Project 1 Redwood Data Report

Elena Wang 1078994

9/19/2021

1 Introduction

1.1 Background

This project is motivated by the new advancement of equipment which is enabled to measure the temporal and spatial value dimensions in large volumes from physical and biological perspectives. The main purpose of this research is generating a sensor network called “macroscopes” to study the spatial variation and temporal dynamics from a coastal redwood tree in a microclimate environment. The research group led by Gilman Tolle designed a wireless micro-weather station in Sonoma California according to the Berkley Motes created by Crossbow and collected data from the wireless micro-scale motes.

In this paper, the author mainly describes how the whole environment works to discuss the process of data collection and applies multidimensional examination to analyze the dataset in order to be more profoundly comprehend the large amount of spatiotemporal data generated by the macroscope system.

Form the result of this research, the whole system can generate enough data to track the microclimate changes in temporal and spatial gradients over time, which means that it is helpful and meaningful for biologist to further study relative theories. Moreover, the new dataset obtained could be used in other ecosystem study and build models to observe the microclimate influences. Therefore, this particular macroscope benefits the observation of complex environmental dynamics and the data produced from this macroscope system verify the work of the sensor network deployment. The analysis could be simplified by the multi-dimensional method.

1.2 Data Collection

The case study of a wireless sensor network was collected over 44 days in early summer in the existence of a 70-meter tall redwood tree, at the density of at regular intervals in schedule: every 5 minutes in time and every 2 meters in space, and captured the sampling every 5 minutes to ensure the variation, which the most satisfy the microclimate dynamics. Nodes were equipped with around 2-meter spacing beginning at 15m from ground level to 70m from ground level to ensure sufficient gradients. The angular location was the west side of the tree since there is a thicker canopy and the most balance given against direct ecological impacts. Radical distance was from 0.1m to 1.0m from the trunk and nodes should be placed as close as possible to the trunk wo ensure the direct impact from microclimate. Each mote was collected by air temperature, relative humidity, and photosynthetically active solar radiation (PAR) including incident (direct) and reflected (ambient) levels and they are hamatop and hamabot respectively. These four variables are the main sources to do further analyze. In addition, PAR would provide the information about energy availability for photosynthesis.

The data generated in the mesh network from wireless download in every 5 minutes was stored in the Sonoma_data_net dataset. Another dataset called Sonoma-data-log dataset was from a local data logging system, which recorded every reading taken by every query until the 512 kB flash chip was full.

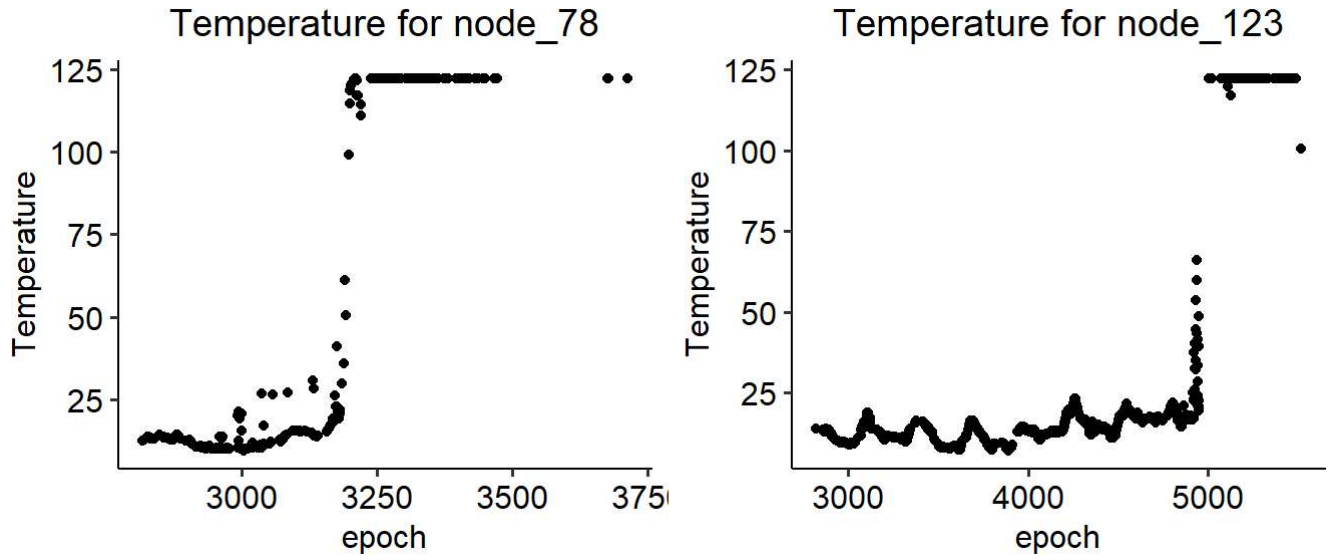
2 Data Cleaning

Data cleaning part is separated into two parts: 2.1 outliers and range conversion, and 2.2 data combination. Since data was collected into two different dataset, we need to combine them together to make sure we have all the data for analysis. And to also analyze the location of sensors, it is better to have a dataset combining three dataset. But before merging the data together, we need to check consistency, range and abnormalities for log and net dataset.

