Solar power forecasts

Spatio-temporal solar power forecasts via regression

by

N. Jaspers

to obtain the degree of Master of Science

at the Delft University of Technology (MSc Sustainable Energy Technology).

Student number: 4629019

Project duration: June 30, 2020 – August 18, 2022

Thesis committee:

Dr. ir. R.A. Verzijlbergh,	TU Delft,	supervisor
Dr. ir. S.R. de Roode,	TU Delft,	supervisor
Prof. dr. ir. Z. Lukszo,	TU Delft,	supervisor

TU Delft

Noud Jaspers

Spatio-temporal solar power forecasts via regression

To my family
and friends.

Preface

What lies here before you is a thesis about spatio-temporal solar power forecasts via regression. This project aimed to build a robust framework for scalable and accurate intraday solar power forecasts using irradiance observations and numerical weather predictions. It was written to fulfil the graduation requirements of the TU Delft's master of Sustainable Energy Technology.

The motivation for this thesis came from my fascination with energy markets and the great opportunity for optimization in them. More and more solar energy is entering the grid, but the dependence on the weather causes mismatches between supply and demand. To solve this problem, we need better forecasts and increased flexibility. This project aims to tackle the former. Remco and Stephan supplied their knowledge and helped to set the direction for this project.

This thesis was written for those that seek a way to implement accurate intraday solar power forecasts in a scalable manner. Through the application of the framework as described in this thesis, it is possible to beat the average market player at the time of writing.

If your aim is to quickly grasp the conclusion of this thesis, then I would like to refer you to the Executive Summary. However, if you aim to implement the framework – or if you just really enjoy the maths behind it – then I would recommend you to read the entire thesis.

On to the acknowledgements on the page after next!

Noud Jaspers
August 2022

Acknowledgements

For my second thesis, it has been quite the journey! I could not have finished this work without the help of some people that I would like to thank here.

I want to thank Remco for guiding me through my thesis for the past two years – I have a clear direction of where I want to go. In addition, I thank Stephan for his enthusiasm, feedback, and support.

I would like to extend my thanks to Linda Kamp for her interest. At some point I disengaged with my thesis, and that interest put me back on track.

I thank my family and Heleen for their support and patience – it has paved my path forward.

To conclude, I thank Zofia, Remco, and Stephan for providing me with the opportunity to defend my thesis. I look forward to the conclusion of this part of my life!

Noud Jaspers
August 2022

Executive summary

Introduction Across the globe, we are facing the challenge of global warming due to the emissions of greenhouse gases (GHGs). This leads to extreme weather, sea level rise, biodiversity loss, and higher death rates. Most of the GHGs come from burning coal or other types of polluting energy sources. Therefore, we need to make the switch to renewable energy (RE).

Solar energy is an RE and available in abundance. The earth receives about 10,000 times more energy from the sun than the rate at which humankind consumes energy. We are able to harvest that energy as well by using photovoltaic (PV) cells. However, the switch to renewable energy entails a paradigm shift.

In the energy systems of today, we burn fuel to follow demand. However, when we are dependent on solar energy, then we have to follow supply instead of demand – the irradiance from the sun fluctuates heavily during the day and there is no sun at night. Therefore, we need to be able to forecast supply and plan demand accordingly. If we do not, then it becomes very difficult to impossible to maintain a stable energy system.

Within that context, this thesis aims to build a regression framework for intraday solar power forecasts that uses numerical weather predictions and observations. To develop this framework, we do a case study of the Netherlands. The goal is to develop an approach that can be used around the globe without the need for new (expensive) measurement infrastructure. It should be reliable, easy to implement, and give state-of-the-art results. We aim to answer the research question: How can irradiance observations and numerical weather predictions be regressed on the spatio-temporal dimension to create accurate intraday solar power forecasts?

Theory Multiple linear regression is a statistical method that allows us to relate a set of predictors, $\mathbf{X}$, to a response, $\mathbf{y}$. First, let us assess $\mathbf{X}$. This matrix has P predictors over the columns and T time steps over the rows. Each column is a vector, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P\}$, and each vector has T elements, $\mathbf{x}_i = \{x_1, x_2, ..., x_T\}$, where i is the index of a predictor.

Second, the response, $\mathbf{y}$, is a vector with T time steps, $\mathbf{y} = \{y_1, y_2, ..., y_T\}$. In regression, we aim to predict $\mathbf{y}$ with $\mathbf{X}$ as $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$, where $\hat{\mathbf{y}}$ is the predicted response and $\mathbf{b}$ is the vector of coefficients with P elements. The vector of coefficients is found by minimizing the squared error,

$$Q = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2.$$

For solar power forecasts, we want to use numerical irradiance predictions and irradiance observations as an input to the regression model. Therefore, if we want to predict y at time step $t + \tau$, then we can build a predictor matrix that contains the last known observation, y_t, and the numerical irradiance prediction for time step $t + \tau$, $x_{t+\tau}$. In that case, our $\mathbf{X}$ has two columns to predict each time step: the lagged observation and the numerical irradiance prediction.

We can expand our regression model if we include more last known observations – also known as lags. For example, if we have two lags, then we include y_t and y_{t-1}. In addition, we can smooth the numerical irradiance predictions over the temporal dimension by including $x_{t+\tau-1}$, $x_{t+\tau}$, and $x_{t+\tau+1}$.

If we have two locations that we aim to predict with each their own predictor matrix, $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively, which contain that location's lagged observations and numerical predictions, then we can concatenate the predictor matrices for both locations, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, to predict each response. By doing so, we acknowledge that there are interdependencies between the two responses. For solar power forecasts, we can imagine that locations in proximity exhibit this form of interdependency. Therefore, we can apply regression over the spatio-temporal dimension.

When we concatenate many predictor matrices, then the amount of predictors for one response becomes very high. To ensure that we only take the predictors that are important, we apply predictor selection. This is done via the introduction of a constraint on the coefficient estimation process, such that the absolute sum of the coefficient vector does not exceed a certain threshold. This results in some elements in the coefficient vector to take a value of zero, therefore negating those predictors from the model. We call this process the least absolute shrinkage and selection operator (LASSO).

To conclude, we can evaluate our regression model's accuracy by calculating the root mean square error (RMSE),

$$\epsilon_{\text{RMSE}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2}.$$

In addition, we can calculate its accuracy against a baseline forecast,

$$\varepsilon = 1 - \frac{\epsilon_{\text{RMSE}_{\text{for}}}}{\epsilon_{\text{RMSE}_{\text{ref}}}},$$

where $\epsilon_{\text{RMSE}_{\text{for}}}$ is the RMSE of our forecast and $\epsilon_{\text{RMSE}_{\text{ref}}}$ is the RMSE of the baseline forecast. We call this metric the skill score (SS), which we denote as our forecast's SS against the baseline forecast. When we express accuracy as

a percentage, then we are referring to the SS.

Experimental setup To do the spatio-temporal regression, we first collected numerical weather predictions and irradiance observations from the ECMWF and KNMI respectively. We did so for all the KNMI's weather stations that measure irradiance (Figure 3.1), which are 32 locations across the Netherlands. We extracted 10 weather parameters from the ECMWF and only one – the global horizontal irradiance (GHI) – from the KNMI. The ECMWF's data contained parameters for irradiance, pressure, wind, temperature, cloud cover, and so forth.

The GHI data from the ECMWF and KNMI was processed to take out its diurnal pattern so that we are left with the pure stochastic process – we call the outcome of this process the clear-sky index. From there, the time steps where there was no sunshine were removed from the datasets. Finally, we calculated some dummy variables and transformed some parameters to help increase the fit between the numerical predictions and observations.

We concluded our experimental setup by defining a simple framework for model training, testing, and validation. That is, we calculate the coefficients, $\mathbf{b}$, of the model on data that spans the period of 2019-09-01 until 2020-08-31, which we call the train set. We tune the model's parameters, such as the amount of lags to include, by calculating the predicted response using the coefficients of the train set on the period of 2020-09-01 until 2020-12-31, which we call the test set. We use this predicted response to calculate the accuracy in terms of RMSE, which we try to minimize as much as possible. Once we have tuned the model's parameters, we validate the model's accuracy on the validation set, which spans the period of 2019-07-01 until 2019-08-31.

Regression framework To develop our regression framework for accurate intraday solar power forecasts over the spatio-temporal dimension, we first defined a framework to increase the fit between the numerical irradiance predictions and the irradiance observations. We used the clear-sky index to do so, which we also use as an input to the spatio-temporal model. This process of correction is called model output statistics (MOS).

The MOS correction uses weather parameters from the ECMWF in combination with some dummy and polynomial predictors to predict the observed clear-sky index in a regression model. We built a separate regression model for each location. Therefore, we do not consider the spatial component – we only use parameters that belong to the location under study. We determine the coefficients per station for the MOS correction on the train set, and we apply them to the test and validation set to produce MOS corrected predictions.

The spatio-temporal model takes in clear-sky indexes of observations and MOS corrected predictions. First, for the observations, we choose how many lags we would like to include, which we denote by p. Second, for the MOS corrected predictions, we choose how many points we would like to smooth on two sides, which we denote by q. Finally, we denote the lead time by τ –

that is, we have observations up to point t and we want to predict for $t + \tau$. We concatenate the predictor matrices for all stations together to predict each station, which therefore incorporates the spatio-temporal dimension. We determine the optimal value for p, q, and other model parameters by using the training, testing, and validation framework.

Results The MOS correction proved to have an SS against to the numerical irradiance predictions of 8% for the test set and 4% for the validation set. When we assess the SS per station for the test and validation set, then we find a weak relation between the two (Figure 5.2). As the test set spans a period of mostly winter, whereas the validation set a period of mostly summer, we find the predicted clear-sky index to be lower on average for the test set than the validation set. Therefore, we find that the MOS correction is more effective when the predicted clear-sky index is low.

For the spatio-temporal model with a lead time of one hour ($\tau = 1$), we tested for different values of lag, p, and smoothing, q, the accuracy in terms of SS against the numerical irradiance predictions. We first ran the model without the spatial component, which we call the temporal model. This model had the best accuracy for a lag of one, $p = 1$, and a smoothing of one, $q = 1$, with an SS of around 25% (Figure 5.3). From there, we added the spatial component, and we found a lag of one and smoothing of one to be most effective as well with an SS around 30% (Figure 5.4). Therefore, we decided to set $p = 1$ and $q = 1$ in further analyses.

When we assess the spatio-temporal model for different lead times up to six hours (Figure 5.5), then we find it to be effective up to a lead time of five hours. Beyond five hours, we find that the MOS corrections are most effective, as the spatio-temporal model identifies patterns that do not exist. In addition, we find the weights of the lagged observations to decrease compared to the MOS corrections as the lead-time increases.

When we analyse the coefficient vector, $\mathbf{b}$, for the spatio-temporal model with a lead time of one hour, then we find stations close to one-another to share high predictive importance (Figure 5.7). Thus, there is a relation between a station's distance and its importance to predict another station. In addition, when we analyse the total importance per station as a predictor for others (Figure 5.8), then we find a pattern where irradiance travels from the south-west across the rest of the Netherlands.

To conclude, we validated the spatio-temporal model with a lead time of one hour. First, we calculated the SS against the MOS corrected predictions, and we find the SS to be 24% for the test and validation set (Figure 5.9(a)). However, we do not find any relation between the SS per station for the test and validation set. This could be due to different weather patterns being at play during the test set – which mostly spans a period of winter – and the validation set – which mostly spans a period of summer. Second, when we calculate the SS against the numerical irradiance predictions (Figure 5.9(b)), then we find the pattern of the MOS corrected predictions coming through

as it serves as an input to the spatio-temporal model. We start to see a very weak relation between the SS per station for the test and validation set, and we observe that the test set outperforms the validation set with an SS of 30% and 25% respectively.

Discussion To determine the accuracy of the spatio-temporal model under different cloud conditions, we conducted a case study on a daily basis for a location in the centre of the Netherlands. We find the spatio-temporal model to perform well when the MOS corrected predictions are uncertain about the thickness and/or the variability of the clouds. In addition, the spatio-temporal model corrects for clear-sky conditions when the MOS corrected predictions are off due to a misestimation of aerosols and/or clouds. Therefore, we find the spatio-temporal model to work well in highly dynamic cloud systems.

To test the accuracy of our spatio-temporal model against state-of-the-art intraday solar power forecast models, we tried to find a case study of the Netherlands for comparison. We did not find such a case study, which confirms that we fill a gap by doing one. Nevertheless, we found comparable models with error metrics that we can validate our spatio-temporal model against. Through such validation, we believe our model to be accurate.

Conclusion We state our main research question: How can irradiance observations and numerical weather predictions be regressed on the spatio-temporal dimension to create accurate intraday solar power forecasts?

As an answer, we propose a five-step approach:

1. collect (numerical) weather predictions and irradiance observations over the spatial (about 50 kilometres apart) and temporal (15 minute to hourly) dimensions;
2. remove the sun's pattern from the numerical irradiance predictions and irradiance observations by calculating the clear-sky index;
3. fit the predictions to the observations via MOS;
4. create a spatio-temporal regression model that takes lagged observations and that smoothes MOS corrected predictions; and
5. apply LASSO for predictor selection.

The model that comes out has an SS between 25% and 30% against numerical irradiance predictions for a lead-time of one hour.

The reason for this model to work as well as it does, is that it extracts the information about the state of the atmosphere from lagged observations, which it applies to the trend of the smoothed MOS corrected predictions. By combining these two sources of information, we create accurate intraday solar power forecasts up to five hours ahead. When we want to go beyond five hours, the MOS corrected predictions prove to be more accurate.

To conclude, our case study of the Netherlands shows that accurate intraday solar power forecasts in highly dynamic cloud systems can be made by

using lagged observations and smoothed MOS corrections in a spatio-temporal context.

Contents

List of Figures

List of Tables

List of Abbreviations

ECMWF European Centre for Medium-Range Weather Forecasts
GHG Greenhouse gasses
GHI Global horizontal irradiance
KNMI Koninklijk Nederlands Meteorologisch Instituut
LASSO Least absolute shrinkage and selection operator
MAE Mean absolute error
MBE Mean bias error
MOS Model output statistics
NN Neural network
NWP Numerical weather prediction
PV Photovoltaic
RE Renewable energy
RMSE Root mean square error
SS Skill score

Part I

Solar power forecasts within a day

1

Introduction

CONTENTS

KEY TAKEAWAYS

Our current energy system is dependent on the burning of fossil fuels, which leads to pollution. Therefore, we have to shift to renewable energy – e.g., solar energy. However, solar energy is not always available, such as at night. Therefore, there is a mismatch between supply and demand.

Solar power forecasts are a way of dealing with the mismatch in supply and demand: by predicting supply, we can plan demand accordingly. However, considering that solar power is highly variable and dependent on the weather, we need high quality solar power forecasts during the day (intraday).

By combining numerical weather predictions with observations over the spatio-temporal dimension, we aim to build a framework for having accurate intraday solar power forecasts. We conduct a case study of the Netherlands.

1.1 Background: solar power is clean but unpredictable

Since the mid-20th century we have observed warming of the earth (Masson-Delmotte et al., 2019). Temperature rise has already resulted in profound changes to our ecological systems, such as:

- increased droughts, floods, and other types of extreme weather;
- sea level rise;
- biodiversity loss; and
- higher death rates.

The root cause of global warming is the emission of greenhouse gases (GHGs) due to human activities, which continue to increase, year after year (IPCC, 2015).

1.1.1 We need to change our energy habits

Consumption of fossil fuels for energy production accounts for 56.6% of all GHG emissions. Considering that all societies require energy services to meet basic human needs (e.g., lighting, cooking, space comfort, mobility, communication) and to drive production processes, energy must be provided with low GHG emissions to mitigate global warming (IPCC, 2012). However, 85% of the world's primary energy supply comes from the combustion of fossil fuels – consumption has been increasing steadily since the Industrial Revolution[1] as Figure 1.1 illustrates.

[1]The Industrial Revolution marked the introduction of machines for labour from 1760 to sometime between 1820 and 1840.

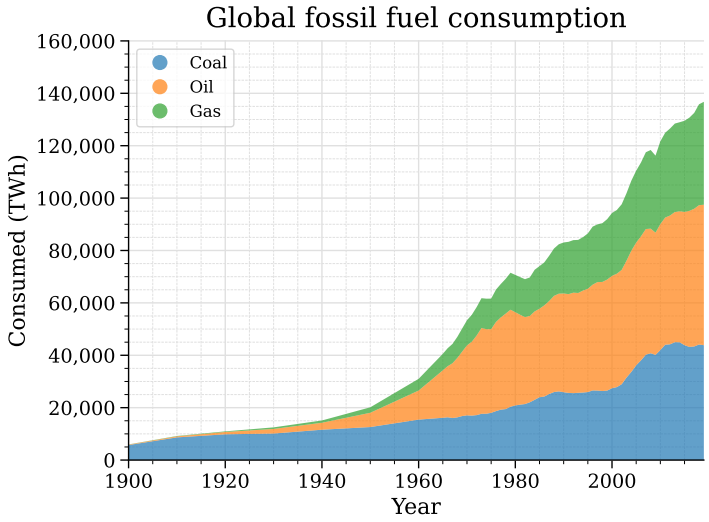


FIGURE 1.1
The trend of fossil fuel consumption since the Industrial Revolution expressed in TWh per year. Data from BP (2020) and Smil (2016).

To have sustainable economic development, energy must be provided without GHG emissions. Renewable energy (RE) serves this purpose (IPCC, 2012). Various types of RE can supply electricity, thermal energy and mechanical energy, as well as produce fuels. RE is any form of energy from solar, geophysical or biological sources that is replenished by natural processes at a rate that equals or exceeds its rate of use. Unlike fossil fuels, most forms of RE produce little to no GHGs.

1.1.2 The abundance of solar energy

In particular, solar energy is the most abundant of all energy resources – the rate at which solar energy is intercepted by the Earth is about 10,000 times greater than the rate at which humankind consumes energy (IPCC, 2012). Solar energy's potential to mitigate global warming is equally impressive. Except for the modest amount of GHG emissions produced in the manufacture of energy conversion devices, solar energy has the potential to displace large quantities of non-renewable fuels. It appears in the form of wind, wave, ocean thermal, hydropower and excess biomass energies.

According to BNEF (2020), EIA (2020), and IEA (2019), renewables are expected to become the global primary energy source by 2050, displacing fossil fuels. The transition to such a system poses numerous challenges when matching supply and demand – the sun only shines during the day but power

is also needed at night. Our current energy system – where we burn fuel to sustain our habits – is not designed accordingly.

A history of solar energy

During the late 1800s, solar collectors for heating water and other fluids were invented and put into practical use for domestic water heating and solar industrial applications, for example, large scale solar desalination (IPCC, 2012). Later, mirrors were used to boost the available fluid temperature, so that heat engines driven by the sun could develop motive power, and hence, electrical power. Also, the late 1800s brought the discovery of a device for converting sunlight directly into electricity – the photovoltaic (PV) cell, which bypassed the need for a heat engine. The silicon PV cell that is used today was discovered around the 1940s.

The modern age of solar research began in the 1950s with the establishment of the International Solar Energy Society (ISES) and increased research and development (R&D) efforts in solar energy (IPCC, 2012). At about the same time, national and international networks of solar irradiance measurements were beginning to be established. With the oil crisis of the 1970s, most countries in the world developed programs for solar energy R&D, which involved efforts by industry, government labs and universities. These policy support efforts have borne fruit: solar energy is enjoying a boom in global production capacity – Figure 1.2 illustrates this fact.

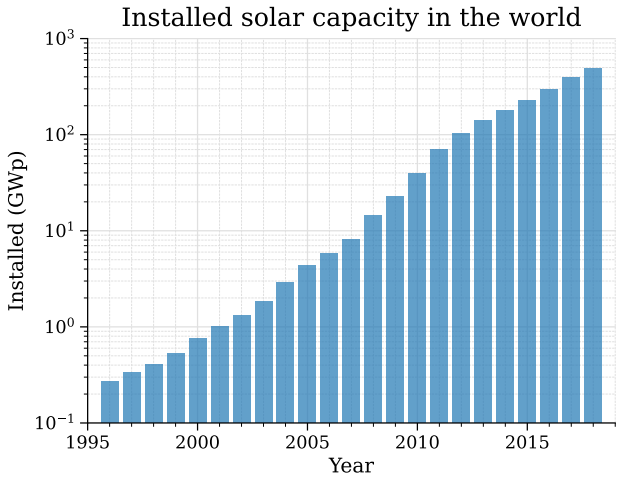


FIGURE 1.2
Global PV capacity since 1996 in a log-scale as expressed in GWp per year. Data from BP (2020).

