Universität Basel

Departement
Umweltwissenschaften

D|U Departement Umweltwissenschaften

Bachelor in Geowissenschaften
Vertiefungsrichtung: Klimatologie

Bachelorarbeit von: Elias Fessler

Accuracy and variability of six global temperature model forecasts in 2018

Betreuung Dr. Roland Vogt

Beurteilung

Universität Basel
Departement Umweltwissenschaften
Meteorologie, Klimatologie & Fernerkundung
Klingelbergstrasse 27 – 4OG
4056 Basel, Switzerland

Universität Basel
Departement Umweltwissenschaften

D|U Departement Umweltwissenschaften

Bachelor's Program in Geosciences

Thesis

Accuracy and variability of six global temperature model forecasts in 2018

Elias Fessler
Sonnenbergstrasse 13
6052 Hergiswil
Switzerland
16-062-085
elias.fessler@bluewin.ch

Abstract. This thesis is a comparison of six different global models, which forecast the temperature for every hour 24 h in advance at 2-m above ground, taken at different terrestrial stations in 2018. Five raw numerical weather forecast models and one reanalysis model are analysed. A time series of one year reporting hourly data from 19150 stations has been used. An identification of four error types under specific conditions was made to locate global patterns and define regions with different accuracy and variability in their predictability. The error increases with a longer distance between the station and the model grid point. Globally, model temperature forecasts have the highest predictability on small oceanic islands and along ice-free coasts, where the air temperature is almost entirely regulated by the sea surface temperature. On the mainland the smallest errors are found in Northern Europe from Northern France to the coast of Scandinavia. The predictability decreases in regions with complex topography and increasing distance from the sea. A different performance depending on the climate zones can be seen.

Bachelor's Thesis (10 credits ECTS)
August 2019
Supervisor: Dr. Sebastian Schlögl
Examiner: Dr. Roland Vogt

Contents

List of Figures

List of Tables

1

Introduction

Meteorologists have set themselves the goal to understand the processes in the atmosphere. Besides organisations, individual people and nations rely on model weather forecasts for all kind of tasks. Temperature is among of the forecasted variables. There are several model types that predict a possible model weather forecast. Raw model weather forecasts are one of them. They are computations of one particular model run. These raw model forecasts can contain large systematic errors. Next to imperfect initial conditions and different model parametrisations there is the error due to representativeness. This error is a result of the fact that the temperature is calculated for an area of a grid cell and not for the specific location of the weather station [19].

This thesis focuses on a global verification of the 2-m above ground temperature forecast of five raw numerical weather forecast models and one reanalysis model over the year 2018 using 19150 stations to represent the world maps and nearly 9000 weather stations for statistical analyses.

It is the objective to give an overview on how the different model errors change under specific conditions. Besides defining regions with possible higher and lower variability, their different biases were also examined. This allows a better understanding of the forecast accuracy in different parts of the world.

In the last few years, current model data have become accessible for the general public. However, archived meteorological data are still not publicly accessible. Thus, the model data was downloaded and stored by meteoblue (meteoblue AG). All model data used in this study are 24-h forecasts. The data of all six global models were only available for the last few years. This is a reason why so few comparisons of these six global forecast models have been done in the past. A standardised comparison on global atmospheric forecast models is still missing.

In this thesis a dependency of the geographical latitude and longitude and the height of the station is considered. Furthermore, besides the horizontal and vertical distances between the station coordinates and the model grid point, an analysis on the climate regions and two different model spreads will be taken into account.

2

Theory

2.1 Development of Numerical Simulations

One of the most important achievements of the last century was the ability to simulate complex physical systems using numerical models. As a result, it is now possible to predict the weather and gain knowledge of the factors that control weather patterns [10].

100 years ago, forecasts were imprecise and unreliable. Observations were scarce and irregular and physical laws were neglected. In 1901, the American meteorologist Cleveland Abbe first proposed a mathematical approach to weather forecasting. Shortly after, the Norwegian scientist Vilhelm Bjerknes proposed a two-step plan: After determining the state of the atmosphere, a forecast could be done on the basis of different laws of motion. The English scientist Lewis Fry Richardson first started studying old weather charts. He started making predictions by observing weather situations of the past and their developments. He advanced algorithms on the basis of Bjerknes work that are notably similar to the algorithms used now. With these computations and mathematical systems, Bjerknes and Richardson laid the foundations for modern forecasting - however, this pre-computer era lacked the necessary computation power [10].

With a better understanding of atmospheric dynamics, advances in numerical analysis, the three dimensional knowledge about the state of the atmosphere with radiosondes, and the development of the digital computer, the scientific community increased its capability to make better weather forecasts [10].

As well as global models that cover the whole planet, there are regional weather forecast models only covering specific regions. Nowadays, a combination of global and local models is common. In 1979 the European Centre for medium-range weather forecasts (ECMWF) started their work with the first operational forecast. In the early 1990s the ECMWF began with ensemble forecasting. Multiple model runs with slightly different initial states deliver differing final conditions. As a result, combined outputs can be used to estimate future atmospheric states. Several outputs close to each other suggest a higher probability and vice versa [10].

Today, the world is clustered into model domains (fields) with a specific side length called spatial or horizontal resolution. These model domains are mostly arranged in rectangles (ICON icosahedral) and are evenly spaced between each other. Besides the surface, the

model domain typically divides the atmosphere vertically into 55 levels of about 14 km [18]. In the last few years, several probabilistic methods based on ensemble forecasting, bias removal and multi-model approaches have been developed. There are increasing numbers of available regional high-resolution weather forecasts, like NEMS (NOAA Environmental Modeling System) and NMM (Nonhydrostatic Meso-Scale Modelling) [18]. They are enhanced with new developments of model physics and parametrisations or post-processing methods. This ensured that the resolution has evolved over about two orders of magnitude in the last decades [20].

2.2 Model Types

There are different types of model data.

Raw data: The raw model output data are the computations of one particular model. The data are the output of one model run with a specific parametrisation and have no post-processing step ('stand-alone') [15].

MOS: The model output statistics (MOS) combines the raw model output and the observations to form a statistical relationship. The raw model output is post-processed using statistics from local historical weather measurements, thereby improving the forecast accuracy [13]. MOS is able to remove systematic forecast errors. In addition to the relation to the season and the forecast hour of the day, MOS usually uses measurements from the day before to compute a forecast [11, 19]. Each station has different MOS equations, but they are derived with the exact same algorithm [20].

Reanalysis: A reanalysis model takes into account a historical climate analysis, measurements, observations and simulations for its model correction [12].

mLM: The meteoblue learning multimodel (mLM) is a multi-model approach which uses actual weather measurements to post-process the numerical forecast output. To forecast, it first compares actual measurements with the different model forecasts to increase the accuracy by making increasing use of artificial intelligence. While MOS is working with one model, mLM includes the comparison of different weather forecast models including ERA5 and as well MOS [16, 17].

With better knowledge of atmospheric initial conditions, better parametrisation of the model computations, and more powerful computer performance - which enables higher resolution – the forecast can be improved [15, 22].

2.3 Models

For the survey six global models were chosen of which four are global forecasting systems of national weather services. Five models are raw model forecasts without removed bias and one reanalysis model. The models have run over the year 2018, computing hourly 24-h forecasts in their specific spatial resolution.

- ERA5:
  Developer: ECMWF (European Centre for Medium-Range Weather Forecasts)
  Spatial resolution: 30 km
  ERA5 is the reanalysis model of the ECMWF [7].
- NEMSGLOBAL (NEMS):
  Developer: meteoblue
  Spatial resolution: 30 km
  NEMSGLOBAL is global model of the NEMS multi-scale models of meteoblue [14].
- GFS05:
  Developer: NOAA and NCEP (National Centers for Environmental Prediction)
  Spatial resolution: 40 km
  GFS05 is the global forecasting system of the United States [21].
- MFGLOBAL (MF):
  Developer: Méteo-France
  Spatial resolution: 40 km
  MFGLOBAL is the global forecasting system of France [5].
- GEM:
  Developer: Canadian Meteorological Centre (CMC)
  Spatial resolution: 25 km
  GEM is the global forecasting system of Canada [4].
- ICON:
  Developer: DWD
  Spatial resolution: 13 km
  ICON is the global forecasting system of Germany [6].

For further information about a model follow the corresponding link in the bibliography.

3

Methods

3.1 Data Review

The data-set was provided by meteoblue and contained hourly temperature measurements from 23460 weather stations world-wide. The measurements were mainly received from the WMO (World Meteorological Organisation) and GDAS (Global Data Assimilation System from NOAA [National Oceanic and Atmospheric Administration]).

The measurement data-set was cleaned up. On all stations the data coverage was calculated. The coverage was calculated as the percentage of the useable data (temperatures between -100 and 70 °C) per year in relation to the maximum possible data quantity of 8760 hourly measurements per year in 2018.

Besides unrealistic temperature values the data-set contained stations with wrong coordinates. Only stations within the worldwide grid from -180° to 180° longitude and 90° to -90° latitude were accepted. Duplicated stations with the same coordinates were removed under the condition that the station with the highest coverage values remained in the data-set. This happened because the data was not provided from one organisation alone. After the clean up the data-set contained 19150 stations.

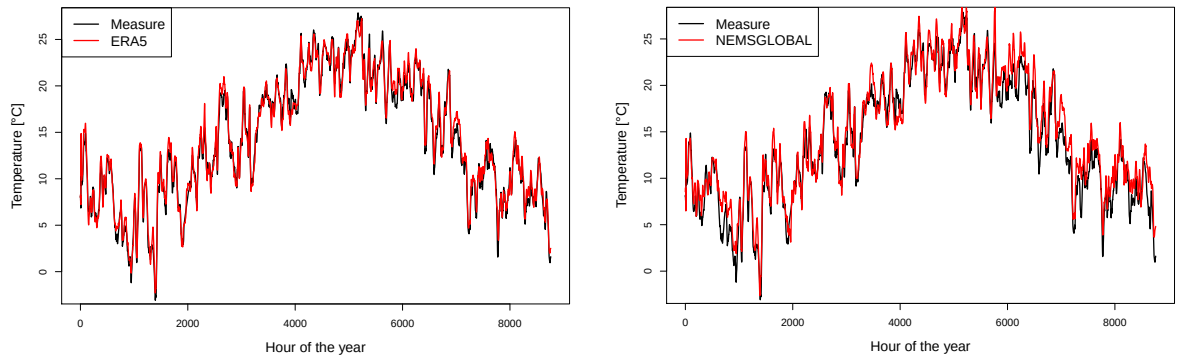

Figure 3.1: Annual cycle of the two model forecasts ERA5 (red, left) and NEMS (red, right) in relation to the annual cycle of the measurements (black) at a station in Southern France (43.83 N 0.02 W, 145 m a.s.l., Coverage: 98.45 %)

Temperature forecasts of the six models were downloaded from the meteoblue interface (history+ advanced access) for 19150 stations. Next to the hourly measured data there were six global temperature forecasts that allowed a global comparison.

Figure 3.1 shows that ERA5 (*left*) has a better overlap of the annual cycle of the measurements compared to NEMS (*right*). For the annual distribution on two models the function moving average "SMA" from the R-package "TTR" was used [25] to smooth the yearly cycle of the data.

With the given coordinates of the stations and the model grid points the difference in height and the horizontal distance was calculated. The difference in height results from the height of the station and the height of the individual models' grid point, both being surface data. It was calculated as

$$\Delta h \; = h_{model} \; - \; h_{measurement} \tag{3.1}$$

with height h in [m]. Thus, a negative height corresponds to a model height being less than the measurement height and vice versa.

The temperature forecasts are located at the coordinates of the intersection points of the individual model grids. With the meteoblue interface taking the next grid point the theoretical maximum horizontal distance is half of the diagonal of the model resolution. The distance Δd can differ whether the corresponding station is near the grid point or at maximum distance $max\ \Delta d$ of:

$$max\ \Delta d \; = \; \sqrt{(a/2)^2 + (a/2)^2} \; = \; \frac{a}{2}\sqrt{2} \; = \; \frac{a}{\sqrt{2}} \tag{3.2}$$

where a is the side length of the grid or spatial resolution.

The difference in degree latitude (Δlat [°]) and longitude (Δlon [°]) was calculated similarly. The latitude or longitude value of the station was subtracted from the models. For the calculations the earth radius r_{earth} was taken as constant with 6370 km. Because of the constant number of km per degree in latitude ($kpdl$) the difference is multiplied with the factor $kpdl = 111.2$ km/°lat.

$$kpdl = 2 * \pi * r_{earth} \; / \; 360 \tag{3.3}$$

$$\Delta LAT = kpdl * \Delta\ lat \tag{3.4}$$

Differing from the latitude distance ΔLAT [km], the longitude distance ΔLON [km] is calculated with the longitude difference Δlon [°] which used the cosine of the geographical latitude lat [°]. Consequently the specific latitude circle of the earth was calculated as

$$\Delta LON = \cos\,(lat) \; * \; kpdl \; * \; \Delta\ lon. \tag{3.5}$$

With the two distances in north-south and east-west direction the direct distance ΔD [km] is calculated with the law of Pythagoras as

$$\Delta D = \sqrt{\Delta LAT^2 + \Delta LON^2}\,. \tag{3.6}$$

The horizontal distances are absolute values. Thus, it is not possible to say whether the station is located north, south, east or west of the grid point. The height difference and the earth's curvature were not regarded here. The distance Δd between the station and the model grid point should have been limited to the half of the diagonal (Formula 3.2). The distance was sometimes a multiple of the theoretical maximum. For verification, the distances were also calculated with the "haversine" function in the R-package "pracma" [2]. The "haversine" function calculates the arc distance between two points on planet Earth. The result was almost the same.

Therefore, it should be stated, that under specific conditions another grid point has been taken instead of the one next to it. In regions where the difference between the height of the model grid and the measurement is large the interface picks another grid point within 3 x 3 grid points. Secondly, the interface takes into account that at coastlines a grid point over land is always taken. Like this a pattern can be seen in Figure 3.2. It shows the stations that have a higher horizontal distance to their corresponding grid point than theoretically possible are mainly located at coastlines and mountainous regions.

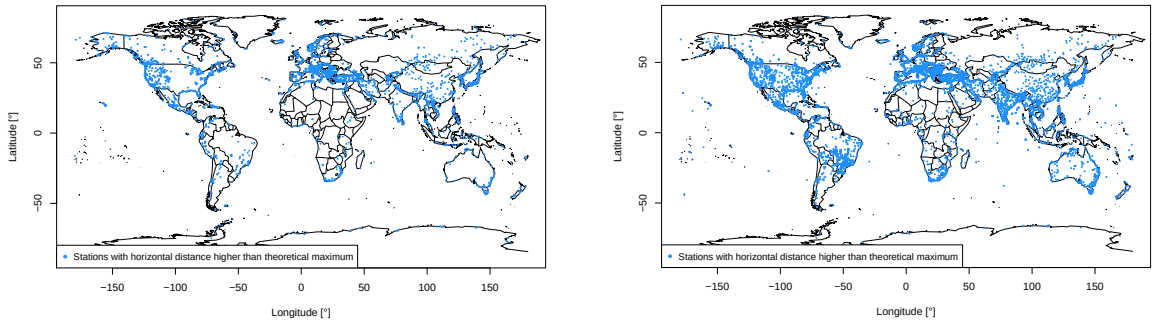

Figure 3.2: Stations (blue) that have a higher distance to their ERA5 grid point (left) and their NEMS grid point (right) than the theoretical maximum regarding the specific spatial resolution (Appendix B.0.3)

In Figure 3.3 the black points are the ERA5 grid points used in this survey. The missing ERA5 grid points have no corresponding weather station inside this thesis and thus are not contained in the data-set. The sector of the map is plotted rectangularly. However, the distances between the points are constant. The green point is the station with the maximum horizontal distance on ERA5 to the next model grid point. The interface of meteoblue has picked the red point (as well ERA5 grid point) and not the nearest grid point south. The result is a higher horizontal distance.

To analyse the dependency on representativeness three different coverage filters were applied. The coverage filters were set on 30, 60 and 90 % (C30, C60 and C90). In the data-set with C30, only the stations with coverage values higher than 30 % were included. The results were different numbers of stations for the following analysis:

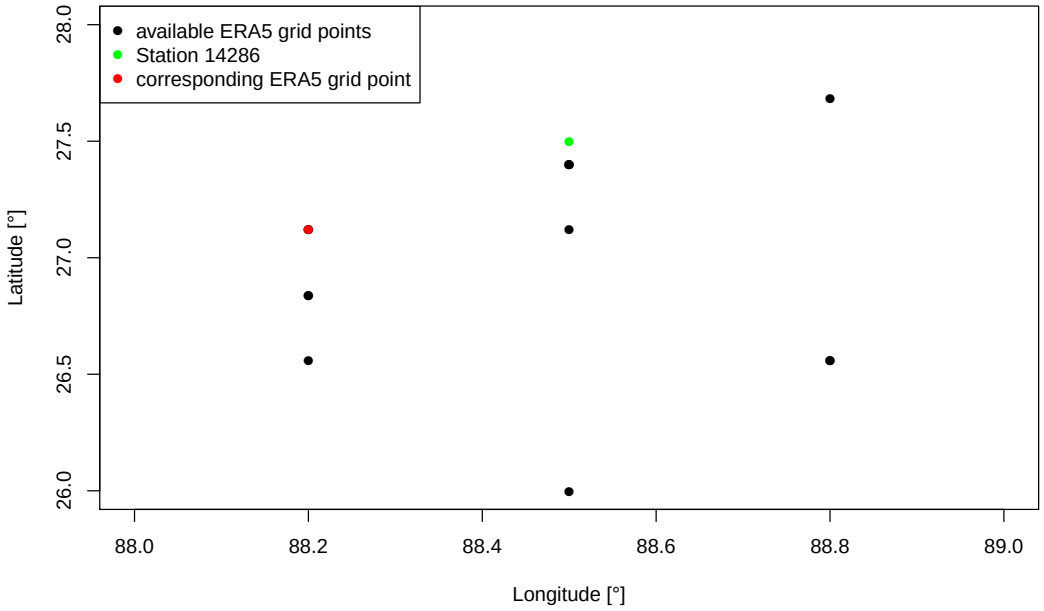

Figure 3.3: Station 14286 in Northern India with the biggest horizontal distance (green) and its corresponding ERA5 grid point (red) in the environment of the other in this data-set used ERA5 grid points (black)

All stations: 19150
30 % coverage: 12515 (C30)
60 % coverage: 8228 (C60)
90 % coverage: 6351 (C90)

Because weather stations are mostly limited to easily accessible locations and land surface there is a mismatch in their distribution. While Europe and North America are densely populated with both people and stations, regions like deserts and the high seas are not covered satisfactorily [23].

