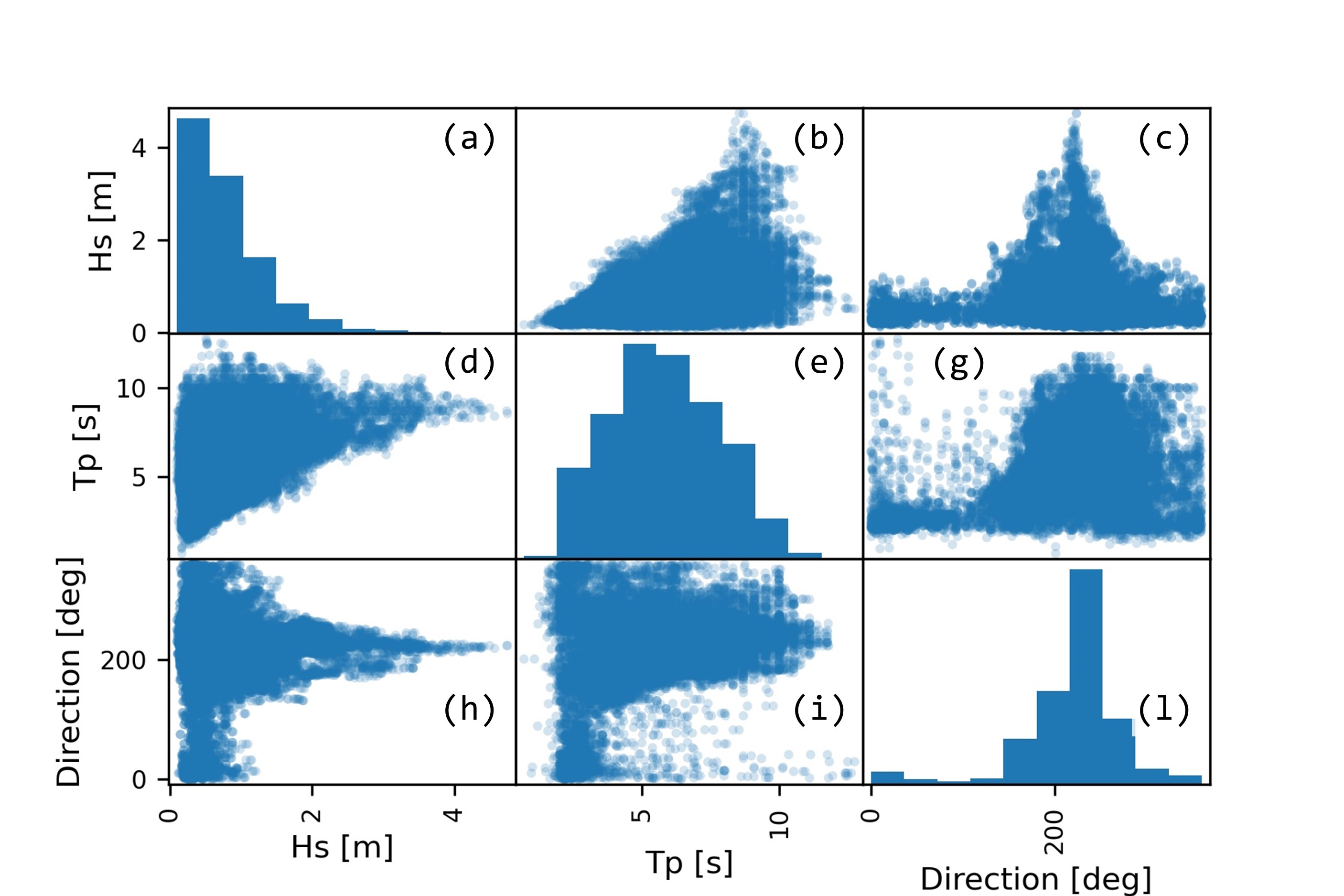
Suppl. Figure 2. Windrose-bar chart of the waves propagation directions. The radius of each circle indicates the counts (e.g., S-W was counted 12940 times).

The waves propagation direction (from which the waves are moving) is shown in Suppl. Figure 2. Most of observations were from South-West, which is certainly a characteristic of the monitored area (see Figure 2). This distribution of the directions indicates that, even if the seismic signal was carrying information about the direction, the ML algorithm would not be able to learn how to detect it. In fact, the dataset mostly contains samples propagating from S-W and very few from the other directions.