2.1 Outliers and Range Conversion

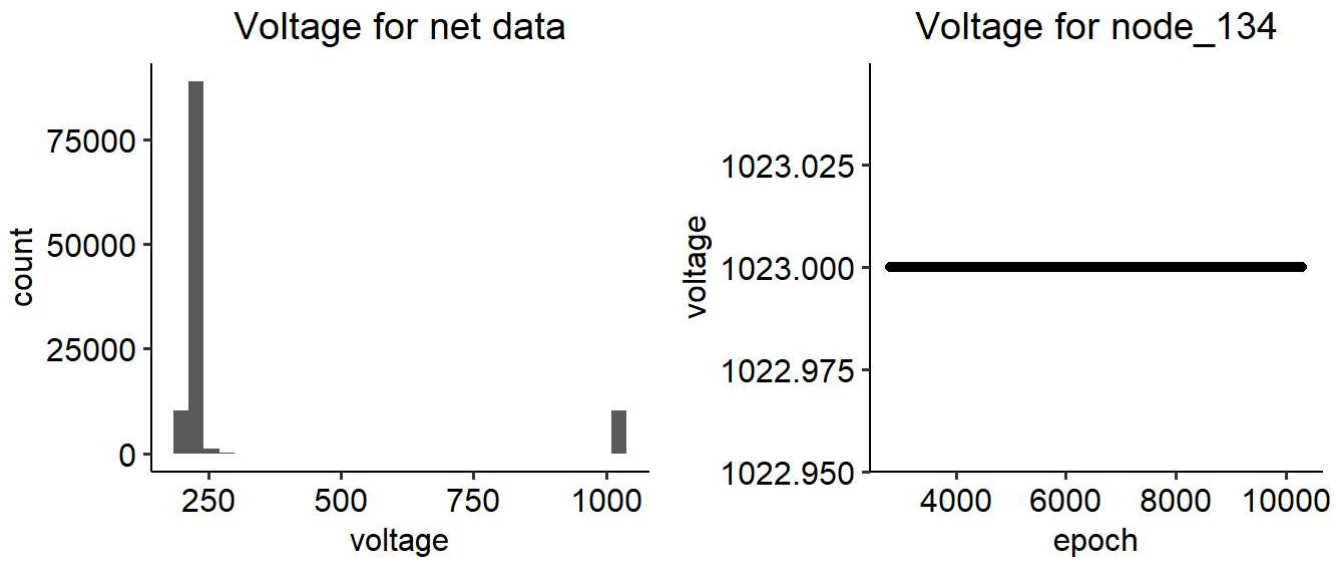
Since `result_time` does not represent the actual date and time for data, we first imported `sonoma-dates` and transferred it into json file to map with epoch. After matching with actual date and time, `sonoma-dates` was combined with both log and net dataset. Before checking abnormalities and outliers, we took a glimpse at data and found that there are 4262 missing values for net data, which took place from 5/14/2004 to 5/26/2004 (most happened at night) and 8270 missing values for log data, which took place from 04/30/2004 to 05/05/2004 (most happened at night). And missing values might influence later analysis, so we chose to remove missing values first.

After removing the missing values, we plotted histograms for each of the variable in two dataset: `humidity`, `humid_temp`, `hamatop`, `hamabot` and `voltage`. For `humidity`, the histogram for log data is extremely left-skewed because of large negative values. So, we filtered nodes in the log dataset whose values for humidity are negative and analyzed them one by one. There are only 3 nodes that have negative humidity values (node_29, node_198 and node_65535). To check if there is any abnormality, we created four histograms to look at the values for `humidity`, `humid_temp`, `hamatop` and `hamabot` for each of nodes. As for node_29, the histogram for both humidity and temperature values is just a horizontal line (a constant), which is wrong since we expect to see some changes as the change of epoch. So, we plan to drop the corresponding humidity and temperature variables for node_29 and set them as NAs. For node_198, there is only one extreme observation that has abnormal values for humidity, hamatop and hamabot, so we need to delete this specific abnormal observation. For node_65535, the entire data for node_65535 only includes one observation among all epochs, which is definitely wrong, so we deleted the entire node. After deleting all abnormal observations and outliers, the histogram for humidity values for log dataset looks better and and has a reasonable range. Then we processed similarly to humidity values in net dataset and found that there are also three nodes whose humidity values are negative (node_78, node_123 and node_141). For node_78, the histograms look quite strange after some epoch since there is a sharp jump and we expect to see at least a smooth change for those values and these values should not be discrete. So we choose to drop the node_78. For node_123, there is also a sharp jump for histograms after epoch 5000, so we choose to drop the node_123.



For node_141, values in the histogram after epoch 9000 go crazy (extreme large) without any smooth change, so we also dropped node_141. After deleting all abnormal nodes and outliers, the histogram for humidity in net dataset looks quite normal and has a similar range even though there are some large values and we need to take a closer look at it. After filtering the nodes whose humidity values are greater than 110, there are only two nodes (node_118 and node_145). For node_145, nothing looks wierd, so we kept it. For node_118, we checked the voltage value and found that it has extreme low voltage value that is close to 0, so it made us be suspicious more about the node and chose to delete it.

For `humid_temp`, the histogram for log dataset looks quite normal and has a reasonable range. While for net dataset, the histogram is extremely right-skewed, which means there are many extreme large values. So, we filtered nodes whose value for temperature is greater than 50 and there is only one node (node_3). And after 36 epoch, there is a sharp jump for temperature values and values go crazy to 100 degrees, so we chose to drop it. For `hamatop`, histograms for both log and net dataset look quite normal, but some values for log data are extremely large, so we filtered the nodes (node_40) whose values for hamatop is greater than 150000. And we found that there are sharp jumps for both hamatop and hamabot values, so we dropped these two values for node_40. What's more, to make the range consistent with the paper, we transfered values for hamatop and hamabot from Lux to PPFD ($PPFD = \frac{hamatop/hamabot}{54}$). For `hamabot`, everything looks normal for both log and net dataset, which is a good sign for previous data cleaning process. Finally, for `voltage`, there are many abnormal extreme small values (close to 0) and extreme large values (close to 1000). After filtering abnormal nodes, we found that those voltage values are constants from the start to the end, which might signal that these sensors read values incorrectly. So, we chose to assign them NA for voltage values.



Furthermore, two dataset have different ranges for voltage values, so we need to transfer them into the same range. After fitting a linear regression method, it is obvious that we could use linear regression formula to transfer voltage values in one dataset to the same range as another dataset.