3.2 Statistical Analysis

For the major analyses four errors have been computed for the hourly data. The Mean Absolute Error (MAE), the Mean Bias Error (MBE), the Root Mean Squared Error (RMSE) and the standard deviation (SD) all in [K] are computed as

$$MAE = \frac{1}{n}\sum_{i,t}|(m_{t,i} - o_{t,i})| \tag{3.7}$$

$$MBE = \frac{1}{n}\sum_{i,t}(m_{t,i} - o_{t,i}) \tag{3.8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,t}(m_{t,i} - o_{t,i})^2} \tag{3.9}$$

$$SD = \sqrt{\frac{1}{n-1}\sum_{i,t}(\Delta T_{t,i} - \overline{\Delta T}_i)^2} \tag{3.10}$$

where m is the model data [°C], o the observation [°C] and n the number of corresponding temperature pairs (model and measurement). ΔT is the temperature difference, the observation temperature subtracted from the model temperature and $\overline{\Delta T}$ the mean of the temperature difference per station. All tables in this thesis are produced with the help of the R-package "stargazer" [8]. The summation is done over time t and then over station i. A negative MBE results when the observed temperature is higher than the model temperature and vice versa.

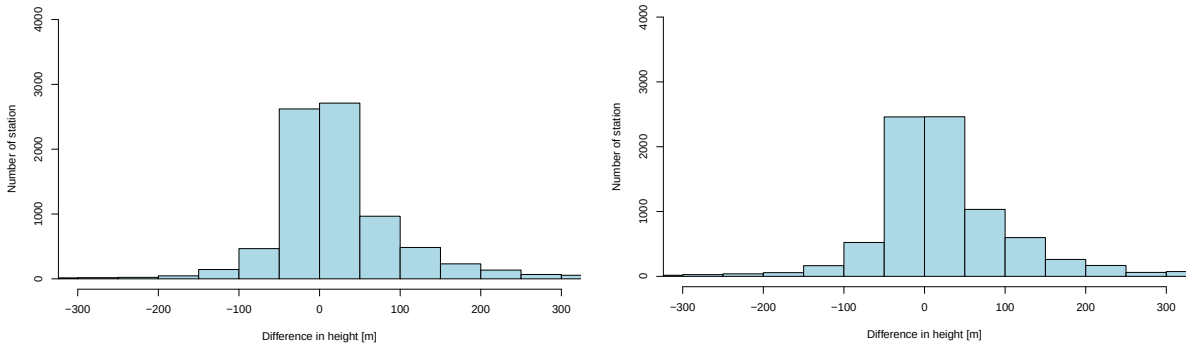

Figure 3.4: Height difference distribution on ERA5 (left) and NEMS (right) (Figure B.2)

Figure 3.4 shows that the majority of the stations has an absolute height difference of less than 100 m.

Additionally, an analysis of daily averages was done. It is worth mentioning that for the daily error analysis the 24-h windows are from 00:00 to 00:00 UTC. Thus, depending on the geographical positions of each station, the 24-h window is shifted. In other words, a day can for example be from 11:00 AM to 11:00 AM.

The whole analysis was done with the "aggregate" function from the R-package "stats" [24]. A comparison of the hourly mean and the daily mean was done. Additionally the model performance predicting the maximum and minimum of a 24-h window was tested.

Note that for the comparison the maximum and minimum of the model forecast in this 24-h window was not necessarily regarded. The model temperatures at the same time where the measurement has its maximum or minimum were taken into account. Besides the absolute value in the window also the exact hour of predicted extrema is relevant. In other words, the absolute maximum forecast could have been accurate but shifted about two hours. Thus, at the specific time of the measurements extrema the forecast is less accurate.

3.3 Downscaling Approach

Because of the height difference between model grid point and station a simple downscaling attempt was applied. The new temperatures were calculated with four different lapse rates Γ: 1.0, 0.8, 0.65 and 0.55 K / 100 m in height to see a possible trend. Hence, a temperature value for a specific hour and station was corrected with the lapse rate multiplied by the difference in height. The downscaled temperature was calculated as

$$T_{downscaled} = T_{model} + \Gamma * \Delta h \tag{3.11}$$

with the raw model temperature T_{model} [K].

A negative difference in height lowers the predicted temperature. The station is higher than the model and thus, the model temperature gets lowered to approach the measurements height.

3.4 World Maps

World maps were plotted to display the analyses for each error and model. Within these maps, it is easier to visualise patterns. Depending upon which errors are plotted first (i.e. starting with large errors and finishing with the small errors, or vice versa) the world map, with the same data, can be interpreted in different ways. If large errors are plotted first, the map shows very good predictability. With inverted plotting order, the world map shows poor predictability.

This is because of high station density in selected regions over the world and the extent of the points representing them; not all stations could be visualised at once. Stations that were plotted first were over-plotted and hence not visible in the corresponding world map. This means that error plots of predictability can be non-representative of the true data. To avoid this, the world was gridded in model grid cells with a horizontal resolution of 2° and 5°. All stations within one of these fields were merged first and then plotted in the centre of the field. The outcomes were tables with 16200 and 2592 values respectively. If a field does not overlap any station, there is consequently no number and thus no coloured point. This technique applies weighting to the global errors.

This was done for the mean, median, maximum and minimum error of the fields, on MAE, RMSE and SD, on the six models on 2° and 5° horizontal resolution maps with all (19150) and C60 (8228) stations, resulting in 288 maps (4 x 3 x 6 x 2 x 2). For the MBE only the mean of the six models MBE was plotted into 2° clustered maps.

Additionally, world maps were generated, where the model with the highest or smallest MBE or the model with the smallest MAE is shown. The fields are coloured to the corresponding model with the lowest MAE and the highest and lowest MBE compared to the others. In other words, the cluster was coloured black if ERA5 had the lowest MAE. The maps could show patterns where some models have the tendency for the highest MBE or the smallest MBE or MAE.

3.5 Climate Zones

Furthermore, the performance of the six models depending upon which climate zone each station is located in is of interest. To classify each station, the Koeppen-Geiger climate zones of the R-package "kgc" were used [3]. These zones were considered in a coarse 5-zone set and a more specific 30-zone set.

The MAE and MBE of all 8228 stations (C60) within a specific climate zone were averaged. Thus, it was possible to give a statement about the predictability in each different climate zone.

3.6 Model Spread

After the climate zones two different model spreads were analysed. The model spread is an index for the predictability of a forecast. It only regards the six different model temperature forecasts without the actual observation. A small temperature difference results if the models all predict approximately the same. Conversely the difference is large when the models do not concur with each other.

For the first model spread 'max-min', the minimum temperature forecast of the six models was subtracted from the maximum at the same hour and station. To know the model with the maximum and minimum temperature was not necessary. The second model spread 'standard deviation' was analysed with the standard deviation of the six models for each hour and station.

The two 2° gridded model spread maps were approached twice but differently. In the end, 4 model spread maps were obtained. In the first approach both model spreads were calculated separately for each station first and later the multiple model spreads were averaged within the grid.

In the second approach the averaging within a grid was done before the two model spread calculation. From all stations in a specific grid all temperatures of a model at a certain time were averaged. These values were used to calculate the 'max-min' and 'standard deviation'. In other words, for example, three stations are within a grid. The three ERA5 temperatures for the first hour are averaged, identical with all six models. From the resulting six values the maximum, minimum and the standard deviation were delivered. The following is similar to the calculation for each station.

4

Results

4.1 Analyses

4.1.1 Overview coverage

In Figure 4.1, *left*, the 2° gridded coverage values are shown. In Europe and North America the temporal coverage is highest. Also Greenland, Australia, Japan, Malaysia and the Arabian Peninsula have high coverages. Regions with low coverages are Brazil, Africa, India, Russia and Siberia. For the white areas not one weather observation is reported in this data set.

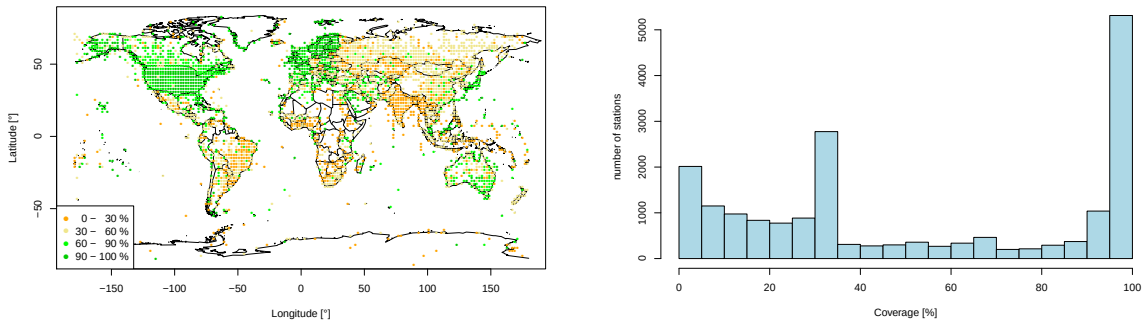

Figure 4.1: Clustered station coverage coloured subdivided into the four coverage classes (left) and coverage distribution of all stations (right)

The coverage distribution is shown in Figure 4.1, *right*. About a quarter of the stations deliver more than 95 % of the hourly measurements. Interestingly, a large portion of the stations only provide one third of these measurements. This is because the afore mentioned stations deliver 3-hourly data. It is important to distinguish the temporal coverage and the temporal resolution. The temporal coverage is again the percentage of the useable data in relation to the maximum possible data quantity of 8760 hourly observations. The temporal resolution can be hourly or 3-hourly. Like this, a 3-hourly data-set with 100 % temporal coverage (2920 observations) automatically does not fulfil the conditions for C60 or C90 in this survey. As a result, about 43 % of the stations comply with C60.

4.1.2 Raw model vs different coverage data

The reanalysis model ERA5 with a twice coarser horizontal resolution performs better than ICON. A higher spatial resolution does not automatically mean more precision, if not only raw model forecasts are considered (Table 4.1, *left*) [1]. Besides the best possible forecast, a stable and reliable bias-removal technique is important [20]. Due to the temperature correction, the reanalysis model ERA5 forecast is best under normal conditions. ICON with its higher spatial resolution performs better than the other raw model forecasts and not quite as good as ERA5. ICON achieves a MAE less than 2 K with coverage 90 % (Appendix A.0.1).

Considering the raw model performances, there is a slight tendency of a smaller errors with higher model resolution. [20]. With an increase of the spatial resolution the error due to representativeness should decrease with the decline of the maximum horizontal distance [19]. Interestingly, GEM gets outperformed by all models despite having the second highest spatial resolution.

Table 4.1: Error comparison [K] *with all (19150)(left) and C60 (8228)(right) stations (Appendix A.0.1)*

	MAE	MBE	RMSE	SD
ERA5	1.7	0.3	2.2	1.9
NEMSGLOBAL	2.4	0.4	3.1	2.6
GFS05	2.6	0.3	3.4	3.0
MFGLOBAL	2.7	-0.2	3.4	3.1
GEM	2.8	-0.8	3.5	3.1
ICON	2.3	-0.1	3.0	2.7

	MAE	MBE	RMSE	SD
ERA5	1.5	0.2	1.9	1.8
NEMSGLOBAL	2.2	0.1	2.8	2.5
GFS05	2.3	0.2	2.9	2.7
MFGLOBAL	2.3	-0.1	3.0	2.8
GEM	2.4	-0.7	3.0	2.7
ICON	2.0	-0.1	2.5	2.4

As can be seen in Table 4.1 and additionally Appendix A.0.1 there is a relation between the errors and the model forecast grid points lying next to stations covering only small parts of the hourly time series. Thus, the errors get smaller if the interpreted data does not contain the stations with small coverage values. By taking the mean, large errors have a major effect on the forecast accuracy and alter the error range. A worse temporal resolution downgrades the correlation. The MAE gets smaller by around 0.3 to 0.5 K comparing C90 to all stations. The MBE decreases by around 0.1 to 0.4 K, the RMSE by around 0.3 to 0.6 K and the SD between 0.1 and 0.4 K.

ERA5, NEMS and GFS05 tend towards warm biases, while MF, GEM and ICON have cold biases. GEM with -0.7 K has cold biases. The lowest standard deviation was found for ERA5, followed by ICON, NEMS, GFS05, GEM and MF.

Regarding this, all the following computations are done with the data for C60. Thus, it can be ensured that stations with a small coverage are removed while enough stations are present to allow a global comparison of the data. Only for plotting the world maps all stations were used. Thus, it can be ensured that remote regions with small coverage values are still represented and visible in the map.

With C60, 86.3 % of the stations have an error less than 2 K with ERA5 (Table 4.2). ICON has 61.7 %, NEMS 50.8 %, GFS05 39.6 %, GEM 36.6 % and MF 35.5 % below 2 K. A higher percentage is within the 2 K range by neglecting stations with lower coverages (Appendix A.0.5).

Table 4.2: Percentages of 8228 stations (C60) per model in a specific MAE range are shown (Appendix A.0.5)

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	19.1	46.8	20.4	7.7	2.9	3.1
NEMSGLOBAL	1.2	9.2	40.4	28.9	10.9	9.4
GFS05	2.2	11.9	25.5	31.9	13.8	14.7
MFGLOBAL	1.9	9.5	24.1	32.6	17.4	14.5
GEM	2.5	11.8	22.3	32.1	13.8	17.5
ICON	5.3	22.2	34.2	22.7	7.9	7.7

4.1.3 Hourly vs daily mean

For an additional test the hourly data set was divided in daily means with the objective of getting an overview of the forecast accuracy on daily mean temperatures.

Table 4.3: Error comparison [K] *on the daily mean forecast with C60 (8228 stations)*

	MAE	MBE	RMSE	SD
ERA5	1.0	0.2	1.3	1.0
NEMSGLOBAL	1.6	0.1	2.0	1.7
GFS05	1.4	0.2	1.8	1.4
MFGLOBAL	1.6	-0.1	1.9	1.6
GEM	1.6	-0.7	1.9	1.5
ICON	1.1	-0.1	1.4	1.2

Comparing the error from the hourly mean (Table 4.1 (*right*)) and daily mean (Table 4.3) it shows that a daily mean temperature is simpler to predict than an hourly mean. This is a result of the smoothing of the mean function. All daily mean forecasts are better than the hourly forecasts, the RMSE decreases up to 1.1 K and the MAE up to 0.9 K. It reconfirms that ICON performs quite as well as reanalysis model ERA5.

4.1.4 Daily maximum vs daily minimum forecast

Besides daily mean, a prediction error on the daily minimum and maximum was computed. In Table 4.4 it is interesting to see that ERA5, NEMS and GFS05 are better in predicting the maximum of a daily window regarding the MAE and RMSE. Compared with MF, GEM and ICON whose MAE and RMSE is smaller, when modelling the minimum temperature.

Not taking into consideration MF and GEM, the worst performing model is NEMS when it comes to the daily mean prediction. However, with a resolution of 30 km NEMS outruns ICON with a 13 km resolution and achieves the second best model in the maximum prediction. All models get worse in predicting the daily extrema. Perhaps the models would have had a more precise forecast if the hour of the measurement extrema had been disregarded.

Table 4.4: Error comparison [K] *on the daily maximum (left) and daily minimum (right) forecast with C60 (8228 stations)*

	MAE	MBE	RMSE	SD
ERA5	1.5	-0.8	1.9	1.5
NEMSGLOBAL	2.2	-0.9	2.7	2.2
GFS05	2.4	-1.5	2.9	2.2
MFGLOBAL	2.7	-2.0	3.3	2.3
GEM	2.7	-2.2	3.3	2.2
ICON	2.3	-1.8	2.8	1.9

	MAE	MBE	RMSE	SD
ERA5	1.8	1.2	2.3	1.7
NEMSGLOBAL	2.4	0.9	3.1	2.6
GFS05	2.7	2.0	3.3	2.3
MFGLOBAL	2.6	1.8	3.3	2.4
GEM	2.4	0.6	3.0	2.4
ICON	2.2	1.7	2.8	2.0

The MBE values show that in all models the daily maximum is predicted too low and the minimum too high. The models do not have the same amplitude as the measurements but are mostly in the range between the minimum and maximum. In this case the models do not exactly forecast the extreme high and low temperatures. Such tables suggest that different models are tuned differently for their forecast. One performs better at the daily mean the other at an extrema. ERA5 and NEMS have a good performance at both extrema. GEM MBE for the daily minimum forecast with 0.6 K is the lowest. ERA5 for maximum forecast is still the best performing model, outrun by NEMS and GEM in minimum forecast. Averaging the MAE, the maximum is slightly better to predict than the minimum. Conversely over the MBE the minimum is better to forecast, primarily due to the good performance of GEM.

