FEDERAL UNIVERSITY OF SÃO CARLOS
CENTER FOR EXACT SCIENCES AND TECHNOLOGY
DEPARTMENT OF MECHANICAL ENGINEERING

EDUARDO SCARDELLATO E SILVA

STATISTICAL CALIBRATION METHODS EFFECTS TO ACCURACY AND UNCERTAINTY REPRESENTATION OF SUB-SEASONAL TO SEASONAL ENSEMBLE FORECASTS

SÃO CARLOS
2025

EDUARDO SCARDELLATO E SILVA

STATISTICAL CALIBRATION METHODS EFFECTS TO ACCURACY AND UNCERTAINTY REPRESENTATION OF SUB-SEASONAL TO SEASONAL ENSEMBLE FORECASTS

Undergraduate Thesis presented to the Department of Mechanical Engineering of Federal University of São Carlos, to obtain the title of Bacharel in Mechanical Engineering.

Advisor: Prof. Dr. Fernando Guimarães Aguiar

SÃO CARLOS
2025

FUNDAÇÃO UNIVERSIDADE FEDERAL DE SÃO CARLOS
COORDENAÇÃO DO CURSO DE ENGENHARIA MECÂNICA (CCEMEC)
Rod. Washington Luís km 235 - SP-310, s/n - Bairro Monjolinho, São Carlos/SP, CEP 13565-905
Telefone: (16) 33519703 - http://www.ufscar.br

DP-TCC-FA nº 52/2025/CCEMec/CCET/R

Graduação: Defesa Pública de Trabalho de Conclusão de Curso

Folha Aprovação (GDP-TCC-FA)

FOLHA DE APROVAÇÃO

EDUARDO SCARDELLATO E SILVA

STATISTICAL CALIBRATION METHODS EFFECTS TO ACCURACY AND UNCERTAINTY REPRESENTATION OF SUB-SEASONAL TO SEASONAL ENSEMBLE FORECASTS

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos – Campus São Carlos

São Carlos, 11 de dezembro de 2025

ASSINATURAS E CIÊNCIAS

Cargo/Função	Nome Completo
Orientador	Fernando Guimarães Aguiar
Membro da Banca 1	Leonardo Marquez Pedro
Membro da Banca 2	Sidney Bruce Shiki

Documento assinado eletronicamente por **Fernando Guimaraes Aguiar**, **Docente**, em 11/12/2025, às 09:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Sidney Bruce Shiki**, **Docente**, em 11/12/2025, às 09:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Leonardo Marquez Pedro**, **Docente**, em 15/12/2025, às 13:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufscar.br/autenticacao, informando o código verificador **2106414** e o código CRC **66CDA57A**.

I would like to dedicate this work especially to my mother, Karina, my father, André, and my brother Ricardo, who, even from afar, have spared no effort to see me happy and pursuing my dreams.

ACKNOWLEDGMENTS

I would like to thank my tutor, Ganglin TIAN, very much for helping me on a day-to-day basis, especially with the analysis to familiarize me with the data that was used for the research, as well as Professor Camille LE-COZ, for also always engaging me critically, especially when it came to finding physical and statistical meaning to the problems

Finally, I would like to warmly thank Professor Fernando for accepting to serve as my tutor in Brazil. His insightful feedback and openness were key to adapt the research and make it both coherent and meaningful.

"Great things are done by a series of small things brought together."
— Vincent van Gogh

This sheet must be replaced by the scanned approval sheet.

RESUMO

Este trabalho investiga o impacto de um método de calibração estatística na precisão e na representação da incerteza de previsões em conjunto na escala sub-sazonal a sazonal (S2S), motivado pela importância dessas previsões para setores como energia, agricultura e gestão de riscos, bem como pelas limitações conhecidas de modelos brutos, que frequentemente apresentam vieses sistemáticos e espalhamento inadequado. A partir de uma revisão bibliográfica sobre técnicas avançadas de pós-processamento estatístico, com destaque para abordagens como Bayesian Model Averaging (BMA), Non-homogeneous Gaussian Regression (NGR) e métodos homogêneos de correção de viés, este estudo foca na avaliação do método de ajuste de média e variância (Mean-Variance Adjustment, MVA), já utilizado em escalas sazonais, investigando se sua simplicidade e baixo custo computacional são suficientes para melhorar previsões S2S em comparação às previsões brutas do modelo estendido do ECMWF e a uma climatologia de referência. O objetivo central é quantificar como a aplicação do MVA afeta a acurácia e a calibração de previsões de velocidade do vento a 100 m de altura. Para isso, utilizam-se dados de reanálise ERA5 como verdade-terreno, além de previsões operacionais e hindcasts do ECMWF para o inverno do Hemisfério Norte sobre a Europa, processados em Python com a biblioteca Xarray, e avaliados por meio de métricas determinísticas (viés, erro quadrático médio – RMSE) e probabilísticas (espalhamento, razão espalhamento–habilidade – SSR, histogramas de postos, Continuous Ranked Probability Score – CRPS e Continuous Ranked Probability Skill Score – CRPSS). Os resultados mostram que o MVA reduz significativamente o viés das previsões brutas e melhora o CRPS e o CRPSS em relação tanto ao modelo bruto quanto à climatologia nos primeiros dias de previsão, especialmente na primeira e segunda semanas, embora o método leve a um certo subespalhamento (underdispersion) do conjunto e a ganhos limitados em termos de SSR, indicando que a incerteza ainda não é representada de forma ideal. Conclui-se que o MVA é uma solução simples e eficiente como primeiro nível de calibração para previsões S2S de vento, melhorando a compatibilidade estatística entre previsões e observações em horizontes curtos a intermediários, mas que, para horizontes mais longos ou para uma representação mais realista da incerteza, é recomendável investigar técnicas de calibração mais sofisticadas e possivelmente multivariadas em trabalhos futuros.

Palavras-chave: Previsão sub-sazonal a sazonal, análise de dados, calibração estatística.

ABSTRACT

This work investigates the impact of a statistical calibration method on the accuracy and uncertainty representation of ensemble forecasts at the sub-seasonal to seasonal (S2S) scale. The study is motivated by the importance of these forecasts for sectors such as energy, agriculture, and risk management, as well as by the known limitations of raw models, which often exhibit systematic biases and inadequate spread. Based on a literature review of advanced statistical post-processing techniques, with emphasis on approaches such as Bayesian Model Averaging (BMA), Non-homogeneous Gaussian Regression (NGR), and homogeneous bias-correction methods, this study focuses on evaluating the Mean-Variance Adjustment (MVA) method, already used at seasonal scales, examining whether its simplicity and low computational cost are sufficient to improve S2S forecasts in comparison with raw forecasts from the ECMWF extended-range model and with a reference climatology. The main objective is to quantify how the application of MVA affects the accuracy and calibration of wind-speed forecasts at 100 m height. To this end, ERA5 reanalysis data are used as ground truth, together with operational forecasts and hindcasts from ECMWF for Northern Hemisphere winter over Europe, processed in Python with the Xarray library and evaluated using deterministic metrics (bias, root mean square error – RMSE) and probabilistic metrics (spread, spread–skill ratio – SSR, rank histograms, Continuous Ranked Probability Score – CRPS, and Continuous Ranked Probability Skill Score – CRPSS). The results show that MVA significantly reduces the bias of raw forecasts and improves CRPS and CRPSS relative to both the raw model and the climatology in the first forecast days, especially in the first and second weeks, although the method leads to some underdispersion of the ensemble and to limited gains in terms of SSR, indicating that uncertainty is still not represented in an ideal way. It is concluded that MVA is a simple and efficient solution as a first level of calibration for S2S wind forecasts, improving the statistical compatibility between forecasts and observations at short to intermediate lead times, but that, for longer horizons or for a more realistic representation of uncertainty, it is advisable to investigate more sophisticated and possibly multivariate calibration techniques in future work.

Keywords: Subseasonal-to-seasonal forecasting, data analysis, Statistical calibration.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ABNT	Brazilian Association of Technical Standards
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	Fifth Generation ECMWF Atmospheric Reanalysis
LMD	Laboratoire de Météorologie Dynamique
MVA	Mean and Variance Adjustment
NGR	Non-homogeneous Gaussian Regression
NWP	Numerical Weather Prediction
RMSE	Root Mean Square Error
SSR	Spread-Skill Ratio
S2S	Subseasonal-to-Seasonal (Forecasting)
BMA	Bayesian Model Averaging
CRPSS	Continuous Ranked Probability Skill Score
JMA	Japan Meteorological Agency
UTC	Coordinated Universal Time

LIST OF SYMBOLS

f	(Re)Forecast value
o	Observation or ground truth (re-analysis)
t	Verification time index $(1, \ldots, T)$
l	Lead time index $(1, \ldots, L)$
i	Latitude index $(1, \ldots, I)$
j	Longitude index $(1, \ldots, J)$
m	Ensemble member index $(1, \ldots, M)$
x_j	Ensemble member to be bias-adjusted
$\bar{x}_e$	Mean of all ensemble forecast members
σ_e	Standard deviation of ensemble forecast members
$\bar{o}_{ref}$	Mean of reference (observation) data
σ_{ref}	Standard deviation of reference (observation) data
x_j^*	Bias-adjusted forecast member
RMSE	Root Mean Square Error
$\text{Bias}_{l,i,j}$	Mean Bias Error at lead time l and position (i,j)
Spread_l	Ensemble spread at lead time l
SSR	Spread-Skill Ratio
CRPS_l	Continuous Ranked Probability Score at lead time l
CRPSS	Continuous Ranked Probability Skill Score
var_m	Variance across ensemble members
$\bar{f}_{t,l,i,j}$	Ensemble mean at time t, lead l, latitude i, longitude j

CONTENTS

1 INTRODUCTION

1.1 Research Motivation and Objectives

The Sub-seasonal to Seasonal (S2S) forecast range, typically spanning from about 3 to 6 weeks, fills the gap between short-term weather forecasts (up to roughly 7–10 days) and long-term climate predictions (from several months to decades). This intermediate range is particularly relevant for applications such as agriculture, water resource management and energy planning, where decisions depend on aggregated conditions over weeks rather than individual days. At the same time, S2S forecasting is intrinsically challenging: the lead time is long enough for much of the memory of the atmospheric initial conditions to be lost, while still being too short for slow components of the climate system, such as the ocean, to fully dominate the predictability (HUDSON et al., 2017).

From a mechanical engineering perspective, this work is closely related to the analysis, design, and operation of energy systems that interact directly with atmospheric flows. Wind is not merely a meteorological variable, but a fundamental physical input for the conception, sizing, and lifetime assessment of wind energy infrastructures. An accurate characterization of wind speed at hub height directly influences key mechanical aspects such as aerodynamic loads acting on turbine blades, structural fatigue, power curve estimation, and drivetrain performance. In this context, improving the accuracy and reliability of wind speed forecasts, particularly at the sub seasonal to seasonal scale, supports better informed engineering decisions that range from structural design margins to maintenance planning and operational optimization of wind turbines.

In addition, this research has a direct and practical impact on wind energy planning and monitoring in France, where wind power represents a major pillar of the national energy transition strategy. Sub seasonal to seasonal wind forecasts are increasingly used by operators and system planners to anticipate production variability, improve grid integration, and reduce operational risks in both onshore and offshore wind farms. By evaluating and improving statistical calibration methods applied to ECMWF ensemble forecasts, this study contributes to more reliable wind resource assessments at time scales that are directly relevant for energy management. As a result, the work lies at the intersection of atmospheric science, data driven modeling, and mechanical engineering applications, reinforcing the role of engineers in connecting physical modeling, uncertainty quantification, and real world energy system performance.

In this context, the present work investigates a statistical calibration technique known as Mean and Variance Adjustment (MVA), already applied in short(and long)range forecasting, but still scarcely explored at the sub-seasonal to seasonal scale. The study focuses on ensemble wind-speed forecasts from the extended-range model of the European

Centre for Medium-Range Weather Forecasts (ECMWF), and uses the ERA5 reanalysis to construct both the observational reference and a climatological baseline. The fundamental goal is to assess how the MVA-calibrated forecasts perform, in terms of accuracy and uncertainty representation, when compared with the original, uncalibrated ECMWF ensemble (raw forecasts); and a simple ERA5-based climatology. To this end, a set of deterministic and probabilistic verification scores is applied to the three approaches, allowing a systematic comparison of their skill and calibration over the S2S range.

