Danmarks Tekniske Universitet

DTU

Assignment 3

Authors

Bao Ngoc Thai - s242504
Erika Samantha Young - s242666
Alex Michel Louis Pardon - s243245

July 14, 2025

Contents

1 Stability

1.1 Simulate 5 realizations of the process

The AR(2) process can be rewritten as:

$$X_t = -\phi_1 X_{t-1} - \phi_2 X_{t-2} + \epsilon_t$$

5 realizations are simulated with $\phi_1 = -0.6$ and $\phi_2 = 0.5$ in Figure 1. The simulations appear stochastic and stationary.

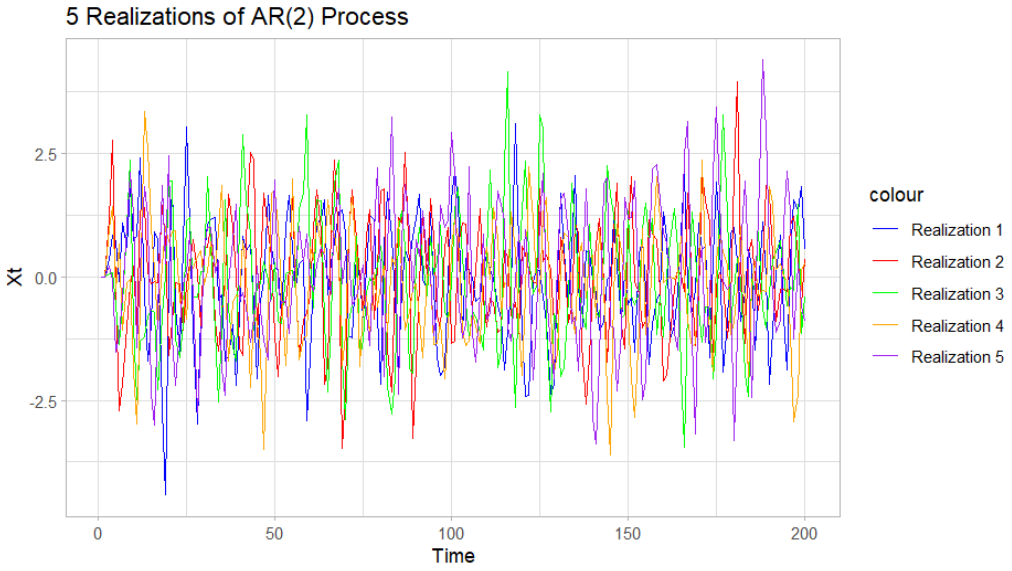

Figure 1: AR(2) process (time series) with n = 200 observations with $\phi_1 = -0.6$ and $\phi_2 = 0.5$

1.2 Calculate the empirical ACF and plot with $\rho(k)$

The theoretical ACF is the true autocorrelation function derived from a known model, given known processes, in this case AR(2). Theoretical autocorrelation function $\rho(k)$ for an AR(2) process satisfies the difference equation:

$$\rho(k) = -\phi_1 \rho(k-1) - \phi_2 \rho(k-2), \quad \text{for } k \geq 2$$

with initial conditions:

$$\rho(0) = 1$$

$$\rho(1) = \frac{-\phi_1}{1+\phi_2} \quad \text{(assuming stationarity)}$$

Empirical ACF is calculated based on the sample data and gives an estimate of autocorrelation at different lags. The empirical autocorrelation function (ACF) at lag k is given by:

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

Where:

- X_t is the value of the time series at time t.
- $\bar{X}$ is the mean of the time series.

- n is the length of the time series.

The empirical and theoretical ACF for the simulations of 1.1 is given in Figure 2. As the process is stationary, the empirical and theoretical ACF are very similar, where the theoretical ACF is around the average of the empirical ACF.

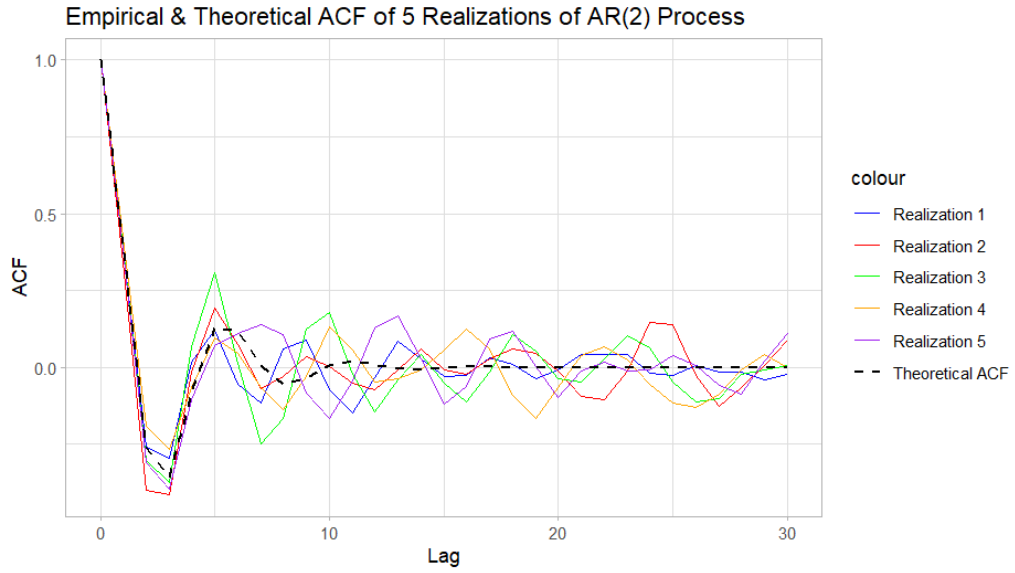

Figure 2: AR(2) process (ACF) with n = 200 observations with $\phi_1 = -0.6$ and $\phi_2 = 0.5$

1.3 $\phi_1 = -0.6$ and $\phi_2 = -0.3$

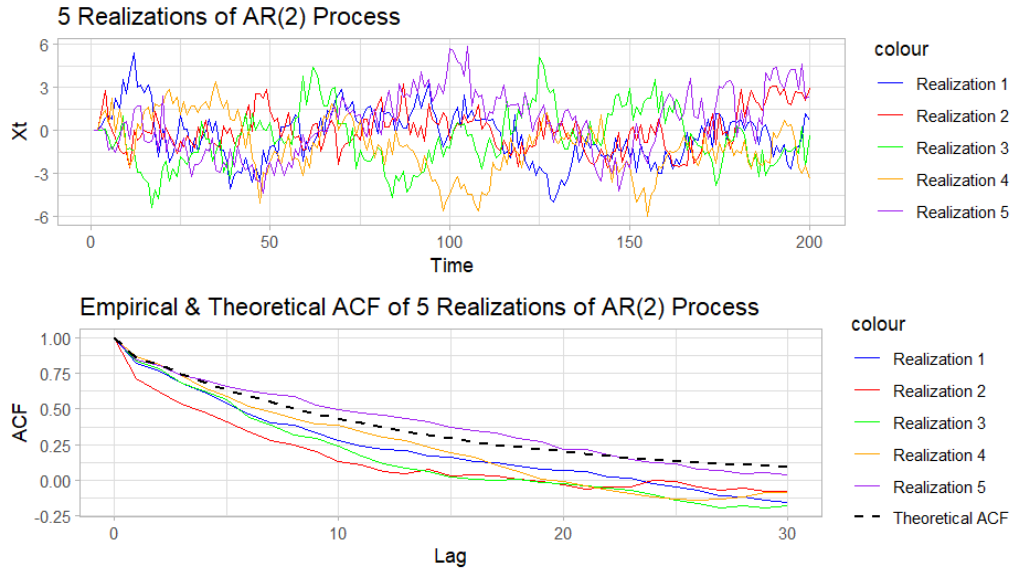

Figure 3: AR(2) process with n = 200 observations with $\phi_1 = -0.6$ and $\phi_2 = -0.3$

An AR(2) process with n = 200 observations with $\phi_1 = -0.6$ and $\phi_2 = -0.3$ along with its empirical and theoretical ACF is presented in Figure 3. Similarly to 1.2, the time series of 5 realizations are stochastic and mostly stationary. Therefore, the empirical and theoretical ACF show similar results.

