Stay Report

To: Zentralanstalt für Meteorologie und Geodynamik — ZAMG, Vienna, Austria

Period: 4th February – 1st March, 2019

Topic: Work on analog-based post-processing method

Supervisors: Mag. Alexander Kann and Irene Schicker, PhD

Introduction

I stayed at the Zentralanstalt für Meteorologie und Geodynamik (ZAMG) for four weeks during which I was working on the analog-based post-processing method applied to a NWP model output for point forecasts. This is the continuation of previous work carried out during stay 1 (13/11-09/12, 2017) and stay 2(02/02-03/03, 2018), where the basic algorithm in Python was written and the usability of the analogs method investigated for Austria. Thus, the method was already tested using the AROME deterministic model (1/1/2015-31/08/2017) and corresponding observations from 265 TAWES sites (1/1/2015-31/10/2017).

PART 1: Deterministic analog forecast: Shifting time window

During this stay results for modified analog-based post-processing method (**mAN**) are compared against previously developed analog-based post-processing method (**AN**) and against AROME deterministic model forecasts (**AR**).

The analog ensemble method

In the analogs method, the best-matching historical forecasts compared to the current prediction (analogs) corresponding to the identical lead time and (point) location can originate at any past date within a defined training period. The quality of the analog (the ''difference'') is evaluated using a pre-defined metric (more information available at previous stay reports, Delle Monache et.al., 2013 and Odak Plenkovic et.al, 2018). The search is localized using a time window centered at the defined time of a day, in order to use limited number of degrees of freedom (as proposed in Van den Dool, 1989). Thus, using an NWP forecast at time *t* at a specified location *x* the metrics include the differences between a set of variables, the predictors, centered at the lead time (i.e. *t+6*) and including one time step before and after (*t+5* and *t+7*), to account for shifts in the NWP forecast and a trend. This part was implemented during the past two stays at ZAMG.

Extending the search window

The steps for the recent stay at ZAMG included broadening the time window mentioned above. Often, depending on the weather situation, similar analogs can be found more than one step before/after the selected lead time. Therefore, we decided to increases the window of opportunity for searching suitable analogs. Thus, the window was enlarged to include more time steps. In total five time steps are included. In previous example the search centered at lead time *t+6,* in addition to the *t+5* and *t+7* also includes the time steps centered at *t+5 (incl. t+4, t+6) and t+7 (incl. t+6, t+8).* For every selected forecast lead time at the given location, three possible metrics are calculated and then compared to find the best match out of these three. Only the best match is used afterwards, in order to avoid choosing subsequent highly correlated analogs. As members of the new analog ensemble (AnEn) the corresponding observations to the best-matching analogs are selected. Thus, in case of the most similar forecast for the selected forecast lead time +6 hours is the one of the search window *t+5*, the matching observations and, thus, the ensemble member, will be the observation corresponding to *t+5*.

Scripts and optimization – deterministic analog forecast: Shifting time window

Building the database for the training period is computationally the most demanding part of the entire analog-based scheme, so for this part the one already built is used. To speed up the process a little bit, it was copied and the stations with more than 50 % data missing were excluded. The database is now called: `MyData2015_6_reduced1.db`, and the included stations are listed as `included_stations` in `IOP.py` module. This database contains all the 2015-2016 data for the analog training. The database is based on sql, created using the python `sqlite3` module. Included in the database are:

- `statnr` – unique station number
- `idate` – initialization date

- `itime` – initialization time
- `fhour` – forecast hour (lead time)
- `rrr` - precipitation
- `tl` – 2-m temperature
- `ff` – wind speed
- `dd` – wind direction (deg)
- `rf` – relative humidity
- `pred` – red. pressure
- `ff_obs` – wind speed observations

The basic scripts for the analog method application (the forecasting of the mean of the ensemble: AN forecasting) were developed during my previous stay. The main script is the `IOP-analogs.py` script. Besides loading the data from the training sql database and writing forecasts to another database the main parts are in the `anen` module from `IOP.py`. The module seeks for 20 most similar analogs, sorted by difference (similarity). All available predictors are used (`rrr, tl, ff, dd, rf, pred`). Each predictor is normalized using the standard deviation, where wind direction is treated as circular variable. The width of the time window used in the analog search is fixed (+/- 1 lead time step). The difference metric is calculated applying the `mymetrics` function and the `groupby` object (grouped by initialization date – one by one forecast). Previously, this function was called once, calculating the differences by using the time window centered at the same lead time. During the recent stay, the scripts were modified (`IOP-analogs2.py, IOP2.py`). The `mymetrics` function is called three times, where the center of the time window for comparison is shifted to one lead time before/after. The data used as an input to `mymetrics` is shifted by the new function `move_time_win`. Outputs of `mymetrics` function are now three arrays converted to the dataframes. The second one (`a_tmp2`) is the ''regular'' time frame, while `a_tmp1` and `a_tmp3` are the shifted ones. These dataframes are joined on indices (`idate`), so numpy array out of all the differences `my_diff` is used to determine minimum values. The algorithm picks the one with the smaller difference out of the three for every fixed historical forecast time. This way it cannot happen that there are more than one subsequential analog chosen within the ensemble. For instance, for the selected current forecast (i.e. 8.7.2017.) at lead time 12:00 it can not happen that two or three analogs date from the same day (i. e. 1.7.2015., centered at 12:00 and centered at 13:00). as the most similar ones out of 17 members. The analog with the best scores is chosen as ensemble member. The reason for this constraint is that they would not be independent members.

By using a moving window the possibility that the most similar situation happened sooner/later in a day is allowed. Thus, the measurement corresponding to the time window center of the most similar historical NWP is chosen as AnEn member, regardless whether the measurement happened exactly at the same time, time step sooner or later. However, this also means that this experiment is not a way to correct time shift error in NWP forecast. The NWP time shift (up to +/- one time step) should have already been corrected by the process of choosing a measurement (contrary to choosing historical forecast). The time window is not widening, but moving in time while maintaining the same width. Therefore the NWP time shift could only be corrected up to +/- one time step, same as when using only one static time window. This experiment is not expected to correct NWP time shift better.

A solution had to be found for cases with missing forecast values at either the selected lead time or before/after. However, joining these dataframes with `how='left'`- option (meaning the shape is set on the data with regular time frame window), this problem is solved.

Now all it is left is to sort the metrics for every lead time and choose N smallest values. Data is grouped by forecast hour, having the same dataframe shape (dimensions) as before in order to use previously developed algorithms. The information whether the analogs come from the same time in a day (`fhour`) or not is lost (not needed). For every forecast hour the analog forecast contains 20 analog members. These are the corresponding observations `'ff_obs'` to the smallest differences. The analog members are numbered `ff1` – `ff20` and saved in dataframe, together with AROME (AR) `ff` forecast, just as before.

During the previous stay the time needed to execute the script for one forecast was approximately 14 minutes for all stations (265). When these adjustments were done, there were some memory issues and the script would crash. Therefore, additional optimization was needed.

The `for` loop in the `IOP-analogs2.py` script that looped over stations was switched with groupby object. The “current” station info was included in the apply lambda function as `x.statnr.unique()[0]`. Also, after making this change, errors have occurred while writing

the data in the `sql` database. It seems that for some reason the columns with timestamp (`'idate'`,`'otime'`) were the reason, and the redefining them solved this problem:
`analogs['idate'] = pd.to_datetime(analogs['idate'], errors='coerce').`
Additionally, the definition of `nwp` variable ("current NWP forecast") is changed from dataframe to numpy array within `mymetrics` function. The time needed for one forecast (all stations) run is now less than 20 minutes which seems reasonable. The time needed to run the script without time window center shifting ("regular analogs") is reduced from 14 minutes to less than 10.

In Fig. 1 an example of forecasting using the regular AN method (left) and using the modified analog-based forecast (mAN) is shown. Ten analog ensemble members are used here with the ensemble mean forecast of the AN shown in red. Observations are denoted by the green line. The example of the modified analogs forecast, mAN, is shown in the right figure. The AN and mAN forecasts were initialized at 8 July 2017 at 00 UTC. One can see that the AN and mAN are very similar, but not the same, indicating that at least some of the most similar historical NWP situations really do happen sooner or later in the day.

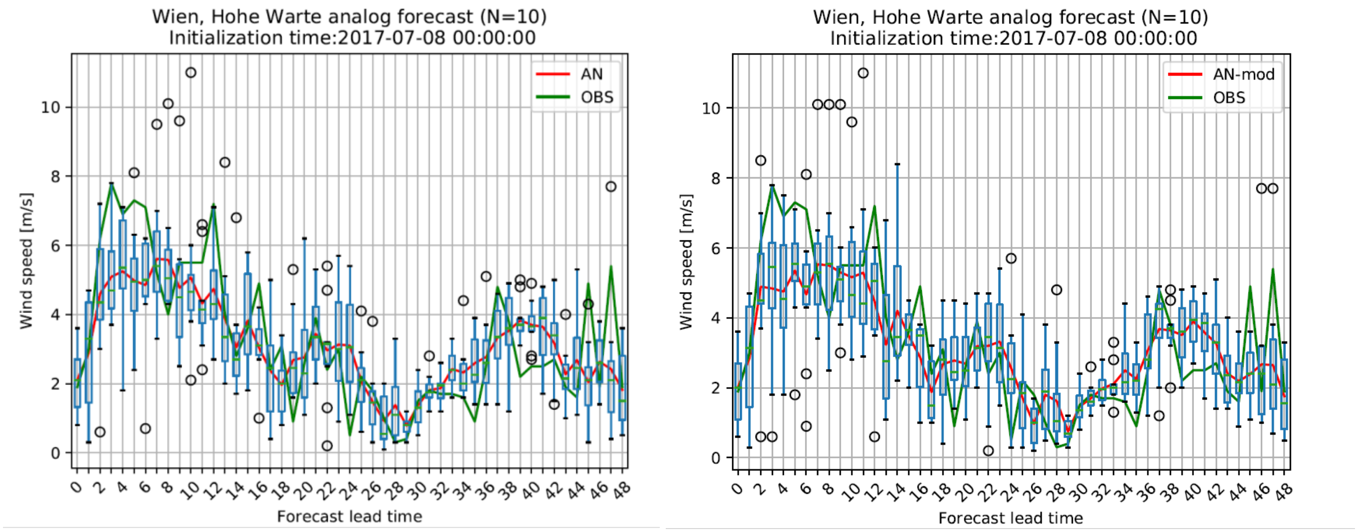


Figure 1. The example of the analog-based forecast for Hohe Warte station initiated at 2017/07/08 (up to 48-h forecast lead time). The ensemble consists of 10 members. The spread of the ensemble is represented by boxplots, where circles represent the outliers. The red line represents AN (left subplot) and modified mAN (right subplot)– forecasting the mean of the ensemble. The results are compared to observations (green line).

Results– deterministic analog forecast: Shifting time window

The algorithm was evaluated for two selected months. One winter, January 2017, and one summer month, July 2017, were chosen. Data of the period 2015 – 2016 was used as training data. The analog method setup for AN and mAN is equal, consisting of 17 analog ensemble members. The procedure at this part is almost exactly the same as during previous stay – the forecasts are joined together with corresponding verifying observations by using `IOP-preplot-merge.py` script. The mean of the ensembles (members 1-17) are calculated and added as a separate forecast '`an`'. The data needed for the verification procedure is saved to `Jan_shifted_results.db` and `Jul_shifted_results.db` databases, while the results for 'regular' analogs (as during the previous stay) are used from `Results_Jan_2017.db` and `Results_Jul_2017.db`. Note: the AR forecast within these databases needs to be checked every time, since during previous stay several database versions were produced. Some of these versions do not have '`ff`' variable adjusted (it should be divided by 10 when loading from AR model files).

The verification procedure is done by using modified `verif` Python package (more info at: https://github.com/WFRT/verif) Now the results for mAN are added to the previous winter-summer comparison of the AR and AN forecast.

It was already shown in the last report that (observed) wind speed, as well as its diurnal cycle is stronger in July than in January. The difference between the AN and mAN forecast is barely noticeable, the distribution looks very similar if boxplots or histograms are used. Therefore, even though all the results are produced as in previous report, they are not shown since they carry no new information.

Since the difference between AN and mAN is very small and can hardly be seen by using boxplots or histograms as before, the comparison of the mean values is now added for bias, root-mean-square-error (RMSE) and correlation coefficient (CC) measures at Tables 1 and 2.

Table 1. Average values of several verification measures for AR, AN and mAN forecasts for all available stations and all lead times during January. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value).

January	AR	AN	mAN
Bias [m/s]	0.4794	-0.1019	-0.1093
CC	0.5732	0.8358	0.8382
RMSE [m/s]	2.1180	1.2890	1.2810
Spread [m/s]	x	1.2730	1.2610
BSS (>5m/s)	x	0.4906	0.4956
CRPS	x	0.6210	0.6169

Table 2. Average values of several verification measures for AR, AN and mAN forecasts for all available stations and all lead times during June. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value).

