Stay Report

To: GeoSphere, Vienna, Austria
Period: November 6th – December 1st, 2023
Topic: Work on machine learning and analog-based post-processing methods
Supervisor: Irene Schicker, PhD

Wind speed post-processing including analog based-method

1. Introduction

I stayed at the GeoSphere for four weeks during which I was working on the machine learning (ML) and analog-based post-processing methods applied to an NWP model output for point-based wind speed forecasts. Machine learning methods have been gaining a lot of popularity for quite some time, both because of their high-performing capabilities, and relatively simple implementation. The goal of this research is to compare the results of two of those methods (neural networks and XGBoost), with the analog method and raw NWP.

Analogies between similar past forecasts are a potentially useful tool when the training dataset is long enough, thus enabling an adequate identification of true analogs. The ML methods, depending on the type, could also be quite sensitive to the training dataset length. The analogs are already relatively extensively researched, and implemented at DHMZ (Croatian Meteorological and Hydrological Service), so they are used here as a baseline method. The goal is to examine whether the ML methods could outperform the analogs in point-based wind speed post-processing. Previously, the point-based analog approach was thoroughly tested as a deterministic approach (Odak Plenkovic et al., 2018) and applied to calibrate the A-LAEF ensemble (Odak Plenkovic et al, 2020). Although some of the ML methods could also be used to generate a probabilistic forecast, the focus of this work is on the performance of deterministic forecasts. Although not a continuation, the inspiration for this work was partly obtained by the work of Bouallègue (2023).

2. Data and algorithms

Dataset

The 2-year dataset of NWP forecasts and synthetic observations is used. The NWP model used is the Croatian 2-km operational ALADIN model. The forecasting range is 72h, and only the 00 run was used in this research. This model variation is based on the non-hydrostatic dynamical core with 87 vertical levels, where the lowest level is 10m. For the analysis, only the closest grid point to the measurement site is used.

For the measurements, only the synthetic dataset was used, both for the training and testing. The synthetic data represent wind speed for 8 wind turbines near Zadar, Croatia. The synthetic dataset was generated by the ERA5 reanalysis for hub height wind speed (80m).

Neural networks

Deep learning is a subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data. It is the part of supervised learning, used for both regression and classification. An example of simple feed-forward neural network is shown in Figure 1.

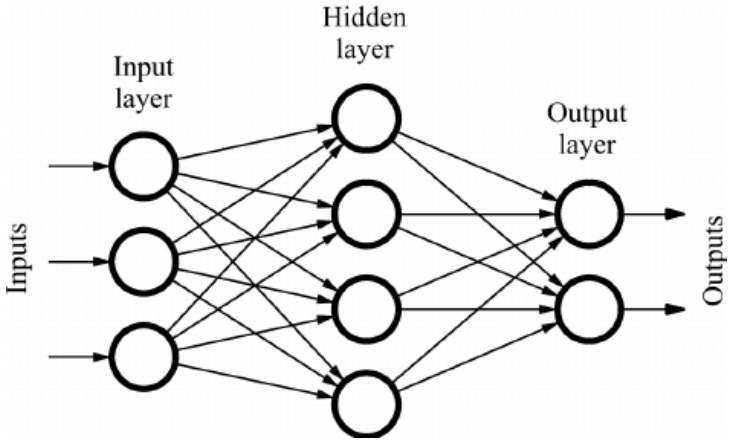

Figure 1. The example of simple feed-forward neural network

Neural networks are sets of algorithms intended to recognize patterns and interpret data. A neural network consists of connected units or nodes called neurons, which loosely model the neurons in a brain. These are connected by edges, which model the synapses in a brain. Each neuron receives signals from connected neurons, then processes them and sends a signal to other connected neurons. The "signal" is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs, called the activation function. The strength of the signal at each connection is determined by a weight, which adjusts during the learning process.

Neurons are aggregated into layers. The signal travels from the first layer (the input layer) to the last layer (the output layer), possibly passing through multiple intermediate layers (hidden layers). A network is typically called a deep neural network if it has at least 1 hidden layer.

XGBoost

XGBoost (eXtreme Gradient Boosting) is an open-source software library that provides a highly optimized gradient boosting framework for various programming languages. The name comes from the engineering goal to push the limit of computation resources for boosted tree algorithms. It is a part of ensemble learning methods, which are building a prediction model by combining the strengths of a collection of simpler base models. The goal is to use multiple simpler models to obtain better predictive performance than could be obtained from any of the constituent model alone.

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms. With XGBoost, trees are built in parallel, instead of sequentially like GBDT. The algorithms train multiple shallow decision trees, with each iteration using the error residuals of the previous model to fit the next model.

Analog method

The analog method has already been extensively used for the NWP post-processing. The details of the method are already described (for example Odak Plenkovic et al., 2018), but the basic idea of the analog method is that if forecasts that are similar to the current prediction can be found in the past, we can infer information about the current forecast error by analyzing the errors of similar historical forecasts. The analog of a forecast for a given location and time is defined as a past prediction that matches selected features of the current forecast. A metric that determines the similarity of two forecasts is proposed by Delle Monache et al. (2011). The verified observations of the best matching analogs are used to produce a new forecast AN, which is defined as the average of the observations. Alternatively, verified observations of best-matching analogs could also be used to form ensemble predictions, or to produce a probabilistic one.

