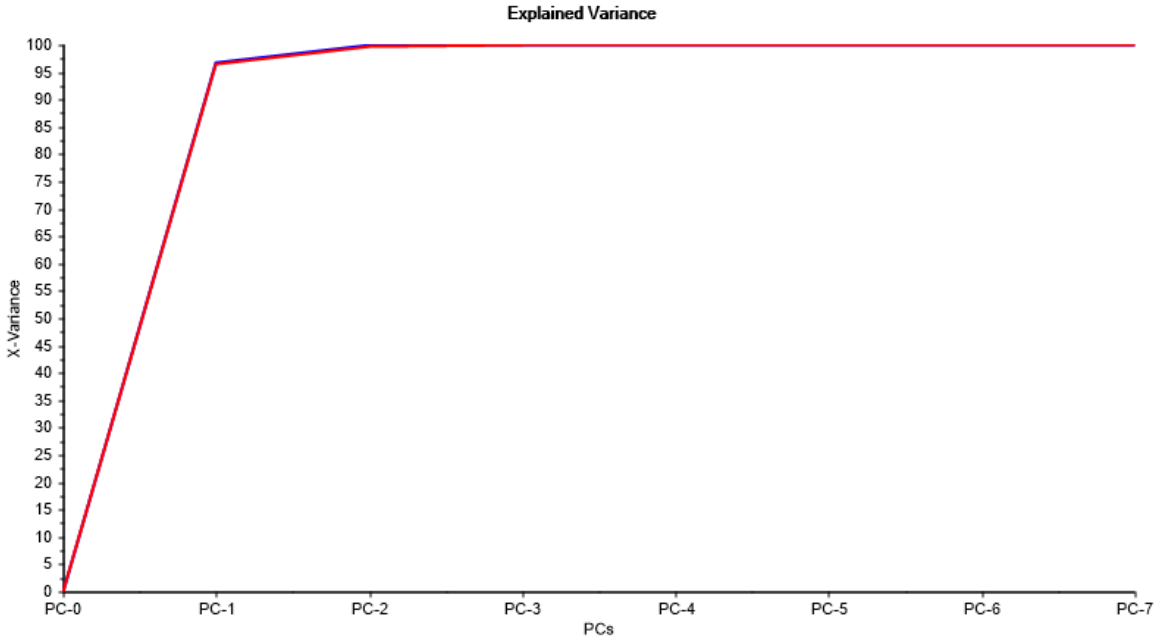
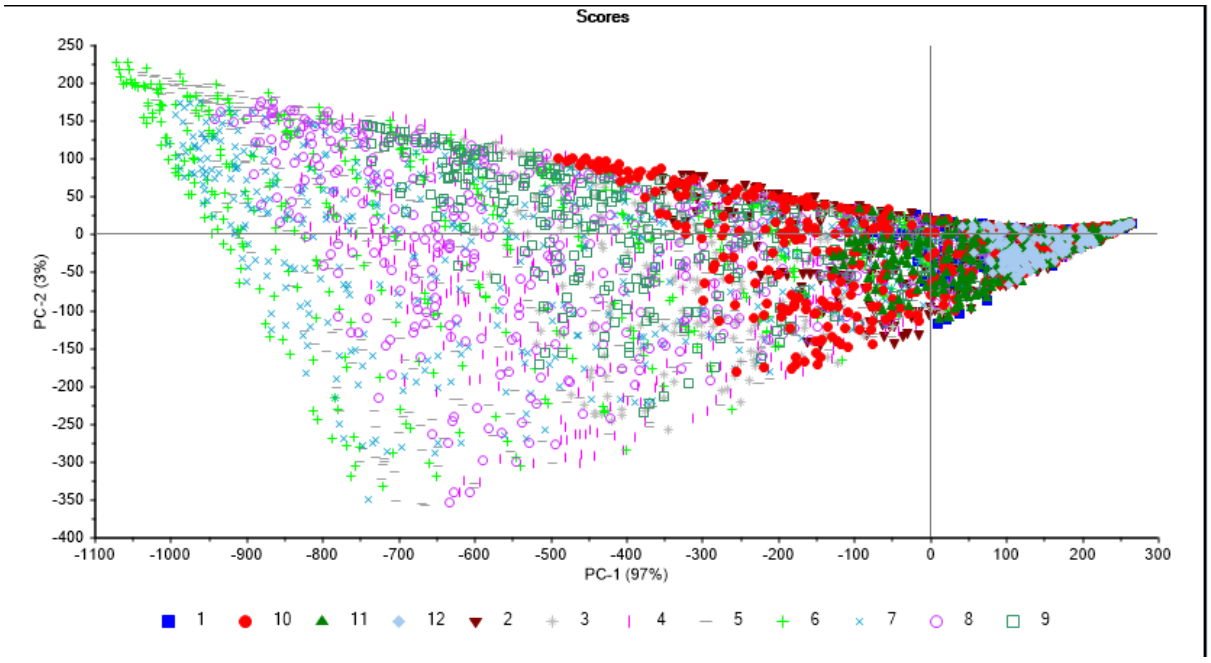
Assignment 5

Interpret scores loadings and explained variance

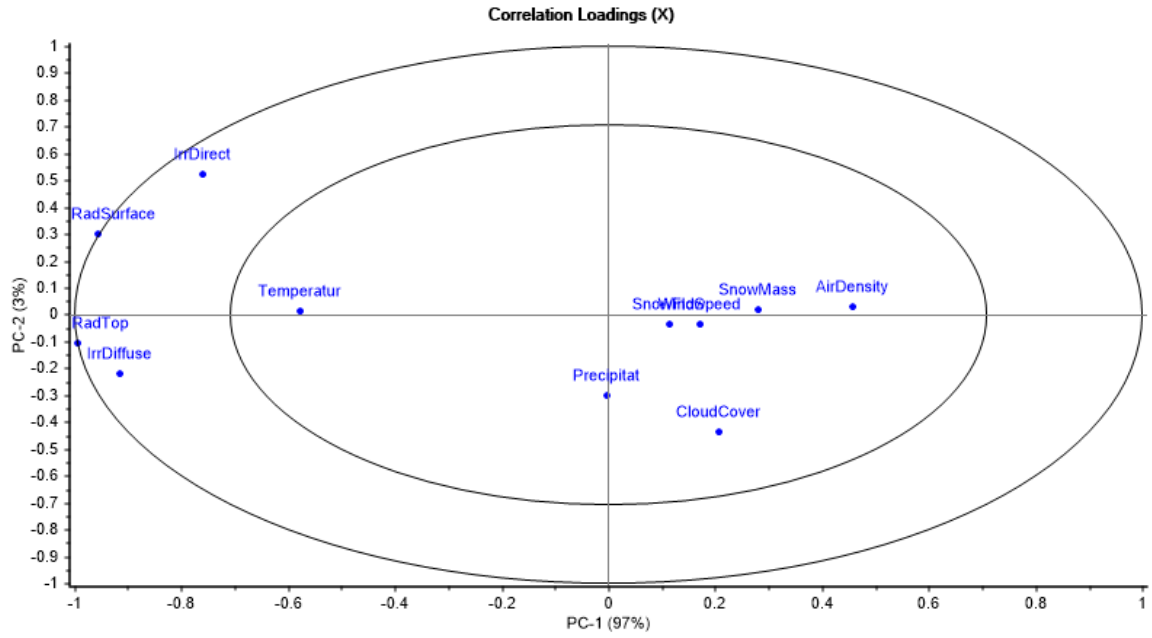
First, this is without any scaling.