July	AR	AN	mAN
Bias [m/s]	0.3466	-0.0831	-0.0801
CC	0.4373	0.7585	0.7599
RMSE [m/s]	1.8020	1.1460	1.1430
Spread [m/s]	x	1.0800	1.0830
BSS (>5m/s)	x	0.3803	0.3836
CRPSS	x	0.5674	0.5664

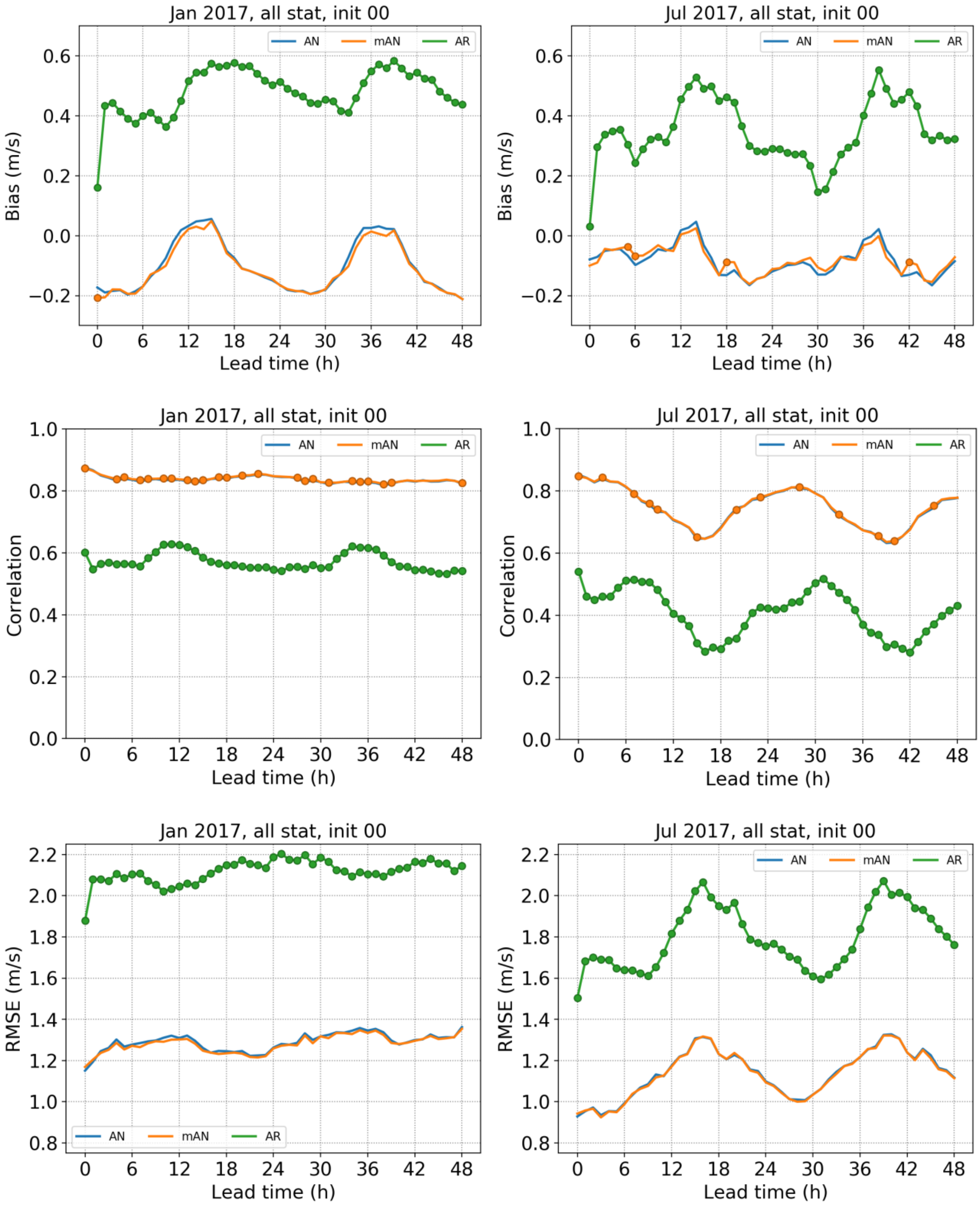


Figure 2. The bias (upper), CC (middle) and RMSE (lower) for the AR, the AN and the mAN wind speed forecasts for January (left) and July (right) 2017.

The average values of bias and RMSE show that the AR error is smaller for July (than for January), but the CC is higher for January. The AN forecast exhibits better results on average than AR, for all the metrics used and both months tested. Further improvements are often achieved with mAN forecast, especially in July.

The results (bias, RMSE and CC) depending on lead time are shown at Figure 2. Both AN and mAN results, compared with the AR results, show a great improvement achieved via post-processing. The difference between the different variations of the analog approach is much less pronounced and rarely significant. However, it can be seen that the bias is somewhat smaller for the mAN forecast. For July it is smaller for majority of the forecast lead times, while the most evident difference is during the afternoon hours in January. For January RMSE also seems to be a bit smaller and CC higher for mAN forecast when compared to the AN forecasts. However, these improvements are not as evident for a summer month.

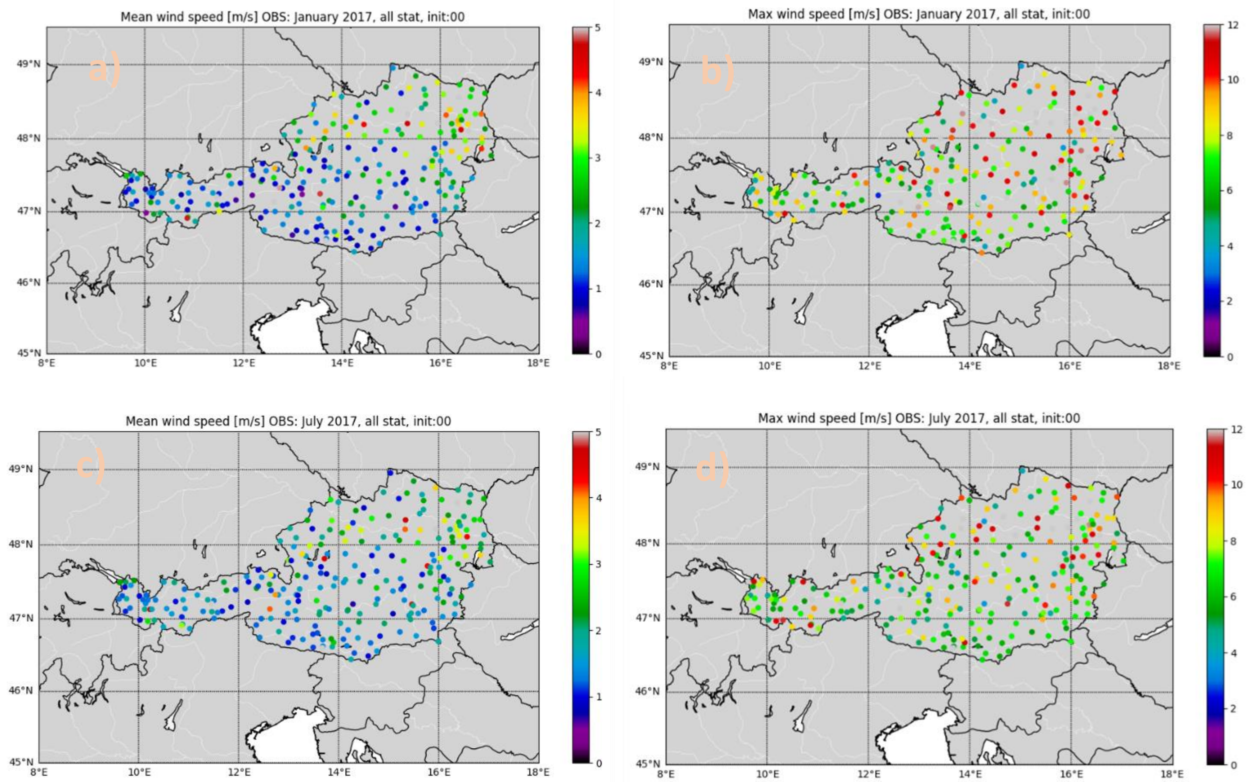

Mean wind speed [m/s] OBS: January 2017, all stat, init:00
a)
Max wind speed [m/s] OBS: January 2017, all stat, init:00
b)
Mean wind speed [m/s] OBS: July 2017, all stat, init:00
c)
Max wind speed [m/s] OBS: July 2017, all stat, init:00
d)
49°N
48°N
47°N
46°N
45°N
8°E
10°E
12°E
14°E
16°E
18°E
0
1
2
3
4
5
0
2
4
6
8
10
12

Figure 3. The spatial distribution of the monthly mean (a, c) and maximum (b, d) of the observed wind speed in the January (upper) and July (down), 2017.

The wind speed was weak and moderate (i. e. < 8 m/s) for both January and July at majority of the stations (Figure 3). The average and the maximum monthly wind speed increases towards north-eastern part (Pannonian plate) for both January and July. The average wind speed across all stations is higher in January (2.76 m/s) than in July (2.27 m/s).

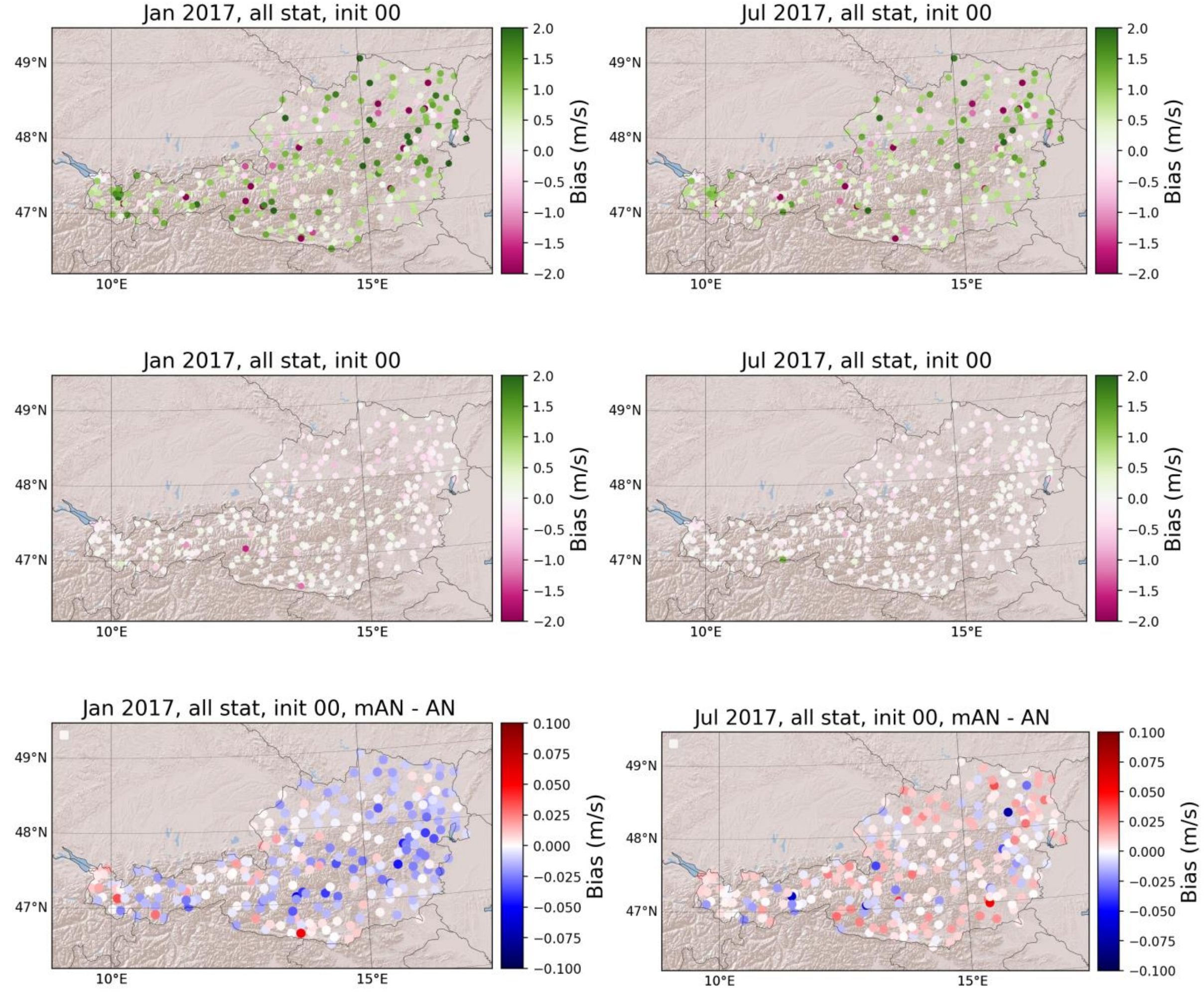

Jan 2017, all stat, init 00
Jul 2017, all stat, init 00
Jan 2017, all stat, init 00
Jul 2017, all stat, init 00
Jan 2017, all stat, init 00, mAN - AN
Jul 2017, all stat, init 00, mAN - AN
49°N
48°N
47°N
10°E
15°E
Bias (m/s)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

Figure 4. The spatial distribution of the monthly mean bias for AR (upper) and AN (middle row) forecast, while the bottom figures show the difference between mAN and AN forecast. All figures correspond to the month of January (left) or July (right), 2017.

The AR forecast bias is slightly positive on average at majority of the stations, especially in January (Figure 4). In both winter and summer month, there is a positive bias in the northeast area (Pannonian plain) and near Warth for the AR forecast. Warth is small high-altitude municipality situated on the complex terrain near the border of the Vorarlberg and the Tyrol in the western Austria. The AN mean bias is smaller than for the AR forecast, showing that the method is able to improve raw NWP forecast. At the locations in the Pannonian plain the bias is slightly negative while at the Alps the bias is positive in January. Differences in bias between the AN and the mAN methods are small, therefore, the differences between these two are shown in the last row in Fig. 4. One can see that the bias is reduced in central Austria during January using the mAN method. When compared to the bias results shown in Figure 2, one can assume that this lower-value bias might correspond to the small positive bias in the AN forecast during winter afternoons.

The correlation coefficient (CC) reduces its value from northeast area towards west and south-west of Austria for any forecast tested (Figure 5). Also, the values are higher for the January than the July. This is regardless of the exact forecast and time of a year. Therefore, it could be concluded that the wind speed is less predictable towards the west and during winters. All forecasts have very low values in the Alps. The CC values as low as shown can suggest very unpredictable month, but also a potential error made in forecasting, loading the data or analysis. However, there is an evident improvement achieved with post-processing for January and especially July for both analog-based variations. Additionally, allowing flexible analog search time window, the results improve within the Alps area the most, especially during the winter month.

