Hemlock removal experiment – Air and Soil Temperature (HF108)

Data standardization and correction, data 2004-2010

This dataset contains temperature data recorded from 2004-2010, at the Hemlock Removal Experiment in the Simes Tract in Harvard Forest. Every hour, the mean, maximum and minimum air and soil temperature were recorded. This dataset is a combination of the old datasets hf108-03 and hf108-04, which were combined, standardized and filtered for questionable data.

1. Data standardization

The old datasets hf108-03 (for 2004 – July 2008) and hf108-04 (for July 2008 – 2010) were first merged into one dataset. During the time period of hf108-03, soil temperature loggers were only present in the mineral soil. Hf108-04 was started when temperature sensors were placed in the organic soil as well. The mineral soil columns of hf108-03 and hf108-04 were therefore merged, and the organic soil variables were filled with NA for the time period of hf108-03.

The hour column was standardized, because midnight was sometimes assigned to the day before (as 2400), and sometimes to the next day (as 000). Midnight is now noted as 2400 and assigned to the previous day, in the whole dataset. There was a double entry at 2008 day 197 at midnight, which consisted of two rows (000 and 2400); the row with hour=2400 was deleted because it contained values that did not match the surrounding data, while the row with hour=000 did. On 2008 day 198, midnight was missing, and a row of NAs was added with hour = 2400.

From 2008 day 157 to 197, the data was recorded four times an hour, instead of once an hour. This data was converted to hourly data by taking the average (for the means per quarter hour), minimum (for the minima per quarter hour), or maximum (for the maxima per quarter hour), of the data points .15, .30, .45 and .00 of every hour.

Two columns were added to the dataset. The column "ym" contains the year and the month number of every row; the column "yw" contains the year and the week number of every row. Week numbers follow the conventional week numbers, with the exception of the weeks at the beginning and end of each year. Week numbers are cut off at the year boundary, such that week 52 or 53 always runs up to December 31st (even if week 1 has officially started already), and week 1 always starts January 1st (even if that was officially still week 52 or 53). Weeks 52, 53 and/or 1 may therefore be longer or shorter than 7 days.

The standardized data was saved as *hf108-03-air-soil-temp.csv*.

2. Filtering the data

Main filter

The dataset was filtered for clear outliers, which seemed to be caused by data logger errors. First the whole dataset was filtered using the first part of the R script *hf108-05-R-QAQC.txt*. The difference between the maximum and minimum value of every hour was used as a method to detect outliers; this difference should not be too big, or negative. For every variable, a reasonable cut off value was set at which the difference would be too big and most outliers would be filtered out. The values that were filtered out were replaced with NA. See the R script for the exact code, and the cut off values that were used for each variable. See Figure 1 for an example of the filtering method.

