Supplementary Material for *"On the Predictability of Land Surface Fluxes from Meteorological Variables"*

N. Haughton G. Abramowitz A. J. Pitman

June 29, 2017

Uncertainty in Cluster-plus-regression models

The Cluster-plus-regression models used in this paper include some uncertainty in their results. This stems from the fact that there is no "correct" initialisation for K-means clustering (we use the common "random selection of samples" initialisation method), and that there is no guarantee of convergence in the clustering algorithm. This means that each time the model is trained, it is slightly different - cluster centres are different, and so different samples are included in the linear regressions for each node, and so the regressions are also all different. This problem is unavoidable with this kind of model (indeed, with nearly all clustering algorithms). Fixed initialisations can be used to ensure identical results on identical training data, but this is only hiding a real aspect of uncertainty.

To check whether this uncertainty was a problem for this study, we replicated two of the models, STH_km27 (3km27 in Best et al. 2015, which our testing indicates suffers from the same problem), and STH_km729, 9 times each over all of the data. For each of the 61 sites each model was independently trained 9 times on 60 other sites. For each of the two models, we also created an ensemble mean of the 9 simulations (STH_km27.mean and STH_km729.mean), and compared these with the other 9 simulations. We then ranked each simulation against S_lin, ST_lin, STH_lin, and STH_km243 as benchmarks (each only trained and run once at each site), and compared how the ranks changed for each site, metric, and variable.

The number of times the ranks changed between the model variants for each combination of variable, metric, and site ($3*10*61 = 1830$) is shown in Figures 1 and 2. The total percentage of non-modal ranks over all combinations, and over the 10 model variants is 6.3% for STH_km27, and 6.7% for STH_km729. The distribution of these changed ranks over the 10 different metrics is shown for the two models respectively in Figure 3 and 4.

We show how this affects the actual PLUMBER-style plots in Figures 5 and 6. While there is clearly some variation between the model, there are not many cases where average ranks change order between the model in question and the benchmark models. Also, this variance is likely to be substantially reduced in most cases where ranks are also averaged over multiple metrics. We would suggest that the variance shown in these figures provides a rough guide for how much uncertainty to expect in the values shown in other PLUMBER-style figures.

Finally, we show an example of how these simulations vary as a time-series, in Figures 7 and 8. This section of time series is arbitrarily chosen from the Tumbarumba site dataset, and shows the periods centred on the 12-hour period with the highest average inter-model variance for each variable. While the noise in this figure appears somewhat worrying, it should be noted that it is the worst case in this time-series. A similar figure for the PLUMBER LSMs is shown in Figure 9, where it is clear that LSM variance is substantially higher, especially for Qle.

It should also be noted that this variance is purely due to uncertainty in the model itself. In each of these simulations, for each site, the model is trained on identical data. If the model were trained on different data each time, it is possible that there would be some additional variance in the simulations.

Lagged Autocorrelation in driving variables

Figures 10 to 14 show the correlations between lagged averages of each met variables, for different length lags. For each plot, the driving variable is averaged over a moving window of each size, for each site, then site data is concatenated, and correlations are calculated over the entire dataset for each lag. High correlations indicate that

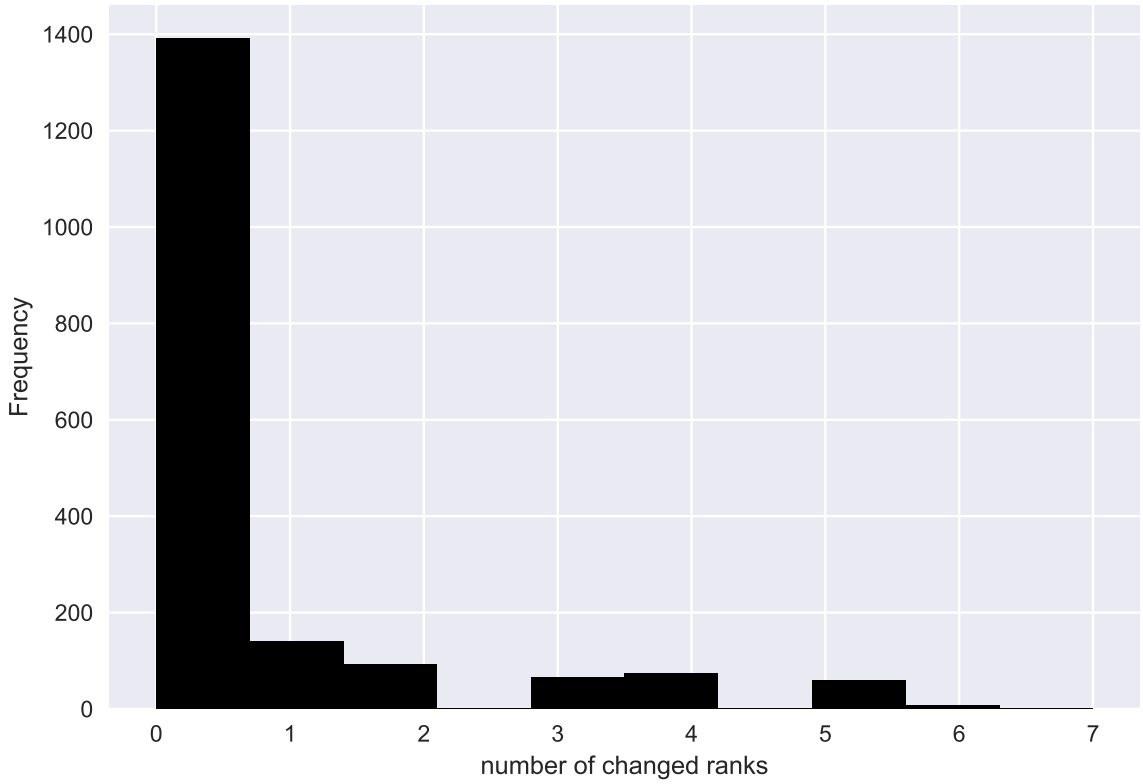

Figure 1: For these figures, rank changes are calculated by taking the mode of the ranks for the 10 model variants, and counting how many of the model's ranks are different to the mode. If all 10 model variants have the same rank, for a particular site, variable, and metric combination, then that combination contributes to the 0 column in the histogram. The theoretical maximum in this case is 8 (if there are 2 models for each of the 5 ranks, then the mode will be arbitrarily chosen, and all other ranks will count toward the "changed ranks".