The RMSE values seem to be slightly higher during January than in July for all the forecasts (Figure 6). The values for the monthly mean RMSE are higher for the AR than for the AN forecasts in both January and July cases. The error seems to be larger in the central Austria and in the south-western part of Austria for the AR forecast (especially for July), while there is a here is no obvious spatial distribution of error for the AN forecasts. The error difference between AN and mAN seem to be very small and there is no obvious spatial distribution.

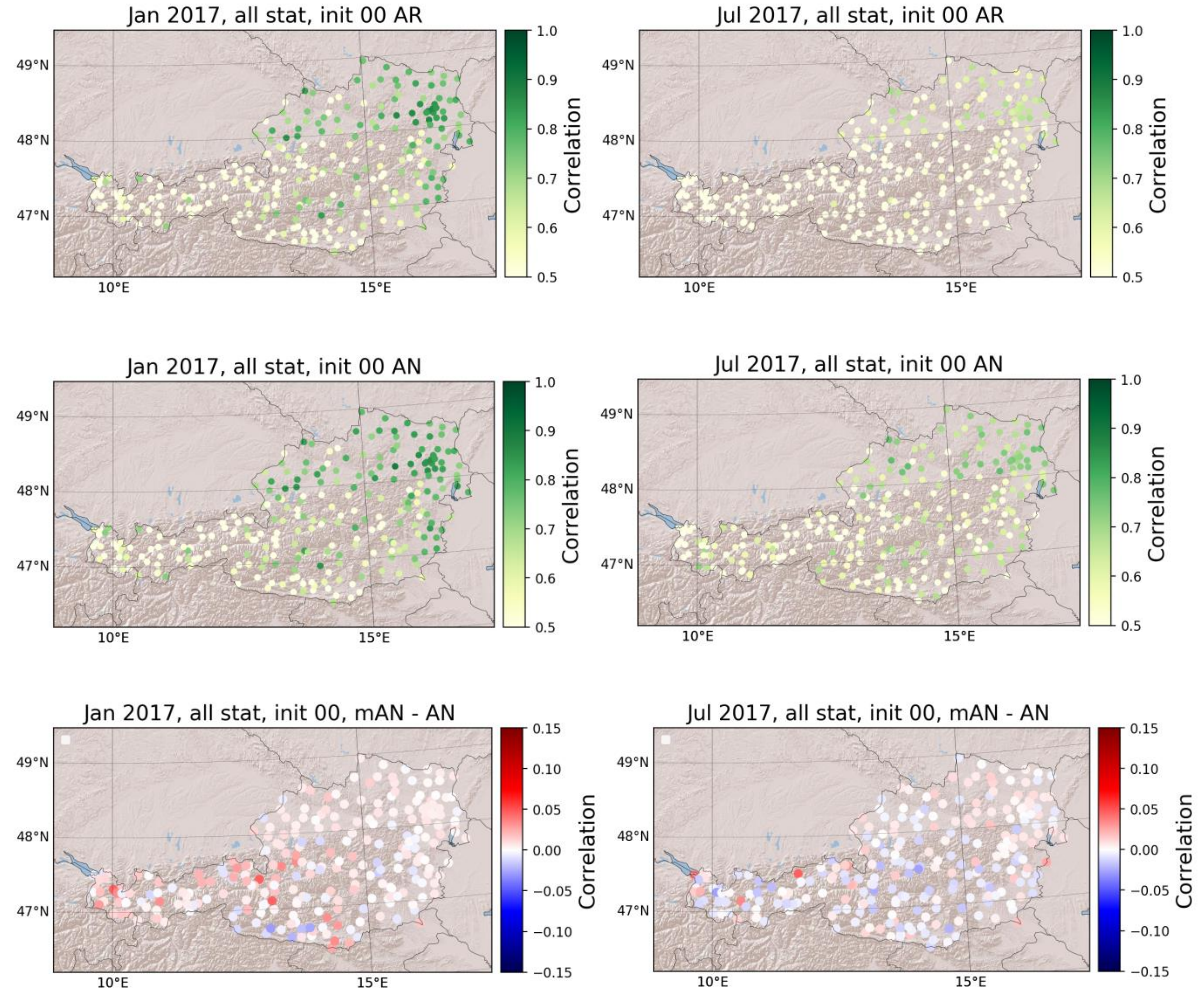
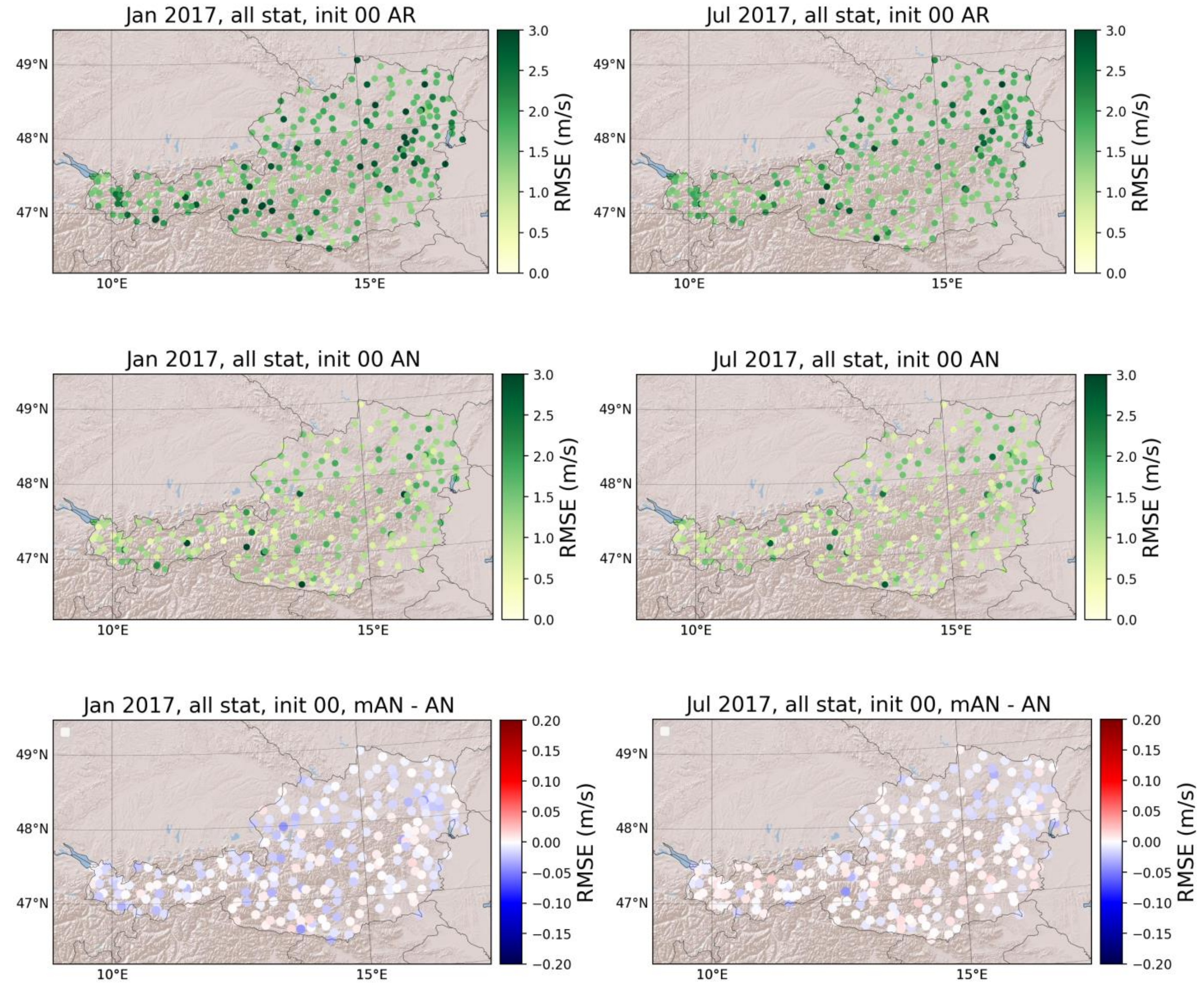

Jan 2017, all stat, init 00 AR
Jul 2017, all stat, init 00 AR
Jan 2017, all stat, init 00 AN
Jul 2017, all stat, init 00 AN
Jan 2017, all stat, init 00, mAN - AN
Jul 2017, all stat, init 00, mAN - AN
49°N
48°N
47°N
10°E
15°E
Correlation
1.0
0.9
0.8
0.7
0.6
0.5
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15

Figure 5. The spatial distribution of the monthly mean correlation coefficient for AR (upper) and AN (middle row) forecast, while the bottom figures show the difference between mAN and AN forecast. All figures correspond to the month of January (left) or July (right), 2017.



Jan 2017, all stat, init 00 AR

Jul 2017, all stat, init 00 AR

Jan 2017, all stat, init 00 AN

Jul 2017, all stat, init 00 AN

Jan 2017, all stat, init 00, mAN - AN

Jul 2017, all stat, init 00, mAN - AN

Figure 6. The spatial distribution of the monthly RMSE for AR (upper) and AN (middle row) forecast, while the bottom figures show the difference between mAN and AN forecast. All figures correspond to the month of January (left) or July (right), 2017.

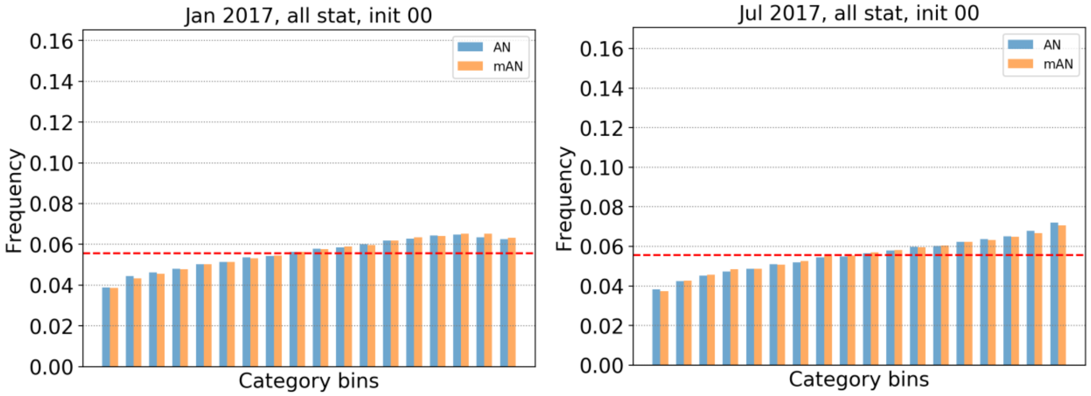


Figure 7. The the rank histograms for AN and mAN probabilistic forecasts at all available stations for January (left) and July (right), 2017.

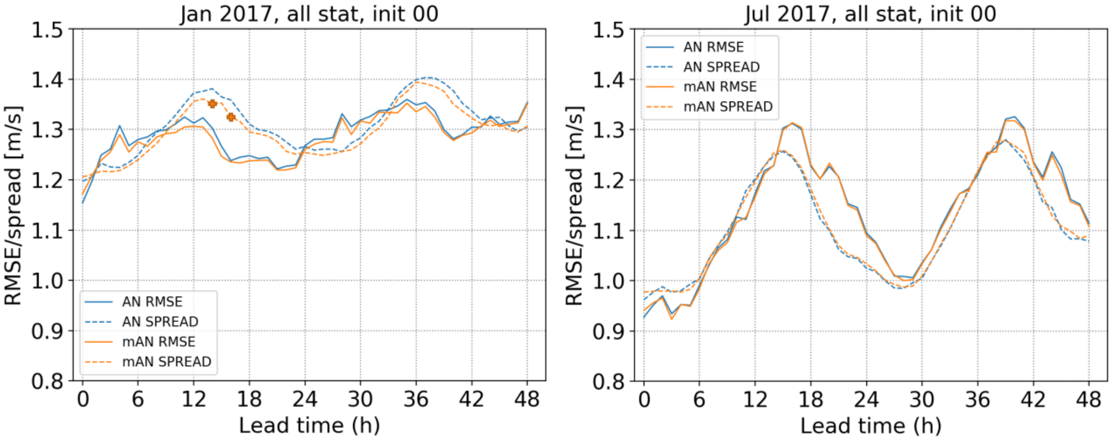
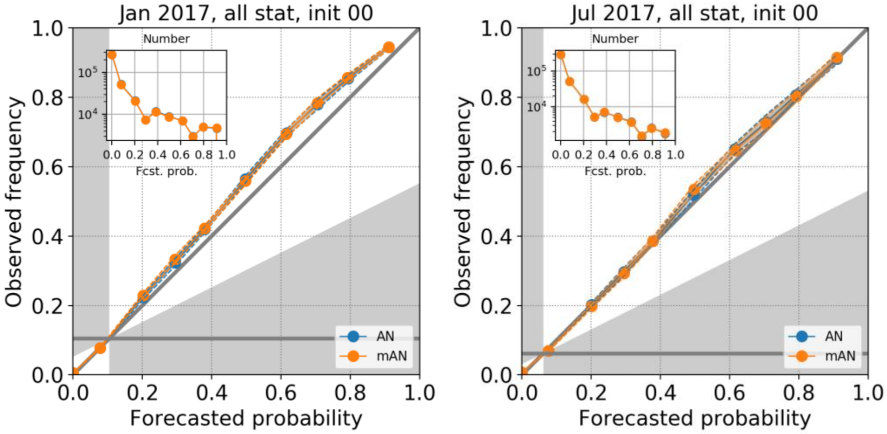


Figure 8. Spread-skill diagram for AN and mAN forecasts at all available stations for January (left) and July (right), 2017



.