Finally, scatter matrix plots can be used to visualize the relationship between variables. For example, Suppl. Figure 3 (d) in supplementary material shows that typically big values of Hs (>3 m) correspond to big values (>6 seconds) of Tp.

|  |  |  |
| --- | --- | --- |
| La Spezia Buoy records (after data augmentation) | | |
|  | Number of records | Percentage |
| Available | 43114 | 82.0% |
| Missing | 9446 | 18% |
| Expected | 52560 | 100% |
|  | | |
| Micro-seismic PSD records | | |
|  | Number of records | Percentage |
| Available | 52049 | 99.0% |
| Missing | 204 | 0.4% |
| Removed during data cleaning | 307\* | 0.6% |
| Expected | 52560 | 100% |
|  | | |
| Simultaneous records | | |
|  | Number of records | Percentage |
| Both Available | 42603 | 81.1% |
| One Missing | 9954 | 18.9% |
| Both Missing | 3 | 0.0% |
| Expected | 52560 | 100% |

Suppl. Table 1. Counting and percentages of available and missing records. The first table is referred to the La Spezia Buoy, the second to the micro-seismic PSDs, and the third to the simultaneous records from the previous two. Simultaneous records are key for training and validating the ML algorithms. \*Micro-seismic data cleaning was performed only on simultaneous records (i.e., when buoy records were available).



Suppl. Figure 3. Scatter matrix plot of the buoy data from 2014. Scatter matrix plots are a powerful tool to inspect data. On the diagonal (a, e, l), the histograms of the variables are shown. Off-diagonal scatter plots (b, c, d, g, h, i) show the relationship between variable pairs. Notice for example in subplot (c), how Hs is always smaller than 1 m when the swell direction is N-E (45°).

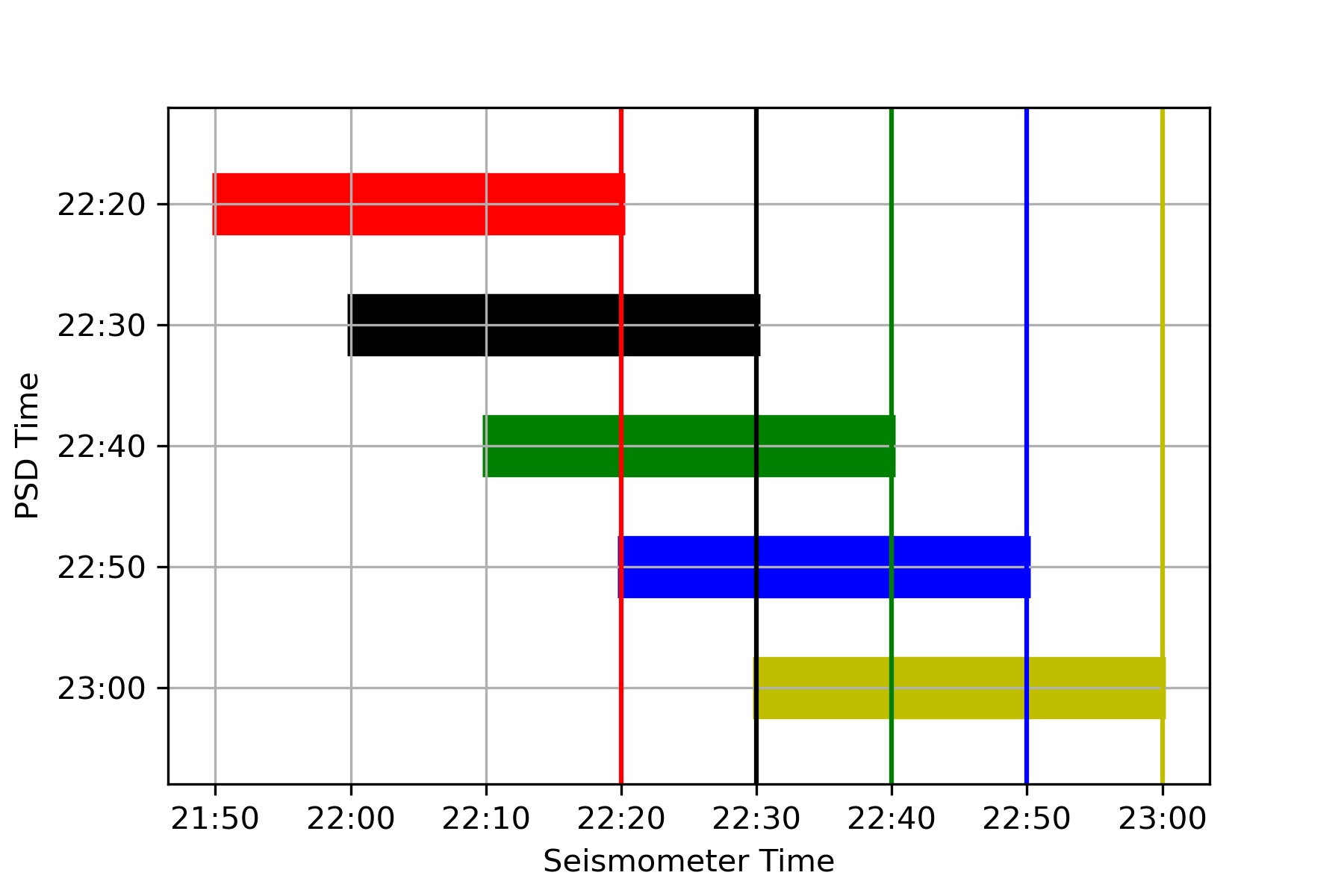
# Micro-seismic power spectral density and spectrogram