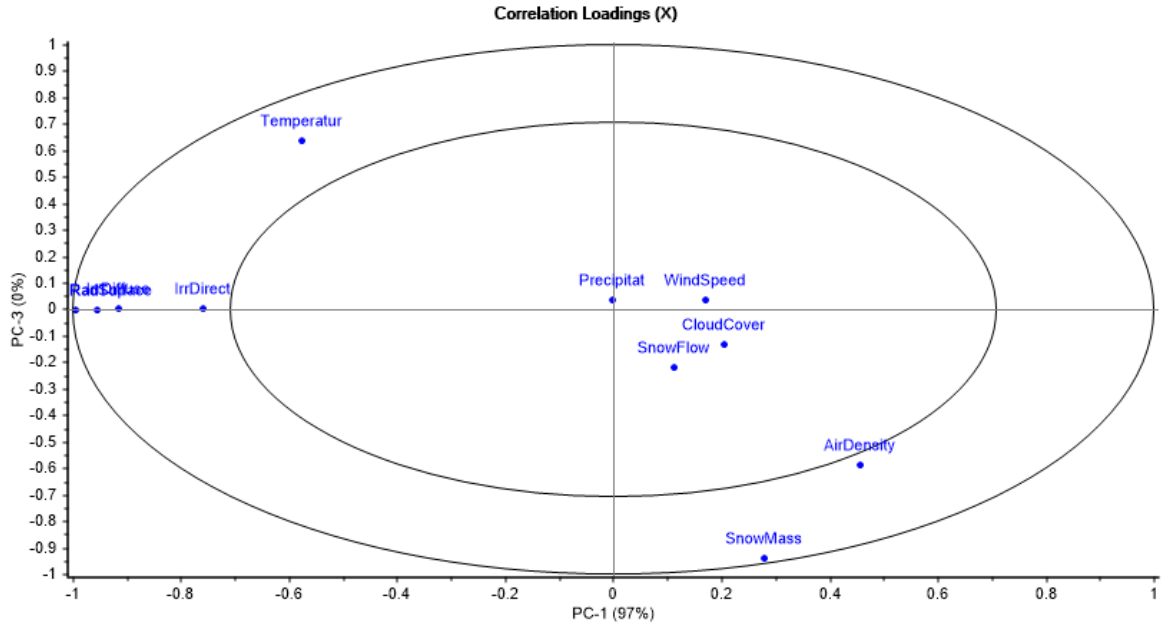
Almost all the variance is explained by the two first PCs. This can indicate that the rest of the PCs can be disregarded in a reduced model. The optimal number of PCs are therefore 2.



From the scores plot we can see that the summer months and early fall are located to the left and the winter months and late fall are to the right. So PC-1 determines the time of year, and tells us that the time of year is one important factor in the data.

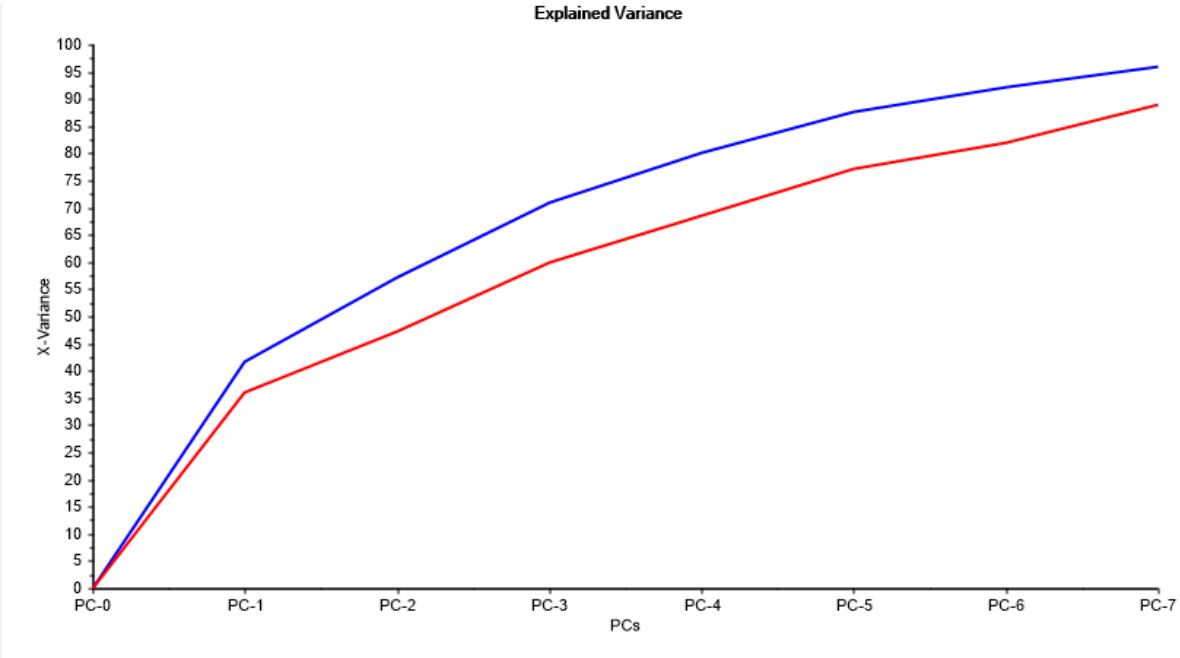


From the correlation loadings plot we see that for PC-1 and PC-2, the relevant variables are primarily RadTop, RadSurface, IrrDirect and IrrDiffuse. Temperature is also close to the 50% circle and could also be relevant for PC-1, especially considering the analysis done above. They all correlate negatively with PC-1.