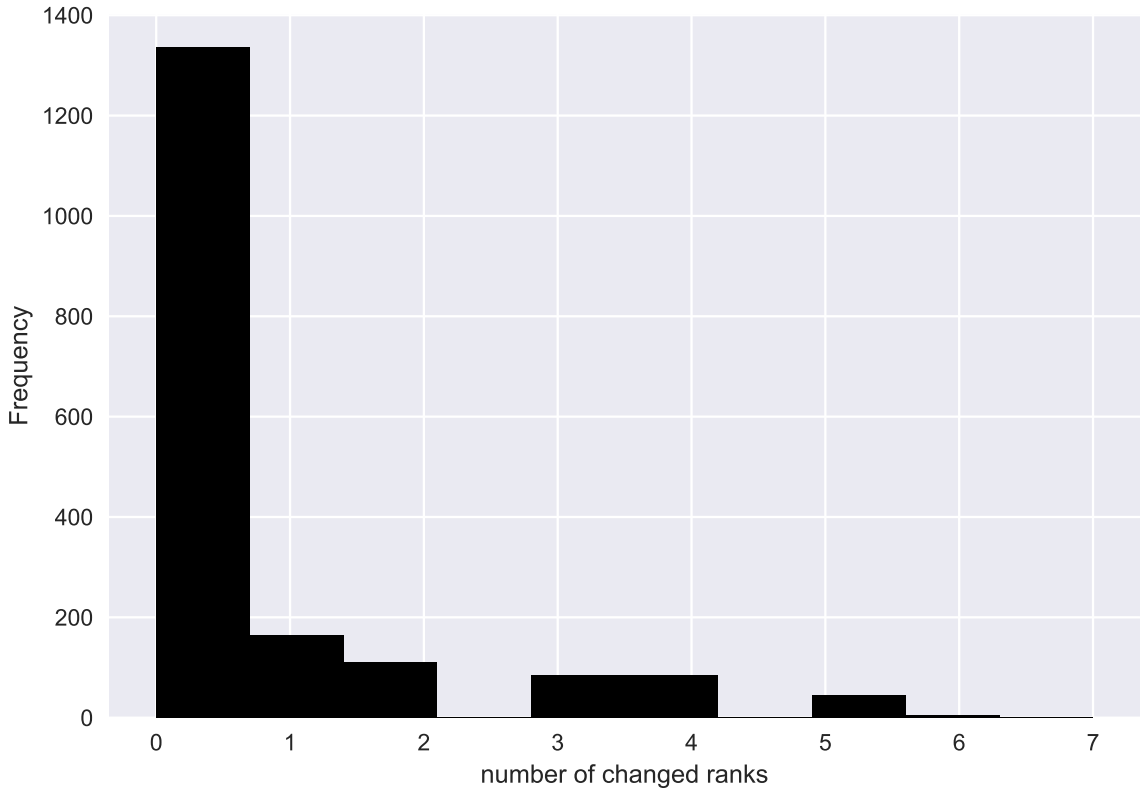

Figure 2: As per Figure 2, for STH_km729.

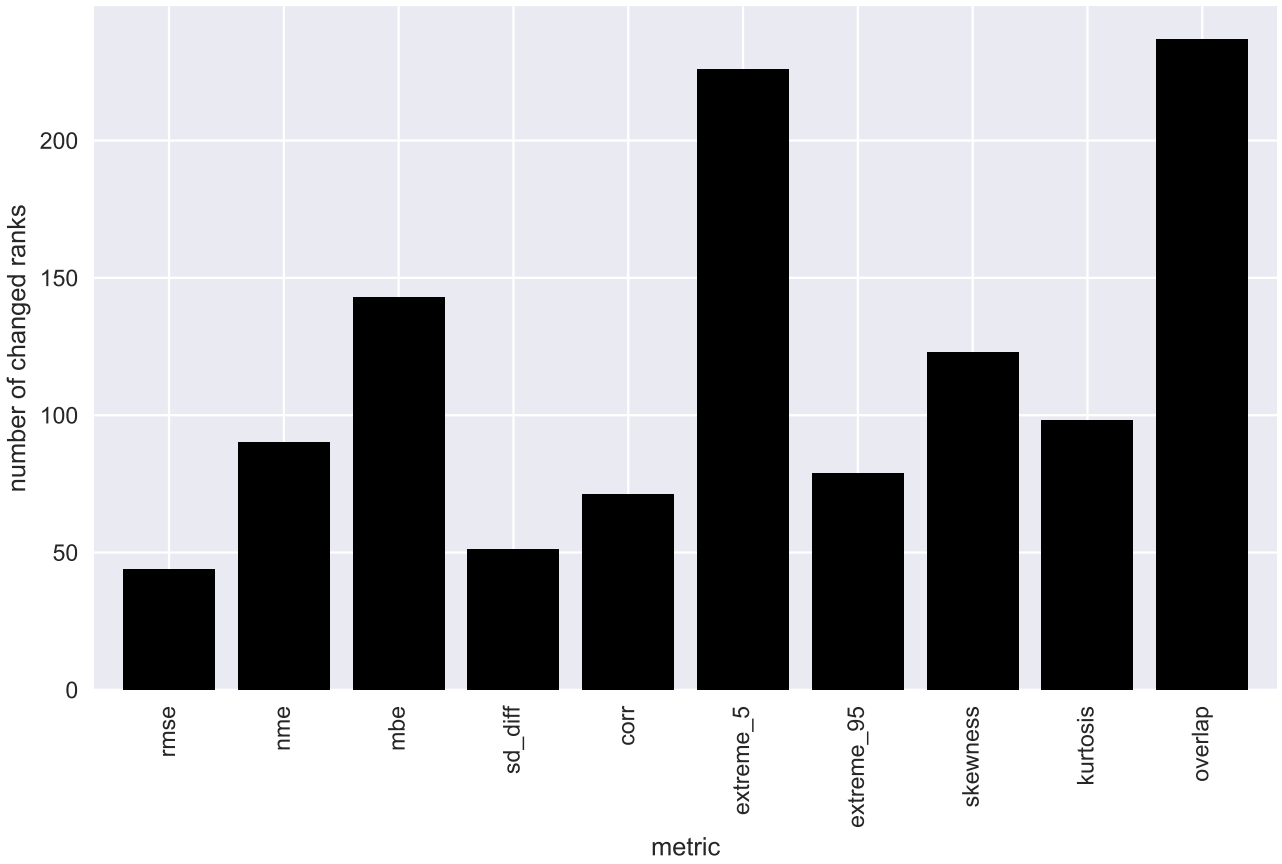

Figure 3: Number of non-modal metrics for each of the different metrics (the total theoretically possible for each bar is 3*8*61 = 1464).