2.2 Data Combination

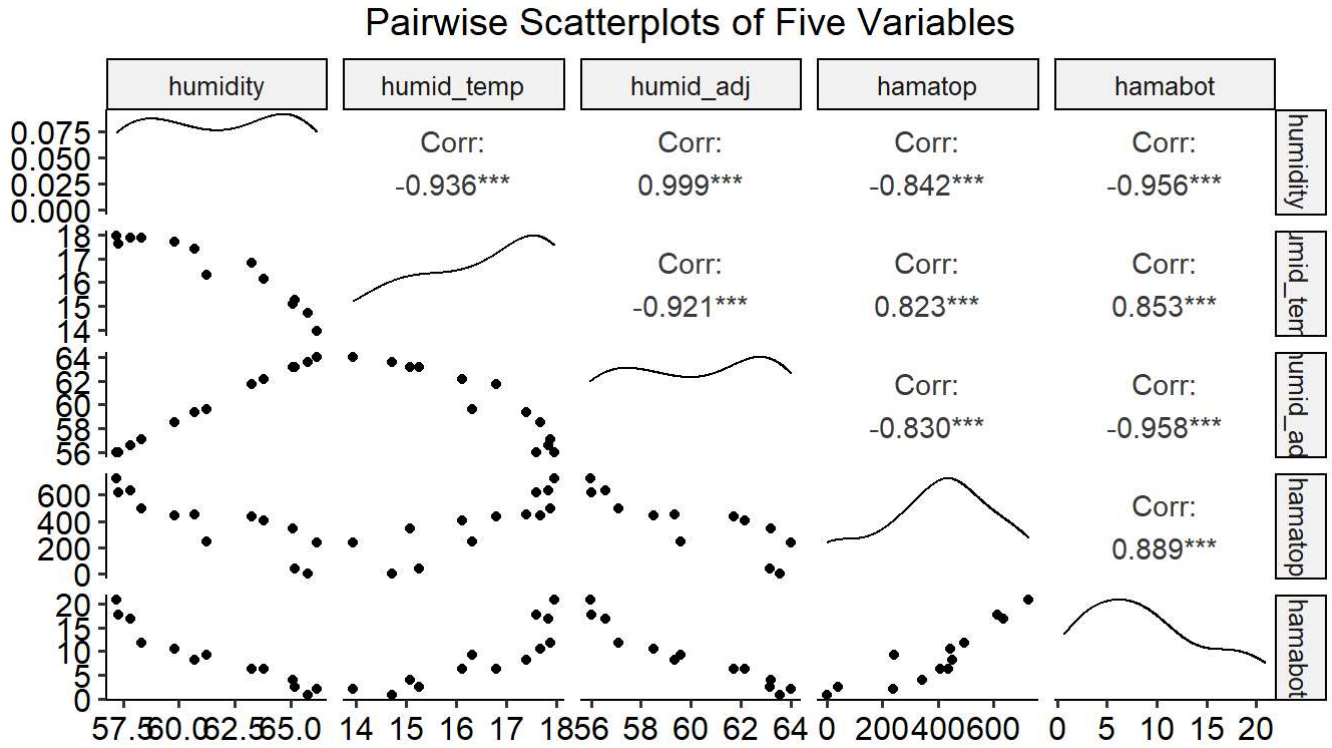
After removing missing values, abnormalities, outliers and transferring data into same range, finally we could merge two log and net dataset together. At first, we chose to use inner join to merge common observations for two dataset and checked if these values for each variable are the same. By creating a `difference` variable, we could calculate the difference between two values for the same observation from two dataset to check if they are the same. And values from two dataset for the same observation are quite close to each other, so we calculated their mean as the true value for this specific observation for each variable in the dataset. Then, we deleted the original

values from both log and net dataset and only kept the mean we mutated before as our true values for `log_net_combined` data. What's more, since we used `inner_join` at first, we need to add complements from both log and net dataset to make sure we have the full combined dataset. So, we used `left_join` to combine log data and log_net_combined data and filtered observations whose `hamabot/hamatop` values for log_net_combined data are NAs, which means that those observations are missing for the combined data and these observations are complements. Then, we did the same thing for net data and found the complement for both log and net dataset. Finally, we merged net, log, net_complement, log_complement and loc data together as our final cleaned combined data called `log_net_loc.csv` .

