Nearest Neighbor Techniques for the global and regional statistical analysis of high temporal resolution CMIP6 surface wind speed with ECMWF reanalysis and satellite winds.

Ponni. MAYA[*1] and George. LAVIDAS [*1]

[*1] Department of Hydraulic Engineering, Delft University of Technology
Mekelweg 5, Delft, 2600AA, Netherlands

Abstract

Regional energy studies should include the effects of climate change in order to improve economic, social, and environmental development. Ocean surface winds have an important role in air-sea interaction, wave modelling, and renewable energy assessment and development. Taking into consideration, there is research carried out for wind speed studies in various regions all around the globe.The IPCC 6th report suggests the high capturing capability of climate models in the newly developed Coupled Model Intercomparison Project version 6 (CMIP6) with CMIP5. This study focuses on historical simulation analysis from CMIP6. The goal is to quantify the variation of 10m height surface winds over the globe and specific regions. The first realization(r1i1p1f1) of all the model simulations available for the CMIP6 at the time of investigation is taken into consideration. The work includes 3 hourly Eastward surface winds (‘uas’; variable in CMIP6) and Northward surface winds (‘vas’; variable in CMIP6) at 10m height from CMIP6. The climate model data are validated by comparing with the corresponding European Center for Medium-Range Weather Forecast (ECMWF) reanalysis wind components, satellite scatterometers, and radiometers datasets. Moreover, the work focuses on the selection of the pre-eminent CMIP6 model historical simulation which is best for reproducing the ‘real datasets’. The regional study is focusing on Japan. The change in seasonal and annual variation in wind speed is carried out with different statistical methods. An optimal machine learning approach is introduced to find the best climate model wind for each ocean basin.

Key words : Energy, Wind, CMIP6, ECMWF, Scatterometers, Radiometers, CMEMS.

1. Introduction

Renewable energy resources are the basic components of social, economic, and environmental development (Danial Khojasteh et al, 2018). Global energy consumption over the last two decades has also increased drastically due to various factors, such as climate change, economic recession, and the reduction of natural resources. This has caused an increase in the demand for renewable energy production in global continents, which would help to mitigate future climate change. Moreover, the EU Commission considers energy efficiency policies to combat the European energy crisis (Calanter P et al, 2022).

Intraseasonal, interannual, decadal, and multidecadal Climate Change studies have been on the rise due to the impact of climate change in different sectors such as energy resources, health, infrastructure, natural hazards, and agriculture (Zittis G et al, 2022). Adaptation and mitigation plans against climate change are necessary for protecting human beings and retaining development. The Intergovernmental Panel for Climate Change (IPCC) includes a long record of climate variables' accretion for climate studies (Zelinka MD et al, 2020).

Within the framework of the IPCC, the Coupled Model Intercomparison Project (CMIP) provides a collaborative platform to access and improve the knowledge of climate change irrespective of the past, present, and future. The 6th

[*] Date of Manuscript Acceptance 2023.1.11

E-mail of corresponding author: p.maya@tudelft.nl

phase of the WCRP Working Group Coupled Modelling (WGCM) on the Coupled Model Intercomparison Project (CMIP6) gives variable specific output data, which has proven to have a better performance than CMIP5 (Yasin Zamani et al. 2019, WCRP). CMIP6 represents the most recent generation of climate models, which are now available. The enormous application of these models leads to multiple statistical comparative studies on different temporal and spatial distributions (Martinez A,, 2022), which are time-consuming (suggest factors that relate ocean surface winds to statistical comparative studies). However, ocean surface winds play a key role in air-sea interaction, wave modelling, and renewable energy assessment and development (Schulz J et al, 2022). Thus a time-efficient analysis is necessary in order to finalise the model outputs which needed much effort from various statistics.

This work introduces decision tree nearest-Neighbour algorithms for the optimised selection of Climate model datasets which are approximately clones. These include K nearest neighbours (KNN) and Decision tree (Song T et al, 2022). This space partitioning data structure algorithm is used to organise points in multi-dimensional space to avoid brute force. For example, testing of the algorithm in different climate model winds at a 10 m height above the sea level are used in this study due to their complex nature.

2. Data and Methodology

2.1 Data

2.1.1 CMIP6 Models

To find the algorithm’ s ability and performance, all the models under specific criteria are taken into account at the time of investigation. The respective models are represented in Table 1. Historically, the available 11 model datasets under the first realisation (*r1i1p1f1*) for particular variables are taken into account. The key variables used for this study are zonal and meridional wind components(‘uas’ and ‘vas’: variable representation in CMIP6). The 3 hourly 10-metre surface winds are preprocessed before each analysis.

After merging the wind components in time, the Eastward surface winds (‘uas’; variable in CMIP6) and Northward surface winds (‘vas’; variable in CMIP6), measured at a height of 10 m from the CMIP6, and the ‘eastward wind’ and ‘northward wind’ from CMEMS satellites, are regridded to a 1 degree x 1 degree (latitude and longitude) grid through bilinear interpolation. This is implemented in the Climate Data Operator (CDO). Every data is set into a uniform calendar using CDO. The leap days are removed from the 11 CMIP6 model data and resampled to 6 hourly for comparison against satellites. For Reanalysis datasets (Era 5) used in this study from ECMWF(section 2.1.3) is preprocessed using the same reggriding procedure and setting the calendar with 3 hourly temporal resolution after preprossessed .

2.1.2 Satellite Merged Dataset

The CMEMS (Copernicus Marine Environment Monitoring Service) surface winds, derived from scatterometers and radiometers, are used for the comparison with the results from this study. These represent “real-time” measurements. The optimised satellite blended data is produced from CMEMS IFREMER (ERS-1 and ERS-2), NASA/JPL (QuikSCAT and RapidScat), EUMETSAT OSI (ASCAT-A and ASCAT-B), CNSA (HY-2A), ISRO (OceanSat-2), from Remote Sensing System (SSM/I SSMIS, and WindSat) and NWP reanalysis. This has a 10 m 6 hourly averaged eastward, and northward wind 10 m components for the period 1992-2014 is ideal for the validation of CMIP6 winds, and is used in this study.