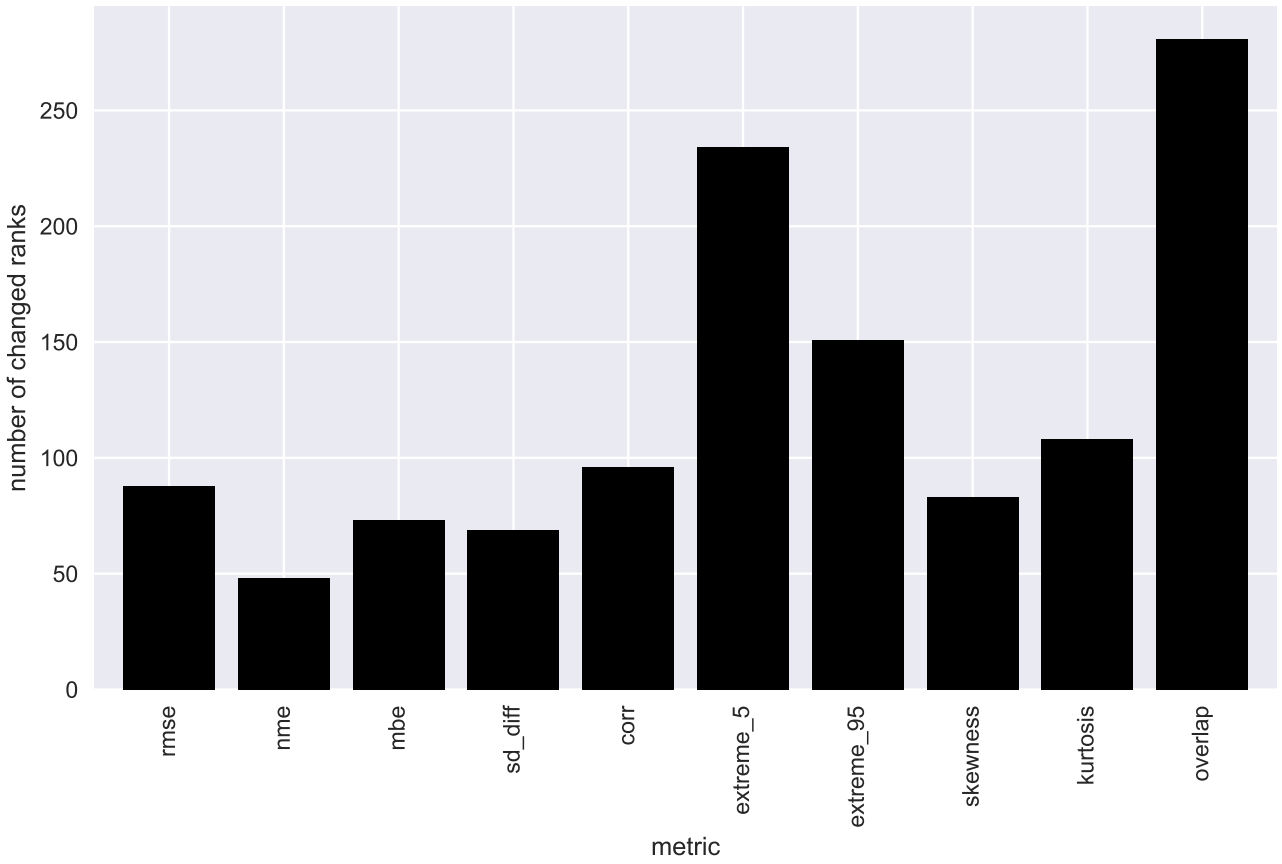

Figure 4: As per Figure 2, for STH_km729.

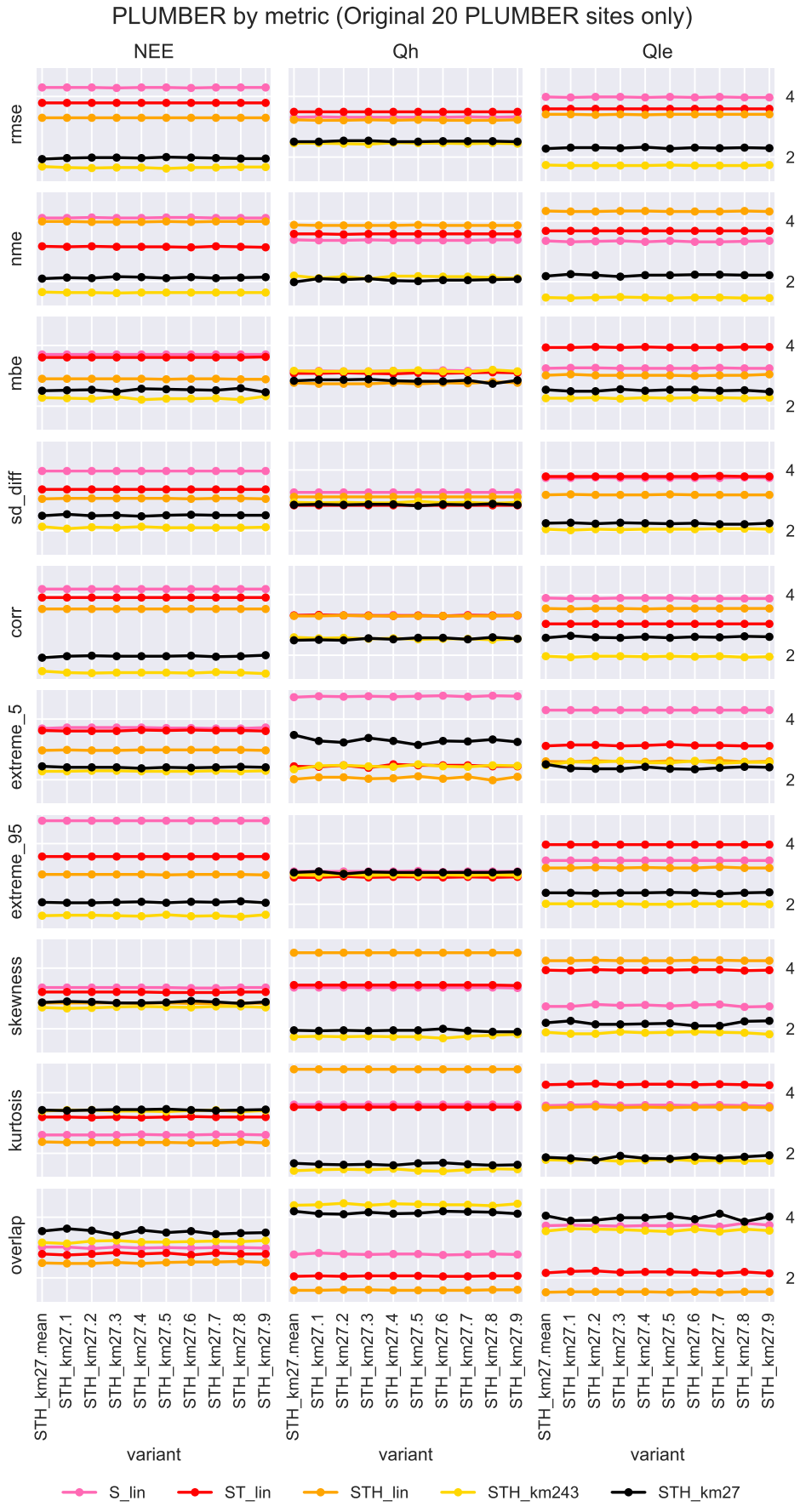

Figure 5: Rank averages over all sites, for each variable and metric, of each of the STH_km27 simulations, plus the mean of all simulations.