4.1.5 Mean vs median error

In addition to the mean, the median of the C60 errors for each model was compared to estimate the impact of possible outliers. The median MAE is at most 0.2 K lower than the mean MAE. Since the differences between the mean and the median (Table 4.1 (*right*) and Table 4.5) are small and the coverage is already included the mean was taken for further analyses.

Table 4.5: Error comparison [K] *on the hourly data set using the median with C60 (8228 stations)*

	MAE	MBE	RMSE	SD
ERA5	1.3	0.2	1.7	1.6
NEMSGLOBAL	2.0	0.2	2.6	2.4
GFS05	2.1	0.2	2.8	2.6
MFGLOBAL	2.2	-0.0	2.9	2.7
GEM	2.2	-0.5	2.9	2.7
ICON	1.9	-0.0	2.4	2.3

4.1.6 Error distribution

The data set has a right skewed distribution. Thus, the mean is bigger than the median. Most of the data are in a common range with some stations that have outliers, see Figure 4.2.

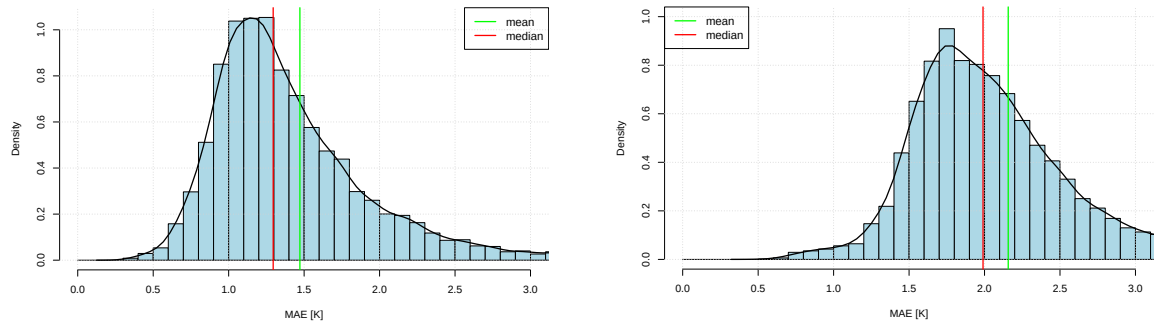

Figure 4.2: Histogram of MAE distribution of ERA5 (left) and NEMS (right): mean (green) and median (red)

In Figure 4.3 the MAE and the RMSE distribution is shown. Note that the stations are sorted according to the error size. Hence, the stations and their error do not correspond between the two graphs nor between each line in one graph. Reanalysis model ERA5 has the smallest error. ICON is the second best performing model followed by NEMS.

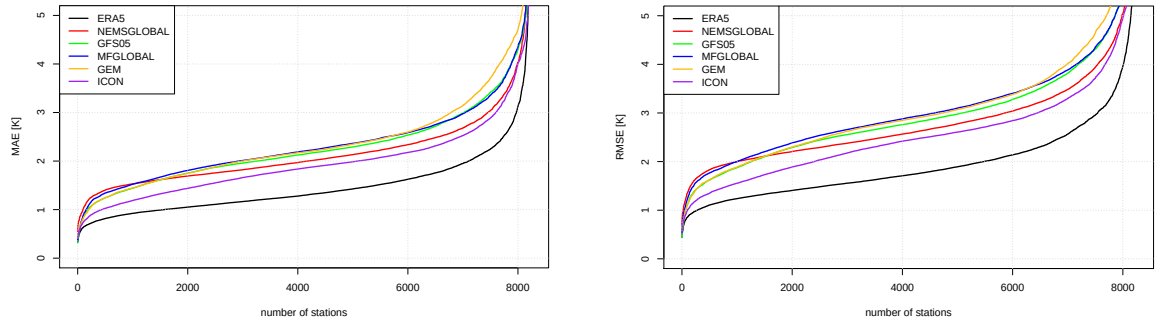

Figure 4.3: MAE (left) and RMSE (right) at 8228 station (C60) for five raw models and one reanalysis model

The advantage of a higher model resolution of GEM compared to GFS05 and MF cannot be seen. Interestingly, GEM has even more stations with a high error and fewer with a small error compared to NEMS, GFS05, MF. The progressions are comparable to the year before [23].

4.1.7 Raw vs gridded data

To avoid over-plotting points in the R-program as shown in Figure 4.4 the world was gridded in model grid cells of 5° and 2° side length. All stations that overlapped were representative for this particular field. Alternatively to the grid, the point size could have been minimised or only particular regions could have been plotted to prevent an overlap.

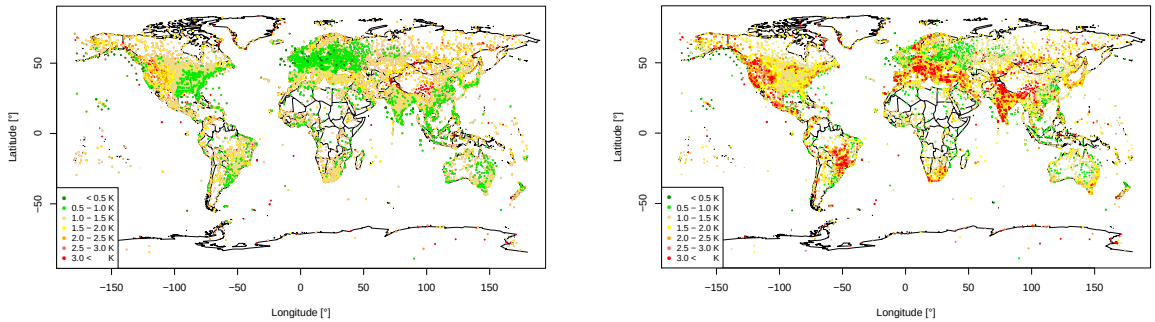

Figure 4.4: Problem of over-plotting: (left) the stations with a small MAE on ERA5 plotted at last, (right) the ones with a high MAE last

By all the stations in one square of 5° side length or 2° respectively ending up in one value the error in regions with a higher station density gets a higher weighting. Regarding Table 4.6, *left*, the mean MAE is constant and only increases on NEMS and MF. The RMSE increase on all six models between 0.1 and 0.2 K.

The increase of 0.1 K on NEMS is seen in both Tables 4.6, the one of MF only in the 5° gridded (*left*). In 2° gridded (*right*) the impact of the cluster is already minimised because it is again nearer to the version without grid. On ERA5 and NEMS the MBE decreases by clustering the world in squares of 5° side length and contrary on the other models.

Table 4.6: Error comparison [K] *with C60 on the 5° clustered (left) and the 2° clustered (right) (Appendix A.0.3 and A.0.4)*

	MAE	MBE	RMSE	SD
ERA5	1.5	-0.1	2.0	1.8
NEMSGLOBAL	2.3	-0.0	3.0	2.6
GFS05	2.3	0.3	3.0	2.7
MFGLOBAL	2.4	-0.4	3.2	2.9
GEM	2.4	-0.8	3.1	2.7
ICON	2.0	-0.2	2.6	2.4

	MAE	MBE	RMSE	SD
ERA5	1.5	0.1	2.0	1.8
NEMSGLOBAL	2.3	0.0	2.9	2.6
GFS05	2.4	0.3	3.0	2.8
MFGLOBAL	2.4	-0.3	3.2	2.9
GEM	2.5	-0.8	3.1	2.8
ICON	2.0	-0.1	2.6	2.5

4.1.8 Raw vs downscaled data

In Table 4.7 the errors after recalculating the model temperatures with a simple downscaling attempt are shown. The MAE remains constant or lowers at most 0.2 K, on GEM. MBE and RMSE decline as well, the standard deviation remains the same.

By plotting all the errors in dependency of the difference in height, it was not easy to see a simple trend. With the aggregated forecast data, the progression of the error with the increasing difference in height can be shown on NEMS (Figure 4.5). In the global temperature model forecast the MBE is significantly dependent on the difference in height between the station and the model grid point. Systematic errors, like these caused by height differences, can be removed with post-processing methods [19]. The reanalysis model ERA5 and NEMS have the smallest decline in MAE (0.3 K) while GEM with 0.6 K has the biggest (Table 4.7). The downscale approach was done on C60.

In Figure 4.5 the raw model MBE and the downscaled MBE are shown. It is observable that stations with a negative difference in height have warm biases and vice-versa (Figure

Table 4.7: Error comparison [K] *of 8228 stations on the downscaled weather forecasts: lapse rate 0.65 K / 100 m (Appendix A.0.2)*

	MAE	MBE	RMSE	SD
ERA5	1.4	0.3	1.9	1.8
NEMSGLOBAL	2.1	0.3	2.7	2.5
GFS05	2.2	0.3	2.9	2.7
MFGLOBAL	2.3	0.1	3.0	2.8
GEM	2.2	-0.4	2.9	2.7
ICON	1.9	0.0	2.5	2.4

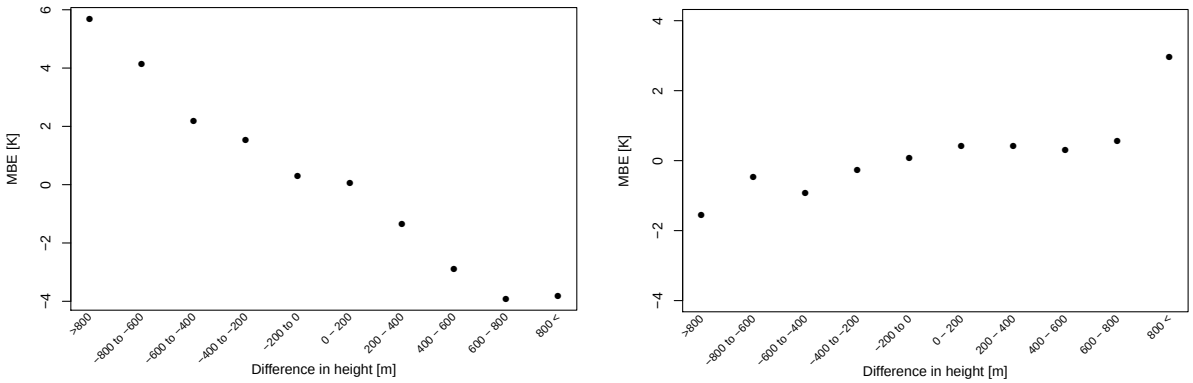

Figure 4.5: MBE shift: The MBE on NEMS in relation to the difference in height before (top) and after (bottom) downscaling the model temperatures: lapse rate 0.65 K / 100 m

4.5, *left*). Thereby the change in temperature per height difference, the lapse rate, is not included. Stations that are 600 to 800 m higher than the model grid point have an MBE of 4 K. The model grid point is lower than the station and thus the temperature forecasts are predicted too high. Working with a lapse rate these biases can be corrected. In Figure 4.5, *right*, the MBE is height corrected. Like this, stations with high positive and negative height differences have approximately the same MBE. The MBE of 4 K of the previous stations drops to -0.5 K. However, a minority of the stations have large height differences (Figure 3.4) and are effectively affected by this downscaling approach and thus, on all stations, the errors only decrease little (Table 4.7).

4.1.9 MAE depending on the horizontal distance

The horizontal distance would theoretically have been limited (Formula 3.2). At least two thirds of the stations per model fulfil these qualifications. Figure 4.6 shows the horizontal distance distribution. While ERA5 has a spatial resolution of 30 km the theoretical maximum would be about 22 km. The maximum distance to the next grid point was 51 km (Figure 3.3). For 16 % of all stations, the interface did not pick the next ERA5 grid point. For NEMS the maximum distance is 77 km and a percentage of 31 %, the distance is further than 22 km. The other models are comparable. 29 % of GFS05, 28 % of MF, 14 % of GEM and 9 % of ICON are not taking the next grid point (Figure B.1).

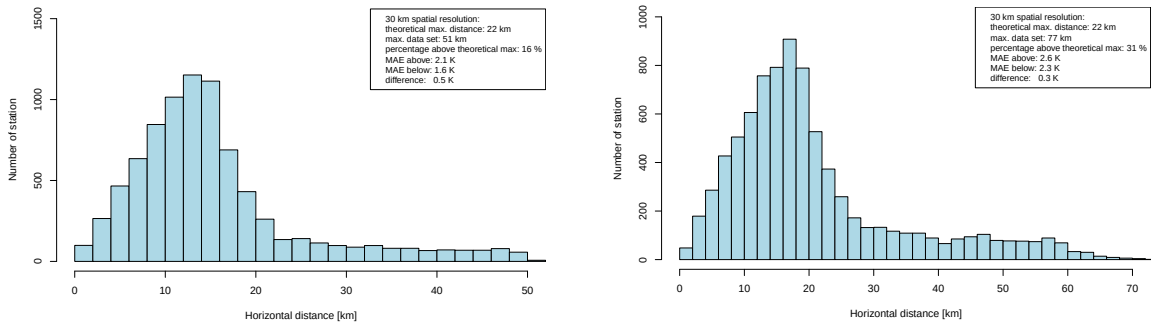

Figure 4.6: Horizontal distance distribution on ERA5 (left) and NEMS (right) including the theoretical maximum distance, the maximum distance in the data set, the percentage of the stations above the theoretical maximum, and the error depending on this benchmark (Appendix A.32, B.0.1)

Table A.32 shows the MAE depending on the access to the next model grid point. This is done for C60 and all stations. If the interface does not pick the next grid point the MAE of these stations is always about 0.2 to 0.6 K higher than of those which are lower than the benchmark. Considering this, the error increases with bigger horizontal distances. The absolute error declines from all stations to coverage 60. Depending on the decreases the difference varies. ERA5 and GEM have with 0.6 K the largest span, ICON with 0.2 K the lowest.

Only the stations with horizontal distances lower than the theoretical maximum, all the MAE decline for 0.1 K (Table 4.1, *right*). Interestingly, for the three models NEMS, GFS05 and MF with about 30 % of the station not going to the next grid point, the error between the two subsets is even smaller compared to ERA5 and GEM with lower percentages. Here, ICON having the lowest percentage values above the theoretical maximum distance and the highest spatial resolution performs best.

4.2 World Maps

4.2.1 Mean Absolute Error (MAE)

Figure 4.7, *left* is the gridded version of Figure 4.4. The trend for higher errors in the Rocky Mountains, India, China and Eastern Russia can be seen in the 2° gridded map. Figure 4.7 shows that on ERA5 low MAE of less than 1 K are mainly found in Northern Europe, Western Russia, Indonesia and Australia. Additionally, ERA5 performs well in the eastern part of North America, large areas of Africa and China. Errors higher than 3 K are found in Alaska, Greenland, the Rocky Mountains, India, the Himalayas, China, Mongolia and Eastern Russia. Comparable results delivers NEMS with a little shifted error range.

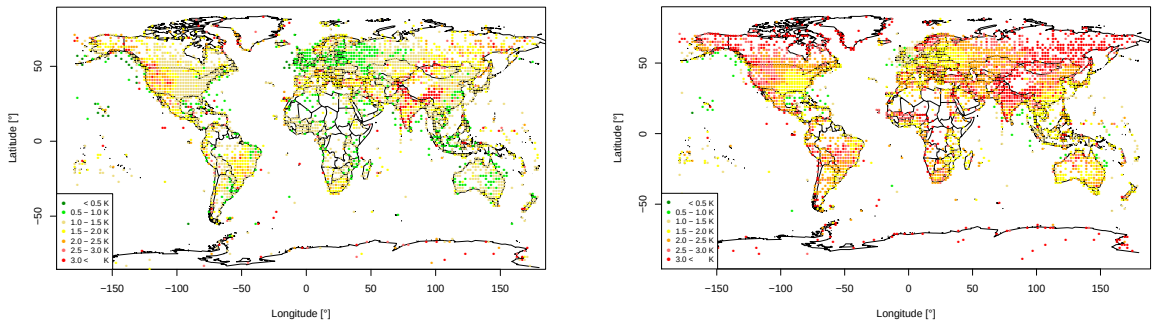

Figure 4.7: World maps of the 2° gridded MAE on ERA5 (left) and NEMS (right) (Appendix B.0.4)

Looking at GFS05, MF, GEM and ICON (Appendix B.0.4), Asia, Australia, Brazil and South Africa have increased MAE. Comparing these four models, ICON outperforms the others. With its higher resolution, it is clearly performing better in complex terrain such as the Rocky Mountains. Regarding this in North and South America, large differences in performance are found in complex terrain. Compared with the other raw models, NEMS has a better performance in Australia and the eastern coast of Asia.

Additionally to the complexity of terrain in different regions there is a trend of decreasing predictability with increasing distance from the sea. On simpler terrain, such as Europe and China, the predictability gradient is weaker than along North America's west coast which is strongly affected by the Rocky Mountains. In continental regions like Siberia that are cut off from the sea and lie next to mountainous regions such as the Himalayas it is hard to exactly predict the cold spells.

4.2.2 Mean Bias Error (MBE)

On ERA5 MBE the global patterns are less pronounced than on MAE. The MBE is quite well-balanced world-wide. ERA5 has cold biases in Alaska, North Canada, Greenland, Norway and Spitsbergen. Furthermore, an underestimation is found in the Northern Rocky Mountains, the Central Andes, the Himalayas, China, Central America, Indonesia, New Zealand and on islands such as Cape Verde and the Canary Islands. Regions with warm biases are the Southern Rocky Mountains, Brazil, India, Northern China, Mongolia, the Philippines, Southern Europe and the far eastern part of Russia.

