UNIL | Université de Lausanne
Faculté des géosciences
et de l'environnement

Bachelor Thesis

Statistical Downscaling of Two-Meters Temperatures in Casablanca, Morocco

Meryam Cherqaoui Fassi

Under the supervision of Prof. Tom Beucler
Under the co-supervision of Graduate Assistant Milton Gomez Delgadillo

Source : www.moroccopedia.com/casablanca/

Contents

List of Figures

List of Tables

Abstract

Statistical downscaling is a powerful tool for numerical weather prediction. Many studies have been conducted to obtain high-resolution local weather from NWP models, especially in Asia and South-America. But the lack of pre-processed data sets in other parts of the world, particularly in Africa, does not provide a solid basis for testing different downscaling methods, and needs to be addressed. In this study we tackle the lack of basis for the city of Casablanca, in Morocco, by providing a clean benchmark to compare different statistical downscaling methods for two-meters temperature measurements from 1979 to 2018. We test two methods of statistical downscaling (interpolation and linear regression) using ERA5 meteorological reanalysis, which are used to question if more complex methods lead to a better predictability. The results show that the interpolation is more accurate, but further methods need to be compared in order to conclude on the question. The scripts used for this study as well as the data set are publicly available at: https://github.com/meryamcherqaouifassi/2021_Bachelor_Thesis

1 Introduction

In the current context of climate change, it is of the upmost importance to understand what shapes the interactions between the atmosphere, the ocean and the land-surface. GCMs, or Global Climate Models, enable to solve physical equations describing theses interactions, and to have a reliable discerning of the climate at a global scale. NWP models, or Numerical Weather Prediction models, use current weather observations to deduce future weather using data assimilation, which is a branch of applied mathematics that combines observed data and a numerical model to optimize the information of both. A data assimilation algorithm "makes a forecast using a model then uses de the data to make a correction of the model" (Burov, 2019). But the solutions are discretized using grid boxes, and so the resolution is coarse, typically unable to predict what can occur at a city level. Yet, the increase in frequency and intensity of climate events, such as heatwaves or heavy rainfalls, requires a good understanding of that level, to be able to predict local impacts and take action in time. Methods have been developed in order to infer the weather from global models, among which dynamical and statistical downscaling.

Dynamical downscaling translates a global model's output to smaller areas, using a high resolution dynamical model called Regional Climate Model, or RCM (Furtado, 2017). This method involves taking the information from a NWP model to feed it into a RCM, that becomes driven by the NWP model boundary conditions, enabling to describe future weather at high resolution and over limited areas of the globe (Furtado, 2017). Although dynamical downscaling has several benefits, including physical consistency, it also has disadvantages. For instance, the method is computationally expensive, meaning that it requires a large amount of computing power and resources (Hong Kanamitsu, 2014). Moreover, if the models used have inherent errors or biases, whether coming from the global or the regional model, those errors can creep into the analysis and make it either false or complicated to correct (Furtado, 2017).

Most statistical downscaling approaches begin by comparing the NWP model output for a particular time period in the past with observations during the same period. By comparing the model with the data, a relationship linking the weather patterns from both can be found, and described using statistics. The statistical relationship found can then be applied to future weather projections to obtain future local data. The main advantage, in comparison to dynamical downscaling, is the little computer power needed for this method (Mullens, 2018). Moreover, a statistical downscaling approach can be crafted for a specific purpose. There are several ways to build a statistical downscaling relationship, and 3 main categories of methods: transfer functions, weather typing, and weather generators (Mullens, 2018).

Transfer function are the simplest methods, and include, among others, linear regression, delta method and bi-linear interpolation. The common aspect is that each of these methods builds a direct relationship between the observations and the NWP model for the same region.

Weather typing methods, also called analogs methods, build a relationship between the model and the observations in a non-direct way: it links "the occurrence of particular weather types to local climate" (Duan, 2013).

Weather generators are different from all the others. They are stochastic models of meteorological variables, often used to downscale climate for impact assessment purposes (Maraun &

Widmann, 2017).
Studies have been led to compare the efficiency of the different methods, and have highlighted the importance of choosing the right method for the right geographic and climatic characteristics. They have also shown that the field of efficiency being different from one method to another, a combination of results from diverse methods is the best way to achieve a goal. Indeed, some methods have strengths where others lack, and vice versa (Liu et al., 2015).
Only a few studies have focused on North Africa. Yet, the growing drought provoked by the increase of temperatures and decrease of precipitations may be a major future challenge. The lack of studies involving statistical downscaling of climate variables in this part of the world makes the data availability and processing challenging, and highlights the need for more research in this field. In this study we address this obstacle by developing and providing the data set to apply statistical downscaling for observed two meters temperatures between 1979 and 2018 in Casablanca. Moreover, we explore some of the transfer functions methods, in the attempt to investigate the following question : how complex does a statistical downscaling method need to be for accurately predicting land surface temperatures in Casablanca, Morocco?
We developed two methods : interpolation and multiple linear regression. We used several interpolation approaches, among which the linear, nearest-neighbor, quadratic and cubic interpolations. We calculated the climatological and persistence baselines to compare their performance with that of the models.
After a description of the study area and the data series (section 2), we detail the methodology used to process the data (sections 3.1) and to apply interpolation (section 3.2) and linear regression methods (section 3.3), as well as the baselines (section 3.4). We then present the results (section 4) and discuss them (section 5).

2 Study Area and Data Series

2.1 Study Area

Casablanca is located on the Atlantic coast of Morocco, in North Africa (Figure 1), 27 meters above sea level. The city has a Mediterranean climate with a strong oceanic influence, that confers mild temperatures and relatively wet winters, as well as moderately hot summers. The average annual temperature is 18.88°C and the year-to-date precipitation is 426.1 mm (Climate Weather Averages in Casablanca, 2022).
In summer, temperatures are generally around 23°C, and hot days are relatively rare. But during periods of wind, caused by the Sirocco air coming from Sahara, the city can record particularly hot temperatures for a few days, up to 40.5°C, especially in august, which is the hottest month (Casablanca Climate, 2022). During that period, patches of morning mist and night dew are quite common.
The area is subject to oceanic disturbances from the Atlantic during the rainy season that begins in October and may continue until May.
In winter, the average temperature is 13,7°C and solid precipitations are almost non-existent (Climate Weather Averages in Casablanca, 2022).