Figure 9. The reliability diagram for AN and mAN probabilistic forecasts of wind speed to exceed 5 m/s at all available stations for January (left) and July (right), 2017.

Finally, mAN forecast is compared to AN forecast using probabilistic verification scores that have not been computed during previous stay. The AN and mAN rank histogram shows slight underestimation for both months tested, which is consistent with slightly negative ensemble mean bias (Figure 7). There are no meaningful differences between AN and mAN at this point. The spread-skill diagram shows that AN and mAN error match the ensemble spread adequately for both months tested (Figure 8). The mAN result shows both error and spread reduction in January, while the mAN result for July is more similar to AN.

The reliability diagrams for AN and mAN are almost indistinguishable (Figure 9), both forecasts showing good resolution and slight under-forecasting, especially for January.

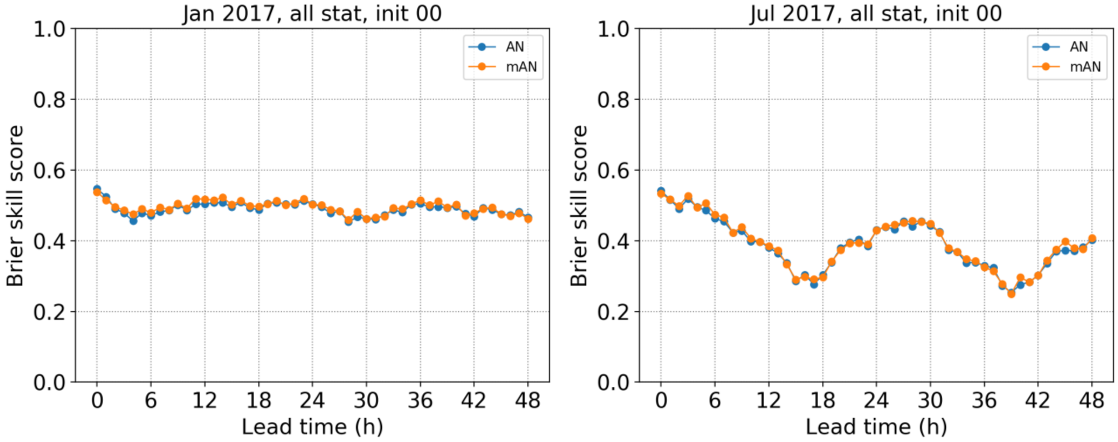


Figure 10. The Brier skill score (BSS) depending on lead time for AN and mAN probabilistic forecasts of wind speed to exceed 5 m/s at all available stations for January (left) and July (right), 2017.

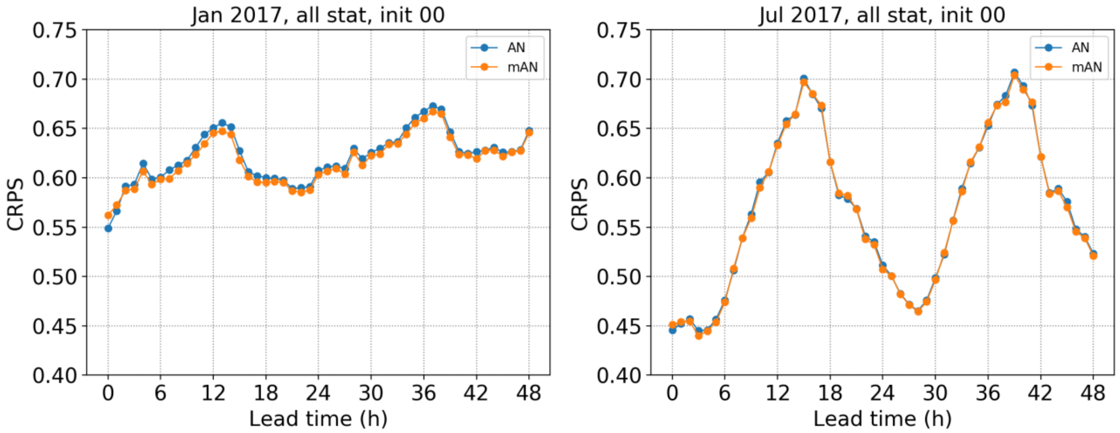


Figure 11. Continuous rank probability score (CRPS) depending on lead time for AN and mAN probabilistic at all available stations for January (left) and July (right), 2017.

The further mAN over AN forecast improvement is indicated by slightly higher Brier skill score (calculated for 5 m/s wind speed threshold) and lower continuous rank probability scores for January (Figures 10 and 11). The differences are small (and probably not significant in this experiment), but extend for the majority of lead times. The differences for the summer month are less pronounced. The BSS dependency on probabilistic forecast threshold (for wind speed to exceed) reveals that mAN and AN forecasts exhibit similar values for low and moderate wind (i.e. up to 10 m/s) (Figure 12). However, the probability of wind speed exceeding 10 m/s or more (this is climatologically less frequent event than low or moderate wind speed) is better predicted by mAN than AN. Even though these results are probably not significant at this point, they indicate that the true benefit from extending the “analog-search pool” (mAN) could be revealed in extended experiment. For instance would have shorter training period and concentrate on climatologically rarer events.

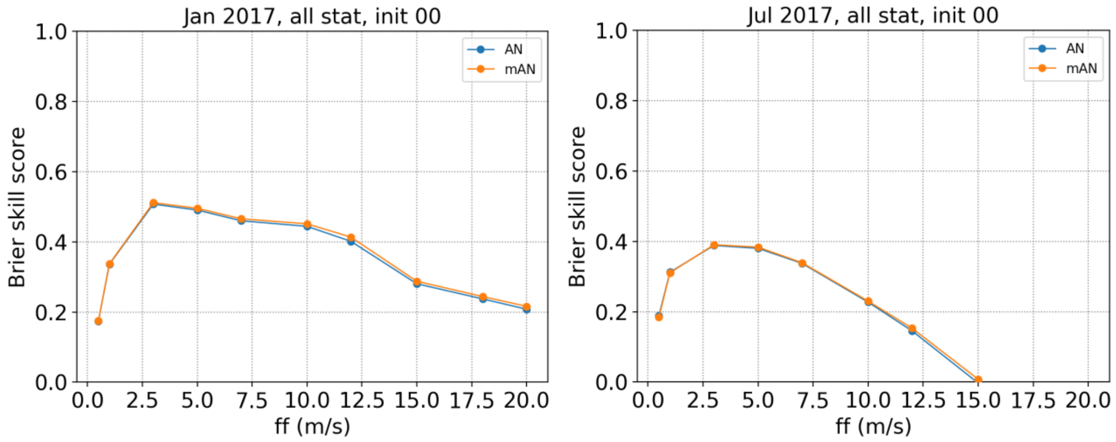


Figure 12. The Brier skill score (BSS) depending on wind speed threshold for AN and mAN probabilistic forecasts at all available stations for January (left) and July (right), 2017.

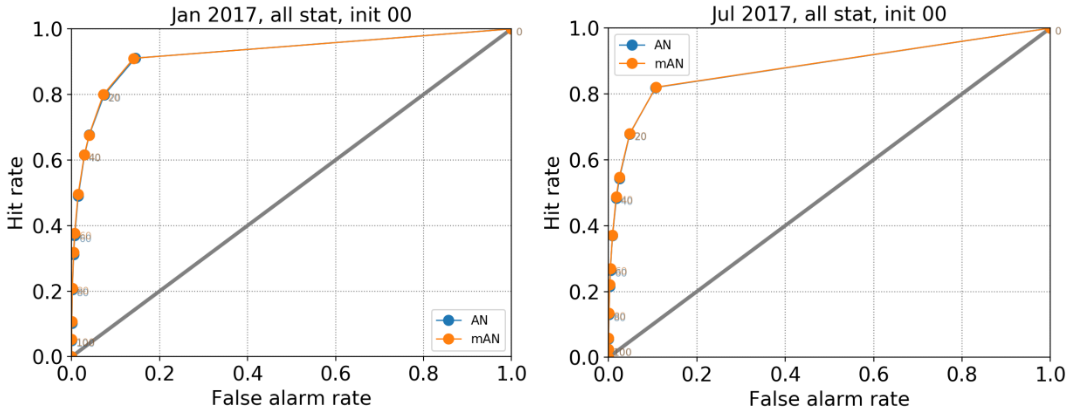


Figure 13. The Relative Operating Curve (ROC) for AN and mAN probabilistic forecasts at all available stations for January (left) and July (right), 2017.

Finally, Relative Operating Curve (ROC) diagram shows that both forecasts discriminate event (wind speed exceeding 5 m/s) from non-event better in January if compared to July. This is partially explained by the fact that this event occurs more often in January (can be seen at the reliability diagrams at Figure 9). There are no apparent differences among AN and mAN ROC curves.

PART 2: Probabilistic analog forecast input: Post-processing LAEF forecasts

After testing the analog forecasts with AR deterministic input, next step is to test several different configurations of LAEF ensemble forecast input. The following configurations are tested:

a) LA_Ws: raw LAEF wind speed ensemble forecast (17 members)

b) AN_Ws: LAEF wind speed ensemble forecast used as predictors (17 predictors)

c) AN_Me: The means of the LAEF ensemble forecast for the wind speed, direction, temperature (2 m), relative humidity, pressure and precipitation (6 predictors)

d) AN_St: The means and the standard deviation of the LAEF ensemble forecast for the wind speed, direction, temperature (2 m), relative humidity, pressure and precipitation (12 predictors)

e) AN_Al: All members of the LAEF ensemble forecast for the wind speed, direction, temperature (2 m), relative humidity, pressure and precipitation (6×17 predictors)

f) AN_11: For every member of the LAEF ensemble forecast analog search includes the for the wind speed, direction, temperature (2 m), relative humidity, pressure and precipitation (6 predictors) among the same member historical forecasts. Therefore, in this member-by-member approach, 6 predictors are used, but the search algorithm is used 17 times for every forecast. This is the most demanding configuration at the moment. The code could be optimized if it would produce the best result. However, since that is not the case, it will not be done.

If more than one meteorological variable is used as a predictor, the predictor variable choice corresponds to the previous case (AR deterministic input).

The list of the stations is now reduced, due to more computational power needed. In total, 29 stations are used, marked with following numbers:

11007,11012, 11025, 11035, 11036, 11060, 11070, 11101, 11108, 11126, 11170, 11180, 11190, 11198, 11204, 11216,11246, 11273, 11290, 11320, 11343, 11344, 11346, 11350, 11358, 11380, 11383, 11384, 11389.

All the lists of stations can be found in `ANENm_stationlist.py` module, and this one is listed as `included_stations_laef_reduced` function.
The time window is fixed and includes one lead time step before/after to include trend. The analog ensemble members are saved up to the 20 members, but testing is carried out for 17 member ensemble as the LAEF ensemble consists of 17 members (16 members plus one control member) and it is close to the optimal size which includes 15 members (as in Odak Plenkovic et al., 2018).

Scripts and optimization - probabilistic analog forecast input: Post-processing LAEF forecasts

Using an ensemble as input for the AN method involves the implementation of several databases for training. The scripts for loading include the main program (depending on the exact input) and the module `ANENm_loading_data.py`. The main programs corresponding to the previously listed inputs are:

a) `ANEN_loading_laef_WS_ens.py`
b) and c) `ANEN_loading_laef_means.py`
d) `ANEN_loading_laef_all_reduced.py`

In the main program the path and name of the LAEF files to be loaded are being adjusted as well as the database name. Then the module `ANENm_loading_data.py` uploads the actual data by using appropriate one out of several functions (starting with “`load`”) depending on the forecast. The data is saved in a form of the numpy array first (due to computational reasons during loading and pasting the data needed). The separate function is made to transform the numpy array to dataframe with predefined column names after all the adjustments are done. These functions are also stored in ANENm_module and start with “`np2df`”. The `ANEN_adjust_db.py` script can be used if the database needs adjustment (i.e. data type or reducing number of stations). The databases containing the reduced list of stations used for testing are:

a) `TrainLaefWind2015_6_reduced.db`
b) and c) `TrainLaefMean2015_6_reduced.db`
d) `TrainLaefAll2015_6_reduced2.db`

Note: the 2015-11-30 LAEF files are corrupted and deleted for all the variables.

The analog ensemble forecasts are carried out using ANEN_analogs.py main program (except AN_11 forecast) and ANENm_analogsearch.py module. The analog ensemble (initial) forecasting dates (and lead times) correspond to the “currently” loaded AR forecast (the one to be post-processed) defined at the beginning of the script. The main program can be easily adjusted to be used in any listed experiment. The differences between configurations are: the input/output database names, predictors used (the name of the input dataframe columns used as predictors are listen in mycol variable), the number of circular variables c (they are always placed at the left side – the first columns) and function names from the module. For producing AN_11 forecasts ANEN_analogs.py is used. It is very similar to ANEN_analogs.py, but it calls the search algorithm for every LA_Ws member separately (using the same module). The module ANENm_analogsearch.py contains two functions – a simpler one adjusted for case a) anen_laef_ws where there are no circular variables included, and a more general one anen_laef_all that works for all the other cases. The output from the module is exactly as it was when using IOP-analogs.py script – 20 members are saved and ordered from the most to the least similar to the current forecast. The results for January and June 2017 are stored and then merged with corresponding observations (IOP-preplot-merge2.py script), then stored again in the databases named “Res_” + forecast name + month. The same is done for the LAEF wind speed ensemble forecast (LA_Ws). The examples for these analog-based forecasts are plotted for the same station and date as at Figure 1 (Figure 14).