3. Results

As already said, the 2-year dataset is used for training and testing. January of 2022 is used as a test dataset for all experiments. For all experiments other than the analogs, the rest of the dataset (the rest of 2022., and the whole of 2021.) is used for training and validation. Considering that the analogs usually perform solid if at least a one-year dataset is used, the analog method uses the whole 2021. for training.

Considering that the number of experiments is quite high, in Tables 1. and 2. are shown all of the neural network and XGBoost experiment details.

Experiment	Learning rate	Width	Depth	Activation	L2 regularization	Droput rate
1_1	0.0001	100	6	selu	0.01	0.1
1_2	0.0001	5	6	selu	0.00001	0.1
1_3	0.0001	10	2	relu	0.1	0.1
1_4	0.0001	5	3	relu	0.1	0.1
1_5	0.0001	50	4	relu	0.1	0.25
2_1	0.001	100	6	elu	0.0001	0.2
2_2	0.01	100	7	softplus	0.001	0.1
2_3	0.001	40	3	relu	0.01	0.1
3_1	0.0001	20	7	selu	0.01	0.25
3_2	0.0001	10	3	softplus	0	0.25
3_3	0.0001	40	3	relu	0.1	0.25
4_1	0.0001	100	6	elu	0.0001	0.25
4_3	0.0001	40	3	relu	0.01	0.1

Table 1. The hyperparameters of neural network experiments: learning rate, width and depth of a neural network, activation function, regularization parameter, and dropout rate.

Experiment	Subset	Reg. lambda	N estimators	Min. child wght.	Max. depth	Learn. rate	Gamma
5_1	1.0	1.0	300	100	7	0.1	0.1
5_2	1.0	0.01	200	100	7	0.2	5.0
5_3	0.8	0.1	150	100	7	0.2	0.0
5_4	1.0	0.1	600	10	7	0.1	1.0

Table 2. The hyperparameters of XGBoost experiments: a subset of the dataset used, lambda regularization term, number of estimators, minimum sum of instance weight needed in a child, maximum depth of a tree, learning rate, and minimum loss reduction required to make a further partition on a leaf node of the tree.

For all of the experiments, the randomized grid-search hyperparameter optimization was performed. A few most promising configurations were chosen and then tuned a little bit more before the final model training.

NWP predictors that are used in the analog method are two-meter temperature, relative humidity, 10-m wind speed and direction, 10-m wind gust, mean sea level pressure, diffuse and global radiation, cloudiness, precipitation, and 80-m wind speed and direction. Additional predictors for experiments 2 and 4 are the time of day and day of the year, while the additional predictors for experiments 1,3, and 5, including the aforementioned time of day and day of the year, are the latitude and longitude of wind turbine.

For experiments 1, 3, and 5, January 2021 is used as a validation dataset both for randomized hyperparameter grid-search, and final model training. For experiments 2 and 4 randomized grid-search is performed using KFold cross-validation with 5 folds. The validation dataset of the final model training of experiment 2 is, again, January 2021, while the validation dataset of the final model training of experiment 4 is set randomly, as 10 random percent of the whole training dataset.

In Figure 2. are presented the results of experiment 1, for 5 different neural network variations. The best results, regarding the RMSE and dispersion error, are variations 1 and 5. These variations also show the best result for bias, while variation 4 shows the smallest bias of standard deviation. Variations 1 and 5, as could be seen from the Table 1, are the ones with the largest number of neurons, meaning that they are the most complex of the analyzed variations. The other hyperparameters do not show clear pattern regarding the performance measures.