2.1.3 Atmospheric Reanalysis

In order to verify the results obtained by the satellite comparison, all the models are compared against the corresponding European Center for Medium-Range Weather Forecast(ECMWF) reanalysis wind components for the period of 1992-2010. Reanalysis datasets are used to evaluate the climatological error anomalies and assess the progress in modelling capabilities in various fields and observational system transitions. The fifth generation of atmospheric reanalysis (Era5), produced by ECMWF, is used in this study. The Era5 has significantly improved quality and quantity compared to previous global reanalysis (Hersbach H et al, 2020).

Table 1 CMIP6 Model and Institution

Model	**Institution**
IPSL-CM6A-LR	Institut Pierre-Simon Laplace, Paris 75252, France
GFDL-ESM4	NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08540, USA
MPI-ESM1-2-LR	Max Planck Institute for Meteorology, Hamburg 20146, Germany
AWI-ESM-1-1-L R	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research,, Germany
MIROC6	Japan Agency for Marine-Earth Science and Technology
IPSL-CM5A2-INCA	Institut Pierre Simon Laplace, Paris 75252, France
MPI-ESM-1-2-HAM	ETH Zurich, Switzerland
GFDL-CM4.gn	NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08540, USA
GFDL-CM4.gr	NOAA, Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08540, USA
CMCC-CM2-SR5	Fondazione Centro Euro- Mediterraneo sui Cambiamenti Climatici, Lecce 73100, Italy
CMCC-ESM2	Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Lecce 73100, Italy

2.2 Methodology

2.2.1 Cosine Similarity

In huge data analysis, cosine similarity is a widely used method to find the measure of similarity between two vectors in an inner product space. The calculation of cosine similarity is represented in equation(1),

$$cos(x,y) = \frac{\frac{<x \cdot y>}{\|X\| \cdot \|y\|} \sum_{i=0}^{n-1} X_i y_i}{\sqrt{\sum_{i=0}^{n-1} \left(X_i\right)^2} \times \sqrt{\sum_{i=0}^{n-1} \left(Y_i\right)^2}} \tag{1}$$

Where

X_i= component of the satellite vector

y_i= component of the model vector

2.2.2 Bias Error

Bias is calculated in each grid point over the selected domain by taking normal error represented in the equation (2)

$$\text{Bias} = \text{Model(i)} - \text{Satellite(i)} \tag{2}$$

2.2.3 Euclidean Distance

The length of the line segment between each point for the model and corresponding with “real datasets” are represented by the Euclidean distance equation(3).

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \left(y_i - x_i\right)^2} \tag{3}$$

2.2.4 Chebyshev Distance

The chessboard distance is (Yongjian et al. 2021) the distance between model and “real measurements” datasets, defined as the greatest of their differences along any coordinate dimension in the vector space.

$$d_{12} = \lim_{m \to \infty} \left(\sum_{i=2}^{n} \left| x_{1i} - x_{2i} \right|^m \right)^{\frac{1}{m}} \tag{4}$$

2.2.5 K Dimensional Tree (KD tree)

KDtree is an effective combination of both KNN and a Decision tree, to structure data in multi-dimensional space. In order to avoid brute forcing, KD Tree works internally to construct the data points in a decision tree-like structure. The structure involves the root node extended downward through branches until it reaches the parental node. This is the widely used approach for the classification and grouping purpose of data sets (Bahzad Taha Jijo et al 2021). In this case, it works as a binary search tree where data in each node works as a k-dimensional data point in space.

3. Result and discussion

3.1 Bias Error between the historical CMIP6 models and merged satellite product

The Bias error for all the 11 CMIP6 models is compared against CMEMS datasets for a temporal resolution for 6hrs over the global ocean basin for a period of 1992-2014 as shown in Fig(1a,1b,1c, and 1d) for the Japan region (latitude 26.1,63.3, longitude =122,157.9). Each figure represents different seasons (DJF- December January February) (MAM- March April May)(JJA- June July August) (SON- September October November). The 11 CMIP6 models encounter a maximum range of bias from 8m/s to a minimum bias range of -8m/s for wind magnitude(Figure 1a,1b,1c, and 1d)).