The filtered data was saved as *simes_microclim_filtered.csv.*

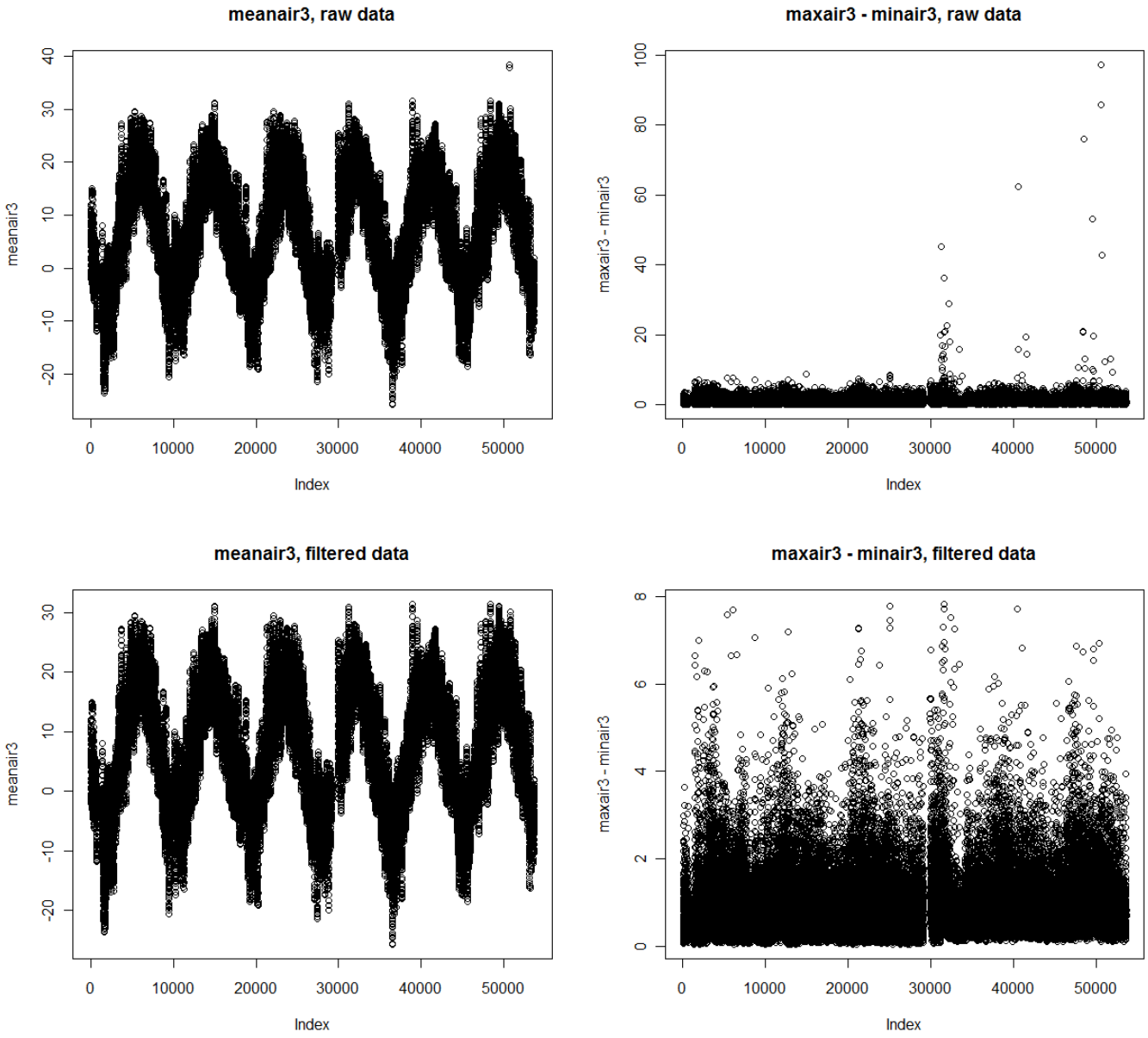


Figure 1

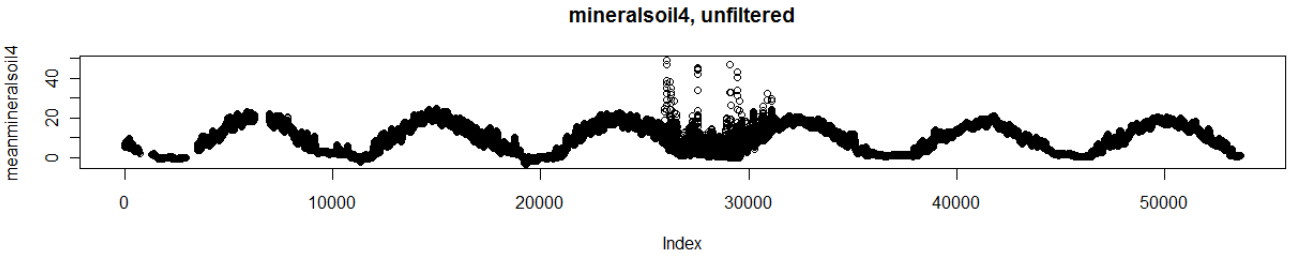
Additional filtering in the winter of 2007-2008