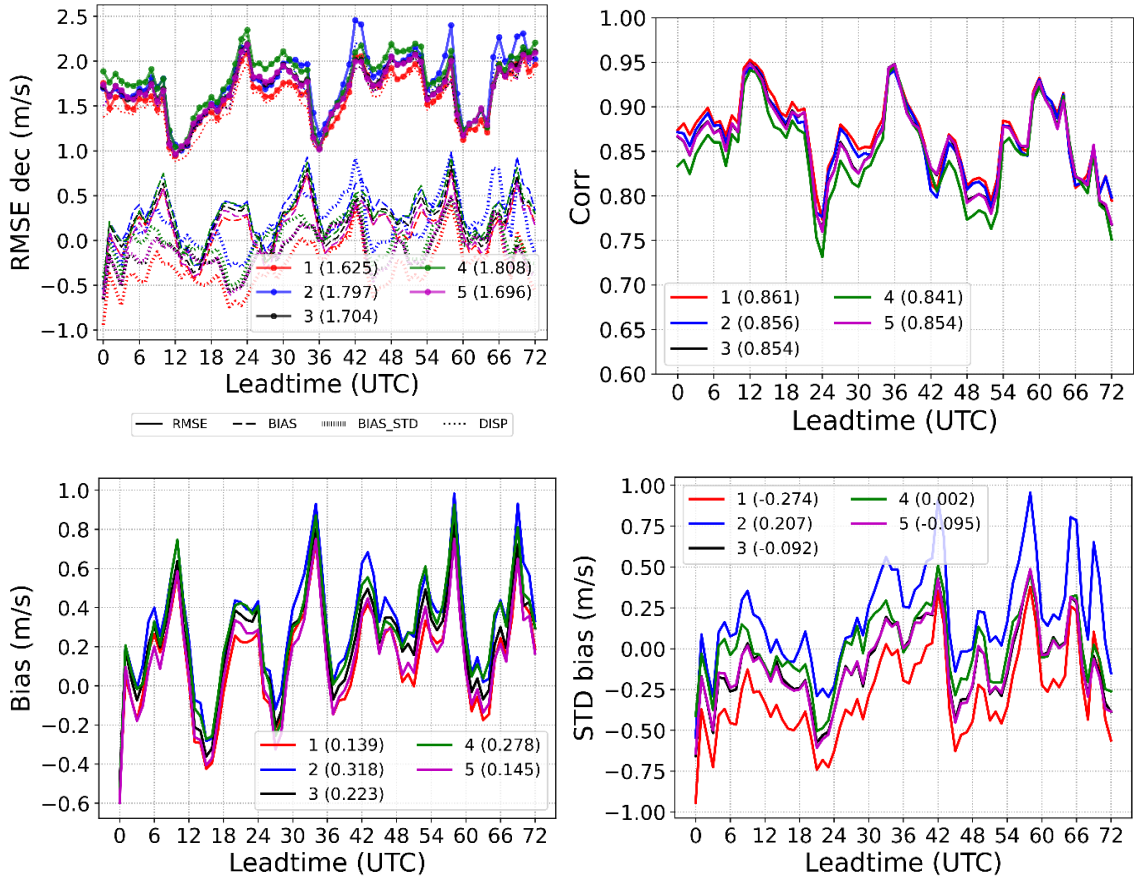

Figure 2. RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for the experiment 1. In the parenthesis next to each experiment variations are shown summary measure values for RMSE, correlation coefficient, bias, and the bias of standard deviation.

On Figure 3. are presented the RMSE decomposition and correlation coefficient for experiment 2, for 3 different neural network variations. For better visibility, the bias and bias of standard deviation for experiment 2 are shown in the appendix in Figure 1. For all of the verification measures, the variation 3 shows the best results, while the variation 2 shows the worst results. In this case, the simplest neural network variation (regarding the number of neurons) shows the best results. Also, this variation uses the largest value of the L2 regularization term.

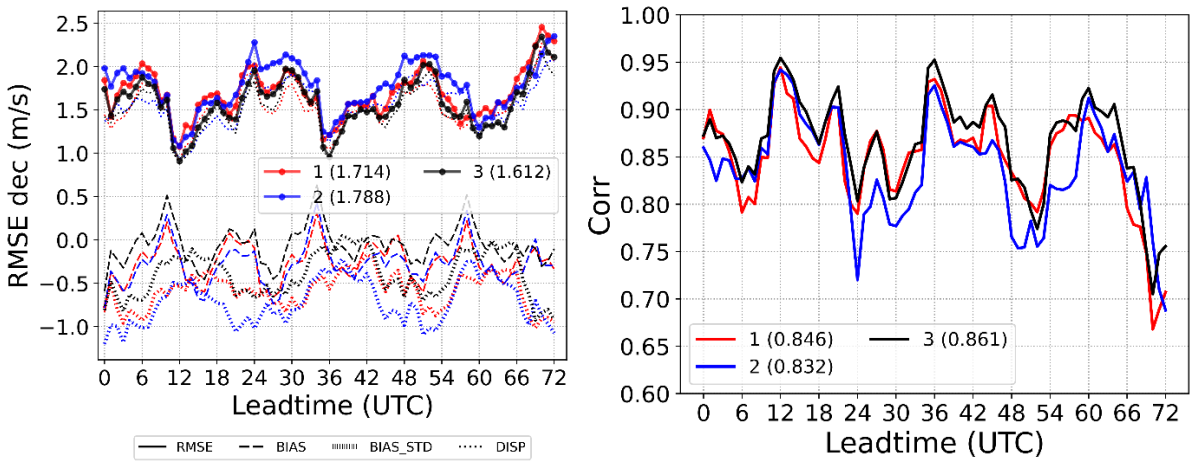

Figure 3. RMSE decomposition and correlation coefficient for the experiment 2. In the parenthesis next to each experiment variations are shown summary measure values for RMSE and correlation coefficient.

The RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for experiments 3 and 4 are shown in the appendix. For both experiments, variation 3 shows the superior results , which is again the variation where the number of neurons in each layer is equal to forty, and there are 3 hidden layers. Although the hyperparameters for these two variations are quite similar, the data split, as already said, is very different.

In Figure 4. are presented results for experiment 5, for 4 different XGBoost variations. The variation 2 shows best results , while the worst results are for variation 4. During the training period, some of the weights were slightly changed in search of better performance. We can see that the best results are achieved when the minimum sum of instance weight needed in a child is equal to 100. The worst results are when that value is set to 10. The maximum depth variable takes value of 7 as optimal in all cases. The largest differences in this experiment are for the gamma and lambda regularization terms.

