Master thesis

Analysis and improvement of the quality of global wind forecasts

Brigitte Häuser
Student number: 4121964

Albert-Ludwigs-University Freiburg
Chair of Environmental Meteorology
In Cooperation with meteoblue

December 8th, 2021

Writing Period

10. 06. 2021 – 10. 12. 2021

Supervisor

Prof. Dr. Dirk Schindler

Prof. Dr. Anke Weidlich

Declaration

I hereby confirm that my master thesis is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.
Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date	Signature

Abstract

In this work, wind speed models are analyzed and compared on a global scale as well as for Europe. The reanalysis model ERA5 as well as the prediction models ICON, MFGLOBAL, GFS05, and models from the NEMS family are validated with metar measurement data. The analysis includes a variety of error metrics for overall and high wind speed performance. Furthermore, in order to improve the forecast skill, two multimodel approaches are implemented. The first approach uses a global multimodel, while the second adjust weightings locally.

ICON has been found to be the prediction model with the highest overall performance, while the performance does not vary strongly among the selected models. On a spatial level, wind predictions near oceans were found to have a worse performance than for continental stations. Furthermore, only NEMSGLOBAL showed a strong spatial distribution of bias (MBE). All models lack performance in the prediction of high wind speeds. This could not be enhanced by multimodeling. However, both multimodel approaches could improve forecasts when all wind speed values were included, enabling simple global prediction improvement as well as optimized model combinations for different locations.

Contents

List of Abbreviations

API	Application Programming Interface
DWD	Deutscher Wetterdienst
E	Random luck
ECMWF	European Centre for Medium-Range Weather Forecasts
EPS	Ensemble prediction system
FAR	False alarm ratio
HSS	Heidke skill score
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MBE	Mean bias error
metar	Meteorological Terminal Air Report
MM	Multimodel
NaN	Not a Number
NWP	Numerical weather prediction
PC	Portion correct
POD	Probability of detection
QC	Quality control
RMSE	Moot mean square error

1 Introduction

Wind speed prediction plays an important role in many aspects of life. In a world threatened by climate change, reducing carbon emissions by phasing out the use of fossil fuels is essential to reduce the large emission burden. In order to cope with the worldwide increase in energy consumption [1], renewable energies must be promoted and further developed. Wind as a renewable energy resource will have a major role in the future energy mix [2]. To find suitable locations for wind turbines and to be able to assess their potential, estimating the wind speed power is needed for different sites which do not have measurement stations. For this, accurate wind speed predictions need to be used. Besides, estimating the danger posed by extreme weather events is relevant for insurances and agriculture and can even influences the evacuation of people. For example, severe storms have caused the most financial damage and fatalities of all natural disasters in the US over the past 40 years [3]. Thus, forecasts of wind speed can not only save money, but have the power to potentially save thousands of lives. Furthermore, the correct estimation of wind is one crucial part in the understanding of the overall structure of the atmospheric system. Wind speed is a relevant input parameter for complete atmospheric forecasts in order to predict the movement of air packages. Therefore, wind has a big influence on parameters like air temperature or precipitation, which themselves are important drivers for human health, agriculture, and insurances.

When creating models, a network of measuring stations is needed so that the quality of the models can be checked. At the same time, models replace the need for an even larger number of measuring stations. Instead of carrying out long-term measurements at each point of interest, such as a potential site for wind turbines, model data for these regions can be used. This saves a lot of money and allows for better area-wide analyses and spatial patterns.

Therefore, wind speed prediction models have great relevance. At the same time, wind speed is a parameter that is hard to predict. Wind is variable and strongly dependent on regional conditions such as the roughness of the surface or surrounding mountains or bodies of water. The prediction is further complicated by measurement inaccuracies and limitations of numerical weather prediction (NWP) models. Mostly three different limiting factors exist for NWP models: The limited computational capacity, unknown initial conditions, and the simplified parameterization of physical processes. The computational capacity limits the accuracy of the prediction: On the one hand, a higher spatial resolution needs higher computational power because more cells have to be calculated. The

correlation of the reduction of the cell width and the increase of the grid cells is exponential: If the grid width of a model reduced from 3 km to 1 km, 9 times as many cells have to be calculated on a horizontal scale. Additionally, grid cells multiply due to the three-dimensionality of the grid. Furthermore, space and time scales are related: spatially smaller processes refer to a shorter time span. If one observes a small spatial resolution, one must also consider smaller time steps. Thus, the required computing power increases strongly with a reduction of the grid.

On the other hand, the amount of input parameters increases the model complexity – this increases the model performance as well as complicates the calculation of it. The unknown initial conditions of the atmospheric state lead to input parameters deviating from reality. Despite a large number of weather measurement stations, the current state of the atmosphere is never fully known. This creates uncertainty in the data assimilation. However, even if there is complete knowledge about a system, processes are often too complex to implement them completely. Therefore, they need to be represented in a simplified way in the forecast models. This decreases the accuracy of the models. Nevertheless, computational capacity is constantly increasing and more and more knowledge about the atmospherical processes develops. At the same time, the quantity and quality of observation increases. This enables NWP models to improve their accuracy. Because many models rely on input parameters from other models, the improvement of one model increases the performance of other models, too. Because models can never copy the reality and always show a simplification, models can differ in the ways of approximation such as the inclusion of environmental factors like the altitude or distance to the sea. Thus, apart from improving the existing models, a better understanding of strengths and weaknesses of different models in a spatial distribution as well as the prediction skill of varying wind speeds would be of great benefit. This work helps to evaluate selected wind speed models and compare them to each other. Further, two multimodel approaches for improving the prediction performance are presented.

2 Research questions

In this work, the reanalysis model ERA5, the global scale prediction models ICON, MF-GLOBAL, GFS05 and NEMSGLOBAL are evaluated. For Europe, the regional scale prediction models NEMS12 and NEMS4 are additionally used. For the evaluation, the following questions should be answered: To what extent does the wind speed prediction performance of the models studied differ? Are there spatial differences in which models forecast better? Important subquestions to help answer these general questions are: How high are the mean absolute error and mean bias error of the individual models for wind speed? What are the detection probability and the false alarm rate for strong wind events?

Moreover, an improvement of model prediction is desirable. For this purpose, the question to be answered in this work is: To what extent can model performance be improved by weighting different models? How much does model performance vary when multimodel weights are optimized regionally versus globally?

3 State of the art

3.1 Numerical weather models

Numerical weather prediction (NWP) models predict an atmospheric state for a specific time and location. NWP models need suitable initial and boundary conditions as input variables such as the observed current weather or predictions from other models. Future weather can be approximated by partial differential equations [4]. In global numerical weather prediction models, the earth is covered by a three-dimensional grid. This grid is mostly rectangular, but there are other approaches like a triangular shape [5]. The grid size determines the spatial resolution of the model. Grid sizes vary greatly among models, and even within a model, there can be differences in spatial resolution. On the one hand, layers close to the ground are usually resolved higher in order to better predict processes close to the ground, and on the other hand, areas of particular interest to the model operator are often spatially resolved higher [6].

Processes, which are smaller than the spatial resolution or shorter than the temporal resolution are called subscale. Subscale processes cannot be resolved, but are parameterized to be included in model prediction. For wind predictions, these are especially processes due to orography, such as mountain-valley winds or turbulence due to changes in ground roughness [7].

Due to the nature of partial differential equations, small errors always arise in the prediction, which grows in the time of forecasting. Therefore, models degrade in performance the longer the weather forecast extends into the future. In addition, forecasting further into the future is complicated by the movement of air masses, which means that larger areas must be known. For example, in Germany, a global model must be used for a forecast of 5 days or more [7].

Nevertheless, numerical weather models show an improving performance over the last decades [4]. This is a result of the technological development as well as a better scientific understanding of atmospheric processes. Satellite data enables data with global coverage so that predictive skill in the Northern and Southern hemisphere is almost equal. However, small differences in the initial state can lead to big changes in the forecast due to the butterfly effect, which describes the unpredictable evolution of a system due to an arbitrarily small change in initial conditions [8]. Therefore, data assimilation is an important part of prediction models.

3.2 Quality control of measurement data

Errors in measurements can occur at many stages, reaching from inaccuracies or failures of the measuring devices to storage problems, e.g. if formats are not recognized. Errors can be classified into random, systematic, and rough errors [9]. Random errors have a zero-centered normal distribution and are unavoidably for all data, independent of the measured value. Systematic errors are distributed asymmetrically, describing a persisting bias in the data. They can have multiple origins, e.g. instrument bias or calibration drifts [10]. Rough errors are caused by instrument malfunctions or are communication-related - even if they generally occur on only a small part of the data, the distortion can be huge [9]. Systematic and rough errors can be flagged, eliminated, and eventually corrected. This procedure is named quality control (QC) and is aimed to ensure the quality of the measurements. Within the QC, there are different options to check data: In the **plausibility check**, one searches for data outside the plausible data range. For wind speed data with limited availability of metadata, an upper limit of 100 $\frac{m}{s}$ can be established as well as 0 $\frac{m}{s}$ as lower limit, because wind speed can not be negative [11].
In a **temporal consistency check**, extreme variability or unrealistic steady behaviors are figured out. For wind speed, periods with a very low variability often result from damaged or frozen instruments or faulty communications between an instrument and the datalogger [12]. In order to identify low variability errors, there are different approaches: Data can for example be flagged in a moving window with a standard deviation below a threshold, which is previously defined [12]. Alternatively, it can be searched for constant data sequences for an unrealistically long period [13].
On the other hand, excessive high variability is typically a consequence of technical issues and is less common than low variability errors [14]. In the step check, the difference between two following observations is calculated. If the difference exceeds a given threshold, both values are flagged [13]. Another approach is the blip test, which looks for spikes and dips, which has the advantage of being able to distinguish the faulty record from the good ones [15].
Spatial checks compare the observations of a site in relation to those at some neighbor location [9]. The observation can be considered suspect when the differences of the values of neighbor measurements exceed a given threshold [14] or they change in a very different form over time.
Duplication error checks identify measurement periods that could be falsely duplicated, either within the measuring period of one station or between different stations.

Last, **typographical error checks** are relevant when the observations were recorded on paper and later digitized. Here, errors existing due to human mistakes by transcribing and transferring are investigated.
Besides these five approaches, it needs chronological sorting of data, removal of repeated dates, unification of measurement units, and standardization of the observation time [14]. All of these tests handle mostly with rough errors. However, detection of systematic errors is partly possible, for example, if there are metadata about changes of the measuring height, location, or the measuring devices [11]. In order to handle systematic errors, instead of flagging, data can be smoothed or the long-term bias is calculated and offsets with the measurement data.

3.3 Multimodels

In order to improve wind speed forecasts, different approaches of using multiple models have been made. In a study that investigated the summer westerly jet, a multimodel approach was used [16]. Four models in high resolution were analyzed as well as the mean predictions of these four models. In this study, all models showed similar results regarding the change of the summer westerly jet during the mid-Holocene. In order to determine the strength of the change, the average of the model prediction was used.
Another approach of multimodeling was applied in a study from Austria [17]. For the forecast, the ensemble prediction system (EPS) with 51 members from European Centre for Medium-Range Weather Forecasts (ECMWF) was the basis. Because the EPS is underdispersive and hence uncalibrated, the members of the EPS are clustered to eight weighted representative members with a maximized inter-cluster spread and a minimized within-cluster spread. With two limited area models, the forecasts were downscaled. With the resulting four ensembles, two post-processes were implemented: model averaging and heteroscedastic censored regression, both indicated more accurate results than the EPS. In order to evaluate the accuracy, different error metrics were calculated like the MAE, continuous ranked probability score, probability transform integrals, and verification rank histogram. However, the authors highlight that the value of a forecast can not be fully evaluated by just the accuracy and calibration measurements.
ECMWF wind speed ensembles, together with professional numerical models, were also used in a study in the interest of creating a multimodel forecast product of wind speed [18]. The multimodel is obtained by two model averaging methods to improve the time resolution and improve the model forecast accuracy. The multimodel showed a decreased

MAE and a higher correlation than both, ECMWF ensemble forecast and the existing numerical forecast products. The MAE was decreased by 24.3 % compared to the numerical model forecast and 11.7 % compared to the ECMWF ensemble. The correlation was 14.5% higher than for the ECMWF ensemble and 12.5 % than the numerical model.
To evaluate European storm events and get robust estimations of extreme value statistical parameters like the return period, an ensemble of coupled global climate models was used [19]. The idea behind it was to enlarge the statistical sample size by utilizing the inter-model variability. The study found that in regions with decreasing return periods of storm events the multimodel the diagnosed trend is more robust for the multimodels than for single models.
In another paper, machine learning was used in order to develop a data-driven multimodel wind forecasting methodology [20]. Two layers were used to achieve this goal: First, multiple machine learning models generated individual forecasts. The inputs for these machine learning models were generated by another machine learning framework. In the second layer, an ensemble of the individual forecasts was made by a blending algorithm, generating deterministic and probabilistic forecasts. Compared to different single-algorithm models, the multimodel framework shows better 1-hour-ahead wind speed forecasts. Their results show increasing forecasting accuracy by up to 30 %. Summarising these investigations, multimodel approaches differed greatly in implementation in some cases, but indicate a potential improvement in predictions.

4 Objectives and hypotheses

The aim of this work is to achieve a global analysis of the reanalysis model ERA5 as well as for the prediction models ICON, MFGLOBAL, GFS05, and NEMSGLOBAL. While the last four models are used for forecasting, ERA5 is a model with high accuracy for comparing the results of the prediction models. Therefore, the first hypothesis is:

H1: ERA5, as a reanalysis model, performs better than the forecast models.

Because prediction models are mainly used by the weather services of the respective countries, these models probably have been optimized for forecasting in their own countries than for other areas. This leads to the second hypothesis:

H2: Prediction models perform better in the regions in which they were developed.

In addition, the regional models NEMS12 and NEMS4 with smaller grid sizes than the related global model NEMSGLOBAL are evaluated for a part of Europe in order to compare regional with global models. The related hypothesis is:

H3: As wind speed models with a lower grid size, NEMS4 and NEMS12 show higher accuracies in predictions than models with bigger grid sizes.

Another aim of this work is not only the analysis, but also the improvement of the prediction skill: By finding the best weighting of the four global prediction models, predictions should be even more accurate than the best used prediction model. In order to account for regional differences and find the best multimodel, an adjusted weighting for each used observation station should be implemented. The fourth hypothesis is:

H4: A site-specific multimodel of optimized weighting can perform better than the best raw model.

An additional objective of this work is to find a global weighting for a multimodel that has a good overall performance and can be also used in regions without a measurement station. The fifth and last hypothesis is:

H5: While the global multimodel has worse accuracy than the regionally adjusted multimodels, the global multimodel still shows better results than the raw models.

5 Methods

5.1 Measurement and model data

5.1.1 Data origin

For the measurement data, wind speed data for 2019 and 2020 in 10-meter height from 6215 weather observations were used. The punctual observations were located on single airfields all over the world, providing their measurements in Meteorological Aerodrome Report (metar) format in UTC time. For the models, the cells where measurement stations were located were identified. In Figure 1, it can be seen that the location of model cells can differ for the same measurement station and different models, depending on the grid size, shape, and position of the model. For these cells, raster data of hourly wind speed were downloaded for 2019 and 2020. Access to measurement and model data was provided by meteoblue. For the measurements, the meteoblue measurements API was used [21]. Model data was accessed by the Python Dataset SDK [22]. Both, measurement and model data were provided in JSON format. Different models were used, the reanalysis model ERA5 as well as the prediction models ICON, GFS05, MFGLOBAL, and NEMSGLOBAL for the global analysis as well as two regional NEMS models for Europe. ICON, MFGLOBAL, and GFS05, which are provided in 3-hour resolution, are interpolated to hourly resolution from meteoblue. All models are described in the following.

