Supporting Information

Big Data Analytics for Long-Term Meteorological Observations at Hanford Site

Huifen Zhou, Huiying Ren, Patrick Royer, Hongfei Hou and Xiao-Ying Yu *

Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99354, USA
* Correspondence: xiaoying.yu@pnnl.gov

Table of Contents

Support Algorithms

Algorithm S1. MK test.

Mann Kendall test is used to check the monotonic trend of the data [1,2] which is insensitive to outliers and doesn't require data to be normal distribute assumption [3]. The time series data x_1, $x_2 \ldots x_n$ are listed in order. If a data point x_i is greater than the previous point x_k ($i > k$), then set the difference of sign(x_i-x_k) as 1. If a data point (x_i) is equal to the previous point (x_k), then set the difference of sign(x_i-x_k) as 0. When a data point (x_i) is smaller than the previous point (x_k), then set the difference of sign(x_i-x_k) as -1 shown in eqn. (S-1).

$$sign\ (x_i - x_k) = \begin{cases} 1 & if (x_i - x_k) > 0 \\ 0 & if (x_i - x_k) = 0 \\ -1 & if (x_i - x_k) < 0 \end{cases} \quad \text{eqn. (S-1)}$$

then, compute the $S=\sum_{k=1}^{n-1}\sum_{i=k+1}^{n} sign(x_i - x_k)$ and get the variance of S in eqn. (S-2),

$$\text{var (S)}=\frac{n(n-1)(2n+5)-\sum_{j=1}^{m} t_j(t_j-1)(2t_j+5)}{18} \quad \text{eqn. (S-2)}$$

where n is the total data observation, m is the number of the tied groups in the dataset and the t_j is the number of data point in the j^{th} tied group and the n>=8. Then the statistical MK test can be obtained using eqn. (S-3),

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{var(S)}} & S > 0 \\ 0 & 0 \\ \frac{S+1}{\sqrt{var(S)}} & S < 0 \end{cases} \quad \text{eqn. (S-3)}$$

In this work, we use the Theil-Sen's slope to evaluate the monotonic trend. The equation of Sen's slope is expressed in eqn. (S-4)

$$f(t) = Qt + b \quad \text{eqn. (S-4)}$$

The f(t) is the function of time with the continuous increase or decrease. Q is the slope, and b is a constant. The estimated slope Q can be obtained by the following method.

First, obtain the slop $m_{ik} = \frac{x_i - x_k}{i-k}$, here i=2, 3, ……, N, and k=1, 2, ……, (N-1), where x_i is the data point of time i; and x_k is the data point of time k, and (i>k) of all data. Then, rank the N values of m_{ik} from the smallest to largest. Lastly compute the median slope using eqn. (S-5)

$$Q = \begin{cases} m_{(N+1)/2}, & N = odd \\ (m_{N/2} + m_{(N+2)/2})/2, & N = even \end{cases} \quad \text{eqn. (S-5)}$$

A positive Q indicates an increasing trend, and a negative Q indicates a decreasing trend of a time series.

Algorithm S2. Random forest.

The random forest classification model is voted by the classification trees. The general algorithm for RF classification is as the following:

a. From m=1 to M:
   i. Randomly select sample Z of size N from the training dataset
   ii. Grow a random forest tree T_m and with the minimum node size n_{min}
   iii. Output the ensemble trees $\{T_m\}_1^M$.
b. The prediction of random forest classification is expressed in eqn. (S-6)

$$\hat{C}_{rf}^{M} = majority\ vote\left\{\hat{C}_m(x)\right\}_1^M \quad \text{eqn. (S-6)}$$

The important RF classification features are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree [4].

Algorithm S3. Outlier detection.

Multiple outliers remove methods are combined to remove the outliers.

Firstly, use the hard thresholds to remove the outliers. For examples, if the temperature is reported over 30℃ during winter, it will be removed. This is because this measurement would be erroneous. The average monthly atmosphere pressure is reported to be around 30 inches [5], so we set the thresholds' range of the pressure is from 710 mmHg to 800 mmHg.

Secondly, constant values may indicate malfunctions of measurement equipment, such values also were removed. For instance, if the constant values last more than 6 hours, then those constant values were removed.

Thirdly, use the moving average for outlier detection [6] as described below.

i. Require: time series S, moving window k, and the constant value B.
ii. For any point x_i in S,

if i+k is less than the length of S, use eqns. (S-7) and (S-8)

$$\overline{x_\iota} = {\sum_{i+1}^{k+i} x_i}/{k} \quad \text{eqn. (S-7)}$$

$$\sigma_{x_i} = \sqrt{\sum_i^k (x_i - \overline{x_\iota})^2} \quad \text{eqn. (S-8)}$$

if i+k is greater than the length of S, use eqns. (S-9) and (S-10)

$$\overline{x_\iota} = {\sum_{i+1}^{length(S)} x_i}\Big/{(length(S) - i)} \quad \text{eqn. (S-9)}$$

$$\sigma_{x_i} = \sqrt{\sum_i^{length(S)} (x_i - \overline{x_\iota})^2} \quad \text{eqn. (S-10)}$$

iii. Check the value $x_i, if\ x_i$ is great than the $(\overline{x_\iota} - B * \sigma_{x_i})$ and less than $(\overline{x_\iota} + B * \sigma_{x_i})$, then we consider this value normal. Otherwise, consider this value as an outlier.