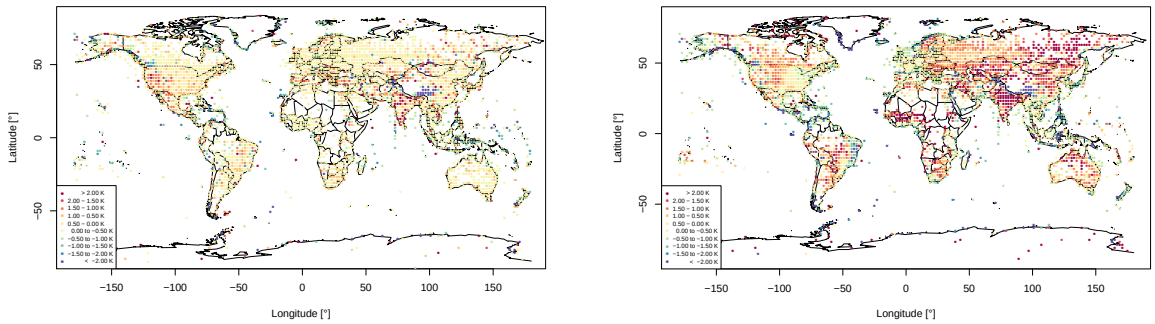

Figure 4.8: World maps of the 2° gridded MBE on ERA5 (left) and NEMS (right) (B.0.5)

When looking at the same error range on NEMS the patterns are clearer. Nearly the entire continental inland across Asia until Europe has, by trend, warm biases. In Alaska, Greenland, Central America, Brazil, Cape Horn, Indonesia, the Himalayas and New Zealand there is a temperature overestimation. Typically, the MBE at coastlines is globally cold biased. Going landward, the inland MBE is getting positive values, like in Brazil, Northern Europe, Africa and America.

GFS05's performance is comparable to NEMS with warm biases over the Rocky Mountains, Western Africa and Asia. On MF warm biases over the Rocky Mountains are found as well. Unlike NEMS and GFS05, MF is quite well-balanced in South America, Africa and Asia, with a tendency to slightly underestimate the temperatures.

Apart from Northern Russia, Central Europe, Western North and South America as well as some coastal regions GEM globally underestimates temperatures. ICON is well-balanced world-wide. Interestingly, New Zealand shows on ERA5 and NEMS predominately cold biases and on ICON, GFS05, MF and GEM warm biases. ICON underestimates the air temperatures for the Asian regions and the west coast of North and South America. In comparison to the other raw model, the global pattern on ICON is smoother and more clearly defined.

4.2.3 Minimum and Maximum MBE and MAE forecast

Figure 4.9 shows the best performing model for each cluster. If ERA5 is included (*left*) it outperforms almost all models globally. In some regions, such as Alaska, the west coast of North America, Greenland, Northern Europe and from the Mediterranean to the Caspian Sea, ICON has the smallest MAE. While NEMS, MF and GEM cannot show specific patterns here, GFS05 is dominant around Central Americas east coast, the Indian Ocean and the Western Pacific. In Alaska, Greenland, Northern Europe, the Mediterranean Sea and along the North American west coast as well as oceanic islands, ERA5 is overtaken by the five raw global forecast models.

Neglecting ERA5, the performances of the five raw model forecasts are shown in Figure 4.9 (*left*). Comparing five raw models, the spatial resolution is partly seen. ICON, with the highest resolution, dominates in Alaska, the Rocky Mountains, Newfoundland, Europe, Russia, Western Central Africa and South Africa. Besides ICON, NEMS has the smallest MAE over the United States, Central America down to Ecuador and Brazil, some parts of Africa, Kazakhstan, Northern China, India, Indonesia and Australia. While GEM still

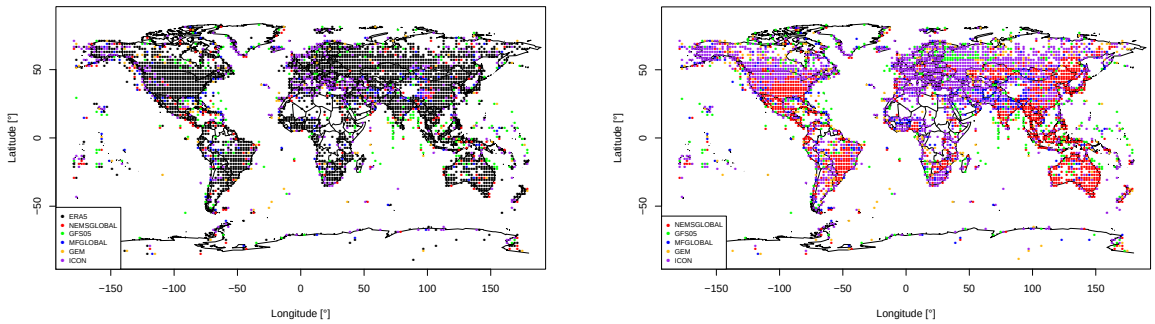

Figure 4.9: Smallest MAE performance regarding all six model (left) and without reanalysis model ERA5 (right) (Appendix B.0.6)

does not show a specific pattern on a global scale, GFS05 additionally performs best in parts of Western Russia. MF has a competitive performance in Central Asia.

In Figure 4.10 (*left*) the model with the highest MBE in a specific cluster is shown. While NEMS and GFS05 dominate the warm biases (Table 4.1) the other models are sparsely represented. ERA5 has the highest MBE over North Eastern America, Central America down to Ecuador, the coast of Brazil, the Alps, Norway, Japan and Indonesia. NEMS has patterns over Canada, South America and Eastern Europe to far Eastern Russia. GFS05 has the warmest biases in Northern Alaska, the Eastern foreland of the Rocky Mountains and the Andes, large areas over Africa, Southern Europe, South Western Asia, continental Australia and on oceanic islands. In Northern America within and next to the Rocky Mountains MF has the highest MBE.

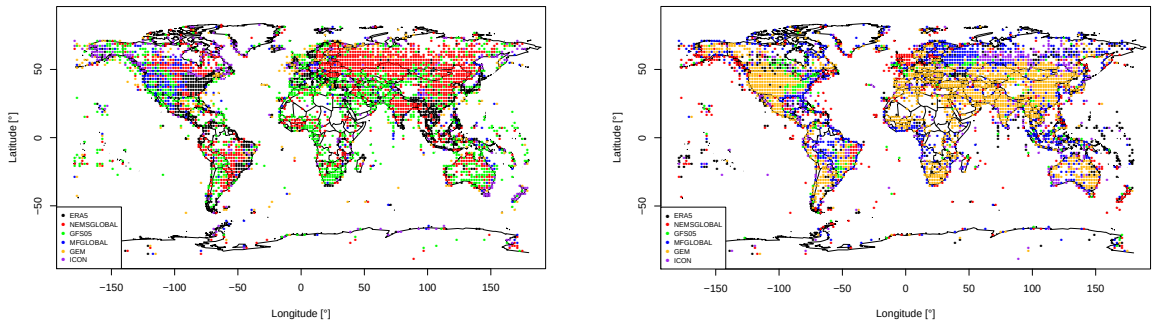

Figure 4.10: Maximum (left) and minimum (right) MBE distribution (B.0.7)

In Figure 4.10, *right*, the lowest MBE is shown. The GEM pattern spreads from Northern Alaska over North America down to Chile. Additionally, most of Southern Europe and Asia as well as large continental areas in Africa and Australia, GEM is represented with the lowest MBE. GEM has the largest MBE absolute value over all stations followed by MF and ICON (Table 4.1, *left*). Besides for along the eastern coast of America and the western coast of Africa, MF has cold biases in Northern Europe, Western Russia and Indonesia. NEMS has the lowest biases at the coast of Alaska and Canada, the North Sea and New Zealand.

4.3 Climate Zones

Table A.33 indicates that ERA5 has the best performance in all 5 climate zones regarding the MAE followed by ICON, NEMS, GFS05, MF and GEM. Neglecting ERA5, NEMS and ICON perform equally well in the equatorial climate (A). See the different climate zone classifications in Figure 4.11. In the arid zones (B) NEMS is slightly better than ICON. In the warm temperate (C), the cold (D) and the polar (E) zones, ICON performs best. While NEMS has an equally good performance as ICON in the tropics and the arid regions, the MAE is strongly increasing from the temperate, to the cold and the polar climate compared to ICON. At the same time, the MAE of NEMS in the polar regions is the highest over all models and climate zones. Besides ERA5, ICON is the only model to perform an MAE under 2 K in the temperate zone. However, all models have their best performance in the temperate zone.

Looking at the fine classification, the Csc and Dwd are empty because there were no stations corresponding to these zones. The high MAE on GFS05, MF and GEM is a result of the bad performances of these models in the polar frost zone (EF) from 4.6 to 4.7 K. The Dwb region is hard to predict. All six models get an MAE of 4 K or higher there. The MAE of MF is the highest overall when focusing the fine climate zones (Table A.33).

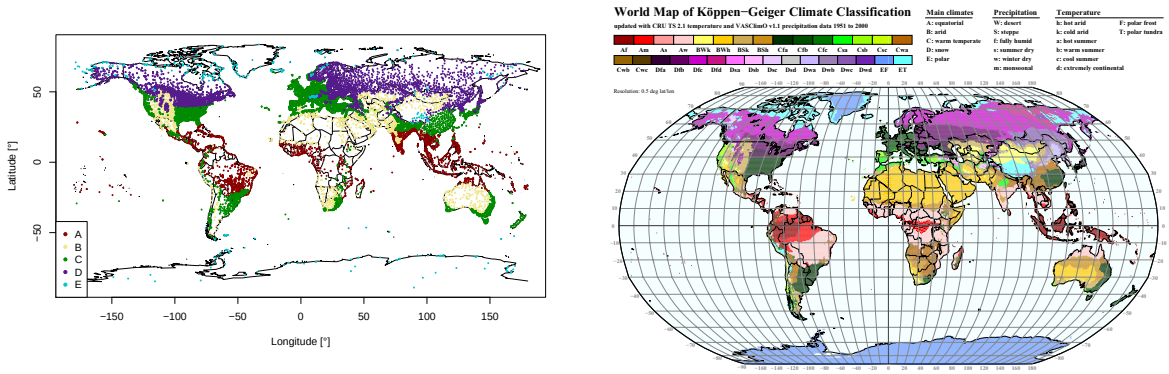

Figure 4.11: All 19150 stations coloured corresponding to the 5 coarse climate zones (left) and the fine classification (30) in the Koeppen-Geiger world map [9] (right)

Looking at the MBE, over the coarse classification ICON has the smallest spread and performs best, followed by ERA5, NEMS, GFS05, MF and GEM (Table A.34). Mostly all models underestimate the temperatures in the equator and polar regions. GFS05 and GEM have more problems predicting in the arid zone than the other four models. All six models underestimate the temperatures in the polar region (E). GEM underestimates all fine climate zones except Dfd. Thus, with that majority of cold biases no compensation on the MBE per climate zone or globally is possible. On MF all but the arid zone have cold biases.

4.4 Model Spread

The two different model spread approaches delivered comparable results. Figure 4.12 shows the second model spread approach of 'max-min' and 'standard deviation'. In Northern Europe and the most oceanic islands the difference between the maximum and the minimum forecast is below 2 K (Figure 4.12, *left*). Additional good model spreads are achieved in North Eastern America, Europe, Western Russia, Indonesia and coastal regions globally. Bigger model spreads appear in complex terrain such as the Rocky Mountains and the Himalayas and continental plains as Siberia, Australia, as well as parts of Alaska, Canada and Greenland.

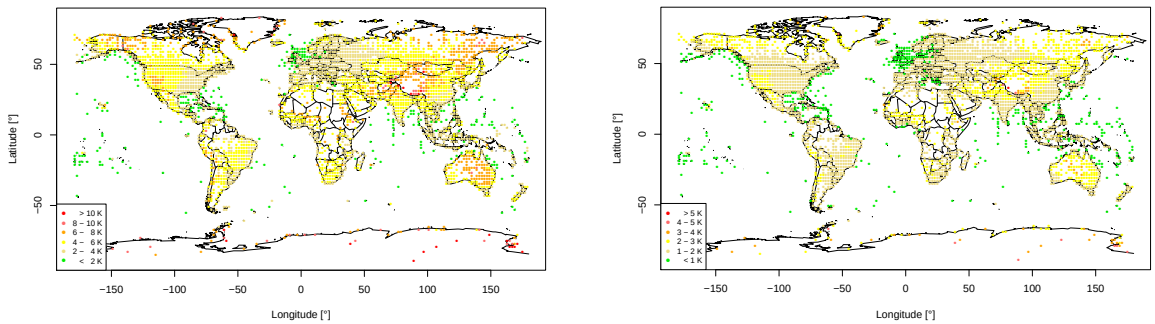

Figure 4.12: Second model spread approach: max-min (left), standard deviation (right) (Appendix B.0.8)

The standard deviation in Figure 4.12, *right*, is another indicator for the predictability. Comparable patterns can be observed. Northern Europe, most oceanic islands and coastal regions globally have a standard deviation of under 1 K. In North and South America, Africa, Europe, Western Russia and Indonesia the standard deviation is still less than 2 K. Similar to Figure 4.12, *left*, the predictability decreases within complex terrain and increasing distance from the sea and thus more continental climate.

5

Conclusions

Table 5.1 points out that

- The reanalysis model ERA5 outperforms all other model temperature forecasts on all different levels of data treatment.
- By only regarding the C60 stations, the error gets smaller because at stations with higher coverage the mean is more balanced than at stations with fewer data pairs.
- A post-processed temperature forecast at coarser resolution is better than a raw model forecast at finer resolution [19].
- Regarding raw model weather forecasts the error decreases with increasing spatial resolution [19].
- A higher resolution does not automatically lead to an increasing accuracy.
- The clustered data get decreased error compared to the raw data. By clustering the data, the weighting of the errors is corrected and this allows a global comparison. The different clusters have no significant impact on the error.

Table 5.1: MAE values for different levels of data treatment [K] *between raw model data, coverage 60, clustered in 5° and 2°, downscaled with lapse rate 0.65 K / 100 m and the daily mean*

	Raw	C60	CL 5	CL 2	DS 0.65	Daily Mean
ERA5	1.7	1.5	1.5	1.5	1.4	1.0
NEMSGLOBAL	2.4	2.2	2.3	2.3	2.1	1.6
GFS05	2.6	2.3	2.3	2.4	2.2	1.4
MFGLOBAL	2.7	2.3	2.4	2.4	2.3	1.6
GEM	2.8	2.4	2.4	2.5	2.2	1.6
ICON	2.3	2.0	2.0	2.0	1.9	1.1

Furthermore, a simple downscaling attempt can improve a model forecast. The reanalysis model ERA5 and NEMS have the smallest decline in MAE (0.3 K) while GEM with 0.6 K has the biggest. On the daily mean forecast all the errors of a 24-h window are smoothed and thus are significantly easier to predict.

Looking at the MBE (Table 5.2) the error progression is comparable to the one on MAE. On ERA5, NEMS, GFS05 and GEM the MBE is constant or decreases. Unlike the latter, MF and ICON have an error increase looking at the clustered calculated MBE. On either model the MBE drops from the 5° to the 2° gridded analysis.

Table 5.2: MBE comparison [K] *between raw model data, coverage 60, clustered in 5° and 2°, downscaled with lapse rate 0.65 K / 100 m and the daily mean*

	Raw	C60	CL 5	CL 2	DS 0.65	Daily Mean
ERA5	0.3	0.2	0.0	0.0	0.3	0.2
NEMSGLOBAL	0.4	0.1	0.0	0.0	0.3	0.1
GFS05	0.3	0.2	0.3	0.3	0.3	0.2
MFGLOBAL	-0.2	-0.1	-0.4	-0.3	0.1	-0.1
GEM	-0.8	-0.7	-0.8	-0.8	-0.4	-0.7
ICON	-0.1	-0.1	-0.2	-0.1	0.0	-0.1

On the meteoblue history+ advanced access the adjacent grid point has not been taken under specific conditions, such as mountainous and coastal regions [personal communication with Mathias Müller, 12.07.2019]. With a longer distance between the station and the model grid point, the error increases (Table A.32). Because weather stations are mostly limited to easy to reach locations and land surface, there is a mismatch in their distribution. While Europe and North America is well equipped with stations, regions like deserts and the high seas are scarcely covered [23]. Clustering the world in grid cells of 2° and 5° side length the error is weighted globally without considering the number of stations.

It is possible to globally define regions with different accuracy and variability in their temperature forecasts and even their biases.

- Temperature forecasts globally have the highest predictability on small oceanic islands and along ice-free coasts.

In these regions the air temperature is strongly regulated by the sea surface temperature [20]. On the mainland the smallest errors are found in Northern Europe from Northern France to the coast of Scandinavia. In Midwestern United States and along the east coast of North America, Western Russia and Indonesia the accuracy and predictability are high. Besides for the simple terrain, the temperature is mainly dependent on the sea surface temperature. The high accuracy and predictability over Europe and North America can be explained by the fact that the observed weather forecast models were developed in these regions [23].

- The predictability decreases in regions with complex topography and increasing distance from the sea.

Mountain ranges, such as the Rocky Mountains, the Andes, the Alps and the Himalayas are more difficult to forecast with spatial resolution limited raw model weather forecasts.

- Continental plains, such as Siberia, Australia, Alaska, Canada and Greenland, with few stations and a long distance to the sea are hard to predict.

In continental regions like Siberia that are cut off from the sea and can lie next to mountainous regions such as the Himalayas it is hard to exactly predict the temperatures of cold spells.

Compared to the other raw models, NEMS has a good performance in Australia and at the eastern coast of Asia. Otherwise GFS05 has a dominance and good performance around Central America's east coast, the Indian Ocean and the Western Pacific. As the model is developed by NOAA the model is possibly tuned to predict tropical thunderstorms. GEM globally underestimates the temperature but has a high performance in predicting the daily minimum.