1.2 Outline of the Paper Structure

This work is structured to explore calibration methods and the evaluation of ensemble forecasting, particularly in the context of numerical weather prediction. The report is divided into several key sections: Introduction, which outlines the objectives, motivation, and structure; Background, providing foundational knowledge on numerical weather prediction, ensemble systems, and calibration methods; Data, describing the datasets used on the research; Methodology, detailing the techniques for mean and variance adjustment and verification methods to measure each (re)forecast and climatology technique; Results, presenting the outcomes through various scores and maps; and Conclusion, summarizing the findings and implications of the study.

2 THEORETICAL BAKCGROUND

Forecasts refer to the predictions or estimations about future events or conditions based on the analysis of current and historical data. These predictions are derived through various methodologies and are applied across numerous fields, including economics, finance, health, and meteorology. Forecasting aims to provide insights into future outcomes to aid decision-making and planning.

With data collection, it is possible to analyse trends and patterns throughout several areas of knowledge and predicting (or at least being aware of) events that are yet to come. In the context of the planet Earth, it is no different. As illustrated in Figure 1, these observations are ingested by a data assimilation system, which combines them with a prior "first guess" provided by a previous model run to produce a consistent estimate of the atmospheric state, known as the initial conditions.

Nowadays, there are different ways and techniques to observe and gather data regarding the weather, such as temperature, precipitation and wind speed. All these observations, mostly made by satellites (BUIZZA, 2021) or even on-ground facilities guarantee that we can collect as much as we can to use this observation to try to forecast future events.

As it is known, all observation methods are subject to errors and misleading information. That is why, once the information is gathered, it must pass through some treatment made mostly from mathematical models. Once the model corrects the observation errors, the analysis is ready to be used and the initial conditions of a system are set.

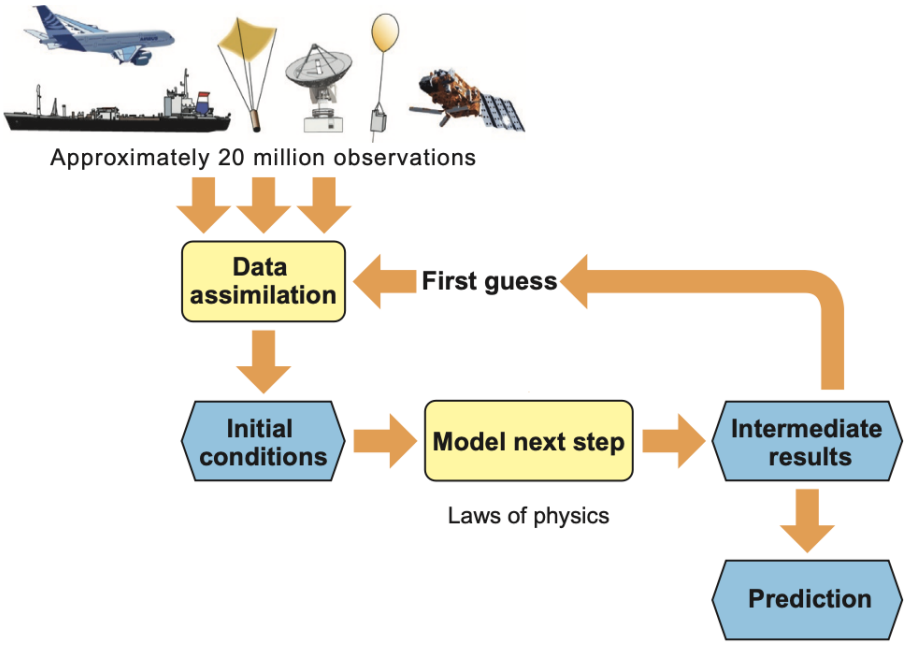


Figure 1 – Simplified schematic of the dynamical weather prediction process. Adapted from (BUIZZA, 2021)

2.1 Importance of Numerical Weather Prediction

Numerical Weather Prediction (NWP) is a scientific method that uses mathematical models to simulate the Earth's atmosphere and forecast future weather conditions.

It plays a vital role in modern meteorology by using numerical models to simulate the atmosphere and forecast the weather, taking all the large amount of observation and, using mostly Newton's laws of physics applied to fluids (see (HOLTON, 2012) to verify equations). Numerical predictions can provide forecasts that affect many aspects of society. These forecasts are crucial for daily weather reports, disaster preparedness, aviation safety, agricultural planning, and much more. As computing power and data accuracy improve, NWP continues to be an essential tool for enhancing our understanding of atmospheric processes and the reliability of weather forecasts, helping to protect lives and property.

2.2 Subseasonal to Seasonal Forecasting (S2S)

Sub seasonal to seasonal forecasting, commonly referred to as S2S forecasting, occupies an intermediate position between short range weather prediction and long range seasonal outlooks. This time window typically spans lead times from approximately two weeks up to two or three months and represents one of the most challenging horizons in atmospheric prediction. Unlike short range forecasts, which are strongly constrained by the initial atmospheric state, and seasonal forecasts, which benefit from slowly varying

climate drivers, the S2S period lies in a transition regime where sources of predictability are weaker, intermittent, and more difficult to exploit.

Short range weather prediction, usually covering lead times from one to seven days, relies primarily on the accurate initialization of the atmosphere through data assimilation. At these time scales, numerical weather prediction models are able to resolve synoptic and mesoscale phenomena such as frontal systems, cyclones, and localized wind events with relatively high accuracy. Forecast skill remains high because errors associated with model dynamics and initial conditions have not yet grown significantly. However, due to the chaotic nature of the atmosphere, this skill rapidly decreases beyond approximately ten days.

At longer lead times, seasonal forecasting focuses on predicting large scale climate anomalies rather than individual weather events. Seasonal outlooks typically provide probabilistic information about deviations in temperature, precipitation, or wind conditions over periods of several months. Their predictability arises from slowly evolving components of the climate system, such as sea surface temperature anomalies, land surface processes, and large scale ocean atmosphere interactions including the El Niño Southern Oscillation. While these forecasts are valuable for long term planning, they lack the temporal resolution required to describe intra seasonal variability.

The S2S forecasting range lies between these two regimes and is characterized by a partial loss of atmospheric initial condition memory combined with an emerging but still limited influence of slow climate drivers. As a result, forecast skill during this period is generally lower and more variable, making the interpretation and practical use of raw model outputs particularly challenging. Nevertheless, S2S forecasts are of high relevance for applications that depend on aggregated weather behavior over several weeks, such as energy production planning, hydrological management, and risk assessment (MARIOTTI; RUTI; RIXEN, 2018).

Figure 2 provides a schematic representation of the evolution of forecast skill as a function of prediction lead time. The figure highlights the high skill associated with short range weather prediction, the reduced and uncertain predictability during the S2S period, and the partial recovery of skill at seasonal scales due to the influence of large scale climate modes.

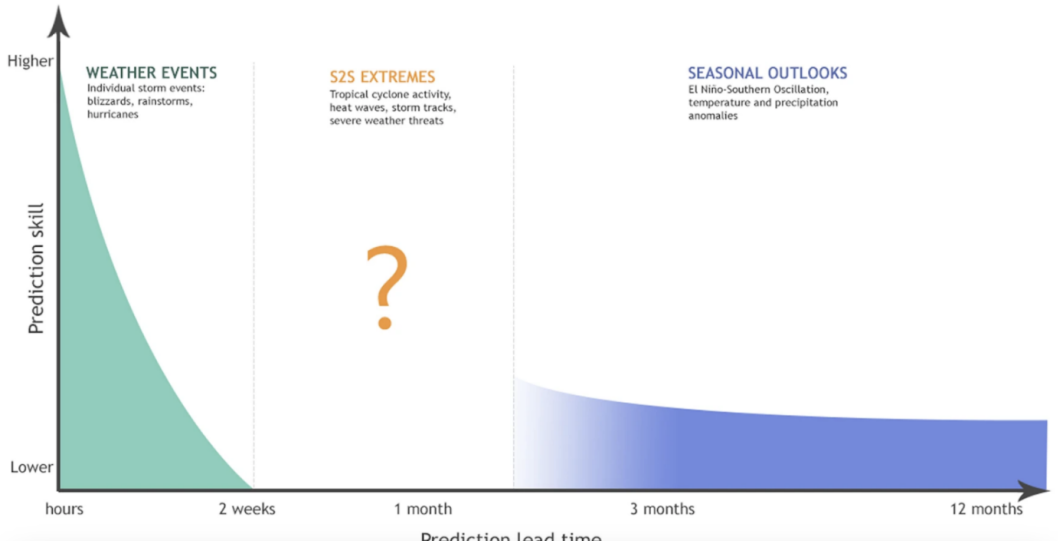


Figure 2 – Conceptual illustration of forecast skill as a function of prediction lead time, highlighting the transition between short range weather prediction, the sub seasonal to seasonal period, and seasonal climate outlooks.

In the context of wind energy applications, the S2S period is particularly critical. While short range forecasts support real time operation and very short term dispatch, and seasonal outlooks inform strategic planning, S2S forecasts address a gap that is directly relevant for maintenance scheduling, resource adequacy assessment, and anticipation of persistent wind regimes. However, the reduced intrinsic predictability at these time scales makes raw ensemble forecasts prone to systematic biases and misrepresentation of uncertainty, reinforcing the need for statistical calibration methods to improve their reliability and practical usefulness.

2.3 Ensemble Forecasting Systems

Ensemble forecasting is a method used in weather prediction that involves running multiple simulations, or ensemble members, to account for the inherent uncertainty in weather forecasting. Each simulation starts with slightly different initial conditions and sometimes uses different model parametrizations to represent the uncertainties in the initial state of the atmosphere and model physics. By comparing the different outcomes from these simulations, it is possible to measure the range of possible future states of the atmosphere, rather than relying on a single deterministic forecast (BUIZZA, 2021). This approach helps in quantifying the forecast uncertainty and provides a probabilistic range of outcomes.

Quantifying forecast uncertainty through ensemble forecasting is vital because the atmosphere is a chaotic system where small changes in initial conditions can lead to vastly different outcomes (WILKS, 2021). By generating a spectrum of possible future states, ensemble forecasting allows forecasters to understand the likelihood of various scenarios, helping them to communicate risks and uncertainties more effectively. This probabilistic

approach is particularly important for decision-making in sectors like agriculture, emergency management, and aviation, where understanding the range of possible weather events can significantly impact planning and response strategies.

2.3.1 Introduction to the Ensemble Forecasting System

The operational ensemble forecasting system builds on these principles by organizing the production of multiple forecasts within a consistent framework. In practice, a central numerical weather prediction model is integrated many times in parallel, with carefully designed perturbations applied to the initial conditions and, in some cases, to selected physical parametrizations. These perturbations are constructed to sample the main sources of uncertainty identified in the observing system and in the model dynamics.The resulting set of forecasts is then post-processed to derive probabilistic products, such as exceedance probabilities or quantiles, and routinely verified against observations.

In global weather forecasting, ensemble forecasting systems are integrated into the operations of major weather prediction centers, such as the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Weather Service (NWS). These systems play a crucial role in improving medium-to long-range weather predictions by accounting for uncertainties and providing probabilistic forecasts that are essential for risk assessment and decision-making processes across various sectors.

2.3.2 Limitations of Ensemble Forecasts

While ensemble forecasting has significantly improved the accuracy and reliability of weather predictions by incorporating multiple model simulations, it is not without limitations. One key issue with raw ensemble forecast (and deterministic as well) outputs is the presence of systematic errors. These errors can appear from inherent limitations in the models, specially caused by the chaotic nature of the atmosphere as lead time increases.

Another significant limitation is the insufficient spread among ensemble members, which can lead to an underestimation of forecast uncertainty. Ideally, an ensemble should capture a wide range of possible outcomes to represent the inherent uncertainty in weather forecasting accurately. However, if the ensemble members are too similar—often due to shared model characteristics or initial conditions—the spread may be artificially narrow. This can result in overconfidence in the forecast, reducing its usefulness for decision-making and other potential usages.

2.4 Calibration

Calibration refers to the statistical compatibility between the forecast and the observation, it means that the forecast is calibrated if the observation cannot be distin-

guished from a random draw from the predictive distribution (THORARINSDOTTIR; SCHUHEN, 2021).