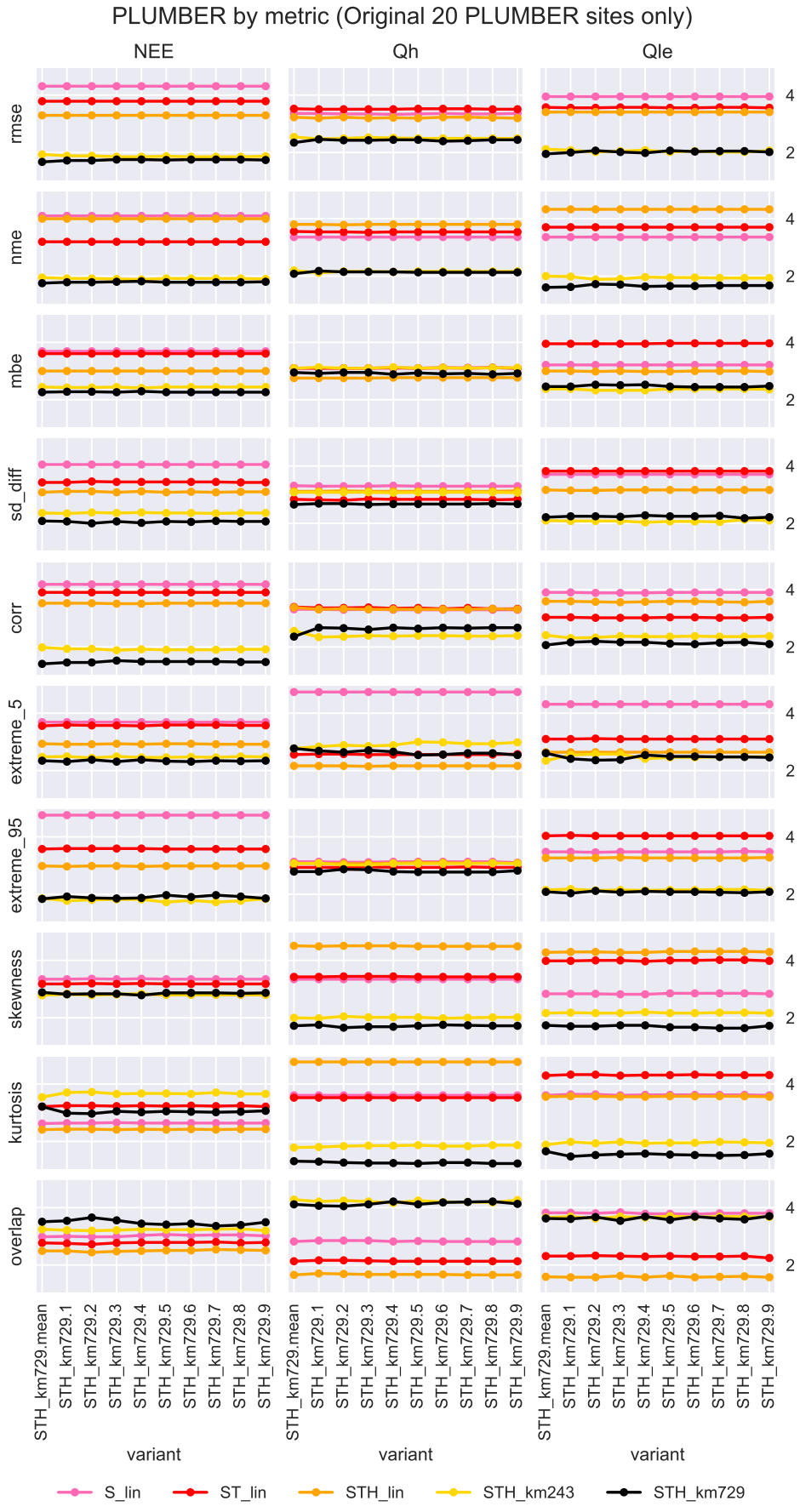

Figure 6: As per Figure 5, for STH_729

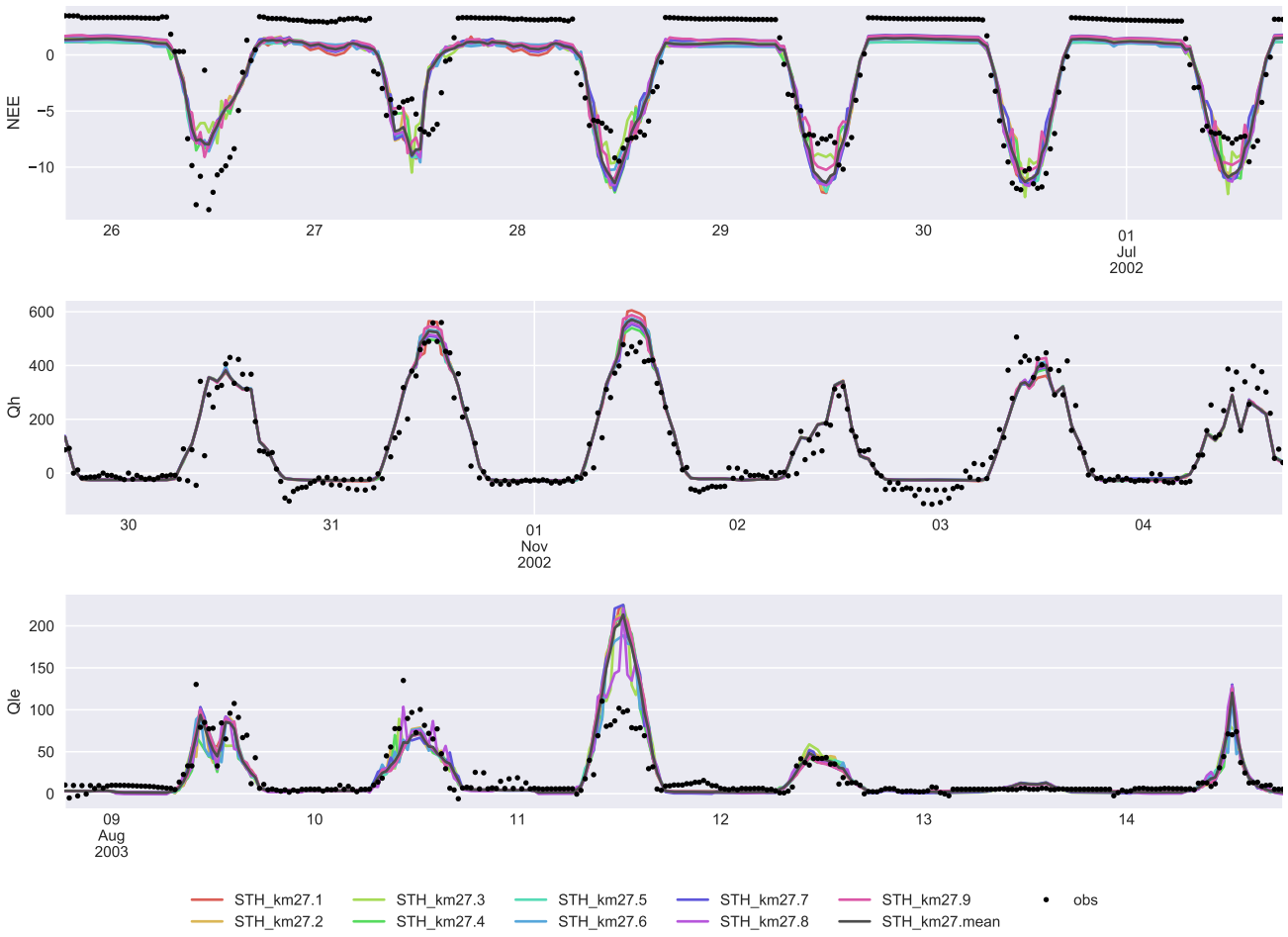

Figure 7: Example time series at Tumbarumba for the STH_km27 model variants and mean, showing variance between individual models. The x-axis is different for each variable, and is centred on the period of highest (worst) inter-model variance for each variable.

