Excercise 2 - Grey-box models and model selection

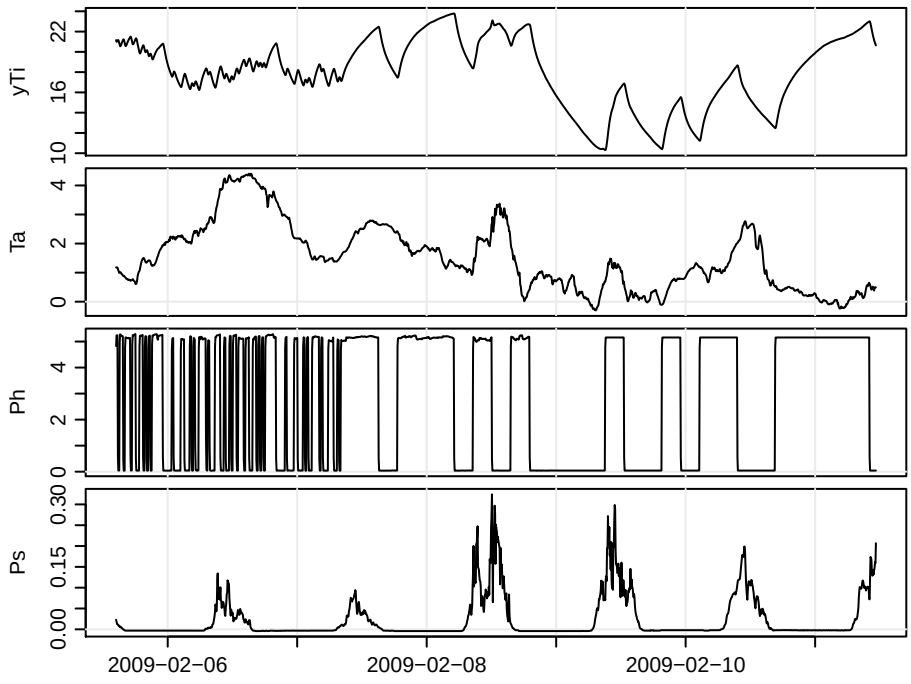
Marco Hernandez Velasco

August 2018

Q1 - Fit and Validate

The exercise is focused on grey-box modelling of the heat dynamics of a building using stochastic differential equations (SDEs). The properties of the PRBS signal and the use of likelihood ratio tests for model selection are considered.

- The data consists of averaged values over five-minute intervals of:
  - Ti (yTi in data) the average of all the indoor temperatures measured (one in each room in the building). The sensors were hanging approximately in the center of each room.
  - Phi-h (Ph in data) the total heat output for all electrical heaters in the building (kW).
  - Ta (Ta in data) the ambient temperature. (notice, that in other material Te is used as the ambient (external) temperature. In this exercise Te is used as envelope temperature).
  - G (Ps in the data) the global radiation (kW/m2).
  - Ws (Ws in data) the wind speed (m/s)



Generate a new object of class ctsm.

Now we can add a system equation (and thereby also a state variable) with the addSystem() function.

- Note that the deterministic part of the SDE is multiplied with dt.
- Note that the stochastic part is multiplied with system noise process **dw1**.
- Note that the variance of the system noise is **exp(p11)**, where *exp()* is the exponential function and *p11* is the parameter.
- Since the variance is strictly positive, but can be very close to zero, it is a good idea to take *exp()* of the parameter, since then *p11* can go from -Inf to Inf but the exponential goes from 0 to Inf, with good resolution towards 0.

```
model$addSystem(dTi ~ ( 1/(Ci*Ria)*(Ta-Ti) + Aw/Ci*Ps + 1/Ci*Ph )*dt + exp(p11)*dw1)
```

Estimate the parameters in the simple model Ti in Equation (1) and see the estimated values with summary(fit). Is the estimation successful (i.e. does the minimization of the negative loglikelihood converge?

The estimate is not successful since one of the estimates (Ria) is close to the lower boundary of the model. Also the estimate of Aw is close the its lower boundary.

```
## Coefficients:
##        Estimate  Std. Error   t value Pr(>|t|)
## Ti0  21.1590625   0.0759270  278.6765   <2e-16 ***
## Aw    1.0068382   0.0415514   24.2312   <2e-16 ***
## Ci    2.7546216   0.0607736   45.3260   <2e-16 ***
## e11 -24.4112015  37.5143135   -0.6507   0.5153
## p11  -1.3586671   0.0171842  -79.0649   <2e-16 ***
## Ria  10.0039623   0.0056182 1780.6258   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Actually, the initial value and the boundary for one of the parameters are poorly set. You can see if the parameter estimates are close to one of their boundaries from the values of df/dPar which is the partial derivative of the objective function (negative loglikelihood).

```
## Coefficients:
##        Estimate  Std. Error     t value    Pr(>|t|)     dF/dPar dPen/dPar
## Ti0  2.1159e+01  7.5927e-02  2.7868e+02  0.0000e+00 -3.0595e-06    0.0036
## Aw   1.0068e+00  4.1551e-02  2.4231e+01  0.0000e+00  3.4174e-06   -2.1531
## Ci   2.7546e+00  6.0774e-02  4.5326e+01  0.0000e+00 -4.2038e-07    0.0000
## e11 -2.4411e+01  3.7514e+01 -6.5072e-01  5.1532e-01  9.1855e-07    0.0002
## p11 -1.3587e+00  1.7184e-02 -7.9065e+01  0.0000e+00  1.4202e-07    0.0000
## Ria  1.0004e+01  5.6182e-03  1.7806e+03  0.0000e+00  1.4461e-05 -637.1910
##
## Correlation of coefficients:
##     Ti0   Aw    Ci    e11   p11
## Aw  -0.01
## Ci   0.07 -0.01
## e11 -0.01  0.03 -0.01
## p11 -0.07  0.00  0.01  0.00
## Ria -0.01  0.06  0.01  0.02  0.01
```

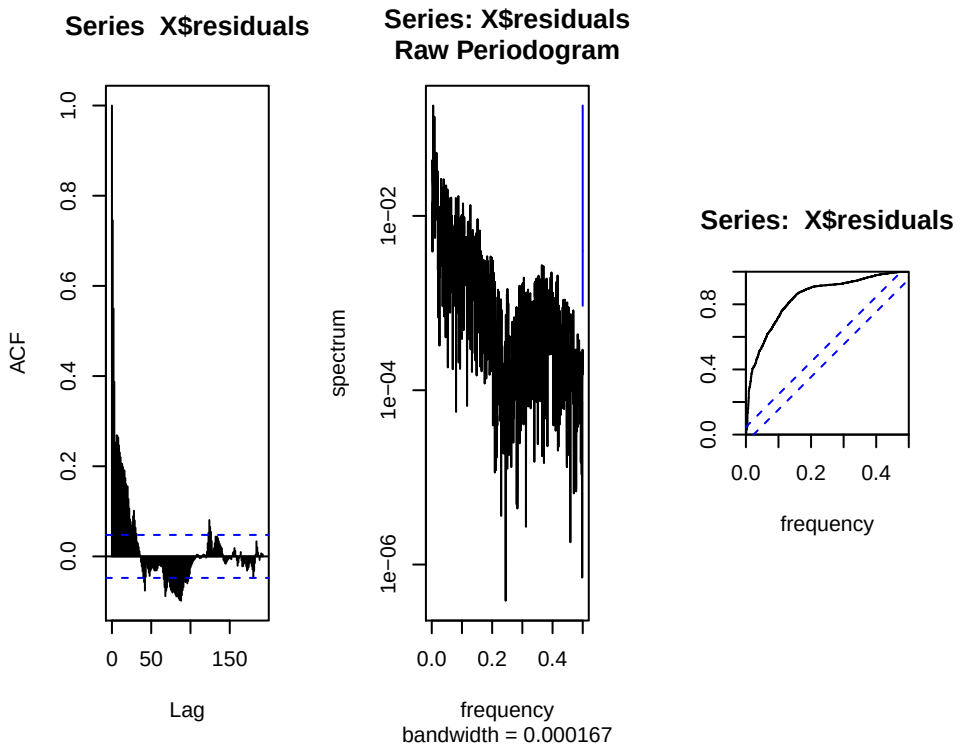
If the value of *dPen/dPar* is significant compared to the value of *dF/dPar* for a particular parameter it indicates that a boundary should be expanded for the parameter.Correct one of the boundaries and re-estimate until the partial derivatives are all very small. Which boundary was not set appropriately?

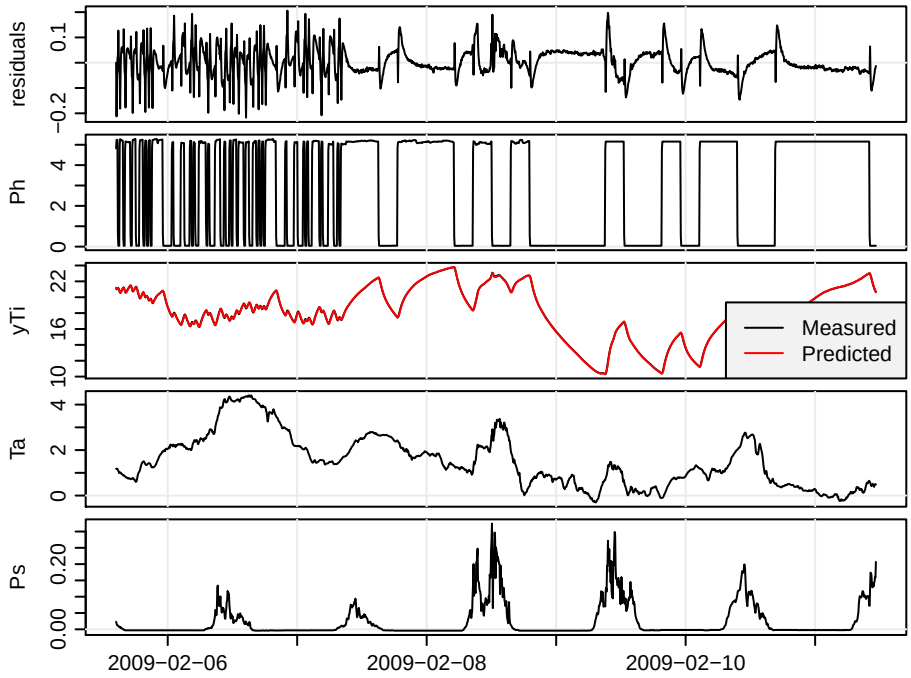
The lower boundary of Ria needs to be corrected since the estimate was close to it. Also for Aw the estimate value is very close to its lower boundary.