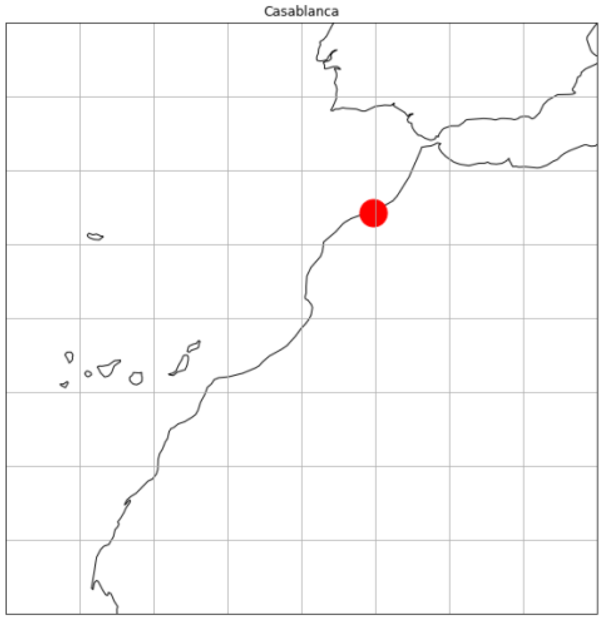


Figure 1: Map of Morocco, dot on Casablanca

2.2 Expected Changes in Climate

According to the fifth assessment report of the Intergovernmental Panel on Climate Change (IPCC, 2014), between 1901 and 2012, the Moroccan Atlantic coast has warmed by 1 to 1.25 °C. As for rainfall, it has decreased by 10 to 15 %. The latest report confirms the increase of the temperatures, or even revises it upward for some areas of the globe (IPCC, 2022). In North Africa, the temperatures are expected to rise up to 5°C in the least optimistic scenario. The consequences could affect both ecosystems and populations in an unprecedented way. Drought may increase the intensity and frequency of fires, and impact human health, natural habitats structure, increase food scarcity and affect the livelihoods of communities that depend on the exploitation of resources.

In order to make decisions aiming at reducing the impacts of global warming across the country, and to preserve the local population, it is important to be able to predict future climatic events in the area.

2.3 Meteorogical Reanalysis

The meteorogical reanalysis we used in this study is ERA5. It is built by assimilating observations into the global model IFS, which is a numerical weather prediction model (NWP model). It is the fifth generation of meteorological reanalysis produced, and it distinguishes itself from the other generations by an increase in vertical and horizontal resolution, and by a large set of

meteorological parameters (Hersbach et al., 2020). ERA5 provides hourly estimates of atmospheric, land and oceanic climate variables. For this study, the selected data are two meters temperatures, over a period of time ranging from 1979 to 2018. Two meters temperature is often used because of its relevance for human comfort and activities.
The statistical downscaling methods we developed use data produced by ERA5 as predictor. We consider that the data from ERA5 are without flaws; we decided to overlook potential bias from the model. The first reason is that ERA5 being the latest generation of reanalysis, it benefits from 10 years of experience in model development and data assimilation, which makes it particularly efficient and a good representation of the IFS model outputs (Hersbach et al., 2020). In addition, we decided to make a "perfect model" approximation to focus on data processing and scaling model development. However, if the biases related to ERA5 were to be addressed, we would use the data related to these uncertainties provided by the model, which we would implement in our calculations.

2.4 Station-Observed Data

We collected data from Casablanca's weather station from 1973 to 2021. They are provided by the Department of Atmospheric Science of the University of Wyoming, and include atmospheric pressure (hPa), geopotential height (m), temperature (°C), dew point temperature (°C), frost point temperature (°C), relative humidity (%), relative humidity with respect to ice (%), mixing ratio (g/kg), wind direction (degrees), wind speed (knot), potential temperature (kelvin), equivalent potential temperature (kelvin), and virtual potential temperature (kelvin). Measures are provided daily at 12Z or 00Z. The geopotential is usually denoted as Φ, with units of m^2, and defined as $\Phi = \int_0^z g dz'$, where $g = 9.81 m.s^{-2}$ is the gravitational acceleration and z is the height in meters. It is used to determine the lowest level of measurements. We arbitrarily choose here to focus exclusively on temperatures, the other variables are therefore not used.

3 Methodology

3.1 Data Pre-processing

For each method, we have pre-processed the ERA5 data and observed data from Casablanca's weather station to fit specific needs, in a way described in the following sections. Yet, some parts of the processing are the same for each method, and are described in this section.

3.1.1 Casablanca Weather Station

Casablanca's data have been downloaded from the University of Wyoming website, as text files, using the *download_script.sh* script in the *utils* folder of the github repository. The original format was a succession of text data frames separated by station information and sounding indices and grouped by months. We have processed them using the *convert.py* script, in order to select the needed variables into one single data frame. We have replaced the blank spaces by missing values, and converted the whole data frame into a .csv file.

We extracted the dates of the measurements, as well as the hours, from the station information, and converted them into datetime64 format, and added them as a feature in the data frame. In order to fit ERA5's temporal range, we have reduced the data to cover the period from 1979 to 2018. We selected two variables: the temperature and the geopotential height. We use the geopotential height as a selection feature; it enables to extract near surface temperatures. The lowest geopotential height is 58 meters, which is more than the height of ERA5 data, and could lead to inaccuracies in model construction.
We have considered extreme temperatures, either abnormally cold or hot, outliers and removed them from the data frame.

3.1.2 ERA5

We have imported ERA5 data as a data set object. First, we converted the temperatures from Kelvin to Celsius degrees, both to fit Casablanca's weather station temperature units, and to make the interpretation of the plots easier. Moreover, we reduced the data to cover the North-Eastern part of Africa, in order to increase the speed of calculations. Indeed, we considered irrelevant to use the whole meteorological reanalysis data for the studied area. As a result, the remaining area covers a longitude ranging from 171 to 173, and a latitude ranging from 33 to 35, which corresponds to 4 points.
Since the weather station provides one measure per day either at 12:00 or 00:00, we reduced the hourly data from ERA5 to match the daily temporal resolution. We paired each ERA5 time with the equivalent in the observed data, to make sure that the dimensions were the same.

3.2 Interpolation

3.2.1 Principle

Interpolation consists in estimating the value of unknown data points based on a known dataset. More specifically, it is the mathematical act of approximating the value in between two or more known values. For statistical downscaling, interpolation uses the measured distances of the four points surrounding the location of study to predict a value for that location.

3.2.2 Interpolation Methods

Multiple methods can be used to interpolate variables. The linear interpolation is the simplest method, and enables to draw a straight line, representing a direct relationship between the know values. To do so, the data curve is fitted using linear polynomials. For a given x between x_0 and x_1, y can be calculated using the equation (1) , which can be rewritten as equation (2).

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{1}$$

$$y = \frac{y_0 \times (x_1 - x) + y_1 \times (x - x_0)}{x_1 - x_0} \tag{2}$$

This formula can be interpreted as a weighted average, the weights being :

$$\frac{x_1 - x}{x_1 - x_0} \qquad \frac{x - x_0}{x_1 - x_0} \tag{3}$$

This method is often used for data forecasting and prediction.
In contrast, the nearest-neighbor method assigns the known value of the closest point to the unknown value. It does not take into account any of the other points.
Quadratic interpolation is the method for which "the function's critical value is bracketed, and a quadratic interpolant is fitted to the arc contained in the interval" (Vandebogert, 2017). In other words, the fitting curve is a second-order polynomial. Cubic interpolation enables to fit the curve using a third-order polynomial. We implemented linear, nearest-neighbor, quadratic and cubic methods using xarray's interp() function.