Looking at the Koeppen-Geiger classification some climate zones are hard to predict such as the Dwb region; conversely Cfb is the zone with the highest predictability (Figure 4.11, *right*).

- All six models have their best performance in the temperate zone and underestimate the temperature in the polar and predominantly tropic regions.

Like in 2017, ICON performs best for a raw model, close to ERA5 [23]. In 2018, ERA5, NEMS and GFS05 tend for warm biases, while MF, GEM and ICON have cold biases. In 2017, MF was underestimating the temperature with cold biases [23]. GEM globally has very cold biases regarding the last two years. The MAE in Table 4.1 (*left*) are comparable with the ones from the previous year [23]. On GFS05, MF and GEM they are slightly higher [15]. The lowest standard deviation is found for ERA5, followed by ICON, NEMS, GFS05, GEM and MF.

6

Outlook

To solve the problem of large systematic errors, post-processing methods like mean bias removal or MOS are used [19]. MOS, reanalysis and mLM profit from error cancellations of the raw model data. On a 24-h forecast mLM performs 0.8 K better than raw 'stand-alone' models and 0.3 K better than MOS and the ERA5 reanalysis model [15, 16, 17]. Thus, besides the best possible forecast, it is mostly important to have a stable and reliable bias-removal technique to study spatial predictability patterns [20]. Multi-models such as mLM include a comparison of different weather forecast models and thus can better estimate and predict a forecast, depending on the present conditions [16, 17].

Verification results are always representative for the period over which the verification was done. Considering that, the results of this thesis, technically speaking, represent the year 2018. In the next few years a verification over several years will be possible. Thus, increasing its representativeness. In addition to an analysis over more years, a seasonal analysis would deliver information on whether the accuracy and variability depends on the season. Besides the temporal, local factors could be considered: distance to the sea (coastal regions), complex terrain (mountain ranges), large areas (continents) and small areas (only one country). Therefore, addressing how the different models would perform depending on the geographical region. For the local analyses other models could be considered. Next to the other model types, such as MOS and mLM, regional models could be compared, including different forecasting horizons. Temperature is only one parameter in atmospheric processes. A similar evaluations could be done for precipitation, dew point, wind speed and solar radiation.

7

Declaration on Scientific Integrity

University of Basel
Faculty of Science

Declaration on Scientific Integrity
(including a Declaration on Plagiarism and Fraud)

Bachelor's Thesis

Title of Thesis *(Please print in capital letters)*:

Accuracy and variability of six global temperature model forecasts in 2018

First Name, Surname: Fessler, Elias
(Please print in capital letters)

Matriculation No.: 16-062-085

With my signature I declare that this submission is my own work and that I have fully acknowledged the assistance received in completing this work and that it contains no material that has not been formally acknowledged.

I have mentioned all source materials used and have cited these in accordance with recognised scientific rules.

In addition to this declaration, I am submitting a separate agreement regarding the publication of or public access to this work.

☐ Yes ■ No

Place, Date: Basel, 30.07.2019

Signature: Elias Fessler

Please enclose a completed and signed copy of this declaration in your Bachelor's or Master's thesis .

April 2018

Bibliography

[1] H. E. Beck, N. Vergopolan, M. Pan, V. Levizzani, A. I. J. M. van Dijk, G. P. Weedon, L. Brocca, F. Pappenberger, G. J. Huffman, and E. F. Wood. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, 21(12):6201–6217, dec 2017. doi: 10.5194/hess-21-6201-2017.

[2] H. W. Bolchers. *pracma: Practical Numerical Math Functions*, Apr. 2019. URL https://CRAN.R-project.org/package=pracma. Visited on 02.08.2019.

[3] C. Bryant et al. *kgc: Koeppen-Geiger Climatic Zones*, Dec. 2017. URL https://CRAN.R-project.org/package=kgc. Visited on 02.08.2019.

[4] CMC. Gem model general documentation. URL http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/DOCUMENTATION/GENERAL/general.html. Visited on 18.07.2019.

[5] CNRM. Weather forcasting model arpege. URL http://www.umr-cnrm.fr/spip.php?article121&lang=en. Visited on 26.07.2019.

[6] DWD. Icon database reference manual. URL https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/icon/icon_dbbeschr_aktuell.html?nn=346840. Visited on 18.07.2019.

[7] ECMWF. Era5 documentation. URL https://confluence.ecmwf.int//display/CKB/ERA5+data+documentation. Visited on 18.07.2019.

[8] M. Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. URL https://CRAN.R-project.org/package=stargazer. R package version 5.2.2.

[9] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel. World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263, jul 2006. doi: 10.1127/0941-2948/2006/0130.

[10] P. Lynch. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444, mar 2008. doi: 10.1016/j.jcp.2007.02.034.

[11] C. Marzban, S. Sandgathe, and E. Kalnay. MOS, perfect prog, and reanalysis. *Monthly Weather Review*, 134(2):657–663, feb 2006. doi: 10.1175/mwr3088.1.

[12] meteoblue. Reanalysis datasets, . URL https://content.meteoblue.com/ru/specifications/data-sources/weather-simulation-data/reanalysis-datasets. Visited on 18.07.2019.

[13] meteoblue. Model output statistics (mos), . URL https://content.meteoblue.com/nl/specifications/data-sources/post-processing/statistics-mos. Visited on 18.07.2019.

[14] meteoblue. meteoblue models, . URL https://content.meteoblue.com/en/specifications/data-sources/weather-simulation-data/meteoblue-models. Visited on 18.07.2019.

[15] meteoblue. Verification of simulations, . URL https://content.meteoblue.com/nl/research-development/processes/verification. Visited on 18.07.2019.

[16] meteoblue. meteoblue learning multimodel (mlm), . URL https://content.meteoblue.com/nl/specifications/data-sources/post-processing/meteoblue-learning-multimodel-mlm. Visited on 18.07.2019.

[17] meteoblue. mml leaflet, . URL https://content.meteoblue.com/de/media/verified-quality/verification/mlm-leaflet. Visited on 18.07.2019.

[18] meteoblue. Weather model theory, 2019. URL https://content.meteoblue.com/nl/specifications/weather-model-theory. Visited on 24.07.2019.

[19] M. D. Müller. Effects of model resolution and statistical postprocessing on shelter temperature and wind forecasts. *Journal of Applied Meteorology and Climatology*, 50 (8):1627–1636, aug 2011. doi: 10.1175/2011jamc2615.1.

[20] M. D. Müller and Z. Janjic. Verification of the new nonhydrostatic multiscale model on the b grid (NMMB): A view on global predictability of surface parameters. *Weather and Forecasting*, 30(3):827–840, jun 2015. doi: 10.1175/waf-d-14-00049.1.

[21] NOAA. Global forecast system (gfs). URL https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs. Visited on 18.07.2019.

[22] T. Palmer. Predictability of weather and climate: from theory to practice - from days to decades. In *Seminar on Predictability of weather and climate, 9-13 September 2002*, pages 1–14, Shinfield Park, Reading, 2003. ECMWF, ECMWF. URL https://www.ecmwf.int/node/11490.

[23] S. Schlögl. Verifications report 2017. Internal global meteoblue study, Oct. 2018.

[24] R. C. Team. *The R Stats Package*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html. Visited on 02.08.2019.

[25] J. Ulrich. *TTR: Technical Trading Rules*, Sept. 2018. URL https://CRAN.R-project.org/package=TTR. Visited on 02.08.2019.

Appendix A

Tables

A.0.1 Error comparison with different coverages

Table A.1: All stations

[K]	MAE	MBE	RMSE	SD
ERA5	1.7	0.3	2.2	1.9
NEMSGLOBAL	2.4	0.4	3.1	2.6
GFS05	2.6	0.3	3.4	3.0
MFGLOBAL	2.7	-0.2	3.4	3.1
GEM	2.8	-0.8	3.5	3.1
ICON	2.3	-0.1	3.0	2.7

Table A.2: Coverage 30

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	0.2	2.0	1.8
NEMSGLOBAL	2.3	0.2	2.9	2.6
GFS05	2.4	0.3	3.1	2.8
MFGLOBAL	2.4	-0.1	3.2	2.9
GEM	2.5	-0.7	3.2	2.9
ICON	2.1	-0.1	2.7	2.5

Table A.3: Coverage 60

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	0.2	1.9	1.8
NEMSGLOBAL	2.2	0.1	2.8	2.5
GFS05	2.3	0.2	2.9	2.7
MFGLOBAL	2.3	-0.1	3.0	2.8
GEM	2.4	-0.7	3.0	2.7
ICON	2.0	-0.1	2.5	2.4

Table A.4: Coverage 90

[K]	MAE	MBE	RMSE	SD
ERA5	1.4	0.2	1.9	1.7
NEMSGLOBAL	2.1	0.0	2.8	2.5
GFS05	2.2	0.2	2.8	2.6
MFGLOBAL	2.3	-0.1	3.0	2.7
GEM	2.3	-0.7	3.0	2.7
ICON	1.9	-0.1	2.4	2.3

A.0.2 Error comparison with different downscale lapse rates

Table A.5: Lapse rate: 0.80 K / 100 m

[K]	MAE	MBE	RMSE	SD
ERA5	1.4	0.4	1.9	1.8
NEMSGLOBAL	2.1	0.3	2.7	2.5
GFS05	2.2	0.3	2.9	2.7
MFGLOBAL	2.3	0.1	3.0	2.8
GEM	2.2	-0.4	2.9	2.7
ICON	1.9	0.0	2.5	2.4

Table A.6: Lapse rate: 0.65 K / 100 m

[K]	MAE	MBE	RMSE	SD
ERA5	1.4	0.3	1.9	1.8
NEMSGLOBAL	2.1	0.3	2.7	2.5
GFS05	2.2	0.3	2.9	2.7
MFGLOBAL	2.3	0.1	3.0	2.8
GEM	2.2	-0.4	2.9	2.7
ICON	1.9	0.0	2.5	2.4

Table A.7: Lapse rate: 0.55 K / 100 m

[K]	MAE	MBE	RMSE	SD
ERA5	1.4	0.3	1.9	1.8
NEMSGLOBAL	2.1	0.2	2.7	2.5
GFS05	2.2	0.3	2.9	2.7
MFGLOBAL	2.3	0.0	3.0	2.8
GEM	2.2	-0.5	2.9	2.7
ICON	1.9	0.0	2.5	2.4

A.0.3 Error comparison on the 5° clustered world

Table A.8: 5° clustered Mean

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	-0.1	2.0	1.8
NEMSGLOBAL	2.3	-0.0	3.0	2.6
GFS05	2.3	0.3	3.0	2.7
MFGLOBAL	2.4	-0.4	3.2	2.9
GEM	2.4	-0.8	3.1	2.7
ICON	2.0	-0.2	2.6	2.4

Table A.9: 5° clustered Median

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	-0.1	1.9	1.7
NEMSGLOBAL	2.3	-0.0	2.9	2.6
GFS05	2.3	0.3	2.9	2.6
MFGLOBAL	2.3	-0.4	3.1	2.8
GEM	2.4	-0.8	3.0	2.6
ICON	1.9	-0.2	2.5	2.4

Table A.10: 5° clustered Maximum

[K]	MAE	MBE	RMSE	SD
ERA5	2.2	0.7	2.8	2.4
NEMSGLOBAL	3.0	0.9	3.8	3.2
GFS05	3.0	1.2	3.8	3.3
MFGLOBAL	3.1	0.5	3.9	3.5
GEM	3.2	0.1	4.0	3.3
ICON	2.6	0.5	3.3	3.0

Table A.11: 5° clustered Minimum

[K]	MAE	MBE	RMSE	SD
ERA5	1.2	-0.8	1.6	1.5
NEMSGLOBAL	2.0	-0.9	2.6	2.3
GFS05	1.9	-0.6	2.5	2.3
MFGLOBAL	2.0	-1.3	2.7	2.5
GEM	1.9	-1.7	2.5	2.3
ICON	1.7	-0.8	2.2	2.0

A.0.4 Error comparison on the 2° clustered world

Table A.12: 2° clustered Mean

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	0.1	2.0	1.8
NEMSGLOBAL	2.3	0.0	2.9	2.6
GFS05	2.4	0.3	3.0	2.8
MFGLOBAL	2.4	-0.3	3.2	2.9
GEM	2.5	-0.8	3.1	2.8
ICON	2.0	-0.1	2.6	2.5

Table A.13: 2° clustered Median

[K]	MAE	MBE	RMSE	SD
ERA5	1.5	0.0	2.0	1.8
NEMSGLOBAL	2.2	0.0	2.9	2.6
GFS05	2.3	0.3	3.0	2.7
MFGLOBAL	2.4	-0.3	3.1	2.9
GEM	2.4	-0.8	3.1	2.8
ICON	2.0	-0.1	2.6	2.4

Table A.14: 2° clustered Maximum

[K]	MAE	MBE	RMSE	SD
ERA5	1.8	0.5	2.4	2.1
NEMSGLOBAL	2.6	0.6	3.3	2.9
GFS05	2.7	0.8	3.5	3.1
MFGLOBAL	2.7	0.2	3.6	3.2
GEM	2.8	-0.3	3.6	3.1
ICON	2.3	0.3	3.0	2.8

Table A.15: 2° clustered Minimum

[K]	MAE	MBE	RMSE	SD
ERA5	1.3	-0.4	1.7	1.6
NEMSGLOBAL	2.1	-0.5	2.7	2.4
GFS05	2.1	-0.2	2.7	2.5
MFGLOBAL	2.2	-0.7	2.9	2.7
GEM	2.2	-1.3	2.8	2.5
ICON	1.8	-0.5	2.3	2.2

A.0.5 Percentage Tab: MAE

Table A.16: Percentage Tab: MAE all stations

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	17.5	42.4	22.2	9.0	3.7	5.1
NEMSGLOBAL	1.5	8.7	34.3	28.0	13.1	14.2
GFS05	2.2	8.4	21.5	27.2	16.2	24.3
MFGLOBAL	2.0	7.3	21.5	27.6	17.4	24.0
GEM	2.2	8.2	19.5	26.2	14.9	28.8
ICON	3.8	18.1	27.7	22.3	11.0	16.9

Table A.17: Percentage Tab: MAE Coverage 30

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	18.5	45.1	21.2	8.2	3.3	3.5
NEMSGLOBAL	0.9	7.9	36.9	29.4	12.4	12.3
GFS05	1.7	9.9	24.3	29.9	15.2	18.8
MFGLOBAL	1.5	7.8	23.7	30.4	18.1	18.3
GEM	1.9	9.8	22.1	29.4	14.7	22.0
ICON	4.0	20.6	31.6	23.2	9.6	10.7

Table A.18: Percentage Tab: MAE Coverage 60

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	19.1	46.8	20.4	7.7	2.9	3.1
NEMSGLOBAL	1.2	9.2	40.4	28.9	10.9	9.4
GFS05	2.2	11.9	25.5	31.9	13.8	14.7
MFGLOBAL	1.9	9.5	24.1	32.6	17.4	14.5
GEM	2.5	11.8	22.3	32.1	13.8	17.5
ICON	5.3	22.2	34.2	22.7	7.9	7.7

Table A.19: Percentage Tab: MAE Coverage 90

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	20.1	47.5	19.6	7.2	2.5	2.9
NEMSGLOBAL	1.0	9.5	41.2	28.7	10.5	8.9
GFS05	2.0	13.0	27.2	33.0	12.8	11.9
MFGLOBAL	1.6	10.5	25.6	33.7	17.0	11.4
GEM	2.2	12.6	23.0	33.7	13.2	15.1
ICON	5.7	23.8	35.6	22.3	7.3	5.1

A.0.6 Percentage Tab: MBE

Table A.20: Percentage Tab: MBE all stations

	<-2 K	-2 to -1 K	-1 to 0 K	0 - 1 K	1 - 2 K	2 K <
ERA5	3.0	5.2	29.3	47.9	11.0	3.7
NEMSGLOBAL	5.2	9.2	24.9	32.5	18.9	9.3
GFS05	5.3	10.2	26.0	33.4	16.1	9.1
MFGLOBAL	8.6	14.0	33.0	30.7	9.4	4.4
GEM	18.1	19.4	34.7	23.0	3.2	1.6
ICON	5.3	12.4	36.7	36.3	6.9	2.4

Table A.21: Percentage Tab: MBE Coverage 30

	<-2 K	-2 to -1 K	-1 to 0 K	0 - 1 K	1 - 2 K	2 K <
ERA5	2.4	4.8	29.8	50.4	10.0	2.5
NEMSGLOBAL	4.6	9.5	26.5	34.6	18.6	6.0
GFS05	3.4	8.9	27.9	36.9	15.5	7.3
MFGLOBAL	6.2	12.5	35.0	33.9	9.1	3.2
GEM	15.2	18.8	36.6	25.6	2.7	1.1
ICON	2.8	10.6	39.5	40.0	5.8	1.4

Table A.22: Percentage Tab: MBE Coverage 60

	<-2 K	-2 to -1 K	-1 to 0 K	0 - 1 K	1 - 2 K	2 K <
ERA5	2.0	4.5	29.6	51.6	10.2	2.2
NEMSGLOBAL	4.5	10.2	29.7	35.9	15.7	4.0
GFS05	3.0	8.9	30.0	37.9	14.7	5.7
MFGLOBAL	5.0	11.3	35.6	35.8	9.2	3.1
GEM	12.7	18.2	40.0	26.1	2.0	1.0
ICON	2.0	8.6	41.4	41.9	5.0	1.1