Supporting Figures

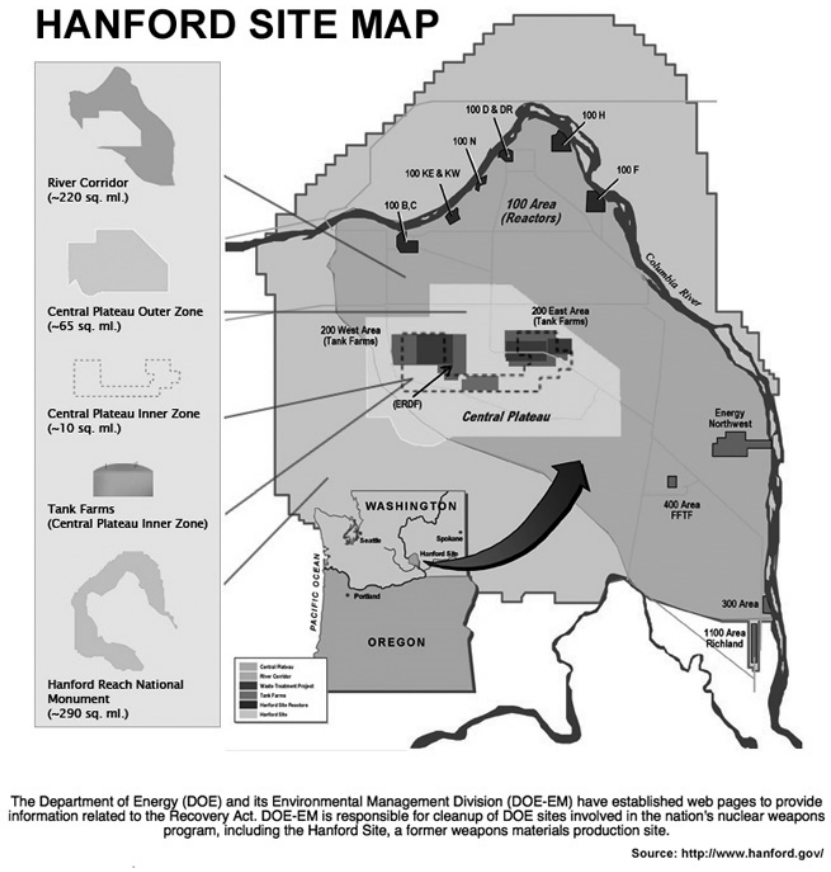

Figure S1. Location of the Hanford Site of the Columbia River and meteorological stations.

Figure S1 represents the map of the Hanford site which is located on 565-square-miles of desert in southeastern Washington State near the Tri-Cities of Richland, Pasco, and Kennewick [7].

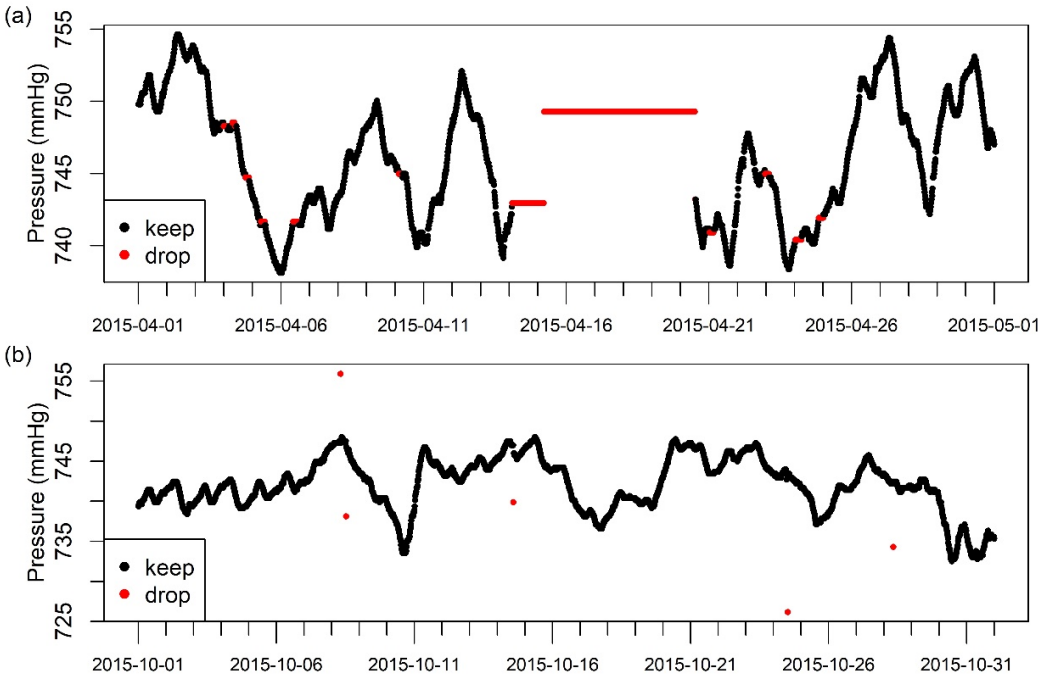

Figure S2. Data outlier analysis of pressure measurement.

Figure S2 gives an example of pressure data outlier analysis. The data points showing no variation highlighted in red are removed (see **Figure S2a**). In addition to visual inspection, the moving average is used to filter additional outliers. Data points that are 3-sigma apart from the average observation points are removed (see **Figure S2b**).