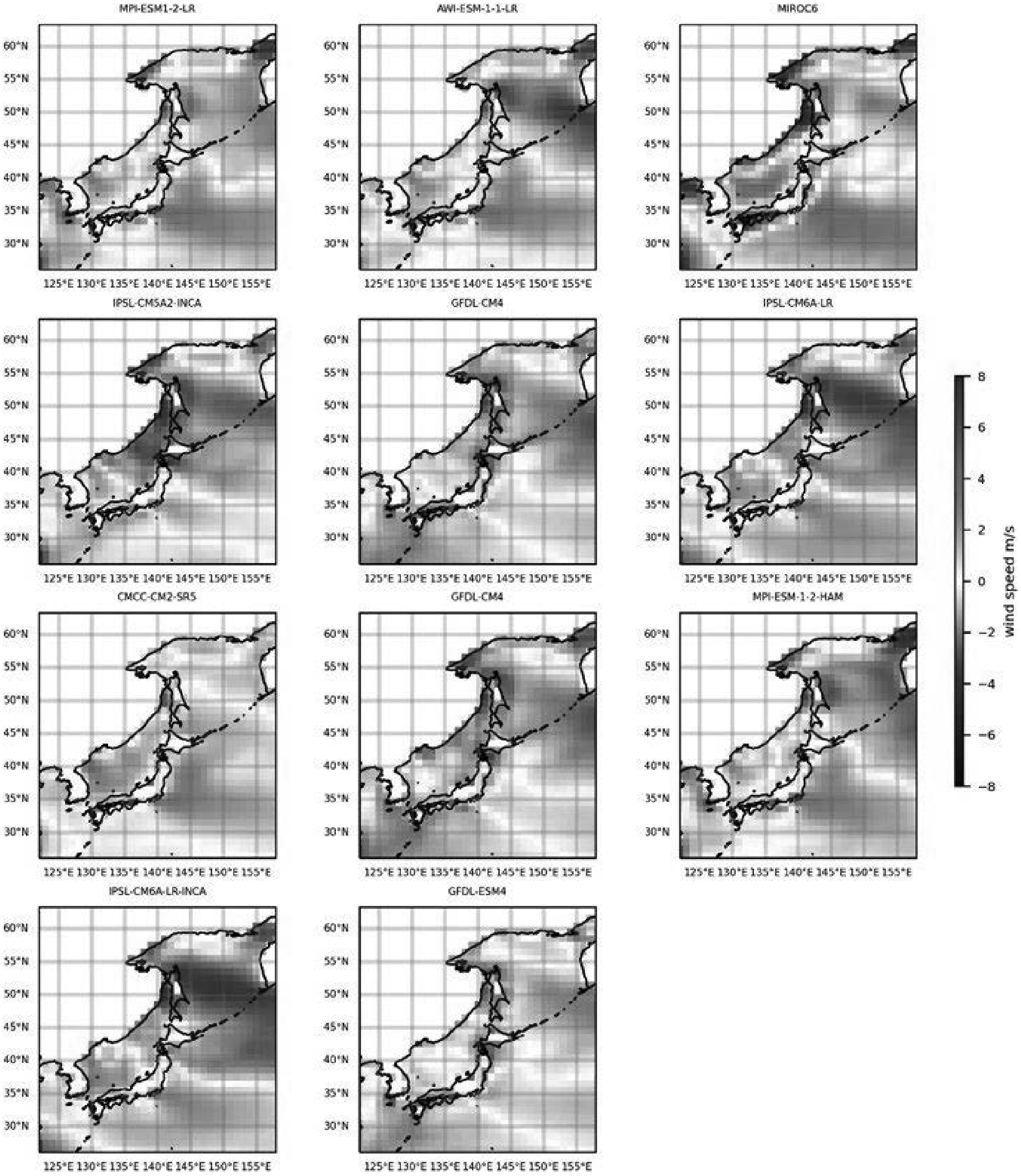

Fig. 1a The bias of DJF wind magnitude(in m/s, shaded) using CMEMS and CMIP6 wind datasets for the period 1992-2014. The corresponding model names are represented above each figure. Before computing all the data points are regridded to a 1 degree x 1 degree (latitude and longitude) grid through bilinear interpolation. set into a uniform calendar using CDO. The leap days removed 11 CMIP6 model data is resampled to 6 hourly for comparison against the satellite merged CMEMS dataset.

For DJF(figure 1a), the maximum positive bias is shown for the models MPI-ESM1-2-LR, AWI-ESM-1-1-LR, MIROC6, MPI-ESM-1-2-HAM, CMCC-CM2-SR5, IPSL-CM6A-LR. The models included in the same family from the Geophysical Fluid Dynamics Laboratory underestimates the satellite product in the selected domain for JJA.11 CMIP6 models overestimate the season MAM for the period 1992-2014(Figure 1b). MPI-ESM1-2-LR, AWI-ESM-1-1-LR, and MPI-ESM-1-2-HAM have a high deviation in the bias for MAM. A higher underestimation of wind magnitude for MIROC6 is shown for the season SON.

In the seasonal analysis over time for 22 years the models arise with contrast discrepancies among each other. The selection of a model among the 11 CMIP6 is harder under these circumstances of results. According to the literature so far multiple statistical analyses are put forward to figure out the best model out of the highly varying datasets for each temporal and spatial resolution.