1.1.3 Solar supply does not follow demand

Solar energy is converted into electrical power via a PV cell. The combination of many PV cells creates a PV module – one panel that can produce sufficient power for human needs. Bundling many PV modules and regulating them together forms a PV system. Such a system relies on sunlight to produce energy, and therefore has stochastic and intermittent behaviour – seasonal effects influence the amount of energy that a PV system produces per day.

Variability of solar irradiance on a second to hourly scale is induced by atmospheric disturbances, most notably clouds, that follow atmospheric movement and affect different locations at different times (Elsinga, 2017). A sudden decrease in electricity production – due to a shadow of a cloud moving across a PV module – can disturb the balance between electrical supply and demand (Notton & Voyant, 2018). To match both, production needs to be forecasted so that consumption can be scheduled accordingly.

An energy system operates at different timescales. Among those timescales, the system should be in equilibrium at all times – that is, supply and demand should be matched. Small deviations can be absorbed by the system via, e.g., batteries, but large mismatches over longer time periods would cause the system to shut down. This problem is solved by forecasting a PV system's power supply and scheduling demand accordingly:

- At the beginning of each day, an accurate forecast for each hour in that day can be made – demand is scheduled for each hour accordingly.
- During the day (intraday), the forecast can be refined as the day proceeds for every hour – flexible demand[2] and battery services are scheduled accordingly.

Therefore, the more accurate the forecast, the lower the system's imbalance costs.

[2]Flexible demand refers to any customer load that need not be on or totally served at all times.

1.2 Scope: solar power forecasts within a day

In an energy system where demand follows production, a solar forecast has to be available at the beginning of each day so that consumption can be scheduled accordingly. As the day proceeds and more information about solar production becomes available, the forecast needs to be updated so that flexible demand can be adjusted accordingly. To create the solar forecast within the day – on an intraday basis – we distinguish between two approaches (Voyant et al., 2017; Yang & van der Meer, 2021):

- an image based approach,[3] which involves:
  - satellite images, or
  - total sky images;[4] and
- a time series based approach,[5] which involves:
  - numerical weather prediction[6] (NWP), or
  - point source measurements (observations).

These approaches are effective over different lead times as given in Table 1.1.

TABLE 1.1
Overview of possible intraday forecast inputs and their lead times (Voyant et al., 2017).

Lead time: Forecast method	**< 15 min**	**15 min - 1 h**	**h - day**
Satellite images		✓	✓
Total sky images	✓	✓	
NWP		✓	✓
Observations	✓	✓	✓

We focus on a time series based approach to create a solar forecast at the beginning of each day, which is updated as the day proceeds. We chose this

[3]Image based techniques relate pixels to clouded/clear parts of the sky. As such, irradiance can be forecasted by tracking the motion of the cloud pixels.

[4]Total sky images come from a bottom-up approach that uses a wide-angle (fish-eye) lens or curved mirror to project the full sky hemisphere onto a finite range, using a digital camera.

[5]Time series based techniques map past measurements and/or other variables to forecast solar power via statistics and/or physical relationships.

[6]Numerical weather prediction uses mathematical equations based on physical relationships of the atmosphere and oceans to predict the weather based on current weather conditions.

approach due to our interest in inference: we would like to understand how numerical predictions and observations from different locations contribute over time to solar power output – we are interested in the spatio-temporal effects of irradiance forecasting.

Time series analysis is compatible with simple statistical methods that can easily be interpreted, whereas an image based approach usually requires complex transformations to first extract data – often in the form of time series – before inference can be done. The extraction of information from satellite images and total sky images is not within our interest. **Therefore, the focus of our research is time series analysis for numerical weather predictions and irradiance observations on the spatio-temporal dimension.**

1.3 Gap: regression framework for numerical weather predictions and irradiance observations

A solar power forecast entails the forecasting of the global horizontal irradiance (GHI) – the total amount of shortwave radiation received from above by a surface horizontal to the ground (Sengupta et al., 2021). We aim to forecast this variable statistically during the day using spatio-temporal irradiance data.

When we aim to forecast one up to a few hours ahead, we mainly find approaches that use observations. These prove to be effective on the short term (Bacher et al., 2009; Voyant et al., 2020), but when moving beyond a couple of hours, NWP becomes more prominent (Sperati et al., 2016; Voyant et al., 2017). The literature describes that to fit NWP well to observations, a process called model output statistics (MOS) – a statistical post-processing technique – is adopted (Glahn & Lowry, 1972; Sperati et al., 2016; Zhang et al., 2022). Therefore, to make accurate intraday solar power forecasts, we can use observations to forecast a few hours ahead and use MOS corrected predictions beyond those few hours.

NWP data is available at an hourly frequency with a spatial resolution of around 9 kilometres (Yang et al., 2022). In addition, irradiance data as measured by weather stations is often available at an hourly frequency with a spatial resolution of 50 kilometres or fewer (López Lorente et al., 2020). We find that the spatial component can contribute to the accuracy of intraday solar power forecasts (Dambreville et al., 2014; Liu et al., 2021). Therefore, to produce accurate intraday solar power forecasts, we can use spatio-temporal data.

Finally, if we aim to decompose our spatio-temporal forecast model to understand how it constructs its forecasts considering the inputs, then we can opt for linear regression due to its interpretability and effectiveness when the inputs are processed to be linear (Hastie et al., 2016). From there, we can create a model that is robust, easy to operate, and fully transparent in how it achieves its results.

Globally, we have to make the switch to a renewable energy system as explained in Section 1.1. Therefore, we need to implement accurate intraday solar power forecasts for the stabilization of renewable energy systems across the globe. If we believe this to be true, then we need to have a scalable framework to produce those forecasts. Such a framework should produce models that have reliable results, are easy to implement, and scale without the need for new (expensive) measurement infrastructure. In the current literature, we find mostly papers that prove the effectiveness of different methods, but we miss the focus on implementation at scale (David et al., 2018). **Therefore, we aim to develop a regression framework that combines numerical weather predictions and irradiance observations over the spatio-temporal dimension to produce accurate intraday solar power forecasts.**

1.4 Contribution: case study of the Netherlands

To build the regression framework, our research encompasses a case study of the Netherlands, where we aim to contribute to the field of solar power forecasts by studying the contribution of spatio-temporal data in a temperate marine climate with cool summers and mild winters. Due to the Netherlands' coastal location, the weather can be highly dynamic – which has a large effect on the presence of clouds and therefore irradiance.

Figure 1.3 shows a plot of numerical predictions of irradiance against observations at Cabauw, which lies in the centre of the Netherlands. Due to the variable weather in the Netherlands, we find the numerical predictions to over-predict medium irradiance conditions. It therefore has difficulty in forecasting periods where there is a high variability in clouds. We aim to combine irradiance observations with numerical weather predictions to create an intraday solar power forecast model that is able to handle highly stochastic (cloudy) situations. Therefore, the contribution of this research is two-fold:

1. We aim to provide a framework for incorporating numerical weather predictions and irradiance observations across the spatio-temporal dimension for accurate solar power forecasts.
2. We aim to provide a case study of the Netherlands to understand how irradiance behaves in a spatio-temporal context.

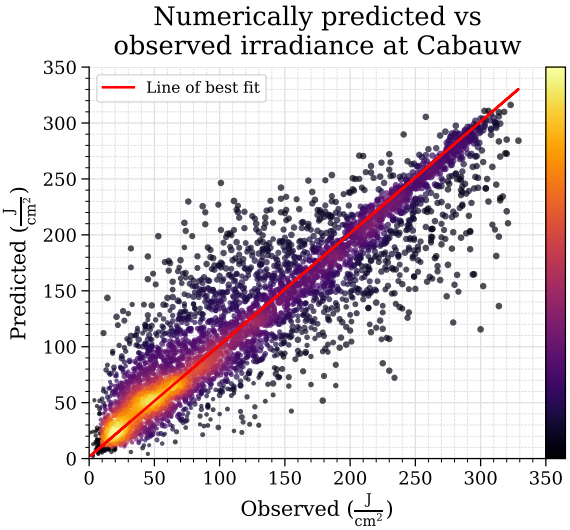


FIGURE 1.3
Scatter density plot – yellow indicates a high density and black a low density – of numerically predicted irradiance against observed irradiance for Cabauw. We find the predictions to overpredict medium irradiance conditions.

1.5 Research questions

The world has to shift to sustainable energy sources to mitigate global warming. Solar energy is a clean energy source, but it has its disadvantages – it does not produce any power at night and fluctuates intensely. To integrate solar power into our electricity system, we have to make accurate forecasts on an intraday basis such that demand can be scheduled accordingly. As such, the stability of the electricity system can be guaranteed.

As we have access to measurements and predictions of solar power over time at different locations, we state our research question as: *How can irradiance observations and numerical weather predictions be regressed on the spatio-temporal dimension to create accurate intraday solar power forecasts?*

The sub-research questions to answer the main question are:

1. What regression techniques for time series are applicable to a spatio-temporal context?
2. How can regression increase the fit of numerical irradiance predictions to irradiance observations?
3. How can regression combine observations and predictions over the spatio-temporal dimension?
4. Does the performance of spatio-temporal regression for solar power forecasts vary spatially and temporally, and if so, why?

1.6 Outline

This thesis is divided in two parts. Part I focusses on gathering the theory, creating the experimental setup, and building the regression framework – it sets the stage for Part II. Part II focusses on the results, its implications, and it sets the context for further applications. We aim to answer the first three sub-research questions in Part I and the fourth in Part II.

For most chapters, we wrote an introduction for clarity about the notations used and for overview of its components. We recommend the reader to return to those introductions if any notations and/or statements are not understood. This thesis builds on these introductions as we move forward, therefore we recommend the reader to read each introduction carefully.

The story of the thesis is as follows. We set out to answer our main research question. To do so, we first identify the theory of regression for time series in Chapter 2. In this chapter, we answer sub-research question one. With the theory defined, we gather the data that we need to apply the theory of regression to the context of solar power forecasts in Chapter 3.

We conclude Chapter 3 with a general framework for the training, testing, and validation of regression models. This enables us to develop a framework for MOS correction in Chapter 4 – we answer sub-research question two. In addition, we develop a framework for spatio-temporal regression in the second part of Chapter 4, which enables us to answer sub-research question three.

We apply and test the developed frameworks in Chapter 5. Here, we evaluate the accuracy of the MOS correction and the spatio-temporal model. We assess the performance of both models under different spatial and temporal circumstances, which enables us to answer sub-research question four. In addition, in Chapter 6, we conduct a case study on a daily scale for one location in the centre of the Netherlands as to deepen our understanding of the spatio-temporal model's accuracy under different cloud conditions. We conclude by validating our model's accuracy through literature.

The goal of this thesis is to answer the main research question, and we do so in Chapter 7. All the preceding chapters contribute to answering that question, therefore, we first identify the answers to all sub-research questions. From there, we formulate a concise answer to the main-research question, which sets the stage for future research.

2

Theory

CONTENTS

KEY TAKEAWAYS

Statistics is concerned with determining the relation between different variables. Therefore, we can use statistics to study the relation between numerical predictions – also known as the *predictors* – and observations – the *response*.

If we aim to predict a response considering multiple predictors, then linear regression is a simple tool that allows us to do so. In addition, linear regression allows us to study the importance per predictor to predict the response.

If we would like to predict the response with lagged values of the response itself, then we can use the theory of autoregression (AR). In addition, we can incorporate spatio-temporal observations by making use of the theory of

vector autoregression (VAR). Finally, we can include numerical predictions as well by simply adding them to the predictor matrix. If we apply smoothing to the numerical predictions, then we can extract their trend to make the model more robust for highly stochastic situations.

To conclude, we introduce different error metrics that can be used to evaluate the accuracy of a model. These error metrics can be used to tune a regression model's parameters so that it performs as expected.

2.1 Introduction

This chapter is concerned with the mathematics behind our regression models. Here, we lay the foundation for all the next chapters.

For the reader to have a clear overview of the different symbols and their equations, Table 2.1 provides an overview.

TABLE 2.1
All symbols used in equations and their definition.

Symbol	Description
μ	The mean (Equation 2.1)
σ^2	The variance (Equation 2.2)
r	The correlation coefficient (Equation 2.3, 2.9)
Q	The least squares criterion (Equation 2.6, 2.14)
t	Index of time step
i	Index of predictor
j	Index of response
T	Total number of time steps
P	Total number of predictors
R	Total number of responses
x	A predictor
y	A response
$\hat{y}$	A predicted response (Equation 2.4, 2.10, 2.18, 2.19)
b	A coefficient (Equation 2.7, 2.8, 2.13)
$\hat{e}$	A prediction error (Equation 2.5)
τ	A time step offset (Equation 2.16, 2.17)
λ	The LASSO coefficient (as used in Equation 2.14)
p	Number of lags for autoregression (as used in Equation 2.18, 2.19, 2.24)
q	Number of points at two sides for smoothing (as used in Equation 2.22, 2.23, 2.24)
ϵ	An error metric (Equation 2.26, 2.28, 2.29)
ε	Skill score (Equation 2.27)

In this chapter, we use matrix notation. Considering Table 2.1, we let $x_{t,i}$ denote the i-th predictor for time step t, where $t = 1, 2, ..., T \wedge i = 1, 2, ..., P$. We let $\mathbf{X}$ denote a $T \times P$ matrix whose (t, i)-th element is $x_{t,i}$. That is,

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,P} \\ x_{2,1} & x_{2,2} & \dots & x_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T,1} & x_{T,2} & \dots & x_{T,P} \end{bmatrix}.$$

When we are interested in a column of $\mathbf{X}$, then we write $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P\}$,

where each column is a vector of length T. Finally, we transpose[1] a matrix with $'$, that is, $\mathbf{X}' = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$.

This chapter was written by consulting the following books:

- Witte and Witte (2017) for the part about statistics in general (Section 2.2);
- Hastie et al. (2016) for the part about linear regression (Section 2.3);
- Hamilton (1994) for the part about autoregression (Section 2.4); and
- Brockwell and Davis (2009) and Chatfield (2003) for the part about data processing (Section 2.5) and error metrics (Section 2.6).

[1] In linear algebra, the transpose of a matrix is an operator which flips a matrix over its diagonal.

2.2 Statistical concepts

To understand what statistical methods exists to convert spatio-temporal weather data to irradiance forecasts, we start by applying the most basic statistical concepts to our case. Those concepts start by first determining the goal of our statistical analysis, which is that we would like to test whether there is a presence or absence of a relationship between two (or more) variables. We conduct an observational study – we focus on detecting relationships between variables that we cannot manipulate nor control.

To find whether there are relationships between variables, we first need to describe these variables and their relationships using mathematical statements. Therefore, we will define those first.

2.2.1 Average

First, we have the average: a number that attempts to describe the middle or central tendency of a set of data. We define the mode, the median and the mean as different kinds of averages. The mode reflects the value of the most frequently occurring one, the median reflects the middle value when ordered from least to most, and the mean, which is the most common, is defined as the sum of all data points divided by the number of data points.

Thus, if we have a process such as solar irradiance that we measure at different time steps, then we can denote any observation by x_t such that we have a set of timely ordered data points as in $\mathbf{x} = \{x_1, x_2, ..., x_T\}$ – also known as a time-series. When we have these time series, we can calculate its mean as

$$\mu_{\mathbf{x}} = \frac{\sum_{t=1}^{T} x_t}{T}, \tag{2.1}$$

where x_t is a predictor at time step t, and T is the total number of time steps in the time series. Here, we define the mean as the balance point for the time series: the sum of all data points expressed as their distance from the mean always equals zero.

2.2.2 Variability

In addition to the average, for us to describe a time series, we are interested in its variability, that is, the amount by which the data points are dispersed in value. We define different measures of variability: the range, which is the difference between the smallest and largest value in the time series; the variance, which is the squared distance of each data point from the mean summed and divided by the number of data points; and the standard deviation, which is the root of the variance.

To calculate the variance for the time series process of $\mathbf{x} = \{x_1, x_2, ..., x_T\}$, we can apply

$$\sigma_{\mathbf{x}}^2 = \frac{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})^2}{T}, \tag{2.2}$$

where x_t is a predictor at time step t, $\mu_{\mathbf{x}}$ is the mean of the time series $\mathbf{x}$, and T is the total number of time steps in the time series. From the variance, we can simply define the standard deviation as $\sqrt{\sigma_{\mathbf{x}}^2} = \sigma_{\mathbf{x}}$.

2.2.3 Correlation

For us to understand whether two variables have any relationship with one-another, we first must examine their relationship graphically. For example, we can examine the relationship between the time series process of predicted irradiance at Cabauw as $\mathbf{x} = \{x_1, x_2, ..., x_T\}$ and the observed irradiance at Cabauw as $\mathbf{y} = \{y_1, y_2, ..., y_T\}$ in a scatter plot as given in Figure 1.3.

We find that there is a linear relationship as given in the graph. The variables are positively related when they have similar relative positions (highs with highs and lows with lows) and they are negatively related when they have dissimilar relative positions (highs with lows and vice-versa). We can define this relationship in a coefficient as in

$$r = \frac{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})(y_t - \mu_{\mathbf{y}})}{\sqrt{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})^2 \sum_{t=1}^{T} (y_t - \mu_{\mathbf{y}})^2}}, \tag{2.3}$$

where x_t is a predictor at time step t, y_t is a response at time step t, $\mu_{\mathbf{x}}$ is the mean of the time series $\mathbf{x}$, $\mu_{\mathbf{y}}$ is the mean of the time series $\mathbf{y}$, and T is the total number of time steps in the time series.

The denominator is always positive, and therefore the numerator tells us whether two variables – or time series processes – are either positively or negatively correlated. If they are positively correlated, then the value of r is positive. If they are negatively correlated, then the value of r is negative. Finally, the value of r is always between -1 and 1, indicating whether the correlation is strong (close to -1 or 1) or weak (close to 0).

2.3 Linear regression

Linear regression provides the framework for us to build a simple model that can relate multiple input variables to one output variable. We explore the concept of linear regression in this section.

2.3.1 Simple linear regression

Simple linear regression is a statistical method that allows us to describe the relationships between two variables:

- x, also known as the *predictor*, explanatory, or independent variable; and
- y, also known as the *response*, outcome or dependent variable.

We call this method of regression simple as its only concerned with one predictor variable. In contrast, multiple linear regression is concerned with multiple – two or more – predictor variables.

Linear regression is not concerned with deterministic (or functional) relationships. A deterministic relationship is one wherein you can exactly calculate the value of y if you have x via a predetermined equation.[2] We are concerned with *statistical relationships*, wherein the relationship between two variables is not perfect – Figure 1.3 is a good example of this.

Let us introduce the following notation to distinguish between our variables:

- x_t is the predictor for time step t;
- y_t is the (observed) response for time step t; and
- $\hat{y}_t$ is the predicted response for time step t.

In that case, $\hat{y}_t$ is defined as

$$\hat{y}_t = b_0 + b_1 x_t, \tag{2.4}$$

where x_t is the predictor at time step t, $\hat{y}_t$ is the predicted response at time step t, b_0 is the constant, and b_1 is the coefficient of the predictor x_t.

When we use Equation 2.4 to make a prediction about the response, then we define the prediction error as

$$\hat{e}_t = y_t - \hat{y}_t, \tag{2.5}$$

[2] An example of such a relationship is the conversion of degrees Celsius to degrees Fahrenheit.

where y_t is the response at time step t, and $\hat{y}_t$ is the predicted response at time step t.

The line that fits best reduces the error to be as small as possible for all data points. One way of achieving this goal is by applying the *least squares criterion*, that is to minimize the sum of squares of the errors as in

$$Q = \sum_{t=1}^{T} \hat{e}_t^2, \tag{2.6}$$

where $\hat{e}_t$ is the prediction error at time step t, and T is the total number of time steps in the time series.

We sum the squares of the prediction error for each time step in order to ensure that positive and negative errors do not cancel one-another out and therefore yield 0.