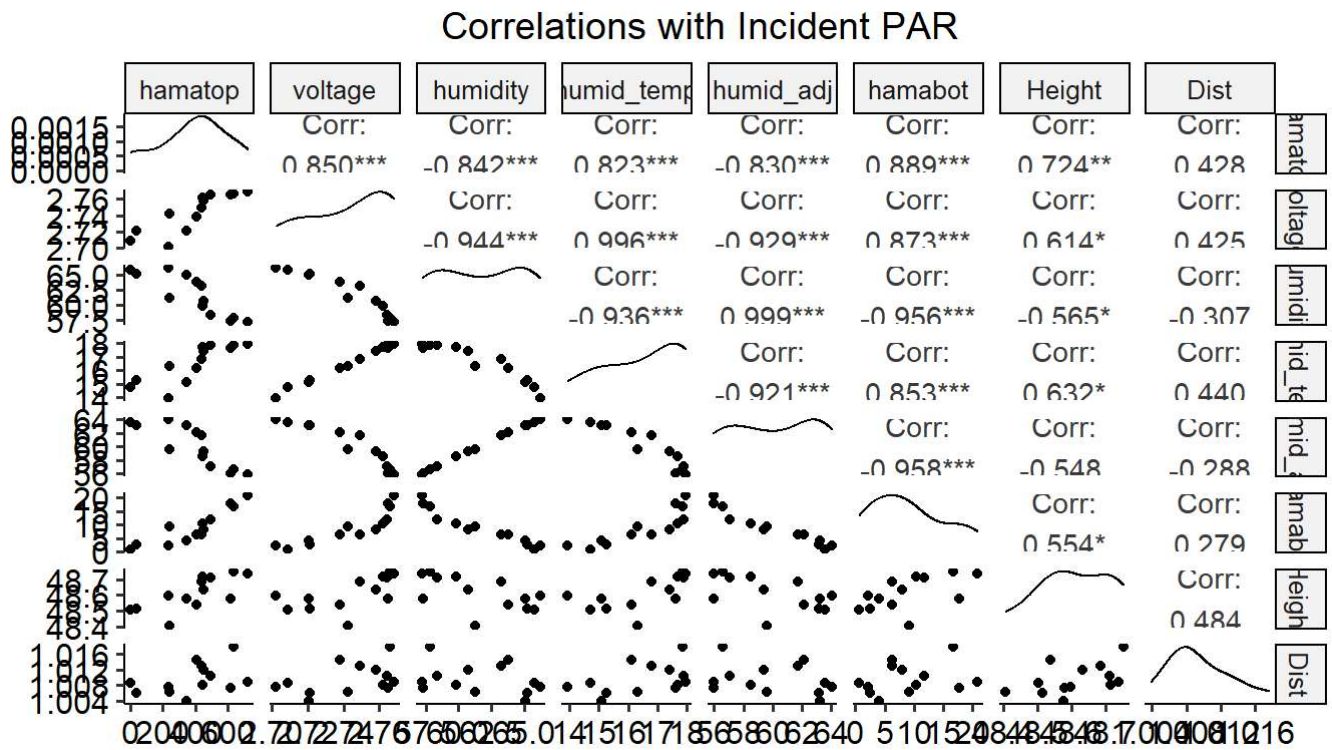
3 Data Exploration

After cleaning dataset, Dates variable is not well organized. For better observation in future data exploration, the first thing is separating the dates as year, month, day, weekday and day and reformatting dates as well-designed dates in `log_net_loc` dataset. In order to do time series in hour, hour variable should also extract from time variables for future evaluation.

To see the pairwise scatterplots, since the dataset is too large, the reasonable range of variables should be determined. Since in the daytime, the sunshine would better reflect PAR, the time chosen is from 8am. to 8pm. every day. To better compare and clearly plot, taking numeric values as average would make more sense. From the the pairwise scatterplots, all of five variables (`humidity` , `humid_temp` , `humid_adj` , `hamatop` , and `hamabot`) have really high correlation between each other. `Humid_adj` and `humidity` have the highest correlation (almost 1), which is reasonable since `humid_adj` is calculated from `humidity` .



To see which variables associated with Incident PAR, Incident PAR are compared by `voltage` , `humidity` , `humid_temp` , `humid_adj` , `hamabot` , `height` , `Dist` , and the result show that have pretty high relationship with these variables except `Dist` .

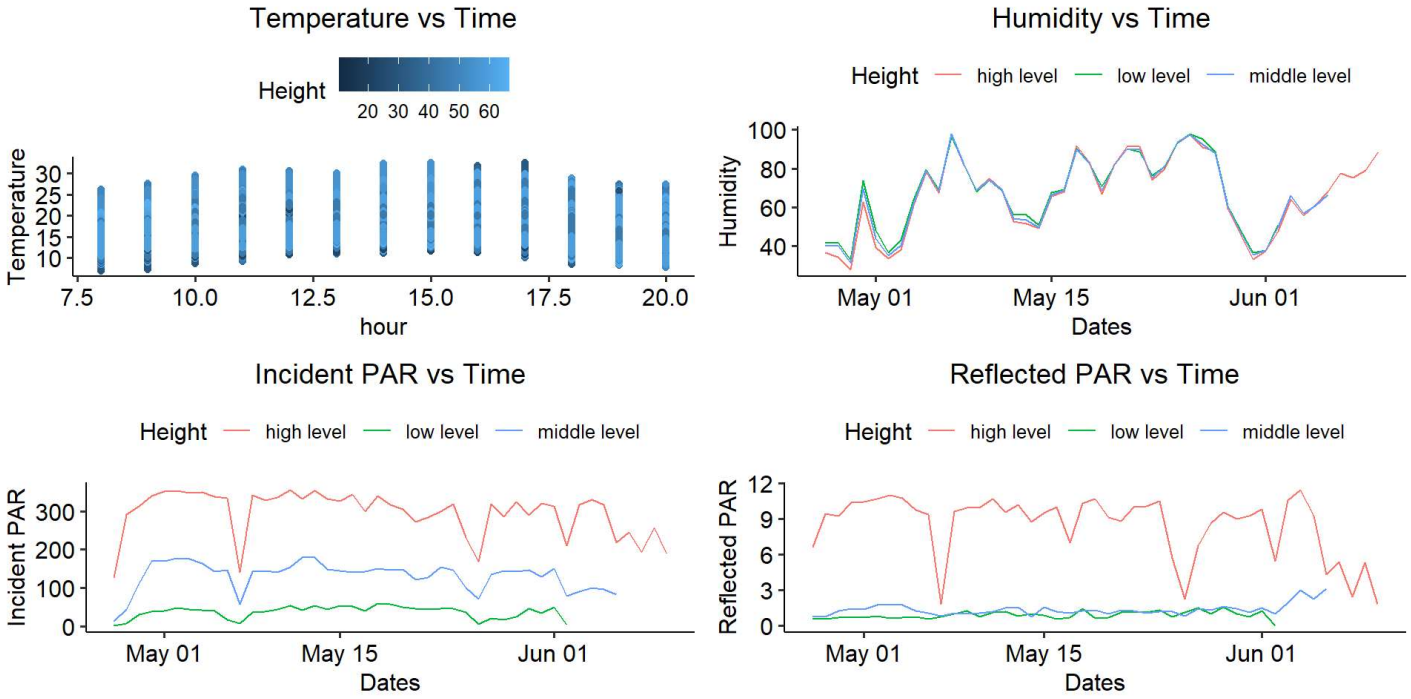


The three dimensions: value, height and time are explored more deeply by the plot graphs with height as color cue to see the temporal trend of values. The first graph is temperature vs time 8 am. to 8pm. since it's normal that the temperature would be regularly changed by day and height. From the plot, temperature have obvious trend that it's lower in the morning and afternoon but higher in the noon, which is normal. According to the height color, height and temperature have positive relationship in general(higher height have higher temperature), which is reasonable since higher node would gain more sunlight. The range of humidity is from 6.9446 to 32.5814 and because of noncontinuous dependent variable, plot is still noncontinuous.