Statistically, calibration addresses the uncertainties and systematic errors present in ensemble forecasting models. These models operate at a lower resolution than the actual atmosphere, which means they cannot capture all the complexities of atmospheric processes perfectly. Moreover, physical processes within the models are often parameterized using simplified equations to describe phenomena that are inherently complex. Without calibration, these simplifications can introduce significant biases and errors into the forecasts. By calibrating the ensemble forecasts, we adjust for these discrepancies, reducing model error and uncertainty, and thus providing a more accurate and dependable representation of future wind conditions at 100 meters height. This statistical refinement is essential for ensuring that decision-makers can rely on the forecasts for critical planning and operational decisions.

2.4.1 Overview of Existing Calibration Methods

One way to divide calibration method on ensemble forecasting is by differentiating Univariate and Multivariate Ensemble Calibration. Univariate Ensemble Calibration focuses on adjusting forecasts for a single variable, such as wind speed at 100 meters high (uv100), to improve the accuracy of predictions for that specific variable. This involves techniques that correct biases and improve the reliability of the probabilistic forecasts for the single variable of interest and it is going to be the reference throughout this work. On the other hand, Multivariate Ensemble Calibration addresses the inter dependencies between multiple variables, calibrating them simultaneously to ensure consistency across the forecasted variables. This approach is essential when variables are interrelated, such as wind speed, temperature, and humidity, providing a more complex forecast.

A Univariate Ensemble Calibration can be approached through homogeneous methods, which apply a uniform correction across all data points. This method assumes that the bias and spread of the forecast errors are consistent throughout the dataset. Homogeneous calibration techniques, such as linear regression or mean bias correction, adjust the entire ensemble forecast by the same amount, based on the average discrepancies observed between the forecasts and the actual observations. This approach is straightforward and is proven to be really effective on seasonal forecasting, specially taking in account the extremely low computing cost (MANZANAS et al., 2019). For example, if the wind speed forecasts at 100 meters almost consistently underestimate by a certain margin a homogeneous correction could systematically adjust the forecasts to improve their accuracy.

The most generally used ensemble calibration methods nowadays have been the Bayesian model averaging (BMA) and non homogeneous regression method (NGR) ((WILKS, 2021)).

The BMA is a statistical method that combines multiple model forecasts by weighting them according to their historical performance, with the weights derived from Bayesian probability. This method accounts for model uncertainty by averaging predictions from different members, each contributing proportionally to their reliability. NGR, on the other hand, is a regression-based approach where the forecast distribution is adjusted by fitting a parametric model (typically, but not necessarily Gaussian) to the ensemble forecasts. NGR is flexible and allows for the correction of biases and dispersion errors in the ensemble.

One of the most popular method, nonetheless, is the Mean and Variance Adjustment (MVA) (MANZANAS et al., 2019). It is a very straightforward method that adjusts ensemble forecasts by comparing them to corresponding analysis (observations). Since this is the actual model used on the LMD to seasonal forecasting and because of its relevance on the scenario, it will be the main method analysed on this report, discussed with further details on section 3.2.1.

3 DEVELOPMENT

3.1 Data

In this work, we're analysing the wind speed at 100 meters high, so, for being able to remark more evidently the differences between each forecast/hindcast, the time of the year selected to evaluate data is from the northern hemisphere winter, from December 3rd to March 31st mostly. In terms of location, due to more relevance, the European continent will be used as base to evaluate the data as well. It is important to saw as well that two different sources will be used, one for the re-analysis data, set as the observation (truth), and other to the forecasts itself.

3.1.1 ERA5 Reanalysis

In this work, the wind speed observations (set as ground truth) used and processed comes from ERA5 (fifth-generation global reanalysis) data. It is a comprehensive reanalysis dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF) that provides detailed information on atmospheric, land, and oceanic variables. This data is essential for understanding past weather patterns and for improving the accuracy of weather forecasts. Theses data are retrieved from January 1979 to 2020 over Europe (34°–74°N, 13°W–40°E) at a spatial resolution of 0.25° and a temporal resolution of 6 h (i.e., instantaneous values at 0000, 0600, 1200, and 1800 UTC) from the Copernicus Climate Change Services' Climate Data Store (RAOULT et al., 2017). In this analysis, we will average the 4 instant values obtained daily so we could capture an average scenario of the wind speed in one day.

3.1.2 ECMWF Forecasts

The operational S2S predictions from the ECMWF model (VITART; COAUTHORS, 2019) are produced by extending the medium range forecasts (i.e., up to 2 weeks) to 46 days 2 times per week (at 0000 UTC on Mondays and Thursdays). These are ensemble predictions resulting from coupled ocean–atmosphere integrations. The ensemble is composed of 51 members (50 perturbed + control).

The raw forecast model dataset, consists the, of five dimensions: 'number' (the members of the ensemble), 'startDay' (the day of initialization), 'LeadTime'(the range of the forecast), 'lat' (latitude), and 'lon'(longitude). This dataset provides more recent forecast data from 2015 to 2021, allowing for an up-to-date assessment of current forecasting capabilities.

3.1.3 ECMWF Hindcasts

Hindcasts, or re-forecasts are essentially backward-looking forecasts. Unlike regular forecasts, which predict future weather conditions, hindcasts involve using current forecasting models to predict past weather events. This process allows verifying the performance of models against known outcomes, providing a benchmark for accuracy and reliability.

The hindcast model dataset includes six dimensions: 'number', representing the ensemble number; 'hdate', indicating the retrospective years corresponding to the hindcasts; 'startDay', the day of the year on which the forecasts are initiated; 'LeadTime', denoting the number of days each forecast predicts; 'lat' for latitude; and 'lon' for longitude. This structure allows for a comprehensive analysis of past weather forecasts and their accuracy over a 20-year period. By comparing these hindcasts with actual historical weather data, we can evaluate the performance of the forecast models and identify any biases or trends.

3.1.4 Climatology

According to the Oxford dictionary, climate is defined as:

"the regular pattern of weather conditions of a particular place"

The climatology involves, then, understanding the long-term patterns and variations in weather over time and space. It examines average weather conditions and their changes over decades, centuries, or even longer periods. Its purpose is mostly related to the predicting of future climate conditions and the understanding of potential impacts of climate change, identifying natural climate cycles and distinguish them from human-induced changes, for example.

In this work, the data from ERA5 reanalysis is used regarding the wind speed (at 100 meters high, noted as uv100) over Europe in the past 35 years. An example of how a pattern can be found between years is to plot a graphic, averaging the wind speed over Europe for all 20 years, as seen in figure 3. The graphic shows that, on average, the wind speed doesn't vary as much over the years. In other words, we see that, for each season, the wind speed over the years follow similar behaviors.

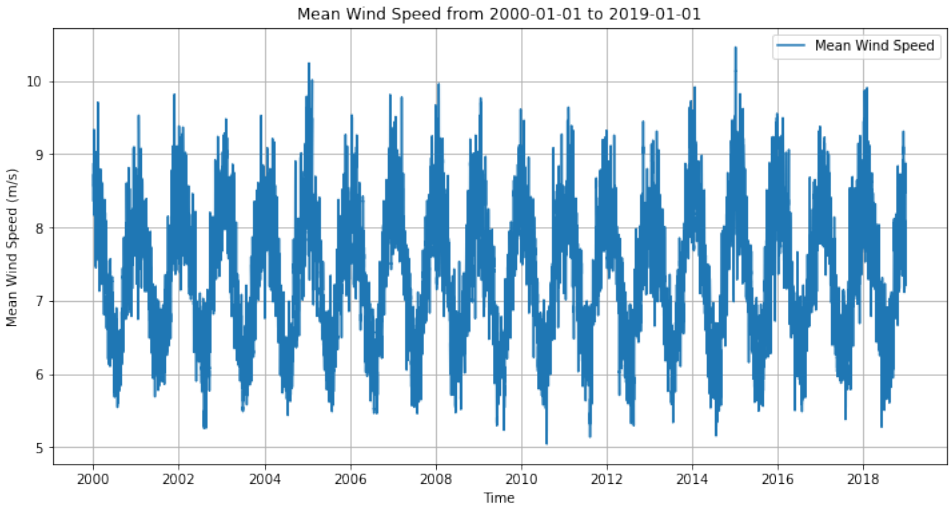


Figure 3 – Mean wind speed in Europe over 20 years

Using this kind of information, climatologists can have an expectation for the wind speed (and many other variables) over each day of the year. For example, on figure 4, it is possible to clearly see the wind variation between the seasons (high speed on winter, low speed on summer) and, when this figure is compared to the yearly graphic, it is evident that, when averaged, the values are not the same (less variable on the yearly average).

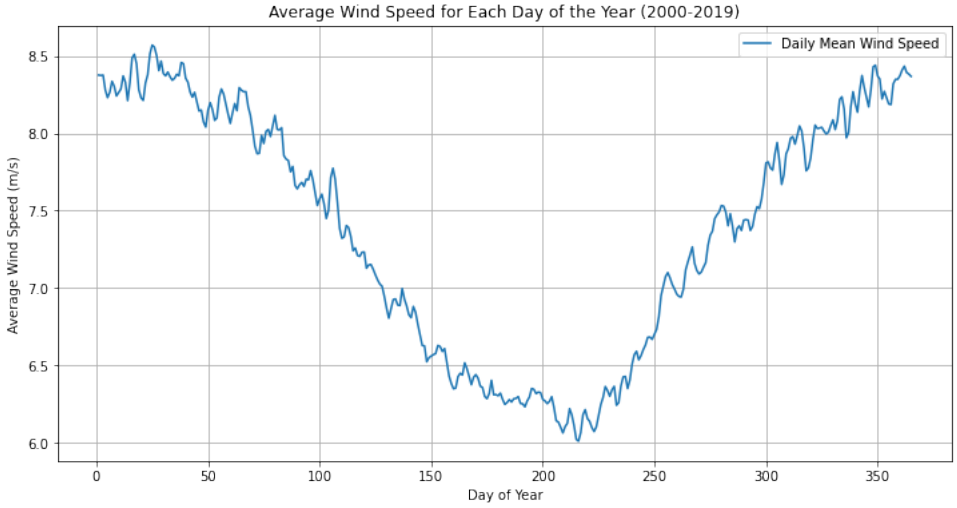


Figure 4 – Mean wind speed over 20 years in Europe for each day of the year

It is important to say that climatology plays a crucial role as a benchmark in evaluating the performance of weather forecasts. By providing a comprehensive understanding of the typical weather patterns and variations over a long period, climatology serves as a baseline against which forecast models can be compared (GOUTHAM et al., 2022). In practice, the accuracy and reliability of a forecasting method are often assessed by comparing its predictive scores with those derived from climatological data. This comparison helps identify the strengths and weaknesses of the forecast model, highlighting areas where it performs better or worse than when just using the historical data.

With that being said, for both forecasts and hindcasts, a climatology is computed so it can be used as a reference for performance and skill of the forecasts.

For forecasts, the climatology is simply computed using the data from the 35 years prior to the forecast. In this case, the forecasts are computed from 2015 and 2016, so the climatology is computed by using each year (from 1979 to 2014) as members. The array is composed by the same number of dimensions as the forecast itself ('startDay', 'LeadTime', 'number','lat','lon').

For the hindcasts, the computation is done the following way: for each hdate, the number dimension is composed by the the 15 years prior to the hdate itself. For example, for the starting day December 1st 2015, for the hdate of 2013, the 'number' dimension is composed by the 15 previous years from the current hdate, in that case, December 1st from 1998 to 2012. The array is composed by the same number of dimensions as the hindcasts itself:'startDay', 'LeadTime','hdate', 'number','lat','lon'.

3.2 Methodology

In this section, the methodology of the paper will be presented. As discussed on subsection 2.4.1, the method that is going to be presented and analysed will be the MVA. To obtain, process and analyse the data, the Python software was used. These datasets are processed and analysed using the Xarray library, a powerful tool for handling multi-dimensional data in Python.

3.2.1 Mean and Variance Adjustment (MVA)

This method is based on the assumption that both the reference and predicted distributions of seasonal wind speed are approximated well by a Gaussian (normal) distribution. The adjustment creates predictions that have the same mean and standard deviation as the reference dataset. The Gaussian assumption is a limitation of the approach because the monthly and seasonal wind speed distribution can be, at times, slightly non-Gaussian. (TORRALBA et al., 2017)

The bias correction scheme can be summarized in this way:

$$x_j^* = (x_j - \bar{x}_e)\,\frac{\sigma_{ref}}{\sigma_e} + \bar{o}_{ref}, \tag{1}$$

Where x_j is the member whose bias needs to be adjusted; $\bar{x}_e$ and σ_e are the mean and the standard deviation, respectively, of all the members of all the hindcasts corresponding to the forecast; and $\bar{o}_{ref}$ and σ_{ref} are the mean and the standard deviation, respectively, of the truth (or observations) corresponding to the forecasts. For the bias adjustment of any given reforecast within a hindcast set, the remaining 19 years of hindcasts are used to adjust the mean and the spread through a leave-one-out approach to prevent overfitting (GOUTHAM et al., 2022).