Table A.23: Percentage Tab: MBE Coverage 90

	<-2 K	-2 to -1 K	-1 to 0 K	0 - 1 K	1 - 2 K	2 K <
ERA5	1.8	4.4	30.5	51.9	9.5	1.9
NEMSGLOBAL	4.7	10.4	31.3	36.0	14.4	3.2
GFS05	2.9	9.0	31.4	38.0	13.3	5.4
MFGLOBAL	4.9	11.0	36.8	35.8	8.7	2.7
GEM	12.9	18.5	40.8	25.2	1.8	0.7
ICON	1.8	8.5	43.0	41.5	4.3	0.9

A.0.7 Percentage Tab: RMSE

Table A.24: Percentage Tab: RMSE all stations

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	3.2	25.1	32.5	18.8	9.0	11.2
NEMSGLOBAL	0.5	1.9	9.8	27.5	24.5	35.7
GFS05	1.2	2.1	8.3	15.7	22.4	50.2
MFGLOBAL	1.0	1.5	6.9	15.0	22.9	52.5
GEM	0.9	2.4	7.8	14.5	20.6	53.8
ICON	1.2	5.9	15.8	20.3	21.6	35

Table A.25: Percentage Tab: RMSE Coverage 30

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	2.7	27.5	33.9	17.9	8.6	9.3
NEMSGLOBAL	0.2	1.3	9.1	29.4	25.9	33.9
GFS05	0.6	2.4	9.6	17.6	25.3	44.3
MFGLOBAL	0.5	1.5	7.8	16.4	25	48.7
GEM	0.6	2.5	9.4	16.6	23.2	47.6
ICON	1.0	6.8	18.3	22.9	23.6	27.4

Table A.26: Percentage Tab: RMSE Coverage 60

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	3.4	28.3	34.8	17.1	8.0	8.3
NEMSGLOBAL	0.2	1.6	10.3	32.3	26.6	28.8
GFS05	0.8	3.3	11.1	18.0	28.5	38.2
MFGLOBAL	0.5	2.1	9.5	16.6	26.9	44.2
GEM	0.7	3.2	11.1	16.4	26.0	42.5
ICON	1.2	9.0	18.9	24.8	24.3	21.6

Table A.27: Percentage Tab: RMSE Coverage 90

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	3.4	30.1	34.5	16.6	7.5	7.8
NEMSGLOBAL	0.2	1.5	10.7	33.1	26.3	28.0
GFS05	0.6	3.5	12.2	19	30.5	34.1
MFGLOBAL	0.4	2.0	10.6	17.5	27.9	41.4
GEM	0.6	3.4	11.9	17.1	27.1	39.9
ICON	1.1	9.9	19.9	26.1	24.5	18.3

A.0.8 Percentage Tab: SD

Table A.28: Percentage Tab: SD all stations

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	4.7	32.2	33.7	15.8	7.0	6.3
NEMSGLOBAL	1.1	2.7	16.1	34.4	23.4	22.2
GFS05	1.5	3.0	11.6	20.0	25.5	38.2
MFGLOBAL	1.1	2.1	9.7	19.6	25.3	41.9
GEM	1.1	3.2	10.7	19.1	24.3	41.4
ICON	1.8	7.4	18.7	22.9	20.9	28.1

Table A.29: Percentage Tab: SD Coverage 30

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	3.8	35.1	33.8	15.1	6.8	5.2
NEMSGLOBAL	0.5	2.0	14.5	35.8	24.4	22.6
GFS05	1.0	3.2	12.6	22.1	27.6	33.5
MFGLOBAL	0.5	2.1	10.4	20.7	26.9	39.2
GEM	0.8	3.3	12.8	21.1	26.3	35.5
ICON	1.3	8.7	20.9	25.0	22.0	22.0

Table A.30: Percentage Tab: SD Coverage 60

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	4.6	36.5	33.9	13.8	6.5	4.6
NEMSGLOBAL	0.6	2.3	15.4	38.5	24.2	18.8
GFS05	1.2	4.2	13.7	22.5	29.8	28.3
MFGLOBAL	0.6	2.7	12.1	21.1	28.1	35.4
GEM	1.0	4.2	14.6	20.3	28.4	31.3
ICON	1.6	11.2	20.8	26.8	22.1	17.3

Table A.31: Percentage Tab: SD Coverage 90

	<1 K	1 - 1.5 K	1.5 - 2 K	2 - 2.5 K	2.5 - 3 K	3 K <
ERA5	4.6	38.7	32.7	13.3	6.1	4.5
NEMSGLOBAL	0.4	2.3	15.9	38.8	23.6	18.9
GFS05	1.0	4.5	14.9	24.0	31.1	24.3
MFGLOBAL	0.4	2.8	13.3	22.0	28.3	33.1
GEM	0.8	4.5	15.4	21.3	29.5	28.4
ICON	1.4	12.3	21.9	28.2	22.0	14.1

A.0.9 MAE in relation to maximum horizontal distance

Table A.32: MAE [K] *of the six global models with C60 (left) and all stations (right) in relation to the maximum horizontal distance (3.2)*

	Coverage 60	all stations
ERA5 distance higher	2.0	2.1
ERA5 distance lower	1.4	1.6
Difference	0.6	0.5
NEMSGLOBAL distance higher	2.4	2.6
NEMSGLOBAL distance lower	2.1	2.3
Difference	0.3	0.3
GFS05 distance higher	2.5	2.9
GFS05 distance lower	2.2	2.5
Difference	0.3	0.4
MFGLOBAL distance higher	2.6	2.9
MFGLOBAL distance lower	2.2	2.6
Difference	0.4	0.3
GEM distance higher	2.9	3.2
GEM distance lower	2.3	2.7
Difference	0.6	0.5
ICON distance higher	2.1	2.6
ICON distance lower	1.9	2.3
Difference	0.2	0.3

A.0.10 Climate zones

Table A.33: MAE [K] *of the six global models depending on the coarse (A-E) and fine (Af-ET) climate zone classification after Koeppen-Geiger*

	ERA5	NEMSGLOBAL	GFS05	MFGLOBAL	GEM	ICON
A	1.5	2.1	2.4	2.4	2.3	2.1
B	1.6	2.3	2.8	2.7	3.2	2.4
C	1.4	2.0	2.2	2.2	2.2	1.8
D	1.5	2.3	2.3	2.4	2.4	2.0
E	2.3	3.4	2.7	3.1	3.2	2.5
Af	1.3	1.9	2.0	2.1	2.0	1.7
Am	1.3	1.8	1.9	2.1	1.9	1.7
As	1.6	2.1	2.7	2.4	2.4	2.0
Aw	1.6	2.2	2.6	2.6	2.5	2.3
BSh	1.4	2.2	3.1	2.9	3.2	2.7
BSk	1.8	2.4	2.9	2.8	3.1	2.5
BWh	1.4	2.3	2.6	2.5	3.2	2.2
BWk	1.6	2.4	2.7	2.4	3.6	2.2
Cfa	1.3	1.9	2.3	2.4	2.3	2.0
Cfb	1.2	1.9	1.9	1.9	1.9	1.5
Cfc	1.6	2.1	1.8	1.9	1.8	1.3
Csa	1.7	2.2	2.3	2.3	2.5	1.8
Csb	1.8	2.4	2.7	2.6	2.7	2.1
Csc						
Cwa	1.9	2.6	2.9	2.7	2.9	2.5
Cwb	2.5	3.7	3.1	3.4	3.6	2.8
Cwc	1.3	1.8	1.9	2.0	3.9	1.7
Dfa	1.2	1.9	2.1	2.2	2.4	1.9
Dfb	1.4	2.2	2.2	2.3	2.2	1.9
Dfc	1.8	2.7	2.5	2.6	2.5	2.1
Dfd	1.4	3.4	3.4	3.4	3.2	3.1
Dsa	1.7	2.5	2.7	2.2	2.7	2.0
Dsb	2.4	2.7	2.4	2.6	3.9	2.2
Dsc	2.3	3.6	3.5	3.8	3.5	2.6
Dwa	1.4	2.3	3.0	2.9	3.5	2.8
Dwb	4.0	4.3	5.0	5.4	4.9	4.7
Dwc	2.0	3.1	3.4	3.0	3.2	2.5
Dwd						
EF	3.5	4.7	3.7	4.7	4.6	3.0
ET	2.2	3.3	2.7	3.0	3.2	2.5

Table A.34: MBE [K] *of the six global models depending on the coarse (A-E) and fine (Af-ET) climate zone classification after Koeppen-Geiger*

	ERA5	NEMSGLOBAL	GFS05	MFGLOBAL	GEM	ICON
A	0.0	0.0	-0.1	-0.6	-0.7	-0.3
B	0.4	0.4	1.2	0.5	-1.5	-0.2
C	0.3	0.1	0.2	-0.1	-0.6	-0.1
D	0.1	0.2	0.0	-0.1	-0.6	0.0
E	-0.7	-0.9	-0.4	-1.4	-1.7	-0.2
Af	-0.2	-0.7	-0.4	-1.1	-0.9	-0.4
Am	-0.1	-0.4	-0.2	-1.1	-0.7	-0.5
As	-0.1	-0.5	-0.5	-0.4	-0.6	-0.3
Aw	0.2	0.3	0.1	-0.3	-0.6	-0.2
BSh	0.3	0.8	1.3	0.2	-1.3	-0.3
BSk	0.6	0.6	1.3	0.9	-1.2	0.0
BWh	0.0	-0.1	1.0	-0.1	-2.0	-0.5
BWk	0.4	0.3	1.2	0.3	-2.5	-0.7
Cfa	0.4	0.1	0.0	-0.3	-0.5	-0.2
Cfb	0.2	0.1	0.0	0.0	-0.4	0.1
Cfc	-0.5	-1.2	-0.5	-0.6	-0.7	-0.1
Csa	0.3	-0.1	0.6	0.0	-1.3	-0.1
Csb	0.4	0.2	0.9	0.5	-1.0	0.0
Csc						
Cwa	0.4	0.7	1.0	0.1	-1.1	-0.2
Cwb	1.4	2.2	0.9	1.4	-0.2	0.3
Cwc	0.3	0.6	-0.5	-0.2	-3.6	0.1
Dfa	0.3	0.1	-0.2	0.1	-0.7	0.0
Dfb	0.1	0.3	0.0	0.1	-0.4	0.0
Dfc	-0.1	0.1	0.3	-0.2	-0.6	0.1
Dfd	0.2	2.3	0.5	-0.1	0.3	0.2
Dsa	0.4	0.9	1.8	0.4	-1.5	-0.2
Dsb	0.0	-0.4	0.3	-0.1	-2.6	0.0
Dsc	-0.6	-1.0	-0.3	-1.8	-1.8	0.0
Dwa	0.2	0.3	0.8	-0.8	-2.0	-1.0
Dwb	-1.2	-0.3	-1.3	-1.8	-1.7	-1.0
Dwc	-0.7	0.3	0.6	-0.6	-1.9	-0.1
Dwd						
EF	-1.4	-1.3	-0.1	-2.4	-2.2	0.0
ET	-0.6	-0.8	-0.4	-1.4	-1.6	-0.3

Appendix B

Figures

B.0.1 Horizontal distance distribution

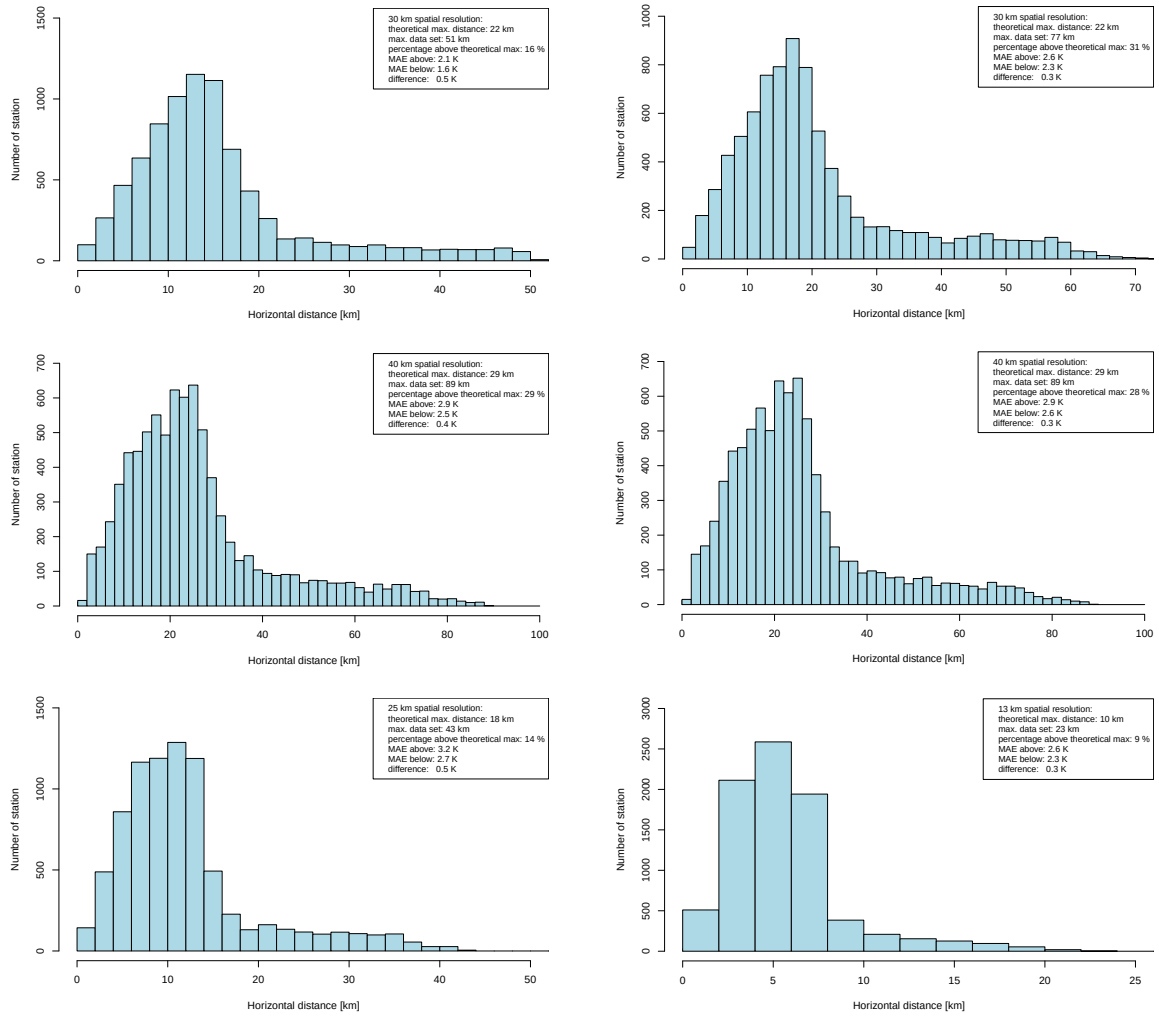

Figure B.1: Horizontal distance distribution of ERA5 (top left), NEMS (top right), GFS05 (mid left), MF (mid right), GEM (bottom left) and ICON (bottom right) including the theoretical maximum distance, the maximum distance in the data set, the percentage of the stations above the theoretical maximum, and the MAE depending on the benchmark (Table A.32)

B.0.2 Height difference distribution

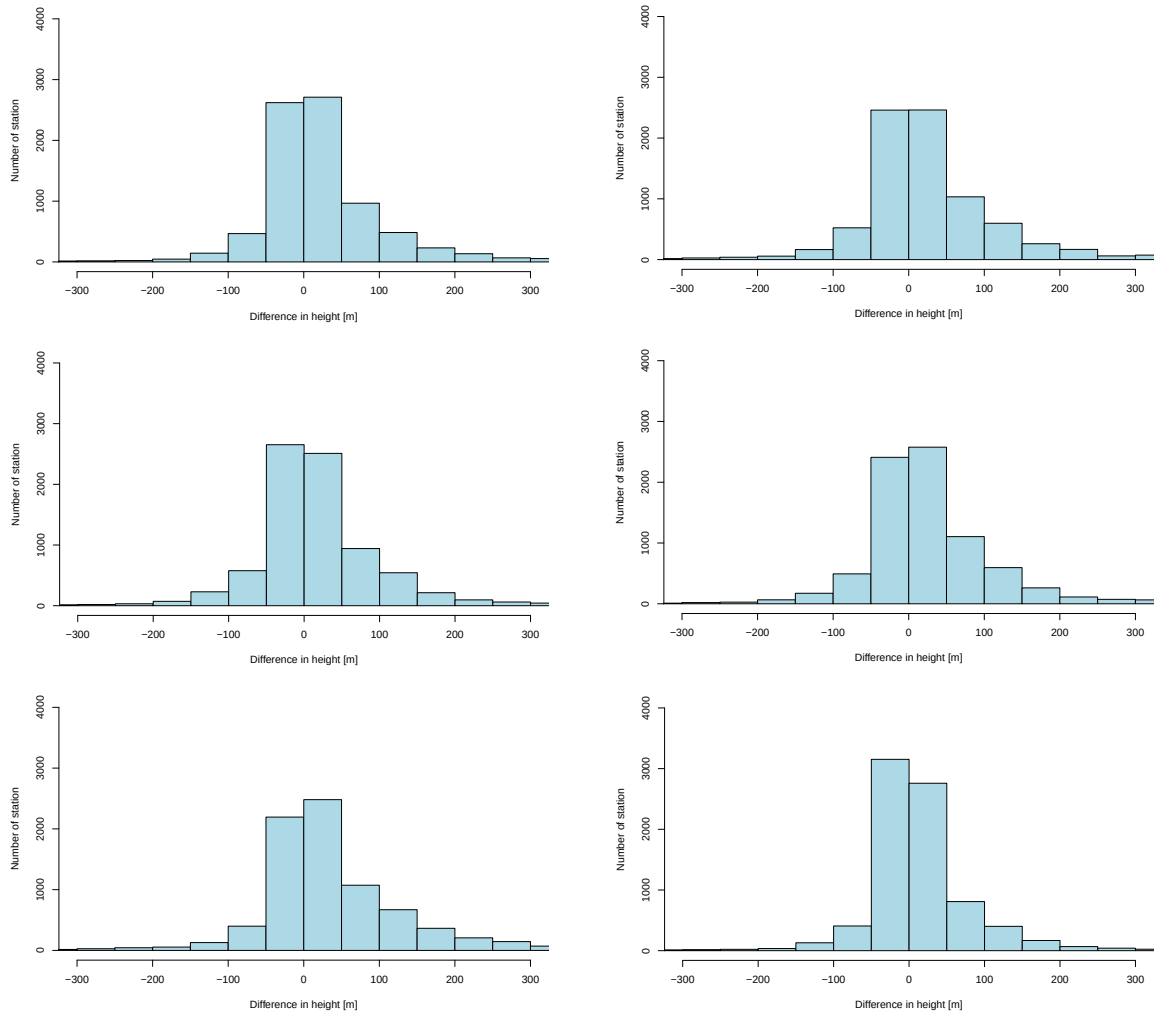

Figure B.2: Height difference distribution of ERA5 (top left), NEMS (top right), GFS05 (mid left), MF (mid right), GEM (bottom left) and ICON (bottom right)