```
## Coefficients:
##        Estimate  Std. Error     t value    Pr(>|t|)     dF/dPar dPen/dPar
## Ti0  2.1159e+01  5.6487e-02  3.7458e+02  0.0000e+00 -2.1683e-03    0.0036
## Aw   7.8897e+00  5.6600e-01  1.3939e+01  0.0000e+00  8.3929e-05    0.0000
## Ci   2.0664e+00  2.6217e-02  7.8819e+01  0.0000e+00 -2.4714e-03    0.0000
## e11 -2.4036e+01  1.7007e+02 -1.4133e-01  8.8763e-01  5.4829e-05    0.0002
## p11 -1.6498e+00  1.6250e-02 -1.0152e+02  0.0000e+00  2.1506e-03    0.0000
## Ria  5.2927e+00  5.8203e-02  9.0935e+01  0.0000e+00  1.4144e-03    0.0000
##
## Correlation of coefficients:
##     Ti0   Aw    Ci    e11   p11
## Aw  -0.01
## Ci  -0.03  0.07
## e11  0.02  0.11  0.15
## p11 -0.01  0.08 -0.09  0.08
## Ria  0.05 -0.34  0.06 -0.09 -0.01
```

The one-step predictions (residuals) are estimates of the system noise (i.e. the realized values of the incremental *dw* of the Wiener process) added together with the observation noise. The assumptions are that the one-step predictions are white noise. Validate if this assumption is fulfilled, by plotting the autocorrelation function and the accumulated periodogram for the residuals. Is the model model suitable, i.e. does it describe the heat dynamics sufficiently?

The model is not sufficient since the residuals seem to be autocorrelated and there is information in the residuals that the model is not capturing. The periodogram should have all "frequencies" represented equally (horizontal lineo of "white noise") but they are not. Also the accumulated periodogram should show equal distribution of "frequencies" (fit in the diagonal line) and the residuls are clearly outside the blue lines (confidence intervals).





It is possible to gain some information about what is missing in the model, with time series plots of the residuals and the inputs. Are there some systematic patterns in the residuals?

The residuals show a systematic pattern, specially after the second day. They are not behaving as white noise. There is a clear pattern that is affected when the heater is turned on/off and the model "re-adjusts" so the residuals change

If yes, do they seem to be related to the inputs? To any specific events in the inputs?
There seems to be a systematic offset in the residuals, specially when the heater is off. We could assume that there is need for another state in order to take into account the thermal dynamics of air (fast changing) separately from the thermal changes of the walls or other objects in the building (slow changing).

Q2 - Extend the simplest model

First the same model but now using the wrapped function Ti where the model is defined.

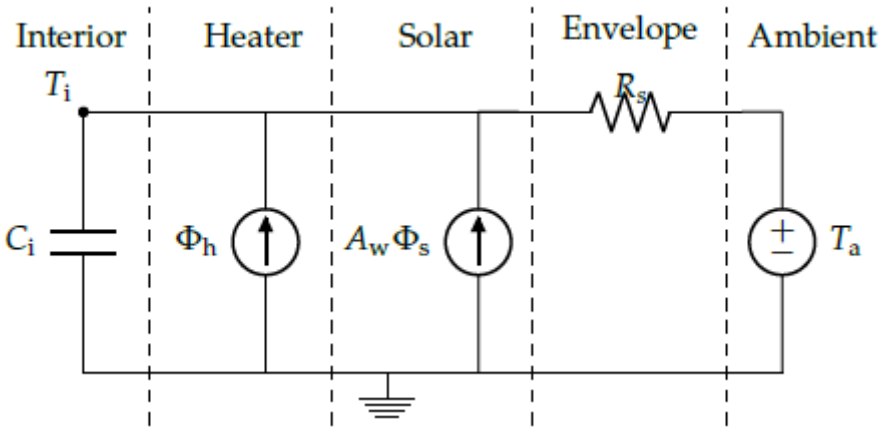
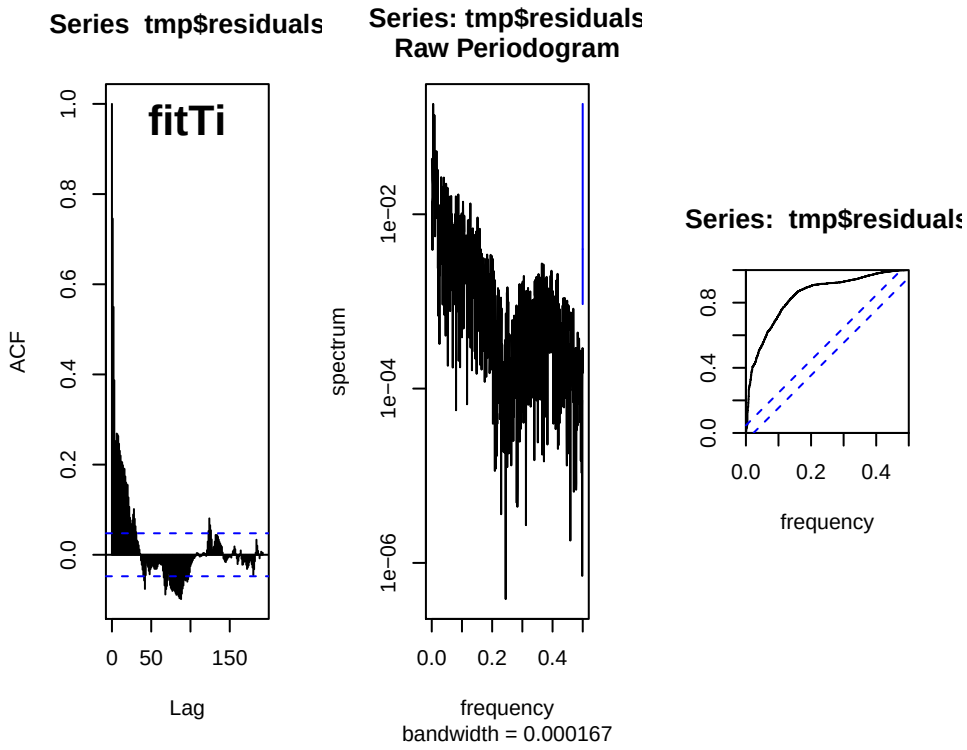
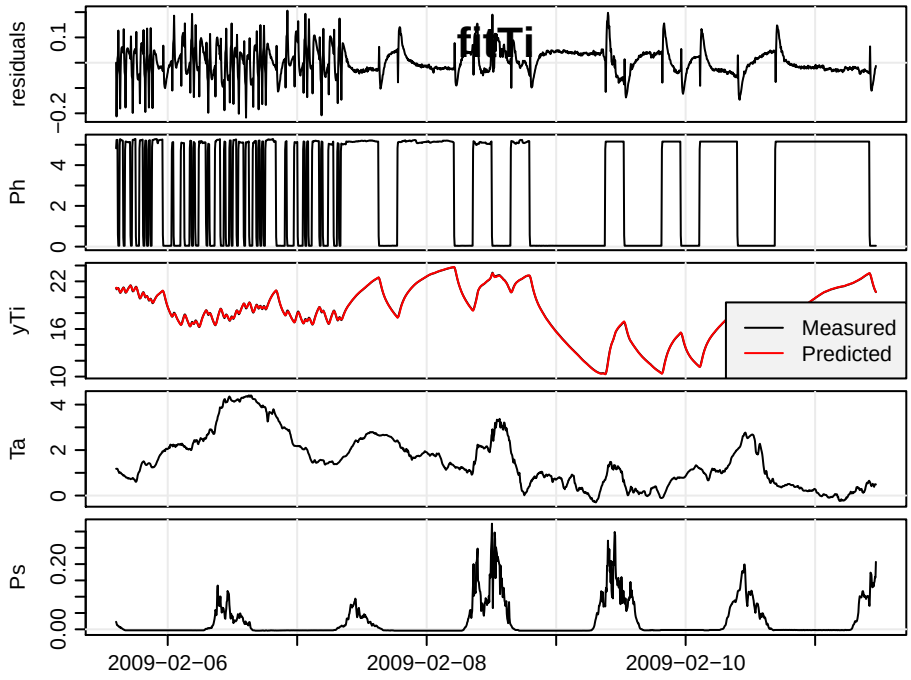


Figure 1: RC-network of the most simple model Ti





TiTe

The most simple model extended with a state in the building envelope TiTe.