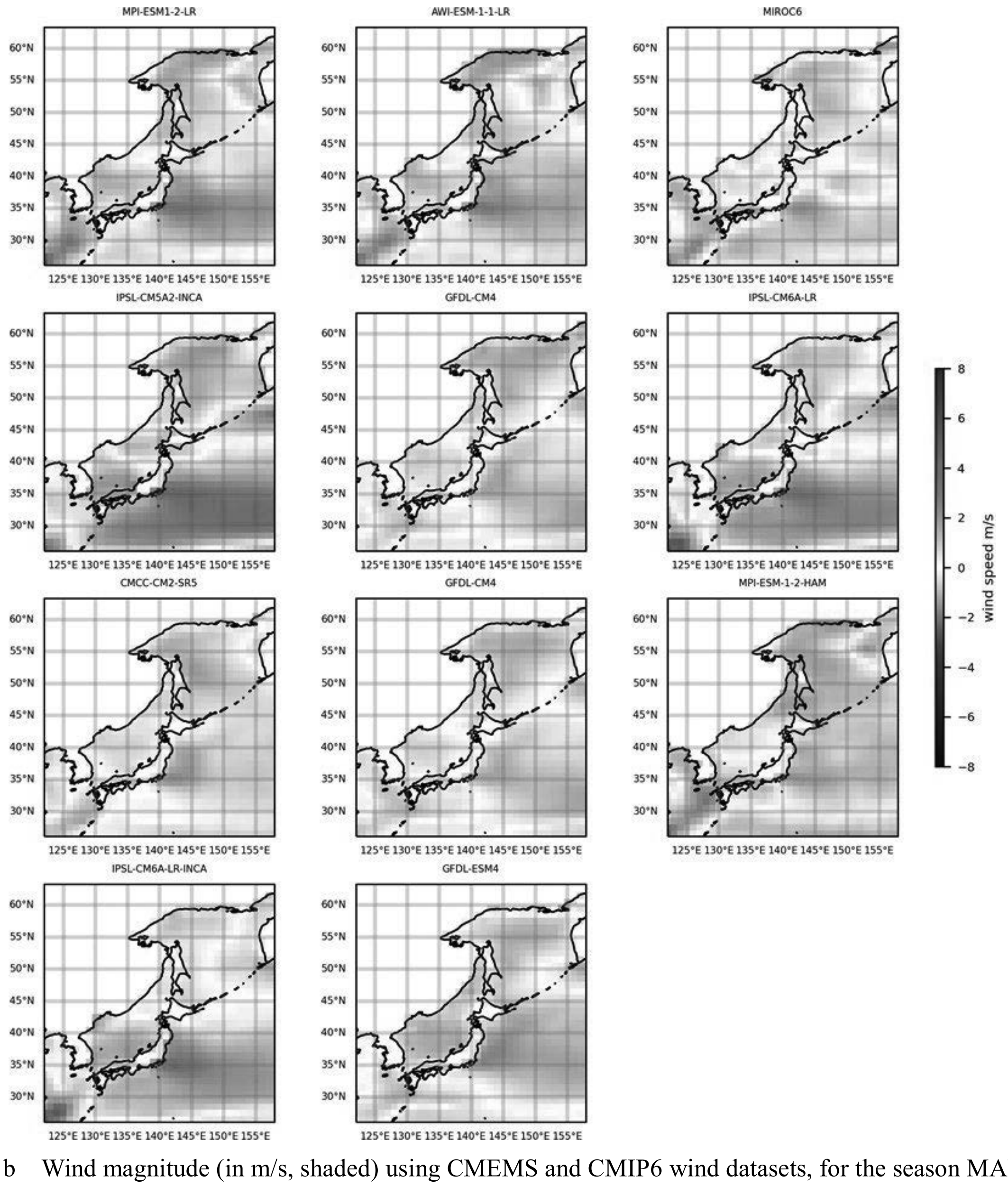

Fig. 1b Wind magnitude (in m/s, shaded) using CMEMS and CMIP6 wind datasets, for the season MAM for the period 1992-2014. The bias (m/s) is individually obtained and it is shaded in each figure.

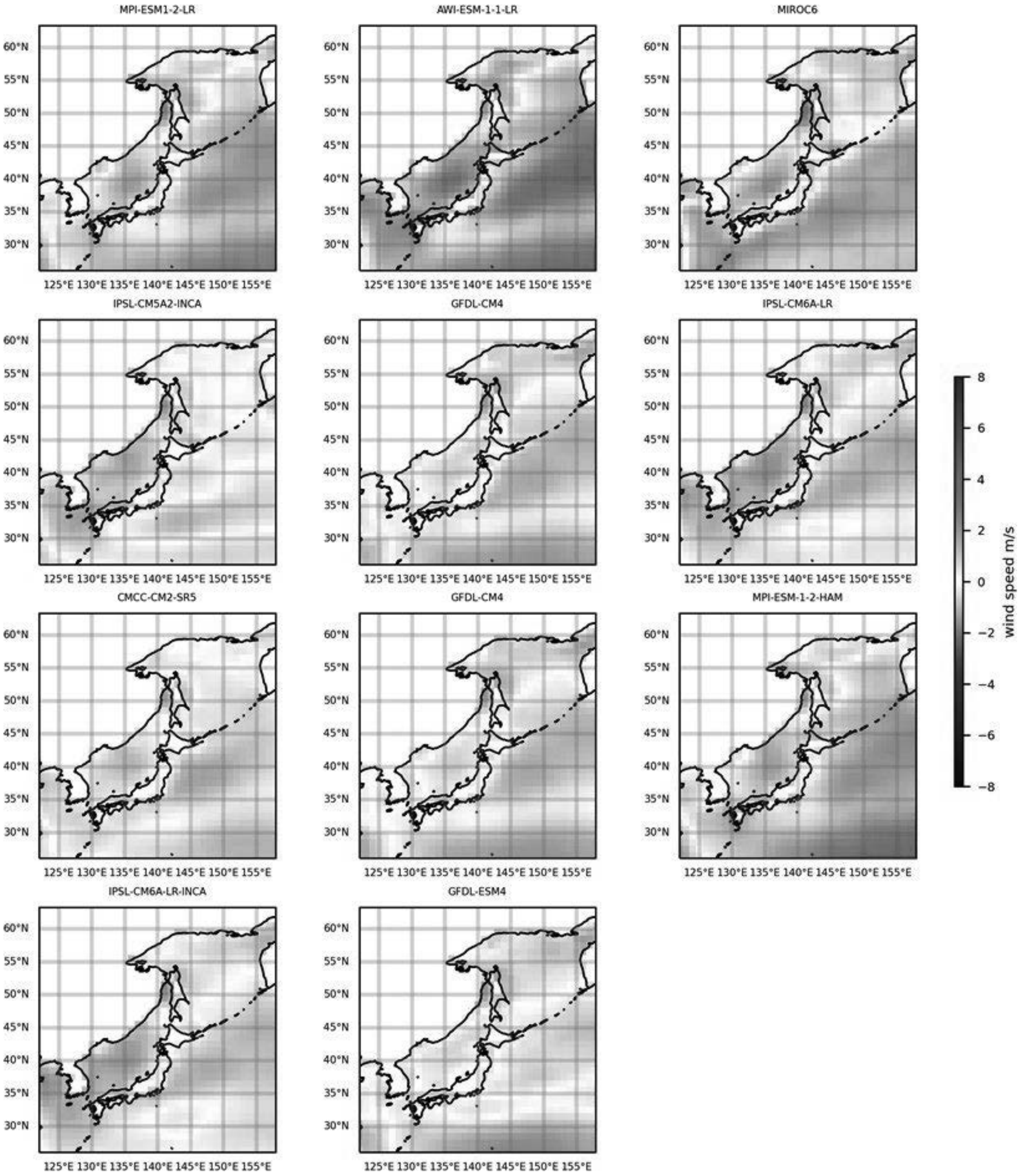

Fig. 1c same as Figure 1a, for the season JJA for the period 1992-2014. The bias (m/s) is individually obtained and it is shaded in each figure.