3.3 Linear Regression

3.3.1 Principle

Linear regression is a mathematical model used to quantify the relationship between two variables. In the general case, making a regression implies having a dependent variable, corresponding to the predicted one, and one or multiple independent variables, that are used to predict. For statistical downscaling, the regression is a multi-linear regression, meaning that we use multiple independent variables to predict. The resulting line represents the best fit, such as the distances between observed values and predicted values, also called residues, are minimized. Unlike the linear interpolation, which requires to pass through all points, linear regression only requires to be as close to all points as possible (Sathwick, 2020). In order to do so, we used the method of least squares, which minimizes the total squared quantity of all the distances to the regression line.
While a correlation analysis only gives the direction and strength of the relationship between the two variables, a linear regression describes it numerically, and thus is more interesting in terms of predictability (Sathwick, 2020). The simple regression model derives from the slope-intercept equation (4), where y is the dependent variable, a the slope, x the independent variable and b

the intercept.

$$y = ax + b \tag{4}$$

For a statistical downscaling method, the equation involves multiple points and corresponds to equation (5), where Y regroups all the observed values of the dependent variable, β_0 is the y-intercept, or bias, β_1 is the slope, or coefficient, X regroups all the observed values of the independent variable and ϵ is the error, or residual term, and describes how far the predicted value is from the observed one.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{5}$$

The model uses the studied variable of n points surrounding the area in order to predict the local variation. For this study, to predict future local temperatures, the number of points we used is n = 4.
The libraries we used for this section give access to several variables that allow to judge the quality of the model. The R^2 score, or coefficient of determination, is a number that quantifies how close the scatter plots points are to the regression line. The closer the coefficient is to 1, the better is the ability of the model to predict.

3.3.2 Process

In order to run the linear regression, we have flattened the two dimension coordinates of the ERA5 model into one dimension points. We use the Scikit-learn library, with the LinearRegression and mean_squared_error functions to solve the model.

3.3.3 Evaluation

One of the biggest challenges when building a statistical method is to make sure the generalization is correct. Indeed, it is important that the model distinguishes random features in the selection of the data set from regularities that actually exist in the population. To do so, the data are usually split into three parts: the training, validation and test sets. The goal of this separation is to optimize the performance of the model across a wide range of conditions, including conditions it has never been trained on. First, the data set is divided into training and test sets. The model is trained using the training set. The test set is meant to be kept apart from the model training, and is used once it is calculated to verify its quality. One way to make the estimation of the generalization error more robust is to use a cross-validation method. A k-fold cross-validation is a method that produces k values of the generalization error estimate, by separating the data into k equal random parts, setting aside the first fold for testing, and train the method on the remaining k - 1 folds (we obtain an estimate of the generalization error), and repeating by setting aside one by one each fold. We obtain the generalization error by computing the average of the errors over all the folds.
We have formed the split and built the cross-validation using sklearn's train_test_split and

cross_validate functions. The training set represents 80% of the data, and the test set the 20% left. The cross-validation is made over 5 groups, randomly split.

3.4 Baselines

3.4.1 Persistence Baseline

A baseline is used to evaluate the skill of a forecasting model. The persistence baseline considers that tomorrow's weather is today's weather. It has been calculated by creating a lag feature that splits the measures in two, the even indexes being concatenated in the first column of a new data frame, and the uneven ones in the second column. After defining a training and a test set, the predictions were calculated. The residuals, mean squared error, mean absolute error, and coefficient of determination are calculated and compared to the linear regression model's, to evaluate its performance.

3.4.2 Climatological Baseline

A climatological baseline is calculated by averaging the observed climate variables over a time period, whether over all times in the training data set or a mean computed for a particular time period, for example weekly or monthly. In the same way as the persistence baselines, the mean squared error, residuals, mean absolute errors and coefficient of determination are used to evaluate the performance of the model. In this study, two different climatological baselines are used, and presented in ascending order of precision:

We calculated the monthly climatological baseline by averaging all the x months temperatures over the studied time period. This means that the input is the entire data set, and the resulting output of the calculation is an array of 12 values, each one corresponding to the averaged temperatures of each of the 12 months. To do so, we converted each datetime64 date, that is in year-month-day hour:minute:second format, into strings and separated them in year, month and day and then grouped them by months. We then averaged the temperatures using the train data set.

The daily climatological baseline averages the temperatures for each of the 365 days of the year. It enables to have an idea of both diurnal and seasonal variations over the studied time period. As for the monthly climatological baseline, we selected the days by splitting the date and we calculated the average by iterating over all the years of the train set.

4 Results

4.1 Visualization of the Data

The following plots show the distribution of Casablanca's station measures (Figure 2) and ERA5 temperatures for the same area (Figure 3). They have a similar shape, with a first peak around 15°C and a second one around 25°C. The density is slightly different and may be due to the number of samples, which is more important for ERA5 temperatures.

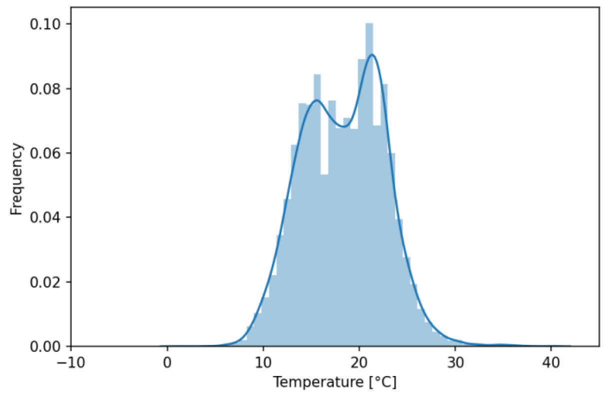


Figure 2: Distribution of Casablanca's measured temperatures

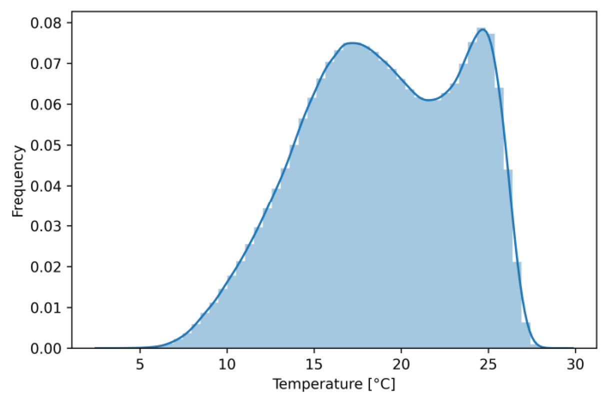


Figure 3: Distribution of Era5's temperatures around Casablanca

The variation of temperatures in time (Figure 4) show a slight cooling between 1980 and 2020, but in reality the studied time period is too small to assert any definitive conclusion.