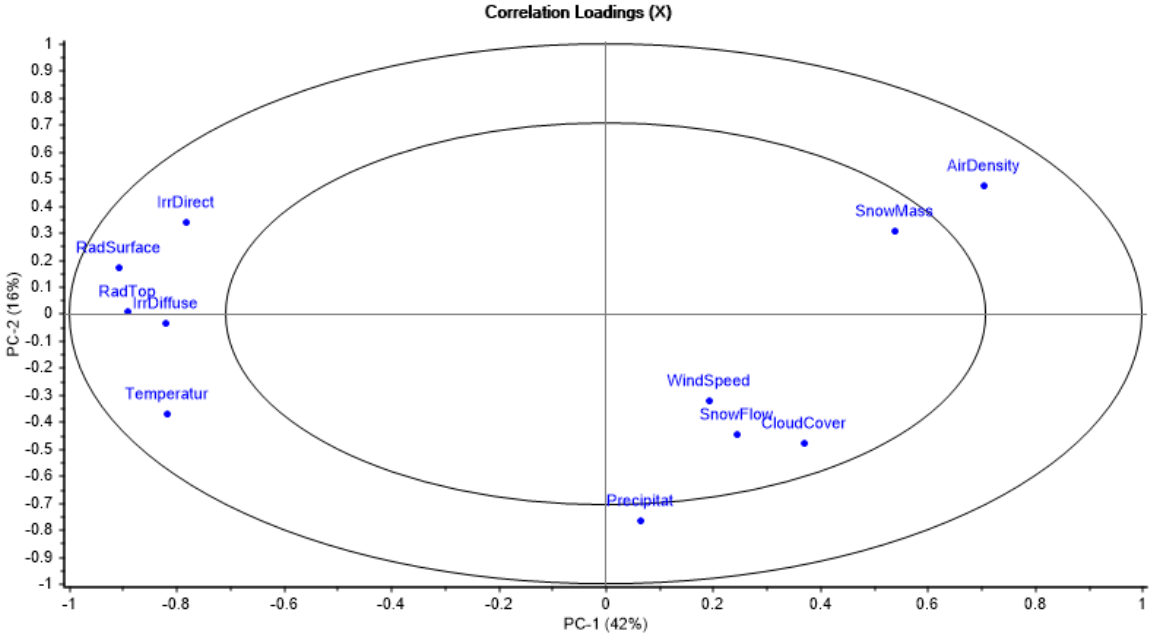


Plotting PC-1 vs PC-3 we can clearly see that Temperature, AirDensity and SnowMass are important in PC-3, and are oppositely correlated. This makes sense as we don't usually see snow and warm weather at the same time.

Now, with scaling to unit variance.

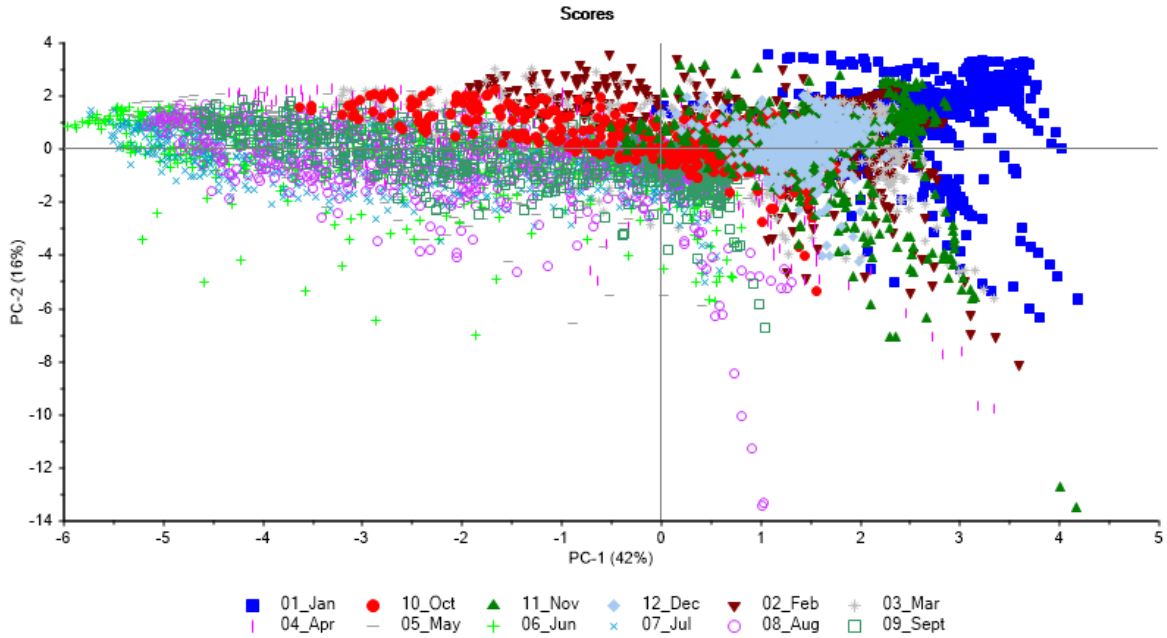


Now it is not so clear how many PCs to choose since each PC explain more now than before. We see that to explain all the variance in the data we likely need all/almost all the PCs.



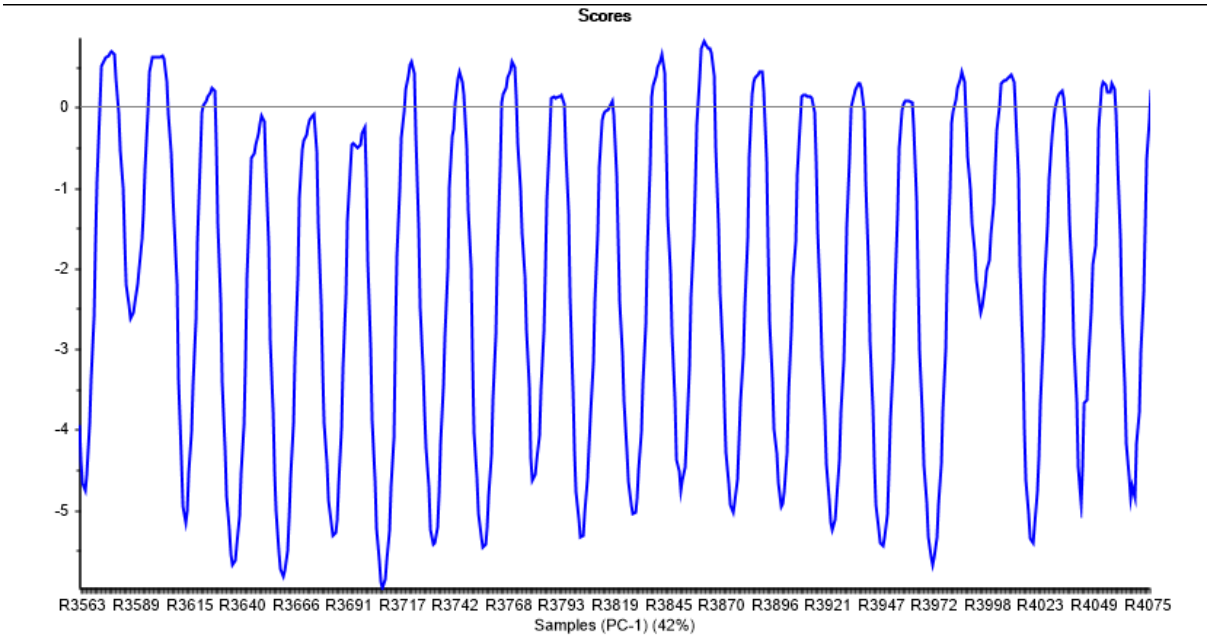
For PC-1 we still see the same variables as above are important, but temperature and AirDensity has also become more important now in PC-1. We also see PC-1, and to some extent PC-2, capturing some of what PC-3 did earlier. Likewise, we see air density and snowmass is inversely correlated to precipitation which I find a little weird (especially since high precipitation means wet weather which should mean high air density), but this might be lack of knowledge on my part.