Results for the best-performing forecasts of each experiment are presented in Figure 5. , as well as the results for raw NWP, and analog-based forecast AN. Considering the already full figure, summary measure values for RMSE, correlation coefficient, bias, and the bias of standard deviation are shown in Table 3. Experiments 1, 2, and 3 show the best results overall, with experiment 2 showing slightly better results than the rest. This could be expected, considering that all three of these experiments use January 2021. as a validation set in the final model training. Experiment 4, which uses random data for validation, shows much worse results in all presented verification measures. This shows that for the best possible results, the validation data distribution should be similar to the test data distribution. The experiment 5, in which the XGBoost is used, shows better results than experiment 4, but not as good as experiments 1, 2, and 3. The XGBoost algorithm shows better results when the dataset is relatively small, while the neural networks work better for large datasets. Considering that the training dataset is almost two years long, the success of the neural networks could be attributed to that reason. All of the experiments outperform the analog method, and especially the raw NWP.

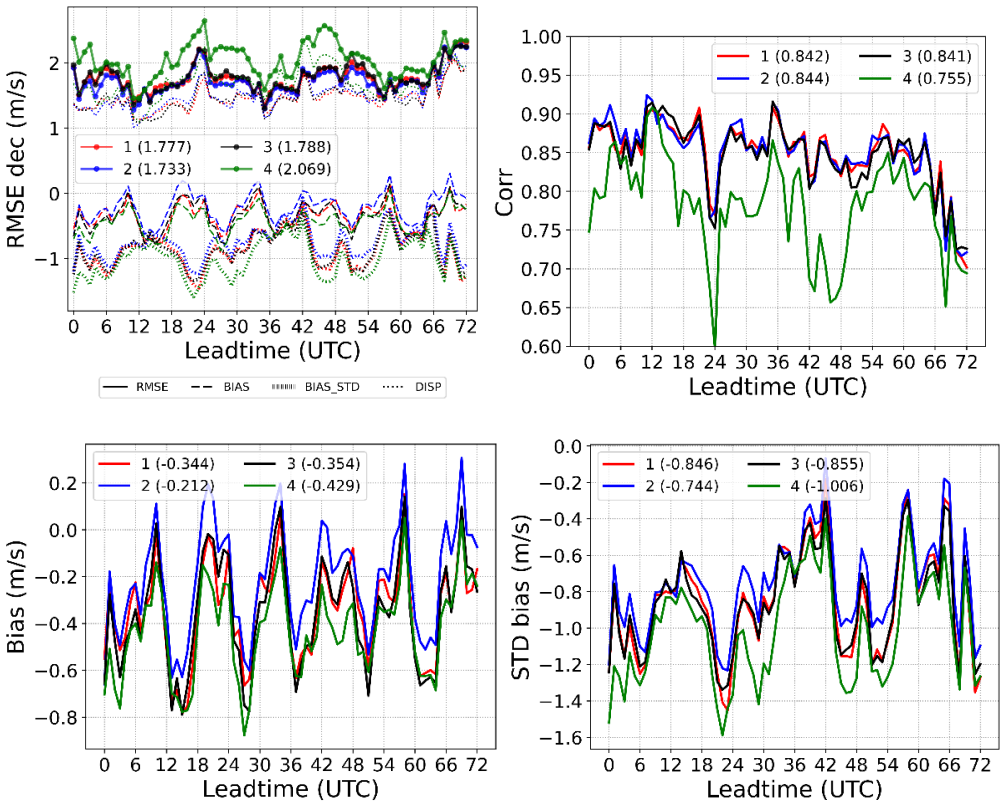

Figure 4. RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for the experiment 5. In the parenthesis next to each experiment variations are shown summary measure values for RMSE, correlation coefficient, bias, and the bias of standard deviation.