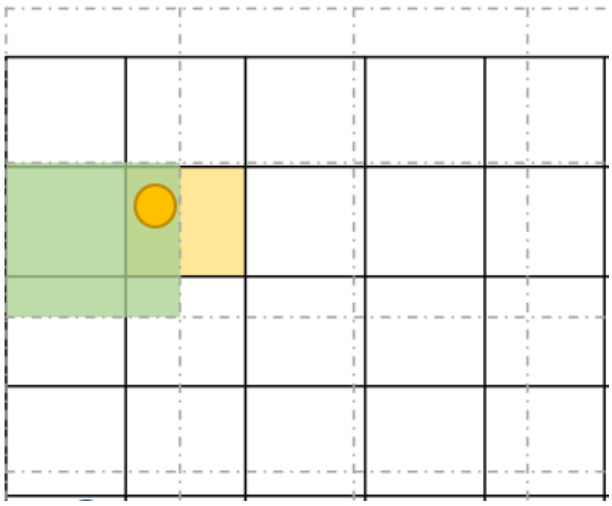


Figure 1: An exemplary measuring station (yellow dot) with two model grids (dashed and complete line). The model grid of the respective models used for the validation at this measuring station is shown in color.

5.1.2 Models

5.1.2.1 ICON

ICON (Icosahedral Non-hydrostatic) is a global prediction model. It is a joint project of the German Weather Service (DWD) and the Max Planck Institute for Meteorology. ICON, unlike other models, does not have a rectangular grid, but a triangular one, each representing an area of about 173 km^2. The triangular grid results in less variability in the area of the grids than rectangular grid cells due to cartographic reasons. The root of the mean area of the spherical triangles is taken as the effective mesh size, which is 13 km for ICON. In Europe, ICON has refined its grid into a 6.5 km grid. ICON is provided in a 3-hourly resolution from the DWD [5]. It covers a period from 2017 to the present [23]. ICON performs well over steep mountain slopes and shows better results than its predecessor at DWD, the Global Model Europe (GME), even though it requires less computer computation power [24]. However, the diurnal cycle is not represented well. Especially, nocturnal wind speeds in the summertime at onshore locations are underestimated by ICON [25].

5.1.2.2 GFS05

GFS (Global Forecast System) is a global prediction model provided by National Centers for Environmental Prediction (NCEP), a department of the U.S. weather service NOAA. GFS is available in a 0.5 °and a 0.25 °grid. GFS05 refers to the 0.5 °grid. The spatial grid varies between 27 and 70 km, dependent on the forecast time: forecasts in two weeks have larger grid cells than those in a few days. In this work, a spatial resolution of 40 km is used. The model is calculated four times a day, providing data in a 3-hour resolution [26]. GFS was found to be able to reflect the diurnal variation of wind speed near the surface under stable weather conditions [27].

5.1.2.3 MFGLOBAL

MFGLOBAL (Météo-France Global) is the global prediction model of the french national weather service Météo-France. The underlying model is ARPEGE, which has a spatial resolution of 40 km for the world (MFGLOBAL) and 11km for Europe (MFEU). It runs every 6 h and generates predictions in a 3-hour resolution [6]. MFGLOBAL is a NWP model with a stretched grid [28].

5.1.2.4 NEMS

NEMS (NOAA Environment Monitoring System) are prediction models from meteoblue. They cover a period from 1984 to the present in hourly time resolution. The spatial resolution is 30 km for the global model NEMSGLOBAL. Besides, there are many regional models with higher resolution like NEMS12 for Europe in 12 km resolution or NEMS4 in 4 km resolution for central Europe. Also, there are higher-resolution models for America, Africa, New Zealand, and Japan. The models run once a day. NEMS does not have an own assimilation system and uses GFS assimilation data instead. The first publishing of the NEMS family was in 2013, being NMM (Nonhydrostatic Multiscale Model) successors [23]. For NMM, wind speed predictions are found to be better for regions with a big distance from the sea. The spatial pattern of the predictions seems to be controlled on large scales since smaller-scale features such as mountain ranges are not depicted very well [29].

5.1.2.5 ERA5

ERA5 is a reanalysis model released by ECMWF as part of Copernicus Climate Change Services [30]. The global horizontal resolution of the hourly data is about 31 km and it covers a period from 1950 on. Reanalysis models use forecast models and historical measurement data in order to create longer-term meteorological data sets. Because they rely on past measurements, they are not used for forecasts. However, they can be used as high-quality models for evaluating the predictions of forecast models. Preliminary ERA5 data is published within 5 days of real-time, quality-assured updates are available within 3 months [31]. Comparison of wind speed observation stations with ERA5 data shows that ERA5 could reproduce the wind speed spectrum range for any location in Europe [32]. In a study, different wind speed models over France are compared [33]. ERA5 performs well, however, it underestimates wind speeds in France, especially in mountainous areas.

5.2 Data preparation and quality control

All data processing and error calculation has been done in Python version 3.9.6, mainly using packages numpy version 1.21.1, pandas version 1.3.1, and geopandas version 0.9.0. The model data was provided to the full hour, while the timestamps of the measurement data varied. For a comparison between them, the timestamp of the measurements was adapted. For the irregular hourly values, the measurements were assigned to the closest full hour. Until minute 29, the timestamp was transformed to the lower hour. From minute 30 on, the timestamp was assigned to the next full hour. Seconds in the timestamp were dropped. A timestamp can be duplicated, for example, 4.40 am and 5.20 am are both assigned to 5.00 am. In these cases, duplicated wind speed data was averaged. Missing hours due to missing measurement data were generated as a timestamp and the corresponding wind speed data was filled with NaN values. By adapting the measurement timestamps, the wind speed data from measurements and models got comparable. As a quality control (QC), wind speed values below zero and above 100 $\frac{m}{s}$ were filtered. Measurement data which did not change in 24 hours points to a measuring error because wind speed is naturally very variable. They were covered with NaNs in the QC. This filtering allowed one NaN in the 24 h window. For more missing data, the data was flagged because there is not enough information about potential changes. The availability of the measurement data is determined before and after the QC. To be used for calculating error metrics and plotting, an observation station needs to have an availability above 30 % after the QC within the observed year.

5.3 Error metrics

5.3.1 Formulas

5.3.1.1 MAE

The mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. It describes an arithmetic average of the absolute errors between prediction and measurement. An MAE near zero is desirable, indicating predictions that are, on average, very close to the measured values. The MAE is calculated as [34]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - M_i| \tag{I}$$

with n as the number of prediction values, P_i as the predictions, and M_i as the observation values.

5.3.1.2 MBE

The mean bias error (MBE) is the arithmetic average of the error of paired observations. Unlike for the MAE, the errors are not absolute. Thus, a negative MBE can occur, showing an underprediction. On the other hand, a positive MBE shows an overprediction - the best value is zero, meaning that the model does not have a bias [35].

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - M_i) \tag{II}$$

with n as the number of prediction values, P_i as the predictions, and M_i as the observation values.

5.3.1.3 MAPE

The mean absolute percentage error (MAPE) is an error metric that expresses the error as the sum of the absolute ratio of the error for each value pair divided by the prediction. The MAPE consideres the fact that errors generally become larger when predicted values are higher. For example, if an observation value for wind speed of 30 $\frac{m}{s}$ would occur and the prediction would be 32 $\frac{m}{s}$, the absolute error would be 2 $\frac{m}{s}$, the absolute percentage error would be 6,67 %. For an observation of 1 $\frac{m}{s}$ and a prediction value of 3 $\frac{m}{s}$, the absolute error would be again 2 $\frac{m}{s}$, but the absolute percentage error would be 200 %.

Therefore, the minimum of MAPE is 0 %, which is the optimum, and range theoretically infinitely far. It is calculated by [36]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{P_i - M_i}{P_i}\right| \tag{III}$$

with n as the number of prediction values, P_i as the predictions, and M_i as the observation values.

5.3.1.4 RMSE

The root mean square error (RMSE) is the root of the mean of the squared differences between prediction and observations. RMSE is always positive, the best value would be zero, showing a perfect fit from the model to the observations. Due to the square, RMSE particularly strongly involves larger errors compared to MAE, MBE or MAPE, being sensitive to outliers. The RMSE is calculated as [34]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - M_i)^2} \tag{IV}$$

with n as the number of prediction values, P_i as the predictions, and M_i as the observation values.

5.3.1.5 Spearman correlation

Spearman's rank correlation coefficient shows the statistical dependence between the rankings of two variables. The best correlation is one, showing that both, observation and prediction values, have the same rank. The lowest correlation is zero. The Spearman correlation, unlike the Pearson correlation, may also be used with non-normally distributed data. Because wind speed follows a Weibull distribution [37], spearman correlation can be used for wind speed data. The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables [38]. If all n ranks are distinct integers, the spearman correlation can be calculated by:

$$r_s = 1 - \frac{6\sum(R(X_i) - R(Y_i))^2}{n(n^2 - 1)} \tag{V}$$

with n as the number of observations and $R(X_i) - R(Y_i)$ as the difference between the two ranks of each observation.

5.3.1.6 FAR

In order to evaluate how well models can predict higher values, error metrics for measurement values above a certain threshold can be calculated. The contingency table in Figure 2 shows the combinations for measurements and model values above and below the threshold.

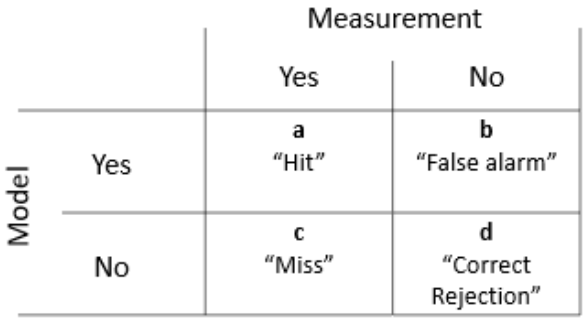


Figure 2: Contingency table for measurements and model predictions. "Yes" means that the model or measurement value is above a certain threshold, "No" means that the value is below the threshold. Inspired by [39].

The false alarm ratio (FAR) describes the probability of the model falsely predicting a value above the threshold. The range is from 0 to 100 %. 0 % would mean, that the model never predicts a value above the threshold when the measurement value is below it. The FAR is calculated by [40]:

$$FAR = \frac{b}{a+b} \tag{VI}$$

with b as the number of false alarms and a as the number of hits.

5.3.1.7 POD

The probability of detection (POD) is an error metric describing the probability of a model prediction above a threshold when the measurement exceeds this threshold. The best value is 100 %, which would mean that the model always has values above the threshold, when the measurement does. 0 % is the worst POD value, meaning that all measurement values above the threshold are missed by the model. It is calculated by [40]:

$$POD = \frac{a}{a+c} \tag{VII}$$

with a as the amounts of hits and c the number of misses.

5.3.1.8 HSS

The Heidke skill score (HSS) describes the superiority to randomly correct estimates. Therefore, it can describe a real skill of predictions. For a HSS of zero, there is no prediction skill at all. A value of one shows perfect skill. The HSS is calculated by the portion correct (PC), the proportion of correctly predicted values out of all values, and the random luck value (E) [39]. PC and E are composed as shown in the following:

$$PC = \frac{a+d}{n} \tag{VIII}$$

$$E = \frac{a+b}{n} \cdot \frac{a+c}{n} + \frac{d+b}{n} \cdot \frac{d+c}{n} \tag{IX}$$

with a as the amounts of hits, b the number of false alarms, c the number of misses, d the number of correct rejections, and n the number of observations $n = a + b + c + d$. The HSS is finally calculated by:

$$HSS = \frac{PC - E}{1 - E} \tag{X}$$

5.3.2 Implementation

For each station and model, error metrics were calculated for 2019 and 2020 separately. MAE, MBE, MAPE, RMSE are calculated as well as the Spearman correlation with all values and the spearman correlation with only value pairs, where the measurement reported a wind speed value over 3 $\frac{m}{s}$. Besides, POD, FAR and HSS were calculated for wind speed thresholds of 5, 15, 20, and 30 $\frac{m}{s}$ in order to evaluate the performance of high wind speed predictions. Because NEMS4 and NEMS12 were only available in Europe, the analysis was split up into two parts: The first part analyses ERA5, ICON, GFS05, MFGLOBAL, and NEMSGLOBAL for all measurement stations worldwide. The second part includes NEMS4 and NEMS12, but only involves stations in Europe, where NEMS4 and NEMS12 data were available.
The error metrics for all stations for both parts were saved in dictionaries with corresponding station ID, latitude, longitude, and the availability of the measurement data.
In order to be able to evaluate the models overall, error metrics were averaged over all stations for an overview.

5.4 Multimodels

In order to improve model predictions and identify spatial strengths of the models, multimodels of weighted models were implemented. Because ERA5 is a reanalysis model and is not available for forecasts, the multimodel approach used only the four global prediction models. For each station, mixed model data was calculated by combining ICON, GFS05, MFGLOBAL, and NEMSGLOBAL. Every model could reach weightings from 0 to 100 % in 10 % steps. Overall, 256 combinations were reasonable for each station, so that the four model weightings added up to 100 %. For all combinations, the MAE was calculated. Two multimodel approaches were implemented: The first approach searched for each individual station the weighting with the lowest MAE and created a different multimodel mixture for each measurement station. The second approach calculated one weighting, which is fixed for all stations. To find the optimal weighting for all stations, the mean of the MAEs for all stations was taken for each weighting combination. This led to a list with averaged MAEs for each weighting combination. By searching for the lowest MAE in this list, the optimized global weighting was found. For both approaches, all mentioned error metrics were calculated for each station with the optimized weighting. In order to evaluate the benefit of additional models in the global weighting approach, combinations of only two or three models instead of all four were calculated additionally. For a global multimodel approach with three models, the weightings of each model after another were set to 0 % in order to find the lowest MAE for each combination of three. The best weighting for the best combination was determined by choosing the smallest MAE of all combinations. For a multimodel with two models, the weightings of two models were set to 0 % at the same time and the weightings for the lowest MAE were identified. This was done for all combinations.

5.5 Plots

For plotting, Python's packages matplotlib version 3.4.2 and seaborn version 0.11.1 were used as well as geopandas version 0.9.0. The shape of the country borders was used from geopandas. The European border was provided by EFRAINMAPS [41]. For the maps, latitude and longitude of the measurement stations were read in as points in the spatial reference system EPSG:4326. The error metrics are plotted as colored points at the station's location. For the world plots, 5180 stations with an availability of over 30 % in 2020 existed. These stations covered each other when drawing the points large enough to recognize the error values. In order to avoid this overplotting, error metrics of stations in a 3x3 °grid were averaged. The averaged point is shown in the middle of the grid. For unifying reasons, single stations were moved to the middle of the grid, too. In order to not let the scaling be distorted by a few extreme values, maximum values of the color scales correspond to the 99% quantile of the respective error metrics overall models, rounded up to 0.5 units. For the correlation, a scale from 0 to 1 was used.

6 Results

6.1 Validation of existing models

6.1.1 Global models

From 6215 global measurement stations, 1035 have availabilities of hourly data below 30 % after the QC in 2020. These 1035 stations are not considered in mean values and plots of the results. For 31 % of the global stations, the QC found values that could be sorted out. Overall, the average availability for hourly measurement data in 2020 for one station is around 72.1 % before QC and 69.9 % after it. Thus, QC sorts out 2.2 % of the measurement data.

