Using Analog Ensembles Algorithms for Multistations

Alexandre Chesneau[1], Carlos Balsa[2], Carlos Veiga Rodrigues[3], Isabel Lopes[4, 5]

[1] Universit–e de Toulouse - Institut National Polytechnique de Toulouse
France alexandre.chesneau@etu.enseeiht.fr

[2] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto
Polit–ecnico de Bragan'ca, Campus de Santa Apol–onia, 5300-253 Bragan'ca, Portugal
balsa@ipb.pt

[3] Vestas Wind Systems A/S - Design Center Porto - Portugal
carlos.rodrigues@fe.up.pt

[4] Applied Management Research Unit (UNIAG), Instituto Polit–ecnico de Bragan'ca
Campus de Santa Apol–onia, 5300-253 Bragan'ca, Portugal

[5] Centro ALGORITMI, Escola de Engenharia - Universidade do Minho,
Campus Azur–em, 4800-058 Guimar~aes, Portugal
isalopes@ipb.pt

ABSTRACT: *In this work, many analog ensembles algorithms were used with the performance of multiple stations. We have deployed many techniques to analyse and benchmark inorder to change the prediction. This issue consists in leading the weather predictions for a location where no data is available, using meteorological time series from nearby stations. Many models are verified and explored. The preliminary one is described by Monache and co-workers, to methods using cosine similarity, normalization, and K-means clustering. Best results were obtained with the K-means metric, wielding between 3% and 30% of lower quadratic error when compared against the Monache metric. Increasing the predictors to two stations improved the performance of the hindcast, leading up to 16% of lower error, depending on the correlation between the predictor stations.*

1. Introduction

Weather prediction using Analog Ensembles (AnEn) is not a recent idea. It was described as early as 1969, by Lorenz [12], who however concluded that such a method would not work. Further works managed to prove the usefulness of this approach in a much more limited scope, in various fields ranging from meteorology to flood study, especially thanks to the decisive contributions of Van Den Dool [9,10].

In the eld of meteorology, a major contribution to the use of the Analog Ensemble method was made by Monache et al (2011) [14], having been refined [13] and applied to a variety of operational situations [6,5,7], showing its accuracy and usefulness in the process.

The Analog Ensemble method is actually a post-treatment procedure, used to improve the accuracy of a meteorological model. The idea is very simple: a model makes forecasts. Alongside these forecasts, are also available historical forecasts - a record of forecasts from the model at past dates. Then, to improve forecasting accuracy, the forecast to improve is compared to historical forecasts. The historical forecasts closest to the current one are kept, and the actual meteorological values observed at these dates are used to improve the forecast value. The name of the method comes from there: past forecasts close to the current forecast are called analogs, and these form an ensemble.

The aim of this paper is to compare the performances of various methods to determine these analogs and to establish which of these methods is the most accurate one. The aim is to discover ways to improve the AnEn method described in the literature. To this aim, the various methods were applied to a hindcasting problem, where the time serie of a meteorological variable in a location was reconstructed using data from other weather stations. Results obtained using the data from only one predictor station, then using two stations as predictors, were compared.

Section 2 of this document contains the methodology with the mathematical formulation, definitions for error quantication, and a description of the data used in this study. The results are presented in Section 3 with subsequent analysis. Section 4 contains the conclusion and final remarks.

2. Methodology

In this section, the data used for the tests are presented, alongside the various methods compared and the tools used to assess the performance of each model.

2.1. Analog Ensembles Overview
The Analog Ensemble (AnEn) method is illustrated in Fig. 1, where the objective is to predict a time-dependent data at a location based on multiple data. These datasets are composed by observations, available only for a limited period of time named Training Period, and a historical dataset available at the Training Period and the time that will be predicted. Usually, the historical dataset is a time-series from a Numerical Weather Prediction (NWP) model used to forecast [13,6] or hindcast [16] meteorological data. In this work, real measurements from meteorological stations were used as the historical dataset.