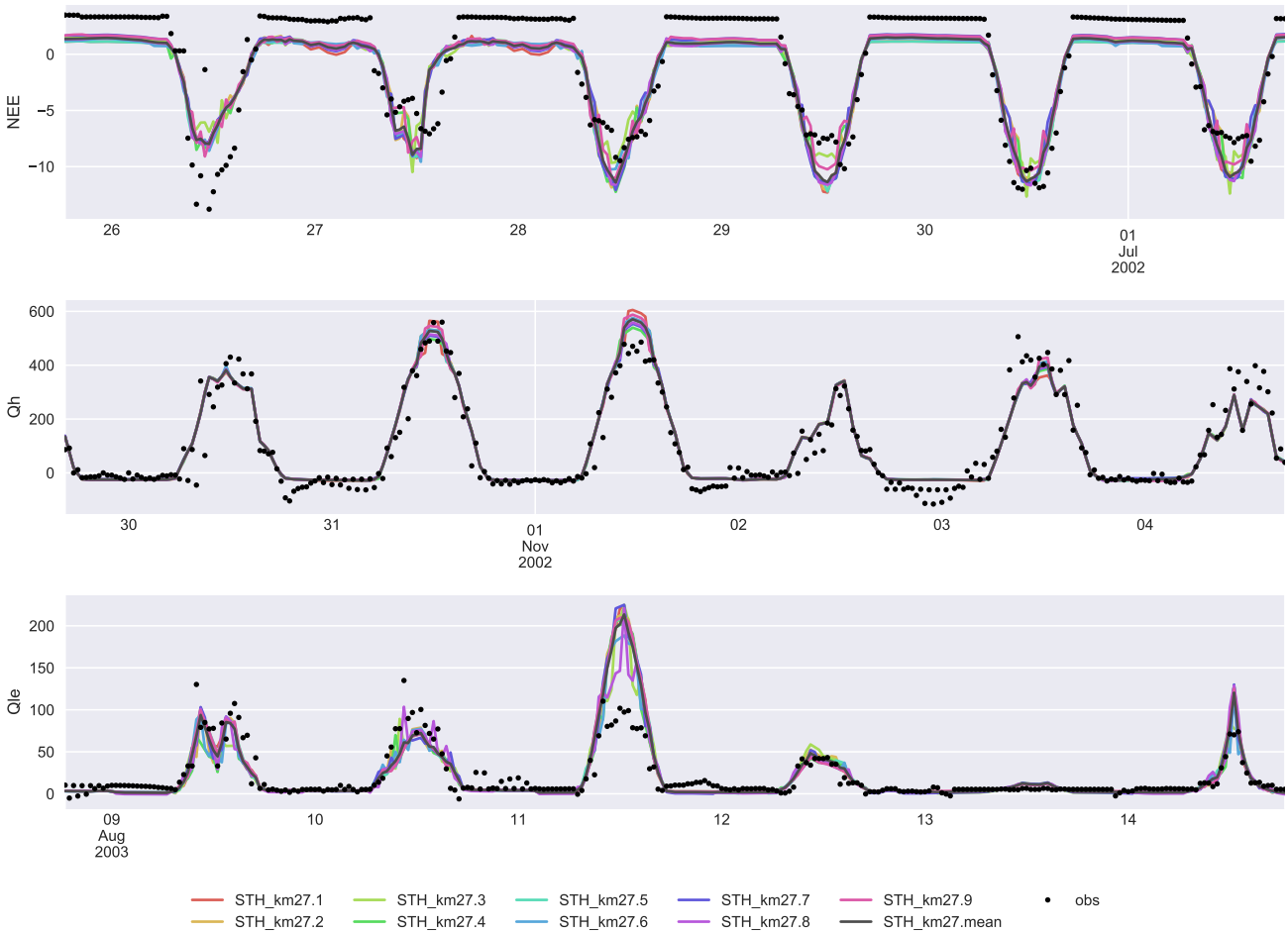

Figure 8: As for Figure 7, but for STH_km729.