We minimize Equation 2.6 by taking the derivative with respect to b_0 and b_1, set to 0, and solve for b_0 and b_1 – we obtain the least square estimates. We find

$$b_0 = \mu_{\mathbf{y}} - b_1 \mu_{\mathbf{x}}, \text{ and} \tag{2.7}$$

$$b_1 = \frac{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})(y_t - \mu_{\mathbf{y}})}{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})^2}, \tag{2.8}$$

where x_t is the predictor at time step t, y_t is the response at time step t, $\mu_{\mathbf{x}}$ is the mean of the time series $\mathbf{x}$, $\mu_{\mathbf{y}}$ is the mean of the time series $\mathbf{y}$, and T is the total number of time steps in the time series.

We find that b_1 is equal to the definition of correlation if $\mathbf{x}$ and $\mathbf{y}$ have equal variance as defined in Equation 2.2 – as in

$$r = \frac{\sum_{t=1}^{T} (x_t - \mu_{\mathbf{x}})^2}{\sum_{t=1}^{T} (y_t - \mu_{\mathbf{y}})^2} \times b_1, \tag{2.9}$$

where x_t is a predictor at time step t, y_t is a response at time step t, $\mu_{\mathbf{x}}$ is the mean of the time series $\mathbf{x}$, $\mu_{\mathbf{y}}$ is the mean of the time series $\mathbf{y}$, b_1 is the coefficient of the predictor x_t, and T is the total number of time steps in the time series.

2.3.2 Multiple linear regression

When we aim to have multiple predictors, we expand our simple linear regression model to multiple linear regression. In formula form, when we have P predictors, it is defined as

$$\hat{y}_t = b_0 + \sum_{i=1}^{P} b_i x_{t,i}, \tag{2.10}$$

where $x_{t,i}$ is the i-th predictor at time step t, $\hat{y}_t$ is the predicted response at time step t, b_0 is the constant, b_i is the coefficient of the predictor $x_{t,i}$, and P is the total number of predictors.

Because of the potentially large number of predictors, we can write multiple linear regression more efficient in matrix form, which results in

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_T \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,P} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T,1} & x_{T,2} & \dots & x_{T,P} \end{bmatrix} \times \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_P \end{bmatrix}, \tag{2.11}$$

where $x_{t,i}$ is the i-th predictor at time step t, $\hat{y}_t$ is the predicted response at time step t, b_0 is the constant, and b_i is the coefficient of the predictor $x_{t,i}$.

Let us introduce the following notation to distinguish between our matrices and vectors:

- $\mathbf{X}$ is the matrix with the predictors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_P\}$;
- $\mathbf{y}$ is the vector with the observed responses $\mathbf{y} = \{y_1, y_2, ..., y_T\}$;
- $\hat{\mathbf{y}}$ is the vector with the predicted responses $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_T\}$; and
- $\mathbf{b}$ is the vector with the estimated coefficients $\mathbf{b} = \{b_1, b_2, ..., b_P\}$.

Thus, we can rewrite the multiple linear regression as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}, \tag{2.12}$$

and solve for $\mathbf{b}$ by minimizing the least squares as in Equation 2.6. We then find

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{2.13}$$

where $\mathbf{X}'$ is the transpose of $\mathbf{X}$, and $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of $\mathbf{X}'\mathbf{X}$.[3] Note that when we take the transpose of $\mathbf{X}$, then $\mathbf{X}' = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$.

When we want to predict multiple responses with the same predictors, we can make use of matrix notation. That is, when we want to predict R responses, we write:

- $\mathbf{Y}$ is the matrix with the observed responses $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_R\} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}'$;
- $\hat{\mathbf{Y}}$ is the matrix with the predicted responses $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_R\}$; and
- $\mathbf{B}$ is the matrix with the estimated coefficients $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_R\}$.

[3] For the full mathematical derivation, we refer to Kong et al. (2020).

2.3.3 LASSO

When we try to estimate the coefficients for a model with many predictors that are heavily correlated, then the estimation of the coefficients might become unstable as the variables embed the same information – this phenomenon is called collinearity. In that case, predictor selection can help to reduce overfitting. The Least Absolute Shrinkage and Selection Operator (LASSO) serves this purpose.

LASSO involves the introduction of a constraint on coefficient estimation (Equation 2.13) so that the coefficients cannot take extreme values. It does so via the minimization of

$$Q = \sum_{t=1}^{T} (y_t - \mathbf{x}_t \mathbf{b})^2 + \lambda \sum_{i=1}^{P} |b_i|, \tag{2.14}$$

where y_t is the response at time step t, $\mathbf{x}_t$ is the vector of predictors at time step t, $\mathbf{b}$ is the vector of coefficients, λ is the LASSO parameter, b_i is the coefficient of the predictor $x_{t,i}$, T is the total number of time steps in the time series, and P is the total number of predictors.

The LASSO also induces predictor selection by shrinking some coefficients to zero. The parameter λ determines the strength of the shrinkage and should therefore be tuned to decrease the error of the model as much as possible. In our study, the LASSO implementation of Friedman et al. (2010) was used as implemented in Statsmodels (2019).

2.3.4 Standardization

When we apply the LASSO, then we put a penalty on the absolute sum of the coefficients (Equation 2.13). However, the magnitude of the coefficients is dependent on the magnitude of the predictor. Therefore, we have to ensure that all our predictors have the same magnitude. In addition, if all our predictors have the same magnitude, then the weights of the coefficients also give an indication of each predictor's importance to predict the response. Therefore, we standardize all our predictors so that they have a mean, μ (Equation 2.1), of zero and a standard deviation, σ (Section 2.2.2), of one as

$$x_t = \frac{x_t - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}}. \tag{2.15}$$

Due to the application of standardization, we can drop the constant, b_0 (Equation 2.7), from our coefficient matrix and the column of ones in the predictors as found in Equation 2.11, as all variables have a mean of zero. We then predict the response, and we apply the inverse of the standardization – that is, we apply the original mean, μ, and standard deviation, σ, from the response back to the predicted response.

2.4 Autoregressive models

Autoregressive models use lagged response variables as predictors in a linear regression model to predict the response. This section explores the use of lagged spatio-temporal observations with numerical predictions in regression to create a statistical forecast.

2.4.1 Autoregression

Consider a time series denoted by $\mathbf{y} = \{y_1, y_2, ..., y_T\}$. The goal is to forecast $y_{t+\tau}$. To do so, we find a function $f_\tau(\mathbf{x}_t)$ that maps the vector $\mathbf{x}_t$ onto $y_{t+\tau}$,

$$\hat{y}_{t+\tau} = f_\tau(\mathbf{x}_t), \tag{2.16}$$

where $f_\tau(.)$ is the function to predict a response at time step $t + \tau$, and $\mathbf{x}_t$ is the vector of predictors at time step t.

For time series that exhibit autocorrelation,[4] it is reasonable for the vector $\mathbf{x}_t$ to consist of the recent history of $y_{t+\tau}$ as

$$\hat{y}_{t+\tau} = f_\tau(\{y_t, y_{t-1}, ...\}), \tag{2.17}$$

where $f_\tau(.)$ is the function to predict a response at time step $t + \tau$, and y_t is the response at time step t. Therefore, the function $f_\tau(.)$ takes the form of a weighted sum of p past values plus a constant as

$$\hat{y}_{t+\tau} = b_0 + \sum_{i=0}^{p-1} b_i y_{t-i}, \tag{2.18}$$

where y_{t-i+1} is the response at time step $t - i + 1$, b_0 is the constant, b_i is the coefficient of the predictor y_{t-i+1}, and p is the total number of predictors in the time series. Equation 2.18 is essentially equal to Equation 2.10, however, the predictors are acknowledged to be lagged values of the response.

To conclude, the AR model is multiple linear regression with its predictors being lagged values of the response. From there, we can solve for $\mathbf{b}$ to calculate $\hat{\mathbf{y}}$.

2.4.2 Vector autoregression

The AR model can be expanded to take into account lags from other time series as well to predict a response. This can be used for the spatial aspect of our forecasts. We do so by extending Equation 2.18 to vector form as

[4] Autocorrelation, also known as serial correlation, is the correlation of a signal with a delayed copy of itself as a function of delay.

$$\hat{y}_{t+\tau} = b_0 + \sum_{i=0}^{p-1} \sum_{j=1}^{R} b_{Ri+j} \mathbf{Y}_{t-i,j}, \tag{2.19}$$

where $\mathbf{Y}_{t-i,j}$ is the j-th response at time step $t-i$, b_0 is the constant, b_{Ri+j} is the coefficient of the predictor $\mathbf{Y}_{t-i,j}$, R is the total number of responses to predict, and p is the total number of lagged responses as predictors.

Equation 2.19 is a vector autoregressive (VAR) model, which can capture the interdependencies of multiple time series at once. In our case, the spatially dispersed time series can be incorporated into one forecast model via this approach. We acknowledge that this is essentially multiple linear regression with lagged and spatially dispersed predictors.

2.5 Data processing for autoregression

To do autoregression on a set of data, first the data needs to be pre-processed to remove any patterns – we want to model the stochastic process. In addition, autoregression can be combined with numerical predictions to create a model that is able to deal with highly stochastic processes.

2.5.1 Stationarity

Consider that we are occupied with a time series that is defined by an equation that relates the value of y_{t+1} to a constant, b_0, in addition to its value in the previous period, y_t, multiplied with a coefficient, b_1, as in

$$y_{t+1} = b_0 + b_1 y_t. \tag{2.20}$$

Equation 2.20 is a linear first-order difference equation.[5] A difference equation relates a variable y_t to its value(s) in a previous period. Equation 2.20 is equal to Equation 2.18 with $p = 1 \wedge \tau = 1 \wedge \hat{y}_{t+\tau} = y_{t+\tau}$. Consider that we would like to calculate the value of y_{t+2} with $\tau = 2$, then

$$y_{t+2} = b_0 + b_1(b_0 + b_1 y_t), \tag{2.21}$$

which can be continued for any value of τ.

We find that different values of b_1 produce different responses of $y_{t+\tau}$ to y_t. If $|b_1| < 1$, the effect of y_t on $y_{t+\tau}$ decays geometrically to zero. However, if $|b_1| > 1$, its effect increases exponentially over time. Therefore, the system is stable when $|b_1| < 1$ as the consequence of any change in y_t eventually dies out. This concept is called stationarity, which indicates that there is no trend and/or (a) seasonal component(s) in the time series. A time series is stationary if $|\sum_{i=1}^{p} b_p| < 1$. Thus, to forecast any time series by its past value(s), it must first be detrended and relieved from its patterns – it must be made stationary.

2.5.2 Smoothing of numerical predictions

Time series can be made stationary by first removing any patterns – we only want to be left with the high frequency fluctuations. We do so by removing the low frequencies, which can be done by *smoothing*.

We can find an approximation for the low frequency signal in a time series by making the assumption that data points nearby in time are likely to be close in value. In that case, taking an average of points around one data point should provide a reasonable estimate of that point's trend.

[5] As it only uses one lag, it is of the first order.

Let q be a non-negative number integer and consider the two-sided[6] moving average,

$$\hat{y}_t = \frac{1}{2q+1} \sum_{i=-q}^{q} x_{t+i}, \tag{2.22}$$

where x_{t+i} is a predictor at time step $t+i$, and q is the number of points to smooth on two sides of x_t. Here, we use the trend from x_t as a prediction for y_t.

In Equation 2.22, the weights are equal for each point in the time series. However, we might want to optimize the weights on prediction of y_t via multiple linear regression (Equation 2.10) as in

$$\hat{y}_t = b_0 + \sum_{i=-q}^{q} b_{q+i+1} x_{t+i}, \tag{2.23}$$

where x_{t+i} is a predictor at time step $t+i$, b_0 is the constant, b_{q+i+1} is the coefficient of the predictor x_{t+i}, and q is the number of points to smooth on two sides of x_t. By using weighted averages, the resulting trend becomes much smoother as data points enter and leave the average gradually. In fact, the weighted moving average acts as a low-pass filter – it filters out the high frequencies.

When we try to predict irradiance by its lagged values, then we can include numerical predictions for the time step that we try to predict as well. If we smooth those numerical predictions, then we can extract the trend from those predictions and correct the lagged observations with that trend. By doing so, we essentially take the high frequencies from the lagged data and the low frequencies from the numerical predictions. Therefore, we create a model that is robust against highly stochastic situations.

If we state that $\mathbf{X}^*$ contains numerical predictions that correspond with $\mathbf{Y}$, then we can combine the VAR model of Equation 2.19 and include smoothing for numerical predictions as

$$\hat{y}_{t+\tau} = b_0 + \sum_{i=0}^{p-1} \sum_{j=1}^{R} b_{Ri+j} \mathbf{Y}_{t-i,j} + \sum_{i=-q}^{q} \sum_{j=1}^{R} b_{R(p+q+i)+j} \mathbf{X}^*_{t+\tau+i,j}, \tag{2.24}$$

where $\mathbf{Y}_{t-i,j}$ is the j-th response at time step $t-i$, $\mathbf{X}^*_{t+\tau+i,j}$ is the j-th prediction at time step $t+\tau+i$, b_0 is the constant, b_{Ri+j} is the coefficient of the predictor $\mathbf{Y}_{t-i,j}$, $b_{R(p+q+i)+j}$ is the coefficient of the predictor $\mathbf{X}^*_{t+\tau+i,j}$, R is the total number of responses to predict, p is the total number of lagged responses as predictors, and q is the number of points to smooth on two sides of $\mathbf{X}^*_{t+\tau,j}$.

[6] Two-sided for approaching a value from its preceding and succeeding data points in a time series.

To conclude, we essentially create a new predictor matrix, $\mathbf{X}$, that consists of lagged observations of the response matrix, $\mathbf{Y}$, and smoothed predictors from the old predictor matrix, $\mathbf{X}^*$.

2.6 Error metrics

We define a regression model's error as

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}, \tag{2.25}$$

which is the vector that includes all errors – Equation 2.5 – over t such that $\hat{\mathbf{e}} = \{\hat{e}_1, \hat{e}_2, ..., \hat{e}_T\}$. This error can be processed in multiple ways to define useful metrics to evaluate a regression model.

2.6.1 Root mean square error

First, we define the root mean square error (RMSE) as

$$\epsilon_{\text{RMSE}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2}, \tag{2.26}$$

which is the rooted mean of the squared errors over a time series with a length of T. When we calculate the coefficients using the least squared criterion as in Equation 2.6, we actually optimize on this error metric. Due to the squared term, the RMSE penalizes large errors more than small ones. Therefore, it provides an insight into the variance of the error.

2.6.2 Skill score

To evaluate how our regression model's forecasts compares with that of a baseline model, we define the skill score (SS) as

$$\varepsilon = 1 - \frac{\epsilon_{\text{RMSE}_{\text{for}}}}{\epsilon_{\text{RMSE}_{\text{ref}}}}, \tag{2.27}$$

where $\epsilon_{\text{RMSE}_{\text{for}}}$ is the RMSE of our forecast and $\epsilon_{\text{RMSE}_{\text{ref}}}$ is the RMSE of the baseline forecast over the same period. Thus, the SS is the reduction in RMSE as a percentage.

2.6.3 Other error metrics

For a regression model, it is important that its predicted responses have consistent statistical properties among them. Therefore, we define two additional metrics to evaluate a regression model. These are used to ensure model stability.

2.6.3.1 Mean bias error

We introduce the mean bias error (MBE) as

$$\epsilon_{\text{MBE}} = \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t, \tag{2.28}$$

which is the mean of the errors over a time series with a length of T. This metric provides an insight into the structural bias of a regression model.

2.6.3.2 Mean absolute error

Finally, we define the mean absolute error (MAE) as

$$\epsilon_{\text{MAE}} = \frac{1}{T}\sum_{t=1}^{T}|\hat{e}_t|, \tag{2.29}$$

which is the mean of the absolute errors over a time series with a length of T. The MAE provides an insight into the overall error, which should have approximately a linear relationship with the RMSE if there are no strong outliers.

3

Experimental setup

CONTENTS

KEY TAKEAWAYS

As the basis for our spatio-temporal analysis, we took the weather stations from the KNMI, which are spread across the Netherlands. For each location, we extracted numerical weather predictions from the ECMWF and irradiance observations from the KNMI, after which we removed diurnal patterns from the KNMI's and ECMWF's irradiance data so that we are left with the pure stochastic process.

From there, we removed the time steps from the dataset where the sun is not above the horizon, and we created some additional predictors to increase the fit between the predictions and the observations.

To conclude, we define a simple process to train, test, and validate our models. First, we create a train set that contains the dates between 2019-09-01 and 2020-08-31 – a full year. Second, a test set that contains the dates between 2020-09-01 and 2020-12-31 – mostly a winter – for tuning the model. Finally, a validation set that contains the dates between 2019-07-01 and 2019-08-31 – mostly a summer – to validate the model's accuracy.

3.1 Introduction

This chapter is concerned with the application of the theory of Chapter 2. Here, we actually gather and process the data to build the predictors, $\mathbf{X}$, and define the responses, $\mathbf{Y}$.

To refer to different parameters that were used in our research, we refer to them by `this` font. In addition, we use bracket notation to refer to specific locations concerning that parameter, e.g., `ssrc[9]` means the clear sky net surface solar radiation downwards from Table 3.1 for the 9th station of Figure 3.1.

For the reader to have a clear overview of all parameters and their sources, we provide a shortlist here:

- The NWP parameters from the ECMWF are given in Table 3.1
- The observations' parameter from the KNMI is given in Table 3.2
- The clear-sky indexes are given in Table 3.3
- The dummy variables and the polynomials of the clear-sky index are given in Table 3.4

3.2 Sources of data

To do the spatio-temporal analysis, first NWP and observed irradiance data had to be collected. Considering that we wanted to analyse the spatio-temporal effects of irradiance, a case-study was conducted for the Netherlands. Data was collected for the locations of the weather stations of the Royal Netherlands Meteorological Institute (KNMI).

The KNMI is a Dutch government institute that specializes in meteorology, climate science and seismology. It hosts a number of weather stations that measure irradiance on an hourly basis across the Netherlands as depicted in Figure 3.1. Therefore, these locations were chosen for the spatio-temporal analysis.

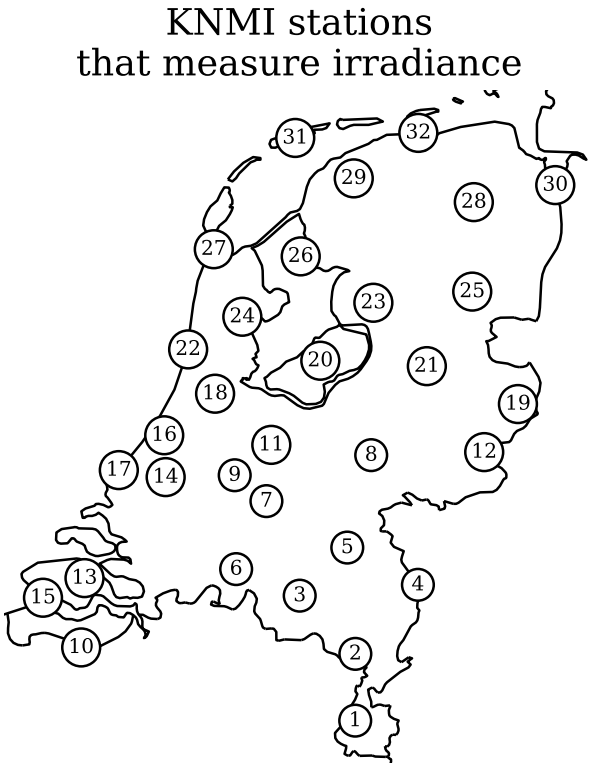


FIGURE 3.1
KNMI stations that measure irradiance across the Netherlands. The numbers are ordered on distance from Maastricht.

In addition to the observed irradiance data from the KNMI, NWP data for

those locations was extracted from the European Centre for Medium-Range Weather Forecasts (ECMWF).

3.2.1 Numerical weather predictions from ECMWF

The NWP data was collected from the European Centre for Medium-Range Weather Forecasts (ECMWF), which is an independent intergovernmental organization supported by most of the nations in Europe. It operates one of the largest supercomputer complexes in Europe and has the world's largest archive of NWP data. One of its hourly forecasts has the highest spatial resolutions, uses the best description of the model physics, and its initial state is the most accurate of the current atmosphere's conditions. This is the High-Resolution[1] (HRES) forecast by the ECMWF.

As we are concerned with intraday forecasts at the beginning of the day, we took HRES data that becomes available each day before 07:00 UTC. The parameters as in Table 3.1 were extracted by Whiffle for each hour.

TABLE 3.1
The weather parameters from the ECMWF extracted for each station.