The AnEn procedure is implemented in three steps for each prediction time. Firstly, in step 1 the corresponding value for the prediction is obtained from the Historical Dataset and this dataset is scanned for analogs matching that value. Analogs are past occurrences deemed close enough to the current prediction, classified as such according to an analog metric. Step 2 consists
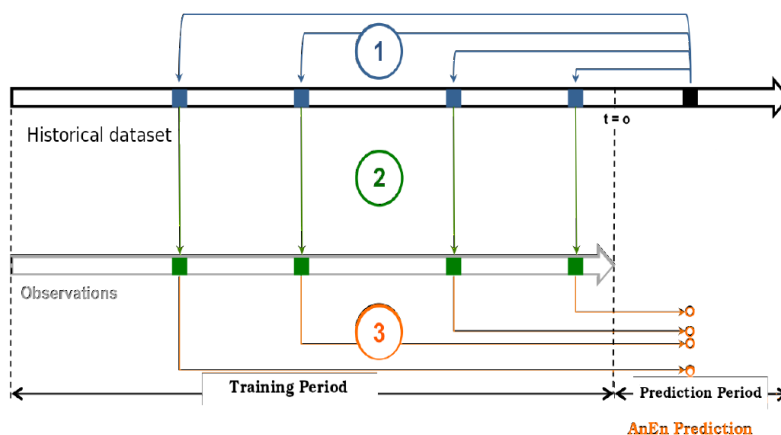

Figure 1. Illustration of the Analog Ensemble method. A prediction of the observations is made for the Prediction Period by finding analogs in the data from the Historical dataset

of matching these analogs with the corresponding real observations at the target station. Step 3 consists of correcting the current prediction with the past values matched to the analogs. The period where historical data from both predictions and observations are available is called the training period. The larger this period is, the better the AnEn method performs [2]. The other period is the prediction period or reconstruction period, in the case of hindcasting. The Analog Ensemble method is very simple, but having an accurate similarity metric is crucial for the success of the forecasting.

2.2 Testing Database

Testing was done using the data from meteorological stations located on the coast of the state of Virginia, USA. These stations were used because their observations are freely available from the United States' National Data Buoy Center [3]. The location of the stations is shown in Figure 2.

The data extended from the years 2012 to 2018. The data from the years 2012 to 2016 was kept as a historical database (training period), and the experiment aimed to reconstruct the data from 2017 to 2018 at one station (reconstruction period). The stations data was time-integrated to samples of six minutes, meaning that they observe the value of meteorological variables ten times per hour. These stations observe 6 different meteorological variables: pressure (PRES), air temperature (ATMP), water temperature (WTMP), wind speed (WSPD), gust speed (GST) and wind direction (WDIR). In many cases time series are not complete, data are missing in many more or less extended periods.
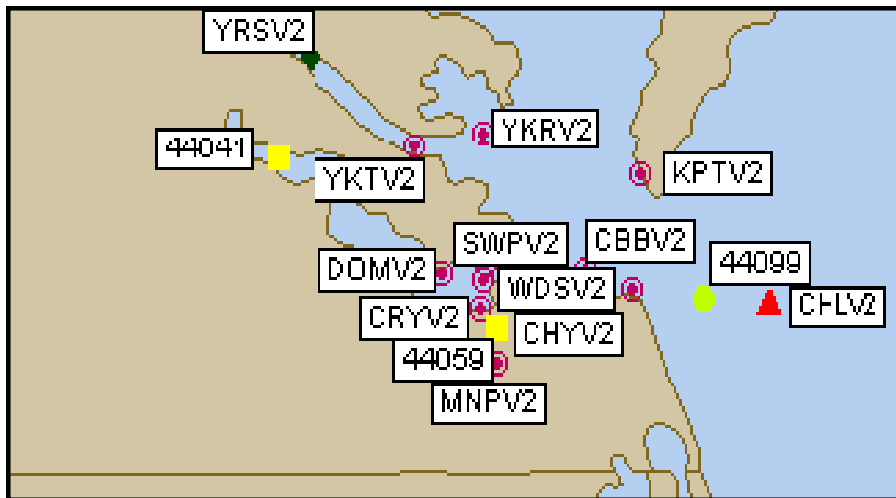

Figure 2. A map of the metheorological stations (adapted from [3])

The idea was therefore to hindcast: at one station where only the meteorological data from 2012 to 2016 (training period) was known, the program had to reconstruct the data from 2017 to 2018 (prediction period), using the other stations, where the full range from 2012 to 2018 was known, as predictors. Based on the AnEn procedure illustrated in Figure 1, the data from 2012-2016 at the target station is the "Observed dataset" whilst the data at all other stations between 2012-2018 is the "Historical Dataset" (comprised of multiple time series), where the "Training" and "Prediction" periods are delimited by 2012-2016 and 2017-2018, respectively.

The advantage of such a setup is that it becomes easy to evaluate the model accuracy because we can compare the estimates obtained with the AnEn method with the real values. There is one important thing to note, however: because the data collected by stations every six minutes over 7 years is huge, it was time consuming to sequentially process all the records in the historical dataset. Instead, the problem was simplifed to predict the weather between 10 am and noon, using analogs of the weather between 10 am and noon. This greatly reduces computing time, while still giving data from different years and dierent seasons, thus very different weather patterns.