In PC-3 there is not too much structure, but in PC-4 we see some inverse relationship between cloud cover and wind speed which makes sense (high wind fewer clouds). It can therefore be relevant to include 4 PCs in this model.



We also do not see the same division of summer and winter months as above in the scores plot.

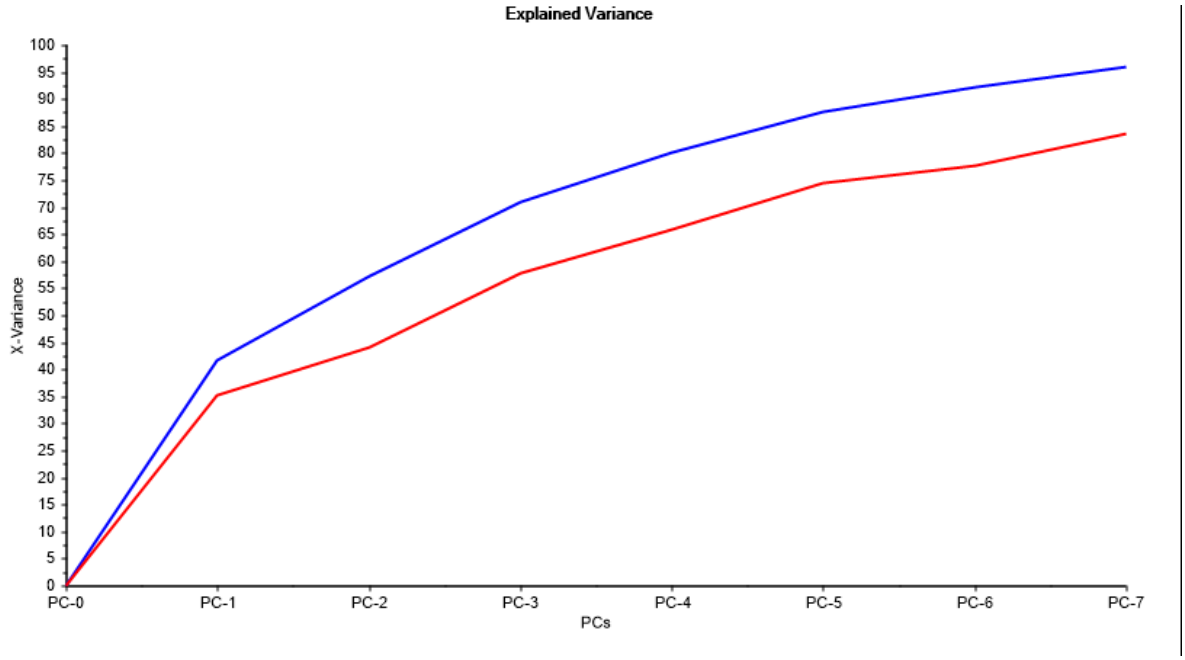
Plotting the scores for PC-1 as a line plot and zooming in gives this



We see clear systematic pattern in PC-1. It is likely the temperature change between night and day since we saw the temperature and the variables related to radiation is high up on the correlation in PC-1 in the correlation loadings.