Name	Description
`cdir`	Clear sky direct solar radiation at surface
`coszenith`	Cosine of solar zenith
`fdir`	Total sky direct solar radiation at surface
`Ms`	10m wind speed
`phis`	10m wind direction
`sp`	Surface pressure
`ssrc`	Clear sky net surface solar radiation downwards
`ssrd`	Surface solar radiation downwards (GHI)
`tcc`	Total cloud cover
`ts`	2m temperature

3.2.2 Observations from KNMI

The KNMI houses a meteorological network in the Netherlands to collect weather data for its forecasts. The meteorological network of the Netherlands consists of 51 automatic weather stations at typical spacing of 50 kilometres, including platforms in the North Sea – 32 of those weather stations collect irradiance data.

The observed surface solar radiation downwards was collected, which is essentially the irradiance that we want to forecast. Therefore, all weather stations that measured surface solar radiation downwards were selected as

[1]The HRES has a resolution of 9 kilometres.

given in Figure 3.1. The parameter as in Table 3.2 was extracted via a self-written Python script from the KNMI's website.

TABLE 3.2
The irradiance from the KNMI extracted for each station.

Name	Description
`Q`	Surface solar radiation downwards (GHI)

3.3 Data processing

Before we were able to use the GHI data[2] in autoregression (Section 2.4), we first had to ensure that the data was stationary as explained in Section 2.5.1. Our aim was to remove patterns from the GHI data so that we can model its stochastic behaviour.

3.3.1 Calculation of clear-sky index

We can imagine that GHI data, which follows the pattern of the sun as in Figure 3.2, can be processed to take out its diurnal pattern, which is described by `ssrc`.

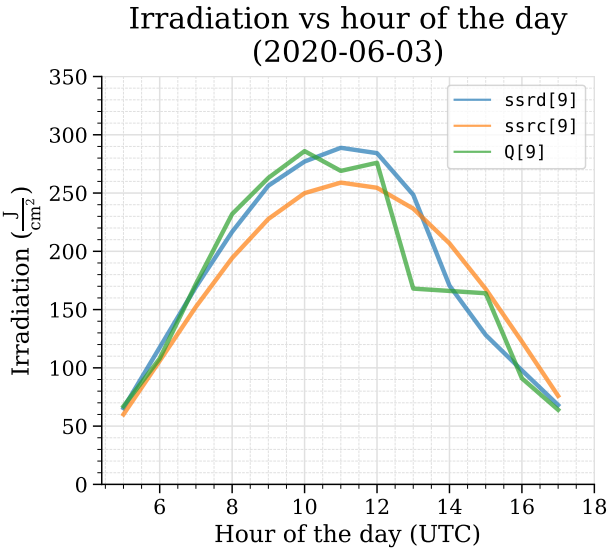


FIGURE 3.2
The `ssrd`, `ssrc`, and `Q` for station number 9 of Figure 3.1 plotted against time for a summer day. We find that there is a parabolic diurnal pattern, which should be removed before time series analysis can be applied as explained in Section 2.5.1.

The ECMWF provides clear-sky data,[3] which it produces as an addition to its irradiance profiles (Hogan & Bozzo, 2018). The clear-sky model calculates what the irradiance would have been if there had been no clouds in the

[2]Global horizontal irradiance data is `ssrd` from Table 3.1 and `Q` from Table 3.2.
[3]Those are the `ssrc` and `fdir` in Table 3.1.

atmosphere. In our research, `ssrd` and `Q` was made stationary with `ssrc`. The result is the clear-sky index.

The clear-sky index (k) is simply calculated by dividing irradiance (I) by its clear-sky irradiance (I_c) as in

$$k = \frac{I}{I_c}. \tag{3.1}$$

As such, the variables as in Table 3.3 were defined.

TABLE 3.3
The clear-sky index as calculated for each station.

Name	Description
`csi_predicted`	$= \frac{\texttt{ssrd}}{\texttt{ssrc}}$
`csi_observed`	$= \frac{\texttt{Q}}{\texttt{ssrc}}$

3.3.2 Selection on solar zenith angle

To calculate the clear-sky index, Equation 3.1 indicates that

- the clear-sky irradiance cannot be zero; and
- the clear-sky irradiance should not approximate zero, as it can take on extreme values when $I \gg I_c$.

Therefore, we want to select a subset of the data to ensure that these two conditions are met. The clear-sky irradiance is not equal to zero when the sun is above the horizon. In addition, the clear-sky irradiance does not approximate zero when the sun is at least a few degrees above the horizon.

The height of the sun compared to the horizon can be expressed by the solar zenith angle (θ), which is the angle between the sun's rays and the vertical direction. Therefore, when it is below 90 degrees, then the sun is above the horizon. In our case, to ensure that the clear-sky values do not approximate zero, we took only those measurements wherein half of the time[4] the solar zenith angle was below 80 degrees[5] as in

$$\mathbf{X}' \wedge \mathbf{y} = \{\mathbf{x}_t \in \mathbf{X}' \wedge y_t \in \mathbf{y} \mid \theta_t < 80\}.$$

We calculated the solar zenith angle via Holmgren et al. (2018).

[4] To ensure that the solar zenith angle is below a threshold for half of the time, the solar zenith angle is taken at half of each step.

[5] The 80 degrees threshold was set based on eliminating outliers in k.

3.3.3 Dummy variables and polynomials

Additional variables were created by using dummy encoding and polynomials. First, dummies help to encode a categorical variable into the regression model. For any hour of the day, h, it takes the form

$$x_{t,h} = \begin{cases} 1 & \text{if hour of the day at } t \text{ equals } h \\ 0 & \text{if hour of the day at } t \text{ does not equal } h. \end{cases} \tag{3.2}$$

The dummy variables were created for each hour of the day in UTC. As the data was filtered on the solar zenith angle, the dummies were created for the hours from 5 until 19.

Second, for the clear-sky index that was predicted by the ECMWF, the squared and cubic was calculated as there appeared a non-linear relationship with the observed clear-sky index as found in Figure 3.3. These proved valuable during the correction of the numerical predictions to better fit the observations. As such, the additional variables as in Table 3.4 were defined.

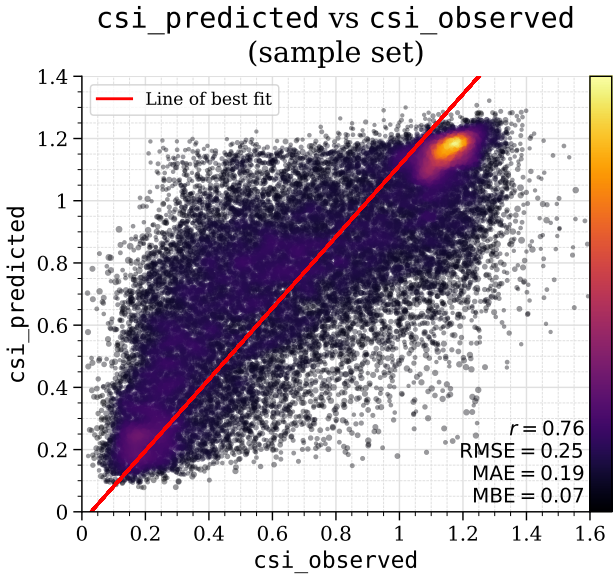


FIGURE 3.3
Predicted clear-sky index against observed clear-sky index for a sample set that spans 2020-09-01 until the end of 2020. We see a non-linear pattern and some bias.

TABLE 3.4
The cubed and squared clear-sky index for each station and additional dummy variables.

Name	**Description**
`dummy_hour_5`	Set to 1 if the hour equals 5 else 0
`dummy_hour_6`	Set to 1 if the hour equals 6 else 0
⋮	⋮
`dummy_hour_18`	Set to 1 if the hour equals 18 else 0
`csi_predicted_squared`	= `csi_predicted`^{2}
`csi_predicted_cubed`	= `csi_predicted`^{3}

3.4 Model training, testing, and validation

To test whether a regression model was actually fitted well, we drew conclusions by calculating the coefficients ($\mathbf{b}$ in Equation 2.13) via a train set and applying those to a test set. The predicted responses for the test set were evaluated by the error metrics of Section 2.6. From there, we used the error metrics to tune the model's hyperparameters,[6] e.g., λ of Equation 2.14. Once an equilibrium for a regression model's hyperparameters was reached, we evaluated if the model actually performed as well as tested by using a validation set. Thus, we have a

- train set to train the model (as we calculate the coefficients of Equation 2.13),
- test set to evaluate increments of the model, and
- validation set to evaluate the model's accuracy.

We define each set to include

$$\Phi_t = \begin{cases} \text{2019-09-01} \leq t \leq \text{2020-08-31} & \text{if train set} \\ \text{2020-09-01} \leq t \leq \text{2020-12-31} & \text{if test set} \\ \text{2019-07-01} \leq t \leq \text{2019-08-31} & \text{if validation set,} \end{cases}$$

and we define the sets as

$$\begin{array}{ll} \mathbf{X}'_{\text{train}} \wedge \mathbf{y}_{\text{train}} = \{\mathbf{x}_t \in \mathbf{X}' \wedge y_t \in \mathbf{y} \mid \Phi_t\} & \text{if train set} \\ \mathbf{X}'_{\text{test}} \wedge \mathbf{y}_{\text{test}} = \{\mathbf{x}_t \in \mathbf{X}' \wedge y_t \in \mathbf{y} \mid \Phi_t\} & \text{if test set} \\ \mathbf{X}'_{\text{val}} \wedge \mathbf{y}_{\text{val}} = \{\mathbf{x}_t \in \mathbf{X}' \wedge y_t \in \mathbf{y} \mid \Phi_t\} & \text{if validation set.} \end{array}$$

Thus, the sets $\mathbf{X}_{\text{train}}$ and $\mathbf{y}_{\text{train}}$ only contain the rows between 2019-09-01 and 2020-08-31, the sets $\mathbf{X}_{\text{test}}$ and $\mathbf{y}_{\text{test}}$ only contain the rows between 2020-09-01 and 2020-12-31, and the validation sets those between 2019-07-01 and 2019-08-31. The number of elements per set are given in Table 3.5.

TABLE 3.5
The number of elements in the train, test, and validation set.

Set	Number of elements
Train	3,369
Test	803
Validation	783

[6] A hyperparameter is a parameter whose value is used to control the coefficient estimation process.

To conclude, we would like to introduce the notation of $\mathbf{b}_{\text{train}}$ if the coefficient vector was calculated with $\mathbf{X}_{\text{train}} \wedge \mathbf{y}_{\text{train}}$, $\mathbf{b}_{\text{test}}$ if the coefficient vector was calculated with $\mathbf{X}_{\text{test}} \wedge \mathbf{y}_{\text{test}}$, and so forth.

4

Regression framework

CONTENTS

KEY TAKEAWAYS

We defined MOS correction as a regression model that uses weather parameters from NWP in combination with some dummy and polynomial predictors to predict the observed clear-sky index. By doing so, and by having a predictor matrix with many variables, the regression model relates those predictors to the clear-sky index while optimizing for accuracy. To execute on this strategy, we first created a dataset for each station, after which we trained, tested, and validated the MOS correction model.

From there, we defined the spatio-temporal model to use lagged observations and to smooth the MOS corrected predictions on two sides. To develop this model, we first created the set of predictors and responses, after which we trained, tested, and validated the spatio-temporal model as we did for the MOS correction model. Note that each station uses the exact same set of predictors in contrast to the MOS correction.

4.1 Introduction

This chapter is concerned with the regression framework for the MOS correction and the spatio-temporal model using the experimental setup of Chapter 3. Here, we actually built the predictors, $\mathbf{X}$, and define the responses, $\mathbf{Y}$.

We use flowcharts to communicate algorithms. Figure 4.1(a) is an example of such a flowchart, which uses the following symbols to communicate events:

- Square is a process
- Circle is a start or an end
- Parallelogram is an input or output of data
- Diamond (or rhombus) is a decision

In addition, we extend our bracket notation from Chapter 3 to matrices. That is, we use bracket notation to sub-select from a matrix the variables that belong to a specific location, e.g., $\mathbf{X}[9]$ means the columns from $\mathbf{X}$ for the 9th station of Figure 3.1 as depicted in Figure 4.1(b).

For the reader to have a clear overview of the regression models' outputs, we provide a shortlist here:

- The outputs from the MOS correction are given in Table 4.1
- The outputs from the spatio-temporal model are given in Table 4.2

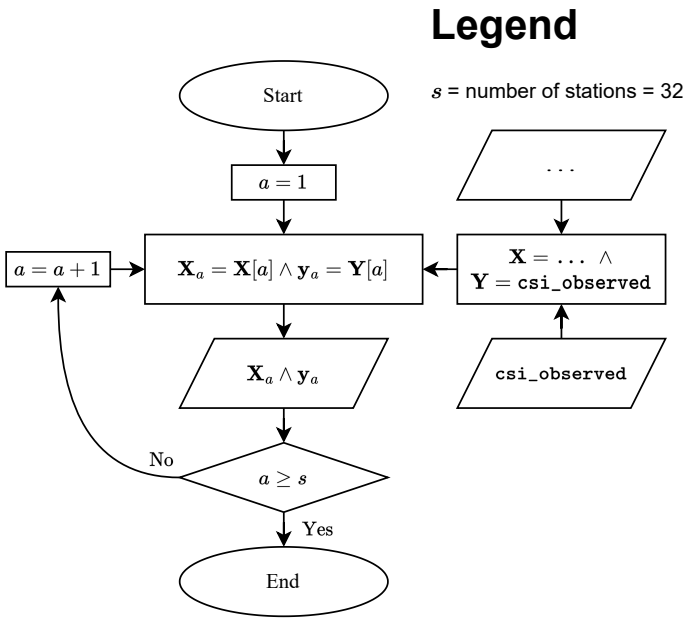
4.2 MOS correction

To ensure that the numerical irradiance predictions, `csi_predicted`, fit well to the observations, `csi_observed`, we first used a statistical process on the numerical predictions called model output statistics (MOS). To do so, we used regression, and we first created a set of predictors and response per station of Figure 3.1. From there, regression was applied to each station's set and each model was tested using the RMSE (Equation 2.26).

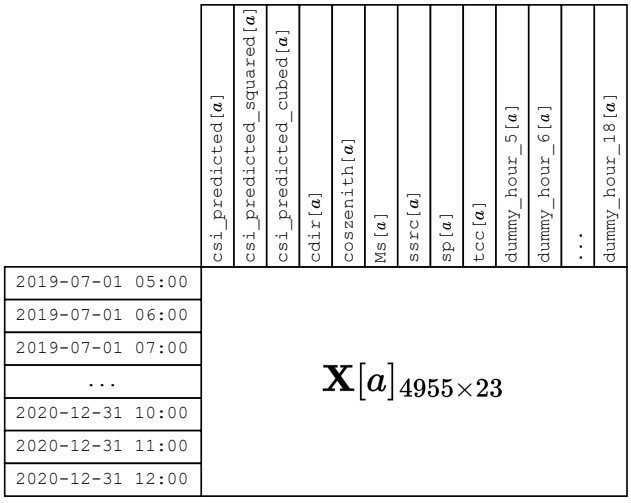
4.2.1 Data processing

For the MOS correction, we applied regression to each station separately as depicted in the workflow of Figure 4.1(a). Therefore, we first constructed a separate set of predictors and response for each station. We looped over each station of Figure 3.1, which is the range of integers from 1 until 32. In the loop, we assigned a to be the iterator. We created for each station, a, a separate set of predictors and response, which we defined as $\mathbf{X}_a = \mathbf{X}[a]$ and $\mathbf{y}_a = \mathbf{Y}[a]$ respectively.

Here, for each station, a, the response $\mathbf{y}_a$ contains the variable `csi_observed[`a`]` and the predictors contain the variables as given in Figure 4.1(b).



(a)



(b)

FIGURE 4.1
Figure 4.1(a) depicts the workflow for the construction of the predictors and response for the MOS correction. We loop over each station to select for each parameter the variables that belong to that station's location. Figure 4.1(b) shows the predictors that were used for each station.

4.2.2 Training, testing, and validation

As a follow-up to the workflow of Figure 4.1, we used the generated sets for each station to build a regression model as depicted in the workflow of Figure 4.2. First, we looped over each station in Figure 3.1, which is the range of integers from 1 until 32. In the loop, we assigned a to be the iterator. We retrieved the predictors, $\mathbf{X} = \mathbf{X}_a$, and response, $\mathbf{y} = \mathbf{y}_a$, for each station. We used these datasets to generate $\mathbf{b}_{\text{train}}$ (Section 3.4), which is the vector of coefficients (Equation 2.13). Furthermore, we did so by applying LASSO (Equation 2.14), and we started with $\lambda = 0$. From there, we applied $\mathbf{b}_{\text{train}}$ to $\mathbf{X}$ to generate a set of predicted responses, $\hat{\mathbf{y}}$ (Equation 2.12). To evaluate the accuracy of the coefficients, we first converted the predicted response, $\hat{\mathbf{y}}$, which is a clear-sky index (Equation 3.1), to irradiance using the clear-sky irradiance, `ssrc`. To conclude, we used the MOS corrected irradiance, $\hat{\mathbf{y}}^*$, and the observed irradiance, $\mathbf{y}^*$, to calculate the RMSE (Equation 2.26) over the test set, $\hat{\mathbf{y}}^*_{\text{test}} \wedge \mathbf{y}^*_{\text{test}}$ (Section 3.4).

We applied iterations to this process wherein we increased the λ value from the LASSO to find an optimum in terms of RMSE. Therefore, we continued in a loop until the RMSE increased instead of decreased, which was quantified by the SS, ε (Equation 3.4). In that case, the model had converged to its most optimal point, and the predicted response was saved to `csi_MOS[`a`]` and `ssrd_MOS[`a`]` as clear-sky index and irradiance respectively. After this was done for all stations, we ended up with two new parameters, `csi_MOS` and `ssrd_MOS`, which are MOS corrected predictions as given in Table 4.1.

TABLE 4.1
The MOS corrected predictions as calculated for each station.

Name	Description
`csi_MOS`	MOS corrected `csi_predicted`
`ssrd_MOS`	MOS corrected `ssrd`

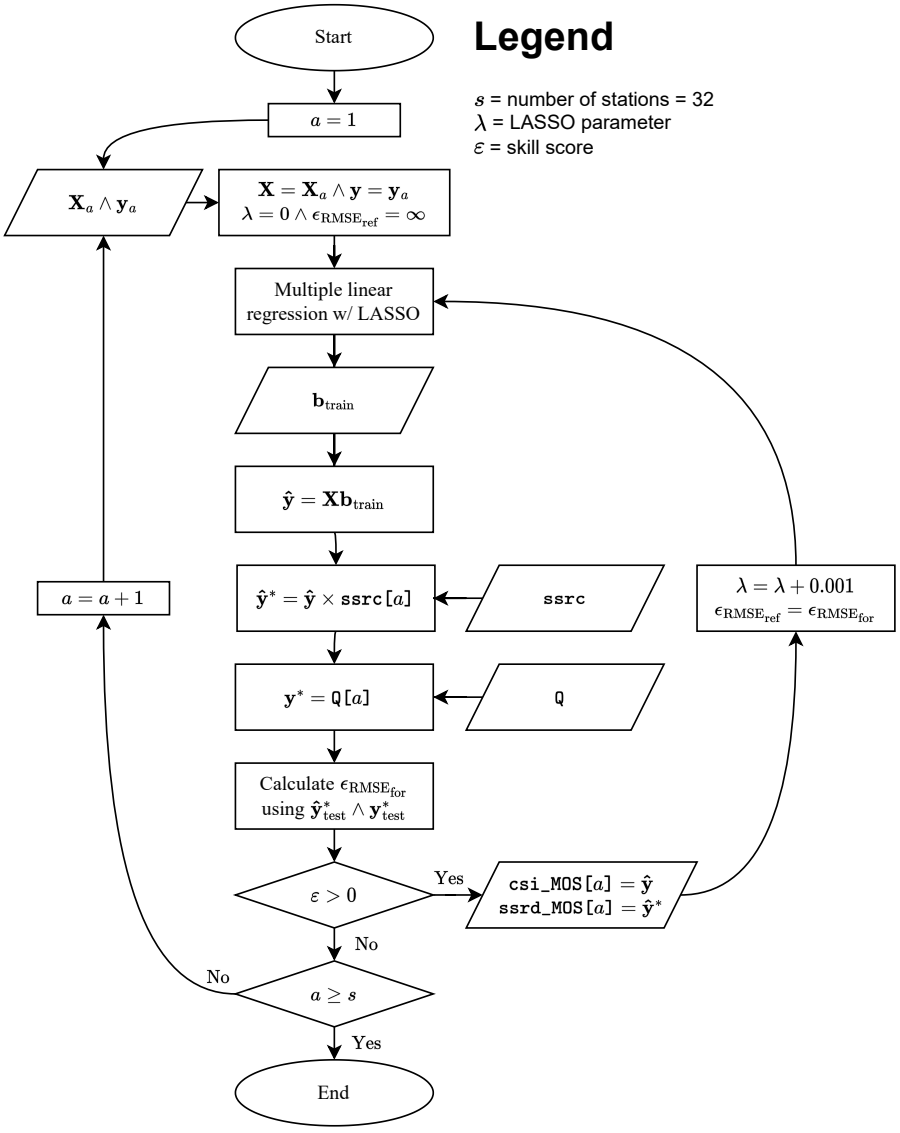


FIGURE 4.2
The workflow for the MOS correction's regression model. The steps that are undertaken within this flowchart are well-described in Section 4.2.2.