For the raw model validation, the MAEs between the wind speed data from measurement stations and the model predictions of GFS05, ERA5, NEMSGLOBAL, ICON, and MFGLOBAL are shown in Figure 3. For all models, over 90 % of MAE values lie between 1 and 3 $\frac{m}{s}$. The highest MAE per station reaches from 12.8 to 15.2 $\frac{m}{s}$, which is not shown in the plot to avoid distortion due to extreme values. ICON and ERA5 show the lowest MAEs over all stations and have very similar values. GFS05 and MFGLOBAL values lie in the middle, while NEMSGLOBAL shows higher MAEs over all stations. Averaged over all stations, MAE reach from 1.453 (ICON) to 1.866 $\frac{m}{s}$ (NEMSGLOBAL). The values for all models can be seen in Table 1. However, that does not mean that NEMSGLOBAL can not have specific stations with better MAEs than other models or ICON is always the best choice. The same value on the x-axis does not show the same station but the same place in the order of sorted MAEs for the respective model. Therefore, a spatial analysis enables a more comprehensive picture.

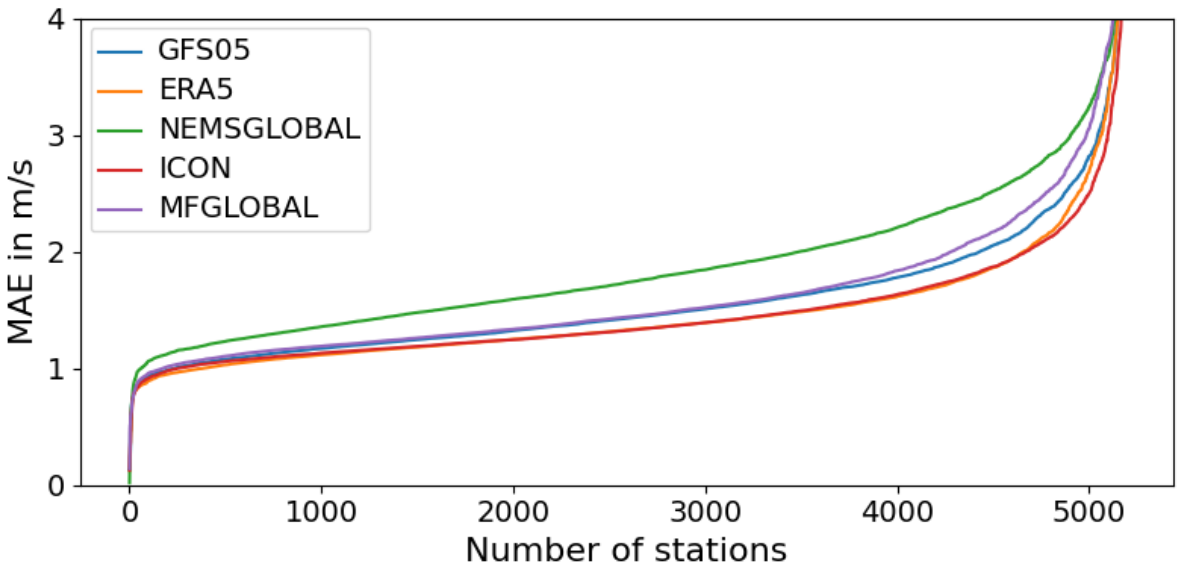
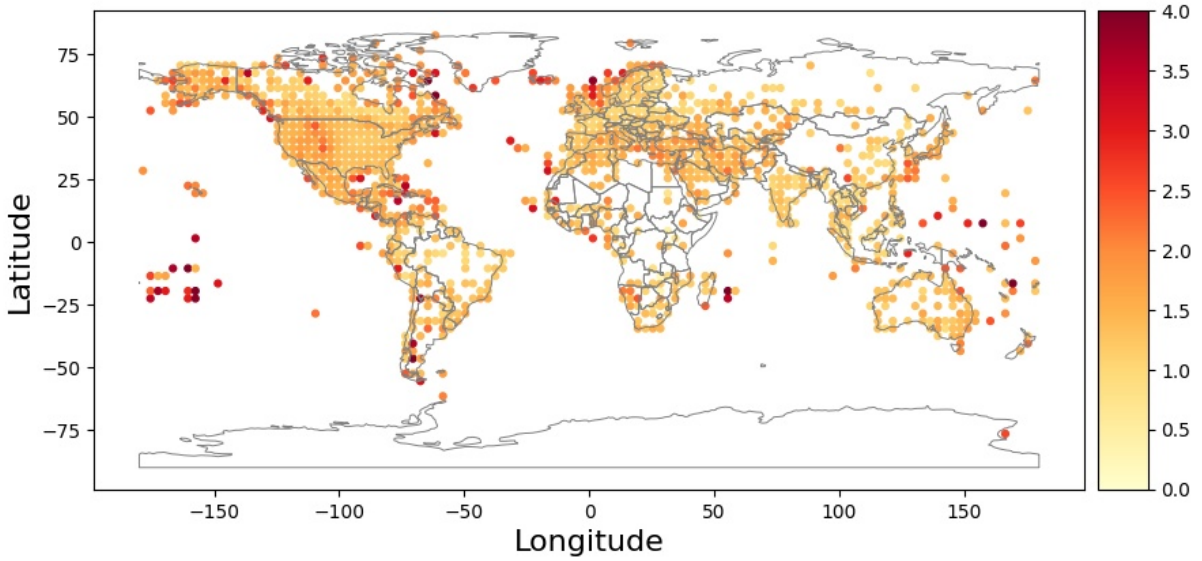


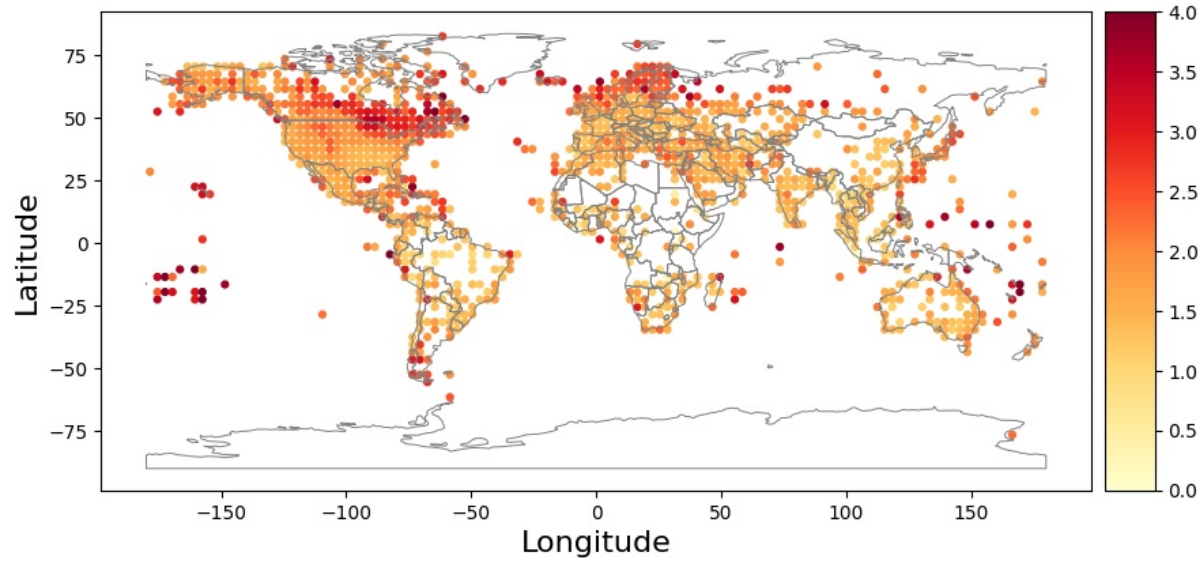
Figure 3: Mean absolute errors for 5180 global measurement stations provided by metar and five models for hourly wind speed in 2020. The MAE values are sorted by size for every model. The y-axis is cut off at 4 $\frac{m}{s}$.

The spatial distribution of the MAEs of the reanalysis model ERA5 and the prediction model NEMSGLOBAL is shown in Figure 4. For both models, values on islands are rather high. While ERA5 shows except that quite evenly distributed MAEs all over the world, NEMSGLOBAL shows high values in Canada and northern Europe. Even though the level of the values differs across the models, the spatial distribution from GFS05, ICON, and MFGLOBAL is quite similar to the ERA5 one. Within USA, all models show slightly higher errors in USA's mountains.

In contrast to the MAE, the MBE shows not only the level of the error but the direction of it. In Figure 5, the distribution of the MBEs for each raw model is shown. While NEMSGLOBAL has a broader distribution with more positive biases, ICON, and ERA5 show higher densities around 0 $\frac{m}{s}$. For all models, there are more big negative values than positive ones. The highest MBE values reach from 3.8 (ICON) to the highest value of 5.7 $\frac{m}{s}$ (NEMSGLOBAL). Lowest numbers reach from -12.8 (NEMSGLOBAL) to -15.2 $\frac{m}{s}$ (MFGLOBAL).



(a) Reanalysis model ERA5



(b) Prediction model NEMSGLOBAL

Figure 4: Mean absolute error in $\frac{m}{s}$ for 5180 global measurement stations of wind speed provided by metar and (4a) the reanalysis model ERA5 or, respectively, (4b) the model NEMSGLOBAL in 2020. In order to avoid overplotting, values are averaged in a 3 x 3 °grid. The color ramp distinguishes values up to 4 $\frac{m}{s}$, after that everything is displayed uniformly.

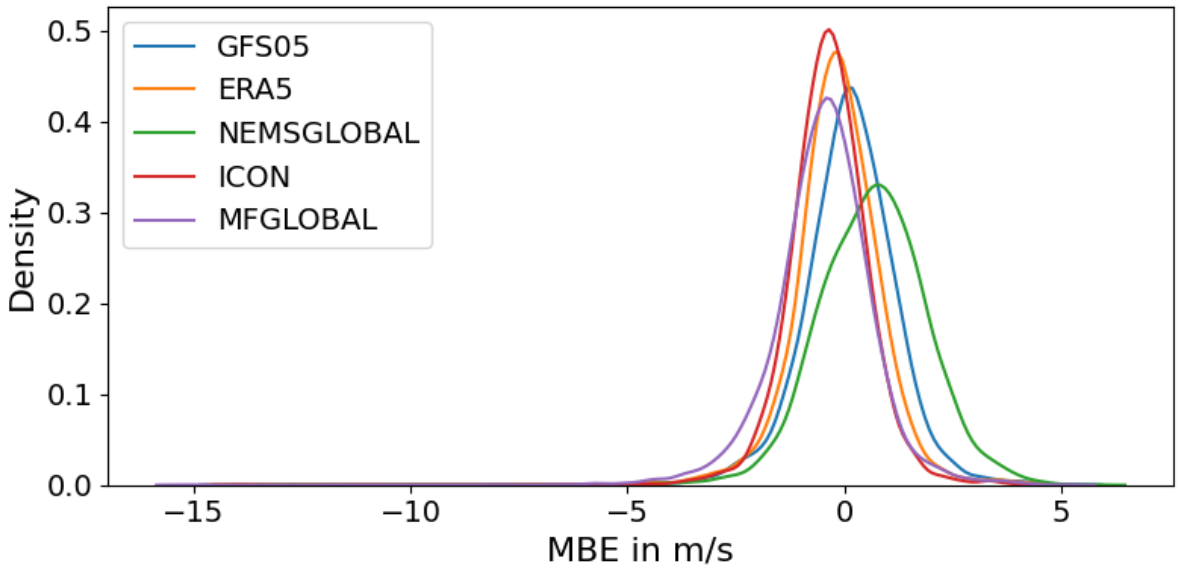
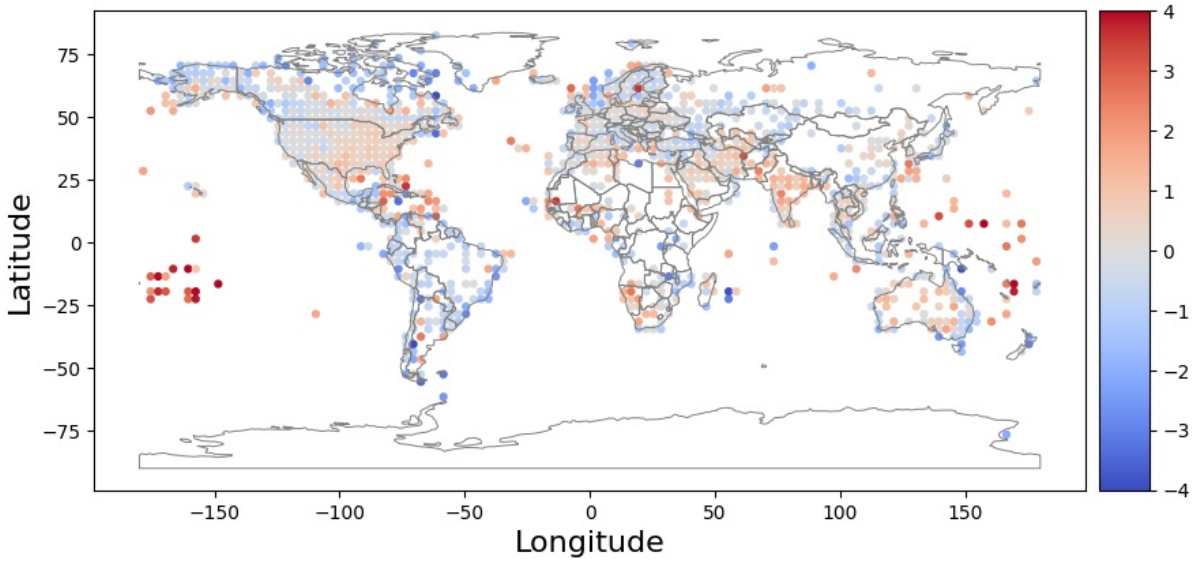
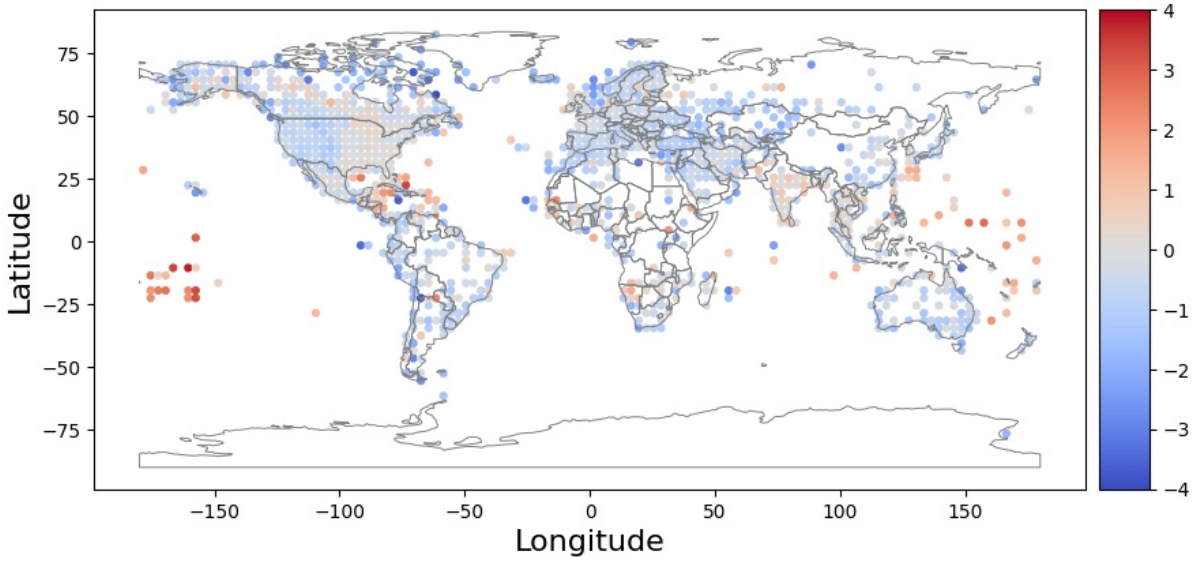


Figure 5: Density plot of MBE in $\frac{m}{s}$ for 5180 global measurement stations provided by metar and the reanalysis model ERA5 and the prediction models GFS05, ICON, MFGLOBAL, and NEMSGLOBAL for hourly wind speed in 2020.

