Evaluation and assessment of the performance of the KNMI Schiphol Kansverwachting (SKV) for Mainport Schiphol with respect to wind direction, wind speed and wind gusts

Research report

Internship - Luchtverkeersleiding Nederland (LVNL)

Author: Thomas Vermeulen (Student number: 1015872)

Luchtverkeersleiding Nederland (LVNL) & Koninklijk Nederlands Meteorologisch Instituut (KNMI)

Knowledge and Development Centre (KDC) Mainport Schiphol & Centre of Excellence (CoE)

WUR supervisor: Gert-Jan Steeneveld

Internship supervisor: Ferdinand Dijkstra

KNMI supervisors: Nico Maat & Hans van Bruggen

MSc research internship - Earth and Environment - MAQ-70824

Wageningen University & Research (WUR)

Meteorology and Air Quality (MAQ)

Tuesday 28th June, 2022

Contents

1 | Introduction

1.1 Scientific context & background information

This research internship report focuses on evaluating and assessing the performance of the KNMI Schiphol Kansverwachting (SKV) with respect to wind direction, average wind speed and wind gusts. The SKV indicates the probabilities of weather situations/factors to occur which are relevant to the aviation sector, such as visibility or wind and is made on the basis of certain weather models. The capacity management and Air Traffic Management (ATM), which is very relevant for the Dutch aviation industry, is strongly dependent on this specific SKV. Inaccurate forecasting can lead to last-minute restrictions which were probably not needed at all in the end. This research evaluates and assesses the performance of the SKV which helps to find solutions to minimise delays and reduce the workload of air traffic controllers. Innovating management and logistics at Mainport Schiphol is essential and relevant for the aviation sector.

Aviation meteorologists are responsible for making the most accurate weather forecasts possible. Information on sudden changes in weather conditions are crucial in order to maintain the safety and efficiency for flight operations. The meteorological observations at airports and forecasts for the relevant airports are produced following a certain code according to regulations from the International Civil Aviation Organization (ICAO) and the World Meteorological Organization (WMO) (Jacobs and Maat, 2005).

The observations at the airport are provided every half hour and are called the "meteorological aviation routine weather report", summarised in the "Meteorological Aerodrome Report" (METAR). The forecast is provided up to several hours in advance and is called the "Terminal Aerodrome Forecast" (TAF). It contains detailed information of the forecasts regarding wind, visibility, cloud-base height, cloud-top height, cloud cover, precipitation and more variables (Jacobs and Maat, 2005). The TAFs are produced several times a day for some civil airports with specific lead times with respect to the issue time of the forecast itself. The TAFs for Schiphol are produced by the forecasters using Direct Model Output (DMO) from Numerical Weather Prediction (NWP) models. This model data is then interpreted and corrected in combination with the latest observations from various sources (satellite data, radar data, observations from several automatic weather stations in the country, etc.) (Jacobs and Maat, 2005).

Important to note is that this model data on which the detailed weather forecast for the aviation sector is based, all have their shortcomings. Surface conditions are being simplified and some small scale effects and heterogeneities cannot be simulated perfectly. In order to partly compensate for this, the Model Output Statistics (MOS) is created, which is a statistical post-processing technique determining the systematical dependencies between local prediction variables (called predictands) and local observations/model variables (called predictor). Because of their relatively high forecasting accuracy, the MOS technique turned out to be a highly valuable interpretation aid for forecast meteorologist and is used for representing the so-called TAF Guidance (TAFG) (Deutscher Wetterdienst (DWD), 2019; Jacobs and Maat, 2005).

In 2003, the so-called Schiphol Kansverwachting (SKV) was introduced, because the TAF procedure did not exactly meet the requirements of the users at the airport and also because of the fact that probabilistic forecasts were considered as rather useful (Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2007). The most essential input for the TAF and SKV is provided by the TAF Guidance, which

can be seen as an automatic statistical post-processing application of numerical weather prediction models running at the KNMI. For the SKV, currently mainly the TAFG of HARMONIE is used. The NWP model output is provided to the KNMI forecaster, who uses this TAFG to construct the local TAF and SKV (Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2007). The first step in order to produce the SKV is to use the model output from operational NWP models and latest weather observations. Previously, HIRLAM was used as model output, but during the last few years this was transferred to HARMONIE. The output of the NWP is in the next step improved, based on several statistical algorithms to produce the TAF Guidance product. In the next step, all the relevant variables for the Schiphol Kansverwachting are grouped, where several adjustments can be made by the aviation weather forecaster. The different steps in order to get to the SKV are visualised in Figure 1. So, in short, the following steps are taken:

- Use model output from operational NWP. Currently HARMONIE is used for this.
- NWP model output is improved based on a statistical algorithm (MOS) as indicated previously to produce the TAF Guidance product.
- The most relevant variables are grouped in the SKV, used by the ATM. Adjustments and remarks are made by the aviation weather forecaster.

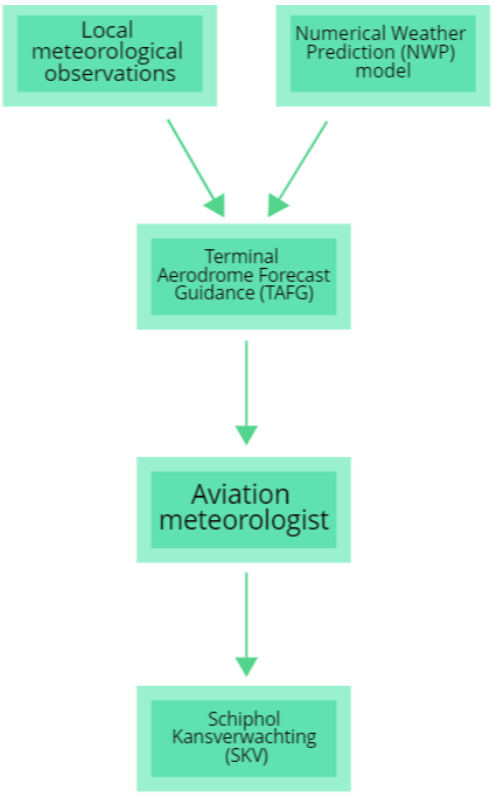


Figure 1: *Schematic overview of the forecast procedure leading to the probabilistic weather forecast called the Schiphol Kansverwachting (SKV) (de Rover et al., 2008).*

Important to note is that the aviation weather forecaster also has the data of other weather models (HIRLAM and ECMWF). The forecaster can use this data, in combination with knowledge about the performance of those models during certain weather conditions to adjust the TAFG for an accurate SKV.

It is crucial for regulating the operational capacity at Schiphol Airport to have accurate, reliable and unambiguous information regarding the weather (Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2007). As indicated in Jacobs and Maat (2005), the impact of surface winds on the aircraft depends on the angle between the wind direction and the exact geographical orientation of the runway. In general, aircraft cannot take off and land when the crosswind and tailwind exceed certain values. Schiphol Airport has several runways with different geographical orientations to cope with the different wind directions. Switching the runways at the proper time in advance can minimise the reduction in airport capacity (Knowledge & Development Centre (KDC) Mainport Schiphol (KNMI), 2015). So, to support the air traffic management in regulating the incoming and outgoing traffic, accurate forecasts related to wind are crucial. Particularly information on wind direction, average wind speed and wind gusts is found to be highly important and forms the basis of this research (Jacobs and Maat, 2005). The SKV is created roughly using lead times of 3 up to 35 hours: the lead times where the air traffic management team is interested in.

1.2 Aim & research questions

Important to note first is that an earlier intern of LVNL, Edwin Kok, focused on the validation of the averaged wind speed with data from 2015 up to and including 2017. A logical next step was to validate again the average wind speed with more recent data and extend this to validations of other aspects of wind: wind direction and wind gusts, as those factors appeared to be very important too for air traffic management (Jacobs and Maat, 2005). The main aim of this research was to evaluate and assess the performance of the KNMI Schiphol Kansverwachting (SKV) for Mainport Schiphol with respect to wind direction, average wind speed and its outliers (wind gusts). The findings can conclude which TAFG of which weather model is most accurate and whether the adjustments made by the aviation weather forecaster in the SKV are making the SKV more accurate or not. The findings can help to improve the forecasts and thus help the air traffic management. Therefore, the following research questions were attempted to be answered:

1. How does the TAF Guidance from different models and the SKV itself compare to each other regarding surface wind direction and how does this performance depend on lead time, period of the year of the valid forecast and on other wind-related variables?
2. How does the TAF Guidance from different models and the SKV itself compare to each other regarding the surface average wind speed and how does this performance depend on lead time, period of the year of the valid forecast and on other wind-related variables?
3. How does the TAF Guidance from different models and the SKV itself compare to each other regarding surface maximum wind gusts and how does this performance depend on lead time, period of the year of the valid forecast and on other wind-related variables?

2 | Data and methods

2.1 Data

TAF Guidance data of several weather models was used, together with the SKV and observational data. More specifically, the following datasets were used with corresponding period which is covered by the specific dataset:

- HARMONIE TAFG, which ranges from 2020/11/04 up to and including 2022/03/31.
- HIRLAM TAFG, which ranges from 2018/01/01 up to and including 2022/03/31.
- ECMWF TAFG, which ranges from 2018/01/01 up to and including 2022/04/03.
- SKV, which ranges from 2017/09/18 up to and including 2021/05/03.
- Hourly observational data and 10-minute observational data from the KNMI automated weather station at Schiphol. This is data from the automated weather station at runway 27 (Buitenveldertbaan), which reduces the uncertainty since the SKV is actually created for the same runway.

As indicated above, the observational data of the KNMI automated weather station is used since it gives the observational data of runway 27 and the Schiphol Kansverwachting is formally written as the forecast for runway 27. Observational data of the METAR was not used, since the wind gusts are only registered when a certain threshold is exceeded.

For the wind direction and average wind speed, the 10-minute averaged wind direction and wind speed of the 10 minutes previous of the valid exact time was used for both SKV, TAFGs and observations to keep it consistent. For the wind gusts, the maximum wind speed measured/forecasted of the 10 minutes previous of the valid exact time was used.

2.1.1 Weather models: HARMONIE, HIRLAM & ECMWF

The HARMONIE weather model has specifically been developed for short-term weather forecasts (up to 48 hours in advance) for a certain region. It has a relatively high resolution (2.5 x 2.5 km) and is often used for the analysis of small-scale meteorological processes. It is a non-hydrostatic convection-permitting model.

Previously, the KNMI used mainly HIRLAM (High Resolution Limited Area Model). This operational synoptic and mesoscale weather prediction model has a lower resolution (11 x 11 km) than HARMONIE and the KNMI decided to switch to HARMONIE a few years ago (Finnish Meteorological Institute (FMI)).

For the medium and long term weather forecasts, the European Centre for Medium-Range Weather Forecasts (ECMWF) is used (Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2020; European Centre for Medium-Range Weather Forecasts (ECMWF)).

The TAFG of the different models have some different lead times, but in this specific study, the lead times which were present in all the TAFGs were used for the validation. The lead times used in this research are described in more detail in the next section.

2.2 Methods

2.2.1 Selecting the data

The analyses which are described below, are executed for the forecasts with several lead times. The data of lead times of 2,3,4,5,6,7,8,9,12,15,18,21 and 24 hours are investigated. The forecast data of the TAF Guidances will be filtered based on those lead times and the dataset will be merged into one dataset and time for which the forecast was valid together with the observations for the corresponding time. This merging ensures that for the verifications of the SKV and TAFGs, the same meteorological situation is validated making the results more reliable. Two important notes have to be made here. The first thing has to do with the fact that the time period of the HARMONIE TAFG is relatively small, because this TAFG was being developed and finished near the end of 2020. Merging will cause that the data of the other TAFGs and the SKV itself will have the same length as the dataset of the HARMONIE TAFG. So, when including HARMONIE, the amount of data decreases significantly. Therefore, two datasets were used: one where HARMONIE is excluded (leading to a merged dataset from 2018-01-01 up to and including 2021-05-03 (the end time of the SKV data)) and one dataset where HARMONIE is included and the SKV is excluded in order to find out the accuracy of HARMONIE based on a considerable amount of data (a merged dataset from 2020-11-04 up to and including 2022-03-31). The first dataset is about two times larger than the second dataset. The second thing that is noteworthy, is that because the amount of data decreases with lead time (due to the number of times the output is saved), certain 'bins' were used to keep the amount of data more or less equal with lead time and with that maintain the homogeneity of the dataset as good as possible. For the lead times of 3 up to and including 9 hours, the bins had a with of 2 hours, meaning for example that for a lead time of 4 hours the lead times of 3 up to and including 5 hours were selected. From a lead time of 12 hours onwards, the width was doubled to maintain more or less the same amount of data and with that homogeneity of the used data.

The Schiphol Kansverwachting contains information in the form of absolute values regarding the wind, no probabilistic information. The SKV shows the wind speed, wind direction and gusts and also the standard deviation of the forecast ensemble of the wind direction and wind speed. Since the TAFG of the different weather models (HARMONIE, HIRLAM and ECMWF) contains also probabilistic information regarding those variables, it is difficult to compare those with the SKV itself. The main idea was thus to compare in the first place solely the error of the absolute values regarding wind direction, average wind speed and wind gust of the SKV and the TAFGs relative to the observational data.

The first part will thus focus on comparing the absolute values of the TAFGs and SKV with observational data and then find out how the errors of the TAFGs and the SKV relative to the observational data compare with each other. So, the first step is comparing the variables wind direction, average wind speed and wind gust of the TAFGs and the SKV with observations for the different lead times mentioned above.

Interesting is to find out whether the errors found, depend on the wind direction itself, the season for which the forecast was valid or even the time of the day for which the forecast was valid. Those three dependencies will also be investigated, since they all form a certain driver for wind to occur (heating/turbulence, which is dependent on season and time of the day for example).

2.2.2 Verification metrics and methods

Several metrics will be used to quantify the validation and will be calculated for every lead time: The Accuracy, Mean Bias Error (MBE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):

$$Accuracy = \frac{Hits}{Total} \tag{1}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{4}$$

The P_i indicates the predicted value (so from the TAFG or SKV), whereas O_i indicates the observational data. For the last variable, the accuracy, several choices where made on when a hit was considered and when not, which was of course on its turn dependent on the variable which was validated. For the wind direction, a hit was considered when the absolute difference between the predicted value and the observed value was less than 10°. For both the average wind speed and the wind gust, a hit was considered when the absolute difference between the predicted value and the observed value was less than 5 knots. Those values were chosen in consultation with Ferdinand Dijkstra and sound in first place rather arbitrary. However, when a certain TAFG (or SKV) performs systematically worse compared to the other ones, this will be found with more than one specific threshold, since the absolute difference is used here. The verification metrics will be used to compare the errors of the different TAFGs and the SKV compared to the observations, but also to investigate the dependency of the error on the variables mentioned in the previous section. In the end, by plotting the verification metrics with lead time, this will give an indication how fast the accuracy of the specific forecast decreases with lead time and whether this differs between the used datasets. Besides the verification metrics used, also regression plots of the predicated values against the observational data will be shown with the corresponding R-squared and equation of the trendline. Furthermore, histograms of error distributions will be made and independent sample t-tests will be run in order to support the validation. Time series (moving averages) of several metrics will be shown too to analyse the order of performance.

2.2.3 Earlier research by Edwin Kok

The previous intern, Edwin Kok, studied only the average wind speed and used data in the period of 2015 up to and including 2017. Edwin found that for the forecasts with a lead time of 3 hours, the bias of the HARMONIE TAFG was lowest (5-6 times lower than the bias of HIRLAM/ECMWF). However, for lead times of 6 hours and more, the HIRLAM TAFG showed the highest accuracy with a bias which is roughly 2 times lower than the bias of HARMONIE and roughly 4 times lower than the bias of ECMWF. In general, the study showed that the forecast quality decreased slightly with increasing lead time as one would expect. However, the error was slightly dependent on the valid season of the forecast. It was found that the forecast of the HARMONIE TAFG tended to overestimate the average wind speed during fall.

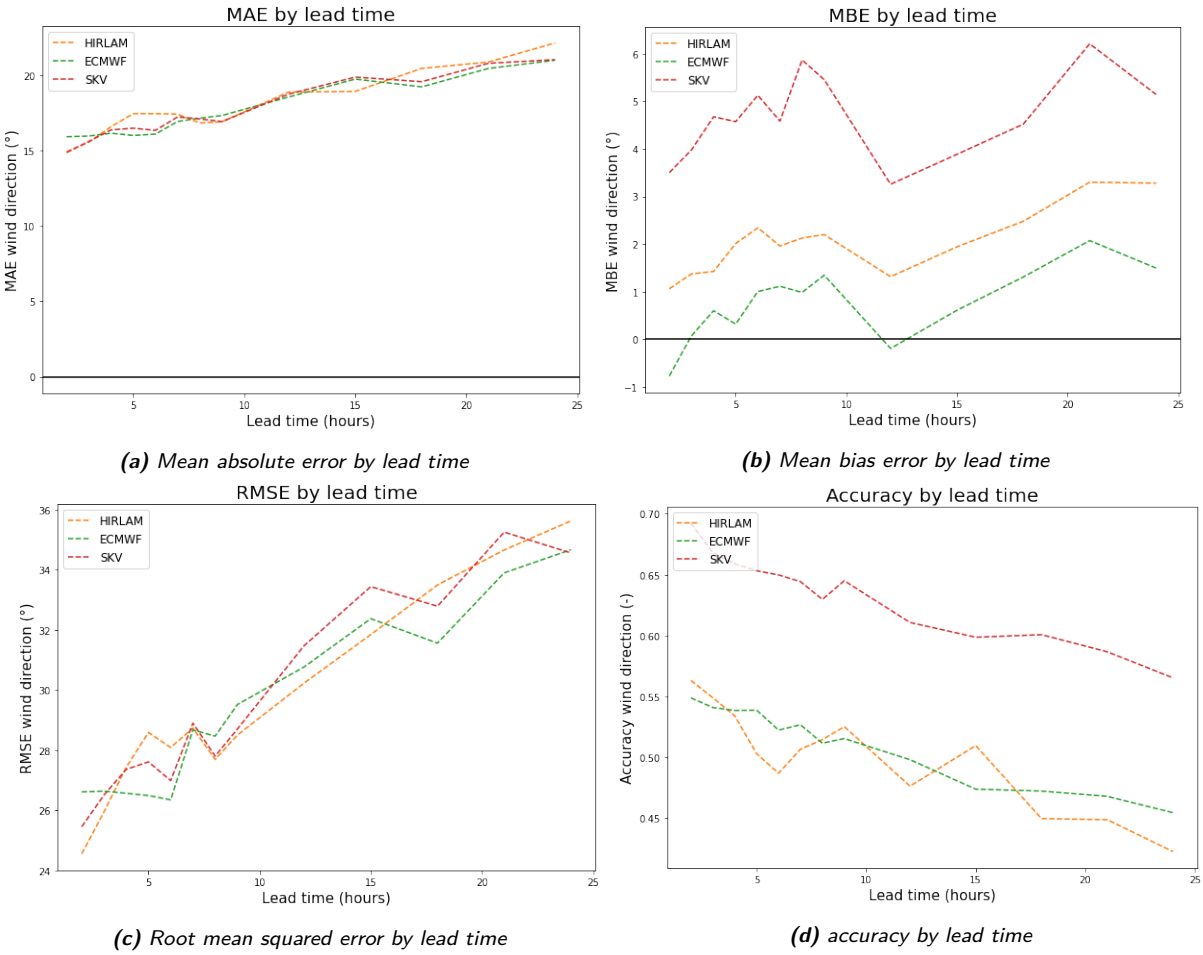
3 | Results and discussion

In this section, the validation of the different variables (wind direction, average wind speed and wind gusts) will be visualised and discussed. Bear in mind that, as indicated in the previous section, two different datasets were used in order to validate as fair as possible by maintaining the data size roughly equal. In one dataset, the TAFG of HARMONIE is excluded and one dataset contains HARMONIE, but does not contain the SKV itself. The discussion follows after showing the results of both datasets.

3.1 Wind direction

3.1.1 Verification metrics

Dataset without HARMONIE



(a) *Mean absolute error by lead time*

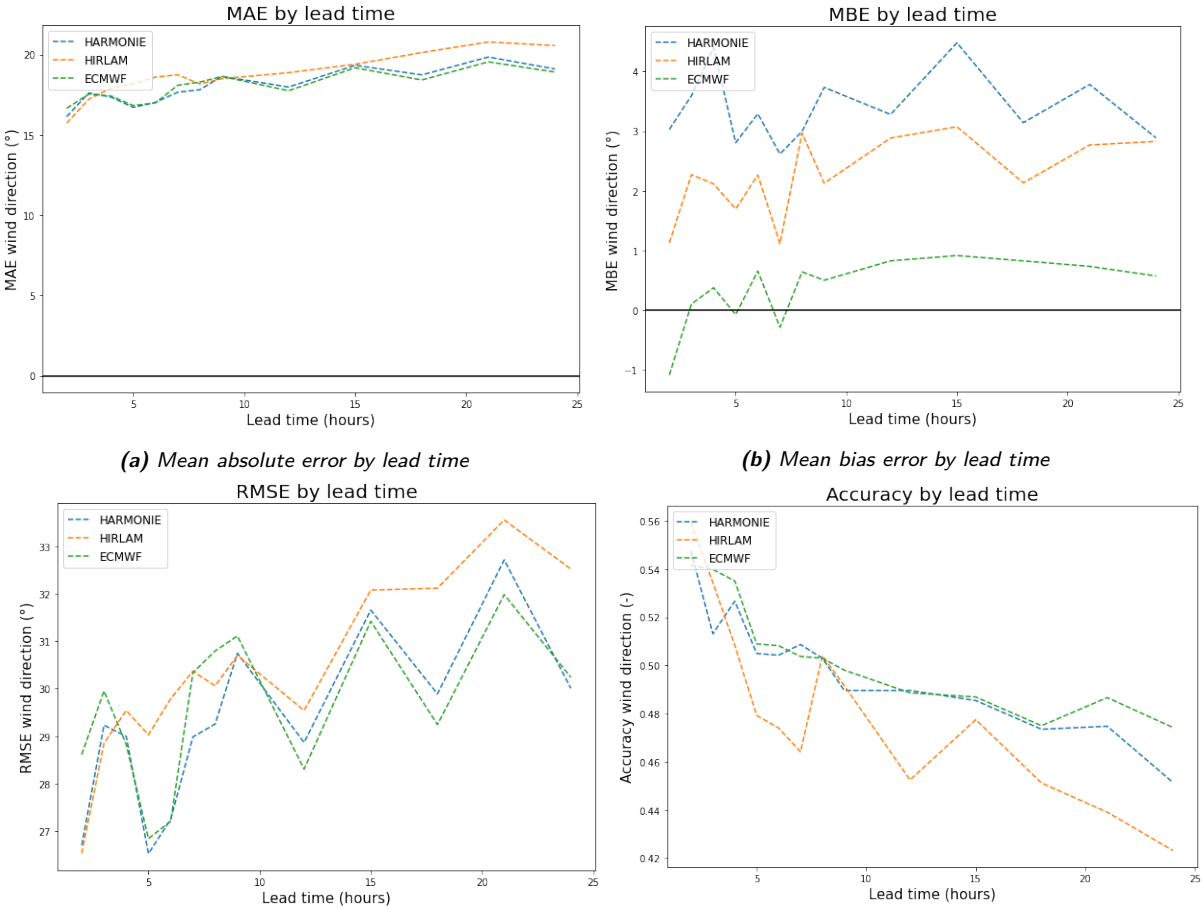
(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

Figure 2: *The four verification metrics used in order to quantify the validation for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

Dataset without SKV



(a) *Mean absolute error by lead time*

(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

Figure 3: *The four verification metrics used in order to quantify the validation for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

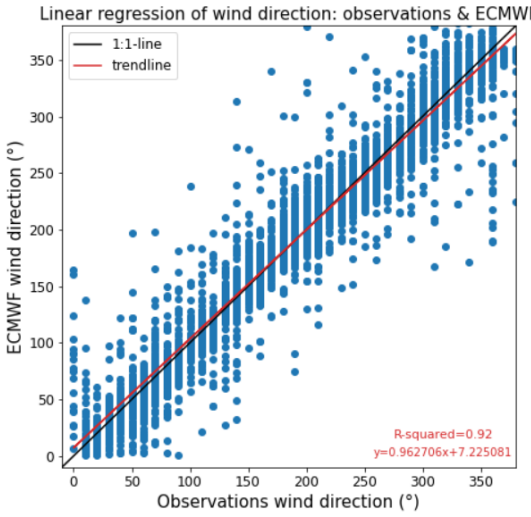
The very first thing that immediately stands out is the fact that the MAE shows values in the range of 15 up to and including 20°, which is rather high. As discussed with the host supervisor, discrepancies of 20° between forecast and observations can make the difference for an air traffic management team to switch to another runway. One other thing that is remarkable, is the fact that when comparing the TAFGs and SKV based on the mean absolute error (MAE), hardly any differences can be observed. Only for the dataset without SKV, it can be found in Figure 3a that in general the MAE of HIRLAM is higher than the MAE of HARMONIE and ECMWF. The mean bias error (MBE) leads to a different view with clearer differences considering the errors of the datasets. It shows that the TAFG of ECMWF has the lowest MBE, followed by HIRLAM and then HARMONIE. When comparing the absolute value of the MBE of the SKV with the one of HARMONIE, it can be concluded that the MBE of SKV is almost for all lead times higher than the one of HARMONIE. So the order ECMWF, HIRLAM, HARMONIE and SKV indicates the order from the lowest MBE to the highest. The third verification metric, the root mean squared error (RMSE), leads to the same view was the MAE did: hardly any differences and for the second dataset slightly higher values for the TAFG of HIRLAM. The last verification metric, the accuracy, leads to a slightly different view. It shows a significantly higher accuracy for the SKV compared to the accuracy for ECMWF and HIRLAM.

This is rather confusing, because the MBE showed the opposite. In the second dataset, the results are a bit more consistent, but still show differences compared to what was found considering the MBE. It shows no large differences between HARMONIE and ECMWF, but a smaller accuracy for HIRLAM.