4.2.3 Example of application

We apply the framework of Figure 4.1 and Figure 4.2 to an example for the reader to gain a better understanding. We assess a simplified case. First, we

do the data processing as described in the workflow of Figure 4.1. Second, we conclude with the model testing, training, and validation of Figure 4.2.

4.2.3.1 Data processing

Say that we have three parameters: `sp`, `csi_predicted`, and `csi_observed`. We have three stations: station 1, 2, and 3. Thus, we have the dataset as given in Figure 4.3. Here, the columns on the top level denote the parameter and on the bottom level the station. The rows denote the timestamps.

	sp			csi_predicted			csi_observed		
	1	2	3	1	2	3	1	2	3
2019-07-01 05:00	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$	$x_{1,7}$	$x_{1,8}$	$x_{1,9}$
2019-07-01 06:00	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$x_{2,7}$	$x_{2,8}$	$x_{2,9}$
2019-07-01 07:00	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,6}$	$x_{3,7}$	$x_{3,8}$	$x_{3,9}$
...									
2020-12-31 10:00	$x_{4953,1}$	$x_{4953,2}$	$x_{4953,3}$	$x_{4953,4}$	$x_{4953,5}$	$x_{4953,6}$	$x_{4953,7}$	$x_{4953,8}$	$x_{4953,9}$
2020-12-31 11:00	$x_{4954,1}$	$x_{4954,2}$	$x_{4954,3}$	$x_{4954,4}$	$x_{4954,5}$	$x_{4954,6}$	$x_{4954,7}$	$x_{4954,8}$	$x_{4954,9}$
2020-12-31 12:00	$x_{4955,1}$	$x_{4955,2}$	$x_{4955,3}$	$x_{4955,4}$	$x_{4955,5}$	$x_{4955,6}$	$x_{4955,7}$	$x_{4955,8}$	$x_{4955,9}$

FIGURE 4.3
An example dataset for MOS correction. We have three parameters for three stations.

From there, we distil the predictors and response for station 1 as depicted in Figure 4.4. In addition, we apply standardization (Section 2.3.4), therefore we do not have a column of constants in $\mathbf{X}$ of Figure 4.4(a).

$\mathbf{X}_1 =$

	sp	csi_predicted
	1	1
2019-07-01 05:00	$x_{1,1}$	$x_{1,4}$
2019-07-01 06:00	$x_{2,1}$	$x_{2,4}$
2019-07-01 07:00	$x_{3,1}$	$x_{3,4}$
...		
2020-12-31 10:00	$x_{4953,1}$	$x_{4953,4}$
2020-12-31 11:00	$x_{4954,1}$	$x_{4954,4}$
2020-12-31 12:00	$x_{4955,1}$	$x_{4955,4}$

(a)

$\mathbf{y}_1 =$

	csi_observed
	1
2019-07-01 05:00	$x_{1,7}$
2019-07-01 06:00	$x_{2,7}$
2019-07-01 07:00	$x_{3,7}$
...	
2020-12-31 10:00	$x_{4953,7}$
2020-12-31 11:00	$x_{4954,7}$
2020-12-31 12:00	$x_{4955,7}$

(b)

FIGURE 4.4
Figure 4.4(a) depicts the predictors for MOS correction as distilled from Figure 4.3. We fit these predictors to the response of Figure 4.4(b).

4.2.3.2 Training, testing, and validation

To calculate the coefficient vector, b_{train} (Equation 2.13), for station 1, we select the rows 784 until 4152, which is the train set. Now that we have the coefficient vector for station 1 as given in Figure 4.5(a),[1] we can simply multiply b_{train} with $\mathbf{X}_1$ to predict the response, $\hat{y}$ (Equation 2.12), which are MOS corrected predictions as given in Figure 4.5(b).

$$\mathbf{b_{train}} = \begin{array}{|c|c|c|} \hline \texttt{sp} & 1 & b_1 \\ \hline \texttt{csi\_predicted} & 1 & b_2 \\ \hline \end{array}$$

(a)

$$\hat{\mathbf{y}} = \begin{array}{|c|c|} \hline & \texttt{csi\_MOS} \\ \hline & 1 \\ \hline \texttt{2019-07-01 05:00} & x_{1,1}b_1 + x_{1,4}b_2 \\ \hline \texttt{2019-07-01 06:00} & x_{2,1}b_1 + x_{2,4}b_2 \\ \hline \texttt{2019-07-01 07:00} & x_{3,1}b_1 + x_{3,4}b_2 \\ \hline \cdots & \\ \hline \texttt{2020-12-31 10:00} & x_{4953,1}b_1 + x_{4953,4}b_2 \\ \hline \texttt{2020-12-31 11:00} & x_{4953,1}b_1 + x_{4953,4}b_2 \\ \hline \texttt{2020-12-31 12:00} & x_{4953,1}b_1 + x_{4953,4}b_2 \\ \hline \end{array}$$

(b)

FIGURE 4.5
Figure 4.5(a) shows the vector of coefficients for station 1. We use this vector to predict the response as given in Figure 4.5(b), which is the MOS correction.

To tune the LASSO parameter, λ (Equation 2.14), we can test the MOS correction by taking from the predicted responses the rows 4153 until 4955, which is the test set, and calculating the RMSE (Equation 2.26). From there, we can tune the LASSO parameter over and over again until we reach the lowest RMSE. If, say, the LASSO calculates that $b_1 = 0$ and $b_2 = 1$, then we can simply remove `sp` from our predictor matrix, $\mathbf{X}_1$, as it has no predictive value.

To conclude, we apply this process as well to station 2 and 3 to end up with

[1] The coefficient vector does not have a constant, b_0 (Equation 2.7), as we apply standardization (Section 2.3.4).

a new parameter, `csi_MOS`. We can then validate our model by calculating the RMSE and other error metrics over the rows 1 until 783, which is the validation set.

4.3 Spatio-temporal regression

We used `csi_MOS` as an input to the spatio-temporal model in combination with `csi_observed` – we regressively smoothed MOS corrected predictions and lagged observations at the same time over the spatio-temporal dimension (Equation 2.24). Our aim was to decrease the RMSE (Equation 2.26) as much as possible.

4.3.1 Data processing

For the spatio-temporal analysis, we used the same set of predictors for each station of Figure 3.1. Therefore, we created one set of predictors and one set of responses instead of doing so separately for each station as was done for the MOS correction. This is depicted in the workflow of Figure 4.6, where we first loop over the amount of lags, p, with the iterator, $i = 0$, to lag `csi_observed` for each $i + \tau$ (Equation 2.24). From there, we loop over the amount of points to smooth on two sides, q, with the iterator, $i = 0$, to include `csi_MOS` for each step in the range of $[t + \tau - q, t + \tau + q]$. Due to the inclusion of lagged observations as predictors, we drop the steps where there are no lags available, which are the first p steps of each day. The same holds for the case where no numerical predictions are available, which are the first and last q steps of each day. We end up with one predictor matrix, $\mathbf{X}$, and one response matrix, $\mathbf{Y}$.

Here, the response consists of the parameter `csi_observed` and the predictors are those of Equation 2.24.

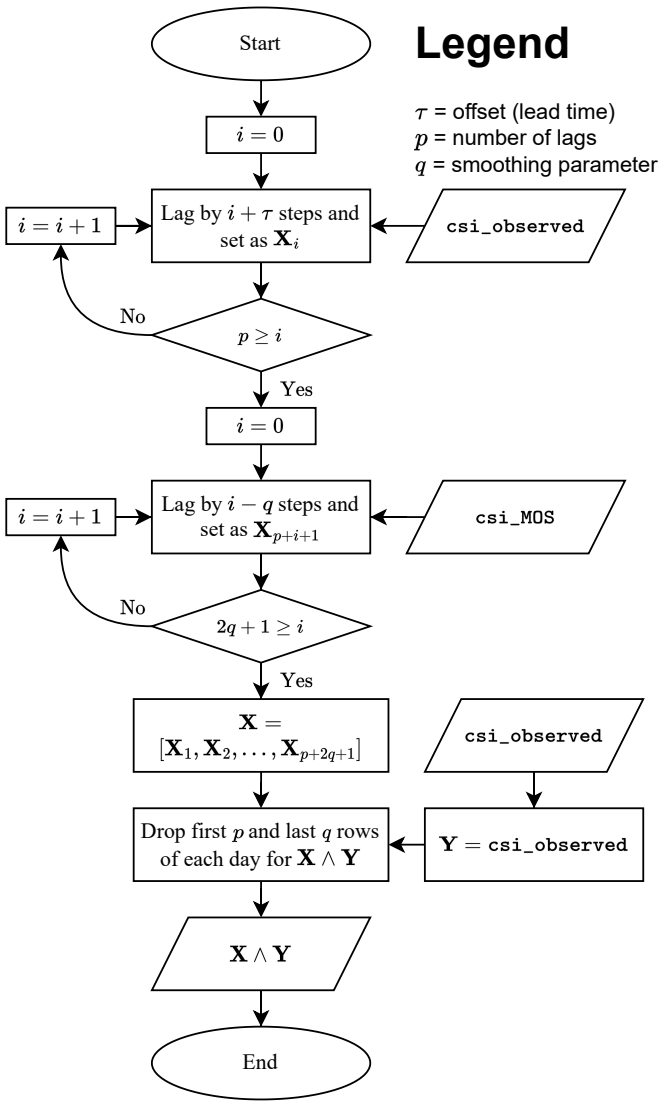


FIGURE 4.6
The workflow for the construction of the predictors and responses for the spatio-temporal regression. The lead time (τ), number of lags (p), and the smoothing parameter (q) can all be played with to develop different versions of the spatio-temporal model.

4.3.2 Training, testing, and validation

As a follow-up to the workflow of Figure 4.6, we used the generated sets to build a regression model as depicted in the workflow of Figure 4.7. As this workflow is almost equal to that of the MOS correction in Section 4.2.2, we only discuss its deviations here.

We retrieved the predictors, $\mathbf{X}$, and response, $\mathbf{y} = \mathbf{Y}[a]$, per station. Thus, we use the same $\mathbf{X}$ to predict each station's response. Again, we used iterations to determine the optimal λ (Equation 2.14) for the LASSO per station. When the model had converged to its most optimal point in terms of RMSE, the predicted response was saved to `csi_SP[`a`]` and `ssrd_SP[`a`]` as clear-sky index and irradiance respectively. After this was done for all stations, we would end up with two new parameters, `csi_SP` and `ssrd_SP`, which contain spatio-temporal solar forecasts as given in Table 4.2.

TABLE 4.2
The spatio-temporal solar forecast as calculated for each station.

Name	Description
`csi_SP`	Spatio-temporal solar forecast as clear-sky index
`ssrd_SP`	Spatio-temporal solar forecast as irradiance

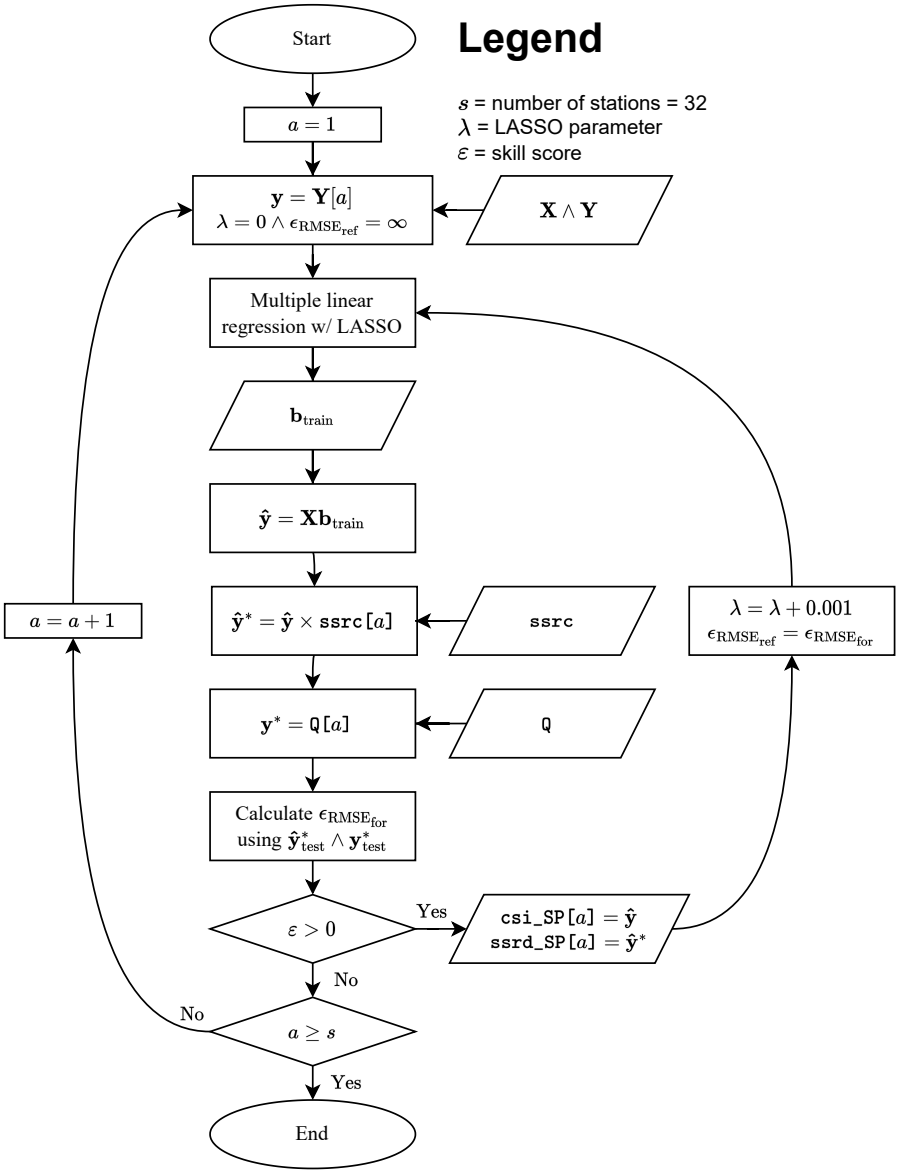


FIGURE 4.7
The workflow for the spatio-temporal regression model. The steps that are undertaken within this flowchart are well-described in Section 4.3.2.

4.3.3 Example of application

We apply the framework of Figure 4.6 and Figure 4.7 to an example for the reader to gain a better understanding. We assess a simplified case. First, we do the data processing as described in the workflow of Figure 4.6. Second, we conclude with the model testing, training, and validation of Figure 4.7.

4.3.3.1 Data processing

Say that we have observations and numerical predictions: `csi_observed` and `csi_MOS`. We have three stations: station 1, 2, and 3. Thus, we have the dataset as given in Figure 4.8. Here, the columns on the top level denote the parameter and on the bottom level the station. The rows denote the timestamps.

	csi_observed			csi_MOS		
	1	2	3	1	2	3
2019-07-01 05:00	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	$x_{1,5}$	$x_{1,6}$
2019-07-01 06:00	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$
2019-07-01 07:00	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$x_{3,4}$	$x_{3,5}$	$x_{3,6}$
...						
2020-12-31 10:00	$x_{4953,1}$	$x_{4953,2}$	$x_{4953,3}$	$x_{4953,4}$	$x_{4953,5}$	$x_{4953,6}$
2020-12-31 11:00	$x_{4954,1}$	$x_{4954,2}$	$x_{4954,3}$	$x_{4954,4}$	$x_{4954,5}$	$x_{4954,6}$
2020-12-31 12:00	$x_{4955,1}$	$x_{4955,2}$	$x_{4955,3}$	$x_{4955,4}$	$x_{4955,5}$	$x_{4955,6}$

FIGURE 4.8
An example dataset for the spatio-temporal regression. We have two parameters – observations and MOS corrected predictions – for three stations.

From there, we create the predictors and response for station 1 with two lags, $p = 2$, a smoothing parameter of one, $q = 1$, and a lead time of one hour, $\tau = 1$, as depicted in Figure 4.9. Here, for $\mathbf{X}$, the columns on the top level denote the parameter, the middle level the station, and the bottom level its time step. We aim to predict for time step $t + 1$, therefore we include the observations of time step t and $t - 1$ and the MOS corrected predictions of time step t, $t + 1$ and $t + 2$. In addition, we apply standardization (Section 2.3.4), therefore we do not have a column of constants in $\mathbf{X}$ of Figure 4.9(a).

We find that there are gaps in our $\mathbf{X}$ for the first p rows of each day and the last q rows of each day. These gaps come from the fact that we do not have any observations and/or predictions for those times. Therefore, we drop those rows from the predictor and response set when we apply regression.

$$\mathbf{X} =$$

	csi_observed						csi_MOS					
	1		2		3		1		2		3	
	t	$t-1$	t	$t-1$	t	$t-1$	t	$t+1$	t	$t+1$	t	$t+1$	$t+2$
2019-07-01 05:00							$x_{1,4}$	$x_{2,4}$		$x_{1,5}$	$x_{2,5}$		$x_{1,6}$	$x_{2,6}$
2019-07-01 06:00	$x_{1,1}$		$x_{1,2}$		$x_{1,3}$		$x_{1,4}$	$x_{2,4}$	$x_{1,5}$	$x_{2,5}$	$x_{1,6}$	$x_{2,6}$	$x_{3,6}$
2019-07-01 07:00	$x_{2,1}$	$x_{1,1}$	$x_{2,2}$	$x_{1,2}$	$x_{2,3}$	$x_{1,3}$	$x_{2,4}$	$x_{3,4}$	$x_{2,5}$	$x_{3,5}$	$x_{2,6}$	$x_{3,6}$	$x_{4,6}$
...													
2020-12-31 10:00							$x_{4953,1}$	$x_{4954,1}$	$x_{4953,2}$	$x_{4954,2}$	$x_{4953,3}$	$x_{4954,3}$	$x_{4953,3}$	$x_{4954,3}$
2020-12-31 11:00	$x_{4953,1}$		$x_{4953,2}$		$x_{4953,3}$		$x_{4953,1}$	$x_{4954,1}$	$x_{4953,2}$	$x_{4954,2}$	$x_{4953,3}$	$x_{4954,3}$	$x_{4954,3}$	$x_{4955,3}$
2020-12-31 12:00	$x_{4954,1}$	$x_{4953,1}$	$x_{4954,2}$	$x_{4953,2}$	$x_{4954,3}$	$x_{4953,3}$	$x_{4954,1}$	$x_{4955,1}$	$x_{4954,2}$	$x_{4955,2}$	$x_{4954,3}$	$x_{4955,3}$		

(a)

$\mathbf{Y} =$

	csi_observed		
	1	2	3
2019-07-01 05:00	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
2019-07-01 06:00	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$
2019-07-01 07:00	$x_{3,1}$	$x_{3,2}$	$x_{3,3}$
...			
2020-12-31 10:00	$x_{4953,1}$	$x_{4953,2}$	$x_{4953,3}$
2020-12-31 11:00	$x_{4954,1}$	$x_{4954,2}$	$x_{4954,3}$
2020-12-31 12:00	$x_{4955,1}$	$x_{4955,2}$	$x_{4955,3}$

(b)

FIGURE 4.9
Figure 4.9(a) depicts the predictors for the spatio-temporal model with two lags, $p = 2$, a smoothing parameter of one, $q = 1$, and a lead time of one hour, $\tau = 1$, as distilled from Figure 4.8. We fit those predictors to the responses of Figure 4.9(b)

4.3.3.2 Training, testing, and validation

For the model selection, we apply the same logic as done in Section 4.2.3.2. We select from the responses, $\mathbf{Y}$, the response for station 1, $\mathbf{y} = \mathbf{Y}[1]$, and we find the coefficient vector, $\mathbf{b}_{\text{train}}$, as given in Figure 4.10(a),[2] and we multiply this vector with $\mathbf{X}$. We end up with the predicted response, $\hat{\mathbf{y}}$ (Equation 2.12), for station 1 as given in Figure 4.10(b). The predicted response only includes the rows for which $\mathbf{X}$ had no gaps. We have left the last three rows of Figure 4.10(b) as an exercise for the reader.