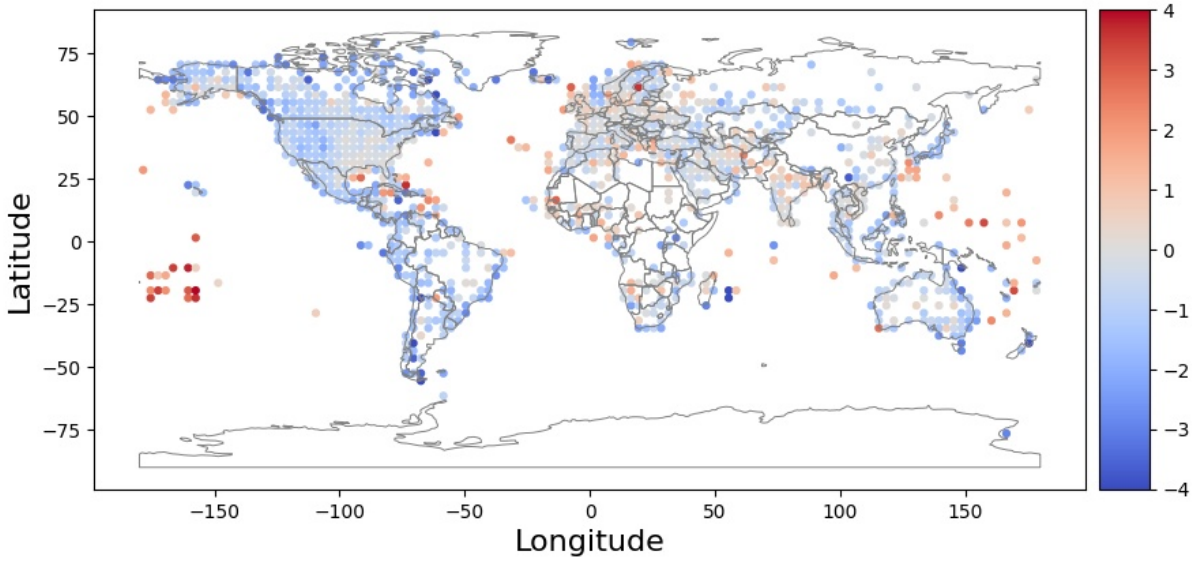
In Figure 6, the spatial distribution of the MBEs of all five models is shown. Here, regions with underpredictions or overpredictions are visible. All five models show a strong overprediction on many islands. For ERA5 and ICON, the spatial distribution is similar and relatively homogeneous. East USA and Canada show low positive MBEs, while the west has low negative values. In Europe, Australia, and South America, more negative values are present than positive ones. Averaged over all stations, ERA5 has a MBE of -0.152 and ICON -0.355 $\frac{m}{s}$. MFGLOBAL shows more negative values in North America and more positive values in Europe. In Australia, there are more positive values. Generally, many stations show very high errors for MFGLOBAL. The mean MBE of all stations is -0.490 $\frac{m}{s}$. For GFS05, a great part of the USA, Canada, Asia, and Australia shows an overprediction. The effect of these positive values can also be seen in the average MBE over all stations, which is 0.130 $\frac{m}{s}$. With that, GFS05 shows the average of MBE nearest null compared to other models. The highest MBE overall station has NEMSGLOBAL with 0.657 $\frac{m}{s}$. NEMSGLOBAL has a high overprediction in Canada, East USA, and North Europe.



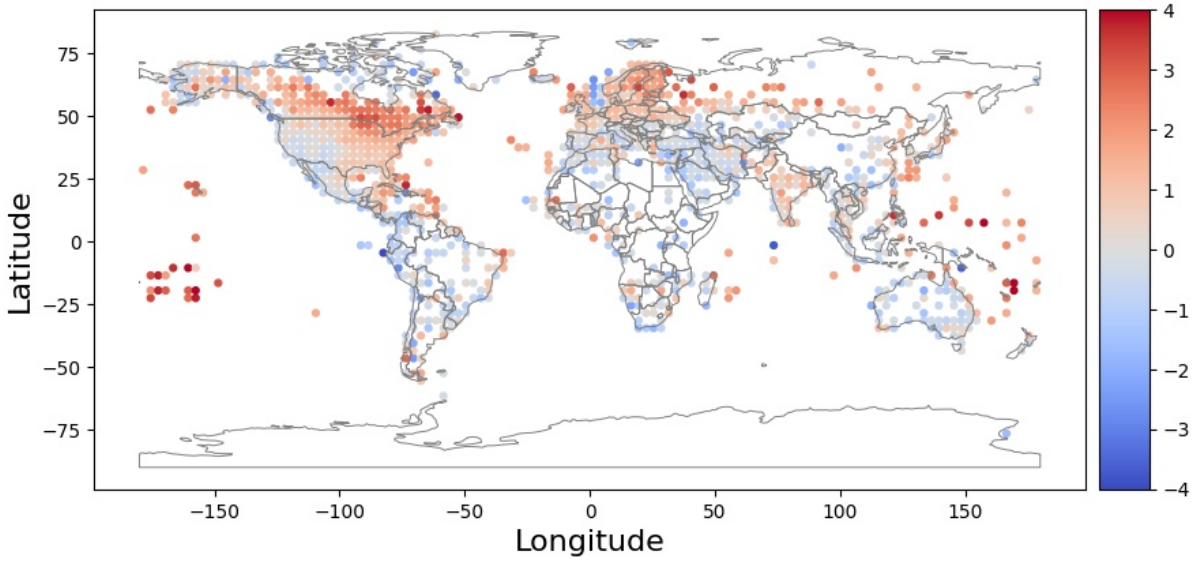
(a) Prediction model GFS05



(b) Prediction model ICON



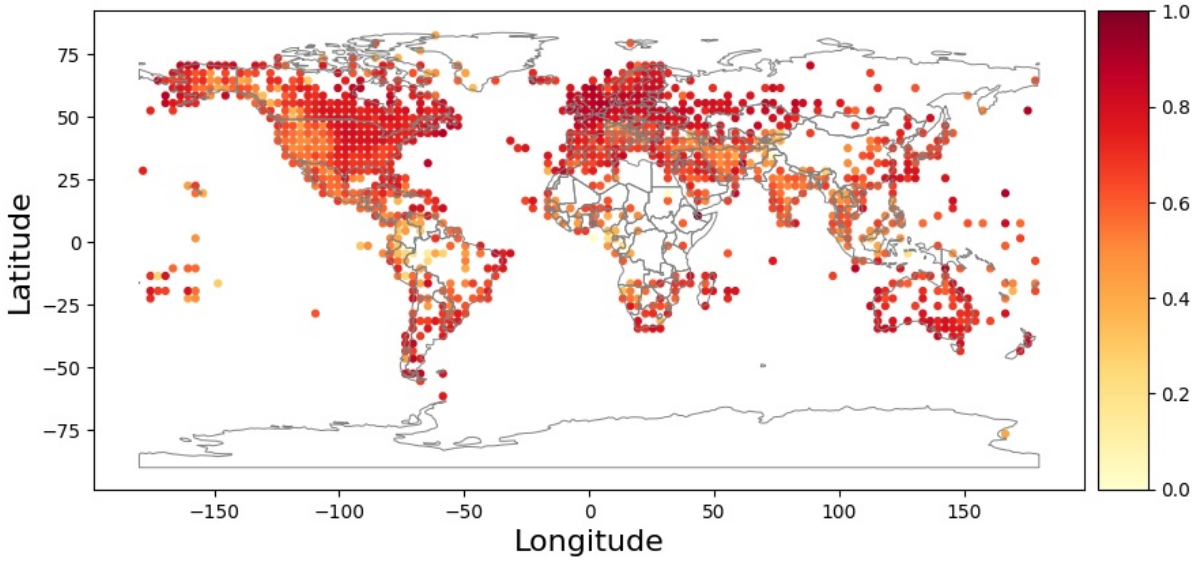
(c) Prediction model MFGLOBAL



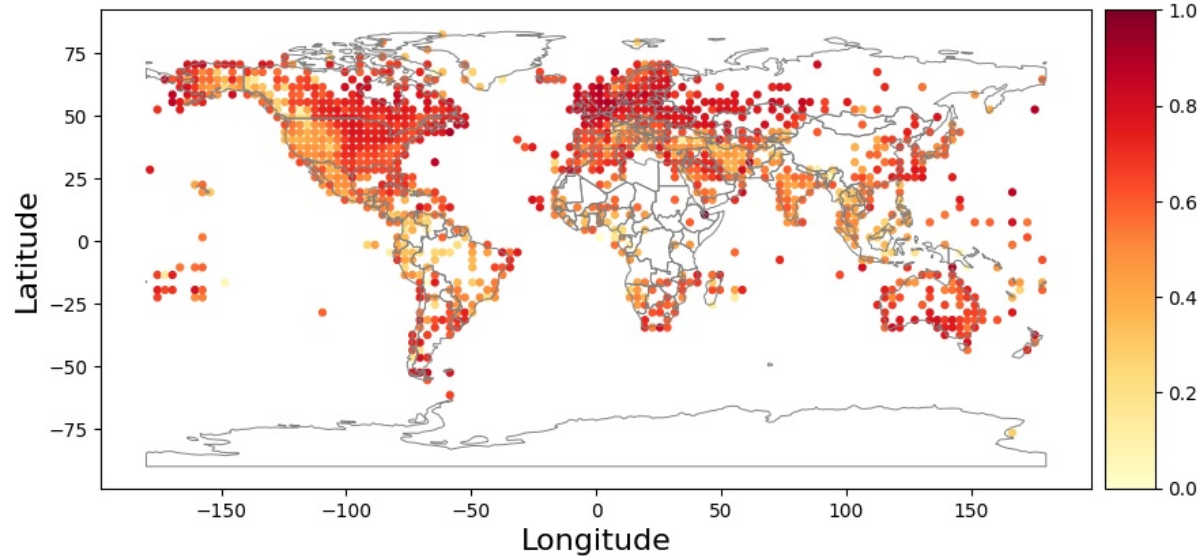
(d) Prediction model NEMSGLOBAL

Figure 6: Mean bias error $\frac{m}{s}$ for 5180 global measurement stations of wind speed provided by metar and the prediction models (6a) GFS05, (6b) ICON, (6c) MFGLOBAL or (6d) NEMSGLOBAL in 2020. Red values show an overprediction of the model, blue values an underprediction. In order to avoid overplotting, values are averaged in a 3 x 3 °grid.

For the RMSE, the lowest value overall stations has ERA5 with 1.860 $\frac{m}{s}$, following by ICON, GFS05, MFGLOBAL and last NEMSGLOBAL with 2.421 $\frac{m}{s}$. All values can be seen in Table 1. The spatial distributions of the RMSE are similar to those of the MAE with high values on islands, over-average values in the USA's mountains for all models, and for NEMSGLOBAL high values in Canada and northern Europe.
MAPE values are for all models relatively homogeneous with two exceptions. First, island stations have often very high MAPEs across all models. Second, NEMSGLOBAL has high values in southern Canada and northern Europe. Mean MAPEs reach from 42.2 % (ICON) to 60.7 % (NEMSGLOBAL).
For ICON and MFGLOBAL, the spatial distribution of the Spearman correlation between the model predictions and the measurements can be seen in Figure 7. Both show very high correlations in Europe and in eastern North America and lower values in Asia and western North America and northern South America. However, ICON has better correlations in areas harder to predict than MFGLOBAL. ERA5 and GFS05 have a similar spatial distribution as ICON, while NEMSGLOBAL and MFGLOBAL are more alike. The mean correlations over all stations reach from NEMSGLOBAL with 0.596 to MFGLOBAL, GFS05, ERA5 to a value of 0.682 from ICON (see Table 1).
Besides the spearman correlation for all values, the correlation for only a part of the values is calculated, too. Here, only those measurement and model pairs are used, where the measurement value is over 3 $\frac{m}{s}$. Generally, the correlation with values over 3 $\frac{m}{s}$ is lower than the correlation with all values. Averaged over all models, the difference is 0.094. The model with the lowest loss is ERA5 with 0.075, the highest one is ICON with 0.105. The spatial pattern of the difference is shown in Figure 8, which is relatively similar for all models. As can be seen, almost in every part of the world, correlation is lower for the tailored values than for the complete value pairs. An exception forms the mountain area of the USA and some seemingly random single stations, where correlation does not change much or even gets a bit higher.



(a) Prediction model ICON



(b) Prediction model MFGLOBAL

Figure 7: Spearman correlation for 5180 global measurement stations of wind speed provided by metar and (7a) the prediction model ICON and (7b) MFGLOBAL in 2020. Darker colors show a higher correlation. In order to avoid overplotting, values are averaged in a 3 x 3 °grid.

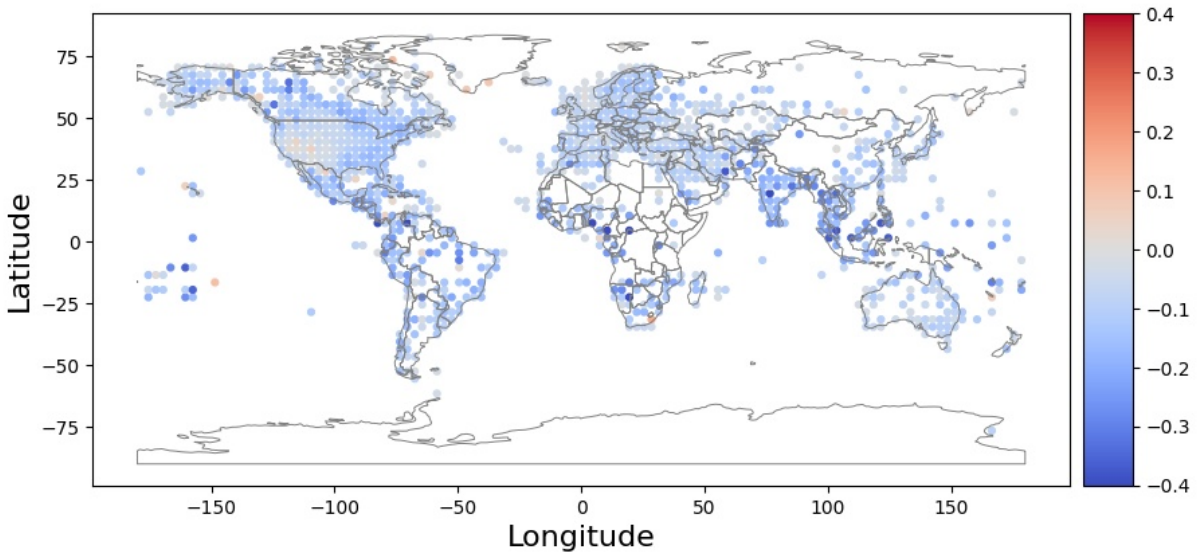


Figure 8: Difference between the spearman correlation with measurement values over 3 $\frac{m}{s}$ and the spearman correlation with all values for 5180 global measurement stations of wind speed provided by metar and the prediction model ICON. Red values indicate a higher correlation for the dataset without values below 3 $\frac{m}{s}$ than for the dataset with all values; blue values show a lower correlation. In order to avoid overplotting, values are averaged in a 3 x 3 °grid.