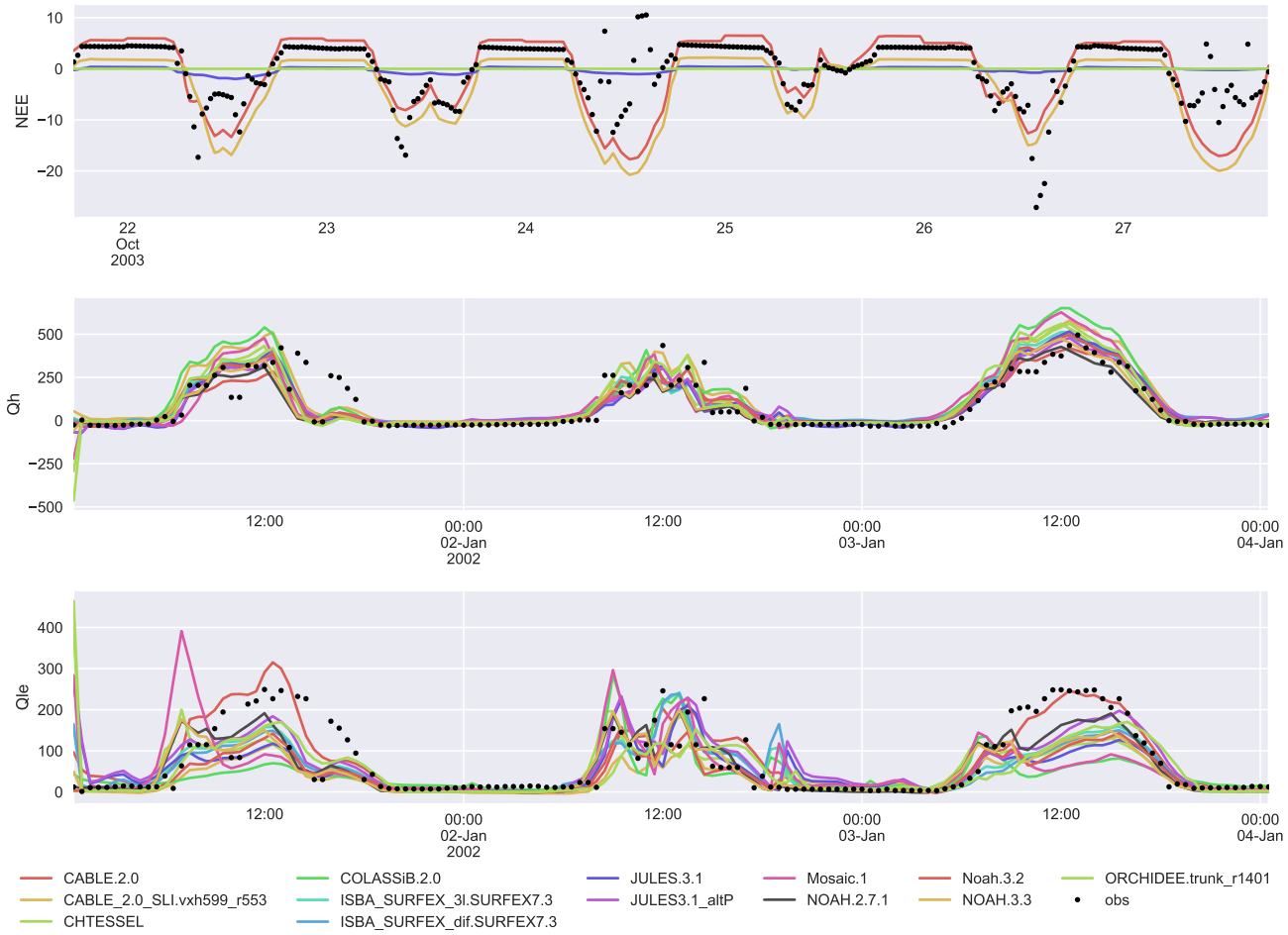

Figure 9: As for Figure 7, but for the 13 PLUMBER LSMs.

averages contain similar information, and are thus likely to not provide significant additional information to models, and might potentially result in multicollinearity if both variables are used in the same model. In some cases, this problem can be avoided by data transformations, such as principal component analysis, or by decomposing one of the variables into separate components, and using one of those components.

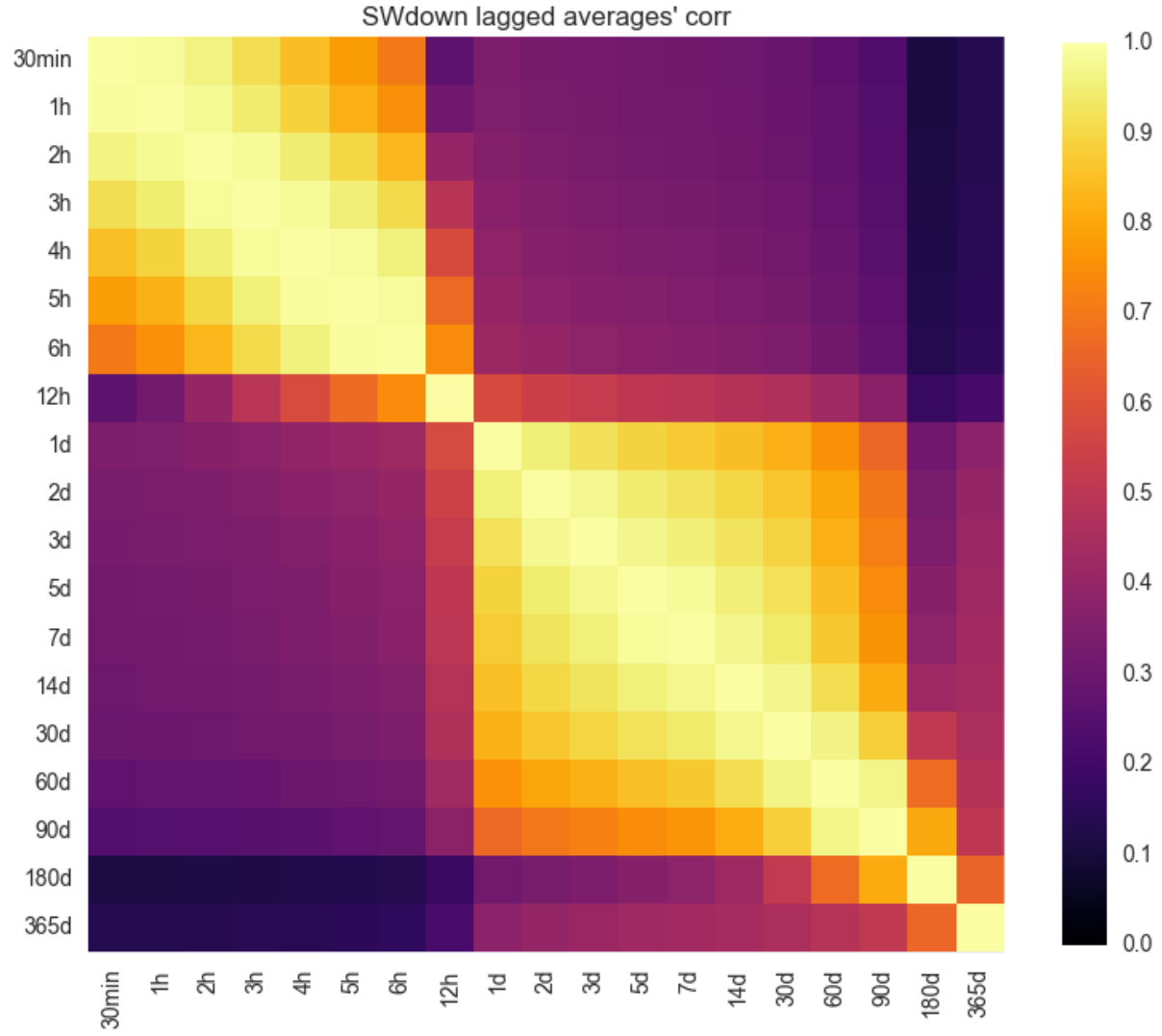

Figure 10: Correlations between moving window averages of SWdown of different lengths. Values are averaged over 61 sites.

Best, Martin J., Gab Abramowitz, Helen R. Johnson, Andy J. Pitman, Gianpaolo Balsamo, Aaron Boone, Matthias Cuntz, et al. 2015. "The Plumbing of Land Surface Models: Benchmarking Model Performance." *J. Hydrometeor* 16 (3): 1425–42. doi:10.1175/JHM-D-14-0158.1.

