

PCA analysis of wind direction climate in the baltic states

By MAKSIMS POGUMIRSKIS*, TIJA SĪLE, JURIS SEŅŅIKOVŠ, and ULDIS BETHERS,
University of Latvia, Riga, Latvia

(Manuscript Received 3 June 2021; in final form 26 July 2021)

ABSTRACT

Wind direction is one of the fundamental parameters of weather. In this study we investigate the wind direction climate 10m above surface level in the Baltic States (Estonia, Latvia, Lithuania). The analysis of wind direction over larger regions is usually hindered by the fact that wind direction is a circular variable, which means that averaged values are meaningless. Here we show how Principal Component Analysis (PCA) can be applied to give a large scale overview of typical wind direction patterns in the region. Here we apply PCA to both observational and reanalysis data. The most significant wind direction patterns are detected in both synoptic scale and mesoscale, and we attempt to link the identified patterns with meteorological phenomena. In addition, the differences in the PCA results between observation and model data are analysed.

The results show that PCA method is successful in identifying and ranking the wind direction climate features, leading to a complete and thorough investigation for the whole region that would be not possible by human researchers analysing individual distributions of wind direction.

Keywords: wind direction, principal component analysis, Baltic Sea, coastal winds

1. Introduction

Wind direction is one of the fundamental parameters that characterises weather. It has a significant impact on human activities on its own, in transport industry or in air pollution modelling (Gupta and Dhir, 2013), but through advection it impacts other weather parameters, such as the temperature, e.g. the temperature in coastal zones depends on whether the wind is from offshore or onshore directions. Information about wind direction is also necessary to retrieve wind speeds over the oceans using satellite data (Koch, 2004).

The Baltic States are located on the eastern coast of the Baltic sea, and, as is typical for mid-latitudes of the Northern hemisphere, western and southwestern winds are prevalent. The wind climate is influenced by the large-scale air pressure fields such as the Icelandic low, the Siberian high and the Azores high (Soomere and Keevallik, 2001).

The direction of large-scale flow often determines weather conditions in the region. In the mid-latitudes north wind is usually related to colder and drier weather, however, south wind is usually related to warmer weather.

During the winter in Europe warmer air usually comes from western directions - from the Atlantic ocean. Colder air comes from eastern directions (Jaagus and Kull, 2011). Most of the cyclones enter the east Baltic region from the southwestern and western directions during both winter and summer (Sepp et al., 2018). The majority of airmasses which bring rainfall come from southwestern and western directions (Jaagus et al., 2009).

During the warm season, increased frequency of shore-parallel southwestern and northern winds have been observed near the western coast of Latvia, which can be explained by a low-level jet mechanism (Sīle et al., 2018).

Compared to other parameters, such as temperature or wind speed, wind direction has received less attention in research literature. One of the possible explanations is that if scalar parameters can be easily interpreted by averaging them over time, there is no similarly simple way of summarizing information about wind direction. Therefore, the analysis of wind direction necessarily means analysing the entire wind direction distribution, which makes it harder to visualize and interpret the results.

The goal of this study is to investigate whether Principal Component Analysis (PCA) can provide an

*Corresponding author. e-mail: maksims.pogumirskis@lu.lv

easy way of summarizing the wind direction climate. Here the Baltic States are used as a study region. PCA is a mathematical method that identifies features in large datasets and has a long history of being used in climatological analysis. For instance, the North Atlantic Oscillation (NAO) index can be defined as the timeseries of the first mode of PCA applied to the sea level pressure anomalies over the Atlantic sector (Rutgersson, 2015). Applications of PCA in studies of wind climatology include classifying and investigating wind field patterns over the USA (Klink and Willmott, 1989) and the Iberian Peninsula (Pedro et al., 2009) and grouping observation stations based on wind gust patterns (Jungo et al., 2002). In more recent studies PCA has been used as a dimensional reduction tool in artificial neural network based weather forecasts (Mezaache et al., 2016) and used for climate model data for the eastern Baltic region to derive climate indices based on monthly mean temperature and precipitation (Bethere et al., 2017).

The goal of this study is to describe and characterize the properties of wind direction distributions in Baltic States and especially Latvia, using both observations from surface stations and modelling data from UERRA reanalysis (ECMWF, 2019). The two main research questions are:

- What are the main patterns and the seasonal cycle of wind direction distributions as seen in PCA results? Can these patterns be related to specific meteorological processes?
- Are the PCAs of model data similar to those of observational data? Can these differences be interpreted as differences between model and observations, that is, can they be used to analyze model performance?

The study used data from the period 1976–2019. First, it was necessary to analyse the changes in observational methodology, then the wind direction distributions from high spatial resolution (11 km) modelling data from UERRA reanalysis were compared with observations using Earth Mover’s Distance (EMD). Further, PCA was applied to both datasets and the main patterns analysed in connection with meteorological processes.

2. Data and methods

2.1. Observation data

For this study the long term near-surface (10 m) wind direction observations carried out by the Latvian Environment, Geology and Meteorology Centre (LEGMC) was used from 22 observation stations (Fig. 1).

The observational data in stations are available since at least year 1966. The initial analysis of the data showed that substantial changes in measurement methodology

have taken place and their impact should be addressed before further analysis. Changes in both the frequency of the observations and in the resolution used for reporting measurements have taken place and, importantly, these changes did not happen at the same time for all stations. The goal of the study is to identify patterns in wind direction distributions using the PCA method, and differences between stations are important. Therefore, it was crucial to make the data as homogeneous as possible to avoid artefacts caused by differences in data resolution.

At the beginning of the period of interest, wind direction observations were performed by human observers every 3 hours. Until October 1976 wind direction in all observation stations was measured using 16 cardinal directions (N, NNE, NE, ...). In November 1976 most observation stations (18 out of 22) transitioned to 10-degree resolution, however, the transition to 10-degree resolution in all stations was finished only by 1989. Between 2001 and 2003 automatic hourly observations with 5-degree resolution were introduced. In 2014 gradual transition to 1-degree resolution began.

As with the introduction of the automatic observations temporal resolution of the observations had increased from 3 hours to 1 hour, adjustments had to be made. To avoid bias towards the more recent period, when observations were recorded once an hour, hourly observations were assigned a third of the weight of 3-hourly observations when calculating the wind speed distribution.

We believe that in this study we can mix together 1-hour and 3-hour data to increase the number of data-points because the typical time scale of wind direction variation is much larger for the processes that we expect to influence our results. Distinction between hourly and 3-hourly data would be necessary if we were to investigate the meteorological processes where the wind direction could change significantly in less than three hours in a nonlinear manner. Such events are quite rare, mostly associated with the passage of meteorological fronts and not the focus of this study. Therefore, we believe that usage of both hourly and 3-hour data is permissible in our case.

It has long been known (Ratner, 1950) that human observers tend to report 8 principal winds (N, NE, E, ...) more often than 8 half-winds (NNE, ENE, ...) when measuring wind direction using 16 cardinal directions. Since this pattern was also observed in the LEGMC data, we decided to exclude the observations that were reported using 16 wind directions rose from further analysis, shortening the analysis period and starting it from November 1976. After this decision, the time series are still more than 40 years long.

For further analysis, all observations were divided into 16 bins, each bin being 22.5 degrees, and bin centres



Fig. 1. Locations of 22 observation stations in Latvia that record wind direction. Observation station Rēzekne (grey marker) was not used in further analysis, since it did not meet the data completeness criteria for the period of study (see text).

corresponding to 16 cardinal directions. Then each month between November 1976 and October 2019 was analyzed for each of the 22 observation stations. Months with incomplete data and months when wind direction was measured using the 16 wind directions rose format were excluded from further analysis. We checked our data completeness against World Meteorological Organization (WMO) guidelines.

According to World Meteorological Organization guidelines (16) wind direction frequency for the particular month should be considered incomplete and therefore excluded, if there are missing observations on 11 or more days during that month. As we are working with hourly data, we considered a day 'missing' if there was at least one missing observation. Additionally, WMO guidelines state that When describing the annual cycle, e.g. calculating the wind direction distribution for January over a multi-year study period, average values should be complete for 80% of the months constituting the study period.

Monthly data were incomplete on average in 7.2% of the months. In all stations, except one, less than 10% of the monthly data were incomplete. The one station that

had 32.8% of its data missing was excluded from further analysis. Such a large number of missing data in a single station is explained by the data being recorded there using 16-wind rose format until 1988. Therefore, we conclude that our observational dataset can be used for further analysis according to WMO guidelines. The exclusion of a single station also did not significantly change the PCA results.

Calm conditions or variable wind conditions were excluded when calculating wind roses, since no particular wind direction is assigned. These conditions were reported 2.3% of the times at the coastal stations and 5.8% of the times at the inland stations (4.7% in average over all the stations). These datapoints are not counted towards 'incomplete' data in the analysis described in the previous paragraph.

