University of Liège
School of Engineering

Milestone 3: Uncertainty

PROJ0016: Big Data Project

Julien Gustin, Romain Charles, Joachim Houyon

November 18, 2021

1 Update from milestone 2

1.1 Model update

Since the last milestone we have done some improvement concerning the model:

We now consider (hopefully) the temperature set, such that when the temperature set is higher than the temperature of the zone we simply **turn on the radiator**, in the other case we **turn it off** but taking into account the thermal inertia that contributes to additional heating of the room with a parameters k s.t. $(0 < k < 1)$.

Here is our updated model:

$$\hat{T}_z(t+1) = \left(\frac{T_a(t) - T_z(t)}{R_z \times C_z} + \frac{Q_z(t)}{C_z} - \sum_{n \in neighbours} \frac{T_z(t) - T_n(t)}{R_{z,n} \times C_z} \right) \times \Delta t + T_z(t) + a \quad (1)$$

Where $Q_z(t) = cp \times m \times (T_r(t) - T_z(t))$ and

$$T_r(t) = \begin{cases} T_w(t), & \text{if } T_s(t) - T_z(t) > 0 \textit{ (radiator on)}. \\ T_r(t-1) \times k, & \text{otherwise } \textit{(radiator off)}. \end{cases} \quad (2)$$

with $(0 < k < 1)$.

(See previous milestones for parameters and notations)

Also concerning equation *2*, we have first tried to add a parameters *tolerance* such that: $-1 \leq tolerance \leq 1$ instead of using 0 but when optimizing our model its value was at the order of 1^{-20} therefore for simplicity and efficiency reason we preferred let the constant 0.

And indeed adding the involvement of the radiator improved a lot our model, here is an example of our predictions for a typical winter week.

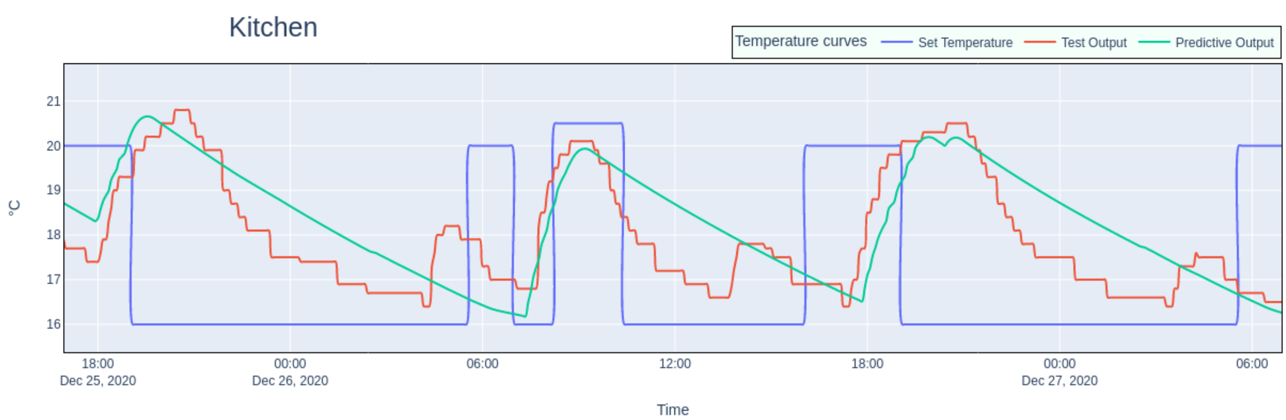

Figure 1. Zoom of a typical Winter day

By looking at this plot, we can see that our model seems to follow the trend even if it is not perfect, mainly because we have made the assumptions of a linear thermal loss of the radiator's heat.

1.2 Distribution of the residuals

Previously, our minimization problem was based on the assumption that the distribution of the residuals was following a `Normal distribution`. However, when drawing the histogram of the residuals of each room, it has been shown that our assumption was not so right.

Another assumption was the `Cauchy distribution`. It improved the results, but we could do better: the `Cauchy distribution` is a simple form of the `Student's t-distribution` with `1 degree of freedom`. Therefore, we decided to minimize the log pdf of a student-t distribution, where `the degrees of freedom df` is estimated.