3.2.2 Verification

Verification methods plays a crucial role when dealing with ensemble prediction systems and statistically postprocessed ensemble forecasts. These verification methods provide evaluations of future weather, which include both the best prediction derived from the ensemble and the associated prediction uncertainty. (THORARINSDOTTIR; SCHUHEN, 2021)

Verification also evaluates the accuracy and reliability of weather forecasts. By comparing the model predictions against actual observations, it is assessed how well the model is performing. This process helps identify any biases or systematic errors in the model, allowing for continuous improvement and refinement of forecasting techniques.

A verification method can be deterministic or probabilistic. Both evaluate differently the performance of a forecast.

For easier nomenclature, the table (1) of symbols and descriptions will be used from now on to describe different variables and parameters of the equations

Symbol	**Range**	**Description**
f		(re)Forecast
o		Ground Truth (re-analysis)
t	$1, \ldots, T$	Verification time
l	$1, \ldots, L$	Lead time
i	$1, \ldots, I$	Latitude index
j	$1, \ldots, J$	Longitude index
m	$1, \ldots, M$	Ensemble member index

Table 1 – Symbols, ranges, and descriptions.

3.2.2.1 Deterministic metrics

Deterministic metrics assess the accuracy of single-valued forecasts. These metrics compare a specific forecast value against the observed value to determine the degree of accuracy. Deterministic metrics are straightforward and are typically used for verifying forecasts that predict a specific outcome without accounting for uncertainty. (RASP et al., 2024)

Common deterministic metrics include:

Root Mean Squared Error (RMSE): Used to assess the accuracy of a forecast by quantifying the average magnitude of the errors between predicted and observed values. Physically, RMSE represents the square root of the average of the squared differences between the forecasted and actual observation, giving a measure of the typical deviation of the forecast from reality. This equation is particularly useful because it penalizes larger errors more heavily than smaller ones, due to the squaring process before averaging, which means

that it is sensitive to outliers or significant forecast errors. The RMSE is defined as

$$\mathrm{RMSE}_l = \sqrt{\frac{1}{TIJ}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}(f_{t,l,i,j} - o_{t,i,j})^2} \tag{2}$$

When applying the RMSE formula that includes a spatial average, as shown in equation (2), it is important to evaluate whether this averaging process is necessary or if a localized analysis might be more informative. Spatially averaging the RMSE can provide a broad understanding of the overall model performance across a region, which is useful for summarizing data into a single metric that can be easily compared across different models or forecast periods. However, this averaging process can obscure localized errors, which may be critical in applications where regional variations are significant. In these cases, it may be more insightful to visualize the RMSE on a map, which would allow us to identify and analyze spatial patterns of error, ensuring that the model's strengths and weaknesses are understood in the context of specific locations.

Mean Bias Error (MBE): Used to measure the average difference between the predicted values and the observed values over a set period. Physically, MBE represents the tendency of a forecasting model to consistently overestimate or underestimate the observed values. A positive MBE indicates that, on average, the forecasted values are higher than the actual observations, while a negative MBE suggests a tendency to predict lower values. This metric is really important because it puts in evidence systematic errors in the model, highlighting whether the (re)forecasts are biased in a particular direction.

It can be computed for each location i,j as

$$\mathrm{Bias}_{l,i,j} = \frac{1}{T}\sum_{t}(f_{t,l,i,j} - o_{t,i,j}) \tag{3}$$

Where $\mathrm{Bias}_{l,i,j}$ is the difference between the (re)forecasted value $f_{t,l,i,j}$ and the observed value $o_{t,i,j}$, averaged over all starting days t. The normalization factor $\frac{1}{T}$ ensures that the bias reflects the average error across the entire period considered.

3.2.2.2 Probabilistic metrics

Probabilistic metrics evaluate the performance of probabilistic forecasts, which provide a range of possible outcomes along with their associated probabilities. These metrics assess how well the forecast probabilities match the observed frequencies and account for the inherent uncertainty in the forecasts.

Spread Skill Ratio: The spread-skill ratio SSR is defined as the ratio between the ensemble spread and the RMSE of the ensemble mean $\overline{f_{t,l,i,j}} = \frac{1}{M}\sum_{m}^{M} f_{t,l,i,j,m}$.

The Spread represents the degree of uncertainty in the forecast, in other words, a larger spread indicates greater uncertainty and a wider range of possible outcomes, while a smaller spread suggests higher confidence with more similar predictions across the ensemble members.

$$\text{Spread}_l = \sqrt{\frac{1}{TIJ}\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}\text{var}_m(f_{t,l,i,j,m})} \tag{4}$$

with var_m being the variance in the member dimension.

It is important to notice that, in the same way as pointed on the RMSE equation(2), the spread is averaged spatially as well. It is crucial to evaluate whether the averaging is necessary or if a more localized analysis is preferable. Spatial averaging can provide a general overview of the forecast uncertainty over a large area, summarizing the overall model performance into a single, more manageable metric. However, this averaging process can mask important local variations in forecast uncertainty, which are often critical in specific applications (severe weather forecasting, for example). In such cases, it is better to plot a map of the Spread and analyze the errors locally.

$$SSR = \frac{\text{Spread}_l}{\text{RMSE}(\overline{f})} \tag{5}$$

A well-calibrated ensemble forecast should have a spread-skill ratio of 1 (FORTIN et al., 2014). If the SSR is greater than 1, it suggests the ensemble is over-dispersive, meaning the spread is too wide and the forecast uncertainty is exaggerated, leading to an overestimation of the range of possible outcomes and if the SSR is less than 1, the ensemble is under-dispersive, meaning the spread is too narrow and the forecast uncertainty is underestimated, potentially missing some of the actual variability in the range of the outcomes. Note that the spread-skill ratio is only a first-order test for calibration (GOUTHAM et al., 2022), this means that the Spread-Skill Ratio provides a basic, initial assessment of how well the ensemble spread matches the forecast uncertainty, but it doesn't capture all aspects of calibration (bias for example).

Rank Histogram: A rank histogram is a graphical tool used in weather forecasting to assess the reliability and calibration of ensemble forecasts. Physically, it represents the distribution of ranks of observed outcomes relative to the ensemble forecast members when they are sorted in ascending order. If the ensemble is well-calibrated, meaning the ensemble members are effectively capturing the true distribution of possible outcomes, the Rank Histogram should appear uniform, indicating that the observations fall equally into each rank bin. Deviations from this uniformity, such as U-shaped or bell-shaped histograms, reveal systematic biases or insufficient spread in the ensemble, such as overconfidence (too narrow spread) or underconfidence (too wide spread) (HAMILL, 2001).

To compute a rank histogram is quite simple. Once with the ensemble forecast and its matching observation, we sort the N ensemble members, for each forecast period, in ascending order. Once with the $N+1$ member rank, we determine where the observed value falls within the sorted list of ensemble forecast, in other words, it involves counting how many ensemble members are less than or equal to the observed value. The rank of the observation will be a number between 1 and $N+1$, representing its position if it

were inserted into the sorted list. We repeat the process for all forecast periods to build a frequency distribution of the ranks of the observed values. Finally, we plot a histogram with the ranks on the x-axis (ranging from 1 to $N + 1$ and the frequency of each rank on the y-axis.

Continuous Ranked Probability Score (CRPS):

The CRPS is computed as the time average of unbiased (conditional), as shown in Equation 6. It captures both the sharpness (the concentration of the forecast distribution) and the reliability (the agreement between forecast probabilities and observed frequencies) of the ensemble forecast. Importantly, CRPS is minimized by any prediction $f_{t,l}$ that aligns with the same component distributions as the ground truth at time t, making it a valuable tool for assessing the quality of probabilistic forecasts.

According to the WeatherBench ((RASP et al., 2024)), CRPS can be computed by averaging over time, latitude, and longitude components, considering the multi-dimensional data of (re)forecasts and climatology $f_{t,l}$ conditional on the initial ground truth o_{t-l}.

$$\mathrm{CRPS}_l := \frac{1}{T}\sum_{t}^{T}\left[\frac{1}{M}\sum_{m=1}^{M}\|f^{(m)} - o\|_{t,l} - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{n=1}^{M}\|f^{(m)} - f^{(n)}\|_{t,l}\right]. \quad (6)$$

The CRPS is negatively oriented (i.e., smallest values indicate more accurate forecasts), and it rewards those forecasts whose probabilities are concentrated around the observation. As the lead time increases, the ability to predict finer-scale features in time and space quickly falls. It has the same units as the physical quantity being assessed (in this case, uv100 wind speeds [m/s]). The CRPS can also be calculated for the observed climatology (GOUTHAM et al., 2022).

The CRPS is very useful for comparing models when the forecasts do not take the form of a standard probability distribution, or if for a given data set such a distribution cannot be perfectly specified (THORARINSDOTTIR; SCHUHEN, 2021). That is why it is going to be a verification tool when applying different calibration methods later in the report.

The Continuous Ranked Probability Skill Score (CRPSS) is a derived metric used to assess the relative accuracy of probabilistic forecasts by comparing them to a baseline, typically climatology. The CRPSS is calculated as shown in the attached equation:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\text{(re)forecast}}}{\mathrm{CRPS}_{\text{climatology}}} \quad (7)$$

This formula indicates that the CRPSS is derived from the ratio of the CRPS of the (re)forecast to the CRPS of climatology. The score ranges from negative infinity to 1, with a CRPSS greater than 0 indicating that the (re)forecasts are more skillful than climatology. A CRPSS of 1 would indicate a perfect forecast, while a negative CRPSS suggests that the forecast performs worse than climatology. This metric is particularly useful for evaluating the effectiveness of different forecasting models.

4 RESULTS

Once the climatology and the MVA correct forecasts and hindcasts were computed, the verification scores were applied to each one of the methods, so it could be described and analysed. This section will be divided into a part containing the curves of forecasts and hindcasts, to a more broad scenario, but with reference to the maps (Appendix A), so we can see with more details, specially in terms of geographical tendencies and behaviors.

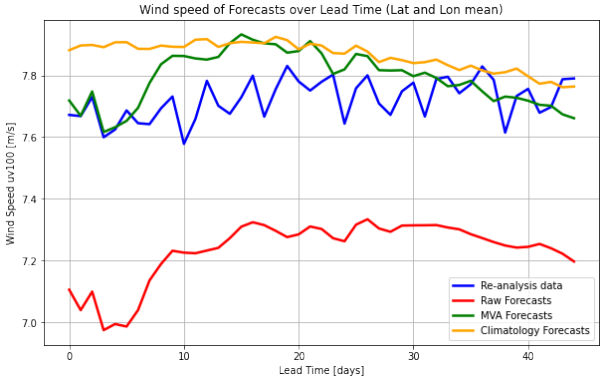
It is important to note that for the analysis, all data was averaged across the start days—specifically, 55 start days (or 55 times 20 in the case of hindcasts), beginning from December 3rd, 2015. This approach was chosen due to computational limitations, ensuring a sufficient number of samples to avoid bias from a short period. Additionally, we evaluated both forecasts and hindcasts to strengthen the robustness of our conclusions. Hindcasts, covering a longer period, provide a broader context for analysis, but since there are differences between forecasts and hindcasts, we also cross-check the results with forecasts to ensure their validity.

4.1 Various scores for raw and MVA-calibrated forecasts and hindcasts

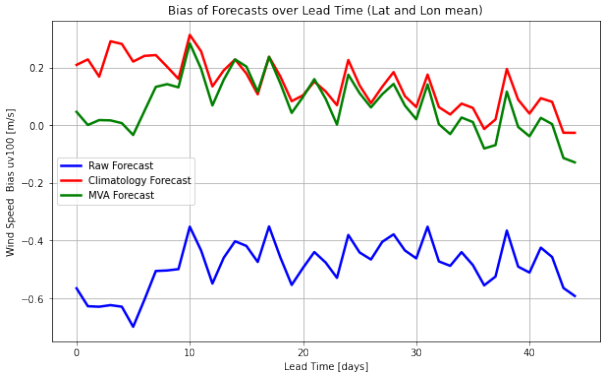
4.1.1 Forecasts

The left graphic 5a shows the wind speed of forecasts at 100 meters high, while the right graphic 5b displays the bias of these forecasts compared to true observations over lead time. The combination of both figures shows that MVA forecast exhibits the smallest bias compared to the Climatology and Raw forecasts. The MVA forecast's bias fluctuates around zero, demonstrating that the adjustments were correctly applied, leading to an less biased prediction. In contrast, the Raw forecast shows a consistently negative bias, indicating a systematic underestimation of the wind speed, while the Climatology forecast tends to have a positive bias, suggesting an overestimation. The MVA forecast's successful reduction of bias suggests that the mean and variance corrections were correctly implemented and have improved the forecast accuracy relative to the uncorrected Raw and Climatology forecasts.