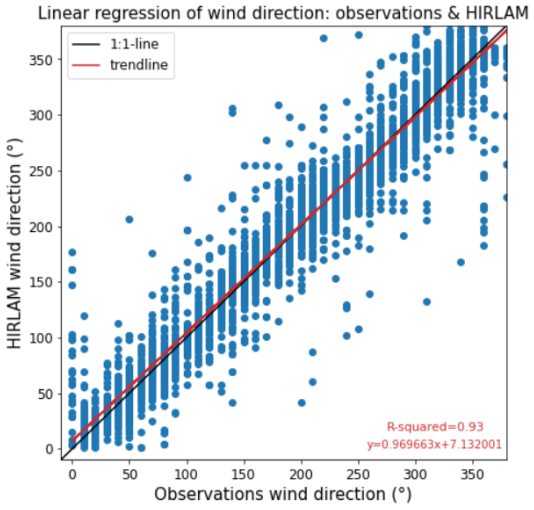
One general thing that can be noticed is the fact that the accuracy seems to decrease with lead time, whereas the MAE and the RMSE show logically an increasing trend with lead time. For the MBE, this effect is however hardly present. The fact that the MBE averages out the positive and negative errors can be the cause of this, just because the fact that the negative and positive bias increase with lead time at the same rate causing hardly a difference in the mean of this bias.

3.1.2 Linear regression

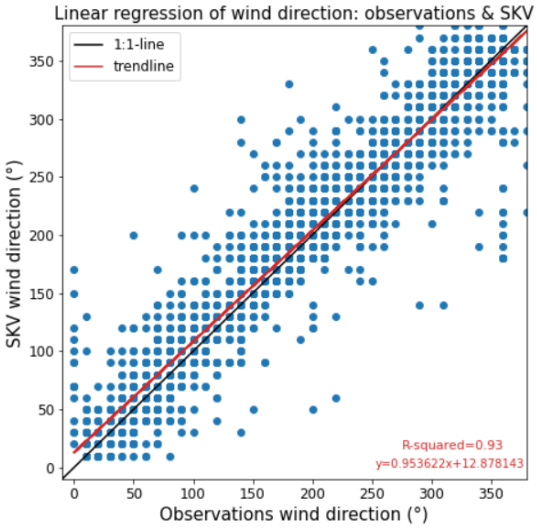
Dataset without HARMONIE



(a) ECMWF vs. observations (lead time of 3 hours)



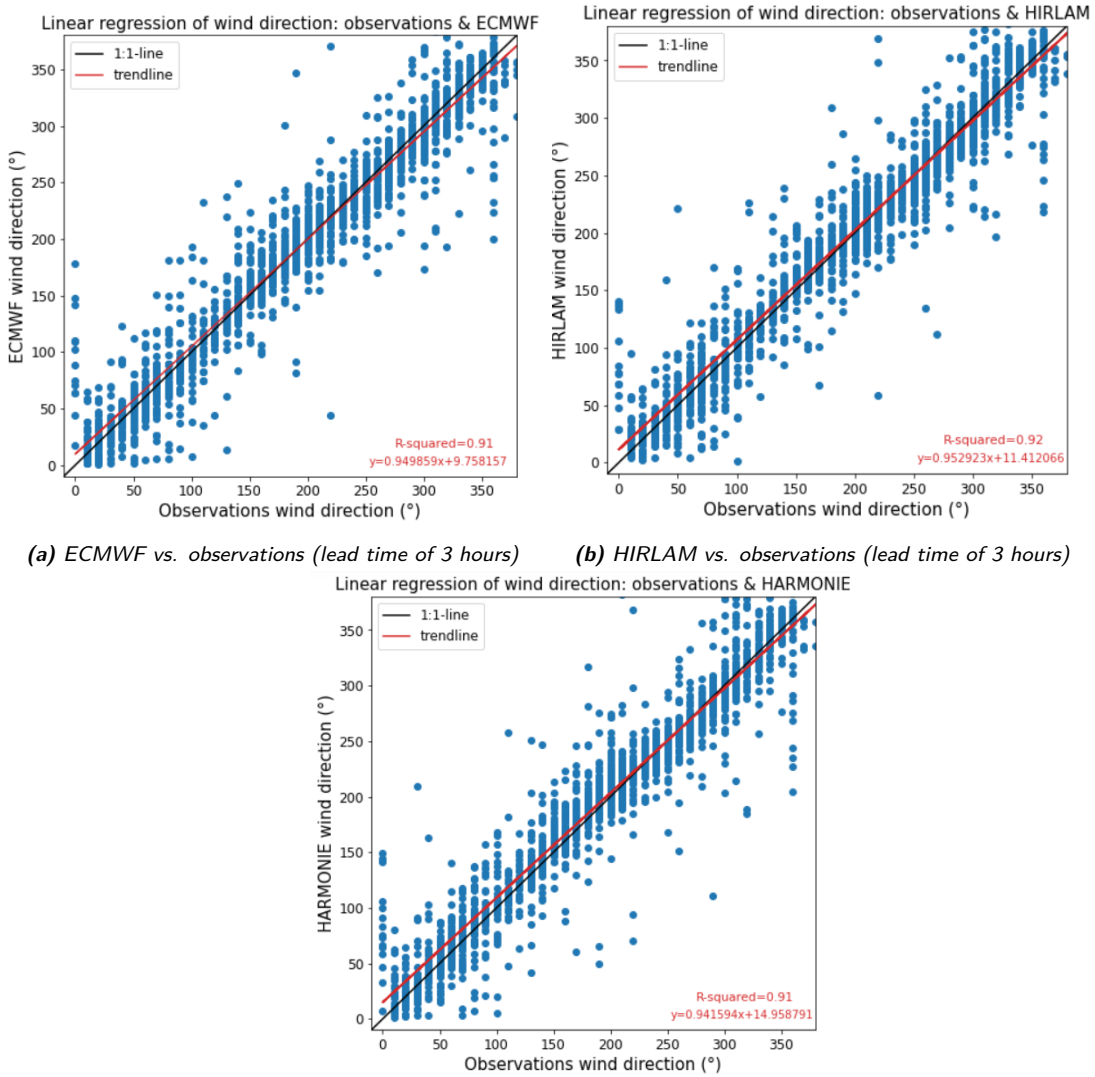
(b) HIRLAM vs. observations (lead time of 3 hours)



(c) SKV vs. observations (lead time of 3 hours)

Figure 4: *The linear regression plots of the wind direction of the TAFG with a lead time of 3 hours of a certain model against observational data for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

Dataset without SKV



(a) *ECMWF vs. observations (lead time of 3 hours)*

(b) *HIRLAM vs. observations (lead time of 3 hours)*

(c) *HARMONIE vs. observations (lead time of 3 hours)*

Figure 5: *The linear regression plots of the wind direction of the TAFG/SKV with a lead time of 3 hours of a certain model against observational data for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

When analysing the linear regression plots showed above, hardly any differences could be observed. For all the linear regressions made, the R-squared is just above 0.90, meaning that 90% of the variance for the wind direction of the TAFG/SKV is explained by the independent variable (observational wind direction). Furthermore, the slopes hardly differ between the linear regressions made. Remarkable is that the differences found using the verification metrics in the previous section hardly show up in the regressions. The linear regression plots are made using a lead time of 3 hours. It is rather remarkable that the differences found using the verifications metrics at a lead time of 3 hours do not show up in those linear regressions. However, in both datasets only the MBE (and for the dataset without SKV the accuracy) showed differences at this specific lead time. Those discrepancies showed in the previous section using the verification metrics are absolutely seen that small, that they do not show up in the linear regression plots.

In Appendix B, a time series of the monthly moving average of both MAE and MBE is shown. The analyses show no clear differences where one certain data source performs dominantly worse considering the wind direction. The moving average of the mean bias error leads partly to a different view since it shows a MBE of the SKV which is dominantly most far away from zero.

A concept that also attracts attention is the fact that this MBE is most of the time positive as the plots in the previous section also showed. This means that that the wind speed is often more backed than forecasted by the models (and SKV). This effect is most prominent present for the SKV as Figure 2b shows. The same effect can be found when analysing the histograms of error distributions in Appendix A. Backing of the wind occurs for instance because of friction due to the existing inhomogeneity of the land surface. More research is needed, but it cannot be ruled out that this effect is underestimated by the models (and not corrected well enough in the SKV). However, backing of the wind occurs also because of cold air advection. However, since it occurs continuously with time as shown in Appendix B, this is definitely not assumed.

Figure 6 below indicates a very important note with respect to the points mentioned for this specific validation. The plot shows the error (bias) compared to the average measured wind speeds. As can be expected, the error in the forecasted wind direction is significantly larger when wind speeds are low. This is also something that needs to be taken in account when discussing the results shown above. The larger deviations shown in the regression plots above are probably all the case during situations with a low wind speed. Due to time limitations, it was not possible to analyse in more detail this validation. However, it would be interesting to do the same analyses using certain ranges of wind speeds. Interesting to note is also that during situations with a low wind speed, when the deviations are relatively high, the wind as variable itself is not very important anymore for the air traffic management team.

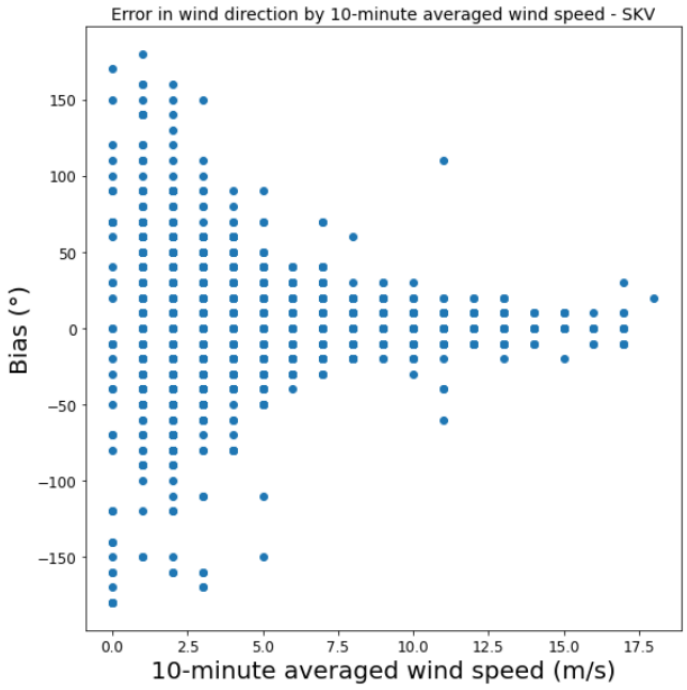
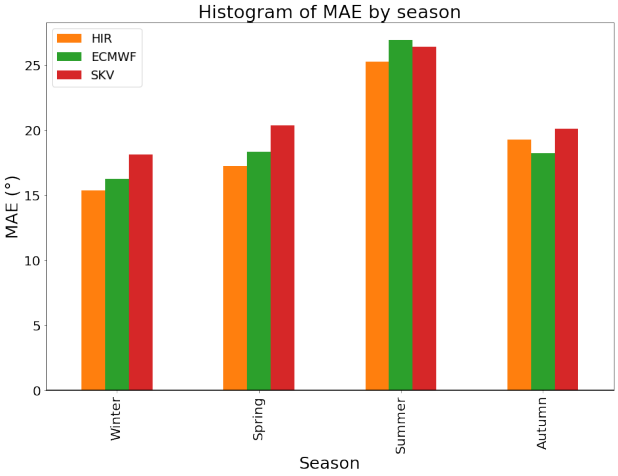


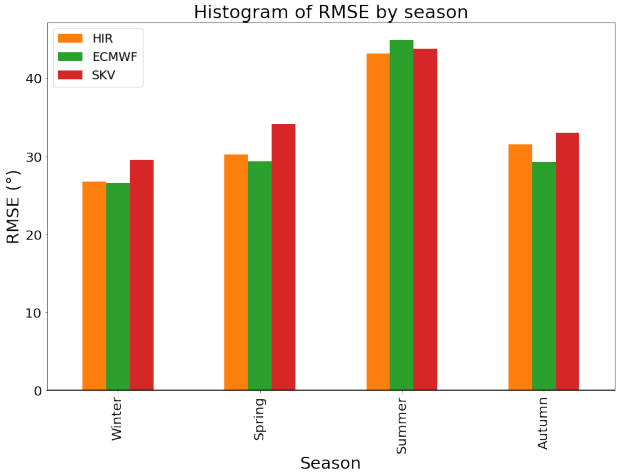
Figure 6: *The bias of the SKV of the wind direction vs. observed 10-minute wind speed (lead time of 3 hours).*

3.1.3 Dependencies

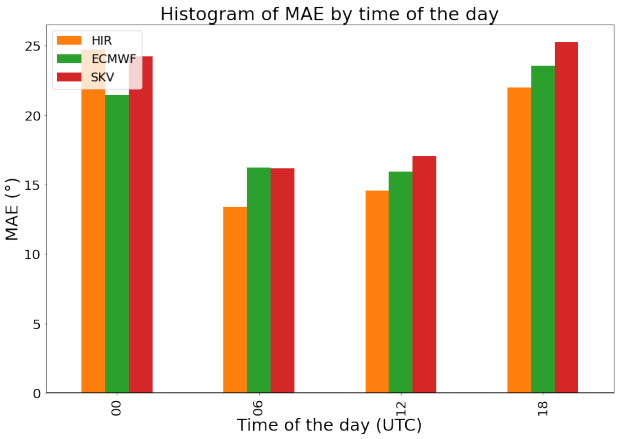
Dataset without HARMONIE



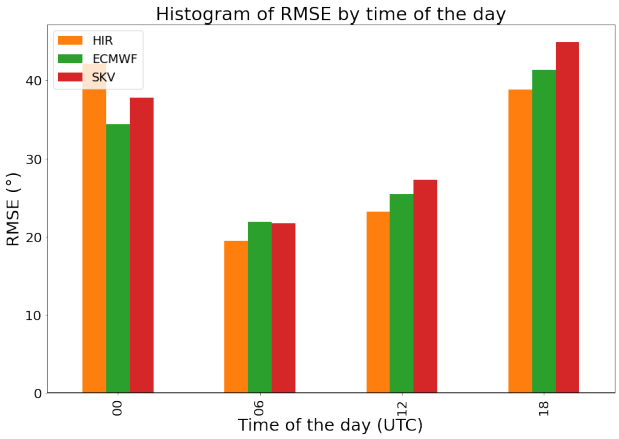
(a) MAE with season (all lead times)



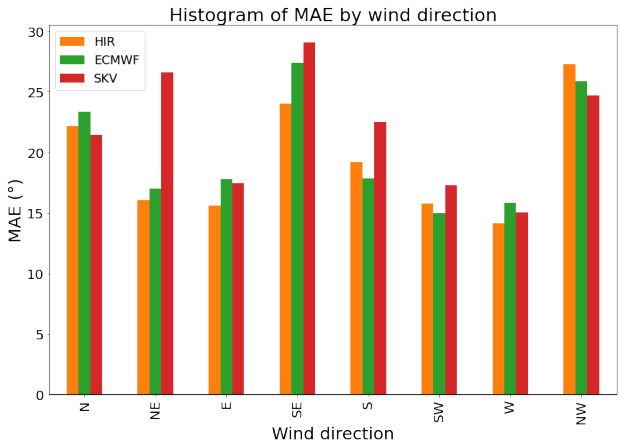
(b) RMSE with season (all lead times)



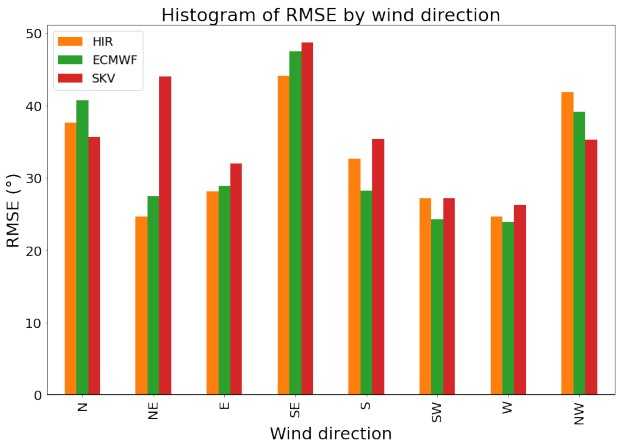
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



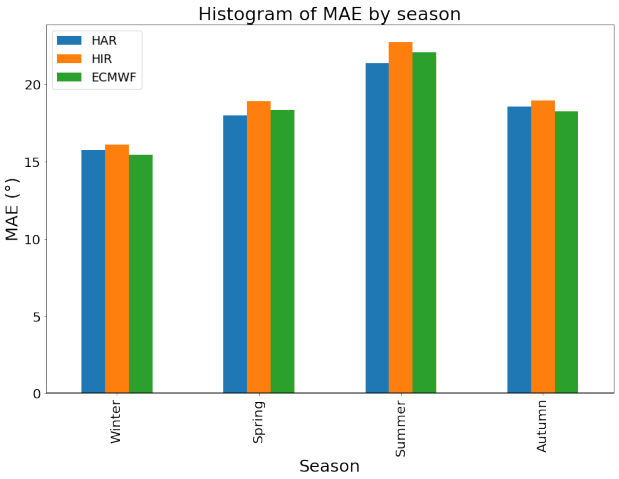
(e) MAE with wind direction (all lead times)



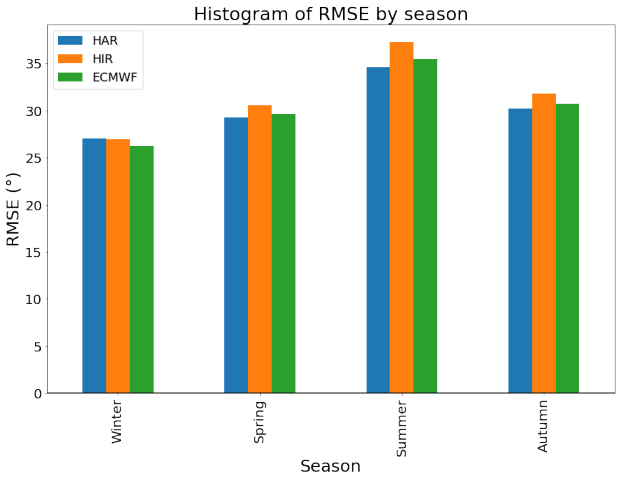
(f) RMSE with wind direction (all lead times)

Figure 7: *The dependency of the MAE and RMSE of the forecasted wind direction on certain factors is shown. Those results are for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

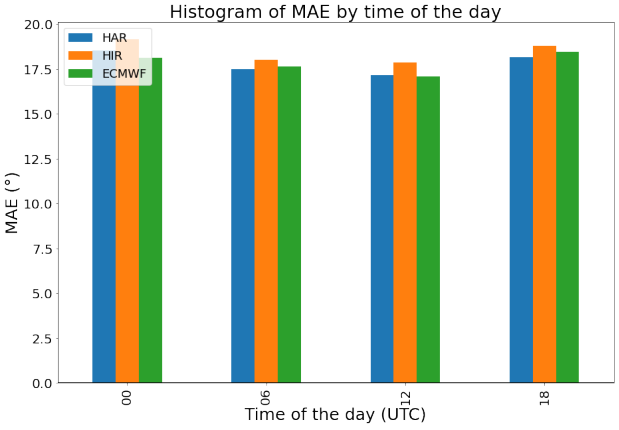
Dataset without SKV



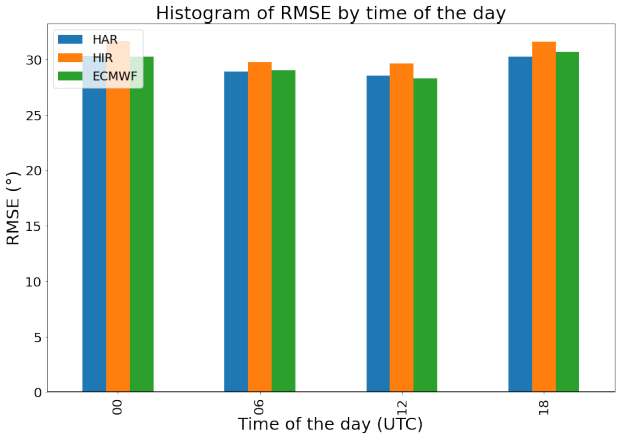
(a) MAE with season (all lead times)



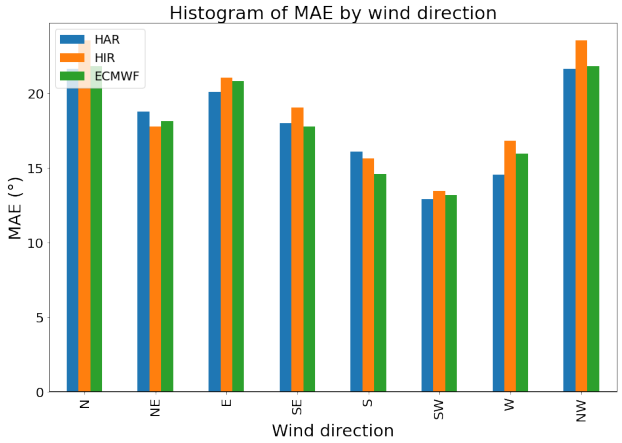
(b) RMSE with season (all lead times)



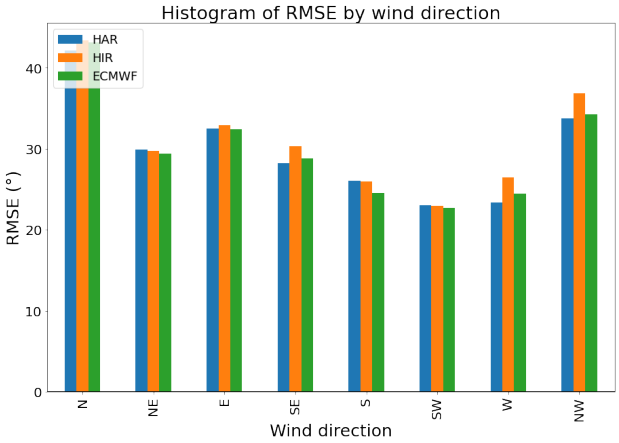
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



(e) MAE with wind direction (all lead times)



(f) RMSE with wind direction (all lead times)

Figure 8: *The dependency of the MAE and RMSE of the forecasted wind direction on certain factors is shown. Those results are for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

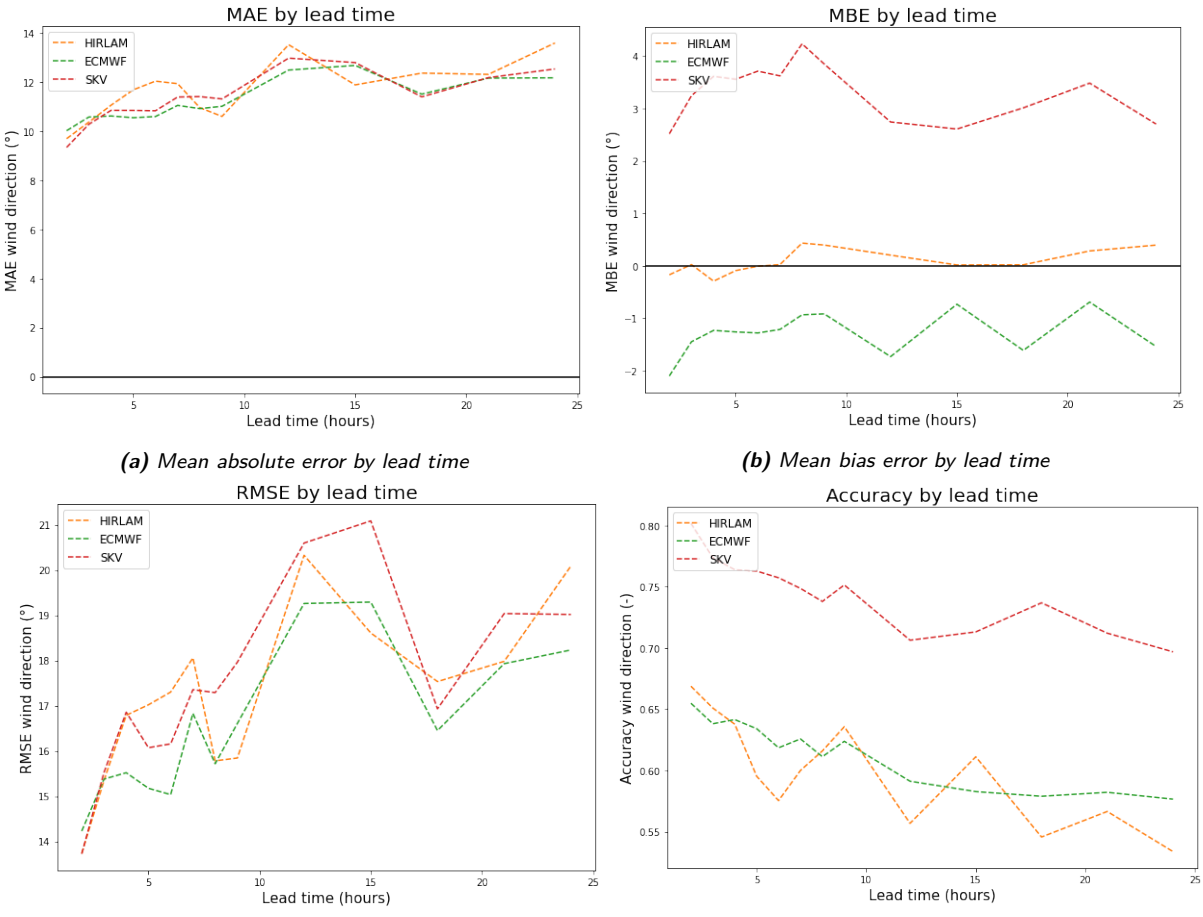
On the previous page, the dependencies of two verification metrics (MAE and RMSE) on the valid season, time of the day of the TAFGs/SKV and the wind direction itself are visualised.

For the dependency on the season, both datasets show similar results. The lowest MAE/RMSE is found during autumn and winter, whereas the spring and especially summer show higher values. A possible explanation for this is the fact that during spring and summer the average wind speed is lower than during autumn and winter, because of the stronger jet stream and stronger average pressure differences near the surface during autumn and winter. When the average wind speed is low, the wind direction is often more variable and therefore the chance of errors to occur and the magnitude of the errors is expected to increase. This is being confirmed by Figure 6. Moreover, the wind direction can be measured less easily when wind speeds are low. For the dependency of the verification metrics on the time of the day, both datasets show different results. The dataset without HARMONIE shows that during evening and night the errors are significantly higher than the errors during early morning and noon. For this specific observation, the same explanation holds. During evening and night, the turbulence is being suppressed and the average wind speed will be lower leading to larger deviations for the forecasted wind direction. During day, the sun will heat the land surface and creates turbulence and with that often a larger wind speed and a less variable wind direction leading to a smaller error. For the next time, it is interesting to find out whether the process of creating/suppressing (mechanic) turbulence is captured well by the models, since this is an important factor for creating wind (energy) (Holtslag et al., 2013). Considering the dependency of the verification metrics on the observed wind direction itself, not a clear pattern could be found. The most occurring wind direction in the region of Schiphol, W/SW, has in general the lowest MAE/RMSE. For a winds of northwest and veering up to southeast, the MAE/RMSE is in general the highest. However, one has to be keep in mind that of course each wind direction occurs more often than the other, so the bars from each wind direction are based on a different amount of data (inhomogeneity).