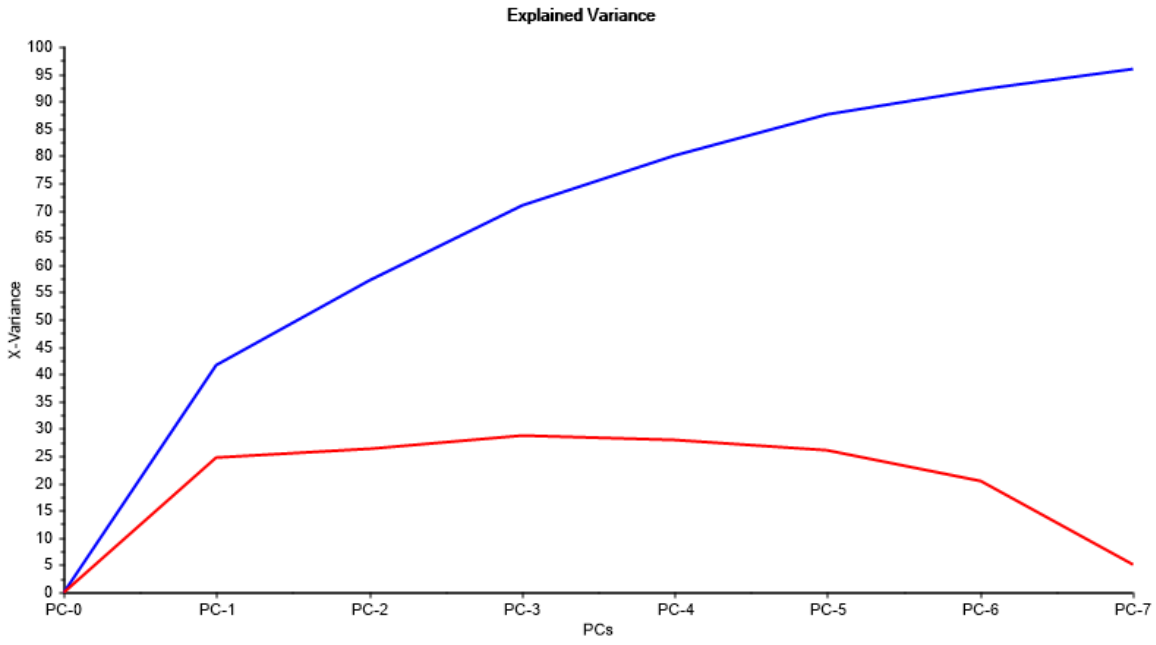
Other cross-validation setups

Using 112233



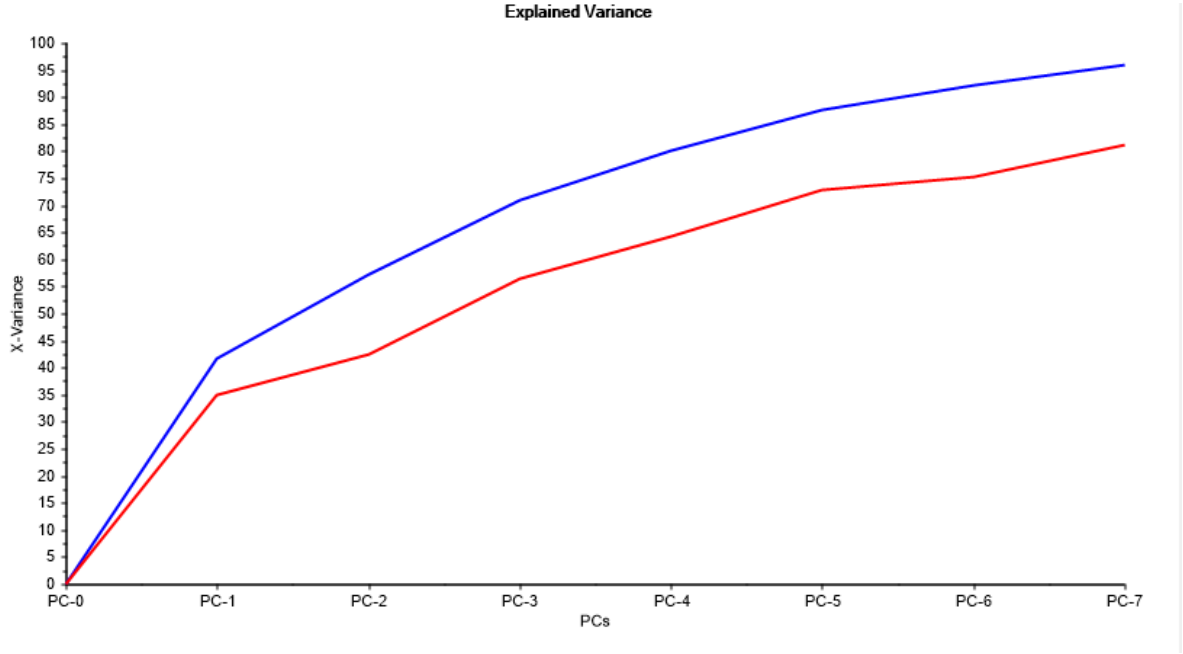
There is little different between this and the previous with random CV.

Using category: IsDayBin



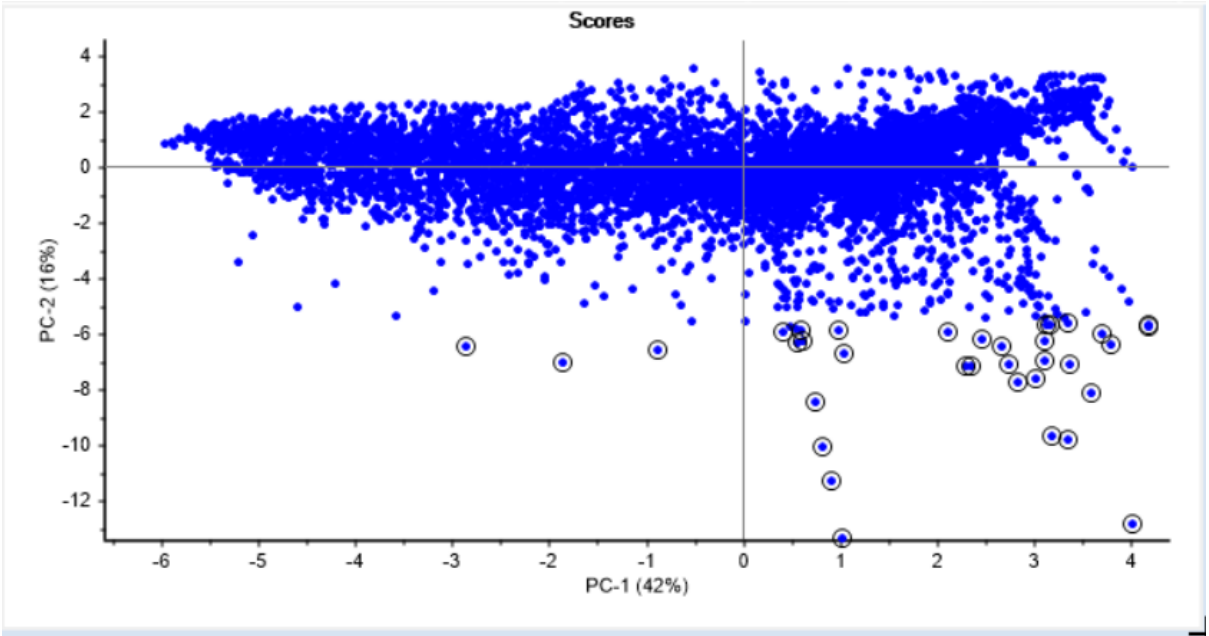
This is terrible compared to the others, and indicate that training on only day or night will seriously degrade the models performance.