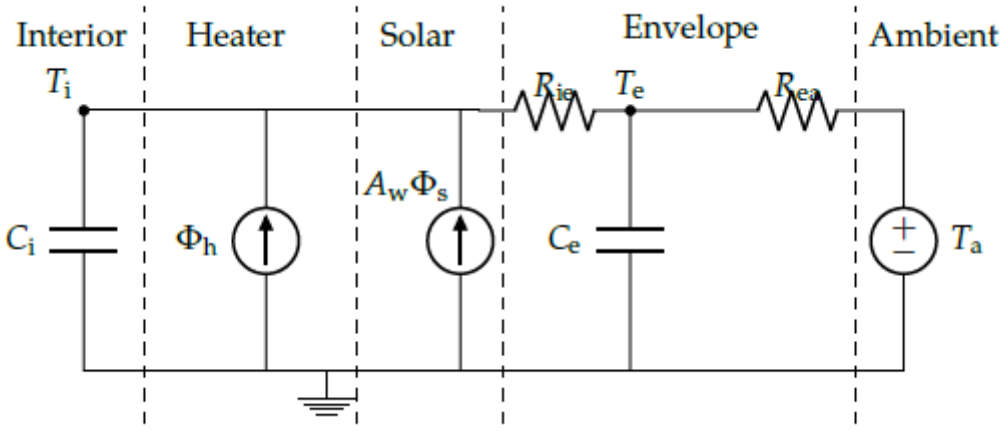
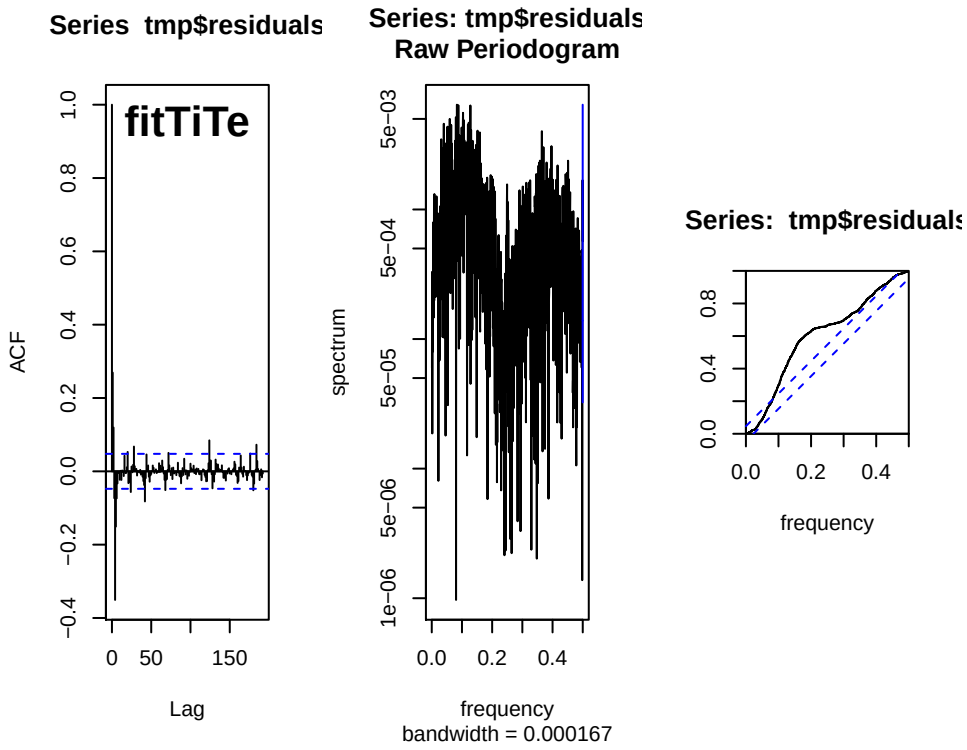
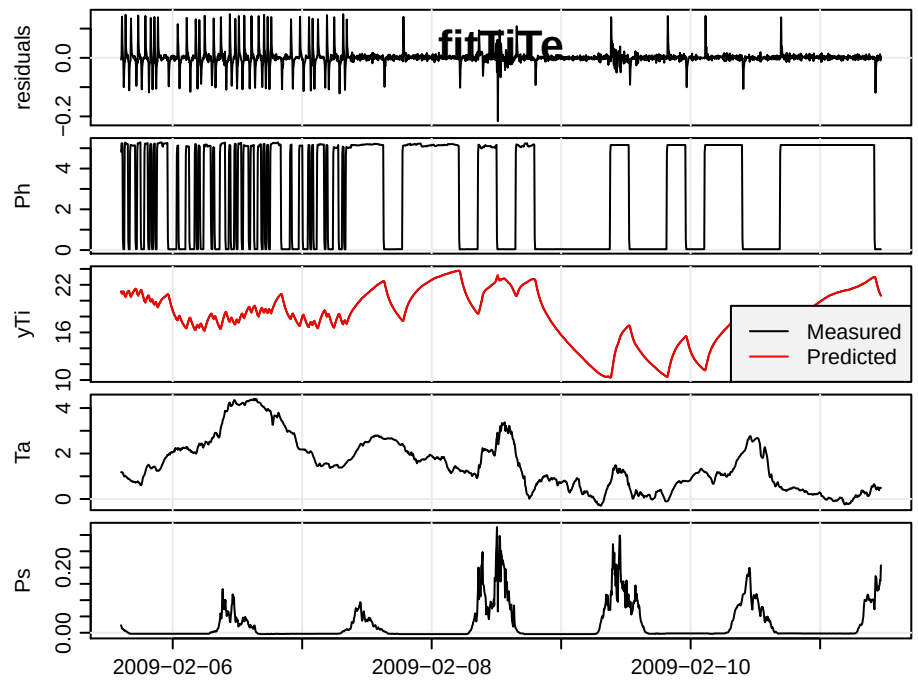


Figure 2: RC-network of the most simple model extended with a state in the building envelope TiTe





TiTh

The most simple model extended with a state in the heater TiTh.

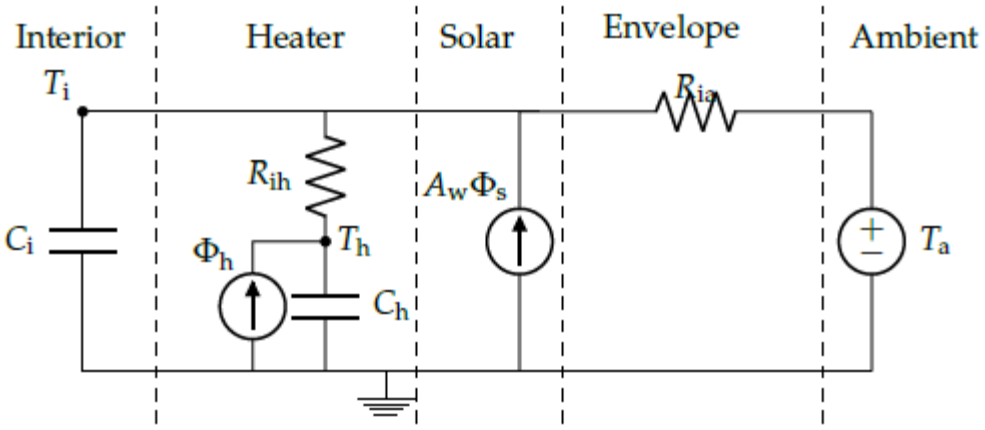
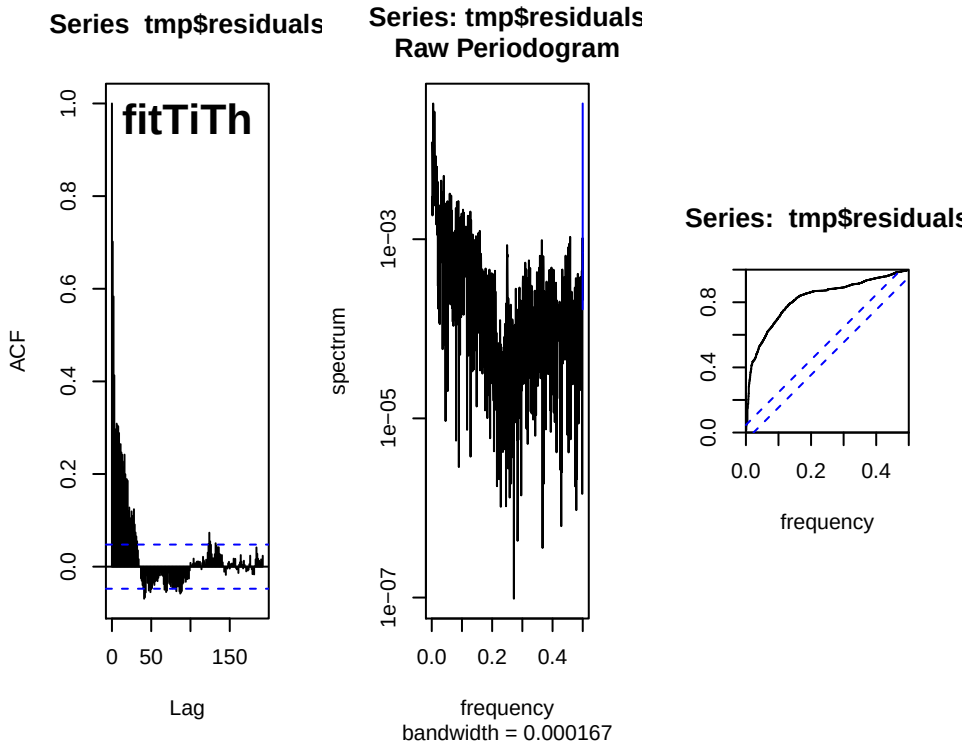
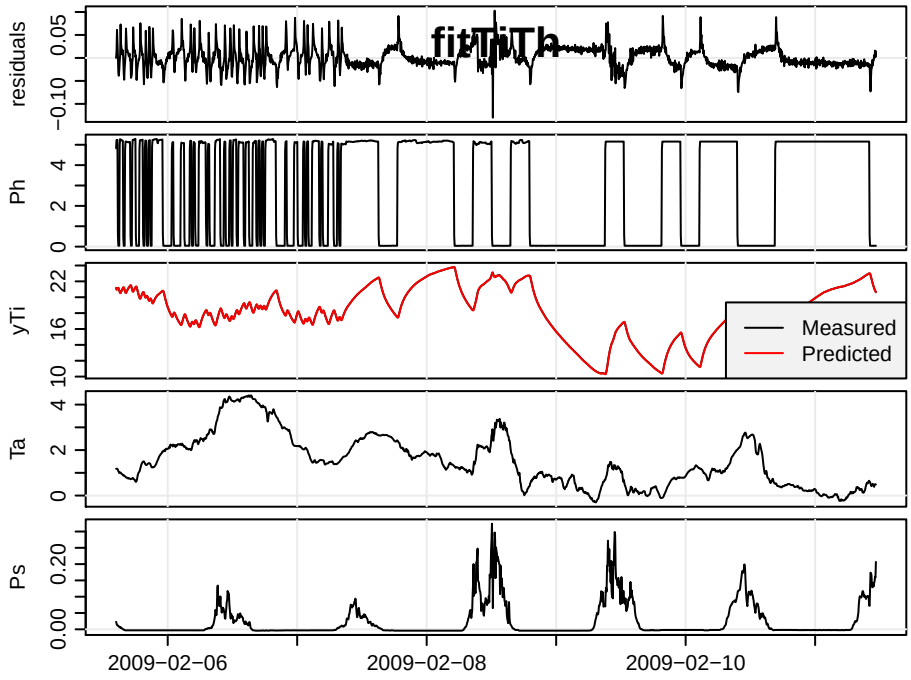


Figure 3: RC-network of the most simple model extended with a state in the heater TiTh.





TiTm

The most simple model extended with a state in which the solar radiation enters TiTm.