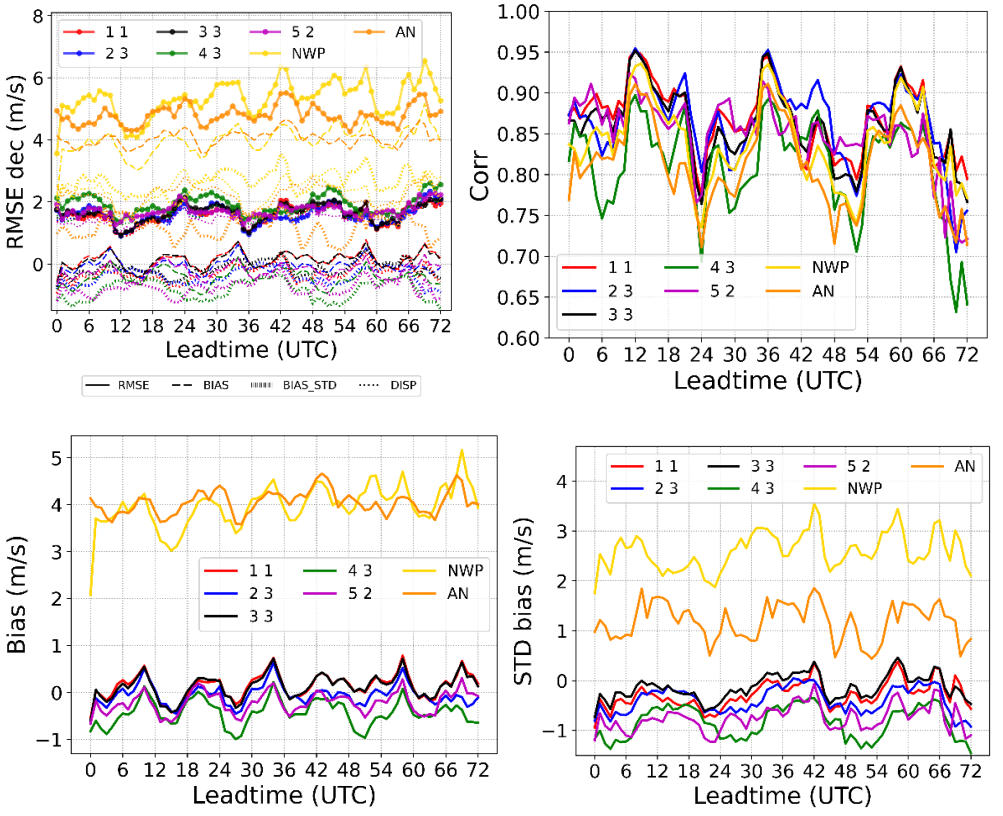

Figure 5. RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for best-performing forecasts of each experiment, raw NWP, and analog-based forecast AN.

Experiment	RMSE	Corr	Bias	Std bias
1_1	1.625	0.861	0.139	-0.274
2_3	1.612	0.861	-0.054	-0.394
3_3	1.680	0.855	0.127	-0.136
4_3	1.950	0.803	-0.430	-0.829
5_2	1.733	0.844	-0.212	-0.744
NWP	5.361	0.834	3.974	2.621
AN	4.807	0.809	4.050	1.188

Table 3. Summary results for RMSE, correlation coefficient, bias, and the bias of standard deviation, for the best-performing forecasts of each experiment. The best forecast is denoted by yellow, and the worst forecast is denoted by the red color.

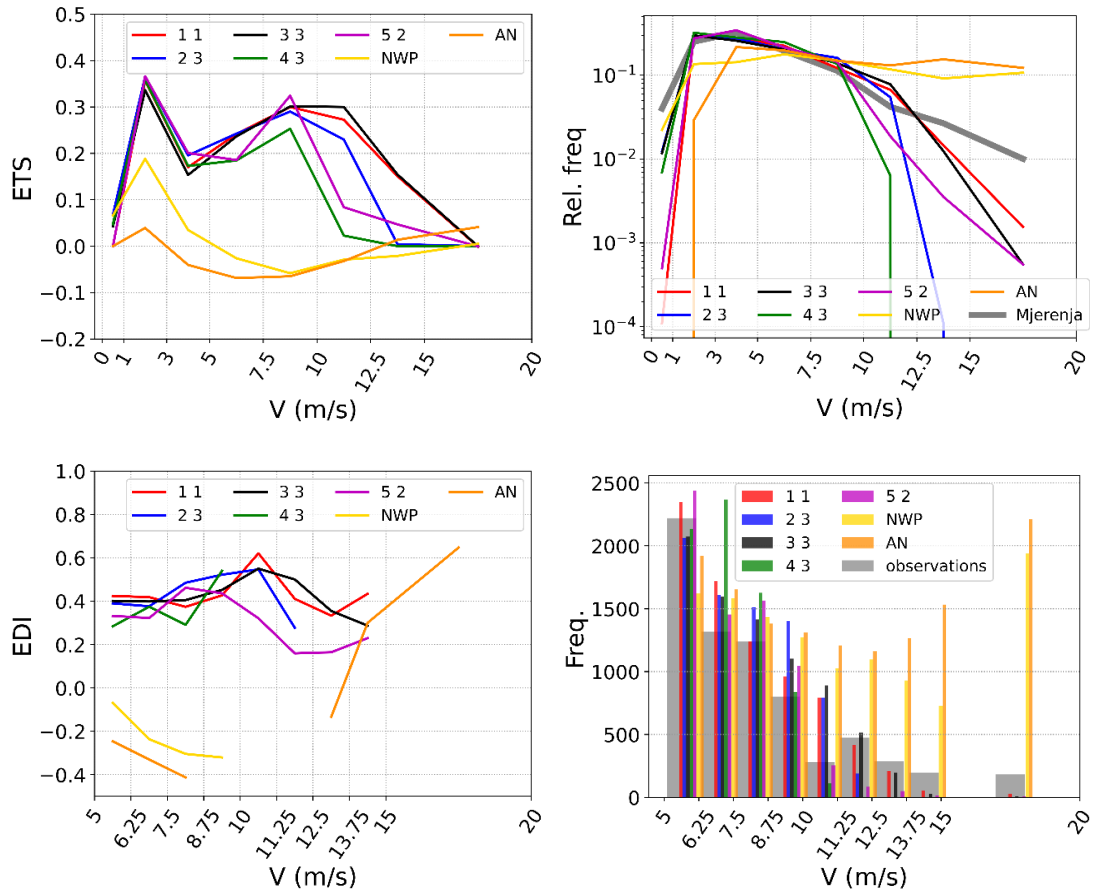

Figure 6. ETS, relative frequency, EDI, and frequency for best-performing forecasts of each experiment, raw NWP, and analog-based forecast AN.