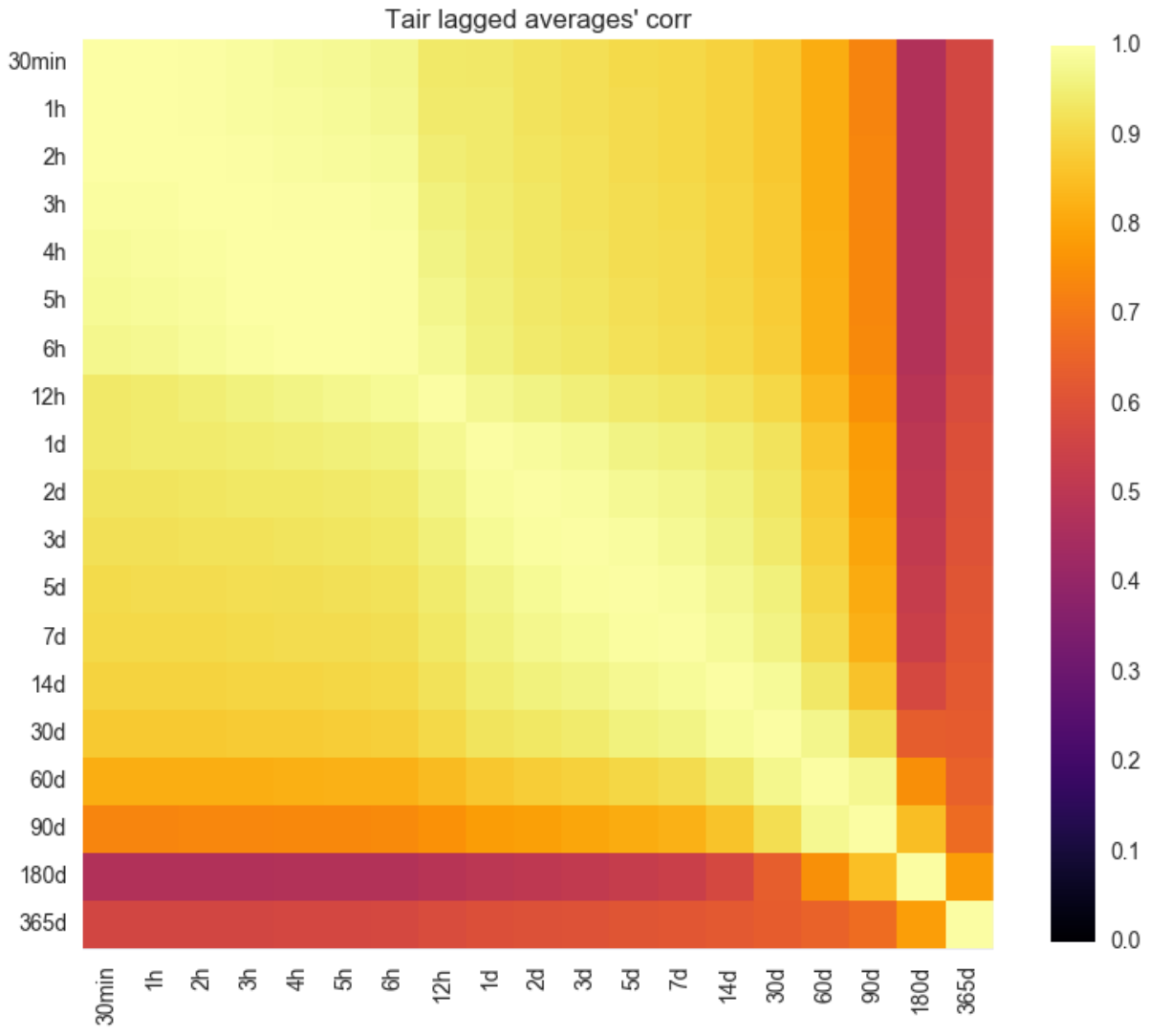

Figure 11: As per Figure 10, but for Tair.

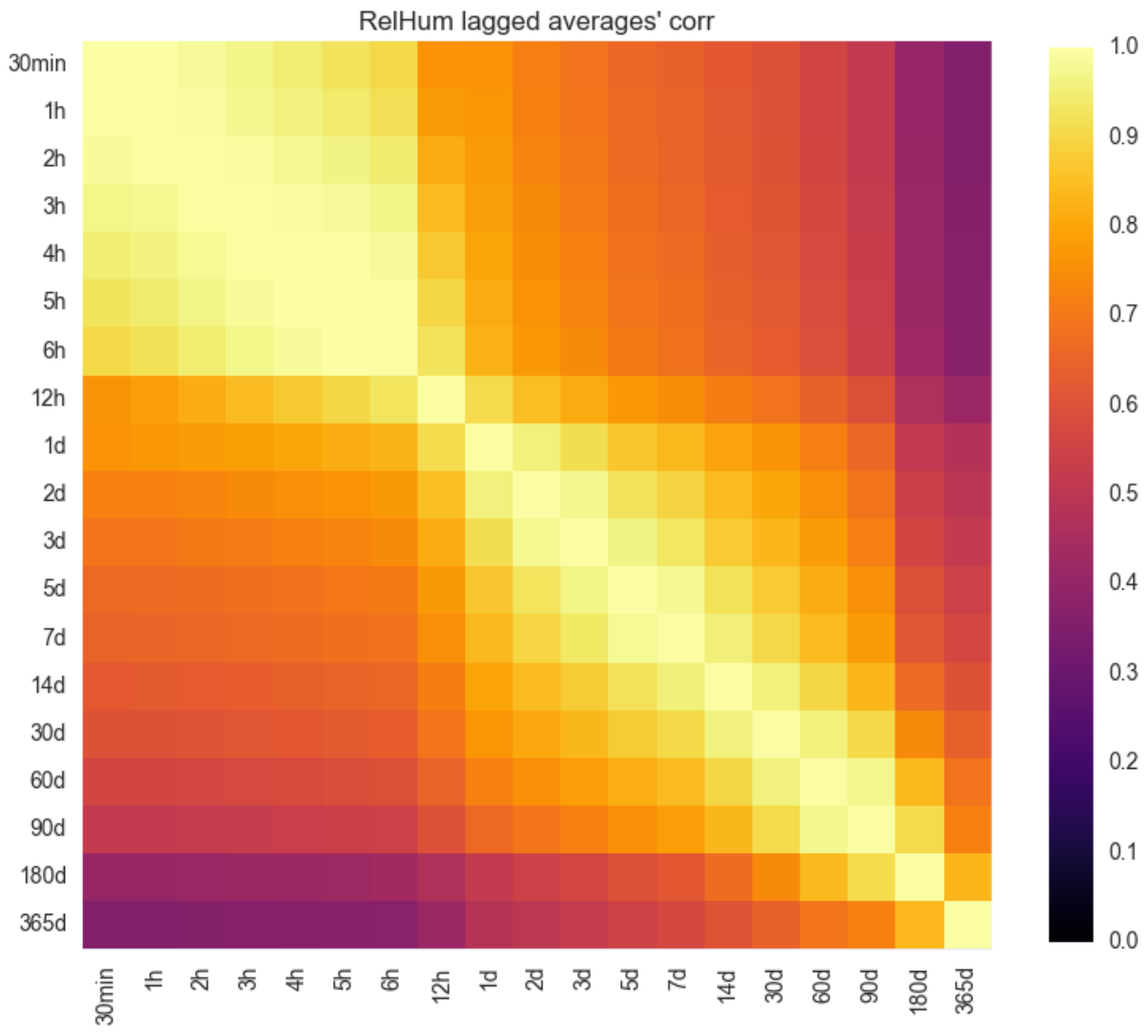

Figure 12: As per Figure 10, but for RelHum.