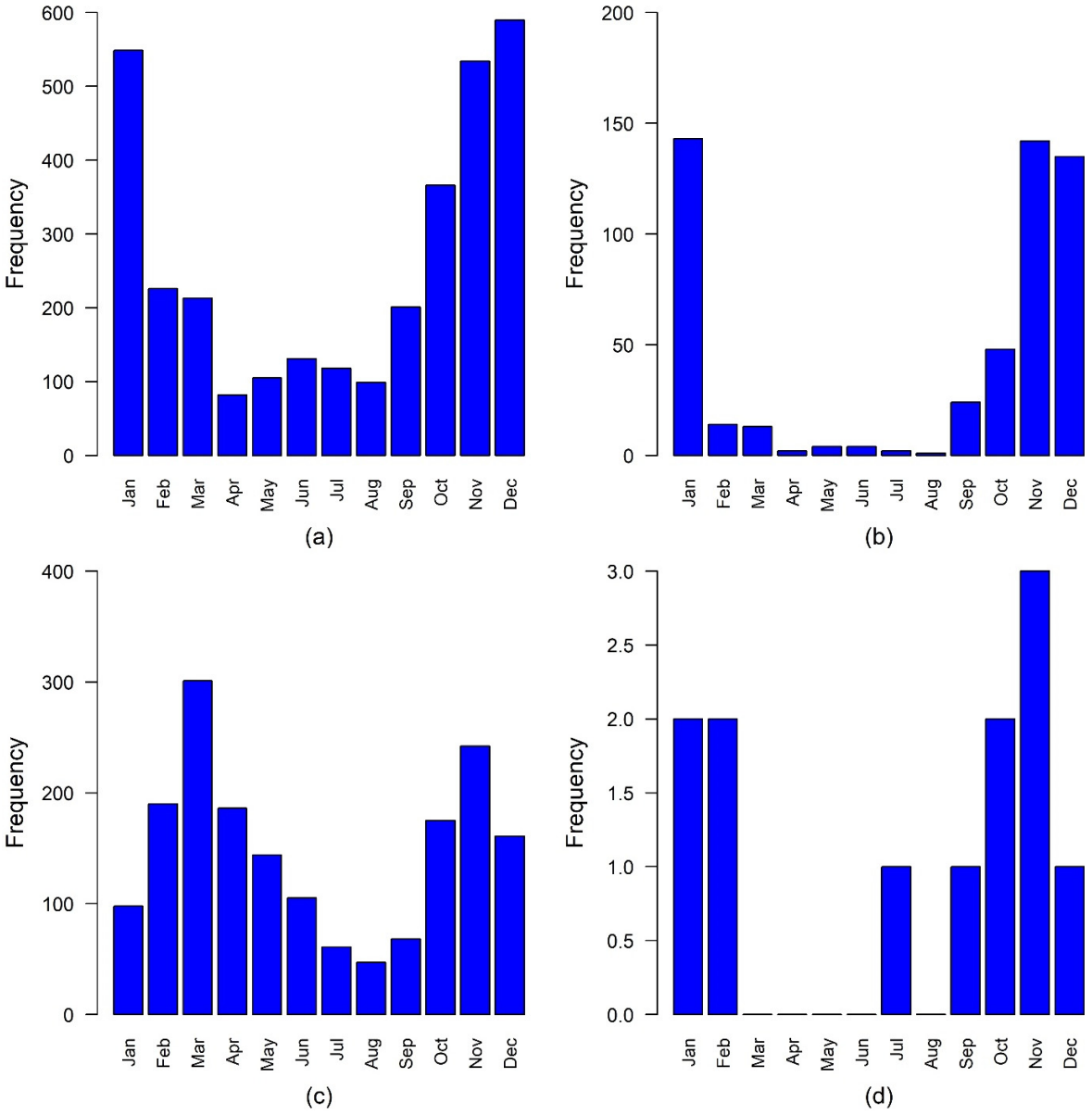

Figure S3. Frequency summary of low wind speed of all sites monthly lasting more than (**a**) 3 h and (**b**) 48 h. Similarly, frequency summary of high wind speed of all sites monthly lasting more than (**c**) 3 hours and (**d**) 48 hours.