In Figure 6. categorical verification results for the best-performing forecasts of each experiment are presented, as well as the results for raw NWP, and analog-based forecast. The ML experiments outperform analogs and raw NWP for most of the categories regarding the ETS measure, while raw NWP shows superior results only for the most extreme category. Similar conclusions can be seen in the relative frequency graph, where the ML methods are better for more common categories, while the raw NWP shows better results for more extreme categories. For category 15-20 m/s the analogs and NWP show higher than observed frequencies, while the ML methods show lower than observed frequencies. The EDI measure also shows superior results of ML experiments for more common events, while the analogs show the best results for more extreme events. However, analogs and raw NWP

strongly overpredict the frequency of those events. Among the ML techniques, experiments 1 and 3, as well as experiment 2, show superior results in comparison with experiments 4 and 5.

On Figure 7. are presented the time series for the whole test dataset, and for one 4-day case for the beforementioned forecasts. It is easy to see that the AN, and especially NWP, show clearly worse results, with almost constant high bias values. All other experiments show relatively similar (and realistic) results, with high correspondence with observations.

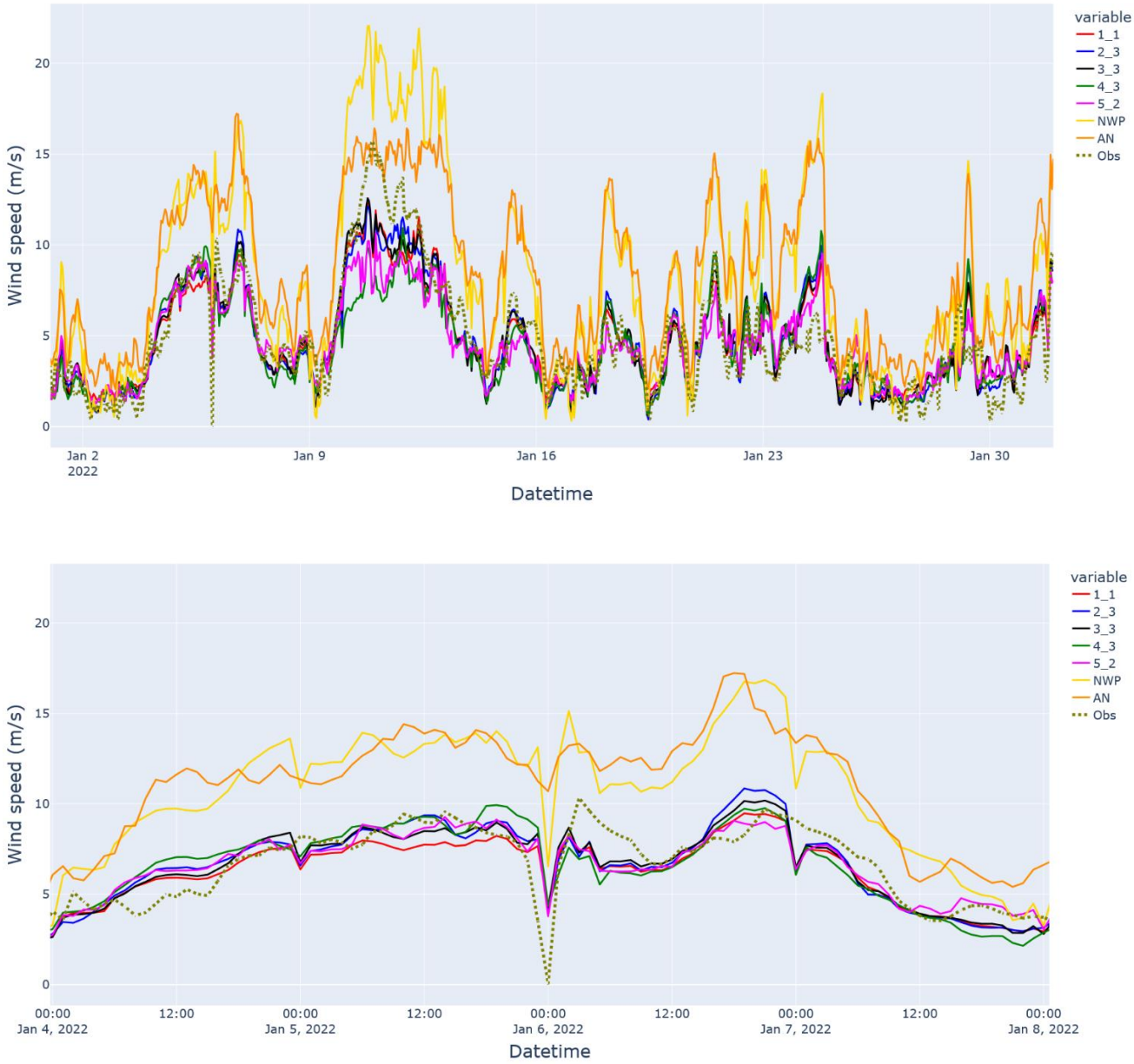

Figure 7. Time series for the whole test dataset (upper) and one 4-day case (lower) for the best-performing forecasts of each experiment, raw NWP, and analog-based forecast AN.