3.1.4 Verification metrics corrected for low wind speed

As indicated in Figure 6, the error seems to be largely dependent on the average wind speed, which is logical. The lower wind speeds are less important for the air traffic management team to deal with. Therefore, the results will give a better overview when the cases with a lower wind speed are filtered out. In consultation with Nico Maat, all the cases with an average wind speed < 7 knots are filtered out and the results of those analyses are shown below.

Dataset without HARMONIE



(a) *Mean absolute error by lead time*

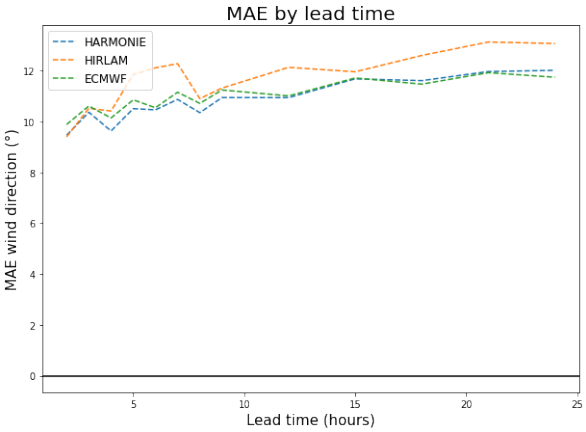
(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

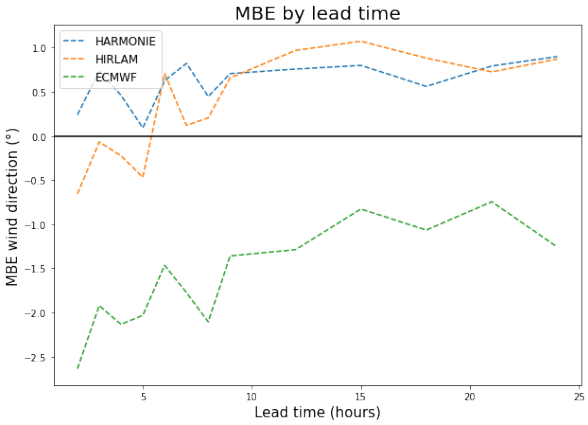
(d) *accuracy by lead time*

Figure 9: *The four verification metrics used in order to quantify the validation for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03. For those plots, only the occurrences with an average wind speed $>$ 7 knots are shown.*

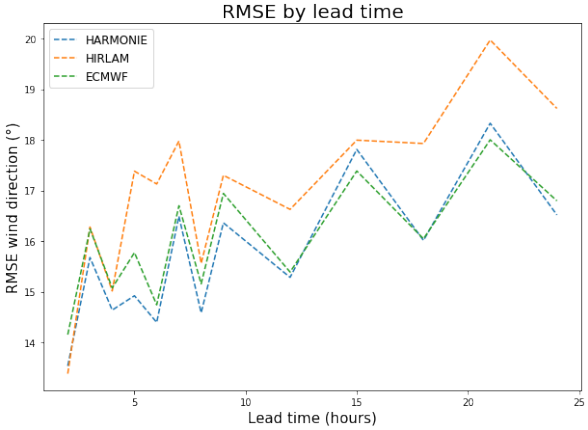
Dataset without SKV



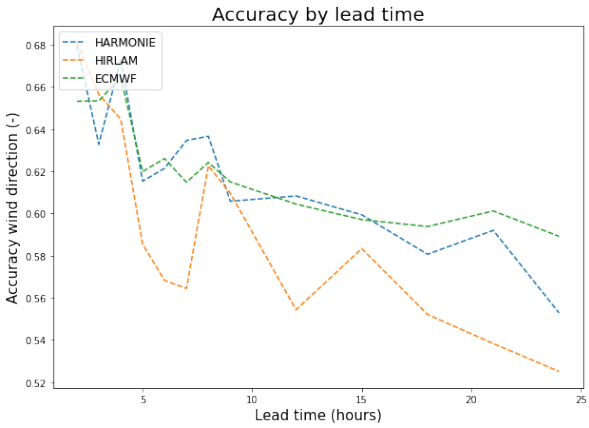
(a) Mean absolute error by lead time



(b) Mean bias error by lead time



(c) Root mean squared error by lead time



(d) accuracy by lead time

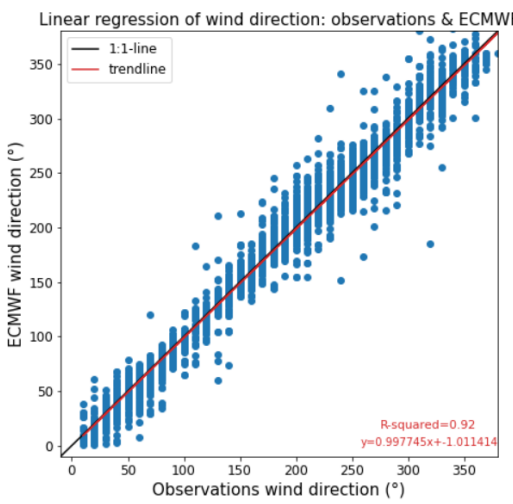
Figure 10: *The four verification metrics used in order to quantify the validation for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31. For those plots, only the occurrences with an average wind speed > 7 knots are shown.*

When filtering out the lower wind speeds (<7 kts), the verification metrics by lead time show quite some different results as found for the original dataset. Still, the SKV shows the highest mean bias error. The TAFG of HIRLAM looks almost perfectly with a MBE around zero. ECMWF seems to 'underestimate' the wind direction with a negative mean bias error. The same order of decreasing MBE was found for the analyses with the original dataset, however, the errors (MBE) are all lower. For the other verification metrics, the same pattern was found. In the second dataset used, without SKV, it can be found that ECMWF again has by far the lowest mean bias error. HIRLAM and HARMONIE are rather comparable with a mean bias error slightly above zero. More or less the same pattern was found for the other verification metrics as was found by using the original (complete) dataset where no filtering based on the observed average wind speed was used.

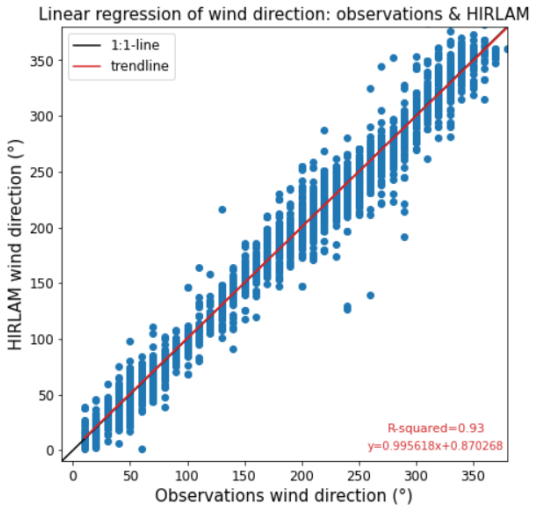
3.1.5 Linear regression plots corrected for low wind speed

Because of the same reason as mentioned earlier, the dataset is filtered based on the average wind speed using the same way as explained previously. Below, the linear regression plots of this specific dataset are shown. Those give a better indication of the error in the wind direction when the average wind speed is not negligible.

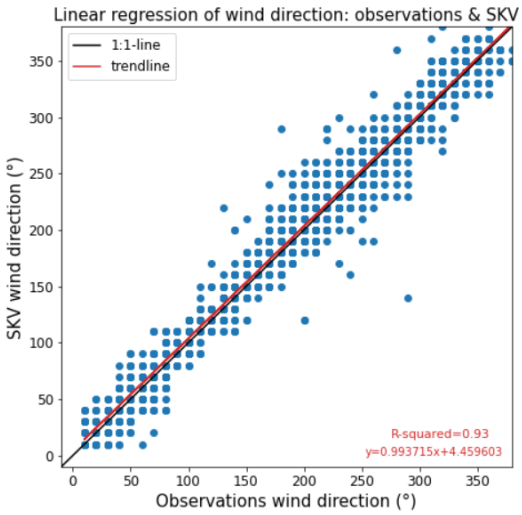
Dataset without HARMONIE



(a) ECMWF vs. observations (lead time of 3 hours)



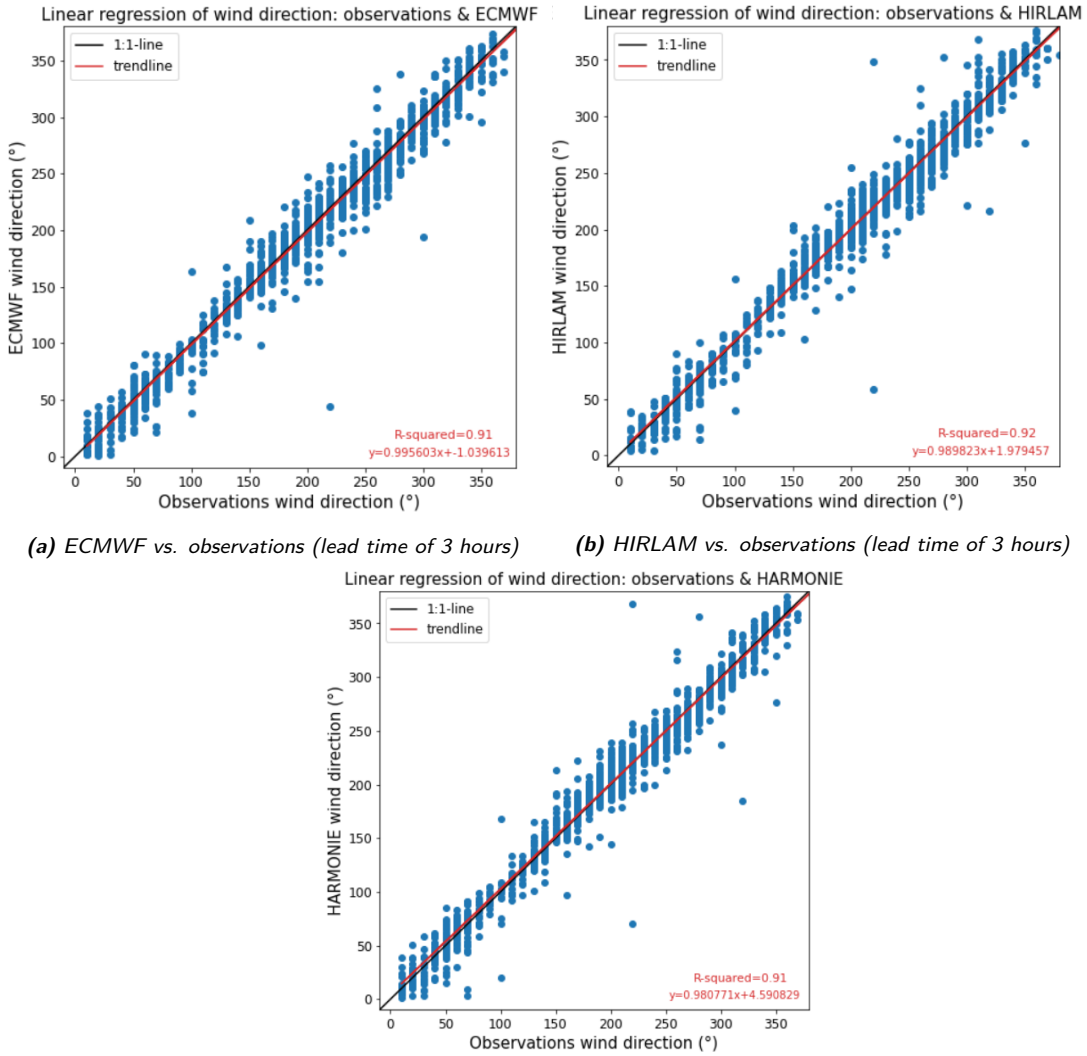
(b) HIRLAM vs. observations (lead time of 3 hours)



(c) SKV vs. observations (lead time of 3 hours)

Figure 11: *The linear regression plots of the wind direction of the TAFG with a lead time of 3 hours of a certain model against observational data for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03. For those plots, only the occurrences with an average wind speed > 7 knots are shown.*

Dataset without SKV



(a) *ECMWF vs. observations (lead time of 3 hours)*

(b) *HIRLAM vs. observations (lead time of 3 hours)*

(c) *HARMONIE vs. observations (lead time of 3 hours)*

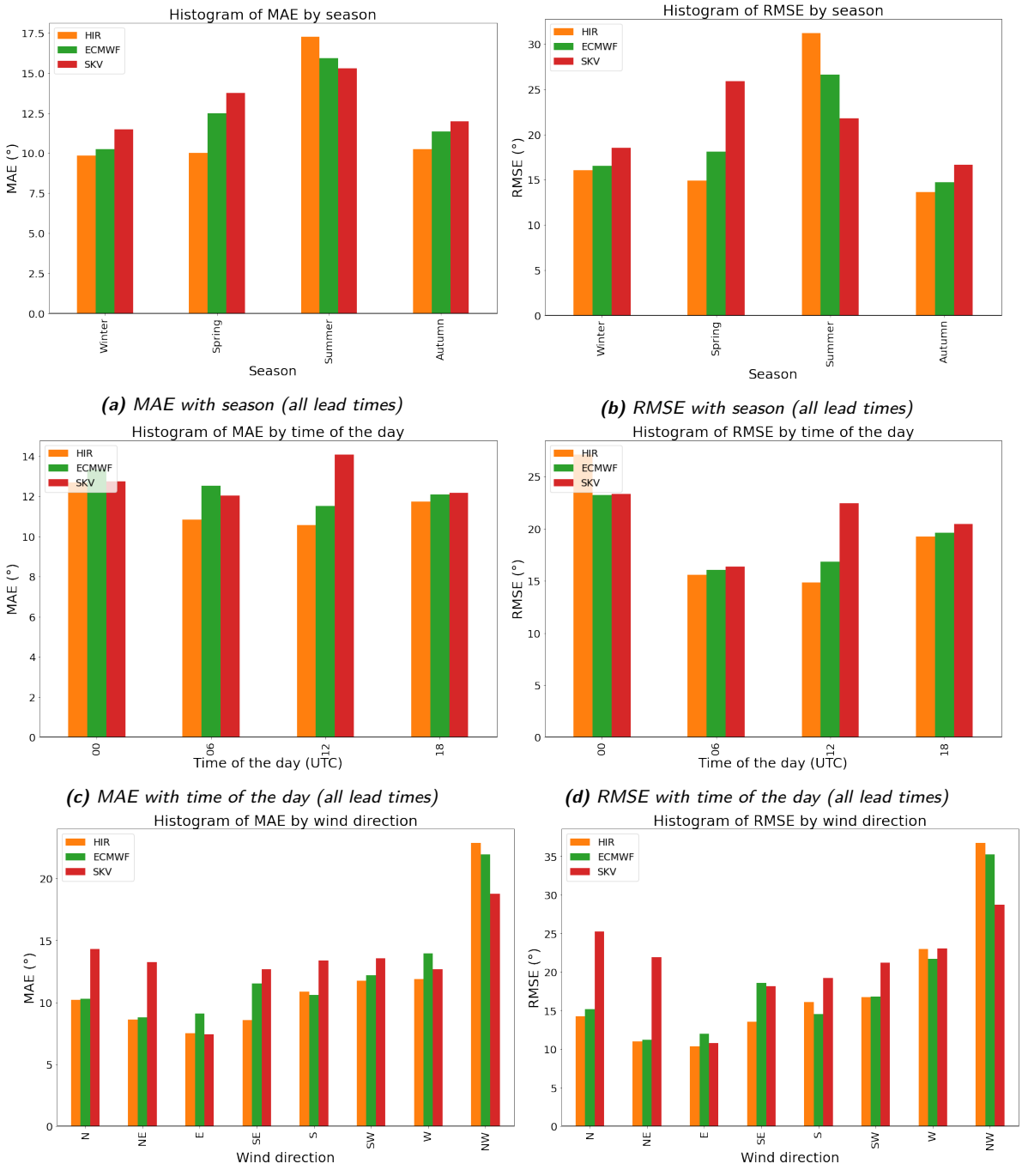
Figure 12: *The linear regression plots of the wind direction of the TAFG/SKV with a lead time of 3 hours of a certain model against observational data for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31. For those plots, only the occurrences with an average wind speed $>$ 7 knots are shown.*

Interesting is that by filtering the cases with a low average wind speed out of the data, the number of points which are located relatively far away from the 1:1-line decrease, which is what was expected based on the correlation found between the error in the forecasted wind direction and the observed average wind speed. Based on the linear regression plots shown above, hardly any differences could be observed between the different TAFGs and the SKV.

3.1.6 Dependencies corrected for low wind speed

Because of the same reason as mentioned earlier, the dataset is filtered based on the average wind speed using the same threshold. Below, the plots of those analyses are shown.

Dataset without HARMONIE



(a) *MAE with season (all lead times)*

(b) *RMSE with season (all lead times)*

(c) *MAE with time of the day (all lead times)*

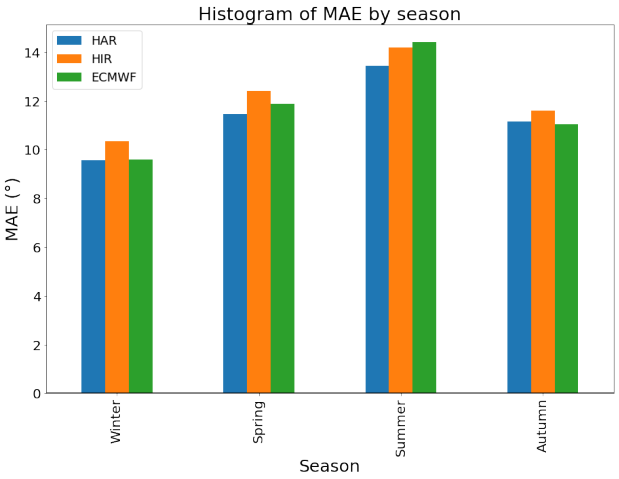
(d) *RMSE with time of the day (all lead times)*

(e) *MAE with wind direction (all lead times)*

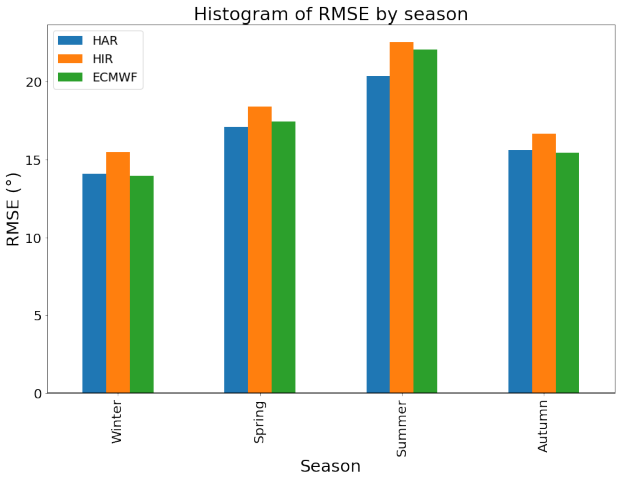
(f) *RMSE with wind direction (all lead times)*

Figure 13: *The dependency of the MAE and RMSE of the forecasted wind direction on certain factors is shown. Those results are for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03. For those plots, only the occurrences with an average wind speed > 7 knots are shown.*

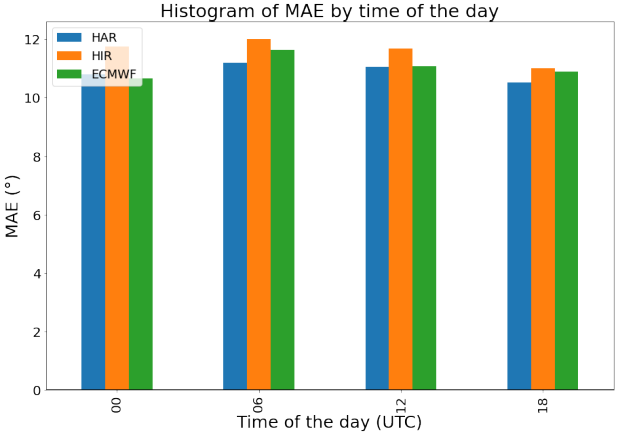
Dataset without SKV



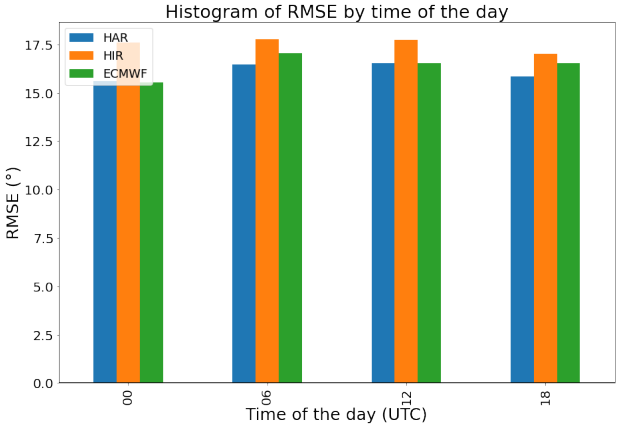
(a) MAE with season (all lead times)



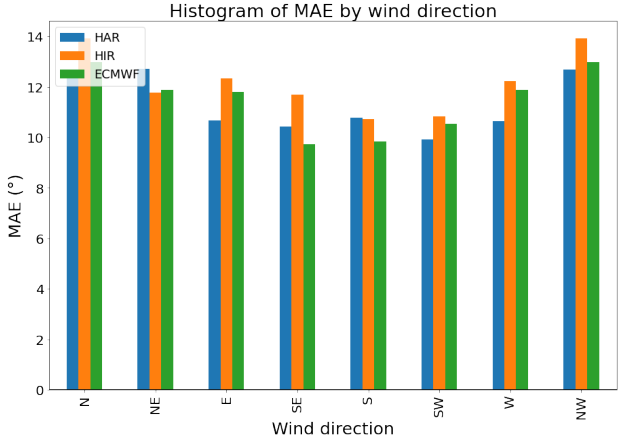
(b) RMSE with season (all lead times)



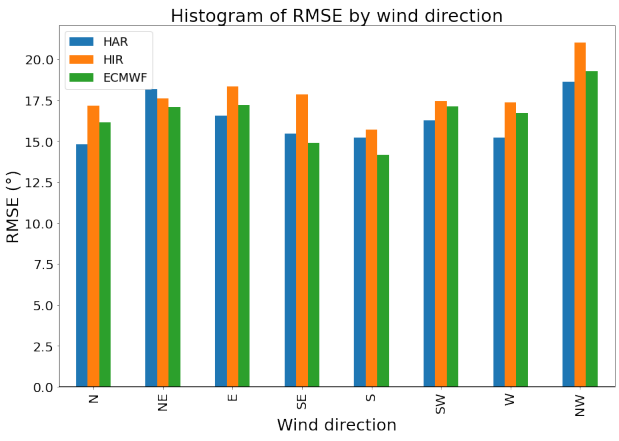
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



(e) MAE with wind direction (all lead times)



(f) RMSE with wind direction (all lead times)

Figure 14: *The dependency of the MAE and RMSE of the forecasted wind direction on certain factors is shown. Those results are for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31. For those plots, only the occurrences with an average wind speed $>$ 7 knots are shown.*

Interesting is that there is still no clear TAFG or SKV which has consistently the highest MAE or RMSE. The thing that attracts the attention is that the effect that the MAE/RMSE is higher during the night and evening, because of the lower wind speeds, is as good as disappeared. This proves that the cause of this finding in the original dataset are the lower wind speeds during the (late) evening and night. Interesting is however that the seasonal effect is still visible. It was hypothesised that this effect was also because of the in general lower wind speeds during spring and summer. However, this effect is not disappeared by filtering the lower average wind speeds out of the dataset. Considering the dependency of the MAE/RMSE on the wind direction, a bit of a clearer pattern can be found. Apparently the errors are higher when the wind is originating from the north(west). However, one has to be keep in mind that of course each wind direction occurs more often than the other in the dataset, so the bars from each wind direction are based on a different amount of data (inhomogeneity).

3.1.7 Statistical tests

In this section, the results of the independent sample t-tests are shown for both datasets of the wind direction of the TAFG/SKV and observational data. Tested is whether the model differs from observations in both directions (two-sided). A significance level of p=0.05 is used. The data used contains the data of a lead time of 3 hours. In the table below, the results of the original dataset is shown. In the second table, the results of the t-test with the dataset where the lower wind speeds are filtered out, are shown.

Table 1: *Table showing the p-value of the independent sample t-tests*

P-value	HARMONIE	HIRLAM	ECMWF	SKV
Dataset without HARMONIE	X	0.512	0.984	0.0560
Dataset without SKV	0.247	0.482	0.971	X

Table 2: *Table showing the p-value of the independent sample t-tests (average wind speed $>$ 7 kts)*

P-value	HARMONIE	HIRLAM	ECMWF	SKV
Dataset without HARMONIE	X	0.992	0.534	0.163
Dataset without SKV	0.830	0.984	0.576	X

For the first table, the tests show no significant differences between the SKV and different TAFGs compared to the observational data. A note has to be made that the P-value belonging to the SKV compared to observational data is on the edge of a confirmed significant difference between the datasets and also way smaller compared to the P-value of the TAFGs. This corresponds with the found result that the MBE of the SKV is by far the highest compared to the error of the TAF Guidances.