This being said, we seek to minimize the following objective function that is the `negative log likelihood of the Student's t-distribution`. That is, for a given cluster, the following minimization problem:

$$\begin{aligned} \min_{\theta} \quad & \sum_{r \in Cluster} \left(\sum_{t=1}^{N} f(T_{rz}(t), \hat{T}_{rz}(t)(\theta), 1, df) \right) \\ \text{s.t.} \quad & 0 < k < 1 \end{aligned} \tag{3}$$

Where $f(x, \mu, \sigma, v)$ is the `negative log likelihood of the Student's t-distribution` of mean μ, scale σ, degrees of freedom v and evaluated for a value x. θ represents all the parameters associated to the cluster and df. Regarding the scale, we kept it equal to 1 because the parameter df already controls the thickness of the tails.

After the optimization process, it is seen that the degrees of freedom df is suggested to be more than 1. Indeed, the value of df was something close to 10, this also confirms that the noise is not Gaussian. Finally, we can see from the histograms of the noise that our current distribution assumption makes sense.

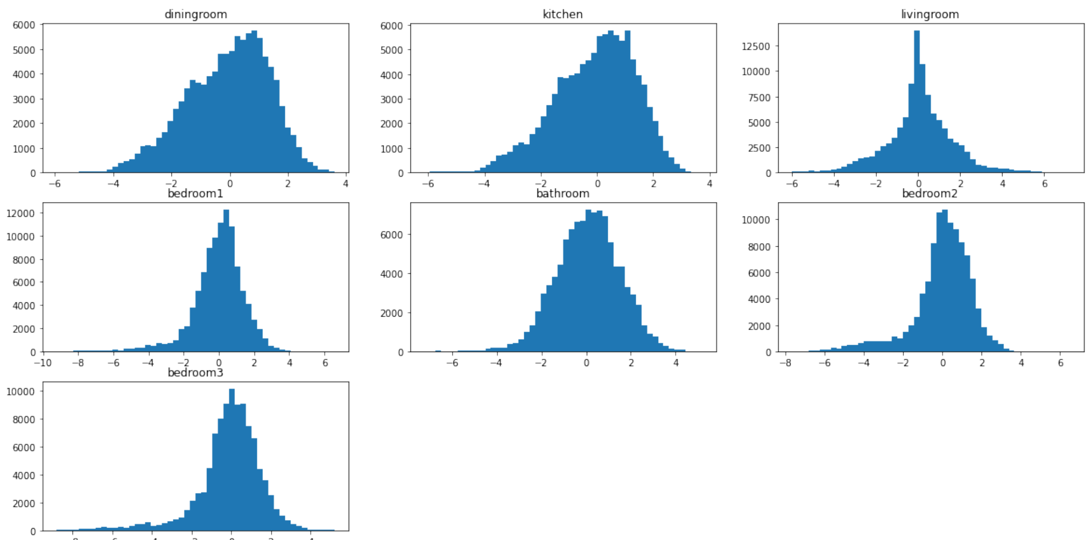

Figure 2. Residual of our predictions given observed temperatures

2 Uncertainty

Where $\theta = \{R, C, k, a, m, df, R_{z,n}\}$ and $D_t = \{T_a(t), T_r(t), T_w(t), T_s(t), T_n(t)\}$ $\forall n \in neighbours$

2.1 Temperature measurements

A stochastic relationship between measurement m_{t+1} and model parameters θ can be expressed as:

$$m_{t+1} = T_z(t+1)(\theta) + \epsilon$$

Where ϵ represent the noise that accounts for measurement error, and recall that the distribution of the noise is assumed to follow a Student's t-distribution.

The thermal model as a Markov Chain tells us that the measurement

$$m_{t+1} \sim P(m_{t+1} \mid \hat{T}(t+1)(\theta))$$

can be written as

$$m_{t+1} \sim t(m_{t+1} \mid \hat{T}(t+1)(\theta), df)$$

2.2 Model parameters

Using Bayes' rule we can represent the posterior density of parameters θ as:

$$P(\theta|m_{t+1}) = \frac{P(m_{t+1} \mid \hat{T}(t+1)(\theta)) \times P(\theta)}{P(m_{t+1})} \tag{4}$$

$$\propto P(m_{t+1}|\hat{T}(t+1)(\theta)) \times P(\theta) \quad \textit{Removing the normalization factor} \tag{5}$$

$P(m_{t+1} \mid \hat{T}(t+1)(\theta))$ is *known* (see 2.1) and concerning the prior distribution of our parameters $P(\theta)$, we have decided to use a uniform distribution *centered* at θ_{mle} (optimized parameters) and bounded on a interval of the order of magnitude of theses optimized parameters. We have made this choice because we are not expert in the field.