2.3 Determination of the Analogs

The determination of the analogs is an important step of the AnEn method. In the present work various methods, or metrics,

have been used to compute the similarity between forecasts.

The first metric used is the metric originally established for the AnEn method by Monache [13], which will be referred to as Monache from now on. It is based on the Euclidean distance. More precisely, it computes the difference between the values of atmospheric variables in the two forecasts, over a window of time.
The formula used is the following:

$$m_{tt'} = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} \sqrt{\sum_{j=-k}^{k} (F_{i,t+j} - A_{i,t'+j})^2} \qquad (1)$$

With the terms being the following:

- F_t is the current forecast, which needs to be improved.

- $A_{t'}$ is a past forecast compared to the current forecast.

- N_v is the number of meteorological variables taken into account when comparing forecasts.

- w_i is the weight given to variable i.

- σf_i is the standard deviation of variable i. This term is used to reduce variables, thus making dierent variables possible to compare. Without this term, computation is impossible and makes no physical sense.

- k is the length of the time window over which the forecasts are compared. Indeed, to compare F_t and $A_{t'}$, we do not only look at the variables at the times t ad t', which are compared, but we look at their evolution over a period of time. The aim is to make

sure the weather pattern in both forecasts is similar.

- $F_{i, t+j}$ is the value of variable i at time $t + j$ in the current forecast.

- $A_{i, t'+j}$ is the value of variable i at time $t' + j$ in the past forecast.

Something of note here is the importance of parameter k: there is no obvious value for this parameter, so a separate study would be needed to determine its optimal value. Considering, the sets F_t and $A_{t'}$ as two vectors of a $2k + 1$ multidimensional space, the Monache metric, presented in Equation (1), can be rewritten as

$$m_1 = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} ||F_t - A_{t'}||, \qquad (2)$$

where $||.||$ represents the Euclidean norm.

An alternative metric consists to normalize F_t and $A_{t'}$ in Equation (2), resulting in the normalized Monache metric presented in Equation (3).

$$m_2 = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} \left\| \frac{F_t}{||F_t||} - \frac{A_{t'}}{||A_{t'}||} \right\| \qquad (3)$$

Normalized vectors all have a norm of 1. The idea behind this reasoning is to look at the global weather pattern present in both forecasts. It can be seen that the basic Monache metric looks not only for a similar weather pattern but also for similar numerical values to the various variables used in the forecast. As a consequence, it will not keep as an analog a forecast behaving exactly like the forecast to improve, but at higher or lower values. Normalization aims at solving this perceived problem. This method is called Normalised Monache, shortened to Normalised.

Keeping in line with the previous idea, the cosine of the angle between the two vectors F_t and $A_{t'}$ can be used such that

$$\cos(\theta) = \frac{A_{t'}^T F_t}{\|A_{t'}\|\|F_t\|},$$

where θ denotes the angle between the vectors F_t and $A_{t'}$. The cosine can then be used to estimate the analog by means of the correlations between the two vectors, as demonstrated by K. Adachi [4]. This is the idea beneath the use of the metric presented in Equation (4).

$$m_3 = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{fi}} \frac{A_{t'}^T F_t}{\|A_{t'}\|\|F_t\|}. \tag{4}$$

It is known that this value behaves like the correlation coefficient, taking values between -1 and 1 with 1 indicating maximum similarity. The idea is, therefore, to replace the Monache metric with the cosine of the angle between the forecasts, keeping as analogs the past forecasts with cosine closest to 1. This is the cosine method.

Lastly, clustering can be applied to this problem. Clustering is the partitioning of data into clusters of similar data, as illustrated by Figure 3. In this case each multidimensional vector $A_{t'}$ is affected to a similar cluster. Clustering is used to create the analog ensembles. By clustering the database of past forecasts, we obtain analog ensembles that can then be used immediately. The only task left is to affect the current forecasting to the good analog ensemble - in other words, to the closest cluster. This
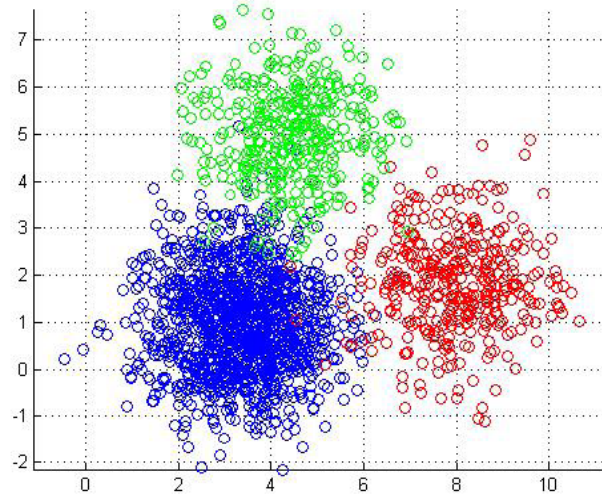