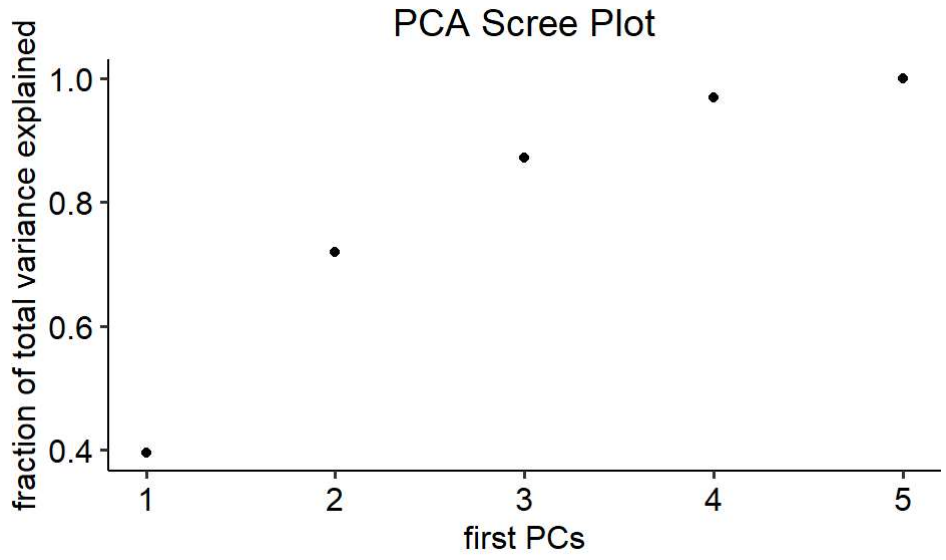
The second graph is humidity vs time. Since time would be obviously changed in days rather than in hour, time range is selected with days during the whole experiment. The value of humidity is taken as average with the same reason above. In order to observe the corresponding height changes, height is separated to different levels: low, middle, and high level with lower than 30, from 30 to 50, and greater than 50 respectively. The plot result tells that the humidity of three levels height has the same tendency. Before May 15th, lower height has higher humidity and higher height has lower humidity, however, in the middle of May, higher height has higher humidity. During the last half month of this experiment, they performed similar. The range of humidity is from 27.84231 to 97.97209 and from the plot graph, humidity may not be continuous since there is a obvious gap at the end of May.

The third graph is showing incident PAR (`hamatop`) and time. To be more accurate, the time is still by days during the entire experiment, so that the whole trend in the experiment would be clearly presented. The conclusion of this graph is that they have same tendencies in different levels of height, and height and incident PAR are positively correlated. The reason of it is that higher mode would obtain more sunshine. Moreover, the incident PAR have regular change in time series. The range of incident PAR is from 1.098938 to 356.325300, which is large range due to different high levels. However, the plot shows that it has continuity.

The last graph demonstrates reflected PAR and time which is by say during the entire experiment. From the line plots below, hamabot in high level is obviously higher than in middle and lower level during the whole experiment. Nodes in middle and low level have very close reflected PAR, although middle PAR is bit higher than low PAR. The range of reflected PAR is from 0 to 11.43551. The continuity is showed in low level and middle level, however, in the high level height, it loss some continuity after June 1st.



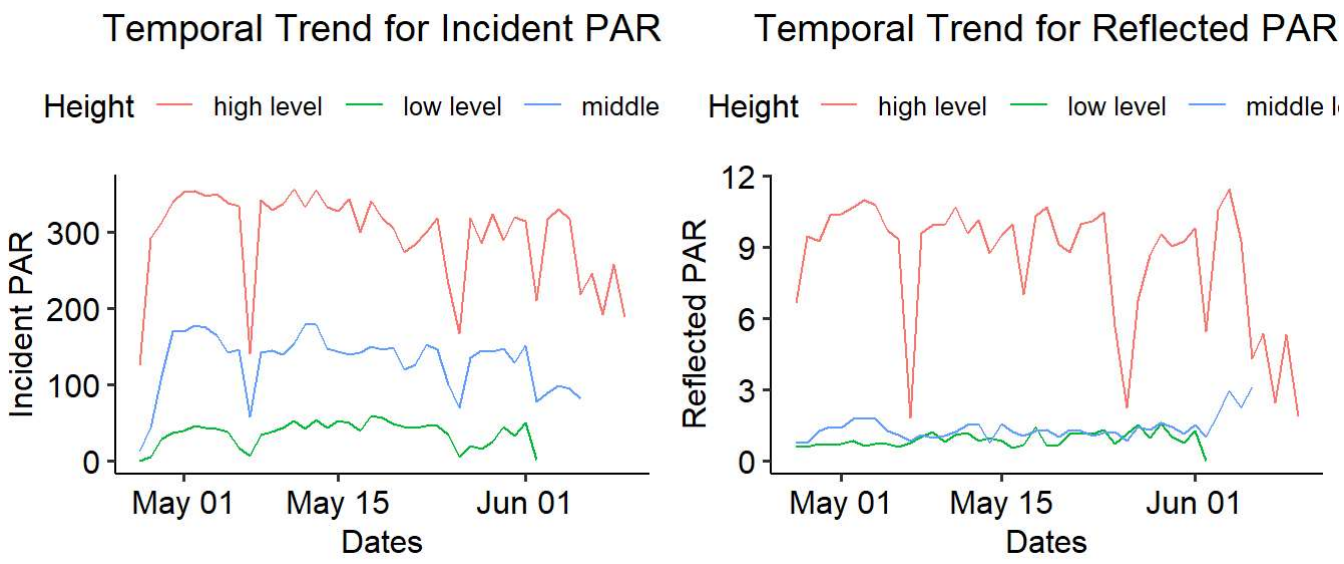
From the PCA analysis and scree plot, this data could be approximated by the lower-dimension representation since from the first three PCs, the fraction of total variance really close to 90%, which means that the first three PCs could have enough information to explain the dataset.