Since the directional resolution of observational data were changed during the analysis period, additional processing had to be carried out to correctly create the monthly wind direction distributions from data with different resolutions. There is a known systematic bias which should be avoided when converting higher resolution (e.g. 10 degree) wind direction observations into 16

wind directions rose format (Lea and Helvey, 1971). When measuring direction with a 16-point compass, the size of each sector is 22.5 degrees. Sectors corresponding to the four cardinal directions (N, E, S, W) contain centres of three 10-degree bins, whereas the other twelve sectors corresponding to intermediate directions contain only two centres of 10-degree bins. Therefore, just mapping every 10-degree to bin to a single 22.5-degree sector will introduce positive bias for the four cardinal directions (N, E, S, W). To remedy that, some of the 10-degree bins should be mapped to two adjacent 22.5-degree sectors at the same time, statistically distributing the observations. The methodology is described in detail in (Lea and Helvey, 1971). A similar method of correction was applied to observations with 5 degree and 1 degree resolution.

Analysis of observational data revealed another inconsistency. During the period where ostensibly the 36 wind directions rose (10-degree resolution) was used by the human observers, 10-40% of reported values ended with ‘5’, e.g. ‘15’. Such abnormalities were present in 27% of months when 10-degree precision was used, suggesting that some observers were in fact using a 72 wind directions rose (5-degree resolution). In comparison, for automated wind direction measurements with 5-degree resolution close to 50% of observations end with ‘5’. To ensure homogeneity, measurements, which ended with ‘5’ before introduction of automatic observations, were equally distributed between adjacent 10-degree bins.

After the bias correction procedure monthly wind direction distributions were calculated for each of the 21 observation stations, with 12 months for each station giving 252 wind roses in total, i.e. data from all Januaries in the study period were aggregated in a single distribution for each station.

2.2. Model data

In this study modelling data from the UERRA reanalysis (Uncertainties in Ensembles of Regional Reanalyses) were used. It has a horizontal resolution of 11×11 km and covers the time period from January 1961 to July 2019. The reanalysis uses the HARMONIE numerical weather prediction model with observational data assimilation. We use hourly timeseries by combining analysis that is available every 6 hours with the first 5 hours of hourly forecast data. Meanwhile, mixed hourly and 3-hourly data were used from the observation dataset. We believe that the usage of hourly model data can be justified with the increased number of datapoints and the fact that almost half of the observations are also with hourly temporal resolution. Our arguments why hourly and 3-hourly data can be mixed together has been set out earlier.

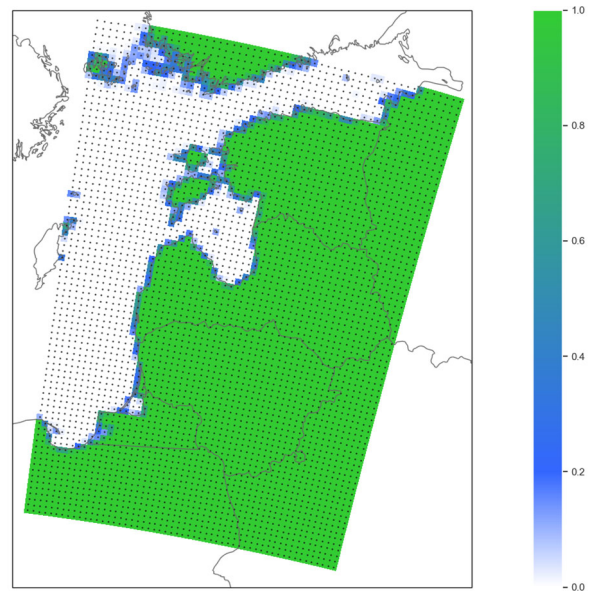


Fig. 2. UERRA reanalysis data used in this study (56×80 gridpoints). Black dots indicate centres of gridpoints. Colours show model data for land/sea coverage of grid cells, 1 – land, 0 – sea.

The UERRA reanalysis covers the whole of Europe, however, for this study a rectangular region (56×80 grid points, the grid is rectangular in the map projection used in the model) on the model grid was selected containing Latvia, Estonia, Lithuania and the adjacent part of the Baltic sea. Figure 2 also shows the representation of land and sea in the model.

Using hourly model data between November 1976 and December 2017 an average wind rose for each grid point and for each of the 12 months of the year was calculated. Datapoints which had wind speed below 1 m/s were excluded to maintain consistency with the observational data. Therefore, on average 8% of the data were excluded over the land and 1.7% over the sea (6% over the whole area of study).

All maps were plotted using Cartopy (Met Office, 2010).

2.3. Diurnal cycles

To better identify the possible meteorological causes for some of the phenomena seen in PCA, for example, the timing of coast perpendicular wind directions that can be associated with sea-breeze, it was necessary to analyze the diurnal cycle of wind direction distributions. Such analysis was carried out for both observational and model data. A two-dimensional histogram was created for each month in each station, where the wind direction

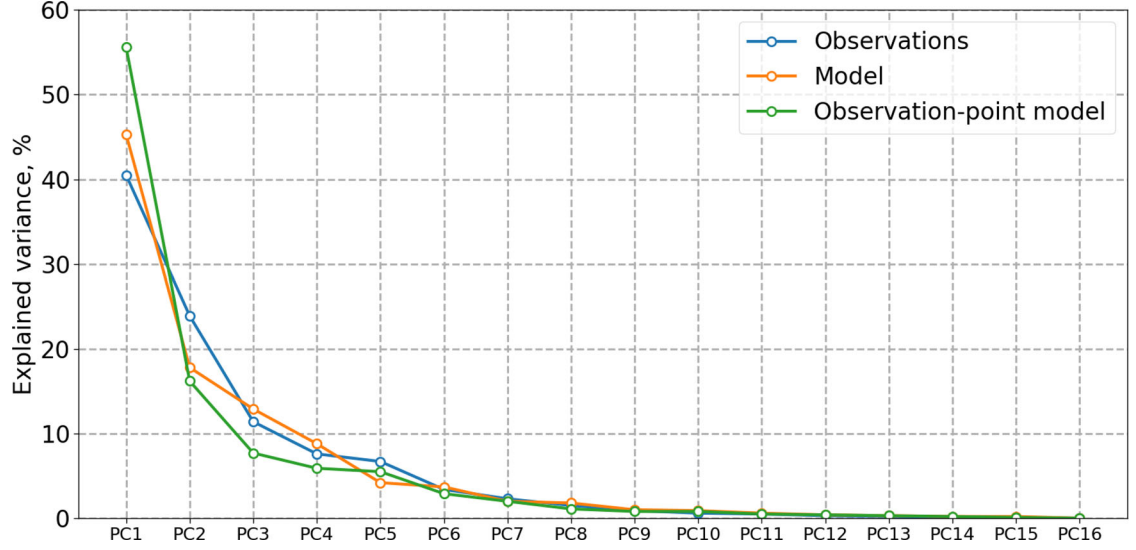


Fig. 3. Explained percentage of variance of the principal components from all three datasets.

probability in each of the 16 direction bins was also analyzed as a function of the time of the day.

At first, the time of the observations was converted to UTC to match the model timezone. Wind direction data were distributed into eight 3-hour bins (e.g. 0:00-2:00 UTC, 3:00-5:00 UTC).

2.4. PCA method and its application to wind direction data

PCA is a dimensionality reduction technique, which finds a correlation in the initial variables and creates new, uncorrelated variables, which are linear combinations of the initial variables. The new variables, called Principal Components (PC), are ranked by the amount of variance which they can explain. The new variables that do not explain significant amount of variance can be excluded from further analysis (Jolliffe, 2002). In this study PCA was performed on model and observation data separately.

The first PCA was performed on the observation data. PCA is typically applied to a $n \times p$ matrix of data, where the columns correspond to p different variables and each of n rows correspond to an ‘observation’ or ‘experiment’ that is interpreted as a datapoint in p -dimensional space. Each of the study-period averaged month for each of the stations (252 wind roses in total – 21 observation stations times 12 months) were the datapoints or ‘experiments’, the initial variables were the 16 compass directions, but the ‘data’ here is the probability of wind blowing from each of the 16-point compass directions in a specific month at a specific station. To simplify – each of the 252 monthly wind roses form a single row in the data matrix. The sum of values of each row is 1 for all rows.

The second PCA was performed on the model data. The general idea here is the same, but each grid-point acts as an observation station, and therefore the matrix is formed by 53,760 rows (wind roses) for each of the 56×80 grid points and for each of the 12 months.

One of the goals of the study is to compare the results of the PCA performed on model dataset with the PCA performed on observational dataset. As the PCA describes the most common features of the dataset, differences in the PCA results could be introduced by the fact that the observations cover only a part of the model region and only over dry land. Therefore, if the goal is to assess the model performance in the station locations using the PCA method, a third PCA needs to be carried out on a subset of the model data, where only the data from grid-points corresponding to observational stations are included. This is the third PCA, where the matrix has the same dimensions as the observational PCA matrix (252 datapoints for each of 16 wind directions), but the data comes from the reanalysis. This dataset is called ‘observation-point model dataset’ further in this study. (For further analysis please see section ‘Comparison Between Model and Observations’).

The PCA method can be interpreted as changing the coordinate system of the data and finding a new coordinate system where the properties of the dataset are easier to analyze. The main features of the dataset should be represented in the new variables (principal components), but how distinct this feature is for each experiment (in this case, for a specific month in a specific station) is being shown by the values of the data in the new coordinate system.