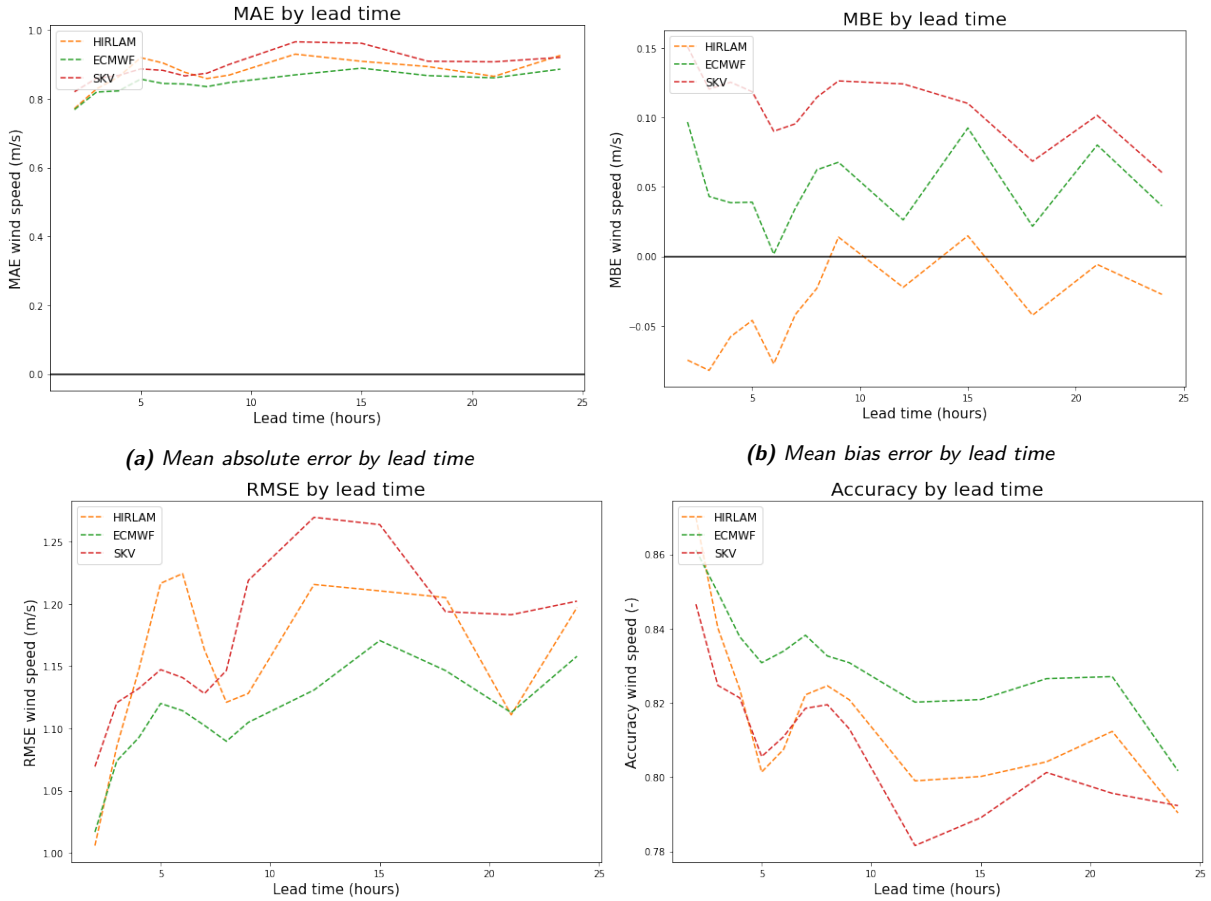
The second dataset shows that the P-values almost for all cases, except for the TAFG of ECMWF, increased. This effect was also expected, since the error decreased by filtering out the cases with a lower average wind speed. Still, the SKV has the P-value which is lowest and thus closest to a significant difference compared to the observations.

3.2 Average wind speed

In this section, the same type of results are shown as in the previous pages, but now for the average wind speed.

3.2.1 Verification metrics

Dataset without HARMONIE



(a) *Mean absolute error by lead time*

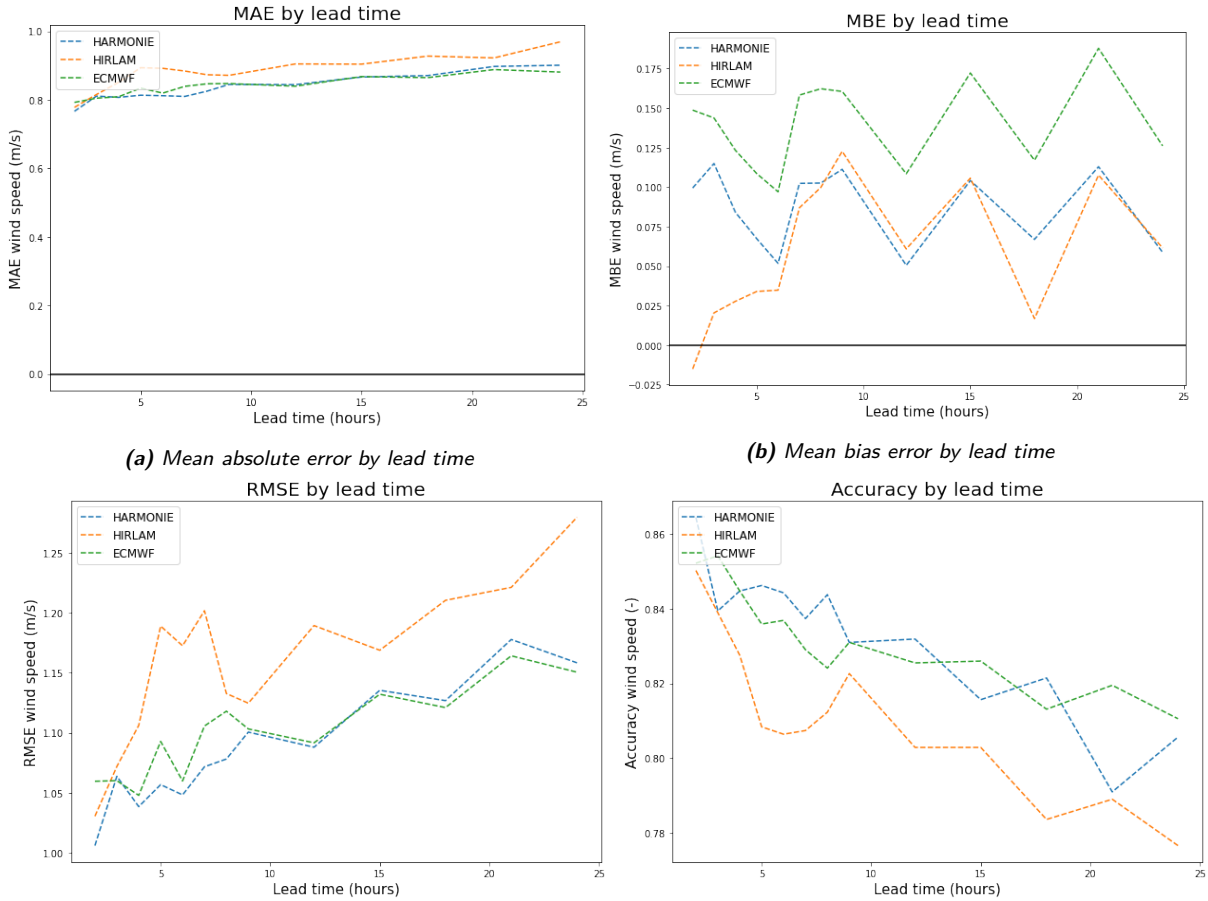
(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

Figure 15: *The four verification metrics used in order to quantify the validation for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

Dataset without SKV



(a) *Mean absolute error by lead time*

(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

Figure 16: *The four verification metrics used in order to quantify the validation for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

The mean absolute error is approximately 1.8 knots. This error is more acceptable than the error of the wind direction (showing sometimes an error of 20°) for air traffic management to deal with as discussed with the host supervisor. As was also found in the validation of the wind direction, the MAE does not show a lot of differences between the TAFGs and the SKV. In the dataset without SKV, Figure 16a, the TAFG of HIRLAM is most of the time the one with the highest absolute error. In the dataset without HARMONIE, SKV is most of the time the one with the highest MAE. However, the absolute differences are small.

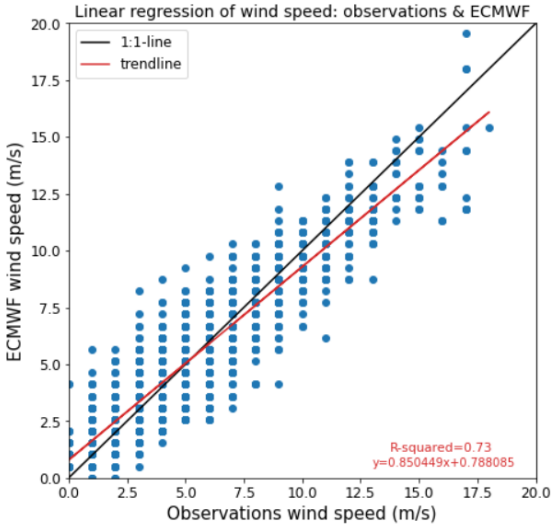
The MBE leads to a slightly different view. The SKV has the highest mean bias error (overestimation), followed by an equal distance from the ideal 'zero-MBE-line' of ECMWF (overestimation) and HIRLAM, which seems to structurally underestimate the average wind speed (a known problem). In the second dataset, the dataset without SKV, the errors are a bit larger when comparing the values on the different y-axes. Again, ECMWF shows the highest deviation, followed by an equal error of HIRLAM and HARMONIE. Interesting is that also here, an overestimation of the average wind speed is dominant. The third verification metric, the RMSE, shows in contrast to what was found in the validation of the wind direction, clearer discrepancies. Most of the time, the SKV has the highest RMSE. The RMSE of ECMWF is slightly smaller than the one of HIRLAM. In the dataset without SKV, HIRLAM has by far the highest

RMSE, followed by an equal RMSE of HARMONIE and ECMWF. Thus, whereas ECMWF and HIRLAM showed a comparable MBE, the RMSE differs where ECMWF/HARMONIE have a smaller RMSE.

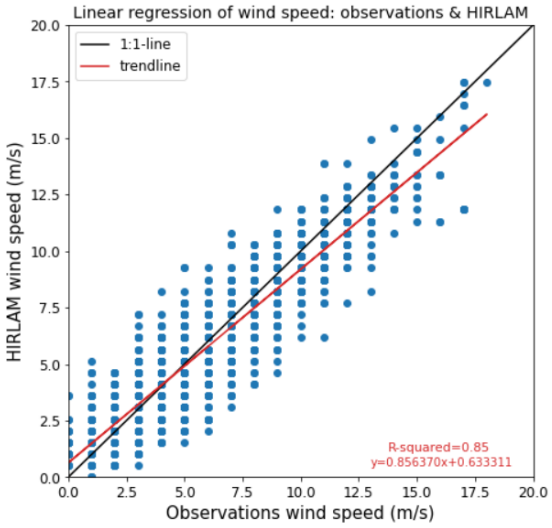
Based on the combination of the MAE and the RMSE, the SKV seems to perform worst, followed by HIRLAM and then ECMWF (and HARMONIE). In contrast to the previous variable which was validated, the accuracy shows what can be expected based on the other verification metrics. ECMWF and HARMONIE have the highest accuracy and the SKV in general the lowest. Summarised, HIRLAM and SKV have a comparable accuracy which is significantly lower than the accuracy of ECMWF/HARMONIE.

3.2.2 Linear regression

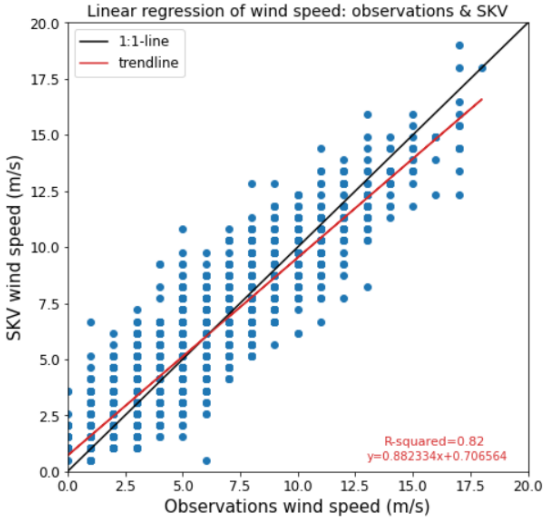
Dataset without HARMONIE



(a) ECMWF vs. observations (lead time of 3 hours)



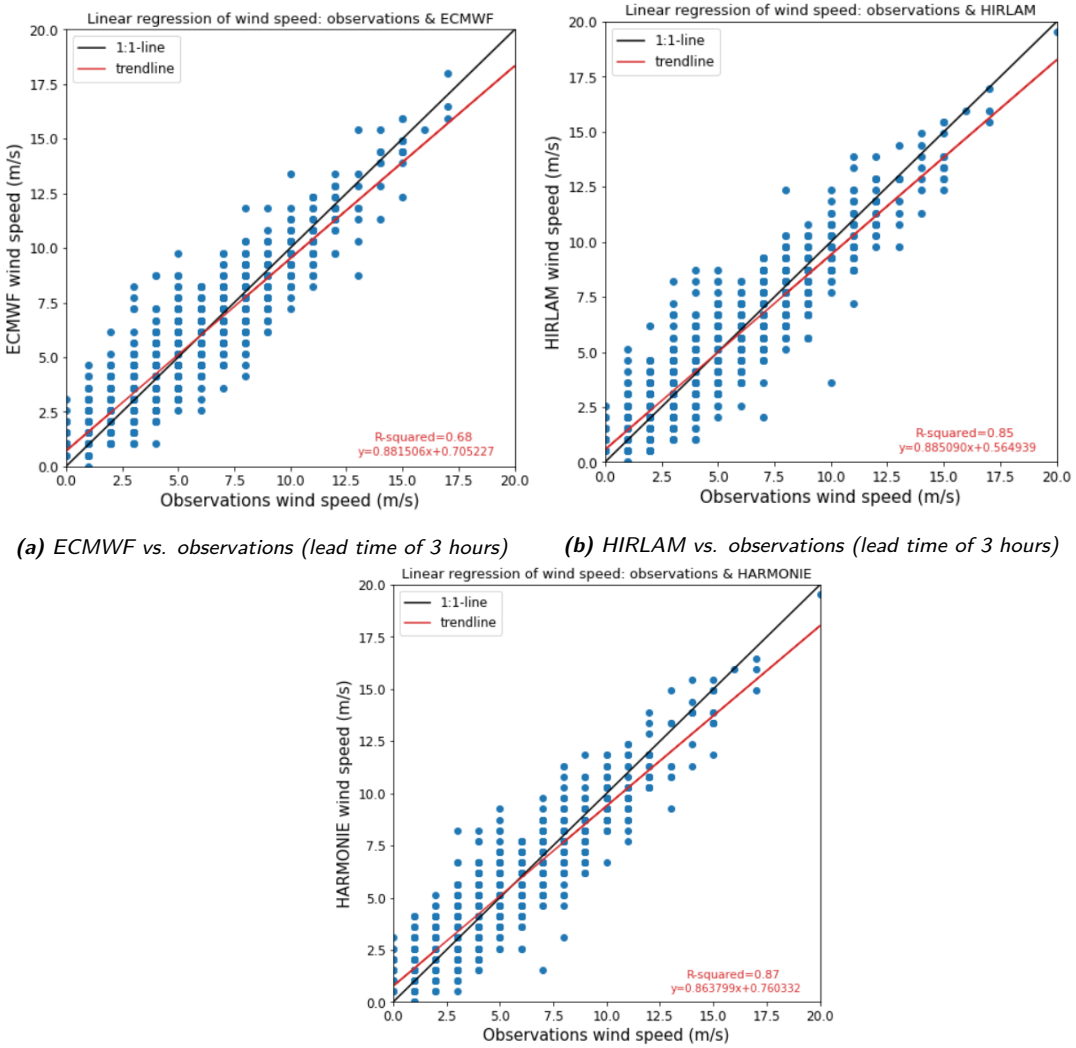
(b) HIRLAM vs. observations (lead time of 3 hours)



(c) SKV vs. observations (lead time of 3 hours)

Figure 17: *The linear regression plots of the average wind speed of the TAFG with a lead time of 3 hours of a certain model against observational data for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

Dataset without SKV



(a) *ECMWF vs. observations (lead time of 3 hours)*

(b) *HIRLAM vs. observations (lead time of 3 hours)*

(c) *HARMONIE vs. observations (lead time of 3 hours)*

Figure 18: *The linear regression plots of the average wind speed of the TAFG/SKV with a lead time of 3 hours of a certain model against observational data for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

The first note that has to be made for the plots shown above is that one point represents more points than other ones. Therefore, the deviations seem relatively large in the first place compared to the MAE found previously. It is clear that for all the TAFGs and the SKV the slope is clearly smaller than 1. In this analysis a linear regression was used, since ideally the line would perfectly fit using a linear regression with a slope of 1. Due to time limitations, no more type of regressions were used. However, it would be logical for this variable to try a polynomial regression on it too since that would seem to fit better. It seems that the lower average wind speeds, roughly in the range of 0 m/s up to 10 m/s, show quite a good linear fit. However, for the very low wind speeds (0 m/s up to 5 m/s), an overestimation compared to the observed average wind speed seems to be present (especially for ECMWF where more points are above the 1:1-line). For larger wind speeds, more and more points seem to end up below the 1:1-line, meaning that the specific TAFG or SKV underestimated the average wind speed. In the dataset without HARMONIE,

this is especially clear for the TAFG of ECMWF. Those discrepancies can be recognised by analysing the R-squared. The lowest R-squared is by far the one of the linear regression with the TAFG of ECMWF, in both datasets. For the SKV and the TAFG of HIRLAM and HARMONIE, the R-squared is comparable. The slope itself is lowest for ECMWF, corresponding to the 'worst' fit shown by the R-squared.

This overestimation was also visible in the analyses of the MBE with lead time. However, considering the magnitude of the error, the other verification metrics showed a bit different results. For the variable validated in this section, the average wind speed, the same holds. Summarised, the combination of MAE and RMSE indicate that the SKV performs worst, followed by HIRLAM and then ECMWF/HARMONIE. The MBE also shows the largest error for the SKV (0.10-0.15 m/s). However, the MBE shows that HIRLAM performs slightly better than ECMWF/HARMONIE (closer to 0 m/s). The linear regression plots as described above show by far the worst fit for ECMWF, followed by a rather equal fit of HIRLAM/HARMONIE/SKV when looking at the R-squared. The R-squared of ECMWF is around 0.70, which is significantly lower than the other ones (around 0.85). The regression plot of the SKV has a slope of 0.88, closest to 1, indicating that the work of the meteorologists can be expected to be useful in order to prevent underestimating larger wind speeds. The general overestimation of the average wind speed (especially of the relatively low wind speeds) is also clearly visible in the histograms of error distributions in Appendix A. The MBE shows in general an overestimation, because the lower wind speeds occur more often than the more extreme ones (which are in general underestimated). In Appendix B, a time series of the monthly moving average of both MAE and MBE is shown. The first thing that attracts attention is that there is not a dominant forecasting data source that performs worst on forecasting the average wind speed, based on those two verification metrics. A rather chaotic pattern characterises the moving average plots of this specific variable. Furthermore, the moving average of the mean bias error does not show a clear over- or underestimation by the different TAFGs (and SKV). As already discussed above, this over- or underestimation was dependent on the actual wind speed itself, which depends on time. Figure 19 shows the dependency of the bias on the observed wind speed of the SKV (the TAFGs showed roughly the same dependency). The general overestimation for the lower wind speeds (positive bias) and the underestimation for higher wind speeds (negative bias) cause the slope to be less than 1 in the regression analyses.

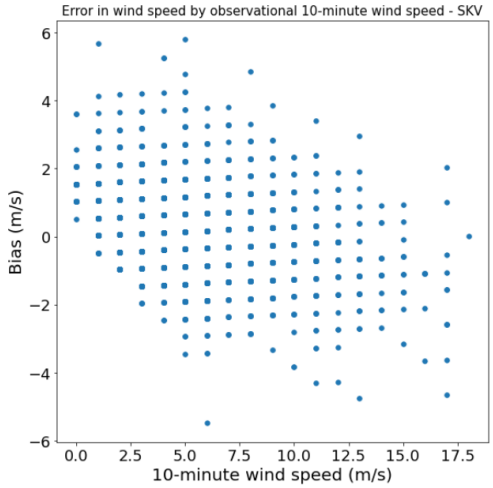
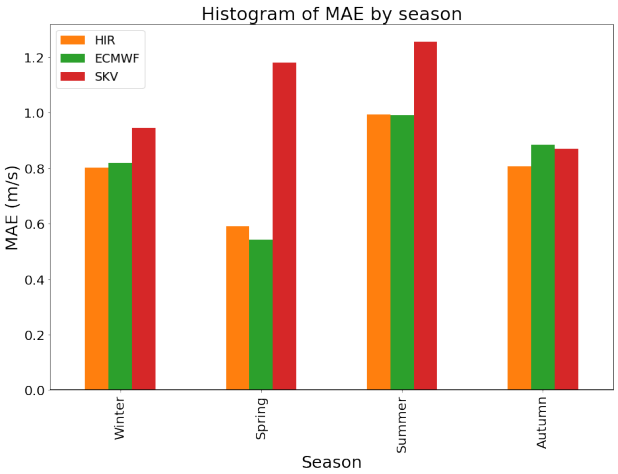


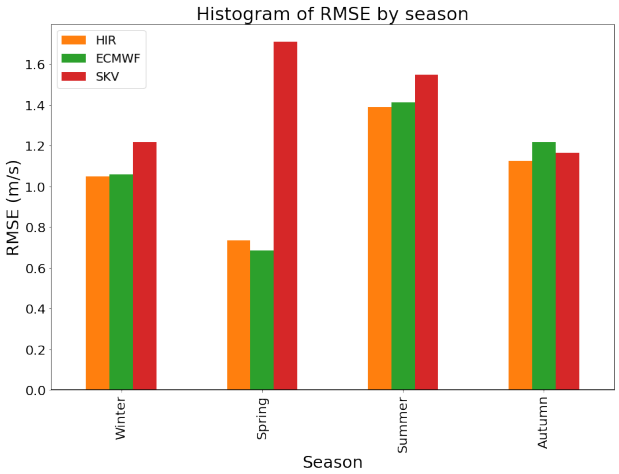
Figure 19: *The bias of the SKV of the average wind speed vs. observed 10-minute wind speed (lead time of 3 hours).*

3.2.3 Dependencies

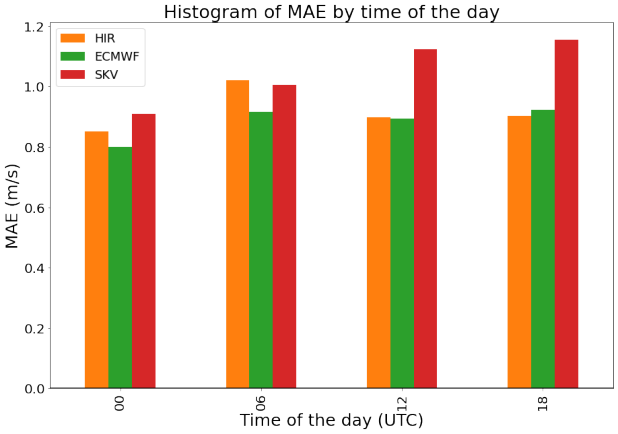
Dataset without HARMONIE



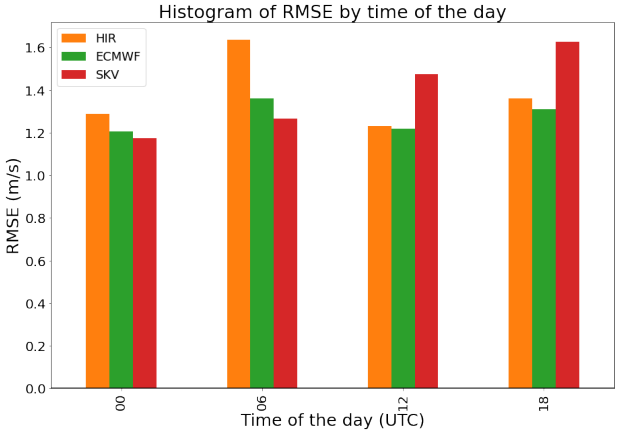
(a) MAE with season (all lead times)



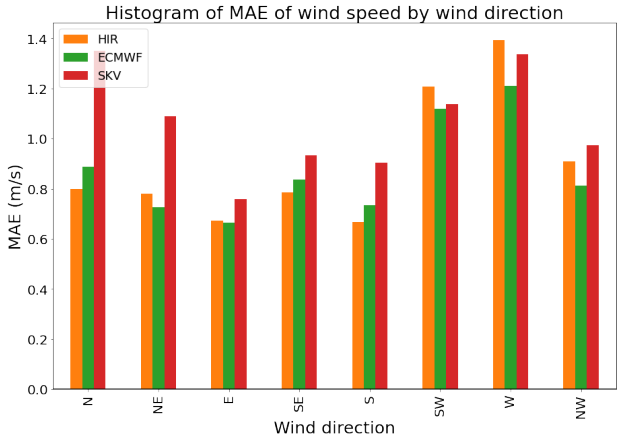
(b) RMSE with season (all lead times)



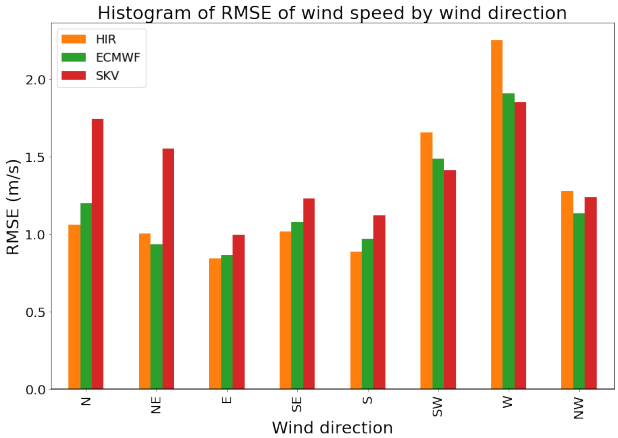
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



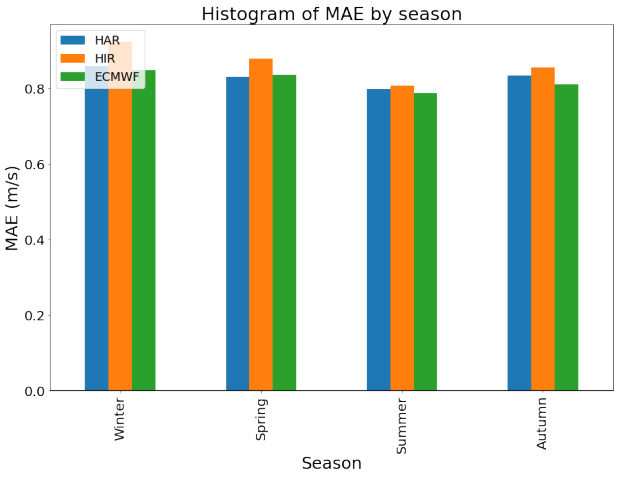
(e) MAE with wind direction (all lead times)