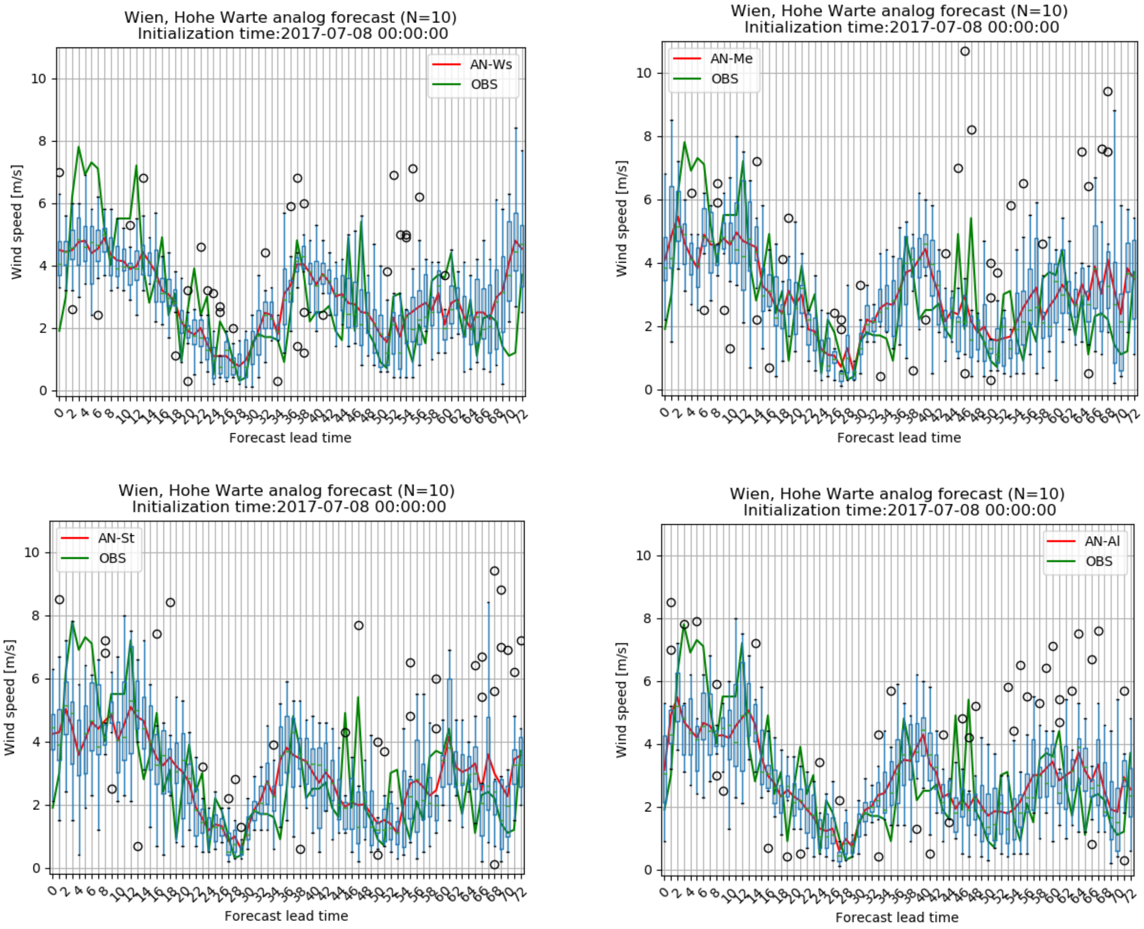


Figure 14. The example of four configurations (AN-Ws, AN-Me, AN-St, AN-Al) of the analog-based forecast using the LAEF input for Hohe Warte station initiated at 2017/07/08 (up to 72-h forecast lead time). The ensemble is consisted of 10 members. The spread of the ensemble is represented by boxplots, where circles represent the outliers. The red line represents analog ensemble mean. The results are compared to observations (green line).

Results - probabilistic analog forecast input: Post-processing LAEF forecasts

The evaluation of the probabilistic input gives new insights into the methodology. All new experiments show an improvement of the original LA_WS forecast with the analog approach (Tables 3 and 4). One can, too, clearly identify differences between the new methods.

Table 3. Average values of several verification measures for LA_Ws and analog-based forecasts (AN_Ws, AN_Me, AN_St, AN_Al and AN_11) for all available stations and all lead times during January. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value).

January	LA_Ws	AN_Ws	AN_Me	AN_St	AN_11	AN_Al
Bias [m/s]	-0.1898	-0.0329	-0.1476	-0.1392	-0.0885	-0.1756
CC	0.4681	0.8275	0.8517	0.8479	0.7783	0.8484
Disp. Err-[m/s]	2.4427	1.6673	1.4541	1.4726	1.8063	1.4596
STD bias [m/s]	-1.1790	-0.3690	-0.6800	-0.6800	-0.6020	-0.7110
RMSE [m/s]	2.7190	1.7080	1.6120	1.6280	1.9060	1.6330
Spread [m/s]	0.7115	1.6330	1.6810	1.6660	1.5790	1.6670
BSS (>5m/s)	0.0212	0.5156	0.5267	0.5226	0.4003	0.5200
CRPS	1.5120	0.8026	0.7720	0.7706	0.9588	0.7799

Table 4. Average values of several verification measures for LA_Ws and analog-based forecasts (AN_Ws, AN_Me, AN_St, AN_Al and AN_11 for all available stations and all lead times during July. The best result among compared forecasts is underlined (the spread is better when closer to the RMSE value).

July	LA_Ws	AN_Ws	AN_Me	AN_St	AN_11	AN_Al
Bias [m/s]	-0.2564	-0.1337	-0.1380	-0.1264	-0.1446	-0.1758
CC	0.3850	0.7265	0.7572	0.7605	0.6814	0.7508
Disp. Err-[m/s]	1.8521	1.4519	1.3240	1.3202	1.4941	1.3346
STD bias [m/s]	-0.9180	-0.4330	-0.5440	-0.5320	-0.593	-0.5600
RMSE [m/s]	2.0830	1.5210	1.4380	1.4290	1.6140	1.4580
Spread [m/s]	0.7720	1.3370	1.3660	1.3920	1.4050	1.3250
BSS (>5m/s)	0.0666	0.3459	0.3730	0.3719	0.2660	0.3633
CRPS	1.1310	0.7364	0.6965	0.6957	0.8049	0.7047

The observed average monthly wind speed is higher in January (2.76 m/s) than in July (2.27 m/s), across all available stations and lead times. Additionally, standard deviation of the wind speed measurements is also higher on average in January (3.03 m/s) than in July (2.19 m/s).

It is possible to specify the source of the error by decomposing the RMSE to the bias of the mean (or simply bias), bias of the standard deviation and dispersion (phase) error (e.g., Murphy 1988; Horvath et al. 2012). In this work STD bias stands for the bias of the standard deviation for the ensemble mean (regardless of the ensemble spread). The average values of LA_Ws ensemble mean bias are small, showing only slight wind speed underestimation of 0.19 m/s in January and 0.25 m/s in July. The standard deviation of the measurements is underestimated by LA_Ws ensemble mean in both January (by 1.18 m/s) and July (by 0.92 m/s). The average CC value for the LA_Ws ensemble mean is higher in January than in July by 0.08. The average RMSE of LA_Ws ensemble mean is 2.72 m/s in January, while it is lower by 0.64 m/s in July (partially due to climatologically lower monthly wind speed). Since average bias of the mean is low, the dominant sources of error are dispersion error followed by the biased standard deviation of the LA_Ws ensemble mean.

Every AN forecast (AN_Ws, AN_Me, AN_St, AN_Al and AN_11) ensemble means exhibit better results on average than AR, lowering all three error sources (measured by Bias, CC and STD bias) during both months tested. After the analog-based post-processing, dispersion error is even more dominant among error sources. AN_Ws is very successful in removing systematic errors (bias and STD bias), especially in January. However, AN_Me, AN_St and AN_Al are more successful than AN_Ws in removing predominant dispersion source of error for ensemble mean. The AN_Me exhibits the best overall results for the ensemble mean for January and AN_St for July. The AN_11 ensemble mean seems to be the least successful among AN forecasts.

The LA-Ws ensemble forecast shows under-spread, with average spread lower than average RMSE for 2.01 m/s (January) and 1.31 m/s (July). The average spread matches average RMSE better after any analog-based post-processing. The AN_St exhibits the best spread among analog-based experiments (the closest value to average RMSE in January and second closest in

June), followed by AN_Me and AN_Al. The AN_11 exhibits the most under-spread ensembles among analog-based experiments.

Brier Skill Score (BSS) is commonly used skill measure for probabilistic forecast of binary event. The binary event is often determined by threshold (i.e. 5 m/s) when evaluating wind speed forecasts. Then, the probability of exceeding that threshold is forecasted and evaluated. The observed frequency of wind speed exceeding 5 m/s event is higher (it is measured in 17 % cases) for January than for July (11 % cases). The 5 m/s threshold is chosen because it is a reasonably high while not being too rare to produce meaningful result. For example, wind exceeds 8 m/s in July only in 2 % cases, resulting in reliability diagrams that are harder to read and interpret. The BSS for LA_Ws probability forecast (for wind speed to exceed 5 m/s threshold) is 0.02 for January and 0.07 for July. The BSS is improved by any analog-based experiment tested in this work. This is especially the case in January, when the underlying climatology shows it is more common event than in July. The best BSS is achieved in AN_Me experiment for both months tested, followed closely by AN_St and AN_Al, while AN_Ws and especially AN_11 are somewhat less skillful.

Continuous rank probability score (CRPS) is a summary metric that can be interpreted as the integral of the Brier score over all possible threshold values for the parameter under consideration. It is negatively oriented (the lower, the better) accuracy measure that is equal to the mean absolute error (MAE) for deterministic forecast, and also has a value of 0 for the perfect forecast. The LA_Ws shows higher CRPS (1.51) for January than for July (1.12). The CRPS value is improved (lowered) by any analog-based experiment, exhibiting better results for July when wind speed is lower on average than in January. The AN_St experiments show the best results in terms of CRPS, followed closely by AN_Me and AN_Al, while AN_Ws and AN_11 are not as successful.

In addition to overall comparison, forecasts are also compared against lead time in terms of several verification metrics. Even when testing the dependency on the lead time, the differences between LA_Ws and analog approach (AN_Ws, AN_Me, AN_St, AN_Al and AN_11 forecasts) are the least pronounced for the bias of the ensemble mean (compared to differences in CC or RMSE). This is partially due to the fact that the bias for the LA_Ws forecast is small to begin with (just a slight underestimation of the wind speed) (Figure 15). The maximum LA_Ws underestimation at noon and early afternoon is almost completely removed by analog approach during January, while the slight LA_Ws underestimation of the wind speed during nighttime is similar among all the tested forecasts. The AN_Ws exhibits the smallest bias of the mean (smaller than the LAEF LA_Ws) for the January, while the other four AN forecasts produce similar result – somewhat smaller bias with less extreme values if compared to LA_Ws. For July, the improvement over LA_WS is more evident for all configurations tested. The first (daily) LA_Ws bias minimum is almost completely removed by all the AN forecasts, while the second minimum (in the evening) is reduced. The differences among AN experiments variations are less pronounced than in comparison with LA_Ws. However, it seems that the AN_WS is the least biased, while AN_Al underestimates the wind speed the most.

Unlike the bias results, where the differences were subtle, the RMSE and CC show great improvement over LA_WS after using the analog post-processing method for both months tested, regardless of the exact AN experiment. The results among AN forecasts seem similar. The AN_Me, AN_St and AN_Al produce somewhat better results than AN_Ws and AN_11 for both months tested. At this point it seems that using more variables than one, but not necessarily all members of the ensemble shows better results. However, additional testing for the spread of these ensembles is something that needs to be done to better distinguish these subtle differences.

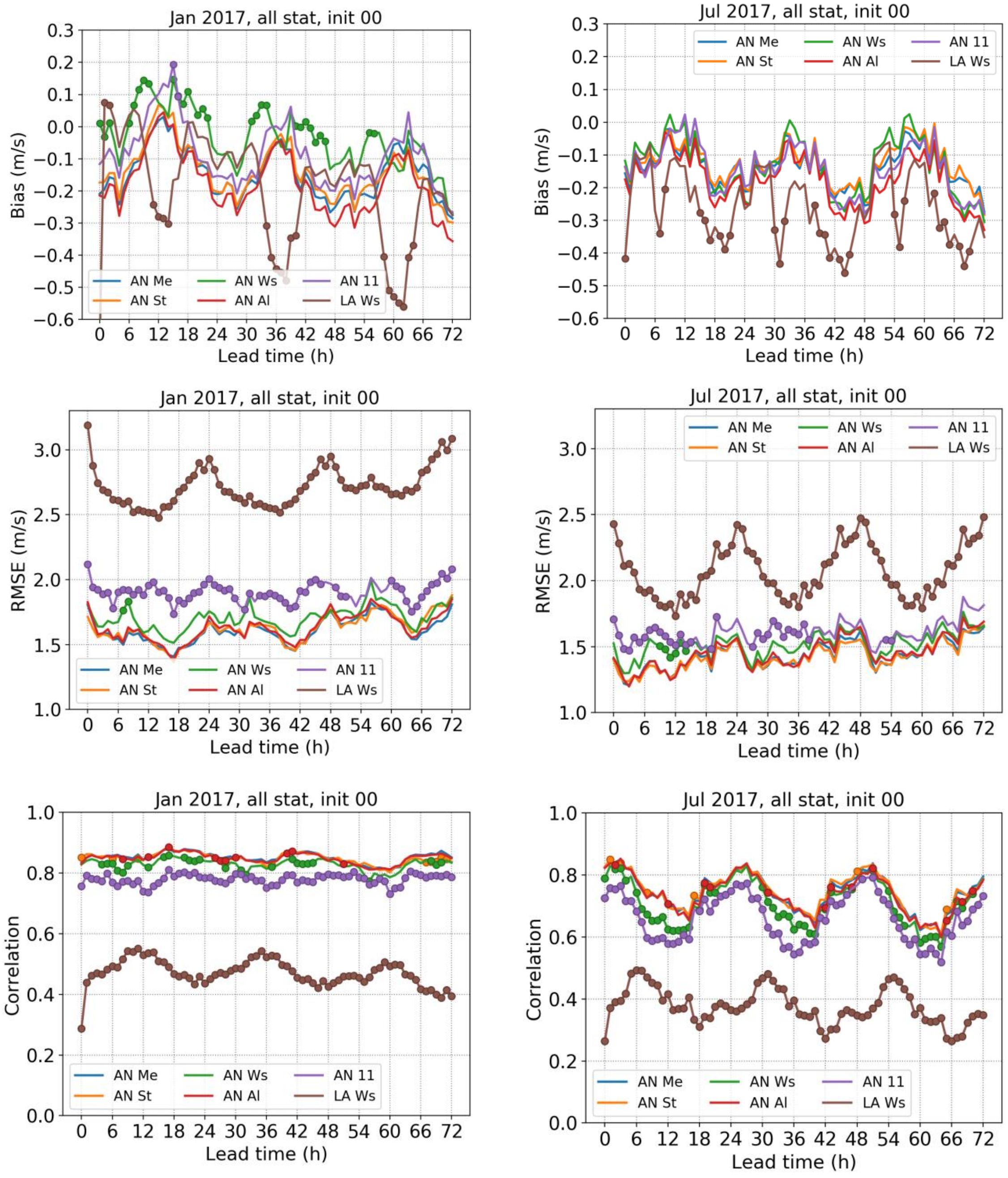


Figure 15. The bias (upper), RMSE (middle) and CC (lower) for the means of the ensemble of LAEF LA_WS forecasts (red line), the five different analog ensemble configurations, at all the stations tested for January (left) and July (right) 2017. The markers are set for the results significantly different from AN_Me forecast (0.05 sig. level).