Observational stations provide information about wind conditions at specific points. If a meteorological

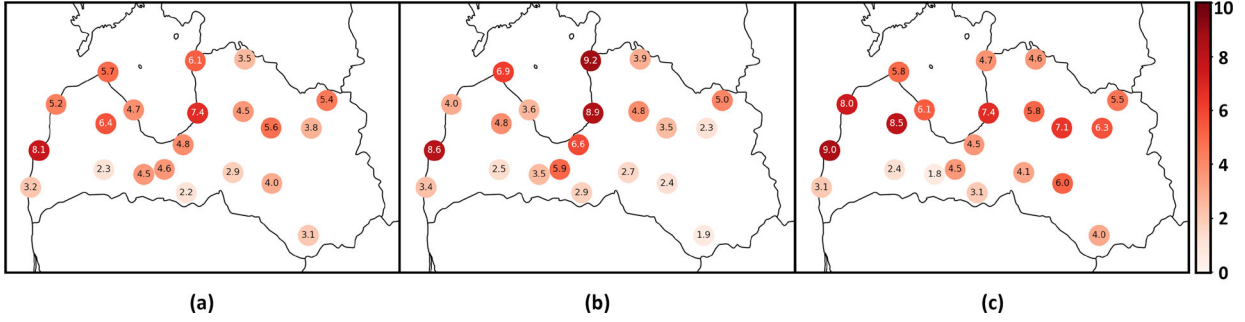


Fig. 4. EMD (measured in **degrees**) between observations and model (a) averaged for the whole year, (b) for December and (c) June. Darker (redder) colours indicate higher (worse) EMD values.

phenomena has a typical scale that is comparable with the distance between the stations, the results from a single observation station can be hard to interpret.

To investigate such cases, the question ‘where in the model dataset can we see features characteristic to observations’ can be asked. The answer can be found projecting the model dataset on the principal components acquired from the observational dataset. Such projection would allow to analyze the spatial extent of features seen in observational dataset.

Therefore, the loadings of the principal components from the observation dataset were applied to the model dataset and plotted onto a map. The goal of the PCA method is to analyze the deviations from the long-term mean wind direction distribution, therefore, at the beginning, the average frequency of each direction in the model dataset was subtracted from each of the wind roses. The resulting data were arranged into rows of a $16 \times 53,760$ data matrix, which was multiplied by a 16×16 matrix containing the loadings of the 16 principal components acquired from the observation dataset in its columns. The resulting $16 \times 53,760$ matrix rows contains values of the 16 principal components from the observation data applied to 53,760 model datapoints.

One of the common questions regarding PCA is choosing the adequate number of principal components to explain variation in the studied variable. The scree plot of the explained variance of the principal components in all three datasets is shown in Fig. 3. While the plot shows no clear number of principal components to be retained, the first two should certainly be kept. The third principal component was retained, since it had a clear physical interpretation (shown in section PCA). Meanwhile, further principal components were hard to interpret. The total explained variance of the first three components was 76.0% in the model dataset, 75.4% in the observation dataset and 79.5% in the model observation-point. These lay in the sensible 70% to 90% cut off range according to (Jolliffe, 2002).

2.5. Comparing model and observations

To evaluate the performance of the model, observations were compared with the results of the model in the grid-point closest to the observation station. Additional consideration is needed when choosing model data corresponding to coastal stations as the model resolution is coarser than the typical distance between the coast and the station, and therefore the nearest gridpoint can be in the sea. As the coastline can significantly influence wind direction, we used the grid-points corresponding to 100% dry land determining it from the model ‘land/sea’ variable that has a value between 0 and 1 (see also Fig. 2).

The averaged values are typically compared between different datasets to analyse scalar variables such as wind speed or temperature. Wind direction is a circular variable and therefore approaches that compare ‘average’ wind direction and metrics derived from that, such as bias, cannot be applied. Metrics that describe differences between the distributions can be used instead and several such metrics are available (Rubner et al., 2000). In this study we choose to use Earth Mover’s distance (Pele and Werman, 2008, 2009) between two distributions that is defined as the minimal cost to transform one distribution into another. Recently EMD has been applied to compare wind speed and direction distributions for wind energy applications (Hahmann et al., 2020). If the value of EMD is 0 then the two distributions are identical, and the larger the value of EMD the more dissimilar the distributions. If two probability distribution functions of a variable are compared, EMD is measured in the units of the variable. EMD was calculated for the annual and monthly distributions of the wind direction over the whole study period.

3. Results

3.1. Comparison between model and observations

The difference between the modelled and the observed wind direction distributions for the whole dataset is

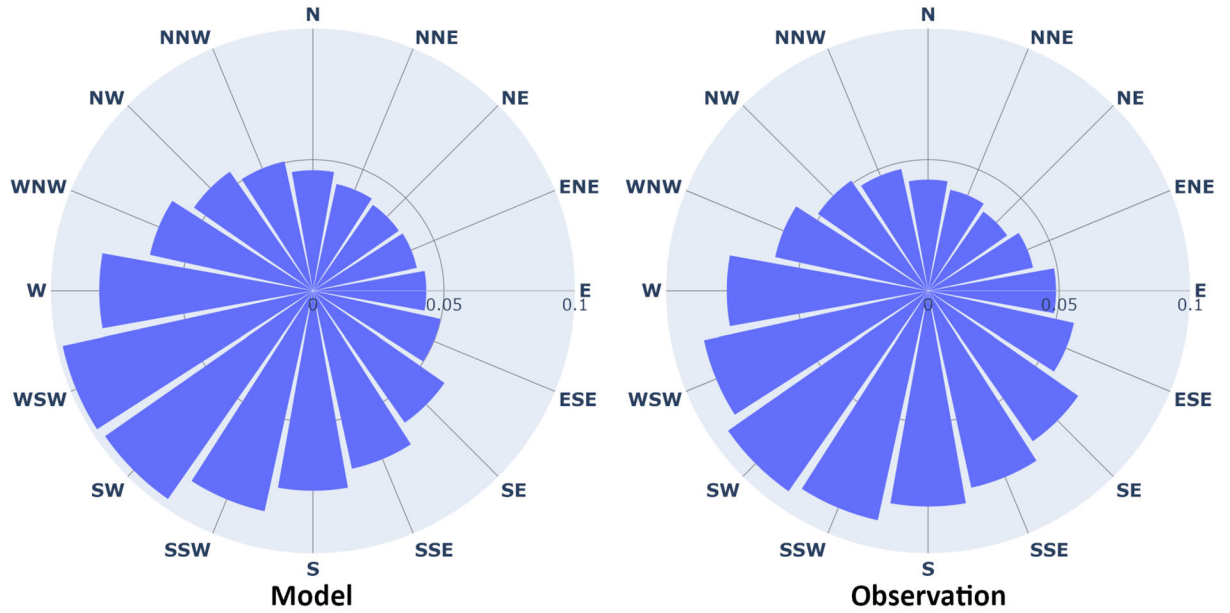


Fig. 5. Average yearly wind roses for model dataset (left) and observation dataset (right).

shown in Fig. 4. Figure 4a shows the EMD values for the whole year, Fig. 4b shows the differences in December, representing the cold season and Fig. 4c shows the differences for June, representing the warm season.

Overall, there is a good agreement between model and observation, with the largest differences (as measured using the EMD metric) not exceeding 10 degrees. The lowest EMD values are inland, in the south of Latvia (Fig. 4a). The difference between inland and coastal stations are even more pronounced in the cold season (Fig. 4b) when coastal stations, especially on the eastern side of Gulf of Riga, show larger (worse) EMD values than those that are further inland.

The disagreement between model and observation does not necessarily mean that the model is wrong. In models, the grid-cell average value of a meteorological parameter is provided, while the observation data can be influenced by the immediate surroundings of the station, such as trees or buildings. Extended analysis of the sources of the disagreement between the models and the observations is beyond the scope of the paper, but analysis of the diurnal cycle representation in stations with the largest EMD scores showed that the differences between model and observation have an annual, and in some cases diurnal cycle. Therefore, they seem to be caused by imperfect representation of mesoscale meteorological processes. In coastal zones the model typically predicts a broader wind direction maximum than the observations.

3.2. PCA

The results of PCA can be interpreted as deviations from the averaged wind regime, where the averaging is done over the whole study region and time period, i.e. the average wind rose for all months and all stations. This average wind rose is shown for both observational and model datasets in Fig. 5. These plots show a distinct frequency maximum for the southwestern (WSW-SSW) winds and with the north-easterly (N-ENE) winds being the rarest, as is typical in the region. The model average wind rose is similar to the observational wind rose and it means that, overall, the model has a good correspondence with the observations and, crucially, the principal components of the observational dataset can be directly compared with the ones for the model.

The loadings of the first four principal components for both observation data, model data and model data in observational gridpoints are shown in Fig. 6. They represent the most prevalent deviations from the mean wind rose. The original wind roses can be reconstructed by a combination of the average wind rose (Fig. 5) and the PCA loadings multiplied by the values of the principal components. The values of the PCA components are shown in e.g. Figs. 7 and 8.