1.4 $\phi_1 = 0.6$ and $\phi_2 = -0.3$

An AR(2) process with n = 200 observations with $\phi_1 = 0.6$ and $\phi_2 = -0.3$ along with its empirical and theoretical ACF is presented in Figure 4. The observations are similar, with a stationary time

series plot and corresponding theoretical and empirical ACF. The ACFs oscillate as the parameters have opposite signs.

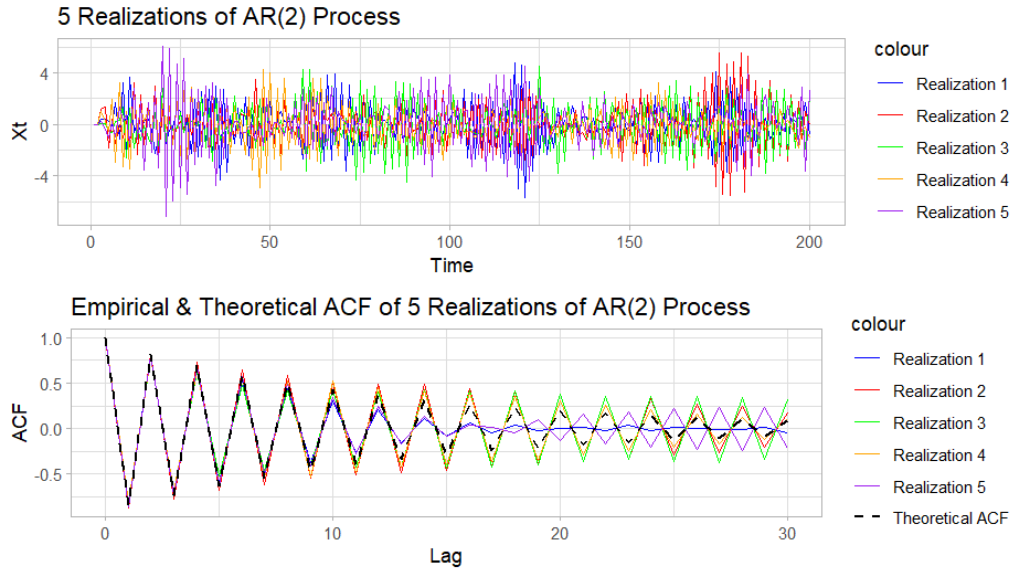

Figure 4: AR(2) process with n = 200 observations with $\phi_1 = 0.6$ and $\phi_2 = -0.3$

1.5 $\phi_1 = -0.7$ and $\phi_2 = -0.3$

An AR(2) process with n = 200 observations with $\phi_1 = -0.7$ and $\phi_2 = -0.3$ along with its empirical and theoretical ACF is presented in Figure 5. The time series plot shows that the simulations are non-stationary. As the theoretical ACF is derived under the assumption of the process being stationary, it does not hold under non-stationary conditions, i.e. the roots of the characteristic polynomial are inside or near the unit circle. This can be observed in the ACF plot.

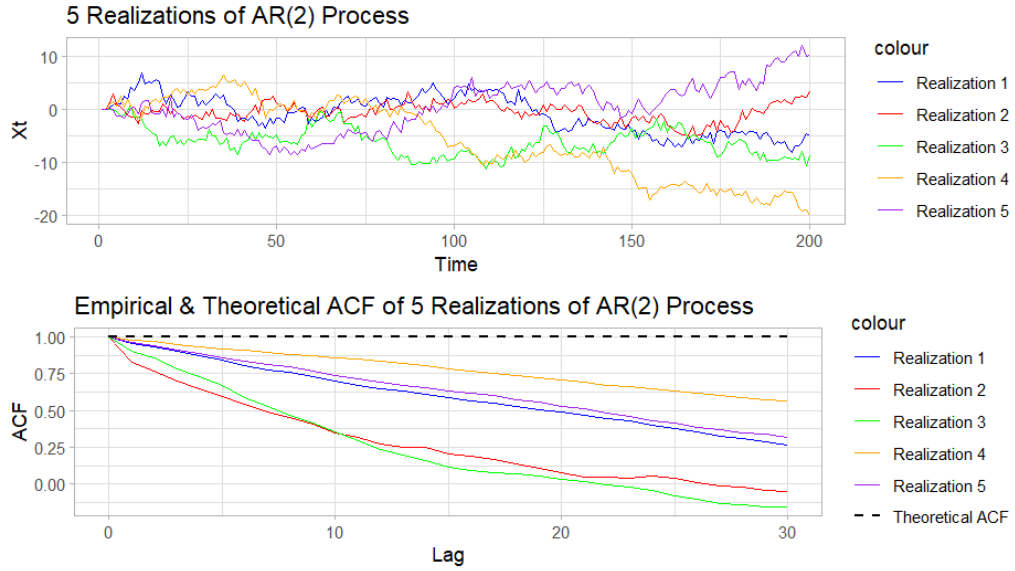

Figure 5: AR(2) process with n = 200 observations with $\phi_1 = -0.7$ and $\phi_2 = -0.3$

1.6 $\phi_1 = -0.75$ and $\phi_2 = -0.3$

An AR(2) process with n = 200 observations with $\phi_1 = -0.75$ and $\phi_2 = -0.3$ along with its empirical and theoretical ACF is presented in Figure 6. The time series plots shows an explosive

process. Similarly to 1.5, as the process is not stationary, the theoretical ACF and empirical ACF differ significantly.

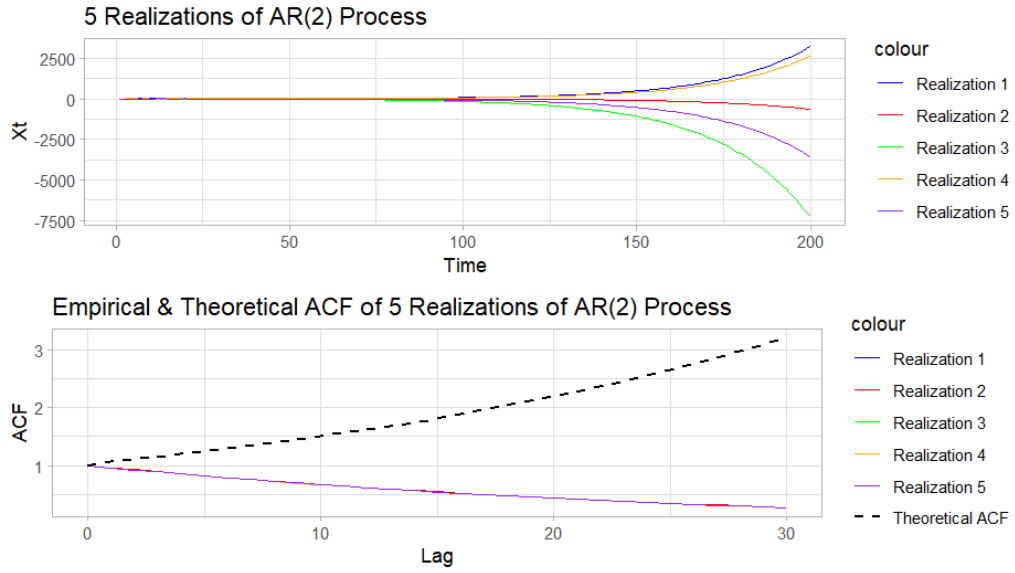

Figure 6: AR(2) process with n = 200 observations with $\phi_1 = -0.75$ and $\phi_2 = -0.3$

1.7 Recommendations: plot time series vs. ACF

Using both the time series plot and the ACF together offers a more comprehensive understanding of the data:

- **Time Series Plot:** Provides an intuitive overview of the data's behavior over time. It helps in identifying trends, level shifts, non-stationarity, or structural breaks that may not be visible in the ACF alone. For example, a non-stationary process might show strong autocorrelation, but without the time series plot, it's difficult to spot drift or variance changes.
- **ACF Plot:** Offers detailed insights into the dependence structure and helps in identifying potential ARMA processes.