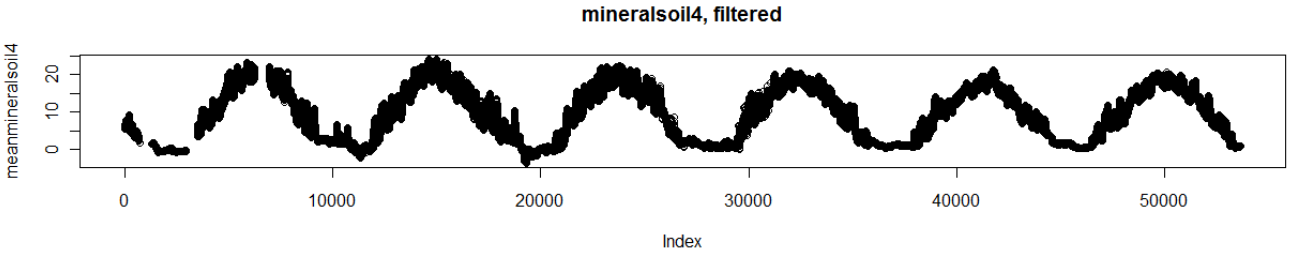
It turned out that for plots 4, 5 and 6 (from the ridge block), there is a period in the winter of 2007/2008 with a lot of outliers, particularly for the mineral soil data. Because these plots share a data logger, the outliers are probably caused by data logger errors. An extra filter was created for this data (see the second part of the R script *hf108-05-R-QAQC.txt*).

The filtering procedure for the mineral soil data was as follows:

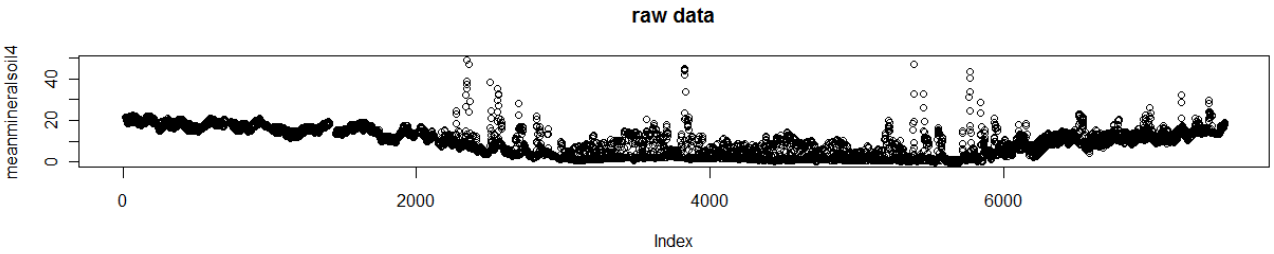
- First, the filter of maximum – minimum values was repeated, but with a cutoff value of 0.6.
- Next, values were filtered out that were too different from values in the corresponding plot in the valley (comparing logged plot 4 with logged plot 2, etc.). If the value in the valley plot was NA, the value of the ridge plot was kept.
- Finally, an extra filter was applied to the period between week 48 of 2007 and week 13 of 2008. This period was very flat in temperature in the other plots, but contained many outliers in the mineral soil plots 4, 5 and 6.

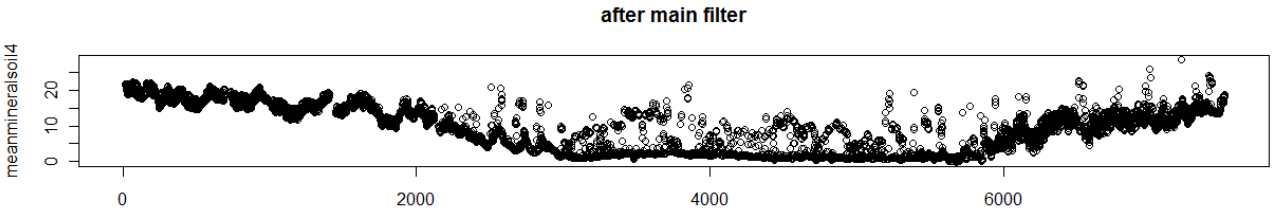
After all these filtering procedures, the results are (for mineralsoil4):