Figure 3. An illustration of the Kmeans clusterisation (adapted from [1])

method was inspired by Gutierrez and co-workers, who used clustering in a forecasting problem [11]. From now on, this method is mentioned as Kmeans.

2.4 Prediction Methods
Making predictions using the Analog Ensemble method is very straightforward in this case. In this case, if one takes a look at Figure 1, there are no NWP predictions and past predictions. The NWP is replaced by data from other stations, data in the past (training period) and in the forecasting interval (prediction period). The principle, however, remains the same. Since there are not forecast to correct in this case, though, but only a prediction to make, this is done easily when used the Monache and K-means metrics like this

$$F_t = \frac{1}{N_a} \sum_{i=i}^{N_a} F_{t_i}. \tag{5}$$

It is simply the mean of the past target variable value at times matching the analogs. In the case of Cosine and Normalised Monache, however, a simple mean is not enough. Because what is looked at is trends and not exactly equal numerical values, there might be a difference between the variable's value at time t and the desired value. Therefore, the equation becomes

$$F_t = \frac{1}{N_a} \sum_{i=i}^{N_a} F_{t_i} + \delta_{tt_i}.$$ (6)

where $\delta_{tti} = A_t - A_{ti}$ to account for the scale difference between the analog and the forecast.

2.5 Using two Stations as Predictors

To improve the accuracy of the hindcast, it is tempting to use data coming from various weather stations as predictors, instead of data coming from just one station. This raises the problem of how to treat these additional data. This problem was solved in two different ways, which were both used in this paper to determine which method is more adequate to handle data coming from various stations.

The first method is called the *dependent stations* variant method. This variant considers the stations to be nothing more than additional predictor variables, and as such compute analogs across all stations at once every time. That it is to say, the observation at time t is deemed to be an analog of the weather at time t' if, and only if, the weather at all stations at the time t' is close to the weather at all stations at time t.

A second idea can be to look for analogs at each station. This is called the *independent stations* variant method. In other words, the metric is calculated at each station independently from one another. The prediction is then made using the mean of the analogs from all the stations. In other words, compared to the first approach, each station forms a disjoint set of data in which analogs are searched separately. Then weights are assigned to each station to form the final set of analogs, so that for example 90% of analogs may come from the study of the data at the first station and the 10% remaining come from the study of the data at the second station.

2.6. Error Assessment

As shown by Chai and Draxler [8], assessing the model accuracy is best done using various metrics. Three metrics are especially useful when trying to assess the performance of a forecasting model. First one is the bias:

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} x_i - y_i,$$ (7)

with n being the number of forecasts, x_i being the forecast and y_i being the truth values. As its name suggests, the bias simply measures the bias of the model: it simply shows the average error compared to the truth. However, it does not really show the behavior of the error. It is useful to determine if the model makes predictions that are lower or higher than the truth, but in itself, it is not enough to know how well the model performs. It only shows the systematic error of the model.

Thus, the Root Mean-Squared Error (RMSE) is also used, computed as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2}.$$ (8)

This error is useful because the squared terms give a higher weight to high error. Thus, the RMSE will be higher if the model makes predictions that are far from the truth, even if these erroneous predictions are few. The RMSE will comparably be lower for a model consistently close to the truth, even if the forecast is still committing an error compared to the truth. It shows the random errors of the model - errors, which happen randomly, not in a systematic way.

The third metric, whose usage alongside the RMSE is recommended by Chai and Draxler [8], is the Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$$ (9)

Compared to the bias, this metric computes the average distance, in absolute value, to the truth. The bias simply computes the average error, but positive errors and negative errors can cover each other. The MAE then gives a somewhat more truthful assessment of the average distance to the truth. A low bias and high MAE means that the model is not really accurate, but that its predictions are sometimes higher than the truth, and sometimes lower. Thus, including the MAE is necessary to really understand how the error is distributed in the forecast since it also shows a systematic error but this time in terms of absolute distance.