While the ACF is critical for identifying model structure, relying on it alone can be misleading. Especially when the underlying data is non-stationary or has irregular patterns.

2 Predicting monthly solar power

2.1 Re-write model and model validation

The equation can be expanded using $X_t = log(Y_t) - \mu$ as follows:

$$(1 + \phi_1 B)(1 + \Phi_1 B^{12})X_t = \epsilon_t \tag{1}$$

$$X_{t+1} + \phi_1 X_t + \Phi_1 X_{t-11} + \phi_1 \Phi_1 X_{t-12} = \hat{\epsilon}_{t+1|t} \tag{2}$$

Figure 7 shows the residual time series as well as the ACF and PACF plots of residuals.

To check if $\{\epsilon_t\}$ is in fact white noise of a variance σ_ϵ^2, we analyze the residuals:

- Mean: $0.0152 \simeq 0$
- Sample variance: $0.0462 \simeq 0.22^2 = 0.0484$

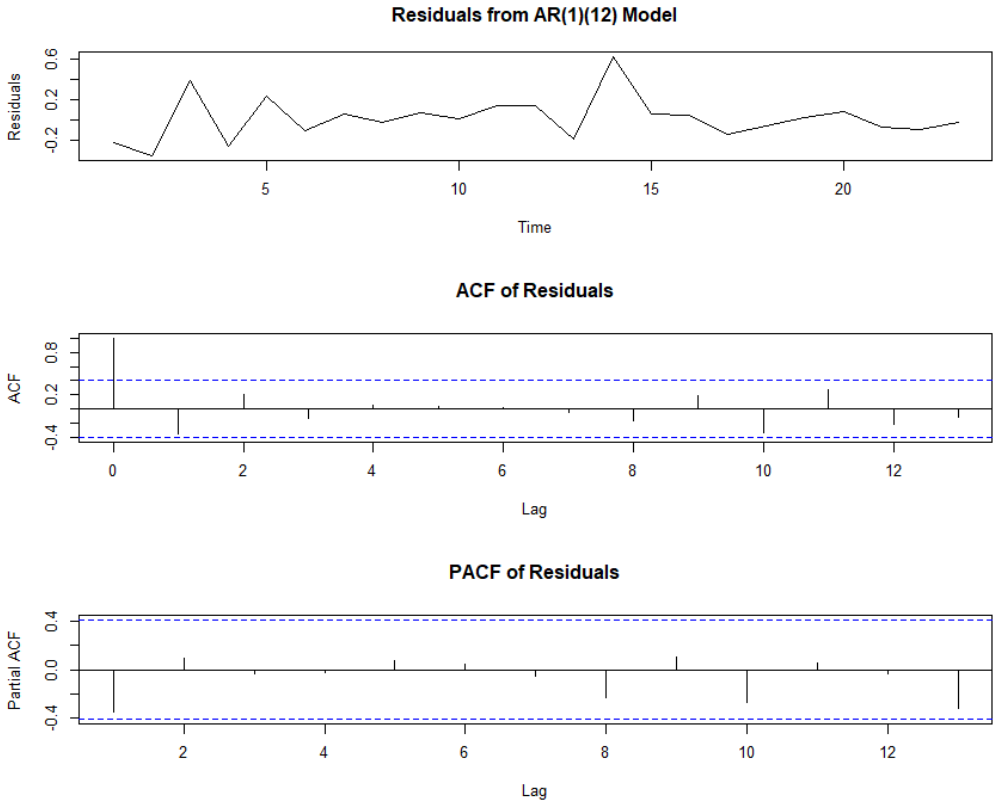

Figure 7: Residuals analysis.

The Ljung-Box test, which evaluates the null hypothesis that the residuals are uncorrelated, gives $p = 0.3817$. As $p > 0.05$, the null hypothesis is not rejected, thus, there is no significant evidence of autocorrelation. This can also be supported by the ACF and PACF plots of residuals: ACF plot show only one significant peak at lag 0 and PACF plot show no significant peaks. This suggests that no AR or MA processes are present in the residuals, hence the residuals are likely white noise.

2.2 Predict next 12 months

The formula used here is:

$$\hat{X}_{t+k|t} = \begin{cases} -\phi_1 \hat{X}_{t+k-1|t} - \Phi_1 X_{t+k-12} - \phi_1 \Phi_1 X_{t+k-13} & \text{for } k > 1 \\ -\phi_1 X_t - \Phi_1 X_{t-11} - \phi_1 \Phi_1 X_{t-12} & \text{for } k = 1 \end{cases} \tag{3}$$

$$\hat{Y}_{t+k|t} = exp(\hat{X}_{t+k|t} + \mu) \tag{4}$$

Table 1 and Figure 8 present the predictions of power generation for the next 12 months.

Table 1: Predictions for next 12 months.

Year	Prediction (in MWh)
2011 January	101.28
2011 February	147.23
2011 March	156.92
2011 April	142.55
2011 May	355.31
2011 June	538.02
2011 July	611.21
2011 August	631.89
2011 September	630.11
2011 October	508.16
2011 November	375.01
2011 December	280.49

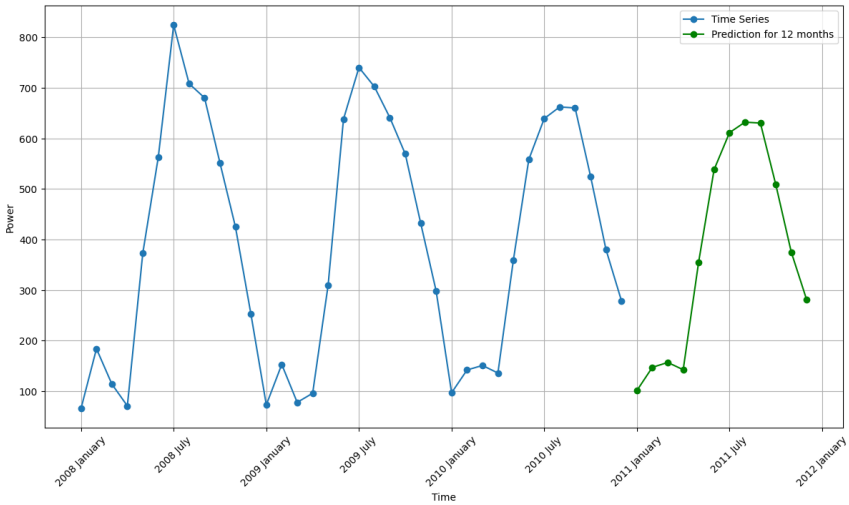

Figure 8: Prediction of the model for the next 12 months

2.3 Prediction intervals

In order to calculate the 95% prediction intervals, the following equations are used. Note that only the AR(1) component is considered, with no seasonality. Thus, first, the AR(1) component of the equation is;

$$(1 + \phi_1 B)X_t = \epsilon_t \tag{5}$$

$$X_t + \phi_1 X_{t-1} = \epsilon_t \tag{6}$$

$$\hat{X}_{t+k|t} = \begin{cases} -\phi_1 \hat{X}_{t+k-1|t} & \text{for } k > 1 \\ -\phi_1 X_t & \text{for } k = 1 \end{cases} \tag{7}$$