4 Interesting Findings

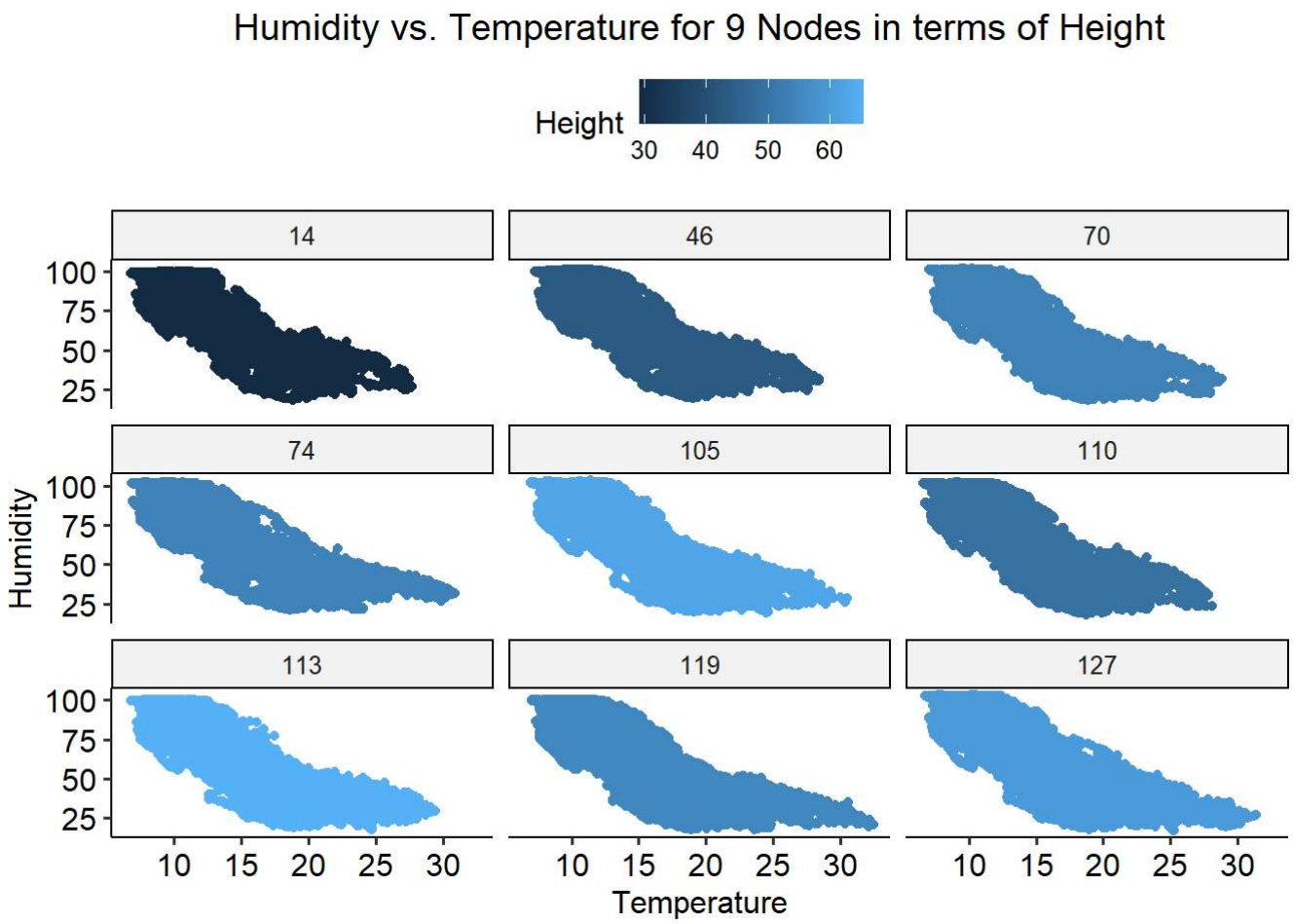
Interesting Finding 1

Sensors with high level of height have the highest values for both incident PAR and reflected PAR. Sensors with middle level of height have higher values for incident PAR than sensors with low level of height. But sensors with middle level of height have similar values for reflected PAR than sensors with lower level of height. What's more, for incident PAR, sensors with different levels of heights have similar patterns, but for reflected PAR, sensors with different levels of heights have different patterns, which means heights influence the temporal trend for reflected PAR but not for incident PAR.



Interesting Finding 2

We are interested in that if heights change the relationship between temperature and humidity so that there might be microclimate. At first, we filtered 9 nodes from different height ranges (3 from height < 40, 3 from 40 < height < 60 and 3 from height > 60) that have more than 6000 observations to make sure we have enough data to analyze. Then, we selected the corresponding humidity, temperature and nodeid variables to combine these nodes together as one dataset. Finally, we plot a scatterplot with facet and color function to see the pattern for each node as well as the pattern for different heights. And we noticed that for each node, there is a similar negative relationship between temperature and humidity (lower temperature means higher humidity). At the same time, in general, the relationship doesn't change for different heights, which means that heights might don't affect the relationship between temperature and humidity and there is no microclimate associated with heights for temperature and humidity.