One last tool used to show an error from a forecasting model is the Taylor diagram, described by Taylor [15]. This diagram shows the proximity between two variables, variables which here are the truth and the prediction. Considering two variables x_i and y_i, each having N components and with means $\bar{x}$ and $\bar{y}$, with the correlation coefficient between them being R, their standard deviations being σ_x and σ_y, then it can be shown that which is the basis of the Taylor diagram representation.

$$\text{RMSE}^2 - \text{Bias}^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y R, \tag{10}$$

2.7 Parameter Selection
Various tests were run to determine which was the most suitable value of k, the time window, and N_a, the number of analogs kept. It has been found that in this testing environment, the best value for k is $k = 20$, which corresponds to a time window of four hours in length (2 hours before the forecast, and two hours after) since the data used made observations 10 times per hour. For N_a, $N_a = 25$ gave satisfying results. Of course, these values may change according to the problem and are not to set in stone but to be adapted to each case.

3. Results

The results of the test can be divided into two parts. In the first time, the importance of the choice of stations was assessed, to see how important choosing the right stations to hindcast the values at another station is. Then, since it is also possible to assign weights to stations, the importance of choosing the right weight was studied.

3.1 Studying the Stations
The first aim of this study is to evaluate how important the choice of the stations used for hindcasting is. To evaluate his, the gust speed (GST) data for the station *ykt* between the years 2017 and 2018 was reconstructed using three different pairs of stations, and the value of GST at these pairs of stations.

Table 1 contains results for an AnEn hindcast whose predictor was based solely on the *mnp* station. The results on table 2 extended the AnEn method to include a second station as a predictor. Pairs were made from stations *mnp*, *dom*, *ykr* and *wds* to assess how consistent the results are across different pairs.

The results in Table 1 show less error for the Kmeans method, while retaining similar bias to the Cosine and Normalised method. This is in line with the results from Table 2, where the Kmeans method consistently shows better performance. A simple application of the Monache metric yielded higher bis (which is also consistent with the results in Table 2). Normalizing the Monache metric shows overall improvements in the bias and RMSE, though more evident in the results from table 2. It is only for the *wds,mnp* pair of predictors that the Monache method shows a superior performance, though the Kmeans method still has lower RMSE.

Comparing results from Table 1 and Table 2. Using *mnp* alone as predictor is worse than using *mnp* and either *dom* or *wds* for the Kmeans and Normalised Monache methods. For the Cosine and Monache methods, it is clearly better to use both *wds* and *mnp* rather than *mnp* alone for hindcasting. However, for these methods, it is better to use *mnp* alone rather than both *dom* and *mnp*.

Comparing the pure Monache metric with the Normalised one, the results in Table 2 show that the latter leads to bias and RMSE reduction. The only predictor pair where this was not observed was *wds*, *mnp*, yet the differences were not meaningful as it corresponds to 4% of higher RMSE. The cosine method behavior resembles the Normalised metric, though with degraded performances in the error metrics. This similarity was expected as both methods find analogs on differences ranging from -1 to 1, i.e. looking solely for relative patterns in the time series.

Method	Bias	RMSE	MAE
Monache	0.51	1.98	1.48
Normalised	0.38	2.05	1.54
Cosine	0.37	2.03	1.53
Kmeans	**0.39**	**1.93**	**1.44**

Table 1. Predicting GST at *ykt* using only mnp station as predictor

Method	First station	Second station	Bias	RMSE	MAE
Monache	*dom*	*ykr*	0.95	2.34	1.75
Monache	*dom*	*mnp*	0.94	2.34	1.74
Monache	*wds*	*mnp*	0.24	1.75	1.33
Normalised	*dom*	*ykr*	0.26	1.79	1.36
Normalised	*dom*	*mnp*	0.68	2.01	1.52
Normalised	*wds*	*mnp*	0.49	1.83	1.43
Cosine	*dom*	*ykr*	0.43	1.81	1.37
Cosine	*dom*	*mnp*	0.87	2.12	1.61
Cosine	*wds*	*mnp*	0.73	1.94	1.51
K-means	*dom*	*ykr*	0.30	1.63	1.22
K-means	*dom*	*mnp*	0.43	1.85	1.38
K-means	*wds*	*mnp*	0.24	1.69	1.28

Table 2. Predicting GST at *ykt* using the different methods and various stations

As it can be seen in Table 2, Kmeans behave in the same way as Normalised Monache and Cosine, which implies that clusterization employs a similar idea as these two methods. Results, however, are noticeably better.