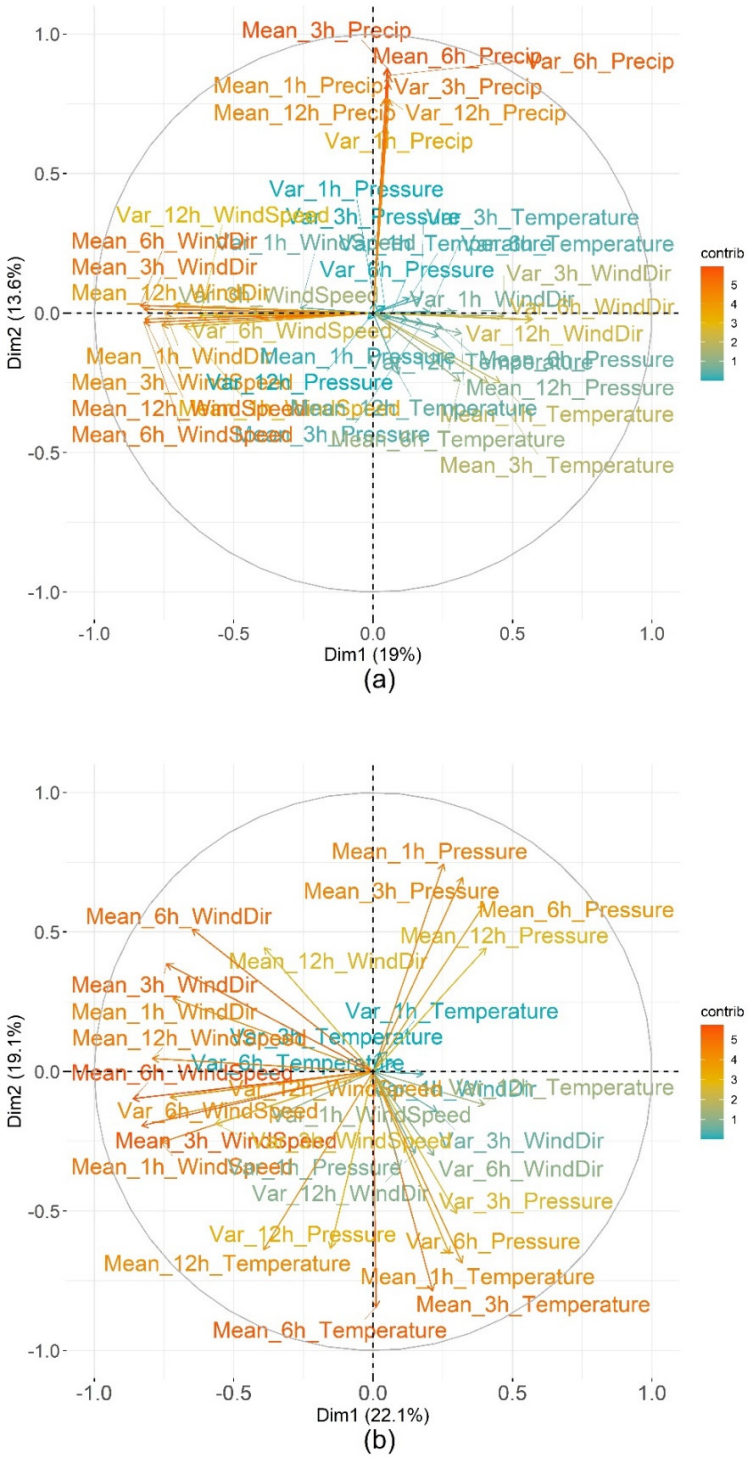

Figure S4. The PCA biplots showing (**a**) no heat wave and (**b**) heat wave events among all sites over 10 years.

As presented in **Figure S4,** the principal component analysis (PCA) biplots with or without the identified heatwave events. Most of the wind variables (e.g., wind directions,

wind speeds) contribute to principal component 1 (PC1) and the precipitation contribute to PC2. PCA results show a linear relationship between those variables. As we described in the paper, the means and variances of 12 h, 6 h, 3 h and 1 h were obtained for each point as shown in **Figure S4a**. PC1 and PC3 explain 32% of the total variance. If there is no heatwave during a period, the correlations between different hours' wind speeds are strong, that is, the same conclusion of the different hours' wind directions. Wind parameters contribute to PC1. Precipitation contributes to the PC2, and the temperature and pressure contribute to PC1 and PC2, respectively. Temperature and pressure have diurnal variations. The relationships among temperature or pressure variables are not as strong as those involving wind parameters. About 41% total variance can be explained from 1-, 3-, 6-, and 12-hour running averages of temperature, pressure, wind speed, and wind direction parameters during heatwave events shown in **Figure S4b**. The variations of the windspeed, wind direction, temperature, and pressure increase. The mean of the temperature contributes more to the variance of heatwave. All measured parameters contribute to PC1 and PC2.