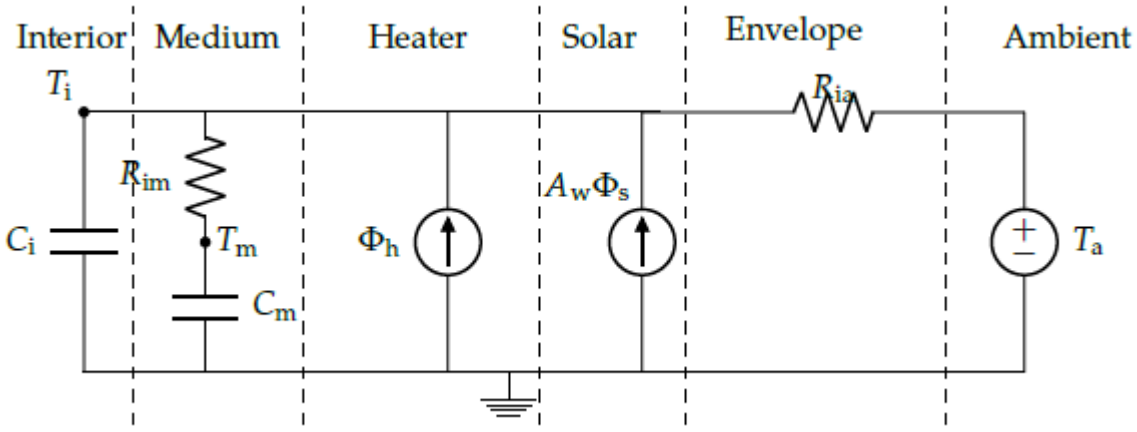
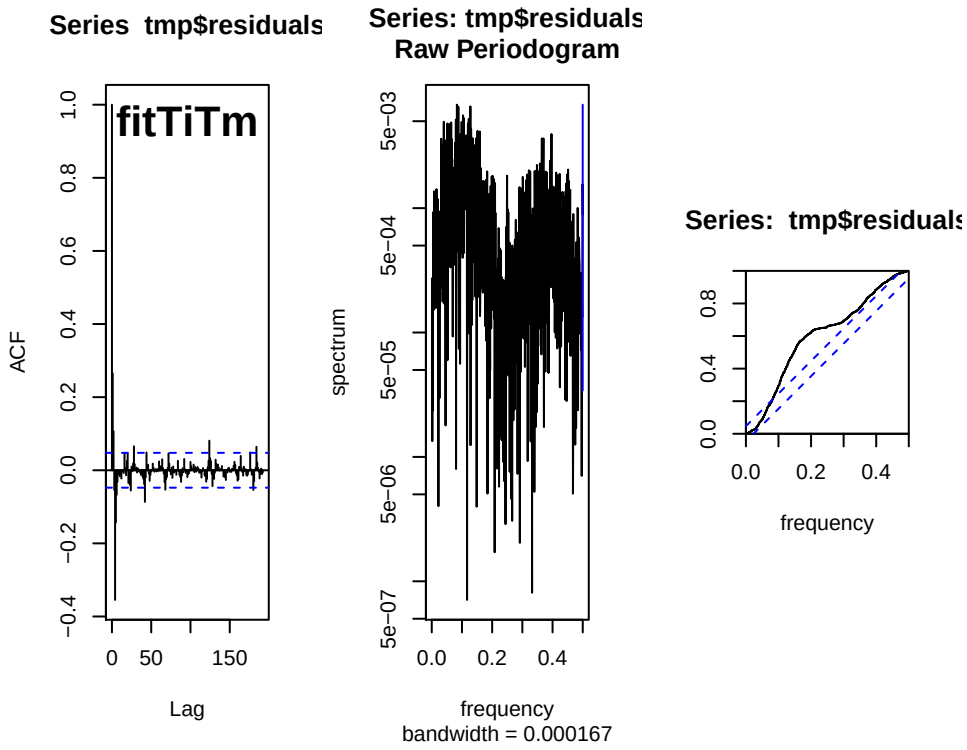
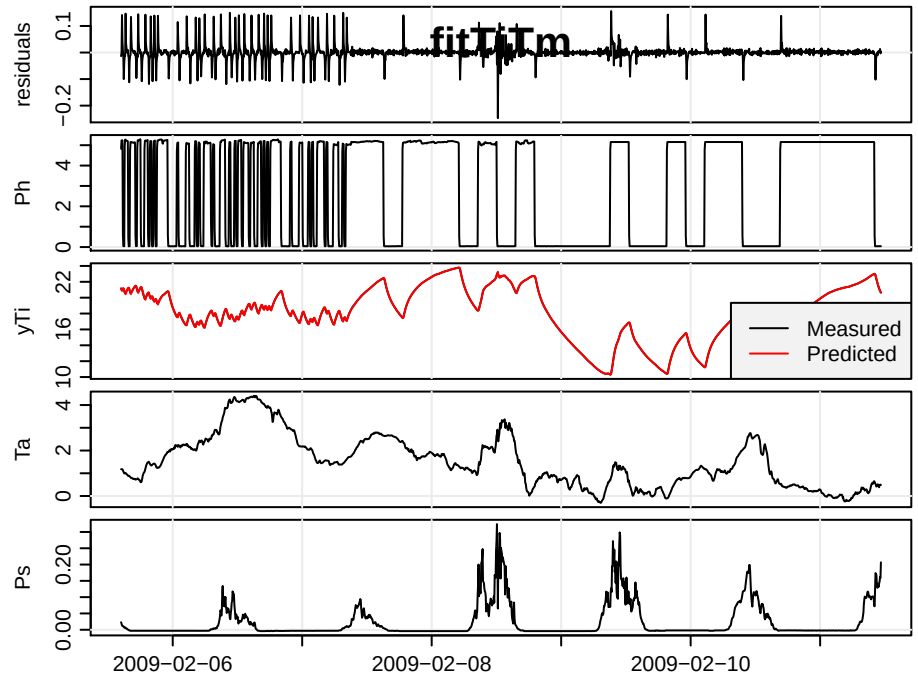


Figure 4: RC-network of the most simple model extended with a state in which the solar radiation enters TiTm.





TiTs

The most simple model extended with a state in the sensor TiTs.

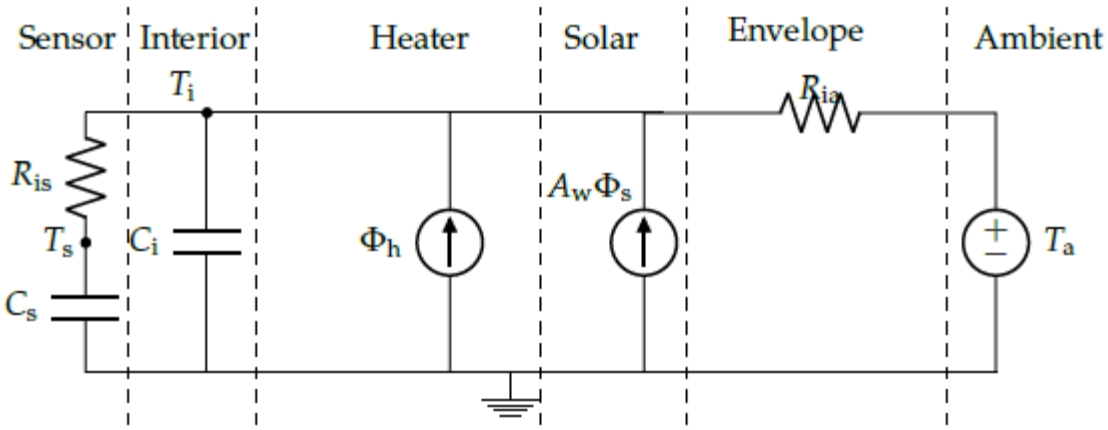
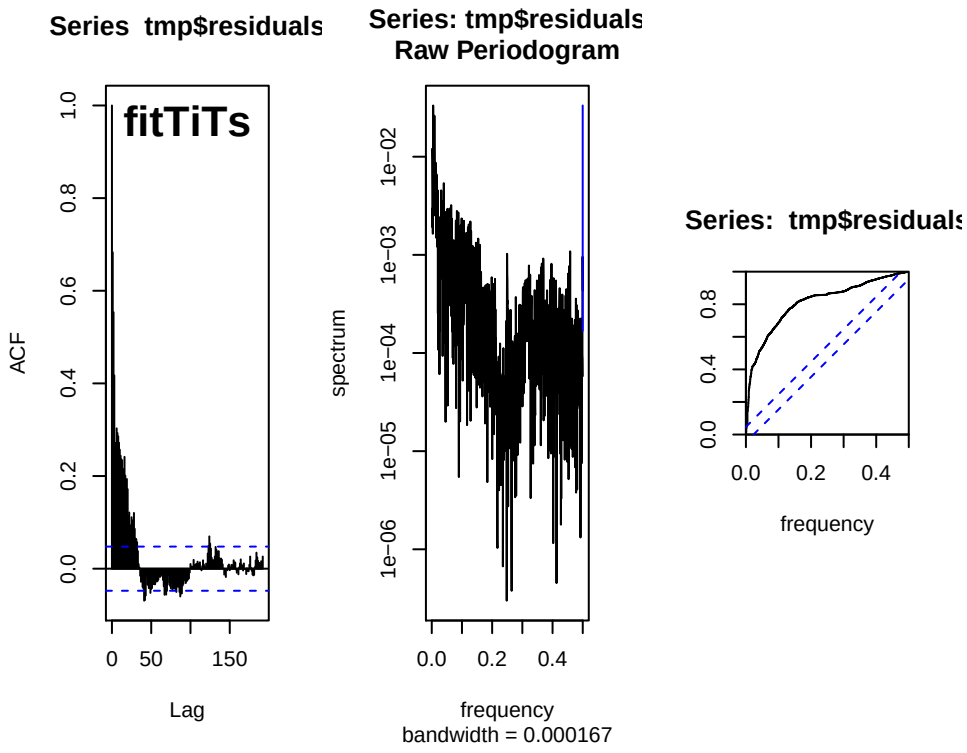
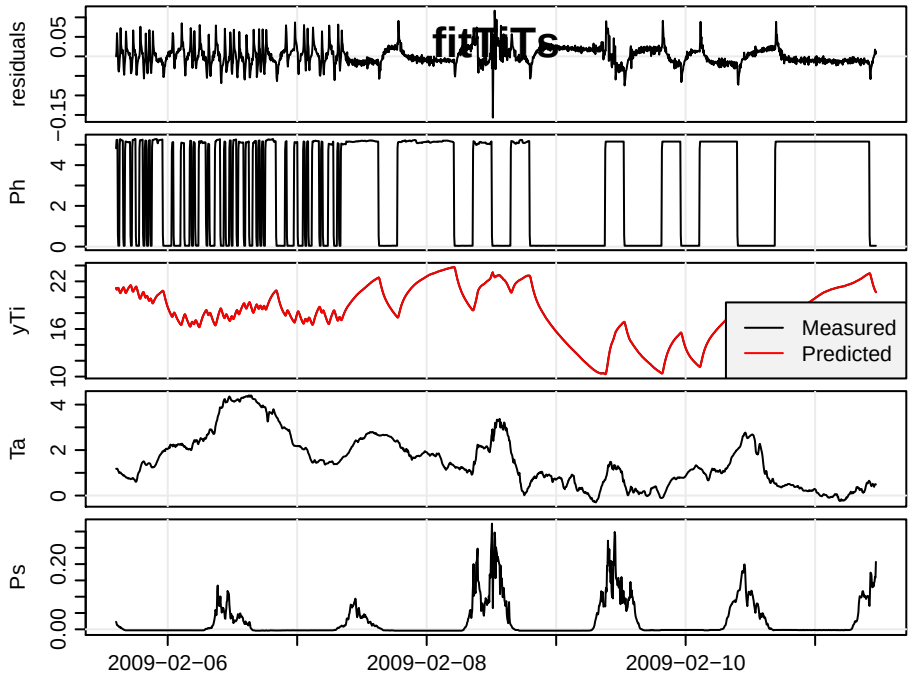


Figure 5: RC-network of the most simple model extended with a state in the sensor TiTs.