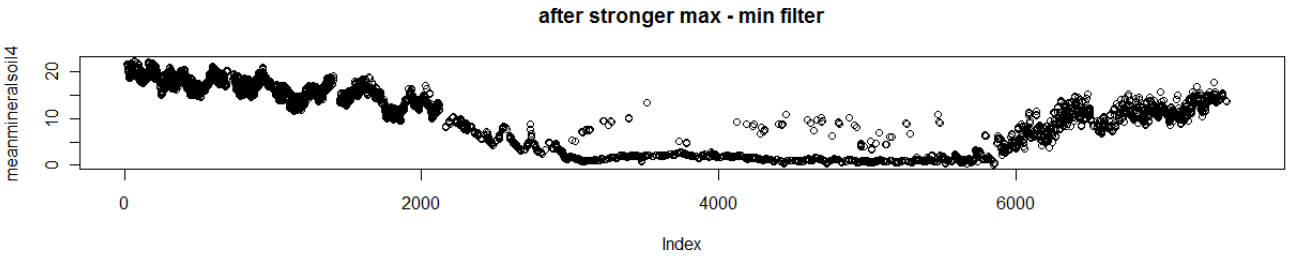


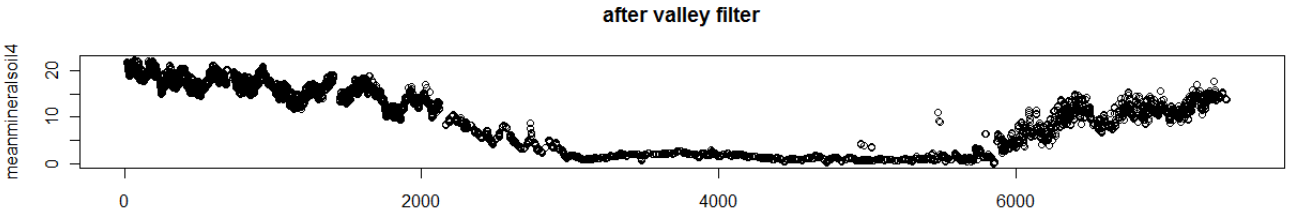


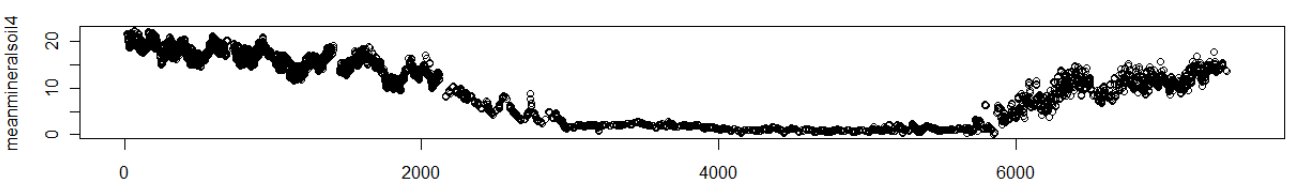
ZOOMED IN TO WINTER 2007 / 2008











The filtering procedure for the air data was similar to the procedure for the mineral soil data, but contained one step less:

- First, the filter of maximum – minimum values was repeated, but with a cutoff value between 2 and 5.
- Next, values were filtered out that were too different from values in the corresponding plot in the valley (comparing logged plot 4 with logged plot 2, etc.). If the value of the valley plot was NA, the value of the ridge plot was kept.
- There was no additional filtering procedure of a specific part of winter 2008/2007, because the air data looked better than the mineral soil data.

See the R script for the cutoff values of all the filters.

Final datafile

The filtered data for winter 2007/2008 was manually inserted into the dataset that was created using the main filter (*simes_microclim_filtered.csv*), replacing the old values in the winter of 2007/2008. The final dataset is called *hf108-04-air-soil-temp-filt.csv.*