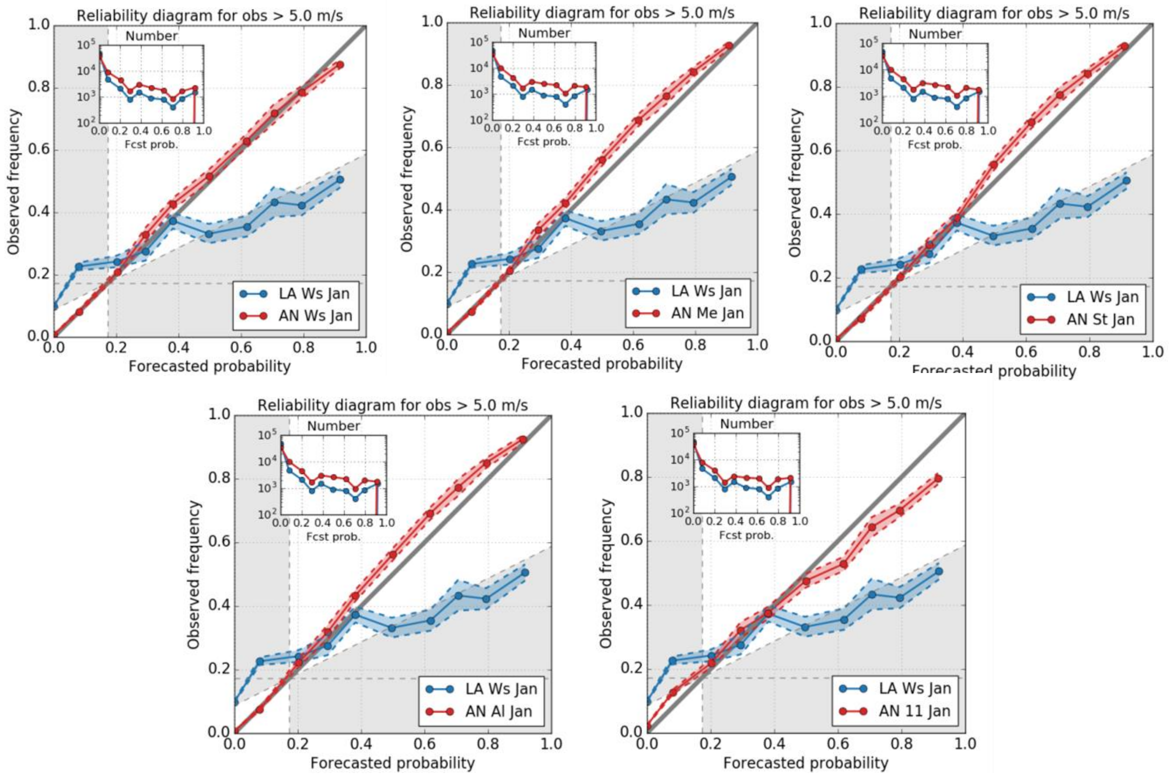


Figure 16. Reliability diagrams for five different AN forecasts, compared to LAEF LA_Ws forecast during January 2017 at all station tested in this study. The dashed lines show 95% confidence interval.

The reliability is tested for probability forecasts that the wind speed will exceed 5 m/s, which happens in 14 % cases tested, more often in January than in July. The LA_Ws ensemble is less reliable for January (Figure 16) when compared to July (Figure 17), and is, too, below the no-skill line for high probabilities forecasted during January. The ensemble based AN forecasts seem to be much more reliable than LA_Ws forecasts in all experiments, showing the skill for any forecasted probability. Small probabilities (i.e. less than 30 % chance for wind speed to exceed 5 m/s) forecasted by the analog approach for all experiments are almost perfectly reliable, while the LA_Ws underestimates the probability. There are only small differences between different AN configurations, especially between AN_Me, AN_St and AN_Al. They slightly overestimate the middle-range probabilities (i.e. 50-80 % probability for the wind speed to exceed 5 m/s). Sometimes the high probabilities (i.e. 90 % chance) are overestimated by the AN.

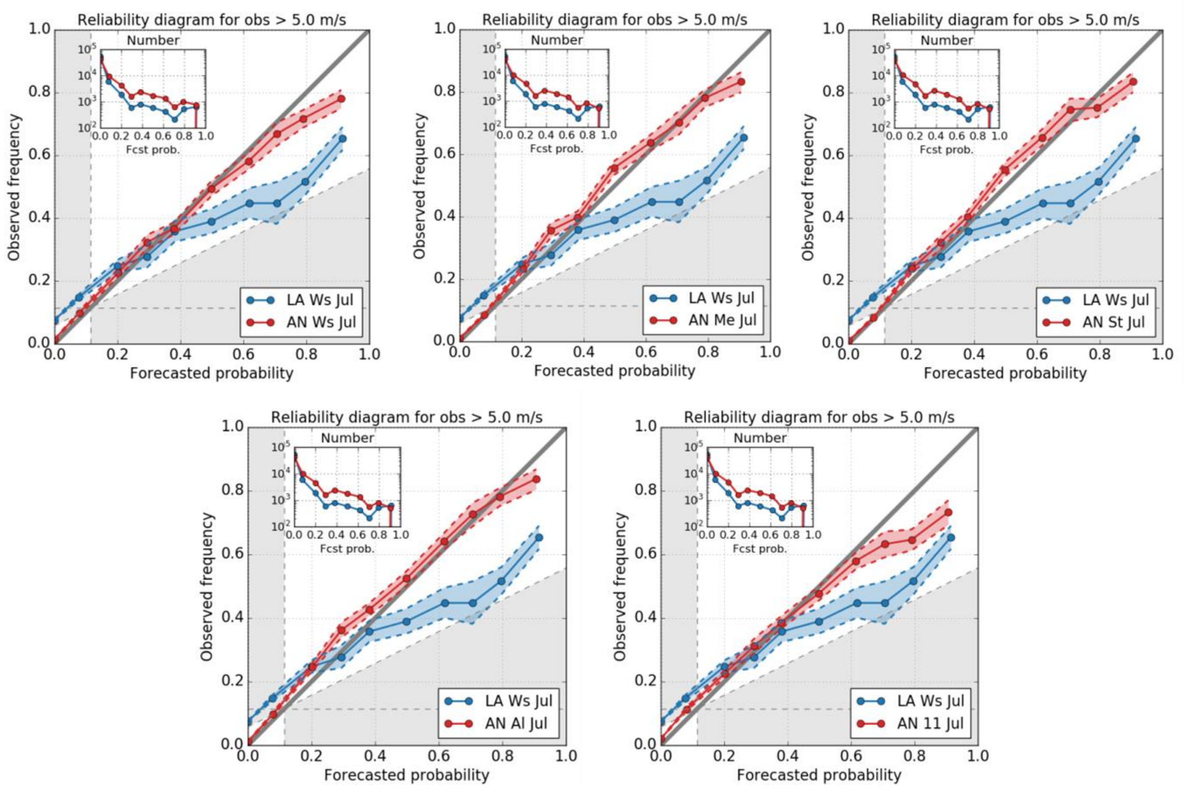


Figure 17. Reliability diagrams for five different AN forecasts, compared to LAEF LA_Ws forecast during July 2017 at all station tested in this study. The dashed lines show 95% confidence interval.

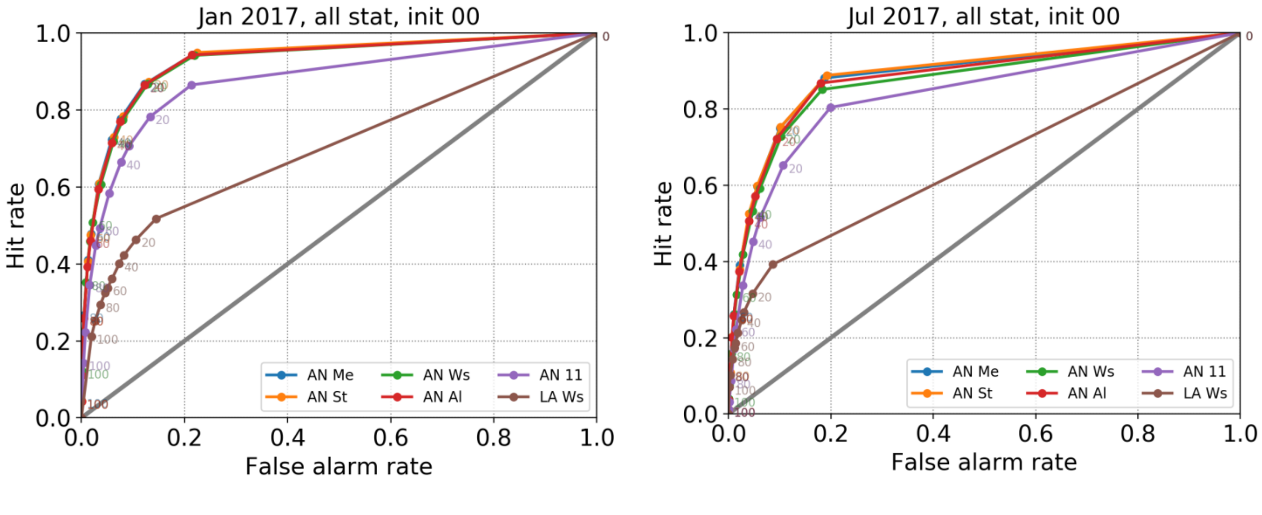


Figure 18. Relative operating characteristic curve for five different AN probabilistic forecasts (for wind speed to exceed 5 m/s), compared to LAEF LA_Ws forecast during January (left) and July 2017 at all stations tested in this study.

method. This is especially the case with the AN_11 configuration which is the least reliable out of the investigated AN forecasts. However, the overestimation of the high probabilities is much less pronounced for any of the AN experiment, compared to LA_Ws ensemble result. Resolution is also higher for any AN forecast, compared to LA_Ws. Additionally, one can notice that the sharpness diagram (upper left corner of the reliability diagram) looks reasonable for both forecasts, but the LA_Ws seem to be a bit sharper than any AN forecast in the experiments performed in this work, indicating higher tendency to forecast extreme probabilities. This is preferable because of better forecast usability if the forecasts are reliable. However, the AN forecasts are more accurate in terms of reliability.

Relative operating characteristic (ROC) curve shows that applying the analog method to the raw LA_Ws improves the discrimination considerably using a threshold of 5 m/s (Fig. 18). Results for January are better than for July indicating that in the convective season most likely a higher resolved NWP model might add some additional information. Similar results are achieved for AN_Me, AN_St and AN_Al forecast, while the AN_Ws and AN_11 are less successful in distinguishing the distribution of the event (wind speed exceeding 5 m/s) from non-event distribution.

The rank histogram shows under-dispersion or LA_Ws forecasts, especially for January (Figure 19). This is not the case for the AN forecasts. The exception is AN_11 forecast, which exhibits some under-spread, also more pronounced for January than for July. The other AN forecasts produce very similar result, showing only slight underestimation of the wind speed values forecasted.