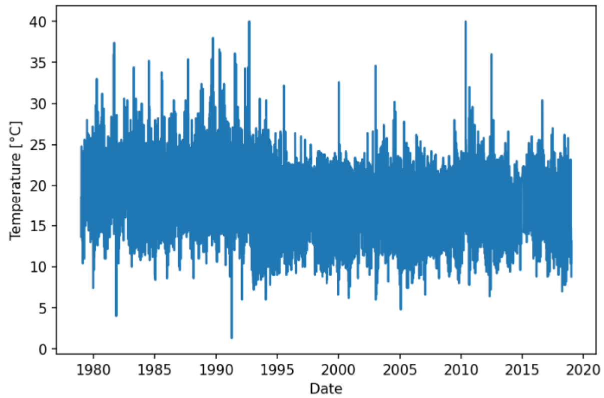
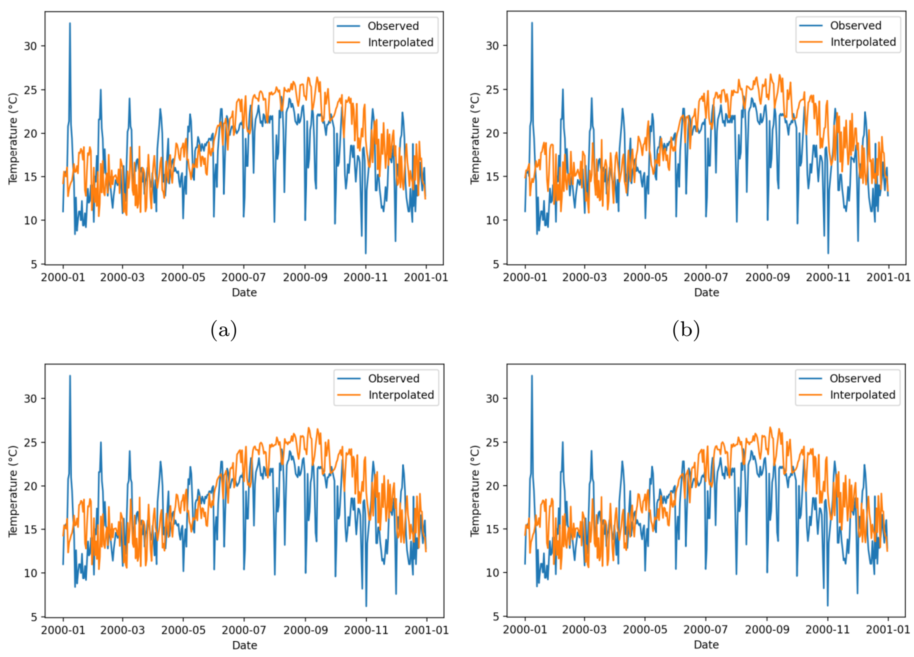


Figure 4: Evolution of Casablanca's temperatures between 1979 and 2021

4.2 Interpolation

4.2.1 Results

The linear, nearest-neighbor, quadratic and cubic interpolations results are depicted in Figures 5, 6 and 8.



(a) (b) (c) (d)

Figure 5: Observed and interpolated temperatures for the year 2000 using (a) linear interpolation (b) nearest-neighbor interpolation (c) quadratic interpolation (d) cubic interpolation

The year 2000 was arbitrarily chosen to illustrate the fitting of the interpolation to the observed temperatures.

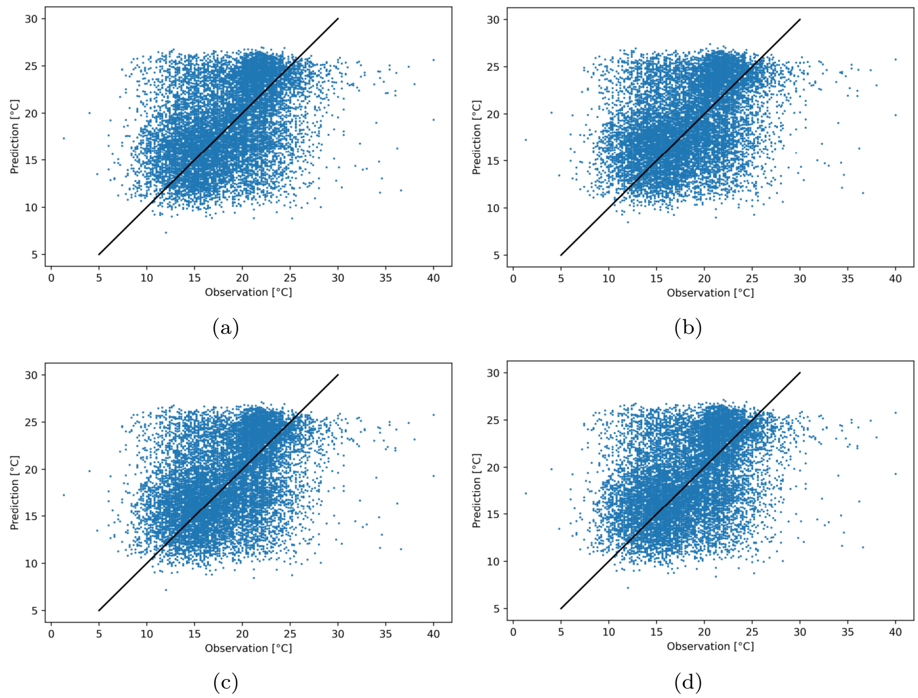


Figure 6: Predicted temperatures as a function of observed temperatures for (a) linear interpolation (b) nearest-neighbor interpolation (c) quadratic interpolation (d) cubic interpolation

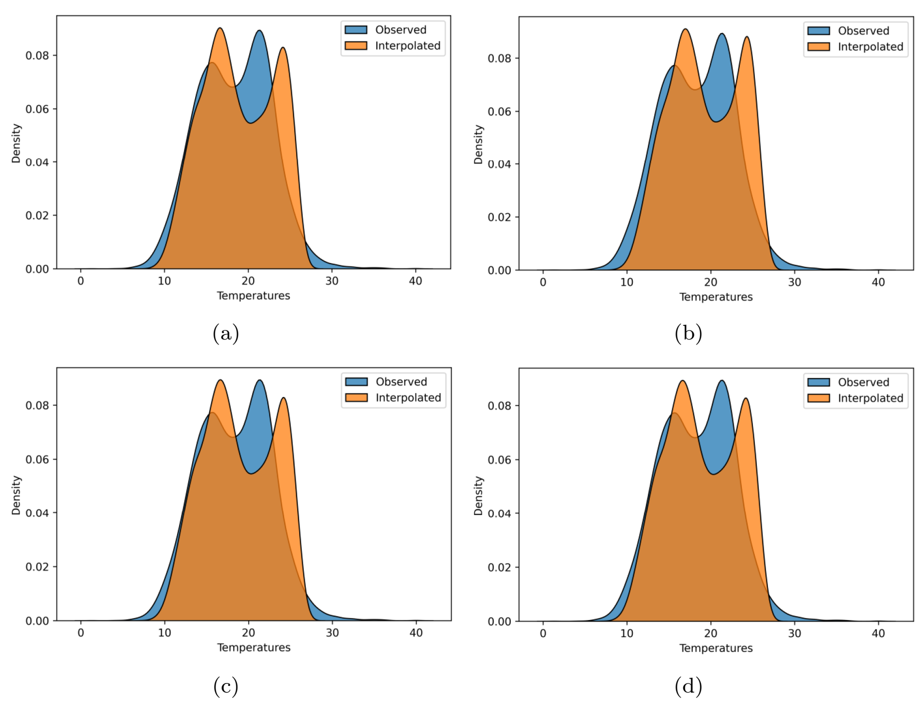


Figure 7: Density plots using (a) linear interpolation (b) nearest-neighbor interpolation (c) quadratic interpolation (d) cubic interpolation

The plots of Figure 7 are two-dimensional density plots, that give a better sense of the model's performance than the scatter plots of Figure 6.