The similarity between the MVA bias and the Climatology bias comes from how the MVA correction works, adjusting forecasts to match the reference data (re-analysis) using climatological means and standard deviations. Since the reference data is very similar to the climatological data, the MVA-corrected forecasts will naturally show similar biases to the Climatology forecast. The MVA corrects the model's climatology, and after two weeks, the influence of the initial conditions appears to diminish, leading the raw forecasts to converge toward their own climatology. Given that the MVA has aligned the model's climatology with the observed climatology, it makes sense that the forecasts and



(a) Wind speed of forecasts at 100 meters high



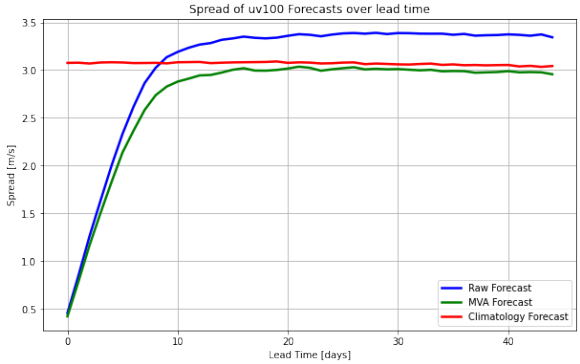
(b) Bias of forecasts

Figure 5 – Wind speed and bias for forecasts.

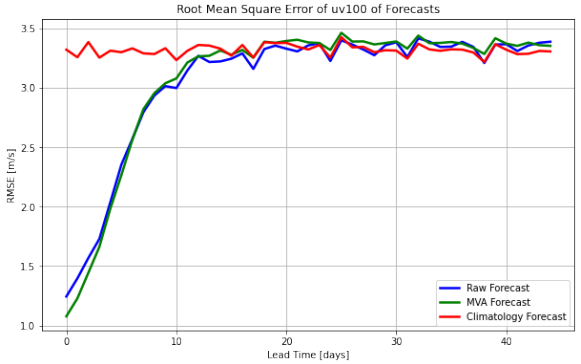
climatology are closely aligned.

When analysing the maps (Figure 17), it reveals that the spatial distribution of bias varies considerably across different regions and forecast methods. In the Raw forecasts (Figure 17a), there is a clear regional pattern where certain areas exhibit a consistent negative bias (blue regions), indicating underestimation of wind speeds. This spatial bias is reduced in the MVA forecasts (Figure 17b), where the bias is more uniformly distributed around zero, confirming that the MVA method effectively corrects for regional biases observed in the Raw forecasts. In contrast, the Climatology forecasts (Figure 17c) tend to show a positive bias (red regions) in specific northern areas, particularly at shorter lead times, suggesting a consistent overestimation in these regions.

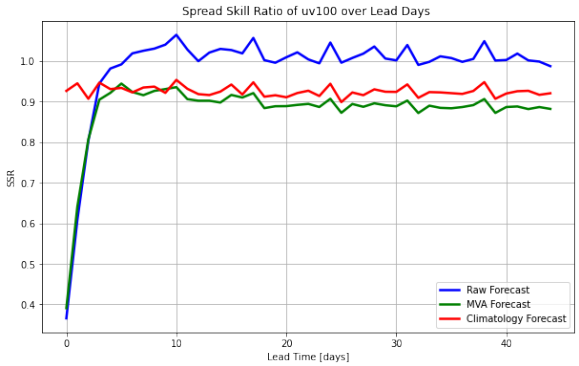
The three graphs show different forecast metrics over lead time: the first graph (6a) displays the spread of forecasts, indicating the range of forecasted values; the second graph (6b) presents the Root Mean Square Error (RMSE) of the ensemble mean, measuring the average difference between the forecasted and observed values; and the third graph (6c) shows the Spread Skill Ratio (SSR), which compares the ratio between the spread and RMSE. They indicate that both raw and MVA corrected forecasts have a lower spread and RMSE at earlier lead days, suggesting higher accuracy and confidence in the short term. As the lead time increases, both spread and RMSE tend to grow, reflecting greater uncertainty and error as the forecast extends further into the future.



(a) Spread of Forecasts



(b) Root Mean Square Error of Forecasts



(c) Spread Skill Ratio of Forecasts

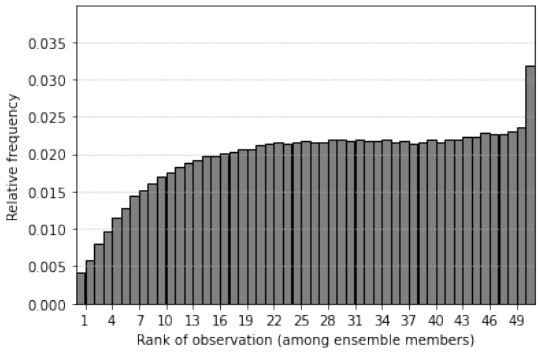
Figure 6 – Metrics of Forecasts: Spread, RMSE and SSR

The lower spread and RMSE at earlier lead days occur because the forecast models are more accurate when they are closer to the initial conditions, where they are more directly based on actual observations and less affected by uncertainties that build up over time. The MVA forecast has a spread and RMSE similar to the Climatology forecast because the MVA correction adjusts the forecasts to match historical climatological data. This adjustment leads the MVA forecast to behave similarly to the Climatology forecast, especially as the forecast period gets longer and the initial data becomes less influential.

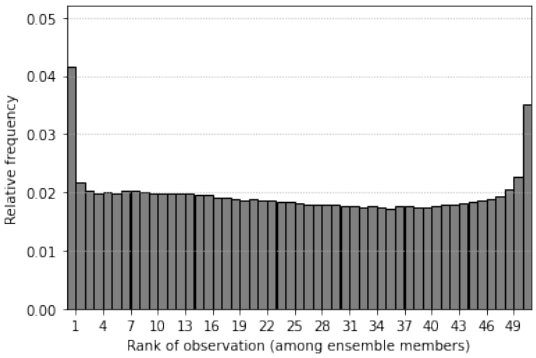
The maps of Spread (Figure 19), RMSE (Figure 18), and SSR (Figure 20a) reveal significant spatial variability, reflecting how forecast performance differs across regions. The Spread and RMSE maps of all forecasts and climatology show that certain areas, as those over the ocean or mountainous regions, tend to have higher uncertainty and error, which increases with lead time.

With the data from the climatology and forecasts, we can plot the rank histogram of forecasts, as seen in figure 7

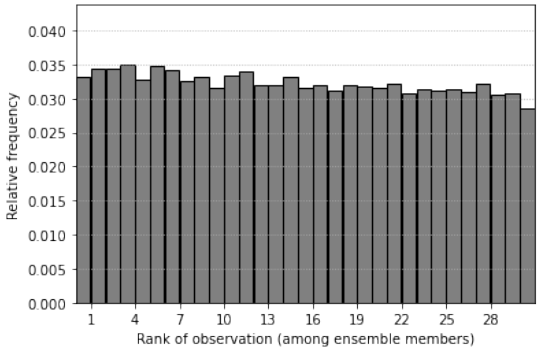
The three Rank Histograms provide insight into the distribution of observed values within the ensemble forecasts and how well these forecasts capture the variability of the data. The first graph (7a), which shows the Rank Histogram for Raw Forecasts, exhibits clearly a negative bias, that is the underestimation of wind speeds. The histogram also shows a distinct dome-shape, indicating overdispersion. Overdispersion means that the



(a) Rank Histogram of Raw Forecasts



(b) Rank Histogram of MVA Forecasts



(c) Rank Histogram of Climatology Forecasts

Figure 7 – Rank Histogram of Forecasts

spread of the forecasts is too large compared to the actual error, causing the forecasts to be too spread out. This overdispersion results from averaging over multiple time instances throughout the day (from the ERA5 data 00h, 06h, 12h, and 18h), which captures a broader range of wind speeds, leading to a larger spread than if forecasts were based on data collected at a single time. This is consistent with the earlier SSR graph (6c), where the spread of the raw forecasts was relatively high, showing that the forecast ensemble members are too dispersed around the observed values.

The second graph (7b, showing the Rank Histogram for MVA Forecasts, has a U-shaped distribution, indicating in that case an underdispersion.The MVA forecasts are still underdispersed even after the bias correction because the MVA method adjusts the forecasts based on the average and variability of both the forecast ensemble and the reference (re-analysis). The correction aims to make the spread of the forecasts match the observed variability by scaling the forecasts accordingly. However, since the reference data has limited variability, the adjusted forecasts can still be too narrowly spread. This underdispersion means that even after the correction, the forecasts don't cover the full range of possible outcomes, resulting in a spread that's too small compared to the actual errors.

The Rank Histogram for Climatology Forecasts (7c) shows a relatively flat dis-

tribution, meaning that the spread of the forecasts. While the shape of the histogram is more uniform and closer to what we would expect in a well-calibrated model, this does not necessarily mean that the climatology is more accurate. The flatter shape simply indicates that the forecasts are evenly distributed around the observed values, but because climatology averages out the variability over a long period, it might still miss specific events or trends, leading to potential inaccuracies.

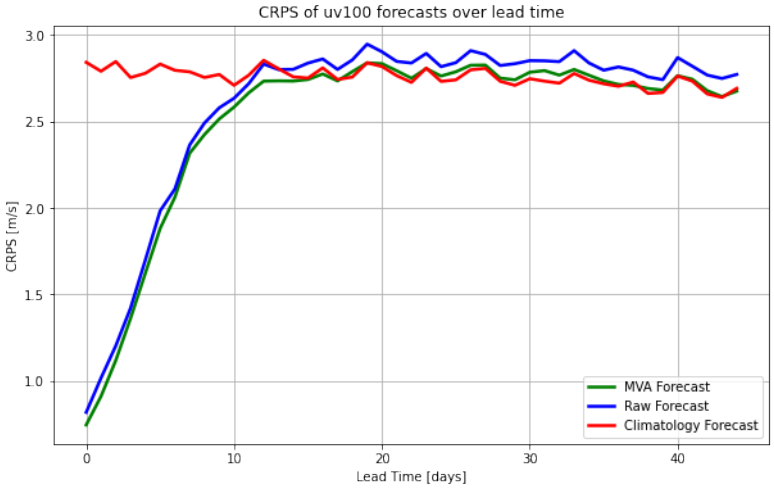
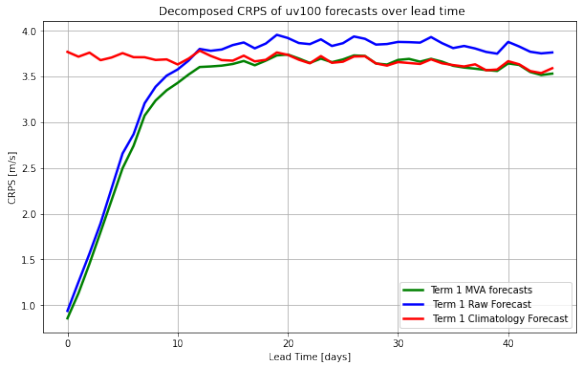
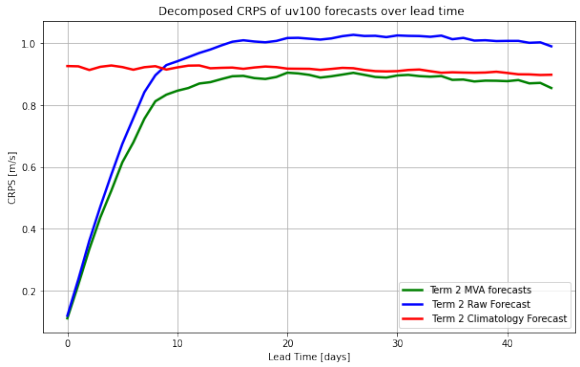


Figure 8 – Continuous Rank Probability Score of Forecasts



(a) First term of CRPS for Forecasts



(b) Second term of CRPS for Forecasts

Figure 9 – Decomposition of CRPS for forecasts

The Continuous Rank Probability Score (CRPS) (8) provides a different perspective on the forecasts compared to the Spread Skill Ratio (SSR)(6c). Since the CRPS reflects how closely the forecast distribution matches the observed data, the MVA forecast shows a better socre than the raw forecast and climatology, which suggests that the MVA correction effectively reduces bias and aligns the forecast distribution more closely with the observations. This improvement is primarily seen in the first term of the CRPS (9a), which measures the difference between the predicted distribution and the observation. By minimizing this difference, the MVA correction improves the CRPS, indicating a more accurate forecast in terms of how well it matches the observed data.