Then in order to find then **log posterior distribution** $P(\theta|m_{t+1})$ we use the MCMC approach to expresses parameter uncertainties.

2.3 Transition model

Our model is assumed to be deterministic. Therefore, the transition between two time-steps is determined by the **equation (1)**.

2.4 Methodology

To capture uncertainty, we use a subset of the dataset (more-less the time-series of one week) during the winter and summer (for looking at when the radiator is on and off). We run MCMC. on a number of walkers that is twice greater than the number of parameters, each walker making 3000 steps. (the first 250 steps are discard as burn-in).

The idea is to have a kind of sliding window of a two weeks to predict the next few days, for now and mainly for evaluation we use the truth values of T_a, T_z and T_w as input for forecasting, but later these value would also be unknown (milestone 4).

Note also that we use the last temperature measured as starting point for the D.

2.4.1 Winter

Here is an example of forecasting in winter for the dining room.

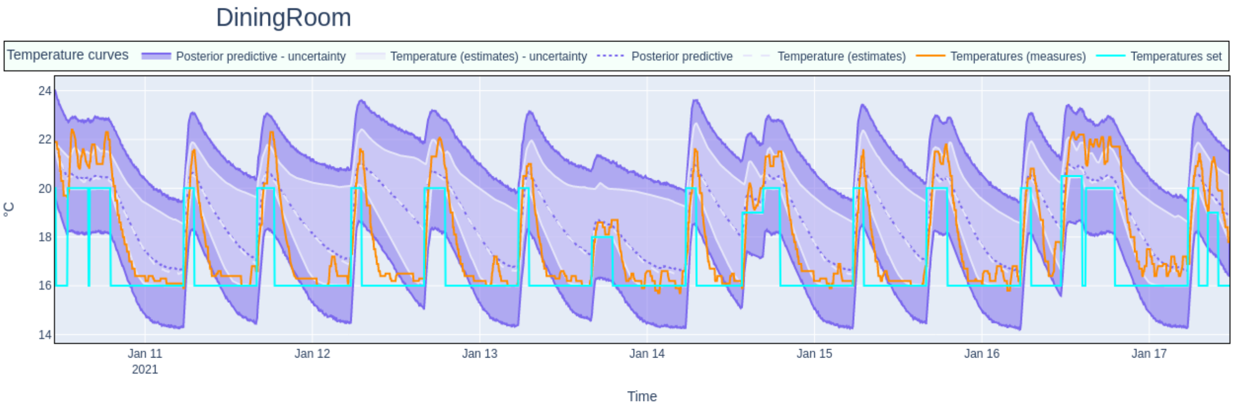

Figure 3. Forecast of a week in January given the precedent week, for the dining room

The results are quite interesting, because we can see that when the **radiator is on** our model is quite confident about what is going on, but as soon as it become **off** it start being less and less confident. This increase of uncertainty is expected, because our assumption about the loss of the radiator is not realistic (loss assumed linear)

2.4.2 Summer

In summer we do not have that much control on what is happening, and relay on the value of parameters R_* and C_* (see 2.4.3).