Next, Eq. (5.149)-(5.151) can be rewritten. First, (5.146) can be rewritten as:

$$\epsilon_t = X_t + \phi_1 X_{t-1} \tag{8}$$

Next, the forecast error derivation can be recursively solved;

$$X_{t+1} - \hat{X}_{t+1|t} = X_{t+1} + \phi_1 X_t = \epsilon_{t+1} \tag{9}$$
$$X_{t+2} - \hat{X}_{t+2|t} = \epsilon_{t+2} - \phi_1 X_{t-1} + \phi_1 \hat{X}_{t+1|t} \tag{10}$$
$$X_{t+2} - \hat{X}_{t+2|t} = \epsilon_{t+2} - \phi_1 (X_{t-1} - \hat{X}_{t+1|t}) \tag{11}$$
$$X_{t+2} - \hat{X}_{t+2|t} = \epsilon_{t+2} - \phi_1 \epsilon_{t+1} \tag{12}$$
$$\tag{13}$$

Therefore, it can recursively be proved;

$$X_{t+k} - \hat{X}_{t+k|t} = \sum_{i=0}^{k-1} (-\phi_1)^i \epsilon_{t+k-i} \tag{14}$$

Thus;

$$\sigma_k^2 = \sum_{i=0}^{k-1} (\phi_1)^{2i} \sigma_\epsilon^2 \tag{15}$$

$$\hat{X}_{t+k|t} \pm u_{\alpha/2}\sigma_k = \hat{X}_{t+k|t} \pm u_{\alpha/2}\sigma_\epsilon \sqrt{\sum_{i=0}^{k-1} (\phi_1)^{2i}} \tag{16}$$

$$\hat{Y}_{t+k|k}^{bound} = \begin{cases} \hat{Y}_{t+k|k}^{up} & = exp(\hat{X}_{t+k|t} + \mu + u_{\alpha/2}\sigma_k) \\ \hat{Y}_{t+k|k}^{low} & = exp(\hat{X}_{t+k|t} + \mu - u_{\alpha/2}\sigma_k) \end{cases} \tag{17}$$

The calculated 95% prediction intervals are added to the previously made predictions in Figure 9. It can be observed that the prediction uncertainty is higher in the summer spike.

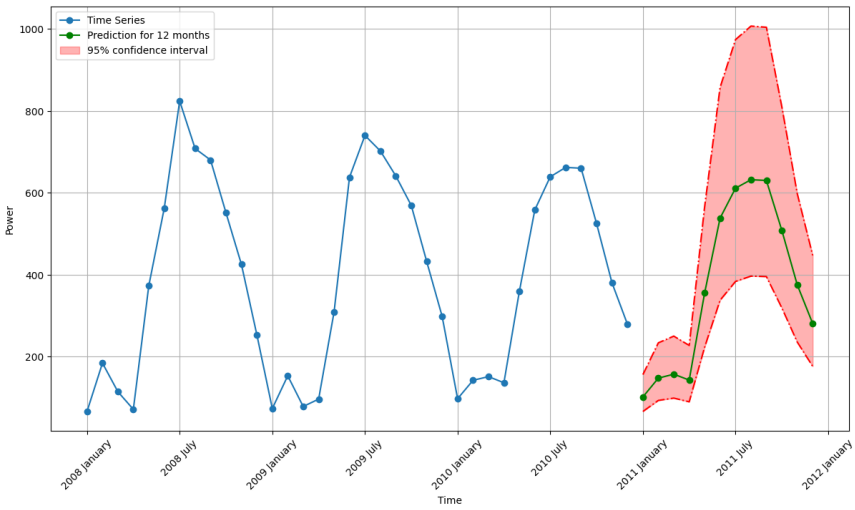

Figure 9: Prediction of the model over time, with confidence interval

2.4 Comments

The forecast from the model shows reasonable results, simulating the seasonal trend of the power generation. Additionally, the residuals analysis suggest the errors are roughly i.i.d, supporting the usability of the model.

The model follows the AR(1)(12) process which assumed to be stationary. However, solar energy

generation could be affected by external 'factors' such as environmental policies, climate change, system upgrades, maintenance, etc. which could potentially push the model to be non-stationary. As such, the forecast here is useful for planning and estimation of solar power generation, but would require continuous re-evaluation should there be any external 'factors' come into play.

Regarding the prediction intervals, these were computed assuming a fixed AR process and thus do not account for potential future structural changes in the data. Therefore, the actual uncertainty in long-term forecasts may be greater than indicated in Figure 9. Additionally, the interval width increases over time ; narrower during the early months of the forecast horizon and wider in the summer months. This is likely due to greater variability in power generation during the summer period, as observed in the training data. This behavior reflects both the accumulation of uncertainty in recursive forecasting and seasonal variation in the underlying process.

3 An ARX model for the heating of a box

3.1 Plot non-lagged time series: Ph, Tdelta, Gv

Figure 10 shows the nonlagged time series of 3 variables Ph, Tdelta, and Gv, from 30/01/2013 to 08/02/2013. Clear inverse patterns are visible between Ph and Gv; peaks in solar radiation correspond to noticeable drops in heating power, suggesting that increased solar input reduces the heating demand. Tdelta remains relatively stable over the observed period, with a potential gradual increase toward early February.

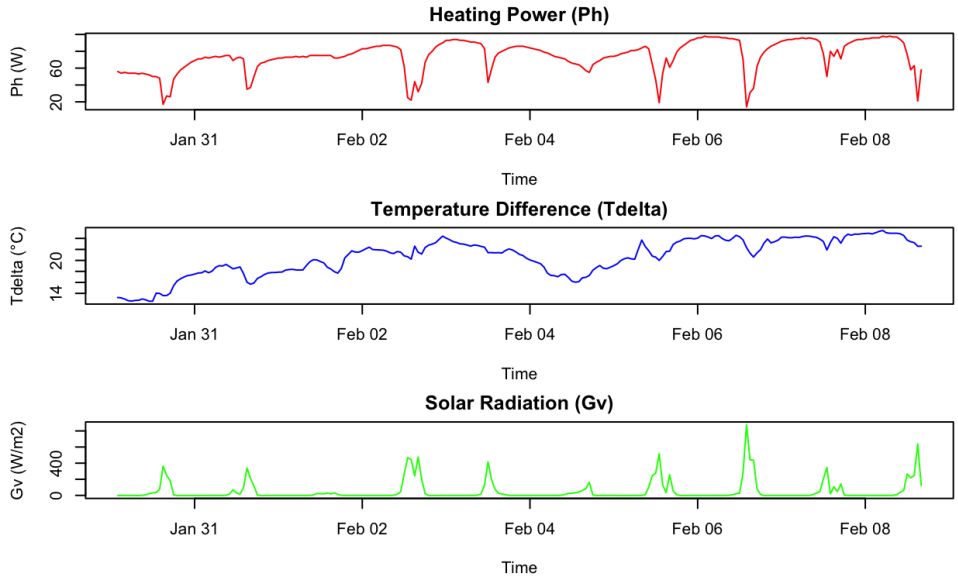

Figure 10: Non-lagged time series: Ph, Tdelta, Gv.

3.2 Split train and test data

The data was split into a model training and testing set, such that '2013-02-06 00:00' is the last data point in the train set (i.e. thour = 1, ..., 167 is the train set).

3.3 Investigate the variables and their relations

Figure 11 shows key visualizations for the data, exploring the dependencies between the output variable Ph and its potential explanatory variables Tdelta and Gv using scatter plots, autocorrelation functions (ACFs), and cross-correlation functions (CCFs).