On the other hand, the SSR shows that the climatology performs better than the MVA in terms of spread and RMSE balance, indicating better reliability. The reason for this is that while the MVA correction reduces bias and brings the forecast distribution closer to the observations (which improves the CRPS), it also reduces the ensemble spread without a proportional decrease in RMSE. This reduction in spread leads to underdispersion, where the forecasts do not fully capture the range of possible outcomes, causing the SSR to be lower for the MVA forecast.

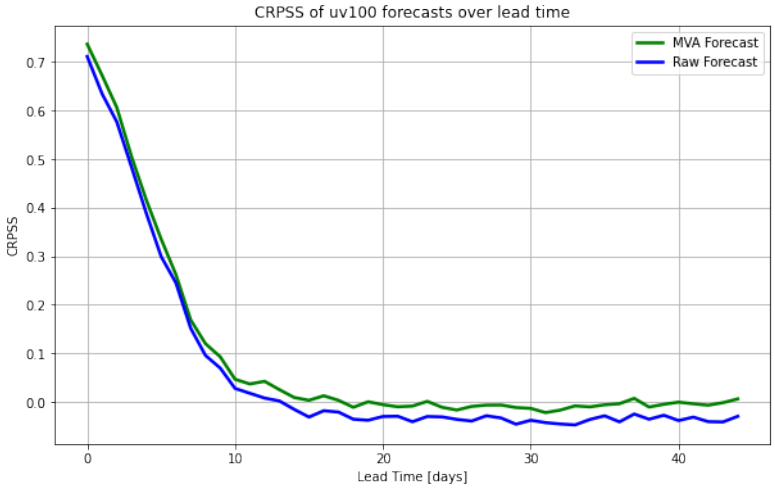
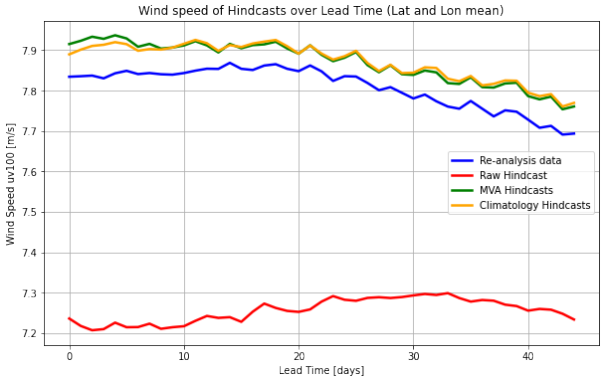


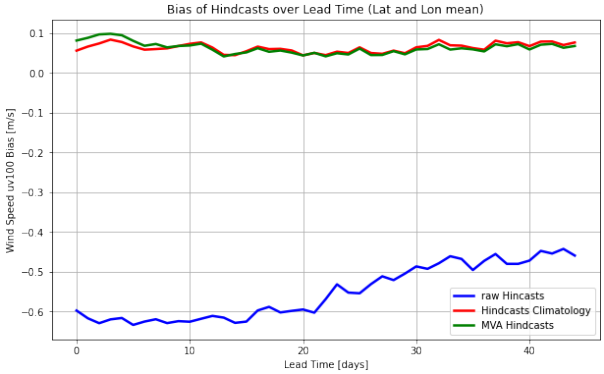
Figure 10 – Continuous Rank Probability Skill Score of Forecasts

When seeing the CRPSS (10), initially, both the MVA and raw forecasts perform better than the climatology, with a CRPSS greater than 0, indicating more skillful forecasts. This advantage persists until around 10 to 15 days into the forecast, where the CRPSS approaches zero, suggesting that the forecasts become as skillful as the climatology from the point that the sub-seasonal to season time scale begins. When comparing the MVA to the raw forecast, the MVA consistently performs slightly better, as indicated by the higher CRPSS throughout the lead time. However, the difference in skill between the MVA and raw forecasts decreases over time, becoming minimal as the forecast horizon extends beyond 20 days.

4.1.2 Hindcasts



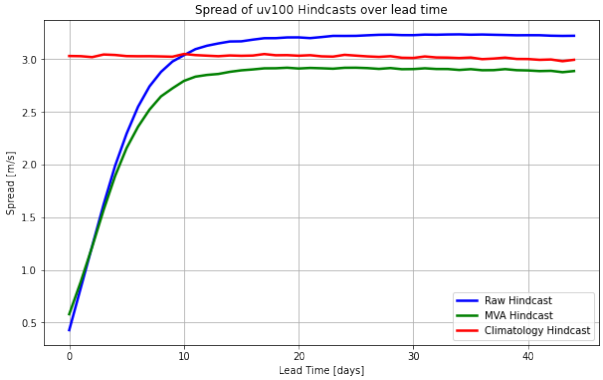
(a) Wind speed of Hindcasts at 100 meters high



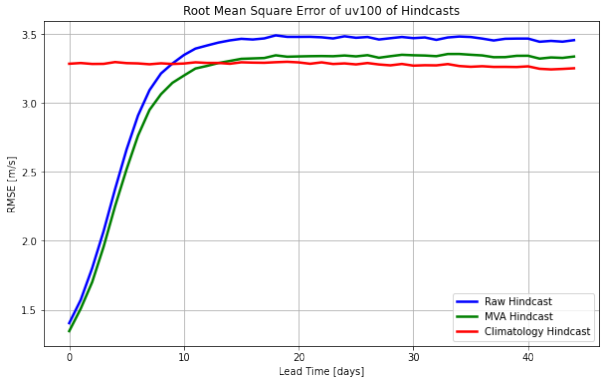
(b) Bias of Hindcasts

Figure 11 – Wind speed and Bias of Hindcasts

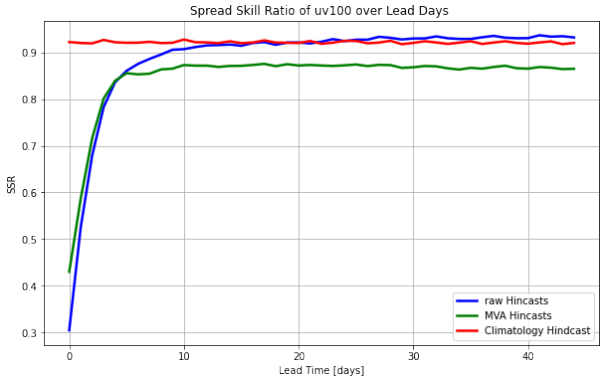
The 11a graph of figure 11shows the wind speed at 100 meters for various hindcasts, with the re-analysis data serving as the true observation. On figure 11b, the raw hindcast data (blue line) consistently underpredicts wind speed compared to the true observation, showing a pronounced negative bias that increases with lead time. The MVA-corrected hindcast (green line) demonstrates a significant reduction in this bias, effectively aligning the hindcast more closely with the true observation through mean and variance adjustments. This indicates that the MVA correction is effective in addressing bias in the raw hindcast.



(a) Spread of Hindcasts



(b) Root Mean Square Error of Hindcasts



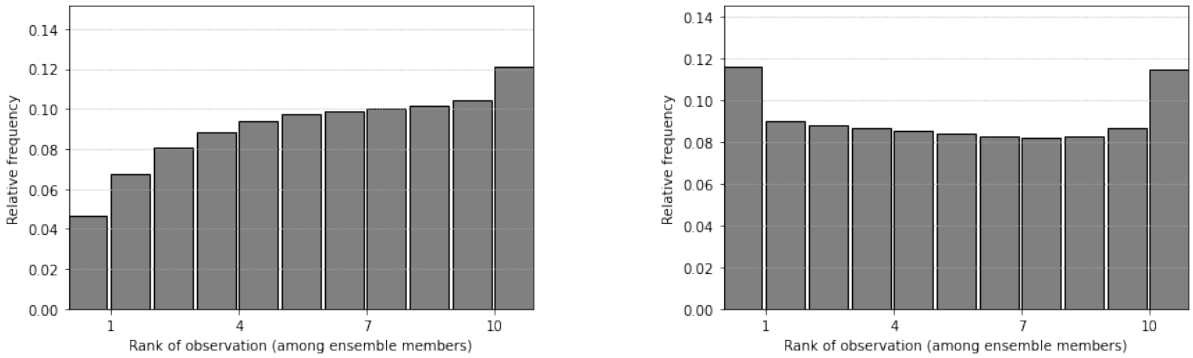
(c) Spread Skill Ratio of Hindcasts

Figure 12 – Metrics of Hindcasts: Spread, RMSE, and SSR

The spread graph (figure 12a) shows that the raw hindcasts (blue line) have the highest spread, followed by the MVA-corrected hindcasts (green line) and the climatology (yellow line), which maintains the lowest spread across all lead times. The RMSE graph (figure 12b) indicates that the MVA-corrected hindcasts have a slightly higher RMSE than the climatology, with the raw hindcasts showing the highest RMSE initially but converging with the others over time. The SSR graph (figure 12c) reveals that the MVA correction does not improve the SSR over the raw hindcasts, bringing it below to the climatology, therefore meaning the hindcast is underdispersed.

The low spread and RMSE for hindcasts in the earlier lead days reflects the

high accuracy of the hindcasts when the model is close to the initial conditions, where uncertainty is minimal. As lead time increases, the spread grows due to the accumulation of errors and uncertainties in the hindcasts and also due to the chaotic nature of the atmosphere. The MVA-corrected hindcast tend to has a spread similar to climatology because the MVA process adjusts the hindcast variance (member by member) to align with the climatological distribution. However, the RMSE of the MVA-corrected hindcast remains slightly higher than that of climatology because, while the correction reduces systematic errors, it cannot entirely eliminate the discrepancies between the hindcasted and observed values, specially as lead time increases.



(a) Rank Histogram of Raw Hindcasts (b) Rank Histogram of MVA Hindcasts

Figure 13 – Rank Histogram of Hindcasts

The Rank Histograms, as on the forecasts, reveal varying levels of dispersion in the hindcasts: the raw hindcasts (Figure 13a) exhibit over-dispersion due to averaging across multiple daily time points, leading to a wider spread in predictions. In contrast, the MVA-corrected hindcasts (Figure 13b), while effectively reducing bias, are under-dispersed, with ensemble members clustering too closely around the mean, indicating insufficient variability captured.