But first, how to interpret the PCA loadings in Fig. 6? The probabilities of wind blowing from different directions are not independent of each other and, for a specific

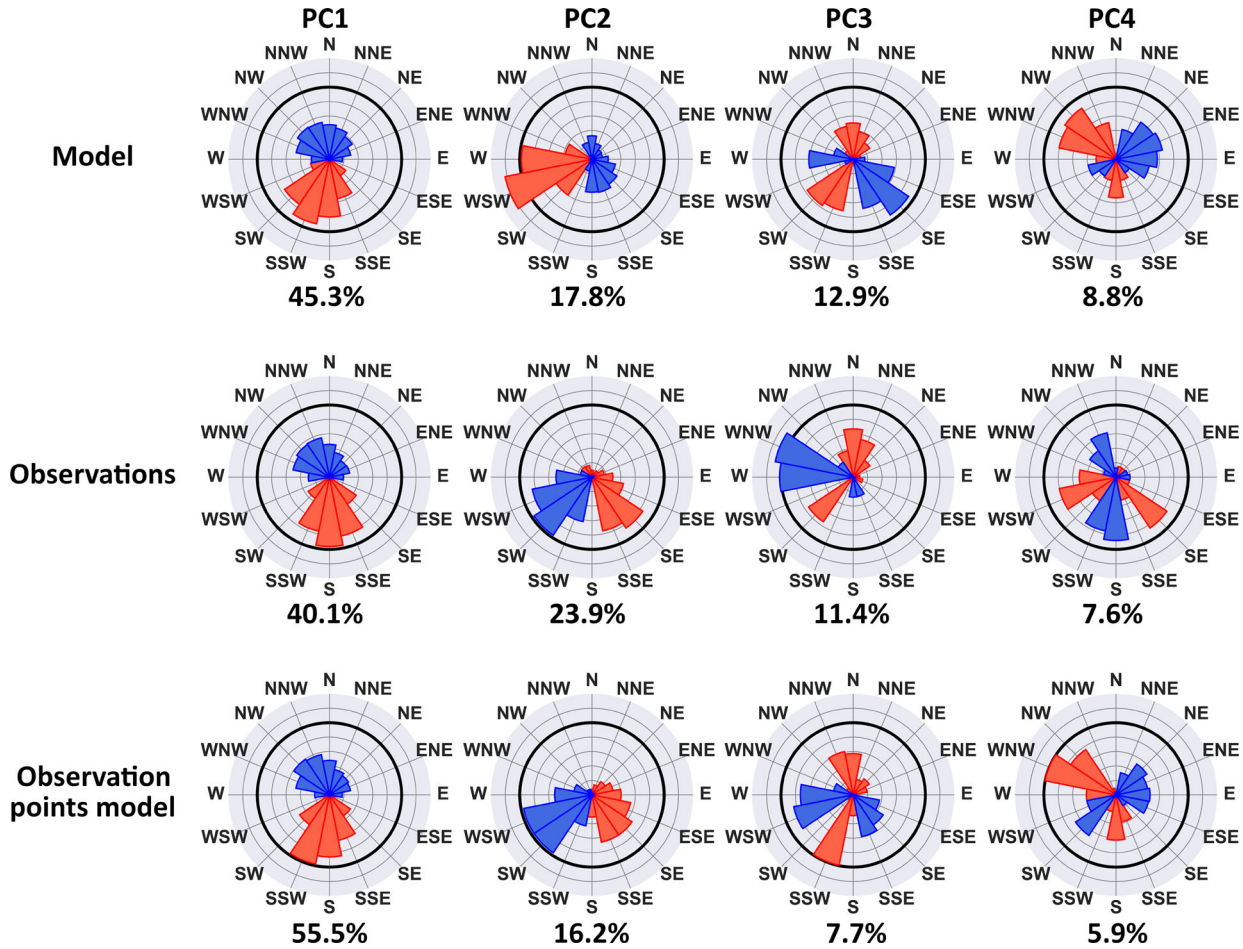


Fig. 6. The loadings of the first 4 principal components for the full model dataset (top row), the observation data (middle row) and the model data from station points only (bottom row). Wind directions are shown on the angular axis, absolute values of the loadings are shown on the radial axis. Grey circles mark increments of 0.1, black circle is at the 0.5 mark. The positive coefficients are red, the negative coefficients are blue. The explained variance of each of the principal components is shown below the loadings.

station and month, an increased probability of wind blowing from a certain direction can be associated with an increased probability of wind blowing from, e.g. adjacent direction. Information about such correlations and anti-correlations is provided by the PC loadings. For instance, PC1 for the model data show that the increased occurrence of SSW wind is correlated with an increased occurrence of SW and S winds, but anti-correlated with an increased occurrence of NW winds. PC1 will be analyzed in detail further in text.

In the case of PC2, significant differences between observational and model dataset can be seen. However, this result cannot be interpreted as proof of the significant differences between the model and the observations, because the PCs for the subset of model datapoints corresponding to the station locations (third row in Fig. 6) are more similar to the observational PC2 (second row). The loadings for PC2 therefore exemplify both the discrepancies in PC

that are caused by differences in spatial coverage (model vs observations) and differences in substance between the model and the observations (seen when comparing the loadings for ‘model in observational grid points’ with loadings for observations).

The PC loadings shown in Fig. 6 describe the most common patterns of wind direction distributions seen in the dataset, but the presence of these patterns for specific stations or months can be investigated by analyzing the values of the principal components. The higher the value of PC, the more pronounced is this pattern. In the next sections the values of PC will be analyzed and a link with meteorological phenomena will be sought.

3.2.1. Interpretation of PC1 results. Positive values of PC1 mean an increased frequency of S winds, whereas negative – an increased frequency of N winds, when compared to the long-term mean wind direction distribution.

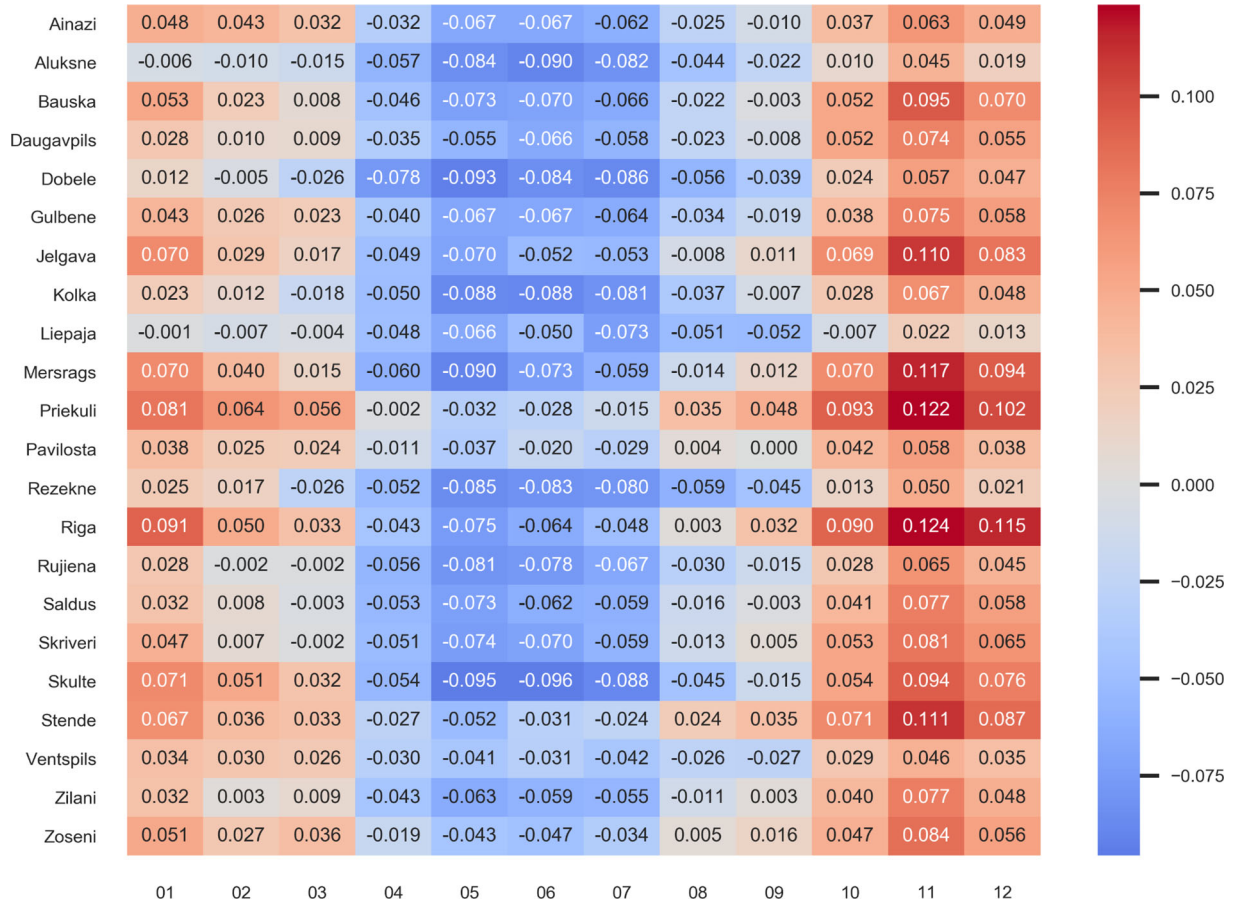


Fig. 7. Principal component analysis of observation data. Values of PC1 in observation stations (vertical axis) throughout the year. Months are shown on the horizontal axis.