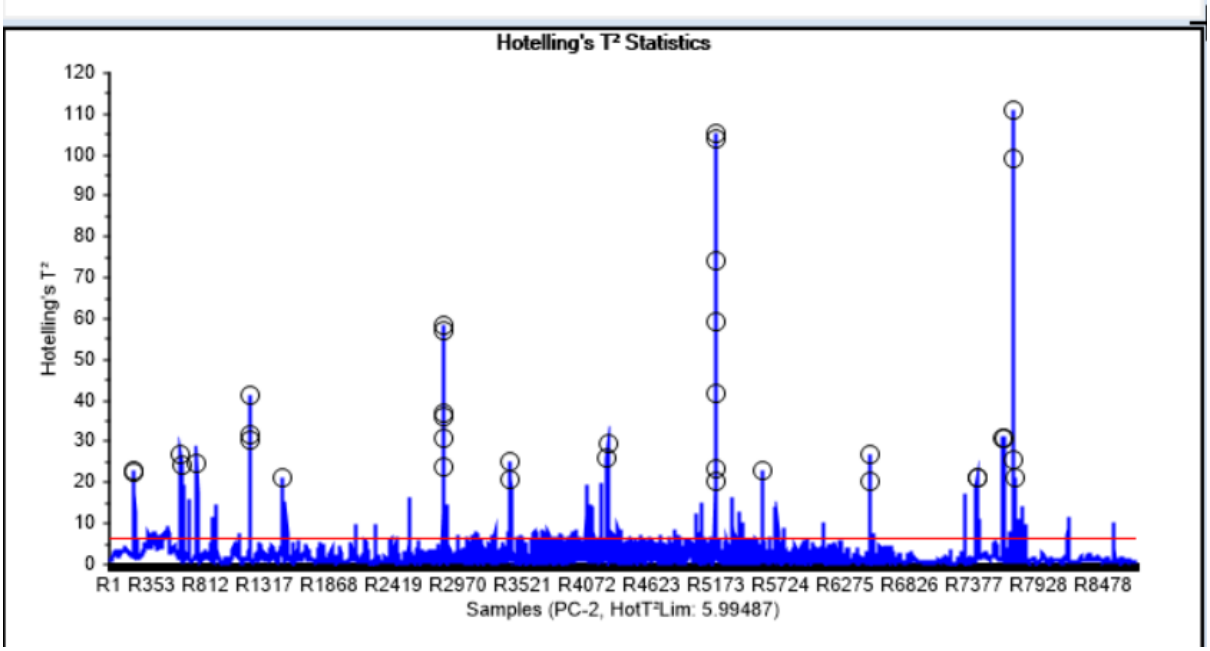
Using category: Month



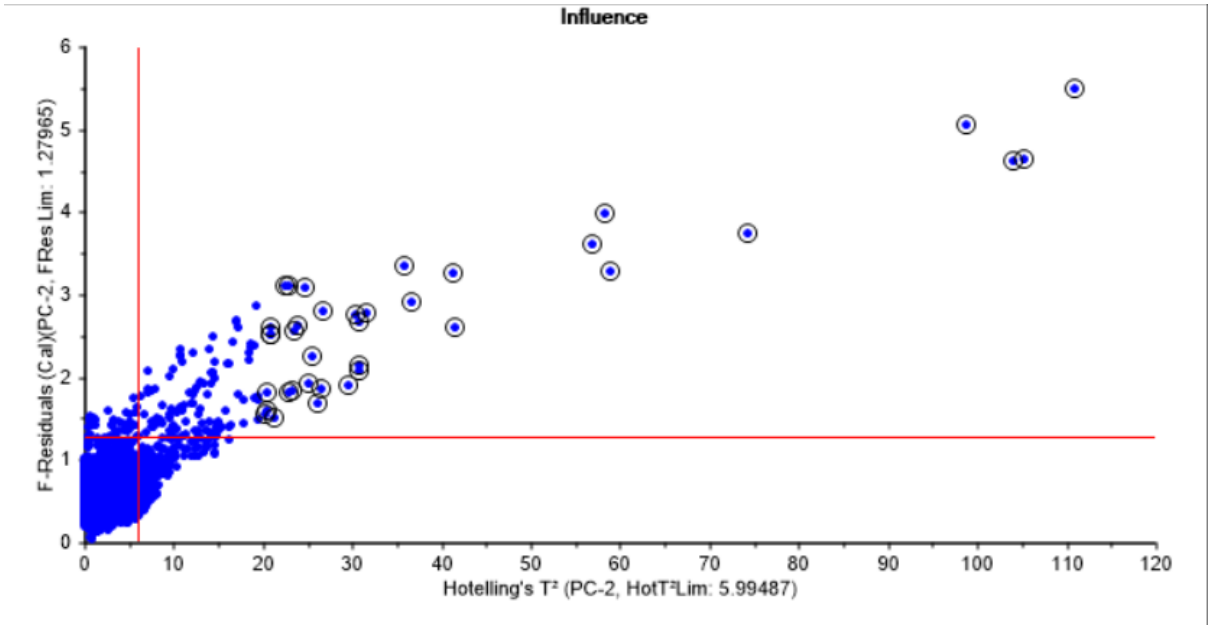
This performs comparably to random and systematic indicating that there is no significant change between months, and making a model for one month and using it for another month will work.

Looking only at the Hotellings T2 for PC-2 we find these as outliers as they exceed the critical value by a lot.