Measurement values over 5 $\frac{m}{s}$ are detected from NEMSGLOBAL with the highest probability of 63.6 %, following from GFS05 with 55.1 %, ERA5 with 47.9 %, ICON with 46.6 % and MFGLOBAL with 37.0 %. For PODs over 15 $\frac{m}{s}$, all values drop strongly. Here, NEMSGLOBAL has the highest value with 20.5 %, followed by GFS05 with 11.6 % - other models show PODs below 7.0 %. The highest POD for measurement values over 20 $\frac{m}{s}$ is shown by NEMSGLOBAL with 8.2 %, while ICON and ERA5 detect the fewest values with 2.6 and 2.1 %. Measurements over 30 $\frac{m}{s}$ were almost never detected in any model - the highest value is 0.3 % by GFS05.
On the other hand, FAR over 5 $\frac{m}{s}$ is highest for NEMSGLOBAL with 45.6 %, too. GFS05 is following with 36.8 %, ICON with 30.2 %, MFGLOBAL with 28.6 % and ERA5 with 27.6 %. For measurements over 15 $\frac{m}{s}$, all values increases. Here, ERA5 shows the lowest FAR with 50.3 %, following by 69.1 % by MFGLOBAL, 74.0 % from GFS05, ICON with 90.3 % and NEMSGLOBAL with 90.7 %. The FAR from measurement values over 20 $\frac{m}{s}$ does not change much to the FAR over 15 $\frac{m}{s}$. For measurements over 30 $\frac{m}{s}$, all FAR values are over 90 %, with the lowest value of 93.4 % from GFS05.
Last, the highest HSS over 5 $\frac{m}{s}$ shows ERA5 with 0.419, closely followed by ICON and GFS05. NEMSGLOBAl has 0.381 and MFGLOBAL shows the worst HSS with 0.310.

For measurements over 15 $\frac{m}{s}$, the HSS decreased strongly for all models. The highest value here is 0.08 for GFS05, the lowest value is 0.037 from ICON. For HSS over 20 $\frac{m}{s}$, the maximum has GFS05 with 0.035. The HSS values of all models can be seen in Table 2.

6.1.2 Regional models in Europe compared to global models

The comparison of regional models with global models is done for a european subset of the global measurement data because the regional models NEMS4 and NEMS12 were only available in Europe for this work. The sorted MAE between the measurements and the respective models is shown in Figure 9. For all models, over 97 % of the MAEs lie between 0.7 and 2.5 $\frac{m}{s}$. NEMSGLOBAL, NEMS4, and NEMS12 have a similar course of sorted MAEs and show the high MAEs compared with other models. The average MAE over all europe stations is lowest for ERA5 with 1.306 $\frac{m}{s}$, following from ICON with 1.333 $\frac{m}{s}$, GFS05 with 1.384 $\frac{m}{s}$, MFGLOBAL with 1.438 $\frac{m}{s}$, NEMS4 with 1.514 $\frac{m}{s}$, NEMS12 with 1.667 $\frac{m}{s}$ and the highest value 1.691 $\frac{m}{s}$ from model NEMSGLOBAL. Lowest MAE values for one station reach from 0.735 (ERA5) to 0.964 $\frac{m}{s}$ (NEMS4), highest values reach from 3.939 (NEMS4) to 4.828 $\frac{m}{s}$(ERA5).

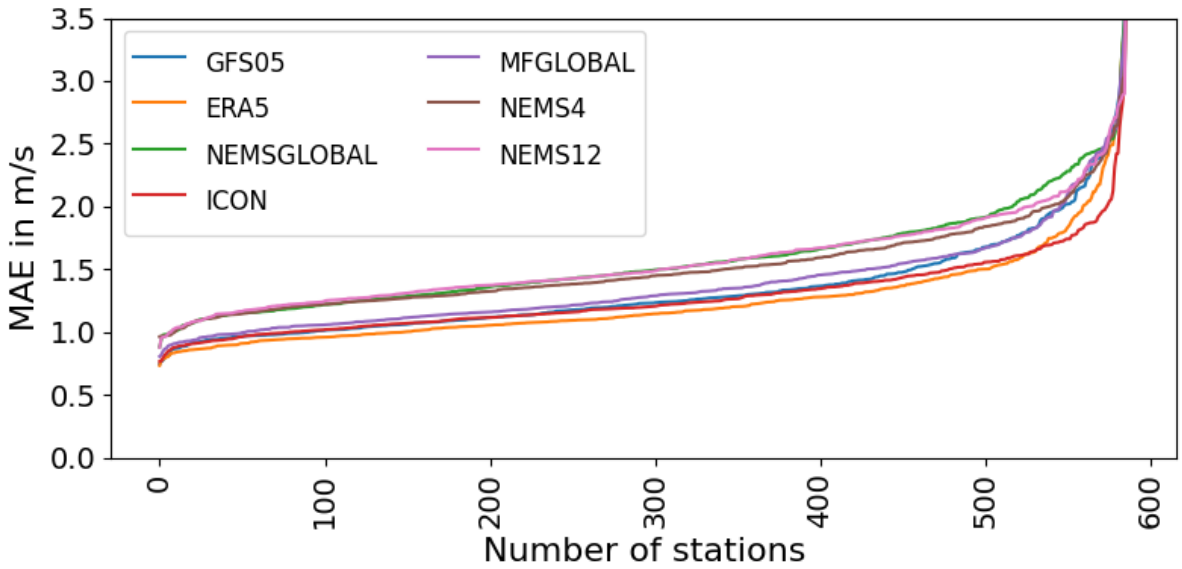


Figure 9: Sorted mean absolute errors $\frac{m}{s}$ for 587 measurement stations in Europe provided by metar and seven models for hourly wind speed in 2020. NEMS4 and NEMS12 are regional models only available in Europe. ERA5 is a global reanalysis model. The remaining are global prediction models.

In Figure 10, the density of MBE values for european stations for all models is shown. The three NEMS models show a shift of the values into the positive range, while MFGLOBAL and GFS05 are quite centered and ERA5 and ICON are shifted to negative values. This fits to the averaged values over all stations: all NEMS-models have positive mean MBEs with 0.585 (NEMSGLOBAL), 0.791 $\frac{m}{s}$ (NEMS4) and 0.838 $\frac{m}{s}$ (NEMS12). All other stations show negative averaged MBEs - MFGLOBAL and GFS05 near null (-0.011 and -0.078 $\frac{m}{s}$) and ERA5 and ICON with -0.346 and -0.529 $\frac{m}{s}$. This is also reflected in the proportion of MBEs above zero: For NEMSGLOBAL, 73.3 % of the stations showed a positive MBE, for NEMS12 89,8 % and for NEMS4 even 93.3 %. For MFGLOBAL and GFS05 it is relatively equally above and below zero: MFGLOBAL has 54.5 % over zero, GFS05 47.7 %. ERA5 has about one third in the positive area (33.2 %) and last, ICON shows only 21.8 % of MBE values in the positive spectrum.

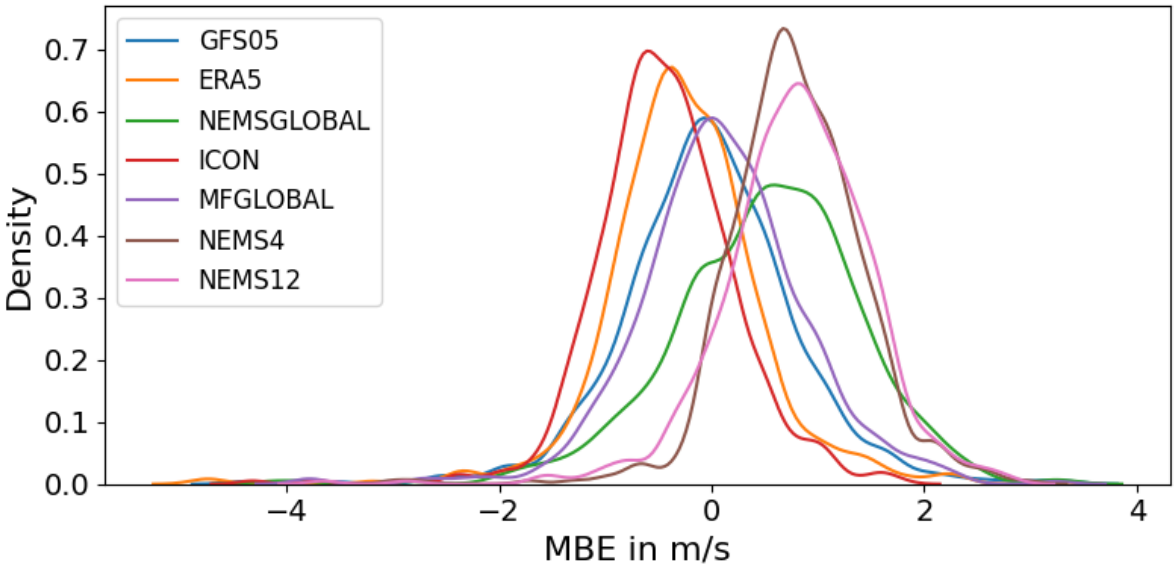


Figure 10: Density plot of mean bias error in $\frac{m}{s}$ for 587 measurement stations in Europe provided by metar and seven models for hourly wind speed in 2020. NEMS4 and NEMS12 are regional models only available in Europe. ERA5 is a global reanalysis model. The remaining are global prediction models.

The spatial distribution of MBEs is visualized in Figure 11 for the model NEMSGLOBAL. As can be seen, values in the northern part of Europe like Germany, Great Britain, Pole, and North France are higher than MBE values in the South like Italia, Romania, Spain, and North France. Besides, a slight West-East pattern can be seen: In the East, values are lower than in the West. This pattern is particularly clear at NEMSGLOBAL, however, all models tend to have a North-South shift in MBEs in Europe with an overprediction in North and an underprediction in South.
The lowest MBE minimum for one station is -4.803 (ERA5), the highest minimum is -3.001 $\frac{m}{s}$ (NEMS4). For the highest value, the differences are bigger: 1.755 $\frac{m}{s}$ has ICON, reaching to 3.326 $\frac{m}{s}$ from GFS05.

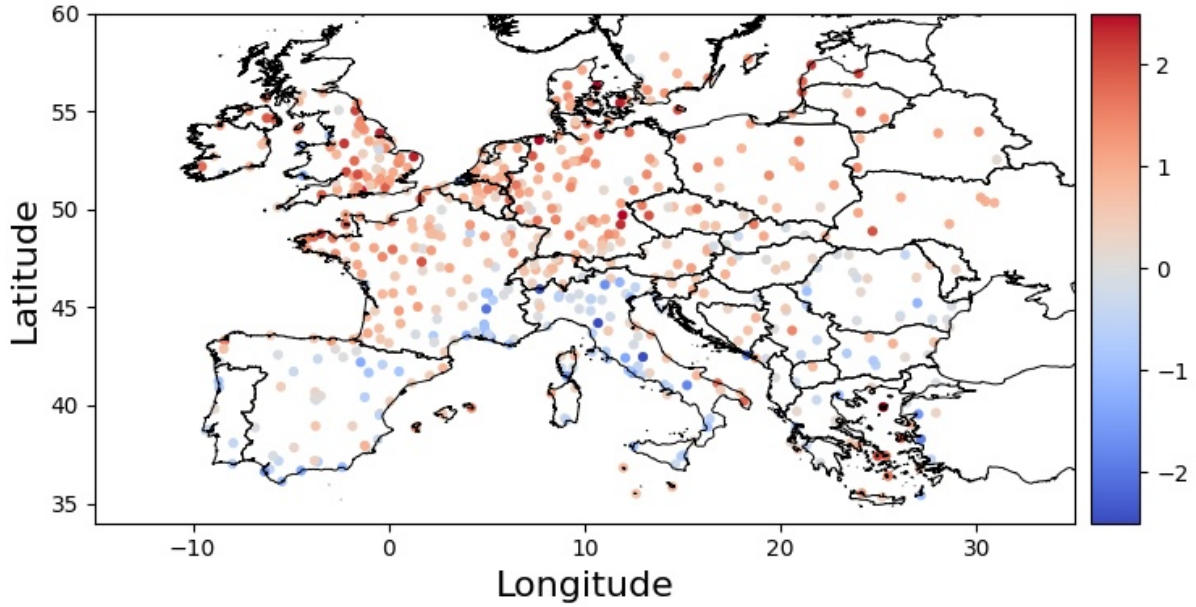


Figure 11: Mean biased error in $\frac{m}{s}$ for 587 measurement stations in Europe provided by metar and the prediction model NEMSGLOBAL for hourly wind speed in 2020.

For NEMSGLOBAL, the distribution of the spearman correlation for all values can be seen in Figure 12. Here, southern parts of Europe like Italia, Southern Swiss, and Bulgaria show much weaker correlations than North Europe like Great Britain, Germany, and North France. For all other models, this pattern is visible in similar strength and shape. For the averaged correlation, NEMSGLOBAL shows the lowest value with 0.644, followed by MFGLOBAL with 0.666, NEMS12 with 0.674, GFS05 with 0.683, ERA5 with 0.697, NEMS4 with 0.703, and the highest correlation of 0.731 from ICON. For all models, there

is a station with a correlation of almost zero (-0.051 for MFGLOBAL to 0.118 for ICON). Looking at the station with the highest correlation among all stations in a model, ERA5 features the highest correlation with 0.926 and NEMS12 shows the lowest correlation with 0.879. The correlation with measurement values over 3 $\frac{m}{s}$ is spatially very similar to the correlation with all values. Here, the weakest model is NEMSGLOBAL with 0.572, followed by MFGLOBAL with 0.593, GFS05 with 0.604, NEMS12 with 0.607, NEMS4 with 0.642, ERA5 with 0.643, and ICON with 0.645. The pattern of the correlations shows parallels to that of the MBE.

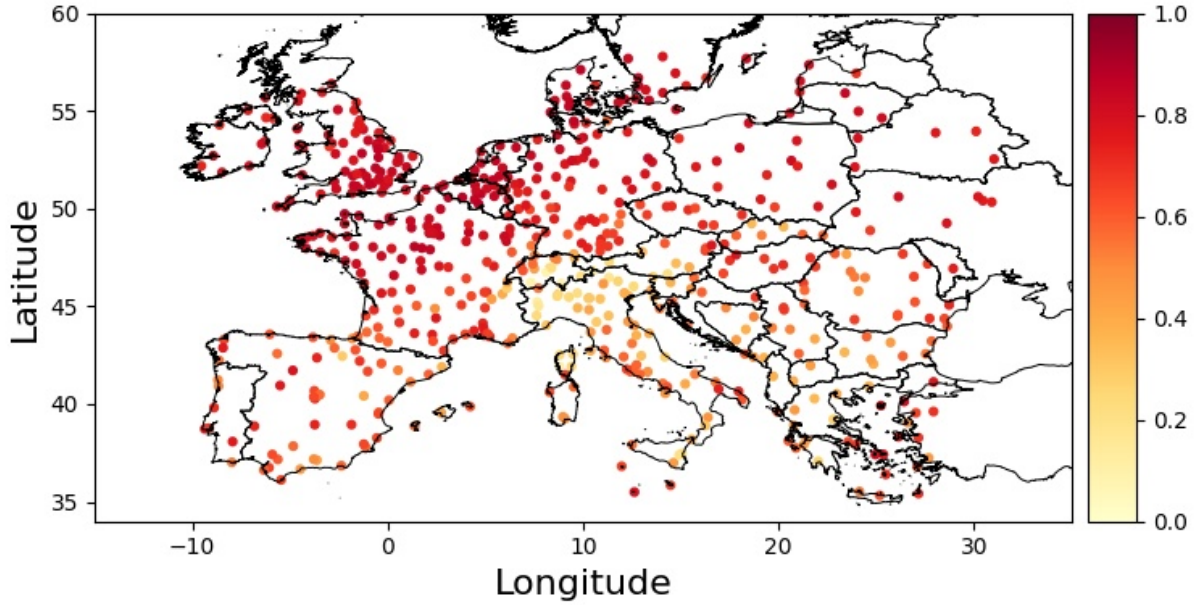


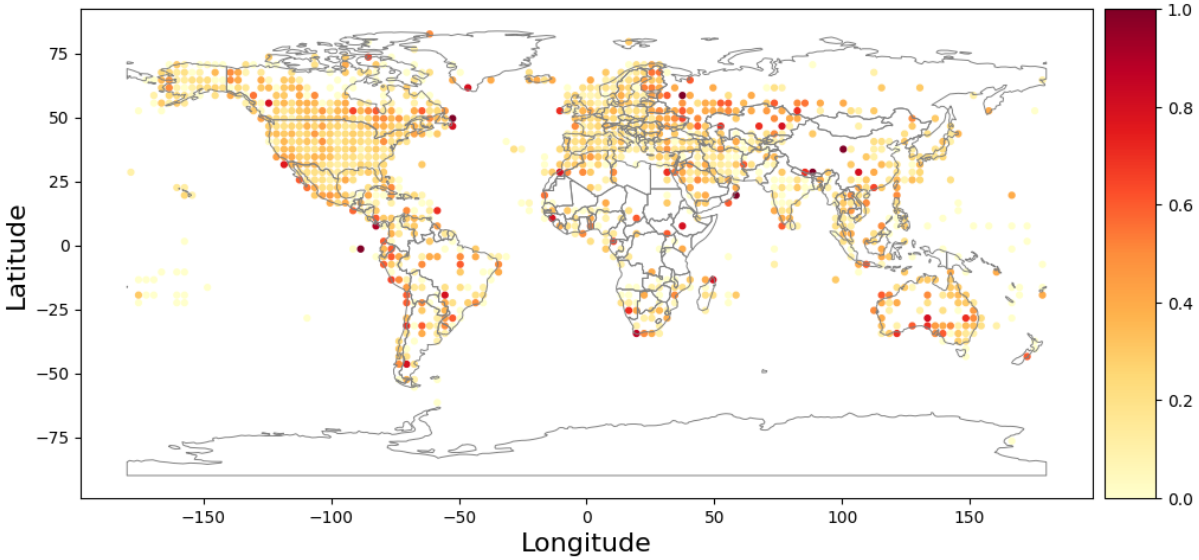
Figure 12: Spearman correlation for 587 measurement stations in Europe provided by metar and the prediction model NEMSGLOBAL. Dark colors show a higher correlation.