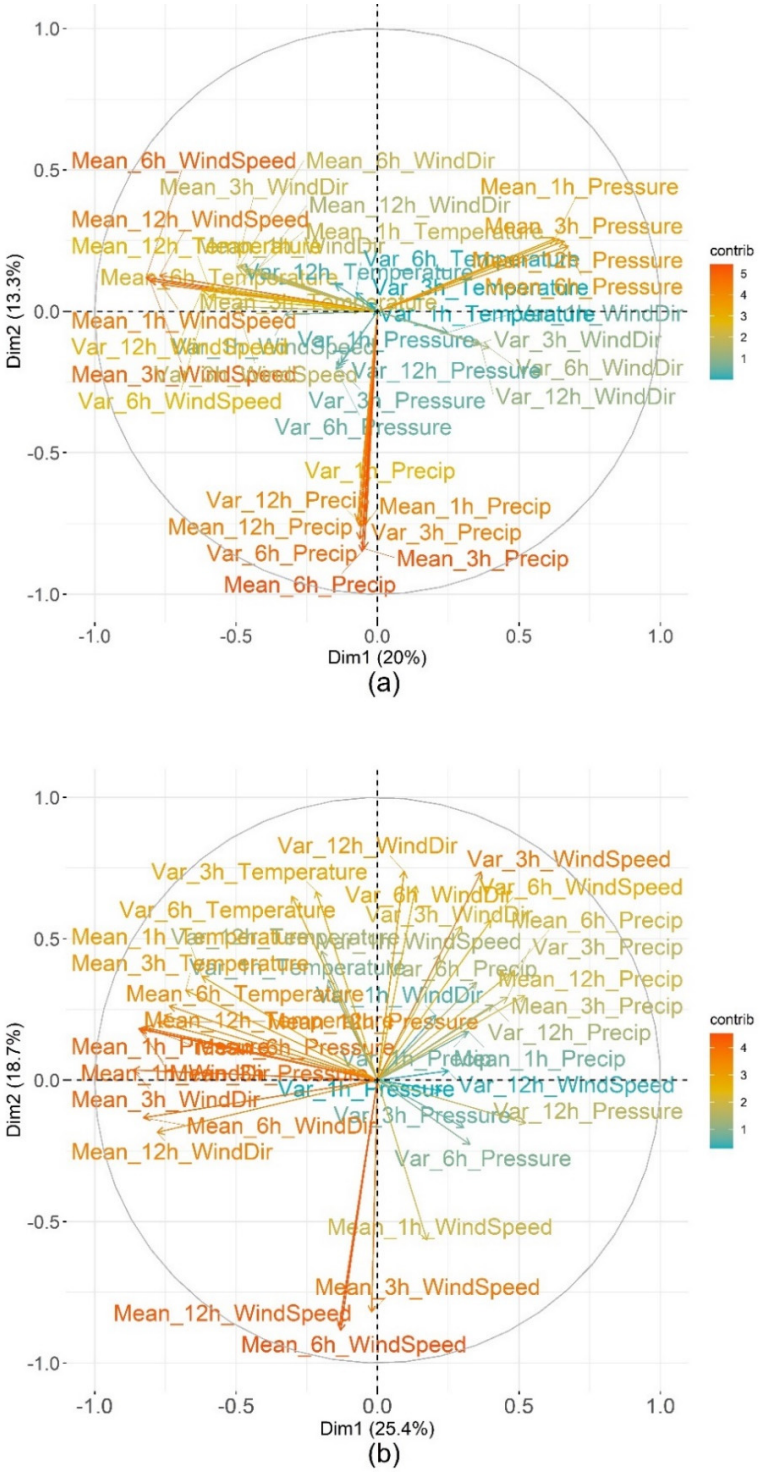

Figure S5. The PCA biplots showing (**a**) no strong wind and (**b**) strong wind events among all sites over 9 years.

The PCA biplots without strong wind periods (**Figure S5a**) and strong wind periods (**Figure S5b**) are depicted in **Figure S5**. **Figure S5a** is significantly different from the

Figure S5b. If there is no strong wind, the wind speeds and the wind directions contribute to PC1, so does pressure. The wind speed and the pressure have a negative relationship. Precipitation contributes to PC2. The means (i.e., 3 h, 6 h and 12 h) have a strong linear relationship among different types of parameters such as the mean of temperature averages for different time intervals. However, during strong wind, the variations of each categorical variable are bigger than those during low or calm wind. The mean values of wind speeds corresponding to different time intervals before the strong wind events contribute to the PC2. In our initial analysis, we learned that the wind speed changed in a short time. The 1 hour before strong wind events' wind speed is different from those when computing the 3-h, 6-h and 12-h intervals. For example, the relationship between the 12-h and 6-h before strong wind are strong. This may indicate that we could use means over different time intervals to relate to the wind impact in the local scale.