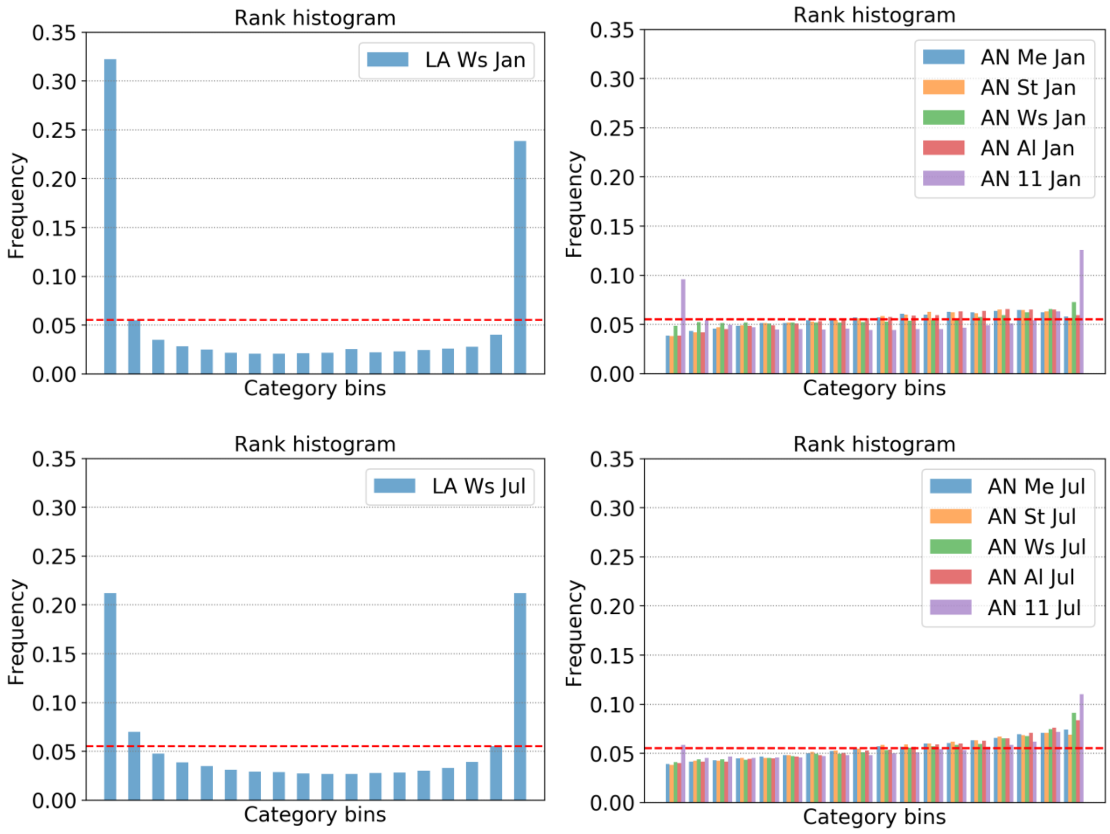


Figure 19. Rank histograms for five different AN forecasts (right), compared to LAEF LA_Ws (left) forecast during January (up) and July (down) 2017 at all station tested in this study.

The spread skill diagrams confirm the LA_Ws forecast are under-dispersive, more during January than in July (Figure 20). The AN_Me forecast is significantly better for both months tested. It shows almost perfect agreement between the RMSE and spread during July, while there it is slightly over-dispersive during January. The other AN forecasts exhibit similar behavior, rarely producing significantly different result. The exceptions are AN_11 and AN_Ws, which for a certain lead times show significantly larger under-spread.

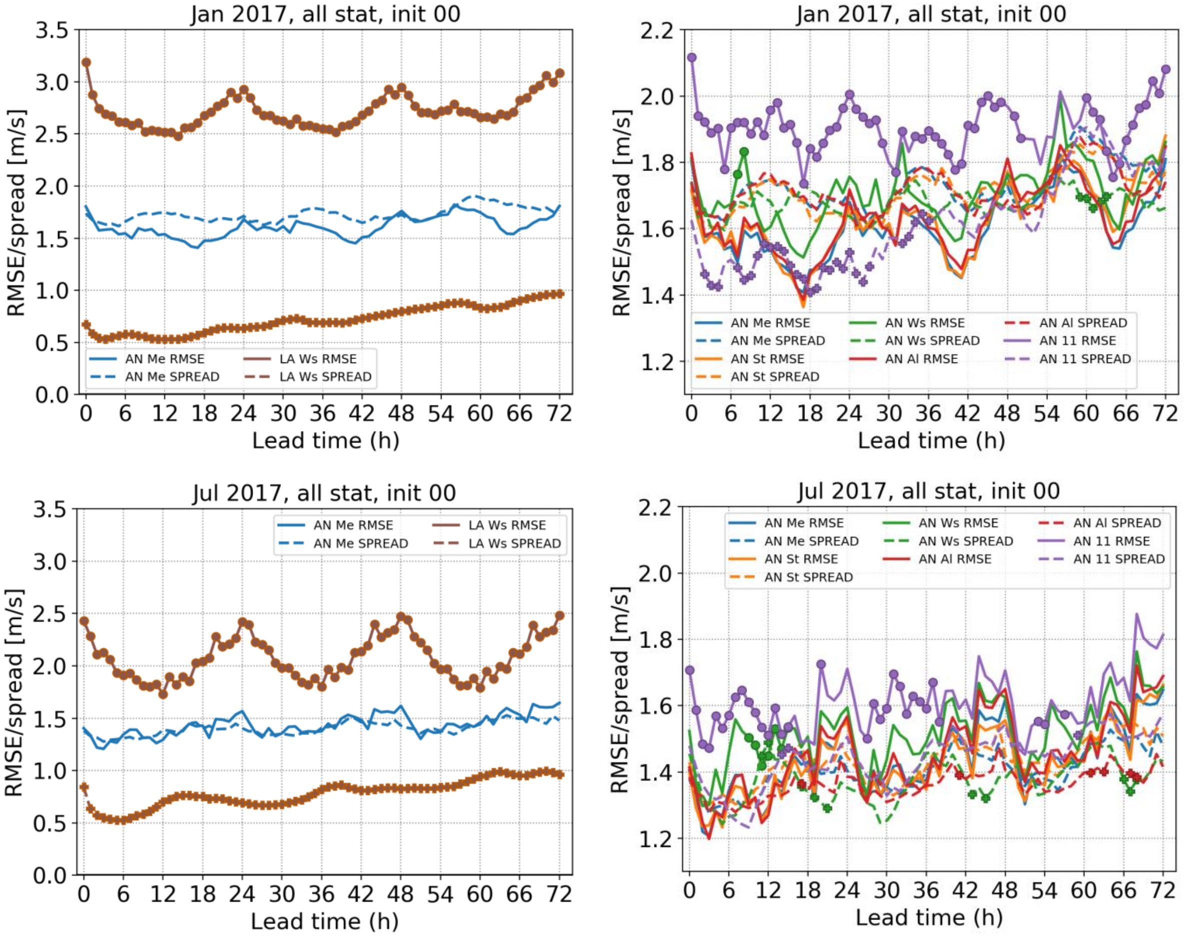


Figure 20. Spread-skill diagrams depending on lead time for AN_Me forecast compared to LAEF LA_Ws (left) and five different AN forecasts (right) during January (up) and July (down) 2017 at all station tested in this study. The markers denote the result significantly different from AN_Me forecast (0.05 sig. level).

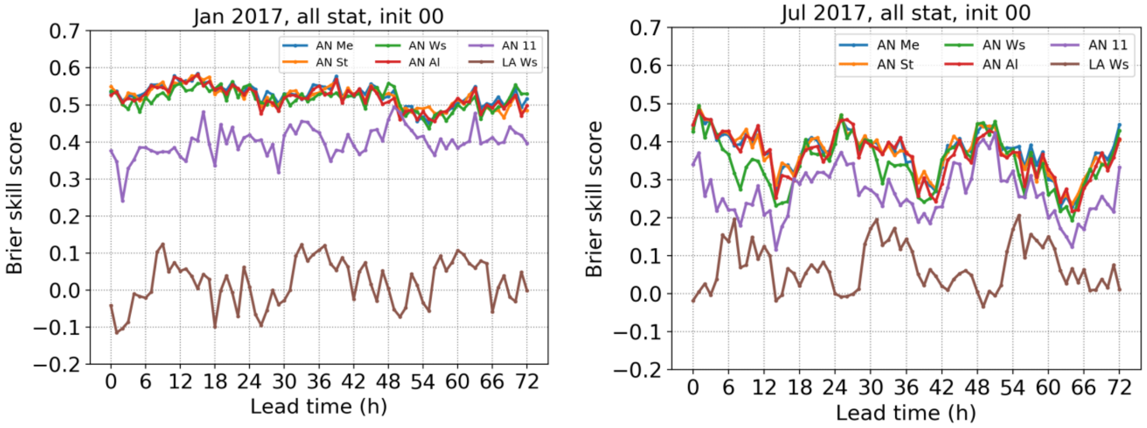


Figure 21. Brier skill score depending on lead time for AN probabilistic forecasts (probability that the wind speed will exceed 5m/s), compared to LAEF LA_Ws during January (left) and July (right) 2017 at all station tested in this study.

Considering the Brier skill score of the probabilistic forecast for the wind speed exceeding 5 m/s, it can be noticed that the improvement of the AN over LA_Ws is noticeable for any of the experiments (Figure 21). The LA_Ws results are worse for January than for July, showing even a negative skill during nighttime. However, the improvement of the AN over LA_Ws forecast is more pronounced for January than for July. The AN results are quite similar, showing an expected decrease in skill for long lead times. The exception is AN_11 forecast, which is somewhat less successful in these experiments, especially for short lead times.

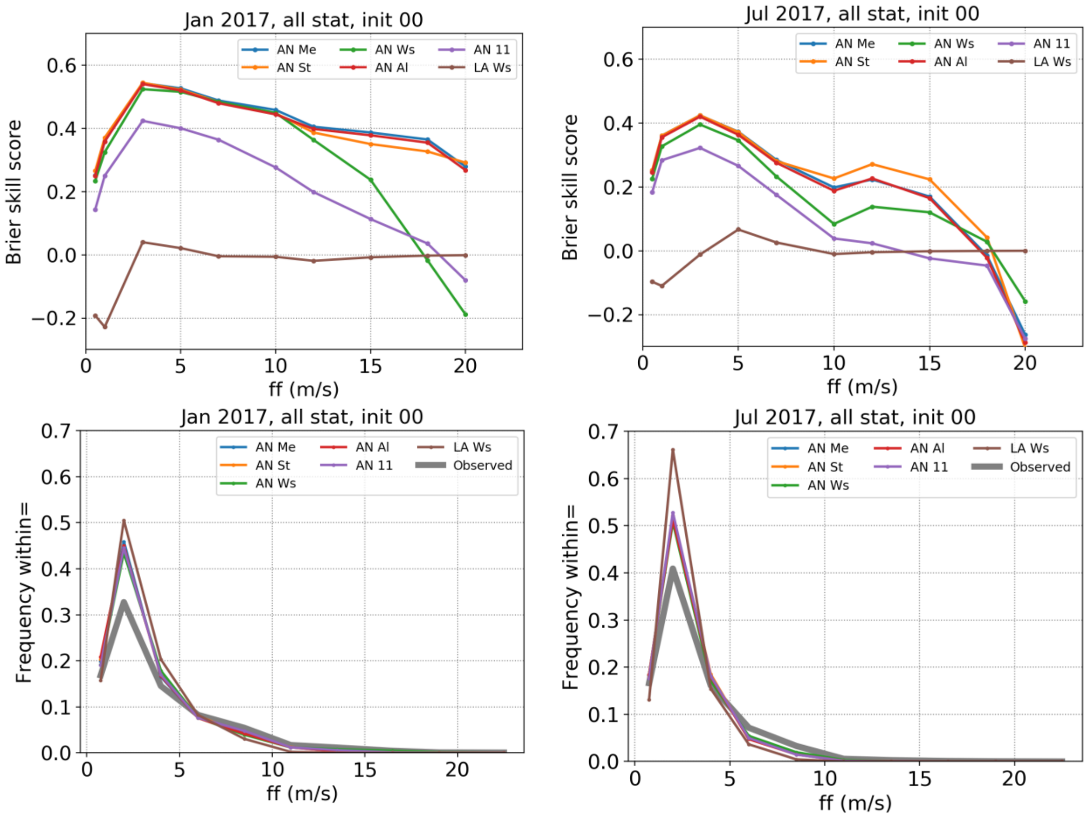


Figure 22. Brier skill score (up) and relative frequency (down) depending on a wind speed threshold for AN probabilistic forecasts compared to LAEF LA_Ws during January (left) and July (right) 2017 at all station tested in this study.

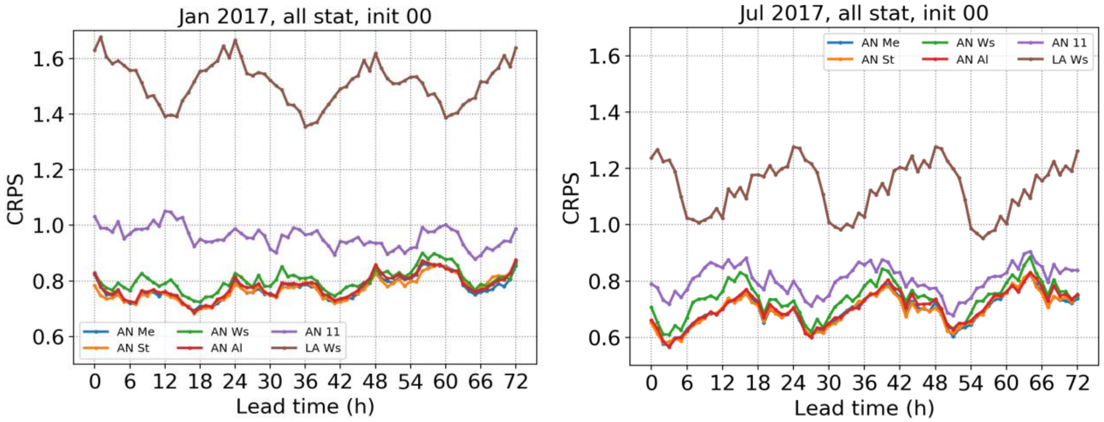


Figure 23. Continuous rank probability score depending on a lead time for AN probabilistic forecast compared to LAEF (LA_Ws) during January (left) and July (right) 2017 at all stations tested in this study.