[2]The coefficient vector does not have a constant, b_0 (Equation 2.7), as we apply standardization (Section 2.3.4).

$\mathbf{b}_{\text{train}} =$

csi_observed	1	t	b_1
		$t-1$	b_2
	2	t	b_3
		$t-1$	b_4
	3	t	b_5
		$t-1$	b_6
csi_MOS	1	t	b_7
		$t+1$	b_8
		$t+2$	b_9
	2	t	b_{10}
		$t+1$	b_{11}
		$t+2$	b_{12}
	3	t	b_{13}
		$t+1$	b_{14}
		$t+2$	b_{15}

(a)

$\hat{\mathbf{y}} =$

	csi_SP
	1
2019-07-01 07:00	$x_{2,1}b_1+\ldots+x_{4,6}b_{15}$
2019-07-01 08:00	$x_{3,1}b_1+\ldots+x_{5,6}b_{15}$
2019-07-01 09:00	$x_{4,1}b_1+\ldots+x_{7,6}b_{15}$
...	
2020-12-01 11:00	...
2020-12-02 11:00	...
2020-12-03 11:00	...

(b)

FIGURE 4.10
Figure 4.10(a) shows the vector of coefficients for station 1. We use this vector to predict the response as given in Figure 4.10(b), which is the spatio-temporal model.

To conclude, we tune the LASSO parameter as done in Section 4.2.3.2, we apply this process to stations 2 and 3 to end up with a new parameter, `csi_SP`, and we validate our model on the validation set.

Part II

Spatio-temporal regression

5

Results

CONTENTS

KEY TAKEAWAYS

The MOS correction improved the accuracy of the numerical irradiance predictions by 8% on the test set. When we compare this to the validation set, we find an improvement of 4%. This is due to the test set spanning a period of winter, whereas the validation set spans a period of summer. The clear-sky index for the test set is lower on average, which is where the MOS correction proves to be most effective.

For the spatio-temporal regression, we find the model with a lag of one and a smoothing parameter of one to perform best. This entails that the model only uses one lagged observation – the one at t – and includes MOS corrected predictions for $t+\tau-1$, $t+\tau$, and $t+\tau+1$ to predict $t+\tau$. For a lead time of one hour, the spatio-temporal model increases the accuracy of the numerical irradiance predictions by 30% for the test set and 25% for the validation set. When set out against the MOS correction, we find the spatio-temporal model to be as accurate for the validation set as for the test set.

By analysing the weights of the stations as a predictor for another station, we find that stations in proximity have higher weights – there is a relationship between distance and weight. When we assess the overall importance per

station as a predictor for other stations, then we find that stations in the south-west of the Netherlands have most predictive value.

5.1 Introduction

This chapter is concerned with the results that come from the application of Part I. Here, we actually study the accuracy of the solar power forecasts that come from the regression framework as defined in Chapter 4.

We use scatter density plots to summarize large datasets in compact plots. Figure 5.1(a) is such a plot, where the colour-bar on the right indicates the density of the scatters. We chose to use these plots as they show us what happens at the places where the scatters are so dense that it becomes impossible to draw conclusions.

We express accuracy for our models in the form of SS (Equation 2.27). When we calculate the SS, we use `ssrd` as the reference forecast unless explicitly stated otherwise. For example, when we would use `ssrd_MOS` as the reference forecast, then we would state that we calculated the SS against `ssrd_MOS`.

5.2 MOS correction

The MOS correction was carried out as described by Figure 4.1 and Figure 4.2. By using the LASSO (Equation 2.14), we selected only the predictors that contributed to a lower RMSE (Equation 2.26). From there, the result, `ssrd_MOS`, was evaluated and compared against `ssrd` over the test set. To conclude, the MOS correction was evaluated over the validation set as well and compared to the test set.

5.2.1 Predictor selection

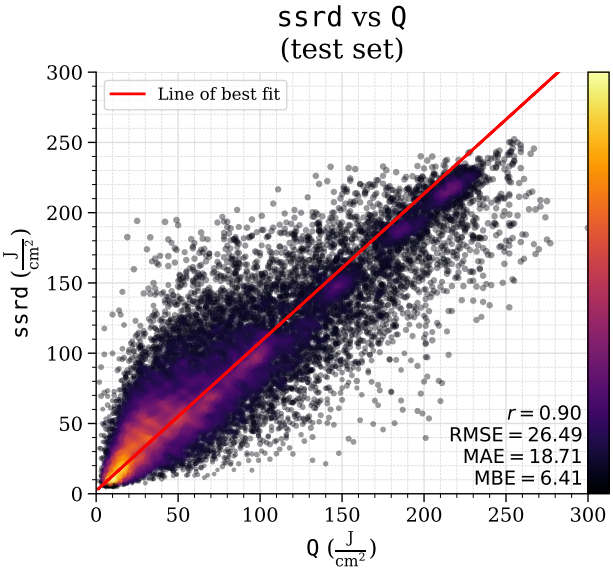
For the MOS correction, the following predictors proved to be useful to decrease the RMSE:

- `cdir`
- `coszenith`
- `csi_predicted`
- `csi_predicted_squared`
- `csi_predicted_cubed`
- `dummy_hour_5` until `dummy_hour_18`
- `sp`

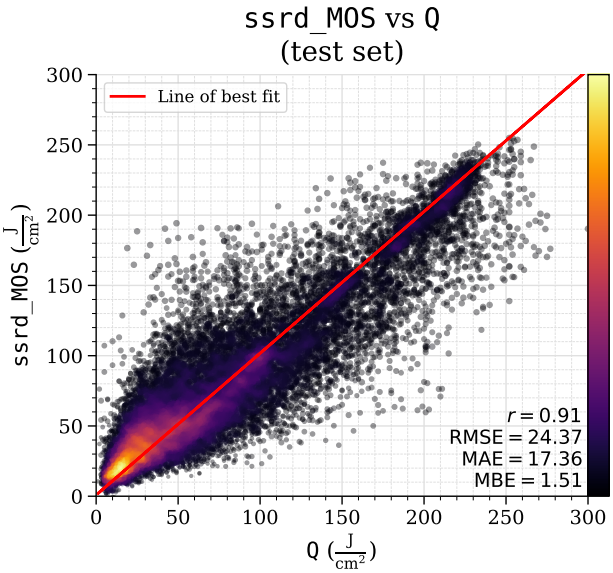
When evaluating the coefficients, we find that `cdir`, `coszenith`, and the dummy variables all act as a general correction on the predicted clear-sky index. The `csi_predicted`, `csi_predicted_squared`, and `csi_predicted_cubed` are used to distil the non-linear relationship between `csi_predicted` and `csi_observed`. Finally, `sp` contains valuable information to increase prediction accuracy. The addition of other predictors increased the RMSE – they did not seem to have any valuable information in them.

5.2.2 Accuracy

The `ssrd` and `ssrd_MOS` were evaluated in Figure 5.1(a) and Figure 5.1(b) respectively. When we compare the two for accuracy, we find an SS of 8% – that is, the RMSE decreases by 8% via the MOS correction. When we assess the scatter density plots of Figure 5.1 more thoroughly, we find that `ssrd_MOS` has less bias and less variance. To conclude, the MOS correction has helped to increase the fit between `ssrd` and `Q`.



(a)



(b)

FIGURE 5.1
Figure 5.1(a) shows `ssrd` plotted against `Q` for all stations in the test set. Figure 5.1(b) shows the applied MOS correction against `Q` in the test set. We find that the MOS correction indeed reduces the overall bias and variance of the forecast.

5.2.3 Validation

When we assess the SS of `ssrd_MOS` against `ssrd` for the test and validation set in Figure 5.2, we find that overall, the validation set performs worse with an SS of around 4% compared to 8% for the test set. In addition, there seems to be a weak relation ($r = 0.42$ of Equation 2.3) for the SS per station between the test and validation.

An explanation for the MOS correction's underperformance on the validation set is that the validation set spans a period of summer (in 2019), whereas the test set spans a period of winter (in 2020). It could be that MOS correction with the chosen predictors is less effective in the summer than the winter. This might be due to the fact that in the winter the predicted clear-sky index, `csi_predicted`, is on average lower than in the summer, which might be where the MOS correction is most effective.

It is also of interest to note that the MOS correction model was trained on a set that mostly spans 2020, where there was unusual weather (van Heerwaarden et al., 2021). The train set had an 18.9% increase in surface irradiance with respect to the 2010 until 2019 mean, which is explained by a low median aerosol optical depth, several exceptionally dry days, and a very low cloud fraction overall. However, as we are using the clear-sky index to make our forecasts, this should be accounted for.

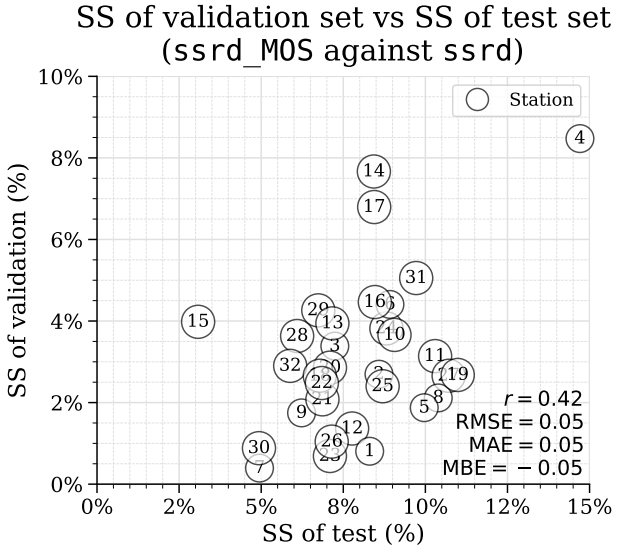


FIGURE 5.2
SS of the MOS correction per station of Figure 3.1 on the test and validation set. The MOS correction is less effective on the validation set than the test set.

5.3 Spatio-temporal model

For the spatio-temporal analysis, we first optimized for a lead time of one hour the lag, p, and smoothing, q, of Equation 2.24 as described in Figure 4.6 and Figure 4.7. Once we found an optimal value for p and q in terms of SS (Equation 2.27), we tested the spatio-temporal model for lead times up to 6 hours. To conclude, we conducted a spatial analysis, which entails the study of the coefficients (Equation 2.13) of the spatio-temporal model.

5.3.1 Selection of lag and smoothing

To test the effectiveness of the spatial component on the lag, p, and smoothing, q, we first ran the model without the spatial component by selecting for each station only the predictors that belonged to that station, $\mathbf{X} = \mathbf{X}[a]$ (Section 4.1). We call this approach temporal regression, or an AR model as described in Equation 2.24 with $R = 1$. From there, we used the entire $\mathbf{X}$ as the predictors for each station, which includes the spatial component as originally described in Figure 4.7. We call this spatio-temporal regression, which is a VAR model as described in Equation 2.24 with $R = 32$.

5.3.1.1 Temporal regression

To test if the spatial component improves the model's accuracy, we first test an AR model – that is, we only use predictors that belong to the station that we aim to predict – for different values of lag, p, and smoothing, q, as in Equation 2.24 with $R = 1$. As this reduces the amount of predictors to $p + 2q + 1$, we do not apply LASSO ($\lambda = 0$ in Equation 2.14), as collinearity becomes less of an issue. Figure 5.3 provides our findings.

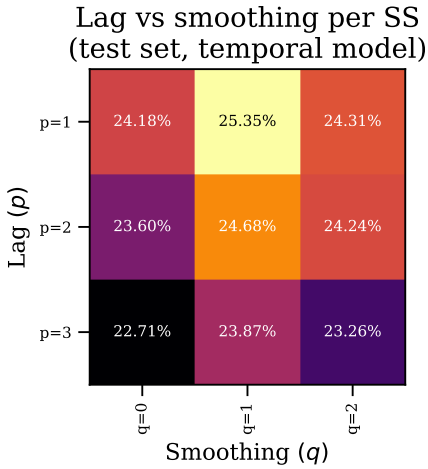


FIGURE 5.3
SS of the temporal model against lag and smoothing for a lead time of one hour. We find that the optimum lies at $p = 1$ and $q = 1$.

The model performs best when there is a lag of $p = 1$ and when the smoothing parameter is set to $q = 1$. When we increase p and/or q from there, we find the regression model to overfit – it misidentifies random occurrences as patterns.

5.3.1.2 Spatio-temporal regression

We now include the spatial component – that is, we use the predictors from all stations to predict each station – for different values of lag, p, and smoothing, q as in Equation 2.14 with $R = 32$. Figure 5.4 provides our findings.

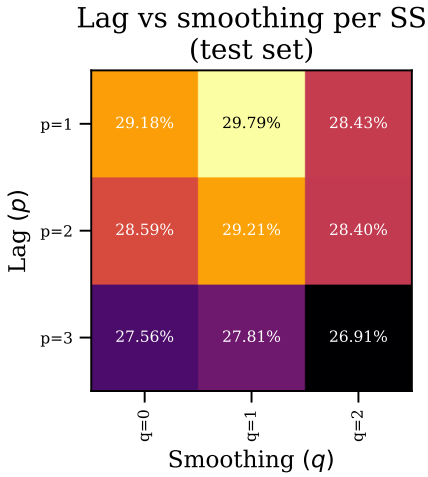


FIGURE 5.4
SS of the spatio-temporal model against lag and smoothing for a lead time of one hour. We find that the optimum lies at $p = 1$ and $q = 1$.

The model performs best when there is a lag of $p = 1$ and when the smoothing parameter is set to $q = 1$. When we increase p and/or q from there, then the amount of predictors becomes larger. As the amount of rows in the train set does not increase, the optimization of the coefficients (Equation 2.13) becomes less accurate due to the decrease in the degrees of freedom.[1] This is visible in Figure 5.4 – when p and/or q increase beyond its optimal point, then the SS decreases.

5.3.2 Performance for different lead times

The conclusion drawn from Figure 5.3 and Figure 5.4 – that $p = 1$ and $q = 1$ has the best SS – holds for all tested lead times (τ in Equation 2.24) up to six hours. Figure 5.5 shows the lead time of the temporal model, spatio-temporal model, and the MOS correction against the SS for $p = 1$ and $q = 1$. As the lead time increases, the SS decreases, up to a point that the MOS correction outperforms the spatio-temporal model – beyond five hours, the spatio-temporal model is ineffective due to the large number of predictors (as explained in Section 5.3.1.2). In contrast, the temporal model outperforms the other models at five hours and beyond due to the smoothing of the MOS corrections over the temporal dimension.

[1] Degrees of freedom encompasses the notion that the amount of independent information you have limits the number of parameters that you can estimate. Typically, the degrees of freedom equals your set size minus the number of parameters you need to calculate.

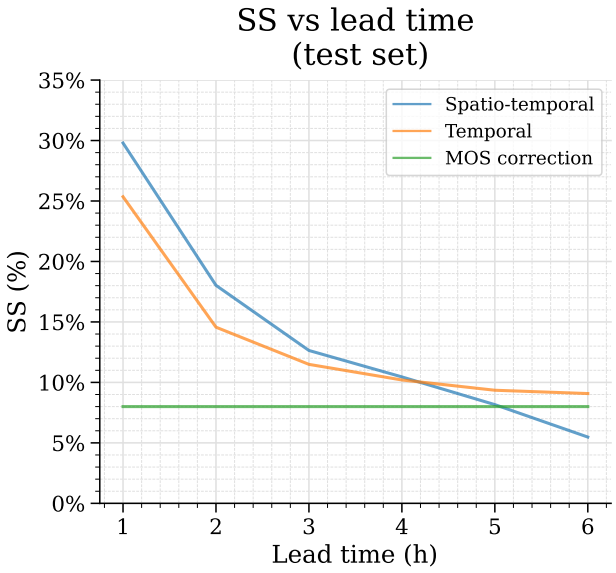
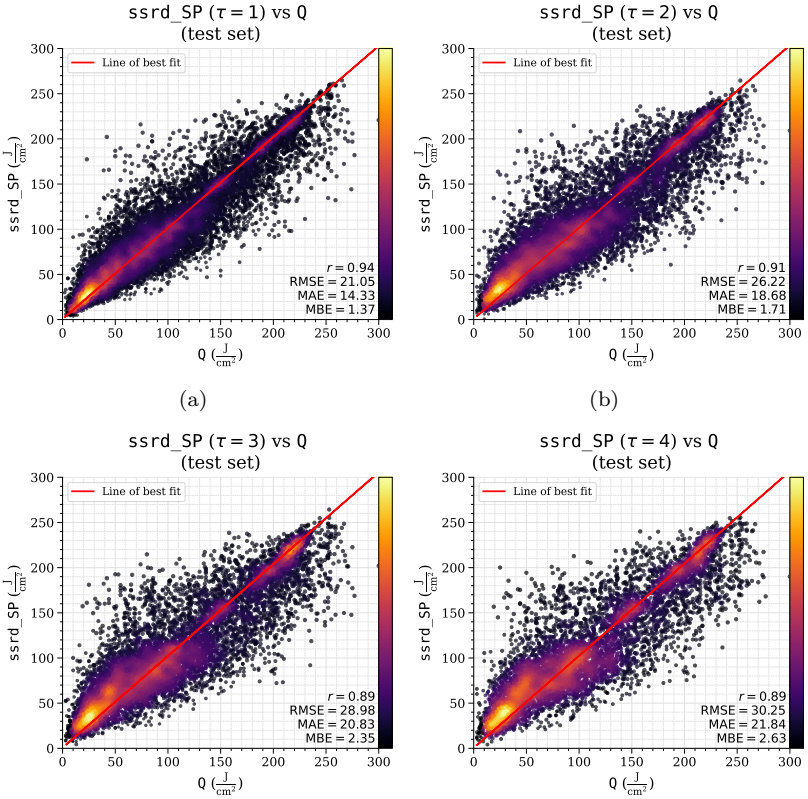


FIGURE 5.5
SS of the the temporal model, spatio-temporal model, and the MOS correction against lead time. We find that the spatio-temporal model is effective up to a lead time of five hours. From there, the temporal model becomes most effective.

Figure 5.6 shows four different scatter plots for the model where $p = 1$ and $q = 1$ with the lead times varying from one hour to four hours. When compared to Figure 5.1, we see that as the lead times increases, the MOS correction becomes more prominent. When we analyse the coefficients, we find that the relative weight of the MOS correction increases compared to the lagged observations as the lead time increases. This indicates that as the lead time increases, the lagged observations have less predictive value.



(a) (b)

(c) (d)

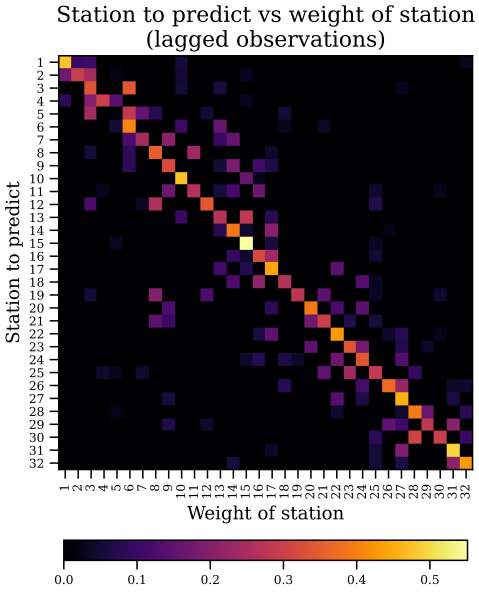
FIGURE 5.6
Scatter density plots of the spatio-temporal forecast against observations for a lead time of one to four hours. We find that the lagged observations are overtaken by the MOS correction as the lead time increases.