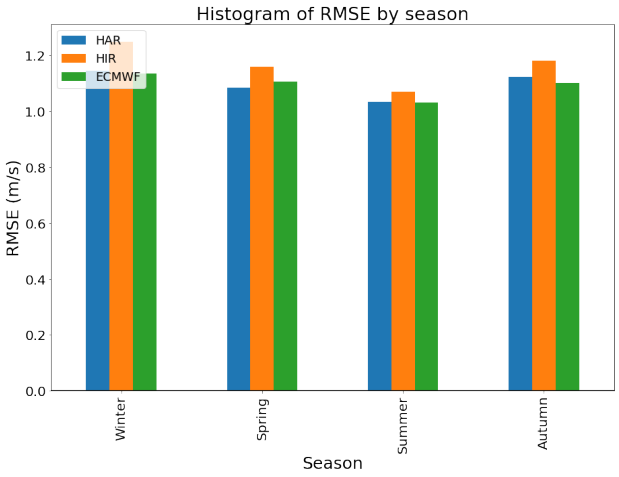
(f) RMSE with wind direction (all lead times)

Figure 20: *The dependency of the MAE and RMSE of the forecasted average wind speed on certain factors is shown. Those results are for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

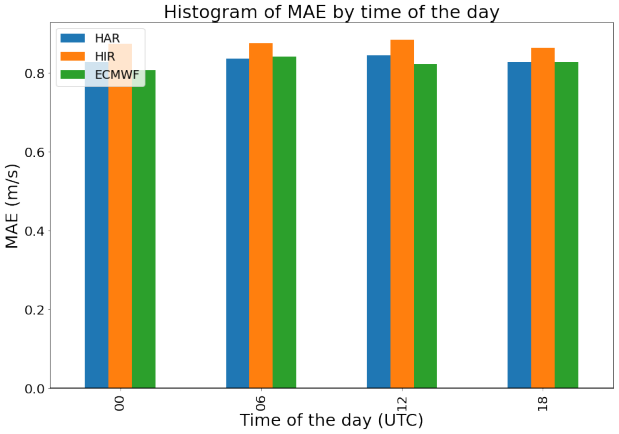
Dataset without SKV



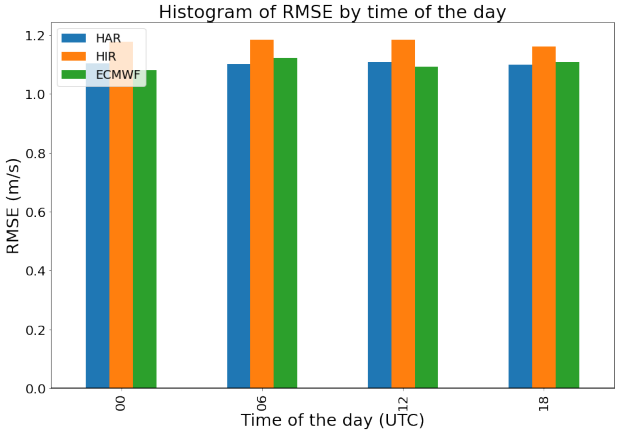
(a) MAE with season (all lead times)



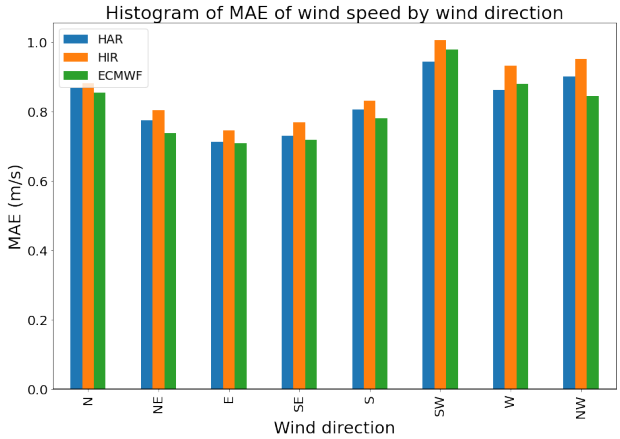
(b) RMSE with season (all lead times)



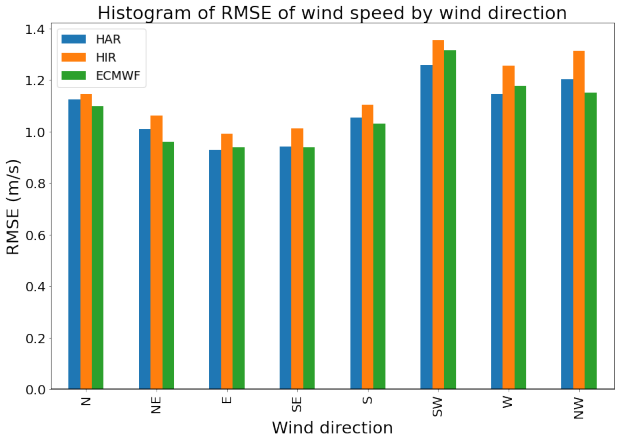
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



(e) MAE with wind direction (all lead times)



(f) RMSE with wind direction (all lead times)

Figure 21: *The dependency of the MAE and RMSE of the forecasted average wind speed on certain factors is shown. Those results are for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

Not a clear dependency of the MAE/RMSE on the valid season is found as was the case for the previously validated variable, the wind direction. The thing that attracts attention is that the SKV has almost for every season (by far) the highest MAE and RMSE, especially during spring and summer. The differences between HIRLAM, HARMONIE and ECMWF are rather small, as is also shown in the plots for the dataset without SKV.

For the dependency of the MAE and RMSE on the valid time of the day, again no clear differences are found. The differences between SKV and ECMWF/HIRLAM seem smaller, but most of the times, the SKV seems to perform worst, followed by HIRLAM.

Considering the dependency of the verification metrics on the observed wind direction itself, not a clear pattern in both datasets could be found too. Interesting is that the somewhat higher errors during a wind from the west/southwest seem to be present in both datasets. Also, larger errors during winds from the north are visible.

3.2.4 Statistical tests

In this section, again the results of the independent sample t-tests are shown for both datasets of the average wind speed of the TAFG/SKV and observational data. Tested is whether the model differs from observations in both directions (two-sided). The same significance level (p=0.05) is used and again the data used contains the forecasting data with a lead time of 3 hours.

Table 3: *Table showing the p-value of the independent sample t-tests*

P-value	HARMONIE	HIRLAM	ECMWF	SKV
Dataset without HARMONIE	X	0.153	0.451	0.0392
Dataset without SKV	0.178	0.813	0.0951	X

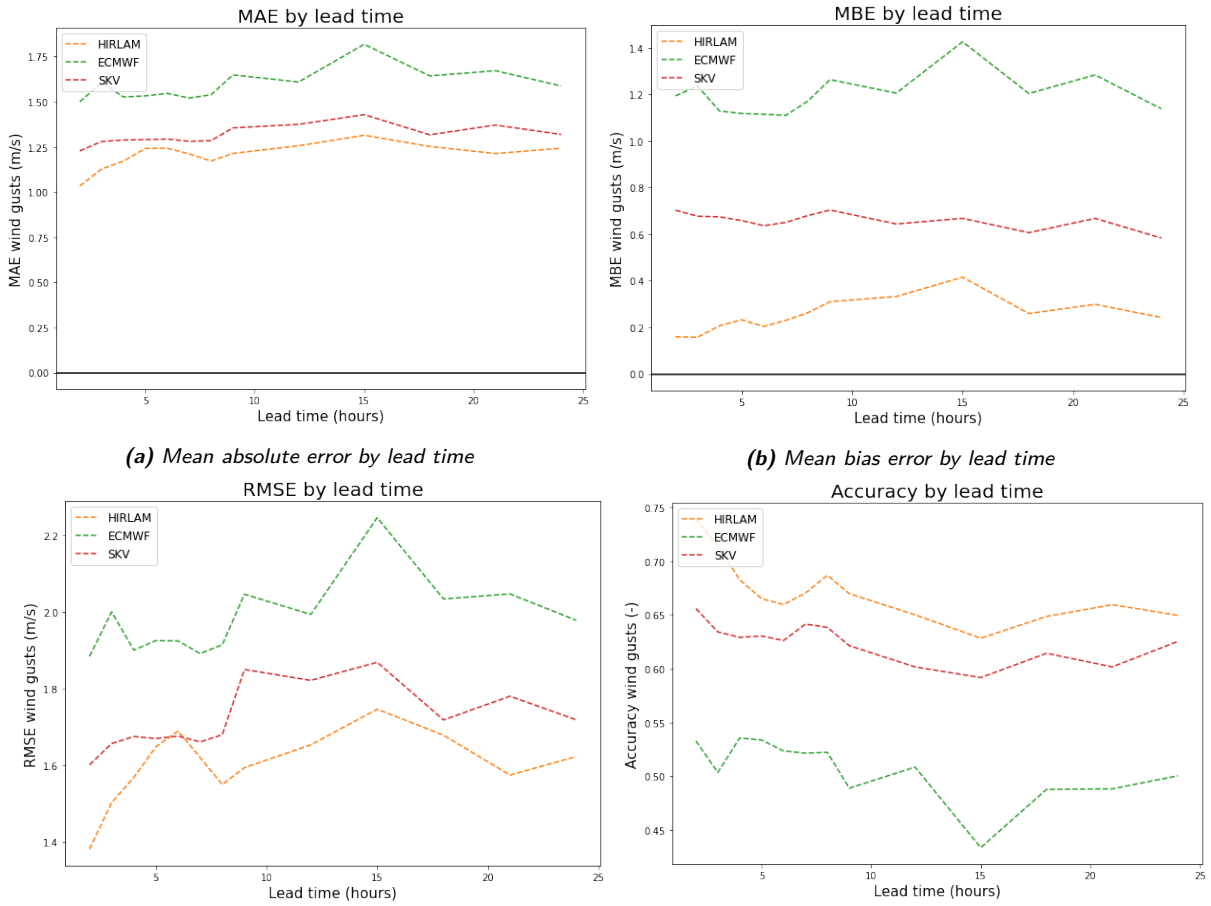
The statistical tests show for all TAFGs no significant difference compared to observational data. For the SKV however, a significant different result is found compared to observational data. The larger deviation of the SKV compared to observational data considering the average wind speed is something that was discussed previously. It was found back in both the MAE, the MBE, the RMSE and the accuracy.

3.3 Wind gusts

In this section, the same type of results are shown as in the previous pages, but now for the maximum wind gusts.

3.3.1 Verification metrics

Dataset without HARMONIE



(a) *Mean absolute error by lead time*

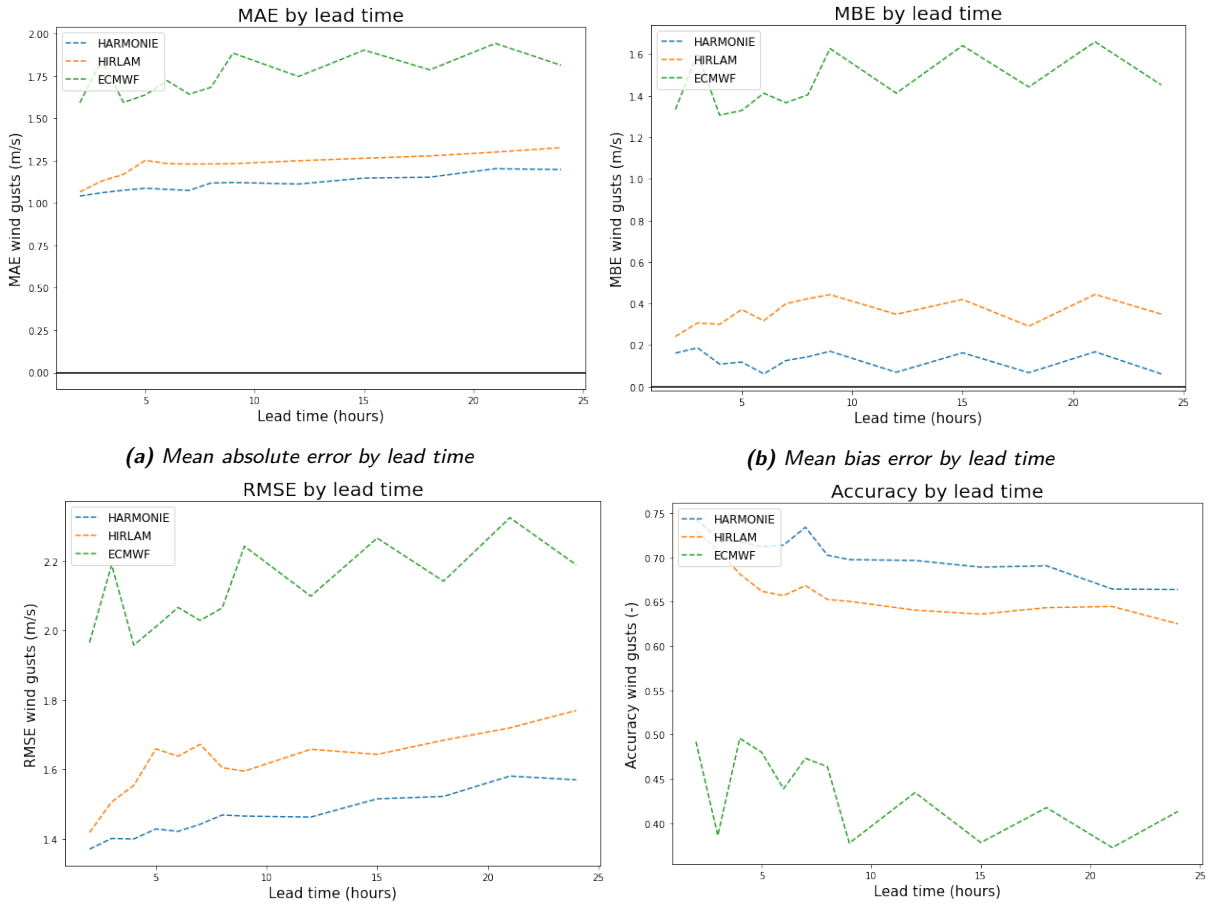
(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

Figure 22: *The four verification metrics used in order to quantify the validation for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

Dataset without SKV



(a) *Mean absolute error by lead time*

(b) *Mean bias error by lead time*

(c) *Root mean squared error by lead time*

(d) *accuracy by lead time*

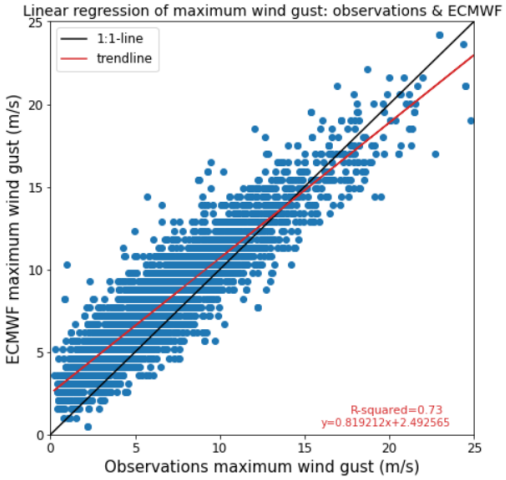
Figure 23: *The four verification metrics used in order to quantify the validation for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

The results of the four verification metrics with lead time show, partly in contrast to the ones of the previous two variables, very consistent results. The mean absolute error ranges from 1 up to 1.75 m/s (2 up to about 3.5 knots). The TAFG of ECMWF shows for all lead times, for both datasets, by far the highest mean absolute error. The same holds for the mean bias error. Considering the mean bias error, it is interesting to notice that the mean bias error is for all lead times and for all TAFGs/SKV positive, meaning that in general an overestimation of the wind gusts is present. The SKV appears to perform better than ECMWF based on those verification metrics. The models for the short-term weather forecasts appear to be most correct considering the forecast of wind gusts for Schiphol. HARMONIE has the lowest MAE and MBE. Exactly the same order of error magnitude can be noticed in the plot of the RMSE and accuracy by lead time. Especially in the second dataset it is clear that ECMWF has by far the least accurate performance when it comes to forecasting wind gusts.

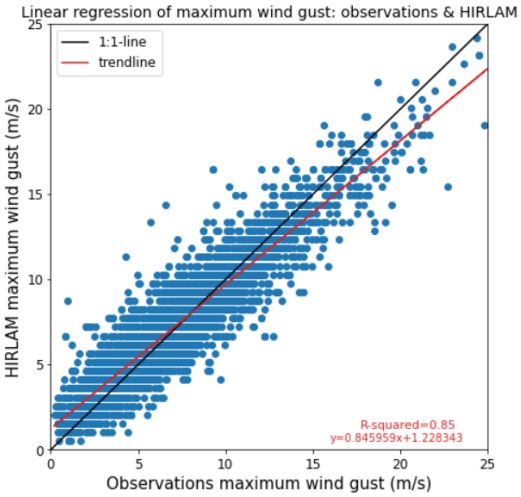
Interesting is that, in contrast to for instance the validation of the wind direction, there is no clear increase of the error by lead time (what one would expect). In the second dataset a small but regular increase of the RMSE with lead time is present and also a slight decreasing accuracy. For the MAE and MBE however, this effect is not visible at all.

3.3.2 Linear regression

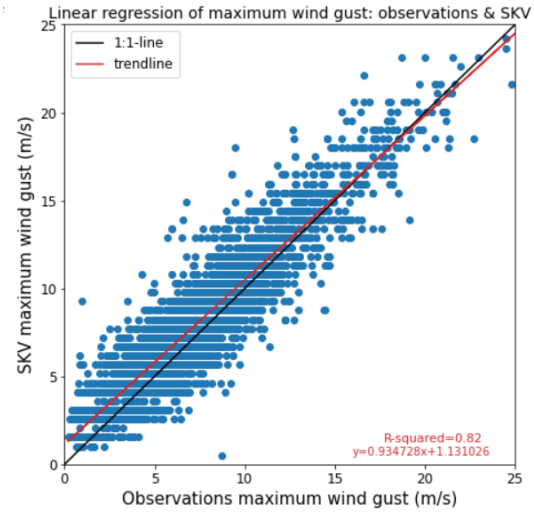
Dataset without HARMONIE



(a) ECMWF vs. observations (lead time of 3 hours)



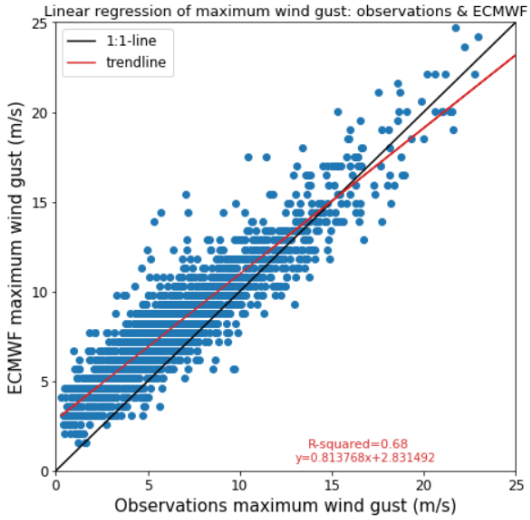
(b) HIRLAM vs. observations (lead time of 3 hours)



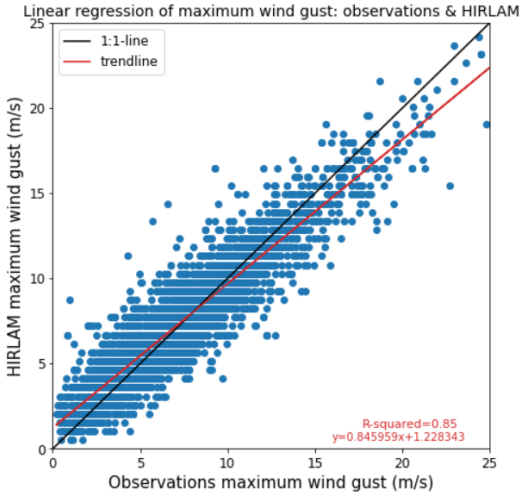
(c) SKV vs. observations (lead time of 3 hours)

Figure 24: *The linear regression plots of the wind direction of the TAFG with a lead time of 3 hours of a certain model against observational data for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

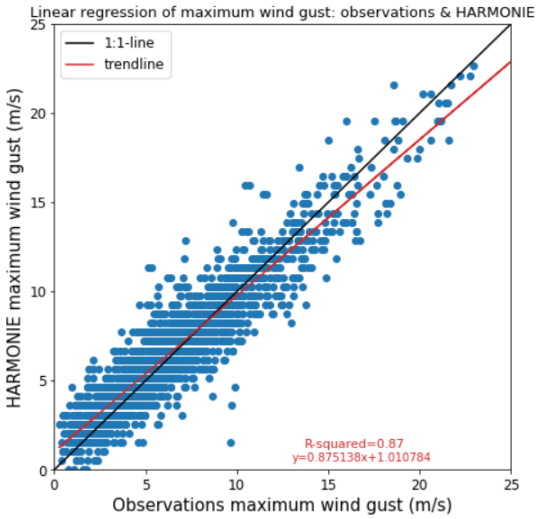
Dataset without SKV



(a) ECMWF vs. observations (lead time of 3 hours)



(b) HIRLAM vs. observations (lead time of 3 hours)



(c) HARMONIE vs. observations (lead time of 3 hours)

Figure 25: *The linear regression plots of the wind direction of the TAFG/SKV with a lead time of 3 hours of a certain model against observational data for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

The regression plots shown above visualise the results which appear to be consistent with the analyses of the verification metrics with lead time. In both datasets, it is clear that ECMWF performs again worst when considering the R-squared which is for ECMWF around 0.70. For HIRLAM/HARMONIE/SKV, the R-squared is around 0.85 with small differences mutually. Also by eye it is clear that for ECMWF most points are above the black 1:1-line, indicating the significant overestimation discussed in the previous section. Especially in the second dataset, this overestimation is clearly visible with almost all the points above the 1:1-line. The general overestimation, also shown in the figures in the previous section, is clearly visible too in the histograms of error distributions provided in Appendix A. The relatively large error of the ECMWF model is also represented by the slope. The slope of ECMWF, in both datasets, shows an error which deviates most from 1. Furthermore, it is interesting that all the slopes are smaller than 1.

It indicates that especially lower observed wind gusts are overestimated, whereas higher observed wind gusts are probably underestimated. In the first dataset, it seems that especially for ECMWF and HIRLAM, this underestimation for the higher wind speeds (>20 m/s) is clearly present. However, for the SKV, this underestimation seems to be corrected. The trendline shows a much smaller deviation (underestimation) from the 1:1-line compared to the TAFGs of ECMWF and HIRLAM. In the second dataset, the dataset without SKV, the same underestimation is found for the higher wind speeds. The magnitude of the error and also the direction (underestimation or overestimation) seems to be dependent on the magnitude of the wind gusts themselves.

In Appendix B, a time series of the monthly moving average of both MAE and MBE is shown. Exactly the same order based on the magnitude of the error is found here. For almost the complete time series, ECMWF performs least accurate. The SKV performs better than ECMWF, but worse than HIRLAM and HARMONIE. The HARMONIE TAFG appears to perform best, also based on those moving averages. Notice also the continuous overestimation of the wind gusts by the different TAFGs and SKV. For HIRLAM and even more for HARMONIE, this effect is less dominant.

Figure 26 below shows the dependency of the error (bias) of the forecasted maximum wind gust on the actual observed wind speed of the SKV (the TAFGs showed roughly the same dependency) with a lead time of 3 hours. It does not visualise a strong relationship, but there seems to be a general overestimation (positive bias) of the wind gusts when the average wind speeds are lower and an underestimation (negative bias) when the average wind speeds are higher. The same pattern was also found in the previous chapter where the average wind speed itself was validated. The MBE in Figure 23b and 22b, but also the moving averages of the MAE and RMSE shown in Appendix B, indicate in general an overestimation, because the lower wind speeds occur more often than the more extreme higher ones. That is why there is a net overestimation of the wind gusts. However, as Figure 26 shows, this is dependent on the magnitude of the wind gusts themselves where the peak winds gusts are often being underestimated.