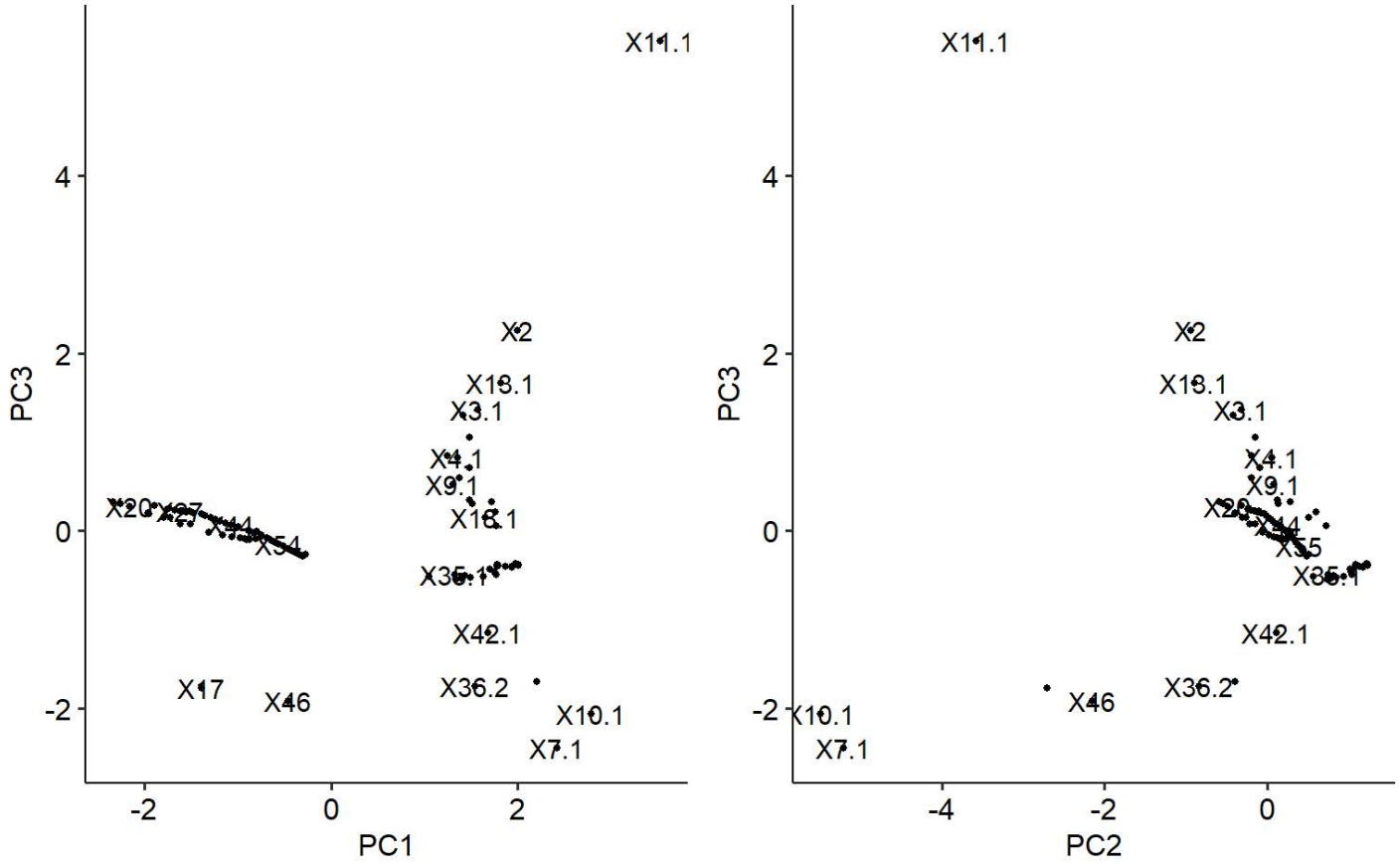
Interesting Finding 3

From PCA analysis above, the number of epoch perhaps would be well predicted from the first three PCs to see which epochs stand out. We would randomly select the first 100 epochs to see which one would stand out compared to in the 1st & 3rd PCs and 2nd & 3rd PCs. In this case, the epoch is the ID identifier and since we have same number of epoch, they should label as .1,.2 etc. Compared to two graphs of PCs, it is good and interesting to see there are some epochs stand out in both of two analysis: epoch 17, 46, 11.1, 10.1 and 7.1, which means that biologist may should adjust these notes and to think about why these notes is abnormal from others in the future more accurate observation.

log_net_loc_Date.epoch <int>	n <int>
2	27
3	62
4	62
5	62
6	62
7	62
8	62

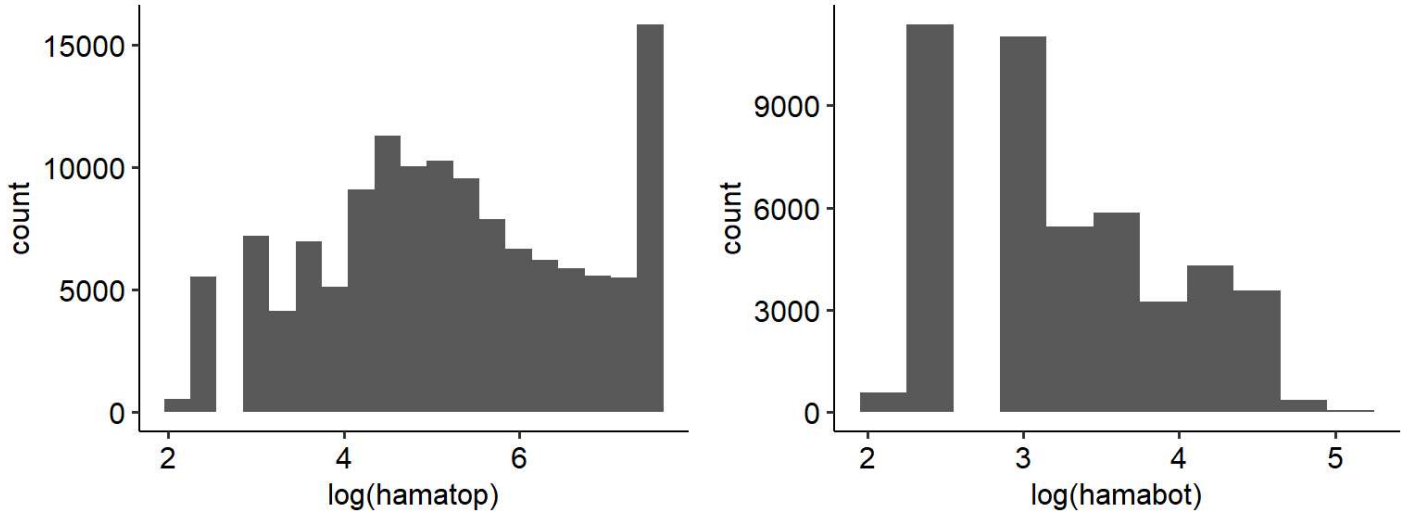
log_net_loc_Date.epoch	n
<int>	<int>
9	62
10	62
11	62