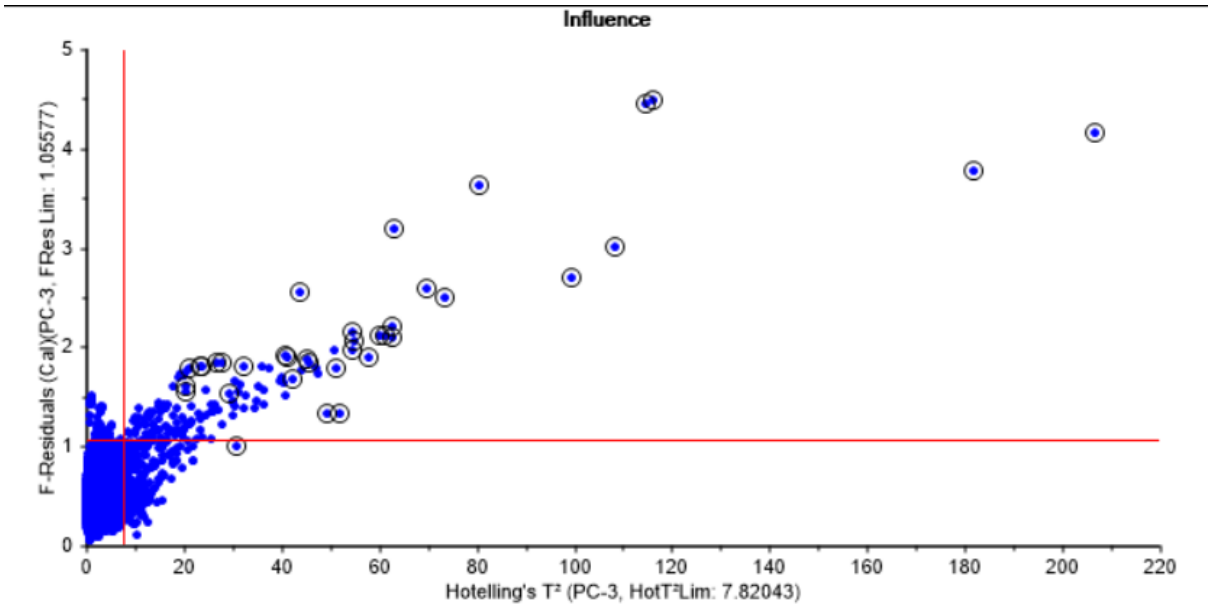




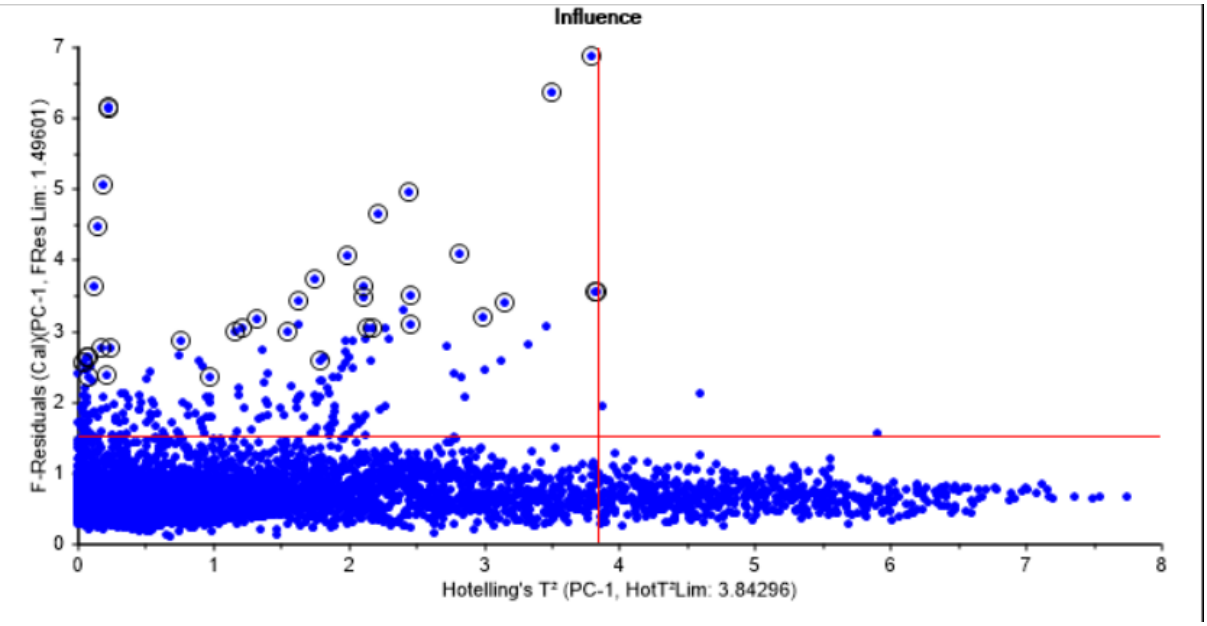
When looking in the influence plot for PC-2 we see the same points being selected, and they are all in the quadrant with large F-residual and large T2, which is where outliers most likely are located.



We also see the same points being selected in the influence plot for PC-3.

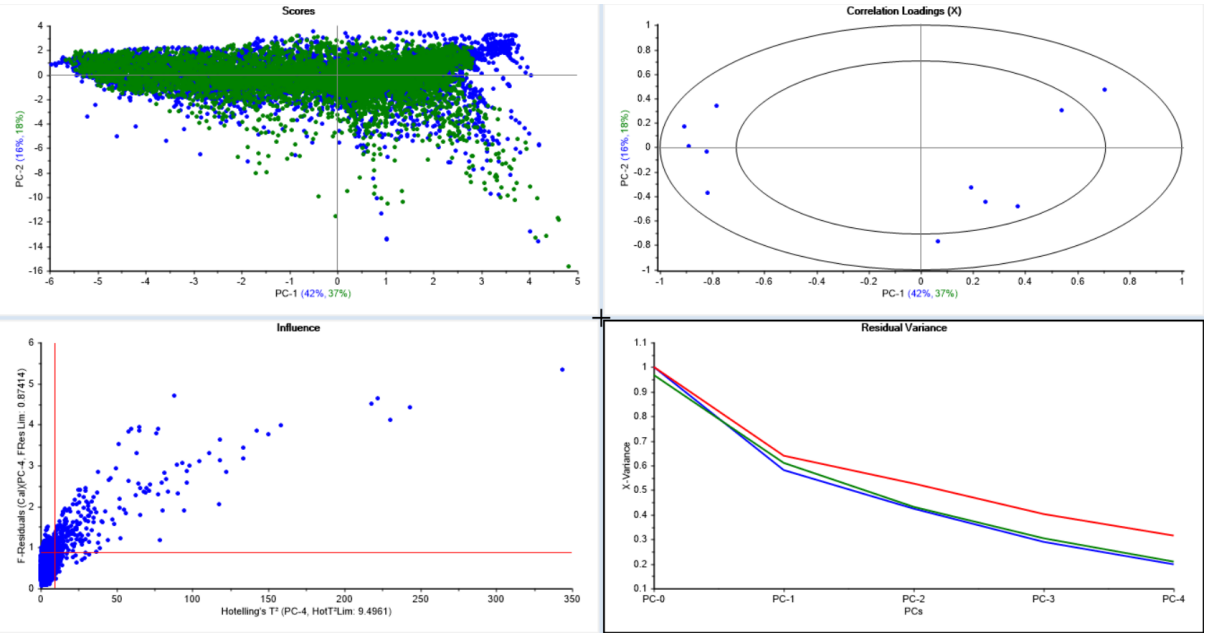


On the other hand, in PC-1, these points have primarily large F-residuals, but they are still probably outliers as we see in the two other influence plots as well as in the scores plot.



Project data onto model

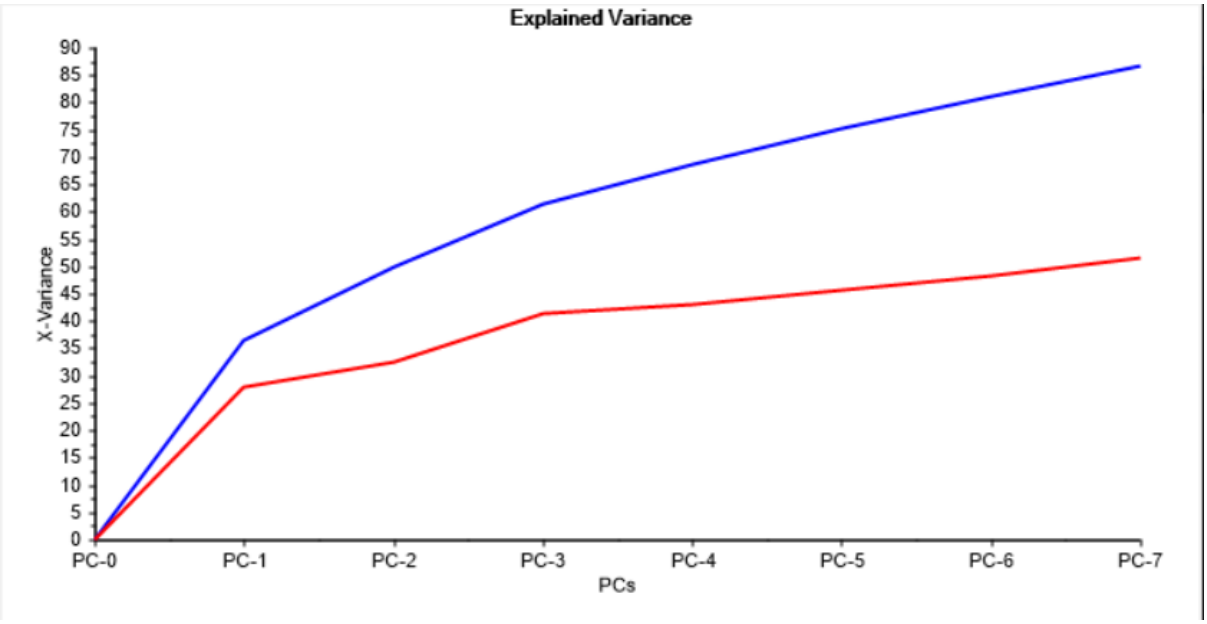
Choose the scale model with random CV and 4 PCs as best model.