Are the extended models improved regarding the description of the dynamics (hint, analyse the residuals)?

The models are improvements to the simplest model, at least based on the residuals analysis, specially **TiTm**. Although it is not optimal yet. It is still needed to optimize based on the likelihood and number of parameters in the model.

Q3 - Selection of model

Use the script to carry out a likelihood ratio test of the simple model to each of the extended models. If the p-values show that more than one of the extended models are a significant improvement, it is suggested to select the extended model with the highest maximum likelihood.

Which of the extensions have the highest likelihood?

The TiTh has the highest likelihood

Perform a likelihood ratio test:

lambda = lik(smallerModel)/lik(largerModel) ,

where the smallerModel is submodel of the largerModel and lambda is chi2(f) distributed with f=dim(smallerModel)-dim(largerModel). Page 20 in Madsen2006.

The p-value is very small indicating that the difference between the models is significant and that the improvement from one model to the other is very unlikely by chance.

Forward model selection: take the selected model and extend it again once more, see the functions with third order models (i.e. models with 3 states).

TiThTe

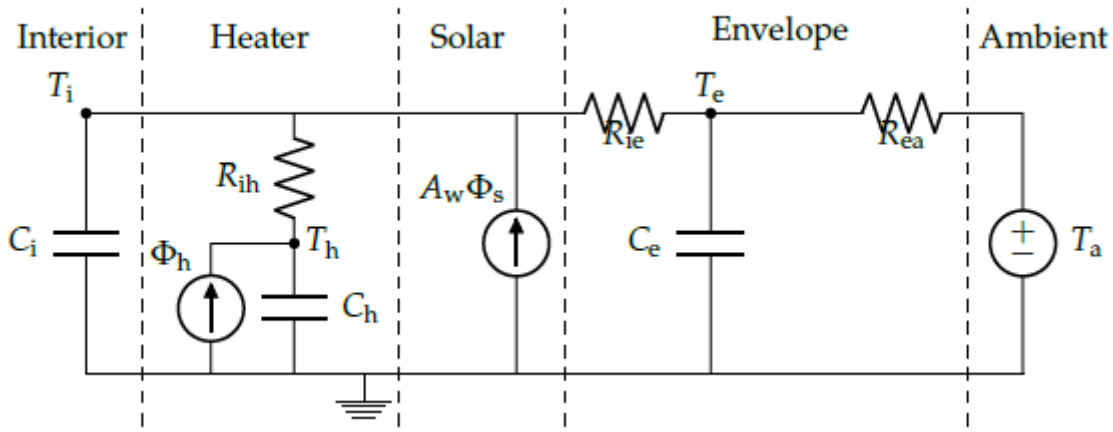
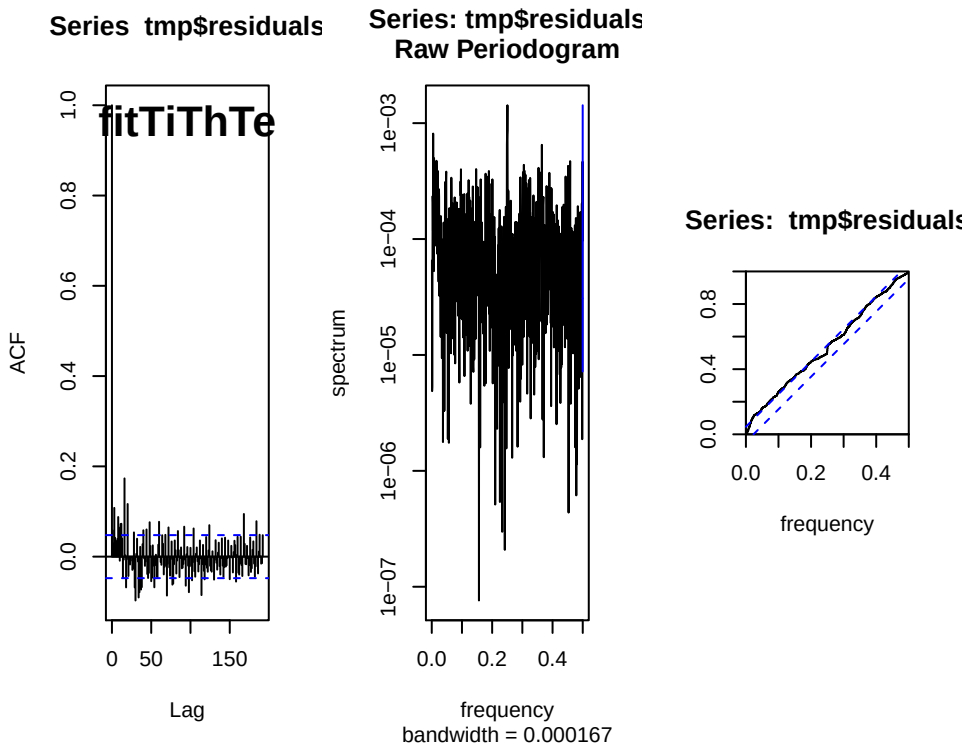
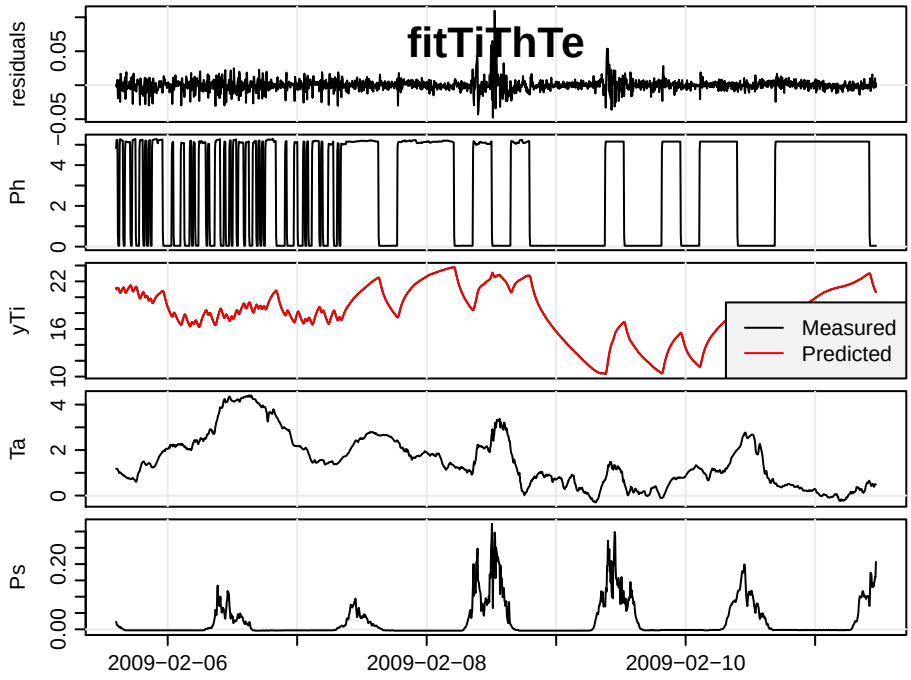


Figure 6: RC-diagram of TiTeTh.