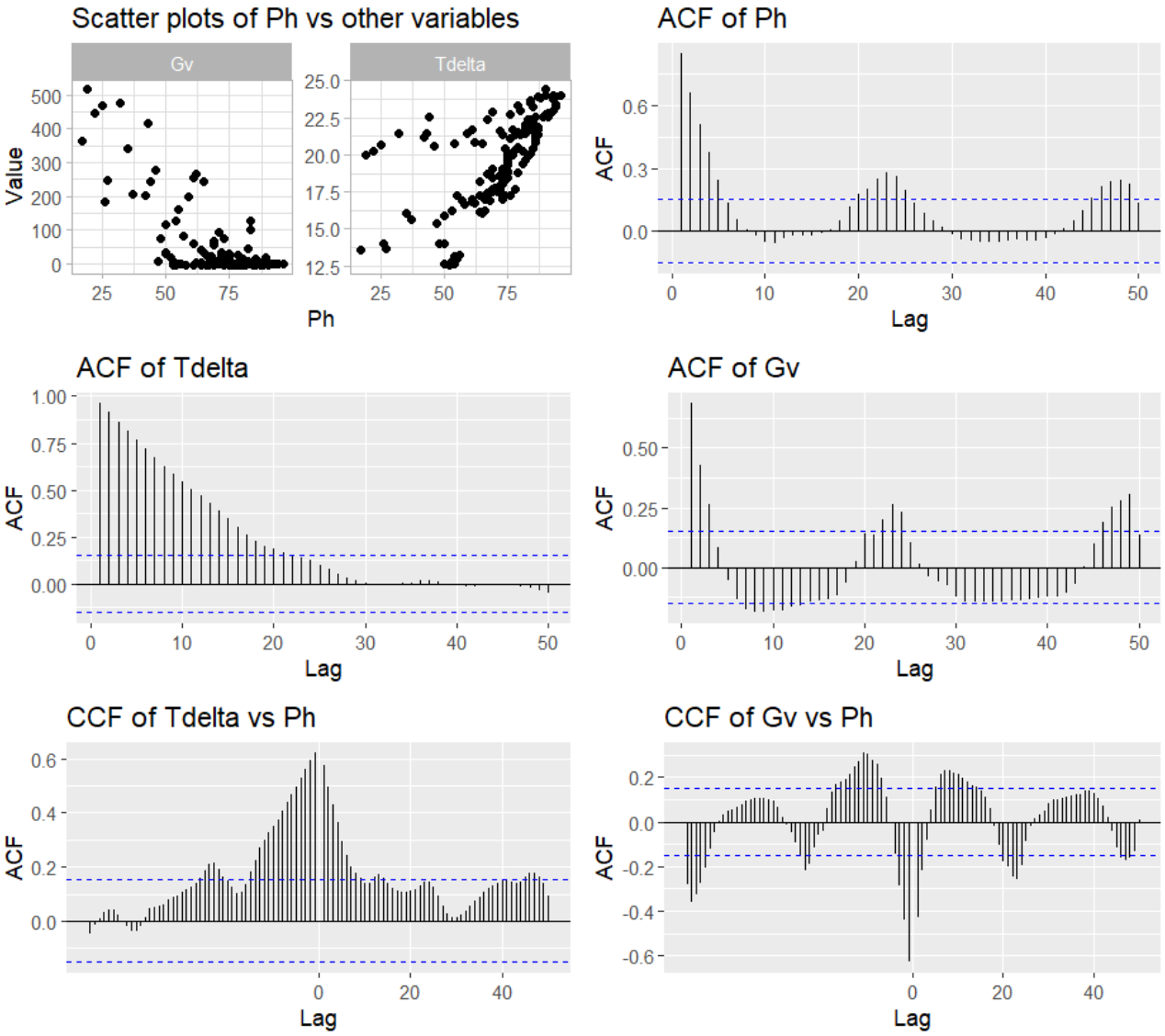

Figure 11: Relations between Ph and Tdelta, Gv.

Observations are as follows;

- The scatter plot shows a negative correlation between Ph and Gv, and a positive correlation can be observed for Ph and Tdelta. This is consistent with the physical intuition that greater temperature differences require more heating, while solar radiation reduces heating demand.
- ACFs for all three variables indicate strong serial correlation, particularly for Tdelta, which shows a very slow decay, suggesting persistent past influence. Ph and Gv show a clear daily periodicity with peaks around lag 24 and 48, reflecting the cycle of heating and solar input.
- CCF(Tdelta, Ph) shows strong positive correlation from lag -10 to about lag 10. This suggests that changes in Tdelta precede and help explain subsequent changes in Ph.
- CCF(Gv, Ph) shows a strong negative correlation at lag 0 and then positive lags on both sides to about lag 10, reinforcing the inverse relationship. This again supports using Gv as an exogenous regressor.

3.4 Estimate impulse response

Linear regression was used to estimate the impulse responses from Tdelta and Gv to the output Ph. This approach is justified by interpreting the system as an approximate finite impulse response model.

$$\text{Ph}_t = \sum_{k=0}^{K} \beta_k^{(T)} \cdot \text{Tdelta}_{t-k} + \sum_{k=0}^{K} \beta_k^{(G)} \cdot \text{Gv}_{t-k} + \varepsilon_t$$

Here, the regression coefficients β_k represent the response of Ph to a unit increase in the corresponding input at lag k, that is, the response to impulse. The Impulse response for both variables are shown in 12.

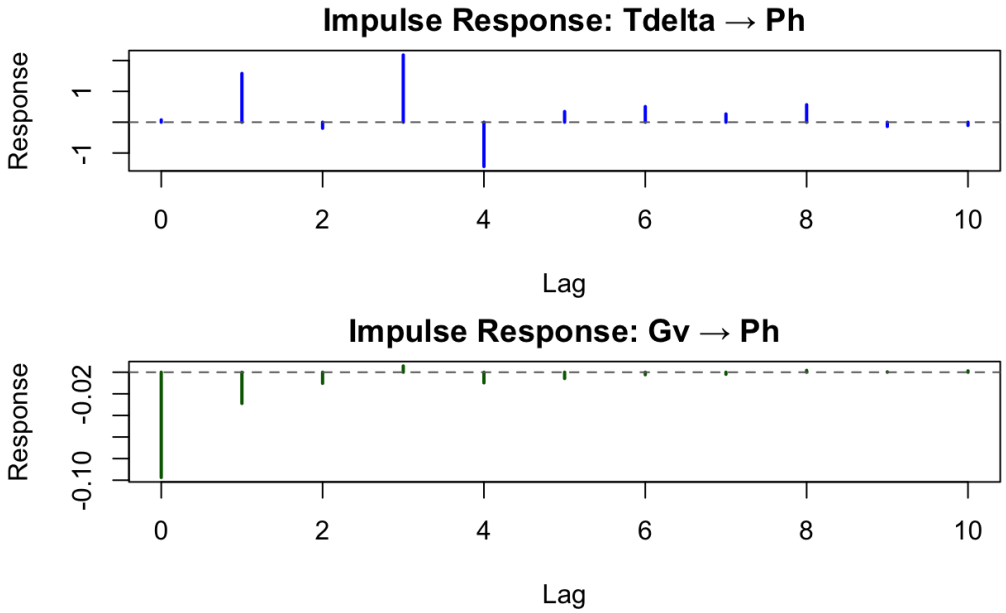

Figure 12: Impulse Response for Tdelta and Gv

Tdelta has a delayed and distributed positive effect on Ph, peaking at lag 3, followed by lag 1, 4, and 8. Therefore, a change in Tdelta today affects Ph up to 8 hours later, but especially after 3 hours. On the other hand, Gv shows a short-lived negative impact, primarily at lag 0.