4. Conclusion

The results show that all of the machine learning experiments outperform the analog method, and especially the raw NWP in most cases. The analogs and raw NWP show relatively higher performance for more extreme events. Among the ML methods, experiments 1, 2, and 3 show better results than experiments 4 and 5. Considering that the analysis is performed using the synthetic dataset, the same analysis should also be performed using real measurements.

5. References

Bouallègue Z, Cooper F, Chantry M, Düben P, Bechtold P and Sandu I (2023). Statistical Modeling of 2-m Temperature and 10-m Wind Speed Forecast Errors. Monthly Weather Review 151(4):897-911.

Delle Monache L, Nipen T, Liu Y, Roux G, Stull R (2011) Kalman flter and analog schemes to postprocess numerical weather predictions. Monthly Weather Rev 139(11):3554–3570. https://doi.org/ 10.1175/2011MWR3653.1

Odak Plenković I, Delle Monache L, Horvath K, Hrastinski M (2018) Deterministic Wind Speed Predictions with Analog-Based Methods over Complex Topography. Journal of applied meteorology and climatology 57:2047-2070.

Odak Plenković I, Schicker I, Dabernig M, Horvath K, Keresturi E (2020) Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain. Quarterly Journal of the Royal Meteorological Society 146:1842– 1860

6. Appendix

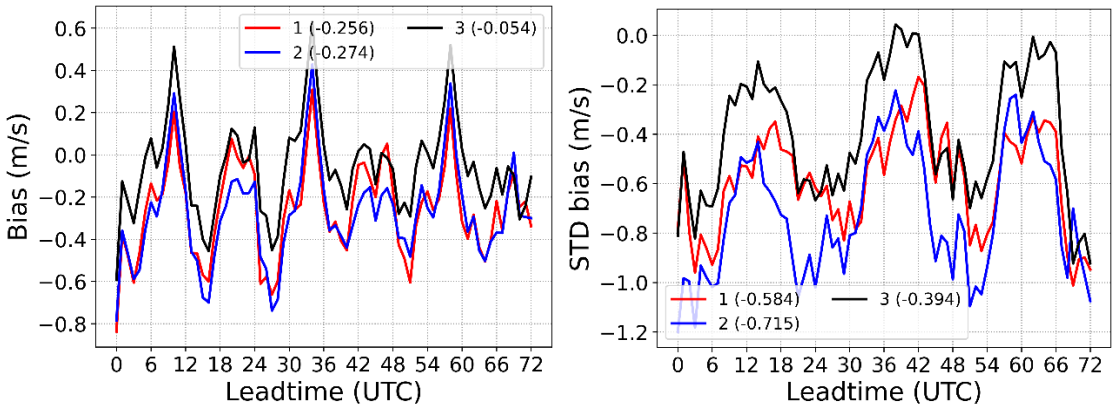

Figure 2. Bias and bias of standard deviation for the experiment 2. In the parenthesis next to each experiment variations are shown summary measure values for bias and the bias of standard deviation.