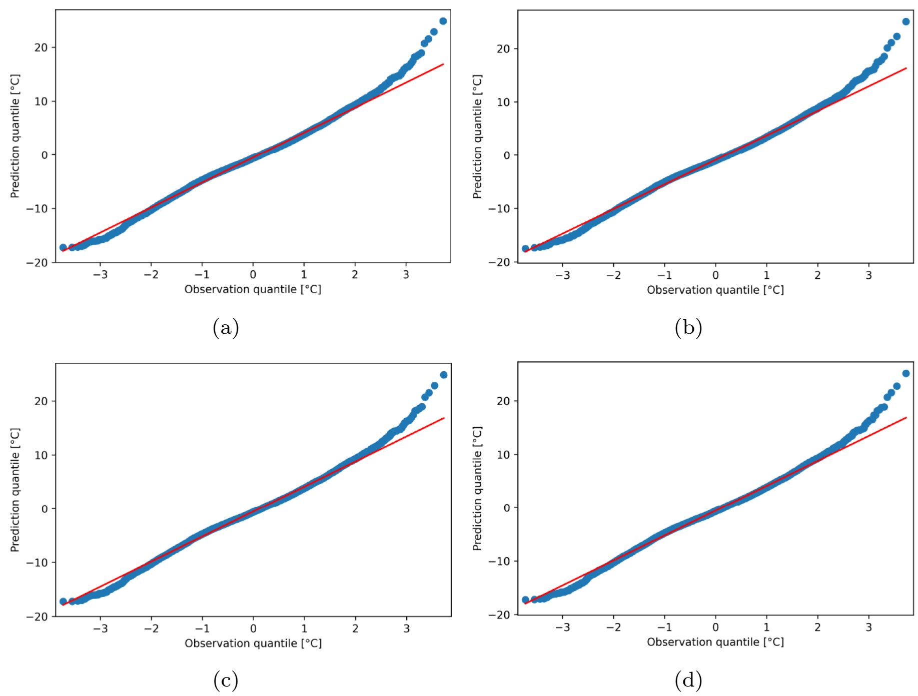


Figure 8: Quantile-quantile plot of predicted and observed temperatures for (a) linear interpolation (b) nearest-neighbor interpolation (c) quadratic interpolation (d) cubic interpolation

Performance metrics were calculated and are grouped in the following table.

	R2 (train/test)	RMSE (train/test)	MAE (train/test)
Linear	-0.183/-0.165	4.696/4.648	3.621/3.605
Nearest-neighbor	-0.188/-0.173	4.706/4.665	3.640/3.637
Quadratic	-0.191/-0.172	4.711/4.663	3.634/3.622
Cubic	-0.191/-0.173	4.712/4.664	3.635/3.623

Table 1: Performance metrics of the interpolations

4.2.2 Interpretation

Figure 5 shows that the differences in the fitting between the four interpolation methods are minimal. Table 1 confirms it: the coefficients of determination are only slightly different, ranging from -0.183 to -0.191 for the training set, and from -0.165 to -0.173 for the test set. The fact that these coefficients are negative shows the very poor predictive ability of the interpolations: it means that the interpolations fit worse than a horizontal line. The plots in Figure 6 show the cloud of measurements, and enable to visualize how well the predicted and observed values relate to each other. In this case, the round shape of the clouds is an indicator of a poor predictive ability for all four methods.

The density plots in Figure 7 display the distribution of both observed and interpolated temperatures. The interpolations show two peaks, similarly to the observed values, but these peaks are shifted to the right, and do not have the same amplitude. It is representative of the poor performance of the model: the points that are poorly interpolated are not outliers, but rather correspond to the average temperatures in Casablanca. The quantile-quantile plots shown in Figure 8 are used to determine the location of the distribution errors. The errors seems equally distributed, and the predictions don't seem particularly over or under the true values. For all four methods, the plot diverges at the end of the line, which means that the interpolations are unable to capture the behavior at the tail ends of the distribution.
The three previous figures show the poor fitting of the four interpolations to the observed values, and confirm what the performance metrics reveal. However, it appears that among the four interpolation methods, the linear method shows better results: the coefficient of determination, mean squared error and mean absolute error are lower. Overall, the test set performs better than the training set. The best performances are, in order, for the linear method, then nearest-neighbor, then quadratic and finally cubic. However it is really a question of determining the least worst method rather than the best.

4.3 Linear Regression

4.3.1 Results

The resulting coefficient, or slope, is an array of 4 values (for the 4 points used for the regression) (Table 2).

0.879	0.454
0.115	0.303

Table 2: Slopes of the multiple linear regression

Performance metrics were calculated and are grouped in the following table.

	R2	RMSE	MAE
Train set	0.163	3.949	3.098
Test set	0.167	3.931	3.089

Table 3: Performance metrics of the linear regression

The regression enables to plot the following graphs :