Feature engineering is key for an effective application of the ML methods. As described in Section 3.3, we used two methods: one based on computing the Power Spectral Density (PSD) of the micro-seismic signal, the other on computing a spectrogram image.

Specifically, we used the *signal.welch* function of the Python Scipy library to compute the PSD in the bandwidth [~0.015 Hz : 2 Hz]. The *nperseg* (length of each segment) parameter of *signal.welch* was set to 2048, providing smooth results within the bandwidth, whereas the other parameters were set as by default. In this way, one PSD was an array of 407 elements as the example shown in Figure 6. The PSD was computed with a period of 10 minutes using 30 minutes wide time-windows (see Suppl. Figure 4): for example, to compute the PSD corresponding to 22:20 o’clock of a certain day, we used all micro-seismic data from 21:50 to 22:20 of that day (see the red bar of Suppl. Figure 4); to compute that corresponding to 22:30, we used the data from 22:00 to 22:30 (black bar); etc.

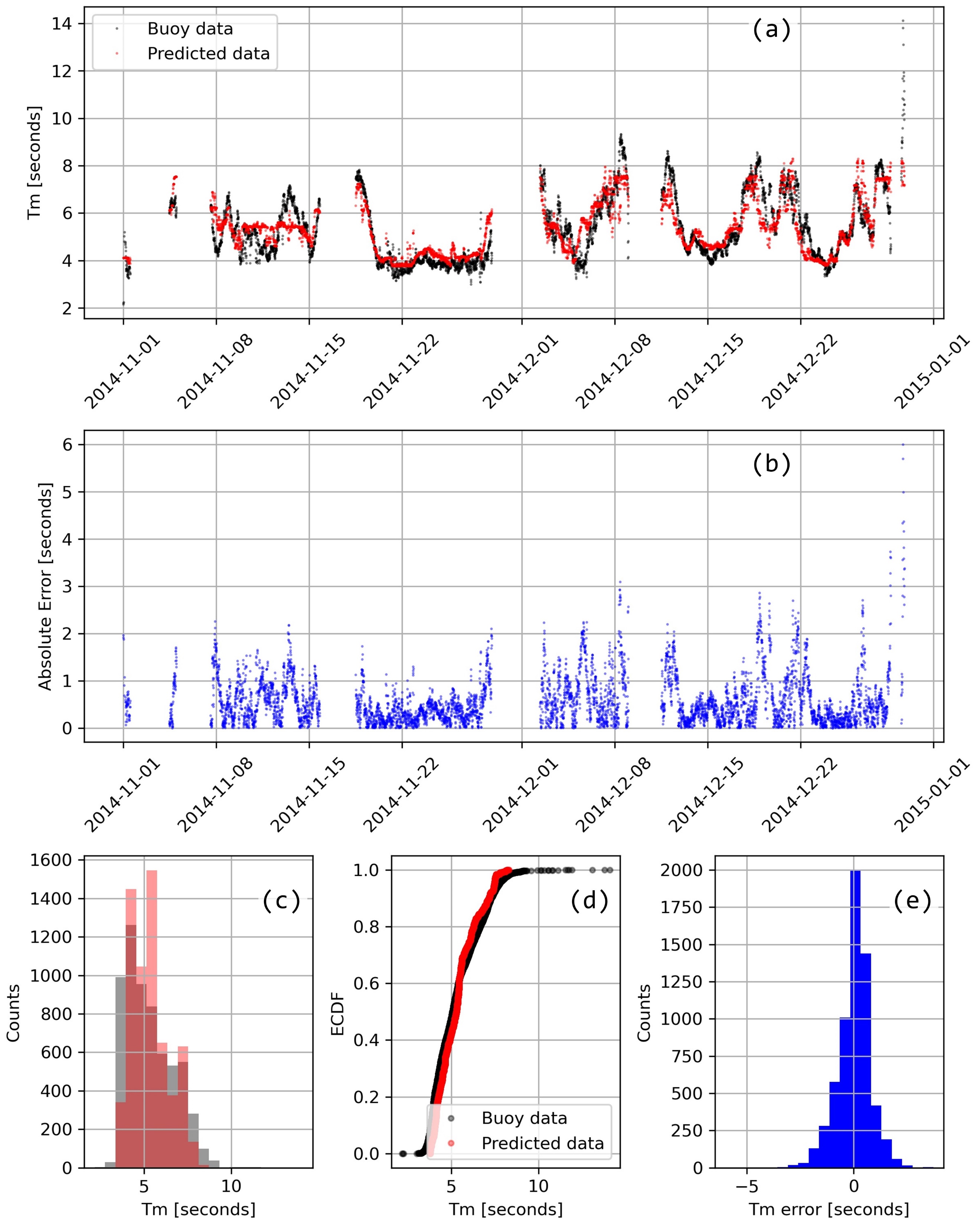
To compute the spectrogram image, we combined the present PSD array with the past PSD arrays to generate one image for each buoy record. In these spectrogram images, each column is the PSD array computed for a specific time. An example is shown in Figure 5, where the PSD arrays corresponding to five timings (i.e., 23:00, 22:50, 22:40, 22:30, and 22:20) are combined to form a spectrogram image with 5 columns and 407 rows. This corresponds to micro-seismic data measured during a time span of one hour and ten minutes (see Suppl. Figure 4).