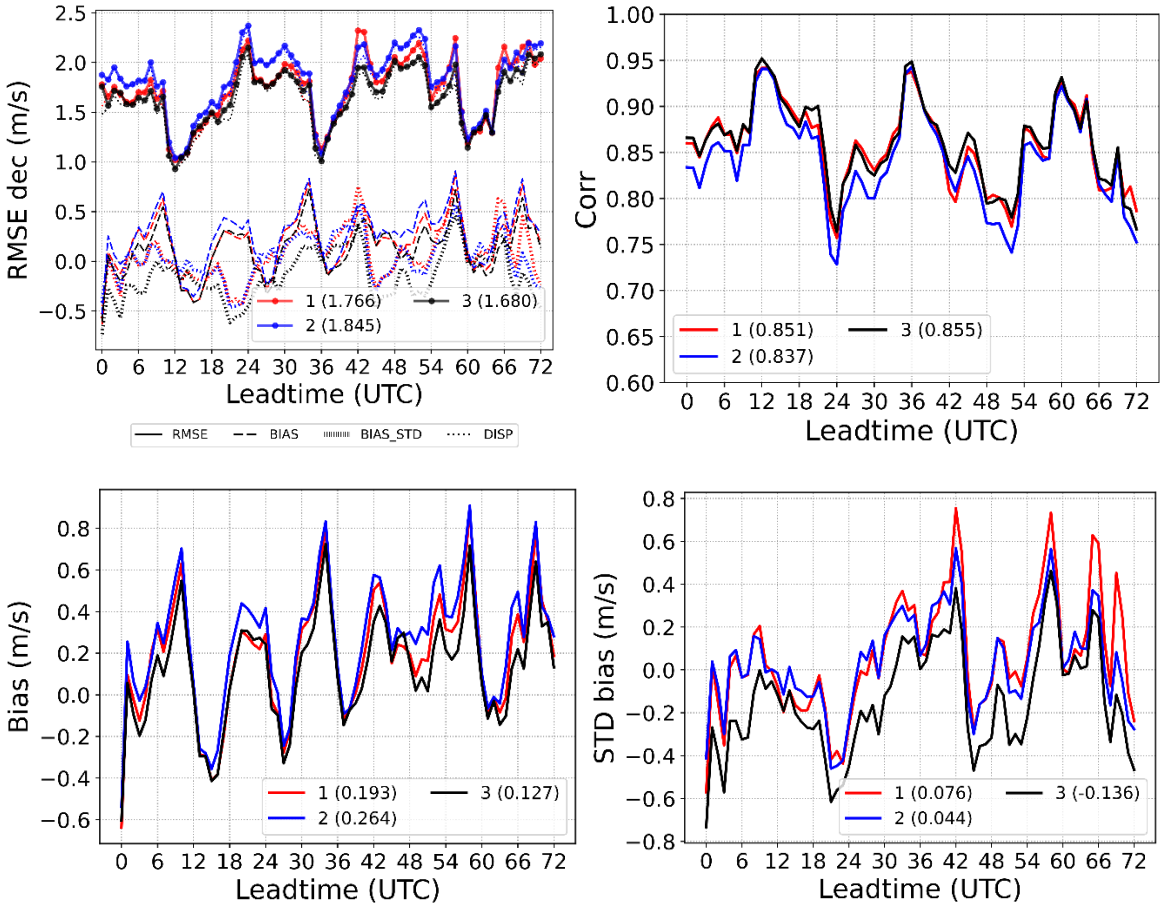

Figure 2. RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for the experiment 3. In the parenthesis next to each experiment variations are shown summary measure values for RMSE, correlation coefficient, bias, and the bias of standard deviation.

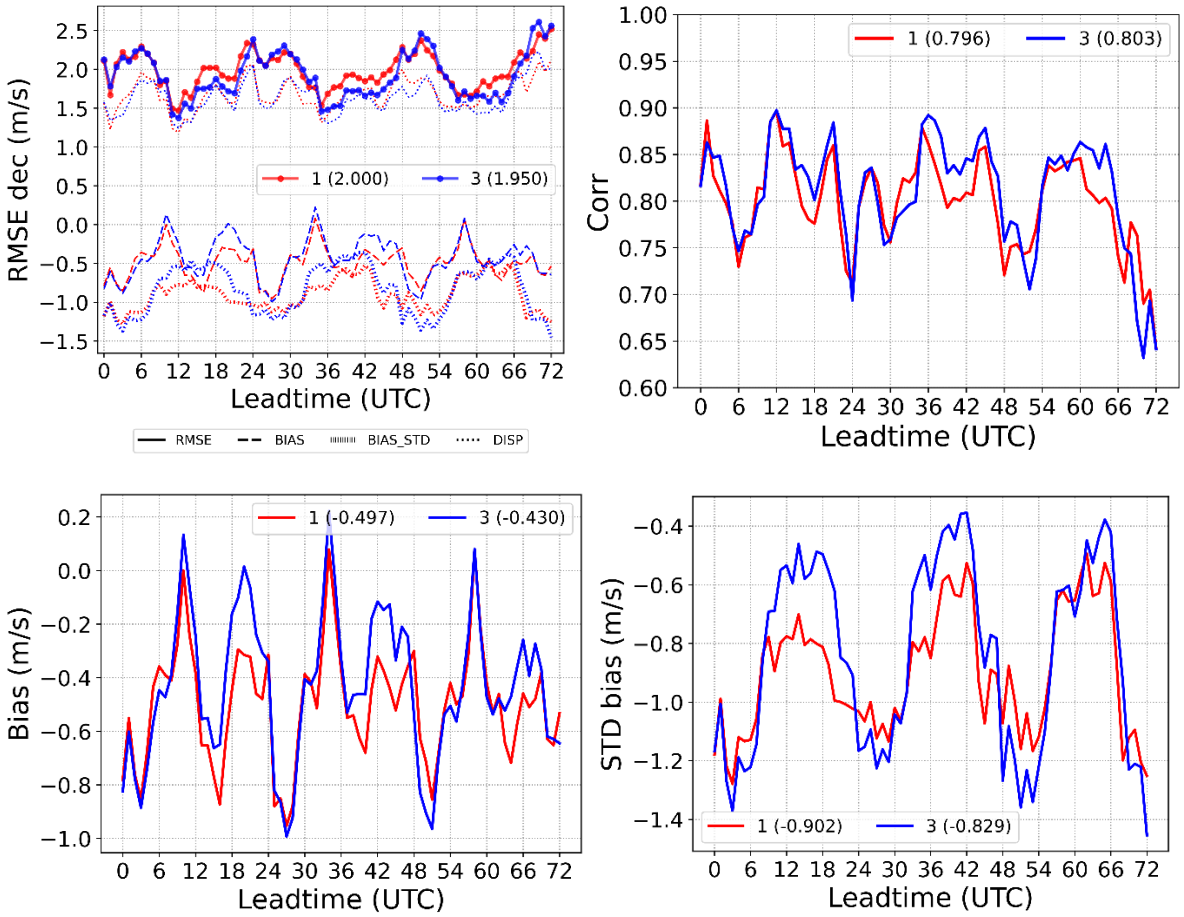

Figure 3. RMSE decomposition, correlation coefficient, bias, and bias of standard deviation for the experiment 4. In the parenthesis next to each experiment variations are shown summary measure values for RMSE, correlation coefficient, bias, and the bias of standard deviation.