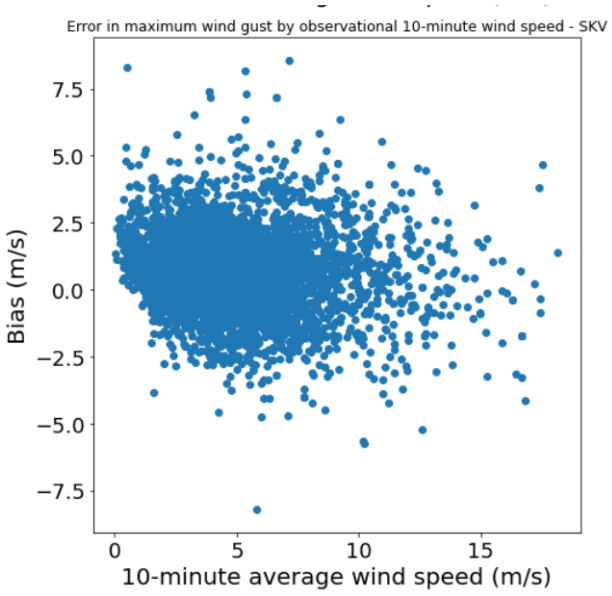
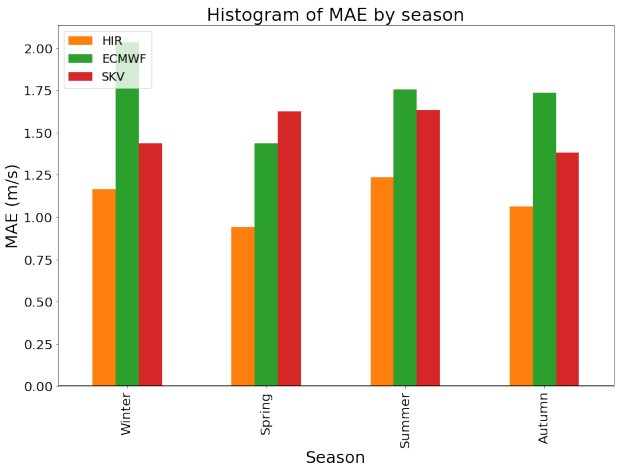


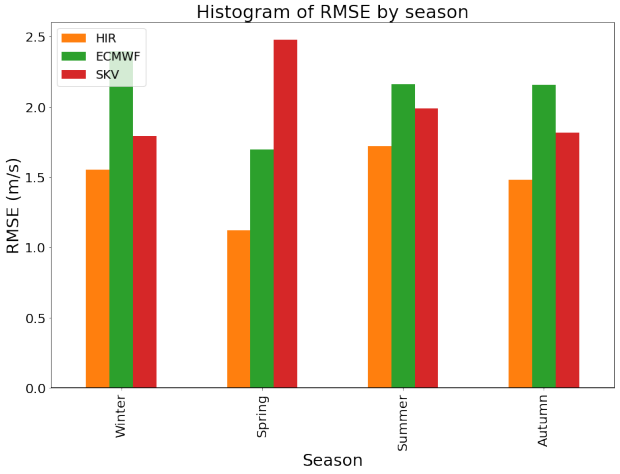
Figure 26: *The bias of the SKV of the maximum wind gust vs. observed 10-minute wind speed (lead time of 3 hours).*

3.3.3 Dependencies

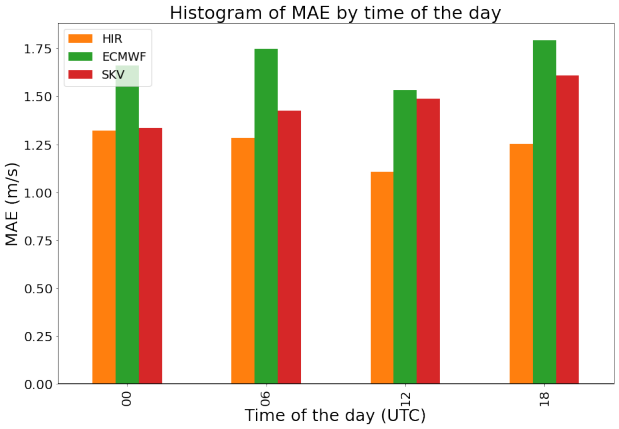
Dataset without HARMONIE



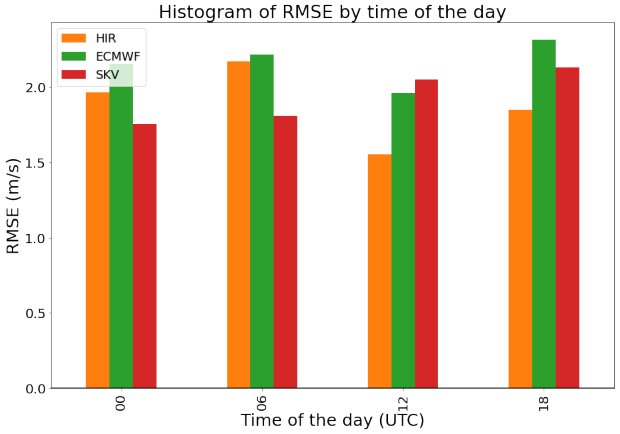
(a) MAE with season (all lead times)



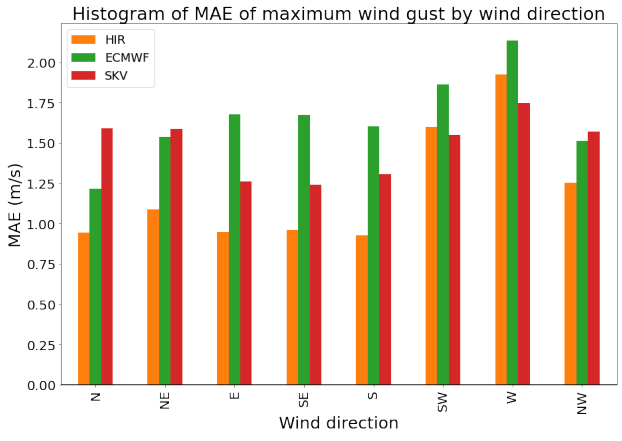
(b) RMSE with season (all lead times)



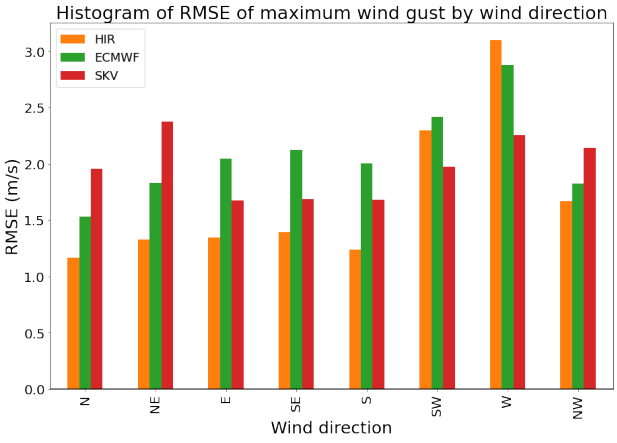
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



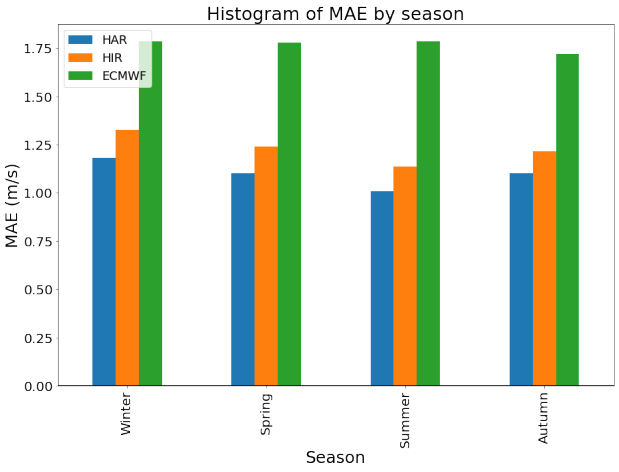
(e) MAE with wind direction (all lead times)



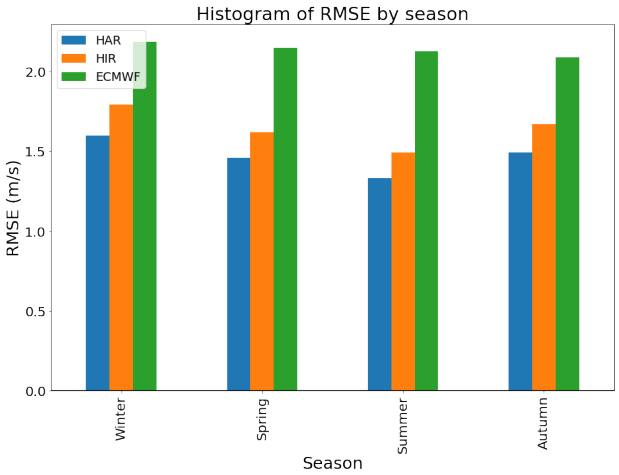
(f) RMSE with wind direction (all lead times)

Figure 27: *The dependency of the MAE and RMSE of the forecasted wind gusts on certain factors is shown. Those results are for the dataset without HARMONIE, so for the period 2018-01-01 up to and including 2021-05-03.*

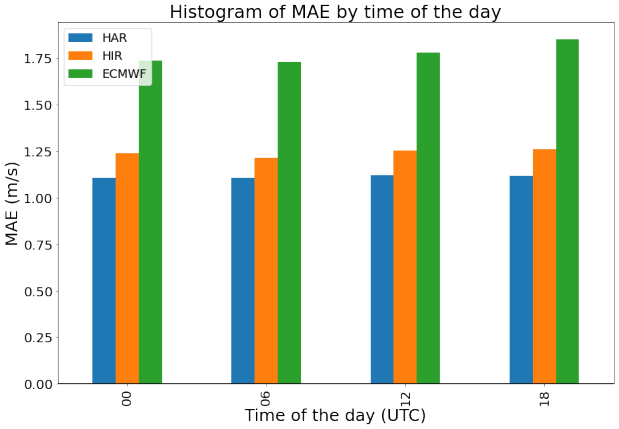
Dataset without SKV



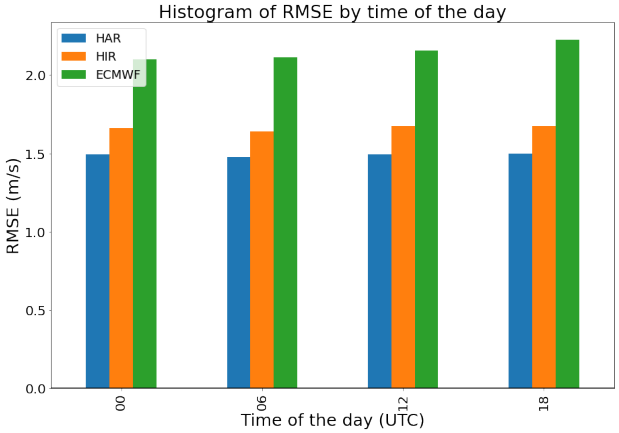
(a) MAE with season (all lead times)



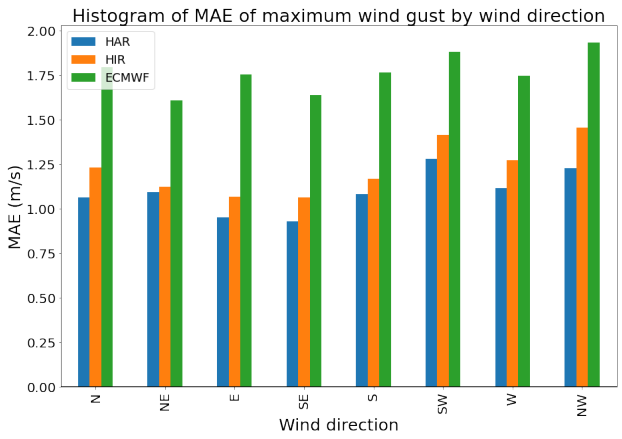
(b) RMSE with season (all lead times)



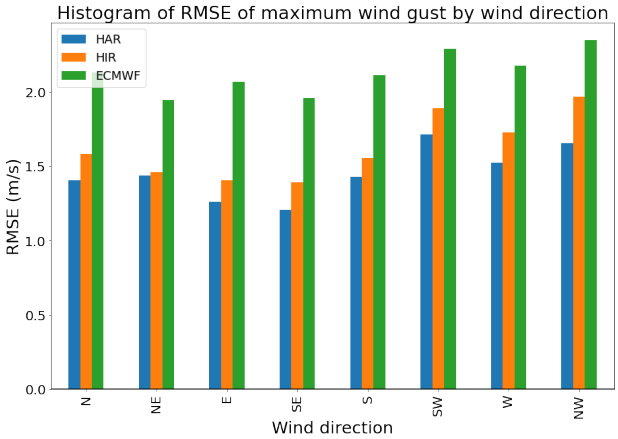
(c) MAE with time of the day (all lead times)



(d) RMSE with time of the day (all lead times)



(e) MAE with wind direction (all lead times)



(f) RMSE with wind direction (all lead times)

Figure 28: *The dependency of the MAE and RMSE of the forecasted wind gusts on certain factors is shown. Those results are for the dataset without SKV, so for the period 2020-11-04 up to and including 2022-03-31.*

The plots on the previous pages show again the dependency of two verification metrics (MAE and RMSE) on the season, time of the day and the wind direction. As was also found in the validation of the average wind speed, not a clear pattern can be found for all the three dependencies analysed. The thing that attracts attention is that ECMWF dominantly has the largest MAE/RMSE. For the dependency of the two verification metrics on the wind direction, a slight pattern seems to be present. In general, a somewhat larger error for winds from the (south)west seem to be present. When the Netherlands experiences higher wind speeds and even extreme wind gusts, most of the time the wind is originating from the (south)west and sometimes northwest. The relatively larger errors in wind gusts for those wind direction probably indicate underestimations as was often the case as shown by the regression plots in the previous section.

3.3.4 Statistical tests

In this section, again the results of the independent sample t-tests are shown for both datasets of the wind gusts of the TAFG/SKV and observational data. Tested is whether the model differs from observations in both directions (two-sided). The same significance level (p=0.05) is used and again the data used contains the forecasting data with a lead time of 3 hours.

Table 4: *Table showing the p-value of the independent sample t-tests*

P-value	HARMONIE	HIRLAM	ECMWF	SKV
Dataset without HARMONIE	X	0.0469	$3.22x10^{-55}$	$3.83x10^{-16}$
Dataset without SKV	0.112	0.00981	$1.84x10^{-3}$	X

The fact that the TAFG of ECMWF performed worst on forecasting wind gusts is clearly represented by the results of the statistical tests too. The lowest P-values, and thus the strongest discrepancy between observations and model data, belong to ECMWF. However, also the SKV has a significant disagreement compared to observations. The same holds for HIRLAM, albeit less strong. As was also discussed using the analyses in the previous sections, HARMONIE performs best on forecasting the wind gusts for Schiphol. The results of the statistical tests support this with clearly no significant difference between HARMONIE and observational data.

4 | Conclusions and recommendations

In the sections below, the conclusions of the results of the validation will be summarised for each validated variable.

4.1 Wind direction

A mean absolute error of more than 15° was found considering the wind direction, which can make the difference for an air traffic management team to switch to another runway or not. However, when filtering out the lower wind speeds (which are not important for the air traffic management team), the MAE decreased to values around 10° to 12°. Between the different datasets however, hardly any differences were present. Both the MAE and RMSE hardly showed any differences between the different the TAFGs and SKV. The MBE however showed clear differences where ECMWF showed the smallest error, whereas the SKV showed the largest.When filtering out the smaller wind speeds, which is more relevant for the air traffic management team, it is found that HIRLAM and HARMONIE have the lowest MBE (around 0°), whereas the SKV clearly shows consistently a positive MBE and ECMWF a (smaller) negative MBE. The order found was supported by the statistical tests, where HIRLAM appeared to perform slightly better than HARMONIE, but significantly better than ECMWF and the SKV. Also the time series (moving averages) in the Appendix confirmed this with smaller differences based on the MAE, but larger differences when analysing the MBE with the largest error in general for the SKV. Furthermore, it was found that the MBE was dominantly positive for the SKV meaning that a stronger veered wind direction was forecasted than was observed. Interesting was the finding that the lowest errors, in terms of MAE and RMSE, were found during autumn and winter. The average lower wind speeds during spring and summer probably cause the more variable wind direction and therefore a larger chance of errors to occur. However, when filtering out the lower average wind speeds, this effect was still visible. Considering the dependency on time of the day, it was found that the error was lower during early morning and noon. This effect disappeared when filtering out the lower observed average wind speed.

4.2 Average wind speed

The mean absolute error for the average wind speed showed very small differences between the models and the SKV, with a slightly preference for ECMWF and HARMONIE as most accurate, whereas the SKV and HIRLAM seemed to perform least accurate. The MBE supports the highest error for the SKV, whereas HIRLAM and HARMONIE show the smallest error. The RMSE and accuracy correspond more with was found based on the MAE. The combination of those metrics leads to the conclusion that the SKV performs least accurate, followed by HIRLAM and then ECMWF/HARMONIE. The time series (moving averages) in the Appendix confirm the, in general, larger error (MBE) of the SKV whereas ECMWF and HARMONIE perform in general more accurate. The least accurate performance of the SKV was supported quantitatively by the results of the statistical tests. The regression analyses showed interestingly a slight underestimation for observed lower wind speeds, whereas the higher wind speeds are in general consistently underestimated, being consistent with slopes smaller than 1. This is partly being corrected nicely already

by the meteorologists in the SKV, since the slope of the SKV is closer to 1 than for the TAFGs on which the SKV is based. Whereas clearer dependencies for the wind direction on valid time of the day, season and the wind direction itself were found, this was not the case for the dependency of the error of the average wind speeds on those factors.

4.3 Wind gusts

The validation of the wind gusts led to a clear conclusion that ECMWF performs by far least accurate. The SKV is found to perform consistently better than ECMWF, but not as accurate as HIRLAM and HARMONIE, which have the highest accuracy for all lead times. Both the MAE, RMSE and MBE support this conclusion. Again, all the MBE-values are positive, meaning that in general an overestimation of the wind gusts is present compared to the wind gusts observed. The conclusions mentioned above are supported by the regression analyses, showing by far the least accurate performance of ECMWF. The SKV performs better, but HIRLAM and HARMONIE perform most accurate. The time series (moving averages) support the strong differences in the magnitude of the error between the different datasets where ECMWF performs by far least accurate, whereas HIRLAM and HARMONIE perform best. The same pattern in the regression analyses is found as was the case with the validation of the average wind speed: an overestimation of the lower wind gusts and a underestimation of the higher wind gusts, quantified by the slopes smaller than 1. The MBE shows in general an overestimation, because the lower wind speeds occur more often than the more extreme ones. The findings mentioned above are clearly supported by the statistical tests showing a relatively strong significant difference of the ECMWF forecasts compared to observational data. Also the SKV shows a significant difference compared to observations. HIRLAM, but especially HARMONIE perform much more accurate. The same as for the validation of the average wind speeds, no clear dependencies on valid season, time of the day and wind direction were recognised.

4.4 Recommendations

Considering the wind direction, the main recommendations would be to base the forecast more on the TAFG of HARMONIE (or HIRLAM), since the verification metrics and statistical tests showed the most accurate performance for this specific model. Also, the overestimation in terms of wind direction is least strong here.

For the average wind speed, the main recommendation is to base the forecast more on the output of ECMWF and HARMONIE. Because of the overestimation of the lower wind speeds and underestimation of the higher wind speeds, it is recommended to investigate in correcting more for this effect. This is partly being done already by the meteorologists since the slope of the SKV is closer to 1 than for the TAFGs on which the SKV is partly based.

For the wind gusts, the discrepancies between the forecasting datasets and the SKV were most obvious. HIRLAM and especially HARMONIE appeared to perform by far best on forecasting wind gusts, so the main recommendation would be to base the SKV more on those models. The same pattern was found as for the average wind speed: an overestimation of the lower wind gusts and an underestimation of the

higher wind gusts. The slope of the SKV is again most close to 1 and clearly more close to 1 compared to the other models, meaning that the correction already takes place quite well by the meteorologists.

Another thing that should be noticed is that the decision of the KNMI to switch as main forecasting model from HIRLAM to HARMONIE seems justified, since HARMONIE shows consistently in general a lower error than HIRLAM does.

All the recommendations and conclusions mentioned in this report are based on this dataset without paying attention to the specific weather conditions themselves in terms of precipitation, incoming radiation and/or (in)stability parameters. Meteorological variables like wind gusts are partly dependent on the atmospheric stability and therefore it would be interesting, if time allows, to find out for example under which stability regimes the bias with respect to observational data is high and which weather models perform better compared to other weather models regarding the TAFG. To quantify the atmospheric stability in the surface layer, for instance the so-called (bulk) Richardson number could be estimated, which indicates the amount of turbulence. Based on this number, several classes could be defined as showed in (Schnelle Jr., 2003). If the data could be obtained on this, or data related to it so that the parameter can be estimated, it would be interesting to find out to which extent the biases of the TAFGs (and SKV) depend on atmospheric stability. It will ideally lead to more detailed suggestions about which dataset can be used best during which meteorological situations (frontal passage, development of thunderstorms, sea breeze circulation, etc.), which could optimise the air traffic management at civil airports further.

Another interesting aspect is the fact that in the SKV the wind-related variables, which are validated in this report, are given as deterministic values and not probabilistic. This made the comparison a bit more difficult. A deterministic forecast is often being drawn from the forecast probability distribution. Within this step, information about the uncertainty in the forecast is being lost which makes it a kind of 'user-dependent' step, since the assessment of the uncertainty in the forecast may be different for everyone. Many extreme weather events are characterised by a low probability of occurrence, but with a relatively high risk. Deterministic forecasts often conceal this information, so therefore these low probability/high risk scenarios should be revealed to the user for an improved decision (Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2007).

4.5 General discussion

The TAF Guidance data used for the analyses in this research is, together with the observational data, merged into a dataset in order to maintain the same meteorological situation which is validated increasing the reliability of the results. As was described in detail in Section 2.2.1, two datasets were used to base the conclusions on a considerable and equal amount of data. The second thing is that because the amount of data decreased with lead time, certain 'bins' were used in order to maintain the homogeneity of the dataset as good as possible. Of course those bins cause uncertainty, since for a lead time of for example 8 hours the lead times of 7 hours up to and including 9 hours were used. By applying the methods as described in this report previously, the inhomogeneity is reduced as much as possible. However, especially when analysing the dependency of the biases of certain variables on the wind direction, this inhomogeneity is still significant since each wind direction occurs not equally often. Besides the fact that this rearranging of forecasting data of the models contains a certain amount of uncertainty, observational data of course

always contains measurement errors and thus uncertainty. However, this uncertainty range is not assumed to influence the conclusions in this report.

Interesting is to compare the results with the results found by the previous intern (Edwin Kok) which were summarised in Section 2.2.3. Edwin performed the validation of only the average wind speed and found that the ECMWF model performed in general worst compared to observational data when judging the bias itself. The bias of the HARMONIE TAFG was lowest (5-6 times lower than the bias of HIRLAM/ECMWF). However, for lead times of 6 hours and more, the HIRLAM TAFG showed the highest accuracy with a bias which is roughly 2 times lower than the bias of HARMONIE and roughly 4 times lower than the bias of ECMWF. This is not perfectly consistent with the results of this research. For the average wind speed, the pattern was quite chaotic with ECMWF performing worst in the dataset without SKV (by judging purely on the MBE). Interesting is the fact that there was in general an overestimation of the average wind speed (when wind speeds are relatively low) was present in Edwin's results too. Apparently this is something that the models still have difficulties with and, as recommended previously, is really something to focus on.

Summarised, it would be a logical next step to continue the reasoning from this report and extend it to research on which forecasting dataset can be used best during which meteorological situations (frontal passage, development of thunderstorms, sea breeze circulation, etc.), which could optimise the air traffic management at civil airports further. It could also be that a certain phenomenon's magnitude is forecasted correctly, but that the timing is off. In such a case, closely monitoring of the actual situation and adjusting the forecast in response might drastically improve the quality. For the validation of the wind gusts, I recommend to split the validation based on the different weather types leading to wind gusts. The process behind creating (convective) wind gusts in the winter (mechanical) or during severe thunderstorms by convection in the atmospheric boundary layer is very different. Therefore, it would be possible that the skill of the different weather models also differs for each weather type leading to wind gusts. Expected is that the more local, small scale short-range weather models (HARMONIE, a non-hydrostatic convection-permitting model, for instance) are more accurate when it comes to resolving local convection during summer leading to severe thunderstorms. Therefore, expected is that this model performs better in forecasting severe wind gusts during severe thunderstorms on warm summer days with an unstable atmospheric boundary layer than for instance a model with a larger resolution like ECMWF, since convection during those days is a very local process. Mechanical wind gusts, purely because of strong pressure differences on a short distance (winter storms), also weather models like ECMWF are probably interesting to look at when it comes to forecasting wind gusts. It depends therefore also strongly on the stability of the atmospheric boundary layer, which is recommended to focus on next time in more detail (Kalverla et al., 2018).

For the validation in general, I would also recommend next time (or for the next intern who is going to work on this interesting topic), to use more specific and complex verification metrics, for instance the Critical Success Index (CSI) or the Hanssen and Kuipers discriminant (HKS) (World Climate Research Programma (WCRP), 2017). I mention this point specifically also because the metrics used, like the MBE, RMSE, are rather sensitive to points which are located on the edges of the distribution (outliers) and give quickly a large deviation to those scores. Next time, for instance the Ranked Probability Score (RPS) could be used and even more variables which take the distribution of the values of the variable to be validated into account. This score is also used specifically by ECMWF (Australian Government - Bureau of Meteorology). Another option to use more complex and valid verification metrics is for instance to scale the

mean absolute error (MAE) with for instance the standard deviation of the complete ensemble to process for example seasonal variability in it.

The validations executed for this specific report are for Schiphol specifically. However, it would be very interesting to find out whether the same results were being found when exactly the same validation was executed for a different location, Cabauw for instance. If this is not the case, then probably some local factors like surface roughness, location of buildings, etc. are causing a more 'local wind climate'. This is something that needs to be taken into account and give the results more background and perspective. A last point I would like to make is about the validity of the independent sample t-tests used in this specific research project. Due to time limitations, I did not execute other kind of tests. However, the variables were not distributed normally and therefore one can argue whether a t-test like this is most relevant. Therefore, for the next intern who will work on this topic, I would recommend to find out which statistical tests could be used more and are most relevant for this specific validation.

Besides the recommendations and content-related discussion points mentioned above, I would also recommend for LVNL and KNMI to start discussing specific goals in more detail. It is my understanding that also the KNMI is partly working on validating and improving the SKV, but that not everything is communicated clearly to LVNL (and other organisations). When discussing the project assignment with my supervisor from LVNL, Ferdinand, it became clear for me that they did not know anything about the projects KNMI was working on. When I had a meeting with Hans and Nico, it became clear for me that the KNMI has worked, and still does work, on the validation of the SKV. This is a very good process and I think the transfer of knowledge can be improved by stronger communication and more contact of the two organisations.

Lastly, I would like to thank Ferdinand Dijkstra, Nico Maat, Hans van Bruggen and Gert-Jan Steeneveld for my interesting internship, their feedback and help during my research project!

5 | References

Australian Government - Bureau of Meteorology, . Ensemble verification metrics. https://www.ecmwf.int/sites/default/files/elibrary/2017/17626-ensemble-verification-metrics.pdf. [Online; accessed 13-May-2022].

Deutscher Wetterdienst (DWD), 2019. Model Output Statistics-MIX (MOSMIX. https://www.dwd.de/EN/ourservices/met_application_mosmix/met_application_mosmix.html. [Online; accessed 23-March-2022].

European Centre for Medium-Range Weather Forecasts (ECMWF), . Forecast charts and data. https://www.ecmwf.int/en/forecasts. [Online; accessed 12-May-2022].

Finnish Meteorological Institute (FMI), . Hirlam Weather Model. https://en.ilmatieteenlaitos.fi/hirlam-opendata-on-aws-s3. [Online; accessed 12-May-2022].

Holtslag, A., Svensoon, G., Baas, P., Basu, S., Beare, B., Beljaars, A., Bosveld, F., Cuxart, J., Lindvall, J., Steeneveld, G., M., T., Van De Wiel, B., 2013. Stable Atmospheric Boundary Layers and Diurnal Cycles: Challenges for Weather and Climate Models. American Meteorological Society , 1691–1706doi:10.1175/BAMS-D-11-00187.1.

Jacobs, A., Maat, N., 2005. Numerical Guidance Methods for Decision Support in Aviation Meteorological Forecasting. Weather and Forecasting - American Meteorological Society 20, 82–100. doi:10.1175/WAF-827.1.