The majority of measured wind speed values during the selected months are in the order of 3 m/s (30-40 % of measurements), while the wind speeds higher than 5 m/s or 10 m/s are rare (Figure 22). However, it is still very important to forecast the latter wind speeds because of their higher impact on people and property, wind energy potential and other fields. For this reason it is important that a probabilistic forecast is consistently good for several different thresholds. LA_Ws forecast is skillful (measured with Brier skill score – BSS) for wind speeds exceeding 5 m/s, but shows much less skill (if any) for higher thresholds (7, 10, 12, 15, 18 and 20 m/s) in these two selected months. The AN probabilistic forecasts improve the result up to 15 m/s or higher for all the AN forecasts up to 15 m/s. Additionally, AN_Me, AN_St and AN_Al in January improve the LA_Ws result for all the thresholds tested (up to 20 m/s). The AN_Me, AN_St and AN_Al forecasts exhibit very similar result, while AN_Ws and AN_11 show a reduced skill. These results reveal a great potential for post-processing usage of analog-approach, even though one needs to be careful with the interpretation, since the number of occurrences of high wind speed (i.e. around 20 m/s) is very small.

The CRPS, as mentioned above, is a great overall measure that takes all the available thresholds into account while assessing the forecast skill. It The CRPS confirms that the LA_Ws forecast exhibits a higher skill during the afternoon than during nighttime and higher during July than during January (Figure 23). The AN forecast do not show a clear diurnal CRPS pattern, but it is more skillful during nighttime than during daytime. The improvement over LA_Ws forecast is greater in January since the LA_Ws is worse than in July. However, the AN results is overall better in July, when the LA_Ws, which also served as input, is better. These results imply that the best results are achieved when the input model is also working better. Also, AN_Me, AN_St and AN_Al show a bit better skill than AN_11 and AN_Ws. These results imply that there is a need to use more than one meteorological variable as predictor. This is due to better ability of the analog method to distinguish different seasonal and synoptic situations. Using a 1 by 1 member analog search would not increase the skills of the raw probabilistic input as one would inherit undesirable properties of the input model such as under-dispersion and lower resolution issues. Finally, it is shown that using basic information of an input ensemble, such as ensemble mean and standard deviation, already improves the forecast skills. Furthermore, it is

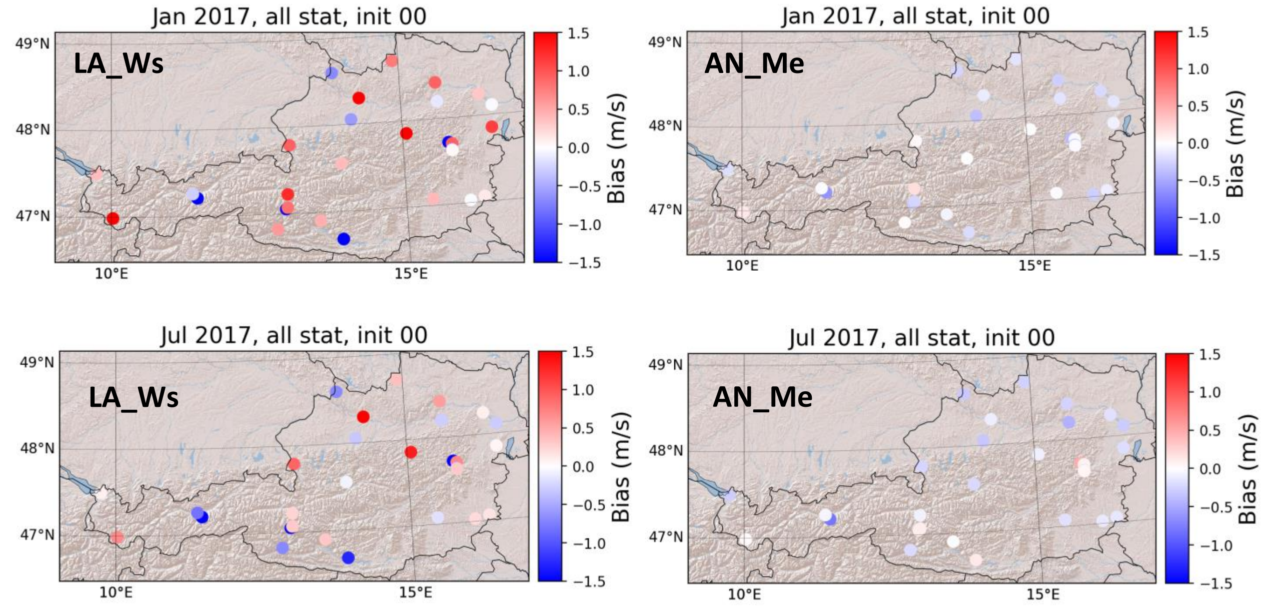


Figure 24. The spatial distribution of the monthly mean bias for the LAEF LA_Ws forecast (left) and AN_Me forecast (right), during January (left) or July (right), 2017.

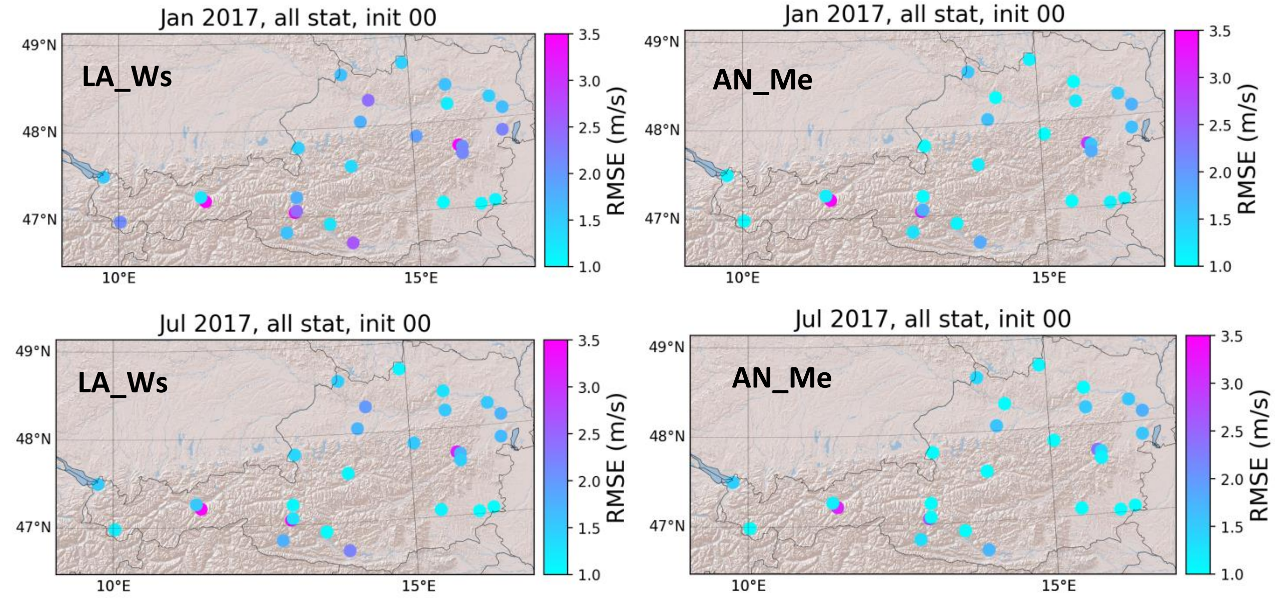


Figure 25. The spatial distribution of the monthly mean RMSE for the LAEF LA_Ws forecast (left) and AN_Me forecast (right), during January (left) or July (right), 2017.

computationally less demanding and produces almost the same result as using the full input spectrum of a raw probabilistic model, i.e. all LAEF members as predictors. Previous results have already shown that the wind speed increases towards north-eastern part (Pannonian plate) for both January and July. The values in January are slightly higher than in July (i.e. Figure 3).

For the LA_Ws forecast, the bias is slightly positive on average at majority of the stations, especially in January (Figure 24). The AN_Me mean bias is smaller in absolute values than for the LA_Ws forecast, showing an underestimation at the majority of stations for both months. The results for other AN forecasts also exhibited very similar findings. The results (CC, mean monthly bias and RMSE) for AN_St and AN_Al are almost undistinguishable from AN_Me resulst, while the AN_Ws and AN_11 are the same or slightly worse. Since they carry no new information, they are not shown from this moment on.

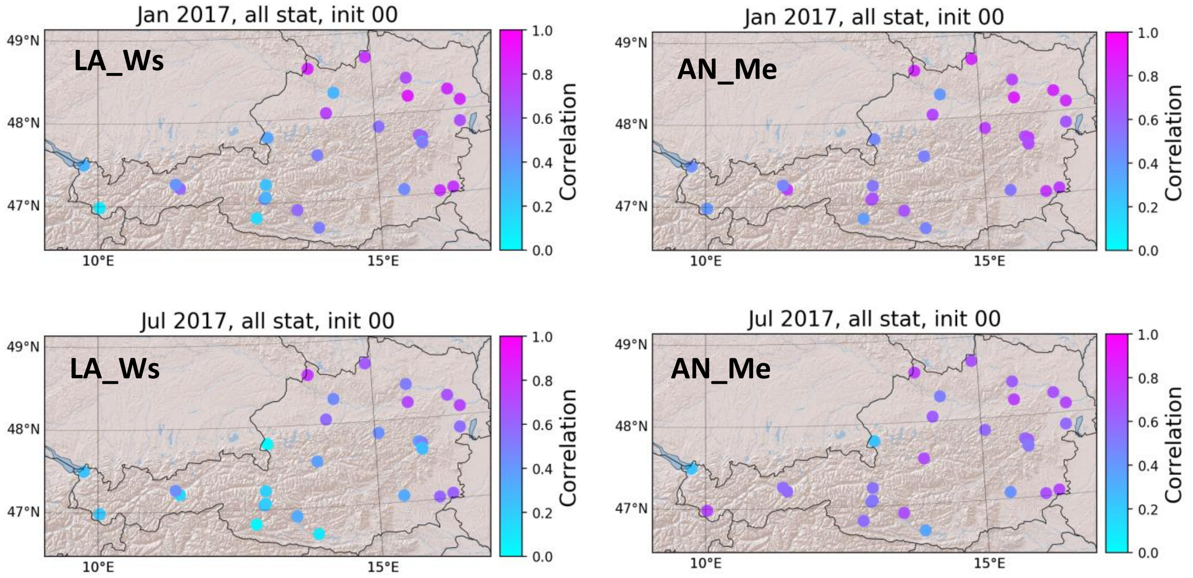

Jan 2017, all stat, init 00
LA_Ws
Jan 2017, all stat, init 00
AN_Me
Jul 2017, all stat, init 00
LA_Ws
Jul 2017, all stat, init 00
AN_Me
49°N
48°N
47°N
10°E
15°E
Correlation
0.0
0.2
0.4
0.6
0.8
1.0

Figure 26. The spatial distribution of the monthly correlation coefficient for the LAEF LA_Ws forecast (left) and AN_Me forecast (right), during January (left) or July (right), 2017.

The RMSE value is slightly higher during January than in July for all the forecasts (Figure 25). The value for the monthly mean RMSE is reduced for the AN forecasts in both January and July cases, if compared with LA_Ws forecasts. There is no obvious spatial distribution of error for the AN forecasts. However, there are large differences for nearby stations situated in a highly complex terrain. Better looking the results for these locations, taking longer period in account might lead to some interesting results.

The CC seems to reduce its value from northeast area towards west and south-west of Austria (Figure 26). Also, the values are higher for the January than the July. This is regardless of the exact forecast and time of a year. Therefore, it could probably be concluded that the wind speed is less predictable towards west and during winters, similar to previous tests. All forecasts lower values in the Alps, as expected. The CC values as low as shown can suggest very unpredictable month, but also a potential error made in forecasting, loading the data or analysis. However, there is an evident improvement achieved with post-processing for January and especially July for all analog-based variations.

Summary-

During this stay results for modified analog-based post-processing method (**mAN**) are compared against previously developed analog-based post-processing method (**AN**) and against AROME deterministic model forecasts (**AR**). The mAN, unlike AN, allows the most similar historical forecast to be found one lead time step sooner or later (time window for analog search is now allowed to shift maintaining the same width). The mAN shows some potential when there is a need to expand the length of the training data (i.e. short training, rare event). The improvement of mAN over AN forecast in this experiment is implicative, but not statistically significant. This is probably because the two-year training is long enough to find the majority of the best historical matches. The differences could be more pronounced if the experiment is extended to include: results achieved with shorter training. The other possibility is to use the same or even longer training, but analyze longer than two months mAN time series, focusing only on rare event.

Several different configurations of LAEF ensemble forecast are tested as input to analog-based post-processing during the second part of the stay. It is shown that using only one predictor variable as input (wind speed LAEF ensemble) already improves the forecast skills and lowers the systematic error of the ensemble mean. Even better results are achieved when using more than one predictor variable. In addition, it is shown that there is no need to use the full input spectrum of a raw probabilistic model, i.e. all LAEF members as predictors. Using basic information of an input ensemble, such as ensemble mean and standard deviation, improves the forecast skills the most among analog-based experiments. Furthermore, it is computationally less demanding while produces very similar result as using the full input spectrum of a raw probabilistic model as predictors.

References:

Delle Monache, L., T. Eckel, D. Rife, and B. Nagarajan, 2013: Probabilistic weather prediction with an analog ensemble. Mon. Wea. Rev., 141, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Horvath, K., D. Koracin, R. Vellore, J. Jiang, and R. Belu, 2012: Sub-kilometer dynamical downscaling of near-surface winds in complex terrain using WRF and MM5 mesoscale models. J. Geophys. Res., 117, D11111, https://doi.org/10.1029/2012JD017432.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Wea. Rev., 116, 2417–2424.

Plenković, I.O., L. Delle Monache, K. Horvath, and M. Hrastinski, 2018: Deterministic Wind Speed Predictions with Analog-Based Methods over Complex Topography. J. Appl. Meteor. Climatol., 57, 2047–2070, https://doi.org/10.1175/JAMC-D-17-0151.1

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. Mon. Wea. Rev., 117, 2230–2247.