For both observational and model datasets, values are spatially homogeneous over the dry land and the Gulf of Riga and slowly varying over time with a distinct annual cycle, reaching their lowest values in May and highest values in November (Fig. 7). Over the Baltic sea the seasonal changes of PC1 are less evident (Fig. 8). This means that in summer there is an increased frequency of N winds and in winter – an increased frequency of S winds for most of the study region. The large spatial extent suggests that the explanation must be sought in the synoptic scale.

To check the veracity of PCA results we plotted the wind direction distributions corresponding to Fig. 8 in points with characteristic PC1 values Fig. 9. The deviation from averaged wind roses seen in Fig. 9 are largely similar to the PC1 loadings seen in Fig. 6, with pronounced maxima from SW direction in winter and more diffused (NW - N - E) winds during the summer. This pattern seems to be less evident in the Baltic Proper, west of Estonian archipelago.

Similar seasonal changes in the large-scale wind directions have been reported in literature. In the cold season,

synoptic scale pressure gradients between the deep Icelandic low and strong anticyclones over Russia and Azores cause the flow over the wider Baltic Sea region to be strongly southwestern (Team, 2008). This mechanism was referred to when describing the increased prevalence of S and SW winds in the cold season in Latvian observation stations (Briede, 2016).

This mechanism could be used to explain the positive values of PC1 component seen in Figs. 7 and 8. However, then two more questions need to be answered. First, if the synoptic scale flow is SW, why are the highest positive values of loadings of the PC1, especially in the observations, mostly from the south? Second, if the SW flow is of the synoptic scale, why are the values of PC1 high over the dry land and close to zero over the Baltic proper?

First, one should remember that the principal components describe deviations from the average flow, which in this case has a strong SW component already (Fig. 5), explaining why PC1 values over the sea can be close to zero. Second, a tentative explanation for both questions

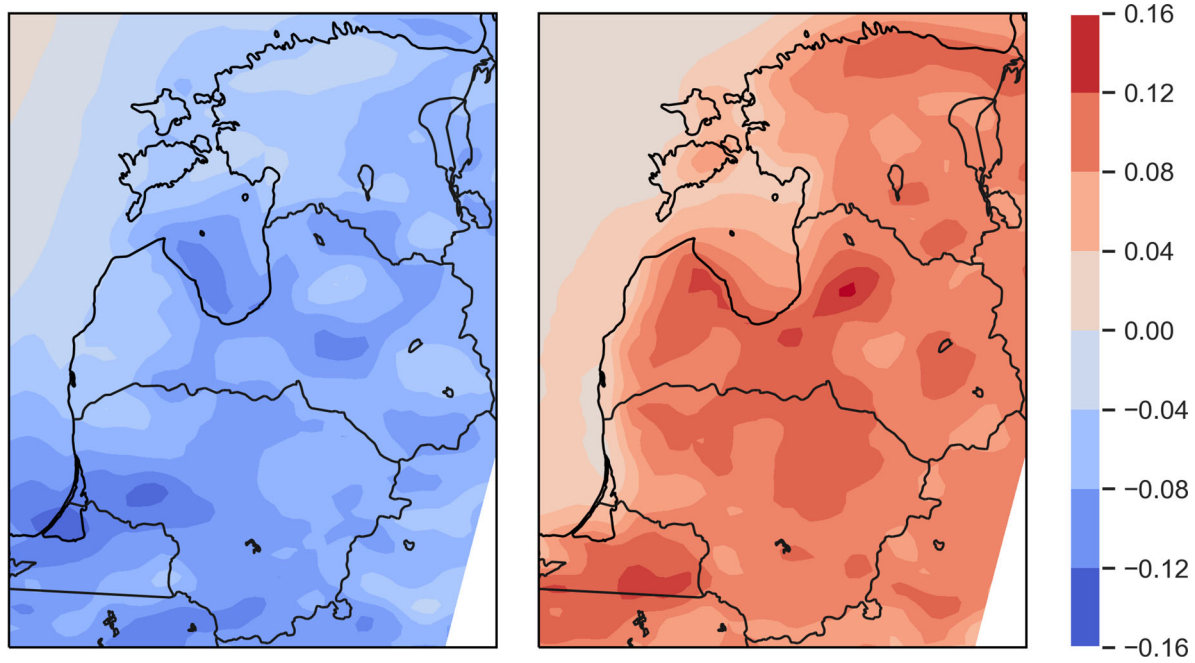


Fig. 8. Principal component analysis of model data. Values of PC1 in May (left) and November (right).

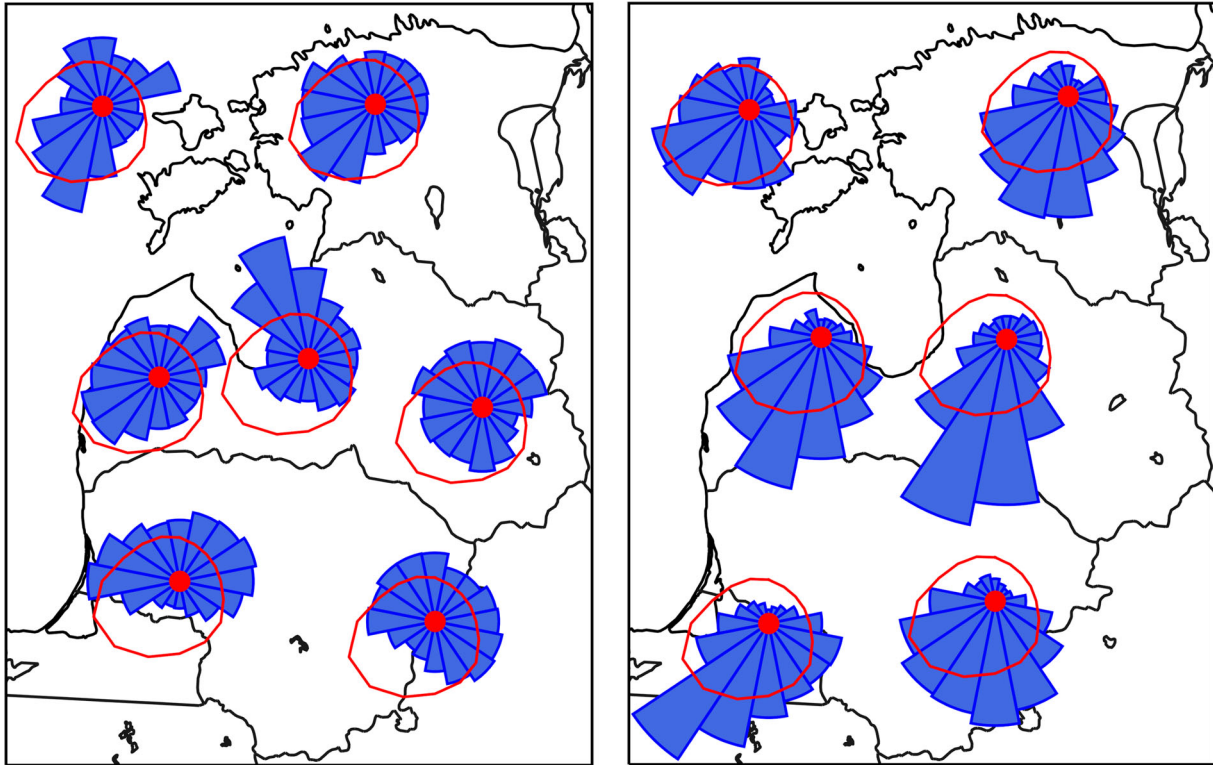


Fig. 9. Wind direction distribution in model data in May (left) and November (right) at the locations marked by red circles. Average yearly wind rose for the region of study is shown with red line.

is the well-known fact that wind changes its direction when flowing over a coastline due to changes in the surface roughness and the balance between the Coriolis

force, the pressure gradient force and surface friction near the surface. As the coastline of the Baltic proper is oriented in a roughly N-S direction, SW wind crossing it

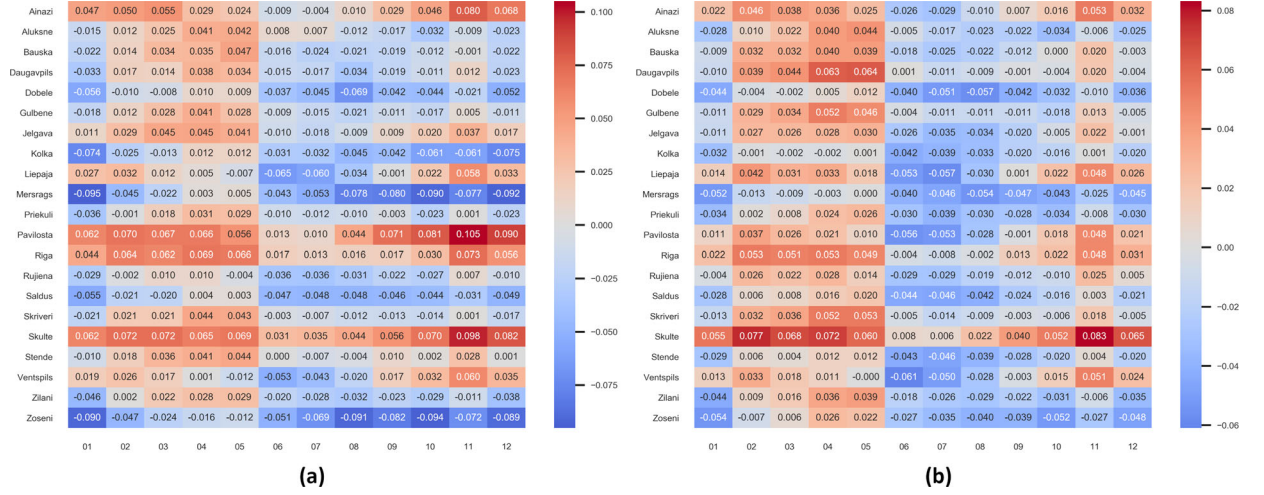


Fig. 10. Principal component analysis of (a) observation data, (b) model data in observational stations. Values of PC2 of observational data stations (vertical axis) throughout the year. (Months are shown on the horizontal axis).

would turn to the left, leading to additional southern winds over the dry land.