Both the PSDs and the spectrograms timeseries were rescaled before feeding them into the ML algorithms. This is key to achieve the best performance as, typically, ML algorithms do not work well if the input varies within a large range. Therefore to reduce it, we used the logarithm of the PSD and we passed the inputs through the function StandardScaler of the Scikit-Learn library of Python.

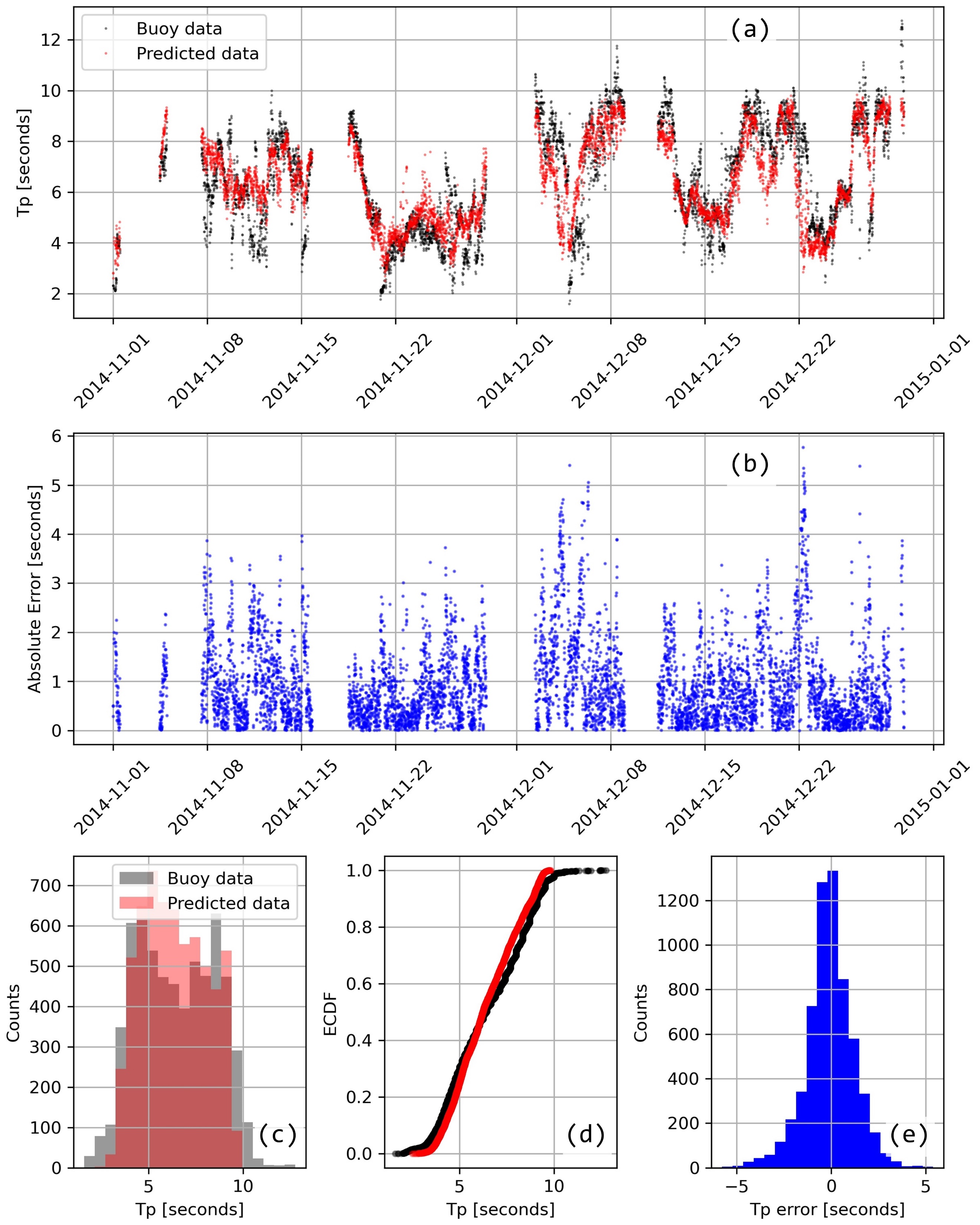


Suppl. Figure . The horizontal bars show the time windows of micro-seismic data used to calculate the corresponding PSD. For example, to compute that at 23:00 o’clock (see Figure 6), the micro-seismic data from 22:30 to 23:00 (yellow bar) were used. Also notice that the horizontal bars of this plot define the time windows used to compute the columns of the spectrogram in Figure 5. Finally, notice that the timings of the last micro-seismic measures of each time window (indicated by the vertical lines) correspond to the timings of the PSD records (indicated on the y axis).

**Supplementary Figures**



Suppl. Figure 5. Mean period (Tm) measured by the buoy (in black) and predicted by the Boosting model (20 pred.) (in red) displayed in the time plot (a), the histogram (c), and the ECDF plot (d). The deviation between the two timeseries is displayed (in blue) in the time plot (b), and the histogram (e).