3.5 Linear regression model

The simple linear model following the below equation was fitted to the data:

$$P_{h,t} = \omega_1 T_{\delta,t} + \omega_2 G_{v,t} + \epsilon_t$$

First, the model was fitted to the data to check its estimation results (top figure). Next, a recursive one-step prediction using this model, starting at 3 data points, was made (second figure). The estimation results of this one-step prediction method, as well as its residuals are analyzed using Figure 13.

- The estimation results, both for the general model and one-step-prediction, generally follow the true values of Ph, successfully capturing daily fluctuations. There is some deviation between the two estimations, especially at early points, but consistent towards the end. However, both tend to underestimate the sharp drops in Ph, especially during rapid changes, indicating that the model may struggle with extreme transitions.
- The residual plot shows that while the residuals fluctuate around 0, they are not purely random. Some structure seems to remain, especially around peak heating events. This suggests that additional dynamics might not be captured by the linear model.
- The ACF of residuals reveals significant autocorrelation at small lags. This indicates that the model fails to capture some temporal dependencies in Ph, and that an ARX model could improve performance.

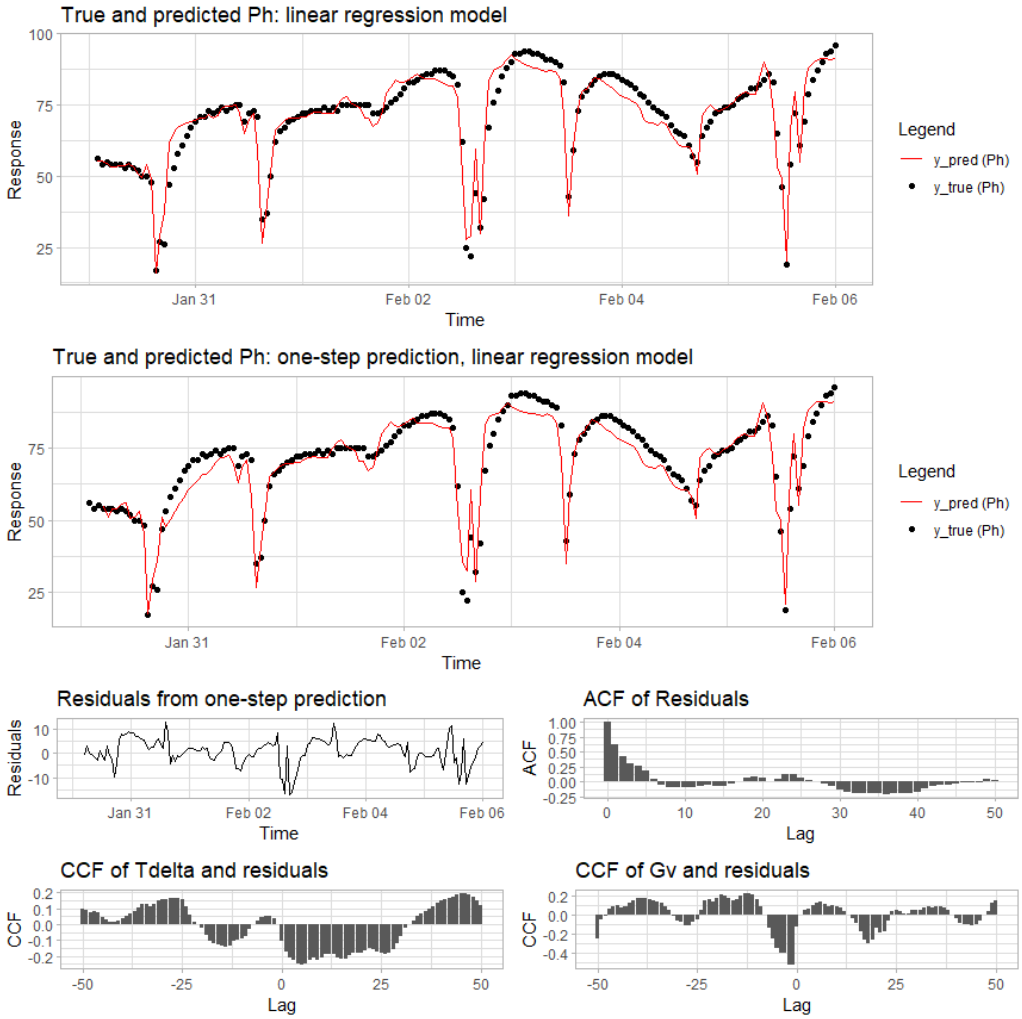

Figure 13: Linear regression model results.

- The CCFs between residuals and the input variables show significant non-zero correlations at multiple lags. This means that lagged values of Tdelta and Gv still contain information that was not fully extracted by the current model.

3.6 ARX first order

The ARX first order model following below equations was fitted to the data:

$$P_{h,t} = -\phi_1 P_{h,t-1} + \omega_1 T_{\delta,t} + \omega_2 G_{v,t} + \epsilon_t$$

Figure 14 shows the results using ARX first order model in a similar order to Figure 13. It can be observed that the model's predictions are closer to the true values, better capturing sudden fluctuations compared to the simple linear regression model. In addition, the model performance is clearly improved looking at the residuals: the residuals' variance is reduced, mostly resembling white noise with ACF having a significant spike at lag 0 only. The CCFs are more balanced and closer to zero across many lags, meaning the ARX model better accounted for the predictor's effects.

3.7 Increase model order

An ARX second order model, keeping the same number of lags for all inputs, follows the below equation. In this section, residuals were not computed using one-step-ahead predictions. Instead, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were calculated based