The situation is more complicated with the negative values of PC1 during the summer indicating an increase in northern winds. The literature describes the pressure differences that cause the SW flow in the synoptic scale to get weaker, leading to NW, W flows in the southern part of the Baltic Sea basin and flows lacking any specific direction in the northern part of the Basin (Team, 2008). This process was used again to explain the increased frequency of W and N winds seen in the Latvian observation stations by Briede (2016). While it is clear how weakening of the synoptic forcing could lead to a decrease in the frequency of SW winds and increase the winds from other directions, the prevalence of specifically N winds requires further research.

Figure 6 highlights some differences between the model and observations reflected in the PC1. First, although the loadings of PC1 are almost identical for both the full-grid and the observation-point model datasets, the observational dataset shows that the maximum of the positive value of PC1 is directly from the south as opposed to WSW seen in loadings for both principal components derived from the model data. More importantly, for the observation dataset the first component explains only 40% of the total variance, while the number is higher (45%) for the full-grid model dataset and significantly higher for the observation-points model dataset (55%). The difference between the observation data and the corresponding-points model dataset indicates that either the model significantly overestimates the strength of the synoptic pattern, or the effect of the large-scale synoptic pattern in each of the stations is weakened by other mesoscale flows not being well represented by the model.

Such flows can include sea and land breezes and flows associated with orography.

3.2.2. Interpretation of PC2 results. The analysis of the next principal component (PC2) differs significantly from the previous analysis. Loadings of PC2 (Fig. 6) show that observation stations used in this study only partially represent the climate of the wider study region because there are significant differences between PC2 directions between full-grid model and observation-points model dataset. Observation stations sample only the climate over dry land, while a significant part of model points are located over the Baltic Sea. Notice that positive values of PC2 for the model correspond to negative values of PC2 from the other two datasets.

Let's start by analysing the pattern seen in the observation and the observation-points model dataset. Its positive values show an increase in SE winds, which is anti-correlated with an increase in SW winds. It is quite similar for both of those datasets with slight differences, i.e. the observation dataset showing a small preference for SE direction over ESE. The most important difference is in the amount of variance explained. Complimentary for the results for PC1, here the situation is reversed – in observational data this pattern is more significant than in the model data at corresponding points, again indicating that the mesoscale phenomena have much greater effect on the synoptic scale in the observation data.

For most stations, values of PC2 (Fig. 10) are slightly positive (SE winds) in spring (Feb – May) and weakly negative (SW winds) for the rest of the year. The increased frequency of SW winds from June to January is in good agreement with the general understanding of SW winds as the most dominant in the region and being caused by synoptic scale processes (Team, 2008). The SW

winds are already the most pronounced in the average frequency (Fig. 5), but as already described when discussing the PC1, the frequency of SW winds can have a seasonal variation, which seems to be also reflected in the values of PC2. The weakening of synoptic scale pressure gradients in spring has been described in literature (Team, 2008). The prevalence of SE winds during the spring is much harder to explain, because, even if the synoptic scale pressure gradients are weakened, it is unclear what synoptic scale pattern should replace them. To get a better understanding of this phenomenon, let's analyze the properties of SE winds seen in observational data. The SE winds in most inland stations have very little diurnal cycle and the winds blowing from SE are comparatively light. This suggests the typical time scale of a synoptic process. A large anticyclone with the centre over northern Russia (the European part of Russia) could create a SE synoptic scale flow over Latvia.

For the sake of completeness, it should be mentioned that in some inland stations SE winds have a diurnal cycle, for instance, Priekule. We suspect that this is related to the orography, because of the location of such stations near the uplands, despite the fact that Latvia is relatively flat. We leave the investigation of this phenomenon as a task for further research.

However, in addition to the general pattern described above, another pattern can be observed: the largest absolute values of PC2 (both positive and negatives) are for coastal stations. Stations located on the eastern shore of both the Baltic Proper (Liepāja, Ventspils and Pāvilosta) and the Gulf of Riga (Rīga, Skulte, Ainaži) show distinctly positive values of PC2 both in spring and in late autumn. The only station on the west side of Gulf of Riga (Mērsrags) shows negative values for autumn and winter.

Analysis of the diurnal cycle of the coastal stations shows that the SE winds in May have a very pronounced diurnal cycle with SE-E winds being present mostly during the night, see Fig. 12. There is little to no diurnal cycle at the inland stations for SE winds in spring (not pictured). In the case of the coastal stations with the land to the east, SE winds are offshore winds. This, together with the timing seen in (Fig. 12), enables us to explain this pattern with a land breeze – mesoscale flow that is oriented from the shore into the sea when the air over the land is colder than the air over the sea. The prevalence of SE winds near the Gulf of Riga for the first half of the year in contrast to the SW dominated storm season in autumn has been previously reported in literature (Kouts (1998), as cited in Soomere and Keevallik (2001)).

Figure 10a shows the values of PC2 for the observational dataset. Figure 10b shows that for the observation-point model dataset the values are similar, but somewhat

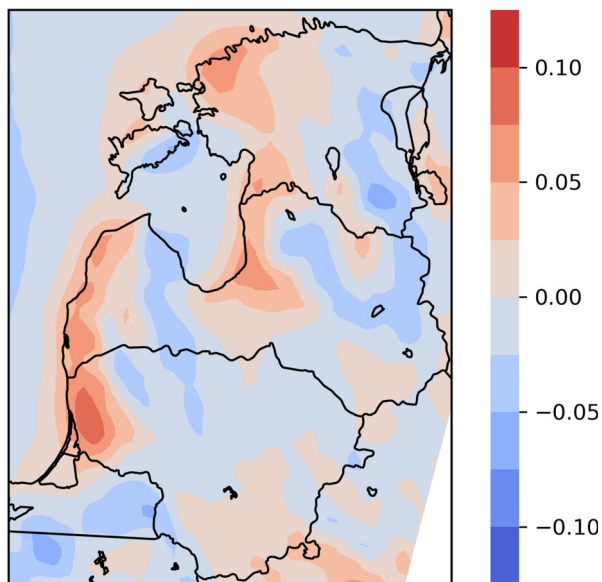


Fig. 11. Values of PC2 derived from observation data applied to the model data in November.

smaller in magnitude. The differences in the PC2 values between the model and the observations for the coastal stations in spring are related to the ability of the model to replicate the coastal flows, their impact on the diurnal cycle and the associated increase in the frequency of SE winds seen in observations. In some stations, e.g. Skulte, where the model is in agreement with the observations, the PC2 values are similar, but in other stations, e.g. Pāvilosta, the model fails to replicate the diurnal cycle and the significantly increased frequency of SE winds seen in observations.

In autumn, especially in November, the PC2 values indicate increased values of SE winds in coastal stations, however, in this season there is no obvious diurnal cycle. To explain this phenomenon it would be useful to see the spatial extent of such a pattern. Therefore, the question ‘where in the model can we see this pattern?’ was asked, using the procedure where loadings from the observational dataset are used together with the model dataset. This procedure is described in detail in the section titled ‘PCA Method and Its Application to Wind Direction Data’.