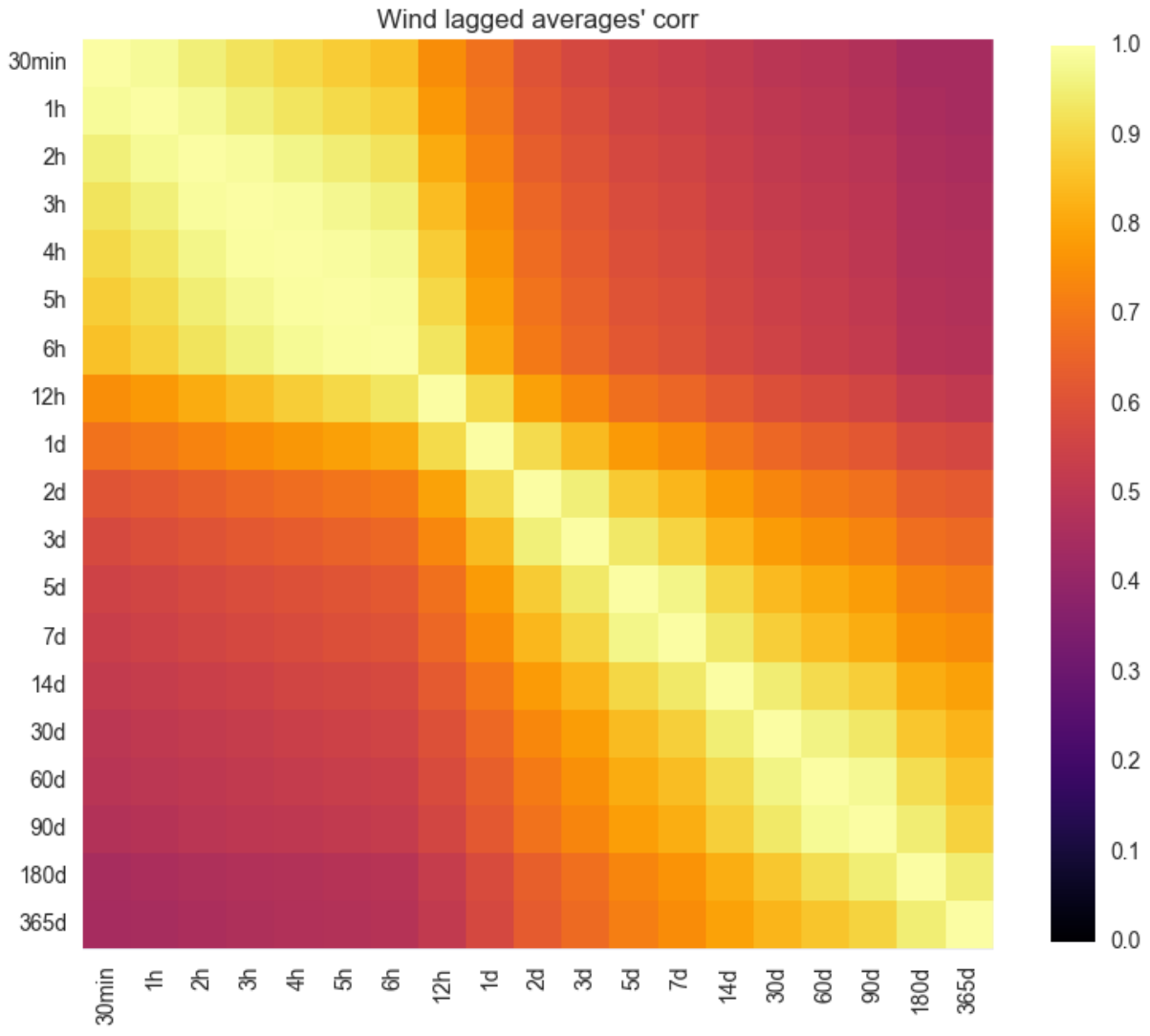

Figure 13: As per Figure 10, but for Wind.

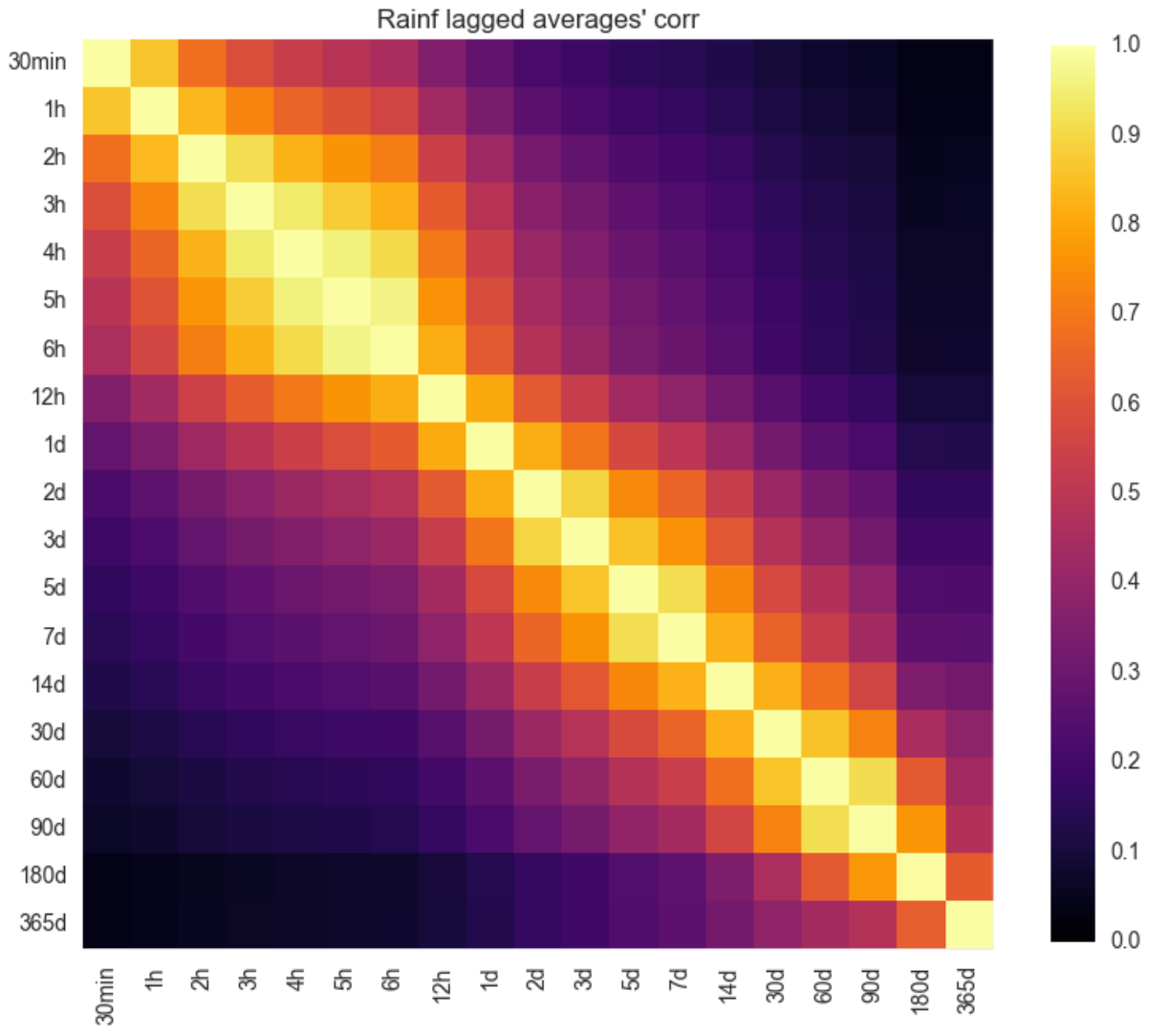

Figure 14: As per Figure 10, but for Rainf.