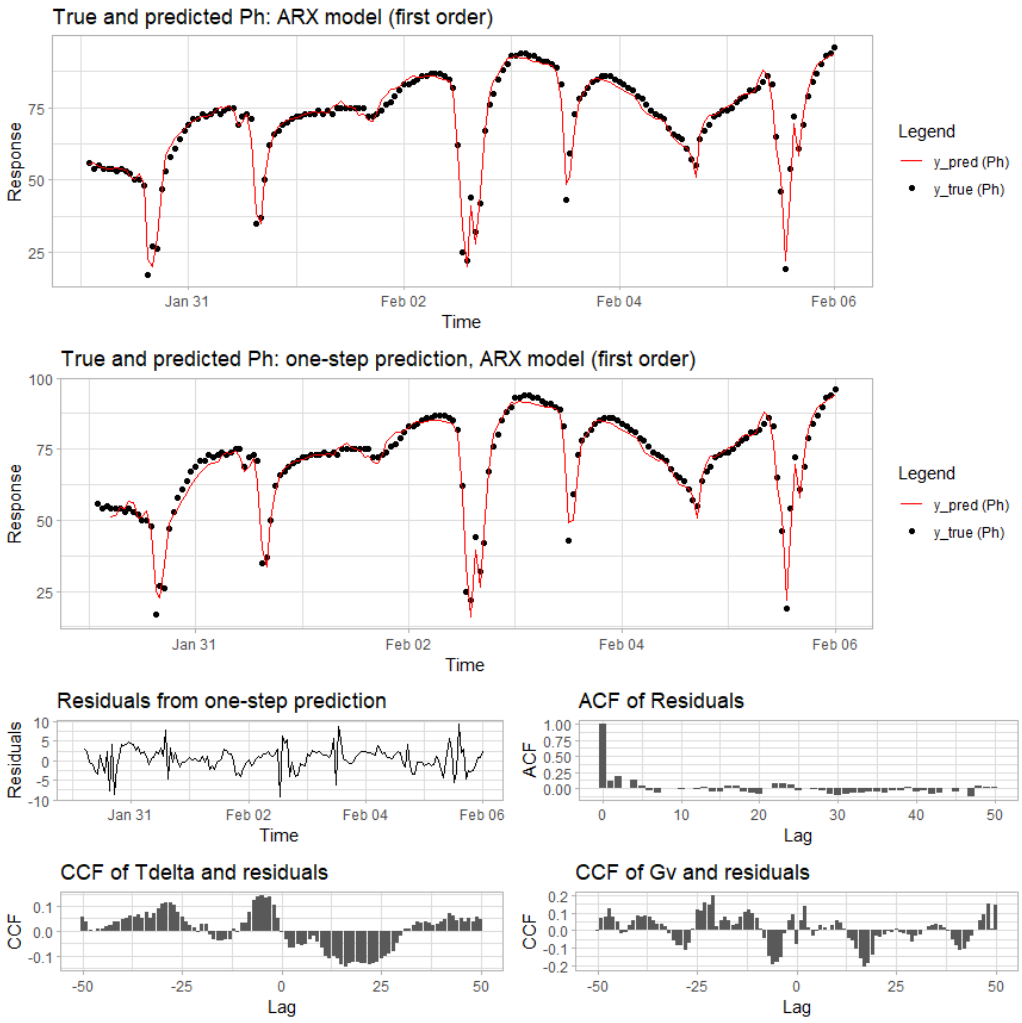

Figure 14: ARX model results.

on the model's overall fit to the entire training dataset.

$$P_{h,t} = -\phi_1 P_{h,t-1} - \phi_2 P_{h,t-2} + \omega_{1,0} T_{\delta,t} + \omega_{1,1} T_{\delta,t-1} + \omega_{2,0} G_{v,t} + \omega_{2,1} G_{v,t-1} + \epsilon_t$$

For a regression model, AIC and BIC is calculated using the below equations:

$$\text{AIC} = 2k + n \ln\left(\frac{\text{RSS}}{n}\right)$$

$$\text{BIC} = k \ln(n) + n \ln\left(\frac{\text{RSS}}{n}\right)$$

Where:

- RSS is the sum of squared residuals.
- n is the number of observations in training dataset.
- k is the number of model parameters which changes depending on the ARX model order.

AIC and BIC are both criteria used for model selection in statistics. The AIC aims to find the model that best explains the data with the fewest parameters. However, it does not penalize complexity as strongly as BIC where BIC takes into account the number of parameters (k) and the number of observations (n). Thus, BIC is more conservative than AIC in terms of model complexity.

Figure 15 shows the comparison between AIC and BIC for different ARX model order from 1 to 10. The two curves indicate that the model performance increases with model complexity until a certain point, after which it decreases again.

AIC is lowest at model order of 6 (lag 6) while BIC has a sharp decrease from model order of 1 to 2 then slowly reaches a plateau point around model of 4-6. This also corresponds to the ACF plot of residuals in Figure 13 where there appear insignificant spikes starting at lag 4 through 6. Based on strict criterion of choosing model order from minimum AIC and/or BIC values, a model order of 6 would be best. However, a lower model order may also be considered, if prioritizing simplicity. Finally, note that this score is only calculated based on the training set.

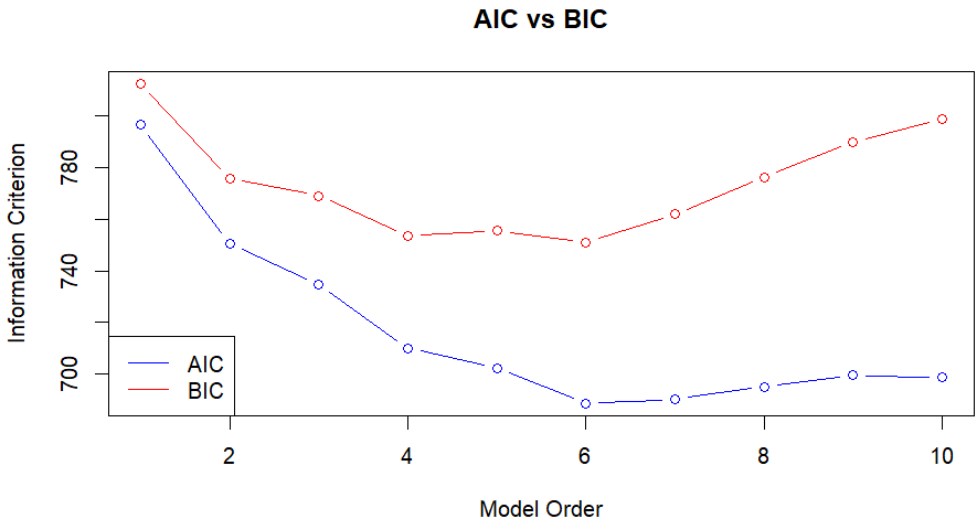

Figure 15: Comparison of AIC and BIC for each model order.

3.8 One-step for test period and calculate the RMSE

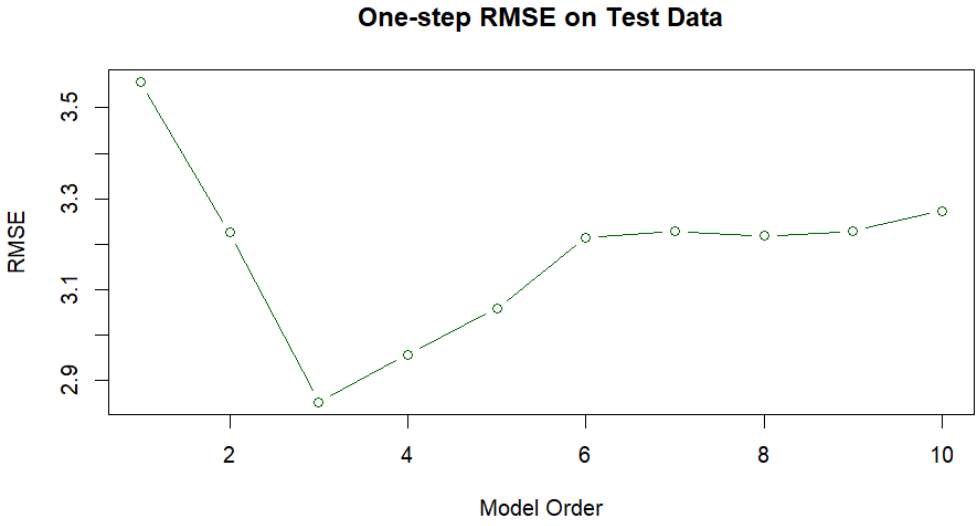

Figure 16: RMSE from test set for each ARX model order.

Figure 16 shows the model performance using one-step predictions evaluated using root mean square error (RMSE) for each model order on the test set. Figure 17 shows the one-step predictions

and residuals for each model order on the test set. Based on RMSE, the model performs best at model order of 3 on the test set, which differs from the analysis on the training set in 3.7.

This illustrates a trade-off between in-sample model selection criteria (AIC/BIC) and out-of-sample predictive performance (RMSE). While AIC and BIC suggest higher model orders, these do not generalize well, leading to worse RMSE on the test set. A simpler model (order 3) achieves better one-step-ahead prediction accuracy, highlighting the importance of evaluating models based on forecasting performance rather than fit alone.