The results of this procedure can be seen in Fig. 11, where the increased frequency of SE winds is mostly associated with the same coastlines as analysed previously, but the model data show that this pattern extends to the territory of Lithuania and Estonia, but is limited to up to 20 km inland from the coastline. This result can again be linked with coastal phenomena. As in November the Baltic Sea is usually not frozen, this can lead to the sea being warmer than the land, which can result in offshore oriented flows. This is the same situation as during the

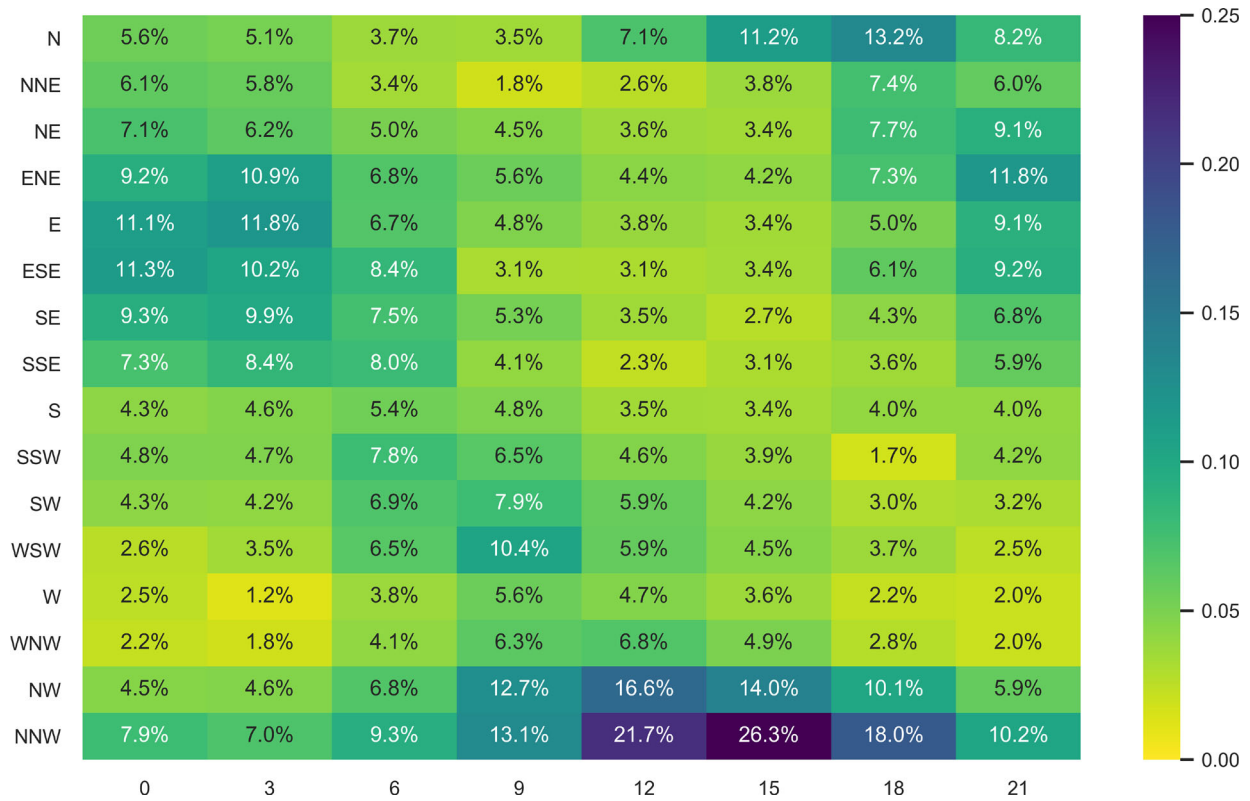


Fig. 12. Wind direction (y axis) probability distribution (shown with color) at the particular time of the day (x axis, UTC) in Skulte observation station, May, observational data.

land breeze that was described earlier. Similar arguments can be made about the increase in SW winds in Mērsrags station, except the model cannot replicate this pattern well.

To conclude, analysis shows that the increased pattern in SW or SE winds seen in the observations cannot be explained by a single phenomenon, but is a combination of different phenomena. It consists of (a) a seasonal cycle for most of the stations associated with large scale synoptic processes (explaining the small positive PC2 in Feb-May, negative for the rest of the year), (b) large values in spring for coastal stations, explained by winds that are present only during night time, can be associated with the land breeze and is only weakly represented in the model data and (c) large values in coastal stations during autumn, especially November, which is again linked with coastal mesoscale phenomena, probably associated with temperature differences between the land and the ice-free Baltic Sea.

As the PC2 loadings for the full-model dataset are significantly different from the observations and the observation-point model PC2, it can be concluded that the observation datapoints are not a representative sample of the wider study region, especially because the model domain includes a large part of the Baltic Sea and

observation data only from Latvia are analysed. As the PC2 is associated with coastal flows, this difference in PC2 can be explained by the fact that, first, the properties seem to be associated with the direction of the coastline and second, the ratio of coastal stations to the number of all stations for the Latvian observation network is much larger than the ratio of coastal grid-points vs all-grid points for the full dataset. The PC2 for the full-model dataset shows an increased frequency of western winds, anti-correlated with the much smaller increase in ESE winds. We leave the analysis of this PC2 as a task for the future.

3.2.3. Interpretation of PC3 results. Let's analyse the next principal component - PC3. High values of the third principal component correspond to an increased frequency of northern and south-western winds. High values of PC3 in the observation data occur between March and August in Ventspils. Model data show an increased value of PC3 between April and August in a strip over the Baltic sea, near the western coast of Latvia (Fig. 13). A recent study (Sīle et al., 2018) reported the presence of coastal jets in this region and time-frame, oriented from north to south, which can be explained by a thermal wind mechanism, usually known for creating large scale

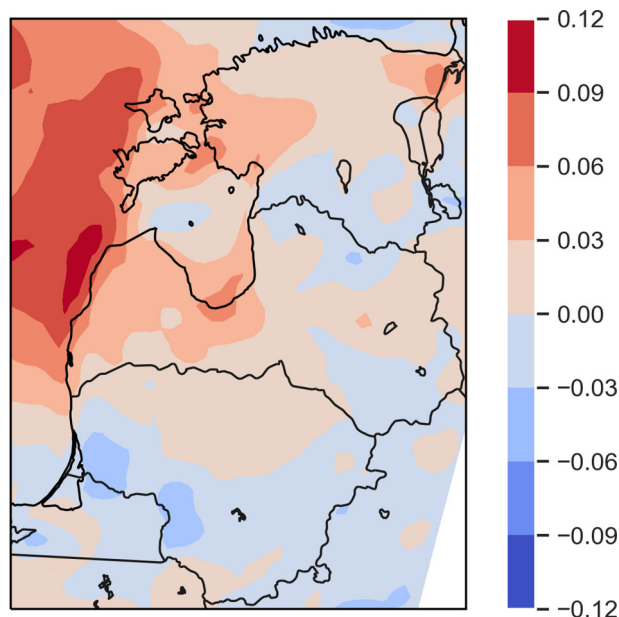


Fig. 13. Principal component analysis of model data. Values of PC3 in June. The highest values occur over the Baltic sea near the west coast of Latvia.

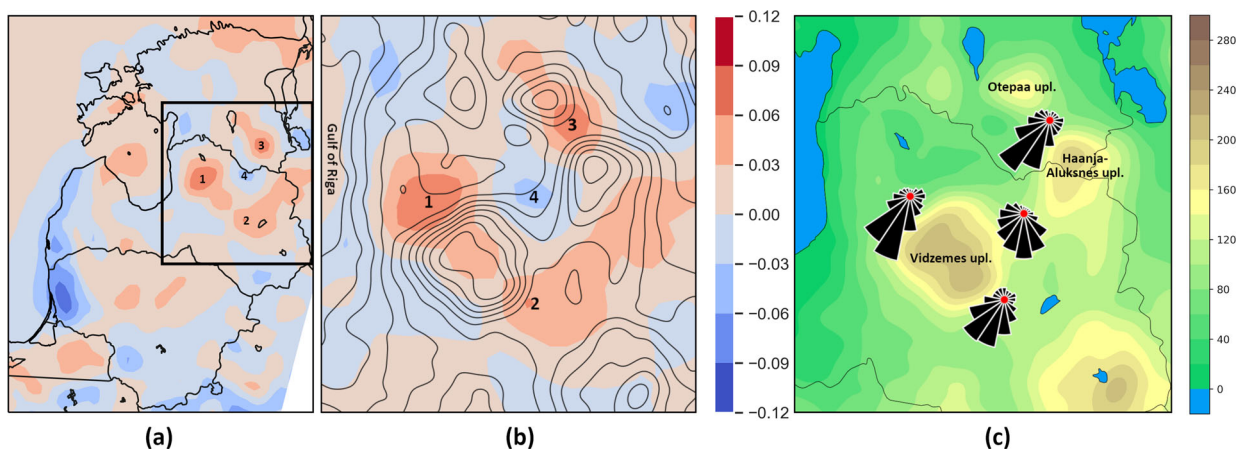


Fig. 14. (a) Principal component analysis of model data. Values of PC3 in December. (b) Enlarged boxed region from (a). Isolines of altitude (acquired from model data) spaced by 20 metres are shown with black lines. (c) Wind roses for December in points 1-4.

coastal low-level jets. It is plausible that these jets are the explanation for the increased values of PC3.

In the cold season a contrasting pattern of high and low values of PC3 is apparent in the Latvia and Estonia border region (Fig. 14a). An enlarged version of the plot of the PC3 values in the region together with orography is shown in Fig. 14b. The height of the terrain is 100 – 300 m above the sea level.

The wind flow in the cold season is typically from the southwest. High PC3 values are associated with regions between two uplands or locations that are on the sides of the uplands looking in the flow direction. Low PC3

values are located downstream from the hills. To better understand this pattern wind direction distributions for December in selected points of high and low values of PC3 were analysed (Fig. 14c). In points 1 and 2 (on the sides of the Vidzemes upland) and in point 3 (between the Otepaa upland and the Haanja-Aluksnes upland) the wind roses showed a distinct prevalence of SW and SSW wind direction, and had high values of PC3. In contrast, in location 4 (between the Vidzemes and the Haanja-Aluksnes upland), where the PC3 value is low, the wind roses were more isotropic, with winds from ESE-NW directions being almost equally likely with only slight

prevalence of SSW winds. Further investigation of this phenomena is outside the scope of this study, however, it is most likely linked to flow interaction with orography.