TiThTs

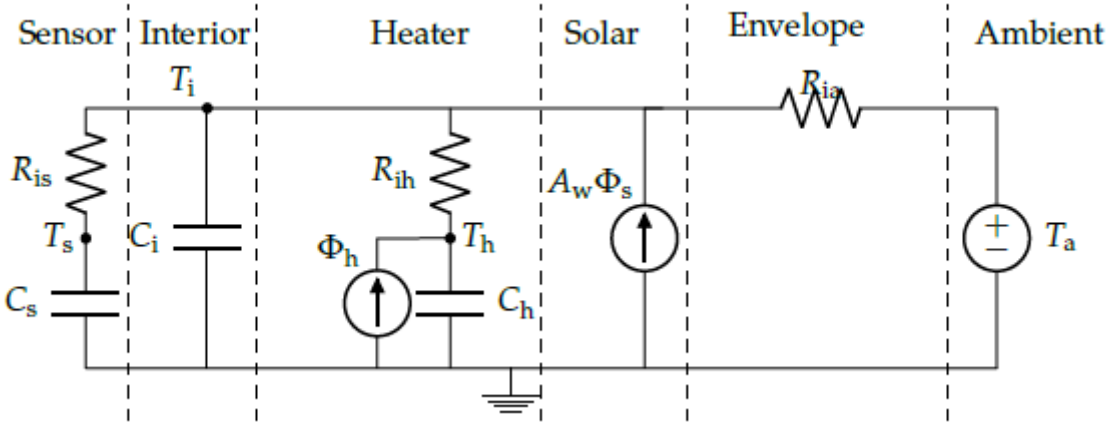
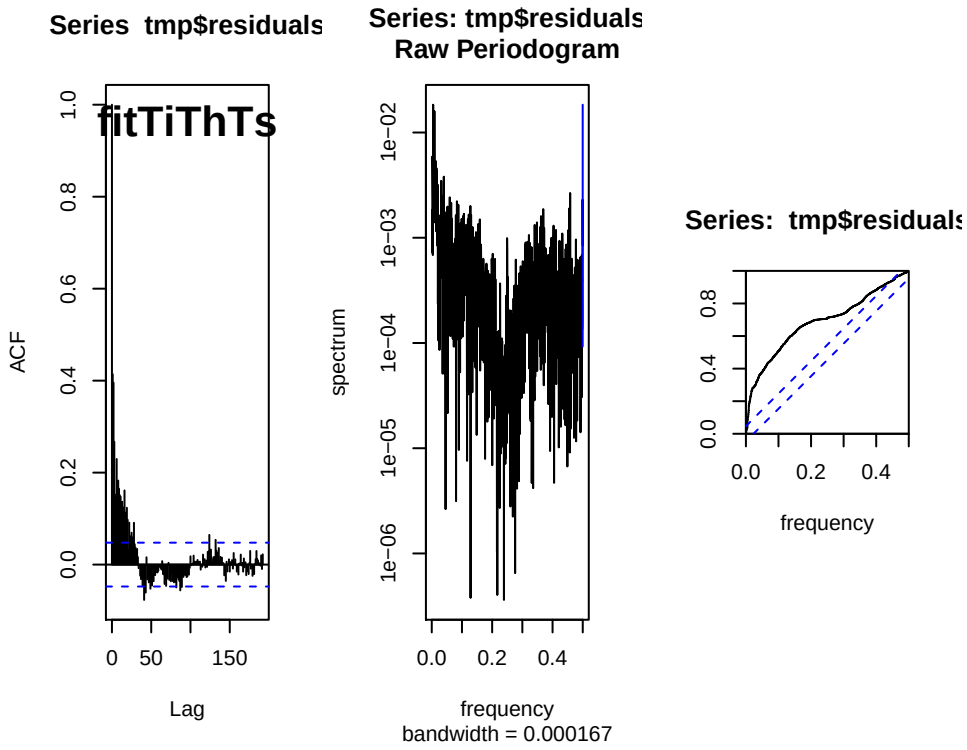
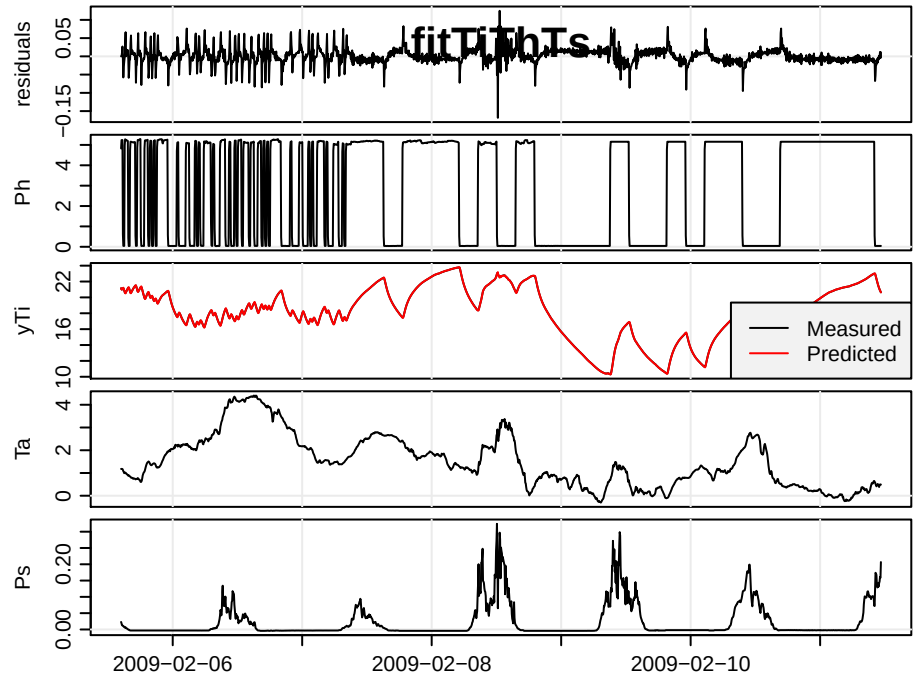


Figure 7: RC-diagram of TiThTs.


Series tmp$residuals
fitTiThTs
ACF
Lag
Series: tmp$residuals
Raw Periodogram
spectrum
frequency
bandwidth = 0.000167
Series: tmp$residuals
frequency


fitTiThTs
residuals
Ph
yTi
Ta
Ps
Measured
Predicted
2009−02−06
2009−02−08
2009−02−10

TiThTm

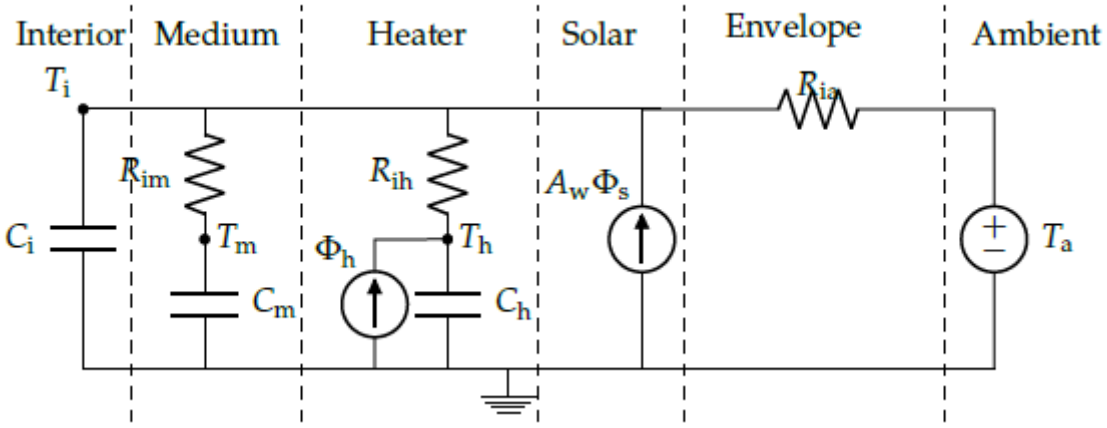
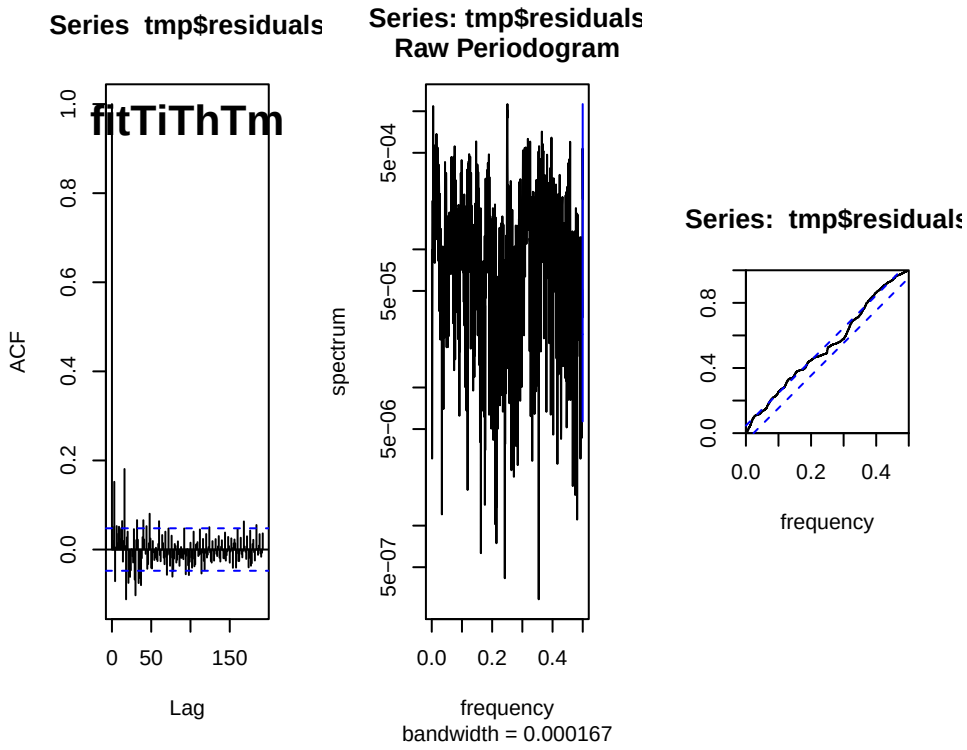
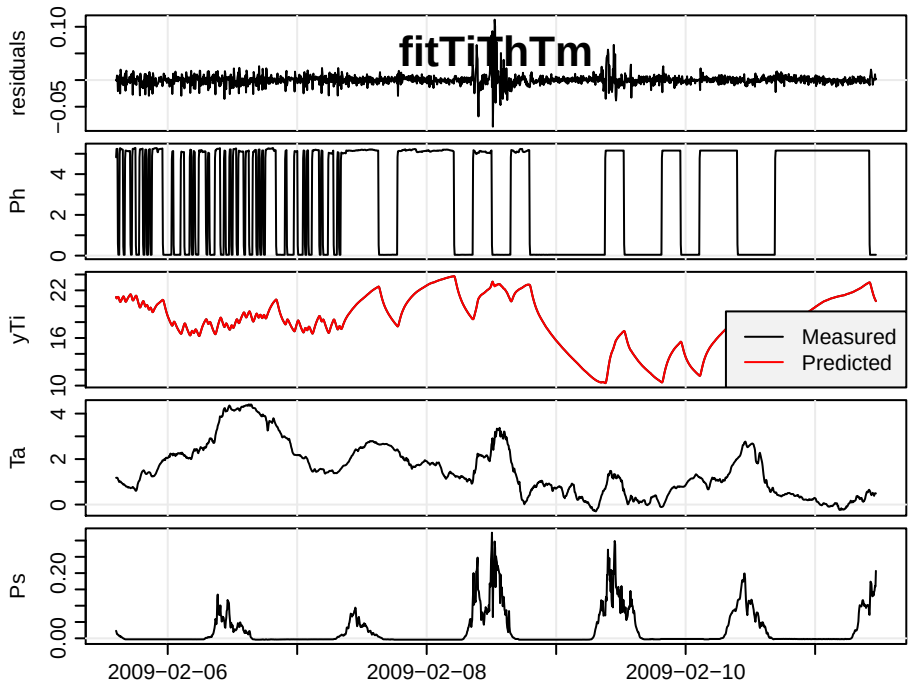


Figure 8: RC-diagram of TiThTm.





Which one of the three extended models should we select?

From the residuals graphs, the TiThTe model seems to be fitting the data better.

From this point the selection and extension procedure should be continued (i.e. models with 4+ states). until no significant extension can be found, however this is beyond the scope of the exercise.