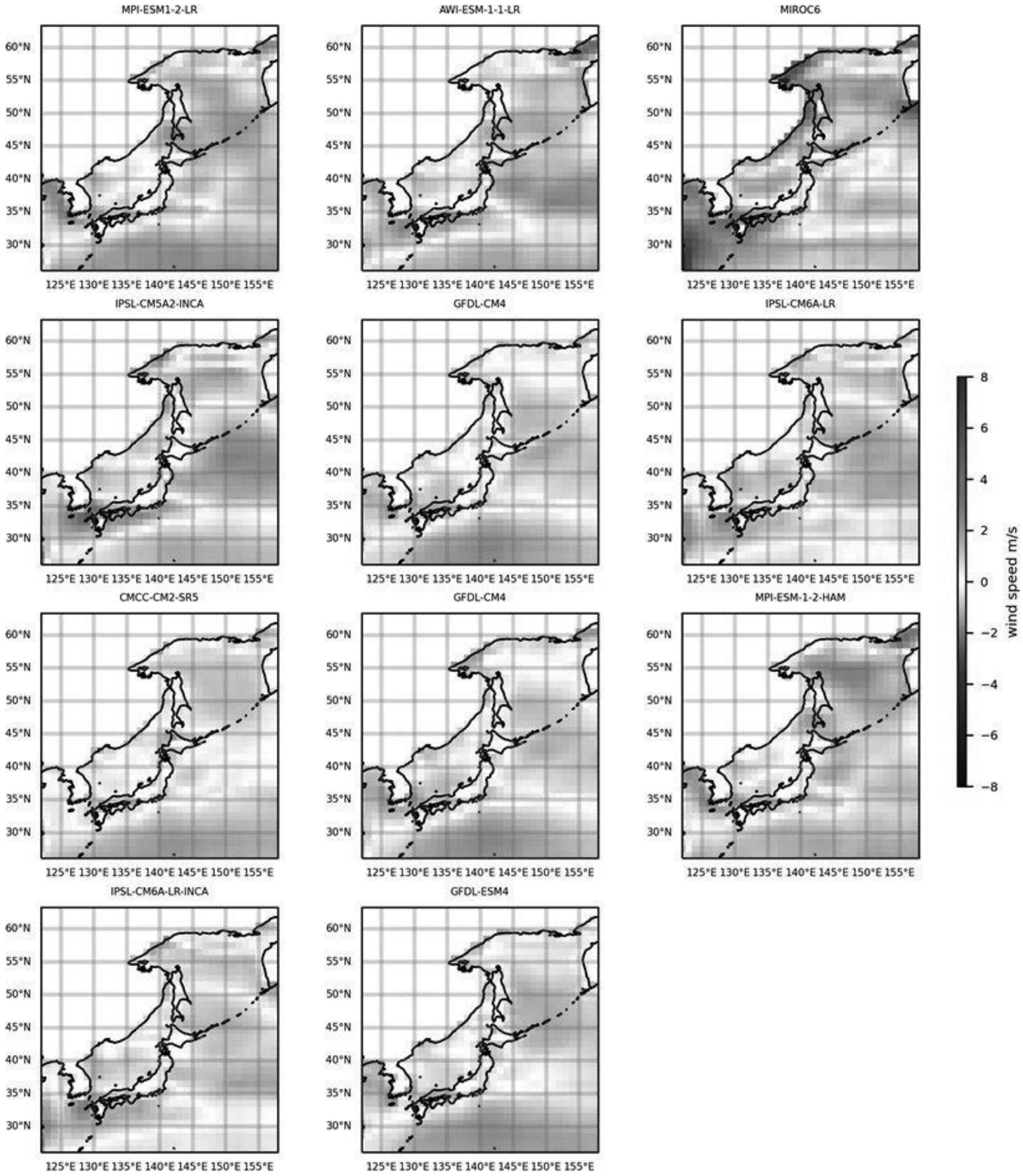

Fig. 1d same as Figure 1a, for the season SON for the period 1992-2014. The bias (m/s) is individually obtained and it is shaded in each figure.

3.2 Cosine similarity(CS) between the historical CMIP6 models and merged satellite product for all the seasons over a period of 1992-2014 for the Global scale.

The research, including the assessment of the CMIP6 datasets, is analysed using the measuring degree concept of Cosine similarity (CS) (Figure 2). The value obtained for all the seasons for the cosine similarity is in a range of 0.976 to 0.987, representing the two vectors pointing in the same direction. Even though the CS values show a high similarity between the CMIP6 models and Satellite data merged products in the inner product space, the difficulty arises for the conclusion in selecting the best model out of the selected CMIP6 models.