4. Conclusions

Long-term observational wind direction data series can contain observer bias, changes in observation methods and data resolution, as well as other inhomogeneities, not all of them documented. All of these inhomogeneities should be addressed before further use of the data.

Comparison of wind direction distributions between model data (UERRA reanalysis) and observations using EMD metric shows that the overall agreement between the model and the observations is good, with differences not exceeding 10 degrees. The largest differences are associated with coastal stations, especially in the cold season.

PCA is a useful tool for identifying the main properties of dataset features, however, its results must be interpreted cautiously, using other information. First, PCA cannot distinguish between different phenomena that coincidentally create an increase in the frequency of the same wind direction. In those cases, additional data must be used to disentangle the effects. This study additionally relies on the diurnal wind direction data, so a task for the future is to apply PCA to such a dataset.

The PCA method identifies an increased overall preference for SW winds which are a well-known characteristic of the region. The next major feature identified by the first PC is the preference for NW winds in the summer and additional S-SW winds in the winter. This feature is present in most observation stations and is more pronounced over the land than over the water. In observational data it is less pronounced than in the model data. It can be explained by seasonal changes in the synoptic scale pressure patterns. In the winter, the Icelandic low deepens and the Azores high and anti-cyclones over Russia are strong. This leads to large pressure gradients and pronounced SW flow. The situation is reversed in the warm season and the pressure gradient is weak. However, the specific reason for increased NW in the summer is a topic for further research.

Additional features identified by further principal components are associated with the coastline, i.e. PC2 shows an increase in offshore winds during the nights in spring, or the whole day in autumn. Features seen in PC3 indicate the presence and extent of low-level jets and features associated with orography but, however, these are less pronounced.

In conclusion, in this study PCA proved useful in identifying features of the wind direction climate and ranking them according to their prevalence. Such analysis could be easily generalised to other regions.

Author contributions

U.B. and J.S. conceived the study. M.P. wrote the code and created the figures. All authors analysed the results. M.P. and T.S. wrote the first draft of the manuscript. All authors contributed to the final version of the manuscript.

Data availability

Observations are available from the Latvian Environment, Geology and Meteorology Centre: <https://www.meteo.lv/en/meteorologija-datu-meklesana/?nid=924>.

UERRA reanalysis data are available from Copernicus Data Store: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-uerra-europe-single-levels?tab=overview>.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

T.S. acknowledges the financial support of the project 'Mathematical modelling of weather processes - development of methodology and applications for Latvia (1.1.1.2/VIAA/2/18/261)'. State Education Development Agency Republic of Latvia. The work of M.P. in this project was financially supported by the project 'Mathematical methods for research excellence (Y5.AZ76).

References

- Bethere, L., Sennikovs, J. and Bethers, U. 2017. Climate indices for the Baltic states from principal component analysis. *Earth Syst. Dynam.* 8, 951–962. <https://esd.copernicus.org/articles/8/951/2017/>. doi:10.5194/esd-8-951-2017
- Briede, A. 2016. *Klimats un ilgtspēja attīstība*. Latvijas Universitāte, Chap. 4., 78–81 isbn 978-9934.18-136-8.
- ECMWF. 2019. UERRA regional reanalysis for Europe on single levels from 1961 to 2019. Online at: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-uerra-europe-single-levels?tab=overview>.
- Gupta, A. and Dhir, A. 2013. Estimation of horizontal pollution potential by calculating impact area for Patiala, Punjab using wind data. *Int. J. Innov. Res. Sci. Eng. Technol.* 2, 2271–2279.
- Hahmann, A. N., Sīle, T., Witha, B., Davis, N. N., Dörenkämper, M. and co-authors. 2020. The making of the New European Wind Atlas - Part 1: Model sensitivity. *Geosci. Model Dev.* 13, 5053–5078. In: url: <https://gmd.copernicus.org/articles/13/5053/2020/>. doi:10.5194/gmd-13-5053-2020
- Jaagus, J. and Kull, A. 2011. Changes in surface wind directions in Estonia during 1966-2008 and their relationships with

- large-scale atmospheric circulation. *Estonian J. Earth Sci.* 60, 220. In: doi:10.3176/earth.2011.4.03
- Jaagus, J., Briede, A., Rimkus, E. and Remm, K. 2009. Precipitation pattern in the Baltic countries under the influence of large-scale atmospheric circulation and local landscape factors. *Int. J. Climatol.* 30, n/a–720. In: doi:.
- Jolliffe, I. T. 2002. *Principal Component Analysis. Springer Series in Statistics*. New York: Springer-Verlag. ISBN: 0-387-95442-2. <http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4>.
- Jungo, P., Goyette, S. and Beniston, M. 2002. Daily wind gust speed probabilities over Switzerland according to three types of synoptic circulation. *Int. J. Climatol.* 22, 485–499. In: doi:10.1002/joc.741
- Klink, K. and Willmott, C. 1989. Principal components of the surface wind field in the United States: A comparison of analyses based upon wind velocity, direction, and speed. *Int. J. Climatol.* 9, 293–308. doi:10.1002/joc.3370090306
- Koch, W. 2004. Directional analysis of SAR images aiming at wind direction. *IEEE Trans. Geosci. Remote Sens.* 42, 702–710. doi:10.1109/TGRS.2003.818811
- Kouts, T. 1998. *Forcing Factors for Hydrography and Currents – Meteorological and Hydrological Variables*. In: The Gulf of Riga Project (funded by The Nordic Council of Ministers). SMHI, Norrköping, Sweden.
- Lea, D. A. and Helvey, R. A. 1971. A directional bias in wind roses due to mixed compass formats. *J. Appl. Meteor.* 10, 1037–1039. In: Oct. > 2.0.CO;2. url: https://journals.ametsoc.org/view/journals/apme/10/5/1520-0450_1971_010_1037_adbiwr_2_0_co_2.xml. doi:10.1175/1520-0450(1971)010<1037:ADBIWR>2.0.CO;2
- Met Office. 2010. *Cartopy: A Cartographic Python Library with a Matplotlib Interface*. Exeter, Devon, 2010–2015. <https://scitools.org.uk/cartopy>.
- Mezaache, H., Bouzgou, H. and Raymond, C. 2016. Kernel principal components analysis with extreme learning machines for wind speed prediction. In *Seventh International Renewable Energy Congress, IREC 2016*. Hammamet, Tunisia, March. <https://hal.inria.fr/hal-01394000>.
- Pedro, A. J., González-Rouco, J. F., Montávez, J. P., García-Bustamante, E., & Navarro, J. 2009. Climatology of wind patterns in the northeast of the Iberian Peninsula. *Int. J. Climatol.* 29.4, 501–525. doi:.
- Pele, O. and Werman, M. 2008. A linear time histogram metric for improved sift matching. In: *Computer Vision – ECCV*. Springer, Berlin, Heidelberg, Oct., pp. 495–508.
- Pele, O. and Werman, M. 2009. Fast and robust earth mover's distances. In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. Sept., pp. 460–467.
- Ratner, B. 1950. A method for eliminating directional bias in wind roses. *Mon. Wea. Rev.* 78, 185–188. 0493(1950)078<0185:AMFEDB>2.0.CO;2. url: https://journals.ametsoc.org/view/journals/mwre/78/10/1520-0493_1950_078_0185_amfedb_2_0_co_2.xml. doi:10.1175/1520-0493(1950)078<0185:AMFEDB>2.0.CO;2
- Rubner, Y., Tomasi, C. and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Computer Vision.* 40, 99–121. url: <https://doi.org/10.1023/A:1026543900054>. doi:10.1023/A:1026543900054
- Rutgersson, A. 2015. Recent change | atmosphere. In: *Second Assessment of Climate Change for the Baltic Sea Basin* (ed. The BACC II Author Team). Springer International Publishing, Cham, pp. 69–97. ISBN: 978-3-319-16006-1. 1007/978 - 3 - 319 - 16006 - 1 _ 4. url: https://doi.org/10.1007/978-3-319-16006-1_4.
- Sepp, M., Post, P., Mändla, K., & Aunap, R. et al. 2018. On cyclones entering the Baltic Sea region. *Boreal Environ. Res.* 23, 1–14.
- Sīle, T., Seņņikovs, J. and Bethers, U. June 2018. Evidence for low-level jets caused by coastal baroclinity at the Kurzeme shore of the Baltic Sea. *Estonian J. Earth Sci.* 67, 149. In: doi:10.3176/earth.2018.11
- Soomere, T. and Keevallik, S. 2001. Anisotropy of moderate and strong winds in the Baltic Proper. *Proc. Estonian Acad. Sci. Eng.* 7, 35–49.
- Team, B. A. 2008. *Assessment of Climate Change for the Baltic Sea Basin*. Springer, Berlin Heidelberg. isbn: 978-3-540-72786-6. url: <https://doi.org/10.1007/978-3-540-72786-6>.
- WMO. 2017. *WMO Guidelines on the Calculation of Climate Normal*. 2017th ed. World Meteorological Organization (WMO). isbn: 978-92-63-11203-3.