TiTeThTs

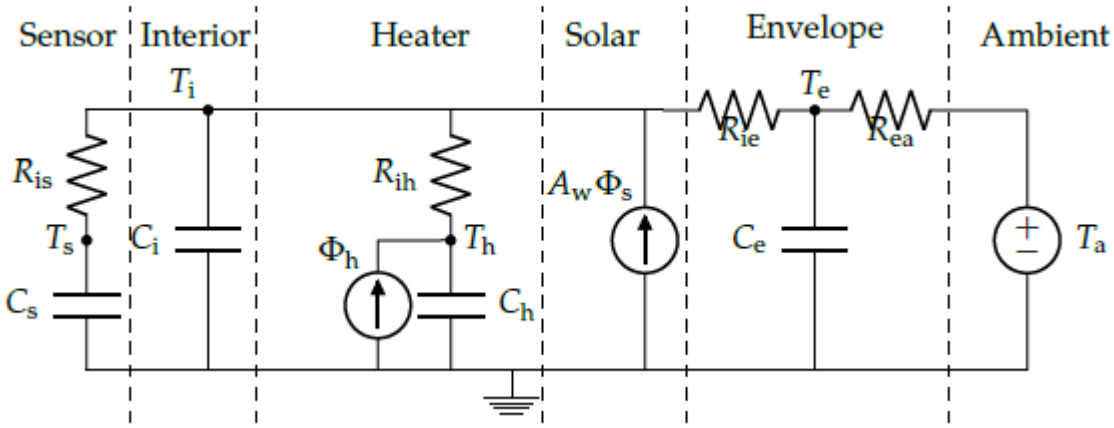


Figure 9: RC-diagram of TiTeThTs

The log-Likelihood ratio test confirms that the TiThTe is the best mextension to the model.

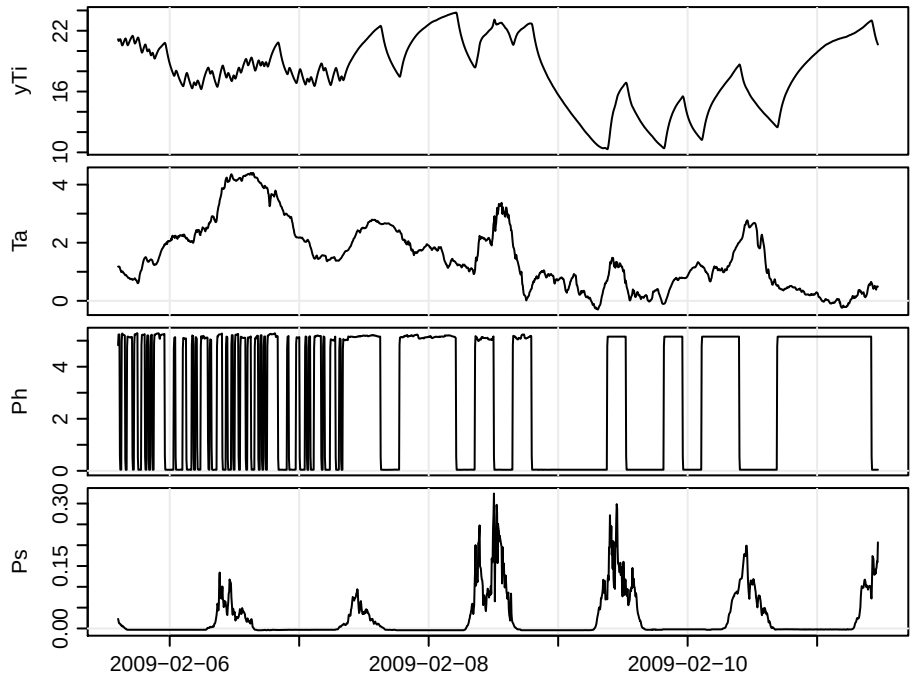
From here we should keep on extending the model, but for now no larger linear models are implemented here. From this points it is also likely that extensions to linear models compensate for non-linear or time-dependent effects.

See the article *"Identifying suitable models for heat dynamics"*, it is included in the .zip file, the performance (i.e. ACF(e_k) etc.) doesn't really change for models larger than TiThTe compared to the larger tested linear model. Hence we should rather look in residuals for non-linear or transformations of the inputs in order to model the effects which are not described well in the current model.

Q4 - Pseudo Random Sequence Signals

This part deals with Pseudo Random Sequence Signals. The function **prbs()** is an implementation of the n-stage feedback registers in the paper (see the function definition in "r/functions/prbs.R).

Do the plotting of the data.



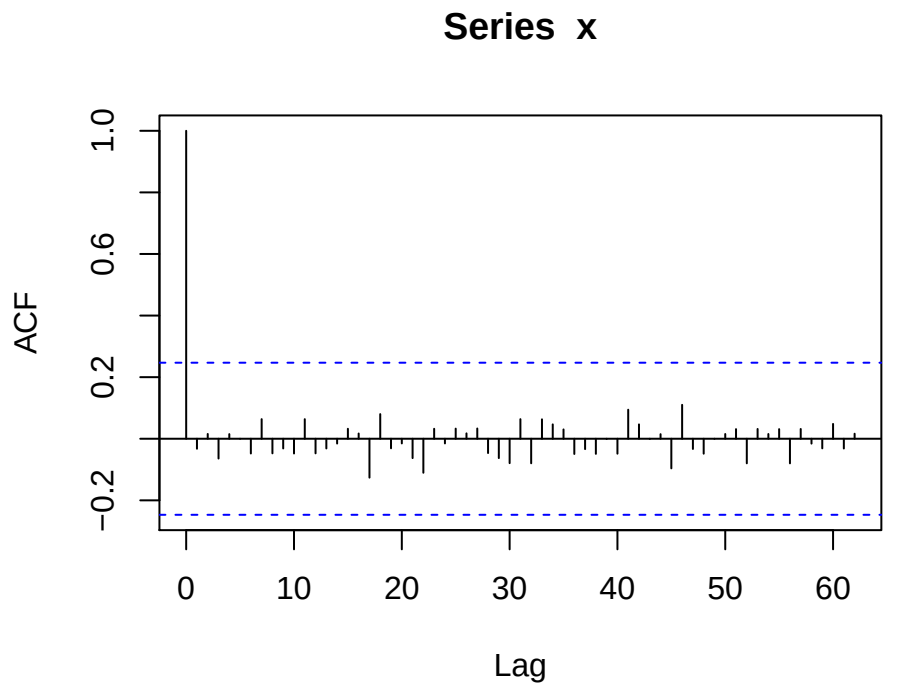
Which of the signals is a PRBS signal? The heater is a PRBS signal, with an on/off pseudo-random behavior.

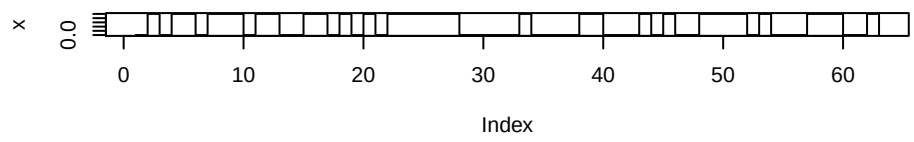
Which of the other signals are highly dependent on the PRBS signal?

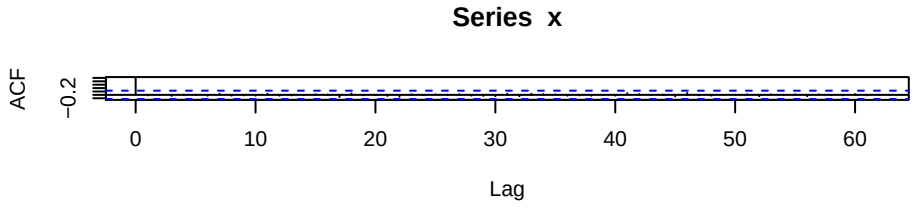
The indoor temperature, yTi, is highly dependent on the PRBS (heater) signal. When the heater is on the indoor temperature rises and when it is off it decreases.

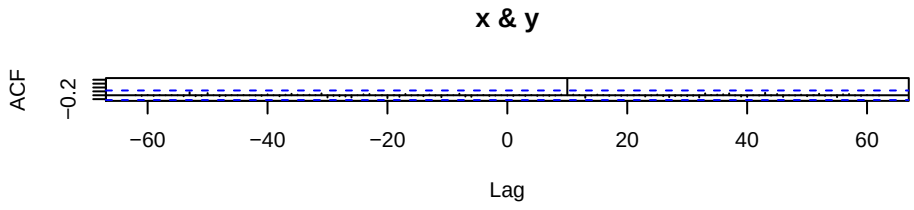
Generate a PRBS signal

Use the function defined in the file "functions/prbs.R", which generates a PRBS signal: + **n** is the length of the register + **initReg** just needs to be some initial value of 1,2,... it is the initial value of the registers and therefore only determines the start of the cycle. + **lambda** is the length of the smallest period in which the signal can change, given in samples