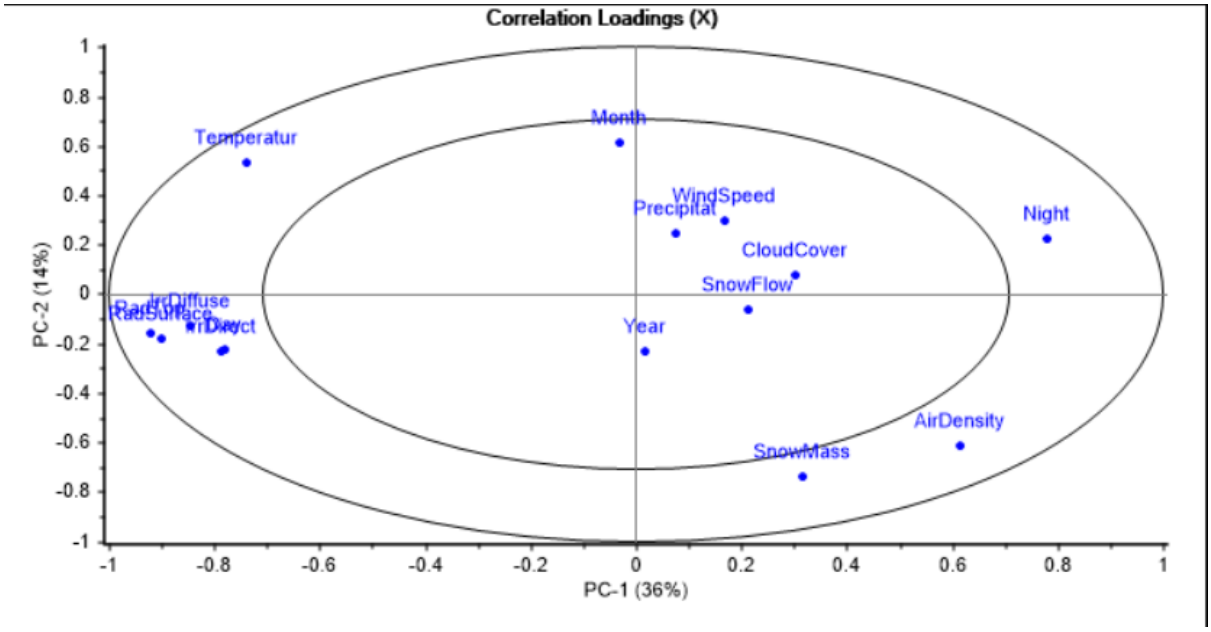
The projection seem to fit the model well with very similar residual variance.

Model with CV over years

Use X variables with the inclusion of day/night/month variables. Also used scaling to unit variance.

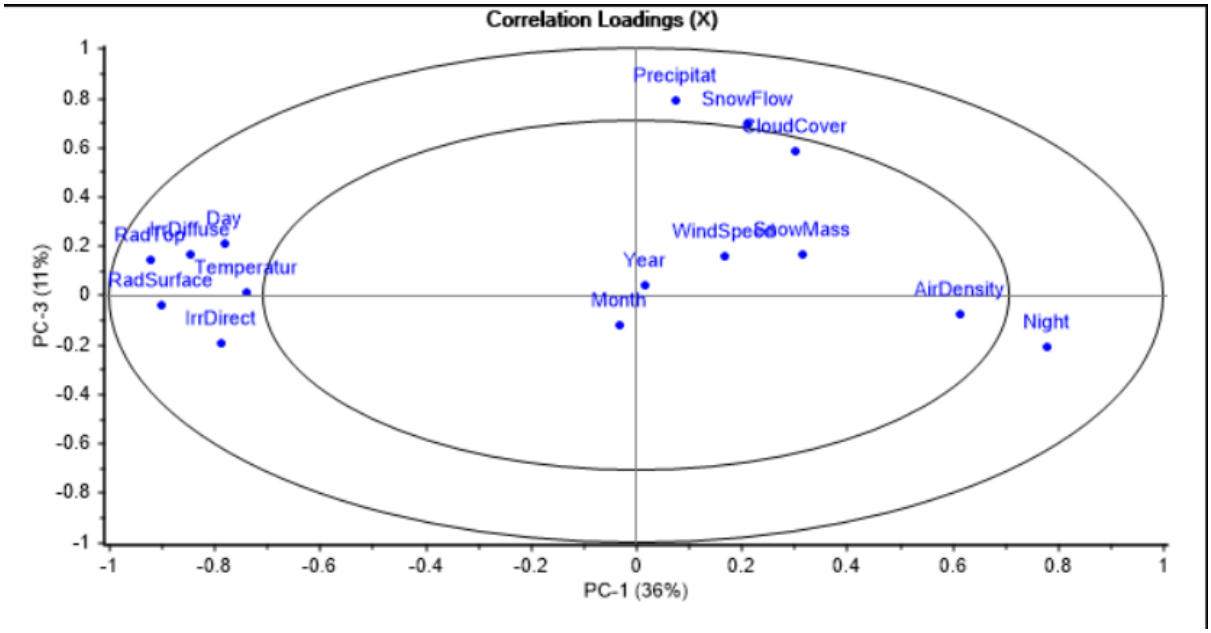


Even though the higher number of PCs explain more of the variance in the training set it seems like it doesn't work too well on the validation, so cross validating across year perhaps does not work as well as random CV, so there maybe be some differences between each year in the model.



PC-1 is now even more clearly the temperature and radiation change from day to night, as we can see where night is directly opposite temperature and the radiation variables.

PC-2 is also the same as discussed above.



PC-3 seem to capture the relationship between cloud cover and precipitation, which makes sense.