Kalverla, P., Steeneveld, G.J., Ronda, R., Holtslag, A.A., 2018. Evaluation of three mainstream numerical weather prediction models with observations from meteorological mast ijmuiden at the north sea. Wind Energy 22, 34–48. doi:10.1002/we.2267.

Knowledge & Development Centre (KDC) Mainport Schiphol (KNMI), 2015. Capacity and runway predictions. https://kdc-mainport.nl/wp-content/uploads/2015/10/150601-To70-Rapport-KDC-Capacity-and-Runway-Predictions.pdf. [Online; accessed 21-March-2022].

Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2007. Low visibility and celing forecasts at schiphol. https://cdn.knmi.nl/system/data_center_publications/files/000/068/115/original/schipholzicht.pdf?1495621044. [Online; accessed 14-March-2022].

Koninklijk Nederlands Meteorologisch Instituut (KNMI), 2020. KNMI stopt met weermodel HIRLAM. https://www.knmi.nl/over-het-knmi/nieuws/knmi-stopt-met-weermodel-hirlam. [Online; accessed 21-March-2022].

de Rover, C., Vogelezang, D., van Bruggen, H., Maat, N., Smit, L., Heijstel, J., Keet, M., Wjingaard, J., ten Hove, R., 2008. Improved Low visibility and Ceiling Forecasts at Schiphol Airport. https://cdn.knmi.nl/system/data_center_publications/files/000/068/181/original/final_report_finalversion_.pdf?1495621069. [Online; accessed 21-March-2022].

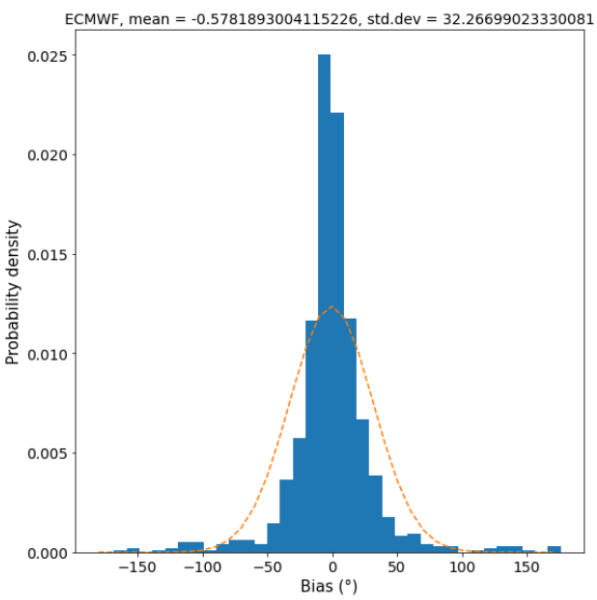
Schnelle Jr., K.B., 2003. Atmospheric Diffusion Modeling. Elsevier - Earth Systems and Environmental Sciences , 679–705doi:10.1016/B0-12-227410-5/00036-3.

World Climate Research Programma (WCRP), 2017. WWRP/WGNE Joint Working Group on Forecast Verification Research. https://www.cawcr.gov.au/projects/verification/. [Online; accessed 23-March-2022].

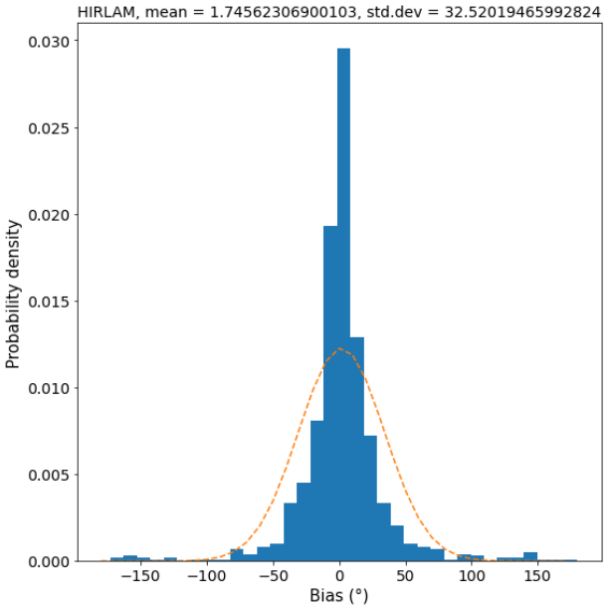
A | Histograms of error distributions

Wind direction

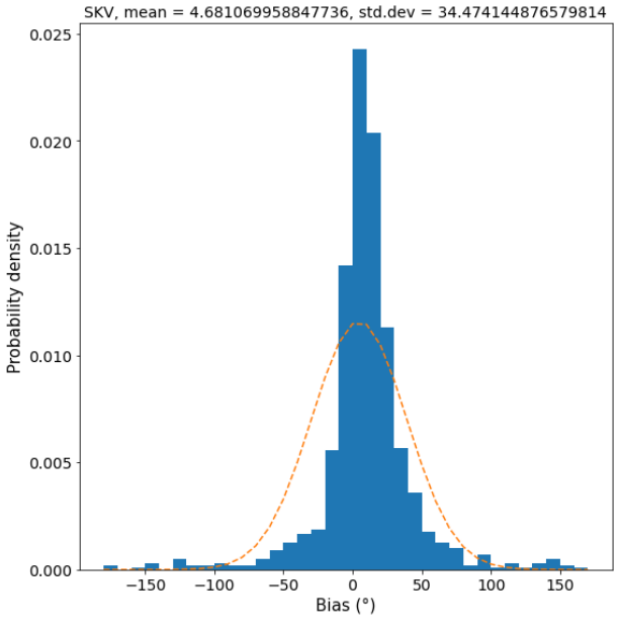
Without HARMONIE



(a) Histogram error distribution ECMWF (all lead times)



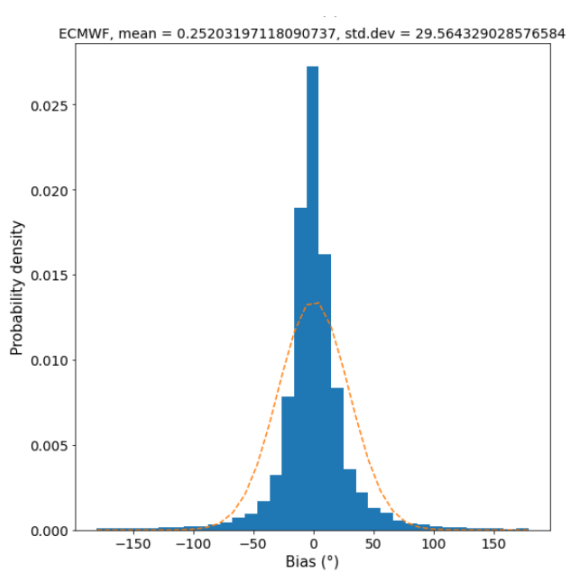
(b) Histogram error distribution HARMONIE (all lead times)



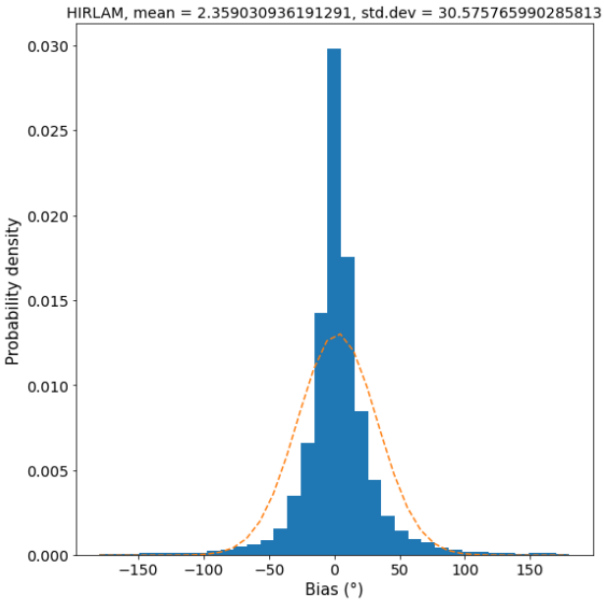
(c) Histogram error distribution SKV (all lead times)

Figure 29: *Histogram of error distributions of the wind direction for the different TAFGs for all lead times for the dataset without HARMONIE.*

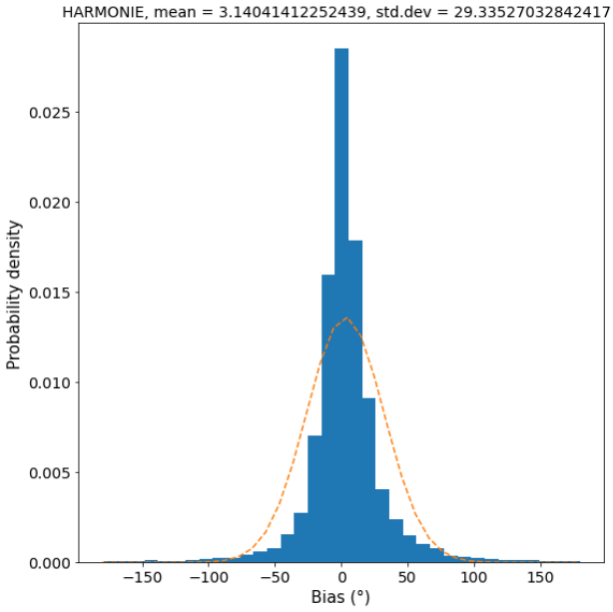
Without SKV



(a) Histogram error distribution ECMWF (all lead times)



(b) Histogram error distribution HIRLAM (all lead times)

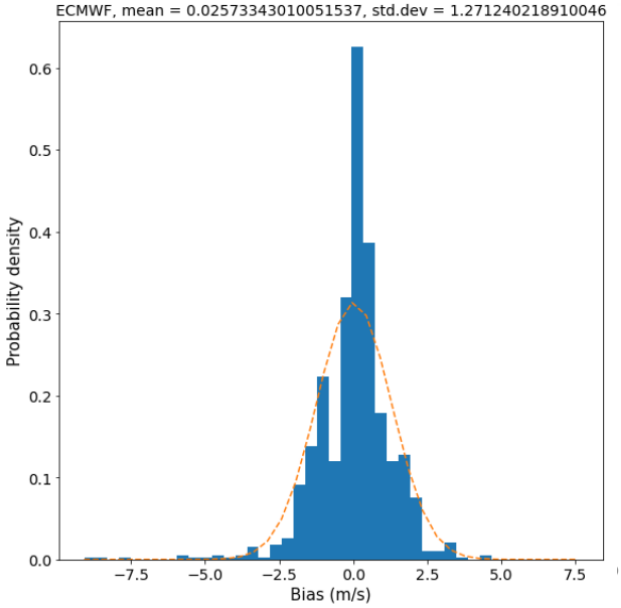


(c) Histogram error distribution HARMONIE (all lead times)

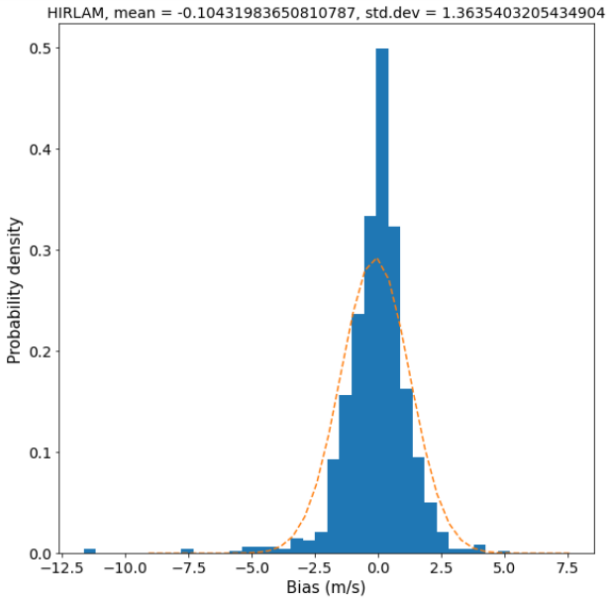
Figure 30: *Histogram of error distributions of the wind direction for the different TAFGs for all lead times for the dataset without SKV.*

Average wind speed

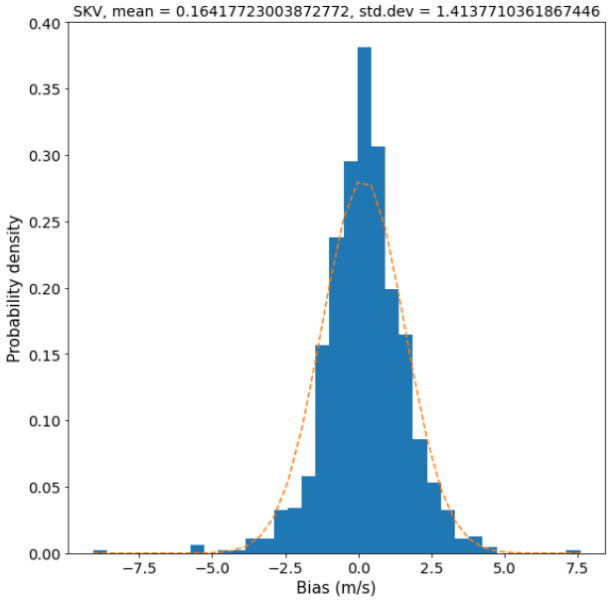
Without HARMONIE



(a) *Histogram error distribution ECMWF (all lead times)*



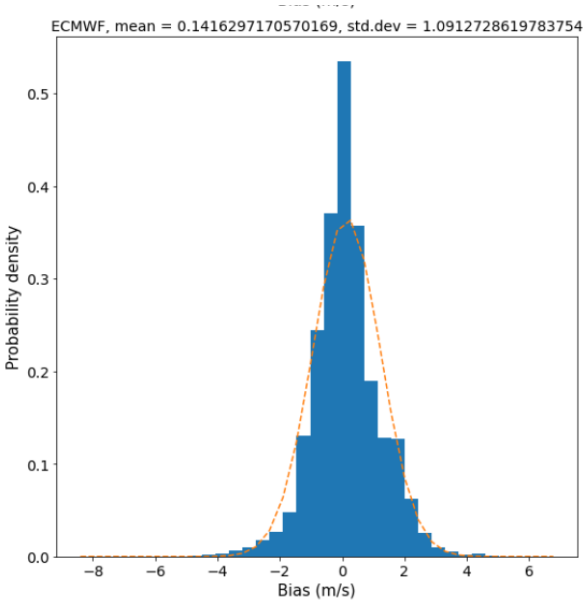
(b) *Histogram error distribution HIRLAM (all lead times)*



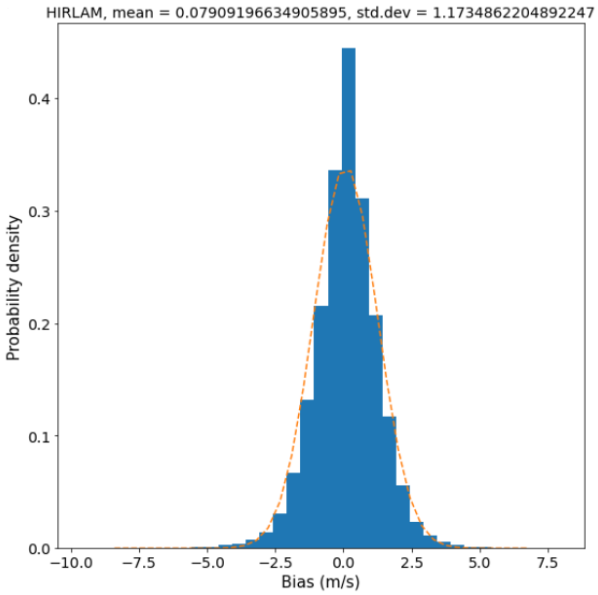
(c) *Histogram error distribution SKV (all lead times)*

Figure 31: *Histogram of error distributions of the average wind speed for the different TAFGs for all lead times for the dataset without HARMONIE.*

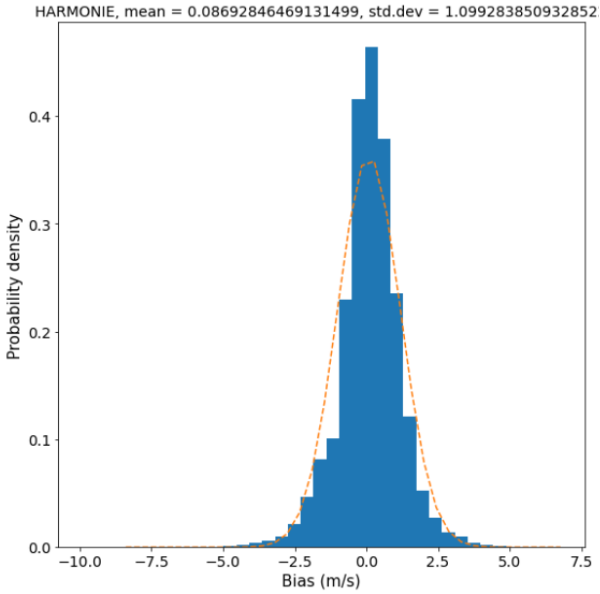
Without SKV



(a) Histogram error distribution ECMWF (all lead times)



(b) Histogram error distribution HIRLAM (all lead times)

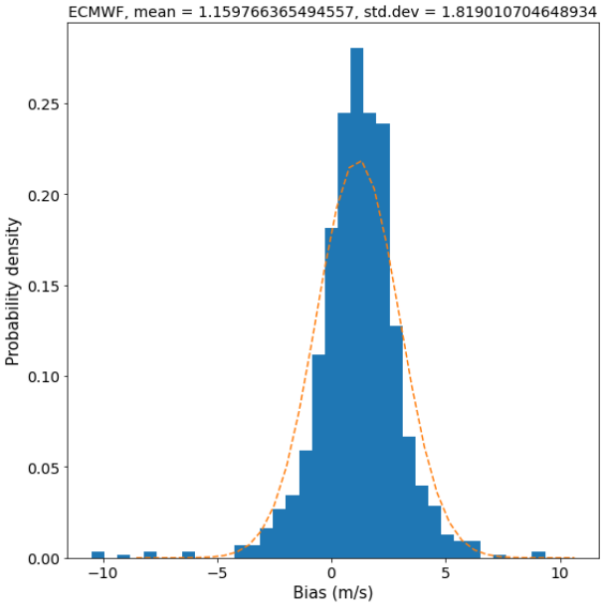


(c) Histogram error distribution HARMONIE (all lead times)

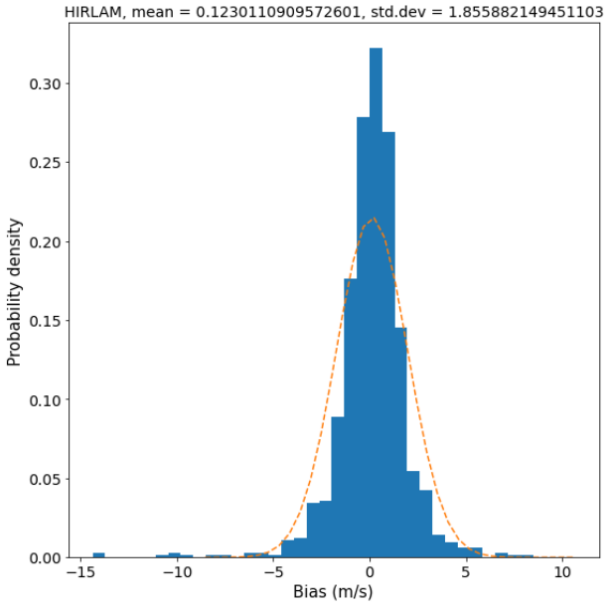
Figure 32: *Histogram of error distributions of the average wind speed for the different TAFGs for all lead times for the dataset without SKV.*

Wind gusts

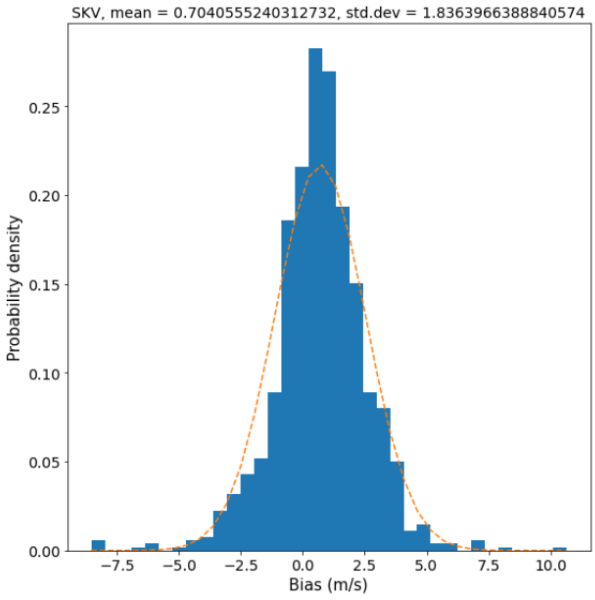
Without HARMONIE



(a) Histogram error distribution ECMWF (all lead times)



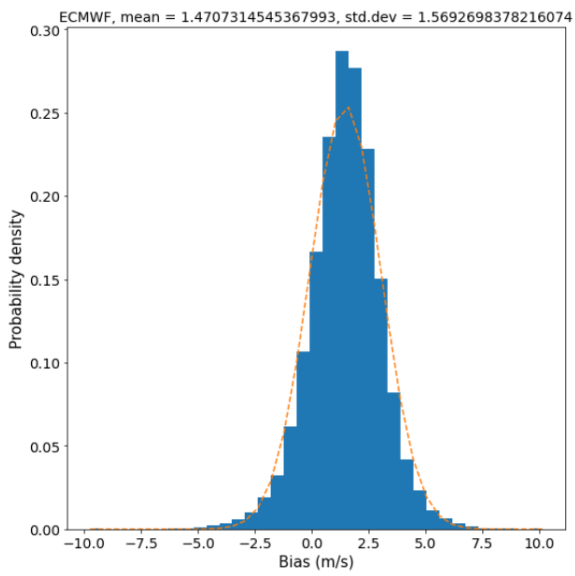
(b) Histogram error distribution HIRLAM (all lead times)



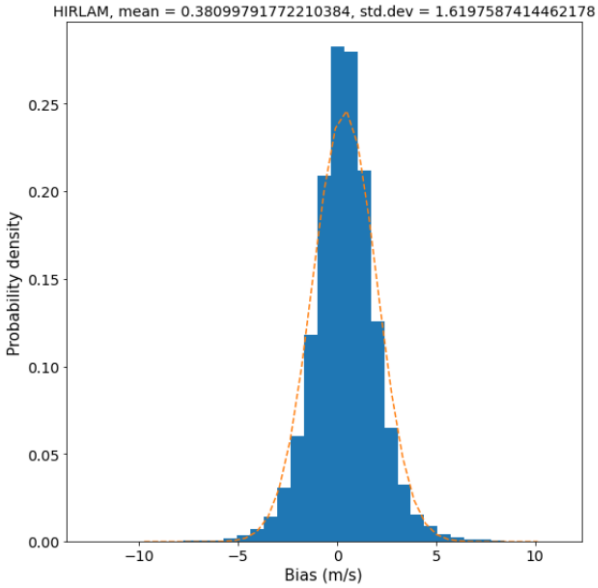
(c) Histogram error distribution SKV (all lead times)

Figure 33: *Histogram of error distributions of the maximum wind gust for the different TAFGs for all lead times for the dataset without HARMONIE.*

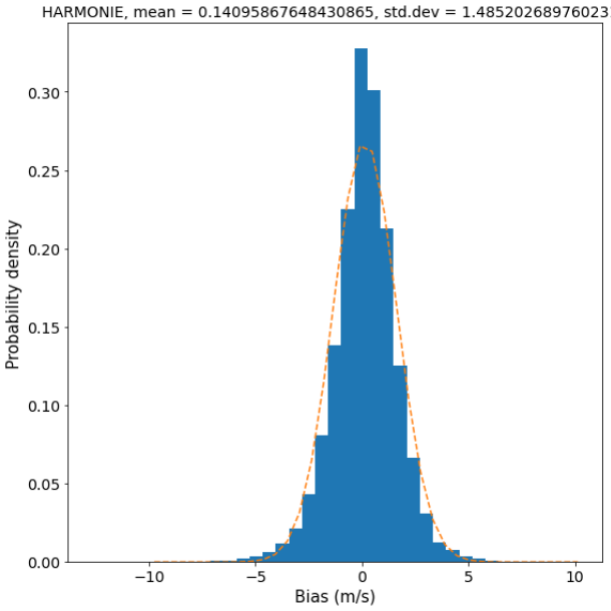
Without SKV



(a) Histogram error distribution ECMWF (all lead times)



(b) Histogram error distribution HIRLAM (all lead times)



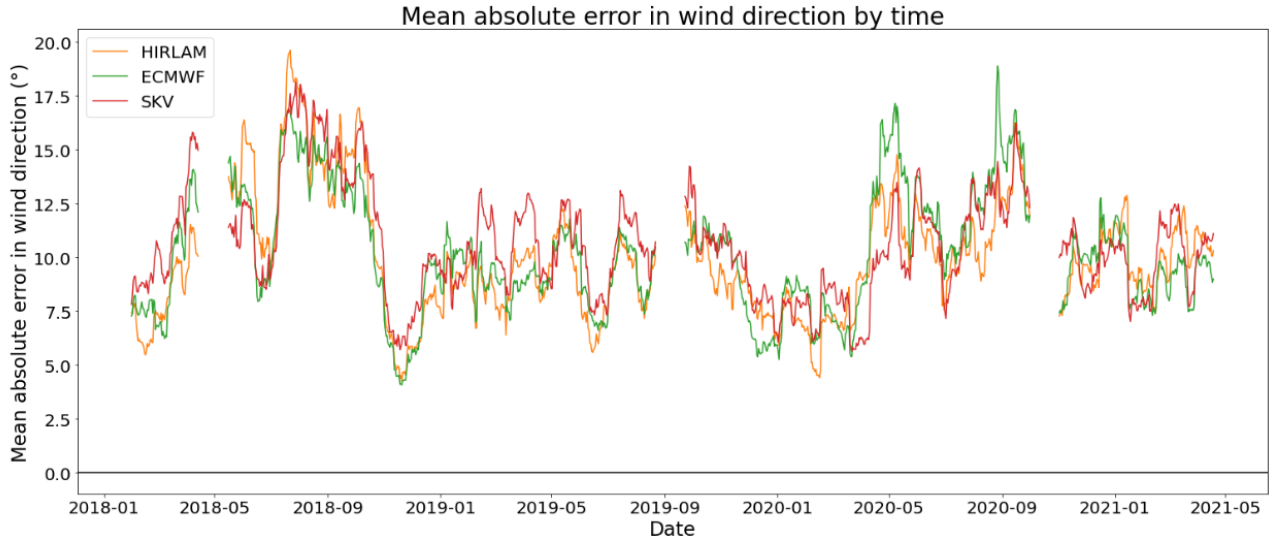
(c) Histogram error distribution HARMONIE (all lead times)