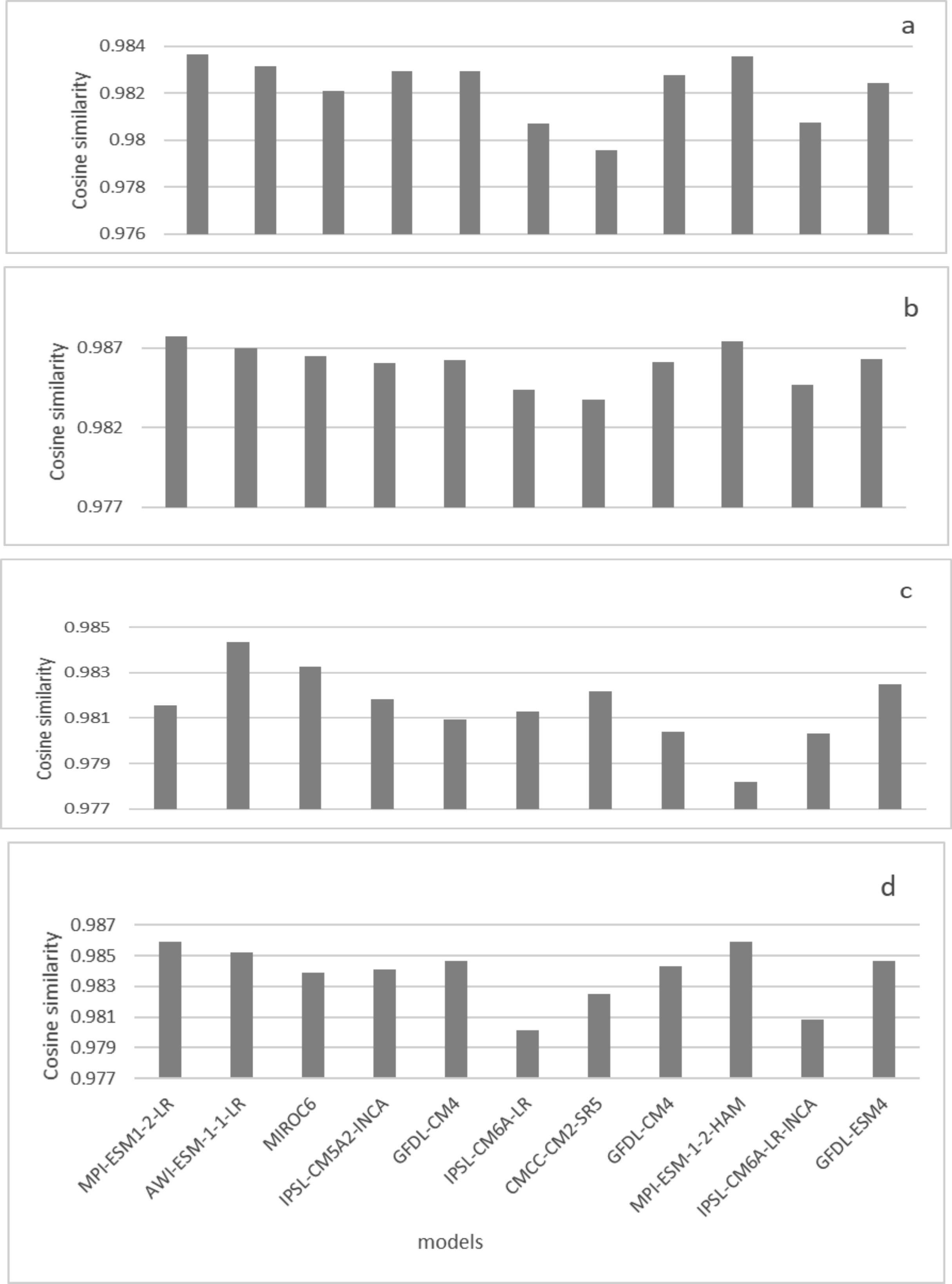

Fig. 2 Cosine similarity of wind magnitude (in m/s) of 11 CMIP6 datasets and CMEMS satellite datasets at 10m above sea level for all the seasons over the period 1992-2014. (a)- DJF, (b)- MAM, (c)- JJA, (d)- SON.

3.3a KDtree Euclidean distance calculation between the historical CMIP6 models and merged satellite product

In this study, the K nearest neighbour is used as a classifier by using the Euclidean distance and Chebyshev distance algorithms. For each new input, the KD tree figured out the nearest neighbour on the basis of Euclidean distance (Figure 3.3a and Figure 3.3b) and Chebyshev distance (Figure 3.3c and Figure 3.3d). Here for the 11 models, the one which is closer to the query has a lesser Euclidean distance than the model it is testing, and so on. Of all the machine learning algorithms, the KNN is one of the widely used ones, because of its easily understandable design and adaptable nature(Shahadat Uddin et al 2022).

The accuracy of the KNN depends on the quality of the data. In this case, we have high-quality datasets due to the structured preprocessing for both the training and query datasets. For large datasets, the prediction stage has a slowing effect and requires high memory for the storage of training datasets, here we use decision tree techniques (KD tree) to make the process faster. The KNN method implemented in the KDTree does not need a training period, because of its simple and powerful nonparametric architecture.

In this case, the value of K is 11, which helps to find the 11 nearest neighbours for each query point. Thereby, we can quantify the difference among them in the Euclidean distance. Similarly, the case is implemented in the calculation of the analysis of the Chebyshev distance (Figure 3.3c and Figure 3.3d).

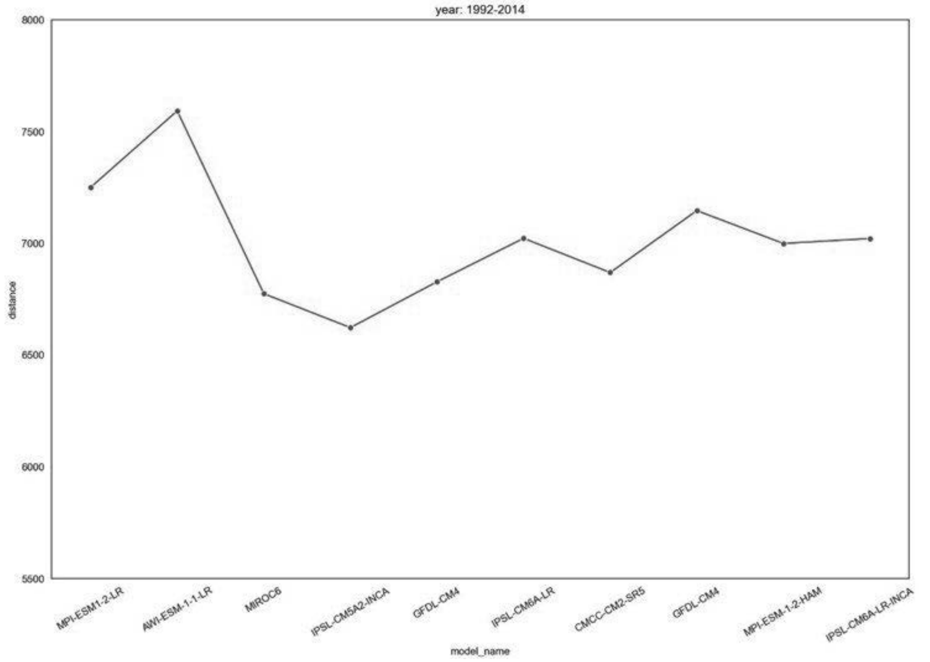

Fig. 3a Kdtree Euclidean distance calculated (m/s) between CMIP6 datasets and CMEMS satellite merged dataset for the period 1992-2014. The points represent each distance of the respective model.