The highest probability to detect a measurement value of over 5 $\frac{m}{s}$ has NEMS4 with 80.9 %, followed by NEMS12 with 77.3 % and NEMSGLOBAL with 67.7 %. The remaining four global models show values below 60 %. For values over 15 $\frac{m}{s}$, NEMS4 shows the highest percentage with 34.2 % which goes down to 5.3 % for ERA5. POD over 20 $\frac{m}{s}$ is 14.1 % for NEMGLOBAL, while ICON and ERA5 do not reach 2 %.
The lowest FAR for measurements over 5 $\frac{m}{s}$ has ERA5 with 20.7 %, followed by ICON with 21.3 % and climbs up to 42.1 % for NEMS12. These values increases for measurements over 15 $\frac{m}{s}$ up to the lowest 43 % for ERA5, which performs much better here than the other models - the second best value has GFS05 with 72.6 %, the worst value has ICON with 93.9 %. For FAR for over 20 $\frac{m}{s}$, the best model ERA5 shows 83.8 % and goes

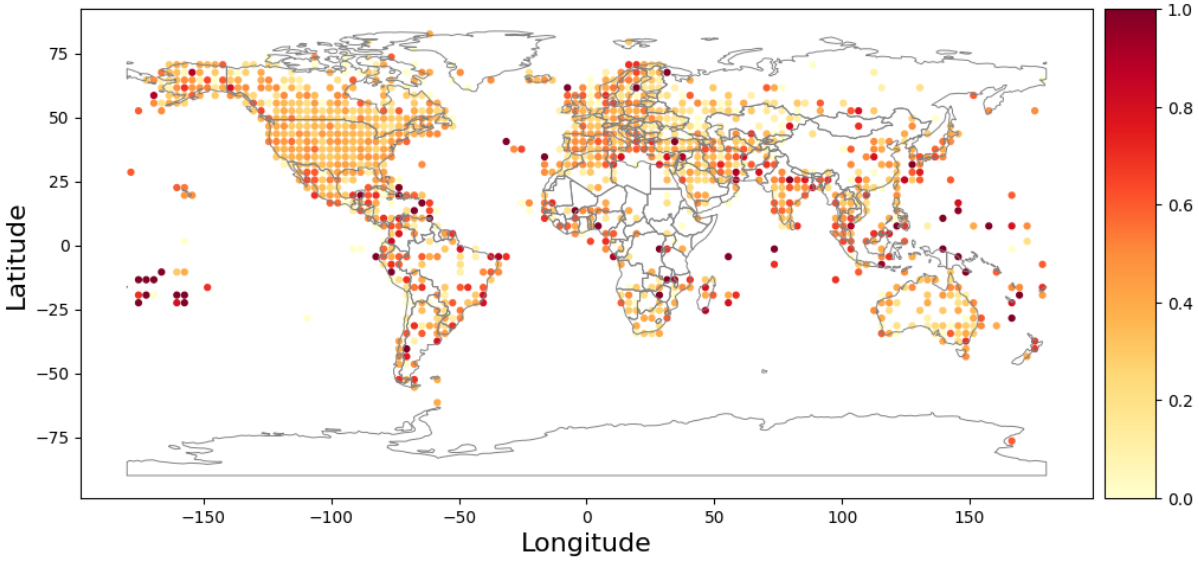
up to 96.5 for NEMS12.
Last, the highest HSS for measurement values over 5 $\frac{m}{s}$ has NEMS4 with 0.534, followed by NEMS12 with 0.495 and ICON with 0.488. The worst HSS shows MFGLOBAL with 0.47, which is really near to the other models. For measurements over 15 $\frac{m}{s}$, best HSS shows NEMS4, again, with 0.131. ICON has with 0.037 the lowest value. For measurements over 20 $\frac{m}{s}$, all values are near to zero with a maximum value from MFGLOBAL with 0.042.

6.2 Multimodels

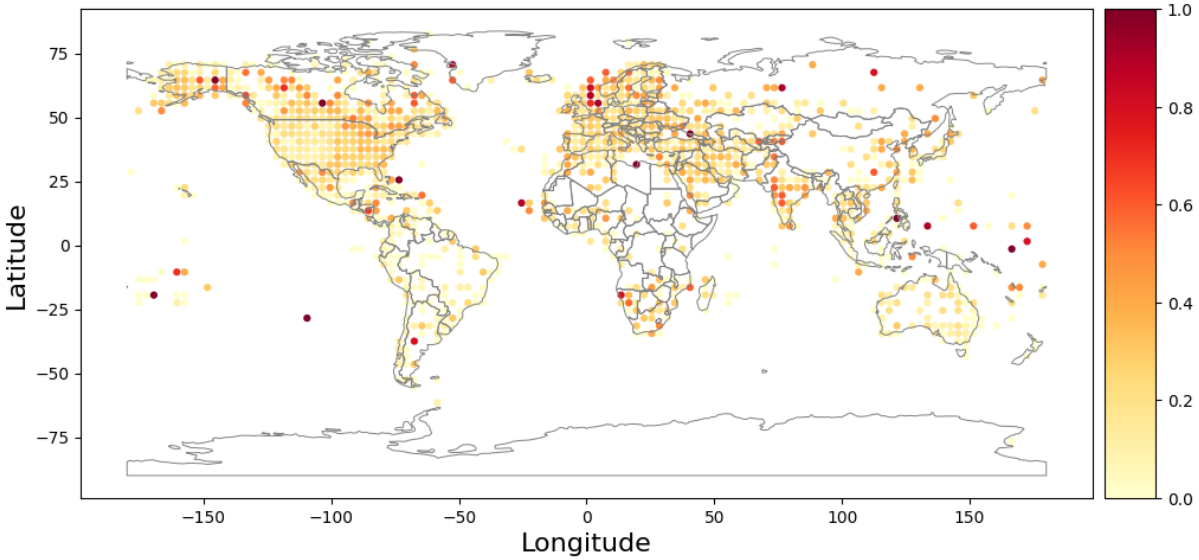
For the multimodel approaches, the four global prediction models GFS05, ICON, MFGLOBAL, and NEMSGLOBAL were weighted in order to minimize the MAE. For the first approach, the best weighting for each station is calculated. The spatial pattern of the station-based weightings of the four models for 2020 can be seen in Figure 13. ICON shows high values in almost every part of the world except Russia and North Canada. Averaged over all stations, ICON's weighting is 38.9 % (48.2 % in 2019) and is with that the model with the highest impact in this approach. The model with the following impact is GFS05 with averaged 23.9 % (21.0 %) of the weight. GFS05 is especially strong in South-West Russia, Kazakhstan, and parts of Australia. MFGLOBAL has 21.8 % (16.5 %) of the averaged weights and has strong areas in North Europe, middle Canada, and East USA. Last, NEMSGLOBAL takes up for 15.4 % (14.2 %), being strong in North Canada.
The second approach is a fixed weight for all stations, where the weight is calculated by the minimized average MAE over all stations. Here, the optimum is found for 2020 (2019) in 50 % (60 %) ICON, 30 % (20 %) GFS05, 10 % (10 %) MFGLOBAL and 10 % (10 %) NEMSGLOBAL.



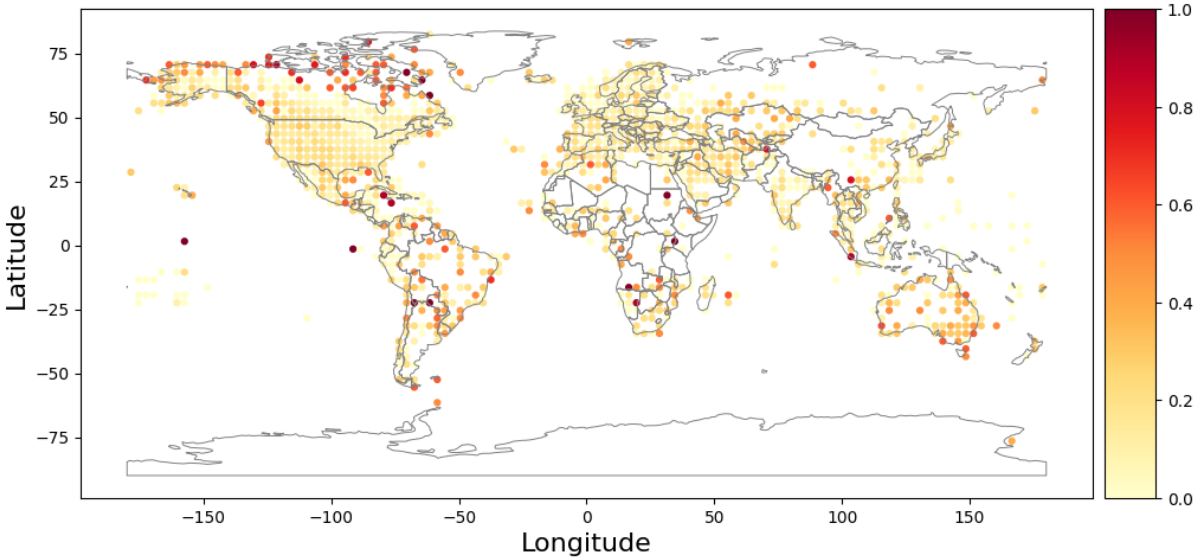
(a) Weights of prediction model GFS05



(b) Weights of prediction model ICON



(c) Weights of prediction model MFGLOBAL



(d) Weights of prediction model NEMSGLOBAL

Figure 13: Model weights for the multimodels with the lowest MAE for the individual stations. Multimodels are composed of the prediction models (6a) GFS05, (6b) ICON, (6c) MFGLOBAL or (6d) NEMSGLOBAL. The MAE is calculated for 5180 global measurement stations of wind speed provided by metar in 2020. Darker colors show a higher weighting of the respective model. In order to avoid overplotting, values are averaged in a 3 x 3 °grid.

Table 1: Mean error metrics for all raw models, a multimodel approach with one best weight for all stations and a multimodel approach with individual weighting per station. Errors are calculated for hourly measurement data from for 5180 global stations in 2020.

Domain	**MAE**	**MBE**	**RMSE**	**MAPE**	**Cor**	**Cor>3** $\frac{m}{s}$
NEMSGLOBAL	1.866	0.657	2.421	0.607	0.596	0.501
MFGLOBAL	1.617	-0.490	2.070	0.474	0.612	0.514
GFS05	1.563	0.130	2.002	0.483	0.635	0.537
ERA5	1.460	-0.152	1.860	0.429	0.652	0.577
ICON	1.453	-0.355	1.925	0.422	0.682	0.577
MM Global Weight	1.369	-0.122	1.768	0.401	0.703	0.598
MM Station-Based	1.259	-0.113	1.644	0.371	0.701	0.596

Because it was optimized to a low MAE, the mean MAE over all stations of the multimodel approaches is lower than all raw models. For the station-based approach, MAE is 1.249 $\frac{m}{s}$, what is 0.194 $\frac{m}{s}$ less than the best raw model ICON. The multimodel with the fixed global weight shows with an MAE of 1.369 $\frac{m}{s}$ worse results than the station-based approach, but better than all raw models.

However, other error metrics get better by multimodelling, too. The averaged MBE is -0.113 $\frac{m}{s}$ for the station-based approach and -0.122 $\frac{m}{s}$ for the global weight. The raw model with the MBE nearest null is GFS05 and has 0.130 $\frac{m}{s}$. The RMSE and the MAPE show their lowest values for the station-based multimodel, too. And the correlation reaches 0.701 for all values and 0.596 for measurements over 3 $\frac{m}{s}$ for the station-based approach. Here, the global weight approach is slightly better with 0.703 and 0.598. All four values are higher than the corresponding correlations of the raw models. All values can be seen in Table 1.

For the probability of detection over 5 $\frac{m}{s}$, the global weight multimodel has a percentage of 51.2 % and the station-based multimodel approach has a value of 50.5 %. With that, they are worse than NEMSGLOBAL (63.5 %) and GFS05 (55.1 %), but better than ERA5 (47.9 %), ICON (46.6 %) and MFGLOBAL (37.0 %). The POD decrease for measurements over 15 $\frac{m}{s}$ to 7.3 % for station-based multimodels and 5.9 % for global weight multimodels. Here, MFGLOBAL lies between them with 6.9 % - ERA5 and ICON show lower values, NEMSGLOBAL and GFS05 have a higher probability. For POD for measurements over 20 $\frac{m}{s}$, the station-based approach shows 3 % and the global weight approach 2.7 %. Now, MFGLOBAL has a higher probability than both with 4 % next to NEMSGLOBAL and GFS05. ERA5 and ICON detect values over 20 $\frac{m}{s}$ less often. Values over 30 $\frac{m}{s}$ are never detected from the multimodels.

Table 2: Mean Heidke-Skill-Score for all raw models, a multimodel approach with a global weighting for all stations and a multimodel approach with individual weighting per station, both combined out of four forecast models. The HSS is calculated for measurement data from for 5180 global stations in 2020.