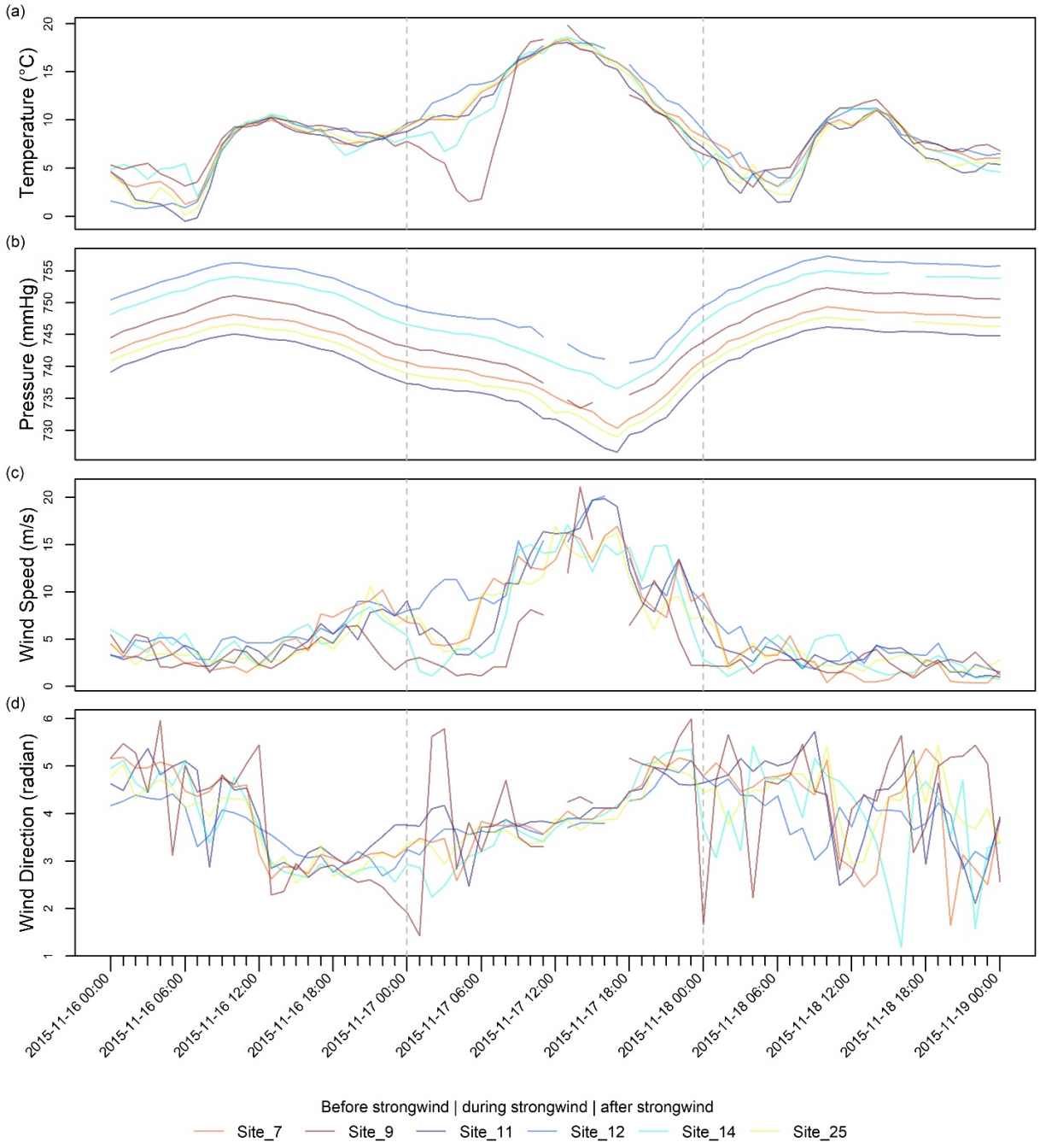

Figure S6. Time series plots of the day before, during, and after the strong wind event: (**a**) temperature, (**b**) pressure, (**c**) wind speed, and (**d**) wind direction. The grey dashed lines indicate when the recorded strong wind event occurs.

The time series plots of before, during, and after the strong wind event time series plot of temperature (a), pressure (b), wind speed (c), and wind direction (d) are depicted in **Figure S6**. This figure gives another example to support the main text discussion of **Figure 5**. During the strong wind period, the air pressure was lower than non-strong

wind periods (See **Figure S6b**). For example, the highest pressure of site 11 is around 757 mmHg and the lowest pressure of site 11 is around 740 mmHg. The maximum pressure difference between strong wind period and non-strong wind period is about 17mmHg. **Figure S6c** illustrates that the maximum wind speed is around 20 m/s (45 mph), and the non-strong wind period the wind speed is around 5 m/s. During the strong wind period, the wind direction is around 4 radians, but in non-strong wind period the wind direction varies from 2.6 radians to 5.1 radians for site 11. For site 9, the wind direction varies from 1.4 radians to 6 radians (see **Figure S6d**).

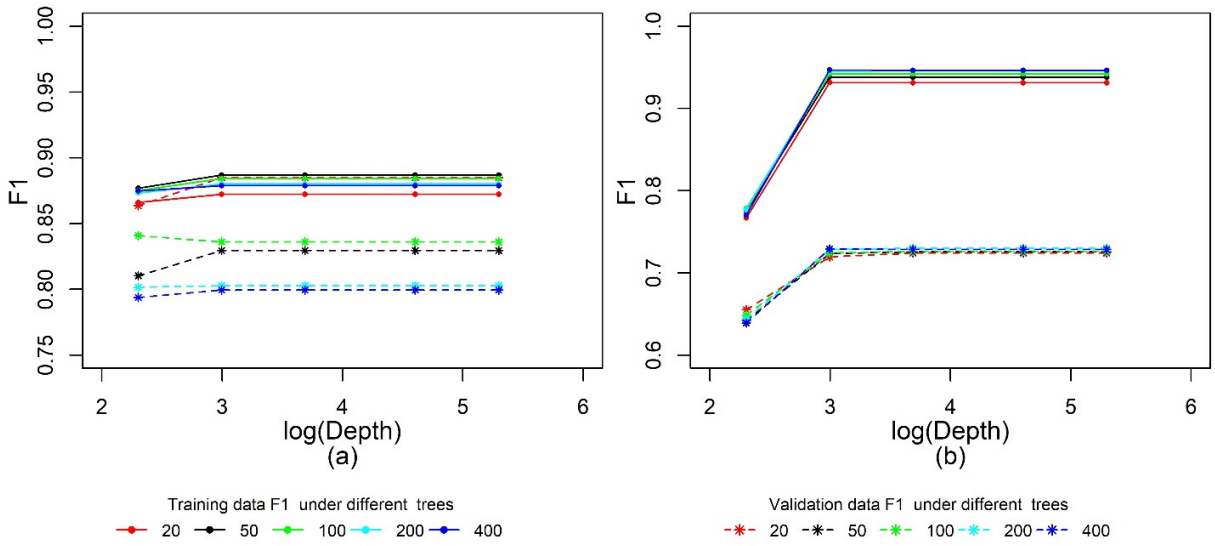

Figure S7. The F1 results of the RF models parameters tuning under different trees: (**a**) the strong wind and (**b**) the heatwave model.

Figure S7 represents F1 scores, which are obtained from the parameters tuning record of the RF models. The F1 score is the weighted average of precision and recall, and this score takes both false positives and false negatives into account. We tuned both the heatwave and strong wind RF models by using different trees, depths, minimum sample splits, and minimum sample leaves. The model depths are 10, 20, 40, 100, and 200, respectively. The numbers of the trees are 20, 50, 100, 200 and 400 that the RF model uses accordingly. The minimum splits are 2, 5 and 10, and the minimum sample leaves are 2, 4 and 6. As we mentioned before, the RF models using the unbalanced dataset, which means more records about normal conditions than extreme events. Therefore, we use the F1 score to evaluate the model performance.