B.0.3 World maps: stations with higher horizontal distance than theoretical maximum

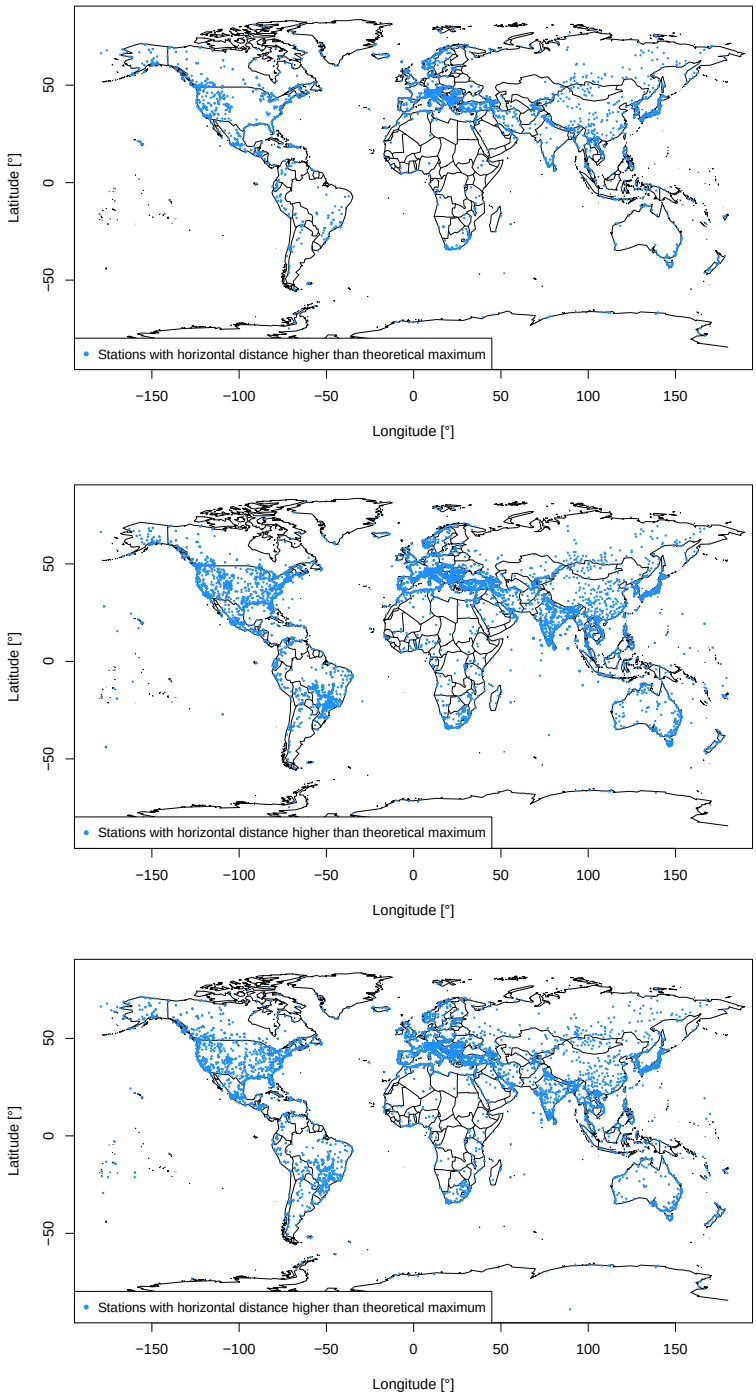

Figure B.3: Stations with higher distance than theoretical height of ERA5 (top), NEMS (middle) and GFS05 (bottom)

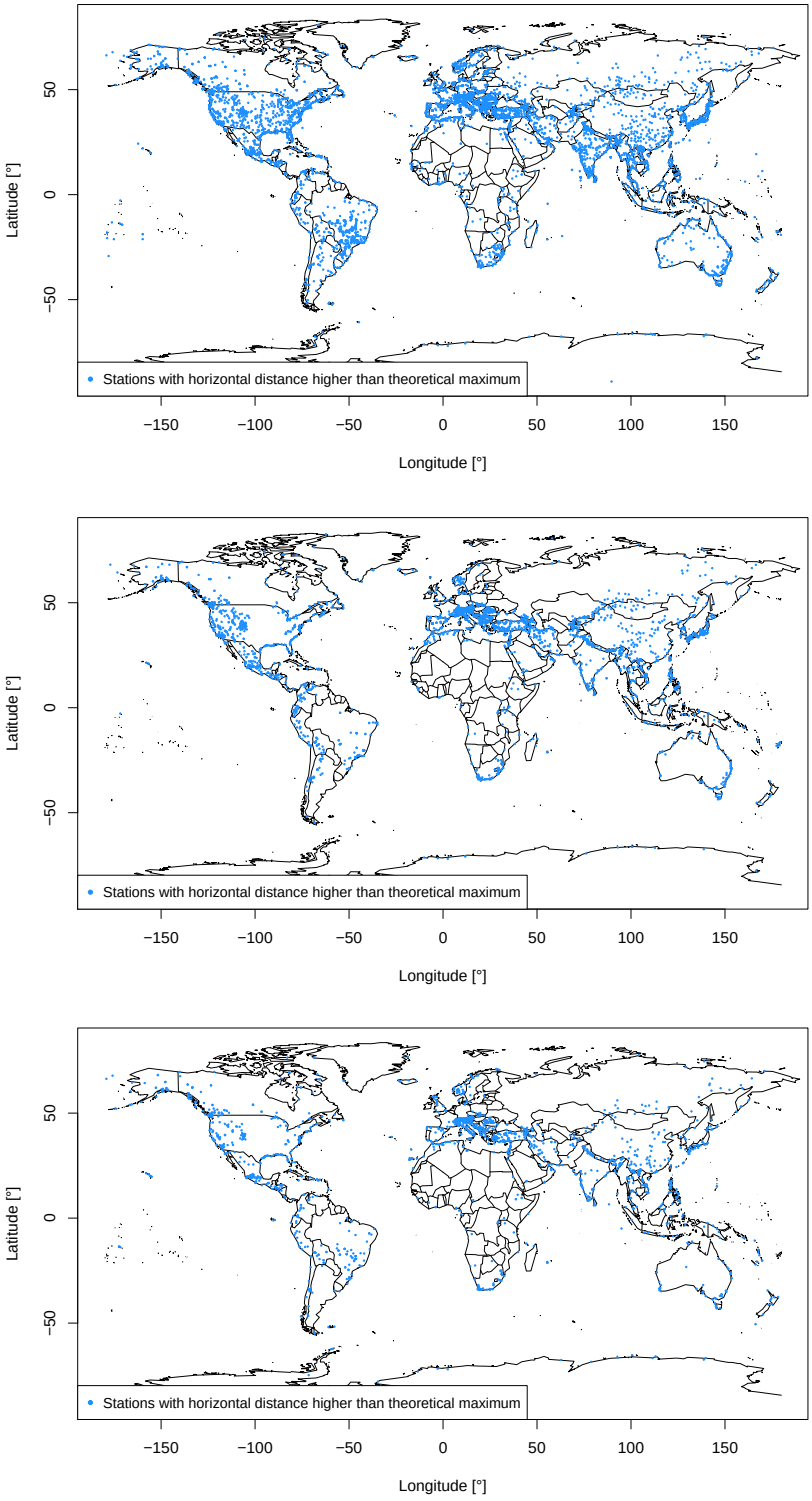

Figure B.4: Stations with higher distance than theoretical height of MF (top), GEM (middle) and ICON (bottom)

B.0.4 World maps: 2° gridded MAE

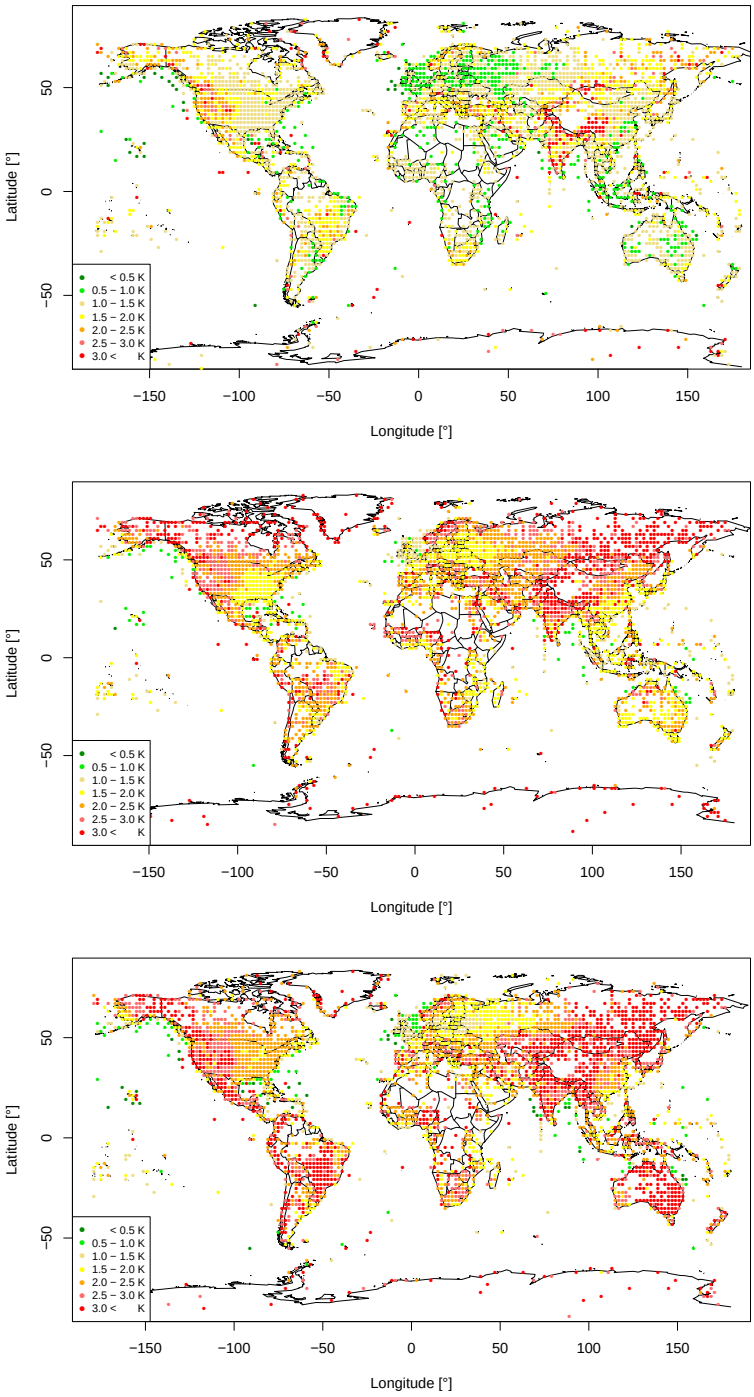

Figure B.5: 2° clustered MAE world maps of ERA5 (top), NEMS (middle) and GFS05 (bottom)

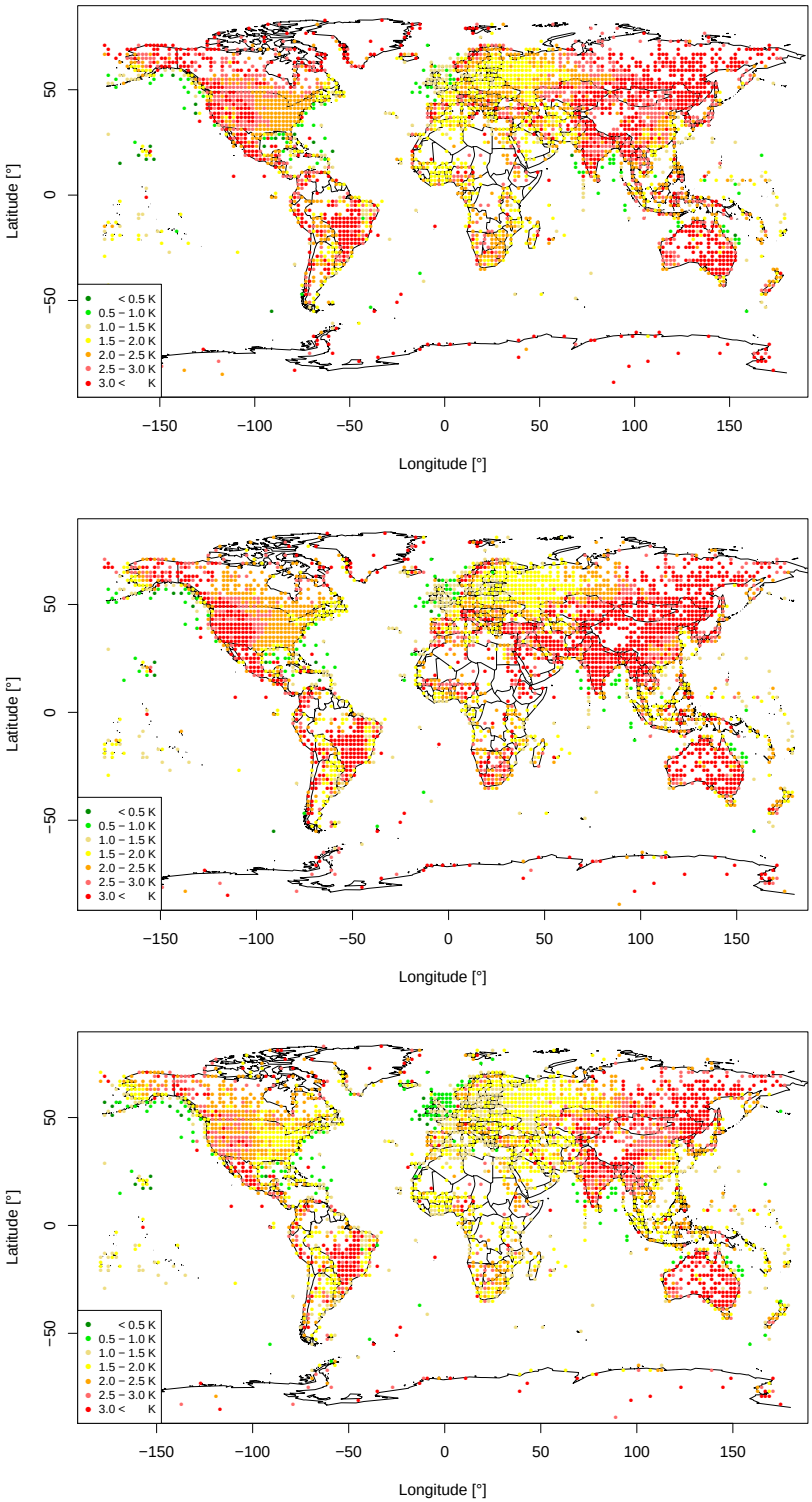

Figure B.6: 2° clustered MAE world maps of MF (top), GEM (middle) and ICON (bottom)