Domain	**>5** $\frac{m}{s}$	**>15** $\frac{m}{s}$	**>20** $\frac{m}{s}$	**>30** $\frac{m}{s}$
NEMSGLOBAL	0.381	0.063	0.023	0
MFGLOBAL	0.310	0.047	0.028	0.001
GFS05	0.412	0.080	0.035	0.001
ERA5	0.419	0.039	0.019	0
ICON	0.413	0.037	0.017	0
MM Global Weight	0.444	0.056	0.023	0
MM Station-Based	0.467	0.080	0.028	0

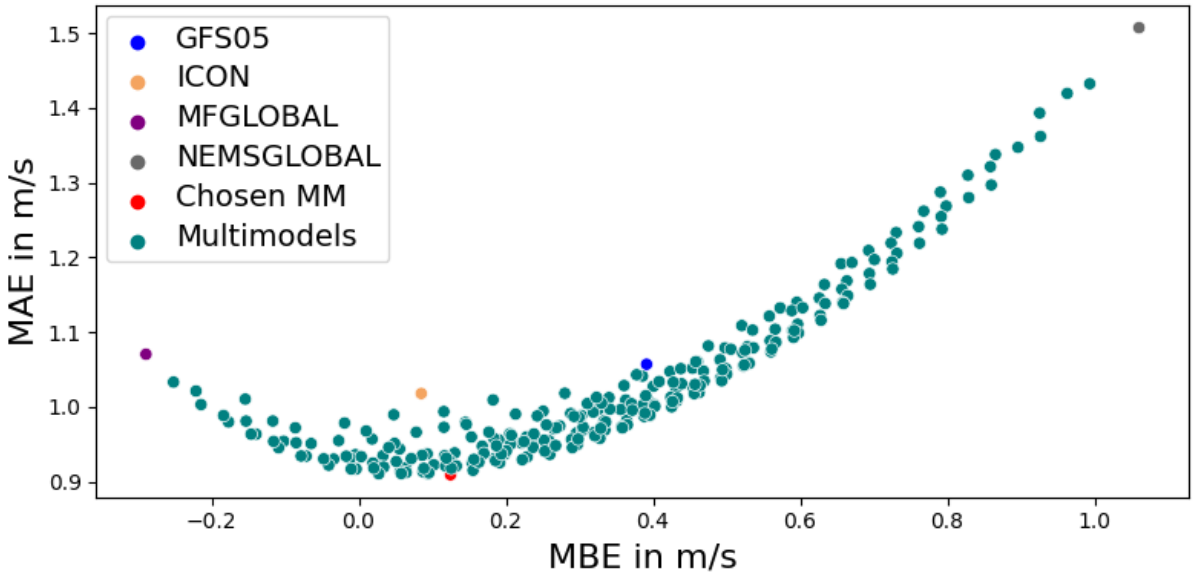
The FAR for the station-based multimodel for measurements over 5 $\frac{m}{s}$ is 27.4 % and therefore better than all other tested models. The probability of the global weight approach is with 29 % a bit higher. NEMSGLOBAL (45.6 %), GFS05 (36.8 %) and ICON (30.2 %) show higher values than both multimodel approaches, MFGLOBAL (28.6 %) and ERA5 (27.6 %) are better than the global weight approach. For values over 15 $\frac{m}{s}$, the station-based multimodel mean increases to 55.7 %, the global weight mean to 59.0 %. Here, ERA5 shows a lower value with 50.3 %. All other models have worse values reaching from 69.1 % from MFGLOBAL to 90.7 % from NEMSGLOBAL. For values over 20 $\frac{m}{s}$, FAR decreases for the multimodels to 54.5 % for the station-based multimodel and 53.2 % for the global weight. ERA5 still is below these numbers with 50.3 %, other models are higher with 67.4 % (MFGLOBAL) to 94.9 % (NEMSGLOBAL). The FAR over 30 $\frac{m}{s}$ is 100 % for both multimodel approaches, like for ERA5, ICON, and NEMSGLOBAL. GFS05 (93.4 %) and MFGLOBAL (97.2 %) are the only two models which correctly predict values over 30 $\frac{m}{s}$, even if they mostly dismiss them, too. The HSS over 5 $\frac{m}{s}$ for the multimodels is higher than for all raw models. The station-based multimodel has 0.467, the global weight multimodel has 0.444. The next best model is the reanalysis model ERA5 with 0.419, next to ICON with 0.413. For measurements over 15 $\frac{m}{s}$, the HSS drop to 0.080 for the station-based approach, which shares with GFS05 the highest value. The global multimodel has 0.056, which lies also below NEMSGLOBAL 0.063, but above ERA5, ICON and MFGLOBAL. For measurements over 20 $\frac{m}{s}$, the multimodels show with 0.028 for the station-based approach and 0.023 for the global weight medium values: GFS05 and MFGLOBAL is better, NEMSGLOBAL on the same level as the global weight multimodel and ICON and ERA5 are worse. For values over 30 $\frac{m}{s}$, the multimodels show a HSS of zero. The HSS values of all models can be seen in Table 2.

Table 3: Best model combinations for a varying number of included forecast models. Combinations are optimized as a global weighting for all 5180 measurement stations. With increasing amount of used models, MAE decreases. MFG is abbreviated for MFGLOBAL, NEMSG for NEMSGLOBAL.

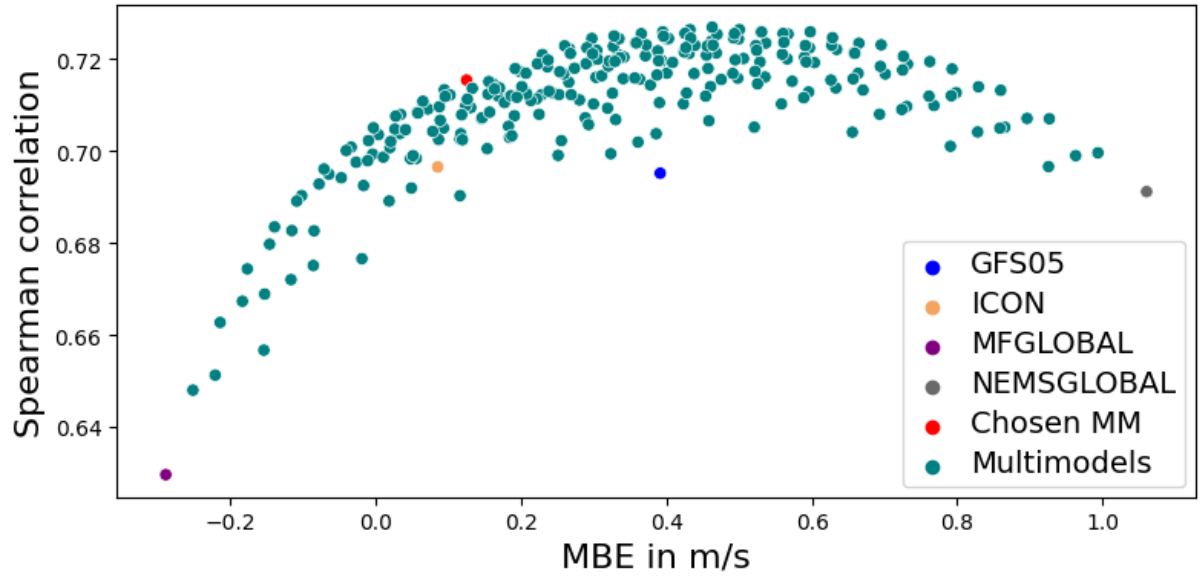
Model amount	MAE	ICON	GFS05	MFG	NEMSG
1	1.453	100 %			
2	1.386	60 %	40 %		
3	1.376	50 %	30 %	20 %	
4	1.369	50 %	30 %	10 %	10 %

In order to evaluate the gain by additional models in the global weight multimodel approach, the averaged MAE for the best raw model as well as the best combinations for two, three or all four global forecast models can be seen in Table 3. MAE is decreasing with an increase in the number of model combinations. If only three of the four models are chosen, abandoning NEMSGLOBAL results in the smallest loss of MAE. If only two models are combined, weighting ICON and GFS05 give the best result. The reduction of the MAE is from the best raw model to a combination of two models is 0.067 $\frac{m}{s}$, while the increase to three models can decrease the MAE only by 0.010 $\frac{m}{s}$. Adding a fourth model, 0.007 $\frac{m}{s}$ can be further reduced.

In order to evaluate the error metrics for all weighting combinations within one observation station, a station was selected randomly for further analysis. The chosen measurement station is located in China. The MAEs, MBEs, and Spearman correlation is shown for all 256 tested weightings in Figure 14. In Figure 14a, the MAEs and MBEs are plotted to each other. Because the chosen multimodel for each station is the one with the lowest MAE, it is at the bottom of the graph. However, this is not the optimum for the MBE. As can be seen, the MBE is relatively low, but there are a couple of possible multimodels with lower MBE values. Besides, ICON has a lower MBE than the chosen multimodel in this example, too, but a worse MAE. GFS05, MFGLOBAL, and NEMSGLOBAL also show lower MAE values and MBEs than the chosen multimodel. In Fig 14b, MBE and Spearman correlation are plotted. As can be seen, the optimum of the correlation does not fit to the lowest MBE value, but is as an MBE of about 0.4 $\frac{m}{s}$. The multimodel with the lowest MAE shows, in this example, a higher correlation than all raw models. However, there are multiple weightings, for which multimodels have a higher correlation. On the other hand, increasing the correlation above the correlation of the chosen multimodel leads to higher MBEs, too.



(a) MAE to MBE



(b) MBE to Spearman correlation

Figure 14: Error metrics for all 256 calculated multimodels and the raw models for a randomly selected station. The red dot shows the multimodel with the lowest MAE for this station (station-based multimodel). In (14a), the MAEs and the MBEs of each multimodel are shown, while in (14b) the MBEs to the Spearman correlations are plotted.

7 Discussion

7.1 Interpretation of wind model validations

The prediction model ICON performed best compared to the other raw models for many error metrics. ICON has the lowest averaged MAE, the highest correlation for all values and values over 3 $\frac{m}{s}$ as well as the lowest MAPE. Therefore, ICON proved to be the best model choice in the majority of the investigated station sites. On the other hand, NEMSGLOBAL turned out to be the raw model with the highest error metrics for most parts of the world. NEMSGLOBAL showed, on the one hand, the worst overall MAE, and on the other hand strong spatial patterns of overall underprediction. Therefore, predictions could be enhanced by adapting the model predictions for regions with general over- or underpredictions. For Europe, a view on the regional model is worthwhile. NEMS4 and NEMS12 showed overall worse prediction skills in MAE, MBE, RMSE and MAPE than ICON, GFS05, ERA5, and MFGLOBAL. However, an effect of the reduced grid size can be seen, if you look exclusively at the NEMS family. NEMS4 with a spatial resolution of 4 km showed the best MAE in the NEMS-family, followed by NEMS12 with the 12 km resolution and NEMSGLOBAL with 30 km resolution. The correlation of NEMS4 for all data and wind speeds above 3 $\frac{m}{s}$ is almost as good as ICON's correlations, while NEMSGLOBAL's correlations are the worst compared to the other models. Therefore, it stands to reason that a reduction in the grid size brings improvements in predictions. Local features, such as mountain-valley winds or roughness can be included in the calculations [42]. Especially in heterogeneous areas, this should have a greater effect, although this has not really been reflected in the NEMS models. Therefore, the third hypothesis H3 can be confirmed.

On a spatial level, all models in Europe do not show a strong spatial pattern for the MAE. Even though there are differences in the values of the MAE, the MAEs of the models are more or less evenly distributed. On a global level, there are differences in performance that seem to depend on the distance to the sea - the less distance, the worse the predictions related to the MAE. This could be a consequence of stronger wind speeds that exist near the water. This is consistent with a validation of the NEMS models by another study [29]. However, it is not the case that regions, where the models were developed, generally got better predictions from these models. Thus, the second hypothesis H2 can be rejected.

For both, Europe and global analysis, ERA5 shows good results. However, ERA5 performs not much better than ICON - in fact, the averaged correlations in Europe of ICON are better than ERA5's correlations, for the world it is very similar. Therefore, the first hypothesis H1 has to be rejected.
But even though there are differences in the prediction accuracy, all analyzed models share the same weakness: higher wind speeds are not predicted well. While the averaged MAE, MBE, RMSE and MAPE show acceptable results, the performance for the prediction of higher wind speeds is very weak for all models. This can already be seen in the lower correlation for wind speeds over 3 $\frac{m}{s}$, which is, very rough, about 0.1 worse than the overall correlation. Even more, it can be seen in the values of POD, FAR and HSS. For example, values over 5 $\frac{m}{s}$ are recognized in only 48 - 63 % of all cases, corresponding to the model. For wind speeds over 20 $\frac{m}{s}$, only 2 - 14 % of the values are discovered from the models. NEMSGLOBAL has the highest POD compared to other models, but this is probably caused by a general overestimation (positive bias) instead of a real prediction skill. This is shown by the high FAR values for NEMSGLOBAL and the mediocre HSS values.
Surprisingly, GFS05, which showed no outstanding results for error metrics like MAE, MBE, and correlation, performs best for higher wind speeds compared to other models. Thus, GFS05 seems to be the right model choice out of the four prediction models, if the prediction of high wind speeds is in the focus. However, GFS05 has the same fundamental problem of underestimation of higher wind speeds, even if it is less pronounced.
Moreover, the underprediction of high wind speeds in all models can be seen in the distribution of the MBE. Negative MBE values reach until -15 $\frac{m}{s}$, which can only occur for high wind speeds. A low correlation and a negative MBE can be indirectly related: A low correlation for Weibull-distributed data like wind speed combined with an acceptable MAE indicate, that the model seems to predict generally low wind speed values. Because most of the time, wind speeds are low, the model's MAE is low. However, bigger changes in measured wind speed are not represented well in the model predictions (low correlation), so that high wind speeds are strongly underestimated (negative MBE). This pattern can be seen in the spatial analysis of correlation and MBE of the different models: Values with low correlation often have negative MBEs. The poor prediction skill for higher wind speeds is particularly problematic because high wind speeds are the ones that can produce high yields in wind use on the one hand, and on the other hand, also bring high risks of storm damage. Thus, low MAE may indicate a good accuracy, but depending on the application area, models with a focus on POD, FAR, HSS, and other skill scores for

higher wind speeds would be better indications for the performance of a model. However, it has to be considered, that a model prediction of 4.9 $\frac{m}{s}$ for a threshold of 5 $\frac{m}{s}$ would be handled as not recognized when the measurement has 5.1 $\frac{m}{s}$, even if the error is in fact really small. Another problem of POD, FAR and HSS is that the values represent only a small part of the dataset. Speaking about one-hour averages especially the observation of high wind speeds of 20 $\frac{m}{s}$ may occur very rarely and for some stations not at all. This makes this error values a tool, but does not empower them for sole interpretation.
Another aspect to be considered for model evaluation is the quality of the measurement data. Metar data have high quality and a quality control was applied, however, errors occur in all observations, as there are many sources of error, such as the inaccuracies of the measuring equipment, non-representativeness of the site, or errors in maintenance by humans. Therefore, comparisons of models must always be taken with caution - if, for example, some models have been adjusted to other high-quality measurement data that have a less exposed location than at the airfield, the model could be misdiagnosed as underestimating.
Last but not least, the location of the grid cells of the models can differ strongly (see Figure 1). Therefore, the same measurement station can be at the same time representative for the grid cell of one model and not representative for the grid cell of another model. For example, a station can intersect with a sea on one side and mountains on the other side. The station can lie on different sides of the model cells. Now, one model will calculate wind speed which is on average representative for the sea, the other model calculates values for the mountain area - totally different wind speeds can result for the same station, even if both models have a good performance. Especially in heterogeneous locations, this will have an effect.

7.2 Multimodels

The station-based multimodel approach provides a spatial analysis of model performances. ICON shows high weightings almost in every part of the world. This is not surprising, being the strongest model in the validation. Interesting, though, is that for almost all stations, a multimodel approach of a weighted mixture of the prediction models is better than ICON or other raw models. This means, that in almost every part of the world, multimodels can improve forecasting.
For NEMSGLOBAL, the spatial pattern is noticeable: In most parts of the world, NEMSGLOBAL does not play a relevant role, but in North Canada, their weighting is very high.