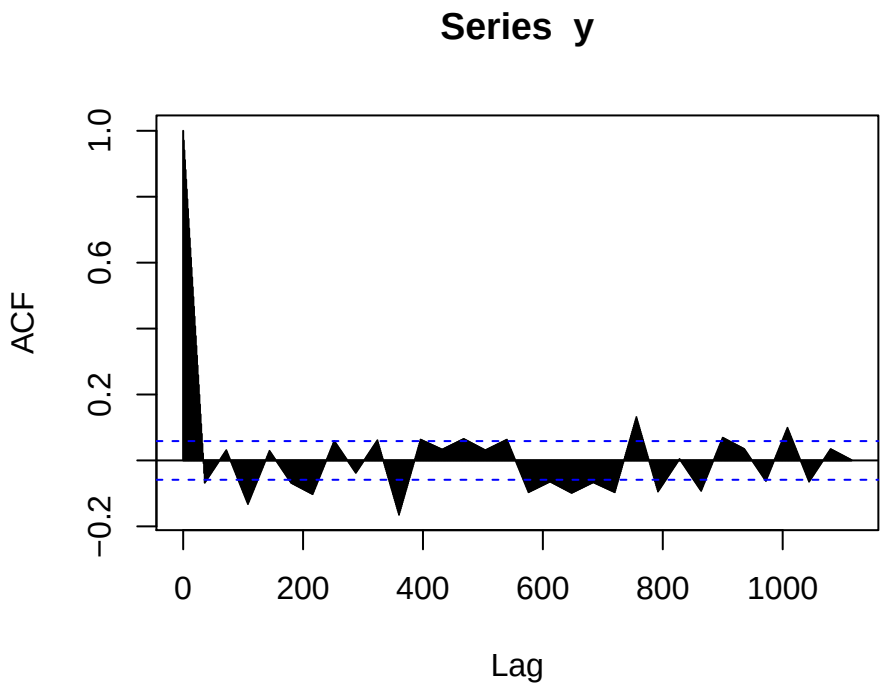


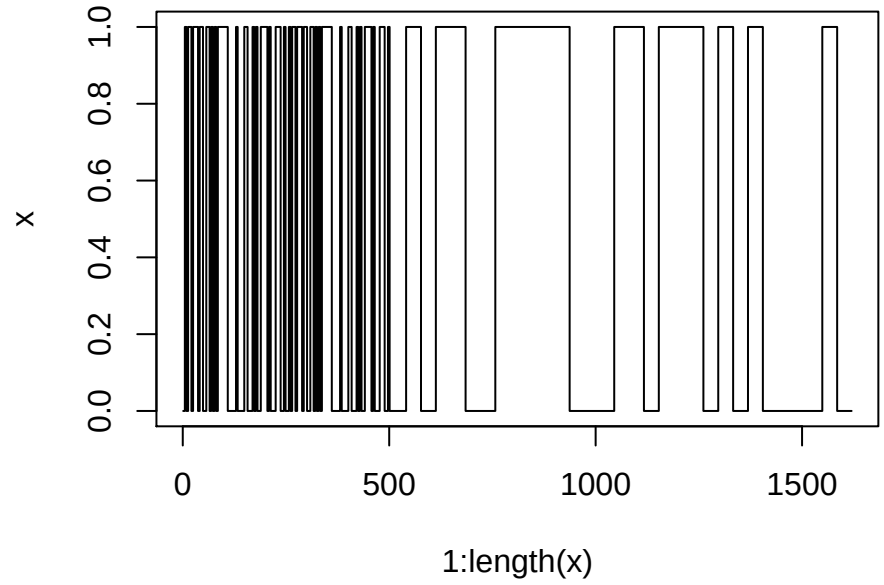




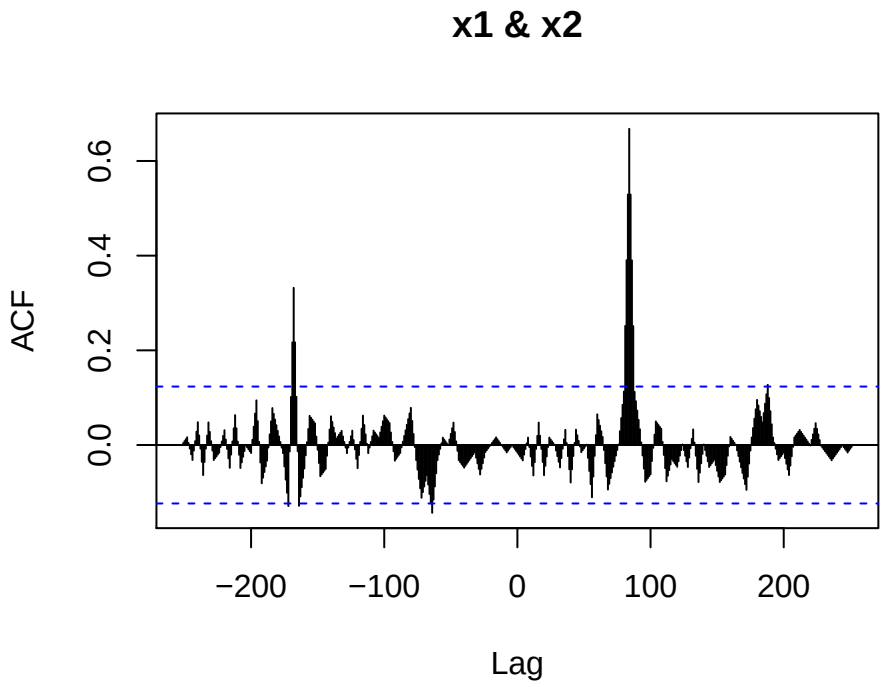


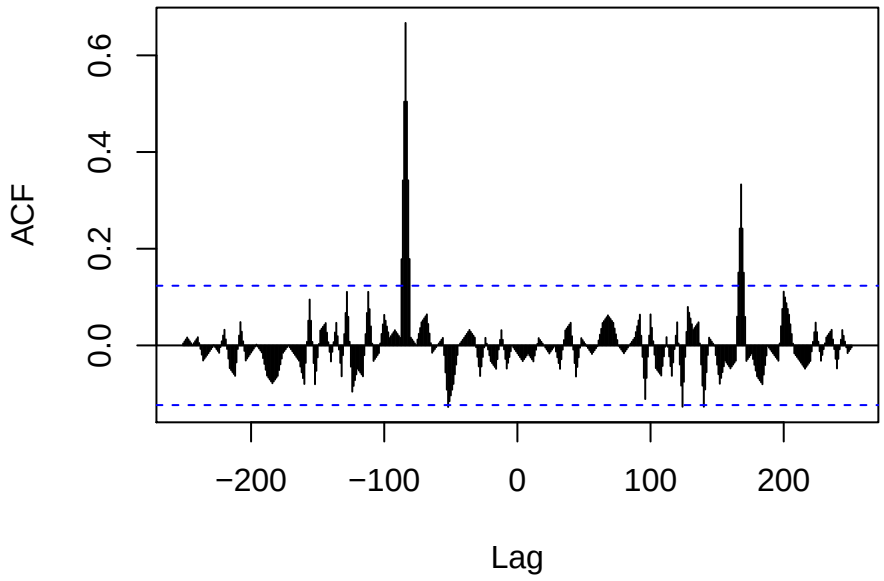
Generate the PRBS signals for the PRBS1 experiment, where a single signal controls all heaters in the building: * n = 6 * lambda = 4 Smallest period in one state for 5 minute sample period is then 4*5min=20min Settling time of the system (T_s) below the period (T_0) in: **lambda * (2^n-1) * 5 / 60 = 21 hours**

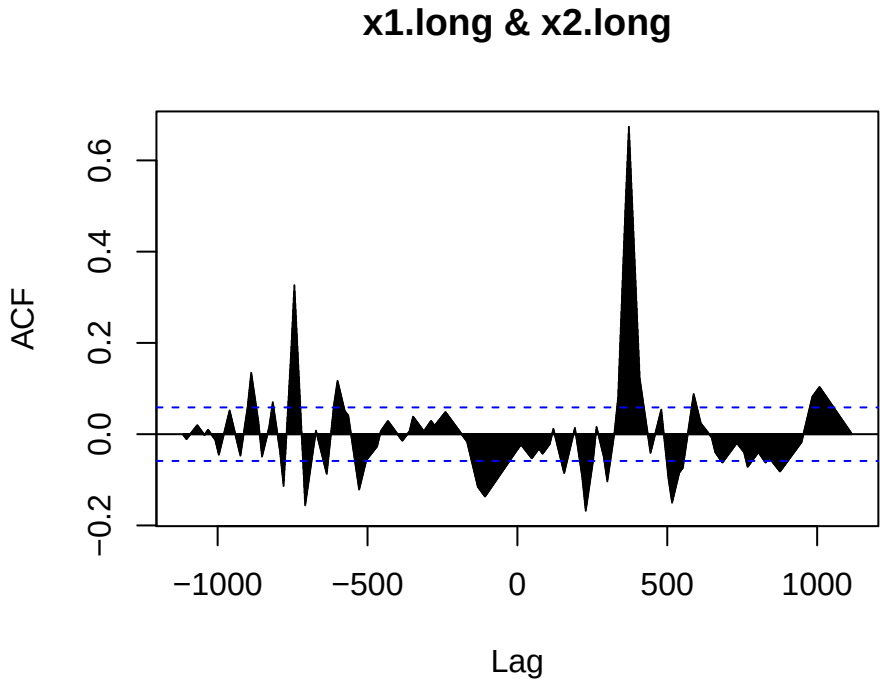


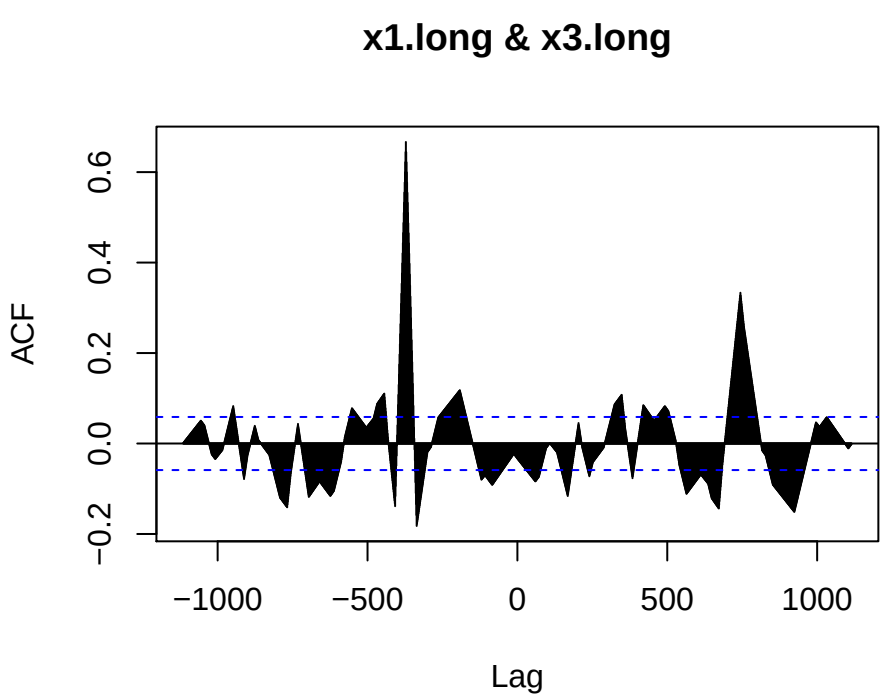


Just for seeing how to make multiple PRBS signals, which are uncorrelated, multiple PRBS signals are simply generated lagging the signals, such that they do not begin at the same time point:


x1 & x2
ACF
0.0
0.2
0.4
0.6
−200
−100
0
100
200
Lag


x1 & x3
ACF
0.0
0.2
0.4
0.6
−200
−100
0
100
200
Lag


x1.long & x2.long
ACF
0.6
0.4
0.2
0.0
−0.2
−1000
−500
0
500
1000
Lag


x1.long & x3.long
ACF
0.6
0.4
0.2
0.0
−0.2
−1000
−500
0
500
1000
Lag


x1
x2
x3
0.0
0.4
0.8
0
500
1000
1500

References

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. *rmarkdown: Dynamic Documents for R*, 2018. URL https://CRAN.R-project.org/package=rmarkdown. R package version 1.10.

Peder Bacher and Henrik Madsen. Experiments and data for building energy performance analysis: Financed by the danish electricity saving trust. 2010.

Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, 2011.

Keith R Godfrey. Correlation methods. In *System Identification*, pages 527–534. Elsevier, 1981.

Rune Juhl. *ctsmr: CTSM for R*, 2018. R package version 0.6.17.

Henrik Madsen. *Time series analysis*. Chapman and Hall/CRC, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.2.1.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2018. URL https://CRAN.R-project.org/package=knitr. R package version 1.20.