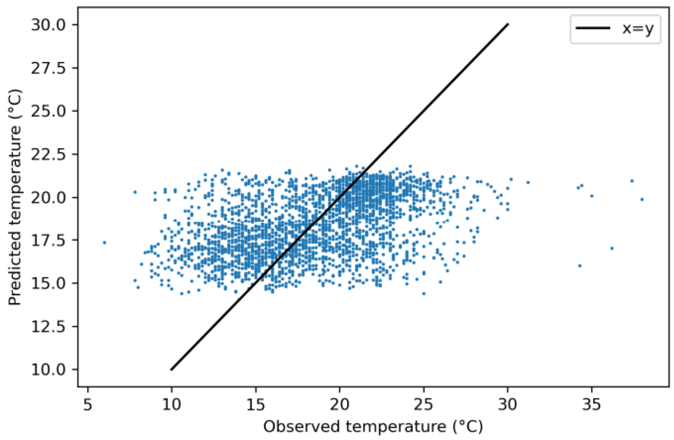


Figure 9: Predicted temperatures as a function of observed temperature

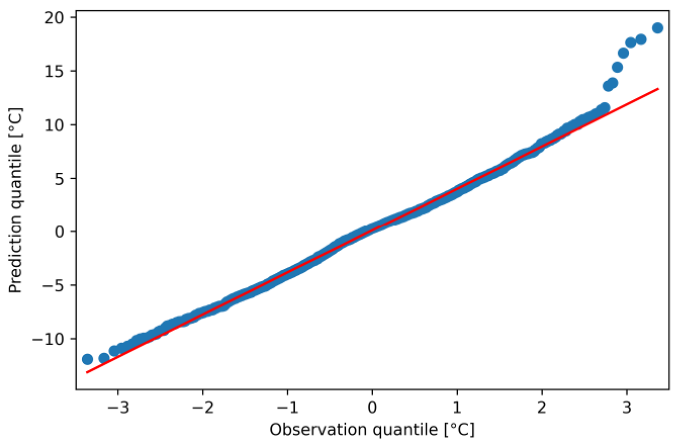


Figure 10: QQ-plot of predicted and observed temperatures

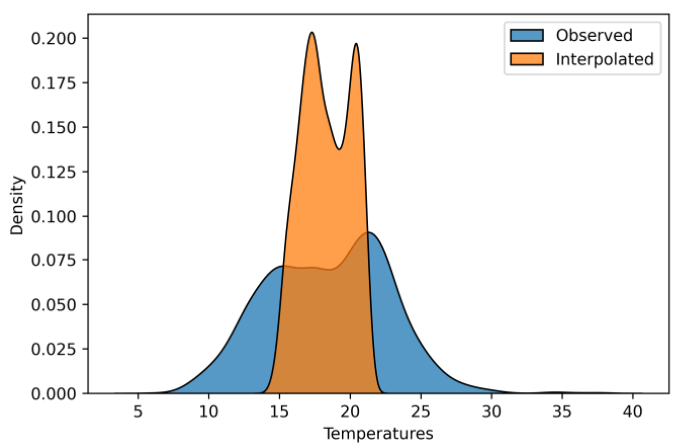


Figure 11: Density plot of predicted and observed temperatures

4.3.2 Interpretation

The value of the R^2 in Table 3 is significantly far from 1 and demonstrates the poor accuracy of the model. The large mean squared error confirms this hypothesis. We can see that the training set performs better than the test set.
Figure 9 is a (y_observed, y_predicted) graph with a x=y line. Both lines can be compared to infer how related the observed and predicted temperatures are. Here we can see that the spreading of the cloud does not follow the x=y axis, which shows once more the poor predictability of the regression.
Figure 10 is a quantile-quantile plot of the observed and predicted temperatures. The plot diverges at the end of the line, which means that the linear regression is unable to capture the behavior at the tail ends of the distribution. For the rest of the plot, the distribution seems to follow the red line. However, Figure 11 shows that the distribution of predicted temperatures is particularly different from that of the observed values, if not radically different. While the observed values are more spread out, ranging from about 5 to 35 degrees, the predicted values seem to be only between 14 and 22 degrees. This figure shows the poor quality of the model.

4.4 Persistence Baseline

4.4.1 Results

Once the persistence baseline computed from the training set, we calculated the residuals and draw the following graphs.

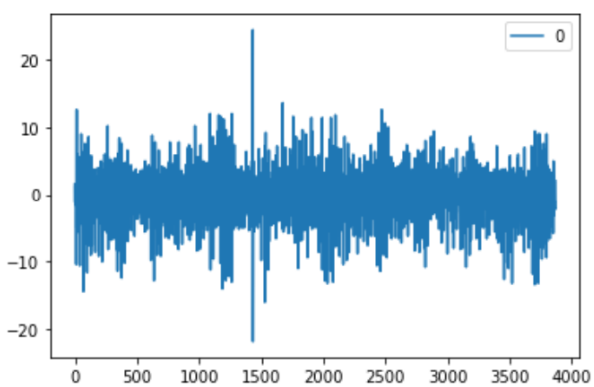


Figure 12: Residuals of the persistence baseline

4.4.2 Interpretation

Figure 12 shows the residuals over time as a line plot. The randomness of the plot around 0 and the lack of cyclic behavior indicate that the model is well fitted. The histogram and density plot in figure 13 shows where the errors are mostly distributed. We can observe that the errors are normally distributed around 0, which confirms that the model is fitted. We can see in the residual quantile plot in Figure 14 that the errors follow mostly the red line. Some outliers are

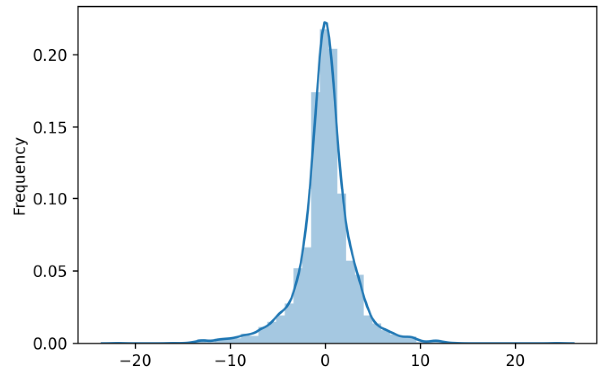


Figure 13: Distribution of the persistence baseline residuals

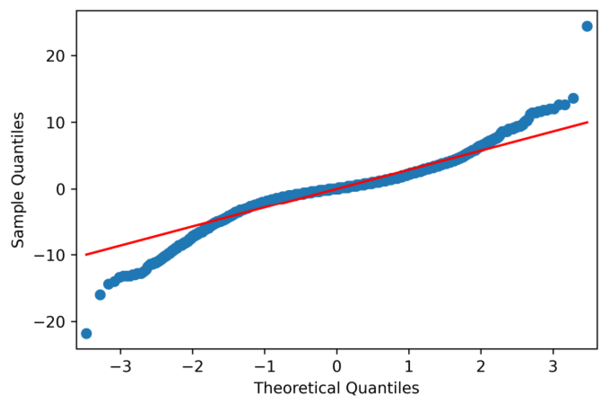


Figure 14: Quantile-quantile plot of the persistence baseline residuals

visible, and would need to be addressed to obtain a better model. But overall, the persistence baseline shows more than sufficient results.

4.5 Monthly Climatological Baseline

4.5.1 Results

The averaged temperatures for each month are grouped in the following table :

	01	02	03	04	05	06	07	08	09	10	11	12
Temp	15.0	16.1	16.7	17.6	19.0	20.2	21.4	21.3	20.6	19.2	16.9	15.7

Table 4: Average temperatures (°C) for the 12 months of the year

The resulting residuals enable to draw Figures 15, 16 and 17.

4.5.2 Interpretation

In the same way as for the persistence baseline, the plot of the monthly climatological baseline residuals (Figure 15) is random around 0. But a cyclic behavior is visible, and might show the