This is explainable with the validation of the MBEs of the models: All models underestimate wind speeds in the North Canada stations. However, NEMSGLOBAL highly overpredicts wind speeds in South Canada. It seems likely that the underprediction of the northern Canada regions in NEMSGLOBAL is so weak because neighboring areas have a high overprediction and this influences the predictions.
GFS05 is the second strongest model, while MFGLOBAL and NEMSGLOBAL's impact is not very high. The impact of the number of weighted models was evaluated for the global weighting. While mixing GFS05 and ICON brings a considerable gain compared to the raw models, the effect of additional models is much weaker. In the optimized global weight, each take 10 % of the weighting, therefore, their presence is improving the predictions. However, the actual MAE improvement of including MFGLOBAL and NEMSGLOBAL is very low. Multimodels composed by several models have the disadvantage that they have a greater risk that one of the used models does not work than raw models or multimodels with fewer models. Nevertheless, they can still make high-quality predictions if one of the models fails temporarily because they have other models that they can use. Therefore, costs and computational effort must be weighed against the additional benefits that arise with additional models.
Averaged over all stations, both multimodel approaches, station-based and global weighting, show a better MAE than all used raw models. The mean of the station-based approach showed the lowest MAE, however, the global weight could decrease the error value, too. Remarkable is, that even if the models are optimized to the MAE, other error metrics improve, too. For both approaches, the averaged MBE is nearer to null than for all raw models, the correlations for all values and for wind speeds over 3 $\frac{m}{s}$ are higher for both multimodel approaches, and the MAPE and the RMSE show also for both multimodel approaches lower values than all raw models. However, in predicting higher wind speeds, the improvement of the multimodel approach is not so clear: For POD and FAR values, both multimodel approaches perform average compared with the other models. For the HSS, which describes the real prediction skill, perform best for wind speeds over 5 $\frac{m}{s}$, but fall off in the course of the increase in wind speeds: For values above 15 $\frac{m}{s}$, the station-based weighting is at least still on par with the best raw model GFS05, for values over 20 $\frac{m}{s}$ both multimodels are lower than GFS05. Nevertheless, these results are acceptable - the averaging of the weighting approach could lead to predictions with worse performance of high wind speed, if the correct high prediction of one model is decreased by other models with more conservative predictions. With that, hypotheses H4 and H5 can be partially confirmed: while the overall predictions can be improved, higher wind

speeds are not necessarily predicted better.
Regarding the error metrics, an optimization for other values than the MAE could be considered. The analysis of all tested multimodels for one station shows that there are multimodels with, for example, a higher correlation or an MBE nearer null, which only can be improved on costs of the MAE. The choice of the optimized error value should be well considered. Another possibility is to optimize to a mixture of different error values, as has already been done in other work [43].
However, both implemented multimodel approaches can be useful: On the one hand, global weighting can help to improve predictions easily. Because the weighting is optimized for more than 5000 stations on a global scale, this weighting could be a good approximation for regions where no stations exist. The analysis with 2019 data showed, that the global weighting is relatively stable over time and is not completely different in another year - that means, that the global weighting can improve accuracy without constant adaptation. However, for a reliable sensitivity analysis of the global weight, data of more years should be investigated. Additionally, one would see if the models change their performance over the years in relation to each other.
On the other hand, station-based weighting can maximize predictions for a specific site, when it is the center of interest. Furthermore, station-based weighting shows spatial patterns in how strong models are in contrast to other models - this can be used, on the one hand, for clustering of weightings as a hybrid of the two approaches and on the other hand for model analysis.

8 Conclusion

In this thesis, the 10 m wind speed predictions of the reanalysis model ERA5 as well as the prediction models ICON, MFGLOBAL, GFS05, and NEMSGLOBAL were analyzed for the whole world. A spatial analysis of different error metrics was made as well as overall performance. On a spatial level, all models showed a trend to predict worse near the sea and better in continental areas. The spatial distributions of the error metrics of the models are similar - an exception is NEMSGLOBAL, which shows a positive bias in North America and Europe. Averaged over all stations, ICON surprises with similar or even better predictions than the reanalysis model ERA5. A big issue for all models is the prediction of higher wind speeds. Wind speeds over 15 $\frac{m}{s}$ were discovered in only 4.5 - 20 % of the cases, regarding the model. The higher the wind speed, the worse the prediction skill.

For Europe, the two models NEMS4 and NEMS12 were added for analysis. These models come from the NEMS family like NEMSGLOBAL but with a higher spatial resolution. NEMS4 with 4 km resolution showed a better performance than NEMS12 with 12 km resolution, which itself performed better than NEMSGLOBAL with 30 km resolution. Therefore, the decrease in grid size resulted in more accurate predictions. However, the reduction of the grid could only partially compensate for the weakness of the NEMS models - ICON, despite a larger grid, performs better than all NEMS models.

Moreover, two multimodel approaches were applied. In both approaches, each weighting of the four global prediction models could get values from 0 to 100 % in 10 % steps when all models added up to 100 %, resulting in 256 combinations. The weightings for the lowest possible MAE are optimized. One approach optimized the weighting for each observation site, the other approach searched the global weighting for the lowest MAEs averaged over all stations. Both multimodel approaches could improve not only the MAE, but further the correlation, MBE, MAPE and RMSE. This enables these approaches for improving predictions. The station-based multimodel can improve forecasts if one is interested in a specific site, the global weighting enables an easy way to improve forecasts globally and can be used in areas without an observation station. However, both multimodels cannot improve the prediction of high wind speeds, which was already a weakness of the raw models.

For further research, the impact of missing processes such as mountain-valley winds in the models could be evaluated by comparing error metrics of weather stations located in different terrain. For example, stations in complex terrain could be averaged and

compared to stations in open terrain. The difference of both errors is then related to the missing process of the model.
For the multimodels, the choice of the error value to be optimized could be reconsidered. For example, the correlation would include high wind speeds more. Optimization for high wind speed measures like the HSS would be possible in order to address the problem of poor forecasting skills for high wind speeds, however, these optimizations would probably lead to a poor overall prediction. Therefore, another approach would be the combination of different error metrics in order to find a comprise between the different challenges.
Besides, multimodels could be further optimized, if the weighting would be analyzed for seasonal changes. Furthermore, regions could be divided into clusters, which, on the one hand, could be representative for the regions and within this range could also be used for regions without stations and on the other hand, could capture the spatial differences better than a global weighting.

Acknowledgments

I want to thank meteoblue for providing the data and the consulting. Especially I want to mention Sebastian Schlögl, which was always open to my questions and ideas. Furthermore, I want to thank Dirk Schindler and Anke Weidlich for examining the master thesis. Special thanks to my boyfriend Manuel and my family for supporting and motivating me and reading my thesis beforehand. And last but not least, I want to thank our dog Mirana, which always cheered me up and remembered me what is important in life when I threatened to sink in work.

References

[1] C. Wolfram, O. Shelef, and P. Gertler, "How will energy demand develop in the developing world?" *Journal of Economic Perspectives*, vol. 26, no. 1, pp. 119–38, 2012.

[2] X. Lu, M. B. McElroy, and J. Kiviluoma, "Global potential for wind-generated electricity," *Proceedings of the National Academy of Sciences*, vol. 106, no. 27, pp. 10 933–10 938, 2009, ISSN: 0027-8424. DOI: 10.1073/pnas.0904101106. eprint: https://www.pnas.org/content/106/27/10933.full.pdf. [Online]. Available: https://www.pnas.org/content/106/27/10933.

[3] NOAA. (2021). "Billion-dollar weather and climate disasters: Summary stats," [Online]. Available: https://www.ncdc.noaa.gov/billions/summary-stats/US/1980-2020.

[4] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, no. 7567, pp. 47–55, 2015.

[5] (Nov. 8, 2021). "DWD Globalmodell ICON," [Online]. Available: https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/%2001_num_vorhersagemodelle/icon_beschreibung.html.

[6] (Nov. 8, 2021). "MétéoFrance MFGLOBAL," [Online]. Available: http://www.umr-cnrm.fr/spip.php?article121&lang=en.

[7] (Dec. 1, 2021). "DWD Numerische Vorhersagemodelle," [Online]. Available: https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/%2001_num_vorhersagemodelle/numerische_vorhersagemodelle.html.

[8] E. Lorenz, "The butterfly effect," *World Scientific Series on Nonlinear Science Series A*, vol. 39, pp. 91–94, 2000.

[9] L. S. Gandin, "Complex Quality Control of Meteorological Observations," *Monthly Weather Review*, vol. 116, no. 5, pp. 1137–1156, 1988. DOI: 10.1175/1520-0493(1988)116<1137:CQCOMO>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/mwre/116/5/1520-0493_1988_116_1137_cqcomo_2_0_co_2.xml.

[10] C. G. N. Wade, "A quality control program for surface mesometeorological data," *J. Atmos. Oceanic Technol.*, vol. 4, pp. 435–453, 1987.

[11] E. E. Lucio-Eceiza, J. F. González-Rouco, J. Navarro, H. Beltrami, and J. Conte, "Quality Control of Surface Wind Observations in Northeastern North America. Part II: Measurement Errors," *Journal of Atmospheric and Oceanic Technology*, vol. 35, no. 1, pp. 183–205, 2018. DOI: 10.1175/JTECH-D-16-0205.1. [Online]. Available: https://journals.ametsoc.org/view/journals/atot/35/1/jtech-d-16-0205.1.xml.

[12] M. Shafer, C. Fiebrich, D. Arndt, S. Fredrickson, and T. Hughes, "Quality assurance procedures in the Oklahoma Mesonetwork," *J. Atmos. Oceanic Technol.*, vol. 17, pp. 474–494, 2000.

[13] D. Meek and J. Hatfield, "Data quality checking for single station meteorological databases," *Agric. For. Meteor.*, vol. 69, pp. 85–109, 1994.

[14] Etor E. Lucio-Eceiza and J. Fidel González-Rouco and Jorge Navarro and Hugo Beltrami, "Quality Control of Surface Wind Observations in Northeastern North America. Part I: Data Management Issues," *Journal of Atmospheric and Oceanic Technology*, vol. 35, no. 1, pp. 163–182, 2018. DOI: 10.1175/JTECH-D-16-0204.1. [Online]. Available: https://journals.ametsoc.org/view/journals/atot/35/1/jtech-d-16-0204.1.xml.

[15] C. A. Fiebrich, C. R. Morgan, A. G. McCombs, P. K. Hall, and R. A. McPherson, "Quality assurance procedures for mesoscale meteorological data," *J. Atmos. Oceanic Technol.*, vol. 27, pp. 1565–1582, 2010.

[16] C. Xu, M. Yan, L. Ning, and J. Liu, "Summer westerly jet during the mid-Holocene: a multi model study," in *AGU Fall Meeting Abstracts*, vol. 2019, Dec. 2019, GC11G–1130.

[17] J. Courtney, P. Lynch, and C. Sweeney, "Statistical Post-Processing of Wind Speed Forecasts to Estimate Relative Economic Value," in *EGU General Assembly Conference Abstracts*, ser. EGU General Assembly Conference Abstracts, Apr. 2013, EGU2013–9680.

[18] H. Y. SHI Lan XU Lina, "Application research on the multi-model fusion forecast of wind speed," *Plateau Meteorology*, vol. 36, no. 4, 1022, p. 1022, 2017. DOI: 10.7522/j.issn.1000-0534.2017.00021. [Online]. Available: http://www.gyqx.ac.cn/EN/abstract/article_6252.shtml.

[19] G. C. Leckebusch, T. Pardowitz, M. G. Donat, D. Renggli, T. Kruschke, and U. Ulbrich, "Multi-model ensemble extreme value analysis of European storm events," in *EGU General Assembly Conference Abstracts*, ser. EGU General Assembly Conference Abstracts, May 2010, p. 12 128.

[20] C. Feng, M. Cui, B.-M. Hodge, and J. Zhang, "A data-driven multi-model methodology with deep feature selection for short-term wind forecasting," *Applied Energy*, vol. 190, pp. 1245–1257, 2017.

[21] (Nov. 26, 2021). "Meteoblue Measurement API," [Online]. Available: `http://measurement-api.meteoblue.com/rawdata/docs#`.

[22] (Nov. 26, 2021). "Meteoblue Dataset SDK," [Online]. Available: `https://github.com/meteoblue/python-dataset-sdk`.

[23] (Nov. 8, 2021). "Meteoblue Datasets," [Online]. Available: `https://docs.meteoblue.com/en/meteo/data-sources/datasets`.

[24] G. Zängl, D. Reinert, P. Rıpodas, and M. Baldauf, "The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core," *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 687, pp. 563–579, 2015.

[25] T. Heppelmann, A. Steiner, and S. Vogt, "Application of numerical weather prediction in wind power forecasting: Assessment of the diurnal cycle," *Power*, vol. 200, p. 0, 2017.

[26] (Nov. 8, 2021). "NOAA GFS model," [Online]. Available: `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00634`.

[27] Z. Zou and Y. Deng, "Comparison of wind speed prediction in wrf model based on ncep gfs and jma gsm forecasting fields," *Water Resour. Power*, vol. 34, pp. 194–197, 2016.

[28] M. Zamo, L. Bel, O. Mestre, and J. Stein, "Improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression," *Weather and Forecasting*, vol. 31, no. 6, pp. 1929–1945, 2016. DOI: `10.1175/WAF-D-16-0052.1`. [Online]. Available: `https://journals.ametsoc.org/view/journals/wefo/31/6/waf-d-16-0052_1.xml`.

[29] 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*, vol. 30, no. 3, pp. 827–840, 2015. DOI: 10.1175/WAF-D-14-00049.1. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/30/3/waf-d-14-00049_1.xml.

[30] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, *et al.*, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[31] (Nov. 5, 2021). "ECMWF," [Online]. Available: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5.

[32] M. O. Molina, C. Gutiérrez, and E. Sánchez, "Comparison of ERA5 surface wind speed climatologies over Europe with observations from the HadISD dataset," *International Journal of Climatology*, 2021.

[33] B. Jourdier, "Evaluation of ERA5, MERRA-2, COSMO-REA6, NEWA and AROME to simulate wind power production over France," *Advances in Science and Research*, vol. 17, pp. 63–77, 2020.

[34] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[35] B. Urquhart, M. Ghonima, D. (Nguyen, B. Kurtz, C. W. Chow, and J. Kleissl, "Chapter 9 - Sky-Imaging Systems for Short-Term Forecasting," in *Solar Energy Forecasting and Resource Assessment*, J. Kleissl, Ed., Boston: Academic Press, 2013, pp. 195–232, ISBN: 978-0-12-397177-7. DOI: https://doi.org/10.1016/B978-0-12-397177-7.00009-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123971777000097.

[36] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.

[37] K. Conradsen, L. B. Nielsen, and L. P. Prahm, "Review of weibull statistics for estimation of wind speed distributions," *Journal of Applied Meteorology and Climatology*, vol. 23, no. 8, pp. 1173–1183, 1984. DOI: 10.1175/1520-0450(1984)023<1173:ROWSFE>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/apme/23/8/1520-0450_1984_023_1173_rowsfe_2_0_co_2.xml.

[38] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, ISSN: 00029556. [Online]. Available: http://www.jstor.org/stable/1412159.

[39] O. Hyvärinen, "A probabilistic derivation of heidke skill score," *Weather and Forecasting*, vol. 29, no. 1, pp. 177–181, 2014. DOI: 10.1175/WAF-D-13-00103.1. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/29/1/waf-d-13-00103_1.xml.

[40] I. T. Jolliffe and D. B. Stephenson, *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, 2012.

[41] (Oct. 13, 2021). "EFRAINMAPS," [Online]. Available: https://tapiquen-sig.jimdofree.com/english-version/free-downloads/europe/.

[42] W. C. de Rooy and K. Kok, "A combined physical–statistical approach for the downscaling of model wind speed," *Weather and Forecasting*, vol. 19, no. 3, pp. 485–495, 2004. DOI: 10.1175/1520-0434(2004)019<0485:ACPAFT>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/wefo/19/3/1520-0434_2004_019_0485_acpaft_2_0_co_2.xml.

[43] C. Jung, D. Schindler, J. Laible, and A. Buchholz, "Introducing a system of wind speed distributions for modeling properties of wind speed regimes around the world," *Energy Conversion and Management*, vol. 144, pp. 181–192, 2017.