These results show a rift between the methods: while Kmeans, Cosine and Normalised Monache all give results following the same trend, Monache's results go in another direction. There is, however, a rational explanation for this behavior: Monache looks for analogs by minimizing the distance between the target variable's value at time t (when the prediction is made) and at time t (the analog). The other methods, however, disregard this distance. Instead, they look for a similar evolution of the weather during the time window. As such, the results imply that at the station *dom*, the weather follows a similar pattern as the station *ykt*, but because of the different location, meteorological values are not the same. This difference in value disturbs the Monache method, but not the other, who look at the underlying weather patterns.

All methods, however, give their worse results with the *dom-mnp* pair, and always by a clear margin, while the *wds-mnp* pair performs well in all cases. This suggests that *wds* is a much better station to predict the weather at *ykt* compared to *dom*, and that *ykr* is a very good predictor station too, since it is able to offset the inaccuracies caused by the use of *dom* as predictor (except in the case of the Monache method, since Monache gives a great emphasis on numerical distance between values). Overall, the *mnp* and *wds* pair is the one giving the lower RMSE overall across all methods.

As expected, results are improved when using two stations, compared to using just one. However, the results for Monache and Cosine methods suggest that the choice of stations is important to really have a gain in performance.

3.2 Studying the Weights
Now that the importance of using the correct stations has been assessed, it became important to evaluate if weighting the contribution from each station could improve on the results, and compare the two approaches described in section 2.5. For this, it was chosen to focus on the wds and mnp pairs, whose results gave the lowest RMSE overall in the previous test. The question is to determine if it is possible to improve these results even further by assigning weights to these stations.

For this purpose, here both the *independent stations* and the *dependent stations* variants were used. The former variant allows weights to be set for each individual station, while the latter does not, which is detailed in section 2.5. As a consequence, in Table 3 tests results showing "–" in the "Weight" column were tested ran with the *dependent method*, while tests with numerical values in the "Weights" column were tested ran with the *independent method*. The target variable was kept the same (GST), for ease of comparison with the previous results. The results are in Table 3.

Considering the results from Table 3 and looking at the stations independently Monache yields the best results, but only if the weights are equal. However setting most weight on *ykr* is better than setting most weight on *mnp*. Normalized Monache however, prefers to have the analogs looked across all the stations at once. Cosine shows no big difference between the dependent and independent methods. The *independent method* performs slightly better when maximal weight is assigned to one station. In agreement with previous results, it appears that independent Kmeans gives best results when *wds* has most of the weight. However, even then it performs clearly worse than dependent Kmeans.

The Figure 4 presents the Taylor diagram for the best case of each method. It is possible to see that the Normalised method

Method	Weight *wds*	Weight *mnp*	Bias	RMSE	MAE
Monache	–	–	0.24	1.75	1.33
Monache	0.1	0.9	0.52	1.95	1.45
Monache	0.5	0.5	0.40	1.78	1.33
Monache	0.9	0.1	0.27	1.74	1.31
Normalised	–	–	0.49	1.83	1.43
Normalised	0.1	0.9	0.61	1.99	1.55
Normalised	0.5	0.5	0.54	1.87	1.45
Normalised	0.9	0.1	0.49	1.83	1.42
Cosine	–	–	0.73	1.94	1.51
Cosine	0.1	0.9	0.05	1.92	1.49
Cosine	0.5	0.5	0.24	1.95	1.51
Cosine	0.9	0.1	0.05	1.93	1.50
Kmeans	–	–	0.24	1.69	1.28
Kmeans	0.1	0.9	1.22	2.56	1.95
Kmeans	0.5	0.5	0.50	2.16	1.66
Kmeans	0.9	0.1	-0.25	2.10	1.65