3.3b Kdtree Euclidean distance calculation between the historical CMIP6 models and Era5

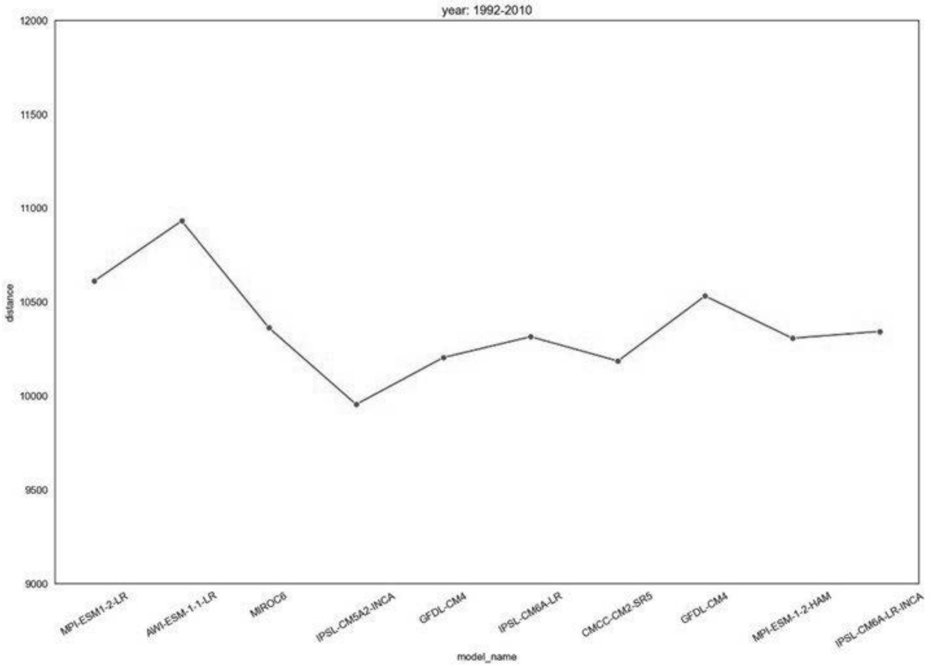

Fig. 3b Kdtree Euclidean distance calculated(m/s) between CMIP6 datasets and Era5 Reanalysis dataset for the period 1992-2010. The points represent each distance of the respective model.

3.3c Kdtree Euclidean distance calculation between the historical CMIP6 models and merged satellite product

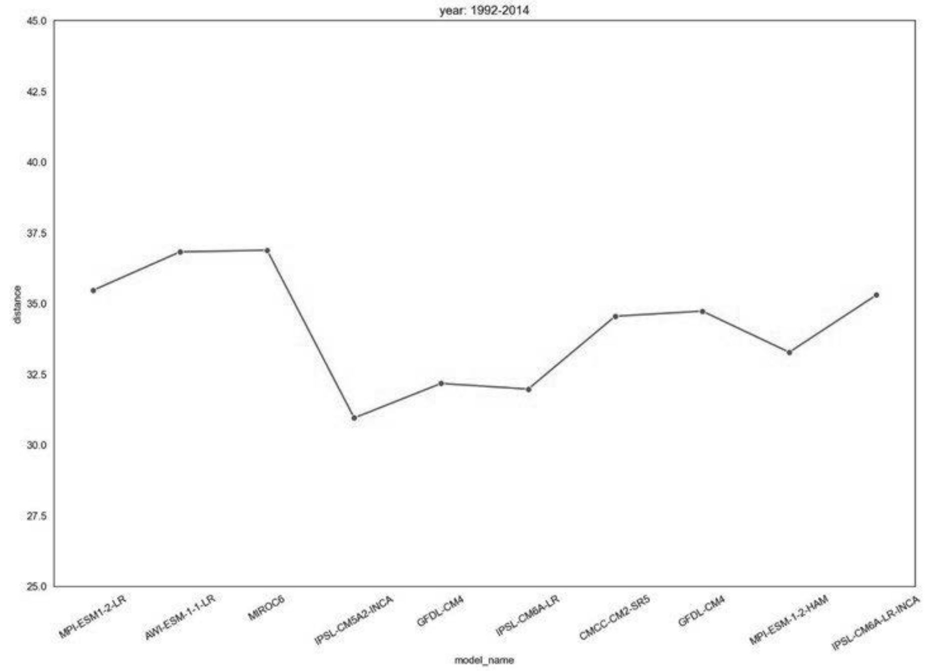

Fig. 3c Kdtree Chebyshev distance calculated(m/s) between CMIP6 datasets and CMEMS satellite merged dataset for the period 1992-2014. The points represent each distance of the respective model.

3.3d Kdtree Euclidean distance calculation between the historical CMIP6 models and Era5