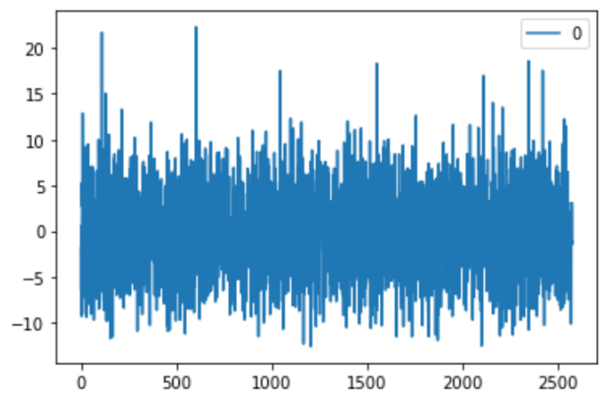


Figure 15: Residuals of the monthly climatological baseline

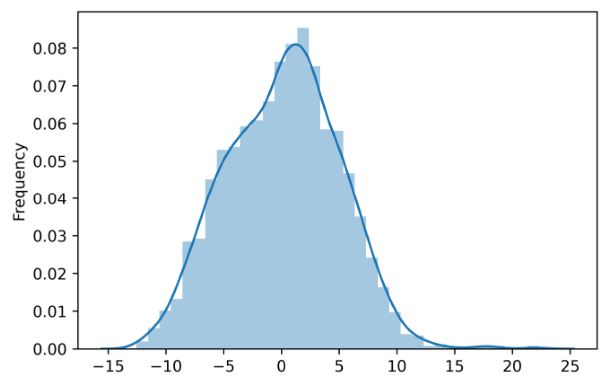


Figure 16: Distribution of the monthly climatological baseline residuals

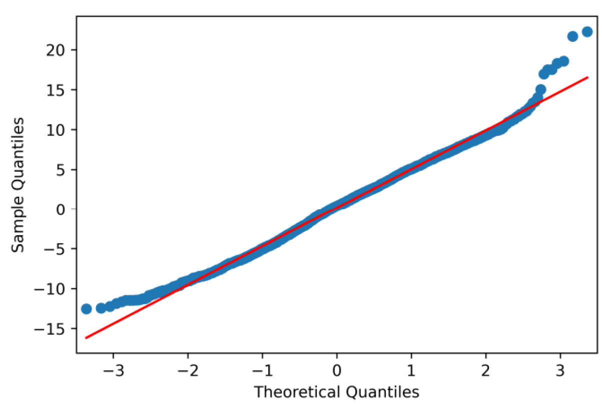


Figure 17: Quantile-quantile plot of the monthly climatological baseline residuals

poor fitting of the model. We can see in Figure 16 that the residuals are not only not normally distributed, but are negatively skewed. This shows once more that the model can largely be improved. But the quantile plot in Figure 17 shows that the relationship between theoretical and sample quantiles follows a straight line, except for a few outliers, which means that the distribution of the quantiles is only slightly different from the idealized Gaussian distribution.

4.6 Daily Climatological Baseline

4.6.1 Results

The calculation and analysis of the daily climatological baseline provide the following plots:

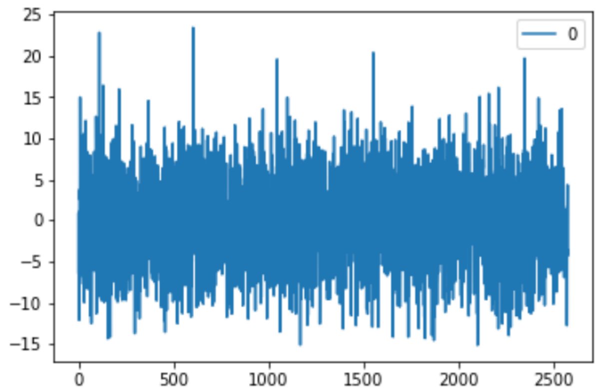


Figure 18: Residuals of the daily climatological baseline

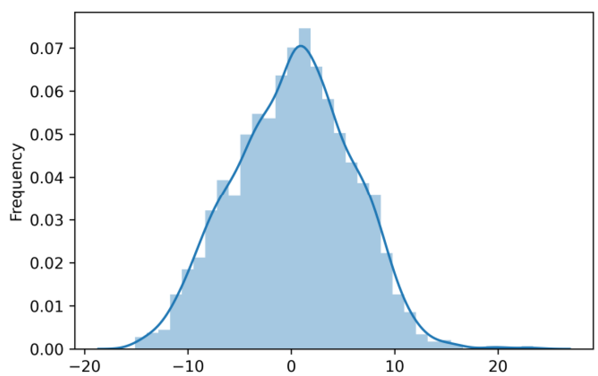


Figure 19: Distribution of the daily climatological baseline residuals

4.6.2 Interpretation

The residuals plot in Figure 18 is random around the y = 0 line, but also shows a slightly cyclic pattern. This can result from the low quality of the model. The distribution of the residuals, drawn in Figure 19 does not follow a normal distribution, but still appears to be equal on both sides of the 0. Therefore, we cannot use it to conclude on the quality of the model. But the quantile plot in Figure 20 shows that the relationship between the theoretical and sample quantiles is almost a perfect line. Overall, we can claim that the quality of the model is good.

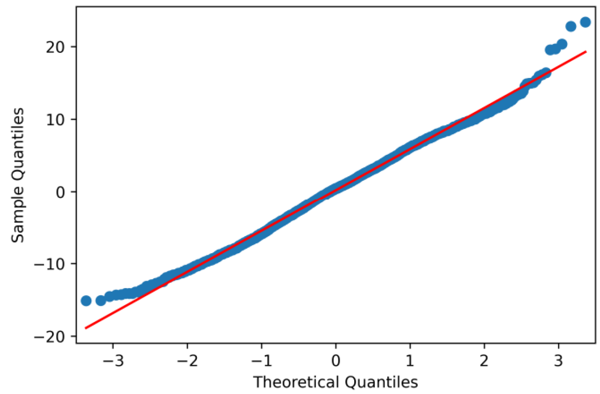


Figure 20: Quantile-quantile plot of the daily climatological baseline residuals

4.7 Performance Metrics

We grouped the performance metrics of the different models and baselines in the following table. The R2 corresponds to the coefficient of determination, the RMSE to the root mean squared error and the MAE to the mean absolute error of the models.

	R2	RMSE	MAE
	train/test	train/test	train/test
Linear interpolation	-0.183/-0.165	4.696/4.648	3.621/3.605
Nearest-neighbor interpolation	-0.188/-0.173	4.706/4.665	3.640/3.637
Quadratic interpolation	-0.191/-0.172	4.711/4.663	3.634/3.622
Cubis interpolation	-0.191/-0.173	4.712/4.664	3.635/3.623
Linear regression	0.163/0.167	3.949/3.931	3.098/3.089
Persistence	0.46	2.99	1.98
Monthly climatology	0.27	4.86	3.94
Daily climatology	0.73	5.67	4.58

Table 5: Performance metrics

5 Discussion

Both the interpolation and the linear regression show insufficient predictive ability. However, the shape of the cloud in Figure 6 seems more stretched, and might show that the interpolation is a bit better at predicting the observed temperatures in the area of Casablanca. It could also be the result of the split between training and test sets, that leads to less points to draw, and so to a less diffuse cloud. However, the randomness of the split in time should not lead to this kind of situation.

The results of the baselines correspond to what could be expected. The best model is the daily climatological baseline, which explains 73% of the variance between predicted and observed temperatures (Table 5). This result means that the temperature of a given day can efficiently

be forecast by assuming it is roughly the same as the average of the temperatures of the same day in the past. The quality of this baseline is confirmed by the analysis of the residuals: the quantile plot is the most fitted to the reference line. This result is probably due to the stability of the temperatures in Casablanca.
The persistence baseline also shows satisfactory results, with a percentage of explained variance of 46% and an average distance between the predicted values from the model and the actual values in the data set of 2.99. This last value is particularly low in comparison with the other RMSEs and shows the good quality of the model. The mean absolute error, which takes into consideration the size of the data set, is also particularly low. The relevance of the persistence baseline could be explained by the low temperature gradient in Casablanca. Indeed, as Table 4 and Figure 2 show, the temperatures in the city are in average comprised between 15 and 21, with only a slight difference from one month to the next, and the distribution is not spread out. This means that considering that tomorrow's temperature is today's temperature is a good approximation, and that the persistence baseline is a good model for Casablanca.
The monthly climatological baseline, on the other hand, lacks predictive ability. The R2 is poor, and the RMSE and MAE are large in comparison with the others. This means that the variation of temperatures within a month is large enough to make the prediction using monthly averages irrelevant.