Suppl. Figure 6. Peak period (Tp) measured by the buoy (in black) and predicted by the Random Forest model (20 pred.) (in red) displayed in the time plot (a), the histogram (c), and the ECDF plot (d). The deviation between the two timeseries is displayed (in blue) in the time plot (b), and the histogram (e).

**Supplementary Tables**

|  |  |  |  |
| --- | --- | --- | --- |
| Config 1 | | | |
| Layer type | Output Shape | | Param # |
| Batch Norm. | (None, 407, 5, 1) | | 4 |
| Conv2D | (None, 102, 5, 32) | | 544 |
| Dropout | (None, 102, 5, 32) | | 0 |
| Conv2D | (None, 26, 5, 16) | | 8208 |
| Dropout | (None, 26, 5, 16) | | 0 |
| Batch Norm. | (None, 26, 5, 16) | | 64 |
| Flatten | (None, 2080) | | 0 |
| Dense | (None, 1) | | 2081 |
|  | | | |
| Config 2 | | | |
| Layer type | Output Shape | Param # | |
| Batch Norm. | (None, 407, 5, 1) | 4 | |
| Conv2D | (None, 204, 5, 32) | 288 | |
| Dropout | (None, 204, 5, 32) | 0 | |
| Conv2D | (None, 51, 5, 32) | 65568 | |
| Dropout | (None, 51, 5, 32) | 0 | |
| Batch Norm. | (None, 51, 5, 32) | 128 | |
| Flatten | (None, 8160) | 0 | |
| Dense | (None, 1) | 8161 | |

Suppl. Table 2. Summary of the sequential CNN models called Config 1 and Config 2. The total numbers of trainable parameters are respectively 10,867 and 74,083.