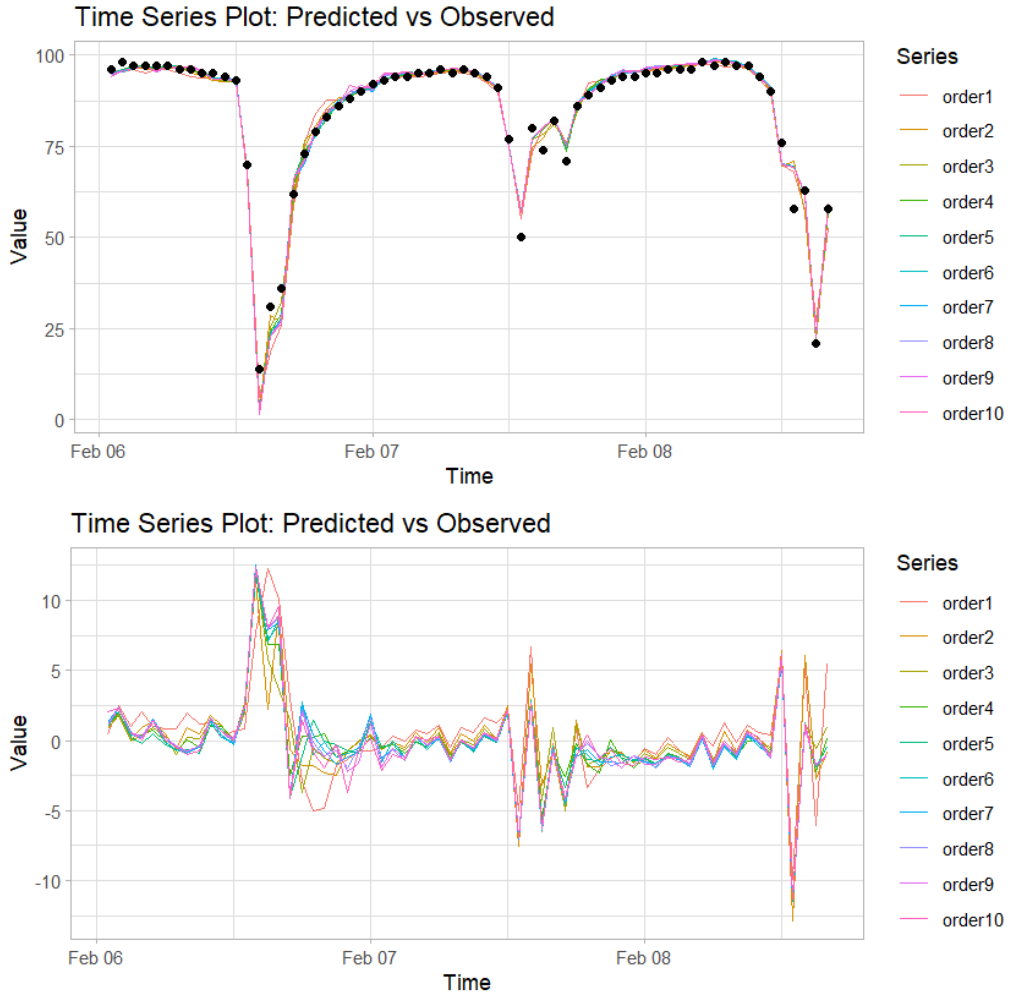

Figure 17: Predictions and residuals on test set for each ARX model order.

3.9 Multi-step predictions

Based on the predictability discussed in 3.8, a model order of 3 with the following equation is used for further analysis:

$$P_{h,t} = -\phi_1 P_{h,t-1} - \phi_2 P_{h,t-2} - \phi_3 P_{h,t-3} + \omega_{1,0} T_{\delta,t} + \omega_{1,1} T_{\delta,t-1} + \omega_{1,2} T_{\delta,t-2} + \omega_{2,0} G_{v,t} + \omega_{2,1} G_{v,t-1} + \omega_{2,2} G_{v,t-2} + \epsilon_t$$

A multi-step prediction (simulation) throughout the entire period, i.e. from beginning of the training period to the end of the test period with the most suited model order, using the observed time series of the inputs (Tdelta and Gv), and the AR lags calculated iteratively is made. This is understood as assuming there are no true values of Ph available and the model trains based observed Tdelta, Gv and predicted Ph iteratively. The multi-step prediction tests how the model performs in a real-world setting where true future outputs are not available. Thus, the first few predicted Ph would have to be initialized (in this case, assuming to be 0) and considered as an 'initial warm-up' period.

Figure 18 shows the results for multi-step prediction using model order of 3. As expected, the

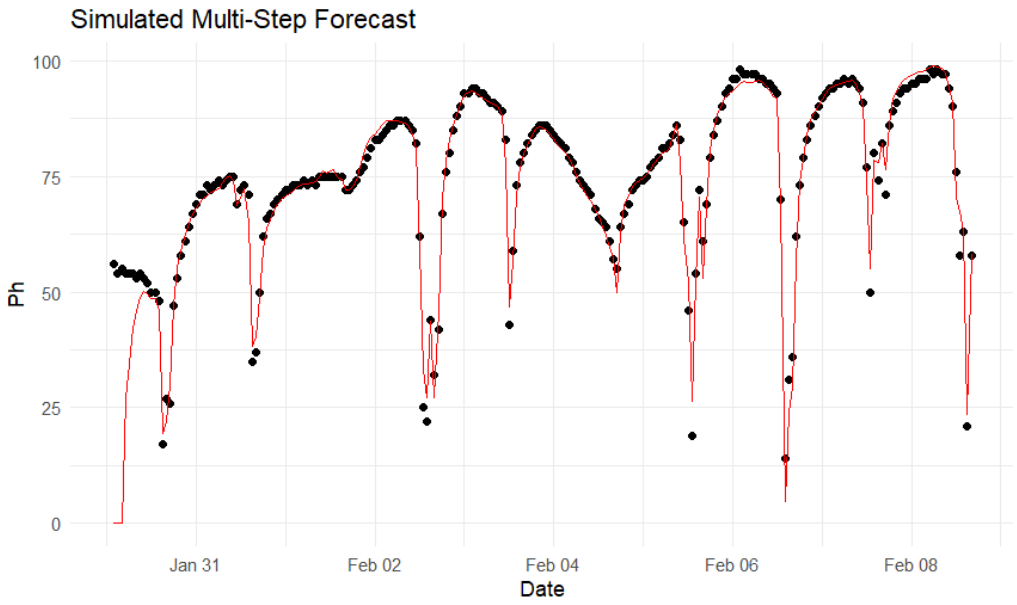

Figure 18: Multi-step prediction with model order of 3, prediction as red line and observations as black dots.

predictions for Ph for the first few time steps are off and should be considered as 'initial warm-up' period. Otherwise, the model could reasonably capture the trend/events in the observed Ph, indicating that the model could be used to predict heating in operational real-time setting such that given continuous measurement for Tdelta and Gv, Ph can be predicted and adjusted.

3.10 Conclusions

The dynamic relationships between heating (Ph), internal/external temperature difference (Tdelta), and solar radiation (Gv) was explored. Through time series plots, cross-correlation analysis, and impulse response estimation, it was confirmed that Tdelta exerts a delayed positive influence on heating demand, while Gv has a short-lived negative effect.

Among the tested models, ARX models of increasing order were compared using both in-sample criteria (AIC/BIC) and out-of-sample predictive performance (test set RMSE). While higher-order models (e.g., ARX(6)) achieved better AIC/BIC scores, they did not generalize as well to unseen data. This illustrates a trade-off; increased complexity may improve in-sample fit but at the cost of robustness and generalization. In contrast, the ARX(3) model provided the most reliable balance, achieving strong one-step predictive accuracy while avoiding overfitting. The ARX(3) model captures both autoregressive dynamics and the influence of external variables using relatively few parameters, making it well-suited for real-time applications as well.

However, this modeling approach also has limitations. It assumes a linear and stationary system, and depends heavily on the accuracy and availability of Tdelta and Gv. Any structural changes in the environment, such as shifts in building usage or sensor faults, could reduce forecast reliability.

Despite these assumptions, the model performs well under realistic conditions. The multi-step simulation results suggest that ARX(3) can be used in real-world scenarios such as smart heating systems, where temperature and sunlight inputs are continuously monitored. With minimal lag structure and solid predictive performance, the model shows promise for deployment in energy-efficient building control systems.