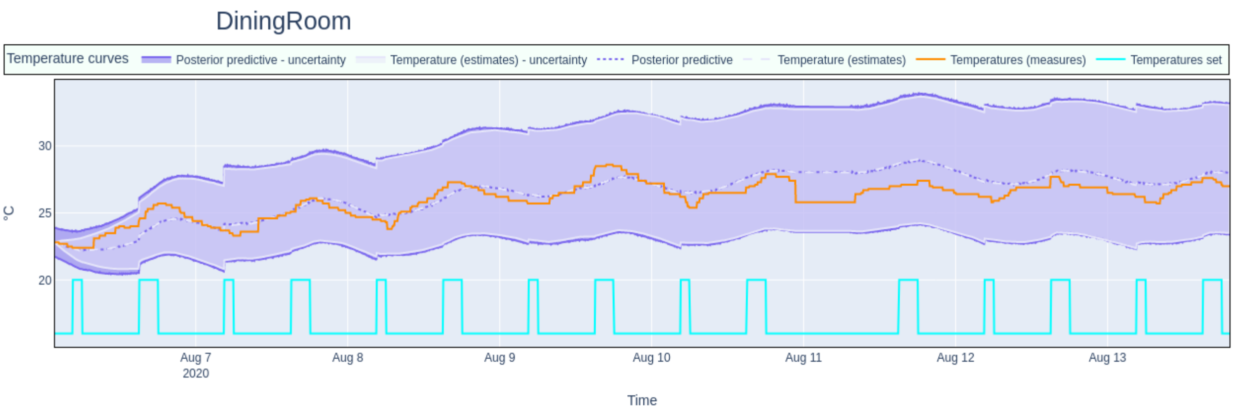

Figure 4. Forecast of a week in August given the precedent week, for the Bedroom3

It is worth noticing that the radiator does not contribute at all because the temperature set is almost always lower than the actual temperature estimated. Therefore, the uncertainty is not captured with these parameters.

2.4.3 Conclusion

As we can see our model is way more confident and accurate in winter than summer, this seems quite logical because in summer you do not really have a control of the temperature of the rooms and you rely mainly of the outside temperature, but also this season is more likely to un-expected things such as keeping the door open for heating the room, close all the door and curtain in order to cold it... While in winter it is the opposite because we are less likely to do what explained before and the heat of the house can be control through radiator (not true for all rooms, see living room).

However it is not such a problem because in this project we are way more interesting of what is happening in winter than summer. It is also worth mentioning that as expected the further we predict the higher is the uncertainty.

2.4.4 Corner plot

By looking at the corner plot for the living room in winter (recall that in milestone 2 we have made cluster of dependent room, and that living room is independent to all other one) we can see that our initials guess are quite good! However values of DF are

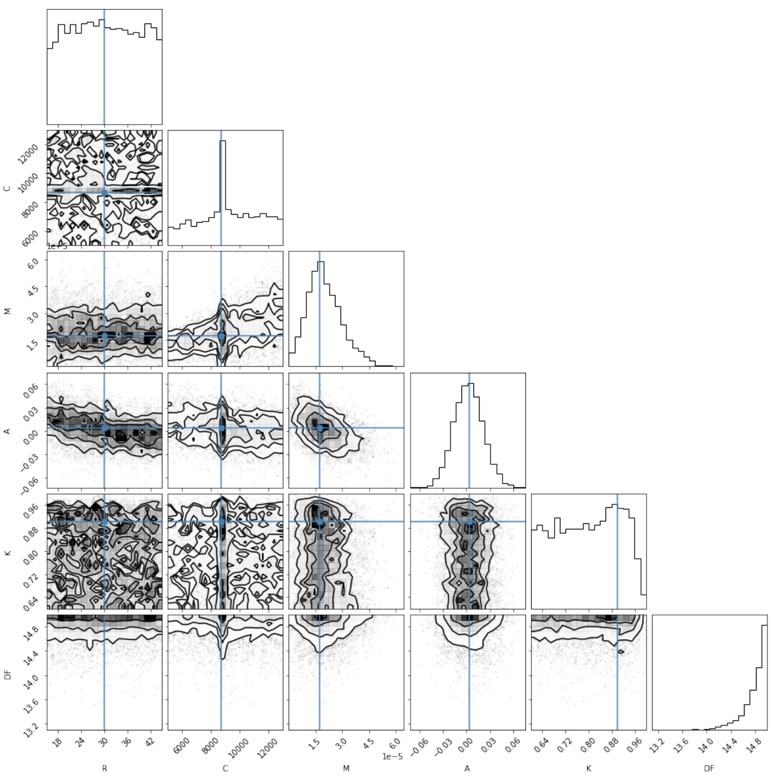

Figure 5. Corner plot of the living room in winter

quite strange, indeed it seems that it want to reach a border meaning that our guess was undervalued but when optimizing with higher value results look quite bad.