Table 3. Predicting GST at *ykt* using the different methods

gives the best results. Its proximity to the truth indicates a high correlation coefficient with it, a low Root-mean Square (RMS) distance to the truth and similar positions on the X-axis shows similar standard deviations - in other words, the forecast obtained by the Normalised method is close to the truth. Predictions from the Kmeans and Monache methods are very close to one another, and also close to the truth, with a high correlation coefficient and a low RMS distance to the truth. However, they are closer to the origin on the X-axis, indicating that their standard deviation is lower than that of the truth. In other words, these methods have troubles following the variations of GST accurately. Cosine is the method, which performs the worst, and its coefficient correlation with the truth is lower than for other methods, and the RMS distance to the truth is higher. However,
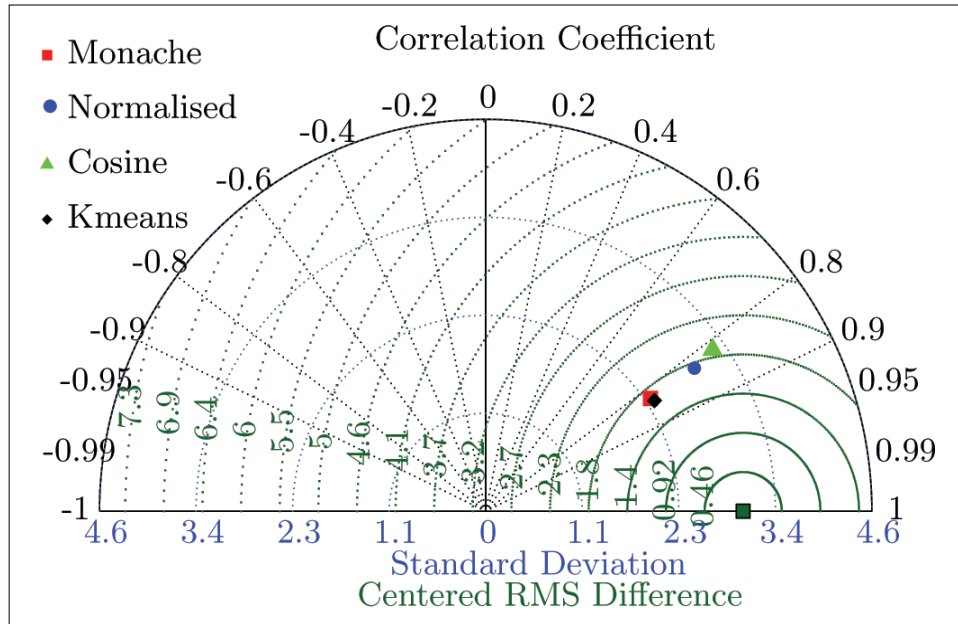

Figure 4. A Taylor diagram comparing the best of each of the four methods. The Square at the bottom represents the truth
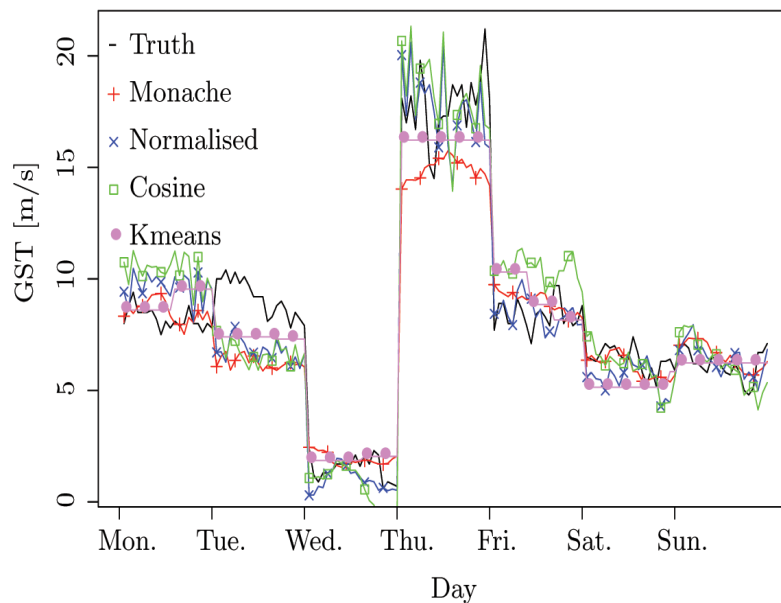

Figure 5. Prevision compared to the truth, for the first week of January 2018

it has the closest standard deviation to the truth, meaning the cosine method is the method that follows the variations of GST the most accurately.

Figure 5 shows the forecasts obtained with the different methods jointly with the truth values. As expected, the forecast by the Normalised method appears to the one closest to the truth. It is interesting to note that while the forecasts by the Monache and Kmeans methods behave similarly, they look rather different from one another. Cosine, as expected, is the furthest one from the truth but displays a lot of variabilities.

Qualitatively the Cosine and Normalized methods give a representation of the time series with higher fidelity, due to additional variance. Quantitatively, however, the additional variance introduces mismatches which result in poorer performance when compared against the Kmeans and Monache methods.