Figure 34: *Histogram of error distributions of the maximum wind gust for the different TAFGs for all lead times for the dataset without SKV.*

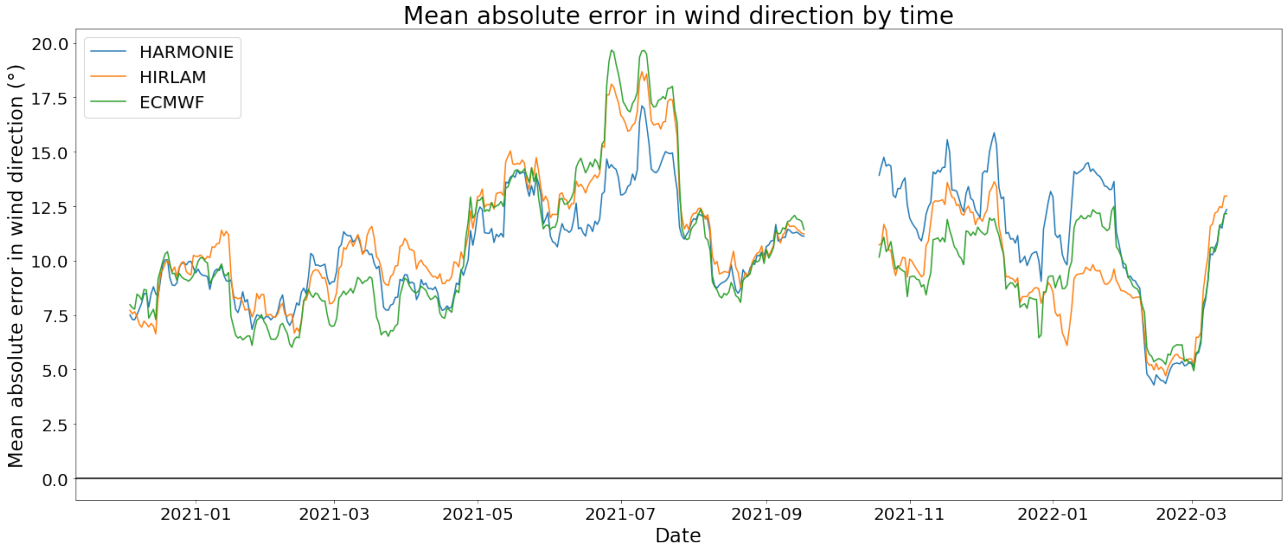
B | Moving averages of MAE and MBE

Wind direction

MAE

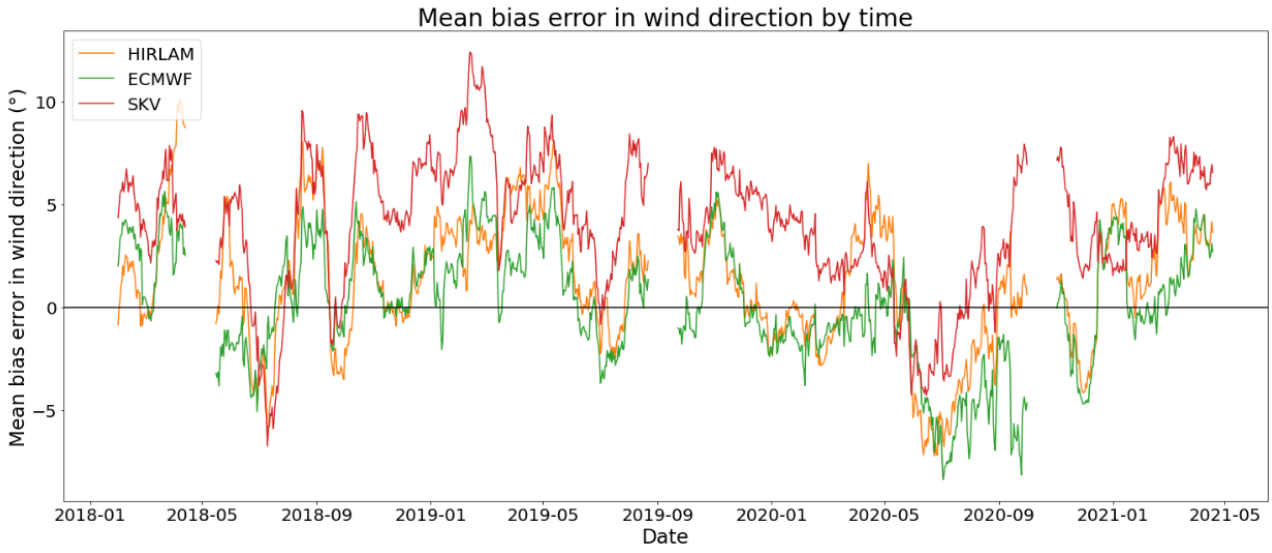


(a) *Moving average (1 month) of MAE of dataset without HARMONIE (lead time of 3 hours)*

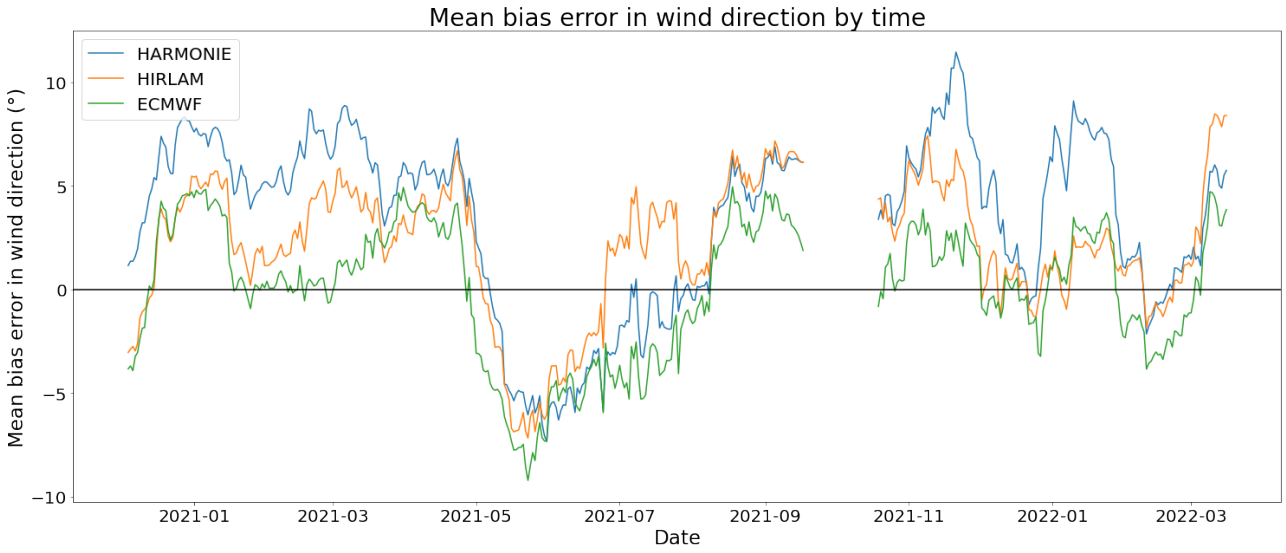


(b) *Moving average (1 month) of MAE of dataset without SKV (lead time of 3 hours)*

MBE



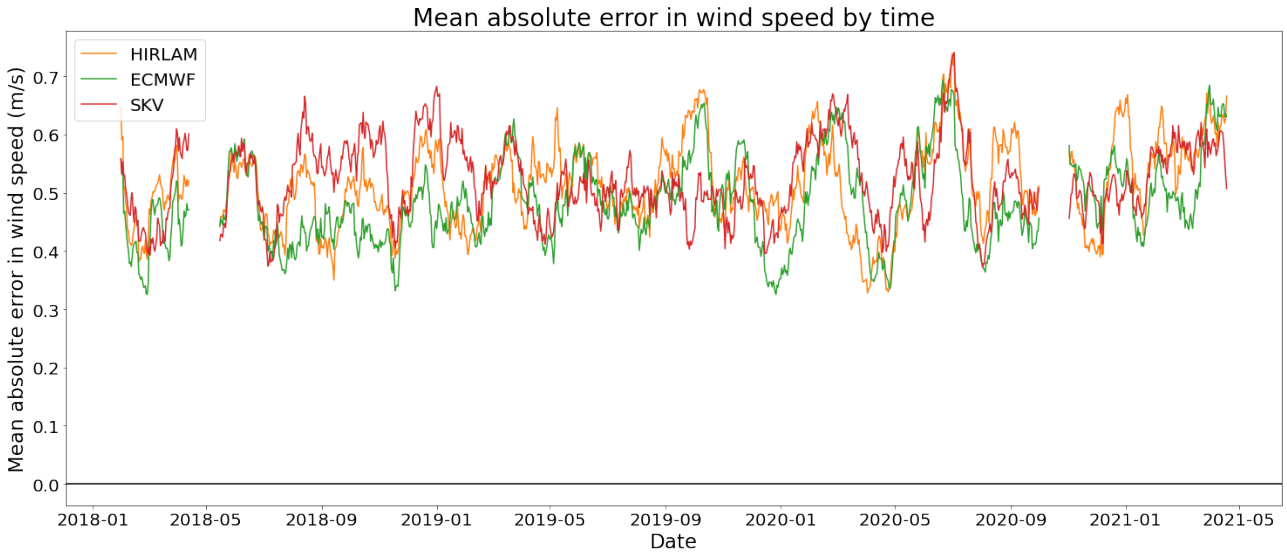
(a) *Moving average (1 month) of MBE of dataset without HARMONIE (lead time of 3 hours)*



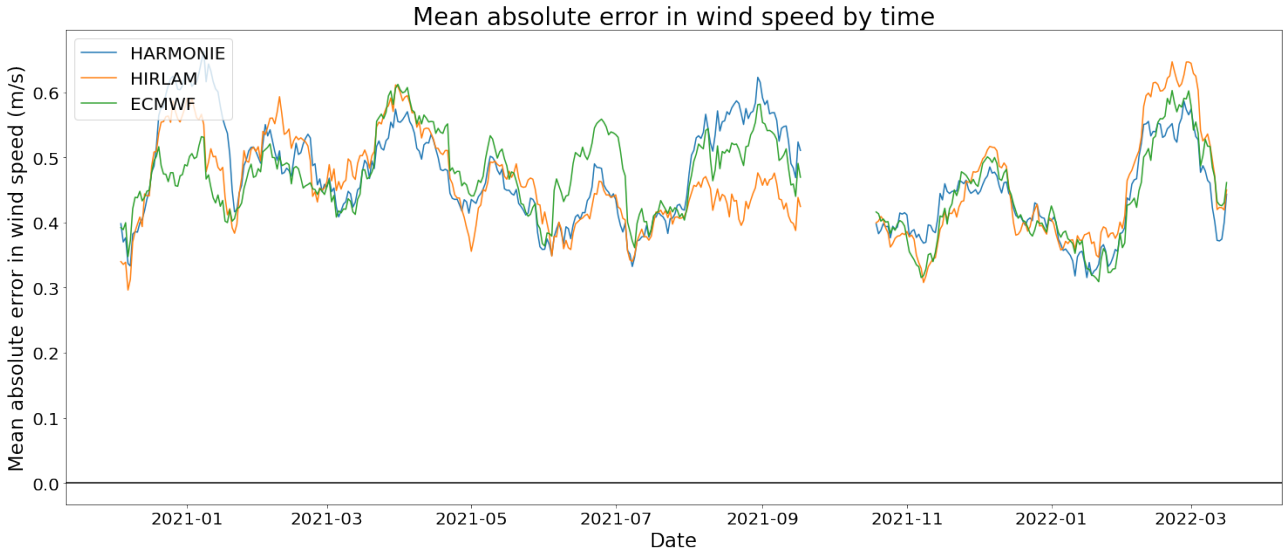
(b) *Moving average (1 month) of MBE of dataset without SKV (lead time of 3 hours)*

Average wind speed

MAE

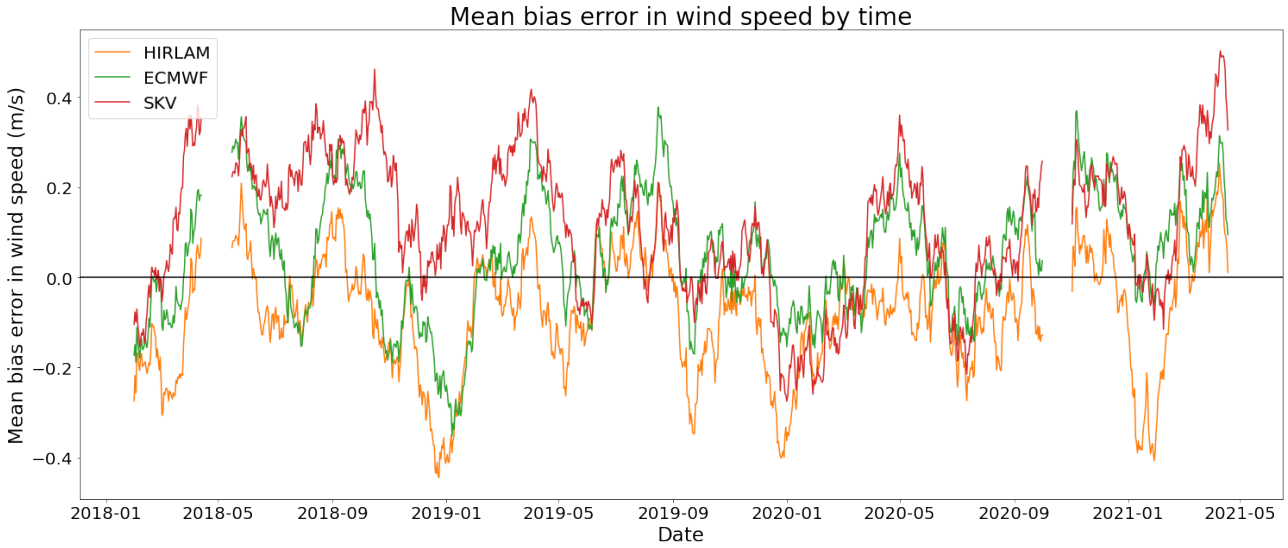


(a) *Moving average (1 month) of MAE of dataset without HARMONIE (lead time of 3 hours)*

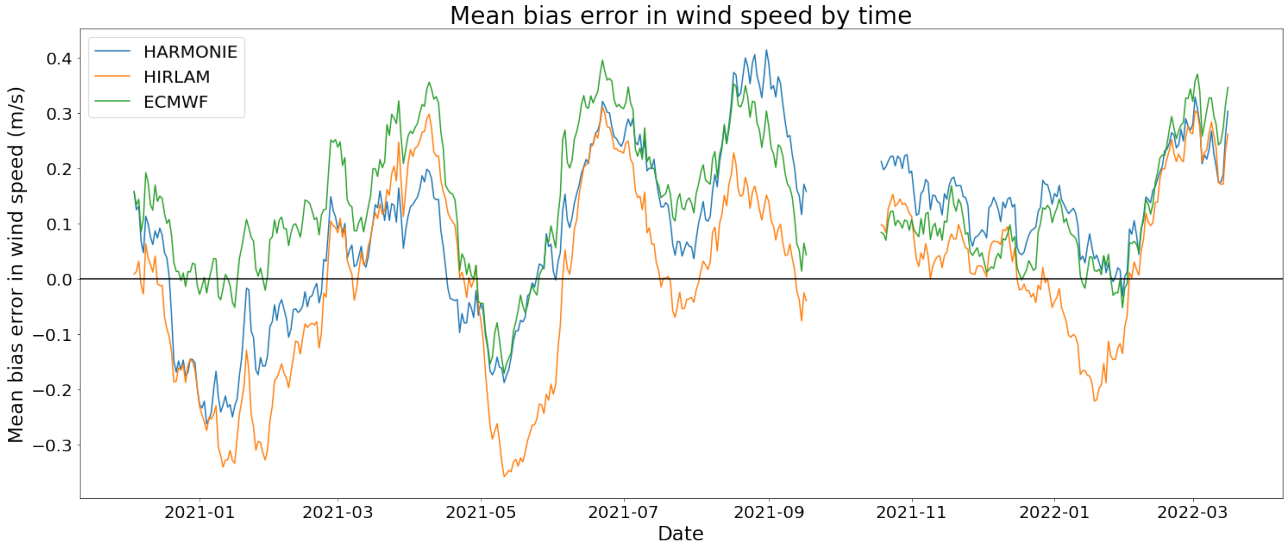


(b) *Moving average (1 month) of MAE of dataset without SKV (lead time of 3 hours)*

MBE



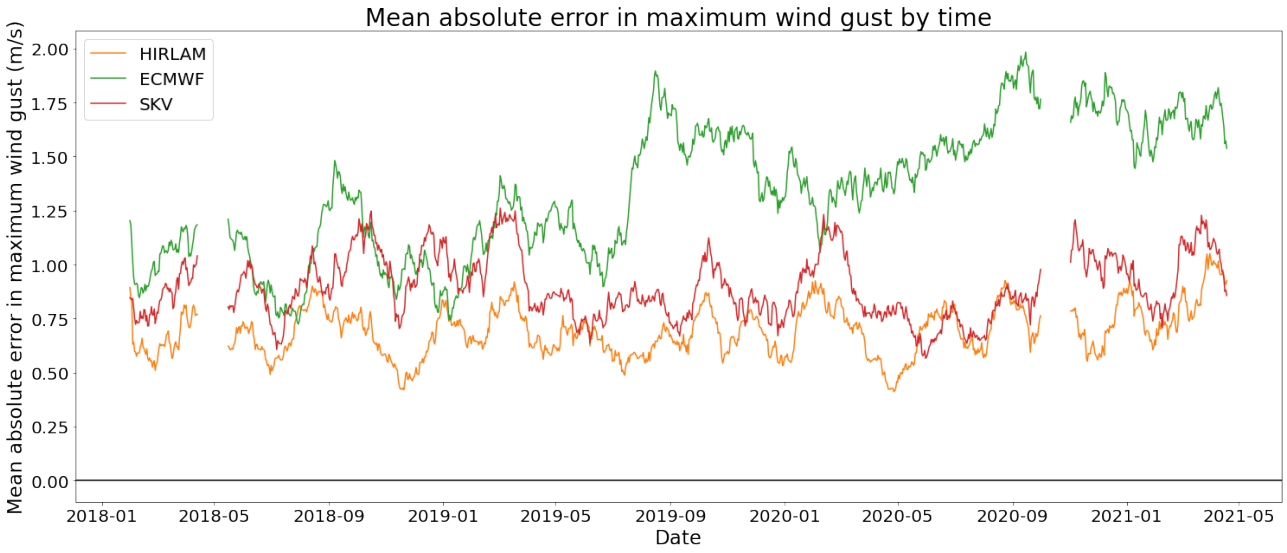
(a) *Moving average (1 month) of MBE of dataset without HARMONIE (lead time of 3 hours)*



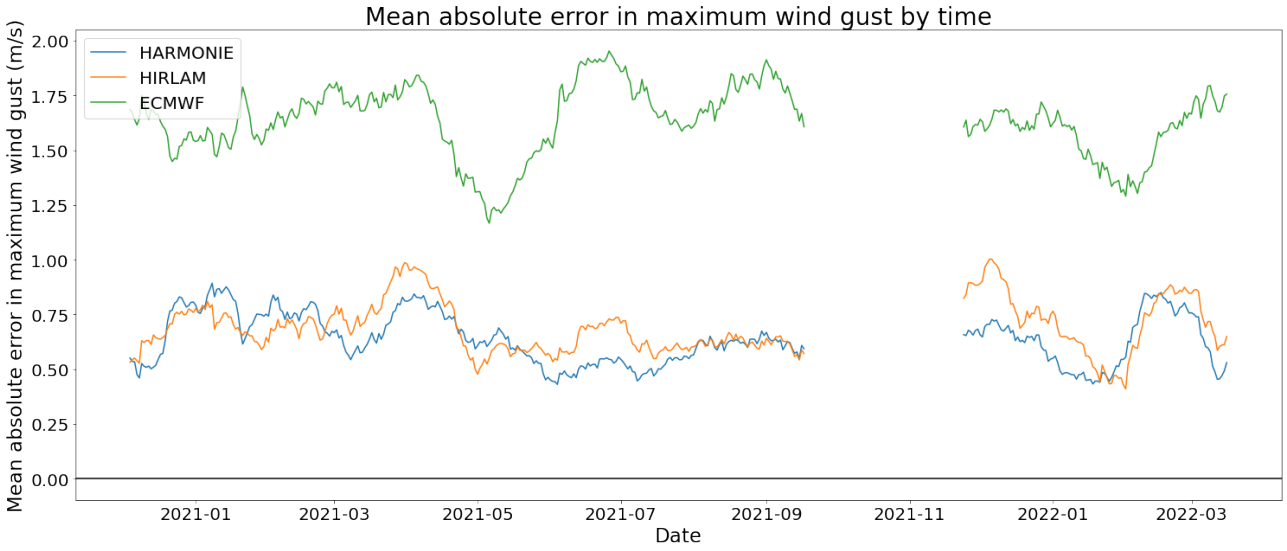
(b) *Moving average (1 month) of MBE of dataset without SKV (lead time of 3 hours)*

Wind gusts

MAE

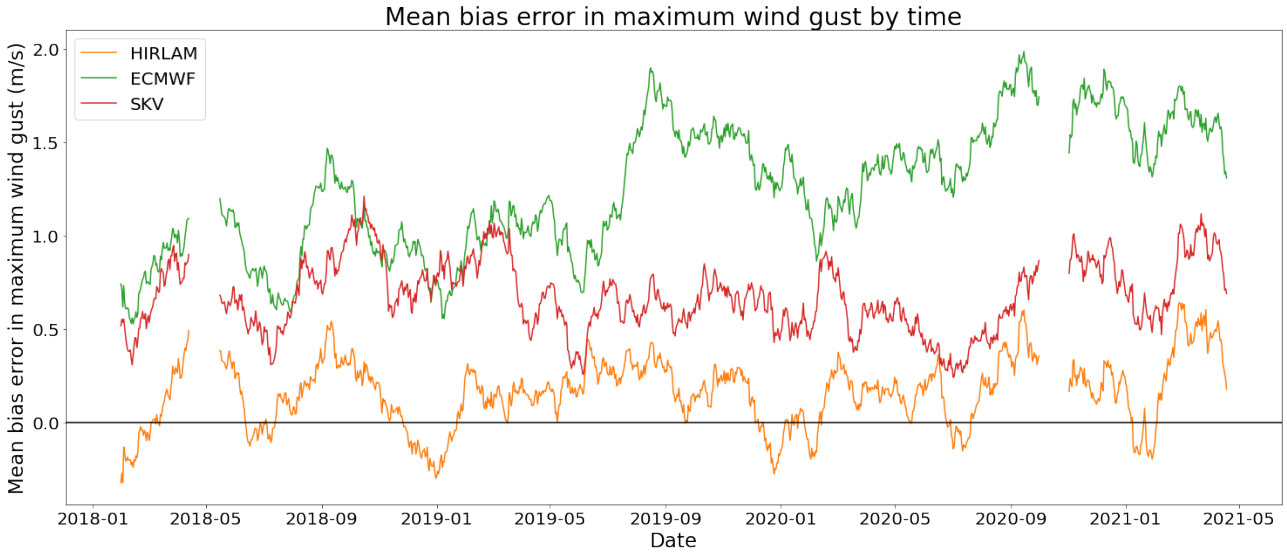


(a) *Moving average (1 month) of MAE of dataset without HARMONIE (lead time of 3 hours)*

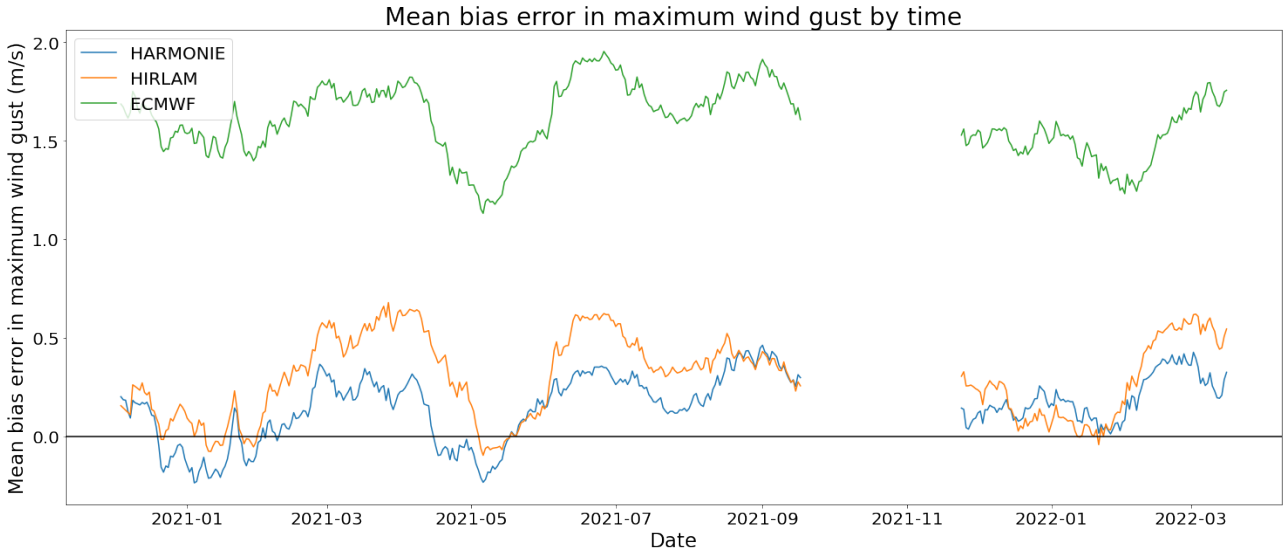


(b) *Moving average (1 month) of MAE of dataset without SKV (lead time of 3 hours)*

MBE



(a) *Moving average (1 month) of MBE of dataset without HARMONIE (lead time of 3 hours)*



(b) *Moving average (1 month) of MBE of dataset without SKV (lead time of 3 hours)*