5.3.3 Spatial analysis

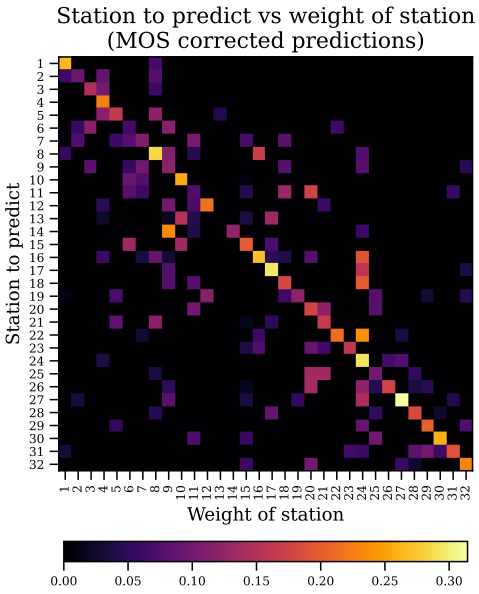
For the spatial analysis, the coefficients (Equation 2.13) of the spatio-temporal model with a lead time of one hour were analysed. The absolute values of the model's coefficients were used as a measure of predictive importance – e.g., if station 1 has a high weight in the coefficient matrix of station 2, then station 1 has high predictive importance for station 2. These coefficients were analysed in a heatmap and in a spatial map for the stations of Figure 3.1.

5.3.3.1 Weights of coefficients per station

Figure 5.7 gives a heatmap of each station's weights as a predictor for another station. Here, we made the distinction between the lagged observations, `csi_observed`, and the predictions to smooth, `csi_MOS`. For the lagged observations, we identify the pattern that stations close to one-another generally share predictive importance. For the predictions, we find that there exists a small pattern related to distance, but it appears to be scattered out more. Thus, the relation between predictive importance and distance is stronger for lagged observations than for MOS corrected predictions.



(a)



(b)

FIGURE 5.7
Figure 5.7(a) shows a heatmap for the weights of the lagged observations per station of Figure 3.1. We observe a pattern where stations close to one-another have increased predictive importance. Figure 5.7(b) shows a heatmap for the weights of the MOS corrected predictions, where the proximity pattern is less prominent.

5.3.3.2 Predictive importance per station

Figure 5.8 gives the sum over the y-axis of Figure 5.7 – it tells us how important each station is as a predictor for others. We find a clear pattern for the lagged observations, where irradiance travels from the south-west across the rest of the Netherlands. For the MOS corrected predictions, there is less of a pattern: stations that have more stations in proximity seem to have higher predictive importance. Thus, the lagged observations seem to indicate a spatial pattern of the weather whereas the MOS corrected predictions simply indicate a spatial average.

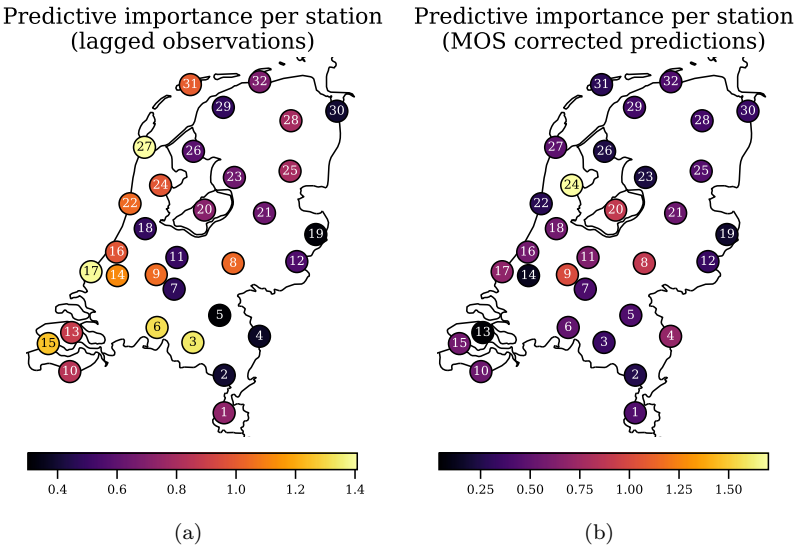


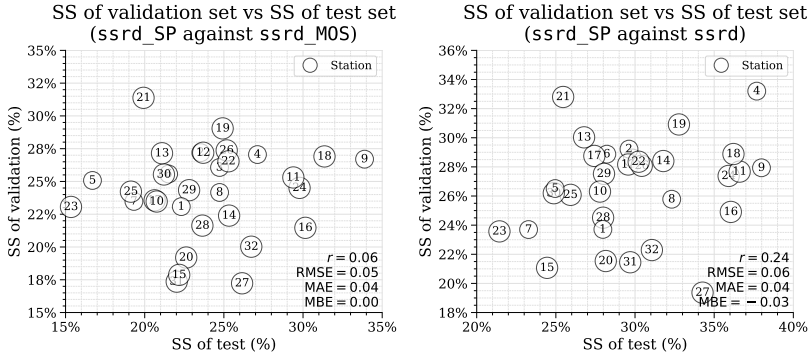
FIGURE 5.8
Figure 5.8(a) shows a spatial map of each numbered station's importance as a predictor using the lagged observations. The south-west is highlighted. Figure 5.8(b) shows such a map using the MOS corrected predictions – there is less of a pattern.

5.3.4 Validation

To test the accuracy of the spatio-temporal model with a lead time of one hour, we calculate its SS (Equation 2.27) for the test and validation set. Figure 5.9(a) gives the SS against `ssrd_MOS` and Figure 5.9(b) against `ssrd`. When we analyse Figure 5.9(a), we find that the spatio-temporal model works as well on the test set as the validation set. However, when we analyse Figure 5.9(b), we find that test set outperforms the validation set, which comes from the

MOS correction (Figure 5.2) that serves as an input to the spatio-temporal model (Figure 4.6).

In Figure 5.9(a), there seems to be no relation ($r = 0.06$ of Equation 2.3) between the SS of the test set and the validation set. Thus, the relation that is present in Figure 5.9(b), where $r = 0.24$, comes from the MOS correction as given in Figure 5.2. The reason for not finding a pattern in SS per station for the spatio-temporal model is most likely due to a difference in weather patterns between the test and validation set (van Heerwaarden et al., 2021). Furthermore, the test set spans a period of winter while the validation set a period of summer. Nevertheless, as the MBE in Figure 5.9(a) is zero, we can rely on the spatio-temporal model to have on average an SS against `ssrd_MOS` of around 24%.



(a) (b)

FIGURE 5.9
Figure 5.9(a) shows the SS of the spatio-temporal model against the MOS correction per station of Figure 3.1 on the test and validation set. The spatio-temporal model has an average accuracy of around 24%. Figure 5.9(b) shows the SS of the spatio-temporal model against `ssrd`. We see the pattern of Figure 5.2 – where the test set outperforms the validation set – coming through.

6

Discussion

CONTENTS

KEY TAKEAWAYS

We test the accuracy of the spatio-temporal model for different cloud conditions by doing a case study on a daily basis for a location in the centre of the Netherlands. We find the model to perform well in situations where the MOS corrected predictions are off about the cloud conditions, thereby greatly increasing the accuracy of the predictions.

To ensure that our results are valid, we compare them against state-of-the-art intraday solar power forecasts. For comparison, we cannot find a case study of the Netherlands in literature, which indicates that we fill a gap by doing one. Nevertheless, we are able to find comparable models with metrics that we can use to validate our model's accuracy against. Through such validation, we believe our model to be accurate.

6.1 Accuracy for different cloud conditions

To determine the accuracy of the spatio-temporal model with a lead time of one hour for different cloud conditions, we did a case study of the most central station of Figure 3.1 – station number 9. We calculated the SS against the MOS correction at a daily scale, and we took the four top performers from the validation set, which has many summer days with long daytime. We used satellite imagery[1] to assess the cloud conditions per day.

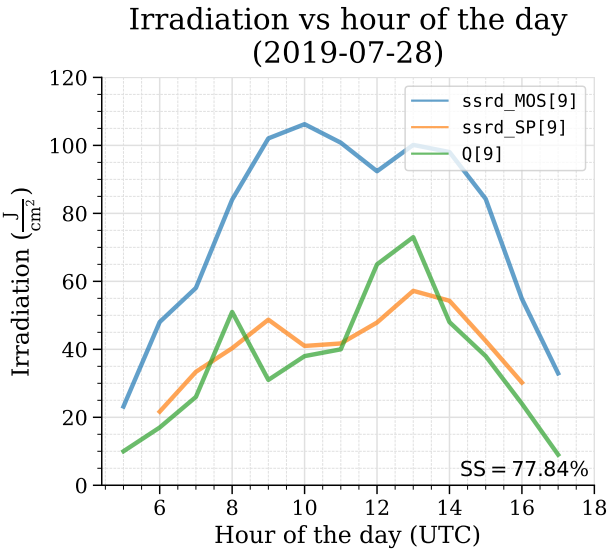
Overall, we conclude that the spatio-temporal model performs well for different cloud conditions. It corrects where the MOS corrected predictions are off about the cloud conditions, thereby greatly increasing the accuracy of the predictions.

6.1.1 Thick cloud deck

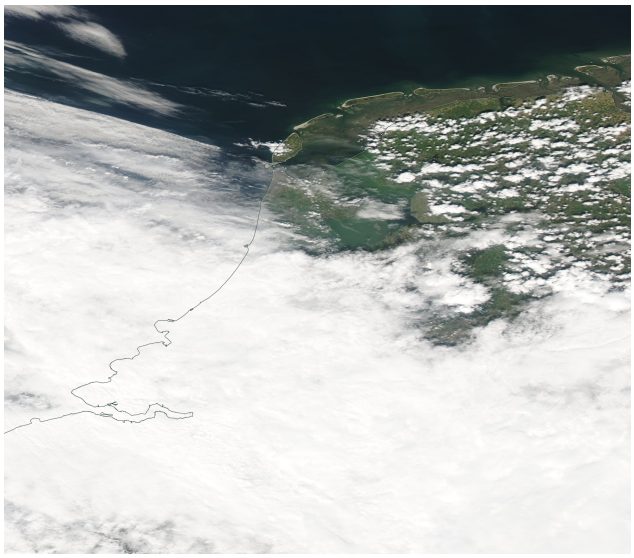
For the first case, we assess the day in the validation set for which the spatio-temporal model has the highest SS against the MOS correction – 78%. Figure 6.1(a) gives the hourly irradiance and Figure 6.1(b) a satellite photo, which was made at 12:45 UTC.

The MOS corrected prediction overpredicts the irradiance quite heavily. The spatio-temporal model corrects for this by taking into account the lagged observations and smoothed MOS corrections over the spatio-temporal dimension. By doing so, it is able to correct the MOS corrected prediction's inability to foresee the thickness of the cloud deck.

[1] We acknowledge the use of imagery from NASA's Worldview application (https://worldview.earthdata.nasa.gov), part of NASA's Earth Observing System Data and Information System (EOSDIS).



(a)



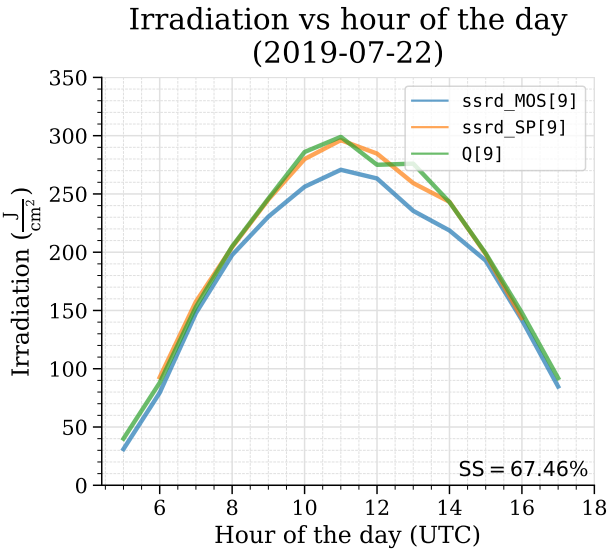
(b)

FIGURE 6.1
Figure 6.1(a) depicts the `ssrd_MOS`, `ssrd_SP`, and `Q` for station number 9 of Figure 3.1 plotted against time for 2019-07-28. Figure 6.1(b) shows a satellite photo for that day made at 12:45 UTC. We find a thick cloud deck, which the MOS corrected prediction underestimates.

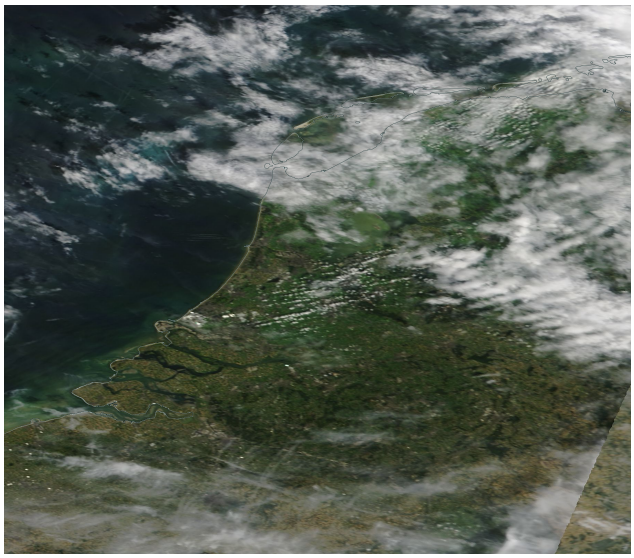
6.1.2 Clear-sky

For the second case, we calculated an SS for the day against the MOS correction of 67%. Figure 6.2(a) gives the hourly irradiance and Figure 6.2(b) a satellite photo, which was made at 11:35 UTC.

This day is close to a clear-sky day. We find the MOS corrected prediction to be off as it overestimates the effect of aerosols and/or clouds. The spatio-temporal model corrects for this as it follows the observations smoothly. When we look at the satellite photo, we see some negligible clouds around the centre of the Netherlands where station number 9 is located, which could have been overestimated by the MOS corrected prediction.



(a)



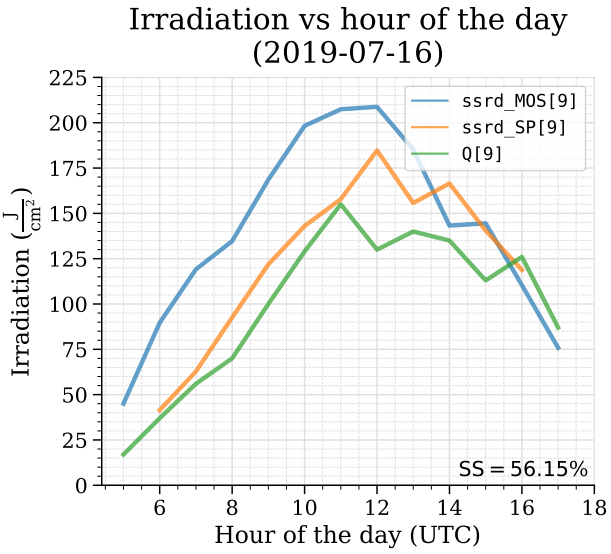
(b)

FIGURE 6.2
Figure 6.2(a) depicts the `ssrd_MOS`, `ssrd_SP`, and `Q` for station number 9 of Figure 3.1 plotted against time for 2019-07-28. Figure 6.2(b) shows a satellite photo for that day made at 11:35 UTC. We find a clear-sky day, whereas the MOS corrected prediction underestimates the irradiance.

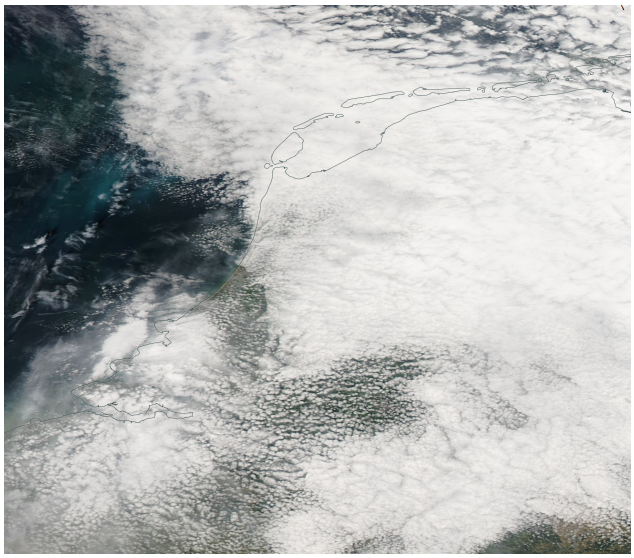
6.1.3 High cloud variability

We assess a day when there was a high cloud variability. The spatio-temporal model performs with a daily SS of 56% against the MOS correction. Figure 6.3(a) gives the hourly irradiance and Figure 6.3(a) a satellite photo, which was made at 12:20 UTC.

Due to the high cloud variability, we find the MOS corrected prediction to overpredict the irradiance, whereas the spatio-temporal model is able to follow the trend quite nicely till midday. From midday forward, the accuracy between the MOS correction and the spatio-temporal model starts to match. When we look at the satellite picture, which was taken around midday, we see a boundary between two types of clouds for the centre of the Netherlands, which is where station number 9 is located. This might explain why the MOS corrected prediction is off, as it is not able to predict that boundary very precisely.



(a)



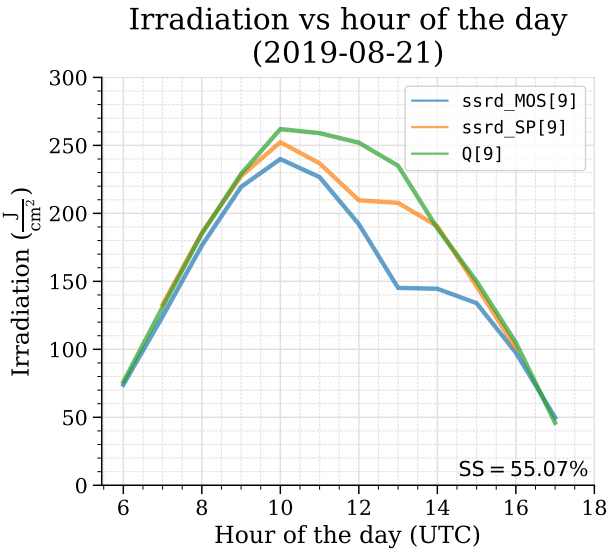
(b)

FIGURE 6.3
Figure 6.3(a) depicts the `ssrd_MOS`, `ssrd_SP`, and `Q` for station number 9 of Figure 3.1 plotted against time for 2019-07-16. Figure 6.3(b) shows a satellite photo for that day made at 12:20 UTC. We find high cloud variability, which the MOS corrected prediction underestimates.

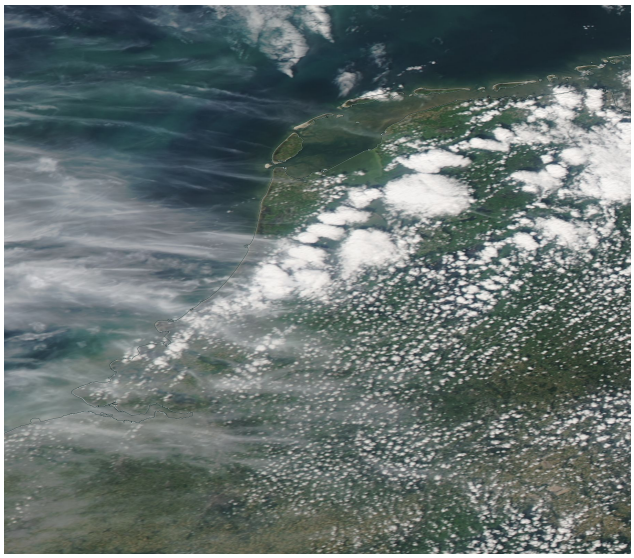
6.1.4 Low cloud variability

To finalize our case analysis, we assess a day when there was low cloud variability. On this day, the spatio-temporal model has an SS of 55% against the MOS correction. Figure 6.4(a) gives the hourly irradiance and Figure 6.4(a) a satellite photo, which was made at 11:55 UTC.

We find the MOS corrected prediction to predict some clouds after midday. However, as we see from the observations, these clouds never occurred. Although we see the effect of these clouds in the spatio-temporal model, we find it to be dampened and shifted. This is what the spatio-temporal model does: it uses lagged observations and smoothed MOS corrections over the spatio-temporal dimension to determine the likelihood of the occurrence of clouds. When we look at the satellite photo, we do find that there were lots of small clouds, which creates the rationale for the MOS correction's predicted clouds after midday.