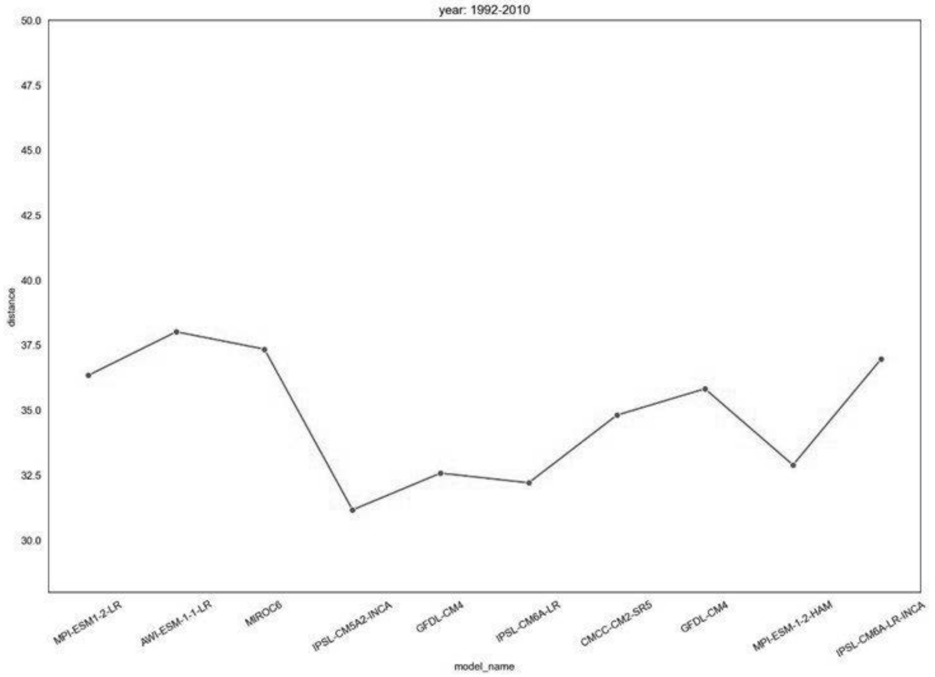

Fig. 3d Kdtree Chebyshev distance calculated(m/s) between CMIP6 datasets and Era5 Reanalysis dataset for the period 1992-2010. The points represent each distance of the respective model.

4. Discussion and Conclusion

This study uses the wind magnitude from historical CMIP6 high temporal resolution (3 hourly) and 1 degree x 1 degree (latitude and longitude) re-gridded spatial data for comparison in Japanese coast against "real datasets". The domain experiences anomalous thermal advection, energetic mesoscale oceanic eddies, and vigorous air-sea interaction, which can affect East Asia (Bunmei Taguchi (2009), Wenyu Zhou (2017). The "real datasets" mentioned includes Era5 and CMEMS. To study the similarity and differences between each model dataset, statistical methods such as Normal Bias and Cosine similarity are tested for all the seasons for a period of 1992-2014. The spatial bias of CMIP6 with CMEMS represents a series of complicated underestimation and overestimation through out different seasons. Even though the cosine similarity for the global regional analysis represents an exemplary result range between 0.977 and 0.978 arriving to a conclusion about the best CMIP6 model is still a challenging task.

For each region, it is well-known that each temporal resolution selection of a model by doing multiple statistics is time consuming. In order to fulfil that gap, a K Dimensional tree is introduced in this research. This method actively searches all the relevant data points in the multidimensional space against the "real measurements" , and denotes the basis of distance calculation. Euclidean distance and Chebyshev distances are calculated for each model dataset, and noted in the Figures 3a, 3b, 3c, and 3) respectively. The model IPSL-CM5A2-INCA has the least Euclidean distance and Chebyshev distance for the years(1992-2014) for CMEMS and (1992-2010) for Era5. The model IPSL is chosen for further studies in the selected regions. Irrespective of the area of interest and the temporal resolution, this nearest neighbouring search(NNS) - which is a combination of KNN and Decision tree techniques - gives the most appropriate model out of the clones, which is very difficult to group, and hence, quantify otherwise.

5. Acknowledgements

The research was a part of the PhD project funded by the Delft University of Technology and the Marine Renewables Energies Lab. The CMIP6 datasets used in the study are archived on the website of IPCC. The Era5 dataset from ECMWF and CMEMS from Ifermer. The authors express the acknowledgement to all the data providers mentioned above.

References

Danial Khojasteh, Davood Khojasteh, Reza Kamali, Asfaw Beyene, Gregorio Iglesias,Assessment of renewable energy resources in Iran; with a focus on wave and tidal energy,Renewable and Sustainable Energy Reviews,Volume 81, Part 2,2018.Calanter P, Zisu D. EU Policies to Combat the Energy Crisis. Global Economic Observer. 2022;10(1):26-33

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz ‐ Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society. 2020 Jul;146(730):1999-2049.

Jijo BT, Zeebaree SR, Zebari RR, Sadeeq MA, Sallow AB, Mohsin S, Ageed ZS. A comprehensive survey of 5G mm-wave technology design challenges. Asian Journal of Research in Computer Science. 2021;8(1):1-20.

Martinez A, Iglesias G. Climate change impacts on wind energy resources in North America based on the CMIP6 projections. Science of The Total Environment. 2022 Feb 1;806:150580.

Schulz-Stellenfleth J, Emeis S, Doerenkaemper M, Bange J, Canadillas B, Neumann T, Schneemann J, Weber I, Zum Berge K, Platis A, Djath B. Coastal impacts on offshore wind farms‒A review focussing on the German Bight area. Meteorol. Z. 2022 Apr 8.

Sediqi MN, Hendrawan VS, Komori D. Climate projections over different climatic regions of Afghanistan under shared socioeconomic scenarios. Theoretical and Applied Climatology. 2019 Song T, Liao H, Subbarayan G. Efficient Local Refinement near Parametric Boundaries Using kd-Tree Data Structure and Algebraic Level Sets. Algorithms. 2022 Jul 13;15(7):245.

Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports. 2022 Apr 15;12(1):1-1.

Yongjian, Sun., Shaohui, Li., Xiaohong, Wang. (2021). Bearing fault diagnosis based on EMD and improved Chebyshev distance in SDP image. Measurement, 176:109100-. doi: 10.1016/J.MEASUREMENT.2021.109100

Zelinka MD, Myers TA, McCoy DT, Po ‐ Chedley S, Caldwell PM, Ceppi P, Klein SA, Taylor KE. Causes of higher climate sensitivity in CMIP6 models. Geophysical Research Letters. 2020 Jan 16;47(1):e2019GL085782.

Zittis G, Almazroui M, Alpert P, Ciais P, Cramer W, Dahdal Y, Fnais M, Francis D, Hadjinicolaou P, Howari F, Jrrar A. Climate change and weather extremes in the Eastern Mediterranean and Middle East. Reviews of geophysics. 2022 Sep;60(3):e2021RG000762