The comparison between the models and the baselines enables to see how the first ones perform compared to the second ones. The linear regression shows poorer prediction ability than all of the baselines.
In terms of complexity, it appears that more elaborate models are not always more adequate or more fitted for downscaling. Indeed, all the interpolations used, while mathematically simpler than the linear regression, show better results in terms of predictability. But the small set of methods developed make the comparison difficult, and any conclusion on the initial question at this stage would be hasty.

6 Code and Data Availability

The scripts written for this study are available at:
https://github.com/meryamcherqaouifassi/2021_Bachelor_Thesis
The scripts written and applied to download and convert the data are in the *utils* folder. The scripts regarding to interpolation, simple linear regression and multiple linear regression are in the *dev* folder. The downloaded and converted data frame of Casablanca's weather station measurements is in the *files* folder. The figures displayed in this paper can be found in the *figures* folder.

7 Conclusion and Future Work

In this paper, the methodology used to obtain a clean data set of the observed two meters temperatures between 1979 and 2018 in Casablanca was presented. The two downscaling methods that were implemented using ERA5 data set were the interpolation and the linear regression. Persistence and climatological baselines were calculated to compare the performance of the models. The results show that both interpolation and linear regression have poor predictability abilities. It demonstrates the limits of simple downscaling methods, and their lack of physical consistency and forecasting relevance.
As for the initial question of complexity, the comparison between the models seems to indicate that a model does not need to be more complex to show better performance metrics in terms of predictability.

Giving the data pre-processing that this work involved, and because we chose to keep our results tractable, only simple downscaling methods were explored. However, other steps in the attempt to overcome the lack of studies on the Casablanca region can be taken.
First, the important variation in temperature gradients between land and sea can be addressed by splitting the data between midnight and midday temperatures, in order to build more robust models. More linear regression models can be calculated, by varying relevant features for two-meters temperatures, such as the pressure, the humidity and the geopotential, that would most likely improve the quality of the regression.
As a next step, more complex downscaling methods can be used, e.g. weather typing, weather generators, or non-linear machine learning methods.
Finally, multiple reanalysis models, using different atmospheric models can be applied and compared, to limit the individual bias of each one.

8 References

Akhter, M. (2017). Statistical Downscaling Based On Multiple Linear Regression Analysis for Temperature and Precipitation inLidder River Basin of India. *IOSR Journal of Engineering, 07*(06), 55-60. https://doi.org/10.9790/3021-0706015560

Casablanca. (2022). In *Wikipédia.* https://fr.wikipedia.org/w/index.php?title=Casablanca&oldid=193019083

Casablanca Climate, Weather By Month, Average Temperature (Morocco)—Weather Spark. (s. d.). Consulté 7 juin 2022, à l'adresse
https://weatherspark.com/y/32760/Average-Weather-in-Casablanca-Morocco-Year-Round

Climat du Maroc. (2022). In *Wikipédia.*
https://fr.wikipedia.org/w/index.php?title=Climat_du_Maroc&oldid=190857453

Climate & Weather Averages in Casablanca, Morocco. (s. d.). Consulté 7 juin 2022, à l'adresse
https://www.timeanddate.com/weather/morocco/casablanca/climate

Colette, A., Bessagnet, B., Vautard, R., Szopa, S., Rao, S., Schucht, S., Klimont, Z., Menut, L., Clain, G., Meleux, F., Curci, G., & Rouil, L. (2014). European atmosphere in 2050, a regional air quality and climate perspective under CMIP5 scenarios. *Atmos. Chem. Phys.*, 22.

Dima Burov. (2019, juin 5). *Introduction to Data Assimilation.*
https://www.youtube.com/watch?v=XwDu5MpeCFk

Duan, J. (2013). *Weather typing methods—Statistical analysis of large-scale climate controls on rainfall.*

GIEC. (2022). *IPCC WGII Sixth Assessment Report.*

Haby, J. (2022). *Z TIME.*
https://www.theweatherprediction.com/basic/ztime/

Hong, S.-Y., & Kanamitsu, M. (2014). Dynamical downscaling: Fundamental issues from an NWP point of view and recommendations. *Asia-Pacific Journal of Atmospheric Sciences, 50*(1), 83-104.
https://doi.org/10.1007/s13143-014-0029-2

IGI Global. (2022). *What is Baseline Climate — IGI Global.*

Jones, N. (2017). How machine learning could help to improve climate forecasts. *Nature, 548*(7668), 379-379.

https://doi.org/10.1038/548379a

Lachgar, R., Badri, W., & Chlaida, M. (2021). Assessment of future changes in downscaled temperature and precipitation over the Casablanca-Settat region (Morocco). *Modeling Earth Systems and Environment*.
https://doi.org/10.1007/s40808-021-01213-5

Liu, J., Yuan, D., Zhang, L., Zou, X., & Song, X. (2015). Comparison of Three Statistical Downscaling Methods and Ensemble Downscaling Method Based on Bayesian Model Averaging in Upper Hanjiang River Basin, China. *Advances in Meteorology, 2016*, e7463963.
https://doi.org/10.1155/2016/7463963

Maraun, D., & Widmann, M. (Éds.). (2018). Weather Generators. In *Statistical Downscaling and Bias Correction for Climate Research* (p. 201-219). Cambridge University Press.
https://doi.org/10.1017/9781107588783.014

Pachauri, R. K., Mayer, L., & Intergovernmental Panel on Climate Change (Éds.). (2015). *Climate change 2014: Synthesis report*. Intergovernmental Panel on Climate Change.

Palmer, T. (2019). The ECMWF ensemble prediction system : Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society, 145*(S1), 12-24.
https://doi.org/10.1002/qj.3383

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11).
https://doi.org/10.1029/2020MS002203

Sirocco. (2022). In *Wikipedia*.
https://en.wikipedia.org/w/index.php?title=Sirocco&oldid=1069154407

South Central Climate Adaptation Science Center. (2016, décembre 22). 2.2.2 *Dynamical downscaling*. https://www.youtube.com/watch?v=GHg7MCaSYEc

South Central Climate Adaptation Science Center. (2018, février 6). 2.2.1 *Statistical Downscaling*. https://www.youtube.com/watch?v=etaMadjy12k

Uncertainty in statistical downscaling of rainfall: Case study of south-east UK. (s.d.). Consulté 13 juin 2022, à l'adresse
https://1library.net/document/y92873jz-uncertainty-statistical-downscaling-rainfall-case-study-

south-east.html

Vandebogert, K. (2017). *METHOD OF QUADRATIC INTERPOLATION*. 22.

Wilby, R. L., Dawson, C. W., & Barrow, E. M. (2002). sdsm—A decision support tool for the assessment of regional climate change impacts. *Environmental Modelling & Software, 17*(2), 145-157.
https://doi.org/10.1016/S1364-8152(01)00060-3

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., & Wilks, D. S. (1998). Statistical downscaling of general circulation model output : A comparison of methods. *Water Resources Research, 34*(11), 2995-3008.
https://doi.org/10.1029/98WR02577