4. Conclusions

In this work meteorological data was predicted at one location based on multiple historical datasets from weather stations. To achieve this, the Analog Ensembles method was applied and several methods were explored, by changing the metric used to determine the analogs in the historical dataset. The prediction horizon was two years, based on a training period of four years of historical and observed time series.

From all these results, it appears clearly that the choice of stations, and how to weight them if a weighted approach is used, has a very important bearing on the hindcasting, and presumably forecasting, accuracy. The problem of selecting stations for hindcasting and forecasting purposes in a non-trivial one, and from these experiments it would appear that the best way to make a viable selection is to simply test hindcasting on known data, to determine which are the stations most suited to forecasting and hindcasting purposes at the target. The use of the K-means metrics leads to an improvement ranging from 3% to 30% of lower quadratic error when compared against the Monache metric. Increasing the predictors to two stations improved the performance of the hindcast, leading up to 16% of lower error, depending on the correlation between the predictor stations. These features show the improvements which can be made on the existing AnEn method.

As future work, one main possibility to explore is about the clustering approach. The results look very promising; however, there are a number of parameters required to be set that were not looked at. Therefore, the number of clusters was left at a basic value. It is possible to tweak this value and see how to best set this value for maximal accuracy, or even control the size of clusters. The K-means algorithm is also a fairly basic clustering algorithm and more accurate algorithms now exist. It would be interesting to look at their performances compared to the basic Kmeans in this case.

It is also possible to look at larger scales, both in terms of a number of variables, of number of stations, or of the distance between stations. Here, this study focused on a rather simple testing environment, with stations located all next to each other. As a next step, it would be interesting to look at the performances of these various approaches at larger scales.

References

[1] Mathworks, https://www.mathworks.com/

[2] National Center for Atmospheric Research, https://nar.ucar.edu/

[3] National Data Buoy Center, https://www.ndbc.noaa.gov/

[4] Adachi, K.: Matrix-Based Introduction to Multivariate Data Analysis. Springer Singapore (2016). https://doi.org/10.1007/978-981-10-2341-5

[5] Alessandrini, S., Monache, L.D., Sperati, S., Cervone, G.: An analog ensemble for short-term probabilistic solar power forecast. Applied Energy **157**, 95–110 (2015). https://doi.org/10.1016/j.apenergy.2015.08.011

[6] Alessandrini, S., Monache, L.D., Sperati, S., Nissen, J.: A novel application of an analog ensemble for short-term wind power forecasting. Renewable Energy 76, 768–781 (2015). https://doi.org/10.1016/j.renene.2014.11.061

[7] Alessandrini, S., Monache, L.D., Rozoff, C.M., Lewis, W.E. (2018). Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Monthly Weather Review,* 146(6), 1723–1744 (2018). https://doi.org/10.1175/mwr-d-17-0314.1

[8] Chai, T., Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geoscientific Model Development,* 7(3), 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014

[9] Van den Dool, H.M.: A new look at weather forecasting through analogues. Monthly Weather Review 117(10), 2230–2247 (1989). https://doi.org/10.1175/1520-0493(1989)117¡2230:anlawf¿2.0.co;2

[10] Dool, H.V.D.: Searching for analogues, how long must we wait? Tellus A: Dynamic Meteorology and Oceanography 46(3), 314–324 (1994). https://doi.org/10.3402/tellusa.v46i3.15481

[11] Guti´errez, J.M., CofiÜno, A.S., Cano, R., Rodr´1guez, M.A.: Clustering methods for statistical downscaling in short-range weather forecasts. *Monthly Weather Review,* 132(9), 2169–2183 (2004). https://doi.org/10.1175/1520-0493(2004)132¡2169:cmfsdi¿2.0.co;2

[12] Lorenz, E.N.: Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences,* 26(4), 636–646 (1969). https://doi.org/10.1175/1520-0469(1969)26¡636:aparbn¿2.0.co;2

[13] Monache, L.D., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K.: Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review,* 141(10), 3498–3516 (2013). https://doi.org/10.1175/mwr-d-12-00281.1

[14] Monache, L.D., Nipen, T., Liu, Y., Roux, G., Stull, R.: Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review,* 139(11), 3554–3570 (2011). https://doi.org/10.1175/2011mwr3653.1

[15] Taylor, K.E.: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres,* 106 (D7), 7183–7192 (2001). https://doi.org/10.1029/2000jd900719

[16] Vanvyve, E., Monache, L.D., Monaghan, A.J., Pinto, J.O.: Wind resource estimates with an analog ensemble approach. *Renewable Energy*, 74, 761–773 (2015). https://doi.org/10.1016/j.renene.2014.08.060