B.0.5 World maps: 2° gridded MBE

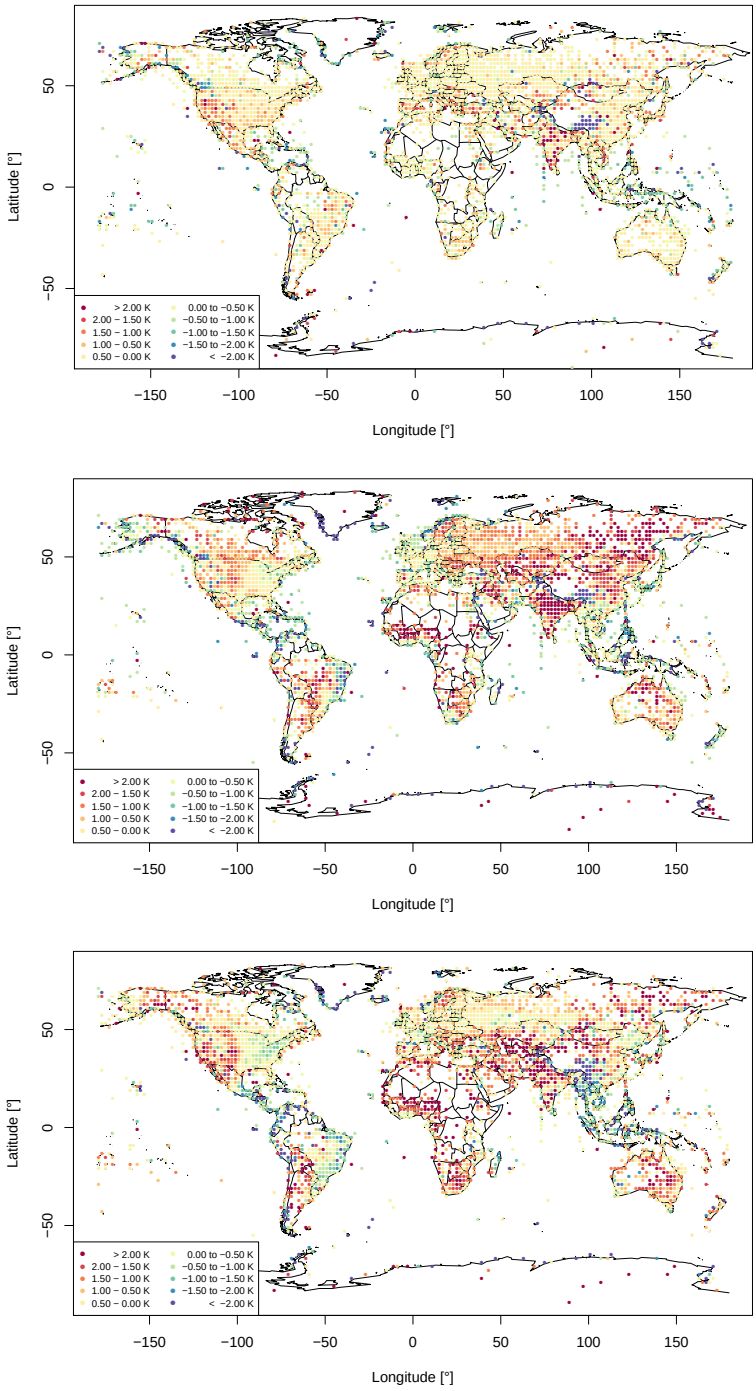

Figure B.7: 2° clustered MBE world maps on ERA5 (top), NEMS (middle) and GFS05 (bottom)

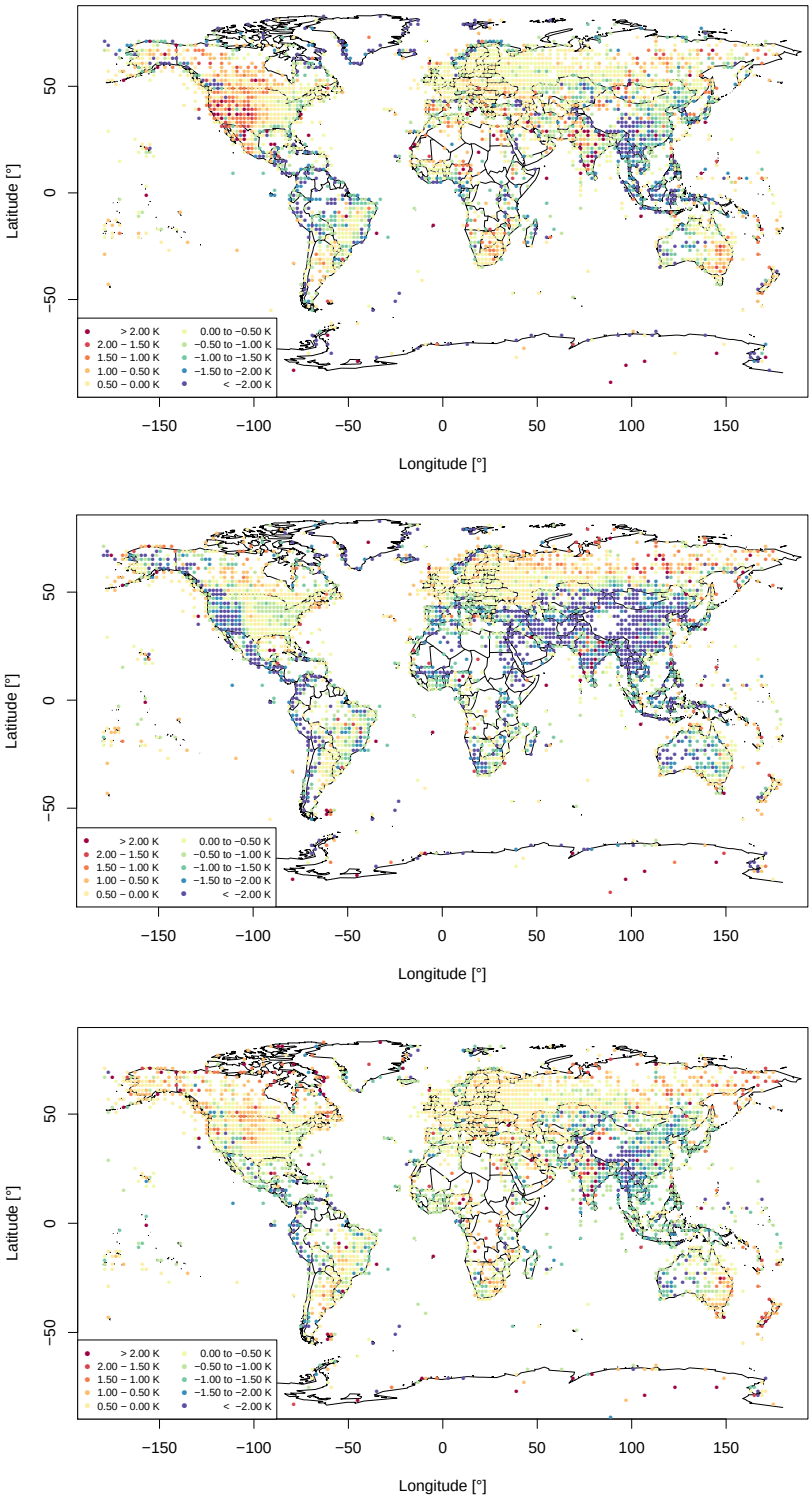

Figure B.8: 2° clustered MBE world maps on on MF (top), GEM (middle) and ICON (bottom)

B.0.6 World maps: 2° gridded Minimum MAE forecast

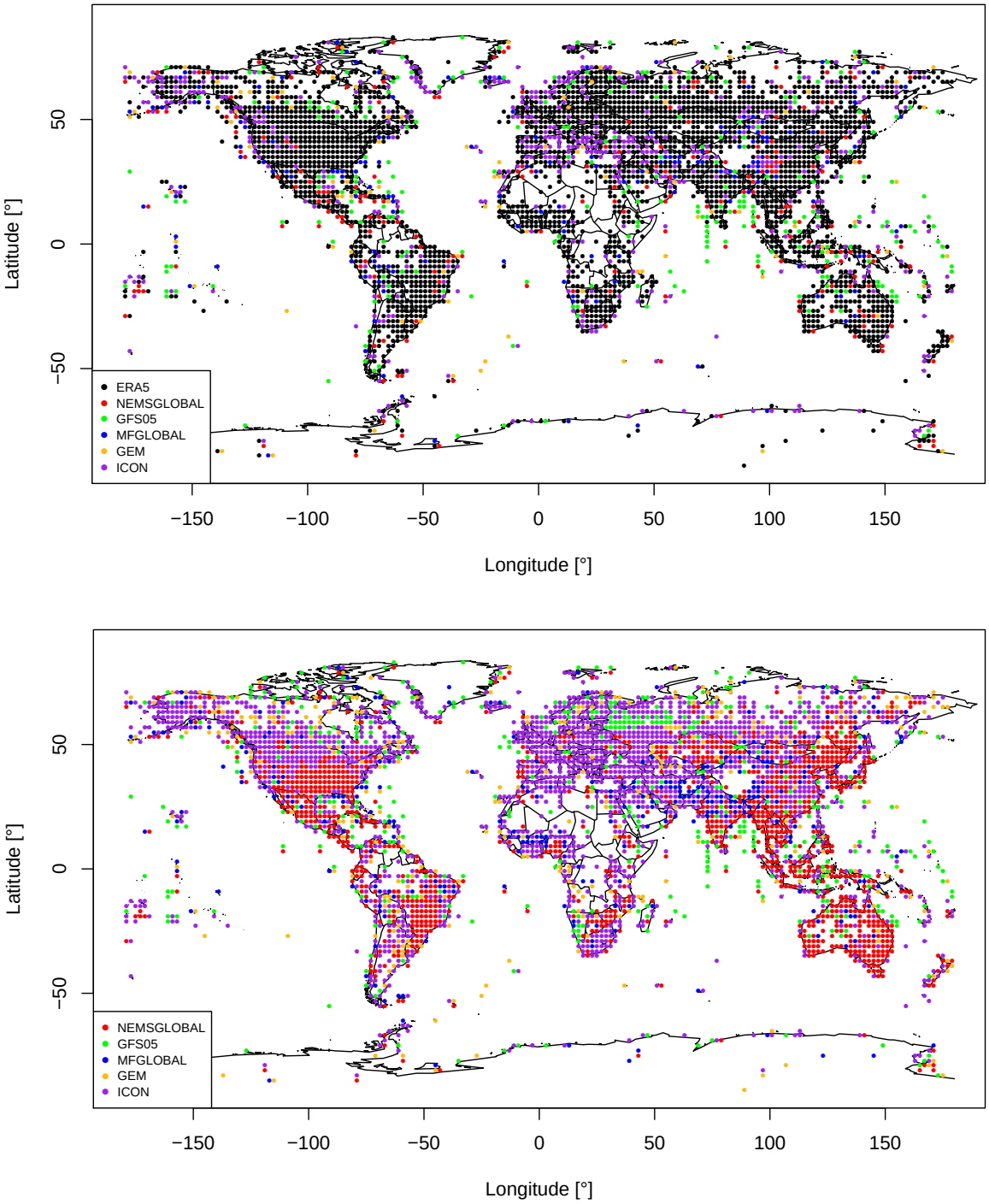

Figure B.9: Minimum MAE distribution on all six global models (top) and minimum without ERA5 (bottom)

B.0.7 World maps: 2° gridded Minimum and Maximum MBE forecast

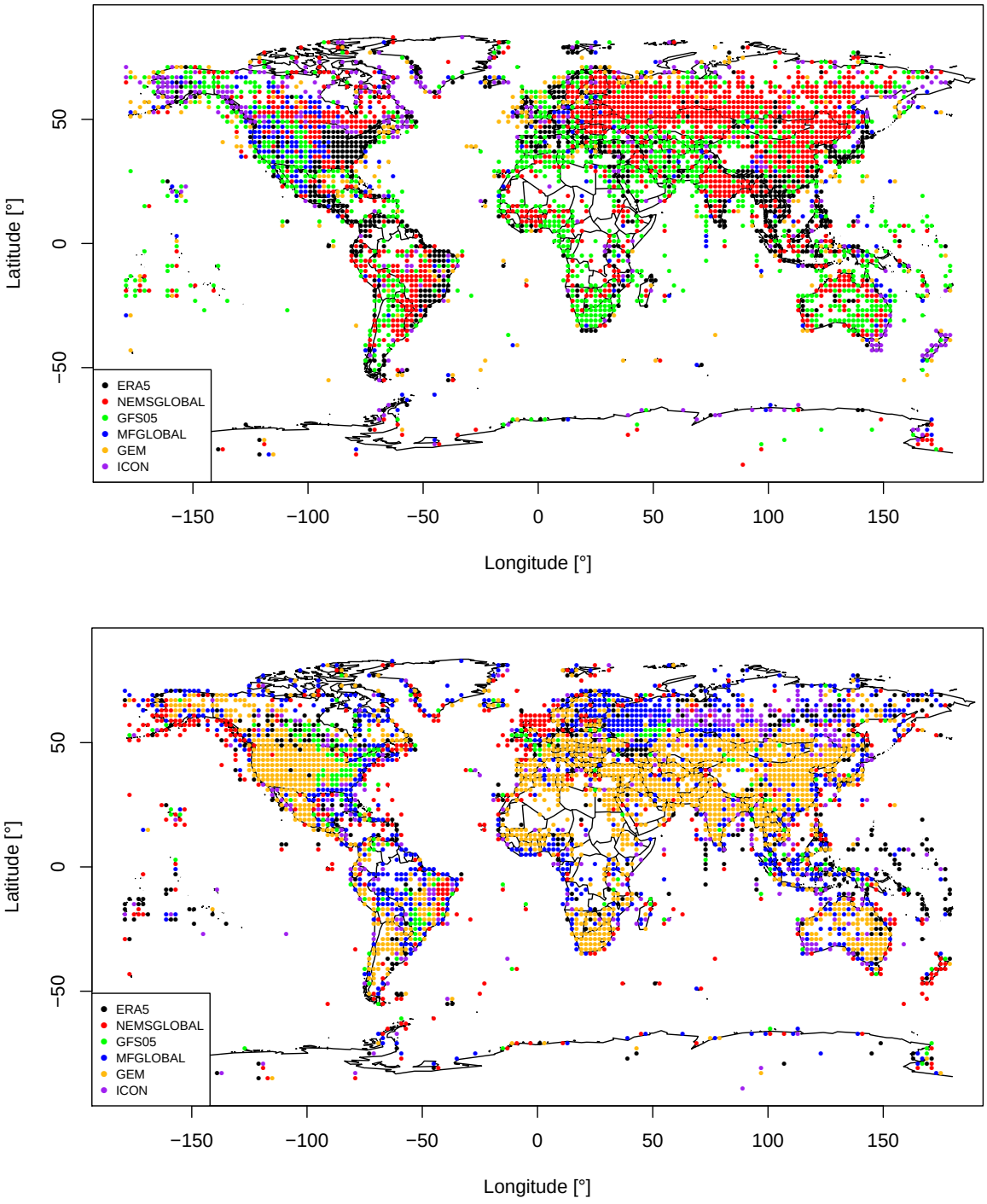

Figure B.10: Maximum (top) and minimum (bottom) MBE distribution on all six global models

B.0.8 World maps: 2° gridded Model spread

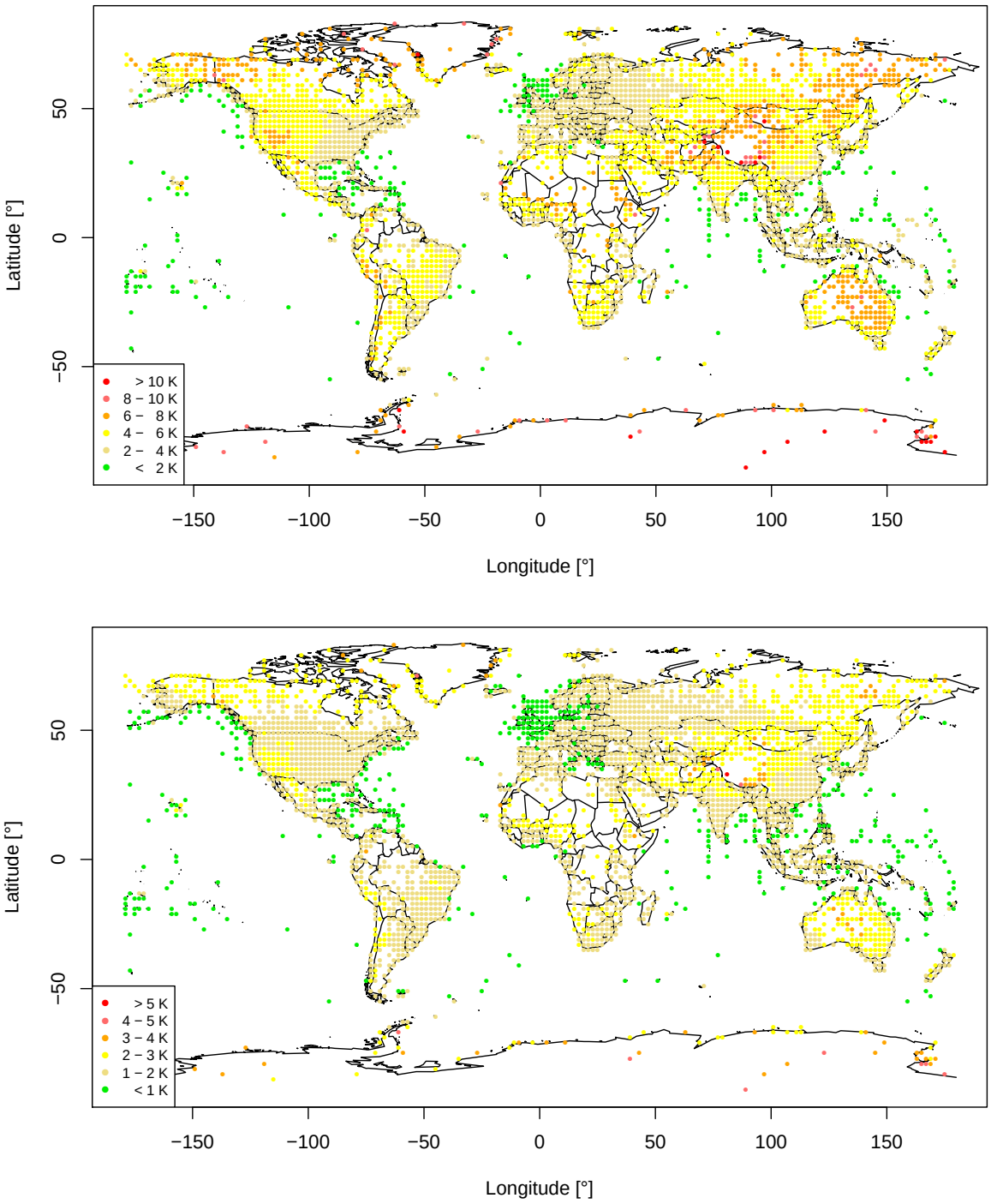

Figure B.11: Both second model spread approaches: max-min (top), standard deviation (bottom)