(a)



(b)

FIGURE 6.4
Figure 6.4(a) depicts the `ssrd_MOS`, `ssrd_SP`, and `Q` for station number 9 of Figure 3.1 plotted against time for 2019-08-21. Figure 6.4(b) shows a satellite photo for that day made at 11:55 UTC. We find low cloud variability, whereas the MOS corrected prediction estimates a higher one.

6.2 Comparison to state-of-the-art

To conclude our discussion, we compare our spatio-temporal model with state-of-the-art intraday solar power forecast models. To do so, we compare against David et al. (2018), which gives a detailed summary of papers that have developed methods for intraday solar power forecasts. It describes and tests state-of-the-art point and probabilistic models to forecast solar power for one hour ahead. Among others, it uses a neural network[2] (NN) for point forecasts. For us to compare our results, the paper expresses the RMSE (Equation 2.26) and the MAE (Equation 2.29) as a percentage of the response's mean, $\mu_{\mathbf{y}}$ (Equation 2.1), which we call the %rRMSE and %rMAE respectively.

The paper concludes that forecasting of the clear sky index is the most effective method when compared to alternatives. It uses one year of data to train the models as we did in our research as well. The difference lies in the fact that their models do not use exogenous variables – that is, numerical weather predictions. Therefore, we only focus on the forecast with a lead time of one hour as the lagged observations are most important in that context.

For point forecasts, the paper finds for the NN with a lead time of one hour an %rRMSE of 22% and an %rMAE of 16%. For our spatio-temporal model with a lead time of one hour, we find an %rRMSE of 23% and an %rMAE of 16%. We cannot compare these results one-on-one, as these models were tested for different locations and different years. However, we can conclude that we have results that are probable.

In Aguiar et al. (2016), which uses an NN with lagged observations, numerical predictions from the ECMWF, and satellite-derived data, we find an %rRMSE of 24% for a lead time of one hour. Again, we cannot compare these results directly to ours, however, we can conclude that our results are probable.

When we try to find papers for direct comparison that conduct a case study of the Netherlands for intraday solar power forecast models, then we cannot find any. Therefore, we conclude that we fill a research gap by doing one.

[2]Neural networks are a set of non-linear algorithms.

7

Conclusion

CONTENTS

KEY TAKEAWAYS

Time series forecasts of solar power can be made by taking lagged observations and smoothing numerical predictions. To improve the fit of the numerical predictions to the observations, the predictions can be pre-processed via a MOS correction – a statistical process that aims to remove systematic errors. We find that the regression of lagged observations in combination with the smoothing of MOS corrected predictions works well to create accurate intraday solar power forecasts as the lagged observations provide the state and the smoothed numerical predictions the trend. The accuracy of the numerical irradiance predictions is improved between 4% and 8% by the MOS correction and between 25% and 30% by the spatio-temporal model.

Our research aimed to make two contributions: to provide a framework for the spatio-temporal regression of observations and numerical weather predictions; and to provide a case study of the Netherlands to understand the behaviour of irradiance over the spatio-temporal dimension. These contributions prove to be useful as a stepping stone for further research.

Topics for further research are:

- the application of spatio-temporal weather forecasting to other contexts;
- the incorporation of other numerical weather models besides the ECMWF;
- the application of non-linear models besides regression; and

- the application of our accurate solar power forecasts to energy trading.

7.1 Answers to the research questions

Solar power forecasts that use regression to combine numerical weather predictions and observations can increase the accuracy of numerical irradiance predictions by around 30% for a lead time of one hour. Hence, we conclude that our regression framework creates accurate intraday solar power forecasts across the Netherlands. The design and implementation of this framework is the focus of this thesis, and therefore the main research question is formulated as: *How can irradiance observations and numerical weather predictions be regressed on the spatio-temporal dimension to create accurate intraday solar power forecasts?*

7.1.1 Sub-research questions

To answer that question, we first address our sub-research questions.

1. What regression techniques for time series are applicable to a spatio-temporal context?

Time series forecasting can be done by using lagged observations. This is due to the fact that those type of observations exhibit a form of correlation. However, time series cannot be used as is – they first need to be stationary, which means that they need to be relieved from any predictable patterns. In the case of solar power, there is a diurnal pattern, which follows the solar zenith angle and can be modelled by the clear-sky irradiance. When we remove this pattern, we are left with the clear-sky index. In addition to using lagged observations, numerical irradiance predictions can be converted to a clear-sky index and smoothed over the temporal dimension as well.

For the spatial dimension, geographically dispersed observations carry value as weather situations move over space and time. For numerical predictions, smoothing can not only be done over the temporal dimension but over the spatial as well. Therefore, in the context of solar power forecasts, regression that includes the spatial dimension for observations and numerical predictions is more accurate than those that do not.

Finally, when the set of spatio-temporal predictors are largely correlated, then predictor selection is required. This can be done by applying the LASSO technique, which enforces a penalty on the coefficients so that some shrink to zero, effectively negating them from the model.

2. How can regression increase the fit of numerical irradiance predictions to irradiance observations?

Model Output Statistics (MOS) serves this purpose. We applied MOS via a multiple linear regression model, where weather parameters from NWP in

combination with some dummy and polynomial predictors were used to predict the response of the observed clear-sky index. We used the clear-sky index as we wanted MOS to purely focus on systematic errors in the stochastic process.

To determine what predictors are useful for MOS, we applied a LASSO for predictor selection. We used a set of error metrics to tune the LASSO as to have it perform as accurate as possible.

We applied MOS to each location – the stations of Figure 3.1 – separately, where we used only predictors for that location. Furthermore, we found LASSO to roughly select the same parameters for each location. The predicted responses of the MOS correction were used as an input to the spatio-temporal model, where they were smoothed over the spatio-temporal dimension.

3. How can regression combine observations and predictions over the spatio-temporal dimension?

To combine the observations and the MOS corrected predictions, we used their clear-sky indexes. First, the future state of the clear-sky index at a location can be approximated by taking a weighted average of that location's observed clear-sky index and others in proximity. In that case, each weight can be interpreted as the probability that the observed clear-sky index will travel to that location. As we sum the clear-sky indexes according to their probabilities, we are able to approximate the future clear-sky index for the location under study.

Second, the MOS corrected predictions are combined over the spatio-temporal dimension by smoothing them over space and time to extract the trend of the clear-sky index. By doing so, we account for spatial and temporal errors in the MOS corrected predictions.

When we combine these processes in one model, then they can synergize as they cancel out each other's errors. The observations tell us something about the state, while the MOS corrected predictions tell us something about the trend. Therefore, by combining both processes in one model, we can improve the accuracy of our forecasts.

4. Does the performance of spatio-temporal regression for solar power forecasts vary spatially and temporally, and if so, why?

We can evaluate the performance of the models over two dimensions: spatially as per station, and temporally as per the test set and validation set.

The MOS correction increases the accuracy of the numerical irradiance predictions by 8% on the test set and 4% on the validation set – also known as its SS. There does seem to be a relation between each station's SS on the test set and validation set, although it is weak. As the test set spans a period of mostly winter and the validation set one of mostly summer, we find that the MOS correction performs best on the test set. This is due to the correction

of a systematic error that occurs when the predicted clear-sky index is low, which occurs more often in the winter than in the summer.

For the spatio-temporal model with a lead-time of one hour, we find that we can increase on accuracy compared to the numerical irradiance predictions by 30% for the test set and 25% for the validation set. The difference in accuracy between the test and validation set is explained as we use the MOS corrections as an input to the spatio-temporal model.

When we compare the spatio-temporal model's accuracy against the MOS correction, then we find it to be as accurate on the test as on the validation set. In that case, there is zero relation between each station's SS on the two sets. We hypothesize that this could be due to different weather patterns over the winter and summer.

When we assess the spatio-temporal model on a daily scale against the MOS corrected predictions, then we find the spatio-temporal model to perform well on days when the MOS corrected predictions are uncertain about the cloud conditions. Therefore, we find the spatio-temporal model to be accurate in highly dynamic cloud systems.

7.1.2 Main research questions

As the sub-questions are answered, we turn to the main question:

> How can irradiance observations and numerical weather predictions be regressed on the spatio-temporal dimension to create accurate intraday solar power forecasts?

We propose a five-step approach:

1. collect (numerical) weather predictions and irradiance observations over the spatial (about 50 kilometres apart) and temporal (15 minute to hourly) dimensions;
2. remove the sun's pattern from the numerical irradiance predictions and irradiance observations by calculating the clear-sky index;
3. fit the predictions to the observations via MOS;
4. create a spatio-temporal regression model that takes lagged observations and that smoothes MOS corrected predictions; and
5. apply LASSO for predictor selection.

The model that comes out has an SS between 25% and 30% against numerical irradiance predictions for a lead-time of one hour.

The reason for this model to work as well as it does, is that it extracts the information about the state of the atmosphere from lagged observations, which it applies to the trend of the smoothed MOS corrected predictions. By combining these two sources of information, we create accurate intraday solar

power forecasts up to five hours ahead. When we want to go beyond five hours, the MOS corrected predictions prove to be more accurate.

To conclude, our case study of the Netherlands shows that accurate intra-day solar power forecasts in highly dynamic cloud systems can be made by using lagged observations and smoothed MOS corrections in a spatio-temporal context.

7.2 Contribution of research

In Chapter 1, we identify two contributions of our research:

1. We aim to provide a framework for incorporating numerical weather predictions and irradiance observations across the spatio-temporal dimension for accurate solar power forecasts.
2. We aim to provide a case study of the Netherlands to understand how irradiance behaves in a spatio-temporal context.

Here, we delve deeper in what these contributions exactly entail.

7.2.1 Regression framework

First, we developed a framework that uses numerical weather predictions and observations to create accurate intraday solar power forecasts. The framework is unique as it focusses on the synergies between the smoothing of predictions and lagged observations over the spatio-temporal dimension. By doing both in one regression model, it is able to produce better results than doing either apart.

Second, due to the standardization of the predictors (Section 2.3.4) and the application of LASSO (Equation 2.14), the framework is able to provide insights about the irradiance patterns for any geographic location over time. Therefore, we provide an opportunity to map those patterns and to better understand the forces that underlie the predictability of solar power.

Finally, by describing in detail the steps necessary to prepare predictions and observations for such a regression, it becomes easy to scale the framework to other locations besides the Netherlands. Our research provides a simple approach that makes use of data that is often already available. By doing so, we provide a solid base for innovation as described in Section 7.3.

7.2.2 Case study of the Netherlands

First, the case study of the Netherlands provides an insight into how irradiance travels across the Netherlands. We are able to define a pattern, where it travels from the south-west across the rest of the Netherlands.

Second, it provides results in terms of an intraday solar power forecast model's accuracy in the Netherlands with a lead time between one and six hours. From there, it can be used as a benchmark to develop new models that might be more accurate.

Third, we provide an analysis for different cloud conditions, which tells us something about the accuracy of the model compared to the MOS corrected predictions. From there, it becomes easier to understand when our framework can add much value and when other solutions might be more appropriate.

7.3 Future research

When we reflect back on the contribution that we aimed to make with our research, then we have succeeded to develop a framework for incorporating numerical predictions and observations over a spatio-temporal context for intraday solar power forecasts. We have tested this framework within the context of the Netherlands, where it proved to be useful to draw conclusions about the behaviour of irradiance in a spatio-temporal context. Nevertheless, as we aim to improve on what we have done within the boundaries of our research, we provide some topics that could be of interest for future research.

7.3.1 Application of our framework to other contexts

As part of our research, we applied our framework to the Netherlands. The Dutch climate has a highly dynamic cloud system, where NWP from the ECMWF seems to struggle to capture those dynamics in irradiance predictions (Figure 1.3). We solved this problem by designing a compatible framework accordingly.

When applying the framework to a different context, we might think of questions such as:

- Could this framework also be applied to other climates, such as in Africa?
- Would such a framework even be useful in those countries when having the goal to stabilize their energy system?
- What type of weather phenomena are difficult to predict via numerical prediction but could be tackled with statistics?

These are questions that address the scalability of our framework, which are therefore suitable for follow-up research.

7.3.2 Incorporation of other numerical methods

The justification for our research is that NWP from the ECMWF overpredicts medium irradiance conditions. However, we could possibly also tackle this problem by using other numerical weather models, such as the Global Forecast System. These models might synergize with one-another, therefore improving on the results that we already have.

Possible questions that come to mind are:

- How can different numerical weather prediction models be combined to create accurate solar power forecasts?
- Do numerical weather prediction models differ in their predictions, and if so, why?

These questions aim to further enhance our framework within the context of the Netherlands via the acquisition of additional data.

7.3.3 Application of other models besides regression

In our research, we used regression as the means to combine irradiance data. To do so, the data was first relieved from its seasonality, after which a linear model could be applied. However, another approach is to leave the data as is while using non-linear models. We might use tree-based methods, support vector machines, or deep learning models.

Considering that a regression based approach can be considered old-fashioned, we might want to address the following questions:

- What type of non-linear models are able to create accurate solar power forecasts from numerical predictions and observations?
- Do non-linear models improve solar power forecasts compared to linear models, and if so, why?

These questions aim to increase the accuracy of the solar power forecasts by model improvement.

7.3.4 The effect of our framework on energy trading

We have developed accurate solar power forecasts to make predictions about the production of PV assets. As these solar power forecasts are more accurate in terms of error metrics, such as RMSE and MAE, we have no idea if these solar power forecasts also comply with a financial incentive – that is, whether it is possible to generate additional profits when deploying these models on energy markets.

To find out whether our accurate solar power forecasts also increase market revenue from PV assets, we might want to answer the following questions:

- What type of solar power forecast maximizes revenue on energy markets for PV assets?
- How can financial risk on the energy market for PV assets be minimized via solar power forecasts?

These questions are useful to understand if the optimization on accuracy for solar power forecasts is valuable in a market context.

Bibliography

Aguiar, L. M., Pereira, B., Lauret, P., Díaz, F., & David, M. (2016). Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. *Renewable Energy*, *97*, 599–610. https://doi.org/10.1016/j.renene.2016.06.018

Bacher, P., Madsen, H., & Nielsen, H. A. (2009). Online short-term solar power forecasting. *Solar Energy*, *83*(10), 1772–1783. https://doi.org/10.1016/j.solener.2009.05.016

BNEF. (2020). *New Energy Outlook*.

BP. (2020). *Statistical Review of World Energy*.

Brockwell, P. J., & Davis, R. A. (2009). *Time series: Theory and methods*. Springer science & business media.

Chatfield, C. (2003). *The analysis of time series: An introduction*. Chapman and hall/CRC.

Dambreville, R., Blanc, P., Chanussot, J., & Boldo, D. (2014). Very short term forecasting of the Global Horizontal Irradiance using a spatio-temporal autoregressive model. *Renewable Energy*, *72*, 291–300. https://doi.org/10.1016/j.renene.2014.07.012

David, M., Luis, M. A., & Lauret, P. (2018). Comparison of intraday probabilistic forecasting of solar irradiance using only endogenous data. *International Journal of Forecasting*, *34*(3), 529–547. https://doi.org/10.1016/j.ijforecast.2018.02.003

EIA. (2020). *Annual Energy Outlook 2020*.

Elsinga, B. (2017). *Chasing the Clouds: Irradiance Variability and Forecasting for Photovoltaics*. Utrecht University
OCLC: 8086870273.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1–22. Retrieved August 7, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

Glahn, H. R., & Lowry, D. A. (1972). The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology and Climatology*, *11*(8), 1203–1211. https://doi.org/10.1175/1520-0450(1972)011⟨1203:TUOMOS⟩2.0.CO;2

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2016, January 1). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd edition). Springer.

Hogan, R. J., & Bozzo, A. (2018). A Flexible and Efficient Radiation Scheme for the ECMWF Model. *Journal of Advances in Modeling Earth Systems*, *10*(8), 1990–2008. https://doi.org/10.1029/2018MS001364 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001364

Holmgren, W. F., Hansen, C. W., & Mikofski, M. A. (2018). Pvlib python: A python package for modeling solar energy systems. *Journal of Open Source Software*, *3*(29), 884.

IEA. (2019). *World Energy Outlook 2019*.

IPCC. (2012, July 1). *Renewable energy sources and climate change mitigation*. Intergovernmental Panel on Climate Change.

IPCC. (2015). *Climate change 2014: Synthesis report*. Intergovernmental Panel on Climate Change. Geneva, Switzerland.

Kong, Q., Siauw, T., & Bayen, A. (2020, December 16). *Python Programming and Numerical Methods: A Guide for Engineers and Scientists* (1st edition). Academic Press.

Liu, G., Qin, H., Shen, Q., Lyv, H., Qu, Y., Fu, J., Liu, Y., & Zhou, J. (2021). Probabilistic spatiotemporal solar irradiation forecasting using deep ensembles convolutional shared weight long short-term memory network. *Applied Energy*, *300*, 117379. https://doi.org/10.1016/j.apenergy.2021.117379

López Lorente, J., Liu, X., & Morrow, D. J. (2020). Worldwide evaluation and correction of irradiance measurements from personal weather stations under all-sky conditions. *Solar Energy*, *207*, 925–936. https://doi.org/10.1016/j.solener.2020.06.073

Masson-Delmotte, V., Pörtner, H.-O., Skea, J., Zhai, P., Roberts, D., Shukla, P. R., Pirani, A., Pidcock, R., Chen, Y., Lonnoy, E., Moufouma-Okia, W., Péan, C., Connors, S., Matthews, J. B. R., Zhou, X., Gomis, M. I., Maycock, T., Tignor, M., & Waterfield, T. (2019). *Global warming of 1.5°C*. Intergovernmental Panel on Climate Change.

Notton, G., & Voyant, C. (2018). Forecasting of Intermittent Solar Energy Resource. In *Advances in Renewable Energies and Power Technologies* (pp. 77–114). Elsevier. https://doi.org/10.1016/B978-0-12-812959-3.00003-4

Sengupta, M., Habte, A., Wilbert, S., Gueymard, C., & Remund, J. (2021, April 14). *Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition* (NREL/TP-5D00-77635). National Renewable Energy Lab. (NREL), Golden, CO (United States). https://doi.org/10.2172/1778700

Smil, V. (2016, December 5). *Energy Transitions: Global and National Perspectives, 2nd Edition* (2nd edition). Praeger.

Sperati, S., Alessandrini, S., & Delle Monache, L. (2016). An application of the ECMWF Ensemble Prediction System for short-term solar power

forecasting. *Solar Energy*, *133*, 437–450. https://doi.org/10.1016/j.solener.2016.04.016

Statsmodels. (2019). *Statsmodels.regression.linear_model.OLS.fit_regularized*. Retrieved July 13, 2022, from https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.fit_regularized.html

van Heerwaarden, C. C., Mol, W. B., Veerman, M. A., Benedict, I., Heusinkveld, B. G., Knap, W. H., Kazadzis, S., Kouremeti, N., & Fiedler, S. (2021). Record high solar irradiance in Western Europe during first COVID-19 lockdown largely due to unusual weather. *Communications Earth & Environment*, *2*(1), 1–7. https://doi.org/10.1038/s43247-021-00110-0

Voyant, C., Notton, G., Duchaud, J.-L., Almorox, J., & Yaseen, Z. M. (2020). Solar irradiation prediction intervals based on Box–Cox transformation and univariate representation of periodic autoregressive model. *Renewable Energy Focus*, *33*, 43–53. https://doi.org/10.1016/j.ref.2020.04.001

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

Witte, R. S., & Witte, J. S. (2017). *Statistics*. John Wiley & Sons.

Yang, D., & van der Meer, D. (2021). Post-processing in solar forecasting: Ten overarching thinking tools. *Renewable and Sustainable Energy Reviews*, *140*, 110735. https://doi.org/10.1016/j.rser.2021.110735

Yang, D., Wang, W., Bright, J. M., Voyant, C., Notton, G., Zhang, G., & Lyu, C. (2022). Verifying operational intra-day solar forecasts from ECMWF and NOAA. *Solar Energy*, *236*, 743–755. https://doi.org/10.1016/j.solener.2022.03.004

Zhang, G., Yang, D., Galanis, G., & Androulakis, E. (2022). Solar forecasting with hourly updated numerical weather prediction. *Renewable and Sustainable Energy Reviews*, *154*, 111768. https://doi.org/10.1016/j.rser.2021.111768