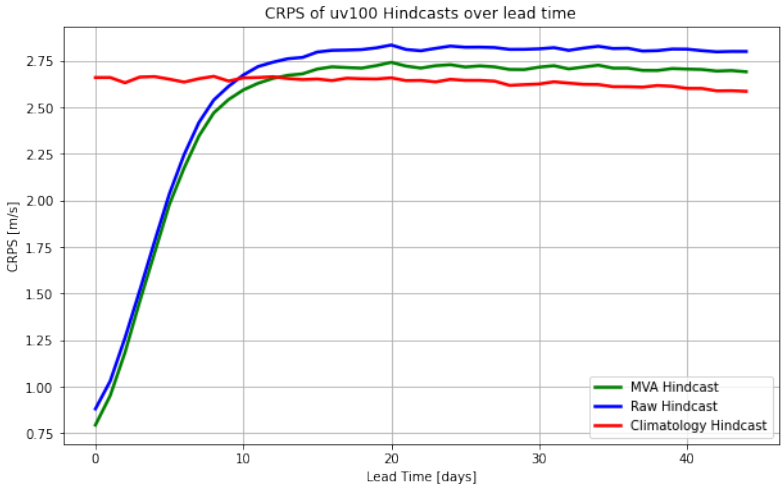
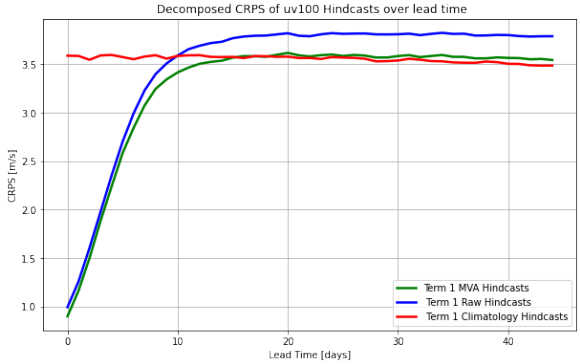
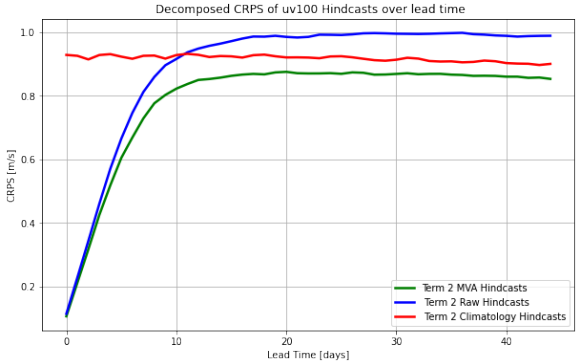


Figure 14 – Continuous Rank Probability Score of Hindcasts



(a) First term of CRPS for Hindcasts



(b) Second term of CRPS for Hindcasts

Figure 15 – Decomposition of CRPS for Hindcasts

Similar to the forecasts, the CRPS of hindcasts and climatology, as shown in Figure 14, highlights the effectiveness of the MVA correction in reducing bias and aligning the forecast distribution more closely with the observed data. The lower CRPS for the MVA-corrected hindcasts compared to the raw hindcasts and climatology indicates that the MVA method effectively minimizes the difference between the predicted distribution and the observations, particularly reflected in the first term of the CRPS (Figure 15a).

However, the SSR graph previously shown (Figure 12c) indicates that the climatology performs better than the MVA-corrected hindcasts in terms of reliability. This discrepancy, as was on the forecasts, arises because the MVA correction, while reducing bias and improving the CRPS, also narrows the ensemble spread without a corresponding decrease in RMSE. This leads to an under-dispersed ensemble, which negatively impacts the SSR by making the ensemble less representative of the true variability in the observations. That way, while the CRPS benefits from the reduced bias (term 1), the SSR suffers due to the inadequate spread in the ensemble.

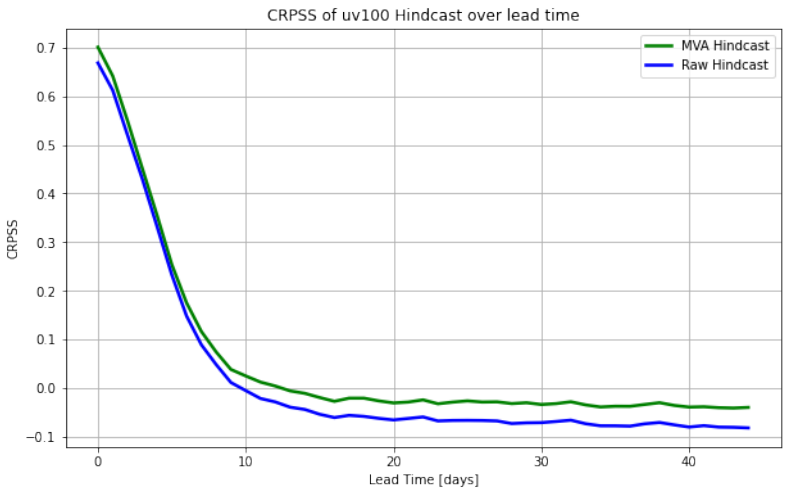


Figure 16 – Continuous Rank Probability Skill Score of Hindcasts

The CRPSS graph (Figure 16) shows that the raw hindcasts barely outperform

climatology, with a CRPSS above zero only until around day 8. The MVA hindcasts perform better than the raw hindcasts, maintaining a positive CRPSS until approximately day 11, indicating some skill over climatology. After these points, both the raw and MVA hindcasts no longer provide a significant advantage over climatology, as their CRPSS values approach or fall below zero. The CRPS maps (Figure 26a) highlight that in certain regions, particularly those with complex weather patterns, the hindcasts—especially in later lead times—may perform worse, as indicated by higher CRPS values.

5 CONCLUSION

The similarity between the MVA-corrected bias and the climatology bias affirms that the MVA correction successfully standardizes the hindcast data to match the long-term climatological trends, since it is based on the same re-analysis data used as the reference for correction. This outcome confirms that the correction method accurately adjusts the hindcasts to align with the re-analysis over many years. However, the residual bias, though minimal, indicates that while the MVA correction is effective, it does not fully eliminate forecast errors, particularly as lead times increases

One potential factor could be the limited amount of data used to compute the mean and standard deviation in the MVA. In this case, only twenty data points corresponding to twenty years were used, which might not be sufficient to capture the full variability and trends, potentially impacting the precision of the MVA correction.

In terms of reliability, the performance of the MVA calibration method does not surpass that of the Climatology approach on a sub-seasonal to seasonal level. This outcome may be attributed to the inherently univariate nature of the MVA method, which might limit its ability to fully capture the complex, multivariate dependencies (for example the spatial correlation between each grid point) in forecasts. While the possibility of data insufficiency was considered, experiments with varying amounts of data did not yield significantly different results, suggesting that the issue likely stems from the methodological constraints of the MVA rather than from data limitations.

When comparing to the raw forecasts, the MVA method successfully improves the accuracy of the forecasts (seen on the CRPS verification scores - Figure 14) on a sub-seasonal to seasonal level. But, when it comes to reliability, the MVA correction does not outperform the raw forecasts. Although the MVA method adjusts the variance across different start days, it unintentionally alters the variance across the ensemble dimension diminishing the spread. This issue is particularly pronounced in cases where the variable in question, such as wind speed in our study, does not follow a Gaussian distribution (TORRALBA et al., 2017). As a result, the MVA correction, which assumes a normal distribution, fails to accurately represent the forecast uncertainty, thereby undermining its effectiveness.

As for possible future work, one alternative would be to consider calibration methods that are not constrained by the assumption of a Gaussian distribution. Approaches such as Bayesian Model Averaging (BMA) and Non-homogeneous Regression (EMOS), as detailed in (WILKS, 2021), offer possible alternatives. These methods are designed to handle non-Gaussian distributions and could be more suitable for variables like wind speed, which do not conform to normality.

REFERENCES

BUIZZA, R. Ensemble forecasting and the need for calibration. In: VANNITSEM, D. S. W. S.; MESSNER, J. W. (Ed.). **Statistical Postprocessing of Ensemble Forecasts**. Reading, United Kingdom: Elsevier, 2021. p. 15–35.

FORTIN, V. et al. Why should ensemble spread match the rmse of the ensemble mean? **Journal of Hydrometeorology**, v. 15, n. 4, p. 1708–1713, August 2014.

GOUTHAM, N. et al. How skillful are the european subseasonal predictions of wind speed and surface temperature? **Journal Name**, 2022. Manuscript received 4 August 2021, in final form 3 March 2022.

HAMILL, T. M. Interpretation of rank histograms for verifying ensemble forecasts. **Monthly Weather Review**, v. 129, n. 3, p. 550–560, 2001.

HOLTON, J. R. **An Introduction to Dynamic Meteorology**. 5th. ed. Oxford, UK: Academic Press, 2012. v. 88. 552 p. ISBN 9780123848666.

HUDSON, D. et al. The subseasonal to seasonal (s2s) prediction project database. **Bulletin of the American Meteorological Society**, v. 98, n. 1, p. 163–173, 2017. ISSN 1520-0477.

MANZANAS, R. et al. Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the c3s dataset. **Climate Dynamics**, Springer-Verlag GmbH Germany, part of Springer Nature, v. 52, n. 1-2, May 2019.

MARIOTTI, A.; RUTI, P. M.; RIXEN, M. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. **npj Climate and Atmospheric Science**, v. 1, n. 4, p. 1–6, 2018.

RAOULT, B. et al. Climate service develops user-friendly data store. **ECMWF Newsletter**, ECMWF, Reading, United Kingdom, n. 151, 2017. Disponível em: <https://www.ecmwf.int/en/newsletter/151/meteorology/climate-service-develops-user-friendly-data-store>.

RASP, S. et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. **Submitted to a Journal**, p. 10–13, 2024. Google Research, Google DeepMind, European Centre for Medium-Range Weather Forecasts.

THORARINSDOTTIR, T. L.; SCHUHEN, N. Verification: Assessment of calibration and accuracy. In: VANNITSEM, D. S. W. S.; MESSNER, J. W. (Ed.). **Statistical Postprocessing of Ensemble Forecasts**. Oslo, Norway: Elsevier, 2021. p. 155–164.

TORRALBA, V. et al. Seasonal climate prediction: A new source of information for the management of wind energy resources. **J. Appl. Meteor. Climatol.**, v. 56, p. 1231–1247, 2017. Disponível em: <https://doi.org/10.1175/JAMC-D-16-0204.1>.

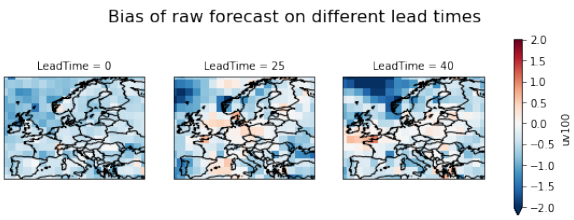
VITART, F.; COAUTHORS. **Extended-range prediction**. [S.l.], 2019. 60 p. Disponível em: <https://doi.org/10.21957/pdivp3t9m>.

WILKS, D. S. Univariate ensemble postprocessing. In: VANNITSEM, D. S. W. S.; MESSNER, J. W. (Ed.). **Statistical Postprocessing of Ensemble Forecasts**. Ithaca, NY, United States: Elsevier, 2021. p. 75–81.