The F1 scores from the strong wind models are presented in **Figure S7a**. The trees will not grow deeper when the depth is greater than 20. When depth is 20, the tree number is 50, both the F1 scores of the training data and validation data will not increase. The F1 score of the training data is about 0.89, and that of validation data is around 0.83. With the increase the depths or trees, the F1 score of training and validation data doesn't change. Thus, we select depth of 30 and tree number of 50 as the optimized model parameters for the heatwave model. **Figure S7b** gives the F1 score history of the heatwave model. When depth is 20 and the number of trees is 200, the F1 score of training data is about 0.95 and the validation data is about 0.73. Meanwhile, we also tuned the models under different minimum sample splits and minimum sample leaves, more details can be seen in **Figure S8** and **Figure S9**.

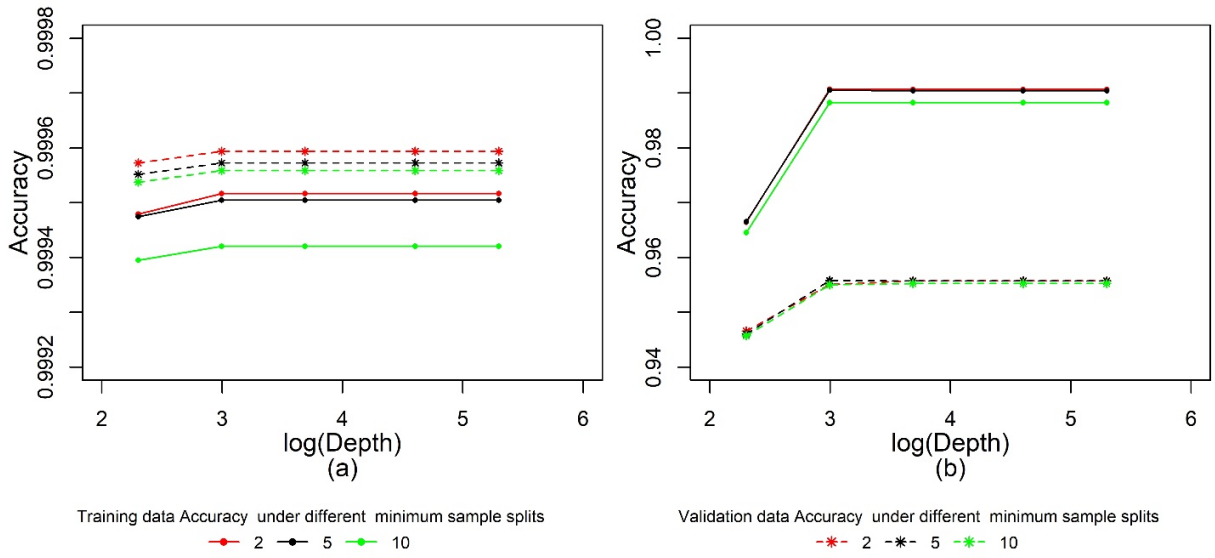

Figure S8. The accuracy of RF models parameter tuning under different minimum sample splits: (**a**) the strong wind and (**b**) the heatwave model.

Taking **Figure S8** as an example, the model accuracy of the strong wind model (see **Figure S8a**) and that of the heatwave model (see **Figure S8b**) yielded a better performance when the minimum sample split was set at 2 for both training dataset and validation dataset.

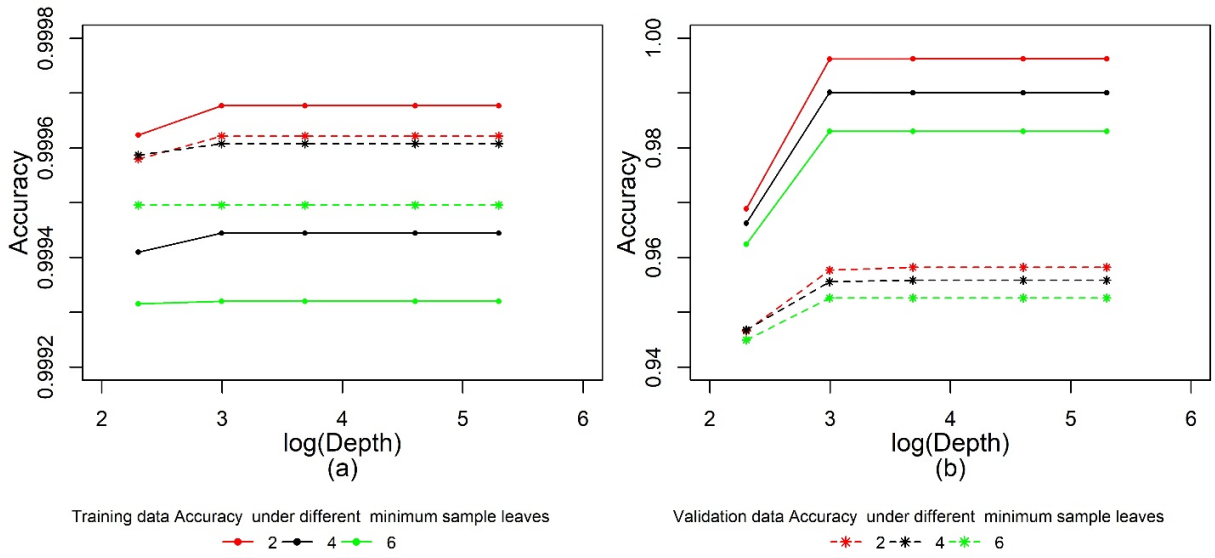

Figure S9. The accuracy of RF models parameter tuning under different minimum sample leaves: (**a**) the strong wind and (**b**) the heatwave model.