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

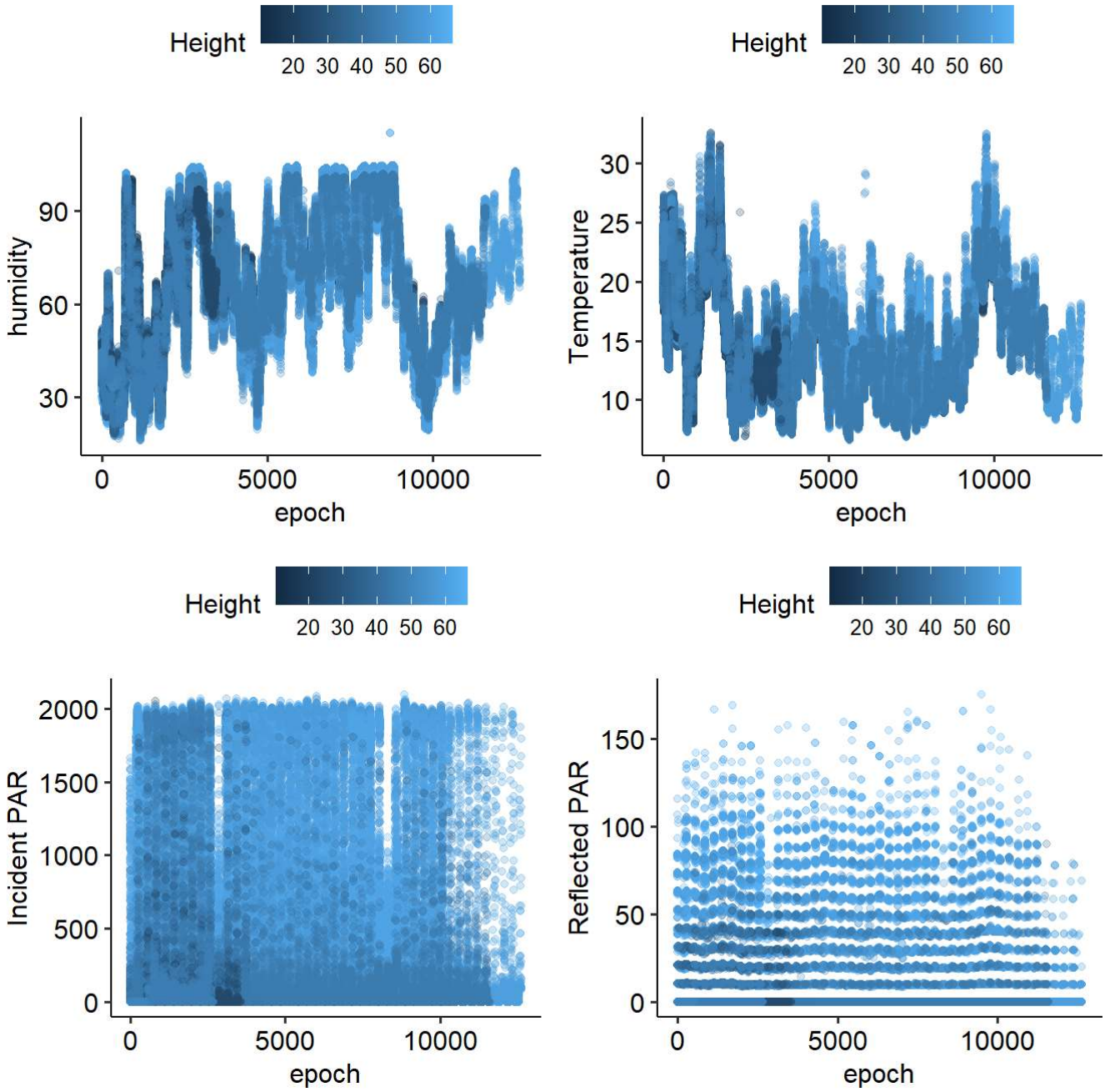


5 Graph Critique in the Paper

Graphs in 3[a] don't show all information about Incident PAR(`hamatop`) and Reflected PAR(`hamabot`) because of extreme right-skewed tails (maybe because of outliers or we need to plot with the log transform of data). Here are the plots for Incident PAR and Reflected PAR after removing outliers as well as transforming data into log form):



For Figure3[c], the author tries to visualize the sensor reading distributions for four variables in terms of the node height (spatial data), for example, what is the distribution for temperature values when the node height is 60.1m. For Figure3[d], they try to visualize the spread of sensor readings from the mean of the distribution in terms of the node height (spatial data). But both of them didn't take time into account, so it is better to visualize distributions in terms of both height and time (spatial and temporal). When time is missing, it is hard to explain the plots since we couldn't assume that all data took place during the same time period, or we will make unfair comparisons for sensor readings. At the same time, those boxplots are too small to read, and different boxplots are too close to each other, so it's hard to see differences between different distributions for sensor readings.



For Figure 4, the author used color to identify different height, which is nice, but I couldn't identify which color represents which height of the node, so there should be some explanations or show the legend for height. At the same time, lines overlapped with each other and it is harder to see the trend for spatial data, which makes the color useless. And there are too many colors, so it's better to categorize similar heights as one color instead of assigning each height a color. Finally, the author also didn't explain why they choose the day 5/1/2004 to plot spatial data and what if the day is totally different from other days (extreme weathers). My suggestions are using time/epoch as x-axis and coloring height to plot each variables, and categorizing similar heights as one color as well as choosing thicker bins and more transparent colors.

For Figure 7, we could combine plots together to better see the difference for two dataset. For example, we could create a new categorical variable called `type`, and if the data comes from log, we write "log" for `type` variable, and if the data comes from net, we write "net" for `type` variable. Then, we are able to combine log data and net data together as a new data including new categorical variable identifies the source of data. When we plot this new data, we would `color/facet` the new categorical variable `type` to specify the source of the data so that we could see data from two dataset together in one plot.