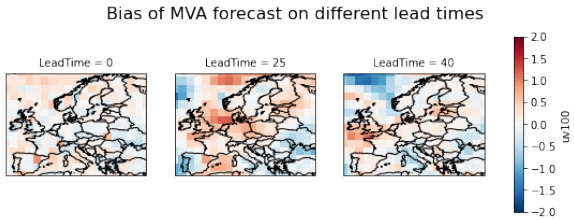
APPENDICES

APÊNDICE A – Maps

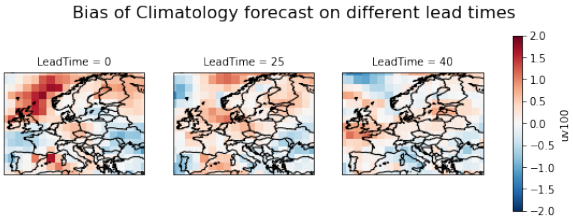
A.1 Forecasts



(a) Bias of Raw Forecasts

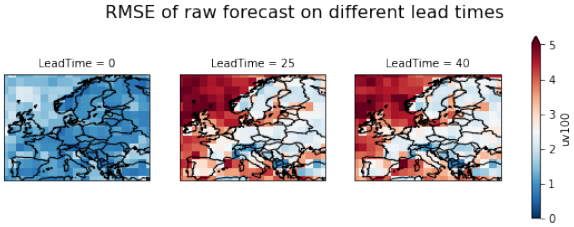


(b) Bias of MVA Forecasts

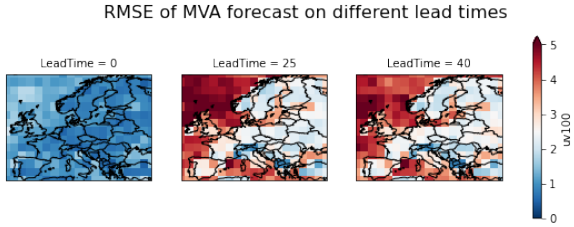


(c) Bias of Climatology Forecasts

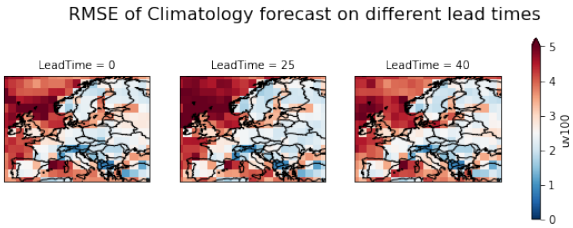
Figure 17 – Bias maps of Forecasts



(a) RMSE of Raw Forecasts



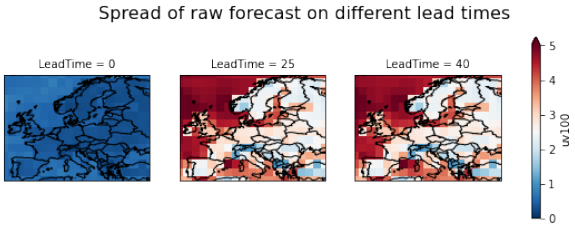
(b) RMSE of MVA Forecasts



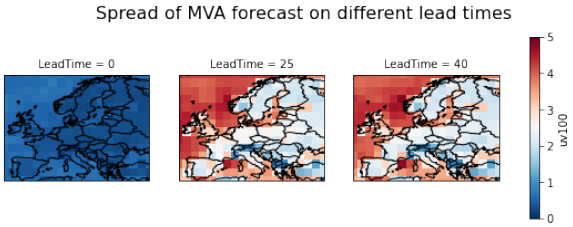
(c) RMSE of Climatology Forecasts

Figure 18 – RMSE maps of Forecasts

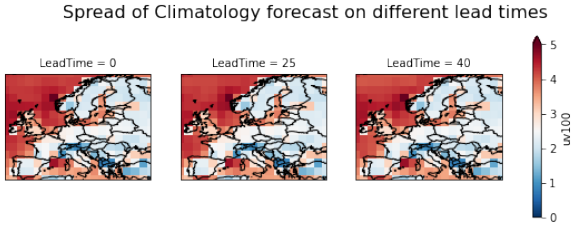
A.2 Hindcasts



(a) Spread of Raw Forecasts

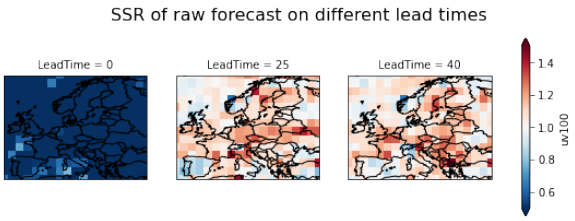


(b) Spread of MVA Forecasts

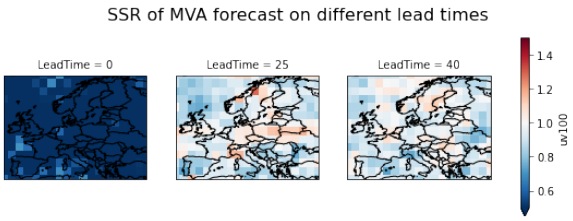


(c) Spread of Climatology Forecasts

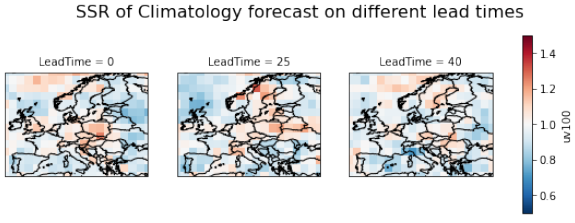
Figure 19 – Spread maps of Forecasts



(a) SSR of Raw Forecasts

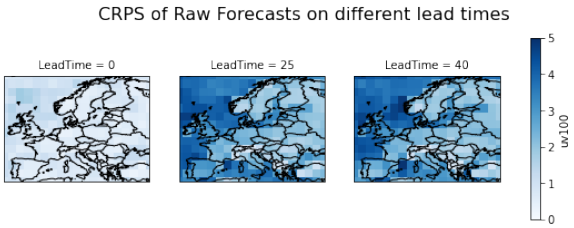


(b) SSR of MVA Forecasts

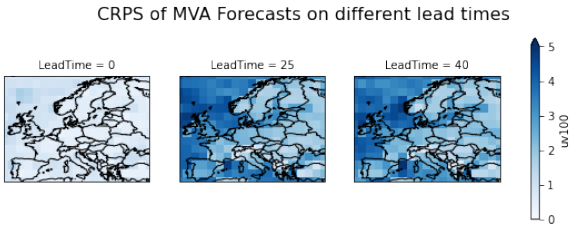


(c) SSR of Climatology Forecasts

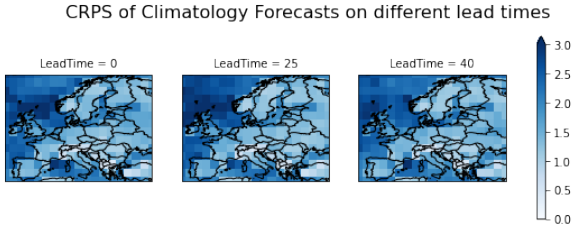
Figure 20 – SSR maps of Forecasts



(a) CRPS of Raw Forecasts

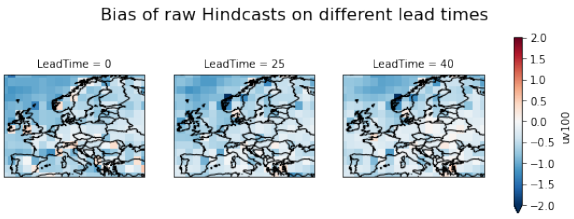


(b) CRPS of MVA Forecasts

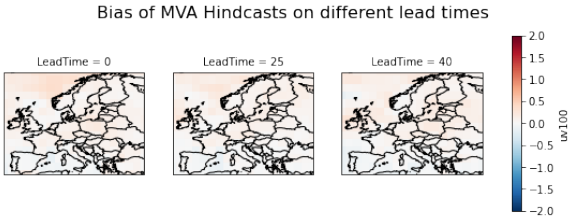


(c) CRPS of Climatology Forecasts

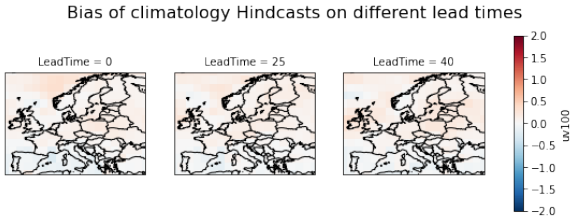
Figure 21 – CRPS maps of Forecasts



(a) Bias of Raw Hindcasts

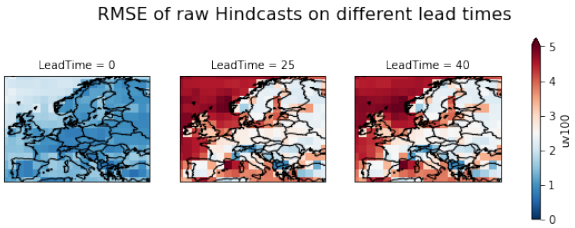


(b) Bias of MVA Hindcasts

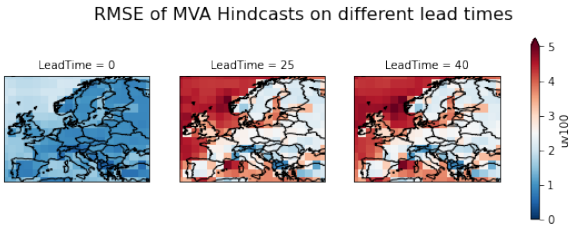


(c) Bias of Climatology Hindcasts

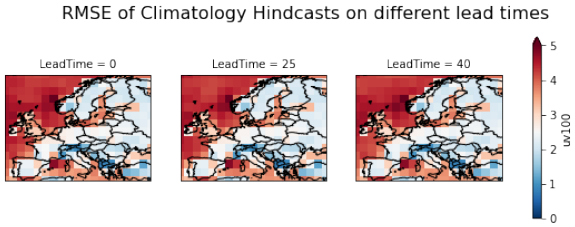
Figure 22 – Bias maps of Hindcasts



(a) RMSE of Raw Hindcasts

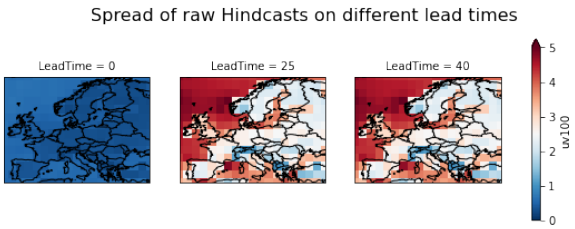


(b) RMSE of MVA Hindcasts

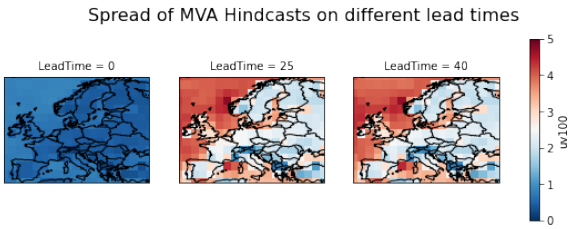


(c) RMSE of Climatology Hindcasts

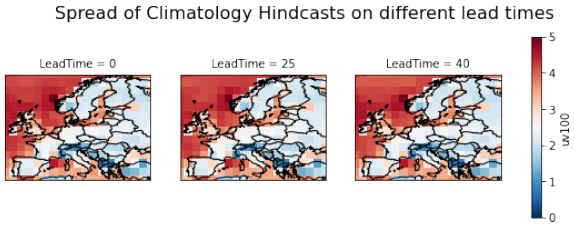
Figure 23 – RMSE maps of Hindcasts



(a) Spread of Raw Hindcasts

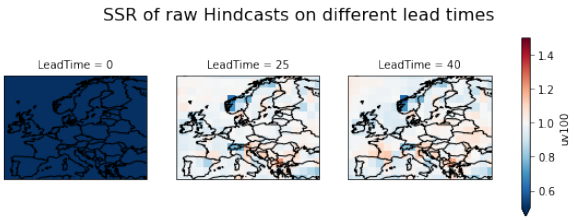


(b) Spread of MVA Hindcasts

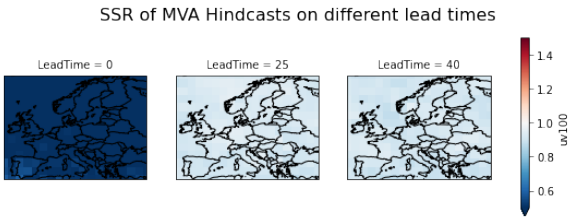


(c) Spread of Climatology Hindcasts

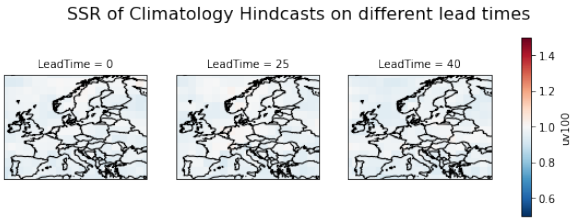
Figure 24 – Spread maps of Hindcasts



(a) SSR of Raw Hindcasts

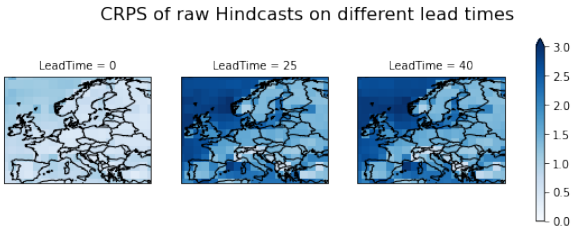


(b) SSR of MVA Hindcasts

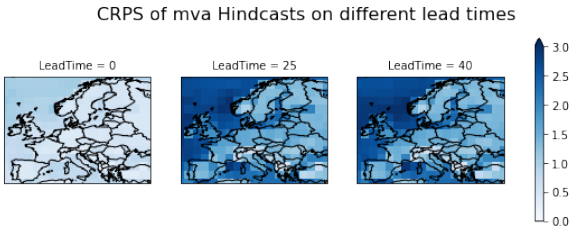


(c) SSR of Climatology Hindcasts

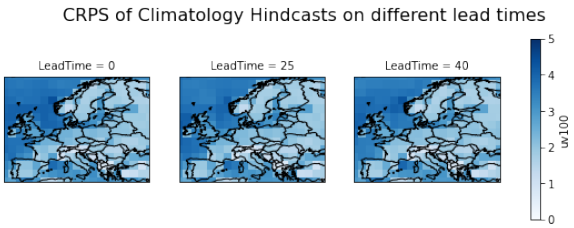
Figure 25 – SSR maps of Hindcasts



(a) CRPS of Raw Hindcasts



(b) CRPS of MVA Hindcasts



(c) CRPS of Climatology Hindcasts

Figure 26 – CRPS maps of Hindcasts