When the value of minimum sample leaves is 2, the model accuracy of strong wind model (see **Figure S9a**) and that of heatwave model (see **Figure S9b**) are better than other values. In this work, we did a comprehensive comparison and chose 20 as the optimized model maximum depths, 50 as the optimized model trees, 2 as the minimum sample leaves, and 2 as the minimum sample splits for the strong wind model. Similarly, we chose 20 as the optimized model maximum depths, 200 as the optimized model trees, 2 as the minimum sample leaves, and 2 as the minimum sample splits for the strong wind model.

Supporting Table

Table S1. Summary of the low wind period.

Start Time	End Time	Lasting Time
10/17/2010 6:00	10/22/2010 13:00	127
11/23/2010 14:00	11/30/2010 6:00	160
10/16/2013 13:00	10/22/2013 14:00	145
11/22/2013 15:00	12/1/2013 2:00	203
12/24/2013 17:00	12/30/2013 13:00	140
1/16/2014 19:00	1/23/2014 8:00	157
11/16/2014 20:00	11/21/2014 20:00	120
1/3/2016 14:00	1/9/2016 2:00	132
1/11/2017 16:00	1/18/2017 1:00	153
1/24/2017 6:00	1/29/2017 8:00	122
12/7/2017 7:00	12/14/2017 9:00	170
12/24/2017 21:00	12/30/2017 5:00	128
1/3/2018 14:00	1/9/2018 19:00	149
10/13/2018 18:00	10/23/2018 21:00	243
12/6/2018 17:00	12/11/2018 17:00	120
10/11/2019 18:00	10/16/2019 19:00	121
11/3/2019 2:00	11/10/2019 9:00	175
12/1/2019 16:00	12/8/2019 13:00	165

Table S1 shows the summary results of the low wind (<5 mph) start time, end time, and lasting period. In this work, we show the low wind periods which last more than 120 hours. During winter, the low wind speed condition happens frequently and usually lasts for a long time.

Table S2. Summary of the no precipitation period.

Start Time	End Time	Last Days	Station No.
6/26/2014 6:00	9/29/2014 19:00	95.5	1
8/15/2018 9:00	11/16/2018 13:00	93.2	15
11/13/2012 2:00	3/27/2013 0:00	133.9	3
7/23/2013 20:00	10/26/2013 12:00	94.7	3
5/23/2015 1:00	8/24/2015 16:00	93.6	4
5/4/2013 19:00	11/7/2013 14:00	186.8	5
1/7/2013 13:00	5/4/2013 12:00	117.0	5
6/17/2014 16:00	10/4/2014 10:00	108.8	5
5/4/2013 19:00	8/24/2013 7:00	111.5	7
5/23/2015 1:00	8/24/2015 16:00	93.6	8

Table S2 shows the summary results of the no precipitation periods' start time, end time, duration, and the station number. In this work, we define the no precipitation or draught periods as those that last more than 90 days.

Table S3. The model evaluation table of Table 2.

Model		Positive Predictive Value (PPV)	Negative Predictive Value (NPV)	Sensitivity	Specificity	Accuracy
Heatwave	Training data	1.0000	0.9991	0.9909	1.0000	0.9992
	Validation data	0.9200	0.9650	0.6419	0.9944	0.9621
	Testing data	0.9568	0.9648	0.6681	0.9967	0.9642
Strong wind	Training data	1.0000	0.9997	0.8679	1.0000	0.9997
	Validation data	0.7857	0.9999	0.9167	0.9997	0.9996
	Testing data	0.9091	0.9998	0.8333	0.9999	0.9997

The model evaluation table is shown in **Table S3**. **Table S3** gives the PPV, NPV, sensitivity, specificity, and the accuracy of the strong wind and heatwave model. Those values can help to learn the model performance. The predictive accuracies are greater than 0.95. The PPVs of the heatwave model are greater than 0.9, which indicates that the heatwave can be predicted.

References

1. da Silva, R.M.; Santos, C.A.G.; Moreira, M.; Corte-Real, J.; Silva, V.C.L.; Medeiros, I.C. Rainfall and river flow trends using Mann–Kendall and Sen's slope estimator statistical tests in the Cobres River basin. *Nat. Hazards* **2015**, *77*, 1205–1221, doi:10.1007/s11069-015-1644-7.
2. Hirsch, R.M.; Slack, J.R.; Smith, R.A. Techniques of Trend Analysis for Monthly Water-Quality Data. *Water Resour Res.* **1982**, *18*, 107–121, doi:10.1029/WR018i001p00107.
3. Tabari, H.; Marofi, S.; Aeini, A.; Talaee, P.H.; Mohammadi, K. Trend analysis of reference evapotranspiration in the western half of Iran. *Agricultural and Forest Meteorology* **2011**, *151*, 128–136, doi:10.1016/j.agrformet.2010.09.009.
4. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
5. Average Station Pressure (in). Available online: https://www.hanford.gov/files.cfm/Monthly_Average_Station_Pressure.pdf (accessed on 20 September 2021).
6. Metcalfe, A.V.; Cowpertwait, P.S. *Introductory Time Series with R*; New York, NY, USA, 2009; pp. 122–126.
7. Hanford Site Map. Available online: http://hanfordproject.com/ (accessed on 20 September 2021).